NIH
National Institutes of Health

USC Stevens Neuroimaging and Informatics Institute

Ci Computation Institute

Institute for Systems Biology

BDDS BIG DATA *for* DISCOVERY SCIENCE

SCHOOL OF NURSING
STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

USC Viterbi
School of Engineering
Information Sciences Institute

Big Data for Discovery Science (BDDS)

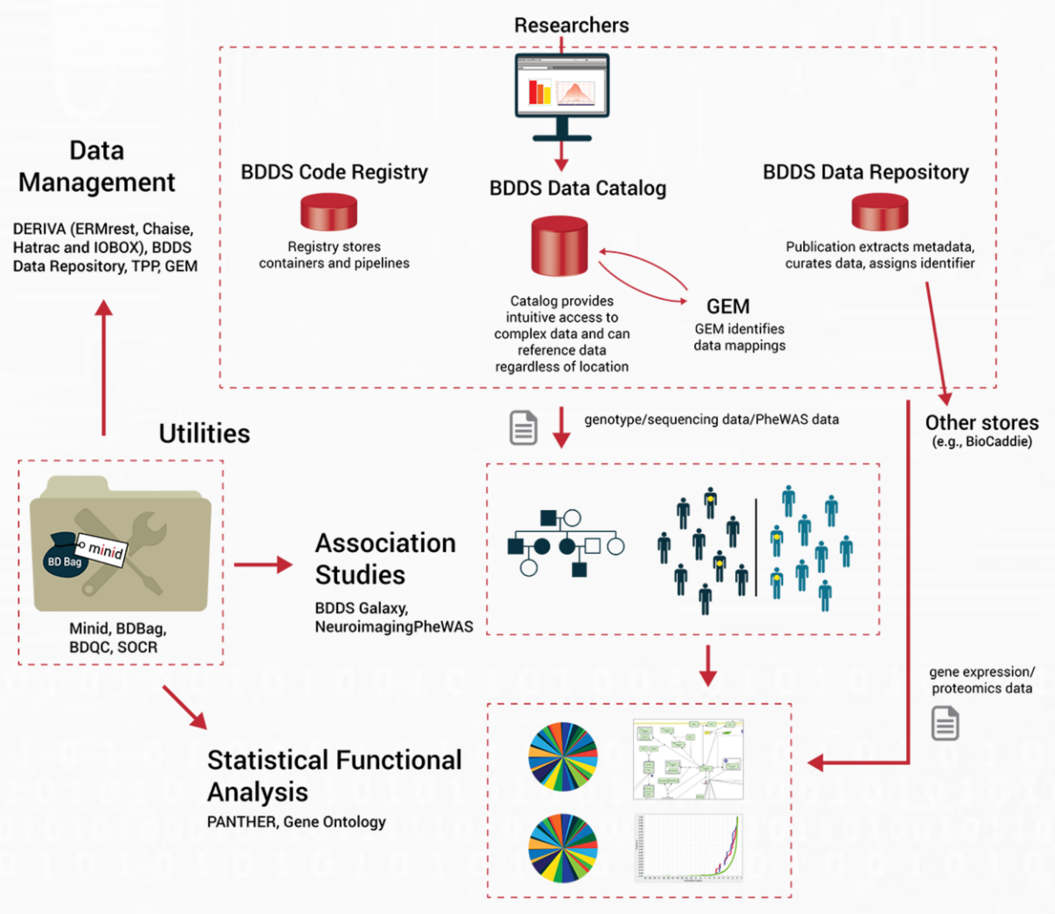
“What can we do today that we couldn’t do before.”

Arthur Toga, PI

Sept. 30, 2016

Santa Rosa, CA

BDDS Platform – Integrated Tools for Discovery



Researchers

Data Management

DERIVA (ERMrest, Chaise, Hatrac and IOBOX), BDDS Data Repository, TPP, GEM

BDDS Code Registry

Registry stores containers and pipelines

BDDS Data Catalog

Catalog provides intuitive access to complex data and can reference data regardless of location

GEM

GEM identifies data mappings

BDDS Data Repository

Publication extracts metadata, curates data, assigns identifier

genotype/sequencing data/PheWAS data

Other stores (e.g., BioCaddie)

Utilities

Minid, BDBag, BDQC, SOCR

Association Studies

BDDS Galaxy, NeuroimagingPheWAS

gene expression/ proteomics data

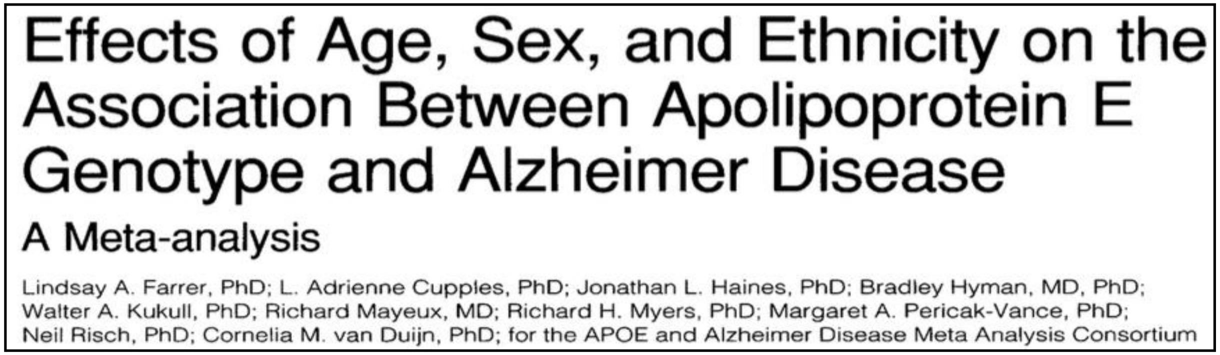
Statistical Functional Analysis

PANTHER, Gene Ontology

BDDS Case Studies

- Data Aggregation
  - Sex/AD
  - Statins
- TReNA
  - Ben Heavner, Ravi Madduri
- PheWAS
  - Carl Kesselman, Lu Zhao
- BDDS Demos this afternoon
  - Dry Creek Valley I

