Andrea Rau, Cathy Maugis-Rabusseau; **Transformation and model choice for RNA-seq co-expression analysis**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 425–436, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw128>

Although a large number of clustering algorithms have been proposed to identify groups of co-expressed genes from microarray data, the question of if and how such methods may be applied to RNA sequencing (RNA-seq) data remains unaddressed. In this work, we investigate the use of data transformations in conjunction with Gaussian mixture models for RNA-seq co-expression analyses, as well as a penalized model selection criterion to select both an appropriate transformation and number of clusters present in the data. This approach has the advantage of accounting for per-cluster correlation structures among samples, which can be strong in RNA-seq data. In addition, it provides a rigorous statistical framework for parameter estimation, an objective assessment of data transformations and number of clusters and the possibility of performing diagnostic checks on the quality and homogeneity of the identified clusters. We analyze four varied RNA-seq data sets to illustrate the use of transformations and model selection in conjunction with Gaussian mixture models. Finally, we propose a Bioconductor package coseq (co-expression of RNA-seq data) to facilitate implementation and visualization of the recommended RNA-seq co-expression analyses.

Aliyu Musa, Laleh Soltan Ghoraie, Shu-Dong Zhang, Galina Glazko, Olli Yli-Harja, Matthias Dehmer, Benjamin Haibe-Kains, Frank Emmert-Streib; **A review of connectivity map and computational approaches in pharmacogenomics**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 506–523, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw112>

Large-scale perturbation databases, such as Connectivity Map (CMap) or Library of Integrated Network-based Cellular Signatures (LINCS), provide enormous opportunities for computational pharmacogenomics and drug design. A reason for this is that in contrast to classical pharmacology focusing at one target at a time, the transcriptomics profiles provided by CMap and LINCS open the door for systems biology approaches on the pathway and network level. In this article, we provide a review of recent developments in computational pharmacogenomics with respect to CMap and LINCS and related applications.

Pora Kim, Peilin Jia, Zhongming Zhao; **Kinase impact assessment in the landscape of fusion genes that retain kinase domains: a pan-cancer study**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 450–460, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw127>

Assessing the impact of kinase in gene fusion is essential for both identifying driver fusion genes (FGs) and developing molecular targeted therapies. Kinase domain retention is a crucial factor in kinase fusion genes (KFGs), but such a systematic investigation has not been done yet. To this end, we analyzed kinase domain retention (KDR) status in chimeric protein sequences of 914 KFGs covering 312 kinases across 13 major cancer types. Based on 171 kinase domain-retained KFGs including 101 kinases, we studied their recurrence, kinase groups, fusion partners, exon-based expression depth, short DNA motifs around the break points and networks. Our results, such as more KDR than 5′-kinase fusion genes, combinatorial effects between 3′-KDR kinases and their 5′-partners and a signal transduction-specific DNA sequence motif in the break point intronic sequences, supported positive selection on 3′-kinase fusion genes in cancer. We introduced a degree-of-frequency (DoF) score to measure the possible number of KFGs of a kinase. Interestingly, kinases with high DoF scores tended to undergo strong gene expression alteration at the break points. Furthermore, our KDR gene fusion network analysis revealed six of the seven kinases with the highest DoF scores (ALK, BRAF, MET, NTRK1, NTRK3 and RET) were all observed in thyroid carcinoma. Finally, we summarized common features of ‘effective’ (highly recurrent) kinases in gene fusions such as expression alteration at break point, redundant usage in multiple cancer types and 3′-location tendency. Collectively, our findings are useful for prioritizing driver kinases and FGs and provided insights into KFGs’ clinical implications.

Yun Zhang, Saurabh Baheti, Zhifu Sun; **Statistical method evaluation for differentially methylated CpGs in base resolution next-generation DNA sequencing data**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 374–386, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw133>

High-throughput bisulfite methylation sequencing such as reduced representation bisulfite sequencing (RRBS), Agilent SureSelect Human Methyl-Seq (Methyl-seq) or whole-genome bisulfite sequencing is commonly used for base resolution methylome research. These data are represented either by the ratio of methylated cytosine versus total coverage at a CpG site or numbers of methylated and unmethylated cytosines. Multiple statistical methods can be used to detect differentially methylated CpGs (DMCs) between conditions, and these methods are often the base for the next step of differentially methylated region identification. The ratio data have a flexibility of fitting to many linear models, but the raw count data take consideration of coverage information. There is an array of options in each datatype for DMC detection; however, it is not clear which is an optimal statistical method. In this study, we systematically evaluated four statistic methods on methylation ratio data and four methods on count-based data and compared their performances with regard to type I error control, sensitivity and specificity of DMC detection and computational resource demands using real RRBS data along with simulation. Our results show that the ratio-based tests are generally more conservative (less sensitive) than the count-based tests. However, some count-based methods have high false-positive rates and should be avoided. The beta-binomial model gives a good balance between sensitivity and specificity and is preferred method. Selection of methods in different settings, signal versus noise and sample size estimation are also discussed.

