

**Abstract Identifying Number: 1006503**

**An Automated Regression Testing Approach for High-Throughput Molecular Profiling Data**

**Erich A. Peterson<sup>1\*</sup>, Horacio Gomez-Acevedo<sup>1</sup>, Jason Liem<sup>1</sup>, Peter Liu<sup>1</sup>, and Donald J. Johann Jr.<sup>1</sup>**

<sup>1\*</sup>Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, 72205

**Background:** The development of complex software pipelines for the analysis of high throughput molecular profiling data is difficult. This is due in part to the number of different software tools coupled together, as well as the complexity and volume of the data generated by many platforms (e.g., Next Generation Sequencing). Additionally, the vast majority of pipeline tools/components are developed and maintained by a wide variety of members in the open source community, and require periodic update, integration, and subsequent re-validation of the pipeline proper. Essential to the testing of a pipeline's validity is the comparison between expected and observed results. To this end, one methodology commonly employed is to compare observed output to a gold-standard, or some known to be valid output. However, the manual inspection and gathering of statistics based on multiple high-dimensional datasets is tedious, and may introduce unfortunate errors.

**Results:** In this research, a method and software package was developed to automate the regression testing of pipeline outputs. The method presented uses a "plug-in" framework, which allows for the addition of any type of structured pipeline output file to be compared during a regression test. Under the aforementioned scheme, a configuration file defines the various attributes within the files to be compared. After execution of the regression test, the software package also presents valuable statistics based on the results. These statistics are output in an easy-to-consume textual format, which allows for the regression testing to be automated and incorporated into existing analysis pipelines validation procedures.

**Conclusion:** An easy-to-use and extensible framework for the automation of regression testing for high-throughput molecular profiling results improves the reliability of complex software pipelines. Automation plays an essential role in this method by reducing potential human error and increasing overall reliability.