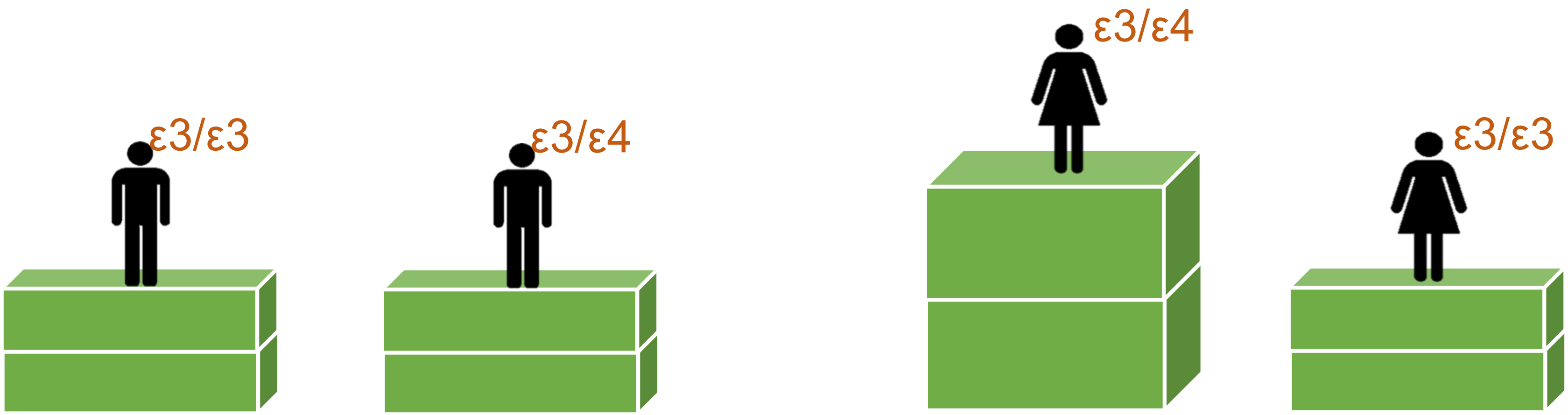
Farrer 1997 Meta-Analysis



Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease

A Meta-analysis

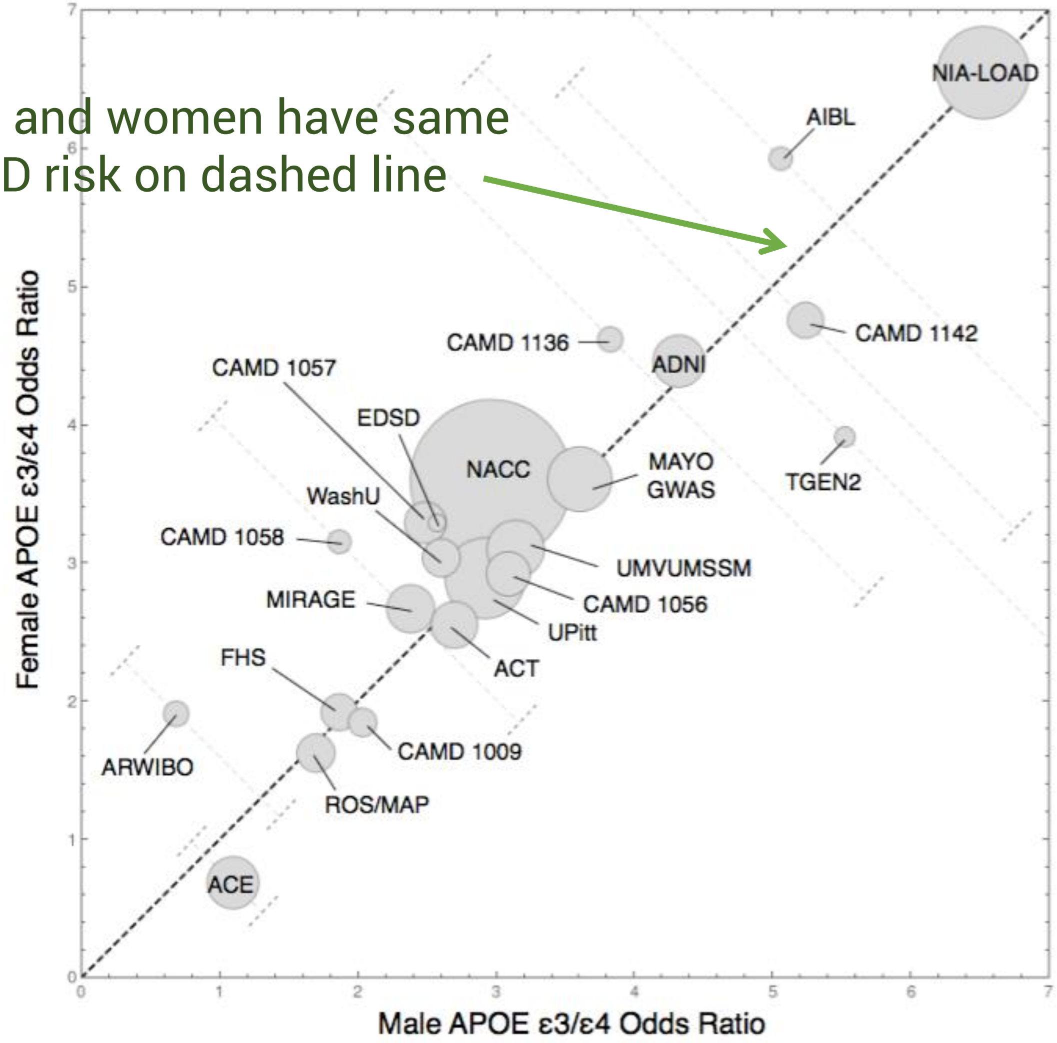
Lindsay A. Farrer, PhD; L. Adrienne Cupples, PhD; Jonathan L. Haines, PhD; Bradley Hyman, MD, PhD; Walter A. Kukull, PhD; Richard Mayeux, MD; Richard H. Myers, PhD; Margaret A. Pericak-Vance, PhD; Neil Risch, PhD; Cornelia M. van Duijn, PhD; for the APOE and Alzheimer Disease Meta Analysis Consortium



OR ≈ 1
(men)

OR ≈ 1.5
(women)

