* [GOfuncR: Gene ontology enrichment using FUNC](https://bioconductor.org/packages/release/bioc/html/GOfuncR.html)

***Steffi Grote***

Max-Planck-Institute for Evolutionary Anthropology • Leipzig • Germany

### Abstract:

GOfuncR performs a gene ontology enrichment analysis based on the ontology enrichment software FUNC. GO-annotations are obtained from OrganismDb or OrgDb packages ('Homo.sapiens' by default); the GO-graph is included in the package and updated regularly (10-Apr-2018). GOfuncR provides the standard candidate vs. background enrichment analysis using the hypergeometric test, as well as three additional tests: (i) the Wilcoxon rank-sum test that is used when genes are ranked, (ii) a binomial test that is used when genes are associated with two counts and (iii) a Chi-square or Fisher's exact test that is used in cases when genes are associated with four counts. To correct for multiple testing and interdependency of the tests, family-wise error rates are computed based on random permutations of the gene-associated variables. GOfuncR also provides tools for exploring the ontology graph and the annotations, and options to take gene-length or spatial clustering of genes into account.

* rgsepd: Gene Set Enrichment / Projection Displays

Karl Stamm

Marquette University Milwaukee, WI

### Abstract

GSEPD is a package for streamlining RNA-Seq data analysis, targeting complex samples with low replicate count such as human tissues, where all factors (metabolic, genetic, etc) cannot be controlled statistically. As a prerequisite, you need only your multiple samples’ count data as a matrix whose columns are samples and rows are RefSeq NM and NR transcript identifiers. A second matrix associates sample identifiers with treatment/condition. Given both datasets, GSEPD will automate differential expression via DESeq2 [DESeq], functional analysis via GOSeq [GOSeq], generate heatmaps of gene expression for significantly differentially expressed genes, and subsets of genes defined by the significantly enriched GO Terms. After gene sets are detected from a differential expression analysis, the results are merged into a novel ‘projection display’ wherein each sample is scored according to each condition’s multidimensional average expression. When the treatment samples are found to have a perturbed expression profile for a particular GO Term (geneset), all samples are scored on an axis ranging from control to treatment condition, and outliers or anomalous samples are readily apparent. Clustering quality of samples in a given geneset-space is quantified by the cluster’s “Validity score” [Validity] and an empirical permutation derived p-value. GO Terms with more genes than samples in your comparison will randomly appear enriched, so the Segregation P-value is used to determine if a GO Term is significantly segregating your samples.

* [CTDquerier: a bioconductor R package for Comparative Toxicogenomics DatabaseTM data extraction, visualization and enrichment of environmental and toxicological studies](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty326/4983065)

[Hernandez-Ferrer C](https://www.ncbi.nlm.nih.gov/pubmed/?term=Hernandez-Ferrer%20C%5BAuthor%5D&cauthor=true&cauthor_uid=29688259)1,2,3,4, [Gonzalez JR](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gonzalez%20JR%5BAuthor%5D&cauthor=true&cauthor_uid=29688259)1,2,3.

1 Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain.

2 Universitat Pompeu Fabra (UPF), Barcelona, Spain.

3 CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.

4 Computational Health Informatics Program, Boston Children's Hospital, Boston, USA.

### Abstract

#### SUMMARY:

Biomedical studies currently include a large volume of genomic and environmental factors for studying the etiology of human diseases. R/Bioconductor projects provide several tools for performing enrichment analysis at gene-pathway level, allowing researchers to develop novel hypotheses. However, there is a need to perform similar analyses at the chemicals-genes or chemicals-diseases levels, to provide complementary knowledge of the causal path between chemicals and diseases. While the Comparative Toxicogenomics DatabaseTM (CTD) provides information about these relationships, there is no software for integrating it into R/Bioconductor analysis pipelines. CTDquerier helps users to easily download CTD data and integrate it in the R/Bioconductor framework. The package also contains functions for visualizing CTD data and performing enrichment analyses. We illustrate how to use the package with a real data analysis of asthma-related genes. CTDquerier is a flexible and easy-to-use Bioconductor package that provides novel hypothesis about the relationships between chemicals and diseases.

