Enabling large scale Genotype-Tissue Expression studies using Galaxy

Princeton BEEHIVE
Biological and Evolutionary Explorations using Hierarchical IntegratiVE statistical models

Genna Gliner[1], Ian McDowell[2], Barbara E Engelhardt[3]

1. Operations Research and Financial Engineering Department, Princeton University; 2. Computational Biology and Bioinformatics, Duke University; 3.Computer Science Department and Center for Statistics and Machine Learning, Princeton University

PRINCETON UNIVERSITY

Abstract

- The Princeton BEEHIVE Group is part of the Genotype-Tissue Expression (GTEx) consortium and develops statistical models and methods for statistical and functional genomics studies including expression quantitative trait loci (eQTL) detection studies, non-coding RNA regulation studies, and allele specific expression (ASE) studies.
- The creation, testing, and deployment of the processing pipelines for each of these different study types requires comprehensive analysis of large datasets through a dedicated pipeline used by all members of the group.
- With the ability to create custom tools and share and modify workflows, **Galaxy** provides a robust framework to develop this pipeline for use across our lab.
- In this poster we chronicle the evolution of the Princeton BEEHIVE Galaxy Pipeline by highlighting how our lab addressed the challenges of tissue specific analysis, data processing and organization, and training lab members to use Galaxy.