Anna Kuosmanen, Tuukka Norri, Veli Mäkinen; **Evaluating approaches to find exon chains based on long reads**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 404–414, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw137>

Transcript prediction can be modeled as a graph problem where exons are modeled as nodes and reads spanning two or more exons are modeled as exon chains. Pacific Biosciences third-generation sequencing technology produces significantly longer reads than earlier second-generation sequencing technologies, which gives valuable information about longer exon chains in a graph. However, with the high error rates of third-generation sequencing, aligning long reads correctly around the splice sites is a challenging task. Incorrect alignments lead to spurious nodes and arcs in the graph, which in turn lead to incorrect transcript predictions. We survey several approaches to find the exon chains corresponding to long reads in a splicing graph, and experimentally study the performance of these methods using simulated data to allow for sensitivity/precision analysis. Our experiments show that short reads from second-generation sequencing can be used to significantly improve exon chain correctness either by error-correcting the long reads before splicing graph creation, or by using them to create a splicing graph on which the long-read alignments are then projected. We also study the memory and time consumption of various modules, and show that accurate exon chains lead to significantly increased transcript prediction accuracy. Availability: The simulated data and in-house scripts used for this article are available at <http://www.cs.helsinki.fi/group/gsa/exon-chains/exon-chains-bib.tar.bz2>.

Hanyuan Zhang, Bruno Vieira Resende e Silva, Juan Cui; **miRDis: a Web tool for endogenous and exogenous microRNA discovery based on deep-sequencing data analysis**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 415–424, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw140>

Small RNA sequencing is the most widely used tool for microRNA (miRNA) discovery, and shows great potential for the efficient study of miRNA cross-species transport, i.e., by detecting the presence of exogenous miRNA sequences in the host species. Because of the increased appreciation of dietary miRNAs and their far-reaching implication in human health, research interests are currently growing with regard to exogenous miRNAs bioavailability, mechanisms of cross-species transport and miRNA function in cellular biological processes. In this article, we present microRNA Discovery (miRDis), a new small RNA sequencing data analysis pipeline for both endogenous and exogenous miRNA detection. Specifically, we developed and deployed a Web service that supports the annotation and expression profiling data of known host miRNAs and the detection of novel miRNAs, other noncoding RNAs, and the exogenous miRNAs from dietary species. As a proof-of-concept, we analyzed a set of human plasma sequencing data from a milk-feeding study where 225 human miRNAs were detected in the plasma samples and 44 show elevated expression after milk intake. By examining the bovine-specific sequences, data indicate that three bovine miRNAs (bta-miR-378, -181\* and -150) are present in human plasma possibly because of the dietary uptake. Further evaluation based on different sets of public data demonstrates that miRDis outperforms other state-of-the-art tools in both detection and quantification of miRNA from either animal or plant sources. The miRDis Web server is available at: http://sbbi.unl.edu.ezp3.lib.umn.edu/miRDis/index.php.

Qi Dai, Chaohui Bao, Yabing Hai, Sheng Ma, Tao Zhou, Cong Wang, Yunfei Wang, Wenwen Huo, Xiaoqing Liu, Yuhua Yao, Zhenyu Xuan, Min Chen, Michael Q Zhang; **MTGIpick allows robust identification of genomic islands from a single genome**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 361–373, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw118>

Genomic islands (GIs) that are associated with microbial adaptations and carry sequence patterns different from that of the host are sporadically distributed among closely related species. This bias can dominate the signal of interest in GI detection. However, variations still exist among the segments of the host, although no uniform standard exists regarding the best methods of discriminating GIs from the rest of the genome in terms of compositional bias. In the present work, we proposed a robust software, MTGIpick, which used regions with pattern bias showing multiscale difference levels to identify GIs from the host. MTGIpick can identify GIs from a single genome without annotated information of genomes or prior knowledge from other data sets. When real biological data were used, MTGIpick demonstrated better performance than existing methods, as well as revealed potential GIs with accurate sizes missed by existing methods because of a uniform standard. Software and supplementary are freely available at http://bioinfo.zstu.edu.cn/MTGI or https://github.com/bioinfo0706/MTGIpick.

Ehsan Motazedi, Richard Finkers, Chris Maliepaard, Dick de Ridder; **Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 387–403, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw126>

Haplotypes are the units of inheritance in an organism, and many genetic analyses depend on their precise determination. Methods for haplotyping single individuals use the phasing information available in next-generation sequencing reads, by matching overlapping single-nucleotide polymorphisms while penalizing post hoc nucleotide corrections made. Haplotyping diploids is relatively easy, but the complexity of the problem increases drastically for polyploid genomes, which are found in both model organisms and in economically relevant plant and animal species. Although a number of tools are available for haplotyping polyploids, the effects of the genomic makeup and the sequencing strategy followed on the accuracy of these methods have hitherto not been thoroughly evaluated.

