Enabling large scale Genotype-Tissue Expression studies using Galaxy

GENNA GLINER, OPERATIONS RESEARCH AND FINANCIAL ENGINEERING DEPARTMENT, PRINCETON UNIVERSITY

IAN MCDOWELL, COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, DUKE UNIVERSITY

BARBARA E ENGELHARDT, COMPUTER SCIENCE DEPARTMENT AND CENTER FOR STATISTICS AND MACHINE LEARNING, PRINCETON UNIVERSITY

Lab Introduction

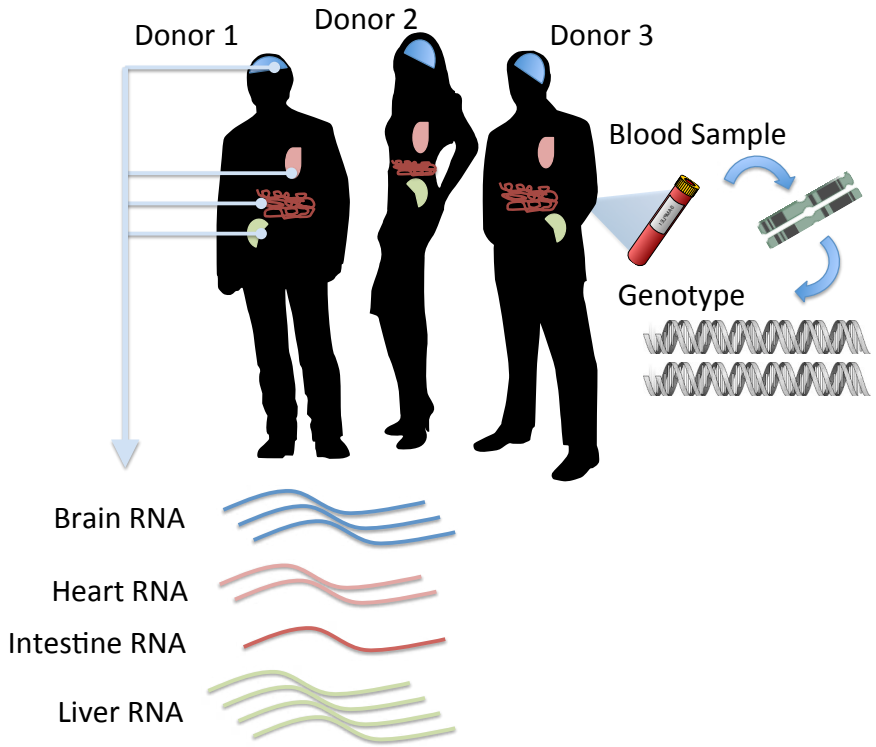
- The Biological and Evolutionary Explorations using Hierarchical IntegratiVE statistical models (BEEHIVE) lab is located at Princeton University Department of Computer Science
- The lab is headed by Professor Barbara Engelhardt
- The Princeton BEEHIVE Group develops statistical models and methods for high-dimensional genomic data
- This includes statistical and functional genomics studies for cis and trans expression quantitative trait loci (eQTL), non-coding RNA regulation studies, and allele specific expression studies

PRINCETON UNIVERSITY

Barbara Engelhardt

The GTEx Consortium

As part of the Genotype-Tissue Expression (GTEx) consortium, the BEEHIVE Lab is involved in processing vast quantities of RNA-sequencing and whole genome sequence data for different statistical and functional genomics studies



Goals

- Provide a databank of samples from multiple human tissues from densely genotyped individuals
- A resource to study human genetic variation and regulation and how it relates to gene expression
- Unique opportunity to analyze this relationship across both tissues and individuals
- The Pilot data release consists of approximately 175 genotyped individuals and over 3000 RNA samples from up to 50 tissues per individual

