POSTER LIST
ORDERED ALPHABETICALLY BY POSTER TITLE
GROUPED BY THEME/TRACK

THEME/TRACK: DATA

Poster numbers: P_Da001 - 130 Application posters: *P_Da001 - 041*

Poster number	EasyChair number	Author list	Presenting author	Title	Abstract	Theme/track	Topics
APPLICATION POSTERS WITHIN DATA THEME							
P_Da001	*773*	*Benoît Carrères, Anne Klok, Maria Suarez Diez, Lenny de Jaeger, Mark Sturme, Packo Lamers, Rene' Wijffels, Vitor Dos Santos, Peter Schaap and Dirk Martens*	*Benoît Carrères*	*A systems approach to explore triacylglycerol production in Neochloris oleoabundans*	*Microalgae are promising platforms for sustainable biofuel production. They produce triacyl-glycerides (TAG) which are easily converted into biofuel. When exposed to nitrogen limitation, Neochloris oleoabundans accumulates up to 40% of its dry weight in TAG. However, a feasible production requires a decrease of production costs, which can be partially reached by increasing TAG yield.We built a constraint-based model describing primary metabolism of N. oleoabundans. It was grown in combinations of light absorption and nitrate supply rates and the parameters needed for modeling of metabolism were measured. Fluxes were then calculated by flux balance analysis. cDNA samples of 16 experimental conditions were sequenced, assembled and functionally annotated. Relative expression changes and relative flux changes for all reactions in the model were compared.The model predicts a maximum TAG yield on light of 1.07g (mol photons)-1, more than 3 times current yield under optimal conditions. Furthermore, from optimization scenarios we concluded that increasing light efficiency has much higher potential to increase TAG yield than blocking entire pathways.Certain reaction expression patterns suggested an interdependence of the response to nitrogen and light supply. Some other reactions showed unexpected regulatory patterns thereby providing prime choice targets for further studies.We concluded that nitrogen limitation directly affects gene expression of nitrogen dependent reactions, while high light generates more energy thereby indirectly propagating into lower metabolism. We also suggest that Phytyldiphosphate synthase acts as a central regulator for pigment synthesis and breakdown.*	*Data/ Application poster*	*Application*
P_Da003	*723*	*Fotis Psomopoulos, Eija Korpelainen, Kimmo Mattila and Diego Scardaci*	*Eija Korpelainen*	*Bioinformatics resources on EGI Federated Cloud*	*Data can be "big" for three reasons – often referred to as the three V's; volume of data, velocity of processing the data, and variability of data sources. If any of these key features are present, then big-data tools are necessary, often combined with high network bandwidth and massive compute systems. As NGS technologies are revolutionizing life science research, established workflows in facilitating the first steps in data analysis are being increasingly employed. Cloud computing provides a robust and cost-efficient solution towards supporting the computational demands of such workflows. In particular, NGS data analysis tools are constantly becoming available as resources within EGI's Federated Cloud. The European Grid Infrastructure (EGI) is the result of pioneering work that has, over the last decade, built a collaborative production infrastructure of uniform services through the federation of national resource providers that supports multi-disciplinary science across Europe and around the world. EGI currently supports an extensive list of services available for life sciences and has been working together with the community to implement further support. The EGI Federated Cloud (FedCloud), the latest infrastructure and technological offering of EGI, is a prime example of a flexible environment to support both discipline and use case through Big Data services. Finally, in addition to providing access to advanced tools and applications, e-infrastructures like EGI, provide the opportunity to create training tools for life science researchers and to create synergies between life sciences and ICT researchers, which is fundamental in moving research forward.*	*Data/ Application poster*	*Application*
P_Da004	*779*	*Mascha Jansen, Rob Hooft, Barend Mons, Celia van Gelder, Luiz Olavo Bonino Da Silva Santos and Marco Roos*	*Mascha Jansen*	*Bring Your Own Data (BYOD) workshops to make life science data linkable at the source*	*Functionally interlinking datasets is essential for knowledge discovery. The 'Bring Your Own Data' workshop (BYOD) has proven an excellent tool for the adoption of techniques to achieve this. It provides a mechanism for data owners who would like to add value to their data by preparing them for data integration and computational analysis, but are unfamiliar with basic techniques to make data Findable, Accessible, Interoperable, and Reusable for humans and computers (FAIR). Using linked data and associated technologies, data owners, domain experts, and linked data experts collaborate to make owner's data linkable and explore possibilities to answer questions across multiple data sources. Momentarily, BYODs play a critical role in establishing a robust and sustainable infrastructure of linkable data sources where the responsibility for FAIR data stewardship starts at the source. We present the organisational roadmap of the three day workshop and the latest insights into making BYODs more productive, including standard objectives to produce FAIR data, refine guidelines, and discover knowledge. Previous BYODs, such as with the Human Protein Atlas, plant breeding data, and data from rare disease registries and biobanks, have shaped the roadmap. Although every BYOD is uniquely tailored, they contain at least a preparatory phase with at least two webinars for data owners and domain experts, an execution phase for the BYOD itself, and a follow-up phase to foster the results of the BYOD by telephone conferences with participants. A BYOD is also a learning experience that helps domain experts to endorse the approach in their domain.*	*Data/ Application poster*	*Application Fundamental*
P_Da006	*417*	*Ken Tominaga, Daisuke Komura and Shumpei Ishikawa*	*Ken Tominaga*	*Classification of digital pathological images using Virtual Adversarial Training with an effective GUI annotation system*	*Automatic cancer detection from digital pathological images has been an important issue in the medical field. Supervised learning has been shown to be effective in the task if we have a large number of labeled training examples (i.e. cancer/non-cancer images). However, the acquisition of labeled data often requires a skilled human agent such as a pathologist and the manual labeling process is costly and time-consuming. To overcome this problem, we have developed a new cancer detection system, which reduces labeling cost and needs only a small amount of labeled data. Key aspects of our system are twofold: 1) Virtual Adversarial Training (VAT), a state-of-the-art training method for semi-supervised learning, was applied to the classification of digital pathological images. VAT needs only a small amount of labeled instances, but performs better than supervised learning algorithms by making use of unlabeled instances, which are easy to obtain. 2) GUI annotation system, which we call Pathology Map, was implemented to help users to easily generate labeled data. Pathology Map uses Google Maps API to display Aperio SVS TIFF files converted into the Google Maps format using VIPS and OpenSlide libraries. We apply our system to Whole Slide Imaging from The Cancer Genome Atlas (TCGA) to demonstrate the effectiveness of our system.*	*Data/ Application poster*	*Application*
P_Da007	*506*	*Raik Otto, Christine Sers and Ulf Leser*	*Raik Otto*	*Comparing characteristic genomic variants allows reliable in-silico identification of Next-Generation sequenced Cancer Cell Line samples*	*Cancer cell lines are a pivotal tool for cancer researchers. However, cancer cell lines are prone to critical errors such as misidentification and cross-contamination which have reportedly caused severe setbacks. Established cancer cell line identification methods compare genotype characteristics obtained during specific experiments (e.g. SNP arrays); characteristic genotype properties of the to-be-identified sample (the query) are matched against the same characteristics properties of the known samples (the references). If a match shows a significant similarity to a reference sample, the query is identified as the reference sample. Such characteristic genotype information can also be derived from NGS data. A query can be identified when the characteristic genotype properties were obtained from Next-generation sequencing of the query and a subsequent comparison to a NGS reference. However, results from different NGS technologies, algorithms and sequencing-approaches, e.g. whole-exome or panel-sequencing, are inherently challenging to compare. SNP-zygosity matching and tandem repeat-counting on such data is in general unreliable due to non-covered loci, SNP-filtering, and zygosity-call divergence caused by differing algorithmic ploidy-settings. Here, we present the Uniquorn method that reliably identifies cancer cell line samples based on NGS genotyping data across different technologies, algorithms, filter-settings and covered loci. Uniquorn compares the query to all references and computes a p-value for the likelihood that an overlap in observed genomic variants is due to chance. Uniquorn was benchmark by cross-identifying 1989 cancer cell line sequencing samples: sensitivity amounted to 96% and specificity to 99%. The R-BioConductor package Uniquorn and the benchmark setup are freely available.*	*Data/ Application poster*	*Application Biotechnology*
P_Da008	*735*	*Arnaud Meng, Lucie Bittner, Stéphane Le Crom, Fabrice Not and Erwan Corre*	*Arnaud Meng,*	*De novo transcriptome assembly dedicated pipeline and its specific application to non-model, marine planktonic organisms*	*De novo assembly corresponds to the reconstruction of a genome or a transcriptome based on sequenced DNA/RNA without any genomic reference. Since the last decade, this powerful approach allows scientists to extend genomic exploration studies to non model organisms, which represent the majority of current living beings/lineages [1]. Bioinformatics constitute therefore a vital step to investigate the genomic dark-matter.Here we introduce our pipeline dedicated to de novo transcriptome assembly and downstream analysis including quality evaluation and in silico biological validation of the transcripts. Our approach is divided in 5 distinct parts: (i) reads quality filtering and cleaning, (ii) de novo assembly with Trinity [2], (iii) quality evaluation via metrics, (iv) likely coding domains prediction and their functional annotation, (v) and in silico validation via sequence similarity networks.As a proof of concept, we processed 54 RNA-seq datasets of Dinoflagellates produced from a consortium, large-scaled sequencing project [3]. As our pipeline relies on multi-threaded components 54 reliable transcriptomes from non-model organismes along with its respectives functionally annotated predicted proteomes were produced in a moderate amount of time. Moreover proteomes were further explored with sequence similarity networks, allowing the analysis of the entire datasets (including unknown sequences) thanks to metadata crossing (e.g. taxonomy, annotation).[1] Hug et al. Nature Microbiology 16048 (2016).[2] Grabherr et al. Nat Biotech 29, 644–652 (2011).[3] Keeling et al. PLOS Biol 12, e1001889 (2014).*	*Data/ Application poster*	*Application*
P_Da009	*364*	*Felipe Albrecht, Markus List, Christoph Bock and Thomas Lengauer*	*Felipe Albrecht*	*DeepBlue: Diving into Epigenomic Data*	*Large volumes of data are generated by several epigenomic consortia, including ENCODE.Roadmap Epigenomics, BLUEPRINT, and DEEP. To enable users to utilize these data effectively inthe study of epigenetic regulation, we have developed the DeepBlue Epigenomic Data Server.With the DeepBlue Epigenomic Data Server, we provide programmatic access to vast amounts ofepigenomic data, in order to facilitate storing, organizing, searching, and retrieving of epigeneticdata.Here, we present a series of tools that build upon the DeepBlue API and enable users not proficientin scripting or programming languages to benefit from our efforts and to analyze epigenomic data ina user-friendly way: (i) an R/Bioconductor package (http://deepblue.mpi-inf.mpg.de/R) integratesDeepBlue into the R analysis workflow. The extracted data are automatically converted toGenomicRanges, which are supported by many related packages for analysis and visualization;(ii) a web interface (http://deepblue.mpi-inf.mpg.de) that enables users to search, select, anddownload the epigenomic data available in DeepBlue; (iii) a web tool for epigenomic datavisualization, named DeepBlue Dive (http://dive.mpi-inf.mpg.de), which is inspired by EpiExplorer and helps researchers to visually compare their own epigenomic data to data already available inDeepBlue; (iv) a web tool, named DeepBlue ML, complementary to DeepBlue Dive, which isinspired by EpiGRAPH and uses LOLA, reporting the enrichment of epigenomic regionsprovided by the user among the experiments available in DeepBlue.DeepBlue and related tools are available at http://deepblue.mpi-inf.mpg.de/.*	*Data/ Application poster*	*Application Fundamental*
P_Da010	*579*	*Dong-Gi Lee and Hyunjung Shin*	*Dong-Gi Lee*	*Disease Causality Extraction from PubMed Literatures*	*Motivation: Recently, the research about human disease network has been successful and become an aid of figuring out relationship between various diseases. In most of the disease network, however, the relationship between diseases has been represented just as association. This incurs a difficulty of finding prior diseases and their influences on posterior diseases. In this paper, we propose a causal disease network that implement disease causality through text mining on biomedical literature.Methods: In order to provide causality between diseases, the proposed method includes two schemes: the first one is lexicon-based causality term strength, which endows causal strength on variety of causality terms based on lexicon analysis. The second one is frequency-based causality strength, which determines the direction and strength of causality based on document and clause frequencies in the literatures.Results: We applied the proposed method to 6,617,833 PubMed literatures, and chose 195 diseases to construct a causal disease network. From all possible pairs of disease nodes in the network, 1,011 causal pairs of 149 diseases were extracted. The resulting network was compared with that of previous study: In both coverage and quality aspects, the proposed method showed outperforming results: it found 2.7 times more causalities and showed higher correlation with associated diseases than the competing method.*	*Data/ Application poster*	*Application*
P_Da012	*363*	*Luca Beltrame, Tony Travis, Luca Clivio, Sergio Marchini and Maurizio D'Incalci*	*Luca Beltrame*	*Distributed file systems for storage and analysis of Next-Generation Sequencing data*	*Analysis of NGS (Next Generation Sequencing) data is a computationally demanding task requiring large amounts of CPU, memory, and disk space. There is also a requirement for high performance data storage systems, resilient to hardware failure, to be connected directly to the computing infrastructure (typically a multi-node cluster) to store large quantities of NGS data reliably.Traditional shared file systems such as NFS (Network File System) do not offer the performance, scalability or cache coherence required by modern NGS data analysis, so alternatives including GlusterFS, Ceph, and Lustre have been developed. However, there is a trade-off between data safety on replicated local storage and degradation of performance across distributed storage. Resilience to hardware failure is typically provided by RAID (Redundant Array of Independent Disks) and redundant storage nodes.Here we describe the evaluation of an alternative file system, RozoFS (https://github.com/rozofs/rozofs) for use with demanding NGS data analysis workloads. We used a synthetic data set (DREAM-TCGA data set 3) to run a complete tumor-normal analysis pipeline ("bcbio", https://github.com/chapmanb/bcbio-nextgen), including base quality recalibration, local indel realignment, somatic variant calling, and structural variants as a benchmark to compare RozoFS with a traditional shared file system (NFS) on two different HPC (High Performance Computing; Cloud4CaRE project) clusters.Our results show high reliability and good performance of RozoFS compared to NFS, in particular, during heavy I/O workloads. These findings indicate that the reliability and robustness of RozoFS's make it a good candidate for demanding NGS analysis workloads.*	*Data/ Application poster*	*Application Fundamental*
P_Da013	*809*	*Tilo Buschmann and Leonid Bystrykh*	*Tilo Buschmann*	*DNA Barcodes Adapted to the Illumina Sequencing Platform*	*The successful completion of multiplexed high-throughput sequencing experiments depends heavily on the proper design of the DNA barcodes. Mutations during barcode synthesis, PCR amplification, and sequencing make decoding of DNA barcodes and their assignment to the correct samples difficult. Previously, we introduced a generalised barcode design for the correction of insertions, deletions, and substitutions which we called the Sequence-Levenshtein distance.However, generalised barcode designs may be wasteful when applied to specific technologies. The Illumina "Sequencing by Synthesis" platform (e.g., Illumina HiSeq/MiSeq) shows a very large number of substitution errors as well as a very specific shift of the read that results in inserted and deleted bases at the 5'-end and the 3'-end (which we call phaseshifts). As a solution, we propose the Phaseshift distance that exclusively supports the correction of substitutions and phaseshifts. Additionally, we enable the correction of arbitrary combinations of substitution and phaseshift errors. Thus, we address the lopsided number of substitutions compared to phaseshifts on the Illumina platform.To compare codes based on the Phaseshift distance to Hamming Codes (correction of substitution errors) as well as codes based on the Sequence-Levenshtein distance (correction of indels and substitution errors), we simulated experimental scenarios based on the error pattern we identified on the Illumina platform. Furthermore, we generated a large number of different sets of DNA barcodes using the Phaseshift distance and compared codes of different lengths and error correction capabilities. We found that codes based on the Phaseshift distance can correct a number of errors comparable to codes based on the Sequence-Levenshtein distance while offering the number of DNA barcodes comparable to Hamming codes. Thus, codes based on the Phaseshift distance show a higher efficiency in the targeted scenario.*	*Data/ Application poster*	*Application*
P_Da014	*584*	*Gokhan Ertaylan, Nadia J. T. Roumans, Roel G. Vink, Marleen Van Baak, Edwin Mariman,Ilja Arts, Theo de Kok and Michael Lenz*	*Gokhan Ertaylan*	*Estimating real cell size distribution from cross section microscopy imaging*	*Microscopy imaging is an essential tool for medical diagnosis and molecular biology. It is particularly useful for extracting information about disease states, tissue heterogeneity and cell-specific parameters such as cell type or cell size from biological specimens. However, the information obtained from the images is likely to be subjected to sampling and observational bias with respect to the underlying cell size/type distributions.Results: We present an algorithm, Estimate Tissue Cell Size/Type Distribution (EstiTiCS), for the adjustment of the underestimation of the number of small cells and the size of measured cells while accounting for the section thickness independent of the tissue type. We introduce the sources of bias under different tissue distributions and their effect on the measured values with simulation experiments. Furthermore, we demonstrate our method on histological sections of paraffin-embedded adipose tissue sample images from 57 people from a dietary intervention study. This data consists of measured cell size and its distribution over the dietary intervention period at 4 time points. Adjusting for the bias with EstiTiCS results in a closer fit to the true/expected adipocyte size distribution with earlier studies. Therefore, we conclude that our method is suitable as the final step in estimating the tissue-wide cell type/size distribution from microscopy imaging pipeline.Availability and implementation: Source code and its documentation are available online. The whole pipeline of our method is implemented in R and makes use of the "nloptr" package. Adipose tissue data used for this study are available on request.*	*Data/ Application poster*	*Application Fundamental Health*

P_Da015	828	Johannes Köster	Johannes Köster	Fully reproducible data analyses with Snakemake and Bioconda	Reproducible and scalable data analyses are crucial to obtain reliable insights from today's high throughput technologies. With the popular workflow management system Snakemake we have previously provided a powerful framework to formalize and execute data analyses on workstations, compute servers and clusters without the need to modify the workflow definition.In Bioinformatics, analyses typically rely on the application of diverse tools and libraries coming from various, sometimes conflicting software ecosystems, and requiring diverse ways of installation. We extend the notion of reproducibility to the definition and automated deployment of software dependencies, and present Bioconda, a distribution of Bioinformatics software for the Conda package manager. Bioconda normalizes and unifies the diverse ways of installing Bioinformatics software and allows the easy deployment and automatic dependency resolution without admin rights. It is growing rapidly and provides over 1400 packages today, from standalone compiled programs like BWA or Cufflinks to R, Python and Perl packages.In addition, we present the integration of the Conda package manager into the Snakemake workflow management system. This turns Snakemake into the first text-based workflow management system that allows to define a workflow together with its software dependencies. In consequence, the installation of the required software in strictly defined versions becomes a part of the automated workflow execution itself.In combination with Bioconda, Snakemake provides self-contained documentation, fully automated deployment and scalable as well as reproducible execution of Bioinformatics analyses.	Data/ Application poster	Application
P_Da016	712	Maryam Soleimani Dodaran, Pernette Verschure, Perry Moerland and Antoine van Kampen	Maryam Soleimani Dodaran	Identification of candidate methylation sites predictive for resistance to tamoxifen treatment using survival analysis of the TCGA breast cancer cohort	Endocrine therapy is a common treatment in women with ER+ breast cancer. However, a large fraction of these patients become resistant to therapy and relapse. The EpiPredict consortium (http://www.epipredict.eu/) aims to uncover the key epigenetic changes underlying endocrine therapy induced resistance. In particular, methylation profiles of breast cancer patients are a prime candidate for the identification of loci linked to therapy resistance. We used the breast cancer subset of The Cancer Gene Atlas (TCGA), one of the major datasets available for studying the role of epigenetics in breast cancer. It contains methylation data of more than 700 breast cancer patients measured on Illumina 450K microarrays. We performed univariate and multivariate survival analysis on the methylation profiles of the primary tumors of tamoxifen-treated patients. Using multivariate Cox proportional hazards models with lasso and elastic net penalties, a reduced set of methylation sites was identified that may be predictive for therapy resistance. We discuss initial results of the methylation profiles for these sites in cell line models of endocrine therapy resistance and consider possible implications of these results for our understanding of (epigenetic) resistance mechanisms in breast cancer.	Data/ Application poster	Application Health
P_Da017	385	Andreas Andrusch, Piotr Wojciech Dabrowski, Jeanette Klenner and Andreas Nitsche	Andreas Andrusch	Identification of pathogen sequences in NGS datasets	NGS-based methods allow for the representative sequencing of all nucleic acids contained in clinical samples with their open view capacities.This enables the analysis of all generated reads for various known pathogens simultaneously but comes at the price of necessary filtering steps for the removal of background reads originating from the patient.Beyond the fact that NGS can extend the diagnostic possibilities provided by PCR, it can also serve as a stepping stone in the detection of novel pathogens.To achieve this we present the newly developed 'Pipeline for the Automatic Identification of Pathogens' (PAIPline) comprises a complete workflow for the pathogen search in NGS datasets, including several steps for the preprocessing and quality control of the raw data to ensure that only information-rich reads will be evaluated.It furthermore includes steps for the assignment of reads to their respective taxons based on reliable, established reference-based algorithms like Bowtie2 and BLAST.Filtering of background reads, contaminants and organisms of low interest as well as the evaluation of ambiguous read information is automatically done before the results are presented.Analysis results are shown in a highly accessible manner, allowing the researcher to gain a quick overview as well as permitting deep analysis.The performance of the PAIPline was benchmarked on real and artificial datasets of known compositions and compared to competing tools.The results and discussed features show that the presented approach is a viable strategy for the identification of pathogen sequences in NGS datasets.	Data/ Application poster	Application
P_Da018	528	Jorge Muñoz, Yuriy S. Shmaliy and Osbaldo Vite	Jorge Muñoz	Improving Confidence Masks to Estimate Genome CNAs Using SNP Array Data	Jitter in the breakpoints of chromosomal Copy Number Alterations (CNA) impacted by noise increase due to typically low signal-to-noise-ratio (SNR). We propose an improvement to the existing Confidence Mask through a Modified Bessel based Approximation (MBA). Function MBA fits the real jitter distribution and decreases error on approximation of jitter probability. We compared MBA and discrete skew Laplace distributions by simulated and single nucleotide polymorphism SNP array measurements and show the differences of confidence masks with both distributions apply to SNP data.	Data/ Application poster	Application
P_Da019	470	Wibowo Arindrarto, Sander van der Zeeuw, Peter van 'T Hof, Wai Yi Leung, Sander Bollen, Jeroen Laros and Leon Mei	Wibowo Arindrarto	Integrated Tracking of Next Generation Sequencing Pipeline Metrics	An enormous amount of sequencing data from various organisms is being generated daily.Depending on the research question, this sequencing data must be passed through a specific data analysispipeline, composed of various tools and scripts. These pipelines usually depend on a number of different externaldata sources, such as genome assemblies and gene annotations. Properly answering the research question meansone must take into account all of these dynamic sources. However, grappling with such a huge amount of data and variationis not a trivial task. We present an integrated solution that centers on Sentinel, a framework for creatingdatabases that track various metrics of a sequencing analysis pipeline run. The framework can in principle be usedto track metrics from a large number of custom pipelines, as long as the pipelines export their metrics as a JSONfile. A JSON schema can also optionally be added to ensure correct processing. The framework is implemented usingthe Scala programming language and is deployed as a web service that exposes a set of programming interfaces. Wedemonstrate a use case of our sequencing core group, where we integrate a Sentinel database with an interactivefront-end for visual exploration of these metrics, enabling quick overview of various metrics and identification ofoutliers. This setup has collected metrics of more than 1,700 RNA-seq samples and will further be expanded to collectmetrics from other sequencing setups with more well-defined ontology-based filtering.	Data/ Application poster	Application
P_Da020	558	Youri Hoogstrate, Alexander Senf, Jochem Bijlard,Saskia Hiltemann, David van Enckevort, Chao Zhang, Remond Fijneman, Jan-Willem Boiten, Gerrit Meijer,Andrew Stubbs, Jordi Rambla de Argila, Dylan Spalding and Sanne Abeln	Youri Hoogstrate	Integration of EGA secure data access into Galaxy	Bio-molecular high throughput data is privacy sensitive and can not easily made accessible to the entire outside world. To manage access to long term-archival of such data the EGA project was initiated to facilitate data access and management to funded projects after completion to enable continued access to these data. Strict protocols govern how information is managed, stored, transferred and distributed and each data provider is responsible for ensuring a Data Access Committee is in place to grant access to the data. Moreover, the transfer of data during upload and download of the data should be encrypted. As part of a TraIT-EGA ELIXIR pilot, here enable the download of EGA data to a Galaxy server in a secure way. Galaxy provides an intuitive user interface for molecular biologists and bioinformaticians to run and design workflows. More specifically, we developed a tool that can download data securely from EGA into a Galaxy server, which can subsequently be further processed. The tool ega_download_streamer is available in the Galaxy tool sheds. This together allows a user within the browser to run an entire analysis, containing privacy sensitive data from EGA, and to make this analysis available in a reproducible manner for other researchers. As proof of concept we have made an RNA-Seq workflow on cell-line data available.	Data/ Application poster	Application ELIXIR Fundamental
P_Da021	741	Junehawk Lee, Junho Kim, Minho Lee and Sangwoo Kim	Junehawk Lee	Machine learning based genetic variant filtration for detecting low-frequency somatic mutations	Recent rapid development of sequencing technologies has enabled examining low-frequency somatic variants. However, current somatic variant calling algorithms are impractical to distinguish true low-frequency somatic variants from prevalent errors emerged during sequencing procedures including library preparation and PCR amplification.To solve this problem, we produced a targeted capture sequencing data of a spike-in sample with 513 true somatic mutations, to discriminate the potential sequencing errors that can be detected as somatic by conventional mutation callers. By using the spike-in sequencing data as a training set, we developed a classifier to separate the possible false positive calls among the calls derived by the conventional somatic point mutation callers. When tested on 660 somatic calls (14 true positive and 646 false positive calls validated by independent amplicon sequencing) with b-allele frequency less than 2% obtained by MuTect algorithm, our classifier successfully filtered out 97% of false positive calls while misclassified 5 true positive calls (35% of total true positive calls). (AUC: 0.91; Sensitivity: 0.64, Specificity: 0.98)	Data/ Application poster	Application
P_Da022	755	Girolamo Giudice, Fatima Sanchez Cabo, Carlos Torroja Fungairiño and Enrique Lara Pezzi	Girolamo Giudice	MAGNETO: augMented functionAl analysis throuGh proteiN intEraction neTwOrk	An essential step in high-throughput data analysis is the biological interpretation through enrichment analysis to identify the over-represented processes and pathways. The major limitation of this approach is that the biological information contained in the molecular interaction network underlying the list of proteins of interest is not taken into account. Since proteins do not act in isolation, their biological effects depend on the neighboring polypeptides they interact with. For this reason, we developed MAGNETO a web server that extracts the maximum-likelihood tissue subnetwork (MLTSN) from the protein-protein interaction network. The MLTSN is highly representative of: (i) the paths connecting the proteins in the starting list and the proteins expressed in the tissue and (ii) the annotations that are likely to appear in the selected tissue. The nodes of the MLTSN represent the testing set for the enrichment analysis against databases such as Gene Ontology, Reactome, KEGG, KEGG drug, DrugBank and others. Our approach allows the discovery and refinement of the biological processes and pathways that usually do not emerge with the standard enrichment analysis. In addition, MAGNETO allows to (i) discover potential new targets for the existing drugs, (ii) to explore the effect of inhibiting a target protein by inhibiting its neighboring peptides and, (iii) to suggest pools of new proteins target for investigational drugs. Finally, MAGNETO implements interactive visualizations of the results that are of great use for interpreting the large amount of data produced as output.	Data/ Application poster	Application Fundamental
P_Da023	780	Bernd van der Veen, Ethan Cerami and James Lindsay	Bernd van der Veen	MatchMiner - An open computational platform for matching patient-specific genomic and clinical profiles to precision cancer medicine clinical trials.	The MatchMiner platform is a developmental effort of Dana-Farber Cancer Institute in collaboration with The Hyve, aiming to accelerate enrollment in precision medicine clinical trials and maximize clinical trial options for all patients.Using genomic, pathological and clinical profiling, a database is created which is allows the MatchMiner engine to search for simple and complex criteria sets defined by investigators in the user interface.MatchMiner is currently being developed in two distinct stages, after which point the entire platform will be made fully open source, and available to other institutions.The first stage of the platform is focused on "trial-centric" matching, enabling clinical trial investigators to create individualized genomic filters, and use these filters to forecast clinical trial enrollment, retrospectively identify new patients for clinical trials, and receive alerts of newly sequenced patients matching specific genomic criteria.The second stage of the platform is focused on "patient-centric" matching, enabling clinicians to view matching clinical trials for their specific patient, based on genomic eligibility and real-time clinical trial enrollment slot availability.In order to maximize adoption amongst clinicians and clinical trial managers, we closely worked together to collect feedback and make necessary design adjustments. MatchMiner will be released to the public and made available open source in Q4 2016 / Q1 2017.	Data/ Application poster	Application
P_Da024	830	Davide Albanese, Paolo Fontana, Alessandro Cestaro and Claudio Donati	Davide Albanese	MICCA 1.X: a state-of-the-art pipeline for amplicon-based metagenomic data processing.	The introduction of high throughput sequencing technologies has triggered an increase of the number of studies in which the microbiota of environmental and human samples is characterized through the sequencing of selected marker genes. While experimental protocols have undergone a process of standardization that makes them accessible to a large community of scientist, standard and robust data analysis pipelines are still lacking. Here we introduce MICCA, a software pipeline for the processing of amplicon metagenomic data that efficiently combines quality filtering of reads, OTU clustering, taxonomy classification, multiple sequence alignment and phylogenetic tree inference. The pipeline can be applied to a range of highly conserved genes/spacers, such as 16S rRNA gene, Internal Transcribed Spacer (ITS) and 28S rRNA. MICCA supports both single-end (Roche 454, Illumina MiSeq/HiSeq ,Ion Torren) and overlapping paired-end reads (Illumina MiSeq/HiSeq). MICCA includes state-of-the-art sequence clustering protocols such as the VSEARCH-based de novo greedy, Swarm, closed and open-reference. Moreover, widely used sequence classification algorithms are available (RDP and consensus-based classifier). A fast and memory efficient implementation of the NAST multiple sequence alignment is implemented since version 1.0. MICCA runs on Linux, Mac OS X and MS Windows (through Docker containers) and it is an open source project. Homepage: www.micca.org.	Data/ Application poster	Application
P_Da025	703	Duong Vu and Vincent Robert	Duong Vu	Multilevel clustering for massive biological data	With the availability of newer and cheaper sequencing methods, genomic data are being generated at an increasingly fast pace. In spite of the high degree of complexity of currently available search routines, the massive number of sequences available virtually prohibits quick and correct identification of large groups of sequences sharing common traits. Hence, there is a need for clustering tools for automatic knowledge extraction enabling the curation of large-scale databases. Currently, there are two approaches on sequence clustering. The first approach employs the idea of the greedy algorithm which has shown to be very efficient in time and memory for clustering large-scale datasets with UCLUST and CD-HIT. However, it does not guarantee a high accuracy for clustering. The second approach is based on pairwise similarity matrices. This is impractical for databases of hundreds of thousands or millions of sequences as such a similarity matrix alone would exceed the available memory. To overcome this problem, we have developed a tool called MultiLevel Clustering that could avoid a majority of sequence comparisons, and therefore, significantly reduces the total runtime for clustering while retaining high accuracy for clustering as current sophisticated approaches. An implementation of the algorithm allowed clustering of all 344,239 ITS fungal sequences from GenBank utilizing only a normal desktop computer within 22 CPU-hours whereas the greedy clustering method took up to 242 CPU-hours.	Data/ Application poster	Application
P_Da026	503	Ian Harrow, Martin Romacker, Andrea Splendiani, Stefan Negru, Peter Woollard, Scott Markel, Yasmin Alam-Faruque, Martin Koch, Erfan Younesi and James Malone	Ian Harrow	Ontologies Guidelines for Best Practice and a Process to Evaluate Existing Ontologies Mapping Tools	The Pistoia Alliance Ontologies Mapping project (http://www.pistoiaalliance.org/projects/ontologies-mapping) was set up to find or create better tools or services for mapping between ontologies in the same domain and to establish best practices for ontology management in the Life Sciences. It was proposed through the Pistoia Alliance Ideas Portfolio Platform (IP3: https://www.qmarkets.org/live/pistoia/home) and selected by the Pistoia Alliance Operations Team for development of a formal business case.The project has delivered a set of guidelines for best practice to build on existing standards. We show how they can be used as a "checklist" to support the application and mapping of source ontologies in particular domains. Another important output of this project was to specify the requirements for an Ontologies Mapping Tool. These were used in a preliminary survey that established that such tools already exist which substantially meet them. Therefore, we have developed a formal process to define and submit a request for information (RFI) from existing ontologies mapping tool providers to enable their evaluation. This RFI process will be described and we summarise our findings from evaluation of seven ontologies mapping tools from academic and commercial providers. The guidelines and RFI materials are accessible on a public wiki:- https://pistoiaalliance.atlassian.net/wiki/display/PUB/Ontologies+Mapping+Resources.Work is in progress to develop our requirements for an ontologies mapping service. We will conduct a survey of Pistoia Alliance members to understand the need for such a service and whether it should be implemented in future.	Data/ Application poster	Application
P_Da027	738	Artaza Haydee, Manuel Corpas, John Hancock and Rafael C Jimenez	Artaza Haydee	PlsCO: A Performance Indicators Framework for COllection of Biological Resource Metrics	Biological communities work across a range of domains and use a variety of biological resources. The selection of a particular resource can be aided by performance indicators to allow investigators to make informed decisions about alternatives. Furthermore, scientists may also need these indicators to justify the funding of a particular resource.When establishing a set of rigorous metrics, an important challenge is knowing the kind of indicators relevant to the scientist. Scientists frequently build their own methods, translating them into programs or scripts. Many of these programs or scripts are lost or forgotten when the project has finished. Hence a large amount of effort is wasted, and valuable methods and conventions that have been developed cannot be reused. We thus propose an approach for bringing together a set of potential measurements and conventions which can be reflected as metrics. Metrics include a variety of measures that provide tangible evidence and intuitive indicators that assess biological resources. Using metrics, a biological community can collect, disseminate and use valuable data essential for its work.We describe PlsCO, a Node.js JavaScript framework for collection/registration, dissemination and reuse of biological resource metrics. PlsCO can be used to: a) provide standard definitions of metrics; b) facilitate software to collect metrics; and c) facilitate the monitoring (by executing each metric's functionality automatically) and analysis of the retrieved metrics. In turn, these metrics data can be used by scientists, funders and academic institutions as performance indicators to assess the impact of biological resources to support decision-making.	Data/ Application poster	Application

P_Da028	715	John Santerre, Rick Stevens, Jim Davis and Fangfang Xia	John Santerre	Platform Based Machine Learning for AMR	Advances in DNA sequencing accompanied by plummeting cost is making sequence -based applications more amenable. Many web platforms are available for analysis (e.g. Galaxy, DNAnexus, OneCodex, etc), but tools that decipher patterns from data are not yet available to biologist as a web platform. Here we present our work building such a system.We are developing tools that enable statistical inference directly from sequencers for web-platforms. We use Random Forests(RF), a naively parallelizable and established Machine Learning algorithm, to produce classifiers that label strains as resistant(RES) or susceptible(SUS) after training. Using K-mers as features, the RF is trained to determine the optimal set of K-mers for classification of a novel strain as RES or SUS. RF provides a quantification of the importance of each K-mer, which allows us to identify the location of key mutations. We show that RF is highly accurate 80% (100 samples) and as high as 95% (3000 samples) in distinguishing between SUS and RES populations of S. pneumoniae and Mycobacterium tuberculosis. We cluster the significant K-mers by gene function from a reference genome and identify the most important features in existing literature for resistance and susceptibility. RF is appears to be robust, and despite lower accuracy on fewer strains (100 vs. 3000) it still is able to correctly identify genes known to be involved in antibacterial resistance. We believe one central outcome of cloud computing in biology will be the full integration of such tools and hope to help usher in that utilization.	Data/ Application poster	Application Biotechnology
P_Da029	593	Myungjun Kim, Yonghyun Nam and Hyunjung Shin	Yonghyun Nam	Prediction algorithm for multi-layered structure of omics	Background: Biological system is a multi-layered structure of omics with genome, epigenome, transcriptome, metabolome, proteome, etc., and can be further stretched to clinical/medical layers such as diseasome, drugs, and symptoms. One of the advantages of omics would be that we can figure out an unknown component or its trait by inferring from known omics components. The component can be inferred by the ones in the same level of omics or the ones in different levels. To implement the inference process, an algorithm that can be applied to the multi-layered complex system is required.Method: In this study, we develop a semi-supervised learning algorithm that can be applied to the multi-layered complex system. In order to verify the validity of the inference, it was applied to the prediction problem of disease co-occurrence with a two-layered network composed of symptom-layer and disease-layer.Results: The symptom-disease layered network obtained a fairly high value of AUC, 0.74, which is regarded as noticeable improvement when comparing o.59 AUC of single-layered disease network. If further stretched to whole layered structure of omics, the proposed method is expected to produce more promising results.	Data/ Application poster	Application
P_Da030	690	Jesse Cj van Dam, Jasper J Koehorst, Peter J Schaap, Vitor Ap Martins Dos Santos and Maria Suares-Diez	Jesse Cj van Dam	RDF2Graph a tool to recover, understand and validate the ontology of an RDF resource	Vast amounts of data are available in the life science domains and its doubling every year. To fully exploit this wealth, data has to be distributed using FAIR (findable, accessible, inter-operable and reusable) guidelines. To support interoperability, an increasing number of widely used biological resources are becoming available in the Resource Description Framework (RDF) data model.RDF triples represent associations: a gene codes for a protein, which has a function associated to a reaction generating specific metabolites. The semantically linked triples, subject – predicate – object, can be joined together to form a knowledge network.Structural overviews of RDF resources are essential to efficiently query them assess their structural integrity and design, thereby strengthening their use and potential. Structural overviews can be derived from ontological descriptions of the resources. However, these descriptions often relate to the intended content instead of the actual content.We present RDF2Graph, a tool that automatically recovers the structure of an RDF resource. The generated overview allows to structurally validate newly created resources. Moreover, RDF2Graph facilitates the creation of complex queries thereby enabling access to knowledge stored across multiple RDF resources. RDF2Graph facilitates creation of high quality resources and resource descriptions, which in turn increases usability of the semantic web technologies.	Data/ Application poster	Application
P_Da031	516	Dushyant Dudhagara, Rahul Rajpara, Jwalant Bhatt and Bharti Dave	Dushyant Dudhagara	Response surface methodology and artificial neural network modeling for fluoranthene degradation using Mycobacterium litorale	Present study aims to investigate fluoranthene degradation by Mycobacterium litorale using computation modeling i.e. response surface methodology (RSM) and artificial neural network (ANN). The effect of various operational parameters as CaCl2 (0.03-0.09 g L-1), KH2PO4 (0.3-0.8 g L-1) and NH4NO3 (0.3-0.8 g L-1) were investigated using two different computation modeling. RSM is the most preferred method for optimization of medium components to date. In last few years, the ANN method has developed as one of the most efficient methods for empirical modeling and optimization, especially for non-linear systems. This study represents the comparative analysis between RSM and ANN for their predictive, generalization capabilities, parametric effects and sensitivity analysis. Experimental data were evaluated by applying RSM integrated with a desirability function approach. In this study, one hidden layer along with the backpropagation algorithm was selected for the proposed ANN model. Consequently, the specific backpropagation algorithm and the number of hidden neurons were optimized. The RSM derived central composite design model, resulted in 51.21% degradation on 3rd day with R2 value 0.9882. The Non linear ANN model predicted 51.28% degradation with 0.9970 R2 value. The root means square error (RMSE) and mean absolute percentage error (MAPE) values were found to be 0.3234 and 0.5715. The results indicated that the ANN model was more precise, consistent and reproducible, compared to the RSM model, proving the superiority of ANN model over RSM model. The study thus opens new avenues for the development of such models for effective remediation strategies for PAHs impacted habitats.	Data/ Application poster	Application Biotechnology
P_Da032	794	Christian Ruckert	Christian Ruckert	Sciobase: A platform for the evaluation of variants from next-generation-sequencing experiments	We developed Sciobase a platform to annotate, evaluate and store variants from next-generation-sequencing experiments. Variants are called using a standard GATK workflow complemented by diverse preprocessing, quality control and visualization programs. Afterwards perl and shell scripts calculate and fetch annotations from multiple public databases and store these together with data from the run output files (e.g. vcf-files, quality reports, links to bam files) into the database. A web front-end allows the visualization and filtering of variants, the analysis of coverage profiles, the creation of reports and the design of primer oligos to validate variants by Sanger sequencing. At the moment we are running three different instances of Sciobase for miscellaneous projects containing about 3000 samples in total. These range from smaller gene panels up to whole genome data. The collection of variants together with phenotype information into a database allows an improved scoring of variants compared to ExAC or 1000 Genomes project frequencies alone.We studied the association between the classification of variants by clinical experts into one of five severity classes and different scoring algorithms used for variant effect prediction.Based on the variants stored in the database so far we identified a small set of variants able to uniquely identify samples. With this set of variants we implemented a SNPshot approach to detect sample swaps. Variants can be analyzed on a single sample basis or compared between different samples. Another module allows the analysis of pedigree data for compound heterozygous variants.	Data/ Application poster	Application
P_Da033	869	Seonho Kim and Hong-Woo Chun	Seonho Kim	Spatial and Contextual EEG information learning for the Diagnosis of Alcoholism	EEG is data source with great potential which is widely being studied for diagnosis of brain disease because it is un-substitutable as well as relatively easy to obtain bio-signal from brain. However, because of many reasons, such as the difficulties in detecting correct sensing positions, in removing noises, in regularizing the strength, etc., technologies still need to be developed for analyzing EEG data. Our research interest lays in early detection of Alzheimer's, or dementia, by using the EEG data, and the actual data from Alzheimer's patients has been collecting this year. In this poster, we present the results of our preliminary tests to identifying alcoholic, instead of dementia, from Alcoholic-Control EEG data obtained from UCI data mining repository in the assumption that the technologies for identifying two disease, dementia and alcoholic, from EEG may not different significantly and can help each other. Our approaches employ various deep learning techniques, such as convolutional neural networks, deep belief neural networks (DBN), LSTMs and their combinations. According to our early experiments, brute force learning with deep belief network with raw EEG data has been yielded greatest performance so far. However, experiments with cNN, LSTM, and their combination shows potentials of enhancement. Our focus is on adding spatial and contextual information of EEG which is not included explicitly in the DBN learning model. We reproduced EEG contour brain map from the raw EEG and the position information of 64 electrodes. Many parameters, such as activation functions and layer numbers, also have been tested.	Data/ Application poster	Application
P_Da034	677	Tammi Vesth, Sebastian Theobald, Inge Kjærbølling, Jane L. Nybo, Ronald de Vries, Igor Grigoriev, Scott Baker and Mikael Rørdam Andersen	Tammi Vesth	The Aspergillus Mine - publishing bioinformatics	Genome analysis is no longer a field reserved for specialists and experimental laboratories are doing groundbreaking research using genome sequencing and analysis. In this new era, it is essential that data, analysis and results are shared between scientists. But this can be a challenge, even more so with no computational specialist. Here we present a setup for analysis and publication of genome data of 70 species of Aspergillus fungi. The platform is based on R, Python and uses the RShiny framework to create interactive web-applications. It allows all participants to create interactive analysis which can be shared with the team and in connection with publications. We present analysis for investigation of genetic diversity, secondary and primary metabolism and general data overview. The platform, the Aspergillus Mine, is a collection of analysis tools based on data from collaboration with the Joint Genome Institute. The Aspergillus Mine is not intended as a genomic data sharing service but instead focuses on creating an environment where the results of bioinformatic analysis is made available for inspection. The data and code is public upon request and figures can be obtained directly from the web-app. This resource will be of great benefit to the Aspergillus community which is in a rapid development in regards to genome sequencing and analysis. At the moment, the service includes analysis of more than 70 genomes, and is expected to double in the next 6 months, with the final goal of the project is the analysis of 300 Aspergillus species.	Data/ Application poster	Application Biotechnology Fundamental
P_Da035	863	Fabio Rinaldi and Lenz Furrer	Fabio Rinaldi	The Bio Term Hub: an integrated resource of biomedical terminology	A coherent, uniform, and unambiguous technical terminology is anessential prerequisite for successful scholarlycommunication. However, in the domain of life sciences, terminology isoften ambiguous and redundant. As an example of the problems createdby the ambiguity of terminology consider the string "cat". A naivesearch in the literature could return several different types ofentities. As well as referring to the animal, it could refer to amedical procedure (computerized axial tomography), and it also acommon abbreviation for a biological process (catalitic activity). Additionally, a search in Uniprot reveals 1346 proteinswhich have a variant of the same string among their synonyms.Almost all life science databases maintain comprehensiveterminologies, in particular the names of the entities that theycurate, yet terminology management is not typically part of their corecompetences. We are creating a unique centralized repository which canfunction as a clearing house for biomedical terminology[1]. Existingterminology from databases is automatically collected and keptsynchronized with them. A web interface provides detailed informationabout each term, and global statistics. For each term, we indicate allentities that have it among their possible names, the databases whereit occurs, and the lexical properties of the term, i.e. statisticsabout polysemy, and synonymy. The primary users of this resource areexpected to be in the biomedical text mining community, where theavailability of rich lexical resources is of crucial importance inorder to achieve accurate analysis of the scientific literature andother textual data.[1] http://pub.cl.uzh.ch/purl/biodb/	Data/ Application poster	Application Fundamental
P_Da036	710	Theo Knijnenburg, Ilya Shmulevich, Sheila Reynolds, Phyllis Lee, Michael Miller, Kelly Iverson, Abigail Hahn, Zack Rodebaugh, Kalle Leinonen, David Gibbs, Varsha Dhankani, Jonathan Bingham, Nicole Deflaux, Matt Bookman and David Pot	Theo Knijnenburg	The ISB Cancer Genomics Cloud	The ISB Cancer Genomics Cloud (ISB-CGC) is one of three pilot projects funded by the National Cancer Institute with the goal of democratizing access to the The Cancer Genome Atlas (TCGA) data by substantially lowering the barriers to accessing and computing over this rich dataset. The ISB-CGC is a cloud-based platform that serves as a large-scale data repository for TCGA data, while also providing the computational infrastructure and interactive exploratory tools necessary to carry out cancer genomics research at unprecedented scales. The ISB-CGC facilitates collaborative research by allowing scientists to share data, analyses, and insights in a cloud environment. The ISB-CGC team includes scientists and engineers from the Institute for Systems Biology (ISB), Google, and CSRA. If you are interested in learning more about the ISB-CGC or would like to propose specific scientific use-cases to our development team, please visit us at www.isb-cgc.org.	Data/ Application poster	Application
P_Da037	454	Georg Summer, Thomas Kelder, Marijana Radonjic, Marc van Bilsen, Suzan Wopereis and Stephane Heymans	Georg Summer	The Network Library: A Framework to Rapidly Integrate Network Biology Resources	Much of the biological knowledge accumulated over the last decades is stored in different databases governed by various organizations and institutes. Integrating and connecting these vast knowledge repositories is an extremely useful method to support life sciences research and help formulate novel hypotheses. We developed the Network Library, a framework and toolset to rapidly integrate different knowledge sources to build a network biology resource that matches a specific research question. As a use-case we explore the interactions of genes related to heart failure with miRNAs and diseases through the integration of 6 databases (STRING-DB for protein-protein interactions, DisGeNet for disease associations, miRDB, TargetScan, DIANA microT CDS and miRTarBase for miRNA-gene targeting). This poster will explore the creation of the network and exemplary analysis using the Network Library, cyNeo4j and Cytoscape.More information about the Network Library and the network creation process is available at bionetlib.wordpress.com.	Data/ Application poster	Application Health
P_Da038	754	Florian Graef, Guilherme Formaggio De Mello and Johanna McEntyre	Johanna McEntyre	The THOR project: Integrating persistent identifiers such as ORCIDs in life sciences data resources	The THOR (Technical and Human infrastructure for Open Research) project (http://project-thor.eu) is a 30-month project funded by the European Commission under the Horizon 2020 programme. In general, THOR aims to extend the integration of persistent identifiers (PIDs) into platforms, services and workflows. The aim is not to build new, standalone services, but to work with existing systems and communities, in this case, the life sciences research community. By creating new and improved integrations of PIDs in the services that researchers and institutions actually use, we aim to ensure that PIDs are usefully embedded in research outputs and activities from the very beginning, with minimal effort for researchers.Life sciences researchers typically publish articles as the major research output, and work by many stakeholders such as the ORCID Foundation, CrossRef, publishers and Europe PMC have gained traction on the integration of ORCIDs into article submission, publication, and distribution systems. Currently there are over 2.5M articles in Europe PMC that have at least one associated ORCID, from around 250,000 unique ORCIDs (i.e. people). The THOR project wishes to capitalise on this adoption in publications, extending into claiming datasets to ORCIDs.We are building services that allow ORCIDs to be integrated into data submission systems, as well as allowing retrospective claiming of data to ORCID records, positioning these contributions alongside articles published and grants awarded.As a first step ORCID authentication has been integrated into the submission forms of the EMBL-EBI resources MetaboLights and EMPIAR.	Data/ Application poster	Application
P_Da039	441	Kumar Parijat Tripathi, Daniela Evangelista, Antonio Zuccaro and Mario Guarracino	Kumar Parijat Tripathi	Transcriptator: a user-friendly graphical interface to functionally characterize novel transcripts and identify non-coding RNA.	Exploring the transcriptomes of interesting non-model organisms in the absence of well-established genome is a difficult task, and inferring biological knowledge from distinct transcriptomic experiments is error prone. In our lab, we develop a Transcriptator web application based on a computational Python pipeline with a user-friendly Java interface. This pipeline uses the web services available for BLAST (Basis Local Search Alignment Tool), Quick-GO and DAVID tools. It offers a tabular report and graphical charts on statistical analysis of functional annotation enrichment and slimming of GO terms. It enables a biologist to identify enriched biological themes, particularly Gene Ontology (GO) terms. It helps in clustering the transcripts based on their common functionalities. Implementation of PORTRAIT (Prediction of transcriptomic ncRNA by ab-initio methods) in our pipeline enables us to identify non-coding RNA in a transcriptome. It helps the user to characterize the de-novo assembled reads, which does not map to genome. Later we investigate the regulatory role of these non-coding RNA on gene transcription. The pipeline is modular in nature, and provides an opportunity to add new plugins in the future. Web application is freely available at: www-labgtp.na.icar.cnr.it/TranscriptatorReference: Tripathi, K. P., Evangelista, D., Zuccaro, A., & Guarracino, M. R. (2015). Transcriptator: An Automated Computational Pipeline to Annotate Assembled Reads and Identify Non Coding RNA. PloS one, 10(11), e0140268.	Data/ Application poster	Application Biotechnology
P_Da040	848	Jennifer Leclaire, Stefan Tenzer and Andreas Hildebrandt	Jennifer Leclaire	triMS5 - storing LC-IMS-MS data sets in HDF5	Mass spectrometry (MS) is a quickly evolving analysis technique with a wide range of applications, including proteomics. Recent innovations such as the integration of ion mobility separation (IMS) and data-independent acquisition (DIA) lead to dramatic increase in both file sizes and complexity of raw data. Typically, the recorded raw data is stored in proprietary vendor file formats. Software packages for the handling of such files are usually closed-source or restricted to Microsoft Windows operating systems.Here, we present triMS5, a file format for storing LC-IMS-MS data based on the Hierarchical Data Format 5 (HDF5), a well-established binary file format for scientific data with various supported programming languages and operating systems. The basic abstraction of HDF5 are array-like data sets which can be further divided into subsets called chunks. Each chunk can be operated on individually, e.g., by subjecting it through natively supported compression filters. Our format combines these mechanisms with a compressed row storage (CSR) strategy to exploit the sparse nature of LC-IMS-MS raw data. To enable efficient range queries, triMS5 uses a multi-dimensional kd-tree to index chunks. Hence, triMS5 allows to access all three dimensions (m/z, retention and drift time) with equal effort, and supports rapid access to signal regions of interest. Compared to the PSI-standard file format for MS raw data, the XML-based mzML, triMS5 approximately halves the file sizes. In its current state, triMS5 is only specified for LC-IMS-MS data but its generic storage layout may also be applied to other data storage challenges in MS.	Data/ Application poster	Application

P_Da041	482	*Parham Solaimani Kartalaei, Maarten-Jan Kallen and Alexander Bertram*	*Parham Solaimani Kartalaei*	*Using R language based bioinformatic workflows as Product-as-a-Service*	*Most scientists use open source tools for development and use of novel analytic methods. Beside the low immediate costs of such tools, scientists benefit from more thorough and transparent testing and validation. The R statistical programming language with the accompanying GNU R interpreter (GNU-R, http://cran.r-project.org) is one of the most successful examples. There are currently over 10.000 packages developed for R with almost 2.000 Biology related packages in BioConductor (http://bioconductor.org), covering most bioinformatic needs and allowing easy development of new analytical workflows.While sufficient for most day-to-day analytic tasks, the current architecture of GNU-R poses limitations in its usability in development of scalable and interactive Product-as-a-Service (PaaS), as it has not been designed for deep integration with web and distributed computing technologies. This is reflected in the current scarcity of PaaS with R-based workflows as back-end. Here we give an overview of the requirements for R based PaaS development and current most promising solutions, while highlighting their strengths and limits.*	*Data/ Application poster*	*Application*
OTHER POSTERS WITHIN DATA THEME							
P_Da043	432	Lingjian Yang, Amanda Williamson, Joely Irlam, Helen Denley, Peter Hoskin, Ananya Choudhury and Catharine West	Lingjian Yang	A network-based approach to derive hypoxia gene signature for bladder cancer patients	Bladder cancer is a common malignancy in the UK. Tumour hypoxia affects the micro-environment, promotes intrinsic resistance to therapy, and is associated with a poor prognosis in bladder cancer. Hypoxia-related RNA-expression signatures have been derived as promising biomarkers for routine clinical application. While such hypoxia gene signatures were successfully proposed for head and neck, breast and lung cancers with strong prognostic values being demonstrated in independent clinical cohorts, there is no bladder cancer-specific hypoxia gene signature. This study, therefore, aimed to derive a novel hypoxia gene signature for bladder cancer patients.A database (n=258) was constructed of genes identified in the literature as hypoxia-related in multiple tumour sites. Publicly available transcriptomic profiles were analysed and a bladder cancer hypoxia gene co-expression network built around the genes of interest from literature by pooling together strong gene-gene interactions. Hub genes (n=17) were identified that collectively reflected the intra-tumour gene expression heterogeneity and then taken forward for validation as a gene signature. Internal cross validation within the training cohort showed the prognostic value of the signature. The signature was independently validated by gene expression profiling samples from a phase III trial cohort where patients were randomised between radiotherapy alone or with hypoxia-modifying carbogen and nicotinamide (CON). Patients stratified as high-hypoxia by the signature benefited from CON (HR for overall survival 0.44, 95% CI 0.24-0.82, P=0.01), while those classified as low-hypoxia derived no benefit. This is the first bladder cancer signature showing prognostic and predictive value in clinical cohorts.	Data poster	Health
P_Da044	580	Fotis Psomopoulos, Athanassios Kintsakis and Pericles Mitkas	Fotis Psomopoulos	A pan-genome approach and application to species with photosynthetic capabilities	MotivationThe abundance of genome data being produced by the new sequencing techniques is providing the opportunity to investigate gene diversity at a new level. A pan-genome analysis can provide the framework for estimating the genomic diversity of the dataset at hand and give insights towards the understanding of its observed characteristics. Currently, there exist several tools for pan-genome studies, mostly focused on prokaryote genomes and their respective attributes. Here we provide a systematic approach for constructing the groups inherently associated with a pan-genome analysis, using the complete proteome data of photosynthetic genomes as the driving case study. As opposed to similar studies, the presented method requires a complete information system (i.e. complete genomes) in order to produce meaningful results.ResultsThe method was applied to 95 genomes with photosynthetic capabilities, including cyanobacteria and green plants, as retrieved from UniProt and Plaza. Due to the significant computational requirements of the analysis, we utilized the Federated Cloud computing resources provided by the EGI infrastructure. The analysis ultimately produced 37,680 protein families, with a core genome comprising of 102 families. An investigation of the families' distribution revealed two underlying but expected sub-sets, roughly corresponding to bacteria and eukaryotes. Finally, an automated functional annotation of the produced clusters, through assignment of PFAM domains to the participating protein sequences, allowed the identification of the key characteristics present in the core genome, as well as of selected multi-member families.	Data poster	Fundamental
P_Da045	655	Andrian Yang, Michael Troup and Joshua Ho	Andrian Yang	A quick and flexible transcriptomic feature quantification framework on the cloud	Major advancement in single-cell capture technology has resulted in the increasing interest in single-cell level studies, particularly in the field of transcriptomics. Current tools designed for transcriptomic analysis are unable to efficiently handle this increasingly large volume of sequencing data generated.To tackle this problem, we have implemented a cloud-based framework for the simultaneous processing of large-scale transcriptomic data. The pipeline utilises state-of-the-art Big Data technology of Apache Hadoop, a MapReduce framework, and Apache Spark, a general purpose data analytics engine, to perform massively parallel alignment and feature quantification analysis of transcriptomic data on a cloud-computing environment which can be scaled to meet user requirements.The default pipeline makes use of STAR for sequence alignment and featureCount for feature quantification. Nonetheless, the pipeline is customisable in terms of choice of parameter and tools for alignment and feature quantification. Our framework also performs RNA-seq data quality control using Picard.We evaluated the performance of the pipeline using a public single-cell mouse RNAseq dataset (869 samples, 1.28T bases) on a 10 node Amazon Elastic MapReduce cluster (320 cores, 2.21TB RAM). The analysis was completed in 6.75 hours, which is 4.3x faster compared to performing the same analysis on an equivalent single computing resource. The pipeline offers the use of low-cost spot instances, providing a saving of 3.32x (US$65.10 spot vs US$216.30 on-demand) for the analysis performed.	Data poster	Fundamental
P_Da046	555	Krzysztof Mnich and Witold Rudnicki	Krzysztof Mnich	A robust approach for discovery of synergistic variables	The biological datasets, like data obtained in gene expression studies or GWAS, are often described with a large number of variables. Identification of the variables that are relevant for the phenomena under investigation is therefore an important initial step of data analysis.Usually it is performed using univariate test for association between descriptive variable and decision variable.However, this approach ignores variables that contribute information on the decision variable only when considered in association with other variables, exhibiting synergy effects.Here we present a methodology to discover such variables, based on the information theoretic approach. The key notion is the weak relevance introduced in [1]. The variable is weakly relevant when it contributes information on decision when added to some other set of variables. We use this definition directly to find whether given variable contributes additional information to a k-tuple of variables.Then we perform analysis of the maximal contribution of given variable in the context of all possible k-tuples. The theoretical distribution for p-value is in this case exponential distribution. The variables with sufficiently small p-values are declared relevant.The methodology was applied to the adaptive immune response in chicken studied in [2]. Significant synergistic effects were discovered for pairs and triplets of variables.Research was supported by the grant from the Polish NSC, grant UMO-2013/09/B/ST6/01550.[1] Kohavi R. John, G. Artificial Intelligence (97), 1997.[2] Siwek M. et al. Animal Genetics (46), 2015.	Data poster	Health
P_Da047	514	Christian Wünsch, Henrik Banck, Jan Stenner and Martin Dugas	Christian Wünsch	AML-Varan – a web-based platform to display and analyze genomic variants from targeted Next-generation sequencing data in clinical practice	Within the past years, many prognostic genetic mutations have been identified that are important to select the best treatment for patients with Acute Myeloid Leukemia (AML). Currently mutation analysis in routine care is done by Sanger sequencing or PCR-based methods, which are suffering from limitations regarding costs, effort or small regions of detection.New NGS methods allow to compensate those shortcomings, but they tend to produce a very large amount of variants with numerous and complex possibilities of annotation. Therefore IT-tools to display and interpret the NGS-data in clinical settings are needed.We analyzed a dataset of 120 targeted-sequencing samples, predominantly from AML patients, with 520 kBp target length. The resulting data was used to implement and evaluate a web-based platform on the basis of MySQL, PHP and JavaScript/Ajax technology, that displays the variants and provides annotation information from ClinVar, COSMIC and CIViC databases.Our software AML-VarAn ("AML Variant Analyzer") is based on a central database that contains 120 samples with a total of 90,000 variants. Raw sequencing results (fastq) or variant lists (vcf) can be imported, and all tables can be exported to csv format.The user interface consists of four display modules: Hotspot regions, Filtered variants, Complete panel and Coverage analysis.The large amount of variants per sample (average 750) showed that an IT-tool is necessary for the analysis of the provided data. Unfortunately the interpretation suffers up-to-now from the fact that annotation of variant pathogenicity (of recent clinical databases) is often incomplete and difficult to validate.	Data poster	Health
P_Da048	798	Francesca Mulas, Chun Zeng, Yinghui Sui, Gene Yeo and Maike Sander	Francesca Mulas	Analysis of Single Cells on a Pseudotime Scale along postnatal pancreatic beta cell development	Single-cell RNA-seq generates gene expression profiles of individual cells and has furthered our understanding of the developmental and cellular hierarchy within complex tissues. One computational challenge in analyzing single-cell data sets is reconstructing the progression of individual cells with respect to the gradual transition of their transcriptomes. While a number of single-cell ordering tools have been proposed, these require knowledge of progression markers or time delineators. Here, we adapted an algorithm previously developed for temporally ordering bulk microarray samples to reconstruct the developmental trajectory of pancreatic beta-cells postnatally. To accomplish this, we applied a multi-step pipeline to analyze single-cell RNA-seq data sets from isolated beta-cells at five different time points between birth and post-weaning. Specifically, we i) ordered cells along a linear trajectory (the Pseudotime Scale) by applying one-dimensional principal component analysis to the normalized data matrix; ii) identified annotated and de-novo gene sets significantly regulated along the trajectory; iii) built a network of top-regulated genes using protein interaction repositories; and iv) scored genes for their network connectivity to transcription factors. A systematic comparison showed that our approach was more accurate in correctly ordering cells for our data set than previously reported methods. Importantly, our analysis revealed never before seen changes in beta-cell metabolism and in levels of mitochondrial reactive oxygen species. We demonstrated experimentally a role for these changes in the regulation of postnatal beta-cell proliferation. In sum, our pipeline identified maturation-related changes in gene expression not captured when evaluating bulk gene expression data across the developmental time course.	Data poster	Biotechnology
P_Da049	561	Agnes Hotz-Wagenblatt, Lin Wang, Renuka Pasupuleti, Christopher Previti and Karl-Heinz Glatting	Agnes Hotz-Wagenblatt	Are you missing important variant information with whole exome sequencing due to coverage problems?	Exome sequencing is widely used in cancer research area nowadays due to its efficiency and cost-effectiveness. Exome sequencing provides relatively high coverage across the coding regions of genome which is essential for detecting variants. But the coverage of the enrichment regions is not uniformly distributed. There are still certain regions which are lowly covered. These regions with inadequate depth may cause problems during variant calling thus give biased biological outcomes.There are two ways that a gene region is not or lowly covered, either by design of the panel or by the sequencing technology. We looked at the Illumina Agilent SureSelect V5 with and without UTRs to analyse the not or lowly covered regions. We checked the design by comparing the target regions as given by Illumina with the annotation of Ensembl V74 and Cosmic V70 (human genome 37). We checked the sequencing technology by analyzing exome data of 17 tumor samples and 12 blood samples (HIPO, Heidelberg Center for Personalized Oncology). Regarding panel design, despite the fact that the general gene coverage is above 90%, about 20 of Cancer Census Genes are only covered less than 50%. Regarding the read coverage of the target regions in tumor and normal data we discovered that only about 100,000 bases (out of 50,390,601) are lowly covered. But in those regions a significant amount of cosmic mutations is localized. About half of those regions have low coverage due to a high GC content. Further analyses will be shown.	Data poster	Fundamental
P_Da050	384	Seyed Ziaeddin Alborzi, Marie-Dominique Devignes and David Ritchie	Seyed Ziaeddin Alborzi	Associating Gene Ontology Terms with Protein Domains	The fast growing number of protein structures in the protein data bank (PDB) raises new opportunities for studying protein structure-function relationships. In particular, as the biological activity of many proteins often arises from specific domain-domain and domain-ligand interactions, there is a need to provide a direct mapping from structure to function at the domain level. Many protein entries in PDB and UniProt are annotated to show their component protein domains according to various classifications (Pfam or CATH), as well as their molecular function through the Gene Ontology (GO) terms. We therefore hypothesize that relevant GO-domain associations are hidden in this complex dataset of annotations. We use as gold-standard all GO-domain associations available from InterPro database and we define GODomainMiner, a novel content-based filtering method to associate GO terms with Pfam domains using SIFTS and the UniProt databases. The GODomainMiner approach associates GO terms with Pfam domains based on the structures and sequences that they share. GODomainMiner finds a total of 20,318 non-redundant GO-Pfam associations for molecular functions in a completely automatic fashion with a recall of 0.96 with respect to the associations present in the InterPro database (1,561 associations). The novel calculated GO-Pfam associations could add value to the description of structural domains of unknown function in Pfam database. They are currently undergoing comparison with the GO-SCOP and GO-CATH domain associations. Moreover, the GODomainMiner resource could be used to annotate thousands of PDB chains or protein sequences which currently lack any GO annotation although their domain composition is known.	Data poster	Fundamental
P_Da051	550	Lilit Nersisyan, Anna Hakobyan and Arsen Arakelyan	Lilit Nersisyan	Association of telomere length with epigenetic regulation of gene expression	Telomere length dynamics plays a crucial role in cancers through variety of yet poorly characterized mechanisms. One of the important issues is to find the association of telomere length with changes in epigenetic mechanisms of regulation of gene expression. Here we have analyzed whole genome sequencing (WGS), RNA-seq, ChIP-seq and DNA methylation data from lung adenocarcinoma cell lines to identify epigenetic modification events linked to gene expression and correlated with telomere length dynamics.The mean telomere length (MTL) was estimated from the WGS data with the Computel software. MTL association with gene expression, DNA methylation and ChIP-seq data was assessed with multivariate linear regression approach.Our data indicated that MTL was individually associated with gene expression, methylation and modification of at least one histone mark for 847, 438, and 105 genes, respectively. 15 genes had both expression and methylation marks, while only two genes (FAM84B, VPS37B) had both histone modification and gene expression marks associated with MTL. Among these 17 genes there were chromatin modifiers (HAT1, METTL16, MLL3), genes implicated in cancers (PLXNA3, FARSA), differentially expressed in telomere elongated cancer cells (FEM1C), or known to be differentiation (PLXNA3) or ageing (VPS37B) dependent. Interestingly, PLXNA3, METTL16 and MLL3 are located very close to the telomere end, implicating a possibility of chromosome position dependent regulation.Altogether, our data have revealed genes presumably associated with telomere length via epigenetic regulatory mechanisms. The causality of the found associations has to be validated, and their role in cancer development is subject to further studies.	Data poster	Health
P_Da052	585	Sarah Elshal, Jesse Davis and Yves Moreau	Sarah Elshal	Beegle 2.0: Yes! We can start from literature mining and end up with disease-gene discovery	Studying our genetic information such that we are able to resolve which genes spell out which diseases is very exciting. Not only does it offer us the chance to better diagnose the diseases, but also cure them in a more effective way. Nevertheless, these kinds of studies are very challenging. They require a lot of literature review, genomic screening, gene association studies, linkage analysis, etc. Previously we have developed Beegle, a generic tool for disease-gene discovery. In a first phase Beegle applies text mining to identify which genes are found to be linked with any given disease of interest. Then in a second phase it applies a genomic data fusion strategy to learn a model and prioritize the whole genome according to how well a gene is predicted to be potentially linked with the original disease of interest. In this poster we would like to present a recent realistic study, which shows that in a two-year span Beegle succeeded to rank at least 36 true novel genes for 20 test diseases in the top 20 ranked genes (top 0.1% of the human genome). We would also like to present a new version of Beegle, which not only presents the user with a better web interface, but it also relies on an updated release of the literature data and a better text mining strategy. Beegle is publicly available at: http://beegle.esat.kuleuven.be/.	Data poster	Biotechnology
P_Da053	395	Sascha Losko, Richard Albang, Hildegard Menke, Verena Schütz, Emiel Ver Loren van Themaat ,Martin Wolff, Kaj Albermann, Klaus Heumann, Hans Roubos and Marco de Groot	Sascha Losko	Beyond Silos: Knowledge Management as the Key to Operational Excellence in Genetic Engineering	In recent years, knowledge management systems and semantic technologies have become standard components of large-scale enterprise software infrastructures — with applications ranging from research, discovery and development all the way to operations. Process optimization and manufacturing greatly benefit from a managed "knowledge/feedback loop". In this talk, Biomax presents its premier knowledge management platform, the BioXM system, which was used to develop a genetic engineering solution together with DSM.Cost-effective DNA sequencing and de novo DNA synthesis have facilitated the emergence and rapid development of modern biotechnology. The development of DNA assembly standards, publicly available part registries for sharing bioparts, and computer-aided design (CAD) tools have been instrumental in accelerating discovery. Applications of modern biotechnology include renewable energy sources and biofuels, industrial enzymes, biosensors, bio-based chemicals, plastics, textiles and other raw materials.The BioXM Knowledge Management system "puts it all together," enabling life scientists to visualize, study, create and alter highly complex pathways and DNA sequences in their genomic context. This allows efficiently building and characterizing parts repositories with respect to information, part function and performance, and using these repositories to help design biological systems targeting the desired functionality in a truly "design → build → test → learn" iterative approach.	Data poster	Biotechnology
P_Da054	737	Bas Stringer, Albert Meroño-Peñuela, Frank Van Harmelen, Sanne Abeln and Jaap Heringa	Bas Stringer	BLASTing the Semantic Web	Life sciences are rapidly adopting Semantic Web technology. An ever-growing amount of databases are (partially) exposed as RDF graphs (e.g. Uniprot, TCGA, Disgenet, Human Protein Atlas…), complementing traditional methods to disseminate biological data. The SPARQL query language provides a powerful tool to rapidly retrieve and integrate (bio)data from different sources. However, the inability to incorporate quantitative reasoning in SPARQL queries inhibits its application in many life science use cases: for example, one may want to find the homologs of a specific protein which are coexpressed in the same tissues. In order to do this, one needs to link up sequence data (e.g. Uniprot), tissue-specific expression data (e.g. Human Protein Atlas) and a quantitative homology detection method (e.g. BLAST).We developed the SPARQL compatible service layer (SCRY), which provides a mechanism for incorporating quantitative data processing within SPARQL queries in a reusable, interoperable manner. SCRY is a lightweight SPARQL endpoint that interprets specific parts of queries as calls to user defined procedures. This allows users to gather input data, derive knowledge from it on-demand, and use the output within a single, reusable query. We demonstrate the power of this approach by finding the tissues which express Hemoglobin β, its homologous proteins, and the tissues which express these homologs in a single SPARQL query.	Data poster	Fundamental

P_Da055	859	Sjoerd M. H. Huisman, Baldur van Lew, Ahmed Mahfouz, Nicola Pezzotti, Thomas Hollt, Lieke Michielsen, Anna Vilanova, Marcel Reinders and Boudewijn P.F. Lelieveldt	Sjoerd M. H. Huisman	BrainScope: interactive visual analysis of brain-wide genome-wide expression data	Molecular neuroscience deals with the activity of genes in the brain, and therefore encompasses the collection and analysis of highly complex datasets. The Allen Institute for Brain Science provides these data, in spatial and spatio-temporal atlases of gene expression. Because of the high number of genes and anatomical regions involved, visualisation of this data is challenging. Current tools often focus either on genes in coexpression modules, or on transcriptional similarities between areas of the brain.We present the BrainScope portal, for visualisation of gene expression data in the brain, which shows both relationships between genes and between samples. It features interactive scatterplots (maps) of genes and samples, made with t-distributed stochastic neighbourhood embedding (t-SNE).The gene map is genome-wide, and is structured according to spatial expression patterns. We show that these patterns are partially driven by cell-type composition, and that genes that cluster together tend to share molecular functions and biological processes.This gene map is linked to the sample map, which shows how anatomical annotation is related to co-expression. Users can select brain regions of interest and find the genes that are highly expressed in these regions.The BrainScope portal visualizes the landscape of gene expression in the brain, both on a global and local level. It is genome-wide and offers the unique opportunity to visually explore relationships both between genes and between anatomical samples in the human brain.	Data poster	Fundamental
P_Da056	671	Jaak Simm, Adam Arany, Hugo Ceulemans and Yves Moreau	Jaak Simm	Broker Macau: joint model building with privacy preservation	We present a method for creating a joint model where involved parties want to avoid explicitly sharing their raw data. In this work we consider P partners who each have a set of input features Xi lying in the same space and partially observed output matrices Yi. Each partner wants to make predictions on its Yi. An example of this setup is where several pharmaceutical companies want to predict compound activities Yi on their assays from chemical structures Xi. The goal of the method is to improve individual models by learning a joint model without sharing private activity matrices Yi.To this end we propose a method of collaborative matrix factorization of Y=[Y1,...,YP] with side information of input features X, where a central broker infers the effect of the side information (chemical structures) without gaining explicit knowledge on the datasets Yi. For that we use Bayesian matrix factorization Macau [1]. The method Broker Macau allows the partners to build a joint model where each partner only learns the factorization of its own matrix Yi and thus is able to make predictions only on its data. With the help of homomorphic encryption system Paillier we ensure that the broker cannot reveal the details of the data.We show empirically that increasing the number of partners improved the accuracies for the individual partners. Additionally, Broker Macau can scale to large datasets of millions of compounds and thousands of assays.[1] https://github.com/jaak-s/macau	Data poster	Health
P_Da057	813	Aurelie Martin, Laurent Naudin and Sébastien Tourlet	Aurelie Martin	Characterization and bioinformatic analysis of a prostate cancer multi-scale network: Gene co-expression, mutome, interactome	This present work is retrospective analysis starting in 2012 in Prostate cancer. Prostate cancer (PCa) is second most frequently diagnosed cancer at 15% of all male cancers and the sixth leading cause of cancer death in males worldwide. There is a need to identify novel therapeutic-based biomarkers or therapeutic strategies for metastatic prostate cancer. In large-scale transcriptome studies (e.g. DNA microarrays, RNA Seq) generate a lot of information on the levels of gene expression. The analysis of large amounts of expression data obtained in different tissues or different experimental conditions used to establish networks of relationships (e.g. co-expression) uniting groups of genes. A major challenge lies in the analysis of these expression systems, both topological level (eg overall structure of the network, identifying areas strongly connected), at the descriptive level (eg definition of metadata related to the experiences and samples). The method presented here builds a specific co-expression network to a disease, prostate cancer, by contextualizing a representative global network of all microarrays published for the human species. The analysis of this network of 6585 genes with 4 centrality measures are the degree centrality, the betweenness centrality, the closeness centrality and clustering coefficient identifies 506 genes of interest. In this study, we are particularly interested in genes coding for transcription factors like proteins (TF) or G protein-coupled receptors (GPCR). We thus find the genes already known to play an important role in the genesis and development of prostate cancer. The analysis was performed without any raw expression data in the prostate cancer indication. We Identified genes as AR, NKX3-1 and MYC already known to play a role in the development of prostate cancer. This Co-expression analysis was performed on 2012, currently among the 61 potential candidates, 20 are still unknown in PCa	Data poster	Fundamental Health
P_Da058	840	Matteo Manica, Roland Mathis and María Rodríguez Martínez	Matteo Manica	CoDON, a learning framework for linking genomics and transcriptomics data to protein expression	In the last two decades, experimental techniques for generating and quantifying high-throughput molecular data have provided unprecedented amounts of data describing different omics levels. However, this ever-increasing availability of information has often failed to translate into new biological insights or actionable clinical statements. The question of how to integrate disparate data types into realistic models of complex biological diseases like cancer remains one of the major challenges.In this work we propose CoDON, a new computational framework that exploits manifold learning techniques inspired by active deep learning research concepts, to learn complex interactions on the genomic and transcriptomic levels that influence protein expression. Such interactions can help us decipher complex molecular mechanisms underlying cancer onset and progression. CoDON uses a neural network architecture that learns a common representation in a reduced feature space through the usage of auto-encoders and an additive layer. This lower dimensional representation is used to estimate the proteomic profiles in a joint training procedure.We employ CoDON on TCGA publicly available RNASeq, CNV, and SNP arrays in order to predict protein patterns from RPPA proteomic arrays.The reduced representation learned by the model enables the deconvolution of highly non-linear molecular interactions in cancer and can be used as a molecular fingerprint to stratify patients. The multi-omics prediction of the protein profiles increases perturbations analysis capabilities, indeed CoDON can be used to investigate the impact of genomic and transcriptomic alterations on the protein level and explore possible targeted therapies.	Data poster	Fundamental
P_Da060	539	Michael J. Pesavento, Pranathi V. N. Vemuri, Caroline Miller, Jenny Folkesson and Megan Klimen	Michael J. Pesavento	Comparison of vascular networks from high resolution 3D whole organ microscopic analysis	Understanding hemodynamics in circulatory systems is a critical component to identifying pathophysiologic states in tissue. Significant progress has been made in vascular network imaging; resolution has increased for high volume methods (eg microCT and MRI), and volume has increased for high resolution methods (eg multi-photon and confocal microscopy). 3Scan's Knife Edge Scanning Microscope (KESM) spans the gap between high volume and high resolution imaging modalities.Bright field images of resin-embedded, whole-organs (brain and pancreas) were obtained from mice following systemic perfusion with India Ink. Images are taken with a resolution of 0.7 um per pixel in XY and a typical slice depth of 5 um in Z, enabling large-scale analysis and comparison of vascular networks of whole organs consisting of up to 5 TB of imaging data in 3D and a maximum physical volume of 50 x 50 x 20 mm. Vascular features are identified via parallelized vessel segmentation and vectorization methods.Comparison of vascular features within a single organ reveals significant differences between the area analyzed within target tissue, largely as a result of the fractal dimension of the vascular network. Comparison of vascular network features between organs yields significant differences between vascular networks that are commensurate with the function of the vascular network for that organ.Rapid throughput analysis of high volume vascular data provides an unprecedented ability to compare vascular features between different vascular networks, as well as identify pathological states within those networks.	Data poster	Biotechnology
P_Da061	545	Charles Labuzzetta, Margaret Antonio, Patricia Watson, Robert Wilson, Lauren Laboissonniere, Jeffrey Trimarchi, Baris Genc, P. Hande Özdinler, Dennis Watson and Paul Anderson	Charles Labuzzetta	Complementary Feature Selection from Alternative Splicing Events and Gene Expression for Phenotype Prediction	A central task of bioinformatics is to develop sensitive and specific means of providing medical prognoses from biomarker patterns. Common methods to predict phenotypes in RNA-Seq datasets utilize machine learning algorithms trained via gene expression. Isoforms, however, generated from alternative splicing, may provide a novel and complementary set of transcripts for phenotype prediction. In contrast to gene expression, the number of isoforms increases significantly due to numerous alternative splicing patterns, resulting in a prioritization problem for many machine learning algorithms. This study identifies the empirically optimal methods of transcript quantification, feature engineering, and filtering steps using phenotype prediction accuracy as a metric. At the same time, the complementary nature of gene and isoform data is analyzed and the feasibility of identifying isoforms as biomarker candidates is examined. Isoform features are complementary to gene features, providing non-redundant information and enhanced predictive power when prioritized and filtered. A univariate filtering algorithm, which selects up to the N highest ranking features for phenotype prediction is described and evaluated in this study. An empirical comparison of pipelines for isoform quantification is reported by performing cross-validation prediction tests with datasets from human non-small cell lung cancer (NSCLC) patients, human patients with chronic obstructive pulmonary disease (COPD), and amyotrophic lateral sclerosis (ALS) transgenic mice, each including samples of diseased and non-diseased phenotypes.	Data poster	Health
P_Da062	767	Kyoko Watanabe, Erdogan Taskesen and Danielle Posthuma	Kyoko Watanabe	Comprehensive functional annotation of GWAS risk loci and candidate gene selection	Genome-wide association study (GWAS) has been applied to a variety of human diseases and traits. As the number of samples is increasing dramatically, statistical power to detect phenotype associated genetic loci is now strong. However, given summary statistics of GWAS, it is challenging to explain underlying biological processes of phenotype due to the complexity to identify true causal SNPs and genes. Additionally, even though incorporation of external data is essential to narrow down to potential candidates which then need to be looked into further details, those resources are spread in different platforms. To overcome those problems, we have implemented the atomized pipeline which annotates a variety of functionality of SNPs within GWAS risk loci (such as deleteriousness and regulatory elements) to functionally map SNPs to genes. The pipeline takes summary statistics of GWAS and returns the list of risk loci, functional SNPs and candidate genes given user defined parameters such as thresholds of P-values, r2, MAF, tissue types and data sources. Results can be queried by SNPs, loci or genes to see detail annotations. Although the pipeline requires a number of parameters, one of the advantages is that it is possible to further filter results and users can easily download only essential information for them. The pipeline has another functionality which can query the list of genes to identify shared functions and co-expression patterns in different tissue types. In the post-GWAS era, this pipeline may play an important role to further understand biological mechanisms associated with phenotypes of interest.	Data poster	Fundamental
P_Da063	465	Byungwook Lee	Byungwook Lee	Construction of database server for Korean patented biological sequences	A recent report of the Korean Intellectual Property Office (KIPO) showed that the number of biological sequence-based patents is rapidly increasing in Korea. We present biological features of Korean patented sequences though bioinformatic analysis. We constructed a web server for Korean patented biological sequences and identified their function with public databases. Our analysis consists of two steps. The first step is a functional identification step in which the patented sequences were mapped into the Reference Sequence (RefSeq) database. The second is an association step in which the patented sequences were linked to genes, diseases, pathway, and biological functions. In this step, we used Entrez Gene, Online Mendelian Inheritance in Man (OMIM), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Ontology (GO) databases. The association between the biological functions and the patented sequences indicated that genes whose products act as hormones on defense responses in the extra-cellular environments were the most highly targeted for patenting.	Data poster	Biotechnology
P_Da064	664	Rudi L. van Bavel, E.J. Blom, Lian Wiggers-Perebolte, Rob Spee, Maarten L. Hekkelman, Remco M.P. van Poecke, Jan van Oeveren and Anker P. Sørensen	E.J. Blom	CropPedia – Integrated database and software interface for gene lead discovery and accelerated breeding	CropPedia is a knowledge platform for integration and visualization of genomics data to enable fast and effective marker development and lead gene discovery. As an in-house web-based software, it allows combining public and private data from multiple crops using public and proprietary tools. These tools include JBrowse for visualization of genome sequences and aligned features, MapViewer for genetic maps and QTLs, VariomicsViewer for SNP data and MongoDB en Solr for fast data storage and retrieval. Advanced features are added for tracking search history in workspaces, doing advanced querying and accumulating gene details in gene passports to assist molecular breeders, trait specialists and bioinformaticians to speed up their molecular breeding.	Data poster	Agro-Food
P_Da066	330	Daniela Borgmann, Serge Weis, Peter Strasser and Stephan Winkler	Daniela Borgmann	Dementia Classification and Recognition Based on Neuropathological, Haematological, and Genetic Data	Despite numerous advances in modern medical research, clinical diagnosis and correct classification of dementia types are still very challenging during a patient's life time, as a decent diagnosis of dementia can only be done by performing neuropathological brain examinations after the decease of the patient. Therefore, a majority of diseased patients is not correctly diagnosed in an early state or in the worst case at no time.We have developed an in-vivo classification system for dementia that combines data sources and relates dementia types to disease-related processes in the brain. In detail, the classification model is based on post-mortem data, namely microscopy images of brain slices of patients (currently used for the diagnosis and classification of dementia), haematological data from patients (blood samples), and genetic data of patients (SNPs). We use post-mortem data as training data for supervised machine learning algorithms and so identify relationships between these features and dementia classifications (which are known post-mortem). The so generated mathematical models will be applied on new data from living patients in order to assign a dementia type and state by only using data available at the patient's lifetime.In our study we analysed data of more than 200 patients suffering from Alzheimer's disease, Parkinson's disease, or Amyotrophic lateral sclerosis, and more than 100 control cases. Using this in-vivo classification system novel correlations between blood parameters, neuropathological features and the state of the disease are detected, and variable interaction networks between the different data collections are identified.	Data poster	Health
P_Da067	844	Aideen Roddy, Anna Jurek, Alexey Stupnikov, Paul O'Reilly, Peter Bankhead, Philip Dunne, David Gonzalez de Castro, Kevin Prise, Manuel Salto-Tellez and Darragh McArt	Aideen Roddy	Development of computational models to study mechanisms of tumour evolution for therapeutic vulnerabilities	Next-Generation Sequencing allows for the in-depth sequencing of genetic materials for the extraction of key aberrant drivers obtained in high throughput. Current analytic approaches in cancer research require sequencing data to be aligned prior to downstream analysis. However, with alternating pipelines required this over-simplifies the complex nature of the cancer landscape and potential therapeutic avenues.We aim to highlight the potential applications of alignment-free clustering in modern research. This approach, which is currently being explored in a phylogenetic context to successfully classify species, involves segmenting the sequencing data into features and obtaining a feature frequency profile for each sample before applying a distance metric (Sims et al. 2009). We aim to develop this concept for the application of sequencing data by applying filtering algorithms to remove non-significant data and creating low-dimensional views using self-organising maps.Thus far, we have applied our method to a cohort of multi-regional longitudinal glioma samples obtained through collaboration with the University of California. Our initial results, using Exome-seq data, show that this approach could be successfully used for clustering analysis in cancer research with the potential to further expand this use into other sequencing modalities. Furthermore, we aim to continue exploring this methodology with the potential to eventually combine multiple sequencing modalities in order to obtain a more accurate interpretation for a true patient passport.The abundance of successful applications of alignment-free analysis in sequencing has revealed the power of this approach showing promise for the future of tumour sequencing analysis in cancer research.	Data poster	Health
P_Da068	317	Jean-Fred Fontaine and Miguel A. Andrade-Navarro	Jean-Fred Fontaine	Disease enrichment analysis for gene sets based on co-occurrences in the literature	Candidate genes derived from high-throughput experiments such as RNA-seq are partly composed by poorly studied genes. Nevertheless, functional enrichment analysis methods can be used to characterize these gene sets with the following idea: if a concept is found more than expected in the annotations of several genes from the input gene set, then the gene set may be related to the function described by this concept. Available software tools offer such computation for various types of concepts such as Gene Ontology terms, protein domains, genomic location, or molecular pathways. Few tools offer this computation for diseases although this is a critical focus of the biomedical literature.In these tools, disease enrichment analysis is computed using gene-disease associations from experimental data on disease causing genes or related molecular pathways. As a significant amount of diseases are associated to few or no genes in such datasets, the tools often fail to return relevant results. This limitation could be addressed by using predicted gene-disease associations to increase the number of genes associated to each disease.We have predicted 40 thousand gene-disease associations from significant (FDR<5%) co-occurrences in biomedical literature from the PubMed database, involving 2214 diseases and 7597 human protein-coding genes. Benchmarks on 20 gene sets known to be associated to diseases show that this method outperforms or performs equally to existing ones in all cases. Contrary to existing methods, parameters can be tuned to increase precision or recall. A web interface and a web service are available at http://cbdm.uni-mainz.de/GeneSet2Disease.	Data poster	Fundamental
P_Da069	656	Amin Allahyar and Jeroen de Ridder	Amin Allahyar	Disease specific network with application in network based outcome prediction	In cancer outcome prediction, biological networks are used to aggregate functionally related genes with added discriminative power and biological relevance. However, recent studies revealed that comparable performance might be achieved using many different biological networks [1]. We aimed to investigate this issue by constructing a candidate synergistic network in which two genes are connected if their integration yields prediction beyond what is attainable individually. This is done by evaluating all pairwise combination of genes. A biological network may perhaps be useful in cancer outcome prediction if it signifies more connections between identified synergistic pairs compare to random pairs.In the next phase, we constructed a new classification problem in which topological properties of two genes (e.g. shortest path etc.) are considered as features and synergistic status between these genes are labels. Using this framework, apart from being able to combine evidences from multiple topological measures, we can exploit any arbitrary number of biological networks.We observed that none of considered biological networks sufficiently resemble the synergistic pairs. Next, we aimed to predict synergistic pairs using topological measures of several biological networks. Our extensive experiments showed that the synergistic pairs can be accurately identified (~85% AUC) using combination of topological measures. Remarkably, our framework identified efficacy of known topological measures (e.g. page rank) as well as several new ones (e.g. eigenvector centrality) that were not recognized before.Reference[1] C. Staiger, et al., Frontiers in Genetics, vol. 4, 2013.	Data poster	Health

