UAMS
University of Arkansas for Medical Sciences

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

Erich A. Peterson, Horacio Gomez-Acevedo, Jason Liem, Peter Liu, and Donald J. Johann Jr.
University of Arkansas for Medical Sciences, Little Rock, AR

DBMI

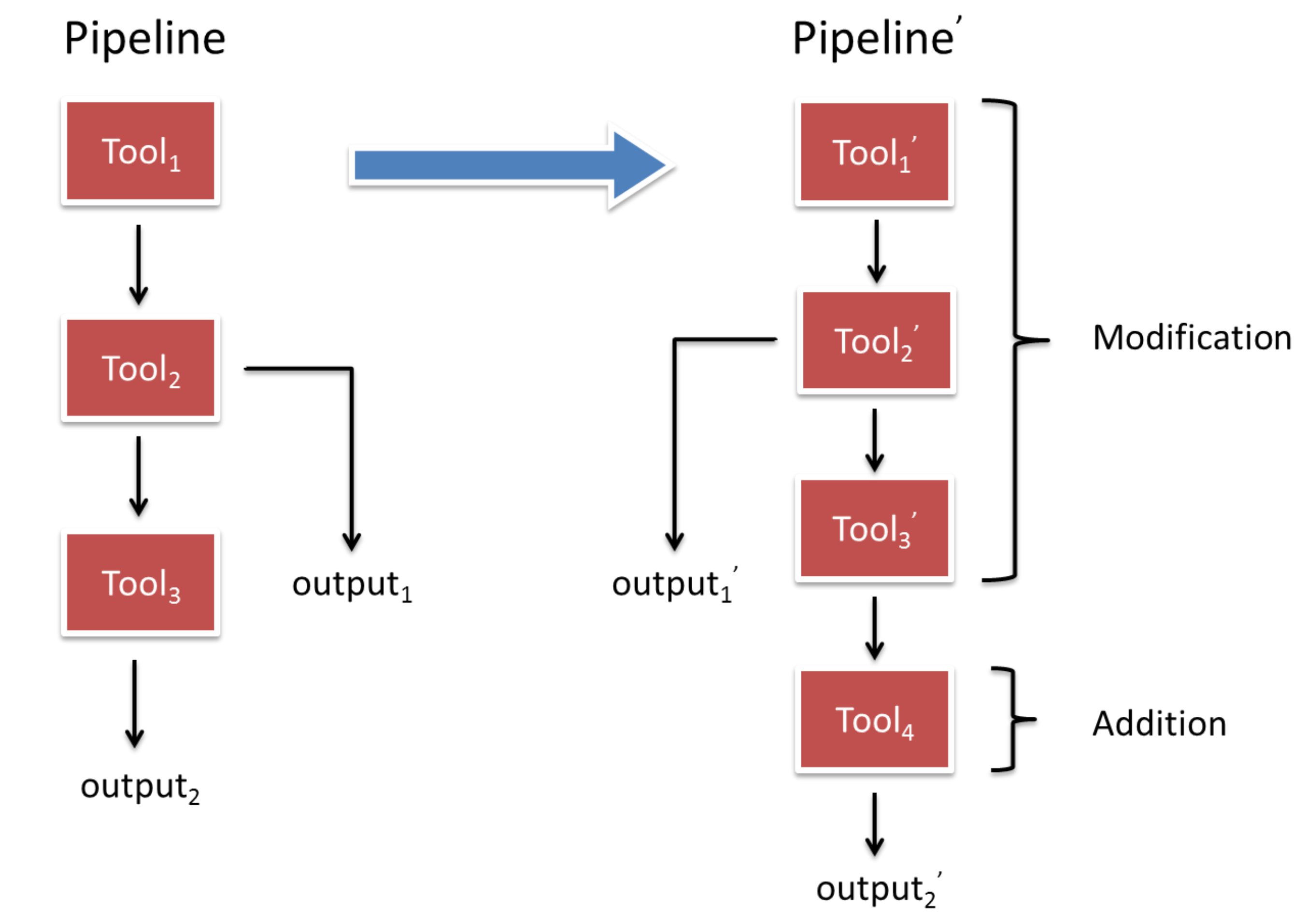
Abstract

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.