More information on GTEc: http://www.gtexportal.org/home/

Motivation

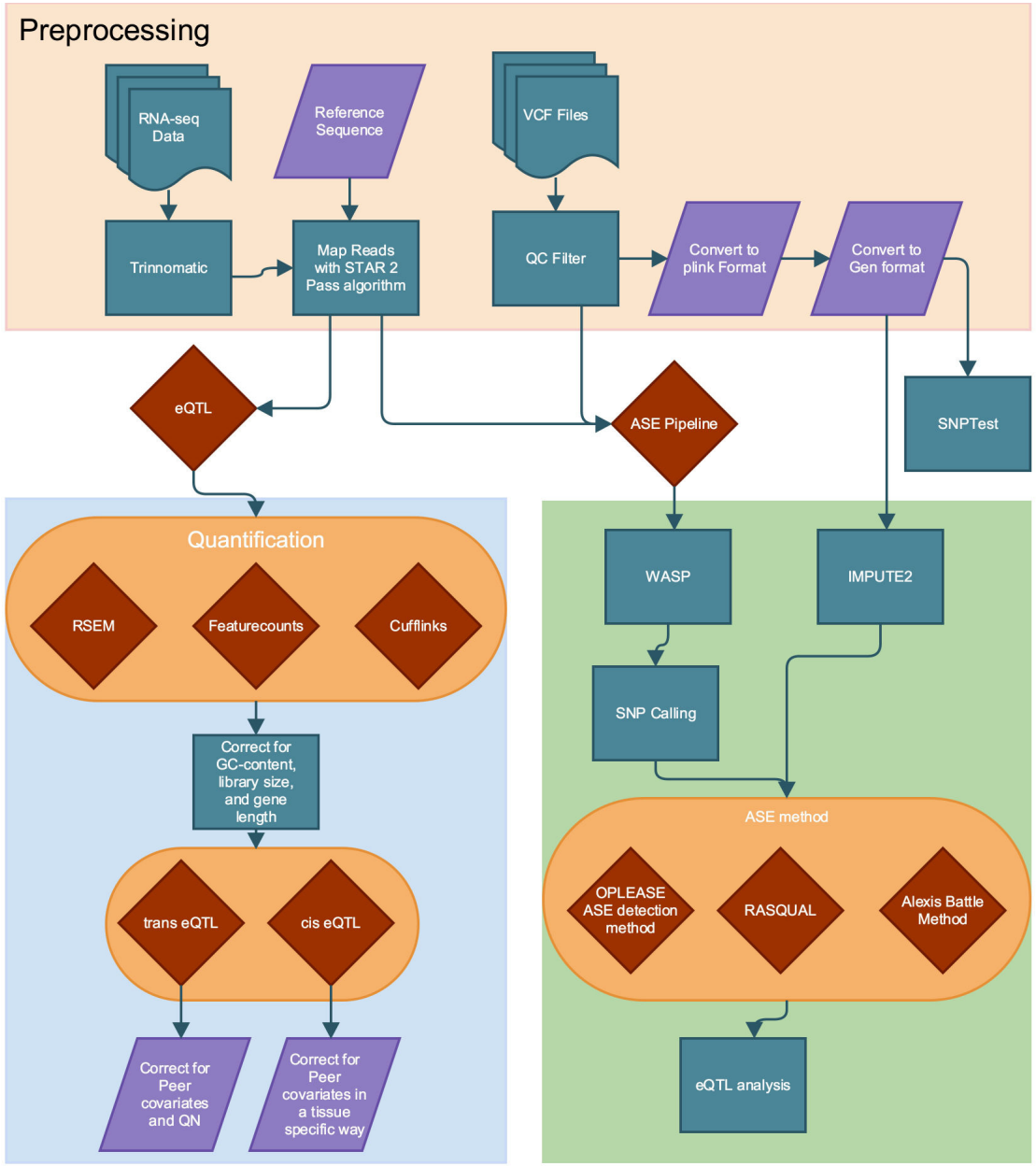
- Complex diseases are often caused by the dysfunction of multiple tissues or cell types (pancreatic islets, adipose, and skeletal muscle for type 2 diabetes)

- **Hypothesis**: Genetic variation affects complex traits and human disease in a tissue specific manner and understanding the role of regulatory variants, and the tissues in which they act, is essential for the functional interpretation of GWAS loci and insights into disease etiology

- The role of the BEEHIVE lab is to develop methods to identify tissue specific eQTL variants and allelic specific expression variants