P_Da070	612	Woong Na, Kijong Yi, Young Soo Song and Moon Hyang Park	Young Soo Song	Dissecting the Role of IgG Subclasses and Complements in Membranous Lupus Nephritis and Primary Membranous Nephropathy	Membranous lupus nephritis (MLN) and primary membranous nephropathy (PMN) are kidney diseases with similar morphology, but distinct etiologies, both affecting glomerulus with immune deposits. Immunoglobulins and complements, main components of the deposits, can be detected using immunofluorescence (IF) microscopy. IF staining patterns for IgG subclasses and complements are different between MLN and PMN, but comprehensive models explaining the complex staining patterns between two diseases were not presented. We investigated 148 cases of IF staining for IgG1, IgG2, IgG3, IgG4, C3, C4, and C1q of renal biopsies, among which MLN and PMN were 53 and 95 cases, respectively. IF staining results were semiquantitative evaluated from 0 to 3 according to the staining intensity of each marker. Principal component analysis and hierarchical clustering showed two diseases can be easily delineated. To investigate the dependence and independence of these markers, after dichotomizing the values into 0 or 1, we evaluated the changes of entropies or mutual information between MLN and PMN. Significant entropy changes were found for all markers except C3, but mutual information were not so in all pairs of the markers, implying the diseases directly influences the production of IgG subclasses and complements, and the interactions between IgG subclasses and complements are robust between two diseases. Interestingly, a first order Markov chain of IgG subclasses could be made according to the mutual information, predicting IgG subclasses were made in the order of IgG3, IgG2, IgG1, IgG4 temporally. Entropy analysis was useful in exploring a part of pathogenesis of MLN and PMN.	Data poster	Health
P_Da071	397	Muhammad Ammad-Ud-Din, Suleiman A.Khan, Disha Malani, Astrid Murumägi, Olli Kallioniemi, Tero Aittokallio and Samuel Kaski	Muhammad Ammad-Ud-Din	Drug response prediction by inferring pathway-response associations with Kernelized Bayesian Matrix Factorization	A key goal of computational personalized medicine is to systematically utilize genomic and other molecular features of samples to predict drug responses for a previously unseen sample. Such predictions are valuable for developing hypotheses for selecting therapies tailored for individual patients. This is especially valuable in oncology, where molecular and genetic heterogeneity of the cells has a major impact on the response. However, the prediction task is extremely challenging, raising the need for methods that can effectively model and predict drug responses. In this study, we propose a novel formulation of multi-task matrix factorization that allows selective data integration for predicting drug responses. To solve the modeling task, we extend the state-of-the-art kernelized Bayesian matrix factorization (KBMF) method with component-wise multiple kernel learning. In addition, our approach exploits the known pathway information in a novel and biologically meaningful fashion to learn the drug response associations. Our method quantitatively outperforms the state of the art on predicting drug responses in two publicly available cancer data sets as well as on a synthetic data set. In addition, we validated our model predictions with lab experiments using an in-house cancer cell line panel. We finally show the practical applicability of the proposed method by utilizing prior knowledge to infer pathway-drug response associations, opening up the opportunity for elucidating drug action mechanisms. We demonstrate that pathway-response associations can be learned by the proposed model for the well known EGFR and MEK inhibitors.	Data poster	Health
P_Da072	706	Lara Schneider, Daniel Stöckel, Tim Kehl, Andreas Gerasch, Michael Kaufmann, Oliver Kohlbacher, Andreas Keller and Hans-Peter Lenhof	Lara Schneider	DrugTargetInspector: An assistance tool for patient treatment stratification	One of the Hallmarks of Cancer is the acquisition of genome instability and mutations. In combination with high proliferation rates and failure of repair mechanisms, this leads to clonal evolution within a tumor, and hence to a high genotypic and phenotypic diversity. As a consequence, successful treatment of malignant tumors is still a grand challenge. Moreover, under selective pressure, e.g. caused by chemotherapy, resistant subpopulations may emerge that in turn can cause relapse. In order to minimize the risk of developing multi-drug resistant tumor cell populations, optimal (combination) therapies have to be determined on the basis of an in-depth characterization of the tumor's genetic and phenotypic makeup, a process that is an important aspect of stratified medicine and precision medicine. To this end, we present DrugTargetInspector (DTI), an interactive assistance tool for treatment stratification. DTI analyzes genomic, transcriptomic and proteomic datasets and provides information on deregulated drug targets, enriched biological pathways and deregulated subnetworks, as well as mutations and their potential effects on drugs, drugs targets, and genes of interest. Using DTI's powerful web-based toolsuite allows users to characterize the tumor under investigation based on patient-specific -omics datasets and to elucidate putative treatment options based on clinical decision guidelines, but also proposing additional points of intervention that might be neglected otherwise. DTI can be freely accessed at https://dti.bioinf.uni-sb.de.	Data poster	Health
P_Da073	553	Asta Laiho, Arfa Mehmood and Laura L. Elo	Asta Laiho	EBSEA: An Exon Based Strategy to Find Differentially Expressed Genes from RNA-seq Studies	A typical goal in RNA-seq studies is to identify differentially expressed genes between distinct sample groups. Conventionally the statistical testing is performed after the data has been summarized at the gene level. However, gene level summary values are prone to bias caused by a single or a relatively few exons with deviant values which are expected to occur, for instance, due to alternative splicing events. Relatively low abundance genes are also easily missed, despite showing systematic changes across their exons. As an alternative strategy, we demonstrate a method in which statistical testing at the exon level is performed prior to the summary of the results at the gene level.To systematically investigate the benefits of the proposed exon-based strategy, we considered two widely-used software packages that are conventionally applied to gene-level read counts (edgeR and limma). However, our testing approach can be combined with any method working on gene-level read count values. In our approach, statistical testing of each exon of a gene is first performed, prior to aggregating the results across the exons to produce gene level statistics. To present the advantage of the approach, we used two publicly available data sets with varying levels of heterogeneity.Our study shows how an exon-based strategy can significantly increase the sensitivity and specificity of the widely used differential expression methods for RNA-seq data over the conventional gene-based strategy. The approach has been implemented in a new R/Bioconductor package EBSEA.	Data poster	Fundamental
P_Da074	634	Witold Rudnicki, Paweł Tabaszewski, Szymon Migacz, Krzysztof Mnich and Andrzej Sułecki	Witold Rudnicki	Efficient Exhaustive Search for Synergistic Informative Variables	We present efficient GPU-based implementation of the algorithm for identification of informative variables in high-dimensional datasets. It performs an exhaustive search of all low-dimensional subspaces of the system in a reasonable time.To this end the variables are discretised using rank of object in given variable to assign the class. The models described with n-tuple of variables are built, n can be {2,3,4,5}. The exhaustive search is performed by generating all possible n-tuples. For each n-tuple several random discretisations are generated and the average information gain is collected for each variable. The variable V is deemed informative if there exist n-tuple of variables {V1,..,Vn-1} such, that adding variable V to the description of the system increases information about the decision variable in a statistically significant way.Algorithm is implemented both on CPU and on GPU. It is implemented as R module, and will be available also as a web server. It can be applied to datasets described with millions of variables containing hundreds of thousands of objects. The exhaustive search of the pairwise synergetic effects for the gene expression data for 1000 objects and 20 000 genes takes less than minute a single GPU, while the 3D search will take less than 24 hours. Even the 4D analysis can be performed within a week on a medium size computational cluster equipped with GPUs.Research was supported by the grant from the Polish NSC, grant UMO-2013/09/B/ST6/01550.	Data poster	Fundamental
P_Da075	352	Abdulrahman Azab	Boris Simovski	Enabling Docker Packaged Tools for HPC	Linux containers, with the build-once-run-anywhere approach, are becoming popular for software packaging and sharing among scientific communities, e.g. life sciences. Docker is the most popular and user friendly platform for running and managing Linux containers. This is proven by the fact that vast majority of containerized tools are packaged as Docker images.A demanding functionality is to enable running Docker containers inside HPC job scripts for researcher to make use of the flexibility offered by containers in their real-life computational and data intensive jobs. The main two questions before implementing such functionality are: how to securely run Docker containers within cluster jobs? and how to limit the resource usage of a Docker job to the borders defined by the HPC queuing system?This position paper presents Socker, a wrapper for running Docker containers on SLURM. Socker enforces running containers within SLURM jobs as the submitting user, as well as enforcing the inclusion of containers in the cgroups assigned by SLURM to the parent jobs. The implementation of Socker is tested on Abel, the HPC cluster at the University of Oslo. The use case is ChIP-Seq workflow with Dockerized tools running on the cluster. We implemented parallelization using MPI for sequence alignment. Socker is proven to be secure and simple to use together with introducing no additional overhead.	Data poster	Fundamental
P_Da076	587	Veronika Weyer-Elberich, Yasmin Abassi, Detlef Schuppan, Ernesto Bockamp and Harald Binder	Veronika Weyer-Elberich	Exploring cell type deconvolution by a weighted regression approach for the resulting groups	Recent gene expression-based deconvolution approaches allow disentangling the different cell types present in tumor samples. This is not only useful for reducing heterogeneity, but the abundance or lack of certain immune cell types, may be biologically meaningful. We consider the lack or subtype variance of T cells for different tumor entity samples, which has been associated with shorter survival. We propose a new algorithm for dividing different cancer patients into two groups according to lack of T cells or other immune cell variance. Specifically, we extract cell-regulated genes that are associated with regulation in other immune cell types and divide the patients into two groups according to these genes. The uncertainty of this partition is examined using a stratified weighted Cox regression approach based on componentwise likelihood-based boosting that provides a prognostic gene signature for patients with a lack of different immune cells in tumor samples. When developing this subgroup signature for some information from the other group is utilized by weighted partial log-likelihood. The effects of different weights and weighting schemes are investigated by resampling inclusion frequencies of genes into the prognostic signatures. Additionally, different cut offs for dividing patients into the two subgroups will be investigated. Applying this to cancer gene expression data, model stability is seen to be improved with intermediate weights. Furthermore, changes in gene selection when changing the weights are seen to reflect the underlying biology. Thus, combination of a deconvolution algorithm with a weighted regression approach is an useful and versatile new bioinformatic tool.	Data poster	Health
P_Da077	623	Alfonso Muñoz-Pomer Fuentes, Wojciech Bażant, Elisabet Barrera, Melissa Burke, Jana Eliasova, Nuno Fonseca, Laura Huerta, Anja Fullgrabe, Maria Keays, Satu Koskinen, Irene Papatheodorou, Amy Tang, Robert Petryszak and Alvis Brazma	Alfonso Muñoz-Pomer Fuentes	Expression Atlas: Functional Genomics Resource at EMBL-EBI	Expression Atlas (http://www.ebi.ac.uk/gxa) contains pre-analyzed RNA-sequencing and expression microarray data for querying gene expression across tissues, cell types, developmental stages and many other experimental conditions, in over 35 organisms including metazoans and plants. Queries can be either in a baseline context, e.g. find genes expressed in the macaque brain, or in a differential context, e.g. find genes that are up or down-regulated in response to auxin in Arabidopsis. All datasets are manually curated to a high standard by in-house curators and processed using standardised analysis methods.As of June 2016, Expression Atlas consists of 2905 datasets, including 95 baseline experiments. All data in Expression Atlas are free to browse, download, reuse and are selected from the ArrayExpress archive of functional genomics data at EMBL-EBI. It is possible to search and download datasets in Expression Atlas into R for further analysis, and we now provide a REST API for access to thousands of pre-analysed RNA-sequencing datasets.Gene expression is shown through heatmaps for one or more genes. Groups of genes belonging to a Gene Ontology category (GO) or Reactome pathway can be queried directly using the GO or Reactome identifier. Latest features allow the exploration of gene co-expression; based on genes with similar expression profiles across tissues, cell lines or other conditions within an experiment. In addition we now allow users to input their gene lists of interest and test the statistical significance of their overlap with available experiments in Expression Atlas through a REST API.	Data poster	Health
P_Da078	808	Peter Walgemoed and Bert Eussen	Peter Walgemoed	Genomic data curation by design	Sharing genomic data globally for all stakeholders from creation to interpretation isa major challenge. Solutions are being developed at the institutional level. To support curation, we have developed a concept where data is tagged from the moment of creation. and can be shared globally. Curation starts with raw data in a lab or with the clinical work-up. The lab is a data collection point but it is driven by its clients (researchers and clinicians). These clients have the responsibility to manage the privacy for their clients (citizens). Therefore data curation is on behalf of the citizen. All procedures and lab services are documented in a trusted, authoring document system (TrustDocA).It will be challenging for citizens to curate their own data. It is likely to grow exponentially and it will become very complex to handle phenotypic, laboratory, treatment, municipal and personal health and lifestyle data. Therefore a trusted and transparent co-operation between institutes is required to curate the data on the citizen's behalf. DATA co-operative not only includes storage and preservation but also creates value by using the data as much as possible.Transparent data collection systems are essential for consortia wanting to share data on behalf of their clients/ citizens as part of a FAIR data policy. Governance should be by design and citizen informed consent implies that a data copy is curated by the DATA co-operative and should be available for future generations.	Data poster	Health
P_Da079	536	Fiona Nielsen and Nadezda Kovalevskaya	Manuel Corpas	Genomic data projects around the world: how to find data for your research	Access to raw experimental research data and data reuse is a common hurdle in scientific research. Despite the mounting requirements from funding agencies that the raw data is deposited as soon as (or even before) the paper is published, multiple factors often prevent data from being accessed and reused by other researchers.The situation with human genomic data is even more dramatic, since, on the one hand, it is probably the most important data to share - it lies at the heart of efforts to combat major health issues such as cancer, genetic diseases, and genetic predispositions for complex diseases like heart disease and diabetes. On the other hand, since human genomic data contains sensitive and personal information, it is often exempt from data sharing requirements.We found out that, on average, researchers use 4-5 genomic data repositories on a regular basis. At the same time, there are many more sources of data available that are often unknown to researchers. We have addressed the most pressing problem for public genomic data, that of data discoverability, by indexing worldwide resources for genomic research data on an online platform (repositive.io) providing a single point of entry to find and access available genomic research data.In this work, we present the overview of genomic data sources around the world and discuss the potential solutions for improving ethical and efficient data sharing.	Data poster	Biotechnology
P_Da080	464	Kedar Tatwawadi, Mikel Hernaez, Idoia Ochoa and Tsachy Weissman	Kedar Tatwawadi	GTRAC: Fast retrieval from compressed collections of genomic variants	The dramatic decrease in the cost of sequencing has resulted in the generation of hugeamounts of genomic data, as evidenced by projects such as the UK10K and the Million Veteran Project (MVP), with the number of sequenced genomes ranging in the order of 10K to 1M. Due to the large redundancies among genomic sequences of individuals from the same species, most of the medical research deals with the variants in the sequences as compared with a reference sequence, rather than with the complete genomic sequences. Consequently, millions of genomes represented as variants are stored in databases. These databases are constantly updated and queried to extract information such as the common variants among individuals or groups of individuals. Previous algorithms for compression of this type of databases lack efficient random access capabilities, rendering querying the database for particular variants and/or individuals extremely inefficient, to the point where compression is often relinquished altogether.We present a new algorithm for this task, called GTRAC, that achieves significant compressionratios while allowing fast random access over the compressed database. For example, GTRAC is able to compress a H. Sapiens dataset containing 1092 samples in 1.1GB (compression ratio of 160), while allowing for decompression of specific samples in less than a second and decompression of specific variants in 17ms. GTRAC uses and adapts techniques from information theory, such as a specialized Lempel-Ziv compressor, and tailored succinct data structures.	Data poster	Biotechnology
P_Da081	510	Valentin Voillet, Phillipe Besse, Laurence Liaubet, Magali San Cristobal and Ignacio Gonzalez	Valentin Voillet	Handling Missing Rows in Multi-Omics Data Integration: Multiple Imputation in Multiple Factor Analysis Framework	In omics data integration studies, it is common that some individuals are not present in all data tables. Missing row values are challenging to deal with because most statistical methods cannot be directly applied to incomplete datasets. To overcome this issue, we propose a multiple imputation (MI) approach in a multivariate framework. In this study, we focus on multiple factor analysis (MFA). MI involves filling the missing rows with plausible values, resulting in m completed datasets. MFA is then applied to each completed dataset leading to m different component configurations. Finally, the m configurations are combined to yield one consensus solution.We assessed the performance of our method, named MI-MFA, on two real omics datasets. Incomplete artificial datasets were created from these data with different patterns of missingness. The MI-MFA results were compared to two other approaches, regularized iterative MFA (RI-MFA) and mean variable imputation (MVI-MFA). For each component configuration resulting from these three strategies, we determined the suitability of the component solution against the true MFA configuration obtained from the original data. The overall results showed that MI-MFA outperformed the RI-MFA and MVI-MFA approaches in nearly all settings of missingness.Two approaches, confidence ellipses and convex hulls, to visualize and estimate the uncertainly due to missing values were also described. We showed how the areas of ellipses and convex hulls increased as variability was added to the data. These graphical representations provide scientists with considerable guidance in order to evaluate the reliability of the results.	Data poster	Agro-Food
P_Da082	400	ChantrioInt-Andreas Kapourani and Guido Sanguinetti	ChantrioInt-Andreas Kapourani	Higher order methylation features for clustering and prediction in epigenomic studies	DNA methylation is an intensely studied epigenetic mark, yet its functional role is incompletely understood. Attempts to quantitatively associate average DNA methylation to gene expression yield poor correlations outside of the well-understood methylation-switch at CpG islands. Here we use probabilistic machine learning to extract higher order features associated with the methylation profile across a defined region. These features quantitate precisely notions of shape of a methylation profile, capturing spatial correlations in DNA methylation across genomic regions. Using these higher order features across promoter-proximal regions, we are able to construct a powerful machine learning predictor of gene expression, significantly improving upon the predictive power of average DNA methylation levels. Furthermore, we can use higher order features to cluster promoter-proximal regions, showing that five major patterns of methylation occur at promoters across different cell lines, and we provide evidence that methylation beyond CpG islands may be related to regulation of gene expression. Our results support previous reports of a functional role of spatial correlations in methylation patterns, and provide a mean to quantitate such features for downstream analyses.	Data poster	Fundamental

P_Da083	795	Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan Yevshin, Haitham Ashoor, Wail Ba-Alawi, Artem S. Kasianov, Yulia Medvedeva,Vladimir Bajic, Fedor Kolpakov and Vsevolod Makeev	Vsevolod Makeev	HOCOMOCO: data integration for building collection of reliable transcription factor binding sites models	The precise locations of transcription factor binding sites (TFBSs) in DNA are needed for solving different problems in functional genomics, e.g. for studying consequences of mutations or polymorphisms. Currently, ChIP-Seq data is the principal data source of TF in vivo binding. Yet, the most variants of this technique do not provide the exact TFBS positions that sometimes can be wrongly coming from DNA bound complexes formed of the test protein and other DNA binding proteins. In vitro technologies, such as HT-SELEX, warrant direct binding, but tend to reveal only a subset of genomic TF binding DNA sites. Currently, the precise location of binding sites can be obtained only with the help of computational methods using TFBS models.We developed a pipeline that integrates multiple ChIP-Seq and HT-SELEX datasets, and validates the resulting models on in vivo data. We used data from 1690 human and mouse publicly available ChIP-Seq experiments, performed in house read mapping and peak calling, combined them with 542 HT-SELEX datasets, and supplied to ChIPMunk motif discovery tools to obtain position weight matrices (PWMs). The resulting TFBS models were subject of manual curation. We constructed the largest up to date collection of PWM models for dozens of human and mouse TFs, and, similarly advanced dinucleotide PWM models for dozens of TFs To facilitate practical applications, all models were linked to gene and protein databases (Entrez Gene, HGNC, UniProt, FANTOM SSTAR, GeneCards, TFclass) and accompanied by pre-computed thresholds for DNA screening. The collection is available at http://hocomoco.autosome.ru.	Data poster	Fundamental
P_Da084	277	Nick Juty, Sarala Wimalaratne, Nicolas Le Novère and Henning Hermjakob	Nick Juty	Identifiers.org: services towards interoperability	The Identifiers.org resolver is purpose built to support the use of HTTP URIs directly for identification and cross-referencing of Life Science data. These URIs can be incorporated in datasets, facilitate usability by tools (for processing and display), and are resolvable by the end user. Moreover, these URIs are free and provide unique, perennial and location-independent identifiers.The information used to provide identifiers.org services is stored in a curated registry of data collections (corresponding to controlled vocabularies or databases). This information includes identifier patterns that are used by the collection, current and legacy physical locations (access URLs) and a record of individual resource uptimes. Consequently we are able to provide services such identifier validation, interconversion services between access URLs and alternative URI schemes, and redirection services to reliable physical locations. We describe these services, as well as our most recent developments.	Data poster	Fundamental
P_Da085	751	Sebastien Tourlet, Frederic Scaerou, Aurelie Martin, Arunthi Thiagalingam, Isabelle Paty, Laurent Naudin and Philip Harris	Sebastien Tourlet	iFiT: an integrative Bioinformatics platform for biomarker and target discovery. A case study in neuroendocrine tumors.	iFiT (Ipsen Focused-on-new biological enTities and biomarkers) is a Bioinformatics platform integrating systems biology functionalities together with semantic & logic-based artificial intelligence within a high-scale computing environment.Key applications are the discovery of potential therapeutic targets as well as the identification of patient stratification candidate biomarkers.Given the limited OMICs characterization of neuroendocrine tumors, the identification of driver genes and pathways is challenging: To help circumvent this paucity of molecular information, iFiT was built on the postulate that co-expressed genes participate in the same biological processes. Furthermore, we fed the platform with curated heterogeneous datasets, pre-clinical and clinical, including molecular and phenotypic information.We focused our search on drugable GPCRs and microRNAs involved in mechanisms such as pancreas islet cells lineage, differentiation, multiplication and hormone secretion. As a result, we identified 42 GPCRs and 10 microRNAs, including well-known NETs-associated genes such as SSTR2 and DRD2. iFiT predicted the driver role of SSTR2 in both proliferation and secretion before the release of the CLARINET study (ESMO 2013). Remarkably, 90% of candidate genes were validated on tumor tissues from 40 GEP-NET patients.In conclusion, iFiT achieves an excellent detection rate, and is proving suitable to uncover hidden information and mine translational knowledge in NET	Data poster	Fundamental Health
P_Da086	340	Sean Robinson, Jaakko Nevalainen, Guillaume Pinna, Anna Campalans, J. Pablo Radicella and Laurent Guyon	Sean Robinson	Incorporating interaction networks into the determination of gene hits with Markov random fields	Associated with a cellular function of interest, high-throughput genomic experiments are used to score individual genes and identify 'hits' (genes with significant scores) likely to be worthwhile targets for further analysis. However, there are many known issues with such an approach. For example, in RNA interference experiments 'off-target effects' and 'siRNA efficiency' are known to lead to false positive and false negative gene hit identification respectively.We present a gene scoring method based on a Markov random field (MRF) to incorporate protein-protein interaction (PPI) networks into the determination of gene hits. We assume that in principle, genes with interacting proteins are associated to the extent that they are expected to exhibit similar behaviour in the experiment. In this way we aim to decrease such false positive and false negative hit results.Two major advantages of the presented MRF method against current methods such as Knode (SANTA) and BioNet are that it easily allows for multivariate scores on the genes as well as multiple hit classes beyond binary 'hit'/'non-hit' corresponding to both positive and negative phenotypes. We show in simulated as well as real data applications that by incorporating the additional PPI network information using an MRF, gene hits are able to be more accurately identified leading to a more effective identification of genes for further analysis.	Data poster	Fundamental
P_Da087	321	Morihiro Hayashida and Hitoshi Koyano	Morihiro Hayashida	Integer linear programming approach to median and center strings for a probability distribution on a set of strings	For a dataset composed of numbers or numerical vectors, a mean is themost fundamental measure for capturing the center of the data.For a dataset of strings, however, a mean cannot be defined,and median and center strings instead of a mean are oftenused as a measure of the center.In contrast to calculating a mean of numerical data, constructingmedian and center strings is not easy, and noalgorithm has been found that is guaranteed to construct exact solutions ofcenter strings.In this study, we first generalize the definitions of median andcenter strings into those of a probability distribution on a set of all strings composed of letters in a given alphabet.This generalization corresponds to that of a mean of numerical datainto an expected value of a probability distribution on a set ofnumbers or numerical vectors.Next, we develop methods for constructing exact solutions of medianand center strings for a probability distribution on a set of strings,applying integer linear programming.These methods are improved into faster ones by using the triangleinequality on the Levenshtein distance in the case where a set ofstrings is a metric space with the Levenshtein distance.Lastly, we perform simulation experiments to examine the usefulness ofour proposed methods in practical applications.	Data poster	Fundamental
P_Da088	372	Vitor C. Piro and Bernhard Y. Renard	Vitor C. Piro	Integrating metagenome analysis tools to improve taxonomic profiling and organism identification	A large and increasing number of metagenomics analysis tools is presently available aiming to characterize environmental samples. Reference-based approaches, the ones that rely on previous genome sequences, are commonly used for this task. They can be classified in two main groups: taxonomic profiling and binning tools. Tools available among these two categories make use of several techniques, e.g. read mapping, k-mer alignment and composition analysis. Variations on the construction of the databases are also common. All this variation creates a complicated scenario to researches to decide which methods to use. Different tools provide good results in different scenarios. We propose an automated method to merge community profiles from several tools, providing a single, reliable and improved outcome. Our method uses the co-occurrence of organisms reported from different methods as the main feature to lead to better community profiling. The intersection of all reported organisms from all tools is analyzed and weighted by the number of occurrences, normalized relative abundances, among other features. By separating those organisms in classes based on features it is possible to apply a guided cutoff and a better selection, keeping the most of true identifications. Merging binning with profiling tools allows us to takes advantage of distinct techniques and improves the final result. In a controlled case, we show that the integrated profiles can overcome the best single profile. Using the same input data, it provides more reliable results with the presence of each organism being supported by a set of tools and metrics.	Data poster	EcosystemsHea lth
P_Da089	833	Jun Cheng, Kerstin Maier, Fabien Bonneau, Žiga Avsec, Patrick Cramer and Julien Gagneur	Jun Cheng	Integrative analysis of mRNA half-life cis-regulatory elements	The stability of messenger RNA (mRNA) is one of the major determinants of gene expression. Although a wealth of mechanisms regulating RNA stability has been described, little is known about how much mRNA half-life is directly encoded in its sequence.Here, using genome-wide mRNA half-life data, we built quantitative models that, for the first time, explain most of the between-gene half-life variation based on mRNA sequence alone for two eukaryotic genomes, Saccharomyces cerevisiae and Schizosaccharomyces pombe. The models integrate known functional cis-regulatory elements, identify novel ones, and quantify their contribution at single-nucleotide resolution. In the well-studied S. cerevisiae, we identified a novel conserved motif that affects around 10% of the protein coding genes and exhibits positional preference within the 3'UTR. We showed that three translation-associated elements are collectively the major determinants of mRNA half-life: codon usage, start codon context and stop codon context. We further examined the dependencies of cis-regulatory elements with respect to mRNA degradation pathways using genome-wide mRNA half-life of 35 S.cerevisiae strains in which different genes mediating mRNA stability were knocked out. We found that the effects of translation-associated elements on mRNA half-lives decrease significantly upon knockout of Ccr4, Not4, Xrn1 and Dhh1. This suggests that the coupling between mRNA degradation and translation depends on the canonical mRNA degradation pathways.Altogether, our results provide a comprehensive and quantitative delineation of mRNA stability cis-regulation and can serve as a scaffold for studying the functionality of known elements as well as for identifying novel ones.	Data poster	Fundamental
P_Da090	860	Yongsoo Kim, Wilbert Zwart, Lodewyk Wessels and Daniel Vis	Yongsoo Kim	Integrative soft multi-way clustering of pan-cancer cell line data to identify context-specific regulation in cancer genome	Regulation in biological systems is highly complex and context-specific. For example, the effect of inhibiting a gene product may depend on the biological context. Thus, it is important to correctly characterize biological contexts in cancer to predict treatment response accurately. We can exploit multi-omics data of tumors and cell lines, such as GDSC 1000 data resource, to better define the contexts and how they modulate response. Here we propose an integrative analysis framework for multi-way multi-omics data based on non-negative PARAFAC (PARAllel FACtors analysis), which is a multi-way extension of non-negative matrix factorization (NMF). Multiple data layers, including mutation, copy number aberration and expression profiles in cancer-related genes are integrated. The obtained factor matrices are used to derive multi-way soft clusters of sets of genes, cell lines and data types, while overlap is allowed between the clusters (i.e. a gene can be involved in more than one clusters). Based on the framework, multi-way clusters that reflect cancer-related contexts are identified from a pan-cancer multi-omics cell line data set (GDSC1000). We interpreted the multi-way clusters in gene oriented, cell line oriented and data type oriented manner. We find that 1) although they are structurally incomparable, there is concordance between the data types, such as copy number loss and decrease in gene expression; 2) some multi-way clusters are specific to one tissue type while others are shared between two or more tissue types; and 3) genes involved in key cancer-related pathways are associated with multiple clusters, indicating frequent aberration of the pathways.	Data poster	Fundamental
P_Da092	595	Ben C Stöver, Sarah Wiechers and Kai F Müller	Ben C Stöver	JPhyloIO: A Java library for event-based reading and writing of different alignment and tree formats through one common interface	Today a variety of alignment and tree file formats exist, some of which well-established but limited in their data model, others more recently proposed offer advanced future-orientated features for metadata representation. Most phylogenetic and other bioinformatic software currently only supports one or few different formats, while supporting many widely-used standards simultaneously would be desirable to achieve optimal interoperability and prevent data loss by external conversions.We developed JPhyloIO, which allows reading and writing of alignment and tree formats (NeXML, PhyloXML, Nexus, Newick FASTA, Phylip, MEGA, XTG, PDE) using a common interface. It is the only currently available Java-library that generalizes between the different data and metadata concepts of all formats, while still allowing access to their individual features. By simply implementing a single JPhyloIO based reader and writer, application developers can easily support all formats in one step and the event-based architecture allows the library to be combined with any application business model design, while still being memory efficient for large datasets.We provide JPhyloIO as a service to the scientific community, which will benefit from simplified development of software that supports various standards simultaneously. Our aims are to increase the interoperability between different (phylogenetic) software tools and to foster usage of more recently proposed formats providing a powerful metadata concept. It is currently integrated in a number of applications and is fully interoperable with our Java-library LibrAlign, which offers powerful components for multiple sequence alignments and attached raw and metadata.Download and documentation: http://bioinfweb.info/JPhyloIO/ .	Data poster	Fundamental
P_Da093	782	Mira Valkonen, Matti Nykter, Leena Latonen and Pekka Ruusuvuori	Mira Valkonen	Learning based detection of early neoplastic changes in histological images	Digital pathology has been rapidly expanding into a routine practice, which has enabled the development of image analysis tools for quantification of histological images. Prostatic intraepithelial neoplasia (PIN) represents premalignant tissue involving epithelial growth confined in the lumen of prostatic acini. To understand oncogenesis in the human prostate, we studied early neoplastic changes in mouse PIN (mPIN) confined in prostate. We implemented an image analysis pipeline for describing early morphological changes in hematoxylin and eosin stained histological images. The model is based on manually engineered features and supervised learning with random forest model. For training, we used a set of mPIN lesions of abnormal epithelial cell growth and glands of normal tissue segmented by an expert. The extracted features include 102 local descriptors related to tissue texture and spatial arrangement and distribution of nuclei. These extracted features provide a numerical representation of a tissue sample and were used to computationally learn a discriminative model using machine learning. The implemented random forest model is an ensemble of 50 classification trees and it uses bootstrap aggregation to improve stability and accuracy. Leave-one-out cross-validation (LOOCV) was used to evaluate the performance of our random forest model. The classification model was able to discriminate normal tissue segments from early mPIN lesions and also describe the spatial heterogeneity of the tissue samples. The model can be easily interpreted and used to assess the contribution of individual features. This feature significance provides information about differences in the histology between normal glands and early neoplastic lesions.	Data poster	Biotechnology
P_Da094	469	Ryohei Suzuki, Daisuke Komura and Shumpei Ishikawa	Ryohei Suzuki	Learning High-level Features of Pathology Images Using Multi-Resolution Convolutional Auto-Encoders	Recent developments of machine learning techniques, especially deep neural network-based approaches, have enabled unsupervised learning of high-level features from images.Trained network is itself useful for providing features to supervised algorithms (e.g., support vector machine), and also known to improve the efficiency of supervised learning of a network with the same topology (pre-training).Pathology images are important target of machine learning with crucial applications such as decision support for medical diagnosis.Although, their extremely high-resolution nature makes it difficult to naïvely apply existing learning techniques to them.To tackle this problem, we present a novel unsupervised learning framework called multi-resolution convolutional auto-encoder.It is based on the idea of stacked convolutional auto-encoder[1] trained to reconstruct the input image as the output, but notable for taking sets of overlapping image patches of different physical scales as the input.The proposed network consists of parallel stacks of convolutional layers for different image scales, and a fully-connected ordinary auto-encoder on the top of the all convolution stacks to integrate the features from all scales.After greedy layer-wise training and whole-network training by error back-propagation, the network learns correlated features across diverse range of sizes (i.e., from cellular to histological differentiation).We show the accuracy of discrimination task between cancer and normal cells using the trained network compared with a set of independently trained convolutional networks without integration layer.References1. Masci, J., Meier, U., Cireçan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction, ICANN 2011, Springer (2011)	Data poster	Fundamental
P_Da095	827	Neetika Nath, Christian Klose, Mathias Gerl, Michal A. Surma, Kai Simons and Lars Kaderali	Neetika Nath	Lipoinformatics – machine learning approach to study lipid profiles	Lipids are the highly diverse class of molecules that are structural components of biological membranes and function as energy reserves and signalling molecules. Within the metabolomics field, shotgun lipidomics, providing absolute quantification and high reproducibility is perfectly suited for bioinformatics approaches to guide the biotechnologies to improve human health.The objective of this study is to develop a robust bioinformatics approach to identify lipid diagnostic biomarkers in human plasma that support the classification of subjects with high or low body mass index (BMI). The second objective of this study is to compare different normalization strategies for lipoidomic data of 326 human subjects with high (BMI > 30) or low (BMI < 25) BMI. We applied a random forest method implemented in varSelRF (R package) executing 1000 bootstrap samples. This yielded the most important features distinguishing high and low BMI. The resulting set of discriminating lipids is selected by the backward stepwise elimination of features with smallest cross-validation error.In our analysis we found no significant differences between normalizations by total lipid content or lipid class. The models were equally good with accuracies close to 0.75 and sensitivities and specificities at 0.72 and 0.75, respectively. Our results suggest that if using random forest for the analysis, the focus of the analysis should be to determine the important features.	Data poster	Biotechnology Health
P_Da096	639	Borong Shao and Tim Conrad	Borong Shao	Lung Cancer Prognosis Classification - the Effect of Data Types, Feature Transformation, Classifiers and Threshold	Biomarker discovery has evolved from analyzing single data type to exploring multiple -Omics data types as well as biological networks. The quality of discovered biomarkers varies among studies as they applied different data integration approaches such as building models on merged data, integrating models built from individual types of data, and utilizing biological networks to transform original features to subnetwork features. We obtained gene, miRNA, and protein expression data and patient prognosis data of lung adenocarcinoma from The Cancer Genome Atlas and compared the predictive capability of these data types by ranking corresponding features and using increasing number of features to build prognosis classifiers. We also mapped gene expression and miRNA expression data to epithelial mesenchymal transition network and transformed original features to 3 nodes subnetwork features, which were then used to build classifiers. In addition, we evaluated the average predictive capability of data as the prognosis threshold varies. Experimental results showed that using the same number of features clinical data obtained the highest classification accuracy while gene expression data obtained the lowest accuracy. When applying correlation feature ranking method together with support vector machine classifier protein data obtained higher prediction accuracy than miRNA data. Regarding feature transformation, we observed that depending on the number of features included, transformed features could achieve higher prediction accuracy than original features. Last but not least, the predictive capability of different types of data changed as prognosis threshold varied. Certain threshold was hard for most of the data types to predict.	Data poster	Health

P_Da098	610	Fanny Georgi, Vardan Andriasyan, Artur Yakimovich, Robert Witte and Urs Greber	Fanny Georgi	MorphoSphere: A deep learning framework to score cancer cell proliferation and oncolytic virus efficacy in spheroid models	Cancer involves uncontrolled cell proliferation eventually leading to life-threatening conditions. Spheroids are self-assembled cell aggregates, mimicking organotypic tissues at micro-scale. They provide significant biological complexity and are used to bridge the gap between single cell studies and animal models. Spheroids respond to cues from their environment in a way that cannot be studied with monolayers of cultured cells. Spheroids can be used to ask questions, such as how oncolytic virus infection affects spheroid integrity and growth. Manifold natural and engineered viruses are known to kill cancer cells by lysis.Here, we introduce a simplified tumor model to understand the parameters controlling oncolytic efficacy of viruses in tumor tissue. We present a platform for high-throughput screening of scaffold-free spheroids inoculated with different viruses. We employ high-throughput live cell imaging and automated image analysis, in conjunction with a newly developed automated deep learning image quantification framework, called MorphoSphere. MorphoSphere monitors spheroid dynamics by measuring morphological and textural features. We employ convolutional neural network-based approaches to automatically classify spheroidicity and viability of cell aggregates. We showcase the anti-tumor potential of viruses in spheroids featuring diverse tumor characteristics. By combining this approach with quasi-tomographic light-sheet microscopy and fully replicating reporter viruses, we correlate the macroscopic tumor killing ability of different viruses with the underlying microscopic phenotype of virus spreading between cells. We find that the efficacy of spheroid killing tightly correlates with the ability of the oncolytic virus to rapidly and deeply spread within the tumor model tissue.	Data poster	Health
P_Da099	448	Dalia Cohn-Alperovich, Alona Rabner,Ilona Kifer,Yael Mandel-Gutfreund and Zohar Yakhini	Dalia Cohn-Alperovich	Mutual enrichment in aggregated ranked lists with applications to gene expression regulation	It is often the case in biological measurement data that results are given as a ranked list of quantities – for example differential expression (DE) of genes as inferred from microarrays or RNA-seq. Recent years brought considerable progress in statistical tools for enrichment analysis in ranked lists. Several tools are now available that allow users to break the fixed set paradigm in assessing statistical enrichment of sets of genes. Continuing with the example, these tools identify factors that may be associated with measured differential expression.A drawback of existing tools is their focus on identifying single factors associated with the observed or measured ranks, failing to address relationships between these factors. For example – a scenario in which genes targeted by multiple miRNAs play a central role in the DE signal but the effect of each single miRNA is too subtle to be detected, as shown in our results.We propose statistical and algorithmic approaches for selecting a sub-collection of factors that can be aggregated into one ranked list that is heuristically most associated with an input ranked list (pivot). We examine performance on simulated data and apply our approach to cancer datasets. We find small sub-collections of miRNA that are statistically associated with gene DE in several types of cancer, suggesting miRNA cooperativity in driving disease related processes. Many of our findings are consistent with known roles of miRNAs in cancer, while others suggest previously unknown roles for certain miRNAs.	Data poster	Fundamental
P_Da100	687	Perla Aurora Troncoso Rey and Wiktor Jurkowski	Perla Aurora Troncoso Rey	Network assisted combined analysis of transcriptomics and metabolomics data	In recent years, the use of high-throughput experiments has become more popular and accessible, increasing the number of studies that are now looking at several aspects of a biological system (e.g. gene regulation, metabolism), typically interrogating and analysing each aspect (i.e. -omics data) independently. However, it is beneficial to use omics datasets in a combined analysis as it could uncover results which would not appear when only using a single omics type.In this work we look at the problem of omics data integration that makes use of biological knowledge as priors in multivariate statistical models. We start with a penalised logistic regression approach for gene selection. This approach is used to analyse transcriptomics data to find the subset of genes that are potentially influential to separate two conditions (e.g. healthy vs disease). This model uses a protein-protein interaction network (represented as an undirected graph) as prior knowledge to identify groups of connected elements that collectively change between the conditions. We study the network's effect for gene selection by testing with networks compiled from different sources and with different topological properties.Subsequently, we explore the challenges of multi-omics integration. We modify the logistic regression model for the combined analysis of transcriptomics and metabolomics data, using protein-protein and metabolic networks as priors. Using metabolic networks poses a challenge due to their more complex interactions (typically represented as directed graphs). Finally, we compare results to corroborate the hypothesis that a combined analysis provides better insight when studying a condition.	Data poster	Health
P_Da101	312	Susanne Schaller, Johannes Weinberger, Sandra Mayr, Thomas Stuettler, Peter Lackner andStephan Winkler	Susanne Schaller	New Developments in ImmunExplorer: From NGS Data Over Machine Learning To Health State Prediction	The human adaptive immune system, represented mainly by the B and T cells and their receptors, plays an essential role in the recognition of potential pathogens such as microorganisms, parasites, and viruses. Knowing the immune repertoire status of individuals is of high importance in basic and medical research, transplantation medicine as well as in diagnosis and treatment of several severe diseases. In the past few years, new high-throughput sequencing technologies emerged, which allow a rapid identification of antibody and T cell receptor gene sequences. Therefore, to properly analyze NGS data in the context of the immune repertoire an immunoinformatics pipeline is required.Here we show a pipeline to analyze NGS data in order to predict the health state of the immune repertoire using the software ImmunExplorer (IMEX). IMEX is a software framework with multiple features, specifically designed for immune repertoire analysis including statistical evaluations, primer efficiency, clonality, diversity, V-(D)-J, or classification analysis. A wrapper for MiXCR has been designed and developed, which enables processing of NGS data in addition to the standard procedure of using IMGT/HighV-QUEST output data for immune repertoire analyses.We present a full immunoinformatics pipeline to profile the immune repertoire of patients and to classify their health states. This pipeline has been used to evaluate a set of patient data by processing NGS data using the newly implemented NGS analyzer, performing clonality and diversity analysis, calculating features based on the preceding analyses and predicting the health states using machine learning approaches all integrated in the software IMEX.	Data poster	Health
P_Da102	341	Kees van Bochove, Reinhard Schneider, Sacha Herzinger, Wei Gu, Venkata Satagopam, Serge Eifes,Riza Nugraha, Gustavo Lopes,Piotr Zakrzewski,Peter Kok,Ward Weistra,Janneke Schoots, Annick Peleraux, Rogerio Martins, Heike Schürmann, Sherry Cao,	Kees van Bochove	Open Source Development Success through collaboration: SmartR in tranSMART	tranSMART is an open source translational research platform used by academic researchers and pharmaceutical companies around the world. The tranSMART Foundation, supported by many of these users, guards the quality of the platform by setting code standards and encouraging collaboration.The Innovative Medicines Initiative (IMI) project eTRIKS is the result of a collaboration between 17 different academic and industrial partners. Each combining their strengths in the development of a platform and services for data staging, exploration and use in translational research.Within eTRIKS one of the academic partners, University of Luxembourg, developed a new visualisation platform for within tranSMART, called SmartR. SmartR is aimed to provide a highly dynamic and interactive way of visualizing and analyzing data within tranSMART. Using recent web technologies it generates interactive analytics within the web browser (figure 1) rather than making use of static images generated by R.Academic and industrial environments put different constraints and requirements on software development. Where academic developments are focussed on proving the validity of a novel innovation, software for industrial research needs to be scalable and reliable.Within separate development projects and hackathons the pharmaceutical companies Pfizer and Sanofi have sponsored the open source bioinformatics software company The Hyve to work with the original developer to upgrade the SmartR visualization platform to be of commercial quality in code and analysis algorithms and allow for easy extension with more workflows (figure 2). By leveraging the trust built in the open source community these competing companies have involved each other in their projects building towards a common goal.Active collaboration is still underway to release the enhanced SmartR as a default plugin with the 16.2 version of tranSMART, which will be released in the second half of 2016.	Data poster	Biotechnology
P_Da103	724	Dilip Durai and Marcel Schulz	Dilip Durai	Optimal normalization of sequence data for de novo transcriptome assembly	Recent developments in sequencing technologies have resulted in generation of huge amount of data in a short span of time. This has generated interest in de novo analysis of the sequences. One of the most common method for de novo analysis is the de Bruijn graph based de novo assembly. A major challengeface by many of the modern assemblers is the high amount of redundant reads in the dataset which results in large amount of memory consumption. We observed that only a certain percentage of reads are required to obtain a high quality assembly. Current heuristics for redundancy removal have a risk of losing kmers which might for connections between two nodes and hence might result in sub-optimal assembly.Here, we consider the problem as a set cover problem and propose a normalization algorithm which calculates the minimal number of reads required to cover all nodes in the de Bruijn graph. Hence, we maintain the connectivity between the nodes in the graph. Upon applying the algorithm to various human dataset we achieved a better reduction as compared to the existing redundancy removalalgorithms. Also the reduction did not compromise on the quality of the final assemblyWe feel that this algorithm will make the process of assembling sequence more efficient especially in an era where the sequencers are producing billions of reads having high error rates and sampling biases.	Data poster	Fundamental
P_Da104	668	Robbin Bouwmeester, Frans M van der Kloet, Martijs J Jonker, Age K Smilde and Johan A Westerhuis	Robbin Bouwmeester	Penalizing miRNA-mRNA correlations based on their association likelihood improves enrichment of relevant terms in B-cell differentiation	MicroRNAs (miRNA) play an important role in post-transcriptional regulation. They can regulate multiple biological processes by either a translational block or by mRNA degradation. Finding the mRNA targets of miRNAs in eukaryotes is not a trivial complement sequence alignment problem. Experimental and in silico evidence of binding between pairs of mRNA and miRNA sequences can be found in so-called target databases.Studies that involve both miRNA and mRNA measurements should benefit from using this binding evidence in the statistical analyses. However, experimental target databases are incomplete in terms of available miRNA-mRNA associations (low sensitivity), while in silico databases detect a large number of false positive associations (low specificity). At this moment, there is no consensus on how these target databases should be used in genome-wide miRNA-mRNA expression analysis.The evidence of miRNA-mRNA associations were obtained from multiple target databases such as miRecords, miRTarBase, RepTar, TargetScan, PicTar and miRWalk. The likelihood of association between a miRNA-mRNA pair was calculated for in silico predictions using experimentally determined pairs as a gold standard. Correlations of miRNA-mRNA expression are penalized based on their association likelihood to reduce spurious associations.The new approach was validated internally using cross validation procedures. Furthermore, external validation was performed using mRNA and miRNA sequencing data from pre-B-cell differentiation cell lines of mice at 6 different time points. The penalized correlations resulted in an increased number of relevant terms in a gene set enrichment analysis compared to filtering with single target databases or combinations thereof.	Data poster	Biotechnology
P_Da105	410	Aliaksei Vasilevich, Shantanu Singh, Aurélie Carlier and Jan de Boer	Aliaksei Vasilevich	Phenotypic space as benchmark of cells fate	It is well known that cell shape has an effect on cell function, and that by manipulating cell shape, we can direct cell fate. Altering the cell shape through surface topographies opens new opportunities for the development of biomedical materials. To obtain a variety of cell shapes, we applied a high-throughput screening approach and determined the cell response to 2176 randomly generated surface topographies. Cell morphology was captured by high-content imaging and we performed image analysis in CellProfiler which generated a large dataset with hundreds descriptors. Importantly, we found biologically meaningful clusters of cells based on cell shape features. In total we identified 28 surfaces based on cell shape diversity – the resulting selected surfaces were observed to have distinct designs. These 28 topographies were further used to reveal how different cell shapes induced by topography affect fundamental cell functions. To investigate this, we have performed various functional assays with hMSCs such as: differentiation, proliferation, migration, apoptosis and protein synthesis. We used these assays to identify surfaces inducing the most unique cell response, and to further narrow down the list of topographies. By performing microarray analysis on cells grown on these surfaces, key target genes involved in surface topography interaction will be identified.The results of this study will lead to new advances in our understanding of how surface cues can influence cell behavior, enabling the improved design of materials for biomedical applications.	Data poster	Biotechnology
P_Da106	407	Electra Tapanari, Dan Bolser, Alessandro Vullo, Robert Petryszak, Christoph Grabmueller, Paul Kersey, Nuno Fonseca, Laura Huerta Martinez and Maria Keays	Electra Tapanari	Plant RNA-Seq data in the Track Hub Registry	There is a plethora of RNA-Seq data submitted by scientific studies worldwide to the European Nucleotide Archive (ENA). We created a pipeline that discovers all the plant RNA-Seq data available in ENA, aligns them to the Ensembl Plants reference genomes and generates CRAM alignment files that are then submitted to ENA as analysis objects. Using the UCSC track hub standard, alignments stored in the CRAM file format can be attached to the Ensembl browser and visualized in the genomic context as track hubs.The Track Hub Registry (THR) is an Ensembl-built platform where track hubs can be registered and automatically linked to supported genome browsers. Plant track hubs were registered using the REST API service of the THR and are updated daily. At the moment there are around 1,000 plant RNA-Seq studies from 37 plant species, corresponding to the same number of track hubs in the THR. The users can filter on their condition of interest and find the relevant track hubs. They can then see the expression levels of that condition in the genome browser.	Data poster	Agro-Food
P_Da107	816	Rabie Saidi, Alexandre Renaux, Tunca Dogan and Maria Martin	Rabie Saidi	PredComp: A tool for comparing and benchmarking protein annotation predictions against UniProtKB	A number of automatic annotation systems are integrated in UniProtKB/TrEMBL to infer functional attributes of proteins. With the continuous development of additional prediction systems in the literature for different academic and industrial purposes, there is a strong need for benchmark comparisons to assess the coverage and quality of these annotations. To facilitate this benchmarking, we have developed PredComp, a public tool to compare various types of functional annotations of a protein set supplied by any method, against annotations provided by systems integrated in UniProtKB/TrEMBL.PredComp covers the main annotation systems present in UniProtKB/TrEMBL including SAAS and UniRule. It summarizes the annotation gain of the systems prediction by highlighting the percentage of entries that previously lacked annotation for a particular predicted feature. Moreover, it classifies the system annotations in comparison to the set of annotations obtained by the systems present in UniProtKB/TrEMBL (collectively and individually per system) as identical, similar, or mismatched (a.k.a. contradicting) annotations. Such classification is useful in quantifying numerically the comparability and correlation between the new system's annotations and those already existing in the database which in turn is useful in validating the new system's predictions intuitively. PredComp provides such information in the form of a hierarchical graphical report that can be navigated to acquire knowledge about the new system's annotation on different comparison dimensions.It's anticipated that the tool will frequently be used by software developers, pharmaceutical companies and the biomedical research community. PredComp is publicly available as a web-server at www.ebi.ac.uk/Tools/pfa/predcomp.	Data poster	Agro-Food Application Biotechnology Health
P_Da108	659	Martin Strazar and Tomaz Curk	Martin Strazar	Predicting alternative splicing from contextual information on splicing factors	Alternative splicing is an integral part of mammalian transcription. The majority of human genes undergo alternative splicing, and improper splicing is often associated with disease. The role of many RNA-binding proteins (RBPs) in splicing remains unclear.The availability of next-generation sequencing assays motivates searching for the "splicing code" [1], a model that can relate multiple cis- and trans- acting factors to differential exon usage.We model differential expression of more than 50.000 human cassette exons upon shRNA knockdown of 153 different RBPs (including SRSF1, U2AF1/2, FUS, hnRNPs family), using data from the ENCODE project [2]. We propose a novel, integrative Bayesian matrix factorization (BMF) method that integrates differential exon usage with side information on exons (RNA sequence, structure, conservation) and RBPs (protein-protein interactions, iCLIP/eCLIP assays) by placing Gaussian Process (GP) priors on latent matrices. Automatic relevance determination is applied to infer the optimal GP covariance structure, which is then used to predict differential exon usage upon knockdown of RBPs with no shRNA knockdown data.The BMF model competes favorably with related techniques in predictive performance and interpretability. It discovers combinations of RBPs important in splicing, which was previously used only indirectly [3]. The model can predict changes in splicing upon sequence mutations or upon introduction of new RBP binding sites, which enables a more mechanistic understanding of alternative splicing.[1] H. Xiong et al., Science, vol. 347 (2015).[2] The Encode Consortium, PLoS Biol. vol. 9, (2011).[3] A. Busch, K. J. Hertel, RNA, vol. 21 (2015).	Data poster	Fundamental
P_Da109	451	Wojciech Lesinski, Agnieszka Kitlas Golinska, Aneta Polewko-Klim, Andrzej Przybylski and Witold R. Rudnicki	Wojciech Lesinski	Predicting Arrhythmia with Random Forest	The study is devoted to development of predictive models of arrhythmia onset using machine learning methods.The input data consisted of 145 ECG signals in the form of RR-intervals.The samples contained both periods of normal heartbeat and periods with onset of arrhythmia.The 33 features describing the signal were obtained using analysis in time domain, frequency domain and by using nonlinear Poincaré maps.The feature relevance was determined using the perturbation importance obtained from Random Forest within cross-validation loop.The 15 most important features were collected in each step of cross-validation.The results of feature selection were stable and repeatable.Then two classes of predictive models were built using selected 15 features.In the first case Random Forest algorithm was applied, using 5-fold cross-validation.The average classification error obtained in 1000 iterations of the procedure was 0.24.For comparison we applied identical procedure for the same set with randomly permuted class labels.In this case the mean classification error was equal to 0.5.What is more, the maximal error obtained in 1000 trials with real data has been lower than the minimal error obtained for the randomised data.The results obtained in the current study are comparable to those obtained for example in [1], however, the lower number of simple features built were used to build models and with lower number of cases. This demonstrates robustness of the approach used in the current study.Özçift, A. (2011). Comp. Biol. Med.	Data poster	Health
P_Da110	366	Sofia Papadimitriou, Andrea Gazzo, Guillaume Smits, Ann Nowé and Tom Lenaerts	Sofia Papadimitriou	Predicting digenic variant effects with DIDA	With the advances in medical genomics, it has been shown that many genetic disorders previously considered to be monogenic, may be attributed to more complex inheritance mechanisms, following instead an oligogenic inheritance model. However, little is still known about the genetic causes of these disorders. The aim of this work is the study of digenic diseases, the simplest case of oligogenic disorders, and the construction of predictive methods that can distinguish variant combinations within two genes leading to disease or not. For this purpose, we exploited the information present in the publicly available DIDA database, whose main entity is a digenic combination (i.e. a combination of variants within two genes) leading to a digenic disorder, combined with information of the involved genes and their associated genetic variants. As a neutral dataset, we used the variant information of healthy individuals from the 1000 genome project, further filtered and annotated to create comparable digenic combinations with those in DIDA. Using these instances, a random forest predictor for digenic combinations was created. Our results reveal that single variant effect predictors on the gene and protein function (such as Polyphen-2) together with Pfam information, as well as differences in the wild type and mutated amino acid properties, are essential for the discrimination of neutral from disease-causing digenic combinations. These results constitute a first step in determining the genetic causes of digenic diseases and open the path for the construction of more advanced predictive tools for complex genetic disorders.	Data poster	Health

P_Da111	471	Hiroki Konishi, Daisuke Komura, Hiroto Katoh, Ken Tominaga, Ryohei Suzuki and Shumpei Ishikawa	Hiroki Konishi	Prediction of antigen-specific immunoglobulins from amino acid sequences using semi-supervised deep learning.	Antibody immunoglobulins recognize and neutralize harmful agents such as pathogens and cancer cells through their binding to antigen molecules derived from the agents. Detection of immunoglobulins that recognize a specific antigen or antigens with shared physiochemical properties (e.g. carbohydrates, proteins and lipid) will unravel the contribution of these antigens to the whole immune response in various disease state. Recently, next-generation sequencing (NGS) technologies have produced unprecedented amount of immunoglobulin sequences. Although these 'immunosequencing' data could be potentially useful for the prediction of antigen-specific immunoglobulins, to the best of our knowledge, no such methods have been developed so far.Here we have developed a new deep learning-based method for the prediction of antigen-specific immunoglobulins by the amino acid sequences obtained from NGS data. Amino acid sequences were converted into a series of numerical index reflecting the physicochemical property scores such as hydrophobicity and used as input of deep learning. Although deep learning has generally achieved superior performance in DNA or RNA analysis over other supervised learning algorithms, it needs enormous amount of labeled data (e.g. immunoglobulin sequence and antigen it recognizes), which is hardly obtained. In order to compensate for the lack of the labeled data, we have taken a semi-supervised learning approach, which improves performance by utilizing unlabeled data as well as labeled data.We have applied the proposed method to simulated and real datasets to show the effectiveness of the method.	Data poster	Biotechnology
P_Da112	399	Konstantin Okonechnikov, Ana Conesa and Fernando Garcia-Alcalde	Konstantin Okonechnikov	Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data	Detection of random errors and systematic biases is a crucial step of a robust pipeline for processing high-throughput sequencing (HTS) data. Bioinformatics software tools capable of performing this task are available, either for general analysis of HTS data or targeted to a specific sequencing technology. However, most of the existing quality control (QC) instruments only allow processing of one sample at a time. This is a major limitation, since sequencing experiments are often conducted using biological replicates and can include multiple conditions. We would like to present the second version of Qualimap, a toolkit for QC of HTS alignment data. Qualimap 2 provides new analysis capabilities that allow multi-sample comparison of sequencing datasets. Additionally, it includes a novel mode for discovery of biases and problems specific to RNA-seq technology based on the redesigned read counts QC mode. In general, Qualimap is a multiplatform user-friendly application with both graphical user and command line interfaces. The results of the QC analysis are presented as an interactive report within the graphical user interface, as a static report in HTML, as a PDF, or as a plain text file suitable for parsing and further processing. Importantly, Qualimap 2 has gathered a community of users who frequently suggest new features and contribute their code. Additionally, large number of the novel features were tested by users. The recent publication describing Qualimap 2 was already cited 10 times and the development of the project remains active.	Data poster	Application Biotechnology Fundamental
P_Da113	318	Jan Koster, Richard Volckmann, Piet Molenaar, Danny Zwijnenburg and Rogier Versteeg	Jan Koster	R2: Accessible online genomics analysis and visualization platform for biomedical researchers	In this era of explosive genomics data generation, there is a growing need for accessible software solutions that can help unlock biological/clinical characteristics from such data. With the biomedical researcher in mind, we developed a comprehensive web-based system called R2 (r2.amc.nl).The R2 platform consists of a database storing both publicly accessible as well as shielded datasets with unified gene annotation, supplemented with a large suite of tools and visualizations that can be used on these data and their associated annotation. As such the user experiences the same look & feel throughout the mining process. R2 also forms a perfect liaison between bioinformaticians and molecular biologists.In the public section, R2 hosts over 90,000 samples. Besides gene expression, the platform is also being employed in the integration, analysis and visualization of aCGH, SNP, ChIP, methylation, miRNA, and whole genome sequencing data.R2 contains a set of interactive inter-connected analyses, allowing users to quickly hop from one view to another. Analyses include, correlation, differential expression, gene sets, gene ontology, transcription factor binding sites, PCA, k-means, Kaplan Meier scans, signature creation etc. Visualizations include, various gene oriented plots, heatmaps, circos, genome browser, Venn, etc.Many parts of the R2 platform are publicly accessible through the portal. The gene expression analysis tools have thus far been used in more than 340 peer-reviewed scientific publications. R2 is also used in many international collaborative efforts involving unpublished datasets. The webservers have been serving over 1.200.000 pages over the past 12 months (April 2016).	Data poster	Fundamental
P_Da114	876	Katerina Taškova and Miguel Andrade-Navarro	Katerina Taškova	Rank aggregation-based prioritization of drug-response genes in toxicogenomic data	Toxicogenomic database are valuable source for analyzing drug response in biological systems, and have been used for identification of gene biomarkers of drug-induced toxicity. In this context, we present comparative analysis involving a comprehensive large-scale toxicogenomic database with the goal i) to compare the concordance of early drug-response genes selected by differential expression analysis via robust rank aggregation methods and rat-to-human orthology mapping with gene candidates from human toxicity literature; ii) to check the extend to which the orthology mapping limits this concordance, and how suitable is the rat animal model for prioritizing human toxicity gene reporters.More precisely, we focused on gene expression time profiles corresponding to a set of 33 mainly toxic drugs across all single-dose experimental scenarios (human and rat primary hepatocytes, rat liver and kidney) deposited in the Open TG-GATEs database. Drugs-wise differential expression-based gene rankings were summarized into one final ranked gene list, that was limited to the human one-to-one orthologs in the rat scenarios. We evaluated the performance of the ranking method against human toxicity gene candidates selected based on gene-and-toxicity co-occurrence analysis of PubMed article annotations. Finally, we compared different ranking schema based on the ROC curve analysis, in order to obtain better concordance between the gene expression-based and literature-based gene candidates.	Data poster	Health
P_Da115	478	Nicola Lazzarini and Jaume Bacardit	Nicola Lazzarini	RGIFE: a ranked guided iterative feature elimination heuristic for biomarkers identification	Current -omics technologies are able to sense the state of a biological sample in a verywide variety of ways. Given the high dimensionality that typically characterises these data, relevantknowledge it's often hidden and hard to identify. Machine learning methods, and particularly feature selection algorithms have proven very effective over the years at identifying small but relevant subsets of variables from a variety of application domains, including -omics data. Many methods exist with varying trade-offs between the size of the identified variable subsets and the predictive power of such subsets.In this work we focus on an heuristic for biomarkers identification called RGIFE: rank-guided iterative feature elimination. RGIFE is guided in its biomarkers identification process by the information extracted from the machine learning models and incorporates several mechanisms to ensure that it creates minimal and highly predictive biomarker sets.We compared our heuristic against 4 well-known feature selection algorithms using 10 cancer related transcriptomics datasets. First we assessed the prediction performance of the heuristic and we compared the number of selected features by each method. Secondly, using a prostate cancer related dataset as case study, we looked at the biological relevance of the identified biomarkers. RGIFE obtained similar performances to widely adopted feature selection methods while selecting significantly less feature. The case study showed the higher biological relevance of the selected features in comparison with the other methods.The RGIFE source code is available at: http://ico2s.org/software/rgife.html.	Data poster	Fundamental
P_Da116	350	Eugenia Galeota and Mattia Pelizzola	Eugenia Galeota	SEMANTIC AWARE RETRIEVAL AND INTEGRATION OF PUBLIC (EPI)GENOMICS METADATA	Integration and reuse of publicly available biological data from high-throughput sequencing platforms relies on the availability of well-organized and clearly described metadata. To this purpose, software tools that enable their annotation with controlled vocabularies, and the quantification of the relationships between studies are indispensable.We developed a user-friendly R package that allows users to easily and efficiently annotate public repositories' metadata with concepts from a multitude of biomedical ontologies. The software also enables the identification of additional coherent samples, using various semantic similarity measures to relate the metadata of a query study with those of other relevant studies.Proving the utility of our approach we applied this software to annotate thousands of Gene Expression Omnibus ChIP-seq metadata in order to retrieve all the human ChIP-seq experiments targeting the Myc transcription factor, associating them to specific disease and tissue/cell-line concepts. We demonstrated how it is possible to study the chromatin modifications associated to the Myc activity, by including independent ChIP-seq experiments targeting a number of epigenetic marks annotated with concepts compatible to the Myc samples. The organization of the samples by ontology-based semantic similarities resulted in patterns of ChIP-seq signals coherent with the biological knowledge on the field. This example illustrated the power of this approach, and the usefulness of combining previously unrelated, while semantically compatible, large-scale datasets.	Data poster	Fundamental
P_Da117	547	Gurnoor Singh, Arnold Kuzniar, Anand Gavai, Richard Gf Visser and Richard Finkers	Gurnoor Singh	Semantic-mining of QTL tables in scientific articles for trait lead discovery	Quantitative trait loci (QTL) are genomic regions associated with traits of interest. QTL contains genes that are candidates for expression of phenotypes (e.g. disease resistance or nutritional value). Many studies nowadays focus on identification of these candidate genes as they assist in, for example: 1) understanding of the molecular mechanism underlining a given phenotype, 2) building better software tools that help in breeding improved cultivars. However, QTL information is mostly captured as tables, in full-text or supplementary material of scientific articles. Traditional text-mining techniques focus on extracting knowledge from unstructured free text and thus cannot extract QTL information. Accordingly, it is difficult to capture an overall picture of QTL for a selected plant species.In this study, we aim to develop a tool which extracts QTL information from heterogeneous tables in full text or supplementary information of a scientific publication. The schema of a table and its meta-data is extracted by taking europmc xml files as an input. Rows, columns and individual cells of a selected table are enriched with annotations based on Trait Ontology, table-caption, table-footers and table-headings. These annotations help in mining and storing the relationships expressed in a table to an Open Linked format based on FAIR Data Principle.The developed system will summarize QTL information. When combined with knowledge from other databases and genome sequences, this tool will lead to a more efficient and an effective-way to perform trait-lead discovery.	Data poster	Agro-Food
P_Da118	732	Richard Lupat, Jason Li, Kaushalya Amarasinghe, Chalini Wijetunge, Jordan Sands and Tony Papenfuss	Richard Lupat	Seqliner: software framework for managing and developing sustainable bioinformatics analysis pipelines in a production environment	With the enhancement of high-throughput sequencing (HTS) data in recent years, the volume of data being generated has increased tremendously and requires a more specialised data processing workflow. A typical HTS sample will go through a series of software or analysis methods, which often referred as 'pipeline'. Some of the biggest challenges for managing these pipelines are: i) Analysis method changes frequently to deal with new data types and for achieving better performances, ii) these software packages are often written by various organisations and using different languages, hence integration between the steps in the pipelines are often difficult, iii) the hardware where these pipelines will run on will vary depending on the use cases and are often upgraded to cope with the demand for quicker turnaround time, iv) the requirements for locking down analysis pipelines for better analysis reproducibility, v) the ability to customise pipelines depending on individual needs, most of the time minor tweaks to small part or parameters of the pipelines.We propose seqliner, a software framework for managing and developing these pipelines. It was designed with a concept of reusable modules, pipelines and configurations file. A module consists of one or more analysis tools that are wrapped around a consistent framework class and will be defined with a certain requirements of inputs and outputs as well as set of parameters that can be configured via configuration files. These modules will serve as building blocks for pipelines and multiple pipelines can be combined to build more complicated pipelines.	Data poster	Health
P_Da119	819	Adem Bilican, Yves Widmer, Simon Sprecher and Rémy Bruggmann	Adem Bilican	Systems Biology of forgetting in Drosophila	Targeted DamID (TaDa) is an efficient technique to perform cell-type-specific (or genome-wide) binding profiling of a protein of interest without individual cell isolation. The TaDa method relies on a construct formed by the DNA adenine methyltransferase (Dam) enzyme from E. coli and a protein of interest with DNA or chromatin-binding capabilities. The binding of the protein of interest to the DNA activates the Dam enzyme resulting in specific Adenine methylation at GATC sites. In the SynaptiX project, we are interested to study transcriptional changes during the process of forgetting. Therefore, we focused on the TaDa technique by studying the binding of the RNA polymerase II, which represents a marker for transcriptional activity. The Dam-POLII itself is under the control of a cell specific promotor. We performed an experiment on two groups of flies: one group that forms memory (paired training) and another group that does not form memory (unpaired training). The experiment was divided in 4 time points [time points] with a total of 64 samples. The samples were sequenced using Illumina technology resulting in approximately 25 million paired-end reads per sample. Based on the overall gene expression changes between the paired and unpaired protocols we identified 33 candidate genes involved in the process of forgetting in Drosophila such as Dop1R2 known to be involved in Alzheimer's disease and amnesia. Finally, these candidate genes will be tested with the RNAi technology to confirm their potential role in forgetting in Drosophila.	Data poster	Health
P_Da120	449	Chul Kim, Boseok Seong, Sang-Jun Yea, Yunji Jang, Seokjong Yu and Hyojin Kang	Chul Kim	The correlation analysis between the user search trends and prescription usage in the traditional Korean medicine	Objective :The purpose of this study is to find out if any correlation between the actual usage of prescription in hospital and the internet search trends exists in the field of Traditional Korean Medicine(TKM). In this study, we chose four TKM prescription, i.e. Ojeok-san, Socheongryong-tang, Hyangsapyeongwi-san, Gumiganghwal-tang.Materials and methods: The prescriptions selected in this study were the top 4 in terms of the annual number of medications (ANM) in TKM clinics and hospitals in Korea. And two representative web search engines, i.e. NAVER and GOOGLE, were selected to collect the web search logs for words related to 4 prescriptions. Then Pearson's correlation coefficient are calculated between collected data.Results: The web search traffic logs were collected for the past seven years (2007~2013) from NAVER and GOOGLE and data for the annual number of medications are download from web site of National Health Insurance Service in Korea. The correlation coefficient between web search traffic logs of prescription terms in NAVER and ANM ranged from 0.770 to 0.923. However the correlation coefficient between GOOGLE and ANM is very low.Conclusion : Because the correlation coefficient between search trend in NAVER(market share : 75% in Korea) and ANM for four prescriptions is all over 0.7, it can be interpreted as a Strong positive correlation. Even if you consider that Internet use is rapidly increasing, the market and interest in TKM is increasing obviously in proportion.	Data poster	Health
P_Da122	518	Małgorzata Wnętrzak, Paweł Błażej and Paweł Mackiewicz	Małgorzata Wnętrzak	The impact of crossover operator on the genetic code optimization performed by Evolutionary Algorithms	There are many theories trying to explain the current organization of the canonical genetic code. One of them postulates that the genetic code evolved to minimize harmful effects of amino acid substitutions and translational errors. A way to verify this hypothesis is to find a code that would be the best optimized under given criteria and compare it with the canonical genetic code. This approach requires effective algorithms to search the huge number of possible alternatives. In this context, Evolutionary Algorithms seem to be such appropriate methods. They are based on mutation and crossover operators, which are responsible for generating the diversity of potential solutions to the optimization problem. They have distinct properties and play different roles in the optimization process. We developed new types of crossover operators dedicated for the genetic code models under the study. To assess the influence and effectiveness of operators in searching the space of potential codes, we applied various combinations of mutation and crossover probabilities under three models of the genetic code. The obtained results demonstrate that the usage of crossover operators can substantially improve the quality of the solutions. The best found genetic codes without restrictions on their structure minimized the costs in polar amino acid requirements about 2.7 times better than the canonical genetic code.	Data poster	Fundamental
P_Da123	684	Lea A.I. Vaas, Janneke Schoots, Stefan Payralbe, Steen Manniche, Kees van Bochove, Cindy Levy-Petelinkar, Claus Stie Kallesøe, Phil Gribbon and Manfred Kohler	Lea A.I. Vaas	The ND4BB Information Centre – general concept and technical challenges	The New Drugs for Bad Bugs (ND4BB) initiative is a series of programs designed to specifically address the scientific challenges associated with antibacterial drug discovery and development. The over-arching concept of ND4BB is to create an innovative public-private collaborative partnership that will positively impact aspects of antimicrobial resistance research which benefit the future discovery and development of novel agents for the treatment, prevention and management of patients with bacterial infections.One important objective of ND4BB is to develop a data repository to provide an information base for research projects focused on antibiotic resistance. All consortia partners contribute data to the ND4BB data hub and collaborate to share data and experience amongst all programme members and the antibiotic research community as a whole.Here we present the technical concepts underlying the ND4BB Information Centre and describe the specific challenges of a data base setup integrating both compound-centric and sample-centric data from multiple providers. The unique strength of the unconventional combination of a commercially available data base system (LSP by Gritsystems, DK) with open source solutions (tranSMART plus service by THE HYVE, NL) resulted in a comprehensive data-warehouse system for research data from preclinical drug research, and is not restricted to antimicrobials.Exemplary workflows will highlight possible types of research questions to be tackled and illustrate major features of the dedicated R-packages facilitating collection, download and data preparation for analysis in R (R Core Team 2016) or other tools like TIBCO Spotfire®.	Data poster	Health
P_Da125	619	Sam Nicholls, Amanda Clare, Wayne Aubrey and Christopher Creevey	Sam Nicholls	Towards an algorithm for extracting exciting enzymes from metagenomic data sets	There has been much interest in investigating the genomic repertoire of microbial communities for compounds of medical or industrial relevance such as small peptides and enzymes.If isolated, they could be exploited in a wealth of scenarios including the refinement of biofuels, production of plastics, creation of new classes of antibiotics or even scrubbing oil from water.However, identification of these from a highly biodiverse microbial community is not a trivial undertaking as metagenomic assemblies regularly underrepresent the true variation present and mask possible novel peptides and enzymes.The problem is: given millions (or billions) of short DNA strings from a microbial community containing multiple species (many of which are unknown or unculturable), how can we identify and assemble the "true" DNA sequences (the haplotypes) of the genes responsible for these "interesting" biochemical reactions?To address this we attempt to identify variants (SNPs) shared by multiple reads (short strings of DNA), aligning to a genomic region of interest. Such shared SNPs represent variation "lost" in the assembly and can be represented by a graph where probabilities of one SNP variant following another can be evaluated from the read frequencies and associated qualities seen in the raw reads. Potential haplotypes can be constructed as a path through this graph: the metahaplome.We introduce a description of the metahaplome problem and demonstrate the difficulties involved in extracting likely haplotypes. We also present precursory work on a probabilistic graph-based approach to find approximate haplotypes to serve as starting points for primer design.	Data poster	Fundamental

P_Da126	492	Todd Taylor, Naveen Kumar and Maxime Hebrard	Todd Taylor	Turning 'big data' into 'small data' through crowdsourced curation: integrating all types of scientific knowledge	'Big data' in the form of scientific media comes in an endless variety of languages and formats, including journal articles, books, images, videos, databases, etc. With textual media, there is often additional information (tables, figures, supplementary data) associated with or embedded in the text. While there are many good resources for browsing, searching and annotating some of this media, there is no single place to search them all at once, and generalized search engines do not allow for the type of comprehensive and precise searches that researchers require. And, as more and more data continues to accumulate, the problem will only grow worse. One could argue that any scientific media that is on the web is therefore connected, but much of it remains offline or is inaccessible and is therefore neither discoverable nor connected.To address these issues, we created iCLiKVAL (iclikval.riken.jp/), an intuitive web-based tool that uses the power of crowdsourcing to accumulate annotation information for all scientific media found online (and potentially offline). Annotations in the form of key-relationship-value tuples (any language), added by users through a variety of methods, can make vast amounts of unstructured data easier to comprehend and visualize by turning it into 'small structured data'. This allows for much richer data searches and for discovery of novel connections by basically integrating all forms of scientific knowledge through common terminology. iCLiKVAL is an open-access database, and all of the annotation data is freely available for text mining and other purposes via our API.	Data poster	Biotechnology
P_Da127	525	Seyed Ziaeddin Alborzi, Marie-Dominique Devignes and David Ritchie	Marie-Dominique Devignes	Using Content-Based Filtering to Infer Direct Associations between the CATH, Pfam, and SCOP Domain Databases	Protein domain structure classification systems such as CATH and SCOP provide a useful way to describe evolutionary structure-function relationships. Similarly, the Pfam sequence-based classification identifies sequence-function relationships. Nonetheless, there is no complete direct mapping from one classification to another. This means that functional annotations that have been assigned to one classification cannot always be assigned to another. Here, we present a novel content-based filtering approach called CAPS to systematically analyse multiple protein-domain relationships in the SIFTS and UniProt databases in order to infer direct mappings between CATH superfamilies, Pfam clans or families, and SCOP superfamilies. These mappings are beneficial to: 1) transfer annotations from one classification scheme to another, 2) investigate annotation consistency between different classifications. CAPS discovers a total of 5,576, 6,618 and 7,623 non-redundant SCOP-CATH, SCOP-Pfam, and Pfam-CATH associations with recalls of 0.968, 0.978, and 0.978, respectively, with reference to the manually curated InterPro database. Overall, CAPS associates 2,549 CATH with 1,817 SCOP Superfamilies, and 3,033 and 3,168 Pfam clans with 2,745 and 2,109 CATH and SCOP Superfamilies, respectively. This corresponds to four times as many SCOP-CATH Superfamily associations as currently exist in Genome3D, and 16 times as many SCOP-Pfam associations as available on the Pfam website, with almost 100\% overlap of both datasets.	Data poster	Fundamental
P_Da128	379	Markus List	Markus List	Using Docker compose for the simple deployment of an integrated high-throughput screening platform	Dealing with massive amounts of biological data is unthinkable without state-of-the-art tools. Over time, these applications have become increasingly complex and can often only be used when a long list of preconditions are met. There are serious issues with the installation and maintenance of tools due to version conflicts, outdated repositories and poor documentation. Moreover, complex tasks require integrating several tools into a workflow. Open platforms like Galaxy and Taverna have emerged to simplify building and operating such workflows. Nevertheless, ensuring the fulfillment of all preconditions remains a critical issue.A solution to encapsulate a tool with its dependencies in container images is Docker. An extension, called Docker compose, further facilitates interaction of several containers in a coherent software configuration. Here, we demonstrate the power of this approach by creating and deploying an integrated high-throughput screening (HTS) platform through Docker compose, which comprises systems for laboratory information management, HTS sample and plate management, HTS data analysis, systems biology analysis, and reverse-phase-protein array data management/analysis, as well as service containers for relational database management, load balancing and for a key-value store.Docker is a promising solution to fully address issues in deploying scientific software even in cases where several tools need to be integrated. In addition, Docker compose allows, for instance, to deploy a complex HTS platform in a single command. We expect that the use of Docker in cases like this will enable the more widespread use of scientific software and free up time spent on dealing with software dependencies.	Data poster	Biotechnology
P_Da129	574	Amjad Alkodsi, Katja Kaipio, Johanna Hynninen, Sakari Hietanen, Rainer Lehtonen, Olli Carpén, Seija Grénman and Sampsa Hautaniemi	Amjad Alkodsi	Whole-genome characterization of pre- and post-treatment high-grade serous ovarian cancer	High-grade serous ovarian cancer (HGSOC) is the most common and aggressive subtype of ovarian cancer, which is the fifth most common cancer-related cause of death in women. While an HGSOC patient typically responds well to first-line chemotherapy, most women suffer a treatment-resistant recurrent tumor and succumb to their disease.We obtained whole-genome sequencing (WGS) data from 15 samples from six HGSOC patients with an average coverage of ~150x. Samples were collected from different metastatic sites both before and after neoadjuvant chemotherapy. Our objective is to gain understanding of the genetic divergence between pre- and post-treatment cancers, tumor heterogeneity among different anatomical sites, and putative drivers of chemo-resistance.We analyzed WGS data for somatic point mutations, short indels, copy number variants and breakpoints of structural rearrangements. TP53 mutations were ubiquitous in all samples as expected. We identified gained somatic mutations in post-treatment tumors in genes not previously reported to be mutated in primary HGSOC. Majority of the detected alterations were shared between all samples from the same patient exhibiting limited divergence between primary and post-treatment tumors as well as between different anatomical sites. A vast proportion of detected breakpoints showed microhomology at break-junctions indicating defective double-strand DNA repair by homologous recombination (HR). We discerned two mutational signatures attributed to aging and HR deficiency, and observed an increased relative contribution of the latter in late private mutations. These results comprehensively characterize the genomic landscape of post-treatment HGSOC and provide a solid basis for identification of mechanisms of resistance in HGSOC.	Data poster	Health
P_Da130	873	Yana Safonova, Alexander Shlemov, Andrey Bzikadze and Sergey Bankevich	Yana Safonova	Y-Tools, a toolkit for analysis of adaptive immune repertoires using immunosequencing data	Reconstruction and analysis of adaptive immune repertoires is an important part of various immunological studies. Modern biotechnologies allow one to perform deep and full length scan of antibodies and TCRs using immunosequencing and mass spectrometry. Analysis of such data raises multiple algorithmic challenges that are still poorly addressed in existing bioinformatics tools. Here we present Y-Tools, a novel multipurpose toolkit for construction and investigation of adaptive immune repertoires using immunosequencing and mass spectra data. Y-Tools include: IgRepertoireConstructor, an algorithm for adaptive immune repertoire construction and immunoproteogenomics analysis; aimQUAST, a quality assessment tool for adaptive immune repertoires; IgSimulator, a versatile repertoire simulator; DiversityAnalyzer, a tool for diversity analysis of adaptive immune repertoires, and AntEvolo, an algorithm for construction of clonal trees and evolutionary analysis of antibody repertoires, IgRepertoireConstructor, aimQUAST, IgSimulator, and DiversityAnalyzer are freely available at Github. AntEvolo to be released in 2016.	Data poster	Health

POSTER LIST
ORDERED ALPHABETICALLY BY POSTER TITLE
GROUPED BY THEME/TRACK

THEME/TRACK: ELIXIR

Poster numbers: P_EI001 - 037 Application posters: P_EI034 - 037

Poster number	EasyChair number	Author list	Presenting author	Title	Abstract	Theme/track	Topics
P_EI001	714	Joan Segura, Daniel Tabas Madrid, Ruben Sanchez-Garcia, Jesús Cuenca, Carlos Oscar Sánchez Sorzano, Ardan Patwardhan and Jose Maria Carazo	Joan Segura	3DBIONOTES: Unifying molecular biology information	With the advent of next generation sequencing methods, the amount of proteomic and genomic information is growing faster than ever. Several projects have been undertaken to annotate the genomes of most important organisms, including human. For example, the GENECODE project seeks to enhance all human genes including protein-coding loci with alternatively splices variants, non-coding loci and pseudogenes. Another example is the 1000 genomes, a repository of human genetic variations, including SNPs and structural variants, and their haplotype contexts. These projects feed most relevant biological databases as UNIPROT and ENSEMBL, extending the amount of available annotation for genes and proteins.Genomic and proteomic annotations are a valuable contribution in the study of protein and gene functions. However, structural information is an essential key for a deeper understanding of the molecular properties that allow proteins and genes to perform specific tasks. Therefore, depicting genomic and proteomic information over structural data would offer a very complete picture in order to understand how proteins and genes behave in the different cellular processes.In this work we present the second version of a web platform -3DBIONOTES- that aims to merge the different levels of molecular biology information, including genomics, proteomics and interactomics data into a unique graphical environment. Current development offers a unified view of three of the most relevant protein databases: UniProt, PDB, EMDB, and ENSEMBL onto which other sources of biological annotations are also provided, such as PhosphoSitePlus, Immune Epitope DB, BioMuta and dSysMap.	ELIXIR poster	ELIXIR
P_EI002	647	Chao Zhang, Sanne Abeln, Jochem Bijlard, Christine Staiger, Youri Hoogstrate, Alexander Senf, Saskia Hiltemann, David van Enckevort, Remond Fijneman, Jan-Willem Boiten, Gerrit Meijer, Dylan Spalding, Jaap Heringa, Susanna Repo, Niklas Blomberg, Andrew Stubbs, Jordi	Chao Zhang	A Systematic Solution to Map Processed Data in tranSMART to Raw Data in Multiple Repositories	With the evolving of high-throughput experimental techniques, large amounts of molecular profiling data are becoming available for regular clinical studies. These data need to be stored, processed, archived, distributed and, more importantly, linked. In ELIXIR pilot project, we focus on connecting the archival storage of such data, with databases that store the processed data and visualised workflow systems that manage the computational pipelines. After perusing the processed data, users often come back to the raw data not only to reconfirm the data processing but also to further explore the raw data in workflow systems like Galaxy. In the Translational Research IT (TraIT) project of Center for Translational Molecular Medicine, raw data are stored in European Genome-phenome Archive (EGA); and processed data, tranSMART. In light of the distinct interpretations of data ontology structures in different repositories, we are aiming to establish the a systematic, flexible and sustainable mapping between the processed data and the corresponding raw data regardless of their repositories.We propose an ontology structure with a few stable mapping concepts -- Study, Sample Collected Experimental Data (SCED) and File -- to connect different repositories, separating the three essential processes: data uploading, data ontology structuring and data retrieving, in such a way that a general upload and structuring methodology can be used for malleable data retrieving. To demonstrate this ontology structure, we create a RDF graph demo using CTMM-TraIT Cell Line Use Case (CLUC) data.	ELIXIR poster	ELIXIR
P_EI003	811	Jon Ison and Registry-Core Bio.Tools Core Team	Jon Ison	bio.tools : tools & data services registry	bio.tools is a registry of bioinformatics software information, sustained by a community-driven curation effort, tailored to local needs and shared amongst a network of engaged partners. Life sciences yield huge data sets that underpin vital scientific discoveries. In support, a plethora of databases and tools are used, in technically complex and diverse forms, across a spectrum of scientific disciplines. The corpus of information for these resources is fragmented across the Web, with much redundancy, and has lacked a common information standard. The outcome is that scientists often struggle to find, understand, compare and use the best resources for the task at hand. bio.tools enables users to easily formulate precise queries and quickly retrieve tools and services that fulfill specific scientific functions and technical requirements. Ontologies are used, such as EDAM for scientific domains, functions, types of data, data formats and identifiers. These support the retrieval of concise, consistent and therefore comparable information, for the convenience of the user. bio.tools can provide a practical catalogue, that will help scientists not only find, understand, compare and select resources, but also use and connect them in workflows. It is anchored within a broader initiative, fostered by the European Infrastructure for Biological Information (ELIXIR), that includes community-driven performance benchmarking, training, and multi-faceted user support.	ELIXIR poster	ELIXIR
P_EI004	744	Carlos Horro, Manuel Corpas, Rafael Jiménez and John Hancock	Carlos Horro	BioCIDER: a Contextualisation InDEx for biological Resource discovery	Life-science resources (i.e., databases, tools, training materials, courses and event information) are many, diverse and dispersed. The 2016 Nucleic Acids Research (NAR) Database Issue reported 1,685 major databases in the molecular biology domain, while the latest NAR Web Server Issue presented 97 new resources for 2015 alone. It is thus difficult for researchers to be aware of and familiar with all current and relevant research assets, the utility of which is further compromised if these are scattered and difficult to find. Discoverability of resources and information can be significantly enhanced if a suggested list of resources is exposed to users in context with the information they are currently browsing. If sufficiently relevant, a list of potential resources placed in an unobtrusive way can provide users with new, advantageous information and save precious time browsing further. To date, there is no life science-focused service that provides contextualised information driving researchers to the discovery of relevant databases, tools, events and training materials. To address this gap, we have developed BioCIDER, a Contextualisation InDEx for biological Resource discovery. BioCIDER automatically collects information (metadata and source description) from a compendium of centralised registries, including the Bio.tools service registry, the iAnn collaborative event dissemination portal, and the ELIXIR training portal. BioCIDER provides an open source BioJS widget that can be embedded in any website.	ELIXIR poster	ELIXIR
P_EI005	820	Matúš Kalaš, Sveinung Gundersen, László Kaján, Hervé Ménager, Jon Ison, Christophe Blanchet, Steve Pettifer, Rodrigo Lopez, Kristoffer Rapacki andInge Jonassen	Matúš Kalaš	BioXSD | BioJSON | BioYAML - Integrated formats for sequence data	BioXSD has been developed as a tree-based data model and an exchange format for basic bioinformatics data, centred around a bio-polymer sequence. BioXSD allows integration of diverse features, information, measurements, and inferred values about a biological molecule or its part, annotated with provenance and reliability metadata, ontology concepts, scientific remarks, and conclusions.BioJSON and BioYAML are the ongoing developments. These exchange formats are based on the same data model as BioXSD, but providing serialisations in JSON and YAML respectively. BioJSON and BioYAML thus enrich the BioXSD family with alternatives to the original XML.As tree-based data formats, BioXSD, BioJSON, and BioYAML are particularly suitable for programming in object-oriented languages, and for use with web applications and web APIs (Web services), while at the same time allowing a reasonable level of human readability.BioXSD|BioJSON|BioYAML are developed together with GTrack (the universal tabular format for sequence features; http://gtrack.no), by ELIXIR Norway and an international community of collaborators (http://bioxsd.org/#Contact). The BioXSD|GTrack family is going to support smooth interoperability between these alternative, universal formats, and between the tools that consume or provide them as inputs or outputs.	ELIXIR poster	ELIXIR
P_EI006	807	Aravind Venkatesan, Julien Gobeill, Jee-Hyub Kim, Francesco Talo, Michele Ide-Smith, Patrick Ruch and Johanna Mcentyre	Aravind Venkatesan	Bridging literature and data using text-mined annotations in Europe PMC	Bio-curation is essential in maintaining high quality information in biological databases. With the exponential growth in data, curators are faced with a challenging task of bridging the gap between raw data and the knowledge it represents. Curators are required to develop or extend structured vocabularies and tag data with functions, molecular interactions and gene-disease associations by manually extracting information from literature. Biological databases are both dependent and required for this process. Therefore, it is critical that the literature is linked effectively to underlying data and related biomolecular databases, aiding curators extract the essence from the articles. Some links to data are currently made available through data citations, however, deeper integration is required to make curation more scalable. To this end, text mining offers a solution by tagging entities, for instance, gene names, functions and ontological concepts, reducing the burden of manual curation.Here we present a new Europe PMC [1] service - SciLite - that allows text-mined annotations from any source or provider to be displayed on full text articles. The goal of this annotation system is to expose text-mining outputs from the community in useful ways for curators, as well as other interested stakeholders. In the context of ELIXIR, this system will support database curation processes and provide a mechanism to make deep links between the literature and data for clear provenance of curatorial statements.References:[1] Europe PMC Consortium (2015). Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic acids research, 43(D1), D1042-D1048 [PMC4383902].	ELIXIR poster	ELIXIR