GTEx Consortium

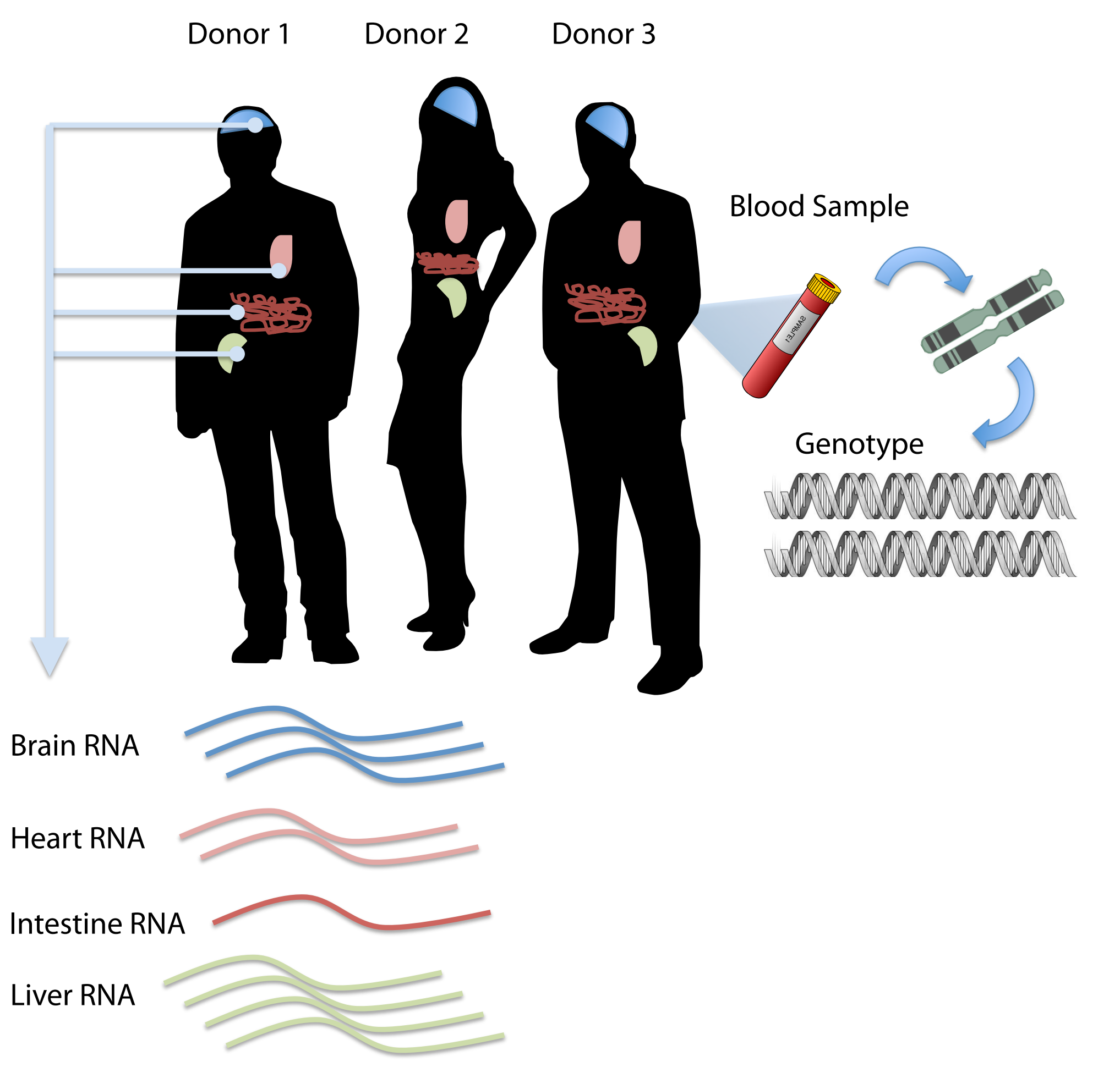


Figure 1 The Genotype-Tissue Expression (GTEx) project aims to provide a databank of samples from multiple human tissues from subjects who are densely genotyped. The goal is to provide the scientific community with a resource to study human genetic variation and regulation and how it relates to gene expression. The GTEx dataset provides a unique opportunity to analyze the relationship between gene expression levels and genetic variation across both tissues and individuals. The Phase 1 data release consists of approximately 450 individuals and over 9000 samples from up to 50 tissues per individual.

An Example Workflow

The following is an example analysis to illustrate how we can use Galaxy in the BEEHIVE lab for tissue specific analysis.

Data Preprocessing:

- In the GTEx data, each donor has a unique ID and each tissue sample collected from a donor has a unique sample ID.
- To compare genetic expression from each individual across tissues, we need to map the sample ID to the individual for each tissue.
- We performed this mapping using the **Join two Datasets** tool.

Tissue Specific Analysis:

- We performed an exploratory analysis to determine the relationship between tissues and gene expression at several known eQTLs documented on the GTEx portal.
- Created a Galaxy workflow to select samples from 8 tissues that contained ≈ 100 samples and the selected genes for analysis.
- Applied the following Galaxy tools: **Correlation**, **Perform Logistic Regression with vif**, **Compute partial R square**, **Compute RCVE**, **Principle Component Analysis**, **Perform Best-subsets Regression**, and **Correlation**.

Tissue	Samples
Whole Blood	177
Muscle - Skeletal	146
Lung	133
Artery - Tibial	118
Thyroid	113
Skin - Sun Exposed (Lower Leg)	109
Nerve - Tibial	98
Heart – Left Ventricle	97