* [ELMER v.2: An R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles](https://www.biorxiv.org/content/early/2018/04/02/148726)

Tiago C Silva, Simon G Coetzee, Lijing Yao, Nicole Gull, Dennis J Hazelett, Houtan Noushmehr, De-Chen Lin,\* Benjamin P Berman\*

1 Department of Genetics, Ribeir˜ao Preto Medical School, University of S˜ao Paulo, Ribeir˜ao Preto, Brazil,

2 Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA, USA

3 Roche Sequencing Solutions, Belmont, CA, USA

4 Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California, USA and

5 Department of Neurosurgery, Henry Ford Hospital, Detroit, MI, USA

## Abstract

Motivation: DNA methylation can be used to identify functional changes at transcriptional enhancers and other cis-regulatory modules (CRMs) in tumors and other primary disease tissues. Our R/Bioconductor package ELMER (Enhancer Linking by Methylation/Expression Relationships) provides a systematic approach that reconstructs gene regulatory networks (GRNs) by combining methylation and gene expression data derived from the same set of samples. Results: We present new ELMER version that provides a new Supervised analysis mode, which uses pre-defined sample groupings and can identify additional Master Regulators, such as KLF5 in basal-like breast cancer. Availability: ELMER (v2.0) is available as an R/Bioconductor package at http://bioconductor.org/packages/ELMER/ with auxiliary data at http://bioconductor.org/packages/ELMER.data/

* [DaMiRseq-an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification.](https://academic.oup.com/bioinformatics/article/34/8/1416/4721783)

Chiesa M1, Colombo GI1, Piacentini L1.

Immunology and Functional Genomics Unit, Centro Cardiologico Monzino, IRCCS, 20138 Milan, Italy.

RNA-Seq is becoming the technique of choice for high-throughput transcriptome profiling, which, besides class comparison for differential expression, promises to be an effective and powerful tool for biomarker discovery. However, a systematic analysis of high-dimensional genomic data is a demanding task for such a purpose. DaMiRseq offers an organized, flexible and convenient framework to remove noise and bias, select the most informative features and perform accurate classification.

Genomewide position-specific scores, such as those estimating conservation, constraint, fitness or mutation tolerance, are ubiquitous in current genome analyses. The diversity of sources and formats of these scores, as well as their size, increase the burden to use them. We present GenomicScores, a Bioconductor package that provides efficient storage and seamless access of genomewide position-specific scores from R, facilitating their use in genome analysis workflows.

* [GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor.](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty311/4987140)

Pau Puigdevall Robert Castelo

Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, 08003, Spain.

## Abstract

**Summary**

Genomewide position-specific scores, such as those estimating conservation, constraint, fitness or mutation tolerance, are ubiquitous in current genome analyses. The diversity of sources and formats of these scores, as well as their size, increase the burden to use them. We present GenomicScores, a Bioconductor package that provides efficient storage and seamless access of genomewide position-specific scores from R, facilitating their use in genome analysis workflows.

Availability and implementation:

GenomicScores is implemented in R and available at <https://bioconductor.org/packages/GenomicScores> under the open source ‘Artistic-2.0’ license.

* [MutationalPatterns: comprehensive genome-wide analysis of mutational processes](https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-018-0539-0).

[Blokzijl F](https://www.ncbi.nlm.nih.gov/pubmed/?term=Blokzijl%20F%5BAuthor%5D&cauthor=true&cauthor_uid=29695279)1, [Janssen R](https://www.ncbi.nlm.nih.gov/pubmed/?term=Janssen%20R%5BAuthor%5D&cauthor=true&cauthor_uid=29695279)1, [van Boxtel R](https://www.ncbi.nlm.nih.gov/pubmed/?term=van%20Boxtel%20R%5BAuthor%5D&cauthor=true&cauthor_uid=29695279)1,2, [Cuppen E](https://www.ncbi.nlm.nih.gov/pubmed/?term=Cuppen%20E%5BAuthor%5D&cauthor=true&cauthor_uid=29695279)3.

Center for Molecular Medicine and Oncode Institute, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

## Abstract

### Background

Base substitution catalogues represent historical records of mutational processes that have been active in a cell. Such processes can be distinguished by various characteristics, like mutation type, sequence context, transcriptional and replicative strand bias, genomic distribution and association with (epi)-genomic features.

### Results

We have created MutationalPatterns, an R/Bioconductor package that allows researchers to characterize a broad range of patterns in base substitution catalogues to dissect the underlying molecular mechanisms. Furthermore, it offers an efficient method to quantify the contribution of known mutational signatures within single samples. This analysis can be used to determine whether certain DNA repair mechanisms are perturbed and to further characterize the processes underlying known mutational signatures.