P_EI007	836	Jose Borbinha, Pedro L Fernandes, Inês Chaves, Bruno Costa, Daniel Faria, João Cardoso, Célia Miguel, Ahmad Nadali,Daniel Sobral, Arlindo Oliveira,Mário J. Silva and Cymon Cox	Arlindo Oliviera	Business Model Canvas for ELIXIR	A business is a system that creates value to customers. Accordingly, a business model defines the business concepts (core are value and resource, but others might also exist), their relationships (concepts interdependency) and their dynamics (how resources are acquired and spent, and value is created and delivered). A team representing the four laboratories integrating ELIXIR PT used the Business Model Canvas to define reference business models for the ELIXIR Hub and for its own context, using as sources of information the ELIXIR web sites. We believe this is a very efficient technique to reach a wide range of stakeholders at the business and political levels.According to the findings, the cost structures for the PT Node and the ELIXIR Hub are similar: value driven, with mainly fixed costs for coordination, technical support, development and maintenance. Human resources and IT infrastructure are also key resources common to both contexts.Specific to the ELIXIR Hub are the key resources of reference data models and vocabularies, related with knowledge management as key activity. Other Hub specific key activities are coordination of Nodes, outreach communication, and dissemination, while key activities specific of the PT Node are training and consulting. Key activities common to both contexts are brand value development.For the PT Node the main agreed costumer segment is the biomedical industry and R&D community, with a main value proposition on woody plants (key data resources for eucalyptus, cork oak, pine and grapevine). For key activities, it emerged the development of analytical software tools.	ELIXIR poster	ELIXIR
P_EI008	818	Niklas Blomberg, Friederike Schmidt-Tremmel, Andrew Smith and Manuela Schuengel	Andrew Smith	CORBEL - Harmonisation of access to Europe's biomedical research infrastructures	The Grand Challenges in health can only be met by translation of biomedical discoveries to new, innovative and cost effective treatments. Biological and medical research that addresses these challenges spans a broad range of scientific disciplines and user communities. The ESFRI Biological and Medical Science Research Infrastructures (BMS RIs) sit at the centre of this movement, providing pan-European access to the specialised research services, instruments, samples and facilities that underpin the revolution in life science research and translation.CORBEL, uniting 11 BMS RIs, aims to establish a collaborative and sustained framework of shared services between the participating RIs. CORBEL addresses the critical need of users - particularly those in large advanced research projects - to seamlessly integrate and leverage specialist services from multiple RIs and national centres. Provision of harmonized accession processes, unified ethical and legal support, joint data management, and coordinated user access to advanced research instruments, facilities and samples will boost R&D - from discovery of basic biological mechanisms to applied medical translation.An Open Call for research projects utilising several RIs will be launched in October 2016. These projects will serve as proof-of-concept studies for the envisaged streamlined access to European RIs and will demonstrate its added value for research as well as for the society.	ELIXIR poster	ELIXIR
P_EI009	772	Rob Hooft, Niclas Jareborg, Frederik Coppens, Heinz Stockinger, Robert Pergl and Brane Leskosek	Rob Hooft	Data Management Planning in ELIXIR	The ELIXIR research infrastructure bundles not only the databases and tools of bioinformatics, but it also brings together life science data expertise. The assembled expertise can form a fantastic resource for researchers making a data management plan (DMP); currently, this expertise is hard to discover and access. Several of the ELIXIR nodes are looking for ways to offer DMP services to their communities. The technical coordinators in these nodes are planning to build these services together.First, we will expose the ELIXIR expertise through a web-based data management planning portal, using existing assets:A web technology platform (Czech Republic) to build and manage hierarchical (context-sensitive) questionnaires.A hierarchical analysis of the landscape of life science data management, in the form of a mind map, and associated explanatory text (Netherlands).The ELIXIR e-learning platform (Slovenia).We will also search for collaborations with others providing tools for data management planning across the sciences. Our portal will allow researchers who are making a DMP to find ELIXIR experts they could consult and appropriate learning resources that can help broaden their knowledge. For data stewards the portal will function as a checklist. Important motto will be: Data Management Planning not because we have to, but because it pays off.In addition to this resource, training is needed on many aspects of DMP and this has been prioritized for 2017 by the ELIXIR Training Platform. As a first step, skills needed for various target groups in the ELIXIR community will be identified.	ELIXIR poster	ELIXIR
P_EI010	770	Alba Gutiérrez-Sacristán, Janet Piñero, Núria Queralt-Rosinach, Emilio Centeno and Laura I. Furlong	Janet Piñero	disgenet2r: An R package to explore the molecular underpinnings of human diseases	DisGeNET is a discovery platform designed to answer questions concerning the molecular mechanisms underlying human diseases (http://www.disgenet.org/). DisGeNET follows the FAIR data principles (http://www.datafairport.org/), and can be explored using a suite of tools that includes a web interface, a Cytoscape app, and a SPARQL endpoint. We present disgenet2r, a novel R package for exploring and analyzing DisGeNET. disgenet2r contains a variety of functions for leveraging DisGeNET using the powerful visualization and statistical capabilities of the R environment. disgenet2r is specially designed to harness the large amount of information contained in DisGeNET, facilitating its analysis and interpretation. By integrating different disease vocabularies, disgenet2r eases the exploration of gene-disease associations from different perspectives. It offers different types of visualization, such as heatmaps and networks, and it is especially well suited to explore genes and variants associated to diseases. To allow answering more sophisticated research questions that need the interrogation of heterogeneous data resources, the disgenet2r package leverages the potential of Semantic Web technologies, without the need of special expertise in this area. This is achieved through a set of functions that connect DisGeNET with other resources present in the Linked Open Data, covering different information such as gene expression, gene function, drug activity, and biological pathways, among others. The disgenet2r package (https://bitbucket.org/ibi_group/disgenet2r) expedites the integration of DisGeNET data with other R packages, and allows the development of complex bioinformatic workflows.	ELIXIR poster	ELIXIR
P_EI011	838	Maxim Scheremetjew, Simon Potter, Dario Vianello, Hubert Denise, Alex Mitchell and Rob Finn	Maxim Scheremetjew	EBI's Metagenomics Pipeline: Moving towards cloud computing	EBI metagenomics (EMG, https://www.ebi.ac.uk/metagenomics/) is a free to use hub for the analysis and exploration of metagenomic, metatranscriptomic, amplicon and assembly data. The resource provides rich functional and taxonomic analyses of user-submitted sequences, as well as analysis of publicly available metagenomic datasets that are held within the European Nucleotide Archive (ENA).The pipeline is capable of providing analysis of extremely large datasets. For example, in 2015, we analysed the oceanographic dataset, Tara Oceans, which is the largest project to be processed by EMG to date, with ~10 Tb of sequence data (~29 billion sequences). The pipeline is also able to provide a high level of throughput: the number of analysed datasets within the resource has grown 6-fold in 2016 (as of 22 July 2016), and now comprises over 55,000 sequence runs, with over 200 billion sequences analysed in total.To address future analysis challenges, as metagenomic datasets grow ever larger, we have continued to refine the pipeline, with the aim of improving scalability and portability. As part of this process, we have begun to investigate ways to deploy the pipeline on computing clouds within the ELIXIR hub, as well as commercial clouds, such as Amazon or Google. Here, we give an overview of the analysis pipeline itself, outline a number of updates that we have made to ensure scalability and discuss the ongoing work to deploy it on the cloud.	ELIXIR poster	ELIXIR
P_EI012	785	Mikael Linden, Michal Procházka, Premysl Velek, Susanna Repo, Tommi Nyrönen and Ilkka Lappalainen	Premysl Velek	ELIXIR Authentication and authorization infrastructure	ELIXIR is developing and deploying ELIXIR authentication and authorisation infrastructure (ELIXIR AAI) - a set of general purpose services that support scientific services to authenticate their end users, and to decide what kind of access permissions users have in the services. The end users can benefit from a single login - no need to remember a multitude of usernames and passwords. A well-organised approach to service login and access also increases information security.The first release of ELIXIR AAI is deployed in the ELIXIR-EXCELERATE project, part of the ELIXIR compute platform and scheduled to be operational in the end of August 2016. ELIXIR AAI integrates to components on the ELIXIR compute platform, such as cloud and data transfer services.	ELIXIR poster	ELIXIR
P_EI013	867	José María Fernández González, Juergen Haas, Salvador Capella, Torsten Schwede and Alfonso Valencia	Alfonso Valencia	ELIXIR-EXCELERATE WP2 Activities	Critical benchmarking of scientific tools and services in the different research communities, like the ones registered in the ELIXIR tools registry bio.tools, provides added value to these communities and their developers. Critical benchmarking is based on objective quantitative quality measures, both in terms of technical reliability as well as scientific quality.At the same time, criteria agreed within a community in the form of periodic assessments is an effective way to encourage new developments by highlighting areas which require improvements and/or new solutions. Motivated by the success of CASP, a number of similar community driven benchmarking experiments have been organized e.g. CAPRI, BioCreative, CAGI, CAFA, etc. These experiments have great value in organizing community discussions around new developments and solutions.However, continuous benchmarking efforts are required to compare the tools performance in a steady way over large common data sets. Several efforts have been designed and implemented to address this need in different research areas e.g. EVA, CAMEO, LifeBench, BioCreative Metaserver, CAFASP, BECalm, etc. Note that some of them have been abandoned or superseded by newer ones.ELIXIR-EXCELERATE WP2 aims to bring together different communities needing periodic and/or continuous evaluation of their tools and services. The main targets are: learning from the different benchmarking efforts in order to find commonalities across different benchmarking experiments; trace guidelines and best practices for future research community efforts in order to avoid common problems and pitfalls; and, if possible, defining a standard workflow. infrastructure which is transferable to other scientific communities.	ELIXIR poster	ELIXIR
P_EI014	806	Stephanie Suhr, Susanna Repo and Niklas Blomberg	Susanna Repo	ELIXIR-EXCELERATE: accelerating the implementation of ELIXIR	ELIXIR-EXCELERATE is a major EU Horizon 2020 grant awarded to ELIXIR to help implement its scientific programme and integrate Europe's bioinformatics resources into a coherent infrastructure. It supports ELIXIR's early implementation phase by i) delivering world-leading data services for academia and industry, ii) increasing bioinformatics capacity and competence across Europe, and iii) completing the management and organisational processes for an efficient distributed ELIXIR infrastructure.Funded through a four year grant of nearly €20 million and including over 50 partners from ELIXIR Nodes, the grant will deliver services for users within five technical Platforms (Data, Tools, Interoperability, Compute and Training), which are informed by four domain-specific Use Cases: marine metagenomics, crop and forest plants, rare disease and human data. The technical and scientific activities are complemented by a Capacity building programme, which supports the organisational and scientific development of ELIXIR Nodes. A complimentary training programme is aimed at increasing competency within partner organisations.The successful implementation of ELIXIR-EXCELERATE will enable sustainable management and re-use of data for millions of users across the globe and improve the competitiveness of European life-science industries by providing academia, SMEs and multinationals with the tools to develop new knowledge, products and services.	ELIXIR poster	ELIXIR
P_EI015	861	Salvador Capella-Gutiérrez, Josep Ll. Gelpi and Alfonso Valencia	Salvador Capella-Gutiérrez	ELIXIR-Spain: Activities overview and future perspectives in the context of ELIXIR-EXCELERATE	The Spanish National Bioinformatics Institute (INB) joins ELIXIR in 2015. This virtual institute, created in 2003, is formed by 10 research nodes which altogether cover a broad range of bioinformatics areas. INB nodes have an internationally recognised expertise in the areas of genomics, proteomics, structural biology, and translational medicine. Moreover, INB has contributed to create and maintain a bioinformatics infrastructure through the involvement of the Barcelona Supercomputing Centre.As the ELIXIR node in Spain (ELIXIR-Spain), the INB coordinates the participation of its nodes in this European core infrastructure. INB brings to ELIXIR the experience of many years of distributed work aiming to design, implement and maintain different services from databases e.g. Apris, DisGeNet, etc; to tools e.g. Babelomics, JORCA, GEM3, FlexServ; etc; to databases such as 3D BioNotes, BIGNaSim, etc, to complex infrastructures like the INB-BSC Genomics Cloud.Specifically, INB is participating in the ELIXIR's tools platform helping to the curation of WP1's bio.tools ontology annotations, and developing a platform for continuous benchmarking of tools (WP2) such as text mining, paralogy and orthology predictions, or multiple sequence alignments, among others. In the context of ELIXIR-EXCELERATE, the INB is contributing to setting the foundations of collaborative long standing infrastructures. In fact, INB co-leads two of the four use cases, centered in human data, developing infrastructure for the management of rare-diseases data (WP8) and maintaining the European Genome-Phenome archive (WP9), the long term repository for sensitive human genomics data.	ELIXIR poster	ELIXIR
P_EI016	34	John Hancock	John Hancock	ELIXIR-UK	ELIXIR-UK is the UK Node of ELIXIR. ELIXIR-UK's current focus is on enhancing training capacity and capability both across ELIXIR and within the UK. Chris Ponting from the Node co-leads the ELIXIR Training platform and the UK's ELIXIR training grant. As part of this award ELIXIR-UK is developing the TeSS training portal, led by Terri Attwood. Carole Goble, ELIXIR-UK Interim Head of Node, co-leads the ELIXIR Interoperability platform and plays an important role in developing links internationally, and especially with the USA, in this area. Susanna-Assunta Sansone leads the BioSharing initiative which is central to ELIXIR's interopaerability activities. John Hancock, ELIXIR-UK's Node Coordinator, manages the Node's activities.	ELIXIR poster	ELIXIR
P_EI017	683	Magnus Palmblad, Arzu Tugce Guler, Anna-Lena Lamprecht, Kristian Davidsen, Jon Ison and Veit Schwämmle	Magnus Palmblad	Functional software annotation and automatic workflow generation for mass spectrometry data processing	Many software utilities operating on mass spectrometry (MS) data have been described in the literature. Finding that which one needs is often hard, however. We have added a number of MS-related terms to EDAM and annotated over 200 software tools currently in the public domain, including those on http://ms-utils.org, in the ELIXIR Tools and Data Services Registry http://bio.tools. The ms-utils.org tools emphasizes modular rather than monolithic design. Such small utilities performing one operation with well-defined inputs and outputs are ideally suited for assembly into scientific workflows. Annotating the ms-utils.org content with EDAM terms elevates it to the biotoolsXSD standard, supporting the exposure of these resources in the bio.tools registry, bringing the utilities to a broader audience. We used these annotations to automatically generate workflows in four use cases using loose programming and the jABC framework plugin PROPHETS. The use cases were selected to represent common data analysis tasks in MS-based proteomics: peptide retention time prediction, protein identification and enrichment analysis, localization of phosphorylation and protein quantitation using isotopic labeling. Automatically generating and running different but logically equivalent workflows allows the user to verify their analysis results. Software and service annotations are also useful to find a replacement for a workflow component that is no longer supported. This is the first demonstration of using the EDAM ontology to annotate mass spectrometry software utilities and generate workflows for MS data processing.	ELIXIR poster	ELIXIR
P_EI018	805	Michael Dondrup, Wei Zhang, Frank Nilsen, Zhaoran Zhou and Inge Jonassen	Michael Dondrup	LiceBase – a species focused resource for sea lice – including an RNAi LIMS and tools for data analysis and genome annotation	We present LiceBase, a model organism database and web-portal for genomics of sea lice and other economically relevant marine genomes. Sea lice are the major pathogens affecting the global salmon farming industry. The annual costs for sea lice management have recently been estimated to exceed €500 millions and the aquaculture industry relies on few medicines for lice control. We have recently sequenced and annotated the genome of the Atlantic salmon louse in collaboration with Ensembl and the EBI; large scale RNA-sequencing and reverse genomics experiments are constantly being conducted. The aim of LiceBase is to provide excellent bioinformatics resources for the analysis, retrieval, and visualization of the sea lice genome and related Omics data to the global research community. LiceBase is closely integrated with other Norwegian Elixir applications such as NeLS (Norwegian infrastructure for Life Sciences) Storage and NeLS Galaxy, allowing users to run computational pipelines.LiceBase is a Norwegian international deliverable to Elixir.LiceBase is freely accessible at https://licebase.org.	ELIXIR poster	ELIXIR
P_EI019	726	David Sehnal, Karel Berka, Lukáš Pravda, Radka Svobodová-Vařeková, Michal Otyepka and Jaroslav Koča	Karel Berka	MOLE 3.0 – remastered tool for detection and analysis of functionally important "void spaces" within biomacromolecules	MOLE is a gold standard in quick geometrical detection of channels and tunnels within biomacromolecular structures. MOLE 2.0 (www.mole.upol.cz) was first tool to come with automatic and user–defined detection of channels and tunnels using Voronoi diagram and Delaunay tesselation representations. New version of MOLE 3.0 also enables detection of pores and better description of individual types of important void spaces within protein structures together with additional increase of speed. Alpha version of MOLE 3.0 is available at http://webchemdev.ncbr.muni.cz/MOLE3/.	ELIXIR poster	ELIXIR

P_EI020	666	Klaas Vandepoele	Klaas Vandepoele	PLAZA 3.0: an access point for comparative and regulatory genomics in plants	Comparative sequence analysis has significantly altered our view on the complexity of genome organization and gene functions in different kingdoms. PLAZA 3.0 is designed to make comparative genomics data for plants available through a user-friendly web interface. Structural and functional annotation, gene families, protein domains, phylogenetic trees, and detailed information about genome organization can easily be queried and visualized. Compared with the first version released in 2009, the number of integrated genomes is more than four times higher, and now covers 37 plant species. The new species provide a wider phylogenetic range as well as a more in-depth sampling of specific clades, and genomes of additional crop species are present. The functional annotation has been expanded and now comprises data from Gene Ontology, MapMan, UniProtKB/Swiss-Prot, PlnTFDB and PlantTFDB. Furthermore, we improved the algorithms to transfer functional annotation from well-characterized plant genomes to other species. Recently, more than 1 million of conserved non-coding sequcenes were added for ten dicot species, which provide detailed information about conserved transcription factor (TF) binding sites for 642 TFs covering 35 TF families. These new data and features make PLAZA 3.0 (http://bioinformatics.psb.ugent.be/plaza/) a versatile and comprehensible resource for users wanting to explore genome information to study different aspects of plant biology, both in model and non-model organisms.PLAZA 3.0: an access point for plant comparative genomics. Proost et al., Nucleic Acids Res. 2015A Collection of Conserved Non-Coding Sequences to Study Gene Regulation in Flowering Plants.Van de Velde et al., Plant Physiol. 2016	ELIXIR poster	ELIXIR
P_EI021	429	Konstantinos D. Tsirigos, Arne Elofsson and Pantelis G. Bagos	Konstantinos D. Tsirigos	PRED-TMBB2: Improved topology prediction and detection of beta-barrel outer membrane proteins	PRED-TMBB was presented for the first time in 2004 and is one of the most cited methods regarding the topology prediction and detection of beta-barrel outer membrane proteins. Here, we present an update to this method, PRED-TMBB2, which contains several new features that improve its performance significantly. The major difference is the incorporation of evolutionary information in the form of Multiple Sequence Alignments (MSAs), which drastically improves the topology prediction capability and makes it able to achieve higher performance compared to all other available methods. At the same time, the single-sequence version of PRED-TMBB2 manages to perform better than almost all methods regarding detection of beta-barrel proteins in large datasets, outperforming even methods that use MSAs and are much slower. The combination of single- and multiple-sequence version of PRED-TMBB2 is something unique and we anticipate it will be of great interest to researchers in this field.	ELIXIR poster	ELIXIR
P_EI022	713	Andrew Nightingale, Jie Luo, Leyla Jael Garcia Castro, Maria Martin and Uniprot Consortium	Andrew Nightingale	Protein Data Services and Feature Viewer Enabling Knowledge Driven Research	Complex biological processes, such as rare heterogenetic diseases, are difficult to discover and interpret. Coupled with the continuous growth and complexity in Biological data there is a requirement to develop tools for data linkage, integration and visualization to facilitate scientific progress that can contribute to essential infrastructures such as those provided by Elixir.In order to respond to this challenge and contribute to the Elixir effort, we have developed REST services and a BioJS component for accessing and visualizing protein data, while also ensuring interoperability with other tools and resources. This will enable users to fully transition from the genome, to the transcriptome and to the proteome and thus facilitate knowledge driven biomedical research. These services use a number of resources including UniProtKB as the source for proteins and functional information and Ensembl for genomic information and have a flexible design that can be extended to incorporate data from further resources.For example, our services include protein mappings to genomic coordinates and variation data, enhanced with proteomics experiments. Following simple instructions a novice user can quickly learn to carry out advanced searches tailored to their scientific needs. Based on these services, we have developed a new interactive visualization BioJS component depicting sequence functional annotations from UniProtKB such as domains, sites, PTMs and variants from multiple sources. This ‘Feature Viewer’ presents curated and large-scale experimental data in an intuitive compact picture with related protein annotations grouped together in zoomable tracks in a similar way to tracks in genome browsers.	ELIXIR poster	ELIXIR
P_EI024	534	Margarita C. Theodoropoulou, Konstantinos D. Tsirigos, Stavros Hamodrakas and Pantelis G. Bagos	Pantelis Bagos	Recent updates in the Database of Outer Membrane Proteins (OMPdb) in 2016	Beta-barrel outer membrane proteins (OMPs) are crucial for the life of Gram-negative bacteria, since they participate in many diverse procedures. OMPdb (http://www.ompdb.org/) is the largest, most complete and well characterized collection of OMPs from Gram-negative bacteria. Our database contains extensive information for each protein (entry) including protein description and classification, sequence, organism name, taxonomy, links to other databases, accompanied with annotation for TM segments and signal peptides. All proteins are classified into families based on function and sequence similarity. Each family (family entry) is extensively described and the information provided are the function of protein members, literature references, a list of proteins with 3D-structure (if any), and the respective seed and full protein alignments. Currently, OMPdb contains 91 families and more than 400,000 proteins. Out of the 91 families, 15 families were built completely from scratch, 16 do not belong to the respective clan of Pfam, while 6 of them are annotated as DUF in Pfam. OMPdb follows the monthly updates of Uniprot through a semi-automated procedure. Users may search the database using Text, Domain and/or BLAST Search and the database can be downloaded in several formats (text, FASTA, XML) through the Download page. We are now in collaboration with Pfam and TCDB in order to cross-link our databases. Finally, our database is coupled with PRED-TMBB2, the best performing algorithm for the topology prediction and detection of OMPs. We believe that OMPdb is valuable tool in the hands of researchers working with this important superfamily of transmembrane proteins.	ELIXIR poster	ELIXIR
P_EI025	740	Diana Domanska and Abdulrahman Azab	Diana Domanska	Software Provisioning Inside a Secure Environment as Docker Containers using STROLL File-system	TSD (Tjenester for Sensitive Data), is an isolatedinfrastructure for storing and processing sensitive research data,e.g. human patient genomics data. Due to the isolation ofthe TSD, it is not possible to install software in the traditionalfashion. Docker containers is a platform implementinglightweight virtualization technology for applying the build-oncerun-anyware approach in software packaging and sharing. Thispaper describes our experience at USIT (The University Centreof Information Technology) at the University of OsloWith Dockercontainers as a solution for installing and running softwarepackages that require downloading of dependencies and binariesduring the installation, inside a secure isolated infrastructure.Using Docker containers made it possible to package softwarepackages as Docker images and run them smoothly inside oursecure system,TSD. The paper describes Docker as a technology,its benefits and weaknesses in terms of security, demonstratesour experience with a use case for installing and running theGalaxy bioinformatics portal as a Docker container inside theTSD, and investigates the use of STROLL file-system as a proxybetween Galaxy portal and the HPC cluster.	ELIXIR poster	ELIXIR
P_EI026	866	Jiri Vondrasek	Jiri Vondrasek	Structural Bioinformatics and Cheminformatics - the major focus of the Czech ELIXIR Node	Building sustainable infrastructure for biological data involves synergy of compatible resources as well as corresponding tools and services. The Czech ELIXIR Node comprises several high level solutions for structural bioinformatics, cheminformatics and genomic data available at the national as well as international level. The portfolio of these tools and services represents advanced scientific methods and results available via progressive technical solutions. A small number of selected tools are presented.For Cheminformatics we introduce solution utilizing Resource Description Framework (RDF) and the SPARQL query language applied on Integrated Database of Small Molecules. In the field of Structural Bioinformatics we present 4 complex tools including PatternQuery – a tool for detection structural fragments in biomacromolecules, (Multi)SETTER – SEcondary sTructure-based TERtiary structure superposition tools, program MOLEonline 2.0 which determines channels and pores in 3D structures of proteins and finally the DNATCO – a tool for DNA conformers assignment.Frequently used tool developed and curated by the Czech Node is Repeat Explorer which is dedicated to discover and identify repeats in NGS data. RepeatExplorer as well as other presented tools are a part of services provided by ELIXIR CZ - Czech research infrastructure for biological information. For other services provided by ELIXIR's Czech Republic Node visit www.elixir-czech.cz/services	ELIXIR poster	ELIXIR
P_EI027	829	Niall Beard, Aleksandra Nenadic, Susanna-Assunta Sansone,Terri Attwood,Carole Goble, Rafael Jiminez,Milo Thurston, Norman Morrison, Celia. Van Gelder and Fredrick Coppens	Niall Beard	Structured Data for Life Science using Schema.org	ELIXIR explicitly supports the FAIR Principles - Findable, Accessible, Interoperable, Reusable - for its data, software, tools, events and training resources. “Finding” has the significant challenge of effective discovery and indexing of web-based resources across all ELIXIR information providers - this is an issue because there has been no agreement within ELIXIR about how to expose such resources in order to make them discoverable.One solution to this problem is to adopt Schema.org mark-up. Schema.org is a community initiative supported by four major search-engine providers: Google, Bing, Yahoo and Yandex. It provides a simple way to publish data in a standard format.If websites publishing life-science training materials, data, tools, profiles etc. were to use Schema.org mark-up, then their websites could be crawled, and the data could be indexed and exposed in searchable portals. However, this approach has challenges. Bioschemas is a newly formed community group in the life sciences to address these challenges, aiming to make the adoption of Schema.org part of a powerful way to discover and collect life-science information. It produces specifications, one for each information type (‘Training Course’, ‘Event’, etc.). Each specification lists the Schema.org minimum properties expected, and the contraints for each property: for example, the ‘Events’ specification that the property ‘topic’ should be an EDAM ontology topic. The specifications are developed openly and are available on GitHub, with the support of existing communities of domain experts. Bioschemas also identifies the types or properties that are needed in the life sciences but not present in Schema.org, and works with the community to encourage the adoption of these types and properties into Schema.org.	ELIXIR poster	ELIXIR
P_EI028	831	Herve Menager and Edam-Core Edam Core Team	Herve Menager	The EDAM Ontology	EDAM is an ontology of well established, familiar concepts that are prevalent within bioinformatics, including types of data and data identifiers, data formats, operations and topics. EDAM is a simple ontology - essentially a set of terms with synonyms and definitions - organised into an intuitive hierarchy for convenient use by curators, software developers and end-users.EDAM is suitable for large-scale semantic annotations and categorization of diverse bioinformatics resources, and also suitable for diverse application including for example within workbenches and workflow-management systems, software distributions, and resource registries. Version 1.15 of EDAM has been released. Contributions and suggestions are welcome!	ELIXIR poster	ELIXIR
P_EI029	786	Ilkka Lappalainen, Jordi Rambla, Serena Scollen, Mikael Linden, Macha Nikolski, J. Dylan Spalding and Susanna Repo	Serena Scollen	The ELIXIR Beacon Project	ELIXIR has partnered with the Global Alliance for Genomics and Health (GA4GH) to light ELIXIR Beacons as primary data-discovery services for genomics. The Beacon will provide a single point of access to the data stored within the Node resources by promoting interoperability and standard technical data access interfaces. The ELIXIR Beacon project defines the Beacon query interface, user authentication and authorization mechanisms and the service security requirements together with the GA4GH. The ELIXIR Beacon reference implementation is fully integrated with the ELIXIR authentication and authorization services. It is designed to work with research consented sensitive human data as well as data from other organisms.The ELIXIR Beacon service has three distinct data access tiers. These tiers are designed to provide increased level of information for Beacon users with access rights based on data security and consent requirements.The Public Access Tier does not require user to authenticate before querying on data such as allele frequencies on national population or non-human data. Registered Access Tier covers allele frequencies on individual cohorts used for constructing national level allele frequencies. Data that require Data Access Committee approval is provided only to approved researchers through Controlled Access Tier. This open web service is designed to be technically simple, easy to implement, and to not return privacy violating information.The ELIXIR Beacon project includes partners from EMBL-EBI, Belgium, Finland, France, Netherlands, Spain, Sweden and Switzerland.	ELIXIR poster	ELIXIR
P_EI030	871	Sveinung Gundersen, Matúš Kalaš, Boris Simovski, Brynjar Rongved, Henrik Skifjeld, Sivert Kronen Hatteberg, Abdulrahman Azab, Osman Abul, Arnoldo Frigessi,Geir Kjetil Sandveand Eivind Hovig	Sveinung Gundersen	The GTrack ecosystem - expressive file formats for analysis of genomic track data	GTrack, BTrack and GSuite are file formats designed to handle genomic track data of heterogeneous types. The file formats are designed to complement each other and work jointly as a complete ecosystem for representation and analysis of most types of data that can be located along a reference genome.GTrack is a tabular format that was developed to provide a uniform representation of most types of genomic datasets, being able to replace common formats such as WIG, GFF , BED, BED-like formats, and even FASTA. GTrack supports all possible track types, mathematically defined as a delineation of possible genomic datasets into 15 different basic informational structures. In addition to common track types such as points or segments, this includes 8 types of tracks usable for analysis of the three-dimensional aspects of DNA.The BTrack format supports the same variety of informational content as GTrack, but in a binary form. BTrack is unique in supporting a collection of multiple tracks stored together in one (possibly compressed) HDF5-based binary file, while still supporting a high level of efficiency.The GSuite format is a unique tabular format that binds together the whole chain of multi-track analysis, from search and retrieval of genomic tracks, through intermediate processing, to analysis.A Python library supporting parsing, conversion and operations is available with a rudimentary API. The BTrack format is supported only in a prototype version.The GTrack ecosystem has, together with BioXSD, been selected as one of four main national deliverables from Norway towards the ELIXIR project.	ELIXIR poster	ELIXIR
P_EI031	762	Frederic B. Bastian, Julien Roux, Mathieu Seppey, Komal Sanjeev, Valentine Rech de Laval, Philippe Moret, Panu Artimo, Séverine Duvaud, Vassilios Ioannidis, Heinz Stockinger and Marc Robinson-Rechavi	Frederic B. Bastian	TopAnat : a new way to understand genomics results using gene expression enrichment in anatomy	TopAnat is an innovative tool to discover where a set of genes is preferentially expressed, and it represents a completely new kind of enrichment analyses. TopAnat is quite similar to a Gene Ontology (GO) enrichment test, which determines the GO terms preferentially associated to a set of genes. In our case, however, the test is applied to terms from an anatomical ontology (Uberon ontology), mapped to genes by expression patterns. This allows to study a new type of property of gene sets, regarding their expression domains. TopAnat is both highly sensitive for detecting organs where genes have an expression biais, and specific to provide the most relevant and precise terms. For instance, we used TopAnat to analyze the expression domains of genes associated with autistic end epileptic disorders in human, from Jabbari and Nürnberg, 2016; TopAnat succesfully determined that these genes were preferentially expressed in some specific brain regions, likely to be associated with these disorders (see http://bgee.org/?page=top_anat#/result/8fce889da7b4519c57925736ed3933032c8122819/). Note that TopAnat is not to be confused with a differential gene expression analysis, where gene expression levels are compared between two conditions, to detect changes in expression. Rather, TopAnat retrieves the anatomical structures where genes are expressed, and for each anatomical structure, tests whether genes from the list of interest are over-associated with this structure, as compared to a background list of genes. TopAnat is available as a webtool (http://bgee.org/?page=top_anat), and as a Bioconductor R package (https://bioconductor.org/packages/release/bioc/html/BgeeDB.html).	ELIXIR poster	ELIXIR
P_EI032	784	Ian Sillitoe, Natalie Dawson, Paul Ashford, Sayoni Das, Su Datt Lam, Jon Lees, Millie Pang and Christine Orengo	Natalie Dawson	Using CATH-Gene3D to explore the impacts of disease-induced genetic variations	CATH classifies 3D structures from the PDB into superfamilies of protein domains that are evolutionarily related. Since protein structure tends to be much more highly conserved than sequence, CATH superfamilies are often able to trace further back in evolution than sequence methods alone. Currently, CATH classifies more than 300,000 domain structures (from ~80% of PDB structures) into ~2700 evolutionary superfamilies.Once these distant structure-based evolutionary relationships have been established, the Gene3D resource uses start-of-the-art sequence comparison technology to augment these superfamilies with more than 50 million protein domain sequences from ~20,000 cellular genomes. Many of these superfamilies contain protein sequences with detailed functional annotations, which enable a deep understanding of the evolutionary mechanisms by which functions evolve. A recent development is the identification of functional families within CATH superfamilies and the establishment of a new function prediction protocol, which has been highly ranked by the CAFA independent assessment.CATH-Gene3D is an endorsed resource of the UK ELIXIR Node and our HMM libraries are used to provide structural and functional annotations in the ELIXIR marine metagenome infrastructure use case. CATH-Gene3D is also a member of the Genomics England Functional Effects Domain, for which we are using our function family data to explore the impacts of disease-associated residue mutations and identify the specific protein domains that are enriched in such mutations. CATH-Gene3D is a member of a consortium of UK structural bioinformatics groups contributing to an ELIXIR training initiative that is developing web-based training workflows to analyse the impacts of mutations.	ELIXIR poster	ELIXIR
P_EI033	477	Eric Bonnet, Yimin Shen, Xavier Benigni, Nizar Touleimat, Jorg Tost, Jean-François Deleuze and François Artiguenave	Eric Bonnet	WBS: a computational pipeline for the treatment of whole-genome high-throughput bisulfite sequencing data	DNA methylation is an important epigenetic mechanism used by higher eukaryotes and is involved in several key physiological processes, including regulation of gene expression, X-chromosome inactivation, imprinting and silencing of germline-specific genes and repetitive elements. Patterns of methylation are maintained through somatic cell divisions and may be inherited across generations. These patterns are altered in many complex human diseases, such as imprinting disorders and cancer. Understanding methylation patterns is therefore of great importance for many biomedical questions.Bisulfite treatment of DNA is method of choice to analyse these patterns. Bisulfite treatment leaves methylated cytosines unaffected. Thus, bisulfite treatment introduces specific changes in the DNA sequence that depend on the methylation status of individual cytosine residues, yielding single-nucleotide resolution information about the methylation status of a segment of DNA. Various analyses can be performed on the altered sequence to retrieve this information. Especially, rapidly falling costs of high-throughput sequencing have made the global analysis of DNA methylation at the whole genome level a viable option.However, there are significant computational challenges associated with the computational treatment of bisulfite generated reads. Here we describe WBS (Workflow Bisulfite), a computational pipeline set-up at the Centre National de Génotypage (CNG) for the analysis of bisulfite whole genome sequencing data. The pipeline is built around standard state-of-the-art tools and workflows for the treatment of bisulfite reads. We describe the organization of the pipeline, performance and possible evolutions in the framework of the analysis of mammalian (mostly human) whole genome DNA methylation patterns.	ELIXIR poster	ELIXIR
ELIXIR/TRAINING							

P_El/Tr034	339	Teresa K. Attwood, Louisa Bellis, Cath Brooksbank, Pedro L Fernandes, Valerie Florance, Rita Hendricusdottir, Lee Larcombe, Patricia M. Palagi, Celia W.G. van Gelder, Allegra Via, Sarah L. Morgan, Gabriella Rustici and Rochelle E. Tractenberg	Louisa Bellis	Assessing the impact(s) of international bioinformatics & computational biology training within ELIXIR & BD2K	Two large-scale initiatives have recently been created – one in the USA (Big Data to Knowledge, BD2K) and one in Europe (ELIXIR) – with emphasis on training and capacity building to promote, respectively, biomedical and life science research in the current, dynamic context of big data and bioinformatics. Definitions and metrics of success and impact for the training being developed (ELIXIR) and marshalled (BD2K) are needed. These should include quantitative and/or qualitative indicators of whether, how and to what extent the training delivered is: (1) successful, based on its stated goals, and (2) aligned with the driving strategies of the above-mentioned initiatives.Determining the impact of training is a challenging task that requires: 1) a definition of "impact"; 2) concrete understanding of the purposes and the corresponding stakeholders to which "demonstrating training impact" might be useful (i.e., how will the outcome of our analyses be used, and how will this affect future decisions); 3) articulation of what types of indicator/metric/measurement of impact are best to use; and 4) determination of the most appropriate strategies to collect such data.BD2K and ELIXIR share a commitment to identify the most reliable and robust indicators, to collect and analyse the relevant data, and to develop and publish guidelines. The two groups have already met, and continue to work to align their efforts, to share and discuss their results, and to promote globally useful definitions and metrics for training impact and success.	ELIXIR/Training poster	ELIXIR Training
P_El/Tr035	841	Brane Leskosek, Eija Korpelainen and Jure Dimec	Maja Zagorščak	Node collaboration through the ELIXIR e-learning platform – follow up	The eLearning platform developed by ELIXIR-SI (EeLP) enables remote execution of the f2f courses, so that teacher can be in one location and students on remote and distributed locations. EeLP offers secure access to the training materials, presentations, exercises and assessment systems in the form of online lessons, discussion forums for teacher and students, as well as single-sign-on using eduGain authentication (ELIXIR AAI is in preparation).In order to gain experience and to develop best practises, we transformed the popular ELIXIR-FI course "RNA-seq data analysis with Chipster" into e-learning format, and successfully executed the course with the teacher in FI, students in CZ and e-learning materials and video conference system (VC) in SI. The communication between the teacher and students was conducted through a two-way VC. Students downloaded the analytics software and data sets from the ELIXIR-FI servers. Guided online by the teacher, the students performed their data analysis tasks and were constantly assessing themselves in the EeLP that also allow the teacher to follow their progress.We are planning the EeLP and e-learning services to be a long-term activity. With simple web based system (https://elixir.mf.uni-lj.si/elearning/) we collect information about f2f courses for which authors are interested in transformation to appropriate e-learning formats. The courses from SE and IT nodes are already being transformed.Experiences so far shows that EeLP combined with lectures over VC is a scalable and cost-effective way complementary to f2f training and capacity building, and it could be used for training researchers, developers and infrastructure specialists.	ELIXIR/Training poster	ELIXIR Training
P_El/Tr036	822	Niall Beard, Terri Attwood and Aleksandra Nenadic	Terri Attwood	TeSS - The Life Science Training Portal	TeSS [1] (ELIXIR's life science training portal) has been in development since early 2015. Following a proof-of-concept (pilot) phase, funding was received (as part of ELIXIR-EXCELERATE) to harden the product and bring it to a production-level service.TeSS aggregates links to disparate training materials and events scattered around the institutional websites of ELIXIR Nodes and other content providers (GOBLET [2], Software and Data Carpentry [3, 4], EBI TrainOnline [5], Genome3D [6], on-course [7], etc.), making them centrally discoverable and searchable. Training resources within TeSS can be collected and arranged into packages and/or training workflows, which are graphic representations of scientific pipelines to organise resources into easily navigable views. Aggregation of training content happens automatically through a set of custom-made nightly-run scraper scripts. Scrapers use a number of techniques to extract information: HTML-scraping and APIs had been the predominent methods, but more recently we have focused on parsing structured schema.org mark-up data. The TeSS team has been heavily involved in the specification definition and promotion of the adoption of a schema.org standard for describing training materials and events online through the BioSchemas [8] group. We are currently developing an integration strategy with other ELIXIR registries (e.g., BioTools [9] and BioSharing [10]) to link training materials to relevant tools, databases, standards and policies.[1] https://tess.elixir-uk.org[2] http://mygoblet.org[3] http://software-carpentry.org[4] http://www.datacarpentry.org[5] http://www.ebi.ac.uk/training/online/[6] http://genome3d.eu/[7] https://www.on-course.eu/[8] http://bioschemas.org	ELIXIR/Training poster	ELIXIR Training
P_El/Tr037	617	Bjoern Gruening and The De.Nbi Special Interest Group Training And Education	Bjoern Gruening	The de.NBI Training Network	The German Network for Bioinformatics Infrastructure (de.NBI) provides a nationwide infrastructure for bioinformatics tools, resources, and training for these tools funded by the German Ministry for Research and Education. Consequently, de.NBI develops and collects educational materials related to bioinformatics. Training activities are focused on supporting and training end users through training courses, webinars, and online training. Life science researchers will thus be enabled to exploit their data more effectively by applying tools, standards and compute services provided by de.NBI. The network has been offering a number of training activities ranging from summer schools, hands-on trainings, hackathons to online teaching activities. In 2016, de.NBI will organize around 40 training courses and thus aims to train around 500 participants.The different training courses are adapted to different levels of users. The range goes from beginners' courses up to expert meetings. Thematically, these courses provide expertise in the application of de.NBI tools and databases, programming skills, data management as well as data interpretation.de.NBI has developed internal standards for monitoring the quality of its educational events. Standardized survey forms permit the comparison of the quality of individual training events and yield invaluable feedback to instructors. In order to scale up the training efforts, we providing its training materials online. de.NBI will join the European ELIXIR network in 2016 and will thus integrate its training activities into the ELIXIR activities. de.NBI training activities are accessible online on the network's website at http://www.denbi.de.	ELIXIR/Training poster	ELIXIR Training

POSTER LIST
ORDERED ALPHABETICALLY BY POSTER TITLE
GROUPED BY THEME/TRACK

THEME/TRACK: GENES

Poster numbers: P_Ge001 - 062 **Application posters: *P_Ge001 - 004***

Poster number	EasyChair number	Author list	Presenting author	Title	Abstract	Theme/track	Topics
APPLICATION POSTERS WITHIN GENES THEME							
P_Ge001	*803*	*Haruka Ozaki and Itoshi Nikaido*	*Haruka Ozaki*	*ATAC2NET: A pipeline for reconstructing gene regulatory network based on ATAC-Seq data*	*Reconstruction of gene regulatory networks are important for understanding cell differentiation, cellular functions, and disease progression. Digital genomic footprinting using DNase I-Seq and ATAC-Seq can profile genomic occupancies of several hundreds of transcription factors in the same biological context at once. Thanks to the convenience of performing ATAC-Seq experiments, genome-wide chromatin accessibility data have been accumulating in public repositories, providing resource for reconstructing gene regulatory networks. However, although several studies evaluated performance of footprint detection programs for predicting TF binding, no systematic evaluations have been performed on computational methods for reconstructing gene regulatory network based on detected footprints. Moreover, it is unclear whether footprint detection programs designed for DNase I-Seq data is also effective for ATAC-Seq data. Here, we systematically evaluated the performance of computational methods for detecting footprints as well as for reconstructing gene regulatory networks using ATAC-Seq data. We showed that prediction performance was affected by properties of transcription factors, DNA amounts, and library sizes, rather than experimental methods used. We found that the reconstructed networks showed cell-type specificities properties, suggesting their biological significance. Based of these results, we developed ATAC2NET, a pipeline for reconstructing gene regulatory network based on ATAC-Seq data. We applied it several ATAC-Seq datasets and we are analyzing in detail the properties of the resulted networks. In addition, we are currently evaluating an alternative network reconstruction approach combining ATAC-Seq and gene expression data.*	*Genes/ Application poster*	*Application Fundamental*
P_Ge003	*802*	*Shishir Gupta, Roy Gross and Thomas Dandekar*	*Shishir Gupta*	*Re-annotation of the ant Camponotus floridanus genome, comprehensive analysis of its immune transcriptome and general reconstruction of ant interactomes*	*The sequencing of several ant genomes within the last six years open new research avenues for understanding not only the genetic basis of social insect species but also the complex systems such as immune responses. To form a better view of the immune repository and study the carpenter ant Camponotus floridanus immune responses against the bacteria, experimental data from Illumina sequencing and mass-spectrometry (MS) data in normal and infectious conditions for larvae and adults are analysed and integrated with bioinformatics approaches such as interactomics. Besides infection induced transcriptome profiling the data generated from Illumina sequencing was used for improving existing annotations, identifying alternative transcripts and functional modules in interactomes. We present our latest pipeline for gene prediction and alternative transcripts. The pipeline uncovered 1928 genes affected by alternative splicing events coding for 4666 alternative transcripts in C. floridanus. Current results allow better structural and functional annotation of C. floridanus genome, annotation of alternative transcripts, characterization of the immune system, transcriptome profiling and infection induced subnetworks of C. floridanus. Moreover, we analyze the protein-protein interactions (PPIs) of C. floridanus immune system with pathogenic bacteria such as Serratia marcescens and with the endosymbiont Blochmannia floridanus. We found that the immune system of C. floridanus is equally rich as in other insects, including diverse antimicrobial peptides and does not rely more on 'social immunity' as other social insects. Furthermore, our results indicate strong activation of immune defense in larvae while protection in adults depends mostly on ROS-mediated immunity.*	*Genes/ Application poster*	*Application Biotechnology*
P_Ge004	*864*	*Antonio Colaprico, Tiago Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S. Sabedot, Tathiane M. Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi and Houtan Noushmehr*	*Antonio Colaprico*	*TCGAbiolinks: An R/Bioconductor package for integrative analysis with TCGA data*	*The Cancer Genome Atlas (TCGA) research network has made public a large collection of clinical and molecular phenotypes of more than 10 000 tumor patients across 33 different tumor types. Using this cohort, TCGA has published over 20 marker papers detailing the genomic and epigenomic alterations associated with these tumor types. Although many important discoveries have been made by TCGA's research network, opportunities still exist to implement novel methods, thereby elucidating new biological pathways and diagnostic markers. However, mining the TCGA data presents several bioinformatics challenges, such as data retrieval and integration with clinical data and other molecular data types (e.g. RNA and DNA methylation). We developed an R/Bioconductor package called TCGAbiolinks to address these challenges and offer bioinformatics solutions by using a guided workflow to allow users to query, download and perform integrative analyses of TCGA data. We combined methods from computer science and statistics into the pipeline and incorporated methodologies developed in previous TCGA marker studies and in our own group. TCGAbiolinks downstream analysis can be divided into 1) supervised analysis, comprising differential expression analysis, enrichment analysis, and master regulator analysis or 2) unsupervised analysis,; comprising inference of gene regulatory network, cluster, classification, ROC, AUC, feature selection, and survival analysis. Using four different TCGA tumor types (Kidney, Brain, Breast and Colon) as examples, we provide case studies to illustrate examples of reproducibility, integrative analysis and utilization of different Bioconductor packages to advance and accelerate novel discoveries.*	*Genes/ Application poster*	*Application*
OTHER POSTERS WITHIN GENES THEME							
P_Ge005	733	Aubin Samacoits, Florian Mueller and Thomas Walter	Aubin Samacoits	3D FISH image simulation framework to develop analysis method for mRNA localization	Many studies have characterized gene expression at the genome-wide level, but focused mostly on expression levels. However, only few studies focus on another key parameter: sub-cellular mRNAs localization. With single molecule FISH (smFISH) it is now possible to visualize individual mRNA molecules and hence investigate their spatial distribution in individual cells. However, to perform these analyses, several computational tools are necessary. First, cells need to be segmented and individual mRNA molecules be detected. While these image analysis tools are already well developed, there currently exists no validated statistical framework for the analysis of the mRNA localization. In such an analysis, the spatial coordinates of mRNAs are mapped into a carefully designed feature space. From this representation, machine-learning analysis will be performed to identify different mRNA localization classes, and eventually group genes according to their mRNA localization. In order to carefully develop and validate these feature sets and the subsequent machine-learning pipeline, an annotated image database with different localization classes is needed. Such validation databases exist already for cell segmentation or protein localization, but not for mRNA localization. Here, we present a virtual cell environment to simulate smFISH with non-random 3D mRNA localization. We base these simulations on experimental data, providing accurate 3D contours for cells and nuclei. Further, mRNAs are simulated considering realistic variations in their intensity and different experimentally observed localization patterns. Taken together, our approach yields realistic smFISH images, which can provide the basis for the development of a machine learning approach for mRNA localization classification.	Genes poster	Biotechnology
P_Ge006	814	Tine Goovaerts, Sandra Steyaert, Jeroen Galle, Tim De Meyer and Wim Van Criekinge	Tine Goovaerts	A mixture model for the omics based identification of monoallelically expressed loci and their deregulation in cancer	Imprinting is an epigenetic phenomenon leading to the expression of a single allele in a parent-of-origin specific manner. Inadequate computational techniques restrict insight in imprinting and diseases associated with imprinting deregulation, such as cancer. Hence, we introduce a mixture model for the identification of monoallelically expressed loci based on large scale omics data and a method to identify samples and loci featured by loss of imprinting.Our rationale is that RNA-seq (or similar omics data) for monoallelically expressed loci will exhibit apparent deviation from the Hardy-Weinberg equilibrium (HWE). As only one allele is expressed or epigenetically modified, heterozygous samples will ideally be recognized as homozygous. The model hence detects those loci in which the observed heterozygous fraction is shifted towards the homozygous fractions. Furthermore, it does not rely on prior genotyping and takes into account sequencing errors and possible partial imprinting. Once imprinted loci have been identified in control data, loci featured by loss of imprinting in the pathology under study can be identified.The model enabled the identification of 140 imprinted SNPs in 113 healthy control samples of the TCGA breast cancer RNA-seq data, corresponding to 53 genes. Some well-known imprinted loci, such as IGF2 and PEG3, were detected. Deregulation of these loci was investigated in 506 breast cancer samples from TCGA. Loss of imprinting - i.e. re-expression of the silenced allele - was observed in 8 of the imprinted SNPs.	Genes poster	Fundamental
P_Ge007	753	Lisa Barros de Andrade E Sousa and Annalisa Marsico	Lisa Barros de Andrade E Sousa	A statistical model for epigenetic regulation of miRNAs	microRNAs are small, non-coding RNAs involved in post-transcriptional gene regulation. Since the dysregulation of only a few miRNAs can affect many biological pathways, miRNAs are thought to play a key role in cancer development and can be used as biomarkers for cancer diagnosis and prognosis. In order to understand how miRNA dysregulation leads to a cancer phenotype it is important to determine the basic regulatory mechanisms that drive miRNA expression. Although much is known about miRNA-mediated post-transcriptional regulation, little is known about the epigenetic control of miRNAs. Here, we performed cell-line specific miRNA promoter predictions and built a classification model for expressed and non-expressed miRNAs. The classification model is based on several epigenetic features, e.g. histone marks and DNA methylation at both, miRNA promoters and around miRNA hairpins. We were able to classify intragenic and intergenic miRNAs with an accuracy of 79% and 85%, respectively, and identified the most important features for classification via feature selection. Surprisingly, we found that DNA methylation seems to have a dual role in regulating miRNA expression at transcriptional level: at promoters, high levels of DNA methylation correlate with transcriptional repression, while around miRNA hairpins high levels of DNA methylation have a positive impact on the expression level of the mature miRNA.	Genes poster	Fundamental
P_Ge008	709	Virag Sharma, Bjoern Langer, Leo Foerster, Pradeep Kiruvale and Michael Hiller	Virag Sharma	A Systematic Approach to Identify Gene Losses using Genome Alignments	Inactivation of protein-coding genes in different species is an important type of genomic change that can explain phenotypic differences among these species. For example, the loss of the Gulo gene in some mammals explains their inability to synthesize Vitamin C. While mutations in gene sequences can be detected from genome alignments, there is no method to systematically detect gene losses in an automated fashion.We have developed a computational pipeline that systematically searches for gene losses across different species without requiring any manual curation. Given a reference species and a genome alignment of the reference species with other species, our pipeline is able to identify the different types of gene inactivating mutations such as frameshifts and in-frame stop codons. To avoid mistaking artifacts for inactivating mutations, we strictly control for assembly gaps, low quality genomic sequences, alignment issues such as processed pseudogene misalignments and changes in gene structures.In order to associate gene losses with phenotypic changes, we applied the pipeline on a multiple genome alignment of 29 species with mouse as the reference and focused on gene losses in mammals which are either completely or partially blind. Our pipeline reports the inactivation of several known and potentially novel genes involved in vision-related functions. We conclude that the pipeline is a valuable tool for the compilation of a gene-loss catalogue for genomes that will be sequenced in the future. Furthermore, the pipeline will provide the basis to systematically link phenotypic changes to genomic changes using approaches like Forward Genomics.	Genes poster	Fundamental

P_Ge009	404	Patrick van den Berg, Stefan Semrau and Nikolai Slavov	Patrick van den Berg	An integrated transcriptomics and proteomics study of embryonic stem cell differentiation	Embryonic stem cells (ESCs) can be differentiated into all cell types of the adult body. In vitro differentiation of ESCs has therefore been used extensively as a model for embryonic development and it is critical for applications of ESCs in regenerative medicine and disease modeling. To differentiate ESCs into well-defined cell types, precise manipulation of gene expression is necessary. The majority of existing work has focused on transcriptional regulation of expression. Here, we study gene regulation at the level of protein turnover (translation and degradation) to discover novel ways to control ESC differentiation. In particular, we extracted mRNA and protein during retinoic acid induced differentiation of mouse ESCs. mRNA and protein abundance were then quantified by RNA sequencing and mass spectrometry, respectively. The measurement of 10 samples during a 96 h differentiation time course allowed us to follow the expression dynamics with unprecedented temporal resolution. We have developed a statistical model that identifies genes that are differentially regulated at the mRNA and protein level. After validation of the identified candidate genes we will unravel the general mechanisms that underlie their regulation.	Genes poster	Fundamental
P_Ge010	549	Peter-Bram 't Hoen, Eleonora de Klerk, Martijn Vermaat, Yavuz Ariyurek, Johan den Dunnen, Stephen Turner and Seyed Yahya Anvar	Peter-Bram 't Hoen	Analysis of PacBio full-length mRNA sequencing data uncovers widespread coupling between alternative transcription start sites, exons and polyadenylation sites	Short read sequencing technologies typically fall short in resolving complete transcript structures. The single molecule long read technology offered by the PacBio SMRT® technology provides reads that are well over the average size of an mRNA molecule and therefore generates complete cDNA sequences from the transcription start site until the polyadenylation site. The analysis of millions of these single-molecule long sequencing reads representing full-length mRNA molecules in MCF-7 human breast cancer cells and three human tissues provides the first opportunity to study coordination of transcription initiation, splicing and polyadenylation. To this end, we tested which alternative RNA features (transcription start sites, exon, polyadenylation site) were present more frequently or less frequently in the same transcript than expected by chance (mutually inclusive or mutually exclusive, respectively). Doing this, we found evidence for mutually dependent selection of alternative transcription initiation, splicing and/or polyadenylation sites in thousands of genes. The coordinated selection of mRNA features was often tissue-specific. Moreover, these events occurred across the entire mRNA molecule, where the selection of a particular transcription start site determined the selection of alternative exons or polyadenylation sites far downstream. A selection of events were subsequently validated by classical RT-PCR followed by Sanger sequencing. We conclude that there is an unprecedented degree of coordination between transcription, splicing and polyadenylation contributing to the transcript diversity observed in different tissues.	Genes poster	Fundamental
P_Ge011	467	Polewko-Klim Aneta, Lesiński Wojciech, Kitlas Golińska Agnieszka, Siwek Maria and Rudnicki Witold	Polewko-Klim Aneta	Application of the random forest method in identification of candidate genes in quantitative trait loci regions for adaptive immune responses of chicken	Current study aims at identification of the genetic markers associatedwith the variation of the variation of the adaptive immune traits in chicken. We have used machine learning methods to construct predictive models for the strength of response for three antibodies: KLH, LPS and LTA. The set of descriptive variables consisted of 384 SNPs preselected as candidates, based on the earlier work. Two procedures based on the Random Forest (RF) classifier were applied. To this end the predictive RF models were built and the relevance was assigned to variables using RF's perturbation importance as a measureof the relevance. The features that consistently show high relevance wereconsidered relevant. The entire procedure was performed within cross-validation loop.The predictive RF models based on these variables explain 11.6% of variance for KLH data, and roughly 3.5% of variance for LPS and LTA data. The procedure applied to a control run where antibody samples were collected before immunisation leads to a model with no predictive power.The number of SNPs identified as relevant in all 300 repeats was 10, 12 and 15 for KLH, LPS and LTA respectively. The respective numbers for 90% threshold are 17, 19 and 19.When the threshold is set at 50% of the numbers are 31, 27 and 30 for KLH, LPS and LTA respectively. Many SNPs identified in the study are common for more than one antigenicresponse. The SNPs identified in the study correspond to the several previouslyidentified genetic markers for immune response.	Genes poster	Agro-Food
P_Ge012	474	Brandon Malone, Ilian Atanassov and Christoph Dieterich	Brandon Malone	Bayesian Identification of Translation from Ribosome Profiling	Motivation:Ribosome profiling via high-throughput sequencing, riboseq, is a promising new technique for characterizing the occupancy of ribosomes on messenger RNA (mRNA) at base-pair resolution. The ribosome is responsible for translating mRNA into proteins, so information about its occupancy offers a detailed view of ribosome density and position which could be used to discover new translated open reading frames, alternative start codons and new isoforms.Contributions:We propose Rp-Bp, a Bayesian approach to predict the translation of open reading frames (ORFs) from riboseq data. In particular, Rp-Bp is useful for identifying novel translated short ORFs (micropeptides) and isoforms with high confidence. We use state-of-the-art Markov chain Monte Carlo techniques to estimate posterior distributions of the likelihood of ORF translation. A second novel contribution is automatic selection of periodic read lengths and ribosome P-site offsets via Bayesian model selection. Furthermore, we develop a competitive reference implementation for prediction based on the chi^2 test, Rp-chi.Results:We empirically demonstrate that our read length selection technique significantly improves sensitivity by resulting in up to an order of magnitude more predictions for Rp-Bp. Proteomics- and QTI-seq validation verifies the high quality of all of the predictions. Experimental comparison shows that Rp-Bp compares favorably to another recent tool for translation prediction. Qualitatively, we show that the method effectively identifies novel micropeptides and isoforms.Availability:The source code for Rp-Bp and Rp-chi is available at https://github.com/dieterich-lab/rp-bp.	Genes poster	Health
P_Ge013	349	Karl Koechert, Jie Cheng, Li Liu, Jose Garcia-Vargas, Barry Childs and Carol Pena	Karl Koechert	Biomarker identification in early clinical development – effective combination of hypothesis driven and data driven approaches in a clinical phase II trial assessing copanlisib activity in non-Hodgkin Lymphoma	Copanlisib, a novel pan-class I PI3K inhibitor with predominant activity against α and δ isoforms, has shown promising single agent activity in a phase 2 study in patients with indolent or aggressive NHL. Tumor gene expression profiling of 24 patients was used with both hypothesis- and data-driven approaches to identify genes or gene-signatures that may be associated with copanlisib treatment efficacy.The hypothesis-driven approach focused on pathways directly associated with copanlisib's mechanism of action, namely the B cell receptor (BCR)- and PI3K-signaling pathways, as well as disease-context pathways associated with e.g. tumor microenvironment. Gene expression of candidate pathways was integrated in a weighted manner to a patient-wise pathway score based on logistic or Cox regression models. Response rates were increased in patients with increased BCR and PI3K score (p=0.06 and 0.07; AUC=0.81 and 0.75, respectively). In addition, progression-free survival (PFS) was longer in copanlisib-treated patients with increased BCR score (HR=0.035, p<0.0001) and increased PI3K score (HR=0.24, p=0.02).The data driven approach used adaptive two way filtering (Cheng et al. 2012) combined with permutation-based cross validation to infer single genes predictive for best response or PFS and identified candidate genes with potential prognostic and/or predictive value, most prominently gene GPR18 (AUC=0.95; HR=5.8, p=0.01).In summary, using two complementary and robust analysis approaches, we have identified genes and gene signatures that are associated with objective response and PFS in this population of copanlisib-treated patients. Durable response to single-agent copanlisib is associated with tumors with activated PI3K/BCR pathways.	Genes poster	Health
P_Ge014	362	Irina A. Eliseeva, Ilya E. Vorontsov and Ivan V. Kulakovskiy	Ivan V. Kulakovskiy	Can transcription determine mRNA translation in mammals? Digging evidence with sequence analysis.	Transcriptional regulation of gene expression can determine mRNA stability and localization in yeast. It is an open question whether there is similar machinery in higher eukaryotes, e.g., whether translational state of a particular transcript can be defined at the transcriptional stage.In higher eukaryotes, the translation of many ribosomal and translational factors genes is controlled by the mTOR pathway that is directly involved in cell proliferation, aging, and oncogenesis. The 5' terminal oligopyrimidine sequence motif (TOP) is the specific feature of many mTOR translational targets. However, many mTOR targets carry improperly positioned non-terminal TOP or lack TOP completely.It is tempting to apply sequence analysis methods to identify transcriptional regulators that may leave imprints on transcribed mRNAs and thus determine forthcoming translational control. We utilized public CAGE and Ribo-Seq data to identify robust mTOR targets in human and mouse and performed sequence motif analysis of the respective promoter regions.Binding sites of several transcription factors were significantly enriched in promoters of the mTOR targets; among those transcription factors there were proteins having RNA-binding activity or direct interactions with other RNA-binding proteins. This suggests a principal role of transcription in mTOR translational control in higher eukaryotes.	Genes poster	Fundamental
P_Ge015	843	Sabrina Krakau, Hugues Richard and Annalisa Marsico	Sabrina Krakau	Capturing protein-RNA interaction footprints from iCLIP-seq data	RNA bindings sites for a protein of interest can now be detected genome-wide and at a high resolution thanks to the development of CLIP-seq technologies. Among these methods, iCLIP provides individual-nucleotide resolution and is particularly powerful for the characterization of protein-RNA interaction landscapes. However, existing methods for the analysis of iCLIP sequencing data suffer from several drawbacks: they do not account for the influence of transcript abundances nor do they model possible sources of technical or computational biases.To improve the analysis of such data we are developing an approach based on a non-homogeneous Hidden Markov model. Individual binding sites are called, taking into account regions enriched in protein bound fragments and the specifics of iCLIP truncation patterns. The underlying statistical framework enables us to simultaneously normalize for RNA abundances and to include as well other external data as covariates (e.g. nucleotide compositions, read lengths, mappability information). We devised a realistic iCLIP read simulation setup, that starts from real RNA-seq data and RNA binding sites, in order to evaluate our methods performance. Additionally we validate our approach using published iCLIP datasets from proteins with known predominant binding regions.Preliminary results on simulated data show that our tool is able to recover binding sites with a good accuracy. Further, on a real iCLIP dataset from the eIF4A3 protein our approach is in general more precise in determining the known binding regions than existing methods.	Genes poster	Fundamental
P_Ge016	792	Gwenneg Kerdivel and Valentina Boeva	Gwenneg Kerdivel	CIMP in adrenocortical carcinomas is associated with high expression of DNMT1 and increased Wnt and Notch signaling pathways activities.	Adrenocortical carcinomas (ACCs) are rare and aggressive endocrine cancer of the adrenal gland that exhibit recurrent genomic aberrations, negatively correlated with overall survival. Recently, a subtype of ACC characterized by a CpG island methylator phenotype (CIMP) has been discovered. CIMP is associated with especially poor diagnosis and one reason for this could be the promoter silencing through hypermethylation of tumor suppressor genes. By now, no drivers of CIMP in ACC have been identified. Using publicly available gene expression dataset of human ACCs from the TCGA and the Cochin Institute (Assié et al. 2014), we showed that DNMT1 expression is significantly increased in High-CIMP patients as compared to Low-CIMP patients, suggesting that DNMT1, rather than DNMT3A/B, could be responsible for the hypermethylation in CIMP tumors. Interestingly, expression of DNMT1 negatively correlates with overall survival. In addition, in patients with low or intermediate CIMP, the expression of DNMT1 allows a better discrimination of patients with good or poor survival. Together these results suggest that DNMT1 expression provides a reliable prognostic value. Moreover, we observed an inverse correlation between DNMT1 and APC expression, associated with an increased activation of Wnt signaling pathway in High-CIMP versus Low-CIMP samples (pathway analysis with ROMA). Not surprisingly, an increased activation of Notch signaling pathway is also observed as it is known to integrate Wnt signaling. Thus, the high aggressiveness of ACCs exhibiting high-CIMP as compared to low-CIMP could be due to the overactivation of these two pathways, known to cooperate in tumorigenesis of several cancer types.	Genes poster	Health
P_Ge017	573	Oren Tzfadia, Tim Diels, Klaas Vandepoele, Yves Van de Peer and Asaph Aharoni	Oren Tzfadia	CoExpNetViz: the Construction and Vizualisation of Co-expression Networks	Motivation:Comparative transcriptomics is a common approach in functional gene discovery efforts. It allows for finding conserved co-expression patterns between orthologous genes in closely related plant species, suggesting that these genes potentially share similar function and regulation.Existing co-expression tools are limited to data from model systems, which greatly limit their utility. Moreover, in addition, none of the existing pipelines allow plant researchers to make use of their own unpublished gene expression data for performing a comparative co-expression analysis and generate multi-species co-expression networks.Results:We introduce CoExpNetViz, a computational tool that uses a set of query or 'bait' genes as an input (chosen by the user) and a minimum of one pre-processed gene expression dataset	Genes poster	Biotechnology
P_Ge018	423	Josef Panek	Josef Panek	Computational modeling of RNA secondary structure using a novel approach	Information about evolutionary conservation of RNAs is employed for RNA secondary structure prediction in pairwise manner. For evolutionarily related RNAs, conserved structural segments are identified using pairwise sequence alignment and their structure is copied from known, experimentally resolved RNA structure into predicted structure. The remaining structural segments, showing week or no conservation, are predicted de novo using a standard prediction algorithm and merged with structure of conserved segments according to their position in the alignment. The presented approach is demonstrated here by modeling of secondary structure of mammalian ribosomal ribonucleic acids, one of the most essential biological molecules, whose structure is extremely large and complex.	Genes poster	Fundamental
P_Ge019	456	Lukasz Kreft, Pieter De Bleser, Paco Hulpiau, Arne Soete, Alexander Botzki and Yvan Saeys	Lukasz Kreft	ConTra v3: a tool to identify transcription factor binding sites across species, update 2016	Transcription factors are important gene regulators with distinctive roles in development, cell signaling and cell cycling, and they have been associated with many diseases. The ConTra v3 web server allows easy visualization and exploration of predicted transcription factor binding sites in any genomic region surrounding coding or non-coding genes. In this updated version, users can choose from nine reference organisms ranging from human to yeast. ConTra v3 can analyze promoter regions, 5'-UTRs, 3'-UTRs and introns or any other genomic region of interest. Thousands of position weight matrices are available to choose from, but the user can also upload any other matrices for detecting specific binding sites. Besides this visualization option, additional new exploration functionality is added to the tool that will automatically detect transcription factor binding sites (TFBSs) having both the highest regulatory potential and the highest conservation scores of the genomic regions covered by the predicted transcription factor binding sites. The regulatory potential is calculated based on the number of predicted TFBSs weighed by their distances to the reported transcription start site of the gene of interest. A typical analysis is run in four simple steps of choosing the gene, the transcript, the region of interest and then selecting one or more transcription factor binding sites for visualization or, alternatively, let ConTra v3 explore the transcription factors most likely regulating your gene of interest. The ConTra v3 web server is freely available at http://bioit.irc.ugent.be/contrav3/index.php	Genes poster	Biotechnology
P_Ge020	338	Maarten van Iterson, Erik van Zwet, Bastiaan Heijmans and Eline Slagboom	Maarten van Iterson	Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution	Association studies on omic-level data other then genotypes (GWAS) are becoming increasingly common, i.e., epigenome- and transcriptome-wide association studies (EWAS/TWAS). However, a tool box for the analysis of EWAS and TWAS studies is largely lacking and often approaches from GWAS are applied despite the fact that epigenome and transcriptome data have very different characteristics than genotypes.Here, we show that EWASs and TWASs are prone not only to significant inflation but also bias of the test statistics and that these are not properly addressed by GWAS-based methodology (i.e. genomic control) and state-of-the-art approaches to control for unmeasured confounding (i.e. RUV and cate).We developed a novel approach that is based on the estimation of the empirical null distribution using Bayesian statistics. Using simulation studies and empirical data, we demonstrate that our approach maximizes power while properly controlling the false positive rate. Finally, we illustrate the utility of our method in the application of meta-analysis by performing EWASs and TWASs on age and smoking which highlighted an overlap in differential methylation and expression of associated genes. We implemented our new method to control for bias and inflation of test statistics in the software bacon available from http://bioconductor.org/packages/bacon/.	Genes poster	Fundamental
P_Ge021	457	Petr Nazarov, Matthieu Gobin, Andrei Zinovyev, Eric van Dyck and Laurent Vallar	Petr Nazarov	Decomposition of transcriptional signal from tumours using independent component analysis	Tumour samples have complex cellular composition and show a high level of heterogeneity. The presence of stromal and immune cells, as well as polyclonality of cancer cells, limits interpretability of collected high-throughput data. Here we investigated and applied Independent Component Analysis (ICA) to decompose mixed signals in RNAseq data.First, we validated ICA approach in silico. Five cancers presented at TCGA repositories were selected: two brain cancers (GMB, LGG), melanoma (SKCM), lung squamous cell carcinoma (LUSC) and breast cancer (BRAC). Synthetic mixtures of their gene expression profiles were generated and then decomposed by ICA. We showed that, in order to obtain a robust separation, special attention to data transformation was needed and multiple runs of ICA were required.Next, we performed an in-depth analysis of 169 GBM and 473 SKCM samples. Gene signatures specific to each independent component were determined and associated to gene ontology categories. We identified components originated from different cell types and biological processes – some common and some specific to each tumour. Strong immune signals, neural tissue development and cell proliferation components were seen in both cancers, whereas components linked to melanin and keratin production – only in SKCM. Involvement of each component in samples was linked to clinical factors by ANOVA. We found a strong statistical connection between some of the components and methylation status. In GBM, many components were linked to Verhaak's tumour subclasses.Therefore, we conclude that ICA can detect cell subpopulations in bulk tissues, and help identifying gene signatures with diagnostic potential.	Genes poster	Fundamental

P_Ge022	637	Konstantina Dimitrakopoulou, Elisabeth Wik, Lars Akslen and Inge Jonassen	Konstantina Dimitrakopoulou	Deconvolution of transcriptome data from heterogeneous tissue samples	Microarray and RNA-sequencing technologies are key components in systems medicine approaches towards our comprehension of disease mechanisms. However, classical approaches for the analysis of expression data from complex tissue samples are highly biased by the heterogeneity and the variability in cell type composition. To facilitate transcriptome-based predictive and prognostic models for human diseases, it is necessary to deconvolve the tissue expression into the component expression profiles of each cell type. Experimental techniques such as cell sorting and laser-capture microdissection can physically separate the defined cell types before gene expression analysis, but they are time and resource demanding and can add additional stress on cells and thus affect their gene expression profiles.In silico deconvolution methods represent an appealing alternative to physical cell separation methods. We employ the linearity assumption in which the expression levels measured from a mixed sample are modeled as the weighted average of expression of the different cell types. In particular, we developed a method to estimate both the cell type proportions and the cell type-specific gene expression profiles directly from the mixed expression data based on representative profiles without requiring prior information on cell type-specific expression signatures or cell type proportions. We assess the performance of our approach on benchmark expression datasets and compare it with state-of-the-art existing methods.	Genes poster	Health
P_Ge023	608	Kristoffer Niss, Lasse Folkersen, Claus Berthelsen, Kirstine Belling and Søren Brunak	Kristoffer Niss	Decreased immune gene expression variation along the colon in non-inflamed mucosa of ulcerative colitis patients	Ulcerative colitis (UC) is an inflammatory disease of the colon believed to occur in genetically susceptible individuals exposed to a combination of environmental and microbial factors. The inflammation typically begins in the rectum and over time transitions along the colon in a proximal direction. This migratory progression suggests that UC-induced inflammation can not take hold of the entire colon at disease onset, but is limited to certain susceptible colonic segments. A comparative analysis of the colonic segments may provide knowledge of the etiology of UC, which is still limited.Mucosal biopsies of healthy donors (n=28) and UC patients (n=53) were taken from 1-6 colonic segments and microarray gene expression analyses were performed, yielding 217 samples. By applying segment-specific scaling to the expression levels of each gene, we constructed patterns that emphasize the gene expression fluctuations along the colon. This was done for each condition: healthy, inflamed UC (iUC) and non-inflamed UC (nUC). KMeans clustering of all genes using the three condition patterns together revealed major expression tendencies and a functional analysis of each cluster demonstrated clear intra-cluster gene relationships.A single cluster was strongly enriched for immune response related genes and had significantly (Q=0.05) different patterns in healthy, nUC and iUC for 324/2013 genes. In this cluster, we observed a clear state-change between inflamed and non-inflamed UC and that the gene expression variations between colonic segments in healthy individuals are reduced in nUC samples, possibly because the expression profile of proximal colon samples have changed.	Genes poster	Health
P_Ge024	409	Nicolas Nahuel Moreyra, Julian Mensch, Juan Hurtado and Esteban Hasson	Nicolas Nahuel Moreyra	Differential expression analysis of cold tolerance adaptation by RNA-seq de novo approach.	Over the last years, the role of temperature-related gene expression in ecological adaptation has been receiving increasing attention. Previous findings of our group identified specific cold adaptations involving energy metabolism and arrest of reproduction in females of the fly Drosophila buzzatii in response to winter conditions. We performed a RNA-seq analysis to investigate changes in gene expression profiles in order to identify the genetic basis of such cold adaptations.The study was conducted by exposing sets of females to three thermal conditions: one involving the cold tolerant flies and two control treatments. We used the Trinity software to generate a de novo assembly from RNA-seq reads. To analyze expression levels of the reconstructed transcripts, we mapped the reads against the transcriptome and then estimated the number of RNA-seq fragments (counts) that mapped to each contig. Transcripts were filtered based on a mean count value cutoff and normalized by the TMM method to get the differences in RNA composition. Thereby, we extracted transcripts that were at least 5.6-fold differentially expressed at a significance of 1E-3 in any comparison. This step allowed us to compare over and under-expression transcript clusters. Based on a well-annotated genome, we made a blastx against D. melanogaster genome. The results showed an over-representation of specific genetic pathways and cellular processes, highlighting the relevance of sugar metabolism, glycolysis and oxidative phosphorylation in the expression of cold tolerance. Future studies will assess the role of natural selection shaping the evolution of differentially expressed genes identified.	Genes poster	Ecosystems
P_Ge025	581	Ole Eigenbrod, Jane Reznick, Damir Omerbasic and Gary R. Lewin	Ole Eigenbrod	Discovering molecular signatures of extreme physiology using African mole-rats	The African mole-rats (Bathyergidae) are a family of subterranean rodents with very unusual physiological traits for mammals. The most famous member of African mole-rats is the naked mole-rat (Heterocephalus glaber), which shows several extraordinary phenotypes like poikilothermy, extreme longevity, cancer resistance and extreme adaptation to low oxygen environments. Additionally, the naked mole-rat and some other Bathyergidae species are insensitive to several noxious substances or algogens (e.g. acid, capsaicin, or mustard oil) [Park et al., PLOS Biology 2008].This study focuses on understanding the sensory phenotypes of at least 8 African mole-rat species, as these closely related species show different patterns of insensitivity to noxious substances. Recently, a sequence motif in the NaV1.7 ion channel of the naked mole-rat was found to be directly connected to its acid insensitivity [Smith et al., Science 2011].We sequenced poly-A selected mRNA from multiple tissues of 8 African mole-rat species. As there are no annotated genomes available for most of the species, we performed de-novo transcriptome assembly to obtain the protein-coding sequences. We developed a bioinformatic workflow to annotate putatively coding transcripts and exclude contaminating or falsely assembled sequences and chimeras. Using this approach, we were able to identify more than 10,000 unique protein-coding transcripts per species. We also directly compared the protein-coding sequences and transcript levels across species boundaries. This approach allows a multivariate analysis of the relationship between gene expression level, sequence variation and extreme phenotypes across this rodent family.	Genes poster	Fundamental
P_Ge026	708	Foivos Gypas, Andreas Gruber, Alexander Kanitz and Mihaela Zavolan	Foivos Gypas	Discovery, annotation and abundance estimation of transcript isoforms from high-throughput sequencing data	Mammalian genes typically have multiple transcription initiation and termination sites and exon forms that are used in a cell type specific manner to generate distinct transcript isoforms. In recent years it has become clear that an improved accuracy of transcript isoform abundance leads to a better understanding of cellular processes, such as, for example, miRNA-dependent gene regulation. A variety of methods have been proposed for the estimation of transcript isoform abundance from RNA-Seq data. We recently showed that many of them have comparable accuracy, but some excel in their efficiency [1]. A main bottleneck in estimating transcript isoform abundance is the availability of a complete and accurate set of transcript sequences. However, methods for transcript reconstruction based on mRNA-seq data are in their infancy and do not seem sufficiently accurate [2], even when a reference annotation is provided. In this work we use heterogeneous sequence data sets to expand the set of annotated transcript forms and thereby improve the estimation of transcript abundance across cell types. Our results have implications for the analysis of gene expression and for the analysis of protein variants in different cell types.1. Kanitz A, et al. Genome Biol. 20152. Hayer KE, et al. Bioinformatics. 2015	Genes poster	Fundamental
P_Ge027	434	Yao-Ming Chang, Arthur Chun-Chieh Shih, Ling Li, Ya-Ting Chang and Chien-Chang Chen	Yao-Ming Chang	Dynamically Genetic Program by Co-regulated TF Groups during the Pressure Overload-induced Cardiac Hypertrophy in Mice	Many heart diseases, such as hypertension, heart failure, and valvular heart disease, are accompanied by the cardiac hypertrophy. Understanding comprehensively what transcription factors (TFs) induce the hypertrophic process and when this process begins after pressure overload will be important in providing potential therapeutic targets in treating cardiovascular diseases. In this study, we collected the whole transcriptome data, including gene and miRNA expression data, isolated from hypertrophic murine hearts subjected to transverse aorta banding surgery (TAB) and without TAB surgery (sham) at five time points among the four weeks, respectively. From three analytical perspectives, we analyzed the whole transcriptome differences, functional distributions of differentially expressed genes, and clustered transcription factor (TF) coexpression network. In the result, we found that the globally genetic change began in the early stage, after cardiac pressure-overloaded, earlier than morphological change; moreover, the globally genetic change returned to a normal level within few days while the cardiac size kept on enlarging. It reveals the inconsistent timing between genetic and morphological changes. In addition, we also identified quite a few of TFs and miRNAs that differentially expressed in different stages and many of them have been also found in literature. Interesting, some miRNAs, verified with cardiac functions previously, were expressed not in the early stage but in dozens of days after TAB that indicates their suppression role in the hypertrophic process. In short, using and analyzing a time course transcriptome data our results can enhance the understanding in the dynamically genetic regulation in cardiac hypertrophy.	Genes poster	Fundamental
P_Ge028	485	Aslihan Gerhold-Ay, Johanna Mazur and Harald Binder	Aslihan Gerhold-Ay	Enhancing prediction performance by using mapping approaches for data integration of RNA-Seq and methylation data	High-dimensional data of next-generation sequencing platforms enable the development of molecular signatures for prediction of clinical endpoints like death or case-control status. The integration of heterogeneous data types can help for better prognostic modelling and to understand the underlying biological mechanism. The challenge for the integration of studies is to connect entities present in RNA-Seq data on gene-expression and methylation data on CpG sites. The optimal allocation problem has not been solved yet. Our aim is to investigate how prediction performance can be used as a measure for finding the optimal mapping of CpG sites to their related genes. For evaluation of our mapping approaches two different data sets were used. To obtain the optimal mapping, we define a window of nucleotides around genes. In a two-step approach, first we use regularized regression approaches to estimate a gene-signature based purely on the RNA-Seq data. In the following step, methylation data of the CpG sites that are falling within a window around the signature genes are used to estimate a new signature. The performance of the resulting signature is used to judge mapping quality. We compare different window sizes and show the distinct effect on prediction performance with respect to the endpoint. Mapping approaches for the integration of high-dimensional data can be powerful tools in medical research for more effective treatment of patients. Thus, prediction performance in the proposed approach could be recommended for the allocation of CpG sites and genes, if biological knowledge provides no clear guidance.	Genes poster	Health
P_Ge029	560	Inken Wohlers, Andriy Mashychev, Marcel Schilling, Christina M. Lill and Lars Bertram	Inken Wohlers	Evaluating the prediction of SNPs with effects on miRNA-mediated mRNA expression using transcriptome sequencing data	MicroRNAs (miRNAs) are short 19-22 base pair RNAs that post-transcriptionally alter the expression of mRNAs. This is achieved by binding to specific regions predominantly located in the 3' UTR within the target mRNA, which decreases protein output. We hypothesize that single nucleotide polymorphisms (SNPs) located in or near the miRNA-to-mRNA binding sites interfere with this process. To assess this hypothesis, we previously developed a bioinformatics pipeline that predicts the putative effect of all variants in dbSNP on miRNA-mediated changes in transcript expression. To this end, it uses miRNA-target sites predicted to reside in 3' UTRs of human transcripts and scores the effects of SNPs nearby. In this work, we utilize transcriptome sequencing data to generate a set of reference SNPs for benchmarking our predictions. These are SNPs for which we know – under the given physiological conditions – whether they are linked to a miRNA-mediated effect on transcript expression. This reference dataset is created from public mRNA and small RNA sequencing data generated from 345 lymphoblast cell lines as part of the Geuvadis project (www.geuvadis.org), which we re-processed and re-analyzed using updated protocols and reference panels. Expression profiles are then probed for those miRNA-mRNA combinations which are significantly correlated, followed by cis transcript eQTL analyses to identify SNPs with allele-specific effects on mRNA expression. A subset of the derived transcriptome-based reference data is used for optimizing the accuracy of our predictions using ROC analysis. The optimal prediction model is subsequently assessed in the remainder of the reference data.	Genes poster	Health
P_Ge030	592	Michaela Bayerlova, Annalen Bleckmann and Tim Beissbarth	Michaela Bayerlova	Evaluation of gene signatures applied to expression data of cancer patient cohorts	A gene signature is a collection of gene markers whose mRNA expression is associated with clinical outcome or can guide treatment decisions. With the advance in large-scale gene expression profiling technologies, multiple gene signatures have been established for further classification of cancer diseases into molecular subtypes. We examined two approaches of signature integration with patient data: 1) We applied a newly derived signature to public patient data to test its prognostic power. 2) We utilized a published gene signature for the classification of newly sequenced metastasis samples of patients.1) We derived a new pathway-based gene signature based on a pro-invasive perturbation of a pathway receptor in breast cancer cells. Subsequently, we tested the pro-invasive effect of the signature genes in expression data of breast cancer primaries with clinical annotation of metastasis events. The tumours were analysed by hierarchical clustering of the signature expression patterns and the identified patient clusters were subjected to Kaplan-Meier analysis of metastasis-free survival. The patient sub-groups showed significant differences in prognosis of metastasis development in breast cancer.2) In the second application, we first generated RNA-seq data of liver metastasis samples originating from colon cancer primaries. A published signature for defining molecular subtypes of colorectal cancer was applied using a nearest shrunken centroid-based method. We investigated whether the identified metastasis subgroups reflect characteristics of the primaries subtypes and evaluated the usage of the primary tumour signature extended to metastatic tissue setting.Furthermore, we critically evaluated signature-based classifications in respect of interpretability and clinical relevance.	Genes poster	Fundamental
P_Ge032	852	Lorena de La Fuente Lorente, Ana Conesa, Manuel Tardaguila, Hector Del Risco, Cristina Marti, Victoria Moreno and Susana Rodriguez	Lorena de La Fuente Lorente	FAIR, Functional Analysis at Isoform Resolution by using long reads technologies	Based on the claimed role transcript variants in conferring functional meaning and the lack of methods to study the functional implications of alternative splicing (AS) and alternative polyadenylation (APA), we have developed a new methodology called FAIR. This methodology will let to address the functional profiling of transcript and protein isoforms at a genome-wide level by using long-reads technologies. Moreover, we have implemented it in a software called Transcript2GO.Therefore, using PacBio and Illumina data, FAIR can generate functional hypothesis about the role of alternative isoforms in our system. First, FAIR allows the functional annotation of each PacBio-resolved isoform which involves the ORF prediction and the annotation of several functional layers: miRNA binding sites, PFAM domains, post-translational modifications, UTR motifs, NMD prediction, repetitive elements, etc. Finally, it applies different statistical methods which combine both expression data and functional annotation over each PacBio-resolved isoform. Among the several included statistical methods, we can highlight the Feature Differential Splicing (FDS) which is able to point out functional elements affected by AS/APA.Using our rich annotation pipeline over a neural differentiation system, we found that nearly all genes expressing several isoforms have them annotated with at least one differential functional label, suggesting that functional profiling at isoform resolution is meaningful. We identified several functions enriched in genes regulated by differential splicing, as well as specific features as miRNAs regulated by DS. Other functional insights of the relationship between function and differential splicing are easily revealed by the tools implemented in Transcript2GO.	Genes poster	Fundamental
P_Ge033	734	Rianne Beukhof, Madelon Engels, Sanne Abeln, Bas Stringer, Maurits Dijkstra, Ted Meeds and Jaap Heringa	Madelon Engels	First among Equals -- Discriminating Driver and Passenger Mutations	Carcinogenesis is typically driven by the accumulation of deleterious mutations. Combined with other clinical observations, these driver mutations allow experts to discriminate between different types of cancer, which is essential to accurately predict prognoses of available treatments, and also to develop new ones. However, many types of cancer cause genetic instability, introducing a multitude of passenger mutations in afflicted cells. Typical passenger mutations have no direct clinical relevance, but their abundance complicates the identification of driver mutations.Our study assesses which features can improve the methods we use to distinguish between driver and passenger mutations. Preliminary results were gathered using exome sequencing data from The Cancer Genome Atlas (TCGA) for four different types of cancer. Mutations in known driver genes occur in regions of the genome with a high evolutionary conservation score more often than expected by chance. Certain types of mutation are also correlated. For example, mutations causing a frameshift are more common in known driver genes, whereas silent mutations are statistically underrepresented. Frequency of mutation, however, appears to have no predictive value when considering the types of cancer separately. We further investigate these trends in a handful of case studies.	Genes poster	Health
P_Ge034	681	Anna Feldmann and Nico Pfeifer	Anna Feldmann	From Predicting to Analyzing HIV-1 Resistance Towards Broadly Neutralizing Antibodies	Recently, combination therapy with broadly neutralizing antibodies (bNAbs) was introduced as a viable new option in antiretroviral treatment against HIV-1, that is capable of reducing viral load under detectable levels for up to 60 days in humanized mice and non-human primates. First clinical trials showed that already a single infusion of one bNAb, 3BNC117, is able to suppress successfully viremia in HIV-1 infected humans and even enhance the antibody responses of the individuals. However, the efficacy of this treatment is also affected by the emergence of resistant strains. Prior to the administration of an antiretroviral bNAb combination therapy to a patient, it has to be ensured that the patient's viral strains are susceptible to the particular bNAbs of the combination. So far, resistance to bNAbs can only be tested in expensive and time-consuming neutralization assays. We propose a non-linear SVM-based model to predict the neutralization susceptibility of unseen viral strains to bNAbs based on the viral envelope sequence. Because non-linear SVM classification results are often difficult to interpret, we offer different visualization techniques to improve the biological interpretability of the results using feature space visualization and motif logos. Learning the important binding sites of the bNAbs, the models are also biologically meaningful and useful for epitope recognition. Moreover, we confirmed a trend towards antibody resistance for the subtype B HIV-1 population and extended the analysis to the global HIV-1 population by predicting the neutralization sensitivity for around 36,000 HIV-1 sequences from the Los Alamos National Laboratory HIV Sequence Database.	Genes poster	Health
P_Ge035	757	Arlin Keo	Arlin Keo	Functional analysis of polyQ genes by examining spatial co-expression across the human brain.	Polyglutamine (polyQ) diseases are inheritable, neurodegenerative disorders caused by an expansion of a CAG repeat tract in the coding region of one of the polyQ disease-associated genes. There are nine polyQ diseases which include Huntington's disease (HD) and multiple spinocerebellar ataxias (SCAs), each with their own causative gene. It is known that a longer CAG repeat tract leads to an earlier onset of the disease, but not all differences in age of onset can be explained by repeat length. Recent studies have shown that interaction among the polyQ genes affects the age of onset in HD and SCAs. In this study we aim to find the functional relations among the nine polyQ genes by analyzing their co-expression patterns across the human brain using the Allen Human Brain Atlas data. This high-resolution spatial microarray data allows the construction of gene-gene networks on a whole brain level as well as on a region-specific level. Genes that co-express with multiple polyQ genes are indicators of interaction between the polyQ genes and potentially play a role in the age of disease onset. Moreover, sets of genes co-expressed with each of the polyQ genes may give rise to the functional relatedness when examining the common functional pathways in which they are involved.	Genes poster	Fundamental