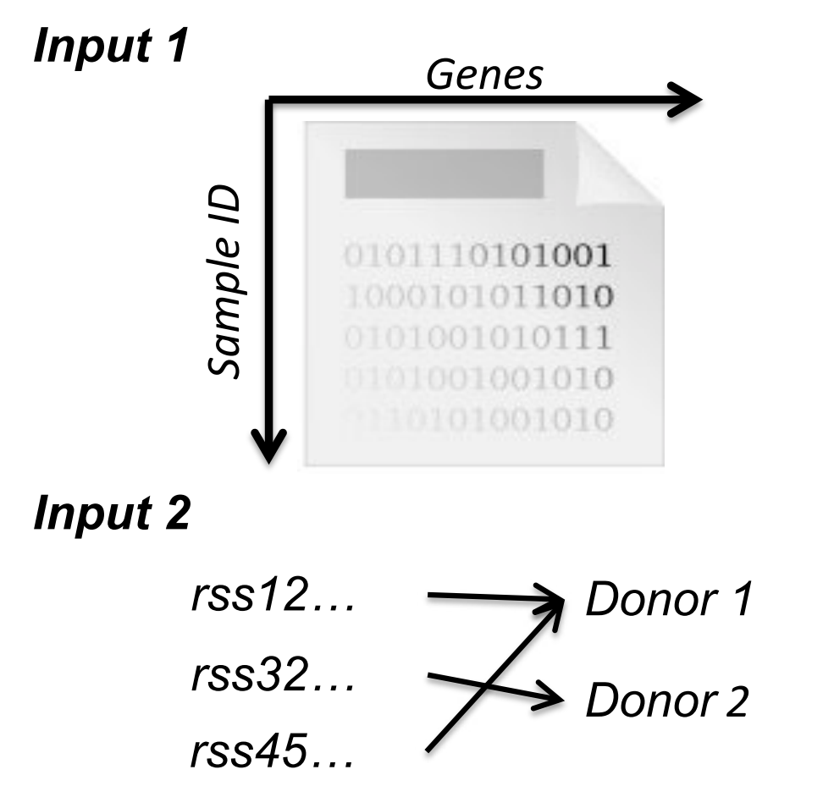
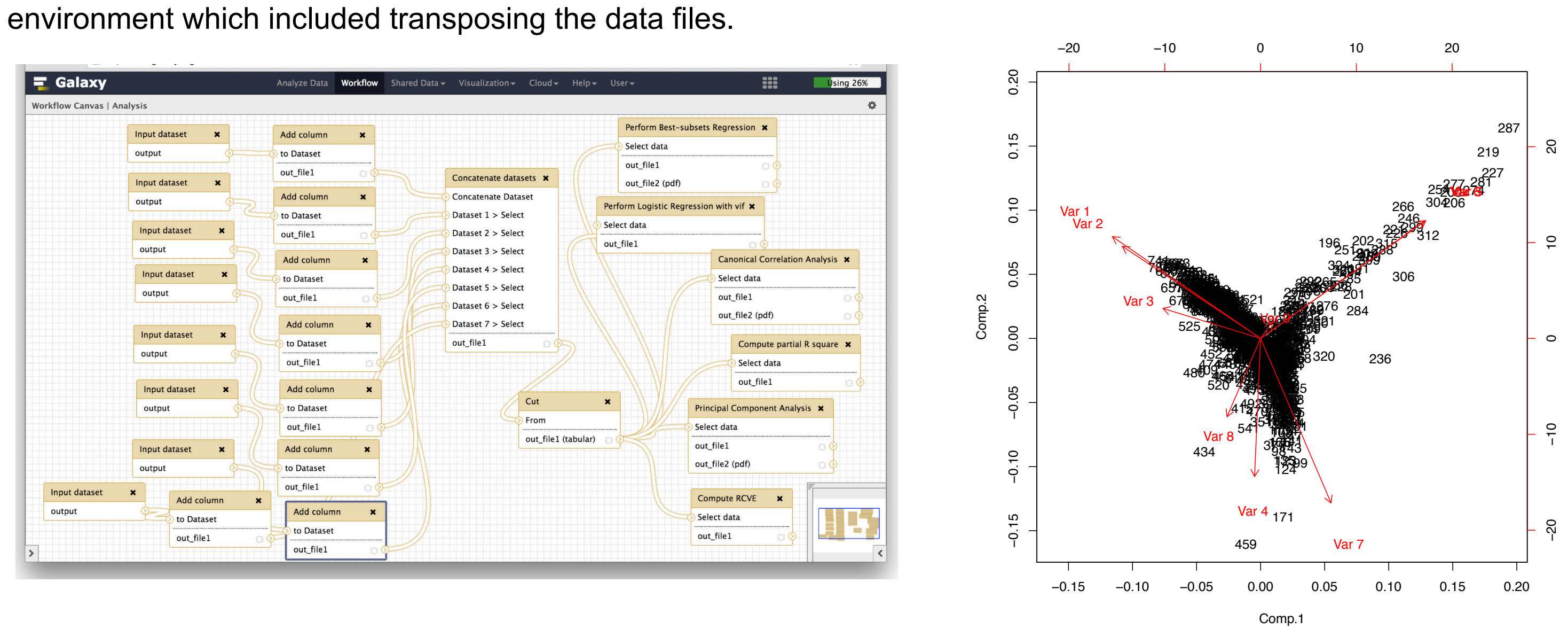


Figure 3: (Left) Table of tissues selected for analysis and the number of samples associated with each tissue. (Right) We used Galaxy to perform the mapping of samples to individuals using the Join two datasets tool. Since the Join two datasets tool performs the join using column labels, we had to perform preprocessing steps outside of the Galaxy environment which included transposing the data files.