Data set comparison



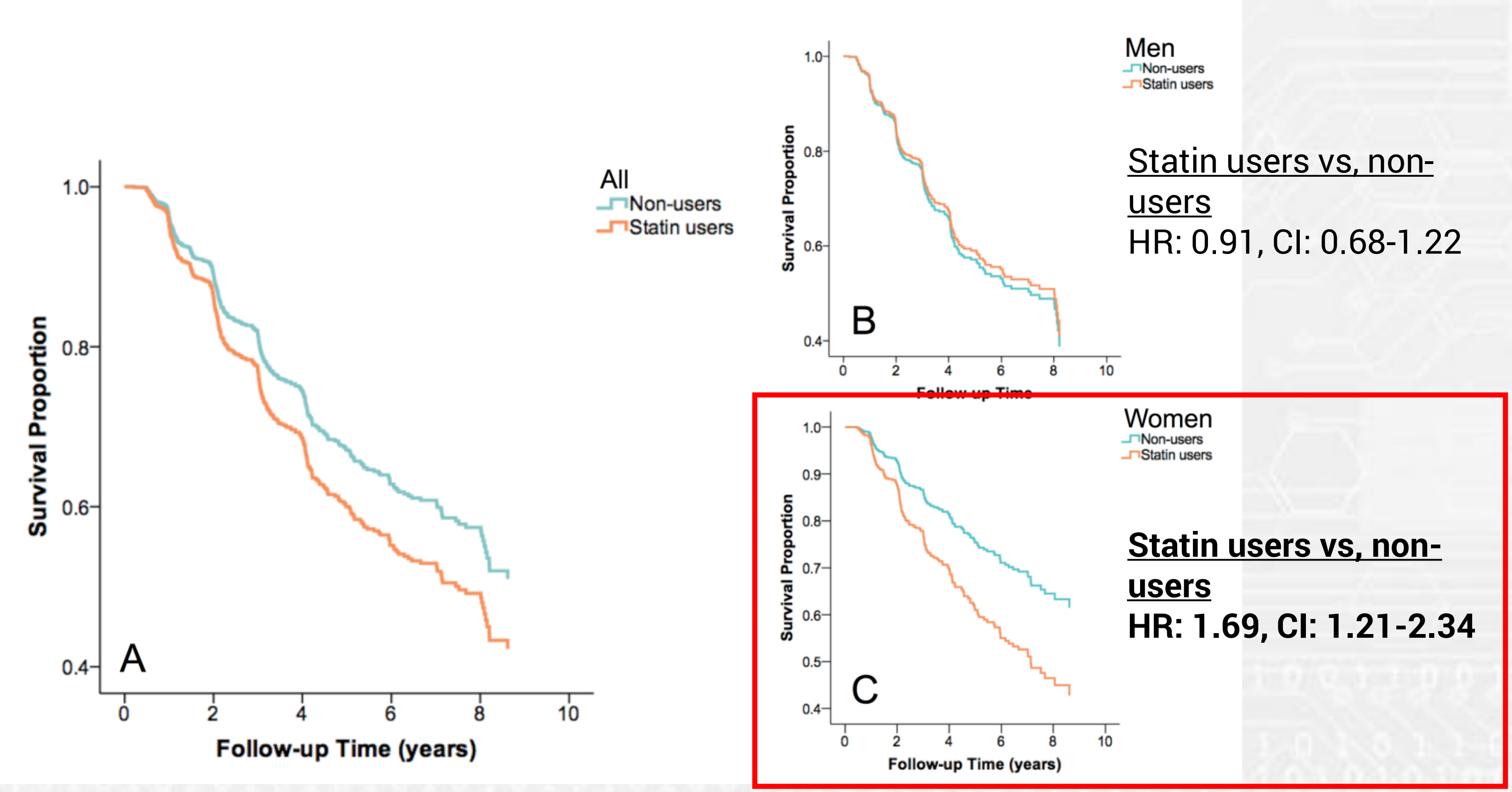
Men and women have same AD risk on dashed line

Background and Objective

- Statins are widely prescribed to treat high cholesterol in men and women
- However, lipophilic statins enter the brain and may impact brain cholesterol that plays a critical role in brain functioning, like estrogen production

- **Objective: to understand the relationship between lipophilic statin use and gender on the brain, cognition, and Alzheimer's disease**