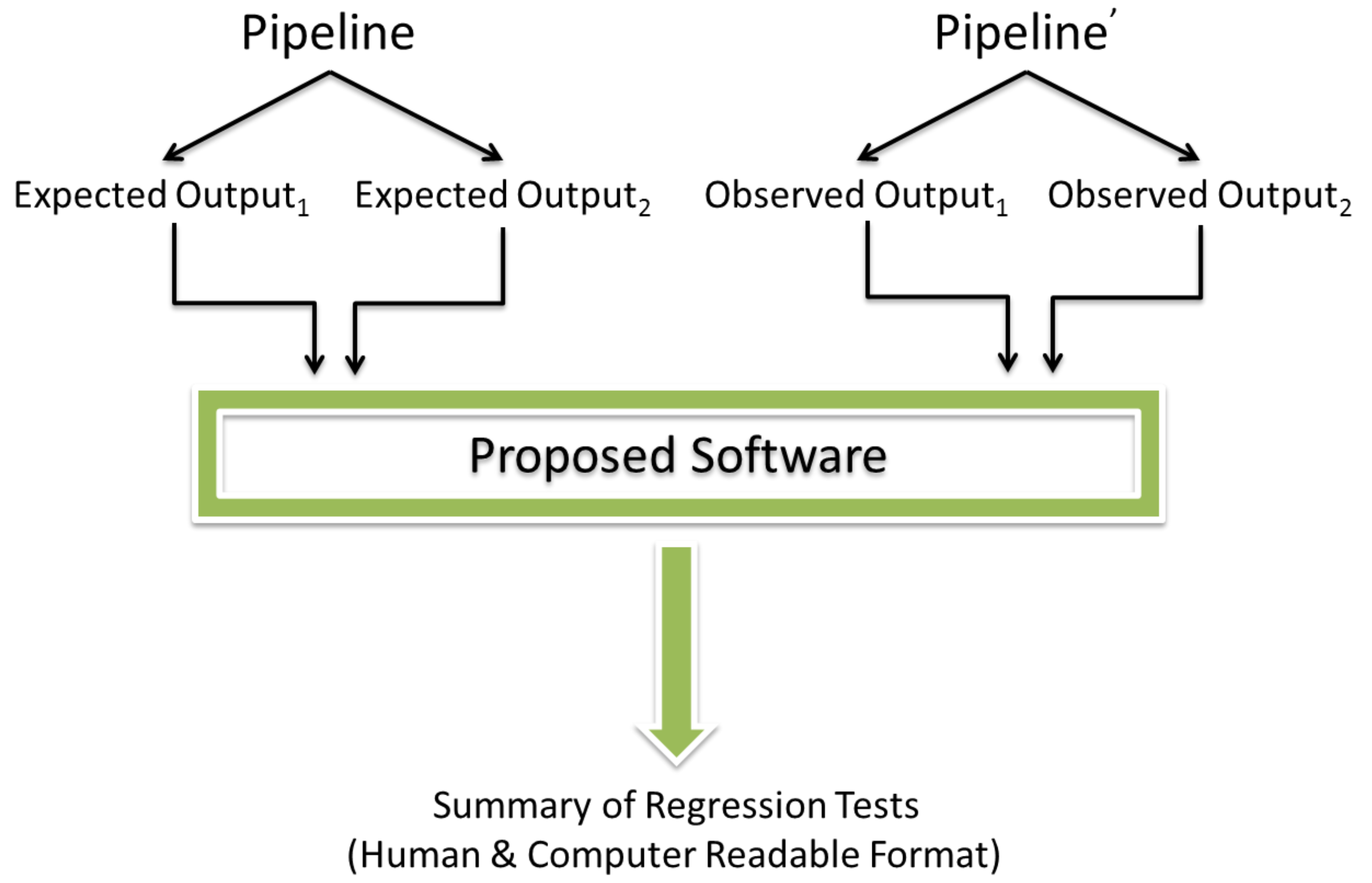
Results: In this research, a method and software package was developed to automate the regression testing of pipeline outputs. The method presented uses a "plug-in" framework, which allows for the addition of any type of structured pipeline output file to be compared during a regression test. Under the 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.

