






**23rd Annual International Conference on Intelligent Systems for Molecular Biology**  
**14th European Conference on Computational Biology**



An Official Conference of  
**iSCB**  
INTERNATIONAL SOCIETY FOR  
COMPUTATIONAL BIOLOGY

Google™ Custom Search

- HOME
- CAREER FAIR
- CONFERENCE ACCOMMODATION
- GENERAL INFO
- PROGRAM
- EXHIBITORS & SPONSORS
- SUBMIT
- COMMITTEES
- CONTACT
- PROMOTION
- PRESS PASS

**Award Winners - ISMB/ECCB 2015**

[Ian Lawson Van Touch Memorial Award](#)  
Outstanding Oral Poster Presentation Prize sponsored by University of California Berkeley Center for Computational Biology  
[F1000Research Poster Awards](#)  
[RCSB PDB Poster Prize](#)  
[Springer Outstanding Poster Award](#)  
[Wiley Poster Prizes](#)

---

**Ian Lawson Van Toch Memorial Award for Outstanding Student Paper**

**TP025: Identification of causal genes for complex traits**  
*Presenting Author:* Farhad Hormozdiani, University of California, Los Angeles, United States  
**Additional authors:**  
Gleb Kichaev, University of California, Los Angeles, United States  
Wen-Yun Yang, University of California, Los Angeles, United States  
Bogdan Pasaniuc, University of California, Los Angeles, United States  
Eleazar Eskin, University of California, Los Angeles, United States

Motivation: Although genome-wide association studies (GWAS) have identified thousands of variants associated with common diseases and complex traits, only a handful of these variants are validated to be causal. We consider “causal variants” as variants which are responsible for the association signal at a locus. As opposed to association studies that benefit from linkage disequilibrium (LD), the main challenge in identifying causal variants at associated loci lies in distinguishing among the many closely correlated variants due to LD. This is particularly important for model organisms such as inbred mice, where LD extends much further than in human populations, resulting in large stretches of the genome with significantly associated variants. Furthermore, these model organisms are highly structured, and require correction for population structure to remove potential spurious associations.

Results: In this work, we propose CAVIAR-Gene, a novel method that is able to operate across large LD regions of the genome while also correcting for population structure. A key feature of our approach is that it provides as output a minimally sized set of genes that captures the genes which harbor causal variants with probability . Through extensive simulations, we demonstrate that our method not only speeds up computation, but also have an average of 10% higher recall rate compared to the existing approaches. We validate our method using a real mouse high-density lipoprotein data (HDL) and show that CAVIAR-Gene is able to identify Apoa2 (a gene known to harbor causal variants for HDL), while reducing the number of genes that need to be tested for functionality by a factor of 2.

The software is freely available for download at [genetics.cs.University of California, Los Angeles.edu/caviar](http://genetics.cs.University of California, Los Angeles.edu/caviar)

[Top](#)




---




**Outstanding Oral Poster Presentation Prize sponsored by University of California Berkeley Center for Computational Biology**



**OP12: Pathway relevance ranking for tumor samples through network-based data integration**  
*Presenting Author:* Lieven Verbeke, Ghent University / iMinds / IBCN, Belgium

**Additional Authors:**  
Jimmy Van den Eynden, Ghent University / iMinds / IBCN, Belgium  
Piet Demeester, Ghent University / iMinds / IBCN, Belgium  
Kathleen Marchal, / iMinds / IBCN, Belgium  
Jan Fostier, / iMinds / IBCN, Belgium

**Abstract:**  
We present a new pathway relevance ranking method that is able to prioritize pathways according to the information contained in any combination of tumor related omics datasets. Key to the method is the conversion of all available data into a single network representation containing not only genes but also individual patient samples. Additionally, all data are linked through a network of previously identified molecular interactions. The performance of the new method is demonstrated by applying it to breast and ovarian cancer datasets from The Cancer Genome Atlas. By integrating gene expression, copy number, mutation and methylation data, the method’s potential to identify key pathways involved in breast cancer development shared by different molecular subtypes, is illustrated. Interestingly, certain pathways were ranked equally important for different subtypes, even when the underlying (epi)-genetic disturbances were diverse. The pathway ranking method was also able to identify subtype-specific pathways. Often the score of a pathway could only be explained by a combination of genetic and epi-genetic disturbances, stressing the need for a network-based data-integration approach. The analysis of ovarian tumors, as a function of survival-based subtypes, demonstrated the method’s ability to correctly identify key pathways, irrespective of tumor subtype. A differential analysis of survival-based subtypes revealed several pathways with higher importance for the bad-outcome patient group than for the good-outcome patient group. Many of the pathways exhibiting higher importance for the bad-outcome patient group could be related to ovarian tumor proliferation and survival.

