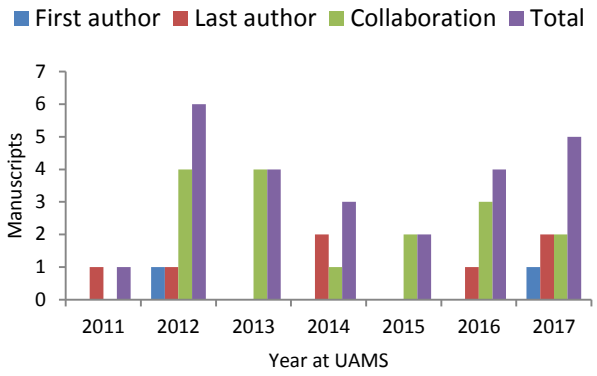
ANNOTATED BIBLIOGRAPHY



Annotated Bibliography Figure 1. Peer-reviewed publications for Dr. Glazko are listed per year since she joined the faculty at UAMS. Tenure clock stopped at 2014 and resumed at 2015.

As detailed in Dr. Glazko's CV, she has published 71 peer-reviewed papers in her field of study. Of these, Dr. Glazko published 47 prior to her faculty appointment at UAMS and 24 as an Assistant Professor at UAMS (Publications Figure 1) in the Basic Scientist Tenure Track.

The five publications presented here were purposefully selected to illustrate how the work that Dr. Glazko started at the University of Rochester Medical Center (URMC), specifically in developing Gene Set (pathway) Analysis methods for omics data, has progressed at UAMS. The work started with a first author paper in *Bioinformatics* (first paper in the list below). Over the years at UAMS the development of Dr. Glazko's initial idea spread in several new directions and resulted in 6 publications that became a valuable contribution in the field, culminating with the publication, in 2017, of the paper about the software package Gene Set Analysis in R. This software package has been downloaded by other investigators more than 6,000 times. Each of the manuscripts is included on subsequent pages. For each publication below, the percent effort from Dr. Glazko, the Journal Impact Factor, and a brief overview are provided.

1. **Glazko G.,** Emmert-Streib, F. (2009). Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics*, **25**:2348-2354.

   80% effort from Dr. Glazko; Journal Impact Factor 5,766

   Overview: In this paper for the first time three different null hypotheses that can be tested by Gene Set Analysis (GSA) approaches were explicitly formulated and it was shown that multivariate GSA approaches outperform univariate GSA approaches when expression profiles are moderately (and highly) correlated. The paper was groundbreaking in a sense that before univariate and different multivariate GSA approaches were blindly applied to omics data without realizing that they do test different null hypotheses and, as a consequence, may have different results on the same data sets. The application of different test statistics to biological data reveals that three statistics (sum of squared *t*-tests, Hotelling's T^2, *N*-statistic), testing different null hypotheses, find some common but also some complementing differentially expressed gene sets under specific settings. This demonstrates that due

to complementing null hypotheses each test projects on different aspects of the data and for the analysis of biological data it is beneficial to use all three tests simultaneously instead of focusing exclusively on just one. The paper was cited 71 times.

2. Rahmatallah Y.* Emmert-Streib F., **Glazko G**. (2012). Gene set analysis for self-contained tests: complex null and specific alternative hypotheses. *Bioinformatics,* **28**:3073-80.

   *Dr. Rahmatallah was a postdoctoral fellow working with Dr. Glazko for 5 years (2011-2016).

   100% effort from Dr. Glazko's laboratory; Journal Impact Factor 5.766

   Overview: In this study for the first time we suggest that the biological interpretability of experimental results under the self-contained Gene Set Analysis (GSA) framework can be increased by applying multivariate generalizations of the Kolmogorov–Smirnov (KS) and Radial Kolmogorov Smirnov (RKS) non-parametric two-sample tests. The KS and RKS tests have important differences from conventional tests: KS is exclusively sensitive to shift and RKS is mostly sensitive to scale alternatives. The analysis of real solid tumor expression data confirms the major trends in the tests' power, as observed in simulations. We found that pathways, detected exclusively by RKS, were tumor-specific, while pathways detected by KS were far less specific. The pathways, selected under different alternatives, potentially should help in interpreting the reasons of underlying phenotypic changes. The intersection of genes in these pathways can constitute a phenotype-specific gene signature, different for different alternatives and important in the follow-up studies. The paper was cited 22 times.