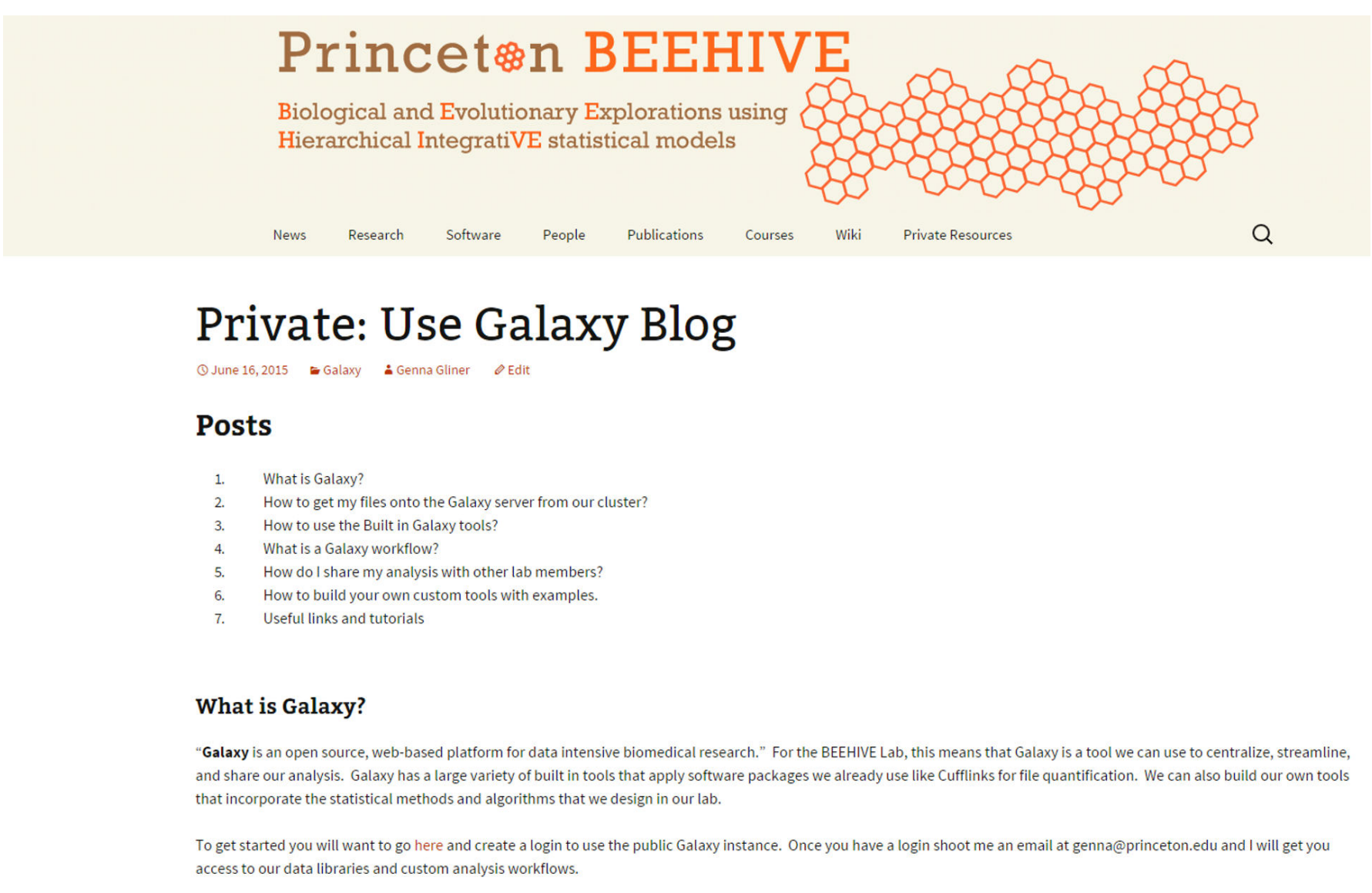


Component	1	2	3	4	5	6	7	8	9
Std. deviation	1.519	1.44	1.038	0.9985	0.9353	0.8651	0.6842	0.6256	0.2473
Proportion of Variance Explained	0.2564	0.2305	0.1197	0.1108	0.0972	0.08316	0.05202	0.04348	0.006793
Loadings	1	2	3	4	5	6	7	8	9
Tissue Label	-0.4447	0.3219	0.1632	0.01687	0.2747	0.03069	0.004727	0.7714	0.01398
EHMT1 (ENSG00000181090)	-0.4149	0.2913	0.05187	0.07422	0.4879	-0.02872	-0.4481	-0.5429	-0.01812
PNPLA7 (ENSG00000130653)	-0.2931	0.09492	0.4531	-0.1592	-0.5388	-0.6078	-0.08592	-0.08409	-0.01628
MAN1B1 (ENSG00000177239)	-0.018	-0.4367	0.2908	-0.04251	0.5652	-0.3916	0.4937	-0.07737	-0.003396
UAP1L1 (ENSG00000197355)	0.4958	0.3725	0.2915	-0.02194	0.1427	-0.06441	-0.03994	0.03403	-0.709
SAPCD2 (ENSG00000186193)	0.4918	0.3716	0.3066	-0.002669	0.1395	-0.08518	-0.05588	0.004924	0.7044
RP11-229P13.19 (ENSG00000238268)	0.2116	-0.5205	0.05723	0.05983	0.1203	-0.2003	-0.7317	0.2955	-0.00602
DNLZ (ENSG00000213221)	-0.1006	-0.2475	0.6598	-0.2388	-0.07407	0.6487	-0.04633	-0.08827	0.002768
FCN1 (ENSG00000085265)	0.03531	0.0518	-0.2514	-0.9518	0.1303	-0.05615	-0.07583	0.02878	0.01501

Figure 4: (Top Left) The Galaxy workflow that was created for analysis. (Top Right and Bottom) Results from the Principle Component Tool applied to the filtered dataset.

Challenges

- The public Galaxy instance has limited analysis tools available.
