**Bioinformatics**

[**Feature Specific Quantile Normalization Enables Cross-Platform Classification of Molecular Subtypes using Gene Expression Data**](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty026/4816109)

Jennifer M Franks Guoshuai Cai Michael L Whitfield (Dartmouth College)

**Abstract**

**Motivation**

Molecular subtypes of cancers and autoimmune disease, defined by transcriptomic profiling, have provided insight into disease pathogenesis, molecular heterogeneity, and therapeutic responses. However, technical biases inherent to different gene expression profiling platforms present a unique problem when analyzing data generated from different studies. Currently, there is a lack of effective methods designed to eliminate platform-based bias. We present a method to normalize and classify RNA-seq data using machine learning classifiers trained on DNA microarray data and molecular subtypes in two datasets: breast invasive carcinoma (BRCA) and colorectal cancer (CRC).

**Results**

Multiple analyses show that feature specific quantile normalization (FSQN) successfully removes platform-based bias from RNA-seq data, regardless of feature scaling or machine learning algorithm. We achieve up to 98% accuracy for BRCA data and 97% accuracy for CRC data in assigning molecular subtypes to RNA-seq data normalized using FSQN and a support vector machine trained exclusively on DNA microarray data. We find that maximum accuracy was achieved when normalizing RNA-seq datasets that contain at least 25 samples. FSQN allows comparison of RNA-seq data to existing DNA microarray datasets. Using these techniques, we can successfully leverage information from existing gene expression data in new analyses despite different platforms used for gene expression profiling.

**Availability**

FSQN has been submitted as an R package to CRAN. All code used for this study is available on Github (<https://github.com/jenniferfranks/FSQN>).

**Bioinformatics**

[**Splice Expression Variation Analysis (SEVA) for Inter-tumor Heterogeneity of Gene Isoform Usage in Cancer**](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty004/4802220)

Bahman Afsari, et al Alexander V Favorov Elana J Fertig (JHU group)

**Abstract**

**Motivation**

Current bioinformatics methods to detect changes in gene isoform usage in distinct phenotypes compare the relative expected isoform usage in phenotypes. These statistics model differences in isoform usage in normal tissues, which have stable regulation of gene splicing. Pathological conditions, such as cancer, can have broken regulation of splicing that increases the heterogeneity of the expression of splice variants. Inferring events with such differential heterogeneity in gene isoform usage requires new statistical approaches.

**Results**

We introduce Splice Expression Variability Analysis (SEVA) to model increased heterogeneity of splice variant usage between conditions (e.g., tumor and normal samples). SEVA uses a rank-based multivariate statistic that compares the variability of junction expression profiles within one condition to the variability within another. Simulated data show that SEVA is unique in modeling heterogeneity of gene isoform usage, and benchmark SEVA’s performance against EBSeq, DiffSplice, and rMATS that model differential isoform usage instead of heterogeneity. We confirm the accuracy of SEVA in identifying known splice variants in head and neck cancer and perform cross-study validation of novel splice variants. A novel comparison of splice variant heterogeneity between subtypes of head and neck cancer demonstrated unanticipated similarity between the heterogeneity of gene isoform usage in HPV-positive and HPV-negative subtypes and anticipated increased heterogeneity among HPV-negative samples with mutations in genes that regulate the splice variant machinery.

**Conclusion**

These results show that SEVA accurately models differential heterogeneity of gene isoform usage from RNA-seq data.

**Availability**

SEVA is implemented in the R/Bioconductor package GSReg.

**Bioinformatics**

# Computational Drug Repositioning using Low-Rank Matrix Approximation and Randomized Algorithms

Huimin Luo Min Li Shaokai Wang Quan Liu Yaohang Li Jianxin Wang (CentralSouth)

**Abstract**

**Motivation**

Computational drug repositioning is an important and efficient approach towards identifying novel treatments for diseases in drug discovery. The emergence of large-scale, heterogeneous biological and biomedical datasets has provided an unprecedented opportunity for developing computational drug repositioning methods. The drug repositioning problem can be modeled as a recommendation system that recommends novel treatments based on known drug-disease associations. The formulation under this recommendation system is matrix completion, assuming that the hidden factors contributing to drug-disease associations are highly correlated and thus the corresponding data matrix is low-rank. Under this assumption, the matrix completion algorithm fills out the unknown entries in the drug-disease matrix by constructing a low-rank matrix approximation, where new drug-disease associations having not been validated can be screened.