P_Ge036	736	Ahmed Mahfouz, Boudewijn P.F. Lelieveldt, Aldo Grefhorst, Lisa T.C.M. van Weert, Isabel M. Mol, Hetty C.M. Sips, Jose K. van den Heuvel, Nicole A. Datson, Jenny A. Visser, Marcel J.T. Reinders and Onno. C. Meijer	Ahmed Mahfouz	Genome-wide co-expression of steroid receptors in the mouse brain: identifying signaling pathways and functionally coordinated regions	Steroid receptors are pleiotropic transcription factors that coordinate adaptation to different physiological states. An important target organ is the brain, but even though their effects are well studied in specific regions, brain-wide steroid receptor targets and mediators remain largely unknown due to the brain complexity. Here, we tested the idea that novel aspects of steroid action can be identified through spatial correlation of steroid receptors with genome-wide mRNA expression across different regions in the mouse brain. First, we observed significant co-expression of six nuclear receptors (Estrogen Receptor alpha, Esr1; and beta, Esr2; Androgen Receptor, Ar; Progesterone Receptor, Pgr; Glucocorticoid Receptor, Gr; and Mineralocorticoid Receptor, Mr) with sets of steroid target genes that were identified in single brain regions. These co-expression relationships were also present in distinct other brain regions, suggestive of as yet unidentified coordinate regulation of brain regions by e.g. glucocorticoids and estrogens. Second, co-expression of a set of 62 known nuclear receptor co-regulators and the six steroid receptors in 12 non-overlapping mouse brain regions revealed selective downstream pathways, such as Pak6 as a mediator for androgen and glucocorticoid receptor's effects on dopaminergic transmission. Third, Magel2 and Irs4 were identified and validated as strongly responsive to the estrogen diethylstilbesterol in the mouse hypothalamus. The brain- and genome-wide correlations of mRNA expression levels of six steroid receptors that we provide constitute a rich resource for further prediction and understanding brain modulation by steroid hormones.	Genes poster	Health
P_Ge037	598	Ge Tan and Boris Lenhard	Ge Tan	Genome-wide prediction of regulatory territories and target genes under complex long distance cis-regulation	Comparative genomics and high-throughput experimental methods like ChIP-Seq have enabled efficient detection of regulatory elements in metazoan genomes. Nevertheless, the assignment of those elements to their target genes has remained a difficult task. Traditional assignment to the nearest gene, or a manual and semi-intuitive process is far from complete, since regulatory regions can be located hundreds of kilobases away from their target genes, sometimes beyond neighboring genes. We previously showed that arrays of conserved noncoding elements span the loci of developmental regulatory genes ("targets") and several other genes ("bystanders"), and define the edges of genomic regulatory blocks (GRBs). We found that the target genes that respond to distal regulatory elements in those regions have specific features that distinguish them from bystander genes in the locus and the genome. In this study, we proposed a robust approach for the automated determination of GRB spans and a machine learning based method for genome-wide detection of target genes. The result is a comprehensive catalog of nearly one thousand human genes likely to be regulated by long-range interactions and the regions harboring their corresponding cis-regulatory elements. The catalog comprises a large number of genes involved in development, transcription and axon guidance. Furthermore, these genes are enriched for genes involved in complex diseases, including cancer and diabetes. The GRB spans and target genes identified in this study provide a rich resource for studying developmental regulation and disease-associated genomic variation.	Genes poster	Fundamental
P_Ge038	450	Charles-Henri Lecellier, Wyeth W. Wasserman and Anthony Mathelier	Charles-Henri Lecellier	Human enhancers associated with immune response harbor specific sequence composition, activity, and genome organization	Enhancers are distal DNA regions involved in the transcriptional regulation of gene expression.The Cap Analysis of Gene Expression (CAGE) technology allows for a precise identificationof active enhancer regions in biological samples by capturing bidirectional RNA transcriptsat enhancer boundaries. Using this technology, the FANTOM consortium recently characterized> 38, 000 human enhancers from about 800 cell and tissue types. This mapping provides us withan unprecedented opportunity to examine enhancers at large scale for specific DNA sequence featuresand functions. We used the distribution of guanine and cytosine nucleotides at enhancer regionsto distinguish two classes of enhancers harboring distinct DNA shape patterns. A functionalanalysis of their predicted protein-coding gene targets highlighted that one class of enhancers wassignificantly enriched for associations with immune response genes. Confirming this result, wefound that this class of enhancers was specifically enriched for regulatory motifs recognized byTFs involved in immune response (e.g. NF-kB). While these enhancers were generally repressedor lowly active, we observed that they were cell type specific and preferentially activated upon bacterialinfection, reinforcing their potential role in immune response. Looking at chromatin capturedata, we found that the two classes of enhancers were lying in distinct topologically-associateddomains and chromatin loops. Taken together, these results suggest that specific DNA sequencepatterns encode for classes of enhancers that are functionally distinct and specifically organized inthe human genome.	Genes poster	Fundamental
P_Ge039	513	Konrad Zych, Chris Maliepaard, Roeland E. Voorrips, Gerrit Gort, Nick de Vetten, Johan C.P. Hopman, Jan M. de Haas, Michiel A. Noback, Ronald Wedema, Jan-Peter H. Nap and Ritsert C. Jansen	Konrad Zych	Improving potato breeding with computational and functional genomics	Potato is one of the most important food crops. Potato is an outbred tetraploid plant making its breeding time-consuming and cumbersome. Including genetic markers in the selection process could greatly improve potato breeding. This approach was successfully used in selection for few monogenic traits (e.g. resistance to Phytopthora infestans). In our study we developed markers for reliable screening for multigenic quality traits like color after frying.We created a large potato population, consisting of two experimental crosses and a panel of cultivars and breeding clones. We performed RNA-Seq on the parents of the crosses in order to extract SNPs, from which we created a 60,000 SNP array. We used this array to genotype our population. We extended the mixture models based genotype calling of fitTetra (Voorrips et al, 2011). We used RNA-Seq data to obtain starting values for the algorithm increasing accuracy of the calling. The resulting genotypes were used together with multi-year high quality phenotypes in association studies. Using multiple levels of correction for population structure and environmental variance and multiple-marker association analysis we elucidated new markers for complex potato quality phenotypes.With our improved algorithm we were able to salvage 20% more high quality SNPs and filter out the lower quality SNPs. As a result, we created one of the most comprehensive genomic resources for potato with more than 30,000 SNPs measured in more than 1,500 samples. Association analysis resulted in a set of markers that could be used by the companies to extend their breeding scheme.	Genes poster	Agro-Food
P_Ge040	406	Saskia Trescher, Jannes Münchmeyer, Christopher Schiefer and Ulf Leser	Saskia Trescher	In-silico Approaches for Estimating Transcription Factor Activity from Transcriptome Data	The regulation of gene expression is indispensable for the adaptability of all organisms. It is predominantly controlled by a complex network of transcription factors (TFs). In order to elucidate regulatory principles between TFs and their putative target genes at different scales, numerous algorithms have been presented. Assessing their performance is an important task and facilitated by the availability of a growing number of transcriptome and TF binding datasets.We report on our result from comparing three different in-silico approaches for identifying the most influential regulators of genes using transcriptome data. Specifically, we compare our re-implementation of the work by Schacht et al. [1] and tools provided by ISMARA [2] and RACER [3]. All of them can integrate information about TF binding (from i.e. ENCODE, TRANSFAC) with sample-specific expression data (e.g. mRNA, DNA methylation, CNV) either in each sample or across phenotypes. The resulting most active regulators vary considerably among the investigated methods. Using different underlying TF-gene networks unveils a notable dependence of TF activity on the number of target genes. In order to resolve these discrepancies we plan to study systematically the influence of network topology and data structure using synthetic input.[1] doi:10.1093/bioinformatics/btu446[2] doi:10.1101/gr.169508.113[3] doi:10.1371/journal.pcbi.1003908	Genes poster	Fundamental
P_Ge041	487	Hyojin Kang, Chul Kim, Boseok Seong and Seokjong Yu	Hyojin Kang	Integrated approach to combine RNA-seq- and Microarray-derived gene co-expression networks in Alzheimer's disease	Gene co-expression networks (GCNs) are graphic representations of genes showing similar expression pattern across tissues and experimental conditions. They can be used to identify functional modules and biologically relevant genes based on guilt-by-association framework. GCNs usually have been constructed using gene expression datasets generated by DNA microarrays, however the recent RNA-seq technology is rapidly replacing microarrays and allows more complete characterization of RNA transcripts. Since very few analyses have been performed on co-expression networks based on RNA-seq, it is important to infer GCNs from RNA-seq data. Moreover, GCNs from RNA-seq data can be combined with microarray-based networks to increase the robustness in meta-analysis. In this study, we collected many different datasets from NCBI GEO including 25 RNA-seq and 2,102 microarray samples derived from human brain in Alzheimer's disease and performed meta-analysis to identify functional modules responsible for the characterization of Alzheimer's disease. First, we established the GCN pipelines using in-house data performed on the same samples by both RNA-seq and microarray to reduce the artificial bias between two platforms. The GCNs were generated using Pearson Correlation Coefficient and meta-analysis was conducted using rank-based method. Then the same pipelines were applied to infer GCNs from Alzheimer's disease samples. The preliminary results show that the GCNs from microarray data provide rich molecular information to gain insight into biological processes and disease mechanism. There is low size overlap between microarray- and RNA-seq-derived GCNs however, GCNs from RNA-seq would complement ones from microarray due to the higher coverage and dynamic range of RNA-seq.	Genes poster	Health
P_Ge042	493	Yi-Wei Lee, Ting-Yu Chang, Hsei-Wei Wang, Oscar Kuang-Sheng Lee, I-Fang Chung and Shung-Haur Yang	Yi-Wei Lee	Integrated database for long non-coding RNA discovery, profiling, and annotation from RNA-sequencing data sets across cancers	Long non-coding RNAs (lncRNAs) are non-protein coding transcripts longer than 200 nucleotides. Recently, with the rapid growth of deep-sequencing technology and the development of computational prediction algorithms, a lot of lncRNAs have been identified in cancers. Therefore, the aim of this research is to identify lncRNAs by analyzing RNA-sequencing data in a clinically meaningful way, as well as to provide a cancer genomics database.We developed a user-friendly database to systematically collect a comprehensive list of lncRNAs from public databases including Ensemble, GENCODE, NONCODE, and lncRNAdb. In addition, there were >32,000 novel lncRNAs assembled from different cancer RNA-Seq data. These novel lncRNAs were filtered by considering a series of steps, such as transcripts length and coding potential score. Furthermore, we provided analysis results for the related genomic information of lncRNAs, such as cellular function and expression profiles. To investigate the association between diseases and de-regulation of lncRNAs, a straightforward query interface enables a user to find a set of potential biomarkers defined by the expression profiles and clinical outcomes.Using our database, an individual can observe the significant aberrant lncRNAs across cancers without knowing the underlying software that supports analyzing RNA-sequencing data and predicting novel transcripts. Our database simplifies visualization of the lncRNA-disease associations and was adapted for accurate selection of biomarkers by considering multiple data sources simultaneously. We believe that it will allow more efficient translation of laboratory discoveries into the clinical context, and will assist in reinterpreting the function of lncRNAs in cancer research.	Genes poster	Biotechnology
P_Ge043	480	Ping-Han Hsieh, Wen-Ting Wang, He Wang, Wei-Jhen Huang and Chien-Yu Chen	Ping-Han Hsieh	Investigating the effect of similar subsequences present in assembled transcripts on RNA-seq quantification for non-model organisms	Transcript abundance analysis based on RNA-seq has been widely adopted to study transcript expression in different physiological conditions or diseases for non-model organisms. Without reference genome or transcriptome, researchers have to perform de novo transcriptome assembly prior to expression quantification. In such situation, accurate quantification is challenging because the assemblers might produce incomplete sequences or incorrect splicing forms, which may mislead the estimation of expression quantities. This study aims to reveal the effect of similar subsequences present in the assembled sequences on the accuracy of expression quantification. To mimic the transcriptome analysis in non-model organisms, we used synthetic data of RNA-seq data generated by Bioconductor polyester. The expression intensities preset in the simulated data were used as the expected answers for performance evaluation. In this study, Bowtie2 and eXpress were used for read mapping and expression quantification, respectively. We observed that the accuracy of quantification decreases as the number of transcripts that share subsequences increases. Similar results were observed on real data where the expression abundances from RNA-seq were compared with that from microarrays for model organisms. On the other hand, for non-model organisms, qPCR data was used to evaluate the quantification accuracy. The results suggested that similar subsequences present in transcripts indeed have a strong influence on quantification accuracy. In the end, we provided practical suggestions on how the reference can be prepared in order to reduce the influence of similar subsequences on RNA-seq quantification for non-model organisms.	Genes poster	Fundamental
P_Ge044	468	Stefan Tomiuk, Jutta Kollet, Michail Knauel, Lena Willnow, Stefan Wild, Silvia Rüberg, Claudius Fridrich, Peter Mallmann, Frauke Alves, Philipp Ströbel, Dominik Eckardt, Andreas Bosio and Olaf Hardt	Stefan Tomiuk	Isolation of primary human tumor cells improves culture of target cells and reduces bias in molecular analysis	Solid tumors are infiltrated by cells of non-tumor origin, including heterogeneous lymphocyte subpopulations, fibroblasts, and endothelial cells. The amount and composition of infiltrating cells is highly variable and patient dependent, which makes analyses of primary tumor samples difficult. We have developed a fast and easy method to isolate untouched human tumor cells from primary tissue. This procedure is based on the comprehensive depletion of cells of non-tumor origin by combining automated tissue dissociation and magnetic cell sorting (MACS® Technology). Here, we have applied the method to isolate human tumor cells from primary and metastatic ovarian carcinoma, as well as thymoma specimens. The purified human ovarian carcinoma tumor cell fraction was further used for the isolation of CD133+ cancer stem cells (CSCs). We performed Whole Exome Sequencing (WES) and gene expression profiling to i) compare genomic characteristics of isolated tumor cells and unpurified samples, ii) identify tumor cell- and CSC-specific expression signatures, and iii) compare the latter expression data with that of ovarian cancer cells (GSE29450), which had been collected by laser capture microdissection (LCM).	Genes poster	Biotechnology
P_Ge045	788	Tareq Malas	Tareq Malas	Meta-analysis of Polycystic Kidney Disease expression profiles defines strong involvement of injury repair processes	Expression profiling experiments are becoming very popular in human disease study and drug discovery. Although they are useful in revealing novel insights about the disease etiology, there are several pitfalls and limitations to their use that need to be addressed. Among these limitations are the experimental and technology-related biases in the data, and the use of general gene annotation databases such as KEGG and Gene Ontology, which jeopardize the functional interpretation of the data. To overcome these limitations in the context of a study of Polycystic Kidney Disease (PKD), we completed a meta-analysis of published PKD expression profiles in combination with our in-house RNASeq study of a Pkd1-mutant mouse model. We included samples from mice, rats and patients, and from microarray and RNASeq platforms to limit experimental and technology based biases. Comparing these datasets we generated a PKD signature that consists of 960 genes, including several known PKD genes. We show the robustness of our signature by significantly distinguishing PKD from WT samples in independent datasets. To define the tissue injury and repair component of PKD, we also integrated experimental data, namely expression profiles of ischemia reperfusion injury samples, and literature data by mining PubMed abstracts for injury repair and gene/protein associations. We discovered that at least 22% of the PKD Signature genes and 40% of functions are implicated in injury repair processes, supporting the hypothesis that PKD is a state of chaotic renal repair.	Genes poster	Health
P_Ge046	691	Alexandra Poos, Andre Maicher, Anna Dieckmann, Marcus Oswald, Roland Eils, Martin Kupiec, Brian Luke and Rainer König	Rainer König	Mixed Integer Linear Programming based machine learning approach identifies regulators of telomerase in yeast	Understanding telomere length maintenance mechanisms is central in cancer biology as their dysregulation is one of the hallmarks for immortalization of cancer cells. Important for this well-balanced control is the transcriptional regulation of the telomerase genes. We integrated mixed integer linear programming models into a comparative machine learning based approach to identify regulatory interactions that best explain the discrepancy of telomerase transcript levels in yeast mutants with deleted regulators showing aberrant telomere length, when compared to mutants with normal telomere length. We uncover novel regulators of telomerase expression, several of which affect histone levels or modifications. In particular, our results point to the transcription factors Sum1, Hst1 and Srb2 as being important for the regulation of EST1 transcription, and we validated the effect of Sum1 experimentally. We compiled our machine learning method leading to a user friendly package for R which can straightforwardly be applied to similar problems integrating gene regulator binding information and expression profiles of samples of, e.g. different phenotypes, diseases or treatments.	Genes poster	Health
P_Ge047	716	Luca Santuari, Gabino F. Sanchez-Perez, Bas Rutjens, Lidija Berke, Viola Willemsen, Berend Snel, Kenzo Nakamura, Dick de Ridder, Ben Scheres and Renze Heidstra	Luca Santuari	Partitioning of PLETHORA target expression domains guides cell differentiation	Organ formation in animals and plants relies on precise control of cell state transitions to turn stem cell daughters into fully differentiated cells. In plants, cells cannot rearrange due to shared cell walls. Thus, differentiation progression and the accompanying cell expansion must be tightly coordinated. PLETHORA (PLT) transcription factor gradients were shown to guide the progression of cell differentiation at different positions in the Arabidopsis root. While well-describedtranscription factor gradients in animals specify distinct cell fates within an essentially static context, the PLT gradient is unique in its ability to control cell differentiation in a growing organ during continuous production and expansion of cells. To understand the output of their gradients we studied the gene set transcriptionally controlled by PLTs. Our work reveals how the PLT gradient regulates cell state by region-specific induction of cell proliferation genes and repression ofdifferentiation. Moreover, PLT targets include major patterning genes and autoregulatory feedback components, enforcing their role as master regulators of organ development.	Genes poster	Fundamental
P_Ge048	563	Tuomo Hartonen, Biswajyoti Sahu, Kashyap Dave, Teemu Kivioja and Jussi Taipale	Tuomo Hartonen	PeakXus: A Comprehensive Peak Calling Software for ChIP-Nexus and ChIP-exo	Novel chromatin immunoprecipitation (ChIP) experiments ChIP-Nexus [1] and ChIP-exo [2] allow studying transcription factor (TF) binding with unprecedented accuracy. True TF binding locations are separated from noise by peak calling softwares.Most peak calling softwares search binding events by creating a model of "true" peaks from the sites with highest enrichment in the ChIP-experiments and then accepting only the peaks resembling this model. It is however known that most TFs bind cooperatively with other TFs, form dimers or interact with other proteins. These different types of binding create different ChIP-Nexus/exo fingerprints. Fitting the peaks to just one model may lead to missing important binding events.PeakXus is a peak caller specifically designed to leverage the increased resolution of ChIP-Nexus/exo experiments. PeakXus is developed with the aim of making as few assumptions of the data as possible to allow novel discoveries. PeakXus supports use of Unique Molecular Identifiers (UMI) [3] to remove PCR-duplicates that can create artefacts closely resembling true ChIP-Nexus/exo binding events. We show that PeakXus consistently finds more peaks overlapping with TF-specific recognition sequences than published methods. PeakXus is available at https://github.com/hartonen/PeakXus.[1] He et al. (2015). ChIP-Nexus enables improved detection of in vivo transcription factor binding footprints. Nature biotechnology.[2] Rhee & Pugh (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell.[3] Kivioja et al. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nature methods.	Genes poster	Fundamental

P_Ge050	496	Mei-Ju May Chen, Yu-Rui Su, Ping Chang, Tai-Rong Hong, Bor-Wei Cherng,Yi-An Tung and Chien-Yu Chen	Chien-Yu Chen	Potential of lncRNA to regulate gene expression through promoter binding in Drosophila Melanogaster	Recent studies have revealed that a novel factor, long non-coding RNA (lncRNA), may also be a key player in gene regulation. However, it remains unclear for most of lncRNAs on how they regulate gene expression. In this regard, this study aims at investigating whether lncRNAs affect gene expression through binding to gene promoters by exploiting sequence reverse complementary. Here, we examined the possibility of this scenario in Drosophila melanogaster. A set of 4,599 fly lncRNAs was collected from FlyBase and recent studies. To identify promoters that might be bound by lncRNAs, we first adopted BLASTn to align lncRNA sequences to the promoter sequences of mRNAs. An lncRNA was reported to have potential of binding promoters if the number of the qualified alignments in promoter regions was significantly higher than that in the whole genome. We proposed that a high binding enrichment score indicates that a lncRNA might regulate some genes through binding to their promoters. The results revealed that 1,070 lncRNA-gene pairs (involving 82 lncRNAs and 410 promoters) were shown with binding potential owing to sequence reverse complementary. We further utilized the developmental transcriptome of D. melanogaster (Nature 471:473-9, 2011) to see whether the expression of these lncRNA-gene pairs were correlated. The analyses showed that the identified lncRNA-gene pairs have significantly higher correlated expression than random pairs. In summary, this study presented that lncRNAs might regulate gene expression through sequence reverse complementary with promoters, and suggested the potential of lncRNA to regulate gene expression through promoter binding.	Genes poster	Fundamental
P_Ge051	388	Christian Groß, Marcel Reinders, Dick De Ridder, Martijn Derks, Mirte Bosse, Hendrik-Jan Megens and Martien Groenen	Christian Groß	Predicting the impact of genetic variation in livestock	In recent years, advancements in functional effect prediction of variants in human genomes have led to several new discoveries and insights in heritable diseases. Methods such as CADD or Eigen incorporate various forms of variant annotation information to compute one generic score of deleteriousness for every DNA sequence variant. Currently, these methods are solely available for research of human genomes. The goal of this project is to develop methods for gene variant evaluation for livestock i.e. poultry, pig and cattle. This would open up the possibility for new approaches to adjust breeding schedules with the aim to achieve breeding goals without accumulating negative inbreeding side-effects. This would increase the overall health of livestock populations and help to reduce unnecessary suffering in animal farming.Numerous research groups are working with livestock genome data but epigenetic information and annotation lag behind, compared to data which is available for human or model organisms like mouse. With this in mind we first conducted a feasibility study by developing a method for sequence variant evaluation in mouse, based on human epigenetic data. By focusing on mouse data we are able to validate the possibility of transferring annotations from highly conserved regions in the human genome to non-human species.	Genes poster	Agro-Food
P_Ge052	646	Jairo Rocha, Jaume Sastre Tomas and Emidio Capriotti	Jairo Rocha	Ranking Putative Cancer Driver Gene Subsets	We develop a score for some subsets of genes that represents the possibility that this subset be associated with a specific type of cancer. The score depends on the correlation of SNP appearance on normal samples with respect to the same correlation on tumour samples. The normal samples include the genomic data from the 1000Genome Project. The tumour samples could be from different types of cancer (lung, colon and prostate cancers) from the TCGA (The Cancer Genome Atlas Consortium). This is the first time that all possible gene pairs (around 20 million) would be considered. The list of pairs most likely related to each type of cancer would be published. Each pair could be a target to be studied deeply by animal models and future therapeutic targets. The genes in a pair with high score should be treated simultaneously as possible cancer drivers. The score can be used to evaluate patients individually.The work carried out by Dr. Emidio Capriotti and other authors who have published it in September 2014 (Bioinformatics) describes a method to assign a score to each gene in the entire human genome and represents the possibility that the gene is associated with a type of cancer (this study used samples of lung, colon and prostate). There are multiple gene candidates but candidate pairs and subsets could be fewer and revealing.Some results are shown as promising.	Genes poster	Health
P_Ge053	481	Audrey Michel, James P. A. Mullan, Stephen Kiniry, Vimalkumar Velayudhan, Patrick B. F. O'Connor and Pavel Baranov	Audrey Michel	RiboSeq.Org for ribosome profiling data analysis and visualisation.	The ribosome profiling (ribo-seq) technique uses high-throughput sequencing to provide Genome Wide Information on Protein Synthesis (GWIPS) by revealing the locations and densities of actively translating ribosomes at a genome-wide level. On RiboSeq.Org (http://riboseq.org/) we provide freely available resources to help researchers analyse and explore ribo-seq data without having to use command-line tools.GWIPS-viz is an online genome browser which hosts over 1000 pre-populated ribo-seq and corresponding mRNA-seq tracks across 20 genomes generated from data from over 70 published studies, thereby enabling cross-study and cross-species comparisons.RiboGalaxy is a Galaxy-based web server where researchers can pre-process, align, analyse and visualize their ribo-seq data. GUI-based tools are provided to determine the strength of the triplet periodicity signal in ribo-seq data, generate metagene and ribosome profiles and carry out differential translation expression analysis using riboSeqR. The RUST suite of tools can be used to quickly characterise ribosome profiling datasets to assess their quality as well as analyze the relative impact of mRNA sequence features on local decoding rates. The RiboTools suite provides functionality for exploring translation in alternative reading frames and stop codon readthrough events. As well as help pages, we provide forums on both GWIPS-viz and RiboGalaxy usage (http://gwips.ucc.ie/Forum/).	Genes poster	Fundamental
P_Ge054	378	Kerem Wainer Katsir and Michal Linial	Kerem Wainer Katsir	Single Cell Expression Data as a Direct Measure for Identifying Human Genes that Escape X-Inactivation	Sex chromosomes pose an inherent genetic imbalance between genders. In mammals, one of the female's X-chromosomes undergoes inactivation (Xi). Indirect measurements estimate 15-25% of Xi genes to completely or partially escape inactivation. The identity of these escaper genes, and their propensity for escape remain unsolved. We applied a direct method to identify escapees based on RNA-Seq from 25 single-cell lymphoblasts and a pooled version. We quantified the differential allelic expression by assigning reads from expressed genes to SNPs with distinct maternal or paternal identities. We confirmed that X-inactivation occurs and is maintained in single cells. Using strict and relaxed protocols, we confidently identified 27 and 35 escaper genes, respectively. Using 30 published datasets, we compiled a genes' catalogue characterized as escapees or inhibited along with a confidence value. The nature of most reported genes (454 in total) as escapees and inhibited is mixed across many biological contexts. We report a strong statistical overlap between escapees identified from single cells and those reported in the literature-based catalogue. We confirmed the usefulness of single cells' expression data for studying allelic bias phenomena. We conclude that escaping X-inactivation is less deterministic than previously reported with only few genes acting as exclusive escapees.	Genes poster	Fundamental
P_Ge055	479	Volodimir Olexiouk, Steven Verbruggen, Jeroen Crappé, Kenneth Verheggen,Lennart Martens and Gerben Menschaert	Volodimir Olexiouk	sORFs.org: a repository of small ORFs identified by ribosome profiling	Micropeptides, defined as translation products from small open reading frames sORFs (<300nt) are becoming widely recognized. This is also demonstrated by recent characterisation of several members of this new group of bio-active players: Toddler, Pri-peptides, Sarcolipin and Myoregulin (Pauli et. al., Science, 2014; Chanut-Delalande et. al., Nat. Cell Biol., 2014; Magny et. al., Science, 2013; Anderson et. al., Cell, 2015). Ribosome profiling, a NGS-technique measuring translation synthesis, enabled the identification of numerous sORFs demonstrating ribosomal occupancy (Ingolia et. al., Science, 2009 and Cell, 2014). Historically, sORFs have been neglected and their discovery could thus potentially provide new and important biological insights. Means to distribute this knowledge are necessary. Our public repository, sORFs.org (Olexiouk et. al., Nucleic Acids Research, 2015), currently holds 263.496 sORFs identified using ribosome profiling, from three model species (human, mouse, fruit fly). Furthermore, sORFs.org includes various tool and metrics to assess the coding potential of sORFs at multiple levels: ribosome profiling analytics, genomic information, experimental information, visualization of data, dataset information and sORF-specific calculated metrics trying to determine their coding potential in different ways (FLOSS, ORFscore, variation analysis, PhyloCSF conservation score) as described by recent literature. sORFs.org provides researchers an easy way to inspect and query sORFs information, facilitating the integration of sORFs into in-house research projects. Moreover, the PRIDE-respin pipeline enables the automatic rescan of MS fragmentation spectra stored in PRIDE using Pladipus (Verheggen et. al., journal of proteome research, 2015), acquiring proteomic evidence for sORFs.	Genes poster	Fundamental
P_Ge056	387	Sivan Gershanov, Shalom Michowiz, Helen Toledano, Orit Barinfeld, Albert Pinhasov, Nitza Goldenberg-Cohen and Mali Salmon-Divon	Sivan Gershanov	Subgrouping of pediatric medulloblastoma using an integrated analysis of MicroRNA-mRNA expression profile	Medulloblastoma (MB), the commonest malignant brain tumor of childhood, is divided into four tumor subgroups representing distinct molecular entities. Subsequently, treatment should be designed according to the specific subgroup. MicroRNAs (miRNAs) are involved in carcinogenesis and tumor progression by regulating post-transcriptional gene expression. However, the miRNA-mRNA regulatory network in MB is far from being fully understood. The aim of the study is to identify novel miRNA subgroup biomarkers for specific diagnosis by analyzing integrated mRNA-miRNA MB transcriptome sequencing. With this aim, integrated whole transcriptome mRNA and miRNA expression analysis was performed on primary tumor samples collected from 10 MB patients. 867 mature miRNAs were identified in at least a single sample, of them 462 were common to all 4 subgroups. 25 (2.5%) of all expressed miRNAs appeared to be significantly differentially expressed between the subtypes. Namely, upregulation of hsa-miR-224-5p and hsa-miR-449c-5p was found exclusively among WNT, while downregulation of hsa-miR-135b-5p characterized SHH. Among groups 3 and 4, hsa-miR-20a-5p was upregulated or downregulated, respectively. RNA-seq from the same tumor samples identified 500 genes that vary between the four subtypes, among which 69 (13.8%) have anti-correlated miRNA-mRNA interactions with the 25 detected miRNA biomarkers. The predicted mRNAs targets of these miRNAs are associated with different signaling pathways, known to have a role in MB biology. Our study demonstrates that miRNAs are highly specific to distinct MB subgroups. Understanding the involvement of miRNAs and their targets in MB related signaling pathways may improve diagnosis and advance the development of targeted treatment for MB.	Genes poster	Health
P_Ge057	825	Joana P. Gonçalves, Jeroen de Ridder and Lodewyk F.A. Wessels	Joana P. Gonçalves	Temporally-aware discovery of regulatory cascades	Temporal transcriptomes expose dynamics of gene regulation and disruptions leading to disease. Many studies uncover functional units by grouping genes with similar transcriptional responses and linking them to transcriptional regulators. Gene grouping is typically obtained through differential expression or (bi)clustering, while regulators are predicted by direct target enrichment based on protein-DNA binding or regulator-target co-expression.Differential expression scores camouflage distinct variations over time. Clustering maintains chronology, focusing on global patterns often associated with broad biological functions. Biclustering achieves increased granularity from locality, but also generates patterns with arbitrary time gaps. Target enrichment ignores joint regulatory effects and co-expression likely captures targets with upstream co-regulation rather than regulator-target relationships.We propose a method that groups coordinated genes based on temporal phenomena: biological tasks may span shorter periods than the experiment, and participating genes likely coordinate mostly in that time; genes are involved in multiple tasks with different partners; some genes exhibit correlated profiles with delays induced by different response times and/or transcriptional cascades. Additionally, we predict regulators from coordinated regulator-target interactions exploring multi-layered paths without co-expression assumptions.We analysed androgen response in LNCaP cells. Our method recovered prostate cancer genes more effectively than traditional approaches. Identified regulatory units accurately characterised known pathways affected by androgen response, including cell proliferation, lipid metabolism, and unfolded protein response. Notably, delays and co-expression-free regulator prediction enabled discovery of time-shifted targets and inhibitory regulations, respectively, which would be missed otherwise. We validated predictions on the regulation of gene groups using public and in-house experimental data.	Genes poster	Health
P_Ge058	250	David Holloway and Alexander Spirov	David Holloway	TRANSCRIPTIONAL BURSTING IN DROSOPHILA DEVELOPMENT: STOCHASTIC DYNAMICS OF PAIR-RULE EXPRESSION	Segmentation of the anterior-posterior (AP) axis of the fruit fly (Drosophila) is first seen in the striped expression patterns of the pair-rule genes, well before the physical appearance of body segments. even-skipped (eve) is one of the best-studied pair-rule genes, forming 7 expression stripes orthogonal to the AP axis, which in turn regulate downstream genes involved in determining unique cell fates for each segment. Transcriptional control specific to particular stripe locations was first shown with eve: a 1.7 kb enhancer upstream of the coding region is sufficient to drive reporter expression in the 2nd eve stripe position (42 %EL, percent egg length). Recent live imaging of an eve stripe 2 reporter has demonstrated the stochastic nature of pair-rule gene expression. We have developed a stochastic model of eve stripe 2 expression, including binding of the enhancer by upstream transcriptional regulators and the initiation and completion of transcript elongation. All parameters in the model are constrained by experimental data. Simulations allow us to test different regulatory possibilities. A simple on-off model for transcriptional initiation does not fit the experimental time series for the stripe centre, indicating that eve has multiple 'on' rates for transcriptional initiation.	Genes poster	Fundamental
P_Ge059	658	Nick Dimonaco, Robert Hoehndorf and Amanda Clare	Nick Dimonaco	Using Gene Ontology annotations to understand lethality phenotypes	Online databases such as FlyBase provide information regarding the genes of model organisms such as Drosophila melanogaster, including a near complete set of gene disruption phenotypes. In most cases, genes contained in these databases are annotated using the Gene Ontology (GO), which provides information about the molecular function, cellular component and biological process. Here, we use these annotations to train a machine learning algorithm that can be used to identify combinations of GO features that lead to accurate and informative predictions for gene disruption phenotypes.The databases of C. elegans, D. melanogaster, M. musculus, S. cerevisiae and D. rerio were queried for genes associated with lethal or viable phenotype classifications. The available annotated genes were then filtered to remove those with phenotypes corresponding to: conditionally lethal, produced by multiple disruptions, allele-specific or not fully characterised. The remaining genes were then categorised into two subsets per organism; a subset of genes characterised as lethal and a subset characterised as viable. The GO terms associated with these two subsets were used to train a decision tree machine learning algorithm within Weka. We also investigated the over-representation of GO terms within these lethal and viable classes.Our results clearly demonstrate that GO terms can be used to successfully describe and predict lethal and viable phenotypes in model organisms. We discuss the causes of lethality and the variation that we found across the five species.	Genes poster	Fundamental
P_Ge060	401	Deepak Karthik, Gil Stelzer, Sivan Gershanov, Danny Baranes and Mali Salmon-Divon	Deepak Karthik	Utilizing the Benford law for unravelling tissue specificity	The reduction in sequencing costs has led to an unprecedented trove of gene expression data from diverse biological systems. Subsequently, principles from other disciplines such as the Benford law, which can be properly judged only in data-rich systems, can now be examined on this high-throughput transcriptomic information. The Benford law states that in numerical data, the proportion of numbers beginning in any given digit is not uniform but rather skewed, with 1 being the most common digit and 9 the rarest. Here we demonstrate that digital gene expression data has a Benford-like distribution when observing an entire gene set. This phenomenon was conserved in a wide range of biological tissues and developmental conditions. However, when obedience to the Benford law is calculated for individual expressed genes across thousands of cells, genes that best and least adhere to the law are enriched with tissue specific or cell maintenance descriptors, respectively. Surprisingly, a positive correlation was found between the obedience a gene exhibits to the Benford law and its expression level, despite the former being calculated solely according to first digit frequency while totally ignoring the expression value itself. These results demonstrate the applicability and potential predictability of the Benford law for gleaning biological insight from simple count data.	Genes poster	Fundamental
P_Ge061	665	Djordje Djordjevic, Kenro Kusumi and Joshua Ho	Djordje Djordjevic	XGSA: A statistical method for cross-species gene set analysis	Gene set analysis is a powerful tool for determining whether an experimentally derived set of genes is statistically significantly enriched for genes in other pre-defined gene sets, such as known pathways, gene ontology terms, or other experimentally derived gene sets. Current gene set analysis methods do not facilitate comparing gene sets from different organisms as they do not explicitly deal with homology mapping between species. There lacks a systematic investigation about the effect of complex gene homology on cross-species gene set analysis.In this work, we show that not accounting for the complex homology structure when comparing gene sets from two species can lead to false positive discoveries, especially when comparing gene sets that have complex gene homology relationships. To overcome this bias, we propose a straightforward statistical approach, called XGSA, that explicitly takes the cross-species homology mapping into consideration when testing gene set analysis. Simulation experiments confirm that XGSA can avoid false positive discoveries, while maintaining good statistical power compared to other ad hoc approaches forcross-species gene set analysis. We further demonstrate the effectiveness of XGSA with two real-life case studies that aim to discover conserved or species-specific molecular pathways involved in social challenge and vertebrate appendage regeneration.	Genes poster	Application Fundamental
P_Ge062	783	Joske Ubels, Erik van Beers, Pieter Sonneveld, Martin van Vliet and Jeroen de Ridder	Joske Ubels	zPFS: a method to identify gene expression signatures to predict treatment specific survival in cancer	Cancer treatments may have heterogeneous response rates. Patient perspectives such as adverse treatment-related events and survival may be improved by selecting the right treatment at diagnosis. This is a major challenge that requires identification of biomarkers, such as a gene expression signature, based on which the best treatment regime can be determined.Here, we propose a new computational method to identify gene expression signatures that predict if a patient is likely to survive longer when receiving a specific treatment as compared to an alternative treatment. Our algorithm exploits tumor cell gene expression data from phase 3 clinical trials in which patients were randomly assigned to the treatment of interest or another treatment.Our method hinges on the notion that potential signature genes and gene sets can be identified by searching for patients receiving different treatments that have a large difference in survival while exhibiting similar gene expression profiles. To identify these we introduce zPFS, a measure for how much larger than expected this survival difference is for a specific patient. This zPFS measure enables identification of signature gene sets and exemplar patients that can be used to predict treatment specific survival.We demonstrate the utility of our method in a multiple myeloma dataset, where patients either received the proteasome inhibitor bortezomib or not. We find fourteen GO categories that can identify patient groups, comprising at minimum 20% of the patients, that have at least a 2-fold lower risk of experiencing an event (p-value < 0.05) when receiving bortezomib.	Genes poster	Health

POSTER LIST
ORDERED ALPHABETICALLY BY POSTER TITLE
GROUPED BY THEME/TRACK

THEME/TRACK: GENOMES

Poster numbers: P_Go001 - 124 Application posters: *P_Go001 - 011*

Poster number	EasyChair number	Author list	Presenting author	Title	Abstract	Theme/track	Topics
APPLICATION POSTERS WITHIN GENOMES THEME							
P_Go001	*832*	*Mahdi Heydari, Giles Miclotte and Jan Fostier*	*Mahdi Heydari*	*Brownie : correcting second generation sequencing errors using de Bruijn graphs*	*Background: Next-generation sequencing (NGS) methods enable the production of huge amounts of sequencing data at a low financial cost. However, the presence of sequencing errors in these data challenges applications like de novo assembly methods, potentially causing a sub-optimal assembly quality. Therefore, several standalone applications have been proposed that are specialized in the correction of these errors in order to provide cleaner data for downstream analysis tools.Description: We introduce Brownie for the correction of errors in sequencing data generated from the Illumina platform. Brownie builds a de Bruijn graph from all reads with a user defined k-mer size and applies graph correction algorithms by taking into consideration both the graph topology and statistical evidence in order to detect erroneous nodes. Subsequently, the input reads are individually aligned to the corrected de Bruijn graph in order to correct them by finding mismatches, insertions and deletions between the reads and the sequence represented by the graph.Conclusion: We applied Brownie to real data sets and we compared the effect of error correction in genome assembly. Reads which are corrected by brownie resulted in a higher quality of contigs in terms of genome coverage and NGA50 where the assembler is velvet. Brownie, Karect and Blue performs equally good by using Spades and Discovar as an assembler.*	*Genomes/Application poster*	*Application*
P_Go002	*761*	*Roy Straver, Erik Sistermans, Marjan Weiss and Marcel Reinders*	*Roy Straver*	*Detection of Copy Number Aberrations in Exome Sequencing Data Based on a Within-Sample Comparison Scheme*	*Whole Exome Sequencing (WES) is currently the primary method of choice for genome-wide variant detection in diagnostics. Although used for SNP detection, WES can be used to discover pathogenic Copy Number Variations (CNVs) as well. Existing methods to detect CNVs on WES data rely on direct comparison of reads per exon to other samples. Aberrated (deleted and/or duplicated) regions are either called through Hidden Markov Models or require a sequence of aberrated exons.We developed a method based on previous work (WISECONDOR) to find aberrated exons in exome data through internal comparison of probe coverage to a set of probes known to behave alike in terms of read depth (z-scores), as learned on a set of known normal samples. If there is enough certainty for a single probe to be aberrated it will be called while a sequence of less clearly aberrated probes may be found using varying windows of z-scores. A segmentation algorithm is applied to determine exact start and end positions of aberrated regions. In a first validation experiment, at least 10 known aberrations in our data were correctly identified, sizes ranging 5 Kbp and up. One sample with 2 deleted exons was correctly identified, including the partial deletion of an exon in a neighbouring gene.Our work provides an alternative and reliable method to find CNVs in exome data, requiring only a set of previously reference exomes. We do not require resequencing of control exomes in the same run.*	*Genomes/Application poster*	*Application Health*
P_Go003	*475*	*Ulrike Löber, Pin Cui, David E. Alquezar-Planas, Yasuko Ishida, Alexandre Courtiol, Dorina Lenz, Kristofer M. Helgen, Alfred L. Roca, Stefanie Hartmann and Alex D. Greenwood*	*Ulrike Löber*	*ERV evolution - A bioinformatics pipeline to investigate retroviral integration in museum koala samples*	*The koala retrovirus (KoRV) is currently invading the genome of the koala (Phascolarctos cinereus). To investigate the invasion process, we examined three different DNA target enrichment techniques to determine the best method for application to ancient DNA samples of museum koala skins. Museum skin are a resource for looking at genetic changes over time directly. Since the genome of the koala is not available, we developed a bioinformatics pipeline to investigate viral integration sites from next generation sequencing data of the target enrichment libraries. We pre-processed the data from ancient, highly degraded DNA ((cutadapt; Martin, 2011), (Trimmomatic; Bolger,Lohse & Usadel, 2014),(Flash; Magoc & Salzberg, 2011)). A pairwise alignment approach (Smith-Waterman, EMBOSS; Rice, Longden & Bleasby, 2000) was used to search and extract viral sequences. Finally, we created integration site consensus sequences using a sequence similarity based (BLAST; Altschul et al., 1990) clustering approach (TRIBE-MCL; Enright, Van Dongen & Ouzounis, 2002). Statistical analysis was performed using R, including the R package spaMM (Rousset & Ferdy, 2014) to make a logistic regression on a Generalized Mixed Effect Model, considering the identity of koalas and insertion sites as two Gaussian random effects. A comparison of integration sites of ten museum koalas shows that the proportion of integration sites shared between any two koalas is quite small suggesting the KoRV invasion process is still at an early stage in koalas.*	*Genomes/Application poster*	*Application*
P_Go004	*537*	*Anjana Ghelani, Pravin Dudhagara and Rajesh Patel*	*Anjana Ghelani,*	*Genome sequence annotation of virgibacillus spp BAB 3240 strain isolated saline desert soil, Gujarat, India*	*In present work, an attempt was done to assemble and annotate the Indian origin strain Virgibacillus sp. BAB_3240 isolated from the saline desert soil sample of little Rann of Kutch, Gujarat, India. The annotation was performed on Rapid annotation subsystem (RAST) server. Genome size was 622,098 bp with 34.5 G+C ratio. Total 1320 coding sequence and 1389 features were reported. Presences of 149 RAST subsystems were recorded. Total 35 subsystems for various stress response genes comprising osmotic stress, heat shock protein and desiccation resistance were reported. KEGG map analysis displayed presence of many pollutant degradation pathway. Comparison of genome analysis was done for Virgibacillus sk37, Virgibacillus alimentarius, Virgibacillus soli and Virgibacillus pantothenticus. With refers to overall genome comparison Virgibacillus sp. BAB_3240 found more relevant to Virgibacillus sk37. Besides RAST annotation of Virgibacillus sp. BAB_3240 was also performed on BaSys and Microscope Genome Annotation platform in which total 2330 and 1423 genes were identified respectively. Data mining of relevant the presence of stress response genes and pathway for degradation of various environmental pollutants. Annotation data and analysis indicate the possibility of explore the stress for pollution control in industrial influent and saline soil environment.*	*Genomes/Application poster*	*Application Biotechnology*
P_Go005	*393*	*Kshitij Tayal, Naveen Sivadasan and Rajgopal Srinivasan*	*Kshitij Tayal*	*GPhase: Greedy Approach for Accurate Haplotype Inferencing*	*We consider the problem of phasing an individual genotype sample given a collection of known haplotypes in the population. We propose a fast and accurate phasing algorithm GPhase that reconstructs haplotype pair consistent with input genotype. GPhase uses the coalescent based mutation model of Stephens and Donnelly (2000), which is also used in the popular PHASE (v1.1) tool. Computing optimal solution under this model is expensive. GPhase uses a greedy iterative approach for fast haplotype estimation with high accuracy. Our algorithm is simple, efficient and has linear time and space complexity. For gene level phasing, GPhase performed consistently better on both real and simulated datasets when compared to state of the art tools such as SHAPEIT, Beagle, MaCH and Impute2. On real datasets, GPhase obtained 100% accuracy for input samples from 24 genes out of total 81 genes. On simulated data, with just 1000 known haplotypes, GPhase achieved 97.5% switch accuracy on samples with more than 1700 markers. Gphase is extremely fast and scales well for gene level phasing using publicly available and large scale haplotype datasets such as HapMap data and 1000 genome data. This has applications in studies on recessive Mendelian disorders where parent data is lacking.*	*Genomes/Application poster*	*Application*
P_Go006	*369*	*Corinne P. Oechslin, Philipp Agyeman, Rémy Bruggmann, Lucy Poveda, Stephen L. Leib and Christian M. Beuret*	*Corinne P. Oechslin*	*Host nucleic acid depletion method increases sensitivity of pathogen detection by metagenomics analysis in surrogate cerebrospinal fluid samples*	*Management of patients with central nervous system (CNS) infections is a challenge since the aetiology by clinical diagnostic assays remains unknown in up to 60% of meningo-/encephalitis cases. We developed a method for host nucleic acid (NA) depletion to be used in cerebrospinal fluid (CSF) samples before applying NGS and bioinformatics analysis with the aim to identify unknown aetiologies of CNS infections.The host NA depletion method consists firstly of a homogenization to release eukaryotic NA while integrity of viruses and bacteria is preserved. Then, host NA is enzymatically degraded and subsequently depleted by a magnetic bead separation. Surrogate CSF samples consisting of artificial CSF, controlled amount of human cells and defined virus and bacteria concentrations were used to evaluate the method. Combined RNA and DNA sequencing of native and host depleted surrogate and patients' CSF samples was run using IonTorrent™ and Illumina® technology. Sequencing data were analysed using a tentative bioinformatics pipeline involving a taxonomic sequence classifier to remove residual host sequences followed by identification of the remaining non host sequences.A 100.000-fold depletion of host NA with minimal loss of viral and bacterial NA was confirmed in surrogate CSF samples by qPCR. This drop of the host NA proportion after depletion is reflected in a higher match rate of pathogens' sequences after bioinformatics analysis compared to native samples of a pathogen dilution series in surrogate CSF samples. Sequencing was proved to be as sensitive as qPCR. Moreover, a patient's CSF, positive diagnosed for HSV, was confirmed by sequencing.*	*Genomes/Application poster*	*Application*
P_Go007	*413*	*Yuuma Hosokawa, Asuka Kitajima, Satoshi Fujii, Midori Iida, Toshimasa Yamazaki, Hiroki Sasaki, Kazuhiko Aoyagi and Takahiro Yamanoi*	*Yuuma Hosokawa*	*Identification of distinct subtypes in colorectal cancer with the survival and the primary sites*	*To detect biological characteristics of colorectal cancer patients, we determined the number of distinct subtypes of the patients by the repetition of NMF (Non-negative Matrix Factorization) and Smyth's LIMMA. Then, we extracted genes having significant differences among the subtypes in the expression level by PAM (Prediction Analysis for Microarrays).Consequently, we classified the patients, who were included in the data from the National Cancer Center Research Institute, into three subtypes (G1, G2, G3). Furthermore, when we focused on the differences among the three subtypes, we found that survival times with G2 (a long tendency) vs G1 and G3 (a short tendency) yielded statistically significant differences (log-rank test, P<0.00181). Moreover, crucial differences were observed in the primary sites with higher cancer risk in the three subtypes: G1 corresponds to ascending colon or cecum; G2 corresponds to transverse or descending colon; G3 corresponds to rectum above or below the peritoneal reflection, or sigmoid colon. When we linked these clinical data with gene expression levels, it was revealed that gene expression levels in G2 are higher than those in G1 and G3 (survival aspect). Similarly, gene expression levels in G1 were higher than those in G3 (primary site aspect). Finally, we succeeded in building a classifier of these patients using the three subtype and the subtype-specific gene. This classifier indicated a high performance in terms of accurate rate (98.9%). The present classifier would lead us to the prediction of survival time and primary sites where the cancer has developed in colorectal cancer patients.*	*Genomes/Application poster*	*Application*
P_Go008	*535*	*Lieven Sterck,Thomas Van Parys, Stephane Rombauts and Yves Van De Peer*		*ORCAE: A wiki-style platform enabling efficient community curation of gene and genome annotations*	*Conducting gene and genome annotation typically relies on diverse information resources going from sequence to expression data depending on whether structural or functional annotation is performed. To help researchers doing gene annotation while having access to these different data types, we developed ORCAE (Online Resource for Community Annotation of Eukaryotes), a web-technology-compliant portal for use in community genome annotation efforts.ORCAE allows browsing and on-the-fly editing of gene descriptions as well as gene structures, moreover all manual curations are immediately visible for other users. The portal will store all the modifications, so for each locus a history of modifications is available.Through its interface, ORCAE offers easy access to precomputed information that greatly facilitates the work of a curator. The gene page offers several informative graphics with a focus on the quality of the gene structure (eg. Multiple alignments of proteins, ...) helping the annotators in improving the automated annotations. Annotators can therefore use the build-in GenomeView interface to easily check/modify gene structures. A unique feature from ORCAE is that the portal is highly dynamic, all the available information (eg. protein similarity, transcript alignments, ...) is immediately updated and presented on the gene page.ORCAE can both be used to coordinate ongoing annotation efforts in the course of the project as well as to present published genomes to the public by acting as a genome portal.Currently it offers public access to 16 eukaryotic genome projects and restricted access to another 32 genomes.ORCAE is available at http://bioinformatics.psb.ugent.be/orcae/.*	*Genomes/Application poster*	*Application*
P_Go009	*564*	*Jane L. Nybo, Tammi Vesth, Jens C. Frisvad, Sebastian Theobald, Inge Kjærbalding, Igor V. Grigoriev, Scott E. Baker and Mikael R. Andersen*	*Jane L. Nybo*	*Ortholog identification in genera of high genetic diversity and evolution*	*In the era of high-throughput sequencing, comparative genomics is vastly used in the discovery of genetic diversity between species. Current comparative approaches are implementing ortholog identification to establish genome annotations, gene or protein evolutions or defining functional features in individual species and groups.In this project we aim to compare the genomes of 300 species from the genus of the filamentous fungi Aspergillus, which represent evolutionary changes over 200 million years.[1]. Present-day's gene ortholog prediction software combine genetic diversity with evolutionary distance. However, when the genetic variation within a single genus is higher than the variation between fish and humans.[2]. identifying true orthologs becomes problematic. Here we tackle the genetic diversity problem by using stringent alignment parameters and lenient linkage criteria, which captures genetic events such as gene duplications, deletions and horizontal gene transfers. The approach is based on BLASTp alignments with optimized cut-offs considering identity and coverage, and orthologous clustering using single linkage.The result of this procedure is a large number of orthologous protein families, that will allow for guilt-by-association-based detailed investigation into mapping of genotype to phenotype. The data will highlight the genomic and functional features defining individual species, clades and the core/pan genome of Aspergillus.[1] Galagan et al. (2005). Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature, 438:1105–1115.[2] Fedorova et al. (2008). Genomic islands in the pathogenic filamentous fungus Aspergillus fumigatus. PLoS Genetics, 4(4), e1000046.*	*Genomes/Application poster*	*Application Biotechnology*
P_Go010	*412*	*Asuka Kitajima, Yuuma Hosokawa, Satoshi Fujii, Midori Iida, Toshimasa Yamazaki, Hiroki Sasaki, Kazuhiko Aoyagi and Takahiro Yamanoi*	*Asuka Kitajima*	*Prognostic feature extraction in colorectal cancer by combining the gene expression data and the clinical data.*	*The purpose of this study is to find out molecular features associated with prognosis of colorectal cancer. For the purpose, we analyzed the gene expression data with approximately 20000 genes and the clinical data for a set of 242 colorectal cancer patients from National Cancer Center Research Institute (Japan). Firstly, we constructed three kinds of random survival forests (RSF) model which predict survival of the patients. The first model was constructed using both the gene expression data and the clinical data of colorectal cancer. The second was constructed with only the gene expression data. The third was constructed with only the clinical data. Secondly, we compared performances of these models and confirmed which data significantly improved the performances. Consequently, the performance of the model using both of gene expression data and clinical data was significantly higher than that using only the gene expression data. Therefore, in the survival analysis of colorectal cancer, it is useful to simultaneously utilize the gene expression data and the clinical data. Finally, we obtained prognostic features associated with colorectal cancer from the survival prediction model using both the gene expression data and the clinical data. This model provided valuable importance in each feature in terms of how much they are related to survival time. This importance led us to predict which features have relation to the prognosis. The present features were mainly chemotherapy, curability, metastasis and some genes.*	*Genomes/Application poster*	*Application*
P_Go011	*705*	*Harmen van de Werken, Job van Riet, Niels Krol, Peggy Atmodimedjo, John Martens, Leendert Looijenga, Guido Jenster, Hendrikus Dubbink and Winand Dirjens*	*Harmen van de Werken*	*SNPitty: Lightweight Shiny-Based B-Allele Frequency Web Viewer For Detecting Loss of Heterozygosity and Allelic Imbalances in Targeted Multi-Gene Panels.*	*Exploration and visualization of next-generation sequencing (NGS) data originating from targeted multi-gene panels is crucial for analysis of genetic aberrations in both research and clinical settings. However, software for simple, robust and dynamic web-based visualization of single nucleotide polymorphisms (SNPs) in the regions of targeted multi-gene panels is lacking. Therefore, we developed a lightweight Shiny-based B-allele frequency web viewer, called SNPitty which is well-suited for interactive visualization and interrogation of single- and multi-sample Variant Call Format (VCF) files. SNPitty is best applicable with data from NGS targeted multi-gene panels to display allelic imbalance with a focus on loss of heterozygosity (LOH) and copy number variation of genetic heterozygous markers, but can also be used for combined mutational and LOH analysis of genes of interest. Moreover, SNPitty is capable of generating predefined reports, which summarize and highlight the targets-of-interest based on LaTeX-templates.Here, we apply SNPitty on a serial dilution series of patient-derived glioma tissue with matched normal tissue to assess LOH and genomic amplification. DNA was sequenced on an Ion-Torrent platform using a diagnostic multi-gene panel targeting known genetic aberrations associated with tumor formation and progression. VCFs were subsequently generated using Torrent Suite™. 1p/19q co-deletion and LOH were clearly indicated by SNPitty for the glioma sample in regards to degree of dilution.SNPitty can be used freely at http://ccbc.nl/SNPitty. Source code and documentation for SNPitty can be obtained under the terms of the GPL-3 open-source license through BitBucket at https://bitbucket.org/ccbc/snpitty A docker image containing SNPitty is available at https://hub.docker.com/r/ccbc/snpitty/.*	*Genomes/Application poster*	*Application Health*
OTHER POSTERS WITHIN GENOMES THEME							

P_Go013	638	Husen M. Umer, Marco Cavalli, Michal J. Dabrowski, Klev Diamanti, Marcin Kruczyk, Gang Pan,Jan KomorowskiandClaes Wadelius	Husen M. Umer	A distinctive mutational pattern at CTCF motifs in cancer	Somatic mutations drive cancer and there are established ways to study those in coding sequences. It has been shown that some regulatory mutations are over-represented in cancer. We develop a new strategy to find putative regulatory mutations based on experimentally established motifs for transcription factors (TFs). In total we find 1,552 candidate regulatory mutations predicted to significantly reduce binding affinity of many TFs in hepatocellular carcinoma. We observe a highly significant mutation rate at CTCF motifs, in particular at base nine of its core motif in hepatocellular, esophageal, gastric and pancreatic cancers. Near the mutated motifs there is a significant enrichment of genes mutated in cancer, tumor suppressor genes, genes in KEGG cancer pathways and sets of genes previously associated to cancer. Experimental and functional validations support the findings. Furthermore, genes located within topologically associated domains have a significant difference in expression with the presence of CTCF mutations. The strategy can be applied to identify regulatory mutations in any cell type with established TF motifs and will aid identifications of genes contributing to cancer.	Genomes poster	Fundamental
P_Go014	777	Pieter Libin, Nassim Versbraegen, Lize Cuypers, Kristof Theys and Ann Nowé	Pieter Libin	A maximum likelihood method for classifying virus sequences	Background:The classification of virus sequences is essential to support epidemiological surveillance and patient care. The "Rega typing framework", an automated classification method that applies Neighbor-Joining (NJ) phylogenetics, has been shown an effective and popular tool to classify various viral pathogens. However, this method has some important limitations: (a) its scoring strategy evaluates the quality of the assignment indirectly, (b) the procedure is non-deterministic and (c) its cubic computational complexity prohibits the use of large reference sets.Methods:An alternative automated procedure for virus classification, based on maximum likelihood (ML) phylogenetic placement (i.e. pplacer), was developed and integrated in the "Rega typing framework". A score, that represents the confidence of the query sequence's location in a particular clade, was composed. The procedure assigns a classification by selecting the clade with the highest score, if that score exceeds a calibrated threshold.Results:The ML method was validated on a large dataset of hepatitis C virus sequences (Los Alamos HCV database, n=20016, >=800 base pairs per sequence) and compared to the NJ method that was applied on the same dataset. This comparison demonstrates a high level of concordance between the results for the ML and NJ method (97.367%).Conclusion:This research demonstrates the potential of phylogenetic placement to classify virus sequences. The method addresses several limitations of the NJ approach: it delivers (a) a score that directly signifies classification confidence, (b) a deterministic classification approach and (c) a linear time complexity with respect to the number of reference sequences.	Genomes poster	Health
P_Go015	721	Kartikay Chadha, Jo Knight and Andrew D Paterson	Kartikay Chadha	A Novel Method to identify Significant DNA motifs in the human genome associated with Alzheimer's disease.	Alzheimer's disease (AD) is a complex disorder influenced by both environmental and genetic factors. Around 47 million people worldwide are living with dementia, most have AD. Genome wide association studies (GWAS) have identified 21 associated loci (Lambert et al 2013).The proposed method is to compare the DNA sequences around the SNPs of interest (for example GWAS hits) (these regions will be referred to as Areas of Interest- AOI) with regions around matched SNPs in the rest if the genome (Areas Not of Interest- ANOI). We aim to identify motifs with significant differences in the frequency.Such motifs have the potential to help us understand the functional role of GWAS hits and identify risk variants in other studies.We are currently investigating AOI from the AD GWAS mentioned above. The most significant SNP at each locus (index SNPs) and all SNPs in high linkage disequilibrium (LD >0.8) are included. AOIs of 200bp around each SNP are defined. Index ANOI SNPs are matched to the AD index SNPs on the basis of allele frequency.We count of all possible DNA motifs (of a predetermined length) in the AOI and ANOI. Next the counts are grouped according to complementary strands matching and directional matching. Finally, statistical tests e.g. Fisher Exact test and Cochran Armitage trend test are performed.We will use this method to analyses other data such as expression quantitative trait loci data from the Genome-Tissue expression (GTex) project.	Genomes poster	Health
P_Go016	618	Matyas Pajkos and Zsuzsanna Dosztanyi	Matyas Pajkos	A novel motif centric protein alignment method	SLIMs (Short Linear Motifs) are common interaction modules that play critical roles in diverse biological pathways. SLIMs usually reside in disordered regions and their short length and weak phenotype makes their experimental discovery challenging. As a result, SLIM mediated interactions are highly underrepresented in current protein networks. This underlines the importance of computational approaches for the discovery of functional de-novo motifs.Currently, there are two main approaches for de-novo motif discovery. Alignment free methods seek to find enriched motif sequences in a group of related sequences. Alignment based methods, like SLIMPrints (1), exploit the specific evolutionary conservation of SLIMs. As functional SLIM sites show stronger evolutionary constraints compared to their disordered sequential neighborhood, this gives the appearance of island like conservation in multiple alignment of orthologues.However, evolutionary approaches rely heavily on good quality sequence alignments covering larger evolutionary distances. Such alignments are often not available for disordered protein segments, which harbor most SLIMs. In order to overcome this major limitation of evolutionary based de-novo motif discovery methods, we propose a novel SLIM specific alignment method. In this approach, the starting scoring is based on motif enrichments within homologous, and alignments are not forced in regions that have no evolutionary conservation. This enables a more accurate detection of evolutionary conservation over larger distances even within disordered segments, making it feasible to detect functional SLIMs within any homologous, from vertebrate to plants.1. Davey et al. Nucleic Acids Research. 2012 Sep 12; 40(21)10628-10641	Genomes poster	Fundamental
P_Go017	371	Pola Smirin-Yosef, Sarit Kahana, Idit Maya, Doron Levi, Lina Basel-Vanagaite and Mali Salmon-Divon	Pola Smirin-Yosef	A study of normal CNV variations in Israeli population	The Israeli population is composed of a collection of diverse ethnic groups. Each group shares specific genetic variations that passed from its common ancestors throughout the generations. Together with pathogenic events, non-pathogenic polymorphism happen to occur in ancestors, subsequently spread into the restricted genomic pool of its descendants. Providing a comprehensive data resource of non-pathogenic CNVs in the Israeli population pregnancies in order to characterize ethnic-specific polymorphism may greatly contribute to the routine genetic counseling done by the geneticists on a daily basis. Chromosomal Microarray Array (CMA) has had a high impact in clinical diagnostics, leading to the discovery of new genomic disorders, and has become an indispensable tool for routine molecular and cytogenetic testing. CMA is a first line diagnostic test for individuals with developmental disabilities, dysmorphic features and congenital malformations as well as fetuses with congenital malformations and abnormal growth.Here we apply a data mining approach on the results of CMA testing performed at the Raphael Recanati Genetic Institute, contains around 9000 tests from individuals, fetuses with clinical abnormalities, and in fetuses from low-risk pregnancies.The use of an extracted ethnicity-based genetic information, in order to detect ethnic-specific CNV polymorphism in the Israeli population will allow geneticists to distinguish between relevant pathogenic genomic aberrations from benign ethnicity-related variations.	Genomes poster	Health
P_Go019	527	Farzana Rahman, Mehedi Hassan, Negusse Kitaba, Abdulsamie Hanano and Denis Murphy	Farzana Rahman	Analysis of the structure, function and evolution of caleosins: a family of multifunctional eukaryotic proteins	The multifunctional calcium-binding proteins termed as caleosins occur almost ubiquitously in two distinct eukaryotic clades, namely Viridiplantae and Fungi. The evolutionary pattern of caleosin gene occurrence is not consistent their descent from a common ancestor because the Fungi, along with animals and many protists, are members of the Opisthokonta, while the Viridiplantae are derived from unrelated green algal predecessors. This suggests that the caleosin genes may have originated in one of the current clades via by horizontal gene transfer from the other.We have studied the variation in caleosin gene and protein sequences across a comprehensive range of plant and fungal species utilising computational methods and pipelines to understand the structure and function of these proteins in detail. Protein structure predictions suggest that the calcium-binding and EF-hand domains are widely conserved across species, while there is considerable variation in the predicted loop region of the structure. While the biological functions of studied proteins have yet to be determined in detail, it is clear that these proteins have several subcellular locations and participate in a range of physiological processes in both plants and fungi, including acting as peroxygenases. One of the most important of these roles appears to be in responses to a range biotic and abiotic stresses, including plant-fungal interactions.In this research, we describe additional studies that have been carried out to shed light on the origin and functions of this intriguing group of proteins.	Genomes poster	Biotechnology
P_Go020	842	Heinz Himmelbauer, Alexandrina Bodrug, J. Mitchell McGrath, Britta Schulz and Juliane C. Dohm	Heinz Himmelbauer	Analyzing the genomes of wild and cultivated beets	Sugar beet is an important crop plant that accounts for roughly 25% of the world's sugar production per year. We have previously shown that sugar beet has a quite narrow genetic base, presumably due to a domestication bottleneck. To increase the crop's stress tolerance, the introduction of desirable traits from wild beets is required. As a first step, we have set out to characterize the genomes of sugar beet and its wild progenitor species, the sea beet. The genome of sugar beet was assembled from 454, Illumina and Sanger sequencing data, followed by integration with genetic and physical maps (Dohm et al., 2014). Efforts to further improve the sugar beet reference assembly are still ongoing, capitalizing on long-read technologies as well as on an optical mapping approach. We have sequenced the genomes of several sea beet accessions from different geographical areas to sample the diversity of the species. Lastly, we have shortlisted beets of differing genetic background for genome sequencing. We expect our work to provide a solid foundation to decipher the genetic makeup of a species, with profound implications for basic plant research, and for molecular breeding.ReferencesDohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, Rupp O, Sörensen TR, Stracke R, Reinhardt R, Goesmann A, Kraft T, Schulz B, Stadler PF, Schmidt T, Gabaldón T, Lehrach H, Weisshaar B, Himmelbauer H. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature 505 (2014), 546-549.	Genomes poster	Agro-Food
P_Go021	569	Jan Grau, Maik Reschke, Annett Erkes, Jana Streubel, Richard D Morgan,Geoffrey G Wilson, Ralf Koebnik and Jens Boch	Annett Erkes	AnnoTALE: bioinformatics tools for identification, annotation, and nomenclature of TALEs from Xanthomonas genomic sequences	Transcription activator-like effectors (TALEs) are virulence factors, produced by the bacterial plant pathogen Xanthomonas, which function as transcription activators inside plant cells. Their DNA-binding domains consist of a series of highly conserved tandem repeats of approximately 34 amino acids (AAs). Each repeat specifies one base of the DNA sequence bound by contacts with the 12th and 13th AAs, termed the repeat variable di-residue (RVD).Due to their repetitive nature, genomes harboring multiple TALE genes are notoriously difficult to assemble. Here we demonstrate that PacBio sequencing reads of sufficient coverage can completely span TALE genes without ambiguity. This advance has allowed us to assemble the genome of Xanthomonas strain Xoo PXO83, harboring 18 TALE genes, into a single contig.In anticipation of a rapid increase in the number of sequenced Xanthomonas genomes, we have developed an automated pipeline for annotating TALE genes and a systematic nomenclature for streamlining their functional analysis. We present 'AnnoTALE', a suite of bioinformatics applications for the analysis, annotation, and grouping of similar Xanthomonas TALEs into classes based on their RVD sequences. Based on these classes, we propose a unified TALE nomenclature that suggests related functionalities, and that elucidates base substitutions responsible for the evolution of of TALE RVDs and, consequently, specificities. Meanwhile, we have incorporated 12 additional Xanthomonas genomes into AnnoTALE, broadening our understanding of TALE evolution.	Genomes poster	Fundamental
P_Go022	484	Jikai Lei and Yanni Sun	Yanni Sun	Assemble CRISPRs from metagenomic data	CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats and Associated Proteins) allows more specific and efficient gene editing than all previous genetic engineering systems. These exciting discoveries stem from the finding of the CRISPR system being an adaptive immune system that protects the prokaryotes against exogenous genetic elements such as phages. Despite the exciting discoveries, almost all knowledge about CRISPRs is based only on microorganisms that can be isolated, cultured, and sequenced in labs. However, about 95% of bacterial species cannot be cultured in labs. The fast accumulation of metagenomic data, which contains DNA sequences of microbial species from natural samples, provides a unique opportunity for CRISPR annotation in uncultivable microbial species. However, the large amount of data, heterogeneous coverage, and shared leader sequences of some CRISPRs pose challenges for identifying CRISPRs efficiently in metagenomic data.Results: In this work, we developed a CRISPR finding tool for metagenomic data without relying on generic assembly, which is error-prone and computationally expensive for complex data. Our tool can run on commonly available machines in small labs. It employs properties of CRISPRs to decompose generic assembly into local assembly. We tested it on both mock and real metagenomic data and benchmarked the performance with state-of-the-art tools.The source code and the documentation of metaCRISPR is available at https://github.com/hangelwen/metaCRISPR.	Genomes poster	Fundamental
P_Go023	768	Denis Baurain, Mick Van Vlierberghe, Arnaud Di Franco and Hervé Philippe	Mick Van Vlierberghe	Automated tools for the generation and interpretation of single gene trees at a broad taxonomic scale.	Identifying orthology relationships among sequences is fundamental in phylogenomics; indeed, those are essential to understand evolution, diversity of life and ancestry among organisms. To build alignments of orthologous sequences, phylogenomic pipelines often start with a step of all-vs-all similarity search followed by a clustering with an algorithm such as "OrthoFinder" [Emms and Kelly (2015) "Genome Biol" 16:157]. For it to be as accurate as possible, proteomes of good quality are needed but their availability is limited to a small subset of the living beings. Therefore, large-scale taxonomic phylogenomic analyses imply the enrichment of preexisting orthologous groups with transcriptomic or genomic data and the need for robust tools for identifying orthologues from heterogeneous sequence data. To this end, we have developed a novel tool, "Forty-Two", along the lines of HaMStR [Ebersberger et al. (2009) "BMC Evol Biol" 9:157], whose aim is to add (and optionally align) sequences to thousands of preexisting multiple sequence alignments (MSA) while controlling for orthology relationships and potentially contaminating sequences. "Forty-Two" uses advanced heuristics based on a multiple Best Reciprocal Hit (multi-BRH) strategy against reference proteomes to distinguish orthologous and paralogous sequences among homologues. It is fully functional and has already been used in two high-profile phylogenomic manuscripts (under review) dealing with the animal tree of life. Here, we present the principles and algorithms underlying "Forty-Two" as well as the results of an extensive test suite of its features, in order to support its release to the public.	Genomes poster	Fundamental
P_Go024	526	Anna Ershova, Ivan Rusinov, Andrei Alexeevski, Sergei Spirin and Anna Karyagina	Anna Ershova	AVOIDANCE OF GATC SITE AS ADAPTATION TO HORIZONTAL GENE TRANSFER IN MIXED BACTERIAL POPULATIONS	Restriction-modification (R-M) systems serve as prokaryotic immunity systems. Notable are high precision of site recognition by restriction endonucleases (REs) and DNA methyltransferases (MTases) and mobility of R-M systems. Often different strains of the same species encode different R-M systems.We identified four species, Streptococcus pneumoniae, Neisseria meningitidis, Eubacterium rectale and Moraxella catarrhalis, whose strains encode mutually exclusive GATC-specific R-M system genes. Namely, MTase genes of Type II R-M systems that methylate GATC are mutually exclusive with methyl-directed Type IIM RE genes cleaving methylated GATC: their simultaneous activity would kill a host cell. However, cases of horizontal transfer of R-M systems between strains with opposite methylation status are confirmed by phylogenetic analysis. Also mutually exclusive systems are encoded in homologous genome regions of different strains.We suggest a possible mechanism facilitating transfer of mutually exclusive R-M systems. Recognition sites of Type II R-M systems are basically avoided in bacterial genomes due to self-toxicity of such R-M systems. Sites of other Types of R-M systems and Type IIM REs typically are not avoided because they are not self-toxic [1]. However, we observed GATC avoidance in all 61 studied genomes of four mentioned species, including 34 genomes that encode Type IIM REs. We suppose that avoidance of GATC sites in bacterial populations that include strains with opposite methylation status of genomes is an adaptation to horizontal transfer of R-M system genes.The work was supported by RNF grant 14-50-00029.1. Rusinov I. et al, BMC Genomics, 2015, 16:1084.	Genomes poster	Fundamental
P_Go026	373	Annika Buerger, Boet van Riel, Frank Rosenbauer and Martin Dugas	Annika Buerger	BasicSTARRseq: a Bioconductor R-package for analyzing STARR-seq data	Self-transcribing active regulatory region sequencing (STARR-seq) was first described in 2013 by Arnold et al. and allows to identify and quantify enhancer regions in non-coding DNA in large scale. The R-package BasicSTARRseq provides routines for quality controls, analysis and visualization of STARR-seq data. The analysis part is mainly covered through the implementation of the computational procedure to call peaks i. e. identify possible enhancers, which was introduced in the above mentioned article. The peak calling is based on comparing sample data with input data of the STARR-seq experiment and computes p-values to estimate the peaks' reliability. By including user chosen parameters, for example two alternative binomial models for calculating the p-value, peak calling can be adjusted to different kinds of data. The procedure can further be adapted to whole-genome or targeted sequencing. Resulting peaks are annotated to allow an easy overview over the results or for further filtering steps. Quality controls and visualization are offered by routines for comparing different replicates, and the comparison of experiment data and target regions. For plausibility checks or further explorative analysis the package also provides some functions to compare output tracks of other analysis (like peak lists of Chip-seq data, but also other data chosen by the user) with STARR-seq data. BacisSTARRseq includes test datasets extracted from the published data of Arnold et al.	Genomes poster	Biotechnology
P_Go027	548	Mathu Malar C, Jennifer Yuzon,Takao KasugaandSucheta Tripathy	Mathu Malar C	Benchmarking the genome assembly of Phytophthora ramorum Pr102 using third generation sequencing technology	Phytophthora ramorum is the causal agent of Sudden Oak Death disease that has killed over a million trees in coastal California. The P. ramorum Pr102 genome was assembled into 65 MB and 2576 scaffolds with 12 MB gaps in 2006. With the help of improved sequencing technology PacBio produced (~435399 reads, coverage 25X) and with the support of Illumina sequences (20942377 reads, coverage 10X) and the Sanger contigs (7589 contigs, ~54.4 MB), we used several combinations of error correction protocols to produce refined assembly. We followed a new three step error correction protocol resulted in 49% (206487 reads ~1.3GB) corrected reads. We have performed several combinations of hybrid assembly for optimizing genome assembly. Simulated jump libraries (6k, 10k insert size) was generated from the error corrected pacbio reads. First, we assembled error corrected reads with Celera assembler resulting in 2735 contigs (77Mb).Then redundans pipeline was used to reduce heterozygous contigs from the Celera assembled contigs along with the simulated jump libraries of the PacBio and the 2006 assembly (56758 reads and 20K insert size).The latest assembly has only 220 gaps with 2005 scaffolds. CEGMA analysis reveals about 95.7% COGs present. Total number of genes predicted using RNAseq data is 19278.The latest assembly was compared with the previous assembly and it has been found that 8.5Mb of gaps are closed consists of 4402 contigs of earlier. As a future work the Avh effector prediction and the synteny among the other Phytophthora's are yet to be understood.	Genomes poster	Agro-Food Ecosystems Fundamental

P_Go028	771	Mattias de Hollander, Victor Carrion, Marcio Leite, Jos Raaijmakers and Eiko Kuramae	Mattias de Hollander	Classification and binning of plant root and nodule metagenomes	The new advances and developments of high-throughput sequencing technologies are increasing the sequence length and depth. This enables construction of full length ribosomal reads and recovery of draft-genomes from metagenome sequences using automated binning methods. facilitating a better understanding of microbial communities in their natural environments based on taxonomic and functional characterization.Here we used 300bp paired-end Illumina Miseq and HiSeq runs from the plant root endosphere and plant root nodules. Reads are quality filtered and assembled into contigs. Gene abundances were assessed by aligning reads to a non-redundant gene catalogue and normalized by gene length and sequencing depth. Functional profiles were constructed with KEGG Orthology, Clusters of Orthologous Protein families and taxonomic classifications were added in order to determine which organisms and functions are enriched in the different treatments.We were able to reconstruct draft genomes of at least 20 endophytic bacterial genera from the endosphere by applying several automated binning methods. These reconstructed genomes have been mined for new biosynthetic pathways and genes involved in endophytic behaviour.Our results suggest that plants can shape the endophytic community and/or recruit endophytic microbes with specific functions from the rhizosphere for protection against fungal infections. Preliminary analysis of the nodule metagenome shows that the majority of the sequences can be classified to the Enterobacteriaceae family, which is known to be involved in nitrogen fixation and it beneficial to plant growth.	Genomes poster	Agro-Food Ecosystems
P_Go029	621	Jaime Castro-Mondragon, Alejandra Medina-Rivera, Samuel Collombet, Denis Thieffry, Morgane Thomas-Chollier and Jacques van Helden	Jaime Castro-Mondragon	Clustering and enrichment of Transcription Factor Binding Motifs within RSAT	Transcription Factor (TF) binding motifs (TFBMs) are classically represented as position-specific scoring matrices (PSSM). High-throughput experiments (e.g. ChIP-seq, Selex-seq) have enabled the discovery of many TFBMs, made available in an increasing number of motif databases, with a high level of redundancy. Another source of redundancy comes from the utilization of multiple motif discovery approaches. In this respect we present here matrix-clustering, a tool to identify, visualize and browse dynamically the groups of similar TFBMs. The clusters are displayed as trees with merged TFBMS at any branch. This tool emphasizes TF binding variability and reduce redundancy. By clustering entire databases (>4000 motifs), we further show that matrix-clustering correctly groups motifs belonging to the same TF family, and can drastically reduce motif redundancy. In parallel, as the location of TF Binding Sites (TFBSs) are relevant for TF biology, we developed two methods to assess global enrichment (matrix-enrichment) and spatial preferences (position-scan) for TFBMs within sets of query sequences. matrix-enrichment measures the TFBSs enrichment at any level of affinity (threshold-free method) and allows to visualize the enrichment in several set of sequences simultaneously. In some cases, TFs exhibit preferential positioning (e.g. relative to ChIP-seq peak summits or Transcription Start Sites). position-scan considers the distribution of TFBSs and reports TFs with positional bias deviated from a control distribution (e.g. flat), thereby revealing enrichment or avoidance of TFBSs at certain regions. Altogether these programs complement and simplify analyses of TFBMs, and are freely available in the RSAT suite (http://www.rsat.eu/).	Genomes poster	Fundamental
P_Go030	629	Corinna Ernst, Eric Hahnen and Schmutzler Rita	Corinna Ernst	CNV Detection on Multi Gene Panels	Targeted sequencing, which is restricted to the exons of genes known or assumed to be implicated in a special phenotype, decreases costs, storage requirements, and computation times significantly in comparison to whole genome and whole exome approaches. Hence, so-called multi gene panel approaches have become a widely-used tool in clinical diagnostics and in large-scale, genome-wide association studies. Targeted sequencing data is typically characterized by strong biases based on local mappability, GC-content, and further factors affecting capture efficiency. Recent studies revealed that existing tools for CNV detection on targeted data – which are mainly designed for the purposes of whole exome approaches exclusively – are not fully able to face these difficulties as they show a notable lack of accuracy and robustness.We present an approach for CNV detection which is tailored to the challenges of multi gene panel analysis. Our method relies on an improved normalization approach and the ability of position-wise examination of read depth. As the length of sequencing targets is restricted to typically much less than 1 million base pairs, multi gene panels allow for the abandonment of read binning due to computational feasibility.	Genomes poster	Fundamental
P_Go031	551	Inge Kjærbølling, Tammi Vesth, Jens C. Frisvad, Jane L. Nybo, Sebastian Theobald, Thomas O. Larsen, Uffe H. Mortensen and Mikael R. Andersen	Inge Kjærbølling	Co-evolution of secondary metabolite gene clusters and their host	Secondary metabolite gene cluster evolution is mainly driven by two events: gene duplication and annexation and horizontal gene transfer. Here we use comparative genomics of Aspergillus species to investigate the evolution of secondary metabolite (SM) gene clusters across a wide spectrum of species. We investigate the dynamic evolutionary relationship between the cluster and the host by examining the genes within the cluster and the number of homologous genes found within the host and in closely related species.Our strategy is to investigate annotated SM genes (SMURF) and through homology (based on BLAST) identify homologs in the genome and their location (inside or outside of clusters). An example case is the analysis of SM cluster families found across several species where the number of orthologs vary. Depending on the phylogenetic distribution of the SM clusters, this case illustrates horizontal gene transfer (HGT) and gene duplication events.Another case is clusters where one gene has one homolog outside the cluster and the rest of the cluster genes are unique to the cluster. This type of case would indicate a recent or ancestral gene duplication event or HGT. The analysis has been performed on 15 genomes (930 gene clusters) and 79 cases were identified. Comparative genomics based on clusters from 50 new Aspergillus genomes will be applied to get an understanding of which cluster evolution occurs in association with the host and which happens within the gene cluster.	Genomes poster	Biotechnology
P_Go032	566	Jonas Ibn-Salem, Enrique M. Muro and Miguel A. Andrade-Navarro	Jonas Ibn-Salem	Co-regulation of paralog genes in the three-dimensional chromatin architecture	Paralog genes arise from gene duplication events during evolution, which often lead to similar proteins that cooperate in common pathways and in protein complexes. Consequently, paralogs show correlation in gene expression whereby the mechanisms of co-regulation remain unclear.In eukaryotes, genes are regulated in part by distal enhancer elements through looping interactions with gene promoters. These looping interactions can be measured by genome-wide chromatin conformation capture (Hi-C) experiments, which revealed self-interacting regions called topologically associating domains (TADs).We hypothesize that paralogs share common regulatory mechanisms to enable coordinated expression according to TADs.To test this hypothesis, we integrated paralogy annotations with human gene expression data in diverse tissues, genome-wide enhancer-promoter associations, and Hi-C experiments in human, mouse, and dog genomes. We show that paralog gene pairs are enriched for co-localization in the same TAD, share more often common enhancer elements than expected and have increased contact frequencies over large genomic distances.Combined, our results indicate that paralogs share common regulatory mechanisms and cluster not only in the linear genome but also in the three-dimensional chromatin architecture. This enables concerted expression of paralogs over diverse cell-types and indicate evolutionary constraints in functional genome organization.	Genomes poster	Fundamental
P_Go033	344	Florian Schmidt, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K. Polansky, Peter Ebert, Karl Nordström, Matthias Barann, Anupam Sinha, Sebastian Fröhler, Jieyi Xiong, Azim Dehghani Amirabad, Fatemeh Behjati Ardakani, Barbara	Florian Schmidt	Combining transcription factor binding affinities with an open chromatin prior for accurate gene expression prediction	The binding and contribution of Transcription Factors (TFs) to cell specific gene expression is often deduced from open-chromatin measurements to avoid cost and labour intensive TF ChIP-seq assays. It is important to develop reliable and fast computational methods for accurate TF binding prediction in open chromatin regions (ORCs). Here, we report a novel segmentation-based method, TEPIC, to predict TF binding by combining sets of OCRs with position weight matrices. TEPIC can be applied to various open chromatin data, e.g. DNaseI-Seq and NOMe-Seq, using either peaks or footprints as input data. TEPIC computes TF affinities as a quantitative measure of TF binding strength and we show that low affinity binding sites predicted in this way improve performance over a simple presence/absence classification. Further, we show that while footprints called from OCRs capture most essential TF binding events, OCR peaks deliver the best prediction performance. Using machine learning techniques, we assessed the importance of individual TFs for gene expression and found that TEPIC scores nearly reach the quality of TF ChIP-seq data. Finally we show that TEPIC predicts all major known key transcriptional regulators in primary human hepatocytes and CD4+ T-cells emphasizing the reliability and applicability of our method.	Genomes poster	Fundamental
P_Go034	411	Yuvia Alhelí Pérez-Rico, Valentina Boeva, Allison Mallory, Angelo Bitetti, Sara Majello, Emmanuel Barillot and Alena Shkumatava	Yuvia Alhelí Pérez-Rico	Comparative analyses of super-enhancers reveal conserved elements in vertebrate genomes	Super-enhancers (SEs) are extensive hyperactive chromatin regions comprising cis-regulatory elements. Mammalian SEs have been described as central players in driving transcriptional networks that define cell fate and differentiation processes (Hnisz et al. Cell 2013, Vahedi et al. Nature 2015, Thakurela et al. Genome Res 2015). Despite their key regulatory functions, it has not been determined if the characteristic features of mammalian SEs are common to vertebrate SEs outside of the mammalian clade. We identified SEs in pluripotent cells and adult tissues of zebrafish and performed interspecies comparisons with mouse and human SEs. Similar to mammals, zebrafish SEs are highly cell- or tissue-specific. However, the genomic distribution of zebrafish SEs differs from that of the mammalian one, as zebrafish SEs are mainly overlapping intergenic sequences. Despite their overall low sequence conservation, a fraction of SEs maintained their association with orthologous genes in the three species analysed. Strikingly, these SEs displayed higher sequence conservation than the SEs without maintained orthologous associations. Moreover, functional dissection of two SEs associated with orthologous genes revealed zebrafish and mouse SE regions acting as enhancers with conserved functions. In addition, analysis of chromatin accessible regions predicted transcription factors regulating pluripotency in zebrafish. Our analyses determined similarities and differences between vertebrate SEs, and provide SE and transcription factor candidates for future functional studies of cellular identity.	Genomes poster	Fundamental
P_Go036	515	Rudy Pelicaen, Koen Illeghems, Luc De Vuyst, and Stefan Weckx	Rudy Pelicaen	Comparative genomic analysis reveals adaptations of Acetobacter ghanensis and Acetobacter senegalensis to the cocoa bean fermentation process	Fermented dry cocoa beans are the basic raw material for chocolate production. The cocoa pulp-bean mass content of the cocoa pods undergoes a spontaneous fermentation process, which is characterised by a succession of yeasts, lactic acid bacteria, and acetic acid bacteria (AAB). Acetobacter ghanensis LMG 23848T and A. senegalensis 108B are AAB species that originate from a Ghanaian spontaneous cocoa bean heap fermentation process. Based on extensive metabolic and kinetic studies, the strains have been indicated in previous studies as interesting functional starter cultures. Whole-genome sequencing of A. ghanensis LMG 23848T and A. senegalensis 108B using 454 pyrosequencing and 8 kb paired-end libraries, followed by assembly using Newbler and PCR-based gap closure, allowed to identify genetic adaptations to the cocoa bean fermentation ecosystem. Automated gene prediction and annotation using the GenDB pipeline was performed, followed by manual curation. Both species possessed the genetic ability for citrate assimilation and displayed adaptations in their respiratory chain. As is the case for many AAB, the missing gene encoding phosphofructokinase in the genomes resulted in a non-functional upper part of the Embden–Meyerhof–Parnas pathway and all genes encoding enzymes of an alternative tricarboxylic acid (TCA) cycle were retrieved. Comparative genome analysis of the cocoa-derived strains A. ghanensis LMG 23848T, A. senegalensis 108B, and A. pasteurianus 386B revealed significant synteny between the genome sequences of A. pasteurianus 386B and A. ghanensis LMG 23848T. Furthermore, 1733 core genes were identified in these strains and A. senegalensis 108B contained the highest number of singletons.	Genomes poster	Agro-Food
P_Go037	324	Mirjam Rehr and Stefanie Göllner	Mirjam Rehr	Comparing alignment and assembly strategies for targeted high-throughput sequencing with barcoded amplicons	Targeted high-throughput sequencing (HTS) increasingly finds its way into clinical applications - where both high sensitivity and high specificity are required. Together with advances in primer and sequencing technology this calls for tailored bioinformatics solutions. Targeted HTS with barcoded amplicons is facilitating alignment and even makes assembly-based approaches manageable.In this work we compare the performance of several alignment and assembly strategies with respect to runtime and quality scores.The analysis is performed on data which derives from leukaemia patients and has been targeted by HaloPlex HS (Agilent) and sequenced on a MiSeq (Illumina).More specifically we compare alignment to whole genome and to targeted regions, with alignment of the amplicon-barcoded reads to respective amplicon regions only. Furthermore we perform an assembly approach of the amplicon-barcoded reads within their joined fitting amplicon regions. For evaluation we consider quality scores like number of (uniquely) mapped reads, mapping quality, insert sizes, and strand bias.Concluding we discuss implications for the downstream analyses of variant calling and outline a clinical variant calling pipeline for targeted HTS data with barcoded amplicons.	Genomes poster	Health
P_Go038	438	Dimitrios Zisis, Paweł Krajewski, Iris Hovel and Maike Stam	Dimitrios Zisis	Comparison of computational methods for 4C-seq NGS data analysis	Circular chromosome conformation capture (4C) is a cost effective and powerful high resolution methodology, which through the use of high throughput sequencing can study DNA contacts made across the genome by a given genomic site of interest (referred to as a “viewpoint” or “bait”). 4C-seq is a technology with a significant advantage because only the sequence of one of the contacting sites of interest needs to be known. Although until now 4C-seq has been used mainly in human, mouse and model plants, there is still plenty of space for further development. During the last years, the deep study of 4C-seq technology resulted in various methods and tools for the analysis of 4C-seq data, with most important being the 4Cseqpipe, FourCseq, FourSiq and recently 4Cker. Their basic algorithms include all steps for the preprocessing of next-generation sequencing reads, the creation of in-silico library of restriction fragments, read alignment, and contact frequency estimation. By studying these methods we identify differences and similarities in the consecutive steps like the treatment of mapped reads (uniqueness), the estimation of fragment coverage and of contact frequencies to avoid bias, and the normalization and statistical analysis. The purpose of this presentation is to compare those four methods in 4C-seq and discuss about the various computational needs which are necessary for this analysis. We compare our algorithm -4CseqR- with the existing ones in terms of the efficiency and interpretation of each step using as example the data obtained in the experiment with Arabidopsis thaliana and FLC locus as the viewpoint.	Genomes poster	Biotechnology
P_Go039	25	Sarah Sandmann, Aniek de Graaf, Bert van der Reijden, Joop Jansen and Martin Dugas	Sarah Sandmann	Confident Variant Calling in NGS Data – A Mission Impossible?	For decades of years Sanger sequencing has been the gold standard in the field of sequencing. The launching of next-generation sequencing (NGS) techniques has reduced time and costs of sequencing. However, data often contains false positive calls and even today Sanger sequencing is still used to validate the called variants in NGS data.Considering three common next-generation sequencers - Roche 454, Ion Torrent PGM and Illumina NextSeq – we developed optimized variant calling pipelines to automatically reduce the number of false positive calls. Combining information of 23 diverse parameters characterizing the called variants we determined individually calibrated generalized linear models (GLMs). The models rely on amplicon-based targeted sequencing data (19 genes, 28,775bp) from seven to twelve patients with myeloid dysplastic syndrome (MDS). Testing of the models was performed using sequencing data from three additional MDS patients.We succeeded in filtering out 76% of the false positive SNVs and 97% of the false positive indels by applying our model approach. An increase in positive predictive values by by factors of 1.07 to 1.27 regarding SNV calling and by factors of 3.33 to 53.87 regarding indel calling could be observed. However, with respect to clinical diagnostics it should be considered that even the optimized results still contain false positive, as well as false negative calls.	Genomes poster	Health
P_Go040	729	Remi-Andre Olsen	Remi-Andre Olsen	De novo genome sequencing as a service	De novo genome sequencing is time consuming and resource intensive. The National Genomics Infrastructure in Stockholm is a publically funded genomics core facility. We have addressed the challenge of providing these methods as a service to a broad variety of research groups in Sweden. In contrast to smaller labs, de novo sequencing at this scale requires a focus on quality control, traceability and efficiency through automation.We present a bioinformatics analysis pipeline, NouGAT, for producing draft genome assemblies. It automates a set of common tasks usually performed in the first stages of de novo sequencing project: read-preprocessing, quality control, parallelized genome assembly and validation of the produced assemblies. All of our software is freely licensed and open source (http://opensource.scilifelab.se). We also present our ongoing work of evaluating new and emerging technologies for de novo sequencing; linked read sequencing by 10x Genomics, long read sequencing by Oxford Nanopore Technologies and the Illumina NeoPrep system.In the period of June 2015 to May 2016, our facility delivered 94 Illumina sequenced NouGAT genome assemblies to our users ranging from microbes to mammals. We show two microbes assembled from low coverage Nanopore data and highlight the case of a bacteriophage we were able to sequence using the NeoPrep system, where all other attempts using different techniques had failed.	Genomes poster	Biotechnology
P_Go041	865	Jasmijn Baaijens, Amal Makrini,Eric RivalsandAlexander Schoenhuth	Jasmijn Baaijens	De novo viralquasispecies assembly	Due to high recombination and mutation rates, viral genomes undergo rapid, significant evolutionary changes in short time. The ensemble of strains that infects a single host is referred to as viral quasispecies. The inherent genetic diversity can decisively hamper their computational exploration. In order to account for this, the primary goal of advanced viral pangenomics should be to develop reference systems based on strain-resolved, rather than consensus sequence, in analogy to curating individual rather than consensus genomes in human pan-genomics. Challenges are manifold. Most importantly, sequencing error rates can interfere unfavorably with strain abundance, which can obstruct error correction.Here, we present an algorithm for de novo viral quasispecies assembly that addresses this. In a first step, we apply the method presented by Valimaki et al. (2012) to construct an overlap graph based on a sound statistical model. We then apply an iterative clique enumeration procedure to merge reads into contigs.We evaluated our method on a lab-mix of five HIV-1 strains at 20000x coverage that was recently presentedas gold standard benchmark. We obtain contigs that cover 95% of the genomes of the strains, at an error rateof < 0.3%, clearly below the sequencing error rate of ~1%. This compares very favorably with assemblyprograms that have been suggested and/or found to work well in closely related setting: SPAdes: 97.4%coverage at 1.5% errors; metaSPAdes: 88.3%, 2.3%; VICUNA (addressing viral consensus genomeassembly): 16.5%, 3.2%. We also perform highly favorably in terms of contig length statistics.	Genomes poster	Health