Results – clinical conversion



Statin users vs, non-users
HR: 0.91, CI: 0.68-1.22

Statin users vs, non-users
HR: 1.69, CI: 1.21-2.34

DNase Footprinting Analysis using BDDS tools and services

National Institutes of Health

USC Stevens Neuroimaging and Informatics Institute

Computation Institute

Institute for Systems Biology

BDDS

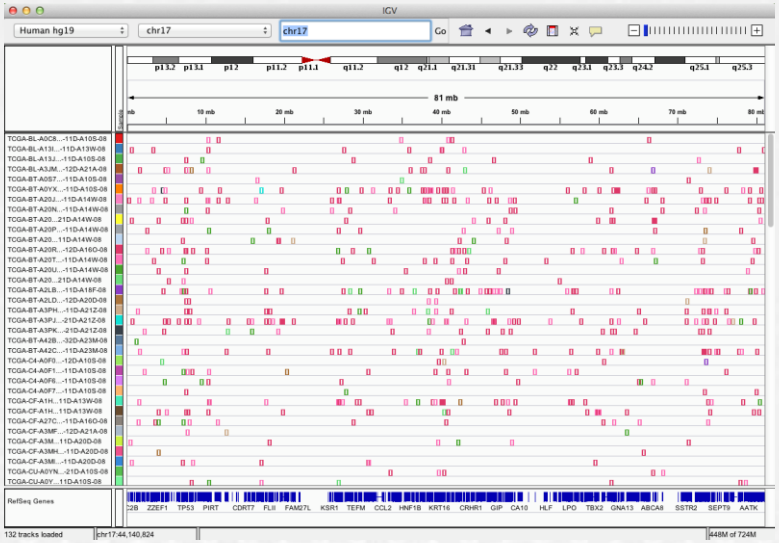
BIG DATA *for* DISCOVERY SCIENCE

SCHOOL OF NURSING STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

USC Viterbi
School of Engineering
Information Sciences Institute

Biological Motivation: Two Puzzling Observations

Genomic Analysis

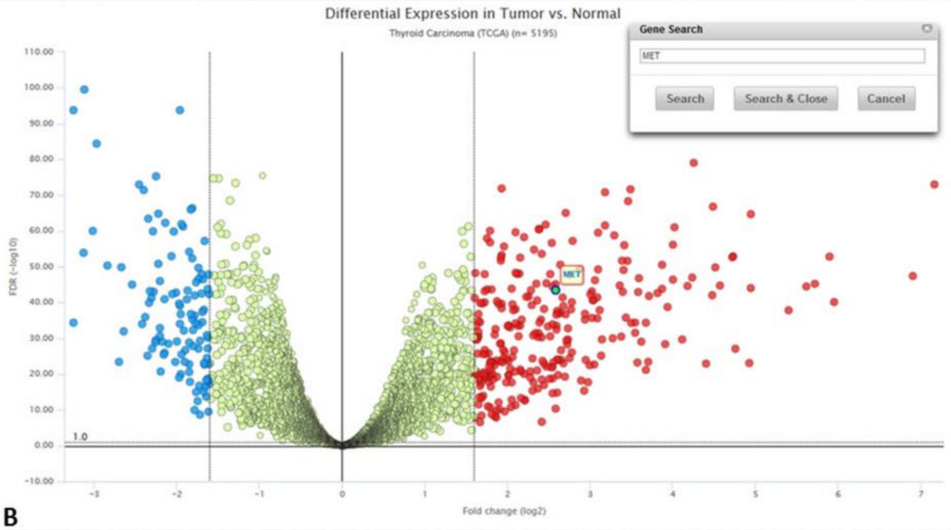


Human hg19 | chr17 | chr17

132 tracks loaded | chr17:44,140,824 | 448M of 724M

IGV
Go
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.33
q22
q23.1
q23.3
q24.2
q25.1
q25.3
81 mb
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
80 mb
Sample
TCGA-BL-A0C8...-11D-A10S-08
TCGA-BL-A13I...-11D-A13W-08
TCGA-BL-A13J...-11D-A10S-08
TCGA-BL-A3JM...-12D-A21A-08
TCGA-BT-A0S7...-11D-A10S-08
TCGA-BT-A0YX...-11D-A10S-08
TCGA-BT-A20J...-11D-A14W-08
TCGA-BT-A20N...-11D-A14W-08
TCGA-BT-A20...21D-A14W-08
TCGA-BT-A20P...-11D-A14W-08
TCGA-BT-A20...11D-A14W-08
TCGA-BT-A20R...-12D-A16O-08
TCGA-BT-A20T...-11D-A14W-08
TCGA-BT-A20U...-11D-A14W-08
TCGA-BT-A20...21D-A14W-08
TCGA-BT-A2LB...-11D-A18F-08
TCGA-BT-A2LD...-12D-A20D-08
TCGA-BT-A3PH...-11D-A21Z-08
TCGA-BT-A3PJ...-21D-A21Z-08
TCGA-BT-A3PK...-21D-A21Z-08
TCGA-BT-A42B...-32D-A23M-08
TCGA-BT-A42C...-11D-A23M-08
TCGA-C4-A0F0...-12D-A10S-08
TCGA-C4-A0F1...-11D-A10S-08
TCGA-C4-A0F6...-11D-A10S-08
TCGA-C4-A0F7...-11D-A10S-08
TCGA-CF-A1H...11D-A13W-08
TCGA-CF-A1H...11D-A13W-08
TCGA-CF-A27C...-11D-A16O-08
TCGA-CF-A3MF...-12D-A21A-08
TCGA-CF-A3M...11D-A20D-08
TCGA-CF-A3MH...-11D-A20D-08
TCGA-CF-A3MI...-11D-A20D-08
TCGA-CU-A0YN...-21D-A10S-08
TCGA-CU-A0Y...11D-A10S-08
RefSeq Genes
C2B
ZZEF1
TP53
PIRT
CDRT7
FLII
FAM27L
KSR1
TEFM
CCL2
HNF1B
KRT16
CRHR1
GIP
CA10
HLF
LPO
TBX2
GNA13
ABCA8
SSTR2
SEPT9
AATK

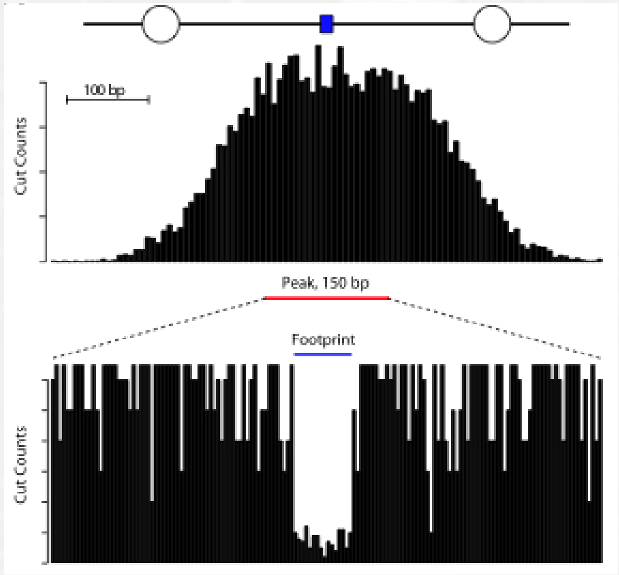
Gene Expression Analysis



Gene Search: MET

Image credits:
http://software.broadinstitute.org/software/igv/MutationData
Nature Methods 13, 9–10 (2016) doi:10.1038/nmeth.3692