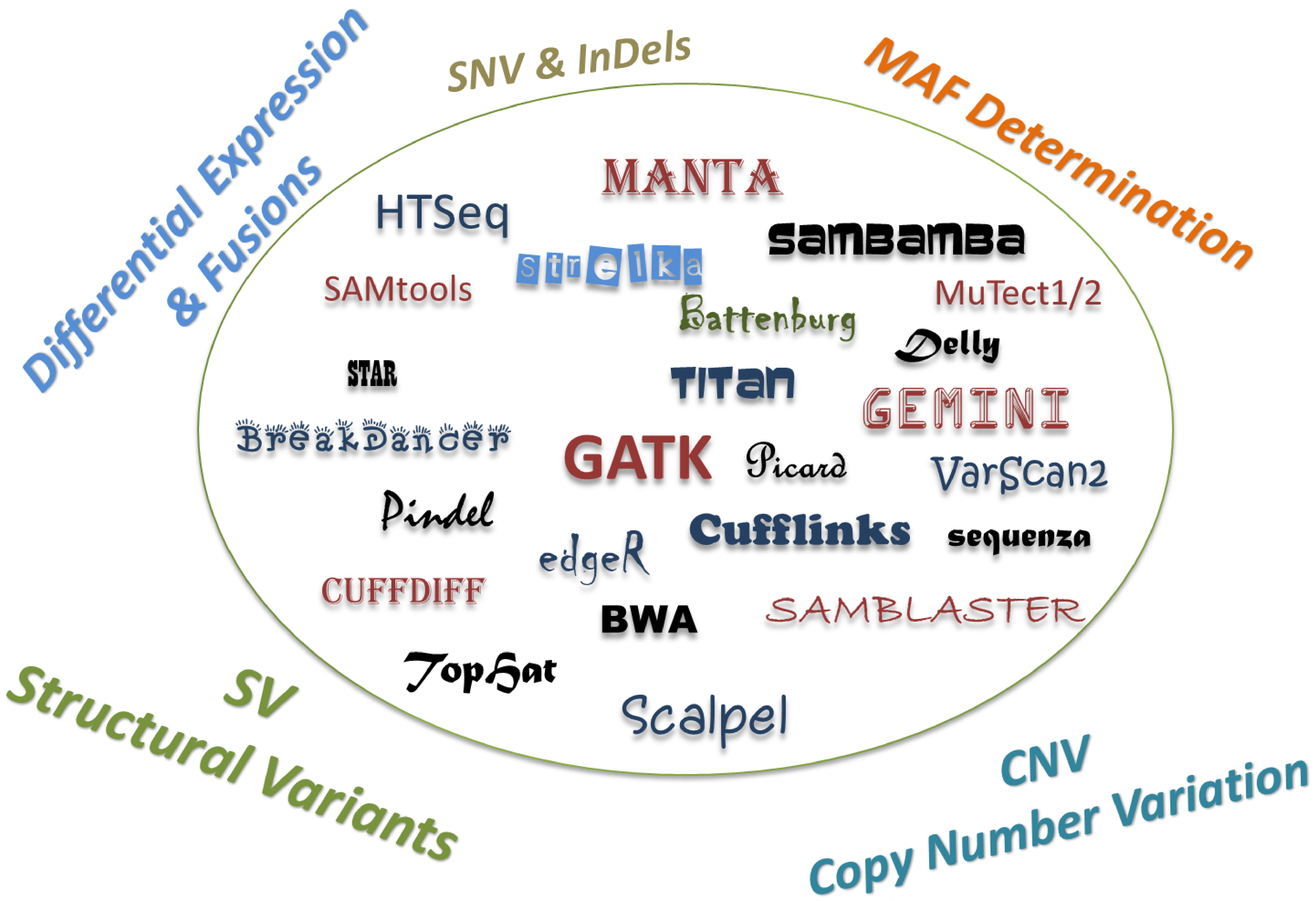
Problem: Pipelines are complex - many tools & steps.
Challenge: How to verify updates and changes?