P_Go042	386	Marita A. Isokallio and James B. Stewart	Marita A. Isokallio	Detecting purifying selection of mitochondrial DNA using a simple next-generation sequencing protocol	Mutations in mitochondrial DNA (mtDNA) are a known cause of several inherited diseases; symptoms of which may occur at any age with varying severity. However, transmission mechanisms of mtDNA mutations are still not fully understood, and the research is further complicated by the lack of methods for targeted manipulation of mtDNA. We use the mtDNA-mutator mouse (Trifunovic et al. Nature 2004) as a model to generate high levels of point mutations into mtDNA. With this model, we have shown a strong purifying selection during germline transmission against amino-acid substitutions on protein-coding genes in comparison to synonymous mutations (Stewart et al. PLoS Biology 2008). However, current methods used to detect mtDNA mutations (e.g. post-PCR cloning and sequencing, Duplex sequencing or circle sequencing) are unable to represent the entire mtDNA, or are laborious, expensive, or of low sensitivity.Here, we improve and combine the existing methods to a simplified, cost-efficient and highly sensitive next-generation sequencing protocol to detect rare mtDNA mutations. We verify the reliability of the improved protocol by sequencing mtDNA from mtDNA-mutator mice and three generations of their descendants. We observe, similar to our previous results obtained by Sanger sequencing, the purifying selection of mtDNA in mouse germline. Furthermore, we extend the previous study by detecting extremely low-level mtDNA heteroplasmy, on whole-mt-genome level, and by revealing purifying selection also in the soma. With the improved protocol, we will clarify the developmental timing of purifying selection in the mouse germline, as well as characterize mtDNA regions essential for replication and transcription.	Genomes poster	Fundamental
P_Go043	361	Kevin Vanneste, Bert Bogaerts, Qiang Fu, Raf Winand, Sigrid De Keersmaecker and Nancy Roosens	Qiang Fu	Development and implementation of a transversal NGS & bioinformatics platform at the Belgian Institute of Public Health: Deployment of user-friendly pipelines for routine use	Despite being a well-established research method, the use of NGS and bioinformatics for routine analysis in a public health setting remains a challenge. The NGS & bioinformatics platform was recently set up at the Belgian Institute of Public Health with the aim of utilizing NGS & bioinformatics for the diagnosis, surveillance, control and characterisation of potentially harmful organisms; and to promote public health genomics by the effective integration of NGS and bioinformatics into clinical use and public health policy. The platform develops solutions and provides data acquisition and analysis tools to complement the WIV-ISP laboratories services (including several national reference centers and laboratories); and to integrate the knowledge of genomics into public health policy. The platform has built up the capacity to generate and analyse NGS data through an in-house Miseq and advanced bioinformatics pipelines and databases. These services are developed under a strict quality system and offered as a high-quality service platform with the aim of being adapted for routine analysis for both surveillance and emergency cases. Specifically, standardized and streamlined pipelines optimized for specific cases are actively researched and developed, and are offered through a user-friendly system based on Galaxy to non-expert users. Expertise is present in regulation and quality control by active contributions to (inter)national workgroups for the development of guidelines and criteria for NGS. Novel solutions are actively being researched and developed with the aim of supporting a proactive public health policy. Lastly, interaction with other high-throughput technologies such as mass spectrometry, are actively being investigated.	Genomes poster	Health
P_Go044	605	Martina Fischer, Benjamin Strauch and Bernhard Y. Renard	Martina Fischer	Differential abundance testing on the strain level in metagenomics data	Rapid advances in NGS technologies massively increased the popularity and potential of metagenomics. Particularly the study of changes in microbial community composition under different conditions is of high relevance due to strong associations with disease and treatment effects.We present a new comprehensive tool including steps from read mapping to accurate differential abundance estimation of individual taxa down to strain level.We build on our previously published metagenomics quantification tool GASiC (Lindner et al., NAR 2013), which conducts reference-based read mapping and constructs a similarity matrix of genomes. This matrix enables the resolution of shared reads and allows estimating even low abundances of taxa with highly similar genome sequences.However, abundance estimates commonly given for taxa in metagenomics data refer to point estimates. Thus no further statistical measures are provided to assess reliability or variance of the estimates. This however plays a crucial role when comparing varying compositions aiming to detect species with differential abundance.We introduce a novel formulation of the problem as a generalized linear model, which resolves absolute mapping counts and delivers abundance estimates along with standard errors. Differential abundance of individual taxa can subsequently be assessed by divergence of corresponding abundance distributions. Further, p-values are calculated reflecting the significance of increased or reduced abundance and a false-discovery-rate (FDR) can be inferred.We demonstrate improved quantitative assessment and statistical identification of differentially abundant taxa in comparison to existing methods. Results are presented on diverse simulated benchmark and real data sets covering different sequencing technologies.	Genomes poster	Health
P_Go045	443	Fatemeh Behjati Ardakani, Nina Gasparoni, Laura Arrigoni, Sarah Kinkley, Matthias Barann, Sebastian Froehler, Peter Ebert, Andreas S. Richter, Gilles Gasparoni, Karl Nordstroem, Florian Schmidt, Stefan Wallner, Jan Hengstler, Kathrin Gianmoena, Cristina Cadenas, Barbara Hutter,	Fatemeh Behjati Ardakani	Distinct epigenetic architectures in bidirectional promoters revealed by single cell analysis	Bidirectional promoters (BPs) are prevalent in eukaryotic genomes. It is poorly understood how the cell integrates different epigenomic information, such as transcription factor (TF) binding and chromatin marks, to determine directionality of gene expression.For example, bimodal distributions of activating histone marks (HMs) are found at BPs, but the question remains unresolved if HMs spread along a BP as part of its regulation.We utilize single cell RNA-seq data and a novel homogeneity score to discover that BP regulation is more complex than previously described. The two genes at a BP may show concordant (homogeneous) or discordant (heterogeneous) expression distributions.Using epigenomic datasets we observe distinct patterns of TF binding and HMs in both groups. New computational models show that these patterns reflect positional preferences of binding TFs that regulate the observed differences in gene expression distributions.Further, we find that the distance between the two transcription start sites (TSS) impacts the correlation of nascent RNA expression, the likelihood of heterogeneous single cell expression, and involvement of upstream enhancer marks in gene regulation.Despite the bimodal distribution of HMs, we observe that the majority of histone marks associated with gene expression occurs downstream of the gene's TSS, except for upstream enhancer marks that are regulated by tissue-specific TFs.Thus, our results unravel an additional layer of complexity in the analysis of BP regulation. This suggests that future studies investigating the associations of regulatory elements in BPs should consider cell heterogeneity as a confounding factor.	Genomes poster	Fundamental
P_Go046	461	Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs and Wyeth Wasserman	Anthony Mathelier	DNA shape features improve transcription factor binding site predictions in vivo	Interactions of transcription factors (TFs) with DNA comprise a complex interplay between base-specific amino acid contacts and readout of DNA structure. traditionally, position-specific scoring matrices (PSSMs) are used to model TF binding sites (TFBSs). Here, we describe an approach that builds upon PSSMs and integrates DNA shape features derived from our DNAshape prediction method. Results from 400 human ChIP-seq datasets show that incorporating DNA shape features (helix twist, minor groove width, propeller twist, and roll) with PSSM sequence-based scores in a machine learning framework consistently improves the accuracy of TFBS predictions. Improvement is also observed when TF flexible models (TFFMs) and a machine learning-based approach are used in lieu of PSSMs. Incorporating DNA shape information is most beneficial for E2F and MADS-domain TF families. Results from the analysis of MADS-domain TFs highlight the importance of propeller twist in a TFBS position-specific manner.	Genomes poster	Fundamental
P_Go048	346	Serghei Mangul, Harry Taegyun Yang, Sagiv Shifman, Eleazar Eskin and Noah Zaitlen	Serghei Mangul	Dumpster diving in RNA-sequencing to find the source of every last read	High throughput RNA sequencing technologies have provided invaluable research opportunities across distinct scientific domains by producing quantitative readouts of the transcriptional activity of both entire cellular populations and single cells. The majority of RNA-Seq analyses begin by mapping each experimentally produced sequence (i.e., read) to a set of annotated reference sequences for the organism of interest. For both biological and technical reasons, a significant fraction of reads remains unmapped. In this work we develop a read origin protocol (ROP) aimed at discovering the source of all reads, originated from complex RNA molecules, recombinant antibodies and microbial communities. Our approach can account for 98.8% of all reads across poly(A) and ribo-depletion protocols. Furthermore, using ROP we show that immune profiles of asthmatic individuals are significantly different from the control individuals with decreased average per sample T-cell/B-cell receptor diversity and that immune diversity is inversely correlated with microbial load. This demonstrates the potential of ROP to exploit unmapped reads to better understand the functional mechanisms underlying the connection between immune system, microbiome, human gene expression, and disease etiology. The ROP pipeline is freely available at https://sergheimangul.wordpress.com/rop/	Genomes poster	Biotechnology
P_Go049	823	Christopher Schröder, Felix Mölder, Christoph Stahl and Sven Rahmann	Felix Mölder	EAGLE: an easy-to-use web-based exome analysis environment	High throughput exome sequencing is a widely used technology for deciphering mutations in the coding regions of a genome at relatively low cost. While bioinformatics analyses of exome sequencing data mostly agree on best practices regarding the analysis steps, called genomic variants depend on the set of parameters and applied filtering.We present EAGLE, a software that combines a best practices variant calling workflow with a web frontend. By storing the called variant information in HDF5 files (instead of SQL databases), EAGLE allows filtering and parameter tuning in almost real time. This enables iterative tuning of thresholds, or the selection of different samples for filtering by medical PIs via the web interface. The web interface presents metadata, annotations, quality control data and statistics to facilitate a comprehensive data analysis on different levels.	Genomes poster	Health
P_Go050	519	Clemens Messerschmidt, Dieter Beule and Manuel Holtgrewe	Clemens Messerschmidt	Efficient and Reliable HTS Data/Sample Consistency Check based on HLA Types	The HLA (human leukocyte antigen) type consists of 6 alleles of the highly variable MHC class I genes,overall more than 11,000 different alleles are known today (Robinson et al., 2014). A combination of alleles willalmost certainly be unique for any individual and therefore can serve as a fingerprint for any humansample. Recent algorithmic progress (Szolek et al., 2014) allows for proper analysis of the highly variable HLAgenes with high-throughput sequencing (HTS). Given a reasonable read coverage, reliable 4-digit HLAtype determination is feasible from WGS, WES as well as RNA-seq data.We propose to use this opportunity for an efficient and reliable consistency check for human HTS data,as it allows to match samples and detect sample swaps. Compared to conventional approaches forswap detection based on matching SNVs our approach has several major advantages: first it is easilyapplicable and reliable even for samples with high mutational burden, e.g., caused by micro-satelliteinstability or POLE1 defects. Furthermore, it is computationally rather inexpensive or even practicallyfree if the research question requires HLA typing, e.g., in cancer immunology.We implemented this idea as a publicly available stand-alone tool as well as part of our in-house analysispipelines, e.g., checking consistency between data generated from tumor material with the correspondingnormal tissue. Another important use case is checking for consistency of pedigrees in rare disease studiesby assessing the consistency of the HLA types of the offspring given the parents' HLA types.	Genomes poster	Health
P_Go051	662	Bartek Wilczynski and Jerzy Tiuryn	Bartek Wilczynski	Efficient method for detection of evolutionarily conserved regulatory elements	Regulatory sequences are frequently more conserved throughout evolution than other non-coding sequences. This is mainly due to the presence of the functional transcription factor binding sites within these elements. However, the evolutionary conservation of functional non-coding sequences is usually less stringent than coding sequences because of the lower constraints on the non-binding sequence parts as well as the possibility of retaining function of the element even after slight rearrangement of the binding sites. We have previously developed a software tool, called Billboard, for detection of such elements using a sliding window approach and a scoring function penalizing non-matching motif occurrences between species and rewarding co-occurrence of motifs within a window length. While this tool gave us interesting predictions of known and novel regulatory elements it was very slow in operation as the scoring function needed to be evaluated on hundreds of random sequences to assess the empirical p-values of the conservation score in true homologous sequences. Here we present an improved version of the method that is based on Gaussian approximation of the background distribution. This method is much faster than the previous implementation and does not show signs of reduced accuracy when compared to the previous Billboard version. Our tests on 246 known enhancers from the RedFly database indicate that we can predict over 90% with less than 30% false positives. This work was supported by research grants awarded by the Polish Ministry of Science and Higher Education [N N519 652740], and by Polish National Science Centre (NCN) DEC-2012/05/B/NZ2/00567.	Genomes poster	Fundamental
P_Go052	631	Sokratis Kariotis, Jeroen de Ridder and Sjoerd Huisman	Sokratis Kariotis	Enhancer-gene networks for the identification of cancer driver genes affected by enhancer mutations	Dynamic and diverse epigenetic modifications on enhancers affect the expression of target genes through DNA looping. Aberrant epigenetic modifications on these regions may result in misregulated gene expression. As deregulated gene expression is one of the important hallmarks of cancer, the study of such genomic regulatory elements is an important field of study in cancer research. As a step towards identifying enhancers with a potential driving role in cancer, we have constructed a enhancer-gene (EG) network by pairing the recently defined enhancer regions with targeted genes based on the correlation between epigenetic mark enrichment and gene expression across a wide range of cell types. The constitutive pairings are subsequently validated in silico using Hi-C measurements that capture the 3D conformation of the chromosomes. The EG-networks are overlaid with known cancer genes and noncoding somatic variation obtained from whole cancer genome sequencing. These networks enable identification of enriched modules that point to cancer drivers that are affected through somatic variations in the non coding genome.	Genomes poster	Health
P_Go053	649	Laura Adams, Christina Boucher, Martin Muggli, Simon Puglisi and Shiho Sugimoto	Martin Muggli	Enzyme Selection for Optical Mapping is Hard	An important ongoing challenge in genomics is the detection of errors in draft genomes. Misassembly errors are caused by sequence reads too short to span repeated genomic regions which then confounds assembly software. High throughput mapping systems, such as those from OpGen, Inc. and Bionano Genomics, generate restriction maps for single DNA molecules on the order of 500 KB long. These maps indicate where specific enzymes nick or cleave the DNA molecules. Such maps then provide long range, structural information for the genome under study. Because they are much longer and generated independently of sequence read data, they can be used to detect assembly errors. Muggli et al., (Bioinformatics, 2015) recently showed that aligning assembled contigs to restriction maps provides valuable information in misassembly detection. However, this only held true with certain enzyme combinations.Map alignment based assembly validation only works well when the restriction site patterns for a given genome and set of enzymes is sufficiently diverse across the genome. Otherwise, misassembled contigs may align by chance to regions of a simpler, more repetitive map.The aforementioned work relies on simulation and exhaustive search across all enzymes to select the most informative maps. In practice, only the reads and assembled contigs are available to select enzymes. Thus, the enzyme selection problem is to ensure the restriction site patterns across a set of contigs are distinct. In this work, we formalize the problem of enzyme selection for misassembly detection, suggest suffix array algorithmic solutions, and analyze their computational complexity.	Genomes poster	Fundamental
P_Go054	596	Teresa Szczepińska and Dariusz Plewczynski	Teresa Szczepińska	Epigenetic marks of the chromatin 3D structure	Combinations of the epigenetic marks along the genome determines patterns of gene expression, DNA replication, and other functions. What is important is that those processes occur in the three dimensional structure of the chromatin and such structure is adding another layer of regulation. Nuclear space consists of general compartments - euchromatin or heterochromatin regions. ChIA-PET and Hi-C experiments give us information about loops and domains within the chromatin structure. On the other hand experiments like ChIP-seq, GRO-seq, Bru-seq, ATAC-seq gives the information about chromatin marks and DNA accessibility. We propose a Bayesian network classifier to discover causative link between chromatin marks and loop placement into euchromatin/heterochromatin region of the nucleus.	Genomes poster	Fundamental
P_Go055	576	Alba Crespi, David Longbottom and T. Ian Simpson	Alba Crespi	Establishing method selection criteria for meta-genomic sequence analysis using high-throughput sequence simulators	The revolution in next-generation sequencing (NGS) technologies has enabled a step-change in the way that sequence data is collected and used in Biology. One field in which the effect has been particularly striking is meta-genomics; the sequencing of mixed source nucleic acid samples. In particular, microbial community characterisation by sequencing is widely used in medical, agricultural and ecological settings to better understand the contribution of these complex cellular communities to system function. These studies have profound implications for human, animal and plant health and disease as well as in diverse areas such as forensic science, environmental pollution monitoring and climate modelling.The increasing quantity of metagenomic sequence data being generated and the diversity of its application area requires highly optimised and computationally scalable solutions to process and interpret these data. We present a comparative evaluation of meta-genomic analysis methods in which we use sequence simulators to generate gold-standard data against which to benchmark the efficacy of the methods. We use our method to develop an approach to estimate errors in taxonomic sequence assignment by perturbing the underlying taxonomic trees used in our simulations. Using the results from these quantitative analyses and considering usability, functionality and compatibility of the methods we present a novel pipeline for metagenomic analysis for both targeted and untargeted studies.	Genomes poster	Fundamental

P_Go056	520	Manuel Holtgrewe and Dieter Beule	Manuel Holtgrewe	Evaluation of structural variant methods for medium-sized deletions in clinical application	For clinical application of short read high-throughput sequencing (HTS) a proper understanding of capabilities and short comings of the methods is essential. Here we address the especially challenging medium size (roughly, 300-50kbp) structural variants (SVs).We improved the annotation of a gold standard data sets for germ line SVs (Parikh et al., 2016) and performed a systematic evaluation of the SV calling methods Delly (Rausch et al., 2012), Manta (Chen et al., 2015), and Lumpy (Layer et al., 2014).Our presentation includes results for the different SV classes and SV sizes using Illumina X Ten and HiSeq 2000 data and highlights strengths and limitations of the methods. Especially for medium size deletions our results in terms of true positive and false discovery rate provide a valuable resource for designing and planing discovery and validation strategies, e.g., in analysis of Mendelian disorders.ReferencesChen, et al. "Manta: Rapid detection of structural variants and indels for clinical sequencing applications." bioRxiv (2015): 024232.Layer, et al., 2014. "LUMPY: a Probabilistic Framework for Structural Variant Discovery." Genome Biology 15 (6): R84.Rausch, et al., 2012. Delly: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012 28: i333-i339.Parikh, et al., (2016). svclassify: a method to establish benchmark structural variant calls. BMC genomics, 17(1), 1.	Genomes poster	Health
P_Go057	748	Ehsan Motazedi, Chris Maliepaard, Richard Finkers and Dick de Ridder	Ehsan Motazedi	Exploiting Next Generation Sequencing to solve the Haplotyping puzzle in Polyploids	We evaluate three recently developed state-of-the-art haplotyping algorithms for polyploids that make use of Next Generation Sequencing (NGS) data, i.e. HapCompass , HapTree and SDhaP, through extensive simulations of random genomes and NGS reads, using tetraploid potato (Solanum tuberosome L.) as the model crop. We investigate the effects of various sequencing parameters and technologies, as well as SNP density, similarity between the homologues and ploidy level on the accuracy and efficiency of haplotyping, and suggest practical guidelines for designing haplotyping experiments using NGS Data.	Genomes poster	Fundamental
P_Go058	633	Claudia Calabrese, Nuno A Fonseca, Alvis Brazma and Oliver Stegle	Claudia Calabrese	Expression QTL mapping in a PanCancer cohort	Expression Quantitative Trait Locus QTL (eQTL) studies represent a key tool to understand the effects of genomic variation on gene expression levels. Here we present some preliminary results of the eQTL analysis carried out within the frame of the PanCancer project, an international collaborative effort to annotate similarities and differences between 30 different tumour types. Whole Genome Sequencing, with both germline and somatic calls, and matched tumour RNA-seq data from more than 1000 TCGA and ICGC cancer patients are available to this purpose.The search for shared patterns of gene expression regulation using cancer-specific molecular features, like somatic variation, and the high heterogeneity of the PanCancer dataset represent the main challenges of this eQTL analysis. To account for batch effects and hidden confounding factors, gene expression values were corrected and used with a linear mixed model, implementing known covariates and genetic kinship inferred from the germline genotype. For the association analysis, common germline SNPs were retained, whereas, to increase the chance to observe a shared somatic genomic variation across the PanCancer cohort, somatic SNVs were aggregated by enhancers and promoters available from the Roadmap Epigenomics Project.Interesting results are emerging from the data, i.e. a set of known cancer-driver genes found in cis and trans-associations with mutated enhancers in more than one cancer study. Further analyses to link the eQTL genomic variation and genes to function are being carried out to shed light on patterns of gene expression regulation in cancer.	Genomes poster	Health
P_Go059	435	Shay Ben-Elazar, Benny Chor and Zohar Yakhini	Shay Ben-Elazar	Extending partial haplotypes to full genome haplotypes using Chromosome Conformation Capture data	Motivation: Complex interactions among alleles often drive differences in inherited properties including disease predisposition. Isolating the effects of these interactions requires phasing information that is difficult to measure or infer. Furthermore, prevalent sequencing technologies limit used in these the essential first step of determining a haplotype to the span of reads, namely hundreds of bases. With the advent of pseudo-long read technologies, observable partial haplotypes can potentially span several orders of magnitude more. Yet, measuring whole-genome-single-individual haplotypes remains a challenge. A different view of whole genome measurement addresses the 3D structure of the genome – with great development of Hi-C techniques in recent years. A shortcoming of current Hi-C results, however, is the difficulty in inferring information that is specific to homologous chromosomes.Results: In this work we develop a robust algorithmic framework that takes two measurement derived datasets: raw Hi-C and partial short-range haplotypes, and constructs the full-genome haplotype as well as phased diploid Hi-C maps. By analyzing both data sets together we thus bridge important gaps in both technologies – from short to long haplotypes and from un-phased to phased Hi-C. We demonstrate that our method can recover ground truth haplotypes with high accuracy, using measured biological data as well as simulated data. We analyze the impact of noise, Hi-C sequencing depth and measured haplotype lengths on performance. Finally, to further demonstrate the importance of phased Hi-C, we use the inferred 3D structure of a human genome to point at transcription factor targets nuclear co-localization.	Genomes poster	Fundamental
P_Go060	704	Franziska Metge and Christoph Dieterich	Franziska Metge	FUCHS - Full circle characterization using RNAseq	Circular RNAs (circRNAs) belong to a recently re-discovered class of RNA species that emerge during RNA maturation by a process called back-splicing. Circular transcripts, as opposed to canonical linear transcripts, from when downstream 5' splice sites are linked to upstream 3' splice sites. Recent advances in next-generation sequencing (NGS) brought circRNAs back into the focus of many scientists. Since then, several studies reported that circRNAs are differentially expressed across tissue types and developmental stages, implying that circRNAs are regulated and not a mere by-product of splicing. Though functional studies have shown that some circRNAs could act as miRNA-sponges, the function of most circRNAs remains unknown. To expand our understanding of possible roles of circular RNAs, we propose a new pipeline that fully characterizes candidate circRNA structure from RNAseq data -- FUCHS.Currently, most computational prediction pipelines use back-spliced reads only to identify circular RNAs. Taking into account all RNA-seq information from long reads (typically > 150 bp), FUCHS reveals additional information about exon coverage, amount of double break point fragments, different circular isoforms arising from one host-gene, and alternatively spliced exons occurring within the same circRNA boundaries.The extra features provided by FUCHS enable the user to perform differential motif enrichment and miRNA seed analysis to determine potential regulators involved in circRNA biogenesis. FUCHS is an easy to use python-based pipeline that contributes to new aspects of the circRNA research.	Genomes poster	Fundamental
P_Go061	512	Yad Ghavi-Helm, Sascha Meiers, Aleksander Jankowski, Jan Korbel, Eileen Furlong	Sascha Meiers	Functional impact of genomic rearrangements on chromatin organization and transcriptional regulation	With chromatin conformation capture-based techniques such as Hi-C it has become possible to study the interaction between cis regulatory elements in the genome (enhancers, promoters, etc.) at a genome-wide scale. Yet our understanding of how these interactions form and under which circumstances they regulate gene expression is only rudimentary. Recent studies investigated somatic chromosomal aberrations or used CRSIPR/Cas9 to edit key regions such as boundaries of topologically associated domains to understand the functional consequences of rearrangements. However, those results remain limited to few exemplary cases.In this ongoing work we used highly rearranged balancer chromosomes in Drosophila melanogaster as a genome-wide model for large-scale genomic rearrangements to investigate how the linear structure of the genome influences chromatin organization and gene expression. We analyzed expression in adult flies as well as embryos and compared the rearranged chromosomes to their normal state in a heterozygous cross, which intrinsically normalizes for trans regulatory effects.Surprisingly, initial results suggest little drastic effects, with some changes in the larger chromatin interaction landscape and hundreds of genes showing moderate differential expression between the two haplotypes. While analysis is still ongoing, we expect this model to yield unique insights into the interplay between chromatin topology and transcriptional regulation in cis.	Genomes poster	Fundamental
P_Go062	834	Leon Kuchenbecker, Knut Reinert and Peter Robinson	Leon Kuchenbecker	Functional T cell receptor beta-chain gene sequence discrimination using SVMs	Adaptive immunity is driven by a highly diverse population of T and B cells expressing unique antigen receptor proteins. The genetic mechanism allowing for this diversity is the somatic recombination of the encoding genes occurring during the differentiation of stem cells into these types of lymphocytes.Targeted enrichment of the recombined genes combined with high throughput sequencing allows for the in depth capture of those immune repertoires. So far, most such immunogenetic sequencing applications use the identified gene sequences only as an identifier for unique clonotypes, e.g. in order to measure repertoire features such as entropy, the clonotype distribution or to track pre-characterized clonotypes.While B cell receptors (or antibodies) are able to directly bind to a target antigen, the T cell receptor (TCR) can only recognize peptides bound to a presenting MHC molecule. The MHC is a highly polymorphic, polygenic protein, which makes a functional assessment of TCRs very hard.In our work, we assess the possibility of a functional annotation of TCR clonotypes using kernelized machine learning techniques, namely Support Vector Machines. Our method is designed to discriminate two subtypes of T cells, cytotoxic (CD8) and helper (CD4) cells, based on the recombined gene sequences acquired by repertoire sequencing. Our approach aims to improve understanding of how the TCR binds the peptide-MHC complex, and also to provide a foundation for future efforts to exploit NGS-based TCR profiling for the characterization of antigen specificity and clinical classification applications.	Genomes poster	Health
P_Go063	538	Rajesh Patel	Rajesh Patel	Genome sequence annotation of Salinicoccus sp BAB_3246 strain isolated from salt Pan, Gujarat, India	In present work genome sequence of strain Salinicoccus sp BAB_3246 from salt pan of little Ran of kutch, Gujarat, India was annotated with Rapid Annotation using Subsystem Technology (RAST). Comparison of genome data was done with Salinicoccus roseus, Salinicoccus carnicancri Crm, Salinicoccus albus DSM 19776, Salinicoccus luteus DSM 17002 and Salinicoccus halodurans H3B36. S. halodurans strain was having the highest genome size of 2,778,379 bp followed by 873,136bp, 713,204bp, 679,606bp, 461,933bp and 342,819bp respectively for S.carnicancri Crm, Salinicoccus sp BAB_3246, S. luteus DSM 17002, S. roseus and S.albus DSM 19776 strain. Maximum 2839 coding sequences were reported for S. halodurans followed by 1691,863, 668, 449 and 334 correspondingly for Salinicoccus sp BAB_3246, S. carnicancri Crm, S. luteus DSM 17002, S. roseus and S. albus DSM 19776 strain. Maximum 73 RNAs was reported for S. halodurans followed by 71 for Salinicoccus sp BAB_3246 and 46 for S. carnicancri Crm strain. Total 27 subsystem annotations were resulted from the RAST based annotation process. Only two subsystems Motility and Chemotaxis; and Photosynthesis were absent in all the six strain. Highest subsystem reported in Salinicoccus halodurans with compare to other strain. Data mining of relevant the presence of stress response genes and operator pathway for degradation of various environmental pollutants. Annotation data and analysis indicate the possibility of explore the stress for pollution control in industrial influent and marine environment.	Genomes poster	Biotechnology
P_Go064	654	Alex Salazar, Marcel van den Broek, Melanie Wijsman, Arthur Gorter de Vries, Pilar de La Torre, Anja Brickwedde, Nick Brouwers, Jean-Marc Daran and Thomas Abeel	Alex Salazar	Genome sequencing and assembly of the biotechnology-relevant yeast strain, CENPK113-7D, using only Oxford Nanopore long-reads shows evidence for a heterogeneous population of cells	CEN.PK113-7D is a haploid strain of Saccharomyces cerevisiae that is used widely in biotechnology because of its robust growth characteristics in industrial settings. Although previous studies have assembled it's genome de novo with short-reads, these assemblies are fragmented requiring biased scaffolding via homology to other completed yeast genomes. In this study, we present one of the most complete de novo genome assemblies of an eukaryotic organism using only sequencing data obtained on Oxford Nanopore Technology's MinION sequencing platform. By sequencing CEN.PK113 7D on a single flow cell, we were able to obtain over 40x coverage of the genome with an average read-length of 10 Kbp—sufficient for a long-read-only assembly. Using Miniasm and Canu, we obtained a 21 contig assembly with an N50 of 712 Kbp of which 11 of the 16 chromosomes were assembled in a single contigs from telomere-to-telomere. This is 36-fold reduction in the number of contigs and a 18-fold increase in the N50 in comparison to a previous short-read assembly of CENPK113-7D. Interestingly, we show evidence of a heterogeneous genomic architecture in CENPK113-7D: one population with a translocation between chromosomes 3 and 8 and another without the translocation. The heterogeneous genomic architecture is supported by read-data and by experimental results. This study does not only provide a valuable resource and insight to the scientific community but also shows the promising capabilities of Oxford Nanopore Technology.	Genomes poster	Biotechnology
P_Go065	583	Jole Costanza, Chiara Ronchini, Margherita Bodini, Luciano Giacò, Anna Candoni, Renato Fanin, Alessandro Cignetti, Corrado Tarella, Antonella Padella, Giovanni Martinelli, Pier Giuseppe Pelicci and Laura Riva	Jole Costanza	Genomics of chemoresistant acute myeloid leukemia	In this work, we investigated the mutational landscape of chemoresistance by performing whole exome sequencing (WES) on the primary, relapse and remission samples coming from 30 acute myeloid leukemia (AML) relapsed patients (between 18 and 73 years of age). We observed that relapsing leukemias have similar median mutation rate per patient to primary tumors (29 vs 32); however, we detected a significant difference in the frequency of transversions between the two conditions (38.32% in primary versus 54.40% in relapse AMLs), indicating that chemotherapy influences the mutational spectrum at relapse. Analyzing this cohort, we confirmed that many of the mutations present in the primary tumor and that persist in the relapse are driver genes involved in chromatin remodeling and methylation (i.e. DNMT3A, EZH2 and ASXL1). In order to understand if the relapse-specific mutations are present in the primary tumors at very low frequency and escaped identification due to the sensitivity limitations of WES, we used Duplex Sequencing to identify mutations at very low variant allele frequency (<1/10000). Indeed, in one patient out of three analyzed up to date, we detected in the primary tumor mutations identified as relapse-specific by WES both in TET2 and KIT at variant allele frequencies lower than 0.005.	Genomes poster	Health
P_Go066	405	Ivo Pedruzzi, Catherine Rivoire, Andrea H. Auchincloss, Elisabeth Coudert, Guillaume Keller, Patrick Masson, Edouard de Castro, Delphine Baratin, Béatrice A. Cuche, Lydie Bougueleret, Sylvain Poux, Nicole Redaschi, Ioannis Xenarios and Alan Bridge	Ivo Pedruzzi	HAMAP - leveraging Swiss-Prot curation for the annotation of uncharacterized proteins	HAMAP (High-quality Automated and Manual Annotation of Proteins) is a rule-based automatic annotation system for the functional annotation of protein sequences. It consists of a collection of family profiles for determining protein family membership, and their associated annotation rules for attachment of functional annotation to member sequences. As well as the annotations themselves, HAMAP rules also specify the conditions under which these annotations may be applied, such as taxonomic constraints or a requirement for key functional residues (identified by structural or other experimental studies), thereby achieving high specificity by coupling predictions to presence of specific residues. Both HAMAP family profiles and annotation rules are created and maintained by experienced curators using experimental data from expertly annotated UniProtKB/Swiss-Prot entries. Part of the UniProt automatic annotation pipeline, HAMAP routinely provides annotation of Swiss-Prot quality for millions of unreviewed protein sequences in UniProtKB/TrEMBL. In addition, HAMAP can be used directly for the annotation of individual protein sequences or complete microbial proteomes via our HAMAP-Scan web interface at http://hamap.expasy.org. Originally developed to support the manual curation of UniProtKB/Swiss-Prot records describing microbial proteins, the scope and content of HAMAP has been continually extended to cover eukaryotic and lately also viral protein families.	Genomes poster	Fundamental
P_Go067	466	Seong-Jin Park, Gunhwan Ko and Byungwook Lee	Seong-Jin Park	HASV: Hadoop-Based NGS Analyzer for Predicting Genomic Structure Variations	The NGS technology produces large scale biologic data sets much cheaper and faster than the previous methods. As it is almost impossible to store or analyze such large scale NGS data with a traditional method on a commodity server, many problems arise. Hadoop is an alternative to this requirement. We aim to address the issues involved in the large scale data analysis on the cloud in bioinformatics. Accordingly, we propose analysis service for predicting genome structural variations associated with diseases by using Hadoop. The result of this study reveals that the system proposed in this study efficiently predicts genomic variations from large scale data sets.	Genomes poster	Biotechnology
P_Go068	749	Przemyslaw Szalaj, Paul Michalski, Zhonghui Tang, Przemyslaw Wroblewski, Yijun Ruan and Dariusz Plewczyński	Przemyslaw Szalaj	Hierarchical modeling of three-dimensional chromatin organization based on ChIA-PET data	Spatial organization of the genome plays an important role in its functioning and is closely related to gene expression level, DNA replication and repair and others. The basic units of this organization are topological domains and chromatin loops. Recent development of advanced chromosome conformation capture (3C) based methods such as Hi-C and ChIA-PET allow to quantify the interaction frequency between distant genomic loci and to infer the 3D chromatin conformation.We developed an algorithm to reconstruct the spatial chromatin conformation based on ChIA-PET data. We base our modeling on underlying biological structures, i.e. chromatin loops and topological domains. First, we employ the weak interactions to create the low-resolution contact maps that we use to position topological domains in relation to each other. Then we take the advantage of the ChIA-PET specificity that allows to target a particular protein in order to identify a set of strong interactions indicating chromatin loops. In our modelling we also consider CTCF motifs orientation and weak interactions between individual chromatin loops. Taken together, this allows us to create reliable models of selected genomic regions, whole chromosomes and even whole genome in a reasonable time. To facilitate the usage of our algorithm we also developed a webserver which allows users to easily generate 3D models, annotate and inspect them using an interactive 3D visualization tool and to produce various statistical plots and heatmaps for the selected regions.	Genomes poster	Fundamental

P_Go070	376	Alla Mikheenko, Vladislav Saveliev and Alexey Gurevich	Alla Mikheenko	Icarus: visualizer for de novo assembly quality assessment	Genome browsers have proven to be instrumental in genomic studies. However, there is still no recognized visualization tool for evaluation of de novo assemblies. We present Icarus – a novel interactive visualizer for assessment and analysis of genomic draft assemblies. The tool is freely available online and as a standalone application, integrated into the tool QUAST (Gurevich et al., 2013), see http://quast.sf.net/icarus.Icarus consists of two types of viewers. Contig Alignment Viewer places contigs according to their mapping to the reference genome, and colors differently correct and erroneous contigs. Misassembled contigs are broken into aligned blocks according to their mappings. Icarus supports all types of misassembly events detected by QUAST (relocations, inversions, etc). If several assemblies are provided, Icarus highlights similar contigs. The viewer can additionally visualize genes, operons, and reads coverage distribution along the genome.Contig Size Viewer places contigs ordered by size. This ordering is suitable for comparing assemblies when no reference is available, and to visualize such metrics as size of the largest contig, contig size distribution, and N50. When the reference is available, erroneous contigs and misassembly breakpoints are highlighted. The user can also click contigs to navigate between representations in both viewers.Icarus is also suited for metagenomic assemblies. It is integrated into MetaQUAST (Mikheenko et al., 2015), and provides visualizations of alignment to each reference genome separately.	Genomes poster	Biotechnology
P_Go071	18	Jens Friis-Nielsen,Jose Mg Izarzugaza and Søren Brunak	Jose Mg Izarzugaza	Identification of Known and Novel Recurrent Viral Sequences in Data from Multiple Patients and Multiple Cancers	The discovery of viruses and other disease-causing pathogens from high throughput sequencing data often requires that taxonomic annotation occurs prior to association to disease. Although this bottom-up approach is effective in some cases, it fails to detect novel pathogens and remote variants not present in reference databases. We propose an alternate approach utilizes sequence clustering for the identification of nucleotide sequences that co-occur across multiple sequencing data instances. Thus, not limited to reported species.We applied the workflow to 686 sequencing libraries from 252 different cancers and 56 controls. We used our pipeline to associate recurrent sequences to the onset of the disease but also to the use of common laboratory kits to identify common methodological or technical artifacts sourcing erroneous conclusions, as we have observed in the recent literature. We provide examples of identified inhabitants of the healthy tissue flora as well as experimental contaminants.	Genomes poster	Fundamental
P_Go072	862	Barbora Hanáková, Eva Budinská and Jan Oppelt	Barbora Hanáková	Identification of subtype specific microbiome from tumour tissue RNAseq data in colorectal cancer.	Colorectal cancer (CRC) is very heterogeneous disease in terms ofprognosis and response to therapy. There is direct and indirectevidence of heterogeneity not only on histopathological level, butalso on molecular level. Understanding of the causes of theheterogeneity is very important for the identification of newpredictive biomarkers, which might be helpful for betterstratification of patients. Despite the huge efforts in the lastdecade, the current molecular predictive and prognostic classifiersare only marginally better than standard clinical risk factors. Thereason why is in intra-tumoural heterogeneity on one side and oninability of molecular profiling to capture several other aspects thatmight be equally important to understanding of CRC heterogeneity.Microbiota has been recently associated with the development ofcolorectal cancer and may be one of the missing pieces in thecharacterization of CRC heterogeneity.The main objective of this study was to correlate microbiome withmolecular subtypes and known clinical variables of CRC.We applied Readscan to the raw RNAseq sequences of the CRC samples fromthe COAD study (The Cancer Genome Atlas) in order to identify non-humansequences in the RNAseq data of tumours. Next, we correlated theidentified bacterial OTU's with CRC molecular subtypes and clinicalvariables.We identified 66 bacterial OTU's specific for molecular andhistopathological CRC subtypes, 23 OTU's correlated with the stage and23 species with the localization of tumour.The work was supported by the project AZV 16-31966A.	Genomes poster	Health
P_Go073	769	Ines Vlahović, Matko Glunčić, Marija Rosandić and Vladimir Paar	Ines Vlahović	Identification of the higher order repeats from T.castaneum to Human and Neanderthal genome using computational Global Repeat Map method	Higher order repeats (HORs) function in species genomes is still mainly unknown. HOR could be classified as regular (head-to-tail "tandem within tandem pattern") and complex, where for regular ones is known that they are a result of recent evolutionary processes in primates. We use our Global Repeat Map method (http://genom.hazu.hr/tools.html) for identification of tandem repeats and HORs. Main characteristic of this method is creation of global repeat map of the investigated DNA sequence by direct mapping of it into frequency domain using complete K-string ensemble [1]. We identified in T.castaneum complex and, surprisingly, regular HORs, not identified previously in insects (only large tandem repeats and complex HOR with different size of primary repeat units were found). Moreover, in human and Neanderthal genome, we identified accelerated HOR structures [2] which are located in NBPF family gene. In addition, we confirm that there are no accelerated HOR structures in NBPF family gene of other primates genomes. NBPF family gene is relevant for human brain expansion and mutation in them, as well as number of variations, lead to neurological disease development (schizophrenia, autism, microcephaly and macrocephaly).[1] Glunčić M, Paar V. 2012. Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Res. 41:e17.[2] Paar V, Glunčić M, Rosandić M, Basar I, Vlahović I. 2011b. Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. Mol. Biol. Evol. 28:1877-1892.	Genomes poster	Fundamental
P_Go074	781	Björn Langer and Michael Hiller	Björn Langer	Identifying the functional role of transcription factors via phylogeny-aware discriminative sequence motif scoring	Many changes of morphological or other complex phenotypic traits result from gene expression changes. Such altered gene expression arises often from changes in cis-regulatory elements. That usually means the loss of important transcription factor (TF) binding sites, because the interaction between TFs and specific sites on the DNA is a key element of gene regulation.The Forward Genomics framework links phenotypic differences between species to their underlying genomic differences by focusing on the loss of a trait in independent lineages. However, its reliance on sequence conservation is a main limitation for its application on regulatory regions. We extend the Forward genomics strategy by taking into account the flexible organization of regulatory regions' functional units, the TF binding sites, in terms of both order and strength. Given a multi-species alignment and a set of regulatory regions, our tool systematically searches for TFs whose changes in binding affinity between species fit the phenotype signature and reports them ranked according to the level of fit. We prove the concept of our approach on both biological data and artificially generated regions. This method will contribute to discovering the transcription factors that are involved in the evolution of phenotypic changes between species.	Genomes poster	Fundamental
P_Go075	408	Nan Du and Yanni Sun	Nan Du	Improve homology search sensitivity of PacBio data by correcting frameshifts	Single-molecule, real-time sequencing (SMRT) developed by Pacific BioSciences produces longer reads than secondary generation sequencing technologies such as Illumina. The long read length enables PacBio sequencing to close gaps in genome assembly, reveal structural variations, and identify gene isoforms with higher accuracy in transcriptomic sequencing. However, PacBio data has high sequencing error rate and most of the errors are insertion or deletion errors. During alignment-based homology search, insertion or deletion errors in genes will cause frameshifts and may only lead to marginal alignment scores and short alignments. As a result, it is hard to distinguish true alignments from random alignments and the ambiguity will incur errors in structural and functional annotation. Existing frameshift correction tools are designed for data with much lower error rate and are not optimized for PacBio data. As an increasing number of groups are using SMRT, there is an urgent need for dedicated homology search tools for PacBio data.In this work, we introduce Frame-Pro, a profile homology search tool for PacBio reads. Our tool corrects sequencing errors and also outputs the profile alignments of the corrected sequences against characterized protein families. We applied our tool to both simulated and real PacBio data. The results showed that our method enables more sensitive homology search, especially for PacBio data sets of low sequencing coverage. In addition, we can correct more errors when comparing with a popular error correction tool that does not rely on hybrid sequencing.	Genomes poster	Fundamental
P_Go076	796	Sweta Talyan, Miguel Andrade-Navarro and Enrique Muro	Sweta Talyan	Improving the prediction of Human processed pseudogenes	Pseudogenes are extant genomic loci that are quite similar to their parental functional genes, but cannot be translated into functional proteins because of deleterious mutations. Pseudogenes are classified as processed, duplicated and unitary, depending on their biogenesis mechanisms such as retrotransposition, DNA duplication and gene decay respectively. Duplicated pseudogenes maintain the parental gene structure and all regulatory regions while the processed pseudogenes retain neither the upstream regulatory regions nor the introns. Recent studies confirm the tissue specific transcriptional activity of more than 13% of all human pseudogenes. For some of those, functional regulatory roles have been found, including being causative of diseases.Currently, psiDr/GENCODE is the standard repository of pseudogene annotations. It is based on Pseudopipe and Retrofinder prediction methods followed by HAVANA manual curations. These methods of ab-initio pseudogene detection and classification were developed at an early stage of the human genome annotation, when little sequencing information from human and other organisms were available: Pseudopipe (Zhang et al. 2003, 2006, Zhang and 2004), Retrofinder (Baertsch et al. 2008) and the method from Torrents et al. (2003). These methods are still the norm and rely mostly on homology.In the wake of data availability, better pseudogenes annotations is essential especially for humans and other model organisms. Towards this, we aim to develop a novel method for pseudogene genome-wide prediction specially processed pseudogenes that takes advantage on information provided by the annotation on all the genomes sequenced till now. Such method will improve the current pseudogene annotation and classification and facilitate better understanding of non coding genome in future.	Genomes poster	Fundamental
P_Go077	315	Erdogan Taskesen, Anniket Mishra,Danielle Posthumaand Yolande Pijnenburg	Erdogan Taskesen	Joint analysis of GWAS with epigenetic data revealed candidate markers in FTD/MND, and convergence in pathways.	The use of Genome-wide association studies (GWAS) have become a standard approach to identify genetic risk variants. However, in Frontotemporal dementia (FTD) only a handful of highly penetrant genetic variants have so far been identified. A currently important open question is the role of epigenetic factors, and whether these converge on biological processes, and as such cause degeneration of the frontal and temporal lobes.In this study we stepwise integrated the DNA-Methylation Profiles (DMP) with SNPs from a FTD GWAS study to detect novel risk-SNPs that may have been missed using conventional methods. We furthermore analyzed whether genetic and epigenetic processes converge on biological processes.Analysis of FTD patients with Motor Neuron Disease (FTD/MND) showed a homogeneous profile with in total 224 unique genes with significantly differential cytosine DNA-methylated levels (PFDR<0.05). Although DMPs are derived from peripheral blood, we demonstrate brain tissue specificity for the detected genes. Moreover, the Prefrontal-Cortex, and the Primary-Motor-Cortex were highly enriched (P<0.05).For the detection of novel candidate genetic markers, we extracted SNPs from GWAS FTD/MND that reached significance under P<0.05. After gene-mapping, we identified significant overlap with the 224 DNA-methylation markers (53 genes, P=0.0005) which indicates non-random behavior of genes that are target in FTD/MND. These genes are described with function in neuron or brain. Moreover, independent pathway analysis of GWAS and DMP genes showed convergence in biological processes. With these results we clearly show that understanding genetic and epigenetic factors are critical for unravelling the road to abnormal neurological development.	Genomes poster	Health
P_Go078	594	Thies Gehrmann, Jordi Pelkmans, Han Wösten, Johan Baars, Anton Sonnenberg, Marcel Reinders and Thomas Abeel	Thies Gehrmann	Karyollele specific expression in Agaricus bisporus	Background: The average cell in the cultivated white button mushroom, Agaricus bisporus, contains six nuclei, each being a copy of one of the two parental nuclei, referred to as the homokaryons of A. bisporus. Genes therefore exist in two different forms, called karyolleles, once in each homokaryon. The two homokaryons of A. bisporus are called P1 and P2. We examine for the first time, the spatiotemporal karyollele specific expression of genes.Methods: Using gene predictions for the genome sequences of both the P1 and P2 homokaryons, we identify karyollele pairs. Unique markers that distinguish them are discovered and quantified in RNA-seq data from different tissues throughout the development of the mushroom.Results: We find that the P1 and P2 nuclei are differentially active in different tissues throughout development. Furthermore, we find that chromosomes in the different nuclei are also differentially active. However, the regulation occurs at the gene level. This is indicated by neighbouring karyolleles on the same chromosome which are upregulated in different nuclei.We find 520 differentially expressed genes throughout development. These genes represent a large variety of functionality, including metabolism and regulatory elements. That the P1 homokaryon is active in specific tissues of the mushroom reveals a complex regulation of development between nuclei. Improving the phenotype of the mushroom may therefore rely upon the selection of traits or even chromosomes that may be active primarily in one homokaryon.	Genomes poster	Fundamental
P_Go079	517	Lionel Morgado and Frank Johannes	Lionel Morgado	Learning sequence patterns of AGO-sRNA affinity from high-throughput sequencing libraries to improve functional sRNA categorization in plants	Loading small RNAs (sRNAs) into Argonaute complexes is a crucial stage in all pathways identified so far in plants that depend on these non-coding sequences. After this step, important mechanisms such as transcriptional and post-transcriptional silencing (PTS) can be activated depending on the specific AGO protein to which a sRNA binds. The use of high-throughput sequencing platforms became common practice nowadays, and has been allowing to capture a huge number of short length sequences which lack functional characterization. Most tools for sRNA function prediction are dedicated to PTS and are characterized by a very high false positive rate. Information concerning AGO-sRNA affinity can contribute to define sets with a higher chance to be biologically active. However, the only way to get an indication on AGO association is via expensive and laborious experimental procedures since no computational tool exists to infer such property. It is known that the key for AGO loading is embedded in the sRNA primary structure, but the patterns that drive this combination haven't been fully explored to date. A Support Vector Machine based approach was employed to identify these marks in large libraries of sRNAs obtained via high-throughput sequencing after immunoprecipitation of AGO proteins from Arabidopsis thaliana. The models trained were afterwards incorporated in a pipeline for biologically functional sRNA detection and categorization based on AGO-sRNA affinity. Further tests show that the inference system can be applied to plants in general owing to the fact that AGOs are well conserved proteins inside the kingdom.	Genomes poster	Biotechnology
P_Go080	600	Kathrin Trappe, Enrico Seiler, Jan R. Forster, Tobias Marschall and Bernhard Renard	Kathrin Trappe	Mapping-Based Horizontal Gene Transfer Detection from Sequencing Data - Enhancing Metagenomic Approaches for Pathogen Identification	Horizontal gene transfer (HGT) is a fundamental mechanism that enables organisms such as bacteria to directly transfer genetic material between distant species.This way, bacteria can acquire new traits such as antibiotic resistance or pathogenic toxins.Current bioinformatics approaches focus on detecting past HGT events by exploring phylogenetic trees or genome composition inconsistencies.These techniques normally require the availability of finished and fully annotated genomes.However, especially in outbreak scenarios where new HGT mediated pathogens emerge, there is need for fast and precise HGT detection. Next-generation sequencing (NGS) technologies facilitate swift analysis of unknown pathogens but, to the best of our knowledge, so far no approach detects HGTs directly from NGS reads.We propose the tools Daisy and Donald, novel mapping-based pipelines for HGT detection from NGS data. Donald leverages metagenomic profiling tools to identify candidate references for the acceptor genome reference (the parent genome of the HGT organism acquiring the HGT sequence) and the donor genome reference (the parent donating the HGT sequence). Subsequently, Daisy determines specific HGT regions relying on established methods from structural variant detection approved for human NGS data.Preliminary results of simulated and real data show that Donald successfully identifies acceptor and donor candidates as such and is able to distinguish non-HGT samples as true negatives. Daisy detects HGT regions with base pair resolution, and outperforms alternative approaches using a genome assembly of the reads. We see our approach as a powerful complement for comprehensive analysis of bacterial genomes in the context of NGS data.	Genomes poster	Health
P_Go081	398	Maxime Hebrard and Todd D. Taylor	Maxime Hebrard	MetaTreeMap: A New Visualization of Metagenomic Phylogenic Trees	Metagenomic samples can contain hundreds or thousands of different species. The most common method to identify these species is to sequence the samples and then classify the reads to nodes along a phylogenic tree. Linear representations of trees with so many nodes face legibility issues. In addition, such views are not optimal for appreciating the read quantity assigned to each node. The problem is exaggerated when comparison between multiple samples is needed. MetaTreeMap adapts a visualization method that addresses these weaknesses.A treemap represents a hierarchy as nested rectangles. Each element of the hierarchy (node) is converted to a rectangle. Each sub-node is then a sub-rectangle. In addition, the area of each rectangle is proportional to the associated quantity (assigned read number). The final result is a tile-like figure where the larger tiles represent the more abundant species in the dataset.Our tool uses treemaps to enhance the display of phylogenic trees and allows researchers to easily browse through depth levels by rank selections, by color changes, by zoom events and search functions. We also display a synchronized spreadsheet (same color and zoom functions). Furthermore, multiples samples can be loaded and visualized at the same time allowing visual and numerical comparison.The goal of this software is to provide the user with the ability to easily display phylogenic trees based on various quantities assigned to the nodes, such as read number, read percentage or other values. The tool can be used online at http://metasystems.riken.jp/visualization/treemap/.	Genomes poster	Ecosystems
P_Go082	688	Francis Blokzijl, Joep de Ligt, Myrthe Jager, Valentina Sasselli, Sophie Roerink, Hans Clevers, Ruben van Boxtel and Edwin Cuppen	Francis Blokzijl	Mutational signatures in normal adult stem cells of different human tissues	Recently, large-scale analyses of tumour mutation data across different cancer types have revealed 30 mutational signatures, which are thought to reflect mutational processes in transformed cells. To understand the extreme variation in age-related cancer risk across tissues, it is essential to determine the activity of mutational processes in normal cells prior to malignant transformation. Here, we determined the mutational load of normal adult stem cells (ASCs) of the small intestine, colon and liver of human donors with ages ranging from 3 to 87 years. To this end, we exploited the organoid culturing system to select and clonally expand ASCs. We performed whole genome-sequencing to determine the mutational loads and subsequently identified mutational signatures using non-negative matrix factorization (NMF). While the tissues studied here exhibit very distinct division rates and cancer incidence, they show comparable annual mutation rates (~40 novel mutations per year). ASCs of the colon and small intestine show a high contribution of a mutational signature that indicates activity of spontaneous deamination of 5-methylcytosines, likely reflecting the high division rate of these stem cells. Liver ASCs show high activity of a mutational signature with unknown etiology. Importantly, mutation spectra of driver genes in colorectal and liver cancer show high similarity to the tissue-specific ASC mutational spectra, suggesting that intrinsic mutational processes in ASCs can initiate tumorigenesis. In addition, we observed increased chromosomal instability in colon ASCs that is characteristic of segregation errors, which could underlie the difference in cancer incidence between colon and small intestine.	Genomes poster	Health

P_Go083	765	Nadezda Volkova, Bettina Meier, Victor Gonzalez Huici, Simone Bertonlini, Peter Campbell, Anton Gartner and Moritz Gerstung	Nadezda Volkova	Mutational signatures of DNA repair deficiencies and cytotoxin exposures in C. elegans	Cancer is caused by alterations in the genome. These alterations can be an effect of combination of environmental factors damaging DNA and deficiencies in DNA repair and replication leading to characteristic mutational spectra.Mutational signatures (Alexandrov et al. 2013) became a very useful tool of cancer investigation in the last years. However, the signatures identified so far mostly represent complex conglomerates of the action of different mutational processes. For many signatures, the link with the underlying mutational processes is still unclear.In this study we used C. elegans as a model organism to present a systematic screen with 9 types of genotoxins under 58 different genetic conditions including single and double knock-outs of DNA repair associated genes. Upon exposure over several generations we used whole-genome sequencing to study patterns of DNA damage.We studied the mutational spectra by analysing different types of genetic lesions including point mutations, indels and structural variants using rigorous quality control procedure. This approach allows us to dissect the precise individual contributions of each factor using zero-inflated negative binomial additive models, and also identify significant genetic and gene-mutagen interactions such as 3-fold increase in mutational burden for pms-2/pole-4 double knock-out and mutational spectra expansion for DMS exposure in rev-1 mutants. In summary, this analysis presents the first systematic catalogue of mutational signatures caused by genotoxins and DNA repair deficiencies.	Genomes poster	Fundamental
P_Go084	375	Natalia Szóstak, Agnieszka Rybarczyk, Maciej Antczak,Tomasz Zok, Mariusz Popenda,Ryszard Adamiak, Jacek Blazewicz andMarta Szachniuk	Natalia Szóstak	New in silico approach to assessing RNA secondary structures with non-canonical base pairs	The remarkable RNA molecules properties and diversity allow them to play important roles in the cellular processes. They can act not only as carriers of genetic information but also participate in the regulation of gene expressions and serve as catalysts in many biological pathways. The function of RNA is strongly dependent on its structure, therefore an appropriate recognition of this structure, on every level of organization, is crucial. One particular concern is the assessment of base-base interactions, described as the secondary structure. It greatly facilitates an interpretation of RNA function and allows for structure analysis on the tertiary level. Computational approaches consider mostly Watson-Crick and wobble base pairs. Handling of non-canonical interactions, important for a full description of RNA structure, is still a challenge.Here we present a novel two-step in silico approach to asses RNA secondary structures with non-canonical base pairs. The knowledge of extended secondary structure can accelerate an advancement of the 3D RNA module concept and improve the module identification and search within available structures. It can also be useful in supporting new solutions to RNA motif discovery problems. Its first application to our on-going analysis of the mechanism of spontaneous degradation of RNA molecules showed improvement in accuracy of prediction of RNA degradants.We believe that our work concerning the recognition of non-canonical interactions in RNA structures will be influential not only for the scientific community but also for clinical and pharmaceutical industry that take into consideration the RNA molecules.	Genomes poster	Fundamental
P_Go086	377	Franziska Singer, Nora Toussaint, Michael Prummer, Falco Kilchmann, Miquel Busquets Lopez, Christian Stirnimann and Daniel Stekhoven	Nora Toussaint	NEXUS: supporting precision medicine with state-of-the art technologies for molecular diagnostics	High-throughput genomics and screening technologies have changed the way biomedical research is performed. The transition from directed testing of a few specific targets, selected based on prior knowledge, to analyzing comprehensive high-throughput data offers tremendous possibilities but also introduces new challenges. Despite the great potential, particularly for the treatment of patients with rare diseases, with tumors lacking known targetable mutations, and of those considered end-of-treatment line, the use of high-throughput techniques to go beyond standard diagnostics is not fully established in the clinics yet. Establishing high-throughput molecular diagnostics for clinical use requires specific protocols accounting for stringent quality control, privacy issues, and thorough process documentation.To this end, NEXUS, a core facility at ETH Zurich, provides state-of-the-art bioinformatics, statistical analyses, and screening of FDA-approved drugs combined with high standards for quality control, data privacy, and reproducibility. We are developing a workflow for the molecular profiling of matched tumor and normal samples from sequencing to clinical decision support. In addition to the identification of somatic variants, our workflow links the detected alterations to possible treatment options, both cancer type-specific and off-label. The analysis results are summarized in a concise and clearly structured clinical report designed to form the basis for discussions in a clinical molecular tumor board.Here, we showcase the designed workflow on samples from the UniversityHospital Zurich. In collaboration with hospital oncologists, researchers at ETH Zurich, and the Genomics Facility Basel, potential targets for off-label therapies could be proposed based on whole-exome sequencing of patient biopsies.	Genomes poster	Health
P_Go087	342	Sneha Mitra and Leelavati Narlikar	Leelavati Narlikar	No Promoter Left Behind: New method that reveals novel promoter architectures from genome-wide transcription start sites	An important question in biology is how different promoter-architectures contribute to diversity in transcriptional regulation. A major step forward has been the development of technologies like CAGE that map transcription start sites at high resolution, genome-wide. However, the subsequent step of characterizing promoters is still done on the basis of established features like the TATA-box or GC-richness. Unfortunately, many promoters cannot be explained by these few elements; de novo motif discovery also fails due to the diverse nature of promoters. E.g. one set of promoters may be characterized by elements A-B-C, another by D-A, a third only by D, and a fourth by E-F. In this scenario, there is little chance that all promoter-architectures will be detected by conventional approaches.We present a new unsupervised machine-learning method---No Promoter Left Behind (NPLB)---that partitions promoters into diverse architectures while simultaneously identifying relevant elements. NPLB identifies novel architectures within various bacteria, fly, and human data-sets, while giving insights into promoter-function. We further show that these architectures have distinct evolutionary signatures, missed by traditional analyses.We believe this work will have an impact comparable to when de novo motif discovery was first developed to identify regulatory elements, because its applicability extends beyond promoter-architectures. The new unbiased way of looking at high-throughput sequence data allows for the identification of regulatory signals associated with any DNA-specified biological event reported at high-resolution. NPLB opens up avenues to learn new biology from high-throughput data, rather than simply validating, albeit at a large scale, what is already known.	Genomes poster	Fundamental
P_Go088	850	Ricard Illa, Diana Buitrago, Laia Codó, Romina Royo, Adam Hospital, Isabelle Heath, Josep Lluis Gelpí and Modesto Orozco	Ricard Illa	Nucleosome Dynamics portal	Nucleosome positioning plays an important role in transcriptional regulation and other DNA-related processes. Here we present NucleosomeDynamics, a new online tool that uses data from MNase-seq experiments as input and allows analysis and visualization of the nucleosome positioning.It uses the R statistical environment on its back end to perform the calculations. Specifically, it uses two libraries, nucleR and NucleosomeDynamics, that were specifically developed for such studies. nucleR allows to efficiently and accurately define nucleosome's location. NucleosomeDynamics, the R library, compares different MNase-seq experiments at a read level and identifies variations in nucleosome occupancy.Additionally, the web portal can compute other nucleosome-related features, like the location of nucleosome-free regions, a classification of transcription start sites based on the properties of the nucleosomes surrounding them, a theoretical prediction of nucleosomes' phasing at a gene level, and and estimation of a stiffness value for each nucleosome.The calculations are accessible in a web portal. The interface allows the user to upload data on the server, select which properties to compute and store the results in a private user workspace. Results can be downloaded as GFF files, BIGWIG files or visualized. For the visualization, we use jbrowse, as fast and embeddable genome browser built completely with JavaScript and HTML5. JBrowse incorporates relevant genome annotations, data from several recent publications in the field and can also incorporate annotation tracks uploaded by the user.The NucleosomeDynamics portal provides a single access point to a complete series of nucleosme occupancy-oriented tools and contributes to a multiscale view of chromatin structure.	Genomes poster	Biotechnology
P_Go089	627	Boris Nagaev, Alexandra Simonova and Andrei Alexeevski	Andrei Alexeevski	Nucleotide pangenome of Brucella highlights evolutionary events	We studied evolution of 55 Brucella genomes that were assembled into two chromosomes. For this purpose nucleotide pangenome (NPG) was constructed by NPG-explorer program (http://mouse.belozersky.msu.ru/tools/npge.html). Brucella NPG consists of 1358 major blocks, which are alignments of long (>100 bp) orthologous fragments with more than 90% identical positions, and 91 unique fragments matching no homologs in other input genomes.Phylogeny of strains was reconstructed by NPG-explorer from 'nucleotide core', i.e. joined alignment of Brucella stable blocks. Stable blocks are major blocks composed of one fragment from each genome such that no duplications of these fragments appear in any genome. Nucleotide core covers 61.2% input nucleotides, it has 96.7% identical positions.Long deletions and insertions were identified using hemi-stable blocks composed of one fragment from each genome of a subset (other genomes lack homologous fragments). Such blocks cover 13.0% input nucleotides. Evolutionary events that gave rise to these blocks were reconstructed by comparison with the phylogenetic tree of strains.Recent duplications and transposable elements were detected using blocks with repeats, which cover 25.4% input nucleotides. Putative events of horizontal transfer from remote taxa were confirmed for certain unique fragments by BLAST search.NPG-explorer identified 76 syntenic regions defined as joins of collinear stable and hemi-stable blocks and/or blocks with repeats.For closely related strains nucleotide pangenomes seem to be preferable to gene based pangenomes. For instance, NPG represents orthologous intergenic sequences and doesn't depend on gene missannotations.The work was supported by grants RSF 16-14-10319 , RFBR 14-04-01693.	Genomes poster	Fundamental
P_Go090	799	Giles Miclotte	Giles Miclotte	OMSim: simulating optical mapping data	Motivation: Optical mapping technologies (Bionano) provide a long range view of the genome, that can not be achieved through more traditional sequencing methods (e.g. Illumina, PacBio, ONT). Generating synthetic data is essential for the development and benchmarking of new tools for data analysis. However, there is no simulation software available for the optical mapping data.Results: We have developed an optical mapping data simulator, OMSim, which simulates Bionano data, based on distributions derived from existing data sources. The simulated data has been extensively tested for compatibility with the Irys software system.Availability: The Python backend and a cross platform graphical user interface are available on the web under the GNU GPL V2 license.	Genomes poster	Fundamental
P_Go091	427	Ramon Diaz-Uriarte	Ramon Diaz-Uriarte	OncoSimulR: genetic simulation of cancer progression with arbitrary epistasis and mutator genes	Forward genetic simulations are widely used in population genetics andcancer research to verify analytic results, to generate data to assess theperformance of statistical methods, and to examine complex models that aremathematically intractable. Several programs and libraries are available,but none of these are well suited if we require:- complete flexibility to specify arbitrary epistatic effects of highorder as well as order effects (fitness of genotype AB depends on whetherA or B is acquired first);- sampling from the population at arbitrary times and with differentresolution (e.g., whole tumor vs. single-cell);- tracking of the complete history of all clones;- large (> 10000) number of genes;- gene-specific mutation rates;- mutator/antimutator genes (genes whose mutation leads to anincrease/decrease in the mutation rate of other genes);- large (> 10^6) asexual populations;- varied models of growth.These scenarios are common in evolutionary genetics studies and cancerprogression models. I have developed OncoSimulR, an R package (that usesC++ underneath for speed), to provide those features. OncoSimulR alsoallows specifying fitness using directed acyclic graphs to definerestrictions in the order of accumulation of mutations.These features make OncoSimulR a unique forward genetic simulation tool,particularly well suited for examining cancer evolution models. OncoSimulRis available from BioConductor(http://bioconductor.org/packages/release/bioc/html/OncoSimulR.html) andgithub (https://github.com/rdiaz02/OncoSimul).	Genomes poster	Fundamental
P_Go092	360	Sjoerd van Hagen, Pieter Lukasse, Sander de Ridder, Fedde Schaeffer, Priti Kumari, James Lindsay, Jianjiong Gao, Benjamin Gross, Zachary Heins, Adam Abeshouse, Hongxin Zhang, Yichao Sun, Robert Sheridan, Onur Sumer, Stuart Watt, Chris Sander, Nikolaus Schultz, Ethan Cerami and	Jochem Bijlard	Open Source Development Success through collaboration: Contributions to cBioPortal	Approximately one year ago the popular cBioPortal for Cancer Genomics was made open source. In this last year its development community has grown and the platform has been extended with many new features. Here we detail some of the contributions The Hyve (Utrecht) has made to the platform, in collaboration with Dana Farber Cancer Institute (Boston), Memorial Sloan Kettering Cancer Center (New York) and Boehringer Ingelheim (BI RCV). The contributions can roughly be divided into three categories: (1) improvement of the data loading pipeline, (2) new data analysis features, and (3) optimizations of the front end.In the data loading pipeline we have introduced a strict separation between the validation step and the loading step. This "separation of concerns" design principle makes the code easier to understand and maintain and simplifies the process of adding new datasets to a local cBioPortal installation. Special effort was spent on making the validator easy to use, which is exemplified by clearer error messages and the generation of a HTML validation report.In the front end we added a whole new pan-cancer view for studies comprising multiple cancer types, added new query options in the Study overview page and added new visualizations to the query results page to support better enrichment analysis of expression (mRNA, Proteins) and co-occurrence (copy number, mutations). We have also implemented integration documentation from the Wiki or Git, and made the portal more customizable (logo, headers, news and FAQ), which is very important for open source software.Last but not least, we have optimized the loading times of the portal to be able to host larger studies, focusing on the most used pages in the application. In the query results page we have successfully shortened the loading times of various analyses.	Genomes poster	Biotechnology
P_Go093	541	Matthias Scholz, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow and Nicola Segata	Matthias Scholz	Pangenome-based computational metagenomic profiling enables strain-level culture-free epidemiology and population genomics studies.	Microbial species comprise strains with largely different set of genes and functional potential. Identifying microbial strains and characterizing their genes is thus essential for pathogen discovery, epidemiology and population genomics. Here we present a novel computational strain-level metagenomic profiling tool, called PanPhlAn [1], for identifying the gene composition and in-vivo transcriptional activity of individual strains from metagenomic and metatranscriptomic samples. PanPhlAn enables both the identification of known organisms and the characterization of previously unseen strains.Applied to the 2011 German E. coli outbreak, we demonstrate the ability of PanPhlAn to recognize outbreak strains and identify their associated virulence and resistance factors. Based on almost two thousand samples, PanPhlAn produced the largest strain-level, culture-free population genomic study of human-associated microbial species. In a large cohort of preterm infants, PanPhlAn enabled the identification of disease-associated strain-level genetic biomarkers [2]. PanPhlAn is available at http://segatalab.cibio.unitn.it/tools/panphlan.References1. Matthias Scholz, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow, and Nicola Segata. Strain-level microbial epidemiology and population genomics from shotgun metagenomics.Nature Methods, 13, 435–438, 2016.2. Doyle V. Ward, Matthias Scholz, Moreno Zolfo, Diana H. Taft, Kurt R. Schibler, Adrian Tett, Nicola Segata, Ardythe L. Morrow. Metagenomic sequencing with strain level resolution implicates uropathogenic E. coli in necrotizing enterocolitis and death in preterm infants.Cell Reports, 14, 2912–2924, 2016.	Genomes poster	Health
P_Go094	524	Cornelia Meckbach, Rebecca Tacke, Stephan Waack, Edgar Wingender and Mehmet Gültas	Cornelia Meckbach	PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information	Transcription factors (TFs) are important regulatory proteins that govern transcriptional regulation. Today, it is known that in higher organisms different TFs have to cooperate rather than acting individually in order to control complex genetic programs. The identification of these interactions is an important challenge for understanding the molecular mechanisms of regulating biological processes.In this study, we present a new method, called Potentially Collaborating Transcription Factor Finder (PC-TraFF) that is based on pointwise mutual information (PMI). PMI is a very useful association measure in the field of linguistics for the detection of word combinations. We adopted the PMI in the field of bioinformatics by considering the genome as a document, the sequences as sentences, and TF binding sites (TFBSs) as words to identify interacting TFs in a set of sequences. Unlike previous methods, PC-TraFF does not require any background set of sequences since it estimates for each TFBS pair the expected levels of background PMI arising from noise of false positive TFBSs using the average product correction. Finally, the signal caused by functional collaborating TFs is separated from the background, enabling the detection of collaborating TFs without the influence of noise.The results of our study show, that PC-TraFF is on the one hand able to identify known collaborating pairs in the sequences under study and on the other hand able to predict novel collaborating TFs thus providing new targets for future experimental validation.	Genomes poster	Fundamental
P_Go095	696	Sebastian Theobald, Tammi Vesth, Jane L. Nybo, Inge Kjærbølling, Jens C. Frisvad, Kristian F. Nielsen, Thomas O. Larsen, Igor V. Grigoriev, Asaf Salamov, Uffe H. Mortensen, Scott E. Baker and Mikael R. Andersen	Sebastian Theobald	Phylogenomic analysis of secondary metabolism genes sheds light on their evolution in Aspergilli	The World Health Organization is reporting a rising number of multiple drug resistant pathogens every year, increasing the need for new drug development. However, current methods for natural product discovery rely on time consuming experimental work, making them unable to keep up with this demand.In the aspMine project, we are sequencing and analyzing over 300 species of Aspergilli, agroup of filamentous fungi rich in natural compounds. The vast amount of data obtained from these species challenges the way we were mining for products and requires new pipelines for secondary metabolite analysis.Natural products are encoded by genes located in close proximity, called secondary metabolic gene clusters, which makes them interesting targets for genomic analysis. We use a modified version of the Secondary Metabolite Unique Regions Finder (SMURF) algorithm, combined with InterPro annotations to create approximate maximum likelihood trees of conserved domains from secondary metabolic genes across 56 species, giving insights into the secondary metabolism gene diversity and evolution.In this study we can describe the evolution of non ribosomal peptide synthetases (NRPS), polyketide synthases (PKS) and hybrids thereof, find possible common ancestors and detect horizontal gene transfer events.Finally, we have performed large scale analysis of gene cluster dynamics and evolution, which provides us with better understanding of speciation in Aspergilli. With this new insights into the evolution of natural products, an application in synthetic natural product assembly lies within our grasp.	Genomes poster	Biotechnology
P_Go097	695	Dmitry Penzar, Mikhail Krivozubov and Sergey Spirin	Sergey Spirin	PQ, a new character-based program for phylogenetic reconstruction	We present a new program called PQ for phylogenetic reconstruction of proteins.PQ uses an original character-based algorithm for scoring a phylogenetic tree. Web interface to the program is available at http://mouse.belozersky.msu.ru/tools/pq/ .The program was tested on thousands of alignments of orthologous proteins from Metazoa, Fungi and Proteobacteria. We compared the ability of PQ and a number of other programs to reconstruct phylogenetic trees that are close to known trees of corresponding organisms.For all tested phylogenetic groups and for all sizes of alignments between 10 and 45 sequences PQ outperforms maximum likelihood program RAxML [1] and maximum parsimony program TNT [2]. Working on small alignments (10-15 sequences) it outperforms distance-based program FastME [3], too. However on 45-sequence alignments of fungal proteins FastME outperforms PQ.The new program can be a good alternative to known programs, especially for analyzing small sets of protein sequences.References:1. A. Stamatakis. Bioinformatics 30(9), 2014.2. P. Goloboff, J. Farris, and K. Nixon, 2003; http://www.lillo.org.ar/phylogeny/tnt/ .3. V. Lefort, R. Desper, and O. Gascuel. Molecular Biology and Evolution 32(10), 2015.	Genomes poster	Fundamental

P_Go098	699	Loukas Mouzos and Thomas Abeel	Loukas Mouzos	Practical approaches for constructing bacterial population reference graphs	Introduction: Cheap sequencing has resulted in hundreds or thousands of individual genomes available for many species. Comparative genomics approaches founded on reference based variant calling likely limit our analytical power due to idiosyncrasies of that reference. An alternative to the single-reference paradigm are population references graphs which seek to encode multiple references in a single representation.We sought to represent hundreds of Mycobacterium tuberculosis (MTB) genomes in a graph-representation, including structural variations and gene annotations.Results: To construct our MTB population reference graph, we used a set of 27 finished and 300 draft assemblies. We then created a set of disconnected graphs corresponding to syntenic regions across the genomes that were multi-aligned using REVEAL. Each graph indicates the variability among the strains in terms of SNPs and indels. The relations among the disconnected graphs indicate the structural variations, particularly inversions. MTB has a relatively simple genome architecture and the vast majority of the global diversity can be represented by less than 10 disconnected graphs that encode several possible inversions.Furthermore, we mapped gene annotations from one well-annotated strain to all others and found good concordance with pre-existing annotations on those other strains.Discussion: Our MTB population reference graph aims to represent all variability of the Mycobacterium tuberculosis species. Contrary to the single reference genome, this data structure reflects the variability of the species in terms of sequence (SNPs and indels) and structural (inversions, deletions) variation, improving the ability to genotype newly sequenced strains.	Genomes poster	Fundamental
P_Go099	343	Daniel Buxton, Nadia Chuzhanova and Jonathan Crofts	Daniel Buxton	Predicting genomic regions linked to schizophrenia using the 3D architecture of the human genome	Schizophrenia is a severe mental disorder with heritability as high as 80% and an incidence of 1% globally. Genome-wide association studies have identified single nucleotide polymorphisms (SNPs) in 347 genes which associate with this disease, but the role of many of these SNPs in the development of schizophrenia is yet to be understood.We hypothesised that there is a network of interacting regions harbouring schizophrenia-associated SNPs which may contain genes, promoters and enhancers. We utilised datasets generated by two chromosome conformation capture techniques, Capture Hi-C and in situ Hi-C, which measure the 3D architecture of the human genome within the cell nucleus. These techniques count interaction frequencies (IFs) between inter- and intra-chromosomal fragments of fixed size, ranging from 1 kb to 1 Mb chromosome regions.Capture Hi-C data was used to locate promoters and enhancers which regulate each gene via looping interactions, and we amalgamated these locations to form extended gene regions (EGRs). These EGRs act as nodes in our 3D interaction network, where nodes are connected by an edge if there is a sufficiently high IF between regions from in situ Hi-C data. Our algorithm found several gene-rich regions which have a high connectivity to the EGRs in our network, with many of these regions containing genes which have previously been found to associate with schizophrenia. We also discovered new gene-containing regions which are enriched in SNPs and have not previously been implicated in schizophrenia.	Genomes poster	Health
P_Go100	562	Andrea Gazzo, Daniele Raimondi, Dorien Daneels, Guillaume Smits, Sonia Van Dooren and Tom Lenaerts	Andrea Gazzo	Predicting oligogenic effects using digenic disease data	Recently DIDA, a unique digenic diseases database fully specifying genes, variants and their properties, was developed (1). Each instance in DIDA, called 'digenic combination', is a combination of two or more variants mapped on two different genes that induce a specific disease. The manner in which the combination generates the clinical outcome differs between instances. We have separated them into two digenic effect classes, "on/off" and "severity". In the former, mutations in both genes are required for the development of the disease. In the latter variants in a single gene are enough to develop the disease while the second increases the severity of the symptoms or affects the age onset. As such the severity class captures monogenic diseases with modified by variants in other genes. We show, using a random forest model, that the genetic and biologic properties related variants and genes in those combinations are sufficient to differentiate between these two classes. The model reaches an accuracy of almost 80%, using a stratified cross-validation. A novel feature relevance analysis that infers decision signatures from the model provides insight into why instances appertain to a specific class. New instances are predicted with an accuracy of 61%. In all, our results show for the first time how to differentiate between true digenic cases and modifiers, which are probably abundant given the heterogeneous nature of all known diseases.(1) Gazzo et al (2016). DIDA: A curated and annotated digenic diseases database. Nucleic Acids Res	Genomes poster	Fundamental Health
P_Go102	502	Malgorzata A Komor, Annemieke C Hiemstra, Thang V Pham, Sander R Piersma, Anne S Bolijn, Pien M Delis-Van Diemen, Marianne Tijssen, Robert P Sebra, Bo Han, Maradith Ashby, Beatriz Carvalho, Gerrit A Meijer, Connie R Jimenez and Remond Ja Fijneman	Malgorzata A Komor	Proteogenomic pipeline for identification of novel biomarkers for colorectal cancer	IntroductionEarly detection of colorectal cancer (CRC) and its precursor lesions (adenomas) is crucial to reduce mortality rates. The fecal immunochemical test (FIT) is a CRC screening test detecting blood-derived protein hemoglobin. However, FIT sensitivity is suboptimal. As adenoma-to-carcinoma progression is accompanied by alternative splicing, we aim to identify proteins derived from alternatively spliced RNA which might serve as candidate biomarkers for CRC detection.Materials and methodsRNA and proteins were isolated from CRC cell line SW480 before and after siRNA-mediated down-modulation of splicing machinery: SF3B1 and SRSF1. To identify splice variants, mRNA was sequenced (Illumina) and analyzed. RNA-seq analysis included quality checks, reads mapping, differential gene expression and differential splicing analysis. In silico results were validated by qRT-PCR. Proteins were analyzed by LC-MS/MS (QExactive). A proteogenomic pipeline was established to enrich the protein sequence database with mRNA-derived splice variants and identify protein isoforms. To further extend the splice-variant database, PacBio Iso-Seq was performed for the siSF3B1- and control-samples.ResultsExpression analysis on RNA and protein level proved that knock-down experiments were performed successfully. RNA-seq analysis revealed hundreds of splice variants, including events described in literature. Proteomics experiments yielded over 6000 proteins per sample, including protein isoforms resulting from alternative splicing.ConclusionsThe proteogenomic pipeline for alternative splicing was established and experimental proof of concept was obtained.In future studies this pipeline will be applied in clinically relevant setting on series of low and high progression-risk adenomas and CRCs. Novel candidates will be evaluated for performance as screening markers.	Genomes poster	Health
P_Go103	872	Marc Hulsman, Marcel J.T. Reinders and Henne Holstege	Marc Hulsman	Removing study-effects present in multi-center exome studies through a probabilistic burden statistic	To elucidate the genetic underpinnings of a complex trait, large sample sizes are required. This is especially true when searching for rare variants. Due to this, more and more exome studies are combining their power by sharing data. However, the use of different sequencing depths and capture kits, combined with non-balanced studies (only cases, only controls) impedes this process, and can easily result in large numbers of false positive results, evident through p-value inflation. Such inflation can be prevented by only considering variants in follow-up analysis that do not have significant differences in their missingness rate across studies. Unfortunately, dependent on the to be combined studies, this will significantly affect the number of available variants, and thereby the statistical power of the combined analysis.Here, we propose a method which solves this problem through the use of probabilistic genotypes calls, which are constructed such that they carry information on the (un)certainty of a call as well as the underlying population frequency. We design a burden test which directly uses this probabilistic information through genotype sampling. Also, we propose a test which filters variants that deviate in frequency across studies, more significantly than what one would expect given common population patterns. Together, we show that this approach significantly reduces p-value inflation, allowing variants with up to 75% missingness to be considered in the burden test.	Genomes poster	Fundamental
P_Go104	486	Elvis Ndah, Veronique Jonckheere, Gerben Menschaert and Petra Van Damme	Elvis Ndah	REPARATION: Ribosome Profiling Assisted (Re-) Annotation of Bacterial genomes	The delineation of genes in bacteria has remained an important challenge because prokaryotic genomes are often tightly packed frequently resulting in overlapping genes. Since deep sequencing of ribosome protected mRNA fragments (Ribo-seq) provides a means to map the positions of translating ribosomes over the entire genome, we here present a de novo approach (REPARATION) - that integrates Ribo-seq data next to biological genome features to delineate the translated open reading frames (ORFs) in bacteria independent of (available) genome annotation.More specifically, our algorithm traverses the entire genome to generate all possible ORFs. Based on a growth curve model to estimate minimum ORF read density and Ribo-seq base pairs coverage thresholds indicative of translation, it then applies a robust random forest model to build classifiers for ORF discrimination. To evaluate the performance of our algorithm we applied it to Salmonella enterica serovar Typhimurium (strain SL1344) using in house Ribo-seq and matching (N-terminal) proteomics data. A database search of the proteomics data against the six frame translation database of the SL1344 genome resulted in the identification of 749 unique N-termini with Ribo-seq evidence. REPARATION was able to pick up translation evidence of 82% of all annotated ORFs that passed the threshold values. Interestingly and despite its high annotation level, we obtained translation evidence of 340 (11%) possible N-terminal extensions (23 matching N-terminal peptides were identified), 240 (7%) truncations (2 N-terminal peptides) and 220 novel ORFs (4 N-terminal peptides). Overall, 92% of all identified N-termini matched the by REPARATION delineated ORFs.	Genomes poster	Biotechnology Fundamental Health
P_Go105	320	Igor Sidorov, Andrey Leontovich, Dmitry Samborskiy and Alexander Gorbalenya	Igor Sidorov	Retrieval of genome-based information from sequence databases using hybrid homology-annotation searches: case of complete RNA virus genomes	Retrieval of biological information is commonly accomplished by scanning databases with query for either annotation matches or significant similarity to target sequences. Accuracy of annotation varies in databases and may compromise both sensitivity and selectivity of annotation-based searches. Similarity-based searches are free of this limitations due to high accuracy of genome sequencing but establishing biologically meaningful similarity thresholds remains unmet challenge for them. Here we describe an approach with improved sensitivity and selectivity of retrieval which combines analyses of annotation and sequence similarities and automatically establishes data-driven similarity thresholds. This approach uses isotonic regression for simultaneous analysis of similarity and annotation matches. It was realized in a computational engine (dubbed HAYGENS, Homology-Annotation hYbrid retrieval of complete RNA virus GENome Sequences). HAYGENS was applied to 13 RNA viral groups of different taxonomic ranks that include many poorly characterized viruses. Sequence alignment profiles of family-specific RNA-dependent RNA polymerase and taxa names were used to query GenBank. Additionally, to retain only complete or nearly complete genomic sequences, the results of hybrid sequence-annotation searches were filtered using original procedure. Comparing to annotation-based searches, HAYGENS' gained sensitivity and selectivity that exceeded 5% for >25000 genomes, with uneven distribution of gains in GenBank divisions. HAYGENS daily updates are available at http://veb.lumc.nl/HAYGENS. With the observed and likely gains in accuracy, HAYGENS could be used for quality assessment of sequence annotations. It may be also useful for transferring annotation in annotation-based databases (GO) and for calculating data-driven family-specific thresholds in sequence profile databases (Pfam).	Genomes poster	Health
P_Go106	676	Paul Kirk, Maxime Huvet, Anat Melamed, Goedele Maertens and Charles Bangham	Paul Kirk	Retroviruses integrate into a shared, non-palindromic DNA motif	Palindromic consensus nucleotide sequences are found at the genomic integration sites of retroviruses and other transposable elements. It has been suggested that the palindromic consensus arises as a consequence of structural symmetry in the integrase complex, but the precise mechanism has yet to be elucidated. Here we perform a statistical analysis of large datasets of HTLV-1 and HIV-1 integration sites. The results show that the palindromic consensus sequence is not present in individual integration sites, but appears to arise in the population average as a consequence of the existence of a non-palindromic nucleotide motif that occurs in approximately equal proportions on the plus-strand and the minus-strand of the host genome. We develop a generally applicable algorithm to sort the individual integration site sequences into plus-strand and minus-strand subpopulations. We apply this algorithm to identify the integration site nucleotide motifs of five retroviruses of different genera: HTLV-1, HIV-1, MLV, ASLV, and PFV. The results reveal a non-palindromic motif that is shared between these retroviruses.	Genomes poster	Fundamental
P_Go107	544	Tsukasa Fukunaga and Michiaki Hamada	Tsukasa Fukunaga	RIblast: An ultrafast RNA-RNA interaction prediction method based on seed-and-extension approach	Long non-coding RNAs play important roles in various biological process such as development and epigenetic regulation. Currently, although more than 25,000 lncRNAs are annotated in Gencode database, most of these lncRNAs are still poorly characterized. To understand the functions of lncRNAs, computational detection of the interaction target RNA for each lncRNA is an essential step. However, existing RNA-RNA interaction prediction tools cannot be applied to the whole human lncRNA dataset because of the high computational costs. Therefore, much faster RNA-RNA interaction prediction software would be needed. Here, we developed an ultrafast RNA-RNA interaction method based on seed-and-extension approach, which is widely used in sequence homology detection tools, and have implemented this algorithm as RIblast software. RIblast discovers seed regions using suffix arrays of queries and a database, and extends both ends of seed regions based on full nearest-neighbor energy model and region accessibility information. To evaluate RIblast performances, we compared prediction accuracy and computational speed of RIblast with those of RNAplex, which is one of the best performing RNA-RNA interaction prediction tool at present. As a result, while RIblast showed a similar prediction accuracy to RNAplex on 109 known batectial sRNA-mRNA interactions, RIblast achieved several ten times acceleration in comparison with RNAplex on a part of human lncRNA dataset.	Genomes poster	Fundamental
P_Go109	774	Emiel Ver Loren van Themaat	Emiel Ver Loren van Themaat	Scalable genome-wide characterization of lactic acid bacteria	With the advance of sequencing and computational analysis techniques the ability to genetically characterize bacterial strains has been extended from single strains to dozens and now to hundreds of strains. Here we present the in silico analysis of hundreds of genome sequences of lactic acid bacteria (LAB) from the DSM collection, mostly Streptococcus thermophilus and Lactococcus lactis species used to make yoghurts and cheese. We have analyzed multiple aspects of these genomes, including (sub)species identification using 16S based taxonomies, core SNP based phylogenomics, plasmid content, undesired genes and their core and pan orthologous gene groups.The genomes were sequenced at high quality using Illumina technology. To create high resolution phylogenomic profiles, core SNPs were identified in whole genome comparisons via conserved K-mers, allowing detailed comparisons of highly similar genomes, but with different phenotypes. These genome-wide SNP profiles - as based on conserved regions - were compared to phage profiles and displayed a high but not a 100% exact correlation, indicating that in addition non-conserved regions are important. Plasmid analysis and pangenomics provide further insights into non-core genes possibly contributing to phenotypes of interest. Undesired genes, like antibiotic resistance genes and biogenic amines, were screened using, a.o., CARD and ResFinder.Overall, the genome sequences were successfully generated and analyzed in a high throughput fashion with a dedicated bioinformatics in-house pipeline utilizing custom, commercial and open-source tools. The genome sequences were used to accurately determine taxonomies, genome pairs via core SNPs and undesired gene content. The core and pan genome analysis provides leads towards functional subgroups and further understanding of the DSM lactic acid bacteria strain collection.	Genomes poster	Agro-Food Biotechnology
P_Go110	725	Francois Boyer, Hend Boutouil, Iman Dalloul, Jeanne Moreau, Jean-Claude Aldigier, Michel Cogné and Sophie Péron	Francois Boyer	Search, identification and quantification of CSR junctions in high-throughput sequencing data using CSReport	B cell development is of major importance to ensure an effective humoral adaptive response. At different stages of development, somatic recombination occurs to either generate a diverse repertoire of B-cell receptors (V(D)J recombination) or to adapt immunoglobulin function (class-switch recombination or CSR). CSR is an intra-chromosomal recombination of the immunoglobulin heavy chain (IgH) locus: double-strand breaks are generated in so-called switch regions. Joining and repair of free DNA ends leads to the expression of a different immunoglobulin isotype. As recombination events imprint the cell's genome, sequencing is a key technique to trace them back and high-throughput technologies (HTS) seem very promising to better characterize CSR in large cell populations. Studies of CSR have, however, never been performed using HTS and the classical method is fastidious. To gain more in-depth knowledge of CSR junctions, we used a HTS-based experimental protocol and to achieve optimal benefit from the large generated datasets, we developed CSReport, a new computational tool which automatically identifies and summarizes sequences that support recombination between two switch regions of the IgH locus. It accurately assigns each segment and returns individual junction structures (blunt junction, micro-homologies or insertions) and break points. By realigning each segment, it ensures high-quality structural information as it is crucial in order to shed light on the underlying repair mechanisms. Using BLAST+ and biopython module, the Python code of CSReport runs in about 30 minutes on a laptop computer for a typical 3-million read filtered library.	Genomes poster	Fundamental
P_Go111	425	Enrique Carrillo-De Santa Pau, David Juan, Felipe Were, Vera Pancaldi, Daniel Rico and Alfonso Valencia	Enrique Carrillo-De Santa Pau	Searching for the chromatin determinants of hematopoiesis	As part of the BLUEPRINT Consortium, we are characterizing the epigenomes of blood cells to understand how changes in chromatin are connected with the different lineage differentiation options. In this work, we present our analyses using hematopoietic stem cells (HSCs), monocytes, macrophages, neutrophils, B-cells (naive from venous blood and tonsil-derived germinal center B-cells) and T-cells (CD4 and CD8), combining hematopoietic samples from BLUEPRINT, ENCODE and NIH Epigenomic Roadmap. We have developed a bioinformatics pipeline to generate a 'chromatin space' where the different cell types are clustered by epigenomic similarity.Our analysis is based on Multiple Correspondence Analysis (MCA), the analog of Principal Component Analysis when working with categorical data. We used our previous approach to deal with protein multiple alignments (Rausell, Juan et al PNAS, 2010) with critical enhancements to deal with millions of regions in the same analysis.The analysis of the orthogonal dimension of the space allows us to identify chromatin determinant regions (CDRs), genomic regions with different epigenomic characteristics between the different groups. Functional enrichment analysis of the neighbouring genes suggests that the chromatin state in this regions could be directly linked with the different cell identities. Our analytical approach allows to combine samples from different sources and identify the regions for which chromatin status associates with cell lineage determination or disease conditions.	Genomes poster	Health
P_Go112	500	Samuel Heron, Owen Dando, Giles Hardingham and Ian Simpson	Samuel Heron	Separation of Mixed Source RNA-Seq Reads by Comparative Genomic Processing	Knowledge of the cell-autonomous and non-autonomous mechanisms operating within biological systems is essential to reveal the underlying molecular processes at work and has been particularly prevalent in functional studies of neurological diseases and cancers. These studies commonly measure gene expression levels in samples, but are often confounded by invasive sample processing. Indeed physical separation techniques for cell mixtures have been shown to trigger stress and apoptosis related genes obscuring the identification of genes of interest and introducing bias.We present a novel approach that alleviates this issue for gene expression quantification using RNA-seq by conducting sequence separation between genomes entirely in silico. Our method takes sequences from a mixed source RNA-seq sample run that contains two different genomes (for example two cell types each belonging to a closely related species, or different strains of the same species) and differentially maps them between the genomes. The mappings for each read are then assessed by alignment quality factors and assigned to one source genome or the other according to tunable selection criteria. Separated read sets can then be quantified using standard differential expression methods for RNA-seq data. Using simulated data for rat and mouse we successfully demonstrate that reads from closely related species can be separated in this manner.	Genomes poster	Fundamental