3. Rahmatallah Y.*, Emmert-Streib F., **Glazko G**. (2014). Gene Set Net Correlations Analysis (GSNCA): A multivariate differential coexpression test for gene sets. *Bioinformatics,* **30**:360-368.

   *Dr. Rahmatallah was a postdoctoral fellow working with Dr. Glazko for 5 years (2011-2016).

   100% effort from Dr. Glazko's laboratory; Journal Impact Factor 5.766

   Overview: Gene set analysis approaches primarily focus on identifying differentially expressed gene sets (pathways). In this study we propose Gene Sets Net Correlations Analysis (GSNCA), a multivariate differential co-expression test that accounts for the complete correlation structure between genes. In GSNCA, weight factors are assigned to genes in proportion to the genes' cross-correlations (intergene correlations). The problem of finding the weight vectors is formulated as an eigenvector problem with a unique solution. GSNCA tests the null hypothesis that for a gene set there is no difference in the weight vectors of the genes expressed in two conditions. In simulation studies and the analyses of experimental data, we demonstrate that GSNCA captures changes in the structure of genes' cross-correlations rather than differences in the averaged pairwise correlations. Thus, GSNCA infers differences in co-expression networks, however, bypassing method-dependent steps of network inference. In summary, GSNCA is a new approach for the analysis of differentially coexpressed pathways that also evaluates the importance of the genes in the pathways, thus providing unique information that may result in novel biological hypotheses. The paper was cited 28 times.

4. Rahmatallah Y.*, Emmert-Streib F., **Glazko G.** (2016). Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. *Brief Bioinform*., **17**:393-407.

   *Dr. Rahmatallah was a postdoctoral fellow working with Dr. Glazko for 5 years (2011-2016).

100% effort from Dr. Glazko's laboratory; Journal Impact Factor 8.399

Overview: Gene Set Analysis (GSA) approaches were initially developed for microarray data and their adaptation for transcriptome sequencing (RNA-seq) is still an active area of research. Here, for the first time, we provide a brief review of several statistically different GSA approaches (competitive and self-contained) that can be adapted from microarrays practice as well as those specifically designed for RNA-seq. We evaluate their performance on simulated and real RNA-seq data. The performance of various GSA approaches depends only on the statistical hypothesis they test and does not depend on whether the test was developed for microarrays or RNA-seq data. Interestingly, we found that competitive methods have lower power as well as robustness for the samples' heterogeneity than self-contained methods, leading to poor reproducibility of results. We also found that the power of unsupervised competitive methods depends on the balance between up- and down-regulated genes in tested gene sets. These properties of competitive methods have been overlooked before. Our evaluation provides a concise guideline for selecting GSA approaches best performing under particular experimental settings in the context of RNA-seq. The paper was cited 16 times.

5. Rahmatallah Y.*, Zybaylov B., Emmert-Streib F., **Glazko G.** (2017). GSAR: Bioconductor package for Gene Set Analysis in R. *BMC Bioinformatics*, **18**: 61.

*Dr. Rahmatallah was a postdoctoral fellow working with Dr. Glazko for 5 years (2011-2016).

100% effort from Dr. Glazko's laboratory; Journal Impact Factor 2.435

Overview: The paper presents the software package GSAR (Gene Set Analysis in R), an open-source R/Bioconductor software package for gene set analysis. It implements self-contained multivariate non-parametric statistical methods and tests a complex null hypothesis against specific alternatives, such as differences in mean (shift), variance (scale), or net correlation structure. These approaches were developed in our previous publications (1-4) and were implemented in GSAR. The methods in the GSAR package are applicable to any type of omics data that can be represented in a matrix format. The package also provides a graphical visualization tool, based on the union of two minimum spanning trees, for correlation networks to examine the change in the correlation structures of a gene set between two conditions and to highlight influential genes (hubs). The package was downloaded more than 6,000 times; that is more than 6000 people have used the GSAR package for their data analysis.