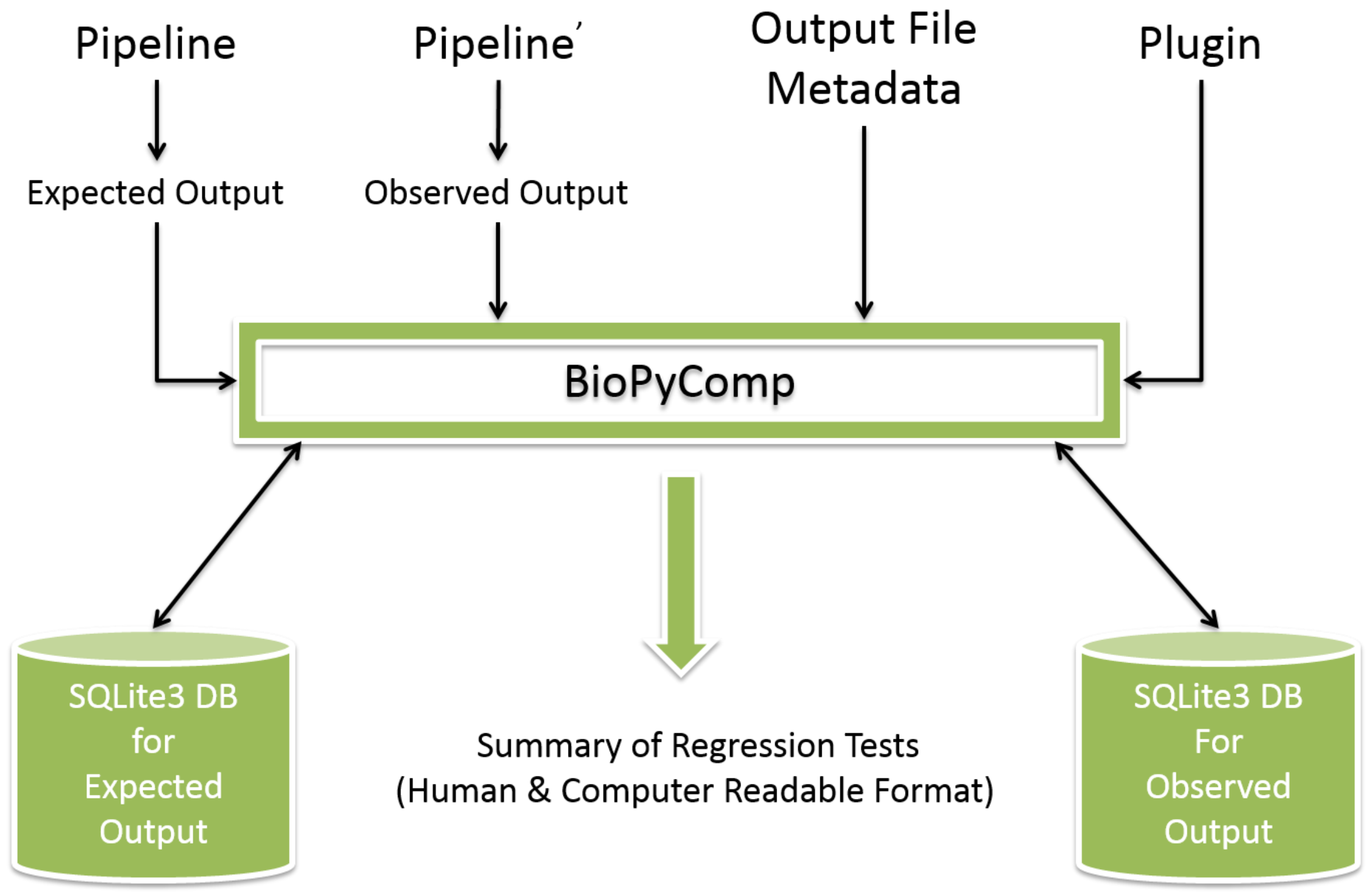


Solution: Use curated *Expected* vs. *Observed* analytical approach. Design flexible plugin software framework. Automate regression testing.



NGS - Many Pipeline Tools & Functions




Pipeline
Pipeline′
Output File Metadata
Plugin
Expected Output
Observed Output
BioPyComp
SQLite3 DB for Expected Output
Summary of Regression Tests
(Human & Computer Readable Format)
SQLite3 DB For Observed Output

Example: *BioPyComp* plugin for Variant Calling Format (VCF) files.

