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

THEME/TRACK: GENOMES

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

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

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

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

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

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

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

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

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

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