### Conclusions

MutationalPatterns allows for easy characterization and visualization of mutational patterns. These analyses willsupport fundamental research into mutational mechanisms and may ultimately improve cancer diagnosis and treatment strategies. MutationalPatterns is freely available at <http://bioconductor.org/packages/MutationalPatterns>.

# [PathwaySplice: An R package for unbiased pathway analysis of alternative splicing in RNA-Seq data.](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty317/4983063)

[Yan A](https://www.ncbi.nlm.nih.gov/pubmed/?term=Yan%20A%5BAuthor%5D&cauthor=true&cauthor_uid=29688305)1, [Ban Y](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ban%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=29688305)1, [Gao Z](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gao%20Z%5BAuthor%5D&cauthor=true&cauthor_uid=29688305)1, [Chen X](https://www.ncbi.nlm.nih.gov/pubmed/?term=Chen%20X%5BAuthor%5D&cauthor=true&cauthor_uid=29688305)1,2, [Wang L](https://www.ncbi.nlm.nih.gov/pubmed/?term=Wang%20L%5BAuthor%5D&cauthor=true&cauthor_uid=29688305)1,2,3.

1 Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL 33136, USA.

2 Department of Public Health Science, Division of Biostatistics, University of Miami Miller School of Medicine, Miami, FL 33136, USA.

3 Dr. John T. Macdonald Foundation, Department of Human Genetics, and John P. Hussman Institute for Human Genomics, University of Miami, Miami, FL 33136, USA.

### Abstract

#### SUMMARY:

Pathway analysis of alternative splicing would be biased without accounting for the different number of exons or junctions associated with each gene, because genes with higher number of exons or junctions are more likely to be included in the "significant" gene list in alternative splicing. We present PathwaySplice, an R package that (1) Performs pathway analysis that explicitly adjusts for the number of exons or junctions associated with each gene; (2) Visualizes selection bias due to different number of exons or junctions for each gene and formally tests for presence of bias using logistic regression; (3) Supports gene sets based on the Gene Ontology terms, as well as more broadly defined gene sets (e.g. MSigDB) or user defined gene sets; (4) Identifies the significant genes driving pathway significance and (5) Organizes significant pathways with an enrichment map, where pathways with large number of overlapping genes are grouped together in a network graph.

# [MethylMix 2.0: an R package for identifying DNA methylation genes.](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty156/4970512)

[Cedoz PL](https://www.ncbi.nlm.nih.gov/pubmed/?term=Cedoz%20PL%5BAuthor%5D&cauthor=true&cauthor_uid=29668835)1, [Prunello M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Prunello%20M%5BAuthor%5D&cauthor=true&cauthor_uid=29668835)2, [Brennan K](https://www.ncbi.nlm.nih.gov/pubmed/?term=Brennan%20K%5BAuthor%5D&cauthor=true&cauthor_uid=29668835)1, [Gevaert O](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gevaert%20O%5BAuthor%5D&cauthor=true&cauthor_uid=29668835)1.

Stanford Center for Biomedical Informatics, Department of Medicine and Biomedical Data Science, Stanford, CA 94305-5479, USA.

Department of Statistics, College of Pharmaceutical and Biochemical Sciences, National University of Rosario, Rosario, Argentina.

### Abstract

#### SUMMARY:

DNA methylation is an important mechanism regulating gene transcription, and its role in carcinogenesis has been extensively studied. Hyper and hypomethylation of genes is a major mechanism of gene expression deregulation in a wide range of diseases. At the same time, high-throughput DNA methylation assays have been developed generating vast amounts of genome wide DNA methylation measurements. We developed MethylMix, an algorithm implemented in R to identify disease specific hyper and hypomethylated genes. Here we present a new version of MethylMix that automates the construction of DNA-methylation and gene expression datasets from The Cancer Genome Atlas (TCGA). More precisely, MethylMix 2.0 incorporates two major updates: the automated downloading of DNA methylation and gene expression datasets from TCGA and the automated preprocessing of such datasets: value imputation, batch correction and CpG sites clustering within each gene. The resulting datasets can subsequently be analyzed with MethylMix to identify transcriptionally predictive methylation states. We show that the Differential Methylation Values created by MethylMix can be used for cancer subtyping.