We developed the simulation pipeline haplosim to evaluate the performance of three haplotype estimation algorithms for polyploids: HapCompass, HapTree and SDhaP, in settings varying in sequencing approach, ploidy levels and genomic diversity, using tetraploid potato as the model. Our results show that sequencing depth is the major determinant of haplotype estimation quality, that 1 kb PacBio circular consensus sequencing reads and Illumina reads with large insert-sizes are competitive and that all methods fail to produce good haplotypes when ploidy levels increase. Comparing the three methods, HapTree produces the most accurate estimates, but also consumes the most resources. There is clearly room for improvement in polyploid haplotyping algorithms.

Marie E Bolger, Borjana Arsova, Björn Usadel; **Plant genome and transcriptome annotations: from misconceptions to simple solutions**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 437–449, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw135>

Next-generation sequencing has triggered an explosion of available genomic and transcriptomic resources in the plant sciences. Although genome and transcriptome sequencing has become orders of magnitudes cheaper and more efficient, often the functional annotation process is lagging behind. This might be hampered by the lack of a comprehensive enumeration of simple-to-use tools available to the plant researcher. In this comprehensive review, we present (i) typical ontologies to be used in the plant sciences, (ii) useful databases and resources used for functional annotation, (iii) what to expect from an annotated plant genome, (iv) an automated annotation pipeline and (v) a recipe and reference chart outlining typical steps used to annotate plant genomes/transcriptomes using publicly available resources.

Guifang Fu, Mian Huang, Wenhao Bo, Han Hao, Rongling Wu; **Mapping morphological shape as a high-dimensional functional curve**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 461–471, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw111>

Detecting how genes regulate biological shape has become a multidisciplinary research interest because of its wide application in many disciplines. Despite its fundamental importance, the challenges of accurately extracting information from an image, statistically modeling the high-dimensional shape and meticulously locating shape quantitative trait loci (QTL) affect the progress of this research. In this article, we propose a novel integrated framework that incorporates shape analysis, statistical curve modeling and genetic mapping to detect significant QTLs regulating variation of biological shape traits. After quantifying morphological shape via a radius centroid contour approach, each shape, as a phenotype, was characterized as a high-dimensional curve, varying as angle θ runs clockwise with the first point starting from angle zero. We then modeled the dynamic trajectories of three mean curves and variation patterns as functions of θ. Our framework led to the detection of a few significant QTLs regulating the variation of leaf shape collected from a natural population of poplar, Populus szechuanica var tibetica. This population, distributed at altitudes 2000–4500 m above sea level, is an evolutionarily important plant species. This is the first work in the quantitative genetic shape mapping area that emphasizes a sense of ‘function’ instead of decomposing the shape into a few discrete principal components, as the majority of shape studies do.

Pietro Hiram Guzzi, Tijana Milenković; **Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 472–481, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw132>

Analogous to genomic sequence alignment that allows for across-species transfer of biological knowledge between conserved sequence regions, biological network alignment can be used to guide the knowledge transfer between conserved regions of molecular networks of different species. Hence, biological network alignment can be used to redefine the traditional notion of a sequence-based homology to a new notion of network-based homology. Analogous to genomic sequence alignment, there exist local and global biological network alignments. Here, we survey prominent and recent computational approaches of each network alignment type and discuss their (dis)advantages. Then, as it was recently shown that the two approach types are complementary, in the sense that they capture different slices of cellular functioning, we discuss the need to reconcile the two network alignment types and present a recent first step in this direction. We conclude with some open research problems on this topic and comment on the usefulness of network alignment in other domains besides computational biology

Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, Yaoqi Zhou; **Sixty-five years of the long march in protein secondary structure prediction: the final stretch?**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 482–494, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw129>

Protein secondary structure prediction began in 1951 when Pauling and Corey predicted helical and sheet conformations for protein polypeptide backbone even before the first protein structure was determined. Sixty-five years later, powerful new methods breathe new life into this field. The highest three-state accuracy without relying on structure templates is now at 82–84%, a number unthinkable just a few years ago. These improvements came from increasingly larger databases of protein sequences and structures for training, the use of template secondary structure information and more powerful deep learning techniques. As we are approaching to the theoretical limit of three-state prediction (88–90%), alternative to secondary structure prediction (prediction of backbone torsion angles and Cα-atom-based angles and torsion angles) not only has more room for further improvement but also allows direct prediction of three-dimensional fragment structures with constantly improved accuracy. About 20% of all 40-residue fragments in a database of 1199 non-redundant proteins have <6 Å root-mean-squared distance from the native conformations by SPIDER2. More powerful deep learning methods with improved capability of capturing long-range interactions begin to emerge as the next generation of techniques for secondary structure prediction. The time has come to finish off the final stretch of the long march towards protein secondary structure prediction.