P_Go113	358	Marc Sturm, Christopher Schroeder and Peter Bauer	Marc Sturm	SeqPurge: highly-sensitive adapter trimming for paired-end short read data	Trimming adapter sequences from short read data is a common preprocessing step in most DNA/RNA sequence analysis pipelines. For amplicon-based approaches, which are mostly used in clinical diagnostics, sensitive adapter trimming is of special importance. Untrimmed adapters can be located at the same genomic position and can lead to spurious variant calls. Shotgun approaches are more robust towards adapter contamination because untrimmed adapters are randomly distributed over the target region. This reduces the probability of spurious variant calls.When performing paired-end sequencing, the overlap between forward and reverse read can be used to identify excess adapter sequences. This is exploited by several published adapter trimming tools. However, in our evaluations on amplicon-based paired-end data we found that these tools fail to remove all adapter sequences and that adapter contamination leads to spurious variant calls.Here we present SeqPurge, a highly-sensitive adapter trimmer that uses a probabilistic approach to detect the overlap between forward and reverse reads of paired-end Illumina sequencing data. The overlap information is then used to remove adapter sequences, even if only one base long. Compared to other adapter trimmers specifically designed for paired-end data, we found that SeqPurge achieves a higher sensitivity. The number of remaining adapters after trimming is significantly reduced compared to other tools. The specificity of SeqPurge is comparable to that of the competing tools. In addition to adapter trimming, SeqPurge also offers quality-based trimming, trimming of no-call (N) stretches, raw read quality-control and error-correction. SeqPurge is available at https://github.com/imgag/ngs-bits .	Genomes poster	Fundamental
P_Go114	731	Andrea Rodriguez-Martinez, Joram M Posma,Nikita Harvey, Jeremy K Nicholson, Marc-Emmanuel Dumas, Jean-Baptiste Cazier, Pierre Zalloua and Dominique Gauguier	Andrea Rodriguez-Martinez	Systems Genetics of Plasma 1H Nuclear Magnetic Resonance Metabotypes Associated with Cardiometabolic Diseases in a Lebanese Cohort	Coronary artery disease (CAD) has a multifactorial aetiology, combining environmental and genetic factors. Epidemiological studies have shown that a number of metabolic conditions are associated with increased risk of CAD. These so-called CardioMetabolic Diseases (CMDs) consist of a cluster of disorders including: type II diabetes mellitus, hypertension, non-alcoholic fatty liver disease, hyperlipidaemia, and visceral obesity. The comprehensive evaluation of the metabolic perturbations observed in CMDs represents a major challenge for accurate diagnosis and personalised healthcare. High-throughput metabolic phenotyping (ie metabotyping) by NMR targets low molecular weight compounds from biofluids or biopsies, which proved to be very successful in diagnosis of CAD, and predicting drug toxicity. Mapping disease-associated metabolites onto the human genome brings new insights in the molecular basis of CMDs and CAD. In order to achieve this, we focussed on a cohort of 1,949 genotyped patients with CAD and CMDs selected from previously studied collection of 8,709 Lebanese subjects with detailed clinical, biological, behavioral and social information available. The geographic location of Lebanon, at the crossroad of Europe, Africa, Asia Minor and the Middle East, makes our study population unique in its genetic characteristics, and represents an excellent opportunity to identify novel alleles regulating metabolite abundance in blood. We profiled 1,949 plasma samples from the cohort by 1H NMR. Metabolome-wide association studies were implemented taking into account demographic and risk factors. We identified several metabolites associated with CMDs, including alanine, histidine, proline, branch chain aminoacids (leucine, isoleucine, valine), lactate, myo-inositol, mannose, glucose, N-acetylated compounds, creatinine, and specific lipoproteins subfractions. These disease-associated metabolites are currently being mapped onto the genome to provide new insights in the genetic landscape of CMD-associated plasma metabolites.	Genomes poster	Biotechnology Health
P_Go115	590	Per K. I. Wilhelmsson, Kristian K. Ullrich and Stefan A. Rensing	Per K. I. Wilhelmsson	TAPscan – An updated genome-wide transcription factor classification workflow	Transcription associated proteins (TAPs) comprise the vast amount of proteins that influence transcription. These proteins are key players in gene regulatory networks and contribute to increasing the potential complexity of gene network circuitry. Here we have updated the work flow constructed by Lang et al. consisting of a set of domain-based classification rules aimed to identify TAPs amongst a given set of proteins. Methods based on the accumulative sequence knowledge of their time are in constant need of revision to stay up-to-date, given the ever increasing number of genomes becoming available. Major updates in workflow subprocesses, such as domain build and search software, are also essential to adopt. With a combination of custom built and existing (PFAM) hidden markov model (HMM) domain profiles a total of 122 TAP families can now be distinguished. This includes, for example, a further diversification of the homeodomain (HD) protein family from previously three to now twelve classes. By using a larger set of published sequences in building our domain profiles and incorporating the now larger amount of available genomes we aim to identify so far not discoverable expansions/gains within the kingdom Plantae (sensu lato).Lang et al. Genome-Wide Phylogenetic Comparative Analysis of Plant Transcriptional Regulation: A Timeline of Loss, Gain, Expansion, and Correlation with Complexity. Genome Biol Evol (2010), Volume 2, 488-503.	Genomes poster	Fundamental
P_Go116	522	Marko Verce, Luc De Vuyst and Stefan Weckx	Stefan Weckx	Taxonomic analysis of water kefir grains and liquor through shotgun metagenomics	Water kefir is a refreshing, fruity drink produced by inoculating water kefir grains into a sucrose solution supplemented with dried figs. Water kefir grains are polysaccharide grains containing microorganisms that ferment the sucrose mainly into lactic acid, acetic acid, and ethanol. In this study, the species diversity of the water kefir microbiota was analysed using shotgun metagenomic sequencing of four samples of a water kefir fermentation process, i.e., both water kefir grains and liquor at two time points. The total number of bases in the four metagenomes after quality control amounted to 1.86 Gbp. The reads were analysed using different tools to decrease the software- and database-dependent biases on the final assessment of the microbial communities present in the samples. The metagenomic reads were assigned to several bacterial genera, most prominently Lactobacillus (mainly L. casei/paracasei, next to L. hilgardii, L. nagelii, L. harbinensis, and L. hordei/mali), Bifidobacterium, Oenococcus, as well as two fungal species, i.e., Saccharomyces cerevisiae and Dekkera bruxellensis. The Bifidobacterium reads most likely belonged to a recently described water kefir-derived species, namely B. aquikefiri, whereas evidence was found that the Oenococcus reads may represent a new species too. The results reflect and support previous culture-(in)dependent research on water kefir. They also demonstrate the merits of using different tools and methods, including metagenome recruitment, to form a more reliable view of the composition of a microbial ecosystem.	Genomes poster	Agro-Food
P_Go117	815	Daniela Beisser, Nadine Graupner, Lars Grossmann, Jens Boenigk andSven Rahmann	Daniela Beisser	Taxonomic assignment of protist metatranscriptome sequences	Next generation sequencing technologies are increasingly applied to analyse complex ecosystems by mRNA sequencing of whole communities. In principle, each sequenced mRNA allows both an assignment of the underlying species and a functional annotation. While the functional information is sufficiently covered by databases such as Uniprot and NCBI the approach is currently limited by incomplete taxonomic references.For an accurate assignment of taxonomic groups to metatranscriptomic reads we build a custom database that comprises all major eukaryotic groups and a stand-alone tool to assign reads with a low false discovery rate.The custom database includes peptide sequences translated from transcriptomes of all relevant taxonomic groups, in total 146 species. We do not attempt to assign sequence reads on species or genera level, but taxonomic groups. The biggest problem is the misassignment of sequences to incorrect species. We therefore perform rigorous filtering, in which we evaluate the distance between the best hit and next hit in another taxonomic group. The developed tool (TaxMapper) is built in a modular way to be applicable stepwise with user-set parameters or as a complete easy-to-use analysis with standard parameters starting from a RAPsearch mapping file to a visualization of community composition.Additionally, we developed a reliable workflow for microeukaryotic metatranscriptome analysis. Written as a rule-based Snakemake workflow, it unites all major bioinformatic steps: preprocessing of raw reads, functional and taxonomic assignment with TaxMapper and statistical analyses. The set-up is generic and can be adjusted to any environmental sample.	Genomes poster	Ecosystems
P_Go118	611	Pawel Blazej, Wnetrzak Malgorzata, Dorota Mackiewicz and Pawel Mackiewicz	Pawel Blazej	The influence of selection at the amino acid level on the synonymous codons usage	There are two main forces that affect varying usage of synonymous codons:directional mutations and divers selection factors. The effectiveness of protein translation is usually considered as the main selectional cause. However, the biased codon usage can be also a by-product of a general selection at the amino acid level, which was showed by Morton (Morton, BR, 2001, Genetics 159:347-358). However, he considered this effect only for four selected mutational processes generating an equal frequency of complementary nucleotides. In order to test the universality of this phenomenon for various mutational processes, we evaluated a wide range of conditions in a mutation-selection model including almost 90,000 stationary nucleotide distributions generated by unrestricted stochastic processes. To determine the conditions in which the impact of selection at the amino acid level on the relative codon usage is minimized and maximized, we applied an evolutionary optimization algorithm. Our results indicate that the intensity of this effect strongly depends on the stationary distribution of the nucleotides and the type of synonymous codon groups. Generally, nucleotide substitution matrices leading to the maximization of this effect generate more adenine and thymine than guanine and cytosine as well as more purines than pyrimidines. The comparison of the simulation results with genomic data demonstrates that this effect is significant and can considerably interfere, especially in AT-rich genomes, with other selections on the codon usage, e.g. translational efficiency.	Genomes poster	Fundamental
P_Go119	694	Maryam Abdollahyan, Fabrizio Smeraldi, Boris Noyvert and Greg Elgar	Maryam Abdollahyan	Transcription Factor Binding Site-based Alignment of Conserved Non-coding Sequences	The identification and functional characterization of regulatory modules in the human genome is a challenging task. Regulatory modules act through the sequence-specific binding of transcription factors and previous studies have demonstrated that co-occurrence of transcription factor binding sites (TFBSs) in close proximity can be a good indicator of regulatory activity. In this study, we analysed the co-occurrence of TFBSs within a set of highly conserved non-coding elements (CNEs) that are associated with the regulation of early vertebrate development. From a computational point of view, analysis of the co-occurrence of TFBSs is complicated by the fact that TFBSs overlap. This rules out the use of classic alignment algorithms (that cannot handle alternative motifs in sequences) or k-mer-based approaches (that count the occurrences of motifs and would enumerate all alternative motifs indiscriminately). Our approach is fundamentally different in that we wrote each CNE as a sequence of symbols, each representing a TFBS identified within that element. We then constructed a graph representation of the CNEs which accounts for the ambiguity due to the overlapping of TFBSs and used a dynamic programming approach to find the optimal alignment between these graphs. We then computed the relative enrichment of short sequences of TFBSs in the alignments of CNEs compared to a background distribution. Our results identify a number of enriched TFBS alignments within CNEs, including a regulatory signature that has been functionally validated in this set of CNEs previously and is associated with hindbrain enhancer activity.	Genomes poster	Fundamental
P_Go120	348	Francesco Pezzini, Daniel Scharf, Ekaterina Shelest and Axel Brakhage	Francesco Pezzini	Transcription factors – histones interplay in regulation of stress response genes	Fungi are known to produce secondary metabolites (SMs). SMs can be synthesized by non-ribosomal peptide synthetases (NRPSs) or poliketide synthases (PKSs) through a complex multi-step process. The genes responsible for the biosynthesis of SMs are often organized in gene clusters - sets of genes which are co-regulated and co-expressed. Usually these clusters are silent but can be activated under particular stress conditions.Epigenetic control plays an important role in regulation of SM gene clusters. However, it is not yet shown if nucleosome occurrence can be one of the factors that influence the expression of gene clusters, and how nucleosome positioning is connected with the availability of transcription factor binding sites (TFBSs), especially for pioneer TFs.Therefore we investigated CCAAT boxes, obiquitous motifs, that are involved in several stress responses. These motifs are a well characterized binding pattern of Hap TF complex, pioneer TF that has a strong structural similarity with histones H2A and H2B and is found in some SM clusters as well. To get insights into the mechanisms of Hap–nucleosome interplay, we constructed deletion mutants for one of Hap subunits, HapC. ΔHapC and wild type transcriptomes were confronted to investigate the occupation of the CCAAT boxes by nucleosomes in known Hap targets and SM clusters. The results help to understand if and how the TF displaces the nucleosome to induce the expression, and what is the impact of this process on the expression of gene clusters.	Genomes poster	Fundamental
P_Go121	701	Ritambhara Singh, Jack Lanchantin and Yanjun Qi	Ritambhara Singh	Transfer String Kernel for Cross-Context DNA-Protein Binding Prediction	This work focuses on sequence-based string classification tasks that aim to accurately predict the DNA binding sites of proteins called transcription factors (TF) in unannotated cell contexts. Previous approaches are unable to perform such accurate predictions, since they do not consider distinctions among sequence segments from annotated (source) and unannotated (target) contexts. We therefore propose a novel method called ``Transfer String Kernel" (TSK) that achieves improved transcription factor binding site (TFBS) predictions using cross-context sample adaptation. TSK maps sequence patterns to a high-dimensional feature space using the discriminative mismatch string kernel framework under SVM. Labeled examples from a source (annotated) context are transferred to a target (unannotated) context by re-weighting source samples adaptively. We have experimentally verified TSK's ability of TFBS identifications for fourteen different TFs under a cross-organism setting. We find that TSK consistently outperforms the state-of-the-art TFBS tools, especially when working with TFs whose sequences are not conserved across contexts. We also demonstrate the generalizability of our model by showing its cutting-edge performance on a different set of cross-context tasks for peptide binding prediction.	Genomes poster	Biotechnology
P_Go122	380	Tommi Rantapero, Minna Ampuja, Alejandra Rodriguez-Martinez, Maaria Palmroth, Matti Nykter and Anne Kallioniemi	Tommi Rantapero	Uncovering gene regulatory basis of differential BMP4 response in breast cancer cell lines	Bone morphogenetic proteins (BMPs) are a group of growth factors that have been shown to have a role in breast cancer progression. It has been shown that BMP4 reduces proliferation in multiple breast cancer cell lines in vitro, while simultaneously inducing migration in a subset of the cell lines. Our study aims to uncover the early BMP4 regulatory target genes and characterize the chromatin landscape in order to gain insight into the underlying basis for the different BMP4 response in breast cancer cell lines.In this study, response to BMP4 stimulation in two breast cancer cell lines MDA-MB-231 (responds to BMP4 by increased migration) and T-47D (responds by decreased proliferation) were studied. RNA-seq and DNase-seq were conducted for both cell lines after 3 h stimulation with BMP4 and untreated control. DNase I hypersensitive sites (DHS, which correspond to regulatory sites within the genome) and differential DHS sites were detected from the DNase-seq data. Furthermore, digital footprinting and transcription binding site prediction were conducted for all DHS-sites.RNA-seq data revealed altogether 92 differentially expressed genes in MDA-MB-231 and 204 differentially expressed genes in T-47D. A subset of differentially expressed genes were selected and validated with qPCR. In addition, a detailed inspection of the open chromatin sites in the promotor regions of upregulated genes in MDA-MB-231 revealed enrichment of several transcription factor binding sites, including SMAD4 which is a known mediator of BMP4 signaling. Further analysis and experiments will reveal a more detailed view of the transcriptional regulation.	Genomes poster	Fundamental
P_Go123	572	Jan Grau, Jens Keilwagen, Michael Wenk, Jessica Erickson, Martin Schattat and Frank Hartung	Jan Grau	Using intron position conservation for homology-based gene prediction	Next generation sequencing has lead to a rapid increase in the number of sequenced genomes. Initial annotation of protein-coding genes in newly sequenced genomes is typically based on computational predictions.Here, we present a homology-based gene prediction program called GeMoMa, which explicitly incorporates the conservation of intron positions. GeMoMa utilizes gene models from a related species and predicts gene models in the genome of an organism of interest. In contrast to transcriptomics-based gene predictions, GeMoMa is capable of predicting rarely transcribed genes. By design, GeMoMa, provides information about putative homologous gene pairs and allows for transferring information about gene function from one organism to another. We apply GeMoMa to several animal and plant species and compare it with state-of-the-art competitors based on available annotations, using RNA-seq data, and Sanger sequencing. Our key findings are: i) Utilizing intron position conservation improves homology-based gene prediction and ii) predictions of GeMoMa can help to improve existing or add new transcripts in annotated genomes. The development of homology-based gene prediction tools has largely stalled during the last years. However, we demonstrate that the inclusion of additional features may substantially improve prediction performance. Hence, our results might trigger the investigation of further features.	Genomes poster	Fundamental
P_Go124	440	Dmitry Ravcheev and Ines Thiele	Dmitry Ravcheev	Utilization of mucin glycoconjugates by human gut microbiota: analysis by comparative genomics	Mucins are high molecular weight, heavily glycosylated proteins produced by epithelium in most animals. In the human intestine, mucins are responsible for forming of the mucus layer. Recent finding demonstrated that alterations in mucin glycoconjugates (MGC) impact on the composition of human gut microbiota (HGM). Here, we present a systematic analysis of HGM encoded systems for degradation of MGC.We applied genomic analysis to 399 HGM genomes microorganisms found in the human gut belonging to the phyla of Firmicutes, Bacteroides, Proteobacteria, Actinobacteria, and Fusobacteria. We analyzed genes required for the degradation of MGC to monosaccharides as well as genes responsible for the utilization of these monosaccharides (fucose, galactose, N-acetylgalactosamine, N-acetylglucosamine, and N-acetylneuraminic acid) as sources of carbon and energy. Genes for utilization of one or more monosaccharides were found in 373 (93%) studied genomes. We found that not all MGC derived monosaccharides could be utilized by the MGC degrading microbes. For instance, only 3 (0.75%) HGM organisms could utilize all five monosaccharides. Additionally, we predicted exchange pathways for MGC derived polysaccharides in the HGM. For example, Lactobacillales genomes have enzymes to separate fucose from the MGC but have no genes for the utilization of this monosaccharides. On the other hand, Bifidobacteriaceae and Enterobacteriaceae have only the utilization pathways but no fucose-separating enzymes. Thus, we propose that HGM organisms are collaborating in the harvesting and utilization of MGC.Taken together, this work substantially expands our knowledge on metabolic interactions between HGM as well as interactions between the HGM and the host organism.	Genomes poster	Fundamental Health

THEME/TRACK: PROTEINS

Poster numbers: P_Pr001 - 080 Application posters: *P_Pr001 - 009*

Poster number	EasyChair number	Author list	Presenting author	Title	Abstract	Theme/track	Topics
APPLICATION POSTERS WITHIN PROTEINS THEME							
P_Pr001	*674*	*Fatemeh Abbasi, Changiz Eslahchi and Reza Hassanzadeh*	*Fatemeh Abbasi*	*A GRAPH THEORETICAL APROACH FOR DRUG TARGET PREDICTION*	*Motivation: The discovery of novel drug targets is a significant challenge in drug development. Many of the currently known drug targets are functionally pleiotropic and involved in multiple pathologies. Several of them are exploited for treating multiple diseases, which highlights the need for methods to reliably reposition drug targets to new indications. So, the identification of interactions between drugs and target proteins is a key area in genomic drug discovery. Therefore, there is a strong incentive to develop new methods capable of detecting these potential drug–target interactions efficiently. Computational methods for novel drug target predictions can greatly reduce time and costs compared with experimental methods.Results: In this work, we present a network based computational approach for novel drug and target association predictions. More specifically, a heterogeneous drug-target graph, which incorporates known drug-target interactions, is first constructed. Based on this graph, drug-drug and target-target similarities, a novel graph based inference method is introduced. Compared with two state of the art methods, 10 fold cross-validation and jackknife results on different data sets, involving targets of enzyme, ion channel, GPCR, nuclear receptor and complete DrugBank indicate that the proposed method can greatly improve novel target predictions.*	*Proteins/ Application poster*	*Application Fundamental*
P_Pr002	*673*	*Changiz Eslahchi, Ali Madi and Changiz Eslahchi*	*Changiz Eslahchi*	*Discovering overlapped protein complexes from weighted PPI networks by removing inter-module hubs*	*Motivation: Detecting known and predicting undiscovered protein complexes from protein-protein interaction (PPI) networks helps us to understand principles of cellular organization and their functions. Nevertheless, extraction of protein complexes from PPI network isn't an easy task. Two major constrains are high noise level and ignoring occurrence time of different interactions in PPI network.Results: An efficient algorithm (IMHRC) is developed based on inter-module hub removal in the weighted PPI network which can detect overlapped complexes. IMHRC by removing some of the inter-module hubs and module hubs, eliminates a meaningful percentage of noise in our dataset and indirectly consider difference occurrence time of the PPI in our network. After removing hubs, some proteins are considered as seeds. Each seed creates a primary cluster. Then removed module hubs are added to the resulting clusters based on the amount of their interactions with other proteins in the clusters. Clusters are then merged based on their overlaps. Consequently, the performance of the IMHRC is evaluated on several benchmark datasets and the results are compared with other state-of-the-art models. The protein complexes that discovered by IMHRC method significantly match with the real data and much better than other methods.*	*Proteins/ Application poster*	*Application Fundamental*
P_Pr004	*847*	*Thomas Kemmer and Andreas Hildebrandt*	*Thomas Kemmer*	*Efficient nonlocal electrostatics computations for proteins using the Julia programming language*	*Electrostatic interactions are a major contributor to protein-protein and protein-ligand interactions. In contrast to other molecular interaction components, they can be significant over medium to long distances and are thus crucial for molecular visibility. Research areas such as rational drug design require accurate estimates of potentials and free energies influenced by electrostatics. One major challenge in this context, however, is the treatment of the solvent the molecules are immersed in, i.e., water in a biological context. Strong simplifications of the structure of such polarizable and highly structured solvents are commonplace to achieve the required computational efficiency, but invariably lead to inaccuracies.Here, we present efficient protein electrostatics computations in a single and easily extensible software package for the cross-platform and open-source Julia programming language. By modelling water in an implicit but nonlocal fashion, we account for correlation of molecular polarization due to the water network around the solute and sustain accuracy without suffering from infeasible runtimes as compared to the explicit case. Our package contains implementations for our own Boundary Element (BEM) solver as well as a reference Finite Element (FEM) solver, both profiting from the good base performance of the Julia language, which can achieve runtimes comparable to C. Additionally, Julia's native and non-native interoperability with other languages such as C, Fortran, R, and Python allows for easy incorporation of our package into existing pipelines.*	*Proteins/ Application poster*	*Application*
P_Pr005	*472*	*Saba Ferdous and Andrew Martin*	*Saba Ferdous*	*Exploration of conformational B-cell epitopes: components to peptide-based vaccines*	*Peptide vaccines have many potential advantages including low cost, lack of need for cold-chain storage and safety. However, it is well known that approximately 90% of B-cell Epitopes (BCEs) are discontinuous in nature making it difficult to mimic them for creating vaccines. We have analyzed the discontinuity of B-cell epitopes by defining extended 'regions' (R, consisting of at least 3 antibody-contacting residues each separated by <= 3 residues) and small fragments (F, antibody-contacting residues that do not satisfy the requirements for a region). Secondly, we have classified region shape as linear, curved or folded. Furthermore, by using molecular dynamics, we have studied mutations in linear and folded (two alpha helices or beta strands connected by hairpin loop) regions that stabilize their conformation: end capping, mutations of hydrophobics (non contacting residues of an epitope) to alanine and glutamine, disulphide stapling and cyclization. We have explored mutations in five linear and five folded epitopes with up to 20 mutant for each of the epitopes. Moreover, to confirm the stability of a stable mutant in the presence of an antibody, it has been simulation with antibody. The stablised epitope mimetics (mutant) will be tested experimentally to check their possibility to use as immunogens for peptide vaccine design.*	*Proteins/ Application poster*	*Application Health*
P_Pr006	*326*	*Anoosha Paruchuri, Huang L-T, Sakthivel R, Karunagaran D and Michael Gromiha M*	*Anoosha Paruchuri*	*Exploring preferred amino acid mutations in cancer and discriminating driver and passenger mutations in Epidermal Growth Factor Receptor*	*Cancer is one of the leading causes of death worldwide. Huge number of somatic mutations get accumulated during cancer development, among which contributes to tumor progression are known as "driver" mutations, whereas most of them are functionally neutral known as "passenger" mutations. Hence, discriminating these mutations has been an active field in cancer research. In this study, we have systematically analysed the effect of these mutations at protein level in 41 different cancer types from COSMIC database on different perspectives: (i) Preference of residues at the mutant positions (ii) Probability of substitutions (iii) Influence of neighbouring residues (iv) Distribution of driver and passenger mutations around hotspot sites and (v) Distribution of silent and missense substitutions. This study reveals the variation of mutations at protein level in different cancer types and their preferences in cancer genes and provides new insights for understanding cancer mutations and drug development. Further, considering the importance of EGFR (Epidermal Growth Factor Receptor) protein based on the number of observed missense mutations in cancer, we have developed a reliable classification model for discriminating driver and passenger mutations in this protein. We grouped the mutations based on secondary structure and accessible surface area and acheived an overall classification accuracy of 80.2%, 81.9%, 77.9% and 75.14% for helix, strand, coil buried and exposed mutants, respectively. We have screened all possible missense mutations in EGFR and suggested probable driver and passenger mutations, which would help in the development of mutation specific drugs for cancer treatment.*	*Proteins/ Application poster*	*Application*
P_Pr007	*368*	*Rakesh Kumar Meena, Sayane Shome and Sanket Thakur*	*Rakesh Kumar Meena*	*In silico prediction of Okra (Abelmoschus esculentus L) encoded micro-RNAs targets, Structure prediction and Molecular docking studies for Okra yellow Vein Mosaic virus.*	*Begomovirus associated symptoms were observed in several Abelmoschus esculentus plants growing in crop fields in India as well as whole world. Protein sequence of the viral coat protein from the yellow vein mosaic virus was collected from NCBI protein database [Accession ID: NP_579972].The nucleotide sequence and the coordinates of an Okra leaf isolate was obtained from NCBI Nucleotide database [Accession ID: KC404826]. The nucleotide sequence was then subjected to sequence search in MiRbase which utilizes BLASTN algorithm to find candidate miRNAs deposited from the database. The miRNA determined in the nucleotide sequence[Accession ID : MI0027065] lies in the interspacer region which further supports our claim for the miRNA candidate for the analysis.3-dimensional coordinates for the viral coat protein and the miRNA candidate was predicted by Modeller software and I-Tasser server (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) . The best suitable structure was determined by validation using SAVS server and observed more than 93% residues in core region. Structural annotation and cavity prediction was carried out by PSI-Phred server and Castp server respectively. The 3-dimensional structure of miRNA candidate was predicted using Chimera software. Molecular docking of the viral coat protein with miRNA candidate was carried out via Autodock4.2 software. Grid coordinates was determined by Autogrid 4 and Lamarckian Genetic algorithm was used for the docking process. Interacting residues participating in the molecular docking was visualized by Chimera and Ligplot+ software. The computational study comprising molecular binding of OYVMV coat protein with miRNA from Okra leaf isolate shows promising results which can be replicated in experimental studies to devise novel therapeutic strategy to treat Okra yellow vein mosaic viral disease.*	*Proteins/ Application poster*	*Application Biotechnology*
P_Pr008	*858*	*Pooya Zakeri, Jaak Simm, Adam Arany, Forough Amini, Mehdi Sadeghi and Yves Moreau*	*Pooya Zakeri*	*Protein Fold Recognition Using Matrix Factorization Technique*	*Most of protein fold predictor machines only cover less than 30 folds, which is far less than protein folds have been identified. Moreover, the typical approaches proposed for protein fold recognition often neglect the relationship between protein folds. These motivate us to formulate the protein fold recognition as a factorization of an incompletely filled binary protein-fold-matrix where the objective is to predict unknown values. Protein fold recognition database such as SCOP can be seen as an incomplete matrix (M×N) where each row is a protein and each column is a protein fold. Then, the SCOP matrix can be modeled by the two smaller matrices P and F, which, when multiplied, approximately reconstruct the SCOP matrix. We propose an extended version of the Bayesian probabilistic matrix factorization [1] with the added advantage of working with multiple protein features as side sources for completing protein-fold-matrix. Accordingly, two sequence-based protein features, including the predicted secondary structure and information extracted directly from position-specific scoring matrices, are incorporated into the proposed factorization model as side information.In order to validate our models in a more realistic task setting, we develop a prospective benchmark, based on the latest version of the SCOP database, which covers about 200 protein folds. The experimental results on our unbiased benchmark show that our proposed model can effectively improve the accuracy of the state-of-the-art protein fold predictors such as GeoFold [2].[1] doi: 10.1093/bioinformatics/btu118,[2] doi: 10.1145/1390156.1390267.*	*Proteins/ Application poster*	*Application Fundamental*
P_Pr009	*606*	*Dhoha Triki, Telli Billot, Benoit Visseaux, Diane Descamps, Anne-Claude Camproux andLeslie Regad*	*Dhoha Triki*	*Study of natural resistance mechanisms of HIV protease-2 (PR2) against protease inhibitors (PI)*	*The therapeutic arsenal against the HIV of type 2 (HIV-2) corresponds to antiretroviral drugs developed for HIV-1. HIV-2 is naturally resistant to some of these drugs. It is therefore important to find new drugs against HIV-2. A solution is to develop specific molecules inhibiting HIV-2 protease (PR2), an enzyme involved in the maturation of virus proteins (Brower et al. 2008). Understand what factors contribute to the efficiency of inhibitors for HIV-1 protease (PR1) and absent from the PR2 can help to improve the PR2 inhibitor design.In this study, we compared a set of 36 structures of PR1 and PR2. They exhibit 48% of sequence identity. Mutations between PR1 and PR2 are primarily located on the elbows regions and few mutations are located in the PI-binding site. We analyzed the effects of these mutations on PR2 structure. First we observed that these mutations seem to modify the PR2 flexibility: PR2 structures have on average higher Bfactor values, meaning PR2 structures are more flexible than PR1. We then noted that these mutations have an effect on the properties of PI-binding sites: those extracted from PR2 structures are less hydrophobic, smaller and more polar than those of PR1. Finally, we observed that these mutations modify PR interface properties: PR2 dimer structures exhibit a lesser energetic stability than PR1 interfaces. To conclude, our study showed that mutations between PR1 and PR2 have important effects on PR. Molecular dynamics simulations could be used to understand the effect of these mutations on the PI-binding mode.*	*Proteins/ Application poster*	*Application Health*
OTHER POSTERS WITHIN PROTEINS THEME							
P_Pr010	389	Patrick Löffler, Samuel Schmitz, Enrico Hupfeld and Rainer Merkl	Patrick Löffler	A Modular Framework to Extend Rosetta Protocols with Multistate Design	Computational protein design (CPD) is a powerful technique to design novel proteins. Many CPD objectives such as design on backbone ensembles, multi-specificity design and the integration of negative design demand the simultaneous optimization of multiple design states. Rosetta is a popular software suite to study and design proteins. Rosetta's protocols consist of specific procedures and a fine-tuned set of parameters to carry out a given task. An example is the use of specific sequence design cycles and catalytic constraints in the enzyme design protocol. At present, the multistate design implementation of Rosetta is a generic approach lacking options to fine tune the calculations in the same manner as specialized single state protocols.We have developed a framework for CPD that integrates multistate design in existing Rosetta protocols while preserving the protocol's original functionality. Our framework consists of two, easily exchangeable components: i) The optimizer searches the sequence space and ii) the evaluator scores the sequences according to the given design task. Currently, we utilize Rosetta's genetic algorithm as an optimizer; protocols for enzyme design and protein-protein interface design serve as evaluators. However, due to the modularity of both components, multistate functionality can be transferred to arbitrary Rosetta applications with little effort. We have benchmarked the above two applications on two datasets consisting of conformational ensembles and achieve an 18 percent performance improvement over conventional methods. As a proof of concept, we have applied our framework to computationally design retro-aldolases which are currently subject to biochemical characterization.	Proteins poster	Biotechnology
P_Pr011	854	Dina Cramer, Luis Serrano and Martin H Schaefer	Martin H Schaefer	A network of epigenetic modifiers and DNA repair genes controls tissue-specific copy number alteration preference	Copy number alterations (CNAs) show a large variability in their number, length and position over cancer types. This variability is clinically relevant as both the amount and length of CNAs (as well as the identity of the affected genes) have a strong impact on patient survival. However, the sources of this variability are not known. We aim to identify genetic and epigenetic factors that contribute to this variability.Analyzing patient data from The Cancer Genome Atlas (TCGA), we have identified proteins that tend to be mutated in samples having few or many CNAs, which we term CONIM proteins (COpy Number Instability Modulators). CONIM proteins cluster into a densely connected subnetwork of physical interactions and many of them are epigenetic modifiers. Therefore, we investigate how the epigenome of the tissue-of-origin influences the position of CNA breakpoint regions and the properties of the resulting CNAs. We find that the presence of heterochromatin in the tissue-of-origin contributes to the recurrence of CNAs in the respective cancer type. We show that these epigenetic states also impact the length of the resulting CNAs, elucidating differences in the mechanisms underlying CNA generation.Therefore, we demonstrate how both the tissue-of-origin epigenome organization and a newly identified class of cancer genes affect the variability of CNA number over patients and cancer types.	Proteins poster	Health
P_Pr012	483	Isaure Chauvot de Beauchene, Sjoerd De Vries and Martin Zacharias	Isaure Chauvot de Beauchene	A new fragment-based docking approach to model protein-bound ssRNA from sequence.	Protein-RNA recognition supports many cellular functions. Abnormal protein-RNA interactions are crucial therapeutic targets in e.g. neurodegenerative diseases and RNA viruses infections. Moreover, synthetic RNA aptamers can be used as protein modulators. The rational design of either aptamers or RNA-protein interaction inhibitors requires atomistic description of protein-RNA complexes. Yet their experimental resolution is arduous, and protein-RNA computational docking is hampered by the high flexibility of RNA single-stranded regions, which mostly provides recognition specificity. The lack of methodology for modeling ssRNA limits all protein-RNA docking methods [2].We developed an original fragment-based approach, predicting ssRNA-protein complexes structure from protein structure and RNA sequence: We (i) cut the RNA sequence in overlapping trinucleotides, represented by sequence-specific ensembles of conformers that we built from known protein-RNA structures; (ii) dock each ensemble on the protein; (iii) select the spatially compatible poses; (iv) assemble them in a realistic conformation.Moreover, we developed and validated an RNA-protein contact predictor, based on statistical analysis of known complexes, which provides (optional) starting points for fragments docking.We applied them on ten complexes with various ssRNA sequences (6-11 nucleotides) and RNA-recognition domains. Without predicting specific contacts, we can identify the RNA binding site more accurately than existing methods [3]. Based on predicting 3-4 contacts, the method allows modeling of bound ssRNA within 1-2Å RMSD [1]. Such crystallographic-like precision, not reached so far , reveals a methodological breakthrough in RNA-protein docking.[1] Chauvot-de-Beauchene et al (2016) NAR 44(10):4565-4580[2] Fulle, Gohlke (2010) JMR 23(2):220-231[3] Chauvot-de-Beauchene et al (2016) PloS.Comput.Biol. 12(1):e1004697	Proteins poster	Biotechnology
P_Pr013	824	Mark Wass, Sarah Jeanfavre, Michael Coghlan, Martin Ridout, Anthony Baines and Michael Geeves	Mark Wass	Adaptation of mammalian myosin II sequences to body mass	The speed of muscle contraction is related to body size; muscles in larger species contract at a slower rate. We investigated the evolution of twelve myosin II isoforms to identify any adapted to increasing body mass. β-myosin head domain had the greatest rate of sequence divergence (0.05% per Myr) and was the only domain where sequence divergence correlated with body mass (0.091% divergence per log mass unit). β-myosin is abundant in cardiac ventricle and slow skeletal muscle. We propose that β-myosin has adapted to enable slower heart beating and contraction of slow skeletal muscle as body mass increased. Additionally, for eight of the twelve myosins, the ratio of divergence in the head:tail domains was significantly different, ranging from 3:1(β-myosin) to <1:2(extraocluar, non-muscle A and embryonic myosin). Our data provide new insights into the evolution of myosin function and indicate distinct evolutionary pressures on head and tail domains in individual isoforms.	Proteins poster	Fundamental

P_Pr014	452	Michal Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Paweł Mackiewicz and Małgorzata Kotulska	Michal Burdukiewicz	AmyloGram:a novel predictor of amyloidogenicity	Amyloids are proteins associated with the number of clinical disorders (e.g., Alzheimer's, Creutzfeldt-Jakob's and Huntington's diseases). Despite their diversity, all amyloid proteins can undergo aggregation initiated by 6- to 15-residue segments called hot spots. Henceforth, amyloids form unique and often zipper-like β-structures, which can turn out harmful.To find patterns defining the hot-spots, we trained predictors of amyloidogenicity based on random forests using n-grams extracted from amyloidogenic and non-amyloidogenic peptides collected in the AmyLoad database.Since the amyloidogenicity may not depend on the exact sequence of amino acids but on more general properties of amino acids in the sequence, we constructed 524 284 reduced amino acid alphabets of different lengths (three to six letters) based on all possible combinations of the handpicked physicochemical properties of the amino acids. The cross-validation of predictors employing the different alphabets revealed the best-performing alphabet with the length of 6 amino acid residues. During analysis we found also 65 n-grams that are the most relevant to the discrimination of amyloid and non-amyloid sequences of which 15 were confirmed experimentally elsewhere.The best-performing predictor, AmyloGram, was benchmarked against the most popular tools for amyloid peptides detection using an external dataset. It has obtained the highest values of performance measures (AUC: 0.90, MCC: 0.63).AmyloGram is available as a web-server: www.smorfland.uni.wroc.pl/amylogram/.	Proteins poster	Health
P_Pr016	531	Dhoha Triki, Mario Cano Contreras, Delphine Flatters, Benoit Visseaux, Diane Descamps, Anne-Claude Camproux andLeslie Regad	Leslie Regad	Analysis of the HIV-2 protease deformation involved by inhibitor binding	HIV-2 is a retrovirus discovered a few years after HIV-1. HIV-2 infections are restricted mainly to West Africa and to some European countries (Valadas et al., 2009 ;Brunet S, et al. 2008). The HIV-1 and HIV-2 genomes differ by about 50% at the nucleotide level. Such differences may be correlated with differential responses to some antiretrovirals such as some protease inhibitors (PIs) (Poveda E et al., 2005, Ren J, et al., 2002). It is necessary to develop new therapeutic molecules specific to HIV-2. One approach is based on the identification of new molecules inhibiting the HIV-2 protease (PR2), a protein involved in HIV-2 protein maturation. To do so, it is important to understand which features are contributing to the PI selectivity and efficiency for the HIV-1 protease (PR1) and absent in PR2. The understanding of the interaction mode between antiretroviral drugs with the PR2 and the PR2 structural deformation implied by the inhibitor binding can help to this task.In this study, we first compared the inhibitor-binding pockets extracted from 19 Xray structures of PR2 (apo and holo forms). In a second step, we analysed the PR2 plasticity using SA-conf tool. This tool analyzes the structural plasticity of a target by comparing the local structures of its different conformations. SA-conf allowed us to highlight the PR2 structural variable regions putatively involved by the PI-binding. This study allowed us to detect residues important for the inhibitor binding in PR2 and to better understand the PR2 deformation implies by the inhibitor-binding.	Proteins poster	Health
P_Pr017	651	Galo Ezequiel Balatti, M. Florencia Martini and Monica Pickholz	Galo Ezequiel Balatti	Antimicrobial peptides mechanisms of membrane lysis and permeation by computer simulations	Antimicrobial peptides (AMPs) are part of the innate immune system, attaching and inserting to the lipidic membranes of external agents among bacteria, fungi, viruses and eukaryotic parasites and killing the cells through a membrane permeation effect.Nevertheles, their molecular mechanisms are not well-known and three different leakage pathways was proposed: the "barrel-stave", the "carpet" or the "toroidal-pore" models. Among AMPs, two peptides obtained from Australian tree frogs, the Aurein 1.2 and the Maculatin 1.2 are proposed as AMPs with different leakage pathways. Here, we carried out extensive Molecular Dynamics (MD) simulations to study the peptide interactions with lipid structures in order to shed light into these mechanisms.. We have used a coarse grain (CG) model within the MARTINI force field[1]. Three simulation replicates were performed, looking to the self-assembly of 1000 lipids (2-oleoyl-1-palmitoyl-sn-glycero-3-phosphocholine, POPC) in the presence of the peptides were performed. Furthermore, we simulate both peptides in a presence of a pre-equilibrated bilayer from two different initial configurations: aqueous phase and inside the bilayer.The simulations results showed two different pathways on the membrane leakage, in good agreement with experimental observations.[2] While Maculatin can form a pore maintaining the structure of the bilayer, Aurein causes the total membrane destabilization and disintegration. A better understanding of AMPs molecular behavior can aim the development of new antimicrobials drugs.[1] X. Periole, S.J. Marrink. Methods in molecular biology 925 (2013) 533-565[2] E.E. Ambroggio et al Biophysical Journal 89 (2005) 1874–1881	Proteins poster	Health
P_Pr018	693	Maria Katsantoni, Tjaart de Beer and Torsten Schwede	Maria Katsantoni	Assessing functional conservation in alternative splice forms	In 75% of human genes, alternative splicing gives rise to more than one transcript per gene. However, little is known about the functional significance these alternative products have. Thanks to RNA-seq technology, human transcriptome data are constantly increasing, which gives a better view of how alternative transcripts are expressed under different conditions e.g. normal and cancer tissue data.In this work we focus on the alternative protein-coding transcripts and what their functional importance may be on the protein level. For this purpose, we combine RNA-seq expression information and functional annotation on the protein level. All available protein-coding transcripts are annotated on the protein level in terms of functional characteristics (e.g. active sites, protein-protein interaction regions and domains). This annotation is based on existing knowledge of one of the proteins of a gene (SwissProt canonical protein isoform) and on evolutionary information.Combining the tissue RNA-seq data with the annotations, we identify cases where the highest expressed isoform is not the canonical isoform and try to characterise how the functional characteristics behave for this set of proteins. This is done via a custom functional conservation score.One of the key findings of this work is the observation of a bimodal distribution of the functional characteristics. That is, there is a tendency for alternative splicing to prefer either inclusion or exclusion of a functional characteristic in contrast to partial inclusion.	Proteins poster	Fundamental
P_Pr019	620	Fabian Sievers and Des Higgins	Fabian Sievers	Benchmarking Multiple Protein Sequence Alignments and the Effect of Guide-Tree Topology	Background: Multiple Sequence Alignments (MSAs) of large numbersof sequences are used in many bioinformatics analyses. However, thequality of progressive MSAs scales badly with the number of sequences.Methods: We show how the quality of MSAs decreases with largenumbers of sequences by benchmarking the quality of the alignment ofembedded reference sequences. One shortcoming of this benchmark isthat only a small fraction of sequences contributes to the qualityassessment of the MSA. We therefore present two schemes that eitheruse contact-map or secondary structure predictions based on the MSA asa measure of quality. These methods use all sequences and columns inthe alignment and are independent of potentially incorrectly curatedreference alignments.Results: The quality of MSAs decreases markedly for all alignersin our study, as the number of sequences is increased beyond a fewhundred sequences. Iteration can increase the useful range ofsequences to something like 1,000 sequences. Using high-quality HMMsseems to have the greatest effect in maintaining MSA quality at thisstage. The usefulness of chained guide-trees and high-qualitybackground HMMs can also be confirmed using our contact-map andsecondary structure prediction methods, which broadly correlate withscores derived from embedded crystal-based reference alignments.References: Fox G, Sievers F, Higgins DG (2015) Using de novo protein structure predictions to measure the quality of very large multiple sequence alignments. Bioinformatics; doi: 10.1093/bioinformatics/btv592	Proteins poster	Fundamental
P_Pr020	382	Po-Chia Chen and Jochen Hub	Po-Chia Chen	Biomolecular structure and dynamics via combined solution scattering experiments & atomistic simulations	X-ray and neutron solution scattering are powerful techniques that are capable of probing the solution behaviour of biomolecules. The measured scattering intensities contain information about the structural ensemble, both in terms of average structure and diversity. However, this information is disguised behind a global average over all conformations and orientations. Thus, measured SAS and WAXS patterns must be interpreted, ideally with independent atomic-level information to alleviate intrinsic ambiguity issues.We previously implemented an explicit-solvent approach in GROMACS to predict the ensemble SWAXS pattern of a biomolecule using molecular dynamics, and demonstrated the necessity of sampling at least picosecond and nanosecond-level freedoms in order to accurately reproduce experiment. Level of sampling depend on the underlying flexibility: from stable folds where accuracy is limited by sidechain and solvent sampling, up to intrinsically-disordered proteins, where accuracy is limited by the sampling of overall conformations. The MD integration also permits the use of SAXS data as constraints, which enables the direct isolation of structures consistent with a target SAXS pattern using related atomic coordinates as the starting conformation.A summary of above functionalities will be presented, along with planned extensions to integrate SANS techniques with contrast variation. We also plan to make capabilities available to integrative modelling workflows on HPC and cloud centers.	Proteins poster	Fundamental
P_Pr021	460	Gergely Gyimesi, Péter Závodszky and András Szilágyi	András Szilágyi	Calculation of configurational entropy differences from conformational ensembles using Gaussian mixtures	The configurational entropy of a molecular system is an important component of free energy that is often neglected in free energy calculations because of the inherent difficulty of the entropy calculation. The commonly used quasiharmonic method is unable to account for multiple basins and anharmonicities in the energy landscape. Here, we present a novel, conceptually simple approach to calculate the configurational entropy difference between two conformational ensembles (typically generated by molecular dynamics or Monte Carlo simulations) of a molecular system. The method estimates the probability density function of the system by a Gaussian mixture, using an efficient greedy learning algorithm along with a cross-validation based stopping criterion. Evaluating the method on conformational ensembles corresponding to substates of five small peptide systems, we found excellent agreement with the exact entropy differences obtained from a full enumeration of conformations. Compared with the quasiharmonic method and two other, more recently developed methods, the Gaussian mixture method yields more accurate results at smaller sample sizes. We illustrate the power of the method by calculating the backbone torsion angle entropy difference between disulfide-bonded and non-disulfide-bonded states of tachyplesin, a 17-residue antimicrobial peptide, and between two substates in the native ensemble of the 58-residue bovine pancreatic trypsin inhibitor. The Gaussian mixture method is a powerful and accurate approach for calculating configurational entropy differences for systems with complex energy landscapes. The program is written in Python and is available from the authors upon request.	Proteins poster	Fundamental
P_Pr025	322	Waqar Ali, Anatol Wegner, Robert Gaunt, Charlotte Deane and Gesine Reinert	Charlotte Deane	Comparison of large networks with sub-sampling strategies	Networks are routinely used to represent large data sets, making the comparison of networks a tantalizing research question in many areas. Techniques for such analysis vary from simply comparing network summary statistics to sophisticated but computationally expensive alignment-based approaches. Most existing methods either do not generalize well to different types of networks or do not provide a quantitative similarity score between networks.In contrast, alignment-free topology based network similarity scores empower us to analyse large sets of networks containing different types and sizes of data. Netdis is such a score that defines network similarity through the counts of small sub-graphs in the local neighbourhood of all nodes.Here, we introduce a sub-sampling procedure based on neighbourhoods which links naturally with the framework of network comparisons through local neighbourhood comparisons. Our theoretical arguments justify basing the Netdis statistic on a sample of similar-sized neighbourhoods. Our tests on empirical and synthetic datasets indicate that often only 10% of the neighbourhoods of a network suffice for optimal performance, leading to a drastic reduction in computational requirements. The sampling procedure is applicable even when only a small sample of the network is known, and thus provides a novel tool for network comparison of very large and potentially incomplete datasets.	Proteins poster	Fundamental
P_Pr026	365	R. Charbel Maroun, Pa Curmi and H El Shanti	R. Charbel Maroun	Consanguinity, genetic disease and molecular simulations	Two siblings born to a consanguineous couple with a previously un-described syndrome were identified. CLDN10 on chromosome 13 stood out as the best candidate gene. Re-sequencing of the coding region of CLDN10 and the flanking splice sites revealed a missense variation c.392C>T (NM_006984); p.S131L in claudin-10b, one of the alternatively spliced isoforms.The claudins are integral membrane proteins involved in the formation of the Tight Junction, which serves as a physical barrier to prevent solutes and water from passing freely through the paracellular space.To provide the molecular basis for this syndrome, we generated 3D models of claudin-10b, a 4-helix bundle. The direct effects of the p.S131L mutation in claudin-10b are a structural destabilization of the 4-helix bundle. In the cell, this should translate in the retention of the newly synthesized protein, given its inability to fold. Addressing of the protein to the plasma membrane should thus be impaired.This prediction was verified experimentally: the WT protein was observed at the plasma membrane after transfection and the label appeared stronger when two adjacent cells were transfected. On the contrary, when cells were transfected with the claudin-10b mutant, the plasma membrane was not labeled and the intercellular space appeared without any fluorescence.	Proteins poster	Fundamental
P_Pr027	424	Olga Zanegina, Evgeniy Aksianov, Andrei Alexeevski, Anna Karyagina and Sergei Spirin	Olga Zanegina	Conserved DNA-protein contacts formed by TATA-box binding proteins	TATA-box binding proteins (TBPs) are components of multiprotein complexes known as TFIID1. These complexes take part in transcription initiation of many genes of Archaea and Eukaryota. TBPs are two-domain proteins, 179–187 amino acid residues in length. In transcription initiation, TBPs specifically bind promoter regions containing sequences 5'-TATAWAWN-3', known as TATA-boxes.At the time, 34 structures of TBP from seven organisms are solved. Among them 25 are in complexes with DNA and 29 contain water molecules. Contacts of TBP with the DNA are highly conserved both in different structures of the same protein and in complexes of proteins from different organisms. We found 22 amino acid residues that form conserved hydrogen bonds and hydrophobic clusters at the DNA–protein interface. We also investigated conserved water molecules, both on DNA–protein interface and on the surface of the unbounded protein. We annotated a functional role of residues participating in the recognition of the DNA. The analysis was performed using services of the database of structures of DNA–protein and RNA–protein complexes NPIDB (http://npidb.belozersky.msu.ru/). The contact distribution on the DNA–TBP interface is in correspondence with the quasi-symmetry of N- and C-terminal domains of the protein. Although most contacts are symmetric, the N-terminal domain is connected with the DNA more tightly forming additional direct and water-mediated hydrogen bonds.As an example of possible applications of our results we predicted possible contacts with DNA of a homologous protein, TBP-L1, whose 3D structure is not available at the moment.	Proteins poster	Fundamental
P_Pr028	851	Václav Mareška and Vojtěch Spiwok	Václav Mareška	Development of the new pharmacophore model: test screening of inhibitors for COX-2 and KAT II	Using pharmacophores becomes an increasingly popular for the searching of new drugs. In comparison with traditional methods, pharmacophore models allow to be fast and efficient tool for virtual screening of large compound databases. Generally, pharmacophore models quantitatively characterize compounds by transformation of their structural characteristics into collective variables. This creates molecule "fingerprints" that can be easily compared. We have tried to design and implement the new pharmacophore model based on: CATS, SQUID and LIQUID models. We have used this model for screening of over 39 million compounds from ZINC database. Nowadays, we test the model for finding of new cyclooxygenase-2 (COX-2) and kynurenine aminotransferase II (KAT II) inhibitors with the same or even better biological activity compared to already known inhibitors. Together with docking calculations, we will test the pharmacophore model for screening another databases, searching new active molecules and will try to improve performance or efficiency of the model.	Proteins poster	Health
P_Pr029	601	Kenji Etchuya and Yuri Mukai	Kenji Etchuya	Environment Factor Depending on Each Sugar Type around O-glycosylation Sites in Mammalian Proteins	Glycosylation is a major post-translational modification and is important for protein folding, function, and enzyme activity. In O-glycosylation, motif residues (usually Ser or Thr) are modified by various kinds of sugars due to each glycosyltransferase in the Golgi body. The resulting sugars each promote a specific biological function and play different roles in living cells. Analysis of each sugar type will enable correlations between sugar type and biological function to be clarified. However, the characterization of the protein's primary sequences around each sugar type was weak and lacked consistency. An analysis of the environmental factors surrounding each sugar type is necessary to clarify the interaction between the glycosyltransferase and glycoprotein. Therefore, the environmental factors, composed of amino acids, were analyzed in this study.The sequence and structural data from mammalian proteins that undergo O-glycosylation was extracted from the UniprotKB/Swiss-Prot 2015_03 and the Protein Data Bank (PDB) release 2015_03, respectively. The physicochemical environment constructed by amino acids around the O-glycosylation sites was investigated by analyzing the amino acid propensities depending on each sugar type within a unit ball in which center was an O-glycosylation site. The propensity of the amino acids was calculated and compared between each sugar types. Significant aromatic residues were found around each sugar, and the correlation between aromatic residues and sugar chains was analyzed. The environmental factors for each sugar type were discussed in this study.	Proteins poster	Fundamental
P_Pr030	355	Kliment Olechnovic and Ceslovas Venclovas	Kliment Olechnovic	Estimation of protein structure quality using contact areas derived from the Voronoi tessellation of atomic balls	In the absence of experimentally determined protein structure many biological questions can be addressed using computational structural models. However, the utility of protein structural models depends on their quality. Therefore, the estimation of both global quality and the quality of local regions of predicted structures is an important and as yet unsolved problem. One of the popular approaches to this problem is the use of knowledge--based statistical potentials. Such methods typically rely on the statistics of distances and angles of residue--residue or atom--atom interactions collected from experimentally determined structures. We present VoroMQA ("Voronoi diagram--based Quality Assessment"), a new method for the estimation of protein structure quality. Our method combines the idea of statistical potentials with the advanced use of the Voronoi tessellation of atomic balls. The new method uses contact areas instead of distances for describing and seamlessly integrating both explicit interactions between protein atoms and implicit interactions of protein atoms with solvent. In addition, VoroMQA utilizes the Voronoi tessellation of balls to describe the orientation of contacts. The method produces scores at atomic, residue and global levels, all in the fixed range from 0 to 1. Also, due to its design, our method evaluates structures of protein complexes as efficiently as monomeric structures. The latest version of VoroMQA was tested on the CASP11 data: the results showed that our method generally performs better than the other available methods using knowledge--based statistical potentials. The software implementation of VoroMQA is freely available as a standalone application and as a web--server.	Proteins poster	Fundamental

P_Pr031	597	Maciej Pajak, Clive R. Bramham and T. Ian Simpson	Maciej Pajak	Exploring spatio-temporal landscape of post-synaptic proteome diversification and functionalisation	Evolution of the post-synaptic proteome (PSP) can be traced back to primitive organisms that lack nervous systems and is thought to be responsible for the emergence of finely-tuned neural system function and behaviour in complex organisms, however these studies have only assessed evolution at the whole protein level. We present an evolutionary analysis of 1461 proteins, the complete human PSP as identified experimentally by Bayes et al. (2011). We focus on selected protein families and complexes, but also analyse the entire set of proteins in search for general patterns of selection spanning multiple subsets of post-synaptic proteins. Our custom analysis framework uses an integrative approach to study selection pressure, aggregating information inferred from models of branch, site, and branch-site selection which allow detection of previously overlooked signals of active diversifying pressure.Firstly, we evaluate the spatial distribution of selection pressure at single amino acid resolution and interpret these in relation to the location of functional domains and post-translational modification sites uncovering domain-level signatures of diversification and revealing strong candidates for downstream functional studies. Secondly, we use bootstrap clustering of PSP elements by their branch-by-branch selection pressure profiles to identify with high confidence distinct temporal patterns of episodic diversification shared by groups of proteins. We map these back to key divergence points in the tree of life allowing a detailed explanation of the rapid development of complex neural function in organisms such as primates, complementing and extending earlier hypotheses.	Proteins poster	Fundamental
P_Pr033	509	Eugenia Polverini, Ilaria Menozzi and Rodolfo Berni	Eugenia Polverini	HIGH STRUCTURAL AND FUNCTIONAL CONSERVATION BUT DIFFERENT LIGAND UPTAKE: THE ROLE OF THE HYDROPATHY PROFILE OF THE PROTEIN SURFACE	Cellular Retinol-binding Proteins (CRBP) type I and II are beta-barrel proteins that show very high structural conservation in spite of a moderately low sequence identity and a different tissue distribution. These retinol carriers play role in the maintenance of vitamin A homeostasis, but exhibit a different affinity for the ligand (100 folds higher for CRBP-I). However, the binding site of the two isoforms is highly conserved.The mechanism of ligand uptake was investigated by means of molecular dynamics simulations, initially positioning the ligand outside the protein. For both CRBPs, the portal region formed by alfa helix II and the two loops between CD and EF strands is involved in the uptake, with a partial unfolding of the helix II. Nevertheless, a different distribution of polar and hydrophobic residues clusters at the surface of the two proteins, in particular at the barrel lid made by helix I and II, favored two different entrance pathways. In CRBP I, the retinol enters the binding cavity through a hydrophobic passage between alpha helix II and CD and EF loops, while in CRBP II the ligand, driven by a few polar interactions, sinks in the hydrophobic region between the two alpha helices. Then, in both cases, several polar residues interacting with OH-group, attract the retinol deeply inside the binding cavity. Therefore, even if the retinol uptake involves the same region, that covers the binding pocket and is intrinsically flexible, the ligand finds the better entrance pathway according to the hydropathy features of the protein surface.	Proteins poster	Health
P_Pr034	797	Tamás Langó, Gergely Róna, Éva Hunyadi-Gulyás, Lilla Turiák, Julia Varga, László Dobson, Nóra Kucsma, György Váradi, János Molnár, László Drahos, Beáta G. Vértessy, Katalin F. Medzihradszky, Gergely Szakács and Gábor E. Tusnády	Tamás Langó	High througput experimental method to improve topology prediction of transmembrane proteins	AbstractTransmembrane proteins play a crucial role in signaling, ion transport, nutrient uptake, as well as in maintaining the dynamic equilibrium between the internal and external environment of cells. Despite their important biological functions and abundance, less than 2% of all determined structures are transmembrane proteins. Given the persisting technical difficulties associated with high resolution structure determination of transmembrane proteins, additional methods, including computational and experimental techniques remain vital in promoting our understanding of their structures, functions and interactions. The topology of transmembrane proteins defines the sequential position and orientation of transmembrane segments and the loops connecting them relative to the inner or outer sides of the membrane. The accuracy and reliability of in silico topology prediction algorithms can be significantly improved by incorporating experimental data as constraints. Therefore, generating such topology data could expedite structural modeling of transmembrane proteins. Here we report a novel, highly optimized high-throughput method for the generation of reliable experimental topology data for transmembrane proteins. Identification of covalently labeled cell surface amino acids by LC/MS/MS allowed the identification of extracellularly located protein segments, which were implemented in an improved computational method to provide accurate and reliable topology models for hundreds of human transmembrane proteins.ReferencesDobson, L., Reményi, I., and Tusnády, G. E. (2015) The human transmembrane proteome. Biol. Direct 10, 31	Proteins poster	Fundamental
P_Pr035	855	Zoran Sucur and Vojtech Spiwok	Zoran Sucur	Homology Modeling and Funnel Metadynamics in the study of oxytocin binding to its GPCR receptor	After being released from neurohypophyseal neurons, in the target tissues oxytocin binds to its GPCR receptor, which has not been studied in detail, yet. G-protein-coupled receptors (GPCRs) belong to very diverse and numerous receptor family, and are involved in vital cell signaling pathways. Different GPCR templates were used for homology modeling, and the best results were obtained for models based on orexin and adenosin a2a receptors. Using Schrodinger software, multiple stable conformations of oxytocin have been identified. In addition, we performed the docking of this hormone to the GPCR model receptor. Further studies of the oxytocin binding modes and its conformational changes upon binding to receptor were performed using Funnel metadynamics, which has proved to be a good technique used for enhancing the exploration of the ligands target binding site.The project was supported by Ministry of Education, Youth and Sports (COST action GLISTEN, CM1207, LD14133, Specific University Research MSMTNo. 20/2014, 21/2014 and 20/2015) and Czech Science Foundation (15-17269S). Computational resources were provided by the MetaCentrum under the program LM2010005 and the CERIT-SC under the program CentreCERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144.	Proteins poster	Health
P_Pr036	571	Tatsuki Kikegawa, Hiromu Sugita, Ryohei Nambu, Noritaka Kato and Yuri Mukai	Tatsuki Kikegawa	Identification of the subcellular localization factors of transmembrane proteins	Transmembrane proteins are typical internal membrane proteins spanning biomembranes including the endoplasmic reticulum (ER), Golgi, and plasma membranes. Their functions are essential to maintain homeostasis via signal transduction, membrane transport, and energy production. Their transmembrane regions usually consist of 10–30 hydrophobic amino acids, which are known as ER-targeting signals called signal-anchors. However, the mechanisms of transmembrane protein localization from ER to other organelles have not been elucidated. Understanding the mechanisms of protein subcellular localization is believed to be crucial for treatment of the incurable diseases resulting from erroneous subcellular localization.In this study, the amino acid propensity around signal-anchors was calculated to elucidate subcellular localization mechanisms of single-pass transmembrane proteins. The transmembrane protein dataset was classified into four groups: plasma membrane proteins, ER membrane proteins, Golgi membrane proteins, and proteins containing KDEL (KKXX) ER retention motif. The results of this analysis suggested that the amino acid propensity was related to the localization mechanisms because a remarkable bias of amino acid propensities was found in each group. These results were applied for predicting protein subcellular localization. The discrimination parameters of each group were evaluated by artificial GFP-signal-anchor fusion proteins. The GFP fusion proteins were expressed in HeLa cells, and the subcellular localization of these proteins was observed by a confocal laser fluorescence microscope.	Proteins poster	Fundamental
P_Pr037	697	Tomas Bastys, Vytautas Gapsys, Nadezhda Doncheva, Hauke Walter, Rolf Kaiser, Mario Albrecht, Bert Groot and Olga Kalinina	Tomas Bastys	Impact of point mutations on inhibitor affinity in HIV-1 protease	HIV (human immunodeficiency virus) protease is one of major targets of antiretrovital therapy, targeted by protease inhibitors (PIs). Through point mutations in protein sequence, a virus population acquires resistance to drugs. Effect of mutation on drug binding can be described in terms of change in drug binding free energy (ΔΔG) or change of the protein half maximal inhibitory concentration, also called resistance factor (RF). Predicting effect of a specific mutation on drug binding is essential for optimizing patient therapy. And understanding the specific mechanisms that influence affinity of the protein towards a PI is of important for development of novel drugs.In this work, we analysed a set of different combinations of known major resistance-associated mutations in HIV protease in complex with different PIs, for which experimental ΔΔG or RF measurements were available. For each combination, molecular dynamics simulations were used to calculate ΔΔG using BennettsAcceptance Ratio method. For a dataset of ten complexes we achieved a correlation coefficient of 0.81 between the theoretical and experimental values of ΔΔG. On a different dataset of eight protease-PI complexes where only RF measurements were available, combining information from ΔΔG calculations for complexes with the same mutations but different PIs, we were able to estimate values of RF, which were in most cases not significantly different from experimental measurements.Partial least-squares regression on molecular dynamics produced predictive models that were able to distinguish dynamics of wildtype and resistant proteases. These models illuminate different mechanisms that contribute to resistance against the PIs.	Proteins poster	Fundamental
P_Pr038	689	Maarten Reijnders, Vitor Martins Dos Santos and Peter Schaap	Maarten Reijnders	Improving functional annotation of microalgal proteins	Microalgae are promising organisms for the production of biobased compounds. However, to make the industrial production of these compounds competitive, we need to understand and improve the metabolic capabilities of microalgae [1]. The first step in understanding is a functional annotation of the proteins encoded in the genome. For a novel species, sequence similarity with proteins of known function from phylogenetic close-by model species is often used to transfer function. However, in absence of well-annotated close-by model species this is not a reliable way of assigning protein functions to microalgae.To reliably assign functions to microalgal proteins we have to go beyond the standard methods. We have designed a pipeline that utilizes multiple existing methods. Individual predictions are combined and compared by scoring for semantic similarities between the gene ontology terms obtained, followed by a machine learning algorithm over all the scores retrieved in the process.The performance of the pipeline on a test set of enzymes showed an improved true positive – false positive ratio compared to existing methods [2]. Additionally, compared to the same existing methods more proteins were annotated, with more annotations per protein.An additional benefit of this method is the modularity. In theory, any other function prediction method can be incorporated in the processes of this pipeline. This allows for continuous improvement of annotation performance.1. M.J.M.F. Reijnders, "Green genes: bioinformatics and systems-biology innovations drive algal biotechnology". Trends in Biotechnology 32.12:617-626, 2-14.2. M.J.M.F. Reijnders, "Algal omics: The functional annotation challenge". Current Biotechnology 4.4:457-463, 2015.	Proteins poster	Fundamental
P_Pr039	760	Eda Suku, Mattia Di Giacobbe, Behnoosh Bahadori, Stefano Capaldi, Mario R. Buffelli and Alejandro Giorgetti	Eda Suku	In silico deorphanization of the GPR3 receptor	Introduction: Alzheimer's Disease is a neurodegenerative disease (ND), characterized by loss of brain connectivity, memory and cognitive functions. Recently, G-protein coupled receptor 3 (GPR3) was identified as regulator of Aβ plaques through the β-arrestin 2 pathway1. GPR3 is an orphan receptor and a deep investigation of its function is still missing. Here we present the identification of two putative GPR3 endogenous ligands and structural insights into the binding pocket using state of the art techniques.Methods: Homology modeling and docking were carried out through the GOMoDo web-server2. The programs OMEGA3 and ROCKS3 were used to perform chemoinformatics searches on ZINC and Human Metabolome Databases.Results: GPR3 model was validated against experimental data on non-endogenous ligands: DPI4 and AF6439445. These molecules were used as starting compounds for chemoinformatics studies. Two endogenous ligands, i.e. beta-carboline and 1-methyladenine, present in different brain pathways and involved in neuronal damage, have been identified. Docking studies of these ligands allowed us to characterize residues putatively involved in receptor-ligand interaction.Conclusions: We aimed to in silico deorphanize GPR3 and characterize its binding cavity. We identified two endogenous ligands and several putative critical residues. Further investigations and experimental informations are being carried out to validate the results and to better characterize the GPR3 function.References:1Thathiah A. et al. Nature medicine (2013): 43-492Sandal M. et al. PLoS One (2013): e740923http://www.eyesopen.com/4Ye C. et al. Journal of Pharmacology and Experimental Therapeutics (2014): 437-4435Jensen T. et al. Bioorganic & medicinal chemistry letters (2014): 5195-5198	Proteins poster	Biotechnology
P_Pr041	628	Dinithi Sumanaweera and Dr. A. Shehan Perera	Dinithi Sumanaweera	In silico prediction of protein function for Saccharomyces Cerevisiae; an ensemble approach	Protein function annotation is vital for identifying disease causative factors and for solving mysteries behind biological system complexities. As manual annotation requires costly and laborious in-vitro methods, in-silico protein function prediction is preferred nowadays. According to literature, one in five yeast mitochondrial proteins are known to be human disease related. We present a weighted heterogeneous data ensemble to classify Saccharomyces Cerevisiae proteins under 'Mitochondrial Organisation' in Gene Ontology (GO). It consists of five euclidean-distance based nearest neighbour models and three affinity-based neighborhood models; utilizing protein properties data, four gene expression datasets and physical/genetic interactions. 239 current GO annotations and 3887 gold standard negative annotations from literature were used to train the base learners. The overall prediction is the weighted average of posterior probabilities outputted by the base models. The weights are determined by a Genetic algorithm (GA) for obtaining the optimal AUC value under ROC. All evaluations were performed using leave-one-out cross-validation for 10 samples; each containing all positive proteins and a random negative protein sample, with 1:1 class ratio. The optimal k parameter for nearest neighbour models was decided as 22 upon empirical results obtained by varying k from 1 to 25. The base models show a substantial level of disagreement with a mean Fleiss Kappa statistic of ~0.2816. The GA-weighted ensemble gives ~14.3% (from ~78.34% to ~89.57%) improvement of the best performing base classifier, whereas only ~13.1% improvement can be seen with an equal-weighted ensemble.	Proteins poster	Fundamental
P_Pr042	533	Fabrizio Pucci, Raphaël Bourgeas, Jean Marc Kwasigroch and Marianne Rooman	Fabrizio Pucci	In-silico prediction of protein thermal stability changes upon point mutations using HoTMuSiC	IntroductionThe ability to rationally modify proteins in order to increase their thermal stability is one of the main goals of protein design, which has interesting applications in a wide series of biomedical and biotechnological processes.We present a newly developed bioinformatics tool that, using as input the three-dimensional (3D) structure of the protein and, when available, its melting temperature (Tm), is able to predict rapidly and accurately the impact of amino acid substitutions on this temperature.MethodsThe key ingredients of our methodology are statistical potentials that are knowledge-driven mean force potentials (PMF) extracted from a dataset of experimentally resolved 3D protein structures. They are linearly combined using an artificial neural network (ANN) with sigmoid activation functions that depend on the solvent accessibility of the mutated residues. If the melting temperature of the protein is known, we use in addition temperature-dependent statistical PMFs that reflect the (melting)-temperature dependence of the amino acid interactions. They are combined using a triple-layer ANN, in which the activation functions of the first layer depend on the solvent accessibility, while those for the second layer are parabolic functions of the protein's number of residues and melting temperature.ResultsThe performance of our method is evaluated in 5-fold cross validation on a dataset of 1626 mutations and yields a root mean square deviation between predicted and experimental ΔTm's of about 4°C. The addition of evolutionary information to the model and the analysis of the relations between thermal and thermodynamic stability changes are also carefully discussed.	Proteins poster	Biotechnology Fundamental
P_Pr044	670	Nesrine Chakroun, Cheng Zhang and Paul Dalby	Nesrine Chakroun	Insights into the intrinsic Stability of a Therapeutic Fragment Antibody by Molecular Dynamics Simulations	Biopharmaceuticals or therapeutically relevant proteins have become one of the fastest growing parts of the pharmaceutical industry. These innovative molecules are more complex than conventional drugs and their processing is much more demanding.The analytical characterization of these new drugs is a fundamental step in the early prediction of their behavior in bioprocesses. This research project aims to develop a framework to improve candidate design and selection at early stages of development by establishing a set of critical analysis and identifying key properties (intrinsic and extrinsic) allowing the prediction of candidates behaviour in large-scale bioprocesses.Our multidisciplinary approach combines the computational analysis (sequence analysis, Molecular Dynamics simulations and docking) and the biophysical characterization of a set of Fragment antibody (Fab) mutants. In particular MD simulations were used to investigate the effects of pH, temperature and mutations in the stability of Fab. This allowed the identification of several key regions and residues in the stability of the molecule which were targeted experimentally to enhance candidate's stability. The effect of formulation was also investigated highlighting the role of electrostatics and salt bridges in Fab stability and folding. Additionally, aggregation kinetics studies were carried out at a wide range of temperature, pH and ionic strength allowing the determination of a model for Fab aggregation	Proteins poster	Health
P_Pr045	849	Gift Nuka, Simon Potter, Siew-Yit Yong, Maxim Scheremetjew, Alex Mitchell, Matthew Fraser and Rob Finn	Gift Nuka	InterProScan 5: Large scale protein function classification	InterPro (http://www.ebi.ac.uk/interpro/) is a freely available resource that is used to classify sequences into protein families and to predict the presence of important domains and sites. InterProScan (https://www.ebi.ac.uk/interpro/interproscan.html) is the underlying software application that allows both protein and nucleic acid sequences to be scanned against InterPro's predictive models (signatures), which are provided by the resource's member databases.Recently, both the Conserved Domain Database (CDD) and Structure-Function Linkage Database (SFLD) have joined InterPro as new member databases. InterProScan has been updated accordingly, incorporating CDD's curated models that use position specific scoring matrices (PSSMs) to represent protein domains, which tend to be more functionally specific than some of the models already used in InterPro. SFLD's hidden Markov models that offer structure-function mapping have also been incorporated. SFLD models allow evolutionary classification of related enzymes according to shared chemical functions to determine conserved active sites.Here, we present these recent developments and performance improvements to InterProScan.Optimisation in the pipeline filters and database query refinements have also resulted in improved throughput for large-scale protein sequence analysis and accelerated InterProScan domain searches by several orders of magnitude.	Proteins poster	Biotechnology Fundamental
P_Pr046	459	Sirawit Ittisoponpisan, Eman Alhuzimi, Michael Sternberg and Alessia David	Sirawit Ittisoponpisan	Landscape of pleiotropic proteins causing human disease: structural and system biology insights.	Pleiotropy is the phenomenon by which the same gene can result in multiple phenotypes. Pleiotropic proteins are emerging as important contributors to both rare and common disorders. Despite this, little is known on the pathogenetic mechanisms underlying pleiotropy and the characteristic of pleiotropic proteins.We analysed disease-causing proteins reported in Uniprot and observed that 12% are pleiotropic (mutations in the same protein cause more than one disease). Pleiotropic proteins were more likely to be essential and have a higher number of interacting partners compared to non-pleiotropic proteins. Moreover, significantly more pleiotropic than non-pleiotropic proteins contained at least one intrinsically long disordered region of over 50 residues in length (p<0.001). Pleiotropic proteins were enriched in deleterious mutations and rare polymorphisms, but not in common polymorphisms. Deleterious mutations occurred mainly in structurally ordered regions. Deleterious mutations occurring in structurally disordered regions were more commonly found in pleiotropic, rather than non-pleiotropic proteins. Finally, we observed that proteins involved in the pathogenesis of neoplasms, neurological and circulatory diseases, and congenital malformations were more likely to be pleiotropic, whereas proteins causing endocrine and metabolic disorders were more likely non-pleiotropic.In conclusion, this study suggests that pleiotropic proteins represent a biologically different class of proteins compared to non-pleiotropic proteins and are an important contributor to human disease. This study provides a better understanding of pleiotropic proteins and their genetic variants, which could greatly aid in the interpretation of genetic studies and drug design.	Proteins poster	Fundamental

P_Pr047	604	Chloé Dequeker, Raffaele Raucci, Elodie Laine and Alessandra Carbone	Chloé Dequeker	Large scale analysis of protein interactions	Protein-Protein Interactions (PPI) are at the heart of processes and their understanding is of utmost importance to facilitate drug design and characterize the mechanisms underlying certain diseases.In this context, our team works on the Help Cure Muscular Dystrophy (HCMD) project, whose aim is to uncover new pathways responsible for the muscular dystrophy by developing a discriminating power over the interacting and non interacting complexes. A complete cross-docking (CCD) has then been realised over 2200 proteins with the help of the World Community Grid (WCG), generating more than 900 billions conformations over 2.5 millions different complexes.In parallel of these computations, our team developed a new method JET² to predict interacting surfaces at large scale (Laine and Carbone, 2015), using different criteria based on residue conservation, physico-chemical properties and the geometrical aspect of the protein structure. JET² has been run over more than 20.000 different chains for which a PDB structure is available.We present new ways to link the two different problems of the prediction of protein interaction sites and the discrimination of interacting partners through optimisation of JET² prediction as well as using different scoring methods. Our work also sheds some light on interactions of proteins with multiple partners, which will be a principal factor in the analysis of the HCMD results.	Proteins poster	Biotechnology
P_Pr048	679	Nicholas Furnham, Natalie Dawson, Syed Rahman, Janet Thornton and Christine Orengo	Nicholas Furnham	Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies	Enzymes, as nature's catalysts, are crucial to life. How they have evolved to undertake their different chemical reactions is of great interest to a wide range of biological disciplines. Over 100 years of detailed biochemistry studies combined with the large volumes of sequence and protein structural data now available, means we are able to perform large-scale analyses to address this question. Using sophisticated tools relating sequences and structures across thousands of genomes though phylogenetic analysis and novel measures of functional similarity we have compiled information on all experimentally annotated changes in enzyme function within 379 structurally defined protein domain superfamilies, linking the changes observed in functions during evolution to changes in reaction chemistry. Using analysis of modifications in reaction chemistry and enzymes active sites we have observed that some superfamilies have changed the reactions they perform without changing catalytic machinery. In others large changes of enzyme function have been brought about by significant changes in catalytic machinery. Interestingly, in some superfamilies relatives perform similar functions but with different catalytic machineries. The collected data and analysis has been developed into a community resource (www.funtree.info). This analysis highlights characteristics of functional evolution across a wide range of superfamilies. It also provides insights that will be useful in predicting the function of uncharacterized sequences as well as the design of new synthetic enzymes.	Proteins poster	Fundamental
P_Pr049	499	Daniele Raimondi, Andrea Gazzo, Marianne Rooman, Tom Lenaerts and Wim Vranken	Daniele Raimondi	Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects	There are many predictors capable of identifying the likely phenotypic effects of single nucleotide variants (SNVs) or short in-frame Insertions or Deletions (INDELs) on the increasing amount of genome sequence data. Most of these predictors focus on SNVs and use a combination of features related to sequence conservation, biophysical, and/or structural properties to link the observed variant to either neutral or disease phenotype. Despite notable successes, the mapping between genetic variants and their phenotypic effects is riddled with levels of complexity that are not yet fully understood and that are often not taken into account in the predictions, despite their promise of significantly improving the prediction of deleterious mutants.We present DEOGEN, a novel variant effect predictor that can handle both missense SNVs and in-frame INDELs. By integrating information from different biological scales and mimicking the complex mixture of effects that lead from the variant to the phenotype, we obtain significant improvements in the variant-effect prediction results. Next to the typical variant-oriented features based on the evolutionary conservation of the mutated positions, we added a collection of protein-oriented features that are based on functional aspects of the gene affected. We cross-validated DEOGEN on 36 825 polymorphisms, 20 821 deleterious SNVs, and 1038 INDELs from SwissProt. The multilevel contextualization of each (variant, protein) pair in DEOGEN provides a 10% improvement of MCC with respect to current state-of-the-art tools. The software and the data presented is available at http://ibsquare.be/deogen.	Proteins poster	Health
P_Pr051	711	Rashmi Hazarika and Vera van Noort	Rashmi Hazarika	Network evolution of MADS-domain protein interaction network	In protein-protein interaction networks, the nodes symbolize interacting proteins while the edges relate to the physical interactions between these proteins. A gain of an edge between two nodes denotes the appearance of a new functionality while losing a subset of their initial interactions symbolizes functional divergence as when duplicate copies of a protein evolve to bind different interaction partners. In this study, we chose the MADS-domain transcription factors which rely on coiled coil interactions. The study of these proteins would help us understand plant evolution better, as proliferation of these proteins and successive diversification of protein functions may explain how modern day Angiosperms evolved. The ancestral nodes for the MADS-domain proteins were estimated, resurrected and their interactions experimentally verified before and after whole genome duplication. The Yeast-2-Hybrid system was used to define the protein-protein interactions of 9 resurrected ancestral MADS-box gene lineages (SEP3, SEP1/2/4, AP1, AP3, PI, AG, STK, SVP and SOC1) before γ triplication. After the γ triplication event, some of these genes duplicated or triplicated, giving rise to a total of 17 post γ triplication MADS-box genes. Using random networks, the various events of Whole Genome Duplications (WGDs) and network dynamics were simulated on an evolutionary time scale using age estimates from literature and the empirical probabilities measured from the actual Y2H networks. The results were directly compared to extant networks of Arabidopsis thaliana and Solanum lycopersicum. We observed a scale free and a highly modular network topology in the simulated networks.	Proteins poster	Fundamental
P_Pr052	775	François Ancien, Maxime Godfroid, Georges Coppin, Fabrizio Pucci and Marianne Rooman	François Ancien	Neural network-based predictions of deleterious human variants derived from protein structures and free energy estimations	Many predictors have been developed to predict the deleteriousness of mutations in the human exome, often exclusively based on the protein sequences and their evolutionary features. However, the explanatory power of these methods in terms of the physical effect that the mutations have on the molecular phenotype is usually quite limited – although such insight is a prerequisite for the development of personalized treatments. Here we analyzed what relevant information the protein structure and stability can add in this context. For that purpose we used a dataset of human variants that are annotated as deleterious or neutral in proteins for which the 3-dimensional structure is available. In a first step we estimated the thermodynamic and thermal stability changes caused by the mutations, using the PoPMuSiC and HoTMuSiC programs, which use artificial neural networks (ANN) and linear combinations of statistical mean-force potentials. These stability changes upon mutations were shown to correlate significantly with the deleteriousness of the mutations: the more destabilizing, the more deleterious, with a balanced accuracy of about 0.6. In a second step, we built on PoPMuSiC and HoTMuSiC to develop a new predictor that focuses on deleteriousness prediction. We implemented different types of ANNs, and in particular probabilistic and echo state networks, in an attempt to catch all the complex information contained in the dataset and to improve the prediction performances. The highest scores, estimated in cross validation, are significantly higher than that of PoPMuSiC and HoTMuSiC and exceed 0.7. This performance is comparable to that of purely evolutionary-based methods, with however the advantage of a better understanding of the biophysical effects that cause the disease.	Proteins poster	Fundamental Health
P_Pr055	381	Olga Voitenko, Andi Dhroso, Anna Feldmann, Dmitry Korkin and Olga Kalinina	Olga Voitenko	Patterns of amino acids conservation in human and animal immunodeficiency viruses	Motivation: Due to their high genomic variability, RNA viruses and retroviruses present a unique opportunity for detailed study of molecular evolution. Lentiviruses, with HIV being a notable example, are one of the best studied viral groups: hundreds of thousands of sequences are available together with experimentally resolved three-dimensional structures for most viral proteins. In this work, we use these data to study specific patterns of evolution of the viral proteins, and their relationship to protein interactions and immunogenicity.Results: We identify extremely conserved and extremely variable clusters of amino acid residues on the surface of proteins from HIV and other animal immunodeficiency viruses. These clusters turn out to be located on the interaction interfaces of viral proteins with other proteins, nucleic acids or low molecular-weight ligands, both in the viral particle and between the virus and its host. In the immunodeficiency viruses, the interaction interfaces are not more conserved than the corresponding proteins on average, and we show that extremely conserved clusters coincide with protein-protein interaction hotspots, predicted as the residues with the largest energetic contribution to the interaction. Extremely variable clusters have been identified here for the first time. In the HIV-1 envelope protein gp120, they overlap with known antigenic sites. These antigenic sites also contain many residues from extremely conservedclusters, hence representing a unique interacting interface enriched both in extremely conserved and extremely variable clusters of residues. This observation may have important implication for antiretroviral vaccine development.	Proteins poster	Fundamental Health
P_Pr056	458	Rosalba Lepore, Agnieszka Obarska-Kosinska, Alfredo Iacoangeli and Anna Tramontano	Rosalba Lepore	PepComposer: computational design of peptides binding to a given protein surface	There is a wide interest in designing peptides able to bind to a specific region of a protein with the aim of interfering with a known interaction or as starting point for the design of inhibitors.Structure-based strategies usually consists in analysing the interacting region from a complex of the target protein with a protein or a peptide trying and identifying a contiguous 'peptide like' region of the partner to be used as starting point. However, if no complex structure is available, one has to recur to de novo design methods and therefore needs to select an appropriate backbone, optimize its relative orientation with respect to the target protein and its sequence (1).To simplify and streamline this process, we developed PepComposer, a computational pipeline for the design of protein-binding peptides that only requires the target protein structure and an approximate definition of the binding site as input. We first select appropriate backbones from monomeric proteins based on previous observations (1) and use a Monte Carlo procedure to design optimal sequences for the identified peptide scaffolds. Peptides are then selected according to the predicted binding energy.PepComposer is fully automatic, available as a web server (http://biocomputing.it/pepcomposer/webserver) and can effectively reproduce known protein-peptide interactions (2).1. Vanhee P., et al. Computational design of peptide ligands. Trends Biotechnol. 2011;29:231-239.2. Obarska-Kosinska A., et al. PepComposer: computational design of peptides binding to a given protein surface. Nucl. Acids Res. (2016) doi: 10.1093/nar/gkw366	Proteins poster	Fundamental
P_Pr057	636	Emilie Neveu, David Ritchie, Petr Popov and Sergei Grudinin	Emilie Neveu	PEPSI-Dock : A Detailed Data-Driven Protein-Protein Interaction Potential Accelerated By Polar Fourier Correlation	Docking prediction algorithms aim at finding the native conformation of a complex of proteins, knowing their unbound structures. Most of the existing predictions the results of a combination of sampling and scoring methods, adapted to different scales. Here we present PEPSI-Dock (Polynomial Expansion of Protein Structures and Interactions for Docking), which improves the first stage of the docking pipeline, being more accurate at the beginning of the docking process, which thus sharpen up the final predictions. Indeed, the method benefits from the precision of a very detailed data-driven model of the binding free energy used with a global and exhaustive rigid-body search space. While being accurate, our computations are among the fastest ones by virtue of the sparse representation of the pre-computed potentials and FFT-accelerated sampling technics. PEPSI-Dock runs in 5-20 minutes on a modern laptop and can be easily extended to other types of interactions.	Proteins poster	Health
P_Pr059	356	Thanh Binh Nguyen and M.S. Madhusudhan	Thanh Binh Nguyen	Prediction of polyproline type II helices receptors	Polyproline type II helices (PPII) are a less common secondary structure of proteins than α helix and β sheet. There is no internal backbone hydrogen bond interaction in this conformation. As a result, the carbonyl and amide groups along the PPII helices prefer to make intermolecular interaction. And hence, PPII mediates many protein-peptide or protein-protein interactions in signalling pathway, immune response, cell-cell communication. There is an abundance amount of proteins which are well-known to bind PPII, including MHC, SH3, WW, EVH1, profilin and GYF domains. These PPII bound proteins share geometry and biophysical features. Using the knowledge from the known PPII-bound families the aim of this study is to detect the PPII binding site in a query protein. This approach could help to identify a new PPII-bound protein.	Proteins poster	Fundamental
P_Pr060	648	Thach Nguyen and Michael Habeck	Thach Nguyen	Probabilistic model for segmentation of protein structures	Motivation: Large-scale conformational changes in proteins are implicated in many important biological functions. These structural transitions can often be rationalized in terms of relative movements of rigid domains. There is a need for objective and automated methods that identify rigid domains in sets ofprotein structures showing alternative conformational states.Results: We present a probabilistic model for detecting rigid-body movements in protein structures. Our model aims to approximate alternative conformational states by a few structural parts that are rigidlytransformed under the action of a rotation and a translation. By using Bayesian inference and Markov chain Monte Carlo sampling, we estimate all parameters of the model, including a segmentation of the protein into rigid domains, the structures of the domains themselves, and the rigid transformations thatgenerate the observed structures. We find that our Gibbs sampling algorithm can also estimate the optimal number of rigid domains with high efficiency and accuracy. We assess the power of our method on severalthousand entries of the DynDom database and discuss applications to various complex biomolecular systems.Availability:The Python source code for protein ensemble analysis is available at:https://github.com/thachnguyen/motion_detection/.	Proteins poster	Biotechnology Fundamental
P_Pr061	793	István Reményi, László Dobson and Gábor E. Tusnády	István Reményi	Profile modeling and multiple sequence alignment of transmembrane proteins	Transmembrane proteins are involved in energy production, signal transduction, cell-cell interaction, cell-cell communication. They are frequent targets for pharmaceuticals, therefore knowledge about their properties and structure is crucial. However, less than 2% of all determined protein structures belongs to transmembrane proteins, thus computational approaches have to be utilized for topology prediction and structure modelling.Analyzing a protein may begin with searching for homologous sequences, namely for entities with statistically significant similarity. There are several methods for homology detection, among which profile modeling exceeds in terms of capability of capturing highly conserved regions. As a result, more accurate alignments can be created from any unaligned set of sequence, and more thorough analysis can be performed. Hidden Markov Models have already been applied to homology search, but their training problem can be considered NP hard. In such a model, to handle the high number of variables, the different training approaches are either local optimization techniques which incorporates task-specific additional (e.g. structure) information, numerous tuning parameters, or just based on an otherwise determined multiple sequence alignment.Previously we have launched the Human Transmembrane Proteome database, which contains topology and structure information about the human α-helical transmembrane proteins. Our aim is to find a general optimization technique to build TM profile model(s) with, to represent the whole proteome to create a starting point for further investigations.The Human Transmembrane ProteomeLászló Dobson, István Reményi and Gábor E. Tusnády (2015)Biology Direct, 10:31	Proteins poster	Fundamental
P_Pr062	720	Diego Alonso-Martinez and Peter Dimaggio	Diego Alonso-Martinez	Profiling the methylome targets of histone lysine methyltransferases	Histone post-translational modifications (PTMs) are epigenetic marks critical in the regulation of gene expression that are regulated by various classes of enzymes including histone lysine methyltransferases (HKMTs). HKMTs catalyse the transfer of a methyl group from S-adenosyl methionine (SAM) to a specific histone lysine target. Due to their overlapping but non-redundant functions, there is current no way to decisively assess which HKMT is responsible for an observed methylation event. This lack of understanding prevents the development of more specific treatments for epigenetic, PTM-related diseases, such as cancer.In this work we propose to engineer the first cellular HKMT methylome profiling assay by combining the classical "bump and hole" approach with extensive bioinformatics and computational modelling of enzyme-cofactor candidate pairs prior to mass spectrometry (MS) validation. A multiple sequence alignment of all post-SET domain containing HKMT sequences alongside a detailed analysis of the crystallographic structure of G9a-SAH (PDB: 2O8J) identified a bulky residue in G9a that could be mutated to create a "hole" for a modified SAM cofactor. Protein folding simulations using Rosetta were used to assess structural validity of the mutant candidates, and corresponding SAM analogues with a "bump" to match the "hole" were designed based on the generated models.In vitro experiments utilising recombinant G9a with our engineered SAM analogues have demonstrated suitable cofactor selectivity against endogenous HKMTs, which supports the feasibility of this approach. This study highlights the importance of computational simulations in the development of more accurate assays to characterise the methylome targets of HKMTs.	Proteins poster	Health
P_Pr063	718	Alexander Smolyakov, Ilya Altukhov, Sergey Gavrilov, Ivan Butenko, Olga Pobeguts, Ilya Kublanov and Dmitry Alexeev	Alexander Smolyakov	Quantitative profiling of membrane-associated proteins in Melioribacter roseus P3M-2	Melioribacter roseus P3M-2 is recently discovered gram-negative bacteria characterized as a new species of Melioribacteraceae family within the Ignavibacteriae phylum. The complete sequence of the M.roseus genome was recently released and showed presence of genes involved in adaptation to the extreme conditions.Currently proteomic studies widely use mass-spectrometry analysis methods. These methods are mostly applied for investigating protein-protein interactions and post-translational modifications, as well as organism proteome inventory; they also offer strategies for quantitative and qualitative proteomic and proteogenomic analysis.In this study cells of M.roseus P3M-2 were grown by aerobic respiration and maltose fermentation at strictly anaerobic conditions. Each culture was grown in three biological replicates, independently, to give a total of six specimens. We carried out an in-depth quantitative proteogenomic analysis of M.roseus P3M-2 based on shotgun LC-ESI-MS/MS data. In total 199,894 tandem mass spectra were obtained. The 1,127 proteins were identified by two and more peptides across all experiments. The quantitative proteome analysis revealed 239 significantly different membrane-associated proteins between cells grown at aerobic and anaerobic conditions.Proteins were classified according to the Gene Ontology annotations BP, CC and MF. 28 GO terms were significantly enriched. KEGG pathway analysis was performed to determine the functional interactions of differentially expressed proteins between aerobic and anaerobic conditions. Wilcoxon test were performed for directed and and undirected regulation. 4 maps were significantly enriched (p.value < 0.05).Using proteogenome approach we have defined 3 novel protein-coding genes, coordinates of 9 genes were reannotated and peptides located in 10 pseudogenes were identified.	Proteins poster	Fundamental