* [IntEREst: intron-exon retention estimator.](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2122-5)

Institute of Biotechnology, University of Helsinki, P.O. Box 56 (Viikinkaari 5), FI-00014, Helsinki, Finland.

### Abstract

#### BACKGROUND:

In-depth study of the intron retention levels of transcripts provide insights on the mechanisms regulating pre-mRNA splicing efficiency. Additionally, detailed analysis of retained introns can link these introns to post-transcriptional regulation or identify aberrant splicing events in human diseases.

#### RESULTS:

We present IntEREst, Intron-Exon Retention Estimator, an R package that supports rigorous analysis of non-annotated intron retention events (in addition to the ones annotated by RefSeq or similar databases), and support intra-sample in addition to inter-sample comparisons. It accepts binary sequence alignment/map (.bam) files as input and determines genome-wide estimates of intron retention or exon-exon junction levels. Moreover, it includes functions for comparing subsets of user-defined introns (e.g. U12-type vs U2-type) and its plotting functions allow visualization of the distribution of the retention levels of the introns. Statistical methods are adapted from the DESeq2, edgeR and DEXSeq R packages to extract the significantly more or less retained introns. Analyses can be performed either sequentially (on single core) or in parallel (on multiple cores). We used IntEREst to investigate the U12- and U2-type intron retention in human and plant RNAseq dataset with defects in the U12-dependent spliceosome due to mutations in the ZRSR2 component of this spliceosome. Additionally, we compared the retained introns discovered by IntEREst with that of other methods and studies.

#### CONCLUSION:

IntEREst is an R package for Intron retention and exon-exon junction levels analysis of RNA-seq data. Both the human and plant analyses show that the U12-type introns are retained at higher level compared to the U2-type introns already in the control samples, but the retention is exacerbated in patient or plant samples carrying a mutated ZRSR2 gene. Intron retention events caused by ZRSR2 mutation that we discovered using IntEREst (DESeq2 based function) show considerable overlap with the retained introns discovered by other methods (e.g. IRFinder and edgeR based function of IntEREst). Our results indicate that increase in both the number of biological replicates and the depth of sequencing library promote the discovery of retained introns, but the effect of library size gradually decreases with more than 35 million reads mapped to the introns.

<https://bioconductor.org/packages/release/bioc/vignettes/IntEREst/inst/doc/IntEREst.html>

* [EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations.](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4625-x)

Division of Theoretical Bioinformatics (B080), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120, Heidelberg, Germany.

Abstract

Background

High-throughput sequencing data are dramatically increasing in volume. Thus, there is urgent need for efficient tools to perform fast and integrative analysis of multiple data types. Enriched heatmap is a specific form of heatmap that visualizes how genomic signals are enriched over specific target regions. It is commonly used and efficient at revealing enrichment patterns especially for high dimensional genomic and epigenomic datasets.

Results

We present a new R package named EnrichedHeatmap that efficiently visualizes genomic signal enrichment. It provides advanced solutions for normalizing genomic signals within target regions as well as offering highly customizable visualizations. The major advantage of EnrichedHeatmap is the ability to conveniently generate parallel heatmaps as well as complex annotations, which makes it easy to integrate and visualize comprehensive overviews of the patterns and associations within and between complex datasets.

Conclusions

EnrichedHeatmap facilitates comprehensive understanding of high dimensional genomic and epigenomic data. The power of EnrichedHeatmap is demonstrated by visualization of the complex associations between DNA methylation, gene expression and various histone modifications.

* [DASC: Detecting hidden batch factors through data-adaptive adjustment for biological effects](https://watermark.silverchair.com/btx635.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAdQwggHQBgkqhkiG9w0BBwagggHBMIIBvQIBADCCAbYGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMqc4qcPXedLGyBbk0AgEQgIIBh5WTnBWCiG7Nkeq2wWea3ZRUIUBJ7BlcNrfA676gkbiBIqgDvZNX0232YKM8iHZ9lFM-7vEOezS_rRT0UKXoliroUUdflayQ6tyqGINEMkVDknOMwIl0YpmKFI9x6NgLNxQtE2f4wKc7GJGEBr_rEQgVwISpWclsdhxxfYnJsIAA8JK9ATahPOonE0kPu__5AZ6Ck7xKny7zMUxK1xBGFWvzwwGUPv8oRhNcDjBl_xB3L3iI94q3scPGW-c-UdpLEG9ClX0Tx_77Box7YsE22o01kYOAmTz3AfqZojTtpRF-lly-85KN81Dj3UG-KCCiKm1Xce1cldUWzIhdCobp1FFMnL8I-52Bg5ko3ExNi5AkUEU8gJb4yG9WPOagAFtV0_8TJnqo1LxwYxIM3WqzOa5Z8hkj8aA8W1EM9_ql-j3pMg0qgU0fIIVlvqTTpP7_nl9gyQcB2ds51sCrJpTyTBe75nM_zwoPsrDkyUtSYvsKCxm2V11Nnb_PgzijVZPEcV6jNFPFMMk).