Valentina Galata, Christina Backes, Cédric Christian Laczny, Georg Hemmrich-Stanisak, Howard Li, Laura Smoot, Andreas Emanuel Posch, Susanne Schmolke, Markus Bischoff, Lutz von Müller, Achim Plum, Andre Franke, Andreas Keller; **Comparing genome versus proteome-based identification of clinical bacterial isolates**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 495–505, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw122>

Whole-genome sequencing (WGS) is gaining importance in the analysis of bacterial cultures derived from patients with infectious diseases. Existing computational tools for WGS-based identification have, however, been evaluated on previously defined data relying thereby unwarily on the available taxonomic information.

Here, we newly sequenced 846 clinical gram-negative bacterial isolates representing multiple distinct genera and compared the performance of five tools (CLARK, Kaiju, Kraken, DIAMOND/MEGAN and TUIT). To establish a faithful ‘gold standard’, the expert-driven taxonomy was compared with identifications based on matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) analysis. Additionally, the tools were also evaluated using a data set of 200 Staphylococcus aureus isolates.

CLARK and Kraken (with k =31) performed best with 626 (100%) and 193 (99.5%) correct species classifications for the gram-negative and S. aureusisolates, respectively. Moreover, CLARK and Kraken demonstrated highest mean F-measure values (85.5/87.9% and 94.4/94.7% for the two data sets, respectively) in comparison with DIAMOND/MEGAN (71 and 85.3%), Kaiju (41.8 and 18.9%) and TUIT (34.5 and 86.5%). Finally, CLARK, Kaiju and Kraken outperformed the other tools by a factor of 30 to 170 fold in terms of runtime.

We conclude that the application of nucleotide-based tools using k-mers—e.g. CLARK or Kraken—allows for accurate and fast taxonomic characterization of bacterial isolates from WGS data. Hence, our results suggest WGS-based genotyping to be a promising alternative to the MS-based biotyping in clinical settings. Moreover, we suggest that complementary information should be used for the evaluation of taxonomic classification tools, as public databases may suffer from suboptimal annotations.

Ping Zhang, Constantinos A Georgiou, Vladimir Brusic; **Elemental metabolomics**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 524–536, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw131>

Elemental metabolomics is quantification and characterization of total concentration of chemical elements in biological samples and monitoring of their changes. Recent advances in inductively coupled plasma mass spectrometry have enabled simultaneous measurement of concentrations of > 70 elements in biological samples. In living organisms, elements interact and compete with each other for absorption and molecular interactions. They also interact with proteins and nucleotide sequences. These interactions modulate enzymatic activities and are critical for many molecular and cellular functions. Testing for concentration of > 40 elements in blood, other bodily fluids and tissues is now in routine use in advanced medical laboratories. In this article, we define the basic concepts of elemental metabolomics, summarize standards and workflows, and propose minimum information for reporting the results of an elemental metabolomics experiment. Major statistical and informatics tools for elemental metabolomics are reviewed, and examples of applications are discussed. Elemental metabolomics is emerging as an important new technology with applications in medical diagnostics, nutrition, agriculture, food science, environmental science and multiplicity of other areas.

Bohdan B Khomtchouk, Edmund Weitz, Peter D Karp, Claes Wahlestedt; **How the strengths of Lisp-family languages facilitate building complex and flexible bioinformatics applications**, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 537–543, <https://doi-org.ezp3.lib.umn.edu/10.1093/bib/bbw130>

We present a rationale for expanding the presence of the Lisp family of programming languages in bioinformatics and computational biology research. Put simply, Lisp-family languages enable programmers to more quickly write programs that run faster than in other languages. Languages such as Common Lisp, Scheme and Clojure facilitate the creation of powerful and flexible software that is required for complex and rapidly evolving domains like biology. We will point out several important key features that distinguish languages of the Lisp family from other programming languages, and we will explain how these features can aid researchers in becoming more productive and creating better code. We will also show how these features make these languages ideal tools for artificial intelligence and machine learning applications. We will specifically stress the advantages of domain-specific languages (DSLs): languages that are specialized to a particular area, and thus not only facilitate easier research problem formulation, but also aid in the establishment of standards and best programming practices as applied to the specific research field at hand. DSLs are particularly easy to build in Common Lisp, the most comprehensive Lisp dialect, which is commonly referred to as the ‘programmable programming language’. We are convinced that Lisp grants programmers unprecedented power to build increasingly sophisticated artificial intelligence systems that may ultimately transform machine learning and artificial intelligence research in bioinformatics and computational biology.