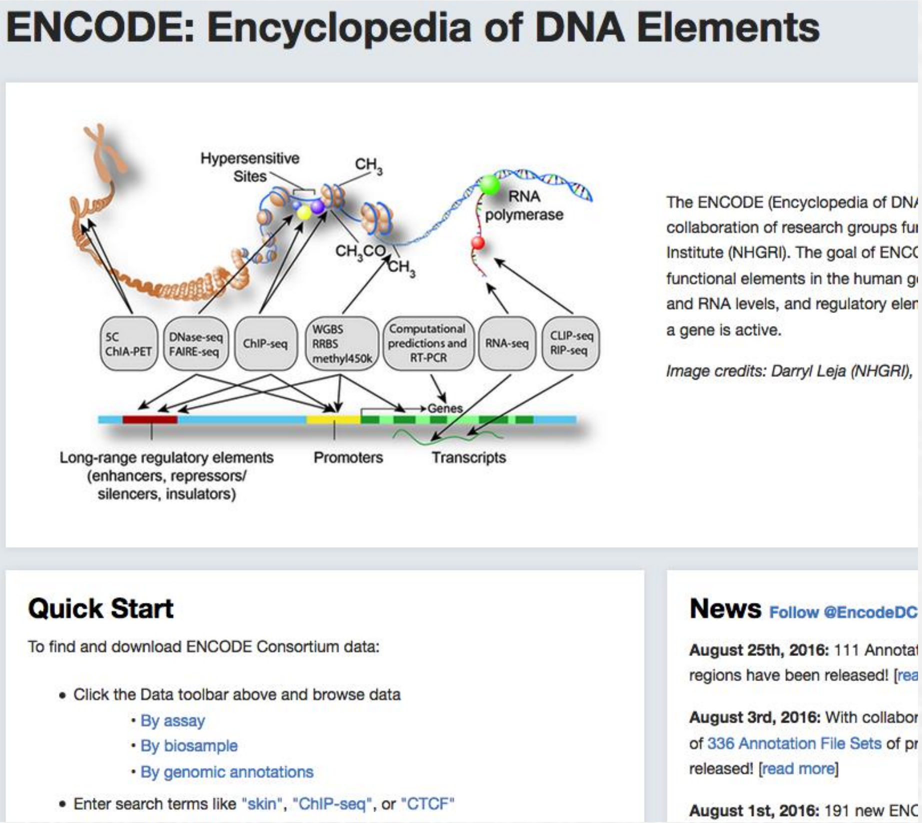
Analysis Plan: TRENA



Vernot et al.
doi: 10.1101/gr.134890.111

- Uniform processing of next generation sequencing data
  - Align to reference genome
  - Identify DNase hypersensitve regions
  - Apply multiple footprinting algorithms to locate putative transcription factor binding sites (TFBSs)
- Evaluate confidence in putative TFBSs
- Use TFBSs as features for machine learning approaches applied to disease-specific research

Primary Data Source: ENCODE



ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DN/
collaboration of research groups fu
Institute (NHGRI). The goal of ENC(
functional elements in the human g
and RNA levels, and regulatory eler
a gene is active.

Image credits: Darryl Leja (NHGRI),

Quick Start

To find and download ENCODE Consortium data:

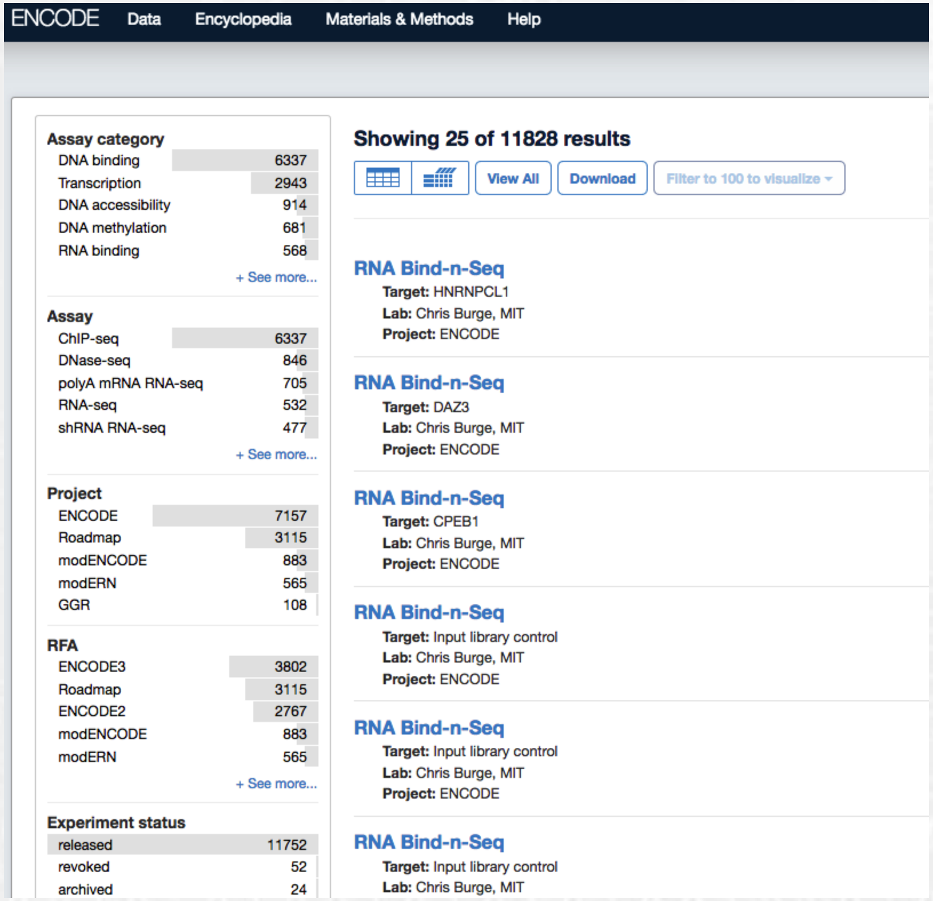
- Click the Data toolbar above and browse data
  - By assay
  - By biosample
  - By genomic annotations
- Enter search terms like "skin", "ChIP-seq", or "CTCF"

News Follow @EncodeDC

August 25th, 2016: 111 Annotat
regions have been released! [rea

August 3rd, 2016: With collabor
of 336 Annotation File Sets of pr
released! [read more]

August 1st, 2016: 191 new ENC



Assay category	
DNA binding	6337
Transcription	2943
DNA accessibility	914
DNA methylation	681
RNA binding	568

+ See more...

Assay	
ChIP-seq	6337
DNase-seq	846
polyA mRNA RNA-seq	705
RNA-seq	532
shRNA RNA-seq	477

+ See more...

Project	
ENCODE	7157
Roadmap	3115
modENCODE	883
modERN	565
GGR	108

RFA	
ENCODE3	3802
Roadmap	3115
ENCODE2	2767
modENCODE	883
modERN	565

+ See more...