**Results**

In this work, we propose a drug repositioning recommendation system (DRRS) to predict novel drug indications by integrating related data sources and validated information of drugs and diseases. Firstly, we construct a heterogeneous drug-disease interaction network by integrating drug-drug, disease-disease, and drug-disease networks. The heterogeneous network is represented by a large drug-disease adjacency matrix, whose entries include drug pairs, disease pairs, known drug-disease interaction pairs and unknown drug-disease pairs. Then, we adopt a fast Singular Value Thresholding (SVT) algorithm to complete the drug-disease adjacency matrix with predicted scores for unknown drug-disease pairs. The comprehensive experimental results show that DRRS improves the prediction accuracy compared with the other state-of-the-art approaches. In addition, case studies for several selected drugs further demonstrate the practical usefulness of the proposed method.

**Availability**

<http://bioinformatics.csu.edu.cn/resources/softs/DrugRepositioning/DRRS/index.html>

**BMC Bioinformatics**

# [Scuba: scalable kernel-based gene prioritization](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2025-5)

Guido Zampieri†, Dinh Van Tran†, Michele Donini, Nicolò Navarin, Fabio Aiolli, Alessandro Sperduti and Giorgio Valle (Padova, Italy)

## Abstract

### Background

The uncovering of genes linked to human diseases is a pressing challenge in molecular biology and precision medicine. This task is often hindered by the large number of candidate genes and by the heterogeneity of the available information. Computational methods for the prioritization of candidate genes can help to cope with these problems. In particular, kernel-based methods are a powerful resource for the integration of heterogeneous biological knowledge, however, their practical implementation is often precluded by their limited scalability.

### Results

We propose Scuba, a scalable kernel-based method for gene prioritization. It implements a novel multiple kernel learning approach, based on a semi-supervised perspective and on the optimization of the margin distribution. Scuba is optimized to cope with strongly unbalanced settings where known disease genes are few and large scale predictions are required. Importantly, it is able to efficiently deal both with a large amount of candidate genes and with an arbitrary number of data sources. As a direct consequence of scalability, Scuba integrates also a new efficient strategy to select optimal kernel parameters for each data source. We performed cross-validation experiments and simulated a realistic usage setting, showing that Scuba outperforms a wide range of state-of-the-art methods.

### Conclusions

Scuba achieves state-of-the-art performance and has enhanced scalability compared to existing kernel-based approaches for genomic data. This method can be useful to prioritize candidate genes, particularly when their number is large or when input data is highly heterogeneous. The code is freely available at <https://github.com/gzampieri/Scuba>

# BMC Bioinformatics

# [Feature selection for high-dimensional temporal data](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2023-7)

Michail TsagrisEmail authorView ORCID ID profile, Vincenzo Lagani and Ioannis Tsamardinos (Greece)

## Abstract

### Background

Feature selection is commonly employed for identifying collectively-predictive biomarkers and biosignatures; it facilitates the construction of small statistical models that are easier to verify, visualize, and comprehend while providing insight to the human expert. In this work we extend established constrained-based, feature-selection methods to high-dimensional “omics” temporal data, where the number of measurements is orders of magnitude larger than the sample size. The extension required the development of conditional independence tests for temporal and/or static variables conditioned on a set of temporal variables.

### Results

The algorithm is able to return multiple, equivalent solution subsets of variables, scale to tens of thousands of features, and outperform or be on par with existing methods depending on the analysis task specifics.

### Conclusions

The use of this algorithm is suggested for variable selection with high-dimensional temporal data.