P_Pr064	728	Francesca Nadalin and Alessandra Carbone	Francesca Nadalin	Residue propensity and local geometry of the interface contacts define the specificity of protein-protein interactions	Obtaining structures of protein complexes experimentally requires a lot of effort. For this reason, reliable methods for modeling PPI in silico are envisaged. Protein docking experiments output a long list of possible conformations, thus properly scoring them is essential for further studies.Previous works showed the applicability of pair potentials to the scoring of docking decoys [Moal 2013].We define new pair potentials as the contribution of two terms: the one derived by the observed contact distribution at the interface, the other representing the likelihood of residues to be located at the interface [Negi 2007].Pair potentials are computed on 230 experimental structures [Vreven 2015] and we call our method CIPS, Combined Interface Propensity for decoy Scoring. CIPS is compared to other methods, respectively based on pair potentials [Glaser 2001, Pons 2013, Mezei 2015] and atomic potentials [Krissinel 2007,Pierce 2007]. CIPS turns out to outperform all tested methods on decoys sets obtained both with all-atom and coarse-grain rigid docking. Further improvement is observed when decoys scoring is done with a combination of CIPS and atomic potentials.Our method is fast, accurate, and robust upon decrease in the level of detail of protein structure representation. This allows important application inthe promising, but not yet deeply explored, field of large-scale cross docking.	Proteins poster	Fundamental
P_Pr065	421	Gabriele Orlando, Daniele Raimondi, Tom Lenaerts and Wim Vranken	Gabriele Orlando	RIGAPOLLO, A HMM-SVM BASED APPROACH TO SEQUENCE ALIGNMENT	Reliable protein alignments are a central problem for many bioinformatics tools, such as homology modeling. Over the years many different algorithms have been developed and different kinds of information have been used to align very divergent sequences. Here we present a pairwise alignment tool, called Rigapollo, based on pairwise HMM-SVM, which can include different types of information in the alignment process. The model is composed by 7 states: a M (match), and six G (gap) states, three for the first sequence and three for the second one. For each amino acid in the sequences, we define an N-dimensional feature vector to describe it. That vector can be defined using any kind of information, from evolutionary (i.e. PSSM) to dynamics predictions. While standard pairwise HMMs require the definition of a finite and discrete alphabet of observable states, our model works directly using these feature vectors (that can be both orthonormal or not orthonormal). We define the emission probability using a SVM trained to discriminate matches from miss-matches and gaps from non-gaps positions. We tested our algorithm on two benchmark datasets of very divergent proteins, one based on Balibase and the other based on Sabmark. Rigapollo improves the quality of the predicted alignments from 20 to 38% respect to the most used state of the art pairwise alignment tools.	Proteins poster	Fundamental
P_Pr066	488	Qingzhen Hou, Paul De Geest, Wim Vranken, Jaap Heringa and K. Anton Feenstra	Qingzhen Hou	Seeing the Trees through the Forest: Sequence-based Homo- and Heteromeric Protein-protein Interaction sites prediction using Random Forest	Motivation: To fulfil biological functions, proteins bind to their partners via specific amino acids.Investigation of the properties and sequential information of these residues is important to reveal the mechanisms of protein-protein interactions and protein functions. These properties, derived from the interacting amino acids at sequence level, are usually exploited as features for machine learning methods to predict protein interacting positions. In this paper, we include two novel features (backbone flexibility and Sequence Specificity) predicted from sequences for protein interface prediction and evaluate the importance of different features using Random Forest.Results: We observe that there is no single sequence feature which enables to pinpoint interacting sites. However, combination of different properties does help the interface prediction. After selecting and integrating multiple features, we developed a Random Forest predictor which is able to distinguish interface and other residues with AUC of ROC plot at 0.72 in our homomeric test-set, which is better than other sequence-based methods. Moreover, when applied to identify interfaces of an independent heteromeric dataset, our method performs slightly better than the best sequence-only predictor. Thus, our predictor trained on homodimeric proteins can not only predict homodimeric interfaces, but is also able to locate interface residues in the heterodimers which suggested that our predictor captures the common properties of both homodimer and heterodimer interfaces.	Proteins poster	Fundamental
P_Pr067	521	Wim Vranken, Daniele Raimondi, Gabriele Orlando and Rita Pancsa	Wim Vranken	Sequence-based prediction of protein early folding residues	We present EFoldMine, a novel protein sequence-based predictor of early folding regions based on the Start2Fold database and the DynaMine predictions of protein backbone rigidity. EFoldMine reaches an AUC of 0.808 for detecting early folding residues, over a 27-fold set of 30 proteins. We observe that first, amino acids involved in amyloid formation have a higher tendency to fold early according to our predictions. Second, there is a weak correlation with folding speed, especially for two-state folders. Third, the predictions especially pick up residues that form extensive contacts in the folded conformation of the protein, less so than residues that become buried. Finally, residues with high covariance signals in the PSICOV contact prediction dataset tend to be in predicted early folding regions. On a proteome scale, the incidence of predicted early folding regions decreases with protein length for a set of human protein domains from PFAM. Overall, our sequence-based early folding prediction provides a novel picture of the residues in the unfolded protein that are inclined to form stable structural elements purely based on local sequence interactions. This view of the statistical behavior of proteins, prior to the formation of highly specific defined interatomic contacts in the folded protein, allows the incorporation of a different kind of information in structural bioinformatics approaches, which are currently mostly based on folded protein structures, and should stimulate further advances in the field.	Proteins poster	Fundamental
P_Pr068	787	Miguel Correa Marrero, Richard G.H. Immink, Dick de Ridder and Aalt D.J. van Dijk	Miguel Correa Marrero	Simultaneous prediction of protein-protein contacts and interaction partners	Protein-protein interactions underlie virtually any biological process. How proteins interact with each other is therefore a fundamental question in biology. However, techniques that give fine-grained information about protein-protein interactions are low-throughput and labour-intensive, which makes the development of in silico approaches attractive.One way to approach the problem is to exploit the phenomenon of coevolution. Protein-protein interaction leads to the coevolution of the interfaces between the interaction partners, meaning that there are correlations between their sequences. From these correlations, one can deduce which residues are involved in the interaction interfaces. This can be done by applying statistical models to multiple sequence alignments of homologs of the proteins of interest.However, one can easily introduce pairs of sequences that have lost the interaction, or paralogs. This introduces noise in the analysis and has limited the application of these coevolutionary approaches. To surpass this obstacle, we are developing a novel approach.Our approach combines traditional correlated mutation analysis with the expectation-maximization algorithm. For each sequence pair in the input alignments, the algorithm will first predict whether they are interacting or not. Using proteins predicted to interact, the algorithm will then predict contacts between columns in the alignment. These two steps are repeated until convergence is reached. This approach is still being tested.	Proteins poster	Fundamental
P_Pr070	845	Sudad Dayl and Ralf Schmid	Ralf Schmid	Structure prediction of the human P2X1 receptor using a homology modelling, ab initio modelling and cross-linking approach	P2X receptors are trimeric ion channels that are activated by the binding of ATP. Each P2X subunit consists of a large extracellular loop, two transmembrane helices, and intracellular amino and carboxy termini. In vertebrates, there are seven genes coding for P2X receptor subtypes. In particular, P2X1 and P2X7 receptors are drug targets for pain management, so structural information for human P2X1 and P2X7 receptors is of great interest. X-ray structures of the zebrafish P2X4 receptor in the closed state and the open state with ATP bound enhanced our understanding of this enigmatic family of ion channel receptors. However, the C and N terminal regions which range from ~ 24-30 and 27-240 residues, respectively were not present in the constructs used for crystallization. To gain insight into the structure of the human P2X1 receptor, we applied a hybrid modelling approach. The extracellular domain and TM helices were homology modelled based on the zP2X4 template (44 % sequence identity). This was combined with fragment-based ab initio prediction for the 20 N-terminal and 20 C-terminal residues of the intracellular domain using ROSETTA with symmetry constraints and anchoring in the membrane. After clustering 10 groups of alternative models were obtained. These clusters of models are validated by site-directed mutagenesis and crosslinking.	Proteins poster	Fundamental
P_Pr071	702	Michael Ringel and Thomas Brüser	Michael Ringel	SubtleP - A new software for subcellular translocation & localization prediction	Protein translocation systems are important for the interaction of microorganisms with their surroundings, especially in host-microbe interactions for instance during infections or in symbiotic-, parasitic- or commensalistic relations. Thus the prediction of these protein translocation systems and their respective substrates might shed light on their functional relevance and facilitate phenotypic screening. Furthermore, by unraveling the functional significance, new targets for antimicrobial drugs may be identified.The identification of protein translocation systems and their respective substrates poses a major challenge for bioinformatics and many algorithms have been devised over the last years to solve this problem. Recently meta-predictors have been developed, which address individual strengths and weaknesses of these algorithms, thus optimizing prediction-accuracy. Due to the multitude of available algorithms and their relevant technicalities it may often be difficult to combine the obtained results and to evaluate their significance in the abovementioned research topics. Therefore, we designed a new user-friendly software taking advantage of many open-source, published prediction algorithms to make these available for proteome wide predictions with an interactive graphical output. This may facilitate the usage of said algorithms by a broader audience. Moreover, the software has been designed in a modular fashion, making well established algorithms available to developers as building-blocks and abstracting basic tasks such as parsing files. Therefore, developers may assemble their own predictions-algorithms upon this infrastructure, thus expediting software development in this field.	Proteins poster	Fundamental
P_Pr072	686	Bálint Mészáros, András Zeke, Attila Reményi, István Simon and Zsuzsanna Dosztányi	Bálint Mészáros	Systematic analysis of somatic mutations driving cancer: Uncovering functional protein regions in disease development	Recent advances in sequencing technologies enable the large-scale identification of genes that are affected by various genetic alterations in cancer. However, understanding tumor development requires insights into how these changes cause altered protein function and impaired network regulation in general and/or in specific cancer types.In this work we present a novel method called iSiMPRe [1] that identifies regions that are significantly enriched in somatic mutations and short in-frame insertions or deletions (indels). Applying this unbiased method to the complete human proteome, by using data enriched through various cancer genome projects, we identified around 500 protein regions which could be linked to one or more of 27 distinct cancer types. These regions covered the majority of known cancer genes, surprisingly even tumor suppressors. Additionally, iSiMPRe also identified novel genes and regions that have not yet been associated with cancer.While local somatic mutations correspond to only a subset of genetic variations that can lead to cancer, our systematic analyses revealed that they represent an accompanying feature of most cancer driver genes regardless of the primary mechanism by which they are perturbed during tumorigeneses. These results indicate that the accumulation of local somatic mutations can be used to pinpoint genes responsible for cancer formation and can also help to understand the effect of cancer mutations at the level of functional modules in a broad range of cancer driver genes.[1] Mészáros B, Zeke A, Reményi A, Simon I, Dosztányi Z.Biol Direct. 2016 May 5;11:23. doi: 10.1186/s13062-016-0125-6. PMID: 27150584	Proteins poster	Health
P_Pr073	473	Dániel Györffy, Péter Závodszky and András Szilágyi	Dániel Györffy	The blind leading the blind: how disordered peptides form an ordered complex	Disordered proteins lack a well-defined three-dimensional structure in their free form in solution but can go through a disorder-to-order transition when binding to their cellular targets. When two disordered proteins form a complex—for example a homodimer—both molecules can become ordered. Because of the huge number of degrees of freedom of a system consisting of two disordered proteins, the computational description of such systems is a serious challenge. We have introduced a two-layer network model to describe the kinetics and the mechanisms of the coupled folding and binding processes occurring during the homodimer formation of disordered peptides. In contrast to the two mechanisms used for the description of ligand binding of proteins, namely induced fit and conformational selection, we distinguish three possible scenarios for the homodimer formation of disordered proteins: (i) rigid docking, where both molecules become ordered before association, (ii) induced fit, where association occurs before the folding of either monomer, and (iii) a mixed type where one of the chains is unfolded while the other is folded when association takes place. Applying our two-layer network model to 2D HP latticemodel dimers andWako–Saitо–Muñoz–Eaton models of several known protein dimers with different experimental behaviors, we found that dimer formation occurs via all three mechanisms for all sequences. The contribution of each mechanism depends on the particular sequence, the type of process (equilibrium or steady-state), and can even vary in time. These results also indicate that dimer formation can proceed by different mechanisms in vivo than in vitro.	Proteins poster	Fundamental
P_Pr074	635	Diego Honda, Sônia Freitas and João Martins	Diego Honda	The Bowman-Birk inhibitor from Vigna unguiculata seeds (BTCI) in complex with Trypsin: a molecular orbital study	BTCI is a Bowman-Birk Trypsin/Chymotrypsin inhibitor from Vigna unguiculata seeds with high biotechnological potential, especially due to its pharmacological characteristics. It presents seven disulfide bonds, which are responsible for its high stability in a broad range of temperature and pH conditions. In this context, it was chosen three semi-empirical methodologies to get chemical insights on structure of the BTCI-trypsin interface and its relationship with inhibition process. To accomplish this objective, we explore the frontier orbitals and their four immediate neighbors. In order to understand the local interactions, we also studied the BTCI and trypsin in vacuum. Likewise, the energy of each disulfide bond of the BTCI was determined. We obtained different behavior for each methodology for trypsin and BTCI, and the BTCI-trypsin complex. However, when we analyzed the interface between those two proteins, all methods are in agreement, pointing out that Cys22 is responsible to maintain the interface conformation during the enzyme-inhibitor interaction.	Proteins poster	Biotechnology Fundamental Health
P_Pr076	632	Flavia Corsi, Alessandra Carbone and Elodie Laine	Flavia Corsi	Towards an accurate prediction of protein-DNA interfaces based on evolutionary information, physico-chemical properties of residues and local geometry of the protein structure.	Protein interactions are essential to all biological processes and they represent increasingly important therapeutic targets. A new method was recently developed for accurately predicting protein-protein interfaces, understanding their properties, origins and binding to multiple partners (Laine & Carbone, PLoS Comp. Biol. 2015). This combines in a rational and very straightforward way three sequence- and structure-based descriptors of protein residues: evolutionary conservation, physico-chemical properties and local geometry. The implemented strategy yields very precise predictions for a wide range of protein-protein interfaces and discriminates them from small-molecule binding sites, permitting to dissect interaction surfaces. The approach is implemented in JET2, an automated tool for sequence-based protein interface prediction that is based on the Joint Evolutionary Trees (JET) method (Engelen …, Carbone, PLoS Comp. Biol. 2009).We developed new strategies for predicting protein interfaces involved in protein-DNA interactions. These interaction surfaces are expected to satisfy characteristics different from those of protein-protein interfaces. We analyzed the evolutionary conservation, physico-chemical and geometrical properties of protein-DNA interfaces and we observed that not only physico-chemical properties but also geometrical patterns holding for protein-protein interactions are not anymore true for DNA-protein interactions. Then, by approaching the question as in JET2, we defined a few new rational heuristics leading to accurate protein-DNA interface identifications.This analysis (based on geometrical descriptors) can be used as the basis for the development of an optimal model of protein-DNA interaction. Other directions are constituted by RNA-protein interaction and small molecules-protein interactions, already partially addressed when analyzing protein-protein interactions (Laine & Carbone, 2015).	Proteins poster	Fundamental
P_Pr077	821	Julia Varga, Laszlo Dobson, Istvan Remenyi and Gabor E. Tusnady	Julia Varga	TSTMP: Target Selection for human TransMembrane Proteins	Transmembrane proteins (TMP) play an important role in living cells, since they are involved in diverse biological processes. Despite the great striving of worldwide structural genomics centres of membrane proteins, there are only around 60 known 3D structures among the human transmembrane proteins (with 2 or more transmembrane segments) and a further 600-700 could be modeled using existing structures. TSTMP database is a resource of human transmembrane proteins considering the existence of an exact 3D structure, or the possibility of modeling structure for the protein using existing 3D structure, or the necessity of a new structure for modeling the protein. The database was built by sorting out proteins from the human transmembrane proteome [1] with known structure and searching for suitable model structures for the remaining proteins by combining the results of state-of-the-art transmembrane specific fold recognition[2] and sequence similarity search[3] algorithms. TMPs were searched for homologues among the human transmembrane proteins to select targets whose successful structure determination would lead to the best structural coverage of the human transmembrane proteome. The database is available at http://tstmp.enzim.ttk.mta.hu.[1] The Human Transmembrane ProteomeDobson, L., Reményi, I. and Tusnády, GE. (2015)Biology Direct, 10:31[2] TMFoldRec: a statistical potential-based transmembrane protein fold recognition tool.Kozma, D. and Tusnády, GE. (2015)BMC Bioinformatics, 16, 201.[3] HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignmentRemmert, M., Biegert, A., Hauser, A. and Söding, J. (2011)Nat. Methods, 9, 173–175.	Proteins poster	Biotechnology
P_Pr078	463	Aram Gyulkhandanyan	Aram Gyulkhandanyan	Two paths of tumors destruction	Currently destruction of cancer cells actively studied in two directions: (i) by method of photodynamic therapy (PDT) and (ii) by acting on receptors of cancer cells leading to prevention of their dimerization. These studies are carried out both via experimental methods and the method of computer modelling (molecular docking). (i) As a damaging agent in a method of PDT are used photosensitizers (usually porphyrins). Photosensitizers accumulate selectively in tumors and upon illumination promoted generating of reactive oxygen species in cells that result to the destruction of cancer cells. (ii) The epidermal growth factor receptor (EGFR) is a membrane-spanning protein that governs major signaling pathways, as a result of its over expression and deregulation goes an aggressive tumor growth. Together with scientists from the University of Nantes we have shown that some small compounds (non-peptide compound nitro-benzoxadiazolyl (NBD)) may bind to dimerization domain EGFR. This causes allosteric activation of receptor, promotes the formation of stable dimers and launching of oncological process. On the other hand via molecular docking and spectral methods we showed high affinity of cationic porphyrins with a number of proteins (serum albumin, hemoglobin, cytochrome c), as well as with low molecular weight compounds (long chain fatty acids). It allows assuming that at complexation of porphyrins (with EGF or with the molecule type NBD) with the extracellular dimerization domains I and III of EGFR and by photodynamic illumination, the active oxygen species can cause destruction of the domains, prevent the dimerization process and cancer launching.	Proteins poster	Fundamental

P_Pr079	700	Erzsébet Fichó, Bálint Mészáros and István Simon	Erzsébet Fichó	Two-state Protein Complexes	Intrinsically disordered proteins (IDPs) lack a well-defined 3D structure. Their disordered nature enables them to fulfill several vital biological roles. Among others they participate in transcription, cell signaling, regulation, and stress-response. Disordered proteins rarely act alone; they are key elements of protein-protein interaction networks, often playing roles in signal transduction. In recent years it became clear that many IDPs are involved in disease development.Protein complexes formed by ordered proteins are well studied; however, the growing number of known disordered proteins and their functions require us to analyze interactions in ordered-disordered and disordered-disordered complexes. While ordered-disordered complexes have also been studied in details in recent years, the "two-state" (disordered-disordered) complexes remain a grey area of protein interactions.These type of interactions are unique because the complexes are ordered, while all participating proteins are unstable when separated. Although, these interactions are vital for the living cells, as of yet there is no available database to collect them. One of our aims is to identify these interactions through bioinformatical approaches. In order to organize our verified results, and to provide a starting point for further research on the background of these two-state protein systems, we also intend to build an online database. Deep analysis of the dataset can lead us to a better understanding of two-state protein complexes. As a long-term objective, it can provide novel pharmaceutical approaches, and can expand our knowledge of pathways mechanisms.	Proteins poster	Fundamental
P_Pr080	332	Alexandre Renaux, Ricardo Antunes, Cecilia Arighi, Andrea Auchincloss, Delphine Baratin, Alan Bridge, Elisabeth Coudert, Béatrice Cuche, Edouard De Castro, John S. Garavelli, Emma Hatton-Ellis, Guillaume Keller, Kati Laiho, Maria Martin, Alistair MacDougall,	Alexandre Renaux	UniRule - Increasing Annotation Depth of Unreviewed Protein Entries in UniProtKB.	UniProt provides a comprehensive and thoroughly annotated protein resource to the scientific community, most notably through the UniProt Knowledgebase (UniProtKB). Within UniProtKB, the reviewed section (Swiss-Prot) contains high quality, manually curated, richly-annotated protein records. In contrast, the unreviewed section (TrEMBL) which makes up 99% of UniProtKB, depends for its annotation on links to other databases and rule-based annotation systems. The use of rule-based annotation is necessary because there is no experimental data available for the majority of the unreviewed protein sequences.UniRule is a rule-based annotation system leveraging the expert-curated data in reviewed UniProtKB to increase the depth of annotation in unreviewed entries. Currently the UniRule system contains over 4,500 rules, which provide annotation for approximately 28% of unreviewed entries.Rules are a formalized way of expressing an association between conditions, which have to be met, and annotations, which are then propagated. InterPro signatures, predictive models for the functional classification of protein sequences, and taxonomic constraints are the fundamental conditions that are used. As a result, UniRule enriches the functional annotation of proteins with nomenclatures, catalytic activities, Gene Ontology terms and sequence features such as transmembrane domains. Data provenance is documented using Evidence Ontology tags.A key feature of the UniRule curation tool is a statistical quality control system which allows curators to evaluate their rules against the reviewed entries, to make sure rules are as accurate as possible. A dedicated space on the uniprot.org website has recently been created to allow users to view and explore UniRule.	Proteins poster	Fundamental
P_Pr082	446	Aytug Kiper, David Ramirez, Susanne Rinné, Wendy Gonzalez and Niels Decher	David Ramirez	Why Kv1.5 blockers preferentially inhibit TASK-1 channels?	Atrial fibrillation and obstructive sleep apnea are responsible for significant morbidity and mortality in the industrialized world. There is a high medical need for novel drugs against both diseases, and here, Kv1.5 channels have emerged as promising drug targets. In humans, TASK-1 has an atrium-specific expression and TASK-1 is also abundantly expressed in the hypoglossal motor nucleus. We asked whether known Kv1.5 channel blockers, effective against atrial fibrillation and/or obstructive sleep apnea, modulate TASK-1 channels. Therefore, we tested Kv1.5 blockers with different chemical structures for their TASK-1 affinity, utilizing TEVC-recordings in Xenopus oocytes. Despite the low structural conservation of Kv1.5 and TASK-1 channels, we found all Kv1.5 blockers analyzed to be even more effective on TASK-1 than on Kv1.5. For instance, the IC50 values of AVE0118 and AVE1231 (A293) were 10- and 43-fold lower on TASK-1. To describe this phenomenon on a molecular level, we used in silico models and identified unexpected structural similarities between the two drug binding sites. Kv1.5 blockers, like AVE0118 and AVE1231, which are promising drugs against atrial fibrillation or obstructive sleep apnea, are in fact potent TASK-1 blockers. Accordingly, block of TASK-1 channels by these compounds might contribute to the clinical effectiveness of these drugs. The higher affinity of these blockers for TASK-1 channels suggests that TASK-1 might be an unrecognized molecular target of Kv1.5 blockers effective in atrial fibrillation or obstructive sleep apnea 1.1.Kiper, A. K. et al. Pflugers Arch. 467, 1081–1090 (2015).	Proteins poster	Health

POSTER LIST
ORDERED ALPHABETICALLY BY POSTER TITLE
GROUPED BY THEME/TRACK

THEME/TRACK: SYSTEMS

Poster numbers: P_Sy001 - 094 **Application posters: *P_Sy001 - 010***

Poster number	EasyChair number	Author list	Presenting author	Title	Abstract	Theme/track	Topics
APPLICATIONS POSTERS WITHIN SYSTEMS THEME							
P_Sy001	*868*	*Hong-Woo Chun and Seonho Kim*	*Hong-Woo Chun*	*Biomedical Big data-based Dementia Prediction*	*Prediction of dementia disease has been tackled with various materials including EEG, fMRI, voice, and ADL (Activities of Daily Living) data. However, there is no approach to combine those data together to predict dementia. This project aims to integrate EEG, pupil reaction, voice, ADL, heart rate, medical history of candidates and their family to predict dementia disease. This poster will show a kind of a progress report. The first step of the project is to develop prediction models for each data and the second step is to combine all data with their proper weights.EEG, voice, pupil reaction, ADL data for 40 MCI (Mild Cognitive Impairment) and 40 NC (Normal Controls) have currently constructed at a hospital . At the same time, publicly available data including EEG, voice, heart rate data for MCI and NC has been collected to develop preliminary prediction models. Because the public datum are not from the same group of people, these public datum cannot be used to develop a integrated prediction model, but they can be used to develop each prediction models. A integrated model will be developed after constructing our own data from the same group of people sooner of later.*	*Systems/Application poster*	*Application*
P_Sy002	*347*	*Meng-Chang Shieh, Nien-Du Yang, Chih-Chieh Chen and Ching-Hsing Luo*	*Meng-Chang Shieh*	*CEPS: A simulation platform for cardiac electrophysiology models*	*Computer simulation and visualization of complex cardiac dynamics have great potential to provide valuable information for cardiac electrophysiology studies. Recently, many cardiac models have been developed to address this issue. Until now, it still lacks an integrated platform for analysis of cardiac electrophysiology computational models.In our study, a simulation platform named CEPS is developed for the study of single-cell cardiac electrophysiology. Our aim is to provide a user-friendly web interface to investigate the ionic mechanisms underlying various physiological and pathological conditions of heart. So far the simulation platform has integrated three cardiac electrophysiological models, including LR91, LRd94, and LRd2000. The simulated results could be indicated and visualized by Java-based graphics viewers, in which the action potential (V), ionic currents (INa, IKr, and IKs) and concentration profiles ([Na+], [K+], and [Ca2+]) are shown and compared. It is also shown that this platform has the ability of simulating the hyperkalemia induced cardiac arrhythmia.CEPS platform would be a useful tool for medical investigators, who are interested in the basic electrophysiology and clinical electrophysiology. The CEPS platform is freely available at the website http://140.117.103.220/~ceps .*	*Systems/Application poster*	*Application*
P_Sy003	*529*	*Emilia Wysocka, Ian Simpson, Matthew Page and James Snowden*	*Emilia Wysocka*	*Dimensionality reduction of rule-based simulation results using intrinsic dimensionality analysis*	*Rule-based (RB) languages such as Kappa and BioNetGen embody a new approach to dynamical modelling in Biology. One of their key advantages is that they can efficiently encode the combinatorial complexity of molecular events commonly found in biological systems. Static and causal analysis of RB model simulation data are usually undertaken using visualisation tools, but more detailed analysis often requires use of bespoke sets of heuristic tests in order to unravel the complex behaviour of molecular species in the simulation. Further, relatively small models can generate very large numbers of molecular species during simulation; identifying which of these dynamic features to focus on in downstream analyses is critical to the refinement and practical use of the model.Problematically, the theoretical development and implementation of model reduction methods has lagged behind the ability to build realistic models and to simulate them at scale. This has greatly hindered principled approaches to the quantitative analysis of dynamical models. We present a potential solution to this problem using a dimensionality reduction technique based on multivariate mutual information (Correlation Explanation, CorEx) that searches for a set of latent variables to best explain correlations within time-series data. We demonstrate the first use of this approach to evaluate our novel Kappa model of DARPP-32 (dopamine and cAMP-regulated phosphoprotein of 32 KDa) phosphorylation; a dynamically regulated process in the integration of neurotransmitter and neuromodulator activity in GABAergic spiny neurons, in response to dopamine and glutamate.*	*Systems/Application poster*	*Application*
P_Sy004	*778*	*Dimitris Manatakis, Andrew Sedgewick and Takis Benos*	*Takis Benos*	*Discovering Causal Associations in Omics and Clinical Data Using Mixed Graphical Models*	*Analyzing multi-modal, biomedical datasets is of paramount importance for precision medicine, discovery of drug combination efficacies and disease cause identification. Probabilistic graphical models offer a promising way to analyze biomedical datasets, since they simultaneously represent the influence graphs and multivariate probability distributions between all variables. Knowing the graphical model structure, one can extract useful information to help in disease prognosis, diagnosis, biomarker selection, patient stratification and gene functional analysis. Our new Mixed Graphical Model (MGM) - Learn algorithm was developed to address a current bottleneck in biomedical data analysis, namely learn directed (causal) graphs over continuous and discrete variables, which most current methods cannot. Here we present an application of MGM-Learn to a complex glioblastoma dataset.*	*Systems/Application poster*	*Application Health*
P_Sy005	*568*	*Hyunjung Shin, Yonghyun Nam, Dong-Gi Lee and Sunjoo Bang*	*Hyunjung Shin*	*Disease Co-occurrence Scoring with Semi-Supervised Learning*	*The disease network have provided insights into establishing relationships between diseases. However, it yet remains as only a map of topologies between diseases, not being able to be a pragmatic diagnostic/prognostic tool in medicine. One way to evolve disease network from bench-to-bedside is to equip a function of scoring that measures the likelihoods of the association between diseases. In this study we propose semi-supervised scoring algorithm for quantifying the probabilities of disease co-occurrence given a primary disease of a patient. In predicting disease co-occurrence on disease networks, the proposed algorithm not only improved the AUC performance up to 0.72 (lifted from random guessing) but also discovered potential disease co-occurrence relations. The results appear to be concordant with the existing literatures on disease comorbidity.*	*Systems/Application poster*	*Application*
P_Sy006	*789*	*Sebastian Thieme, Jesper Romers and Marcus Krantz*	*Sebastian Thieme*	*Improvements in reconstructing biological signalling networks based on rxncon*	*Living organisms are complex systems of interacting components. A crucial step to understand those complex biological systems is the construction of biological networks that reflect our current knowledge of the system. The scope and coverage of different network reconstructions can differ, but they have one aim in common - to convert the knowledge into a mathematical model enabling computational analysis to find possible inconsistencies and gaps. While reconstruction methods for metabolic networks are well established, only a few methods exist for reconstructing cellular signal transduction networks. Here, we present a method – rxncon – enabling a systematised and condensed reconstruction of signal transduction networks. This method has two aspects. On the one hand, we developed a language for reconstructing biological networks. The language addresses the issue, that states are combined in signal transduction networks, which create a large number of specific states, generating highly complex structures. Due to the context-free grammar in the language and the description of the data on the same level of detail as biological findings we can largely avoid the combinatorial complexity. On the other hand, we developed a framework for interpreting and exporting this knowledge into different mathematical models and visualisation formats - bridging large scale network reconstruction and classical mathematical modelling approaches. Hence, rxncon has the potential to reconstruct, validate and simulate genome scale signalling models.*	*Systems/Application poster*	*Application Fundamental*
P_Sy007	*625*	*Ilona Liesenborghs, Jan S.A.G. Schouten, Lars M.T. Eijssen, Martina Kutmon, Theo G.M.F. Gorgels, Chris T. Evelo, Henny J.M. Beckers and Caroll A.B. Webers*	*Ilona Liesenborghs*	*Molecular pathway analysis in human trabecular meshwork cells after treatment with corticosteroids*	*Introduction: Corticosteroids, used for the treatment of many different diseases in ophthalmology, cause an elevation of the eye pressure in 18-36% of patients. This may cause loss of visual field and eventually blindness (corticosteroid-induced glaucoma). The pathogenic mechanism is not completely understood, however, the trabecular meshwork seems to play an important role. To gain more insight into the pathogenic mechanisms, we performed pathway analysis of publicly available microarray datasets in which the gene expression of human trabecular meshwork cells treated with and without dexamethasone are compared.Methods: A search for relevant microarray datasets was conducted in ArrayExpress and Gene Expression Omnibus (GEO). Four datasets were included (GSE16643, GSE37474, GSE65240, and GSE6298). Quality control and pre-processing were performed with ArrayAnalysis.org and pathway overrepresentation analysis and visualization with PathVisio. Pathways with a Z-score >1.96, a permuted p-value <0.05, and ≥3 changed genes were considered significantly changed. Pathways that were significantly changed in at least three out of four studies were further investigated.Results: Pathway analysis of the above-described datasets showed respectively 21, 19, 39, and 21 significantly altered pathways. Five pathways were significantly affected in three out of four datasets. Some of the pathways are already known to be associated with glaucoma or corticosteroid-induced effects, for example, the complement activation pathway (WP545).Conclusion: Pathway analysis and visualisation are powerful tools to gain more insights into the mechanisms of corticosteroid-induced glaucoma.*	*Systems/Application poster*	*Application Health*
P_Sy008	*626*	*Jan S.A.G. Schouten, Ilona Liesenborghs, Martina Kutmon, Lars M.T. Eijssen, Theo G.M.F. Gorgels, Henny J.M. Beckers and Carroll A.B. Webers*	*Jan S.A.G. Schouten*	*Molecular pathway analysis in patients with primary open angle glaucoma*	*Introduction: Glaucoma is one of the most prevalent causes of visual impairment and blindness worldwide. It causes a progressing neuropathy of the optic nerve, resulting in loss of visual fields and eventually blindness. The most common form is primary open angle glaucoma. The pathogenic mechanism is not completely understood, however, the trabecular meshwork seems to play an important role. In order to improve this insight, we performed pathway analysis of a publicly available transcriptomics dataset.Methods: A search for relevant microarray datasets in which the gene expression in human trabecular meshwork cells in patients with and without primary open angle glaucoma is compared, was conducted in ArrayExpress and Gene Expression Omnibus (GEO). Dataset GSE27276 was selected for further analysis. Quality control and pre-processing were performed with ArrayAnalysis.org, pathway overrepresentation analysis and visualization with PathVisio. Pathways with Z-score >1.96, permuted p-value <0.05, and ≥3 changed genes were considered significantly changed.Results: Pathway analysis showed 14 significantly altered pathways. The most significant altered pathways are complement activation (WP545), inflammatory response pathway (WP453), and focal adhesion (WP306).For example, in the complement activation pathway, in particular the classical pathway, the genes are upregulated, indicating activation in primary open angle glaucoma. Studies in other tissues revealed that multiple triggers, in addition to known immunoglobulins, are able to activate this pathway. Further research on this pathway may identify new insights in the pathogenesis.Conclusions: Molecular pathway analysis can give us new insights into the pathogenesis of primary open angle glaucoma.*	*Systems/Application poster*	*Application Health*

P_Sy009	*660*	*Vincenzo Belcastro, Carine Poussin, Stephanie Boue, Yang Xiang, Florian Martin, Julia Hoeng and Manuel Peitsch*	*Vincenzo Belcastro*	*The Systems Toxicology Computational Challenge: Markers of Exposure Response Identification – Insights gained*	*Humans are constantly exposed to chemicals (e.g. pollutants and pesticides) that may trigger harmful molecular changes. Risk assessment in the context of 21st century toxicology relies on the elucidation of mechanisms of toxicity and the identification of markers of exposure response from high-throughput data. The development of relevant computational approaches for the analysis and integration of these large-scale data remains challenging.The purpose of sbv IMPROVER (www.sbvimprover.com/) is the crowd-sourced verification of methods in systems biology via computational challenges. The latest challenge (2016) aimed to address questions on the identification of exposure response markers in human blood enabling to discriminate between (A) exposed and non-exposed subjects, and (B) subsequently between formerly exposed and never exposed subjects (sub-challenge1) as well as the translatability of those markers between species (sub-challenge2). Participants were provided with human and mouse blood gene expression datasets to develop human-specific and species-independent gene signature-based models for class label prediction of independent test samples. Anonymized participant's predictions were scored according to a predefined methodology approved by an independent scoring review panel of experts.Twenty-three teams worldwide participated in at least one sub-challenge. Most of the teams provided highly accurate predictions (pval < 0.05) for the first task (A), while prediction performances were much lower for task B. Different classes of machine learning methodologies were applied including Linear Discriminant Analysis, and Random Forest. A small set of features were common to top 3 ranked submissions. The challenge outcome and lessons learned will be shared with the computational scientific community.*	*Systems/Application poster*	*Application Biotechnology Health*
P_Sy010	*800*	*Mugdha Srivastava, Sybille Dühring, Stefan Schuster and Thomas Dandekar*	*Mugdha Srivastava*	*Understanding of the metabolic interplay between host and fungi by combining metabolic modeling and game theory*	*Aspergillus fumigatus is a prevalent opportunistic pathogen in immune-compromised patients. Virulence traits are multifactorial. This includes the capacity of A. fumigatus to grow and adapt to the host environment, evade the immune system and damage the host. The present work aims to understand the metabolic interaction between the human host and A. fumigatus by combining metabolic modeling and game theory. A metabolic model of A. fumigatus was created with special emphasis on the shared metabolites such as iron. Elementary mode analysis was performed to predict the robustness of the model. To study the conflict and cooperation between the host and fungi for the acquisition of iron, gene expression data was used to identify the metabolic pathways effected during iron starvation and iron replete conditions. Costs and benefits for each effected pathway were calculated using a growth equation and elementary modes. Subsequent steps include the application of game theory to understand and assess the strategies applied by both the host and fungi for the acquisition of iron (calculate Nash equilibria, their stability, risk of change and potential of fungal adaptation).Understanding metabolic strategies used by the host and A. fumigatus will improve therapies in A. fumigatus infection in risk patients. This includes identification of novel protein targets for antibiotics (we currently assess fungal-specific pathways) as well as early markers of strategy change. In particular, blood invasion of A. fumigatus has to be prevented in risk patients and novel early markers to detect such a change towards sepsis and invasion are desirable.*	*Systems/Application poster*	*Application Biotechnology Health*
				OTHER POSTERS WITHIN SYSTEMS THEME			
P_Sy012	640	Jan Bert Van Klinken, Ayse Demirkan, Harish Dharuri, Peter Henneman, Aaron Isaacs, Cornelia van Duijn, Peter-Bram 't Hoen and Ko Willems van Dijk	Jan Bert Van Klinken	A functional validation of human genome-scale metabolic models	Genome-Scale Metabolic Models (GSMMs) are increasingly used for the interpretation and integration of omics datasets. The results of these analyses heavily rely on the comprehensiveness of the GSMM and on how well it is linked to external databases. To test the coverage of human GSMMs, we created a compendium of inherited metabolic diseases and related biomarkers by extracting phenotypic data from OMIM, KEGG DISEASE and genome-wide association studies. Subsequently we assessed the ability of Recon2.04 and HMR2.0 to predict these associations based on network distance and compared performance to gene-set based pathway analysis.We found that Recon2.04 covered 77.9% of the genes and 62.2% of the gene-metabolite interactions, while HMR2.0 covered 86.3% of the genes and 68.1% of the interactions. In comparison, pathway databases covered 93.4% of the genes and 72.5% of the interactions, showing that both GSMMs contained gaps with respect to classical pathway knowledge. Inspecting the results, we found that missing links were mostly due to absent reactions, which mainly involved lipoprotein metabolism and regulatory pathways. Therefore, we extended Recon2.04 by including missing reactions and importing signalling cascades, allosteric interactions and cofactor data from UniProt and Reactome, which greatly increased its coverage (94.5% of the genes; 86.6% of the gene-metabolite interactions).Concluding, current human GSMMs are only partly able to explain biomarkers in inherited metabolic diseases and contain gaps with respect to classical pathway knowledge. Careful manual curation of these models is vital to increase their coverage and, consequently, to enable their use in genetic and epidemiological studies.	Systems poster	Fundamental Health
P_Sy013	422	Lieven Verbeke, Jimmy Van Den Eynden, Piet Demeester, Jan Fostier and Kathleen Marchal	Lieven Verbeke	A multi-purpose network-based data integration strategy for tumour analysis	The study of cancer, a highly heterogeneous disease with different causes and clinical outcomes, requires a multi-angle approach and the collection of large multi-omics datasets. We present MUNDIS, a MUlti-purpose Network-based Data Integration Strategy that, unlike any other method, provides a unified approach for unsupervised subtype classification, driver gene prioritization and network delineation. Key to the method is the conversion of all available data into a single comprehensive network representation containing not only genes but also individual patients. Additionally, prior knowledge can be incorporated by adding previously identified molecular interactions to this network representation. We demonstrate the performance of MUNDIS by applying it to ovarian and glioblastoma tumour datasets from The Cancer Genome Atlas. By integrating mRNA, copy number, mutation and methylation data, MUNDIS was able to identify molecular subtypes of ovarian and glioblastoma cancer that are highly predictive for patient survival. Additional in-depth analysis of these subtypes identified by MUNDIS demonstrates the method's ability to provide a mechanistic insight in the underlying biological processes.	Systems poster	Health
P_Sy014	462	Robin Haw, Guanming Wu and Lincoln Stein	Robin Haw	A Refreshing Look at the Reactome Functional Interaction Network.	The Reactome Functional Interaction (FI) network was developed to significantly enlarge protein coverage for high-throughput data analysis by merging curated pathways in Reactome and other reaction-network databases with protein pairwise relationships from other public sources. We have extended the FI network to encompass interactions between transcription factors and their targets from the ENCODE data sets, and miRNAs and their targets from miRecords and miRTarBase. The current version of the FI network contains 327,867 functional interactions, covering over 12,000 SwissProt identifiers (about 60% of total human genes).ReactomeFIViz, the Cytoscape app based on the Reactome FI network, provides intuitive, user-friendly and rich graphical interfaces for researchers to fulfill pathway and network-based data analysis to discover clinically-relevant disease biomarkers. Using a set of genes, or a gene expression data set, users can carry out network-based analyses by constructing a FI sub-network, search for network modules, and annotate the sub-network or its modules. Users can also visualise Reactome knowledgebase pathways using Cytoscape, either in their native pathway diagram view or expanded FI network view. Using either visualisation approach, users can perform pathway enrichment analysis on a set of genes, and check genes in identified pathways. For prognostic biomarker discovery, users can perform survival analysis using the univariate Cox proportional hazard model using a built-in command. We have recently developed a method to convert Reactome pathways into probabilistic graphical models (PGMs) by adopting the PARADIGM approach to allow users to create predictive models of the effect of perturbing multiple genes on pathway activities.	Systems poster	Health
P_Sy015	652	Adam Kozak, Dorota Formanowicz and Piotr Formanowicz	Adam Kozak	A semi-quantitative Petri net model of oxidative stress in atherosclerosis	Atherosclerosis is a complex disease process of endothelium which affects significant part of population in different age. Beginning of this process is related to endothelium inflammatory process and one of the key factors in its progression is oxidative stress. In this work we present an extended model of oxidative stress in atherosclerosis progression which includes influence of asymmetric dimethylarginine (ADMA), chronic kidney disease (CKD), anti-oxidants, rich cholesterol diet, mast cells degranulation, inflammation and immune system response (including process of forming foam cells which finally build atherosclerotic plaque). The model is based on Petri net and includes some quantitative information, particularly describing well known biochemical relationships but also relationships describing influence of ADMA and superoxide radical anion. A structural analysis of the model was performed. Moreover, comparison of biological conclusions was made for weighted and unweighted versions of the model. Analysis focused on t-invariants and the influence of changes in quantitative information on biological conclusions.This research has been partially supported by the Polish National Science Centre grant No. 2012/07/B/ST6/01537.	Systems poster	Fundamental
P_Sy016	455	Thanh Phuong Nguyen, Laura Caberlotto, Jochen Schneider and Thomas Sauter	Thanh Phuong Nguyen	A systems medicine approach to elucidate the comorbidity network in metabolic diseases	Over the past decades the molecular background of the phenotypic variability in metabolic diseases (MDs) has been studied and a spectrum of relations between clinical syndromes and molecular features has been identified. Although some genes have emerged as important players in the pathogenesis, the precise molecular machinery involved in MDs remains largely unknown. Our aim is to elucidate phenotypic interdependencies and comorbidities of MDs. We present a systems medicine approach to modelling a comprehensive comorbidity network of MDs. In addition to the conventional sharing-gene analysis, we computed network-based separation measure to discover disease modules and then quantified the disease comorbidities using shortest distance analysis. We firstly curated a wide range of MDs from the MESH and the CTD database. Based on 1,261 MD-related proteins, a protein interaction network of 3,958 proteins and 8,645 interaction was reconstructed from the HPRD database. Filtering out diseases with less than 5 associated proteins, we modelled a weighted comorbidity network of 83 meta-nodes representing MDs and 3,403 meta-edges representing disease-disease associations. We inferred 535 significant putative comorbidities corresponding to 76 diseases. In the obtained network, among top 20 central diseases ranked by betweenness and degree centralities, we found Diabetic Cardiomyopathies, Hepatitis, Vitamin A Deficiency, Hypoglycemia, Non-alcoholic Fatty Liver Disease (NAFLD), and Metabolism Inborn Errors. The investigation of the NAFLD comorbidity highlighted adipocytokine and PPAR signaling and cytokine-cytokine receptor interaction pathways. The comorbidity network analysis of MDs allowed the identification of molecular underpinning for MD interdependencies that could represent promising targets for drug therapy in MDs.	Systems poster	Health
P_Sy017	323	Yoshiyuki Asai, Takeshi Abe, Li Li and Hiroaki Kitano	Yoshiyuki Asai	A versatile platform for multilevel modeling of physiological systems: integration of time series data	The importance of systematic software support to develop physiologically detailed, largescale models has been well recognised recently, as models keep increasing in size and complexity for possible applications to medicine. There are several pioneering efforts to develop technologies in that direction, such as SBML (Systems Biology Markup Language) and CellML among others, which are XML-based formats to describe models of biological and physiological systems. Applications, such as CellDesigner for dealing with SBML models, OpenCell for simulations of models written in CellML, and others, have also been developed. The main purpose in developing these languages and applications is to establish common communication foundations, and to enhance the exchange of models among collaborators.In this stream, a modeling platform, PhysioDesigner, has been developed with a model descriptive language PHML (Physiological Hierarchy Markup Language) which can cooperate with SBML, and has partial compatibility with CellML. PhysioDesigner equips a graphical user interface, allowing users to build models with hierarchical structures characteristic of physiological systems. Besides of developing computable models, time series data integration to mathematical models is of interest to researchers, but was not paid careful attention so far in terms of software supports, even though they are often used in bio-physiological simulations. Here, recent extensions to PhysioDesigner are introduced, including the integration of time series data into dynamical systems based on differential equations. Time series data can be obtained either experimentally or by simulation. Concomitantly, the simulator, Flint, has also been expanded to support models including time series data.	Systems poster	Application
P_Sy019	853	John Reid and Lorenz Wernisch	John Reid	Branching Gaussian processes for analysis of cell fate from single cell expresssion	We present a novel branching process model based on Gaussian processes. We show how to perform exact inference in the model using belief propagation. We apply the model to single cell data from the mouse embryo and demonstrate how it identifies genes that are markers for particular cell fates.	Systems poster	Fundamental
P_Sy020	667	Yin Cai, Julius Hossain, Jean-Karim Heriche, Antonio Politi, Birgit Koch, Malte Wachsmuth, Bianca Nijmeijer and Jan Ellenberg	Jean-Karim Heriche	Building a dynamic protein atlas of human mitosis	Cell division requires that the activities of hundreds of proteins be tightly regulated in space and time but the dynamics and interactions of the hundreds of proteins required for mitosis in human cells is incomplete and fragmented. While live cell imaging is a powerful tool for studying the distribution and dynamics of proteins, it has not been used to map large sets of proteins that carry out dynamic functions such as mitosis due to lack of systematic and quantitative approaches. To address these issues, we generated a 4D image data-driven canonical model of the morphological changes during mitotic progression in human cells. We show that this model can be used to integrate the dynamic distribution of 3D concentration data for any number of mitotic proteins recorded by automated quantitative fluorescence live cell imaging. Mining the integrated data allowed us to automatically identify sub-cellular localizations and quantify the timing and abundance of protein fluxes between sub-cellular structures. Our integrated experimental and computational method enables building a 4D protein atlas of the dividing human cell.	Systems poster	Fundamental
P_Sy021	812	Federica Eduati, Victoria Doldán, Bertram Klinger, Thomas Cokelaer, Anja Sieber, Fiona Kogera, Mathurin Dorel, Mathew Garnett, Nils Blüthgen and Julio Saez-Rodriguez	Federica Eduati	Cell-type specific parameters of signalling pathway models as biomarkers for drug sensitivity	Therapies targeting specific molecular processes are major strategies to treat cancer. Genomic features have been associated with response to drugs, rendering them biomarkers for drug sensitivity. However, our ability to stratify patients based on these features is still limited. As drug response is a dynamic process affecting largely signal transduction, we investigate the association between cell-specific dynamic signalling pathways and drug sensitivity.A signaling network was derived from literature and public databases and was used to generate cell-type specific models based on logic ordinary differential equations using CellNOpt for 14 colon cancer cell lines. For each cell line, model parameters were optimised using phosphoproteomics data of 14 proteins upon 42 perturbations, and L1 regularisation was used to induce sparsity of the network. The model parameters (pathway interactions) were then used as features to predict drug sensitivity response (from the Genomics of Drug Sensitivity in Cancer project) to a panel of 30 drugs. Resulting associations were compared with those between mutations and drug response: we found associations for 19 drugs, for 6 of which there is no genetic marker. We also used the associations between pathway interactions and drug response to predict potentially interesting drug combinations, some of which are supported by the literature and some will be experimentally tested.Our results suggest that cell-specific signalling models can be used to understand efficacy of pharmaceutical drugs and as potential biomarkers, even when no mutation can do so.	Systems poster	Health
P_Sy022	650	Manuel Valenzuela, Alejandro Acevedo, Raul Conejeros and German Aroca	Raul Conejeros	Characterization of continuous cultures of Scheffersomyces stipitis on its phenotypic phase plane.	Scheffersomyces stipitis has been extensively studied because of its ability to ferment xylose to ethanol. However, this fermentation depends on oxygen availability instead of carbon source uptake rate. Available genome scale models of the yeast allows the exploration of metabolism in biomass and ethanol production capabilities, and kinetic models of its cultures account for a subset of conditions that can be compared with the potential shown by the metabolic models. This work focuses on determining reachable metabolic states within the phenotypic phase plane and compare them with those obtained using an existing kinetic model for the continuous culture of S. stipitis and with experimentally obtained data. Computational methods considered in-silico analysis of S. stipitis using genome scale metabolic models, in order to evaluate biomass and ethanol production, and to build phenotypic phase planes. Continuous cultures of S. stipitis NRRL Y-7124 were performed in YPX culture media, considering 10, 20 and 40 g/L of xylose in the feed stream. Dilution rates of 0.2, 0.3 and 0,4 h-1 and a kLa of 9 h-1 were used. Preliminary experimental results were mapped over the phenotypic phase plane, being in a wider range than those predicted by the kinetic model of the continuous culture.	Systems poster	Fundamental
P_Sy023	383	Bastian Hornung, Bartholomeus van den Bogert, Vitor Martins Dos Santos, Peter J. Schaap and Hauke Smidt	Bastian Hornung	Characterizing and understanding the rumen microbiota	Omics based approaches have seen a shift from single organism to meta-omics that focus on microbial communities to elucidate their composition and function. One such community is the rumen microbiota, which has gained attention because of its Archaea that contribute with methane to the greenhouse effect. Here we investigate the rumen activity with a focus on methane metabolism.Rumen fluid samples were collected from 12 Holstein dairy cows, which were assigned to 4 diets (grass silage, maize silage, 33:67 and 67:33 mixtures). Methane measurements were conducted in respiration chambers after which a rumen fluid sample was collected. RNA was sequenced on Illumina HiSeq 2500. Reads were cross-assembled into one transcriptome and differential expression (DE) analysis was performed.DE analysis showed that only a small fraction of the assembled proteins had a consistent change between the diets. 10% of these proteins belonged to the Archaea, of which 1/3 were related to methane metabolism, and were less expressed with increasing maize. Methanogenesis from formate in Methanobrevibacter smithii was reduced with the maize diet. The formate formation by members of Clostridiales was reduced, indicating that M. smithii is dependent on these organisms. Furthermore, methanogenesis from methanol by M. stadtmanae was affected.Our study confirms that transcriptome analysis can show how feeds can affect the rumen. Metabolic profiles shifted when cows were fed different feeds, which lowered the methane metabolism and is in agreement with lower methane. This study indicates that a maize-enriched diet lowers methanogenesis and shows a potential of reducing greenhouse gasses.	Systems poster	Ecosystems

P_Sy024	739	Italo Faria Do Valle, Giulia Menichetti, Giorgia Simonetti, Marco Manfrini, Danielle Fernandes Durso, Antonella Padella, Carolina Terragna, Cristina Papayannidis, Carla Adriane Ramos Segatto Fontoura, José Carlos Merino Mombach, Giovanni Martinelli, Gastone Castellani and	Italo Faria Do Valle	Combined genomics and transcriptomics for multi-tumor drug targeting	A comprehensive molecular perspective of tumor samples may contribute to the understanding of similar genomic profiles across cancer types. This understanding may enable us to repurpose therapies from one cancer to another. We hypothesized that the study of gene-gene expression relations across cancer types can reveal clusters of tumors. These clusters may have characteristic gene signatures that provide multi-tumor drug targets, prognostic markers, and a molecular taxonomy for effective cancer categorization. We retrieved the expression data of 760 genes from 2378 samples across eleven tumor types present in the TCGA database. We integrated gene expression profiling, somatic mutational landscape and clinical information in a network environment based on protein interactions and cancer-related pathways. First, we clustered tumor types based on their transcriptomic profiling resemblance. For each cluster of tumors, we retrieved a gene signature by combining the ranking of a network node centrality measure and the information of somatic mutated genes in the vicinity of the signature genes in the protein-protein interaction network. The gene signatures of the tumor clusters presented four main biological processes: NF-KB signaling pathway, chromosomal instability, DNA metabolism and proteasome. Combining the gene signatures with the available clinical information allowed us to stratify samples according their survival outcome, highlighting their biological relevance. The genes signatures also contained genes that have been tested as therapeutic targets for specific tumors, according the ClinicalTrials.gov website. Finally, we propose a set of genes that may be used as drug targets for multi-tumor therapeutic strategies.	Systems poster	Health
P_Sy025	565	Vardan Andriasyan, Artur Yakimovich, Robert Witte, Fanny Georgi, Ivo Sbalzarini and Urs Greber	Vardan Andriasyan	Computational modelling of human adenovirus egress	Human adenoviruses (HAdVs) infect respiratory, ocular, and digestive organs, and cause lethal outcomes in immune-compromised patients and infants. Genetically modified HAdVs are broadly used as gene therapy tools, and oncolytic vectors. HAdV enters through receptor mediated endocytosis, and delivers its DNA genome to the cell nucleus for replication. During the late stages of HAdV infection, pseudo-crystalline subviral particles and virions assemble into discrete clusters in the nucleus. These clusters are highly dynamic depending on their size, although their mode of motility is unknown. The clusters are released from the nucleus to the cytoplasm upon disruption of the nuclear membrane at late stages of lytic infection, and clusters transmit infection to neighboring cells by extracellular mass transfer. Here we use live-cell imaging and biophysical modeling to explore a plausible link between the dynamics of subnuclear clusters, nuclear disintegration and cell lysis. We quantify the dynamics of viral clusters and host cell chromatin using live cell confocal and holographic imaging. Using a top-down approach, we incorporate these data into a partial differential equation model towards predicting forces exerted from viral clusters onto the nuclear envelope and plasma membrane. We identify conditions required for nuclear and cell lysis tuning features of viral clusters. Insights into underlying mechanisms of HAdV spreading from infected to neighboring cells can help facilitate the rational design of new antivirals and combinatorial drug-enhanced viral oncolysis.	Systems poster	Health
P_Sy026	742	Andras Hartmann, Susana Martinez, Sascha Zickenrott, Satoshi Okawa and Antonio Del Sol	Andras Hartmann	Constraint Based Reconstruction of Gene Regulatory Networks	The increasing amount of data produced by current high-throughput technologies allows for a better understanding of disease-related phenotypic traits and yet the molecular mechanisms stabilizing them are predominantly unknown. The inference of gene regulatory networks (GRNs) has potential to significantly enhance the discovery of disease-related and context-specific interactomes[1]. However, GRN inference remains a major challenge due to the combinatorial explosion of solutions and the underdetermined nature of the problem, which can be overcome by adding more context-specific information, such as gene expression, chromatin conformation or covalent modifications. Even though integrating multiple layers of information has been attempted before, all methods are tailored to specific needs[2,3].In this work an Integer Linear Programming approach is proposed for inferring context-specific Boolean GRNs. In order to guide the optimization to a more reliable GRN, both context-independent and context-specific information is encoded into constraints. In contrast to previous methods that explore the search space using heuristics, the proposed methodology guarantees global optimality, significantly reduced runtime, and scalability.The optimization-based approach was validated on various human cell lines and outperformed previous methods in all cases when comparing the amount of retained cell-line specific ChIP-Seq interactions. We believe that our modular framework will significantly enhance the discovery of disease motifs and drug target prediction by providing more reliable context-specific GRNs.References[1]Madhamshettiwar PB et. al, Genome. Med. 4,41(2012)[2]Karlebach G and Shamir R, Nat. Rev. Mol. Cell. Biol. 9,770(2008)[3]Hecker M et. al, BioSystems 96,86(2009)	Systems poster	Health
P_Sy027	374	Maryam Nazarieh, Thorsten Will, Mohamed Hamed, Christian Spaniol and Volkhard Helms	Maryam Nazarieh	Constructing and analyzing disease-specific or developmental stage-specific transcription factor and miRNA co-regulatory networks	TFmiR is a freely available web server for integrative analysis of combinatorial regulatory interactions between transcription factors, miRNAs and target genes that are involved in disease processes in human. To better characterize the cellular processes at molecular level from a network perspective in normal and disease conditions in human and also now in mouse, we have extended the published version of TFmiR by various new features such as the construction of tissue-specific networks. Besides disease processes, the successor of TFmiR can now also be applied to identify regulatory motifs associated with the transitions between different developmental stages from the sets of genes and miRNAs provided by the user. One particular challenge in studying gene regulatory networks is to identify the main drivers and master regulatory genes that control such cell fate transitions. In addition to common topological measures and by considering tissue-exclusive genes, we reformulate this problem as an optimization problem of computing a Minimum Connected Dominating Set (MCDS) for directed graphs. MCDS is applied to the well-studied gene regulatory networks of E. coli and S. cerevisiae and to a pluripotency network for mouse embryonic stem cells. The results show that the MCDS captures most of the known key player genes identified so far in the model organisms. Moreover, this method suggests an additional small set of transcription factors as novel key players for governing cell-specific gene regulatory networks. This set can also be investigated with regard to diseases.	Systems poster	Fundamental
P_Sy028	444	Stefano Vavassori, Karsta Luettich, Marja Talikka, Justyna Szostak and Julia Hoeng	Stefano Vavassori	Construction of a computable biological network model focused on goblet cell hyperplasia/metaplasia in the lung	One of the biggest challenges of the 21st century toxicology is the comprehensive and unambiguous analysis and interpretation of large scale data sets. Our systems toxicology approach employs biological network models to gain a deeper mechanistic understanding of exposure effects in the respiratory tract. Here, we present a computable network model that describes the biological signaling pathways regulating the increase in the number of large airway goblet cells (GC), clinically known as goblet cell hyperplasia/metaplasia (GCH/M). GCH/M leads to mucus accumulation in the lungs which is one of the key features of chronic bronchitis and other obstructive lung diseases.The GCH/M network model was constructed using a semi-automated knowledge extraction workflow (BELIEF) that allows the transformation of unstructured information available in the literature (and published datasets) into a structured, cause-effect, scientific representation in Biological Expression Language (BEL). The network model contains causal relationships from over 40 scientific publications and model focuses on EGFR signaling and, in part, IL13 signaling sharing some effectors with the EGFR pathway (e.g. Foxa2, ROS activation of EGFR via neutrophils).This work is part of a wider effort to build of Adverse Outcome Pathways (AOPs) for respiratory disorders. The ultimate goal is to use the network model to quantify key events in the mucus hypersecretion AOP, which will be instrumental for research applications such as drug development and toxicological risk assessment of exposure to airborne toxicants.	Systems poster	Health
P_Sy029	846	Gaia Zaffaroni, Luc Grandbarbe, Alessandro Michelucci and Antonio Del Sol	Gaia Zaffaroni	Context-specific gene regulatory network to identify key genes in differentiation of NSCs to astrocytes	Bovine serum can induce neural stem cells (NSCs) differentiation to astrocytes with high efficiency, yet its precise molecular function is still unclear. Cell differentiation is characterized by a large scale reprogramming of gene expression patterns, in which transcription factors (TFs) play a leading role by acting differentially on multiple targets. In this study, a Boolean gene regulatory network model was applied to identify TFs driving the differentiation of NSCs to astrocytes. First, all potential interactions occurring among the TFs that are differentially expressed during the differentiation process were extracted from literature. This network was pruned with a genetic algorithm in order to obtain a gene regulatory network whose attractor states represent the Booleanized expression levels of the TFs at the initial and final stage of the differentiation. Then, topological analysis of this contextualized network, including the identification of network stability motifs, was performed. Systematic perturbations on the network were also performed, to identify TFs that cause the biggest change towards the differentiated cell state. The TFs identified with these analyses were compared with the results from two previously published network-based tools for key TF predictions. Our approach is based on gene expression data of the initial and target cell types, and therefore can give results that are more specific to the differentiation under study compared to other approaches that can only be applied on extensive collections of previously published data. The experimental validation confirmed the role of our best candidate TF in the differentiation of NSCs to astrocytes.	Systems poster	Health
P_Sy030	661	Mark Alber, Shixin Xu, Zhiliang Xu, Oleg Kim, Rustem Litvinov and John Weisel	Mark Alber,	Coupled multi-scale modeling and experimental study of platelet adhesion and blood clot deformation and rupture	Two models for studying adhesivity of a platelet to fibrin and stability of a blood clot under blood flow will be described. Importance of the study is underscored by the fact that detached fragments of unstable blood clot (emboli) can occlude downstream branches of the blood vessel, leading to vascular obstruction, or emboli can end up in lungs with deadly consequences. First, a two-state kinetic modeling approach will be described for studying fibrin (or fibrinogen) - platelet integrin binding which was calibrated using experimental data for the αIIbβ3 integrin - fibrin single molecule studies. The model describes unbinding kinetics of αIIbβ3 integrin - fibrin (or fibrinogen) complex from two possible states (low binding affinity and high binding affinity) as observed in experiments. Transition between the two states is assumed to be at equilibrium. Given a pulling force acting on the αIIbβ3 integrin - fibrin (or fibrinogen) complex, the model calculates the probability the bond breaking. The second novel model is a continuum multiphase model for simulating deformation and rupture of blood clot under different shear flow condition, which takes into account interactions between different fully and partially activated platelets, fibrin network and plasma. The blood clot is treated as a viscoelastic material. Simulation results show in detail how the rheological response of the blood clot to the flow is determined by mechanical and structural properties of its components. Model simulations predict that the permeability and porosity of the shell region profoundly affect the stability of the blood clot.	Systems poster	Health
P_Sy031	418	Michele Caselle, Laura Cantini,Santo FortunatoandEnzo Medico	Michele Caselle	Detection of gene communities in multi-networks reveals cancer drivers.	Multi-Networks represent the most effective way to study functional regulatory patterns originating from complex interactions across multiple layers of biological relationships. Such a multi-network approach is mandatory when complex pathologies like cancer are addressed. In this poster weproposе a new, original, multi-network-based strategy, which we recently published in Scientific Reports (2015) 5-17386, to integrate different layers of genomic information and use them in a coordinate way to identify driving cancer genes. The multi-networks that we consider combine transcription factor co-targeting, microRNA co-targeting, protein-protein interaction and gene co-expression networks. The rationale behind this choice is that gene co-expression and protein-protein interactions require a tight coregulation of the partners and that such a fine tuned regulation can be obtained only combining both the transcriptional and post-transcriptional layers of regulation. To extract the relevant biological information from the multi-network we studied its partition into communities. To test of our proposal we applied it to a set of expression data for gastric, lung, pancreas and colorectal cancer and identified from the enrichment analysis of the multi-network communities a set of candidate driver cancer genes. Some of them were already known oncogenes while a few are new. The combination of the different layers of information allowed us to extract from the multi-network indications on the regulatory pattern and functional role of both the already known and the new candidate driver genes.	Systems poster	Health
P_Sy032	419	Jiajia Xu, László Kupcsik, Dirk Inzé and Christian Hermans	Jiajia Xu	Different copies, different roles – paralogous genes in lateral root gene regulatory network of oilseed rape	Nitrogen fertilization is often overused for maximizing Brassica napus (oilseed rape) yield but this raises environmental concerns. We aim to gain better knowledge on lateral root development processes in order to draw strategies to redesign root system architecture (more branched to avoid soil nitrate leaching). In this study, seedlings of Darmor accession were grown on vertical agar plates. We applied N-1-naphthylphthalamic acid (NPA) to block auxin transport, followed by 1-naphthalene acetic acid (NAA) to induce auxin response. That way, the lateral root initiation was synchronized. Time-series of RNA-seq data (24 sampling points within 72 h after NAA application) were used for constructing a lateral root gene regulatory network (LR-GRN). It was modularized based on the degrees and clustering co-efficiencies of each node, i.e., gene. We then compared it with the Arabidopsis LR-GRN constructed by Lavenus et al. (2015) (Plant Cell 27, 1368–1388). Conservations and differentiations between the model and crop species networks were pin pointed. For example, ARF7 played important roles in both LR-GRNs, while PLT5 and PLT1 were the hub regulators respectively in Brassica and Arabidopsis. On average, there are six copies of each Arabidopsis gene in Brassica. This enable us to examine the different roles of different copies after duplication, e.g., ARF7-A1 and ARF7-C3 (first and third copies in A and C genomes respectively) showed up in the same module while ARF7-A2 showed up in another module indicating different functions of different copies of ARF7 in Brassica.	Systems poster	Agro-Food
P_Sy033	498	Ian Walsh, Christopher H. Taron and Pauline M. Rudd	Ian Walsh	Digestor: a software tool to determine relative abundance of glycans from exoglycosidase digestons	Mammalian protein glycosylation pathways are complex and result in a wide diversity of glycan structures attached to many different glycoproteins. Glycoproteins and glycolipids dominate cell surfaces and are the first level of cell communication with the environment, affecting disease pathogenesis as well as providing disease specific markers [1]. Enzymatic digestion of glycans with specific exoglycosidases can be used to improve the interpretation of glycan structural data obtained from Mass Spectrometry (MS) and Liquid Chromatography-MS (LC-MS). Exoglycosidases are enzymes that catalyze sequential removal of monosaccharides from the non-reducing end of glycans. When coupled with highultra performance liquid chromatography (H/UPLC) analysis, glycan profiles obtained after exoglycosidase digestion show reproducible and predictable shifts in the retention times enabling detailed structural assignment [2]. Interpretation of exoglycosidase treatment data can often be difficult and time consuming. It requires manually reconciling the mobility shifts of glycan peaks in an H/UPLC profile in response to treatment with one or more enzymes. Depending on sample complexity and the number of enzymes used in the analysis, this could mean manually interpreting hundreds of peak shifts. To address this issue, we are developing Digestor a software tool which can automatically or semi-automatically annotate H/UPLC exoglycosidase digestions, with the ultimate goal to determine relative glycan abundances in a given sample.REFERENCES1.Pinho, S.S. and C.A. Reis, Nature Reviews Cancer, 2015.2.Mariño, K., et al., Nature chemical biology, 2010. 6(10): p. 713-723.	Systems poster	Biotechnology Health
P_Sy034	766	Åsmund Flobak, Tonje Strømmen Steigedal, Barbara Niederdorfer, Liv Thommesen, Martin Kuiper and Astrid Lægreid	Åsmund Flobak	DrugLogics: Logical models for drug screen prioritization	Multi-drug precision oncology is in need of approaches that enable drug combination prioritization, since the combinatorial explosion renders traditional trial-and-error screening approaches ineffective. Our computation-assisted approach contributes by highly efficient prediction of drug responses while relying only on characterizing the experimental system (cell line, tumor) at baseline conditions.Logical models are derived from cancer signaling topologies, calibrated to particular cell types or tumors by steady state biomarkers from unperturbed cells. Based on a proof-of-concept model (Flobak et al PLoS Comp Biol, 2015) we now explore a pipeline for automated causal network topology assembly, logical model parameterization and model ensemble evaluation. Prior knowledge is integrated from cell signaling databases, and multi-omic data is integrated to describe patterns of signaling activities characterizing an experimental system. Genetic algorithms are employed to optimize logical equations to obtain models where the attractor recapitulates steady state biomarkers from biological assays. Model simulations suggest prioritization of drugs that can disrupt disease phenotypes, restoring regulatory anti-survival phenotypes present in healthy cells.The pipeline correctly classified 20 of 21 combinations as synergistic or non-synergistic, with one novel synergy validated in vivo. When applied to a manually curated topology, models automatically parameterized predicted five synergies (four true positives, no false negatives) when normalized to topology-intrinsic synergies. In ongoing work, model predictions are challenged with a dataset of 171 drug combinations (19 individual drugs) across 8 cell lines.Our prototype computational-experimental pipeline demonstrates the potential to economize pre-clinical drug combination synergy discovery and to provide clinical decision support for personalized therapy.	Systems poster	Health
P_Sy036	835	Hung-Cuong Trinh and Yung-Keun Kwon	Hung-Cuong Trinh	Edge-based sensitivity analysis of signaling networks by using Boolean dynamics	Motivation: Biological networks are composed of molecular components and their interactions repre-sented by nodes and edges, respectively, in a graph model. Based on this model, there were many studies with respect to effects of node-based mutations on the network dynamics, whereas little attention was paid to edgetic mutations so far.Results: In this paper, we defined an edgetic sensitivity measure which quantifies how likely a con-verging attractor is changed by edge-removal mutations in a Boolean network model. Through extensive simulations based on that measure, we found interesting properties of highly sensitive edges in both random and real signaling networks. First, the sensitive edges in random networks tend to link two end nodes both of which are susceptible to node-knockout mutations. Interestingly, it was analogous to an observation that the sensitive edges in human signaling networks are likely to connect drug-target genes. We further observed that the edgetic sensitivity predicted drug-targets better than the node-based sensitivity. In addition, the sensitive edges showed distinguished structural characteristics such as a lower connectivity, more involving feedback loops and a higher betweenness. Moreover, their gene-ontology enrichments were clearly different from the other edges. We also observed that genes incident to the highly sensitive interactions are more central by forming a considerably large connected component in human signaling networks. Finally, we validated our approach by showing that most sensitive interactions are promising edgetic drug-targets in p53 cancer and T-cell apoptosis networks. Taken together, the edgetic sensitivity is valuable to understand the complex dynamics of signaling networks.	Systems poster	Fundamental
P_Sy037	428	Konstantine Tchourine, Richard Bonneau and Christine Vogel	Konstantine Tchourine	Explicit Modeling of Differential RNA Stability Improves Inference of Transcription Regulation Networks	Despite many years of research and the availability of large-scale datasets, modeling RNA transcription and predicting transcriptional regulatory interactions on a systems level in eukaryotes remains a challenging problem and requires modeling changes in RNA abundance due to both the regulation of synthesis and degradation. Even Saccharomyces cerevisiae has several hundred putative TFs and ~6,000 potential targets, rendering the theoretical regulatory interaction space enormous. Furthermore, eukaryotes are marked by extensive promoter regions, many response pathways, and additional regulatory layers, e.g. RNA decay, which further confound gene expression regulation. For these reasons, even the best network inference algorithms have so far performed very poorly in yeast. The Inferelator is a state-of-the-art transcription regulatory network inference algorithm that assumes a linear ordinary differential equation model of transcription factor - DNA regulatory interactions and features an RNA decay term that is assumed to be constant across all genes and conditions. However, experimental evidence indicates that RNA decay rates vary significantly across both genes and conditions. We show that allowing RNA decay rates to vary within the Inferelator framework results in an improved efficiency of recovering known TF-DNA regulatory interactions in S. cerevisiae. Furthermore, we show that RNA decay rates that optimize the performance of the Inferelator recapitulate experimentally measured trends in RNA decay rates in yeast. In particular, using an improved gold standard of interactions in yeast, our model makes accurate gene- and condition-specific predictions about RNA decay rates that have not been previously reported.	Systems poster	Fundamental

P_Sy038	663	Yuriy Hulovatyy, Huili Chen and Tijana Milenkovic	Tijana Milenkovic	Exploring the structure and function of temporal networks with dynamic graphlets	The increasing availability of temporal real-world networks, while opening new opportunities, has also raised new challenges for researchers. Namely, despite a large arsenal of powerful methods that already exist for studying static networks, these methods cannot be directly applied to temporal networks. Clearly, both static (those studying the aggregated network) and static-temporal (those studying a series of the results for individual snapshots) approaches overlook temporal information that is important for studying a dynamic system. We develop such a strategy that aims to fully explore inter-snapshot information. We base our methodology on well-established graphlets (subgraphs), which have been proven in numerous contexts in static network research. We develop new theory to allow for graphlet-based analyses of temporal networks. Our new notion of dynamic graphlets is different from existing dynamic network approaches that are based on temporal motifs (statistically significant subgraphs). The latter have limitations: their results depend on the choice of a null network model, and choosing a good null model is non-trivial. Our dynamic graphlets overcome the limitations of the temporal motifs. Clearly, accounting for temporal information helps. We apply dynamic graphlets to temporal age-specific molecular network data to deepen our limited knowledge about human aging.	Systems poster	Biotechnology
P_Sy039	495	Maria Victoria Aguilar Pontes, Julian Brandl, Adrian Tsang, Mikael R Andersen and Ronald de Vries	Maria Victoria Aguilar Pontes	Expression data integration in an Aspergillus niger genome-scale metabolic model	Filamentous fungi include important species used in industrial applications. One of the main representatives is Aspergillus niger, an industrial workhorse used for enzyme and metabolite production. In order to achieve its full potential, deeper knowledge of the metabolism is needed. We propose a new metabolic network, improving the previous version of Andersen et al. (2008), based on literature and transcriptome data. This network will be used as a model to predict A. niger intracellular carbon metabolic fluxes during growth under different conditions. To evaluate the model, predicted results will be compared to RNA-seq data under the same conditions as well as other experimental results collected from literature.Our aim is to create a model that will give us new insights on carbon metabolic pathways in A. niger and obtain leads to improve industrial processes. This model will also enable us and other researchers to study carbon utilization by fungi in more detail.	Systems poster	Biotechnology
P_Sy040	698	Safiye Celik, Benjamin Logsdon, Stephanie Battle, Charles W. Drescher, Mara Rendi, R. David Hawkins and Su-In Lee	Safiye Celik	Extracting a low-dimensional description of multiple gene expression datasets reveals a potential driver for tumor-associated stroma in ovarian cancer	Patterns in gene expression data conserved across multiple independent disease studies are likely to represent important molecular events underlying the disease. We present a novel graphical model learning approach, INSPIRE, to extract highly coherent and biologically relevant modules of co-expressed genes and the dependencies among the modules from multiple expression datasets that may contain different sets of genes. INSPIRE addresses five big challenges:1) INSPIRE enables use of multiple datasets that are not "synchronized" due to discrepancies between which genes are measured (e.g., platform-specific differences).2) Combining datasets leads to increased power to detect reliable patterns in expression data because of a greater pooled sample size.3) Modeling expression data in a much lower-dimensional space results in more power to learn patterns, even when the number of genes is much greater than the sample size – the norm for gene expression data.4) INSPIRE identifies robust (i.e. conserved across datasets) modules and module dependencies.5) INSPIRE naturally models the module dependencies capturing more complex interactions among pathways, cell populations, or other biologically driven modules.Through extensive simulated and real data analysis, we demonstrate INSPIRE is a great practical trade-off between model complexity and model parsimony when understanding biological pathways. We show that INSPIRE infers more accurate models than existing methods to extract low-dimensional representation of expression data. Applying INSPIRE to nine ovarian cancer datasets followed by experimental validation leads to a new marker and potential driver of tumor-associated stroma, HOPX, whose module suggests a previously unknown mechanism underlying tumor-associated stroma.	Systems poster	Health
P_Sy041	328	Silvia Gerber, Reinhard Guthke and Jörg Linde	Silvia Gerber	Genome-wide gene regulatory network in the opportunistic human pathogenic fungi Aspergillus fumigatus	Aspergillus fumigatus is an opportunistic human pathogenic fungus, which can cause systemic infections that may lead to death in immunocompromised hosts. Since still little is known about genes involved in virulences, it is important to find essential genes for developing new medication. For this aim central genes (hubs) are interesting network features, as they play crucial roles for information and energy transport and thus, are potential drug targets.For the identification of these hubs via topological analyses, a large scale gene regulatory network was inferred based on public, as well as unpublished RNA-Seq data. As the number of available expression data is still insufficient, information from various other sources was collected to complement the expression data. A linear regression algorithm, based on LARS [1] and adaptive Lasso [2], was utilised, which was already applied to infer genome wide gene regulatory networks before [3]. Despite, the low measurable quality of the networks, robust hubs, i. e. genes with a high number of outgoing interactions in multiple networks, were found and analysed.In this study, we inferred the first genome-wide gene regulatory network for Aspergillus fumigatus. Six reliable hubs were found, which showed a certain robustness concerning various parameters. These hubs include genes which are important for the stability of the cytoskeleton and RNA metabolism as well as putative transcription factors. Thus, they are potential drug targets.[1] Efron et. al. The Annals of Statistics (2004)[2] Zou. Journal of the American Statistical Association (2006)[3] Altwasser et. al. Front. Microbiol. (2012)	Systems poster	Fundamental
P_Sy042	570	Simone Daminelli, Josephine Thomas, V. Joachim Haupt, Claudio Duran, Michael Schroeder and Carlo Vittorio Cannistraci	Carlo Vittorio Cannistraci	How network topological models influence drug-target prediction	The identification of drug-target interactions (DTIs) is important for understanding drug mode of action, infer new indicationsand identify possible side effects. Nevertheless, it is still a challenging task especially if we consider its formal definition aslink-prediction problem in complex networks. Moreover, since novel drug-target validation is a costly and time consumingendeavour, a reliable evaluation of predictors performance is an open problem.In this work we compare state-of-the-art supervised methods and topology-based models for drug-target interaction prediction.Besides, we consider prediction algorithms both based on bipartite network projections as well as recently proposed topologicalmodels, based on the Local Community Paradigm (LCP).We analyze 5 gold standard DTIs networks and provide an exhaustive performance evaluation based on two validationframeworks. Additionally, we include a novel independent benchmark set of both positive and negative drug-target interactionsdefined by the value of their experimental chemical affinity. Finally, we investigate differences and similarities of the novelpredictions derived from methods inspired by different principles.Our results show that drug-target networks have enough topological information to identify highly reliable predictions, withcomparable performance to state-of-the-art supervised methods which exploit additional knowledge. Surprisingly, a detailedanalysis of novel predictions revealed that each method prioritize distinct true interactions, suggesting that a combination of avariety of methodologies motivated by diverse principles might improve the current drug-target discovery strategies.	Systems poster	Health
P_Sy043	630	Erika Tsingos, Burkhard Höckendorf, Thomas Sütterlin, Stephan Kirchmaier, Lázaro Centanin, Niels Grabe and Joachim Wittbrodt	Erika Tsingos	How tissues coordinate growth in an organ: Insights from modeling clonal lineages in fish	The continuously growing eye of fish presents the perfect model system to explore how different tissues coordinate proliferation in an organ. The neural retina and surrounding retinal pigmented epithelium (RPE) share a bipartite stem cell niche. Strikingly, labeling the progeny of individual stem cells in medaka fish (Oryzias latipes) reveals heterogeneous lineages that differ between neural retina and RPE. Why do these tissues grow differently, and how can heterogeneous lineages be reconciled with homogeneous organ growth? To answer these questions, we simulate a 3D virtual eye in a computational cell-based model implemented in the platform EPISIM.In the virtual eye, a monolayer of spherical cells is affixed to an expanding hemispherical surface. Cells only proliferate within a narrow ring at the base of the hemisphere; divisions occur with a random chance. This simple model shows that the distinct clonal pattern of neural retina and RPE results from different growth modes. While cells in the RPE passively proliferate in response to an expanding scaffold, the neural retinal cells control organ growth pace. Moreover, neural retinal stem cells in vivo preferentially choose a biased division axis that hints at a role in regulating eye shape. By exploring various proliferation and growth modes, the model highlights a role of the retina in controlling eye growth, eye shape, and retinal architecture. By tweaking these parameters, evolution can calibrate the eye to perfectly adapt to the animal's ecological niche.	Systems poster	Fundamental
P_Sy044	759	Adel Ait Hamlat, Alessandra Carbone, Thierry Jaffredo, Pierre Charbord and Charles Durand	Adel Ait Hamlat	Hub centered deconvolution improves gene regulatory network reconstruction	Gene Regulatory Networks (GRN) are graphical models used to describe cellular systems by representing in vivo interactions within a set of genes actors in this system. These interactions are shown as oriented edges from a regulator gene to a regulated one, the level of expression of the first controlling the rate of transcription of the second. One important characteristic of GRN is that a small portion of the nodes has a high connectivity while the majority is connected to few other genes. The highly connected nodes in a GRN, called hubs, are of a high biological interest as they can be targeted for therapeutic purposes.Here we describe HubNeD (HUB NEtwork Deconvolution) a novel method that exploits the topological properties of GRN to reconstruct them from steady state expression profiles. HubNeD takes as input the Pearson correlation matrix computed from these profiles and works in two steps: first, hubs are inferred from strongly correlated communities. The whole network is then reconstructed by decreasing scores for the interactions between two genes if both of them are strongly correlated to a hub. As these interactions are most likely indirect, this deconvolution step reduces the rate of false positives thus leading to better performances.HubNeD showed better performances on typical Yeast in silico benchmarks compared to classical methods (based on correlation, regression or mutual information). HubNed was also applied to real transcriptomic data of murine stromal cell lines (generated from us and collaborators, and collected form the literature) and provided consistent results.	Systems poster	Fundamental
P_Sy045	756	Paul Ashford, Anna Hernandez, Todd Greco, Anna Buch, Beate Sodeik, Ileana Cristea, Kay Grünewald, Adrian Shepherd and Maya Topf	Anna Hernandez	HVint: A strategy for identifying novel protein-protein interactions in herpes simplex virus type 1	We present HVint, the first dedicated resource of intra-viral protein-protein interaction data for herpes simplex virus type 1 (HSV-1). HSV-1 is one of the most studied members of the human herpesviruses, a group of human pathogens with notorious impact on word-wide public health. To populate the HVint database, binary protein-protein data was collated from five external resources. The coverage of the initial interactome was further increased by means of a computational strategy; interactions from homologous human herpesviruses were mapped to the HSV-1 interactome using orthology relationships. Thus, HVint is not only a centralised resource of known protein interactions for HSV-1 but is also a tool for highlighting potential novel interactions. The latter can be an important asset for prioritising target interactions to test experimentally. The reliability of all interaction data included in HVint was assessed under a standardised scoring scheme that considers several aspects modulating the reliability of an interaction, including the number and type of lines of evidence available for a each interaction. An independent experimental analysis was conducted on a subset of the predicted interactions to assess the power of the implemented computational method. The results support a number of our predictions and contribute to formulate new hypothesis on the nuclear egress and early envelopment pathways. Our computational framework for data integration has been as simplified as possible, making the protocol readily applicable to other species. Finally, a user-friendly web interface was developed to provide intuitive access to all the interaction data in HVint for future users.	Systems poster	Fundamental
P_Sy047	543	Tsukasa Fukunaga and Wataru Iwasaki	Wataru Iwasaki	Importance of considering simple factors in C. elegans behavioral analysis	With rapid advances in genome sequencing and editing technologies, systematic and quantitative analysis of animal behavior is expected to be another key to facilitating data-driven behavioral genetics. The nematode Caenorhabditis elegans is a model organism in this field. Several video-tracking systems are available for automatically recording behavioral data for the nematode, but computational methods for analyzing these data are still under development. In this study, we applied the Gaussian mixture model (GMM)-based binning method to time-series postural data for 322 C. elegans strains and revealed that the occurrence patterns of the postural states and the transition patterns among these states have a strong relationship with each other, which relationship must be taken into account in the computational identification of strains with interesting behavior. Based on this observation, we identified several strains that exhibit atypical transition patterns but use wild-type N2-like postural states. Surprisingly, we found that two simple factors---overall acceleration of postural movement and elimination of inactive conditions---explained the behavioral characteristics of strains with very atypical transition patterns; therefore, computational analysis of animal behavior must be accompanied by evaluation of the effects of these simple factors. Finally, we discovered that the npr-1 and npr-3 genes have closely related functions that were not predictable by sequence homology, proving that our data-driven approach can reveal the functions of genes that have not yet been characterized.	Systems poster	Fundamental
P_Sy048	394	Jennifer Scheidel, Leonie Amstein, Börje Schweizer, Jörg Ackermann and Ina Koch	Jennifer Scheidel	In silico knockout experiments based on Petri net models	The knockout analysis is a worthwhile method to observe the effect of a specific protein on the systems behavior. Mathematical modeling provides the possibility for in silico knockouts. Often only a small fraction of knockout results obtained from a systematical in silico knockout analysis was experimentally investigated. Besides the standard Petri net analysis techniques, such as covered by transition invariants, and the biological interpretability of each transition invariant [1], in silico knockout experiments are useful for model verification and experiment planning.We developed a new tool called SiKnock to perform and visualize in silico knockout experiments. Based on Petri net models we introduce a new concept of in silico knockout analysis to ensure the correct prediction of the systems behavior. SiKnock provides single, double, and multi knockout analysis, visualizes the results as a knockout matrix and provides a graphical user interface. We applied the method to study the autophagic degradation pathway of the pathogen Salmonella Typhimurium. We compared the knockout results with published knockout or knockdown experiments and ensured the biological correctness of the model structure. We found knockout behavior known in literature and generated new hypotheses for experiments, for example, knocking out the autophagy receptor NDP52 (nuclear dot protein 52 kDa) predicted no influence on the recruitment of OPTN (optineurin) to ubiquitinated Salmonella Typhimurium.Reference[1] Koch I, Reisig W, Schreiber F. (2011). Modeling in systems biology: the Petri Net approach (Vol. 16). Springer Science & Business Media. Germany.	Systems poster	Fundamental
P_Sy049	396	Laura Cantini, Emmanuel Barillot, Francois Radvanyi and Andrei Zinovyev	Laura Cantini	Independent Component Analysis unveils the landscape of multi-omics pancancer data	Recent advances in high-throughput technologies have enabled the comprehensive characterization of various cancer types at multiple omic levels. Extracting relevant biological knowledge from this huge amount of information represents a remarkable opportunity in cancerology. However, this achievement is limited by the presence in the data of various overlapping biological factors linked to the tumor cells or to the tumor microenvironment and non biological factors linked to sample processing or data generation.To deconvolute these factors, Independent Component Analysis (ICA), originally developed to solve the blind source separation problem, is perfectly suited. In this work ICA is applied to transcriptomic and methylomic data obtained from 32 different tumor types. Each data matrix is thus decomposed into a number of components, each of which is characterized by an activation pattern both across genes and across samples.Our analysis identified multi-cancer-shared and single-cancer-specific components. Using colorectal cancer (CRC) as a paradigm, we showed that our approach can significantly contribute to the puzzling problem of CRC subtypes identification and characterization. Indeed, the recently published CRC consensus subtypes were consistently retrieved in our analysis and new molecular insights concerning these subtypes were highlighted. Notably, other signals of promising interest, not included in the already known CRC subtypes, were detected. Among them, of particular interest is a component jointly regulated by STAT1 and IRF4.Ongoing analysis aims at comprehensively characterize all the identified pan-cancer components and to integrate the results obtained with methylomic and transcriptomic data to get more insights on cancer complexity.	Systems poster	Health
P_Sy050	540	Lingfei Wang and Tom Michoel	Lingfei Wang	Causal inference of genetic regulations, impaired by confounders and saved by alternative tests	The causal inference of genetic regulations is believed to provide accurate predictions through a series of tests with genotype and gene expression data. We computed the analytical null distribution for every test, and reduced computation time from hours to seconds. The remarkable speedup enabled statistical evaluations of causal inference on Geuvadis and DREAM challenge datasets, only finding that the independence test is widely impaired by confounders and feedback loops, whilst the secondary linkage test can also fail with weak regulations. Correspondingly, we proposed alternative, composite tests to infer genetic regulations, which are demonstrated to outperform existing methods in speed and accuracy. We have implemented the tests and released the package "Findr" at https://github.com/lingfeiwang/findr.	Systems poster	Fundamental
P_Sy051	532	Mustafa Alshawaqfeh, Ahmad Bani Younes and Erchin Serpedin	Erchin Serpedin	Inferring Microbial Interaction Network Using a Stochastic Generalized Lotka-Volterra Model	Inferring the microbial interaction networks (MINs) and modeling their dynamics are critical in understanding the mechanisms of the bacterial ecosystem and designing antibiotic and/or probiotic therapies. Recently, several approaches were proposed to infer MINs using the generalized Lotka-Volterra (gLV) model. Main drawbacks of these models include the fact that these models only consider the measurement noise without taking into consideration uncertainties in the underlying dynamics. Furthermore, inferring the MIN is characterized by the limited number of observations and nonlinearity in the regulatory mechanisms. Therefore, novel estimation techniques are needed that address these challenges.This work proposes SgLV-EKF: a stochastic gLV model with extended Kalman filter (EKF) algorithm to model MIN dynamics. In particular, SgLV-EKF employs a nonlinear stochastic dynamic model rather than the conventional gLV model to infer MIN. SgLV-EKF was compared with Nelder's and Stein's algorithm on two synthetic data-sets and one real data-set. The first data-set models the randomness in measurement data. The second data-set incorporates uncertainties in the dynamics. Whereas, the third data-set is a real time series generated by an infant's gut. SgLV-EKF outperforms the existing algorithms in terms of robustness to measurement noise, modeling errors, and tracking the dynamics of the MIN. Particularly, SgLV-EKF provides consistent accuracy irrespective to modeling errors whereas Nelder's algorithm diverges and Stein's algorithm infers parameters that lie in the unstable region of the dynamic system. The execution time of SgLV-EKF is comparable to Stein's algorithm, and is tens of times faster than Nelder's algorithm.	Systems poster	Ecosystems