Expected Output

Variant Position	MAF	Variant Base
345678	20%	G
1234567	45%	T
123456789	60%	C

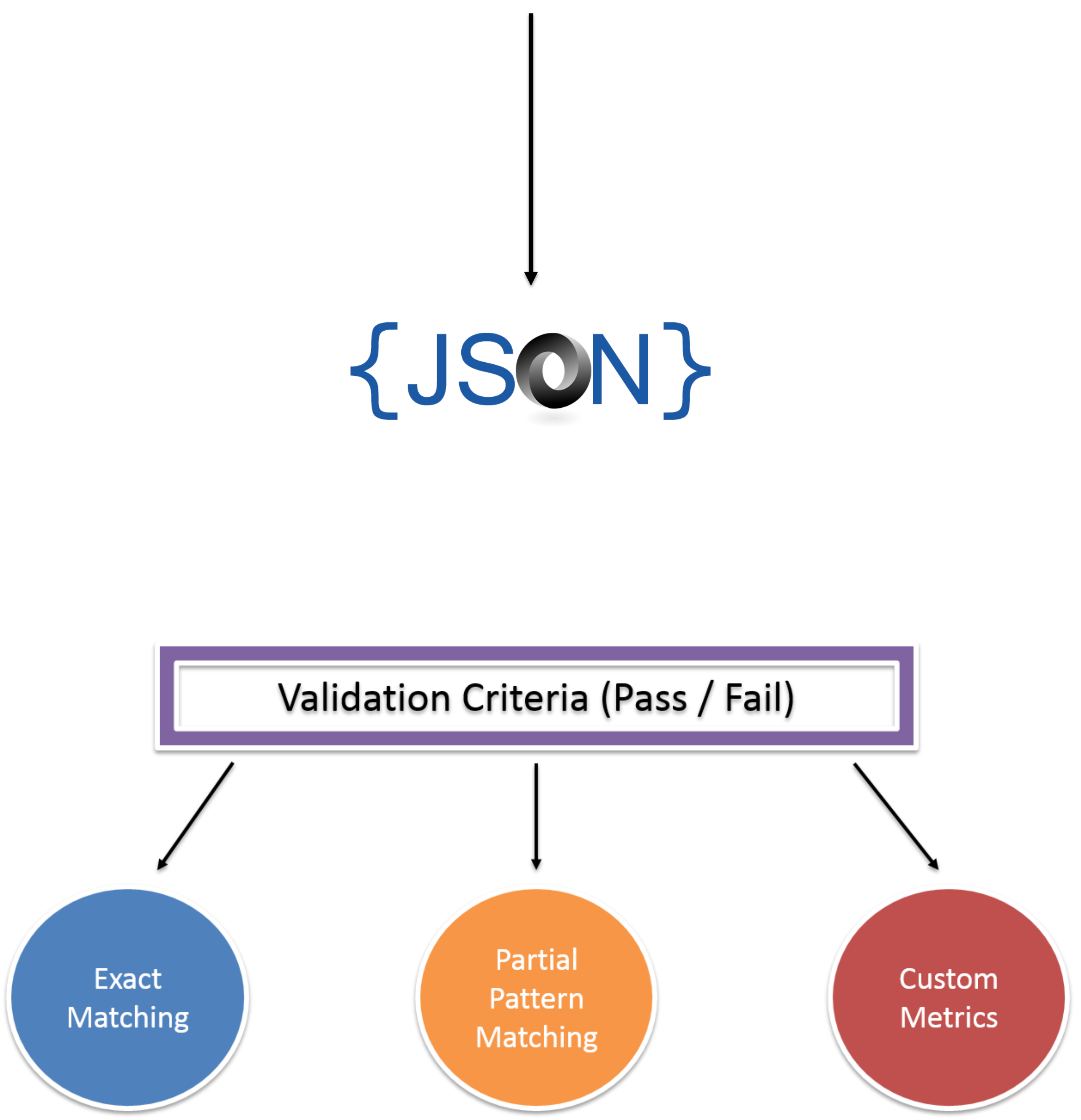
Observed Output

Variant Position	MAF	Variant Base
345678	22%	G
1234567	45%	T
3456789	30%	T

Output File Format Metadata:
- Variant Position (integer)
  - Values must be same
- MAF (float)
  - Absolute difference less than 5%
- Variant Base (string)
  - Value must be the same

Example Output

Distinct Records in Expected Output	Distinct Records in Observed Output	Matching Records	Percent Matching
3	3	2	66%



Future Directions:
- Open source code
- Add additional plugins for other file formats