Galaxy and GTEx

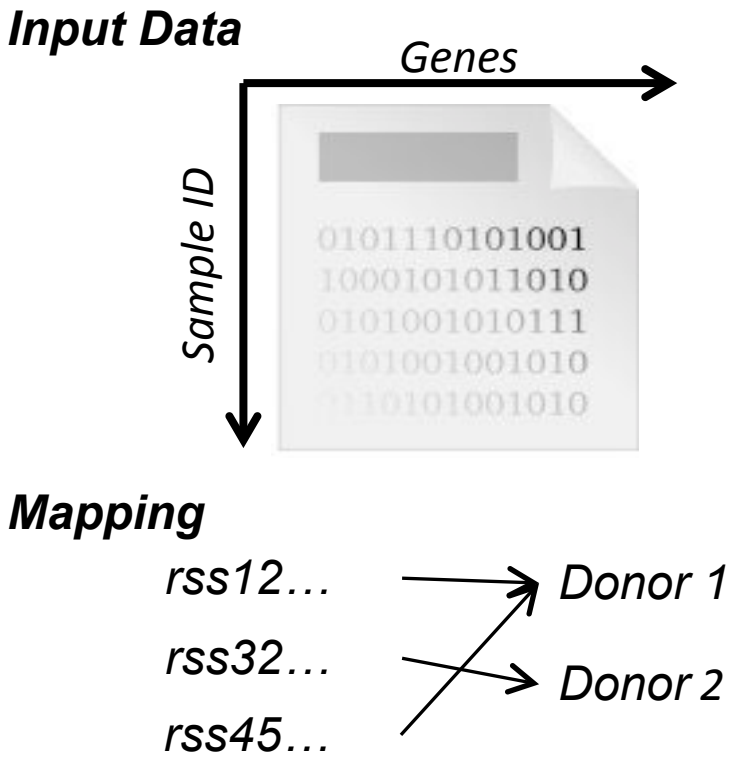
With the ability **to create custom tools** and **share and modify workflows**, Galaxy provides a robust framework to develop our GTEx analysis pipeline for use across our lab.



Tissue Specific Analysis in Galaxy

Data Preprocessing:

- In GTEx Pilot data each donor has a unique ID and each sample from a donor has a unique sample ID
- We to mapped the sample ID to the individual for each tissue using the **Join two Datasets** tool
- We choose 8 tissues that contained approximately 10 samples



Tissue Specific Analysis in Galaxy

Data Preprocessing:

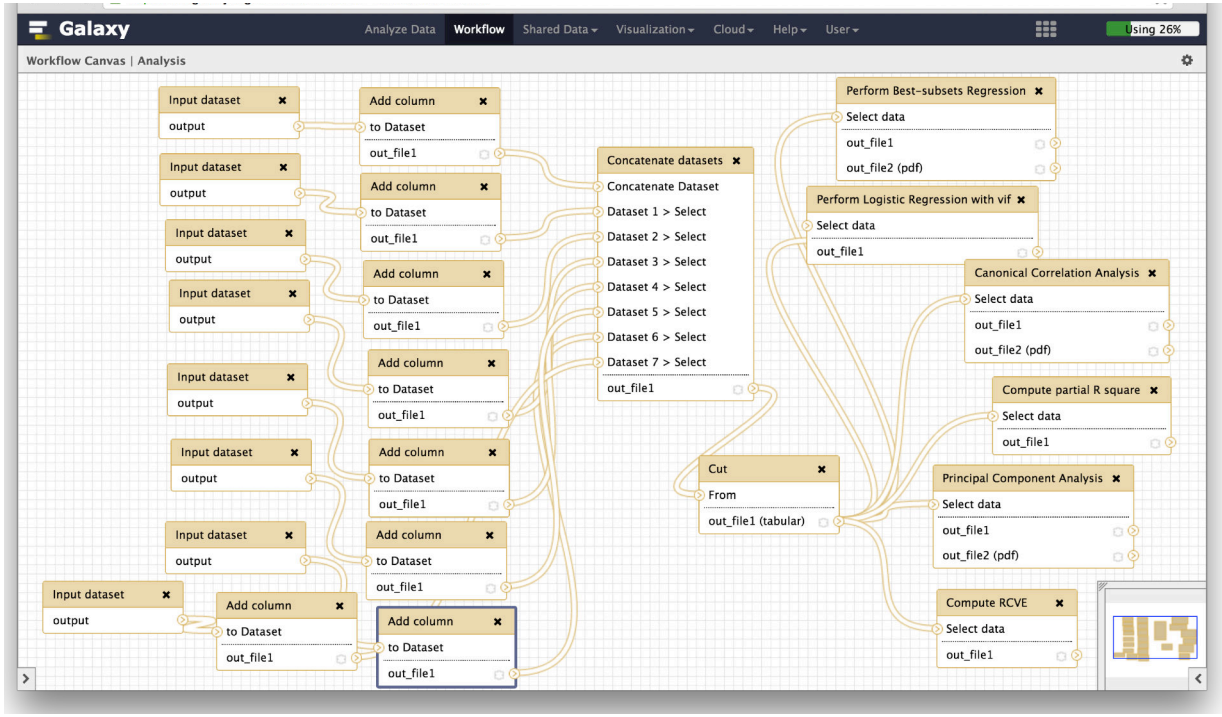
- In GTEx Pilot data each donor has a unique ID and each sample from a donor has a unique sample ID
- We choose 8 tissues that contained approximately 100 samples
- We to mapped the sample ID to the individual for each tissue using the **Join two Datasets** tool