Haidong Yi1,†, Ayush T. Raman2,3,†, Han Zhang1,4,\*, Genevera I. Allen5 and Zhandong Liu2,3,\*

1 College of Computer and Control Engineering, Nankai University, Tianjin 300350, China,

2 Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030, USA,

3 Department of Pediatrics, Neurological Research Institute, Baylor College of Medicine, Houston, TX 77030, USA,

4 Tianjin Key Laboratory of Intelligent Robotics, Nankai University, Tianjin 300350, China and

5 Department of Statistics, Rice University, Houston, TX, 77030 USA

## Abstract

Motivation

Batch effects are one of the major source of technical variations that affect the measurements in high-throughput studies such as RNA sequencing. It has been well established that batch effects can be caused by different experimental platforms, laboratory conditions, different sources of samples and personnel differences. These differences can confound the outcomes of interest and lead to spurious results. A critical input for batch correction algorithms is the knowledge of batch factors, which in many cases are unknown or inaccurate. Hence, the primary motivation of our paper is to detect hidden batch factors that can be used in standard techniques to accurately capture the relationship between gene expression and other modeled variables of interest.

Results

We introduce a new algorithm based on data-adaptive shrinkage and semi-Non-negative Matrix Factorization for the detection of unknown batch effects. We test our algorithm on three different datasets: (i) Sequencing Quality Control, (ii) Topotecan RNA-Seq and (iii) Single-cell RNA sequencing (scRNA-Seq) on Glioblastoma Multiforme. We have demonstrated a superior performance in identifying hidden batch effects as compared to existing algorithms for batch detection in all three datasets. In the Topotecan study, we were able to identify a new batch factor that has been missed by the original study, leading to under-representation of differentially expressed genes. For scRNA-Seq, we demonstrated the power of our method in detecting subtle batch effects.

* [scmap: projection of single-cell RNA-seq data across data sets.](https://www.biorxiv.org/content/biorxiv/early/2017/11/29/150292.full.pdf)

Vladimir Yu Kiselev1 , Andrew Yiu1 and Martin Hemberg1 1 Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Single-cell RNA-seq (scRNA-seq) is widely used to investigate the composition of complex tissues since the technology allows researchers to define cell-types using unsupervised clustering of the transcriptome. However, due to differences in experimental methods and computational analyses, it is often challenging to directly compare the cells identified in two different experiments. Here, we present scmap (http://bioconductor.org/packages/scmap), a method for projecting cells from a scRNA-seq experiment onto the cell-types or individual cells identified in other experiments (the application can be run for free, without restrictions, from http://www.hemberg-lab.cloud/scmap).

* [Metaviz: interactive statistical and visual analysis of metagenomic data.](https://www.ncbi.nlm.nih.gov/pubmed/29529268)

Wagner J1,2,3, Chelaru F1,2,3, Kancherla J2,3, Paulson JN4,5, Zhang A1, Felix V6, Mahurkar A6, Elmqvist N3,7,8, Corrada Bravo H1,2,3.

Department of Computer Science, University of Maryland, College Park, MD

Large studies profiling microbial communities and their association with healthy or disease phenotypes are now commonplace. Processed data from many of these studies are publicly available but significant effort is required for users to effectively organize, explore and integrate it, limiting the utility of these rich data resources. Effective integrative and interactive visual and statistical tools to analyze many metagenomic samples can greatly increase the value of these data for researchers. We present Metaviz, a tool for interactive exploratory data analysis of annotated microbiome taxonomic community profiles derived from marker gene or whole metagenome shotgun sequencing. Metaviz is uniquely designed to address the challenge of browsing the hierarchical structure of metagenomic data features while rendering visualizations of data values that are dynamically updated in response to user navigation. We use Metaviz to provide the UMD Metagenome Browser web service, allowing users to browse and explore data for more than 7000 microbiomes from published studies. Users can also deploy Metaviz as a web service, or use it to analyze data through the metavizr package to interoperate with state-of-the-art analysis tools available through Bioconductor. Metaviz is free and open source with the code, documentation and tutorials publicly accessible.

