Validate Workflow v0.9 (Atmosphere Images tutorial)

The Validate Workflow

This page is designed to aid users in navigating the Validate Workflow.

What is Validate?

"The purpose of Validate is to provide information on both SNP effect size estimation and identifying SNP capability performance for various GWAS and QTL tools. The eventual goal is two-fold: 1 Publish information about the performance of different tools for different types of simulation parameters (such as population structure and different levels of heritability) somewhere easily viewable for iPlant users. Essentially, we hope to show researchers when best to use a tool as compared with another. 2 Provide a pipeline or workflow for testing installed tools. This is to encourage iterations of the first goal."

- Dustin Landers (The architect of the original Validate program)

The workflow consists of several pieces of software called genome wide association study (GWAS) tools, and software to analyze the GWAS tool performance.

More specifically, the workflow includes:

- Simulate: A Python-based simulation software that also outputs the known-truth phenotypes for a given population
- Multiple GWAS tools:
    - FaST-LMM: GWAS analysis tool designed for large data sets, more specifically used to test all SNPs in a data set for statistical significance
    - GEMMA: A GWAS analysis tool specializing in standard linear mixed models and variations thereof
    - QxPak: A versatile statistics package specializing in statistical genomics and quantitative trait loci (QTL) analyses.
    - PLINK: Open-source software designed to convert data into usable formats and to perform basic, large scale analyses efficiently.
- Winnow: A Python-compatible known truth testing tool for genome wide association studies (i.e. a tool that evaluates other GWAS tools)
- Demonstrate: An R script that produces human-readable visual output from the results files produced from Winnow

How to get started

1. It is highly recommended that you watch the webinar "Getting Started with iPlant" given monthly by Jason Williams as an introduction to iPlant, and some of its features.
2. There are a series of accounts which need to be setup and software which needs to be downloaded before getting started. Follow this link and return after you have followed the instructions for setting up accounts.
3. You can acquaint yourself with Atmosphere here, generally though, atmosphere allows you to access a virtual machine where all of the necessary programs have been installed to run the workflow. This lab has several atmosphere images which have been launched although you will likely only need the validate image unless you want to work with a specific tool. Validate 0.9 is available as an Atmosphere image under the name Validate Workflow v0.9.
4. It can also be helpful to, Check here to learn more about stampede and check here to learn more about the Agave API. Stampede is housed at TACC and is the world's largest supercomputer dedicated to science. The Agave API is a tool for creating and implementing apps into stampede. The workflow can be operated exclusively on stampede, however, this process is under development and the following pages will be for use in atmosphere.

Learn about allocations
Learn about CyVerse's allocation policies here.

Next Steps

After looking through and completing the above you are ready to begin, you can either start at the simulate page,found here, or you can continue to scroll through this page to learn more about the validate project, and find links to useful information.

To learn more about the various offerings of the iPlant collaborative please check out the main page for getting started with the iPlant collaborative.

If you are interested in further developing the Validate workflow check here or here for a more statistically oriented guide.

When working in the atmosphere images terminal it can be helpful to know how to use iCommands. These are essentially commands to move data to and from the data store.

You may also want to learn how to install R packages into atmosphere, instructions for which can be found here .

For viewing the source code and additional information on any of these programs, please check the main Github repository.

Many other forms of Statistical method comparison software exist, such as DSCR. A basic comparison between the function and intent of DSCR and the Validate workflow can be found here.

The Validate workflow is still in development and we are testing it currently. If you notice any issues or have any comments we would greatly appreciate them!
Please contact us at labstapleton@gmail.com. Thank you for using our tools!

Particularly large datasets may require instances with more memory. Some GWAS tools can be computationally demanding, and if an instance lacks sufficient memory, the process may be "killed" mid computation. In our experience, filesets in excess of 1GB need at least 4GB of memory to guarantee processing.

Change Log

Version 0.9

- Revamped Winnow code structure and added options for p-value adjustments and covariate integrations
- Demonstrate visualization options have been updated for use with ggplot2 to include more colorful and descriptive graphics. In addition, all Demonstrate functions have been merged into a single R package: DemoMPlot.

Version 0.7

- Revamped repository structure on Github for easier access to source code, easier issue reporting, and more thorough, interactive documentation
- Added export function to Simulate for converting population data to PEDMAP format for use in other analysis programs
- Added gamma distribution option for Simulate with specifiable parameters
- Installed other GWAS tools for further customization: GEMMA, QxPak, FaST-LMM, and PLINK are all available on the instance with the freedom to use your own GWAS tool if desired
- Added additional Demonstrate function, *Demonstrate2*, for data without either heritability or population structure separation
- Created more detailed outputs with Demonstrate2 function including false/true positive histograms, scatterplots separated by analysis method, and comparison tables for specificity, sensitivity, and precision

Version 0.5

- Official introduction of Validate as a workflow rather than an individual program
- Added FaST-LMM for GWAS analysis of data
- Added Demonstrate for data visualization
- Renamed Validate program Winnow
- Added simulation data (Dong Wang data and syngenta data) to test analysis tools
- Winnow as Python-compatible software introduced to the workflow

Version 0.3

- New version of the Validate written in R released on Atmosphere Spring, 2014 – Included major expansion of features. Including ability to handle different file types (for truth files), some automatic file transformation (for truth files), ten new performance measures, and corrected a bug that forced Validate to fail when only a single result/output file was being validated. (Under Github repository name "ktaR")
- Released on DE Fall, 2013 – First release of the R version of validate. Version was an Alpha but most of all functionality was available. (Under Github repository name "ktaR")