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

Tissue Specific Analysis in Galaxy

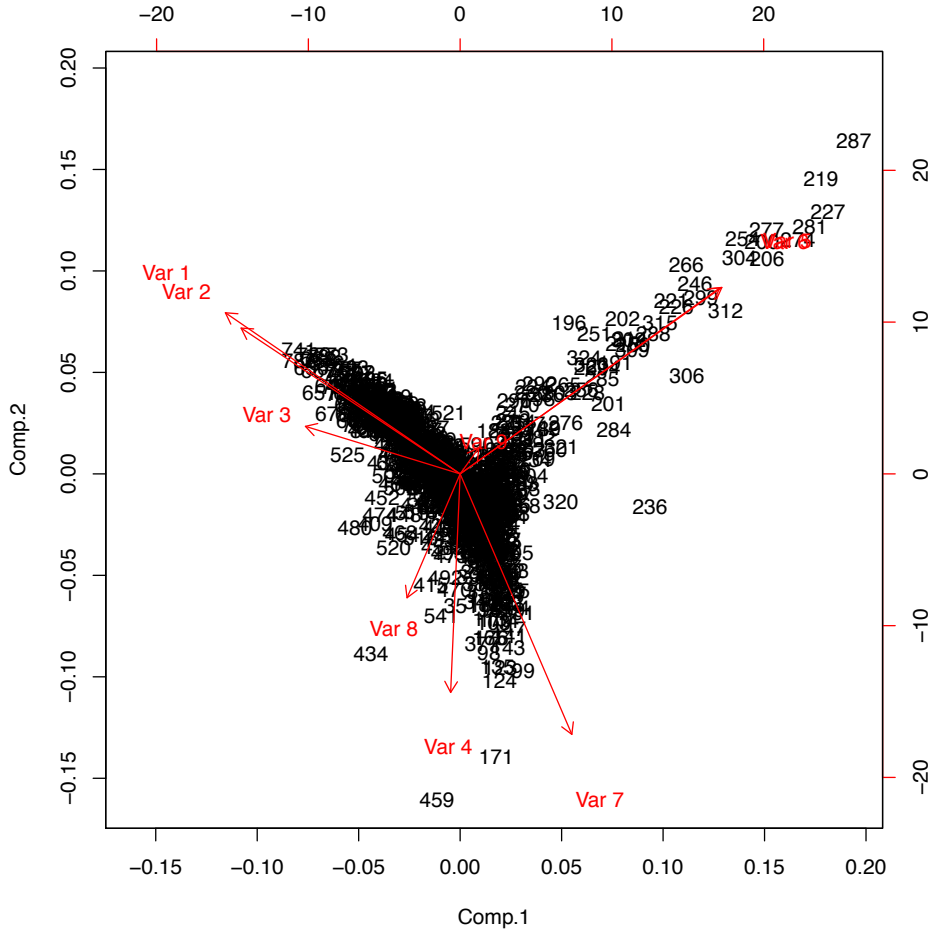
Tissue Specific Analysis:

- Applied tools to analysis the relationship between tissues and gene expression at several known eQTLs documented on the GTEx portal.
- 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 Specific Analysis in Galaxy

Sample output from the **Principle Component** tool



Future of Galaxy in the BEEHIVE Lab

- Incorporate our complete data **processing** and **analysis** pipeline into a private Galaxy instance
- The public Galaxy instance has limited analysis tools availabl and custom tools can only be used on a private Galaxy instance
- Although there are many resources available, developing Galaxy tools and creating specialized analysis is nontrivial
- I have developed a blog accessible to the BEEHIVE group members to assist them in performing tissue specific analysis in Galaxy

Acknowledgements

Thank you to:

- Thee Broad Institute and GTEx consortium
- The Galaxy community for my scholarship to attend GSS2015
- an McDowell for creating our data processing pipeline
- My lab mates in The Princeton BEEHIVE Group