<https://www.iscb.org/ismbeccb2015-program/award-winners>

**OP14: Novel brain-specific miRNA discovery using small RNA sequencing in post-mortem human brain****Presenting Author:** Christian Wake, Boston University, United States**Additional Authors:**

Adam Labadorf, Boston University, United States  
 Alexandra Dumitriu, Boston University, United States  
 Andrew Hoss, Boston University, United States  
 Richard Myers, Boston University, United States

**Abstract:**

MicroRNAs (miRNA) are short non-coding RNAs that regulate gene expression mainly through translational repression of target mRNA molecules. More than 2700 human miRNAs have been identified and some are known to display tissue-specific patterns of expression. Here, we use high-throughput small RNA sequencing to discover novel and possibly brain-specific miRNAs in 94 human post-mortem prefrontal cortex samples from patients with Huntington's disease and Parkinson's disease and normal neuropathology. Using a custom analysis pipeline, we identified 66 novel miRNA candidates that originate in both intergenic and intragenic regions of the genome. 21 of the candidate miRNAs show sequence similarity with known mature miRNA sequences and may be novel members of known miRNA families, while the remaining 45 may constitute previously undiscovered families of miRNAs that are specific to the brain. In a small number of these novel miRNAs, preliminary differential expression analysis between neurodegenerative disease and normal samples identified differences in expression. These results suggest that a portion of these novel miRNAs may not only be unique to brain, but may have a role in the neurodegenerative disease processes.

[Top](#)**F1000Research Poster Awards****OP15: Computationally efficient approach for novel transcript discovery across large RNA-seq dataset reveals glioblastoma-associated lncRNAs****Presenting Author:** Maria Laaksonen, BioMediTech, University of Tampere, Finland**Additional Authors:**

Antti Ylipää, 1) BioMediTech, University of Tampere 2) Department of Signal Processing, Tampere University of Technology, Finland  
 Janne Seppälä, BioMediTech, University of Tampere, Finland  
 Tommi Rantaperö, BioMediTech, University of Tampere, Finland  
 Kirsi Granberg, 1) BioMediTech, University of Tampere 2) Department of Signal Processing, Tampere University of Technology, Finland  
 Matti Nykter, BioMediTech, University of Tampere, Finland

**Abstract:**

Availability of RNA-sequencing data from human tumors and normal tissues has resulted in discovery of hundreds of tissue specific transcripts. Uncovering novel transcripts typically requires computationally expensive de novo transcriptome assembly and combination of assemblies across samples have proven challenging. To be able to search for new transcripts from large RNA-seq cohorts, we developed a computational approach that directly identifies unannotated genomic loci that are variably expressed within a sample set, or differentially expressed between two sample sets. These loci are then subject to gene structure analysis, allowing identification of full transcript structures in data driven manner. Our approach was validated by re-discovering a set of well annotated genes. We were able to correctly re-build known gene structures and identify the typical structural features of protein coding genes even when only a single exon of the gene was given as input.

We applied our approach to RNA-seq data of 169 primary glioblastoma samples from The Cancer Genome Atlas (TCGA). We identified 53 unannotated transcripts that did not contain good quality open reading frames, indicating that they were lncRNAs. The expression of 20 out of 22 high confidence lncRNAs was validated by PCR in at least one glioblastoma cell line. Clinical association analyses in the TCGA glioma cohort revealed that a subset of lncRNA expression profiles associates with patient survival, tumor grade and/or IDH1 mutation status. The functional analysis of lncRNA knockdowns was performed in glioblastoma cells to evaluate their significance in disease aggressiveness.

