1. **beachmat: A Bioconductor C++ API for accessing high-throughput biological data from a variety of R matrix types.**

Biological experiments involving genomics or other high-throughput assays typically yield a data matrix that can be explored and analyzed using the R programming language with packages from the Bioconductor project. Improvements in the throughput of these assays have resulted in an explosion of data even from routine experiments, which pose a challenge to the existing computational infrastructure for statistical data analysis.

The beachmat API uses C++ classes to provide a common interface for data access from R matrix representations. For all representations of a given data type (e.g., integer, double-precision, character strings), we define a base class with common methods for data access. Each specific representation is associated with a derived class that provides customized implementations of the access methods. The intention is for a user to pass in an R matrix of any type, in the form of an RObject instance from the Rcpp API ([Fig 1](http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006135#pcbi-1006135-g001)). beachmat then constructs an instance of the appropriate derived class, returning a pointer to the base class. This pointer is the same regardless of the representation and can be used in downstream C++ code to achieve run-time polymorphism.

 We also demonstrate how beachmat can be incorporated into the code of other packages to drive analyses of a very large scRNA-seq data set.

1. **Applications of Bayesian network models in predicting types of hematological malignancies.**

Network analysis is the preferred approach for the detection of subtle but coordinated changes in expression of an interacting and related set of genes. We introduce a novel method based on the analyses of coexpression networks and Bayesian networks, and we use this new method to classify two types of hematological malignancies; namely, acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS). Our classifier has an accuracy of 93%, a precision of 98%, and a recall of 90% on the training dataset (*n* = 366); which outperforms the results reported by other scholars on the same dataset. Although our training dataset consists of microarray data, our model has a remarkable performance on the RNA-Seq test dataset (*n* = 74, accuracy = 89%, precision = 88%, recall = 98%), which confirms that eigengenes are robust with respect to expression profiling technology. These signatures are useful in classification and correctly predicting the diagnosis. They might also provide valuable information about the underlying biology of diseases. Our network analysis approach is generalizable and can be useful for classifying other diseases based on gene expression profiles.

Our previously published ***Pigengene*** package is publicly available through Bioconductor, which can be used to conveniently fit a Bayesian network to gene expression data.



1. **Gene expression profiles and pathway enrichment analysis of human osteosarcoma cells exposed to sorafenib.**

Sorafenib is an inhibitor of a variety of tyrosine kinase receptors used to treat various cancers including hepatocellular, renal cell and thyroid carcinoma. It has been shown to change various targets associated with osteosarcoma, but the detailed mechanism remains unclear. In order to identify key genes, enriched pathways and important modules during the exposure of human osteosarcoma cells to sorafenib, data for gene expression profiles ([GSE53155](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53155)) were downloaded from the GEO database. In total, 61 differentially expressed genes (DEGs) were identified by the R bioconductor packages. Functional and enrichment analyses of DEGs were performed using the DAVID database. These revealed that DEGs were enriched in biological processes, molecular function and KEGG pathway of inflammatory immune response and angiogenesis. A protein–protein interaction network was constructed by string and visualized in cytoscape, and eight genes were selected as hubs: *IL8*,*CXCL2*,*PTGS2*,*FOS*,*CXCL1*,*C3*,*EHMT2* and *PGF*. Subsequently, only one cluster was identified by mcode, which consisted of six nodes (*CXCL1*,*CXCL2*,*PTGS2*,*FOS*,*C3* and *PGF*) and nine edges. *PGF* was the seed gene in this cluster. In conclusion, the results of this data mining and integration should help in revealing new mechanisms and targets of sorafenib in inhibiting osteosarcoma.

1. **POST: a framework for set-based association analysis in high-dimensional data. Projection onto the Orthogonal Space Testing (POST)**

POST is a suitable for testing the association of gene-sets with many phenotypes.

POST is proposed as a general procedure that can robustly evaluate the association of a gene-set with several different types of phenotypic data (categorical, ordinal, continuous, or censored). For each gene-set, POST transforms the gene profiles into a set of eigenvectors and then uses statistical modeling to compute a set of z-statistics that measure the association of each eigenvector with the phenotype. The overall gene-set statistic is the sum of squared z-statistics weighted by the corresponding eigenvalues. Finally, bootstrapping is used to compute a p-value. POST may evaluate associations with or without adjustment for covariates.