Experiment status	
released	11752
revoked	52
archived	24

Showing 25 of 11828 results

RNA Bind-n-Seq
Target: HNRNPCL1
Lab: Chris Burge, MIT
Project: ENCODE

RNA Bind-n-Seq
Target: DAZ3
Lab: Chris Burge, MIT
Project: ENCODE

RNA Bind-n-Seq
Target: CPEB1
Lab: Chris Burge, MIT
Project: ENCODE

RNA Bind-n-Seq
Target: Input library control
Lab: Chris Burge, MIT
Project: ENCODE

RNA Bind-n-Seq
Target: Input library control
Lab: Chris Burge, MIT
Project: ENCODE

RNA Bind-n-Seq
Target: Input library control
Lab: Chris Burge, MIT

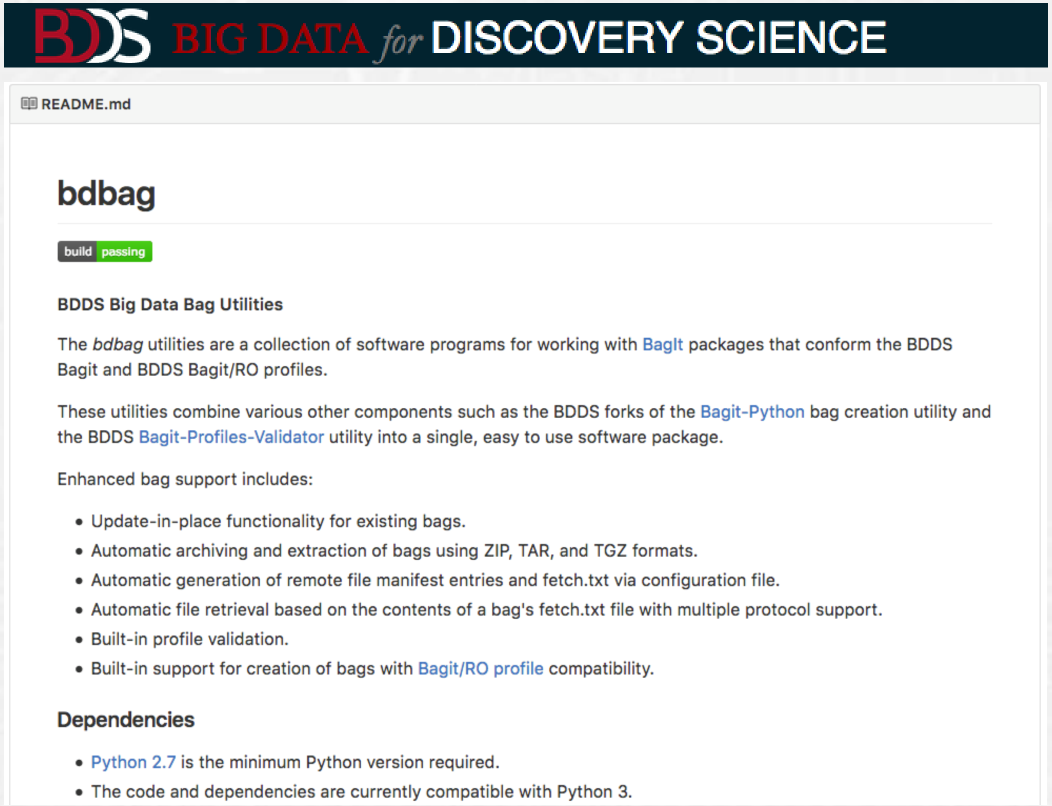
www.encodeproject.org

Big Data Challenges: TRENA

- Gathering raw data from public repository
- Identifying raw data objects
- Transferring data objects
- Scalable data analysis
  - Integrating data from disparate sources
- Providing results for downstream analysis

BDDS Solutions:

Enabling TReNA – BDBag



BDDS BIG DATA *for* DISCOVERY SCIENCE

README.md

bdbag

build passing

BDDS Big Data Bag Utilities

The *bdbag* utilities are a collection of software programs for working with BagIt packages that conform the BDDS Bagit and BDDS Bagit/RO profiles.

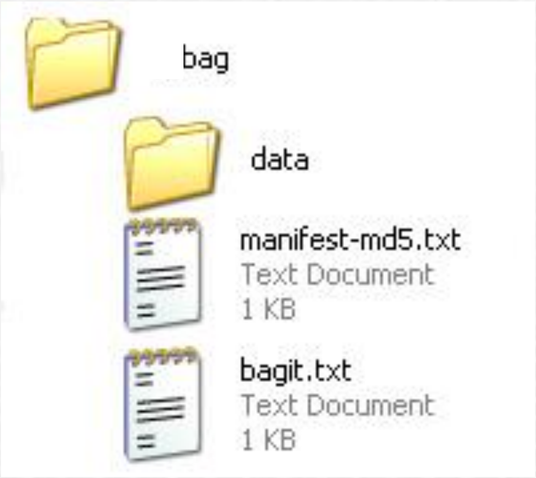
These utilities combine various other components such as the BDDS forks of the Bagit-Python bag creation utility and the BDDS Bagit-Profiles-Validator utility into a single, easy to use software package.

Enhanced bag support includes:

- Update-in-place functionality for existing bags.
- Automatic archiving and extraction of bags using ZIP, TAR, and TGZ formats.
- Automatic generation of remote file manifest entries and fetch.txt via configuration file.
- Automatic file retrieval based on the contents of a bag's fetch.txt file with multiple protocol support.
- Built-in profile validation.
- Built-in support for creation of bags with Bagit/RO profile compatibility.