P_Sy052	657	Krzysztof Gogolewski, Weronika Wronowska, Bogdan Lesyng andAnna Gambin	Krzysztof Gogolewski	Inferring molecular pathways heterogeneity from transcriptional data	Motivation: RNA microarrays and RNA-Seq are nowadays standard technologies to study the transcri-ptional activity of the cell. Most studies focus on tracking transcriptional changes caused by specificexperimental conditions. The obtained information about up and down-regulated genes are inferredfrom the behavior of relatively large population of cells. Even assuming perfect sample homogeneity,different subpopulations of cells can exhibit diverse transcriptomic profiles as they follow different regula-tory/signaling pathways.Results: In this study we propose a novel computational method to infer the proportion between cells thatentered the cell death pathway and those that actively proliferate as a reaction to imposed experimentalconditions. Our method applied to interpret RNA microarray data can also be adapted to detection of othermolecular processes, and in particular can be easily extended to RNA-Seq data. Specifically, we investigatethe influence of C2 ceramide and poly(ADP--ribose) polymerase--1 inhibitor (PJ34) on the viability of neuroblastomacells. Our results shown neurotoxic effect of ceremide which was increased by PJ34. Currently we conduct a seriesof biological assays for further validation of our computational method.Conclusions: The presented methodology complement standard approaches for inferring the regulatorynetwork from transcriptomic data, and could be particularly useful for the analysis of cancer cell lines.	Systems poster	Biotechnology
P_Sy053	607	Laura Follia, Giulio Ferrero, Niccolò Totis, Chiara Riganti, Francesco Novelli, Gianfranco Balbo, Marco Beccuti and Francesca Cordero	Laura Follia	Inspecting Energy Realising Pathways by combination of genomics data and mechanistic approach.	In systems biology a great effort is devoted to study the aberrant signalling pathways enhance cancer progression and cancer metabolism. In Pancreatic Ductal AdenoCarcinoma (PDAC) the protein Alpha-Enolase plays a key role in its metabolism: it is generally overexpressed and it is associated with tumour progression through the Warburg effect.We start from a functional characterization of Alpha-Enolase investigating its role in energy realising pathways (ERP) in PDAC cells using the mechanistic version of the metabolic model. To perform kinetic simulations it is necessary to determine all the kinetic parameters. Unfortunately, because of economical and technical reasons only few of them are well characterized.To cope with this problem we model ERP using Stochastic Petri Net with Uncertainty: a formalism that we have recent proposed to study complex system with unknown parameters. Indeed, from SPNU it is possible to derive a system of Ordinary Differential Equations in which unknown parameters are characterized through an Optimization Problem.The objective function used in the ERP model encodes the Warburg Effect: the cancer phenotype that we want to reproduce.We also enrich the ERP model with the integration of genome data characterizing PDAC patients. Indeed, deep sequencing technologies lead to the production of a high volume of biological data that are used to profile the patients from a genomic point of view.The mutations and copy number variants of genes involved in the glycolysis pathway are integrated in our PDAC metabolic model inspecting the effects of the main mutations, amplification, and deletion events.	Systems poster	Health
P_Sy054	745	Robert Sehlke, Luke Tain, Manopriya Chokkalingam, Nazif Alic, Nagarjuna Nagaraj, Matthias Mann, Christoph Dieterich, Andreas Beyer and Linda Partridge	Robert Sehlke	Integration of two Drosophila insulin mutant models reveals tissue-specific mechanisms of longevity	Drosophila insulin-like peptides (DILPs) are upstream regulators of the IIS pathway in flies. Down-regulation of this conserved pathway increases lifespan, with longevity being dependent on the central transcription factor dFOXO, and other factors, such as the widespread intracellular endosymbiont Wolbachia. Reduced IIS-activity, and a number of other interventions, extends lifespan and affect many history traits. Disentangling longevity-specific downstream processes from the many pleiotropic phenotypes of such interventions presents an on-going challenge, requiring the joint analysis of several complementary, system-wide data sets.We investigated the effects of reduced insulin signalling, through DILP2,3-5 abrogation, integrated along two axes of additional factors: Presence and absence of Wolbachia, and dFOXO-minus versus wild type background, respectively.To that end, we collected shotgun-proteomics and transcriptomics data in eight experimental groups, from up to five individual tissues: Fat body, gut, brain, muscle, and ovaries. To delineate transcriptional and post-transcriptional regulation, the RNA polysome fractions of selected experimental groups were sequenced. We proceeded to identify proteins responding to reduced IIS signalling in a dFOXO-dependent or Wolbachia-dependent manner. Leveraging a priori information from protein-protein interaction networks via network propagation [3,4], robust functional categories of the response were identified. Our analysis suggests Wolbachia influences host translation machinery in the fat body, and that insulin signalling acts tissue-specifically on proteostasis and metabolism to mediate longevity.	Systems poster	Fundamental Health
P_Sy056	730	Ezequiel Iván Juritz, Max Schobert, Kenneth Timmis, Fernando Danilo Gonzalez-Nilo, Lothar Jansch, Dieter Jahn and Jose Manuel Borrero de Acuña	Ezequiel Iván Juritz	Integrative protein network from universal stress protein UspK	Universal stress proteins (Usp) enhance bacterial survival rates when exposed to certain stress agents. Despite their importance, the exact biological role of several universal stress proteins in P.aeruginosa is unknown. We isolated and identified protein interaction partners of the most abundant universal stress proteins in P.aeruginosa: UspK. P.aeruginosa cells were grown under pyruvate fermentation, oxygen limited and denitrifying conditions. Samples were taken at day 1 and day 4. UspK, crosslinked with its interaction partners, was purified from the formaldehyde treated P.aeruginosa cells by affinity chromatography and its interaction partners were determined by LC-MS/MS. Each experiment was performed in triplicate.We downloaded the full interactome of P.aeruginosa from the STRING database and crosslinked this information with the results obtained from our experiments.A network was generated using the interaction information as derived from the STRING database, the protein function (as protein COG codes), the increase or decrease of the protein abundance under stress conditions (sample day 1 vs. day 4) and, finally, we identified the location of the regulators within the network.From the 1,457 detected proteins, 63% were upregulated in stress conditions, 32% were downregulated and 5% showed no significant variation. We detected a 10-fold increase from day 1 to day 4 when considering all UspK-interacting partners, in accordance with the function of the studied protein, involved in stress responses pathways. The obtained network shows the grouping of proteins that share COG codes which can be related to diverse pathways of UspK.	Systems poster	Biotechnology
P_Sy057	497	Kirstine Belling, Francesco Russo, Anders Boeck Jensen, Marlene Danner Dalgaard, David Westergaard, Niels Erik Skakkebæk, Anders Juul and Søren Brunak	Francesco Russo	Klinefelter syndrome comorbidities induced by increased X gene dosage and altered protein interactome activity.	Klinefelter syndrome (KS) (47,XXY) is the most common male sex chromosome aneuploidy. Diagnosis and clinical supervision remain a challenge due to varying presentation and insufficient characterization of the syndrome. Here we present a study combining health data-driven epidemiology and molecular level systems biology to improve the understanding of KS and the molecular interplay influencing its comorbidities. Using health registry data from the entire Danish population covering 6.8 million patients a total of 78 overrepresented comorbidities were identified from Danish hospital patient records. The extracted comorbidities included both clinically well-known (e.g. infertility and osteoporosis) and still less established KS comorbidities (e.g. pituitary gland hypofunction and dental caries). Three approaches were applied to identify key underlying molecular players in the KS comorbidities: (A) Differential expression analysis and identification of coexpressed modules using data generated on peripheral blood, (B) Identification of central hubs in a KS comorbidity network based on known disease proteins and their protein-protein interactions, and (C) Identification of dosage perturbed protein complexes in the KS comorbidity network. Together these approaches pointed to novel aspects of X-chromosome related mechanisms, including perturbed Cytokine-cytokine interaction and Jak-Stat pathways, immune system alterations and disturbed functionality of leptin and erythropoietin signalling in KS. This work presents an extended epidemiological study that links KS comorbidities to the molecular level and identify potential causal players in the disease biology underlying the identified comorbidities.	Systems poster	Health
P_Sy058	582	Bernhard Steiert, Jens Timmer and Clemens Kreutz	Bernhard Steiert	L1 regularization facilitates detection of cell type-specific parameters in dynamical systems	Motivation: A major goal of drug development is to selectively target certain cell types. Cellular decisions influenced by drugs are often dependent on the dynamic processing of information. Selective responses can be achieved by differences between the involved cell types at levels of receptor, signaling, gene regulation, or further downstream. Therefore, a systematic approach to detect and quantify cell type- specific parameters in dynamical systems becomes necessary.Results: Here, we demonstrate that a combination of nonlinear modeling with L1 regularization is capable of detecting cell type-specific parameters. To adapt the least-squares numerical optimization routine to L1 regularization, sub-gradient strategies as well as truncation of proposed optimization steps were implemented. Likelihood-ratio tests were used to determine the optimal regularization strength resulting in a sparse solution in terms of a minimal number of cell type-specific parameters that is in agreement with the data. By applying our implementation to a realistic dynamical benchmark model of the DREAM6 challenge we were able to recover parameter differences with an accuracy of 78%. Within the subset of detected differences, 91% were in agreement with their true value. Furthermore, we found that the results could be improved using the profile likelihood. In conclusion, the approach constitutes a general method to infer an overarching model with a minimum number of individual parameters for the particular models.Availability: A MATLAB implementation is provided within the freely available, open-source modeling environment Data2Dynamics (Raue et al., 2015). Source code for all examples is provided online at http://www.data2dynamics.org/.	Systems poster	Health
P_Sy059	490	Theo Knijnenburg, Gunnar Klau, Francesco Iorio, Mathew Garnett, Ultan McDermott, Ilya Shmulevich and Lodewyk Wessels	Lodewyk Wessels	Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy'	A central challenge in modern biology is to create models that bridge the gap between the molecular level on which interventions can be designed and the cellular and tissue levels on which the biological phenotypes are manifested. Because of the interaction between biological components, single-predictor models are generally not accurate enough to model a biological phenotype. On the other hand, machine learning approaches, such as Elastic Net and Random Forests produce complex multi-predictor models that are hard to interpret and not amenable to the generation of hypotheses that can be experimentally tested. As a consequence, such models are not likely to further our understanding of biology. There is an urgent need for approaches that build small, interpretable, yet accurate models that capture the interplay between biological components and explain the phenotype of interest.We present 'Logic Optimization for Binary Input to Continuous Output' (LOBICO), a computational approach that infers small and easily interpretable logic models of binary input features that explain a continuous output variable. Applying LOBICO to a large cancer cell line panel, we find that logic combinations of multiple mutations are more predictive of drug response than single gene predictors. Importantly, we show that the use of the continuous information leads to robust and more accurate logic models. LOBICO implements the ability to uncover logic models around predefined operating points in terms of sensitivity and specificity. As such, it represents an important step towards practical application of interpretable logic models.	Systems poster	Health
P_Sy060	624	Pauline Traynard, Adrien Fauré, François Fages and Denis Thieffry	Pauline Traynard	Logical model specification aided by model checking: application to the mammalian cell cycle regulation	Understanding the temporal behaviour of biological regulatory networks requires the integration of molecular information into a formal interaction model. Logical modelling, based on Boolean or multilevel frameworks, abstracts from precise quantities and offers a versatile framework to delineate the main dynamical properties of such networks. Logical models are particularly easy to define, simulate and compose. However, formal analysis of asynchronous dynamics faces a combinatorial explosion as the number of regulatory components and interactions increases.We use model checking techniques to verify sophisticated dynamical properties, expressed as temporal logic queries, resulting from the model regulatory structure in the absence of kinetic assumption. We demonstrate the power of this approach by analysing and revisiting a model of the molecular network controlling mammalian cell cycle progression (Fauré et al. Bioinformatics, 2006).It enables a systematic analysis of model properties, the delineation of model limitations, and the evaluation of various refinements and extensions based on recent experimental observations. In particular we detail the inhibitory role of Skp2, and further emphasizes the multifunctional role for the cell cycle inhibitor Rb. The resulting model accounts for the main irreversible transitions between cell cycle phases, the sequential activation of cyclins, and is consistent with documented perturbations (e.g. combinations of loss- or gain-of-function mutations).This core cell cycle model can then be used as a module in more comprehensive cellular models for cell cycle deregulation underlying cancers.	Systems poster	Fundamental
P_Sy062	602	Jessica Hu, Francesco Russo, Jose Maria Gonzalez-Izarzugaza and Søren Brunak	Jessica Hu	Mechanisms of non-oncogene addiction	During oncogenesis, cancer cells harbor vast amounts of genetic alterations including point mutations, deletions, amplifications, rearrangements and translocations. Some of these genetic alterations provide the cell with oncogenic properties such as unlimited proliferation, self-sufficiency, angiogenesis, metastasis and resistance to apoptosis etc. To achieve these new phenotypes, the cancer cell is put under numerous stresses such as mitotic, proteotoxic, metabolic, oxidative, DNA damage and replication stress. It has therefore been hypothesized that cancer cells are more dependent on stress support pathways for their survival than normal cells. This mechanism has been termed non-oncogene addiction (NOA). Genes and pathways involved in NOA are not directly oncogenic, but secondary processes, which are critical to maintaining a stressful environment, rendering them as potent candidate drug targets.An example of NOA that has been exploited therapeutically is the BRCA2-PARP complex. It has been shown that breast cancer cells with a deficiency in the BRCA2 gene are highly dependent on the DNA stress protein poly ADP-ribose polymerase (PARP). Inhibiting PARP can push the DNA-damage stress response of the tumor cell towards a point of lethality, which means that PARP-inhibitors can be used in cancer treatment for people with BRCA1 and BRCA2 deficiency. However, few studies have identified the entire spectrum of NOA genes or investigated their mechanisms using computational approaches.In this project, we characterize NOA gene mechanisms using a systems biology approach and develop a machine learning approach that can predict NOA genes to unravel novel therapeutic targets for precision cancer treatment.	Systems poster	Biotechnology
P_Sy063	743	Sander Rodenburg, Michael Seidl, Francine Govers and Dick de Ridder	Sander Rodenburg	Metabolic network construction of the Phytophthora infestans – tomato pathosystem	The metabolism of a pathogen reflects its relation with its host, as many pathogens lack essential metabolic reactions themselves, but instead exploit metabolites of their host. Therefore, reconstructing a genome scale metabolic network for a pathosystem (i.e. pathogen - host system) can provide insights into their relationship at the metabolic level, to facilitate the development of novel control methods. In our case, we study the interactions between the notorious plant pathogenic oomycete Phytophthora infestans and its host, tomato. Using network analyses and constraint-based modelling, we aim to identify the exchanged metabolites between this pathogen and host. This will provide the foundation for an integrative genome-scale interaction model of this pathosystem at the metabolic, protein and miRNA level.The metabolic networks of P. infestans and tomato are reconstructed based on the KEGG database. Flux balance analysis will be used to find essential reactions that characterize the metabolism of both species, and will reveal what metabolic reactions are involved in infection. Time-series transcriptome data of both species will allow us to infer context-specific flux distributions during the infection process. Currently, draft metabolic networks have been constructed for both P. infestans and tomato1, and network inconsistencies such as network gaps and stoichiometric errors are being resolved. In the near future biomass compositions must be determined and further model refinement will predict inter- and intra-species transport reactions.1. Yuan, H. et al. (2015) Plant J.	Systems poster	Agro-Food
P_Sy064	874	Weronika Wronowska, Krzysztof Gogolewski, Marcin Kostecki and Anna Gambin	Weronika Wronowska	Metabolomic scale analysis of the mechanism of C2 ceramide induced cell death	Ceramide, a bioactive sphingolipid, is known to stimulate the cell death and suppress the cell proliferation. However, there are conflicting reports about the mechanism and the nature of Ceramide induced cell death. Ceramide has been proven to induce neuronal cells apoptotic death through the mitochondrial dependent pathway. In the contrary, ultrastructural analysis of Ceramide treated NB16 neuroblastoma cells revealed 75 % loss of cell viability mainly due to the development of necrotic cell death. Determination of cell death type that prevails in the above mentioned cases, is of particular importance for the selection of adequate bioassays for the measurement of cell viability.We studied the influence of C2--ceramide, which is the exogenous cell-permeable C2--ceramide, on viability of Neuroblastoma SH-SY5Y cell lines. Transcriptomics data from Affimetrics Human Gene 2.1 ST array was analysed. Using a modified iMat method for transcriptomics data integration with global reconstruction of Human metabolome (Recon2), we have generated a case specific model that allows us an analysis of metabolic bases of ceramide induced death.We have demonstrated that inhibition of cell growth is related to general disregulation of lipids metabolism as well. We also used our neuroblastoma metabolic model to analyse the effects of the inhibition and activation of transcriptomic factors selected using IPA analysis of transcriptomics data in particullar SMARCA4, NUPR1, and KDM5B. The results of our analysis are currently under biological validation.	Systems poster	Fundamental
P_Sy065	316	Aliaksei S Vasilevich, Shantanu Singh, Aurélie Carlier and Jan de Boer	Dennie Hebels	Mining for osteogenic surface topographies using machine learning techniques	The TopoChip, a microtopography screening platform, enables the assessment of cell response to 2176 unique topographies in a single high-throughput screen (Unadkat, PNAS, 2011; Hulsman, Acta Biomater, 2015; Reimer, Sci Rep, 2016). Here, we show that surface topographies can be used to modulate the ALP expression in human mesenchymal stromal cell (hMSCs), an early marker of osteogenesis. More specifically, cell response to topography was captured by high-content imaging (Hulsman, Acta Biomater, 2015) and multiparametric "profiles" of cellular response were obtained. Multiple replicates of each topography were used to estimate the median level ALP expression, and we were able to successfully find surfaces that resulted in high and low ALP expression. To predict the cellular response based on surface topography parameters machine learning methods were used. The data were split into training and testing sets in a 3:1 proportion respectively, focusing on 100 high- and low-scoring topographies. In the training step, we performed a 10-fold cross-validation to obtain optimal parameters for each classifier. The caret package in R was used to perform the analysis. We tested several classifiers and identified random forest as most precise, which obtained an accuracy of 96% in distinguishing between high and low ALP expression, on the held-out test set. In summary, the combination of our screening methods and machine learning algorithms open new avenues to design surfaces with desired properties for variable applications. Our next step will be to find a surface topography that induces maximum ALP expression based on our screening data.	Systems poster	Health
P_Sy066	575	Mathias Cardner, Nathalie Meyer-Schaller, Gerhard Christofori and Niko Beerenwinkel	Mathias Cardner	Modelling gene regulatory networks during EMT	Metastasis causes an overwhelming majority of cancer deaths. Epithelial–mesenchymal transition (EMT) of tumour cells has been suggested to play a crucial role in metastasis. In the epithelial state, cells tend to be stationary, whereas in the mesenchymal state, cells are invasive and migrate through the bloodstream. Supposing that EMT is the only mechanism of metastasis, then metastasis could in principle be prevented by inhibiting EMT. However, recent findings indicate that metastases can form without EMT in lung and pancreatic cancer, but that EMT nevertheless contributes to chemoresistance.In this project we analyse the signalling network of transcription factors and micro RNAs during epithelial–mesenchymal transition of mouse mammary cells. While the cells are induced to undergo EMT, the transition is blocked at intermediate stages by RNA interference against a set of transcription factors and miRNAs. With the experimental data gathered during these system perturbations, we apply Nested Effects Models to infer the signalling network through downstream effects of the interventions. Furthermore, kinetic data on the unperturbed EMT allows us to estimate at what stage during the EMT process a certain RNAi intervention halts the transition. The robustness of the inferred network is assessed using a bootstrap technique, which yields a measure of confidence in the estimated signalling between transcription factors and micro RNAs. Permutation tests are used to evaluate the statistical significance of the resulting network, assessing how well it explains the data compared to random or permuted networks.	Systems poster	Fundamental

P_Sy067	505	Céline Hernandez, Wassim Abou-Jaoudé, Romain Roncagalli, Bernard Malissen and Denis Thieffry	Céline Hernandez	Modelling of T cell co-inhibitory pathways to predict anti-tumour responses to checkpoint inhibitors	In recent years, it has been recognized that T cells have a reduced ability to eliminate cancer cells and that expression of co-inhibitors at their surface accounts for their compromised function. By blocking the functions of these co-inhibitors, therapeutic antibodies (checkpoint inhibitors) have become standard treatment for metastatic melanoma, leading to a revival in the study of T cell co-inhibitors. However, our understanding of the immunobiology of T cell co-inhibitors and of their harmful role during anti-tumour responses is incomplete.To overcome these limitations, we aim at defining at the system-level the mechanisms through which co-inhibitory molecules such as PD-1 and CTLA-4 impede T cell functions. To reach our goal, we combine high-throughput analysis with computational methods in order to map TCR co-signalling pathways and predict cell responses to perturbations. First, we focus on the development of comprehensive annotated molecular maps through both manual curation of scientific literature and automated queries to public databases. These maps will be used as a support to analyse high throughput data, which will be in turn used to refine them. Next, these maps will be translated into a sophisticated logical model recapitulating the observed cellular behaviour. Finally, this model will be used to predict cell response to single or multiple perturbations, paving the way to the delineation of novel experiments.This integrated system-level view of the mechanisms of action of key T cell co-inhibitors in cancer will further provide a rationale for designing and evaluating drugs targeting T cell co-inhibitory pathways in anti-cancer immunotherapy.	Systems poster	Health
P_Sy068	447	Andreas Hillmann, Martin Crane and Heather Ruskin	Andreas Hillmann	Moving HIV Treatment Interruption Modelling to a new level - a computational approach	Antiretroviral Therapy remains the only effective remedy for HIV infection to date. Different drug types can be used to block the viral replication cycle although a cure is unattainable, due to persistence of viral reservoirs and pockets. Treatment interruptions, though inevitable and ad hoc in many cases, yield unpredictable risks in terms of viral rebound and emergence of drug resistant mutations.To achieve better understanding of the effect of treatment interruptions on infected organisms, a model is constructed, based on the Cellular Automata (CA) formalism, derived from earlier work of Zorzenon dos Santos et al. (2001). The work focuses on lymph tissue, in which major harbours for HIV infected cells occur in the susceptible cells of a fixed matrix. A regular grid, representing a section of lymph tissue, is populated with susceptible and infected cells. Neighbourhood interaction and propagation to adjacent sites is permitted, based on both deterministic and stochastic rules. Treatment response is realised using alternative model rules and response functions, taking into account specific mechanisms of different drug classes. Perturbations of the emerging structures of infected cells, subjected to the several treatment options, are observed. A new method for quantification of the viral reservoir is presented, based on spatial properties of the CA, and augmenting mean field approximation. Additionally, the impact, of different treatment initiation time steps and interruption schedules, is analysed. Finally, implications are assessed, of different update schemes for model behaviour and performance, and their extension to large-scale simulations.	Systems poster	Fundamental Health
P_Sy069	588	Ferran Briansó, Teresa García-Berrocoso, Joan Montaner and Alex Sánchez	Ferran Briansó	Multivariate Methods for the Integrative Analysis of Transcriptomics and Proteomic Data in a Study on Ischemic Stroke	Ischemic stroke is one of the main causes of death and disability, whose genetic risk is likely to be multigenic and influenced by environmental factors. For that reason, an integrative, multi-omics approach can be very useful to gain deeper knowledge of its genetic components. In this project, human brain tissue samples have been processed to obtain protein and gene expression values. First each type of data has been analyzed independently, using standard bioinformatics protocols, to select features separating affected and non-affected tissue. From the resulting lists of selected features, two distinct approaches for projection-based multivariate analysis have been applied to characterize the two groups. Annotations to standard biological databases (Gene Ontology) have been used as a method for merging information in a common space. The first approach used Regularized Canonical Correlation Analysis and Sparse Partial Least Squares, with the R mixomics package, to provide a visualization of individual relationships between features. It could also be used for variable selection, but did not allow the addition of biological information. The second approach used Multiple Coinertia Analysis and Gene Set Analysis, with the mogsa R package, for presenting samples, features and its associated biological information in a common projection space. However it could not perform variable selection. In summary, both approaches have been able to show distinct but complementary aspects of relations between genes and proteins that could not have been unveiled separately, which is the main goal of these type of integrative omics data analysis.	Systems poster	Health
P_Sy070	431	Natalia Rubanova and Nadya Morozova	Natalia Rubanova	Network analysis of genome-wide loss-of-function screens and its application to cancer research	Genome-wide loss-of-function screens use RNAi (RNA interference) technique to systematically induce individual sequence-specific gene knockdown followed by a readout assay specific to biological process to assess the phenotypic outcome. The result of the screen is the list of genes (hit list) that consists of the most important genes for the biological process. However the functional role of up to 50% of the genes in the hit list might be unknown. The most probable explanation is that there still exist unknown pathways in the process that could be triggered by silencing these genes.We developed a new systems biology tool to predict these pathways. The aim of the tool is to identify short paths (simple chain graphs) that are most likely belong to the biological process of interest in a particular biological system (cell type). The search is done in a global integrated network that consists of human protein-protein interaction, transcription factor interaction, miRNA-gene interaction, transcription factor-miRNA interaction, metabolic data. The likelihood is calculated based on the path's type, strength and over-represented rate among all pathways from each gene in the hit list to so called "final molecular key points" that are molecular events responsible for phenotypic realization of the process.We applied this tool to the analysis of human lung adenocarcinoma screen and human muscle differentiation screen and found previously unknown paths for these processes.	Systems poster	Fundamental
P_Sy071	870	Mushthofa Mushthofa, Martine De Cock and Kathleen Marchal	Mushthofa Mushthofa	Network-based prediction of cancer drug response	Cancer is a complex disease driven by different types of genomic aberrations that give rise to different subtypes of cancer. Due to these diverse possible genetic mechanisms by which the cancer phenotype arises, different ways of inhibiting / inducing death on the cancer cells are needed. This fact gives rise to the different kinds of cancer drugs available for these different subtypes. Through the recent development of genome-sequencing technology, it has become increasingly possible to facilitate personalised medical treatment for cancer patient by obtaining the genomic data of the patient's tumour sample predicting which (types of) drugs will most likely give the optimal result based such data.However, the problem of finding a model that can reliably be used to predict such results is far from obvious. Many approaches have been proposed in which we can identify the relevant genomic biomarkers to predict the response of a certain drug, based on known response data.In this work, we investigate a computational model to predict the drug response of cancer cells which integrates genomic and transcriptomic features of the cells, as well as prior knowledge such as genes and proteins interactions. Given a set of cells with known features and response towards a particular drug, this network-based method derives a set of features most likely to be predictive of the drug's response. We investigate the application of such method to publicly-available data, including the Cancer Cell Line Encyclopedia and the Genomics of Drug Sensitivity in Cancer.	Systems poster	Health
P_Sy072	879	Alexander Spirov, Ekaterina Myasnikova and David Holloway	David Holloway	NONLINEAR MODEL FOR MODULAR GENE EXPRESSION CONTROL, APPLIED TO SPATIAL PATTERNING OF THE DROSOPHILA GENE HUNCHBACK	Genes are frequently regulated in complex manners, necessitating modelling approaches which go beyond linear 'gene-to-gene' interactions and address the modularity of cis-regulatory regions and alternate transcription initiation sites. In particular, sharp expression patterns indicate that gene regulation involves nonlinear transcription factor kinetics. We propose a methodology for approaching this problem, using the example of the multiple CRMs and two transcripts (P1 and P2) found in the Drosophila hunchback (hb) gene, one of the first genes expressed in the embryo. We develop a differential equations model for transcription which takes into account the cis-regulatory architecture of the gene.Non-linearity problem is addressed through biologically substantintated mechanisms. For example, gene regulation is described by the Hill-like activation function taking into account the binding cooperativity, or more complicated concentration dependent response, such that the type of regulation changes depending on the regulator concentration.With the experimental evidence for independent control of expression through different CRMs, and for the different expression of P1 and P2 transcripts, we use a building-up approach of independent components for a model of hb expression. The model is first fitted to wild-type, and then extended to account for the mutant data.Our approach offers a means whereby the knowledge we have of the hb cis-architecture can be turned into a dynamic model for expression patterning. This allows us to both explore the dynamics of gene regulation at a realistic level (nonlinear gene-gene interactions), as well as producing preliminary predictions of transcript-specific patterning which could be corroborated by future experiments.	Systems poster	Fundamental
P_Sy073	420	László Kupcsik, Jiajia Xu, Guangyong Zheng, Xin-Guang Zhu, Dirk Inzé and Christian Hermans	László Kupcsik	Oilseed rape co-expression network reveals mechanisms of root architecture adaptation	Brassica napus (oilseed rape) is an allotetraploid plant with a large (840 Mb) genome. This increasingly important cash crop has poor Nitrogen Use Efficiency (NUE). Our rationale is to ameliorate NUE by redesigning a more branched root system that explores a larger soil volume in order to prevent fertilizer run-offs. We tackle that challenge with a predictive breeding by transcriptomics. The strategy identifies genes whose expression levels are correlated with lateral root proliferation across different nitrate supplies.The root transcriptome response of six doubled haploid accessions with contrasting lateral root branching degrees, were analysed. One-week-old seedlings were grown in hydroponics at 2 mM nitrate and transfered to 0.2 or 20 mM for 24 h. Two low and high N co-expression networks were built from >6,000 differentially regulated transcripts, using GPU computing (two Nvidia Tesla K20m accelerators with 4992 cores under CUDA framework).The resulting network modules were characterised by gene ontology terms. For some modules, the gene expression patterns greatly differed according to the root morphology. In particular, a module related to amino acid metabolism is currently investigated (some genes involved: PAB2, LIF2, UMAMIT31). The dataset also allows us to examine the paralogous gene specialisation in an allotetraploid crop genome.	Systems poster	Agro-Food
P_Sy075	878	Friederike Ehrhart, Kristina Hettne, Marco Roos, Leopold G. Curfs and Chris T. Evelo	Friederike Ehrhart	Pathway analysis of Rett syndrome omics data; an example of knowledge driven molecular data analysis in the rare disease domain	Although being a rare disease, Rett syndrome (RTT) is one of the most important neurodevelopmental disorders in females. RTT females are generally suffering severe intellectual disability and motor impairments. Cause of RTT is a mutation in one gene, MECP2, a central signalling gene which acts as global and gene specific transcription regulator, chromatin shaper, responsible for alternative splicing, and epigenetic imprinting (http://www.wikipathways.org/instance/WP3584). Using systems biology approaches to reveal mechanisms of rare diseases is highly interesting because it draws attention on the whole system instead of single endpoints. Pathway and network analysis approaches bring experimental data together with existing knowledge. In this application poster we demonstrate the use of bioinformatics tools and methods to visualise changes in brain tissue transcriptome samples from RTT females and controls and we add metabolomics information to see how this reflects the transcriptomics results in regulation of glutathione and lipid metabolism.	Systems poster	Health
P_Sy076	577	Dries De Maeyer, Bram Weytjens and Kathleen Marchal	Dries De Maeyer	PheNetic: Integration analysis of parallel omics data sets using multiple evidence networks	As more and more data is gathered from experimental biology, functional interpretation and analysis from these results becomes harder and harder. This not only because of the vast size of the generated data but also the generation of parallel data sets and the integration of multi omics data sets. Analyzing these data requires combining these results in the light of public knowledge of the molecular mechanisms observed in the experiment. To this end biological networks pose ample opportunities for integrating not only parallel results but also multi omics datasets. To this end we have developed and successfully applied the PheNetic framework over the last years which we made available as a web server (De Maeyer, 2015) and was applied to prioritize causal mutations from driver mutations in evolution experiments (De Maeyer, 2016). Here we present an extension of the framework which utilizes high performance compute clusters to better integrate multi omics data from parallel experiments in a query driven approach which can be defined by the user. In specific we illustrate how multiple types of biological networks can be integrated to interpret complex multi omics data sets resulting in a better explanations of experimental data.D. De Maeyer, B. Weytjens, L. De Raedt, K. Marchal, Network-based analysis of eQTL data to prioritize driver mutations. Genome Biol. Evol. , 1–36 (2016).D. De Maeyer, B. Weytjens, J. Renkens, L. De Raedt, K. Marchal, PheNetic: network-based interpretation of molecular profiling data. Nucleic Acids Res. 43, W244–W250 (2015).	Systems poster	Biotechnology
P_Sy077	880	Simone Lederer and Tjeerd Dijkstra	Simone Lederer	Predicting Compound Synergy in the DREAM Challenge	Synergy occurs when drugs combined are more effective than each drug by itself. The aim of the AstraZeneca-Sanger Drug combination Prediction DREAM Challenge is to predict a synergy score of pairs of drugs, given their individual effect and high-throughput data from cell lines.There are 118 drugs tested pairwise in 85 cancer cell lines. 11,759 pairs were screened which is large relative to 91 the number screened in a previous challenge [1].Next to the drug name, target, drug chemical structure descriptions and mono-therapeutic information, genomic, epigenomic and transcriptomic data on the cell line was provided. For prediction of the synergy score, we use both linear regression and kernel regression. This dual regression approach allowed us to use features that could be calculated directly for each drug-pair cell-line combination and more complex information that could only be expressed as a similarity (between cell lines or between pairs of drugs). In detail, we first used ordinary linear regression with the mono-therapeutic parameters as features. Secondly, residuals from the first step are modeled using Gaussian Processes with sub-kernels that capture similarity between cell lines and between pairs of drugs based on their chemical structure and the pathways used.We found the maximal effect of drugs to predict synergy: drugs with stronger maximal effects are more likely to show a stronger synergy score.[1] M. Bansal, et al. A community computational challenge to predict the activity of pairs of compounds. Nat. Biotechnol., 32(12):1–12, 2014.	Systems poster	Fundamental
P_Sy078	430	Otoniel Rodríguez-Jorge, Linda Aimara Kempis-Calanis, Darely Yarazeth Gutiérrez-Reyna, Oscar Ramírez-Pliego, Wassim Abou-Jaoudé, Morgane Thomas-Chollier, Angélica Santana and Denis Thieffry	Otoniel Rodríguez-Jorge	Predictive logical modelling of TLR5 and TCR cooperation for CD4 T cell activation.	Toll-Like Receptor 5 (TLR5) recognises the flagellin monomer, a component of the flagella of many bacteria. Flagellin is being evaluated as a vaccine adjuvant given its ability to induce pro-inflammatory signalling cascades in a variety of cell types. In T cells, flagellin directly provides a co-stimulatory signal to the T cell receptor-mediated (TCR) signals leading to proliferation and IFN-γ production. This study aim to model the cross-talk between TLR5 and TCR signalling pathways leading to CD4 T cell activation. We used the software GINsim to generate and analyse the models. First, we constructed distinct logical models for TCR and TLR5 signalling pathways based on published information and high-throughput data. Next, we validated these models using experimental data obtained in our lab. Then, we reduced these models and merged the reduced versions to obtain a model accounting for the cross-talk between the two pathways. We perform a dynamical analysis of these different models to delineate the specific effects of the cross-talk between TLR5 and TCR pathways on CD4 T cell activation. We then stimulated highly purified naïve CD4 T cells by cross-linking the CD3 molecule, in the presence or absence of flagellin, and evaluated the activation of IKKαβ, c-JUN and CREB by flow cytometry. Experimental data was used to further improve our merged model. The resulting model provides novel insights in the effects of flagellin co-stimulatory signals on CD4 T cell activation.	Systems poster	Fundamental Health
P_Sy079	523	Emre Guney	Emre Guney	ProXide: Proximity based drug side effect detection	Drug safety issues remain as one of the major bottlenecks in drug development, contributing to more than 20% of the clinical trial failures. Though effective, experimental screening of drugs for large scale adverse effect detection is currently unattainable. Computational methods, relying on drug and side effect similarity to train classifiers, offer a cost-effective alternative but typically fail to provide a universal solution over different data sets. In this study, we present, ProXide, a purely interactome topology based drug side effect prediction method. ProXide uses the network-based proximity of drug targets to side effect modules (proteins likely to induce the side effects) to quantify the likelihood of the drug-side effect association. Our analysis of 819 FDA approved drugs and 537 side effect modules in the interactome shows that proximity can discover known drug side effects with prediction accuracycomparable to similarity-based approaches. Furthermore, combined with drug chemical and target similarity, proximity based adverse effect detection is robust against data incompleteness and outperforms any single method individually. We demonstrate how ProXide can pinpoint novel drug-side effect associations on several case studies.	Systems poster	Health
P_Sy080	707	Katarzyna Rzosińska, Dorota Formanowicz and Piotr Formanowicz	Katarzyna Rzosińska	Quantitative model of processes associated with formation of atherosclerotic plague based on continuous Petri net	Formation and stabilization of atherosclerotic plaque is a complex and still not fully understood process. Recent studies over the course of inflammatory states have revealed, inter alia, the existence of a subpopulation of monocytes and two functional phenotypes of macrophages - M1 and M2. Faced with this knowledge essential for formation and stabilization of atherosclerotic plaque seem to be disturbances of blood vessel homeostasis. This results in a tilting of monocyte-macrophage axis towards differentiation of macrophages M1, infiltration of macrophages to the inflammation site and hyperproliferation as a result of stabilization of atherosclerotic plaque.For a better understanding of the processes and factors affecting the inflammatory process in atherosclerosis we have applied a system approach and created a model of this process using continuous Petri nets. These nets are an extension of the classical Petri nets where a marking of a place is a real number instead of an integer.The use of continuous Petri nets for modeling and analysis of the processes related to atherosclerotic plaque formation allowed to describe some crucial properties of the studied biological system and to draw interesting conclusions on the basis of the formal analysis of the model.This research has been partially supported by the Polish National Science Centre grant No. 2012/07/B/ST6/01537.	Systems poster	Health

P_Sy081	614	Sara Ciucci, Yan Ge, Alessandra Palladini, Víctor Jiménez Jiménez, Luisa María Martínez Sánchez, Susanne Sales, Andrej Shevchenko, Steven W. Poser, Oliver Otto, Maik Herbig, Andreas Androutsellis-Theotokis, Jochen Guck, Mathias J. Gerl and Carlo Vittorio Cannistraci	Sara Ciucci	Revealing discriminative network functional modules in omic sciences: an easy and fast unsupervised multivariate method	Recent advances in high-throughput techniques made available a large number of omic datasets and consequently required the development of network-inference methods, to describe the biomedical systems under analysis. Precisely, reverse-engineering or inferring networks are the process of identifying associations between omic entities behind the complexity of a biosystem. However, the usually employed correlation-based network methods only highlight linear associations between omic features, but do not pinpoint the main actors that are responsible for the perturbation of the system under analysis. On the other hand, building correlation networks between significant molecules pre-selected by means of a univariate statistical test is an 'over-the-counter' solution that neglects the multivariate and collective mechanisms at the basis of the omic system complexity.In this study, we developed a new unsupervised multivariate algorithm named PC-corr, to reveal linear discriminative correlation network modules, where combinations of features contribute to distinguish two or more conditions under analysis. PC-corr is based on a preliminary unsupervised sample exploration by Principal Component Analysis (PCA) of the omic dataset, hence it is particularly suited for the analysis of small-size datasets or pilot studies in which knowledge uncertainty represents an important issue.Our results demonstrate that PC-corr method can be used as a valid and fast tool for multivariate inference of discriminative associations between the variables of an omic dataset and consequently for identifying the most relevant omic network modules. PC-corr can thus represent a new tool in precision medicine for the definition of combinatorial and multiscale biomarkers in complex omic data.	Systems poster	Health
P_Sy082	501	Thierry Lombardot, Anne Morgat, Kristian Axelsen, Lucila Aimo, Nevila Nouspikel, Steven Rosanof, Joseph Onwubiko, Elisabeth Coudert, Nicole Redaschi, Lydie Bougueleret, Ioannis Xenarios and Alan Bridge	Thierry Lombardot	Rhea, an expert curated resource of biochemical reactions for enzyme annotation and genome-scale metabolic modeling	Rhea (www.rhea-db.org) is a comprehensive and non-redundant resourceof expert curated biochemical reactions designed for the functionalannotation of enzymes and the description, analysis and reconciliationof genome-scale metabolic networks.Rhea describes enzyme-catalyzed reactions (the IUBMB enzymenomenclature list), transport reactions and spontaneously occurringreactions using species from the ChEBI (Chemical Entities ofBiological Interest) ontology of small molecules. Rhea reactions areextensively curated with links to source literature and are mapped toother publicly available metabolic resources such as MetaCyc/EcoCyc,KEGG, Reactome and UniPathway.Rhea reactions are used as a reference for the reconciliation ofgenome-scale metabolic networks in the MetaNetX resource(www.metanetx.org) and also serve as the basis for the computationalgeneration of the library of lipid structures and analytes inSwissLipids (www.swisslipids.org). External resources that use Rheainclude the EMBL-EBIs Enzyme Portal and MetaboLights resource as wellas the Microscope platform for genome annotation developed byGenoscope.Here we describe recent and forthcoming developments in Rhea, whichinclude the development of a new website, substantial growth of Rheathrough sustained literature curation, and the addition of parent-childreactions relationships to complement the IUBMB enzyme classification.At the time of writing, Rhea (release 73, of May 2016) includes 9179unique reactions involving 8044 unique reaction participants, curatedfrom 8798 unique PubMed identifiers.	Systems poster	Fundamental
P_Sy083	642	Marek Blazewicz, Giovanni Felici, Aleksandra Swiercz, Daniele Santoni, Marcin Jaroszewski, Agnieszka Zmienko and Marta Kasprzak	Marek Blazewicz	Searching for Common Patterns in Biological Datasets	In the recent decades the need for efficient and automated algorithms toprocess and analyze the outcome of biological experiments is constantlygrowing. In the Post-Genomic Era the information is no longer a bottleneck inanalyzing the genomes or seeking for solutions of diseases. The Internet inthe present time is filled with enormous amount of information shared byscientists. The main challenge has shifted now to perform efficient analysisof datasets taken from different experiments and conducted by various groups ofscientists. Ideally, the process should minimize the effort of domain experts.In presented research authors have focused on optimizing the process of dataintegration. The presented algorithm is able to recognize common patterns indifferent biological networks and to find relations between genes preserved ondifferent levels of biological processes. In this research authors have mainlyfocused on discovering functionally related groups of genes. The implementationof the algorithm is available as an open-source application written inlow-level C code for fast and efficient execution. The application is animprovement of authors former work [1].[1] Daniele Santoni et al., An Integrated Approach (CLuster AnalysisIntegration Method) to Combine Expression Data and Protein–Protein InteractionNetworks in Agrigenomics: Application on Arabidopsis thaliana, OMICS: A Journalof Integrative Biology. January 2014	Systems poster	Biotechnology
P_Sy084	609	Iryna Nikolayeva, Kevin Bleakley, Anavaj Sakuntabhai and Benno Schwikowski	Iryna Nikolayeva	Simple enough biomarkers predict a complex disease phenotype	During dengue virus outbreaks, many hospitals are overcrowded with patients due to potential complications that occur in 5% of the patients several days after hospital admission. Being able to predict at admission which patients will develop complications would make it possible to focus limited medical resources on those patients that require them.Based on clinical and transcriptomic data from blood serum in 41 patients at hospital admission, we find simple, yet biologically powerful predictors of dengue complications from omics data.Specifically, we use a generalization of linear models that describe the disease severity using an ensemble of monotonic functions of pairwise transcript measurements. Our implementation allows, for the first time to our knowledge, genome-wide screening and goes beyond classical linear and logistic models; it allows to model relations such as "AND" and "OR" between genes. Features are easier to interpret. And our ensemble model allows us to control the complexity of our predictor.We present the methodology, results from our genome-wide screen for biomarkers for dengue severity, and compare its predictive performance to the state-of-the art biomarker prediction methods.	Systems poster	Health
P_Sy085	437	Kevin Schwahn, Romina Beleggia, Nooshin Omranian and Zoran Nikoloski	Kevin Schwahn	Stoichiometric correlation analysis: towards principles of metabolic functionality from metabolomics data	Motivation: Recent advances in metabolomics technologies have resulted in high-quality (time-resolved) metabolic profiles with an increasing coverage of metabolic pathways. These data profiles represent read-outs from often non-linear dynamics of metabolic networks. Yet, metabolic profiles have largely been explored with regression-based approaches that only capture linear relationships, rendering it difficult to determine the extent to which the data reflect the underlying reaction rates and their regulation.Results: Here we propose an approach termed Stoichiometric Correlation Analysis (SCA) based on correlation between positive linear combinations of non-linearly transformed metabolic profiles. The non-linear transformation is due to the observation that metabolic networks can be modeled by mass action law and its derivatives. Using time-resolved metabolic profiles from Arabidopsis thaliana and Escherichia coli, we show that SCA can be used to quantify the difference in regulatory couplings underlying the stringent response in these model organisms. By using SCA with data from natural variation of wild and domesticated wheat accessions, we demonstrate that that the domestication is accompanied by loss of regulatory couplings. Therefore, application of SCA to metabolomics data from natural variation in wild and domesticated populations provides a mechanistic way to understanding domestication and its relationto metabolic networks.	Systems poster	Fundamental
P_Sy086	353	Junli Liu, Marc Knight and Keith Lindsey	Junli Liu	Systems Biology of calcium and hormone signalling in plant cells	Calcium and hormone signalling systems are two important systems in regulating many aspects of plant development. Experimental evidence accumulated over many years has shown that different environmental stimuli may induce different changes in cellular calcium concentration. However, little is known about how different calcium signatures are decoded to produce specific gene expression responses. Similarly, experimental evidence accumulated over many years has shown that the quantitative properties of the auxin gradient in Arabidopsis root are important in regulating different developmental features. However, little is known about how formation of auxin gradient is regulated by other hormones for controlling plant development. Here we show that systems approaches can be used to integrate experimental data into systems models, to examine the actions of calcium and hormone signalling systems in plant cells and to reveal the underlying regulatory mechanisms. We use two examples we have been developing to demonstrate the applications of systems approaches to signalling systems in plant cells. First, we study how different calcium signatures are actually decoded by a transcription factor, CAMTA, to produce specific gene expression responses. We establish information flow from calcium signatures to CAMTA-regulated gene expression responses by combining experimental data with systems modelling. Second, we develop a spatiotemporal hormonal crosstalk model that describes the integrated action of three hormones (auxin, ethylene and cytokinin). We reveal that a hormonal crosstalk network regulates the emergence of patterns and levels of hormones and gene expression in wild-type and mutants.	Systems poster	Agro-Food
P_Sy087	758	Juan Carlos Higareda-Almaraz, Michael Karbiener, Florian Pauler, Stephan Herzig and Marcel Scheideler	Juan Carlos Higareda-Almaraz	Systems-level network analysis in white-to-brite adipocyte conversion	Obesity with more than 600 million obese adults has reached pandemic levels worldwide1. A positive energy balance, with energy intake exceeding energy expenditure, leads to an increase in adipose tissue mass and consequently to obesity. Adipose tissue is a complex organ which has traditionally been divided into two distinct types: white adipose tissue (WAT); and brown adipose tissue (BAT)2. A new type of thermogenic adipocytes, called "brown-in-white" ("brite"), has been recently discovered. Brite adipocytes are able to burn fat and carbohydrates via non-shivering thermogenesis and are derived from white adipocytes upon cold exposure3, therefore opening the door to novel therapeutic approaches against obesity. However, the regulatory mechanisms that govern white-to-brite adipocyte conversion remain to be elucidated.Our objective is to explore the genomic network that governs the white-to-brite adipocyte conversion in human. By analyzing the human adipocyte transcriptome with a RNA-Seq approach, we currently identify the signature of differentially expressed genes which are involved in the transition from white to brite adipocytes. By using differential network biology, we have identified regulators that might be involved in the early "britening" program and act as pleiotropic genes.1. WHO. (2014); Fact sheet N°311.2. Cannon and Nedergaard. (2004), Physiol Rev, 84:277-359.3. Lee et al. (2012). Cell metabolism,15:480-91.	Systems poster	Health
P_Sy088	603	Joanna Ziobro, Paweł Błażej and Paweł Mackiewicz	Joanna Ziobro	The increase in the antigen concentration may lead to recovery of patients – computer simulation studies	The dynamic development of medicine and inventing new treatments requires modifying current and creating new models describing immunological reactions of human organism. The immune response is a complex set of defensive reactions which includes the antigen recognition, its neutralization and elimination. There are two mechanisms of immune response which are interdependent: cellular and humoral response.We analyzed model which describe the humoral type of human immune response. The Marchuk's model describes reaction between antibody and antigen. This model takes into account the delay of proliferation of lymphocytes in respect of the antigen presentation [1]. We analyzed the stability of the stationary states describing the healthy state of an organism and a chronic disease. The stability of the first state does not depend on the delay. However, the delay can be important in the case of chronic state.We presented a method analyzing the system of delay differential equations [2]. This type of differential equations describes well the complex biological processes such as responses to infection that often occurs with some delay. The system of delay differential equations has several features complicating the analysis more than in case of the systems of ordinary differential equations. We studied the stability of the chronic state depending on the time of delay and initial antigen concentration.[1] G.I. Marchuk, R.V. Petrov, A.A. Romanyukha, G.A. Bocharov, J. theor. Biol., 1991, 151, 1-69[2] F.M. Asl, A.G. Ulsoy, J. Dyn. Syst. Meas. Cont., 2003, 125 (2), 215-223	Systems poster	Biotechnology
P_Sy089	494	Lucila Aimo, Robin Liechti, Nevila Hyka-Nouspikel, Anne Niknejad, Anne Gleizes, Lou Götz, Dmitry Kuznetsov, Fabrice David, F. Gisou van der Goot, Howard Riezman, Lydie Bougueleret, Ioannis Xenarios and Alan Bridge	Alan Bridge	The SwissLipids knowledgebase for lipid biology	Lipids are a large and diverse group of biological molecules involved in membrane formation, energy storage, and signaling. The lipid complement or lipidome of an individual cell, tissue or organism may contain tens of thousands of lipid structures, whose composition is tightly regulated in response to changes in cellular signaling and nutritional status. The perturbation of lipidome composition in diseases such as cancer, hypertension, allergy, diabetes and degenerative diseases highlights the growing importance of lipids as biomarkers and potential diagnostic tools. While modern analytical methodologies such as high-throughput tandem mass-spectrometry provide a high-level overview of lipidome composition, a more complete understanding of the biological roles of lipids requires the integration of lipidomic data with other types of biological knowledge. To facilitate this task we have developed a knowledge resource for lipids and their biology – SwissLipids. SwissLipids provides a hierarchical classification that links mass spectrometry analytical outputs to over 300,000 potential lipid structures, as well as expert-curated data on enzymatic reactions, interactions, functions, and localization, along with supporting links to primary literature and source text. SwissLipids uses reference namespaces and ontologies such as UniProtKB, ChEBI, Rhea and Gene Ontology (GO), and links to matching structures in other reference metabolite databases such as LIPID MAPS and HMDB. In summary, SwissLipids provides a reference namespace for lipidomic data publication, exploration and hypothesis generation. SwissLipids is updated daily with new knowledge and all data is freely available to search or download from http://www.swisslipids.org/.	Systems poster	Fundamental
P_Sy090	719	Quentin Da Costa, Fabienne Guillaumond, Claire Rioualen, Sadia Belonbi-Djefaflia, Victoire Gouirand, Julie Roques, Isabelle Crenon, Eric Mas, Sophie Vasseur and Ghislain Bidaut	Quentin Da Costa	Time-dependent interactome – transcriptome analysis of PDAC tumorigenesis	Pancreatic ductal adenocarcinoma (PDAC) is the most intractable with a 5-year survival below 6 months and therefore represents the most fatal disease among solid cancer.Because of PDAC hallmarks, pancreatic cancer cells must harbor metabolic pathways essential to maintain, under fuel source limitation, cellular bioenergetic and integrity allowing tumor cell growth and dissemination. Based on transcriptomic profile of metabolic changes occurring during PDAC progression in mouse models, we propose to identify metabolic pathways contributing to these tumoral processes.Time –ITI (Interactome-Transcriptome Integration, Garcia et al., 2012), a network-based analysis algorithm, was developed to identify differentially expressed gene modules by integrating mouse global protein-protein interaction network (interactome) on top of gene expression measured on 4, 6 and 9 weeks old mice pancreas with Affymetrix oligo arrays. Then interactome deregulated regions (subnetworks) were identified by individually considering each gene as a potential seed and aggregate recursively its neighbors on the basis of a score measuring the differential of gene expression for a given timepoint. Neighbors are aggregated until subnetwork score could not be improved. Differentially expressed subnetworks identified for each timepoint are then combined in order to obtain a global view of selected subnetwork during tumorigenesis. We statistically validated subnetworks by measuring null distributions of scores for random interactomes and gene expression.Time-ITI features a reporting tool for results analysis. Examples of deregulated subnetworks in PDAC metabolism, (lipids, amino acids and glycosylation) will be presented.	Systems poster	Health
P_Sy091	727	Adrien Fauré and Takeyuki Tamura	Adrien Fauré	Towards a new definition of circuit functionality	In the wake of the seminal work of René Thomas, the notion of functionality in a regulatory network has focused on the asymptotic behavior "generated" by simple positive and negative circuits: Thomas' conjectures state that a positive circuit is necessary for multistationarity; a negative circuit, for sustained oscillations; and functionality has been loosely defined as the property of a positive or negative circuit that produces the corresponding behavior.More precisely, in the logical formalism, a circuit is said to be functional when all its arcs are functional, and its sign is the product of the signs of the arcs. Different definitions of circuit functionality then arise depending on the region of the state space where the arcs are required to be functional (1). Unfortunately, current definitions only allow proof of Thomas conjectures for very restricted conditions, if at all. Moreover, major questions remain unanswered, including the full decomposition of a network into functional modules and the very definition of what "generate" exactly means.In an attempt to clarify those issues we are currently investigating the possible role of symmetry in the dynamics of a model as a marker of functionality. Focusing on the behavior of a circuit in pairs of mirror states, we introduce new definitions and conjectures to connect the presence of circuits in a logical regulatory graph with symmetric patterns in the state transition graph.(1) Comet et al., 2013. Bull Math Biol. 75(6):906-19. Doi: 10.1007/s11538-013-9829-2	Systems poster	Fundamental
P_Sy092	442	Juris Viksna, Alvis Brazma, Karlis Cerans, Dace Ruklisa and Thomas Schlitt	Juris Viksna	Towards experimental validation of models of gene regulatory networks	We have previously developed a model of the GRN of lambda phage. This model is based on hybrid system (HSM) formalism [1] and allows to predict the rearrangements of genome that lead to altered biological behaviours. Here we describe three rearrangements of lambda phage genome that, according to the predictions of the model, should lead to biological behaviours that are different from the known possible behaviours of non-mutated types. Such behavioural differences are experimentally measurable, thus essentially we are proposing experiments that allow to validate the correctness of the developed HSM model. We also assess the practical feasibility of performing such experiments.The current lambda phage HSM model is derived from a very detailed semi-formal description [2] and a number of earlier mathematical formalizations developed by us. The analysis of this model shows that it allows only two 'attractors' corresponding to the known biological behaviours of lysis and lysogeny. However, these stable behaviours are expected to change, if the ordering of affinities of binding sites are modified. In comparison with our previous work, here we present:- description of the possible behaviours of mutated versions of lambda phage and how these behaviours can be determined experimentally;- description of three suggested genome rearrangements that should lead to observably different behaviours;- assessment of feasibility of performing the suggested genome rearrangements.[1] A.Brazma et al. Modeling and analysis of qualitative behaviour of gene regulatory networks. LNCS 7699, 51-66, 2015.[2] H.McAdams, L.Shapiro. Circuit simulation of genetic networks. Science 269, 650-656, 1995.	Systems poster	Fundamental
P_Sy093	391	Theresia Conrad, Olaf Kniemeyer, Thomas Krueger, Sebastian G. Henkel, Axel A. Brakhage, Reinhard Guthke and Joerg Linde	Theresia Conrad	Transcriptomic and proteomic response of Aspergillus fumigatus to caspofungin	Aspergillus fumigatus is one of the most common human pathogenic fungi and causes a wide range of infections. One therapeutic option is the use of the lipopeptide antifungal drug caspofungin. It specifically targets the fungal cell wall by inhibiting the synthesis of the polysaccharide β-(1,3)-D-glucan [1]. Caspofungin exposure induces a compensatory stress response including the adaption of the gene expression and consequently, the protein synthesis and secretion. This study aims to detect potential relationships between the fungal transcriptomic and proteomic response to caspofungin.The transcriptome of the A. fumigatus strain A1163 was measured by RNA-Seq at 0h, 0.5h, 1h, 4h and 8h after caspofungin treatment. Proteome samples, taken at 0h, 4h, 8h (synthesised proteins) and at 0h, 8h (secreted proteins) after treatment, were analysed by liquid chromatography-mass spectrometry. Significantly, differentially regulated mRNA, synthesised and secreted proteins were considered to analyse the shared response of the different cell response levels to caspofungin and the shared association with several stress response pathways.The comparison of the different response levels shows that the overlap of these can only account for a small part of the overall response. However, the pathway analyses demonstrate the association of the mRNA, synthesised and secreted proteins with shared, caspofungin-associated pathways. Some of these pathways can only be significantly associated by combining transcriptome and proteome.Hence, in contrast to the separate consideration of response levels, the combination of different levels provide a deeper insight into the overall fungal response to caspofungin.[1] Altwasser et al. PloS One, 10(2015):e0136932.	Systems poster	Fundamental

POSTER LIST
ORDERED ALPHABETICALLY BY POSTER TITLE
GROUPED BY THEME/TRACK

THEME/TRACK: TRAINING
Poster numbers: P_Tr001 - 017

Poster number	EasyChair number	Author list	Presenting author	Title	Abstract	Theme/track	Topics
P_Tr001	591	Oscar Torreño Tirado, Oswaldo Trelles, Michael T. Krieger and Alex Upton	Michael T. Krieger	An overview of training in the Spanish ELIXIR node	The Spanish National Bioinformatics Institute ('Instituto Nacional de Bioinformática' (INB)) is part of the Carlos III Health Institute ('Instituto de Salud Carlos III', ISCIII). The mission of the INB is to provide bioinformatics support to Spanish research institutions and companies.The INB has actively participated in the creation of ELIXIR. It acts as a transmitter of ELIXIR developments for the benefit of national projects, and promotes the use of INB systems and tools at European level. The Bitlab group, part of the Computer Architecture Department of the University of Malaga, is one of the INB nodes and acts as the training coordinator of Spain in ELIXIR. The INB is heavily involved in training, providing its expertise in organising training events. The training collaboration is bidirectional, with ELIXIR providing materials and certifying them to ensure the quality of the training sessions.A wide range of training courses have been offered in the last year across the whole node. This includes courses in NGS, R programming, proteomics, genomics, and Galaxy. This highlights the breadth in the training offer, with a number of training courses planned in the node for the coming year. This includes a two-day High Performance Computing (HPC) workshop at the University of Malaga in October 2016. The first day will provide an introduction to HPC with introductory practical exercises, whilst the second day will present HPC use cases from the bioinformatics and biomedicine domains. Along with the other planned courses, this demonstrates INB's continued commitment to bioinformatics training.	Training poster	Training
P_Tr002	445	Vera Matser, Cath Brooksbank, Rossen Apostolov, Adam Carter, Alexandre Bonvin, Mark Abraham and Emiliano Ippoliti	Vera Matser	Applying competency profiling of user groups to develop a training programme in Computational Biomolecular Research	Life Science research has become increasingly digital and has a direct influence on our daily life in areas such as health and medical applications, drug discovery, agriculture and food industry. It is one of the largest and fastest growing communities in need of high-end computing, leading to an increasing number of life science researchers who are not computing experts but who need to use complicated computationally intensive biomolecular modelling tools.BioExcel is a newly launched Centre of Excellence for Biomolecular Research aimed at supporting these academic and industrial researchers in the use of high-performance computing (HPC) and high-throughput computing (HTC).To make sure that the biomedical research communities can fully profit from the training offered through the new Centre of Excellence we will be determining the training needs for three user groups (Entry Level Users, Expert Users and System Administrators) by drafting a competency profile. The competencies have been determined with the aid of the community and sent out for wider consultation.To enrich the competency profile we will, for each competence, define what an individual will need to know and what skills they need to have to exhibit competence in a specific area, as well as list what behaviours are suited and unsuited to an individual with that particular competency - so that individuals can assess their own competence in each area and select appropriate training.The competencies will be mapped against existing training and new training courses and material will be developed where gaps are revealed.	Training poster	Training
P_Tr003	746	Janick Mathys, Christof De Bo and Alexander Botzki	Janick Mathys	Bioinformatics Training at VIB: laying the cornerstones for life scientists to survive in data-intense biotech research	Set up in response to the increasing importance of bioinformatics in biotechnology research, VIB's BioInformatics Training and Service (BITS) facility provides trainings, software support and services that contribute to the generation of useful biological knowledge. The facility gives basic and intermediate trainings to the life sciences research community in Belgium, focusing on the fields of bioinformatics, statistics, omics data analysis, programming and support of the software that the VIB offers to its members.Most trainings are part of a training track that teaches researchers to perform complete bioinformatics analyses of the data they generate. The trainings are very hands-on with a focus on applications, many allow participants to work on their own data. To support this practical approach we work in small groups on laptops provided by VIB. The training material (slides, tutorials and exercises) are available on our wiki web site for consultation and we capture the lectures of some courses as videos.To increase the visibility of the BITS training courses, all training web sites are semantically tagged according to bioschemas' Life Science training material specification. In collaboration with the ELIXIR training coordinator group, these metadata are scraped from our web site and deposited in the TeSS training portal (https://tess.elixir-uk.org/).Currently, the BITS training team is setting up a Belgium ELIXIR training network and participates in organising some ELIXIR-BE training courses like Data Carpentry course in November 2016. To even further grow our training network, we organise experience and trainer exchanges within the Core For Life alliance (http://coreforlife.eu).	Training poster	Training
P_Tr004	615	Sandrine Perrin, Victoria Dominguez Del Angel, Jonathan Lorenzo, Jean-François Gibrat and Christophe Blanchet	Victoria Dominguez Del Angel	Cloud Computing Training at French ELIXIR node (French Institute of Bioinformatics)	Cloud Computing presents a new approach to allow the development of elastic, distributed and highly scalable resources.The French Institute of Bioinformatics set up a Cloud Computing Infrastructure which offers services, software, database and computing resources.Education and Training are key components of the IFB-Infrastructure. IFB-core, the national hub of IFB offers training courses to educate the community on how to use the IFB Cloud for analyses and methodological developments in bioinformatics.IFB-core offers 3 training modules to teach life-science scientists to adopt the IFB Cloud.The modules build progressively to cater for the needs of general and advanced audiences.1) In the "Cloud basic usage" module, the attendees learn to deploy the appropriate application in the cloud for analyzing their data. This module is dedicated to non-users of the command line interface. Demonstration on available applications e.g. Galaxy, RStudio and Virtual-desktop technology.2) In the "Cloud advanced usage" module, the attendees learn to deploy complex bioinformatics applications, including multiple virtual machines in a cluster, to install new integrate public data collection, and manage data with NFS virtual disks. We demonstrate automatic installation tools, such as Approver, Docker and how to build a cluster with SGE, Spark or Torque.3) In the "development of the appliances" module, developers learn how to create appliances according to a guideline of good practices. All created appliances will increase the Catalogue. Developers are accompanied during the creation of the appliance.The modules are regularly scheduled throughout the year.	Training poster	Training
P_Tr005	837	Kim Gurwitz, Shaun Aron, Sumir Panji, Suresh Maslamoney, Pedro Fernandes, David Judge and Nicola Mulder	Kim Gurwitz	Distance-based online Bioinformatics training in Africa: the H3ABioNet experience	Africa is not unique in its need for basic Bioinformatics training for individuals from a molecular biology background. However, unique logistical challenges in Africa, most notably access to Bioinformatics expertise and Internet stability, must be addressed in order to meet this need on the continent. H3ABioNet (www.h3abionet.org), the Pan African Bioinformatics Network for H3Africa, has therefore developed an innovative, free Introduction to Bioinformatics course taking these challenges into account. A distance-based learning model has been selected for this 3 month course (July-September 2016) to increase access to expert African and European Bioinformatics trainers covering several Bioinformatics topics, including: Databases and Resources; Genomics; Linux; Sequence Alignment; and Phylogenetics. Classrooms with a total of >350 participants are hosted at 19 institutions, across 11 African countries, in order to provide local administrative and academic support. Classroom selection was based on certain infrastructure criteria, including: computer resources; Internet access; and availability of local teaching assistants. Although lectures are delivered live to remote sites via an online platform, to ensure that classroom success does not rely on stable Internet, classrooms can watch pre-recorded and pre-downloaded lecture videos, as well as work through practical assignments on the lecture content, during biweekly contact sessions. Lecture recordings are available on the course website http://training.h3abionet.org/IBT_2016/. While trainers are available via video conferencing to take questions during contact sessions, online 'question and discussion' forums, hosted on the course management platform, are also available. This distance based model, developed for a resource limited setting, could easily be adapted to other settings.	Training poster	Training
P_Tr006	685	Teresa K Attwood, Pamela Black, Marie-Claude Blatter, Cath Brooksbank, Pedro L Fernandes, Nicola Mulder, Patricia M Palagi, Gabriella Rustici, Maria Victoria Schneider and Celia W G van Gelder	Pedro L Fernandes	GOBLET's Bioinformatics Learning, Education and Training Activities	The Global Organisation for Bioinformatics Learning, Education and Training (GOBLET: http://mygoblet.org) was established to provide a global, sustainable support structure to foster international communities of bioinformatics trainers and trainees. The activities of GOBLET are carried out through committees, which have independent but overlapping focus areas. The Learning, Education and Training (LET) Committee primarily focuses on providing resources for bioinformatics trainers. Here we describe some of the recent activities and resources developed by the LET Committee: (i) A set of consensus descriptors for training materials to ensure that materials are consistently described with a minimum, standard amount of information. This brings a strong improvement in discoverability, shareability and traceability of training materials. (ii) The development of core competencies together with the ISCB Education Committee, and how these can be used to elaborate bioinformatics curricula and training materials appropriate for different audiences. (iii) Our e-learning activities in bioinformatics from the perspective of discoverability of existing e-learning materials and the development of new materials. For these activities we partnered up with other networks and organisations with similar goals.	Training poster	Training
P_Tr007	791	Sarah L Morgan, Richard Grandison, Katrina Costa, Lee Larcombe and Cath Brooksbank	Cath Brooksbank	Providing bioinformatics training for established researchers	The EMBL-EBI training programme provides face-to-face and online learning opportunities focused on accessing public biodata, analysing large data sets and interpreting the results of bioinformatics experiments. Although our major audience is early-stage researchers, we receive frequent requests from experienced researchers needing to enhance their own bioinformatics competency and enable their labs to benefit from data-centric approaches to research. Finding training appropriate to their busy schedules and specific needs has proved challenging, so we have developed two new courses tailored specifically to the needs of (1) bench-based industrial discovery scientists and (2) principal investigators.Bioinformatics for discoveryWorking with the EMBL-EBI Industry programme and with support from the BBSRC we have developed a blended learning module to enable discovery biologists to incorporate bioinformatics-based approaches into their research projects. A two-day face-to-face workshop is followed by an online, workflow-based component that must be completed within 6 months. During this period trainees continue to interact with each other and their instructors via discussion boards and scheduled online discussionsBioinformatics for Principal InvestigatorsAs more early-stage researchers make bioinformatics a major component of their research projects, PIs of classically bench-based research labs have turned to us for guidance on principles and challenges of data acquisition and analysis, and on how best to support their teams. In June 2016 we delivered our first course aimed specifically at PIs. The course covered the fundamentals of bioinformatics and data management combined with discussions on various options for enhancing bioinformatics competency in their teams.	Training poster	Training
P_Tr008	402	Antonio Fabregat, Konstantinos Sidiropoulos, Guilherme Viteri, Florian Korninger, Steven Jupe, Phani Garapati, Peter D'Eustachio, Lincoln Stein and Henning Hermjakob	Antonio Fabregat	Reactome: A curated knowledgebase of biomolecular pathways	Reactome (http://www.reactome.org) is a free, open-source, curated and peer-reviewed knowledgebase of biomolecular pathways. Its aim is to provide intuitive bioinformatics tools for visualisation, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education.Pathways are built from connected "reactions" that encompass many types of biochemical events. Reactions are derived from literature and must cite a publication that experimentally validates them. Pathways are authored by expert biologists and peer reviewed before incorporation into the database. 9,584 reactions in Reactome cover 9,238 human gene products (12,527 including IntAct interactors), supported by 22,838 literature references.Users can search for proteins or compounds and see details of the complexes, reactions and pathways they participate in. Pathway diagrams allow users to examine the molecular events that constitute the steps in pathways and to view details of the proteins, complexes and compounds involved.Different forms of pathways analysis can be performed with the Reactome analysis tools. Users can submit a list of identifiers for overrepresentation analysis or submit quantitative datasets, such as microarray data, for expression analysis. Results of these analyses are overlaid onto the Pathways Overview and Diagram Viewer for easy navigation and interpretation.Interaction data from multiple resources can be used to expand pathways. Interactors from IntAct are included by default in the search feature and can be taken into account in the analysis service. Finally, pathways or all Reactome content can be downloaded in many formats including TSV, CSV, PDF, SBML, BioPax and PSI-MITAB.	Training poster	Training
P_Tr009	403	Konstantinos Sidiropoulos, Antonio Fabregat, Guilherme Viteri, Florian Korninger, Peter D'Eustachio, Lincoln Stein and Henning Hermjakob	Guilherme Viteri	Reactome: New services and widgets to ease third-party integration	Reactome (http://www.reactome.org) is a free, open-source, curated and peer-reviewed knowledge base of biomolecular pathways. It aims to provide intuitive bioinformatics tools for visualisation, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education. Thus, the mainstays of its software development are usability and responsiveness from the user's point of view, likewise modularity and reusability from the developer's side.Reactome offers web services and widgets (http://goo.gl/koRvhp) to facilitate integration in third-party software. One service provides database access while the other performs overrepresentation and expression analysis as well as species comparison. Widgets for the Pathways Overview and Pathway Diagrams are provided for JavaScript and GWT. Both widgets overlay the results of the Analysis Service.Protein-protein or protein-chemical interactions can be used to extend pathways beyond Reactome's curated content. IntAct is the default resource but all other PSICQUIC databases can be selected and in addition, users can submit custom interactions. Interaction data from IntAct are also included in the Reactome main search and the Analysis Service, helping users identify pathways of interest.In summary, Reactome has facilitated data integration by providing easy-to-use services and reusable widgets. Several resources such as OpenTargets (https://www.targetvalidation.org/), ChEBI (https://www.ebi.ac.uk/chebi/), BluePrint (http://dcc.blueprint-epigenome.eu/), PRIDE (http://www.ebi.ac.uk/pride/archive/), PINT (http://sealion.scripps.edu/pint/) and IP2 (http://goldfish.scripps.edu) have already integrated these services and widgets.	Training poster	Training
P_Tr010	416	Thanh Le Van, Matthijs van Leeuwen, Ana Carolina Fierro, Dries De Maeyer, Jimmy Van den Eynden, Lieven Verbeke, Luc De Raedt, Kathleen Marchal and Siegfried Nijssen	Thanh Le Van	Simultaneous discovery of cancer subtypes and subtype features by molecular data integration	Motivations:Subtyping cancer is key to an improved and more personalized prognosis/treatment. The increasing availability of tumor related molecular data provides the opportunity to identify molecular subtypes in a data-driven way. Molecular subtypes are defined as groups of samples that have a similar molecular mechanism at the origin of the carcinogenesis. The molecular mechanisms are reflected by subtype-specific mutational and expression features.Data-driven subtyping is a complex problem as subtyping and identifying the molecular mechanisms that drive carcinogenesis are confounded problems. Many current integrative subtyping methods use global mutational and/or expression tumor profiles to group tumor samples in subtypes but do not explicitly extract the subtype-specific features.We therefore present a method that solves both tasks of subtyping and identification of subtype-specific features simultaneously. Hereto our method integrates mutational and expression data while taking into account the clonal properties of carcinogenesis. Key to our method is a formalisation of the problem as a rank matrix factorisation of ranked data that approaches the subtyping problem as multi-view bi-clustering.Results:We introduce a novel integrative framework to identify subtypes by combining mutational and expression features. The incomparable measurement data is integrated by transformation into ranked data and subtypes are defined as multi-view bi-clusters. We formalise the model using rank matrix factorisation, resulting in the SRF algorithm.Experiments on simulated data and the TCGA breast cancer data demonstrate that SRF is able to capture subtle differences that existing methods may miss.	Training poster	Training
P_Tr011	747	Sarah Morgan, Teresa K Attwood, Brane Leskosek, Gabriella Rustici and Allegra Via	Brane Leskosek	Surveying training provision, needs and capacity across ELIXIR nodes and EXCELERATE use-cases to map skill transfer routes in Europe	Across Europe, the availability of bioinformatics training opportunities varies greatly. Whilst the need for bioinformatics competencies (and hence training) is well recognised, the ability to provide such training is not yet well developed in all countries. In this context, ELIXIR training is vital for promoting the transfer of skills from nodes where training is more developed to those where specific competencies are needed. In order to develop a broad-based ELIXIR training programme, we must first draw a map of training needs, offerings and capacity across European countries.In 2014, the Society for Experimental Biology, in association with members of GOBLET (Global Organisation for Bioinformatics Learning, Education & Training), surveyed bioinformatics training needs amongst life scientists worldwide. In the context of ELIXIR, we want to gain a node perspective, and a deeper, more detailed view of current bioinformatics training provision, needs and capacity. To this end, the first task of the Training Platform's Train-the-Trainer subtask was to launch a survey across all ELIXIR nodes to determine 1) which nodes have the greatest need to increase their training capacity; 2) the subject areas where capacity needs to be built; 3) which nodes are currently delivering training initiatives; and 4) the course delivery methods employed, including the development/use/need for VM and cloud-based systems and e-learning.The outcome of our analysis will be used to delineate a map of training activities and demands across ELIXIR, and to draft a white paper reporting our recommendations to optimise the transfer of skills between ELIXIR nodes.	Training poster	Training

P_Tr012	790	Rafael Hernández-De-Diego, Tomas Klingström, Hadrien Gourlé, Etienne P. de Villiers, Ana Conesa and Erik Bongcam-Rudloff	Hadrien Gourlé	The eBioKit, a stand-alone educational bioinformatics platform	Bioinformatics skills have become essential for many research areas; however, the availability of qualified researchers is usually lower than the demand, a situation that especially affect developing countries. For many developing countries, bioinformatics has been a strategic area of investment in life science. Initial efforts in developing countries have generated hubs of excellence located in the bigger or more affluent countries. Extensive training is however necessary to provide the research professionals with the necessary skills to analyze the virtual mountains of data generated by modern research.The eBiokit was developed as a response to the lack of reliable internet connections and the short time available to visiting researchers conducting hands-on training at workshops or short courses. The eBiokit is a portable bioinformatics educational platform whose main purpose is to eliminate the dependence on the Internet, by offering locally a wide range of services, tools and databases widely used in genomic research, as well as documentation and training material. The architecture of the eBioKit has demonstrated to be an excellent balance between portability and performance, making the eBioKit a great educational tool but also providing small research groups with a platform to incorporate bioinformatics analysis in their research. The eBioKit has proven itself to be an excellent teaching platform in training activities for the African Bioinformatics Network (H3ABioNet) as part of the initiative H3Africa, the SAnBio initiative, The Biotechnology for Central Africa (BeCA) hub, the International Glossina Genome Initiative, Institute of Biochemistry, Molecular Biology and Biotechnology (IBMBB), and many others.	Training poster	Training
P_Tr013	559	Youri Hoogstrate, Saskia Hiltemann, Dave Clements, Bjoern Grüning, Andrew Stubbs, Hans-Rudolf Hotz and Galaxy Training Network	Leon Mei	The Galaxy Training Network: centralizing resources for galaxy trainings	The Galaxy Training Network is an international initiative supporting and developing all aspects of training around the Galaxy analysis platform for biomedical research. Scalability is a recurring challenge in all aspects of high-throughput computational biology, including training. There is far more demand for training than can be met by just in-person training by the core Galaxy Team. The Galaxy Taining Network supports the project by providing resources and centralizing the training efforts. As member of GOBLET (http://www.mygoblet.org/), the Galaxy Training Network takes part in the global coordination of Bioinformatics training.This poster will highlight resources that are available for teaching bioinformatics software in Galaxy and for using and administering Galaxy itself. The Galaxy Training Network unifies core project and community training efforts under one umbrella so that existing training resources become more easily and centrally available, and it makes it easier for new arrivals to get up to speed with training in their locations and communities. We will also highlight directories of tutorials/worked exercises, including up to date sample data, slide sets, videos, the new Galaxy Tours functionality and computational resources such as shared virtual machine images and Amazon Web Service Machine Images.	Training poster	Fundamental Training
P_Tr014	682	Gregoire Rossier and Patricia M. Palagi	Gregoire Rossier	The SIB PhD Training Network: an initiative to gather, connect and train PhD students in Bioinformatics	The SIB Swiss Institute of Bioinformatics created in 2007 the SIB PhD Training Network (TN), a community support for PhD students carrying out their research in bioinformatics or computational biology in Switzerland. The TN aims to foster interactions and exchanges among PhD students and to train them in the most up-to-date methods necessary for their doctoral research.Every year, the TN coordinates several training activities, e.g. graduate level courses, where the network members have a registration priority. Furthermore, we organize annual events such as an international seasonal school, usually held in the Swiss Alps, the “Best Practices in Programming” workshop and the TN Retreat. Every two years the “Bioinformatics in the Chalet” workshop offers a unique and challenging experience of building up from scratch a bioinformatics research project. All these are opportunities for students to exchange ideas about their research projects, to seek feedback and help from their peers, and for networking and developing new collaborations. Most of the TN training activities are part of the SIB Training courses’ portfolio, which can be found at www.sib.swiss/training.The SIB PhD Training Network was a pioneer PhD program in Switzerland and it is still unique in its domain in the country. It has seen near 350 students since the creation of the Network, and counts today close to 230 active members. Students and supervisors recently evaluated the pertinence of the TN and the conclusions of the survey will be presented in this poster.	Training poster	Training
P_Tr015	692	Diana Marek, Gregoire Rossier, Geoffrey Fucile, Walid H. Gharib, Frédéric Schütz, Marie-Claude Blatter and Patricia M. Palagi	Diana Marek	The SIB Swiss Institute of Bioinformatics Training Group: Supporting the development and sustainability of effective bioinformatics training	The SIB Swiss Institute of Bioinformatics has an extensive offer of bioinformatics training courses, involving computational biology methods, statistics, machine learning, computing techniques, and the analysis, management, and reproducibility of biological data. The significant increase in the number of SIB groups has expanded SIB’s resources and expertise, thus offering an opportunity to broaden the scope, scale, and diversity of SIB’s training portfolio. Our courses respond to an increasing demand for bioinformatics training towards ensuring that the Swiss and international scientific community make the best use of bioinformatics and SIB resources.The SIB Training Group teaches, coordinates, and supports courses in close collaboration with SIB members and international partners. In 2015, SIB ran over 50 events, training nearly 1000 participants. These achievements were made possible through a complete planning and organisational framework that fully supports trainers. It includes: an analysis of objectives and requirements, design of course context, content and format, definition of learning objectives and teaching strategies, promotion, an efficient registration system including online payment, a reactive helpdesk for participants, systematic assessment of course quality and learning outcomes, and smooth handling of all logistical/organisational aspects. Our group employs this very efficient training platform to encourage and facilitate participation of SIB members in training activities. Researchers contributing to SIB training can thus increase the visibility and impact of their research activities without the burdens of course logistics and organization. Through this collaborative effort, SIB’s training platform stays at the forefront of developments in bioinformatics to offer sustainable and effective training programs.	Training poster	Training
P_Tr016	672	Patricia M. Palagi, Erik Bogcam-Rudloff, Pedro Fernandes, Eija Korpelainen, Fran Lewitter, Gabriella Rustici, Maria Victoria Schneider, Celia W.G. van Gelder and Teresa K. Attwood	Patricia M. Palagi	Train-the-Trainer: GOBLET’s initiative to increase the provision of bioinformatics training in NGS	GOBLET is a global organisation that coordinates, shares and supports bioinformatics training activities worldwide, aiming to plug critical skills gaps, ultimately to facilitate the advancement of health- and life-science research. The focus of GOBLET’s Train-the-Trainer initiative is on setting up effective training courses to help plug known skills gaps, especially in the area of NGS data analysis. This initiative will help to share bioinformatics training expertise, experience and resources; train bioinformatics and life-science specialists; support life-science research; promote collaborations among scientists worldwide; build capacity in developing and developed countries.The programme will consist of workshops, which will take place on different continents (e.g., South America, Africa, Asia) and are expected to co-locate as satellite events to major conferences. Each workshop is organised around two main topics: 1) how to exploit NGS data, and 2) how to set up and deliver excellent training courses. Trainers (members of GOBLET, expert in the field) will teach on a volunteer basis. Workshop participants will commit to replicate the workshops at least once, driving an exponential effect.To deliver this ambitious project, GOBLET is seeking partners and sponsors interested in increasing the provision of bioinformatics training in the area of NGS, either as GOBLET collaborators to customise the programme to specific communities or to fund workshops at given locations. To learn more about this project, see http://www.mygoblet.org/content/fund-raising, contact frc@mygoblet.org, talk to us in GOBLET’s booth and visit this poster!	Training poster	Training
P_Tr017	776	Celia van Gelder, Sanne Abeln, Rita Azevedo, Luiz Olavo Bonino Da Silva Santos, Jeroen Engelberts, Rob W. W. Hooft, Mateusz Kuzak, Leon Mei, Marco Roos, Merijn van Rijswijk, Andrew Stubbs and Jaap Heringa	Celia van Gelder	Training efforts in the Netherlands: combining forces to provide data – related training for the life science research community	In this era of big data, new skills and competences are needed for life scientists, technologists and data experts. Many people with heterogeneous backgrounds have to be trained. By combining the education expertise present in the Netherlands we work towards establishing a comprehensive, internationally acclaimed and sustainable training and education course portfolio for Life Sciences Research & Technology with a focus on training in new technologies and data integration and stewardship. Our efforts cross bridges between disciplines, application domains, European research infrastructures (ESFRIs and e-infrastructures).Examples of our activities include trainings and training collaborations in:•Bioinformatics and Systems Biology (BioSB Research School)•Software and Data Carpentry (DTL, Netherlands eScience Center, SURFsara)•Metabolomics (DTL and Netherlands Metabolomics Centre)•Proteomics (DTL, Netherlands Proteomics Community)•NGS (DTL NGS Interest Group, Metagenomics Platform)•Galaxy (DTL, ELIXIR-NL, VU Amsterdam, LUMC, ErasmusMC, BBMRI, TraIT)•Bring Your Own Data (BYOD) workshops (DTL, ELIXIR-NL)•Data Stewardship, Data Management, FAIR Data training (DTL, ELIXIR-NL, LERU, Elsevier)•Defining competences and skill sets (DTL, ELIXIR-NL, EDISON)•HPC and Cloud (DTL, ELIXIR-NL, SURFsara)•Generic data science (ELIXIR-NL, RDA-CODATA, EDISON)For all this training areas, we are not only collaborating in the Netherlands, but are also actively engaging and aligning efforts with collaborators in Europe (e.g. VIB in Belgium and SIB in Switzerland) and abroad (e.g. GOBLET and RDA-CODATA). Furthermore, ELIXIR-NL is co-leading the ELIXIR Training Platform. All our training activities are open to all and we welcome new collaborations.	Training poster	Training