In evaluating the association of 875 biological processes with the time to relapse of pediatric acute [myeloid](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/myeloid) leukemia, POST identified the well-known [oncogenic](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/carcinogenesis) WNT [signaling pathway](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/signal-transduction) as its top hit. These results indicate that POST can be a very useful tool for evaluating the association of a gene-set with a variety of different phenotypes. We have developed an R package named POST which is freely available in [Bioconductor](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/bioconductor).

<https://www.sciencedirect.com/science/article/pii/S1046202317304978?via%3Dihub>

1. **bcSeq: An R Package for Fast Sequence Mapping in High-throughput shRNA and CRISPR Screens.**

**Abstract**

CRISPR-Cas9 and shRNA high-throughput sequencing screens have abundant applications for basic and translational research. Methods and tools for the analysis of these screens must properly account for sequencing error, resolve ambiguous mappings among similar sequences in the barcode library in a statistically principled manner, and be computationally efficient. bcSeq, an open source R package that implements a fast and parallelized algorithm for mapping high-throughput sequencing reads to a barcode library while tolerating sequencing error. The algorithm uses a Trie data structure for speed and resolves ambiguous mappings by using a statistical sequencing error model based on Phred scores for each read.

<https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty402/5001385>

<http://bioconductor.org/packages/bcSeq/>.

1. **Using meshes for MeSH term enrichment and semantic analyses**.

MeSH (Medical Subject Headings) is the NLM controlled vocabulary used to manually index articles for MEDLINE/PubMed. MeSH terms were associated by Entrez Gene ID by three methods, gendoo, gene2pubmed and RBBH. This association is fundamental for enrichment and semantic analyses. meshes supports enrichment analysis (over-representation and gene set enrichment analysis) of gene list or whole expression profile. Meshes supports more than 70 species and provides high quality visualization methods to help interpreting analysis results.

**Availability**

meshes is released under Artistic-2.0 License. The source code and documents are freely available through Bioconductor

<https://www.bioconductor.org/packages/meshes>

<https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty410/5001391>

1. **RnaSeqSampleSize: real data based sample size estimation for RNA sequencing.**

### Abstract

#### BACKGROUND:

One of the most important and often neglected components of a successful RNA sequencing (RNA-Seq) experiment is sample size estimation. A few negative binomial model-based methods have been developed to estimate sample size based on the parameters of a single gene. However, thousands of genes are quantified and tested for differential expression simultaneously in RNA-Seq experiments. Thus, additional issues should be carefully addressed, including the false discovery rate for multiple statistic tests, widely distributed read counts and dispersions for different genes.

#### RESULTS:

To solve these issues, we developed a sample size and power estimation method named RnaSeqSampleSize, based on the distributions of gene average read counts and dispersions estimated from real RNA-seq data. Datasets from previous, similar experiments such as the Cancer Genome Atlas (TCGA) can be used as a point of reference. Read counts and their dispersions were estimated from the reference's distribution; using that information, we estimated and summarized the power and sample size. RnaSeqSampleSize is implemented in R language and can be installed from Bioconductor website.

A user friendly web graphic interface is provided at

<http://cqs.mc.vanderbilt.edu/shiny/RnaSeqSampleSize/>

#### CONCLUSIONS:

RnaSeqSampleSize provides a convenient and powerful way for power and sample size estimation for an RNAseq experiment. It is also equipped with several unique features, including estimation for interested genes or pathway, power curve visualization, and parameter optimization.

1. **Next-generation Sequence-analysis Toolkit (NeST): A standardized bioinformatics framework for analyzing Single Nucleotide Polymorphisms in next-generation sequencing data**

NeST (NGS-analysis Toolkit), a modular consensus-based variant calling framework. NeST uses a combination of variant callers to overcome potential biases of an individual method used alone. NeST consists of four modules, that integrate open-source bioinformatics tools, a custom Variant Calling Format (VCF) parser and a summarization utility, that generate high-quality consensus variant calls. NeST was validated using targeted-amplicon deep sequencing data from 245 Plasmodium falciparum isolates to identify single-nucleotide polymorphisms conferring drug resistance. The results were verified using Sanger sequencing data for the same dataset in a supporting publication.NeST offers a user-friendly pipeline for variant calling with standardized outputs and minimal computational demands for easy deployment for use with various organisms and applications.

<https://www.biorxiv.org/content/early/2018/05/21/323535?%3Fcollection>=