**OP17: Low concordance of differential DNA methylation analysis methods****Presenting Author:** Helen McCormick, Victor Chang Cardiac Research Institute, Australia**Additional Authors:**

Eleni Giannoulidou, Victor Chang Cardiac Research Institute, Australia  
 Jennifer Cropley, Victor Chang Cardiac Research Institute, Australia  
 Catherine Suter, Victor Chang Cardiac Research Institute, Australia

**Abstract:**

DNA methylation is one of the most widely used markers for the study of epigenetic contributions to phenotypic variation and disease. There are several methods for analyzing genome-wide DNA methylation data in common use, but there has been no rigorous evaluation of their performance. We have performed a systematic assessment and comparison of four packages: MethySig, methylKit, eDMR and DSS, using an empirical dataset of 12 reduced representation bisulphite sequencing libraries (6 test, 6 control). Surprisingly, we observed very low concordance among these commonly used model-based and binomial test-based approaches: using equivalent pre-processing and filtering parameters for each method, we found that the four methods identified significant differentially methylated cytosines at a concordance rate of less than 1%. Similarly low levels of concordance were observed with identification of differentially methylated regions using tiled data. Our study highlights the need for systematic approaches to reliable differential methylation analysis via data simulation. This concept of simulation will be discussed in the context of the growing implementation of epigenomic data in human medicine.

**OP19: Human paralog genes share regulatory elements and co-localize in the three-dimensional chromatin architecture****Presenting Author:** Jonas Ibn-Salem, Johannes Gutenberg University, Germany**Additional Authors:**

Miguel Andrade-Navarro, Faculty of Biology, Johannes Gutenberg University Mainz, Germany

**Abstract:**

Paralog genes arise from gene duplication events during evolution. The resulting sequence similarity between paralogs often leads to proteins of similar structures and functions in common pathways. Therefore it might be useful for the cell to have paralog genes co-regulated. However, since paralog genes often show also slightly different functions, for example alternative domains, it might be also useful for cells to exclusively express only one out of several paralogs for a specific function or response. Eukaryotic genes are regulated by binding of transcription factors to distal enhancer elements which perform looping interactions to the transcription machinery at gene promoters. We hypothesised that paralog genes share common regulatory mechanism that allows co-regulation and exclusive expression.

To test this hypothesis, we integrated paralogy annotations with genome-wide data-sets of enhancer-promoter associations and genome-wide

chromatin interaction data from Hi-C experiments in human cells.

With carefully sampled control data sets that take linear co-localisation of paralogs into account, we show that paralog gene pairs share a significant amount of common enhancer elements. Furthermore they are located significantly more often in the same topological association domain than expected and therefore cluster not only in the linear genome but also in the three-dimensional chromatin structure of the nucleus.

Together our results indicate that human paralog gene pairs share common regulatory mechanisms. We will further integrate expression data from different tissues and functional annotation of genes to support our findings that paralog genes tend to be expressed either collectively or exclusively depending on the cells functional needs.

**OP31: Detecting small structural variants with SoftSV using soft-clipping information**

**Presenting Author:** Christoph Bartenhagen, Institute of Medical Informatics, Germany

**Additional Authors:**

Martin Dugas, Institute of Medical Informatics, Germany

**Abstract:**

Numerous tools for the detection of structural variations (SVs) have been developed over the last years, including our own contribution called SoftSV. But there still remains a gap between small indels, which can be detected by gapped alignments, and large SVs (many hundred or thousand bp), which can be reconstructed by paired-end reads or read-depth information. Filling this gap remains difficult and often demands special algorithms for split-read alignments directly at the breakpoints, which only a few of the published tools do for this range of SVs.

We initially developed SoftSVs for large SVs and now expanded our approach to small and medium-sized deletions, tandem duplications and inversions (starting at 20bp). Similar to large rearrangements, we detect their exact breakpoints under the premise that no threshold filters SVs with low support or reads with low mapping quality or ambiguous mappings. Our greedy approach exploits any kind of soft-clipped alignment and reconstructs the breakpoint sequence just by comparing the soft-clipped reads at the start and end of an SV.

Using simulated and four real datasets from the 1000 Genomes Project, we evaluate the sensitivity and precision of SoftSV and four other tools. Our results show that sensitive and reliable SV detection is subject to many different factors like read length, coverage and SV type. SoftSV achieved sensitivities and PPVs between 80% and 100% consistently for all SV types on simulated datasets starting at 75bp reads and 10-15x sequence coverage, without requiring any parameter configuration by the user.