Dependencies

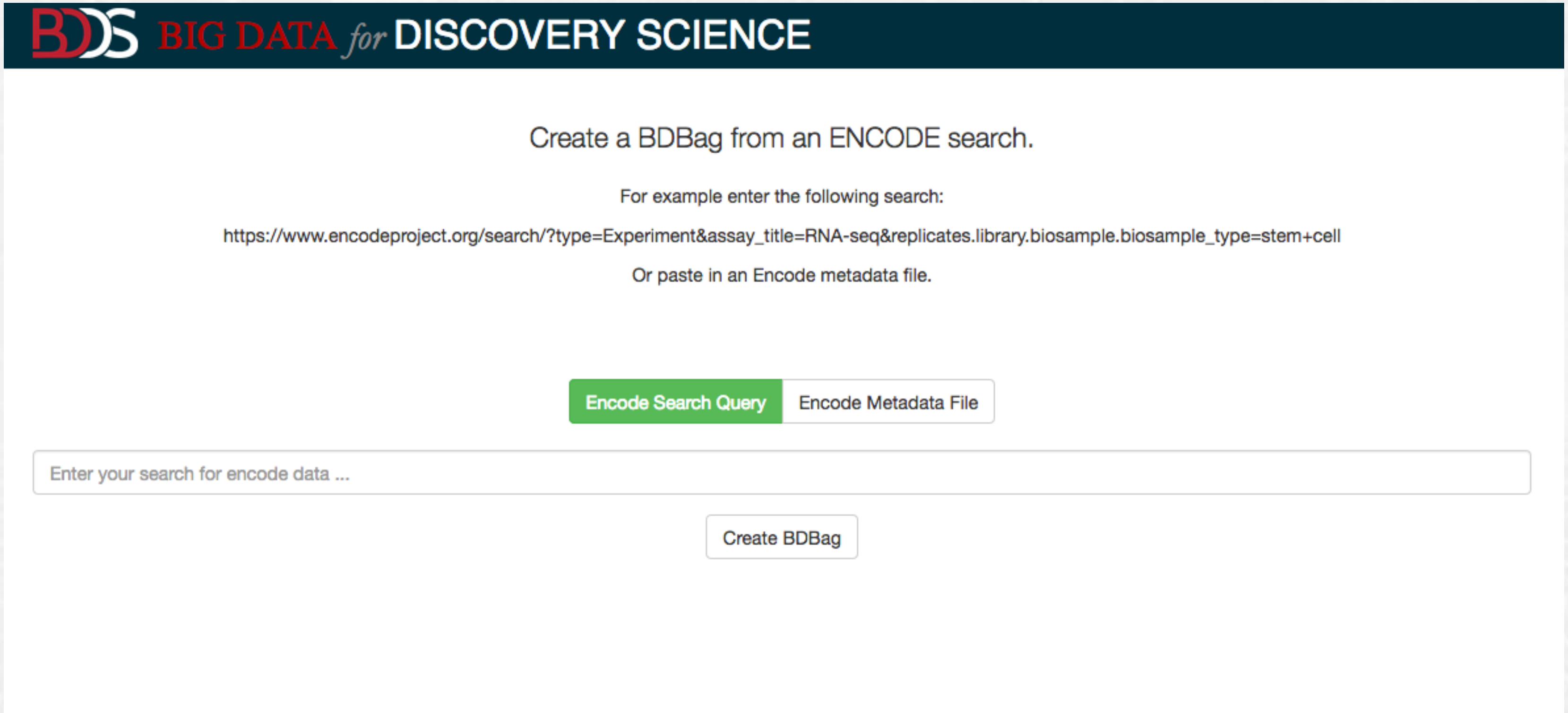
- Python 2.7 is the minimum Python version required.
- The code and dependencies are currently compatible with Python 3.



https://github.com/ini-bdds/bdbag

BDDS Solutions:

Enabling TRENA – ENCODE to BDBag



BIG DATA *for* DISCOVERY SCIENCE

Create a BDBag from an ENCODE search.

For example enter the following search:

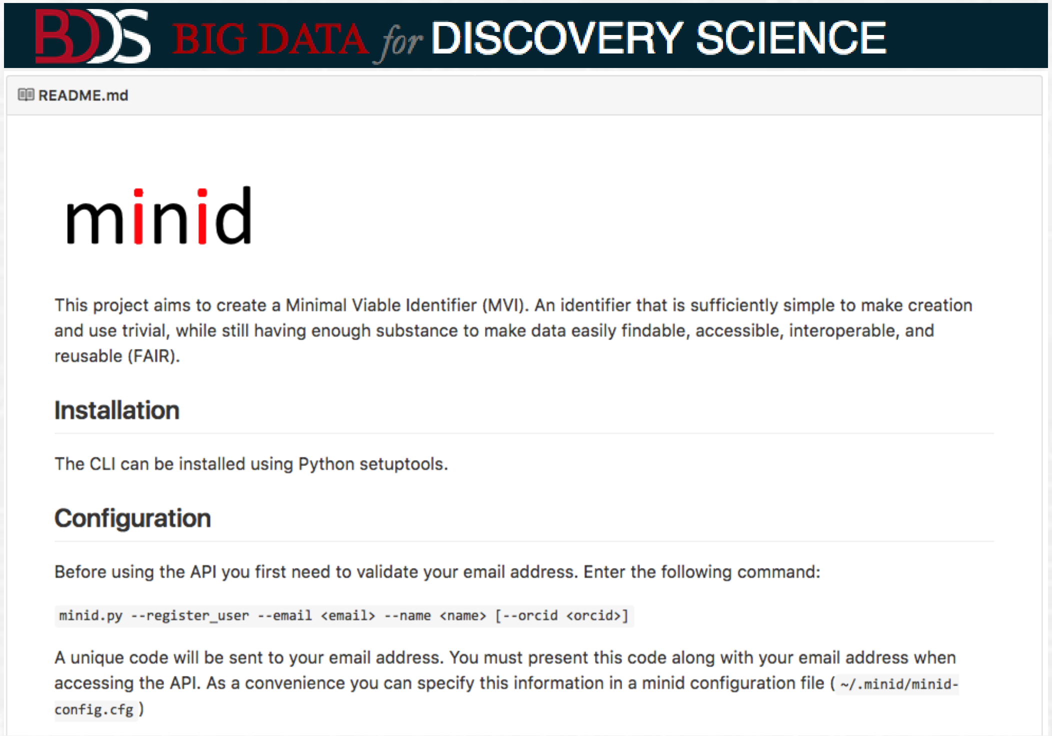
https://www.encodeproject.org/search/?type=Experiment&assay_title=RNA-seq&replicates.library.biosample.biosample_type=stem+cell

Or paste in an Encode metadata file.

http://encode.bdbag.org/

BDDS Solutions:

Enabling TReNA – minids



BIG DATA *for* DISCOVERY SCIENCE

README.md

minid

This project aims to create a Minimal Viable Identifier (MVI). An identifier that is sufficiently simple to make creation and use trivial, while still having enough substance to make data easily findable, accessible, interoperable, and reusable (FAIR).