# [HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor](https://www.biorxiv.org/content/biorxiv/early/2018/04/17/299115.full.pdf)

Lucas Schiffer1,2, Rimsha Azhar1,2, Lori Shepherd3 , Marcel Ramos1,2,3 , Ludwig Geistlinger1,2, Curtis Huttenhower4,5, Jennifer B Dowd1,6 , Nicola Segata7 , Levi Waldron1,2

1. Graduate School of Public Health and Health Policy, City University of New York, New York, NY

2. Institute for Implementation Science in Population Health, City University of New York, New York, NY

3. Roswell Park Cancer Institute, University of Buffalo, Buffalo, NY

4. Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA

 5. The Broad Institute of MIT and Harvard, Cambridge, MA

6. Department of Global Health and Social Medicine, King's College London, London, UK

7. Centre for Integrative Biology, University of Trento, Trento, Italy

## Abstract

Phase 1 of the NIH Human Microbiome Project (HMP) investigated 18 body subsites of 235 healthy American adults, to produce the first comprehensive reference for the composition and variation of the "healthy" human microbiome. Publicly available data sets from amplicon sequencing of two 16S rRNA variable regions, with extensive controlled-access participant data, provide a reference for ongoing microbiome studies. However, utilization of these data sets can be hindered by the complex bioinformatic steps required to access, import, decrypt, and merge the various components in formats suitable for ecological and statistical analysis. The HMP16SData package provides count data for both 16S variable regions, integrated with phylogeny, taxonomy, public participant data, and controlled participant data for authorized researchers, using standard integrative Bioconductor data objects. By removing bioinformatic hurdles of data access and management, HMP16SData enables epidemiologists with only basic R skills to quickly analyze HMP data.

# [CAGEfightR: Cap Analysis of Gene Expression (CAGE) in R/Bioconductor](https://www.biorxiv.org/content/biorxiv/early/2018/04/28/310623.full.pdf)

Malte Thodberg, Axel Thieffry, Kristoffer Vitting-Seerup, Robin Andersson, Albin Sandelin

Department of Biology, University of Copenhagen, DK 2 Biotech Research and Innovation Centre, University of Copenhagen, DK

## Abstract

We developed the CAGEfightR R/Biconductor-package for analyzing CAGE data. CAGEfightR allows for fast and memory efficient identification of transcription start sites (TSSs) and predicted enhancers. Downstream analysis, including annotation, quantification, visualization and TSS shape statistics are implemented in easy-to-use functions. The package is freely available at https://bioconductor.org/packages/CAGEfightR

* [IMMAN: an R/Bioconductor package for Interolog protein network reconstruction, Mapping and Mining ANalysis](https://www.biorxiv.org/content/biorxiv/early/2018/04/17/069104.full.pdf)

Minoo Ashtinai1,2,7, Payman Nickchi1,3,7, Soheil Jahangiri-Tazehkand4,5,7, Abdollah Safari3\*, Mehdi Mirzaie6\* & Mohieddin Jafari1,4\*

1 Drug Design and Bioinformatics Unit, Medical Biotechnology Department, Biotechnology Research Center, Pasteur Institute of Iran, 69, Pasteur St, 13164, Tehran, Iran,

2Department of Computer Science and Statistics, Faculty of Mathematics, K. N. Toosi University of Technology, P.O. Box 16315-1618, Tehran, Iran,

3 Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, V5A 1S6, Burnaby, BC, Canada,

4 Department of Computer Science, Shahid Beheshti University, Evin, Tehran, Iran.

5 School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Shahid Lavasani St, PO Box 19395-5746, Tehran, Iran,

6 Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Jalal Ale Ahmad Highway, PO Box 14115-134, Tehran, Iran,

Summary:

IMMAN is a software for reconstructing Interolog Protein Network (IPN) by integrating several Protein-protein Interaction Networks (PPINs). Users can unify different PPINs to mine conserved common networks among species. IMMAN is designed to retrieve IPNs with different degrees of conservation to engage prediction analysis of protein functions according to their networks.