SoftSV is freely available at <http://sourceforge.net/projects/softsv>

**OP36: Site-specific evolution of selected post-synaptic protein complexes**

**Presenting Author:** Maciej Pajak, University of Edinburgh, United Kingdom

**Additional Authors:**

Martin Dugas, Institute of Medical Informatics, Germany

Clive R. Bramham, University of Bergen, Norway

T. Ian Simpson, University of Edinburgh, United Kingdom

**Abstract:**

Sequence conservation analysis of proteins belonging to the post-synaptic proteome (PSP) has previously revealed that key synaptic protein classes are present in primitive organisms preceding the emergence of nervous systems.

Recent studies suggest that evolution of the PSP may be responsible for the emergence of complex neural system function and behaviour but these analyses assess evolution only at the whole protein level.

We have developed an analysis workflow that integrates codon-resolution selection pressure estimates with domain and motif data to allow refinement of our understanding of domain-centric functionalisation of the PSP.

We show the application of this workflow to the Activity-regulated cytoskeleton protein (Arc) complex, a set of 26 Arc interacting proteins. Arc is highly conserved among placental mammals and plays a significant role in the post-synaptic density as a major regulator of long-term synaptic plasticity, the presumed molecular correlate of memory and learning.

Maximum likelihood phylogenetic inference for proteins of the Arc interactome, followed by site-by-site selection pressure analysis using a fixed effect likelihood methodology reveals a small set of positively selected sites as well as many regions under strong negative selection pressure. Mapping of these sites onto both known and predicted binding domains and post-translational modification sites allows inference of key domain-level functionalisation events during Arc complex evolution and provides a rational basis for prioritising regions for functional studies.

**RCSB PDB Poster Prize**

**OP30: Determining the winning SH3 coalition: how cooperative game theory reveals the importance of domain residues in peptide binding**

**Presenting Author:** Ashley Conard, United States

**Additional Authors:**

Elisa Cilia, Université Libre de Bruxelles, Belgium

Tom Lenaerts, Université Libre de Bruxelles, Belgium

**Abstract:**

Cell signaling relies on protein-protein and protein-peptide interactions involving signaling domains, which typically recognize specific peptide motifs. For instance, SH3 domains bind preferably to proline-rich amino-acid motifs. Phage-display experiments allow one to determine those motifs and whether surface or core domain mutants gain or loose preference for peptide motifs. Here, we present an approach utilizing the Shapley Value (SV) from Cooperative Game Theory to determine the importance of seven residues in the Fyn SH3's hydrophobic core. The core positions and the residues in those positions represent the players of a cooperative game in which the worth of each coalition is measured through its capacity to discriminate the binding and non-binding mutants for certain classes of peptides. The players (positions or residues) can be seen as the features of SH3 mutants in a binary classification task. Essentially, we use a feature selection method based on the SV to assign a pay-off to each core position and residue. We quantify their importance to promote peptide binding as well as their joint effects, and their interactions, represented through networks. Our results provide novel insights suggesting that the Fyn SH3 domain must contain different signatures of amino acids to promote binding to various peptide classes. This analysis highlights residue importance for proper domain function, which helps scale conservation profiles (e.g. WebLogo) by adding functionally relevant properties. These detailed pieces of information contribute an effective and novel approach to understanding the role core residues play, next to normally investigated binding-site residues, in binding specific peptides.

[Top](#)

**Springer Outstanding Poster Award****OP13: ContiBAIT: An R Package for Genome Finishing Using Strand-seq****Presenting Author:** *Kieran O'Neill, British Columbia Cancer Agency, Canada***Additional Authors:**Mark Hills, British Columbia Cancer Agency, Canada  
Peter Lansdorp, British Columbia Cancer Agency, Canada  
Ryan Brinkman, British Columbia Cancer Agency, Canada**Abstract:**

Strand-seq is a method for directional, low-coverage sequencing of DNA template strands in single cells. Taken together, strand-seq data from cells from the same organism provide genomic distance information. This can be used to improve the quality of early-build reference genomes made up of many contigs with no bridging sequence, firstly by grouping contigs from the same chromosome together, and secondly by ordering contigs within chromosomes. We present ContiBAIT, an R package for performing these tasks.

