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, Elja Korpelainen, Kimmo Mattila and Diego Scardaci*	*Elja 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 TrailT-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 (catalitiocativity). 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, Phyliss 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 sequence 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 BainScope 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 centality, 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 Ozdinler, 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 retrieve 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 anTities 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 distributionon 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	EcosystemsHealth
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 naively 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 visualization 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 challengefaced 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