- While there are many instructional videos available, building Galaxy tool wrappers and including the proper libraries is nontrivial.
- Since we are seeking to perform analysis across tissues and individuals, we need to create tools that can work across columns and rows.
- To overcome the challenges specific to working with the GTEx data in Galaxy, I have developed a blog accessible to the BEEHIVE group members.



Princeton BEEHIVE
Biological and Evolutionary Explorations using Hierarchical IntegratiVE statistical models

Private: Use Galaxy Blog

June 16, 2015 Galaxy Genna Gliner Edit

Posts

1. What is Galaxy?
2. How to get my files onto the Galaxy server from our cluster?
3. How to use the Built in Galaxy tools?
4. What is a Galaxy workflow?
5. How do I share my analysis with other lab members?
6. How to build your own custom tools with examples.
7. Useful links and tutorials

What is Galaxy?

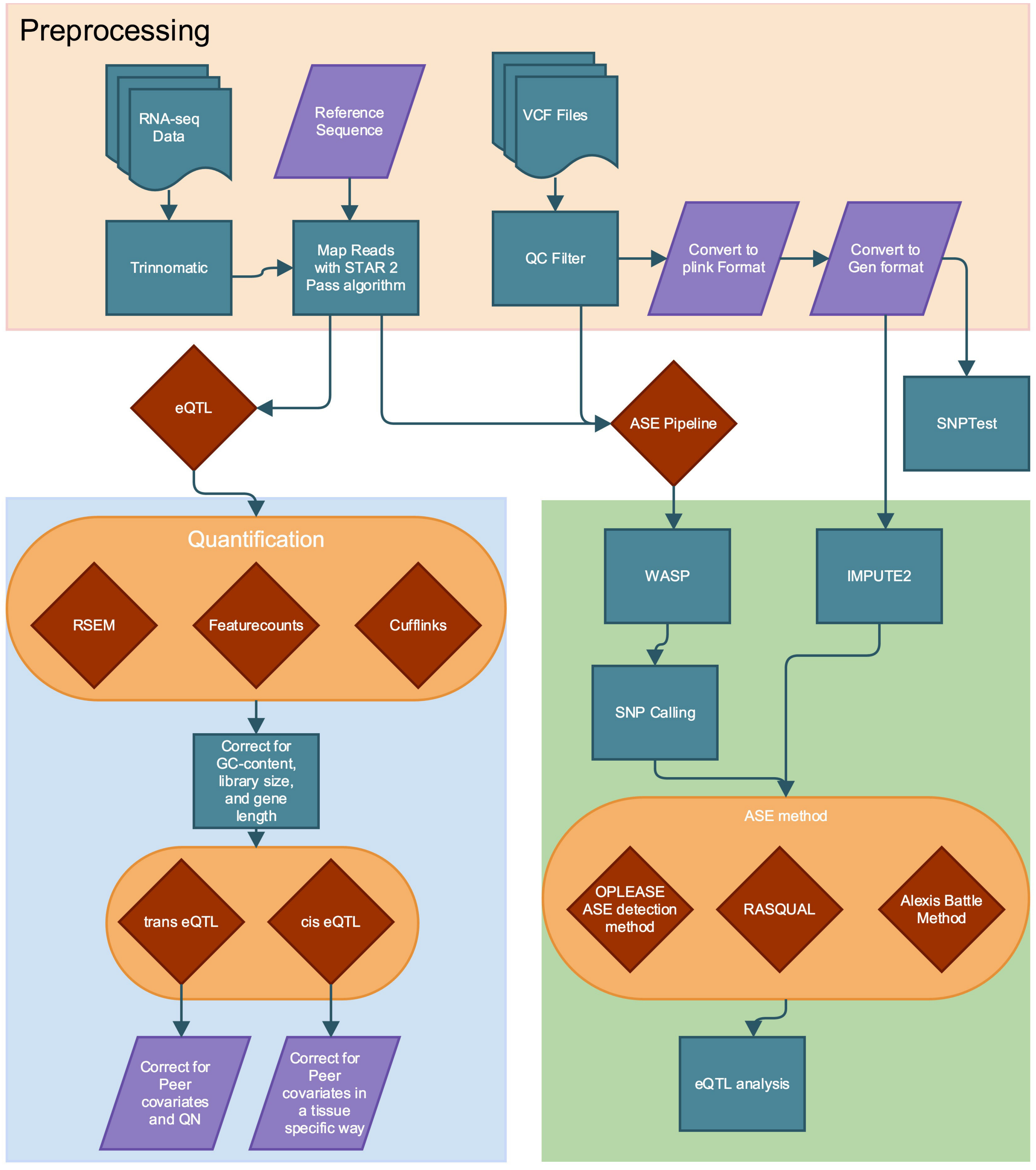
"**Galaxy** is an open source, web-based platform for data intensive biomedical research." For the BEEHIVE Lab, this means that Galaxy is a tool we can use to centralize, streamline, and share our analysis. Galaxy has a large variety of built in tools that apply software packages we already use like Cufflinks for file quantification. We can also build our own tools that incorporate the statistical methods and algorithms that we design in our lab.