For grouping contigs into chromosomes, contiBAIT uses a custom clustering method based on a Chinese restaurant process. Contigs are then reoriented using a greedy algorithm which optimises for global inter-contig distance. Contig groups showing close strand similarity following reorientation are merged.

For ordering contigs within a putative chromosome, ContiBAIT computes the strand distance between all pairs of contigs. The problem then becomes one of finding the lowest-weight Hamiltonian path over the contigs, which can be reformulated into a travelling salesman problem. ContiBAIT then finds the best ordering of contigs using the TSP package.

To validate contig clustering, we applied ContiBAIT to an early build of the mouse genome (mm2), with coordinates lifted over to mm10. ContiBAIT was able to assign most contigs with sufficient read depth for strand-seq analysis to the correct chromosome (median F-measure=0.91).

To validate contig ordering, we applied ContiBAIT to artificial contigs sampled from mm10, of sizes 1MB, 500kB and 250kB. Some chromosomes were well-ordered (Pearson's rho=0.99), while others had large sections locally well-ordered but incorrectly ordered relative to each other.

[Top](#)**Wiley Poster Prizes****OP16: Tau Protein Related Acetylation of Histone 3 Lysine 9 in the Human Brain****Presenting Author:** *Hans-Ulrich Klein, Harvard Medical School, United States***Additional Authors:**Cristin McCabe, Broad Institute, United States  
Jishu Xu, Brigham and Women's Hospital and Harvard Medical School, United States  
David Bennett, Rush University Medical Center, United States  
Philip DeJager, Brigham and Women's Hospital and Harvard Medical School, United States**Abstract:**

Accumulation of tau proteins and amyloid- $\beta$  peptides in the brain are two hallmarks of Alzheimer's Disease (AD). Recent studies suggest that epigenetic mechanisms are likely to play a key role in the pathogenesis of AD. Here, we studied genome wide the active mark H3K9ac using ChIP-seq in 669 post-mortem human brain samples to detect alterations of the epigenome induced by tau. RNA-seq was performed for 500 samples to assess the effect on transcription. We considered modifications of local H3K9ac domains as well as large genomic regions and distinguished alterations primarily associated with tau from those with amyloid.

We identified 26,384 H3K9ac domains which primarily occurred at promoters (15,225) and enhancers (8,071). H3K9ac levels at promoters were positively correlated with transcription, even though H3K9ac alone was not sufficient for transcription. Tau protein loads were significantly associated with H3K9ac levels in 5,980 domains and had a much broader impact than amyloid (610 domains). Domains positively associated with tau showed a strong enrichment ( $p < 10^{-16}$ ) for binding sites of CTCF, which regulates chromatin structure. Indeed, we found large genomic regions showing concordant tau associated increases in H3K9ac. Average transcription in these regions was consistently up-regulated. Strikingly, effect sizes within the regions were highly correlated with the regions' proportion of open chromatin.

Our results demonstrate a genome wide change in chromatin structure in AD, which is mediated by tau. Tau is known to cause heterochromatin relaxation in Drosophila models. CTCF could be a key factor in the pathogenic process of chromatin opening.

**OP18: Computational method for detecting patterns of epigenetic changes from time series ChIP-seq data****Presenting Author:** *Petko Fizev, University of California, Los Angeles, United States***Additional Authors:**

Jason Ernst, University of California, Los Angeles, United States

**Abstract:**

Histone modifications associate with important regulatory regions such as promoters and distal enhancers that control the expression of genes. Time-course genome-wide maps of these epigenetic marks have become available in a growing number of biological settings including stem cell reprogramming and differentiation, adipogenesis, cardiac development, circadian rhythms, embryogenesis and lymphocyte development. However, our understanding of the underlying cellular processes remains limited, because the current bioinformatics tools often fail to utilize fully the temporal aspects of this data. Here, we present a novel computational method for systematic detection of major classes of spatio-temporal patterns of epigenetic changes. The method takes as input data from a series of chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) experiments for a single histone mark that are performed at consecutive time points during a given biological process. The method uses a probabilistic mixture model that explicitly models the spatio-temporal nature of the data to identify regions for which the mark either expands or contracts significantly with time or holds steady. Furthermore, it incorporates information about replicate experiments at each time point, which can increase the accuracy of the method. We present applications of the method on publicly available data from T-cell development, which help in understanding the underlying regulatory dynamics during this process.

[Top](#)