To get started you will want to go here and create a login to use the public Galaxy instance. Once you have a login shoot me an email at genna@princeton.edu and I will get you access to our data libraries and custom analysis workflows.

Figure 2: This blog will serve as the main repository for all things related to Galaxy within the BEEHIVE Lab and will chronicle our efforts to integrate our complete data processing and analysis pipeline into Galaxy.

Future Plans

- Incorporate our complete data **processing** and **analysis** pipeline into Galaxy.
- This requires implementing our own tools and software packages that are not already included in the Galaxy toolshed.
- Our current pipeline has many opportunities to incorporate Galaxy workflows:
  1. Creating a workflow for the Preprocessing step to automatically process new files released by the GTEx Consortium.
  2. Implementing different workflows for each eQTL procedure will automate data processing.
  3. Creating separate workflows for each allelic specific expression (ASE) detection algorithm will allow us to directly compare methods.
- By implementing our pipeline in Galaxy, the BEEHIVE lab will have the flexibility to insert novel data processing algorithms or adjust our pipeline by developing Galaxy tools and incorporating them into the existing workflows.



Preprocessing

RNA-seq Data

Reference Sequence

VCF Files

Trinnomatic

Map Reads with STAR 2 Pass algorithm

QC Filter

Convert to plink Format

Convert to Gen format

eQTL

ASE Pipeline

SNPTest

Quantification

RSEM

Featurecounts

Cufflinks

Correct for GC-content, library size, and gene length

trans eQTL

cis eQTL

Correct for Peer covariates and QN

Correct for Peer covariates in a tissue specific way

WASP

IMPUTE2

SNP Calling

ASE method

QPLEASE ASE detection method

RASQUAL

Alexis Battle Method

eQTL analysis

Figure 5: The current data processing and analysis pipeline used by the BEEHIVE Lab. Once it is incorporated into Galaxy, we will have a flexible framework to process incoming data, implement new analysis techniques, and compare these new techniques to existing methods.

Acknowledgments

GG, IM, and BEE were supported by the Genotype-Tissue Expression (GTEx) consortium under the NIH grant (R01) MH101822 (+S2). Attendance at the GCC2015 for GG was supported by the GCC2015 scholarship grant.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplements to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822).

Contact: genna@princeton.edu