Installation

The CLI can be installed using Python setuptools.

Configuration

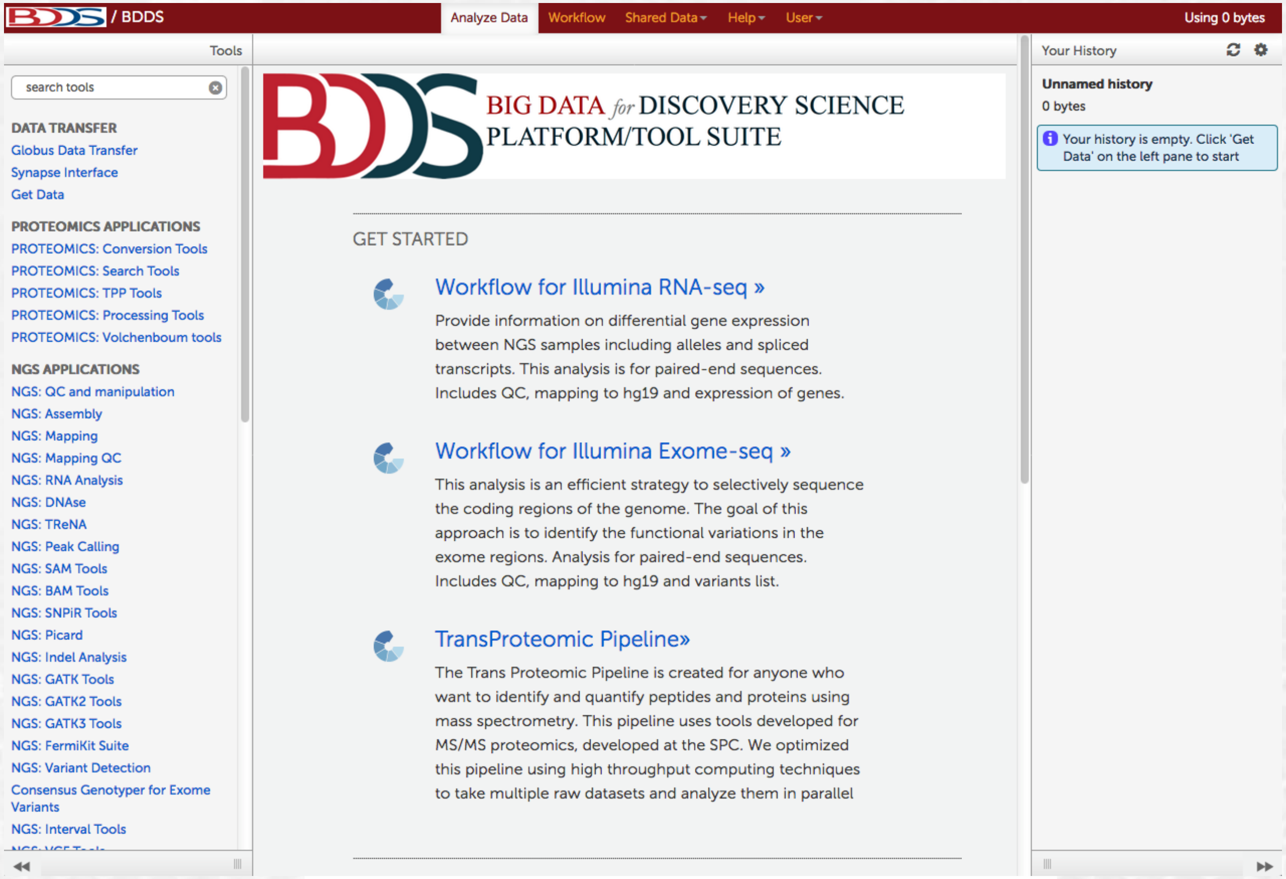
Before using the API you first need to validate your email address. Enter the following command:

```
minid.py --register_user --email <email> --name <name> [--orcid <orcid>]
```

A unique code will be sent to your email address. You must present this code along with your email address when accessing the API. As a convenience you can specify this information in a minid configuration file (`~/.minid/minid-config.cfg`)

https://github.com/ini-bdds/minid

BDDS Solutions:
Enabling TReNA – BDDS Galaxy



DATA TRANSFER
Globus Data Transfer
Synapse Interface
Get Data

PROTEOMICS APPLICATIONS
PROTEOMICS: Conversion Tools
PROTEOMICS: Search Tools
PROTEOMICS: TPP Tools
PROTEOMICS: Processing Tools
PROTEOMICS: Volchenboum tools

NGS APPLICATIONS
NGS: QC and manipulation
NGS: Assembly
NGS: Mapping
NGS: Mapping QC
NGS: RNA Analysis
NGS: DNAse
NGS: TReNA
NGS: Peak Calling
NGS: SAM Tools
NGS: BAM Tools
NGS: SNPiR Tools
NGS: Picard
NGS: Indel Analysis
NGS: GATK Tools
NGS: GATK2 Tools
NGS: GATK3 Tools
NGS: FermiKit Suite
NGS: Variant Detection
Consensus Genotyper for Exome Variants
NGS: Interval Tools
NGS: VCF Tools

BDDS BIG DATA *for* DISCOVERY SCIENCE
PLATFORM/TOOL SUITE

GET STARTED

Workflow for Illumina RNA-seq »

Provide information on differential gene expression between NGS samples including alleles and spliced transcripts. This analysis is for paired-end sequences. Includes QC, mapping to hg19 and expression of genes.

Workflow for Illumina Exome-seq »

This analysis is an efficient strategy to selectively sequence the coding regions of the genome. The goal of this approach is to identify the functional variations in the exome regions. Analysis for paired-end sequences. Includes QC, mapping to hg19 and variants list.

TransProteomic Pipeline»

The Trans Proteomic Pipeline is created for anyone who want to identify and quantify peptides and proteins using mass spectrometry. This pipeline uses tools developed for MS/MS proteomics, developed at the SPC. We optimized this pipeline using high throughput computing techniques to take multiple raw datasets and analyze them in parallel

Unnamed history
0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

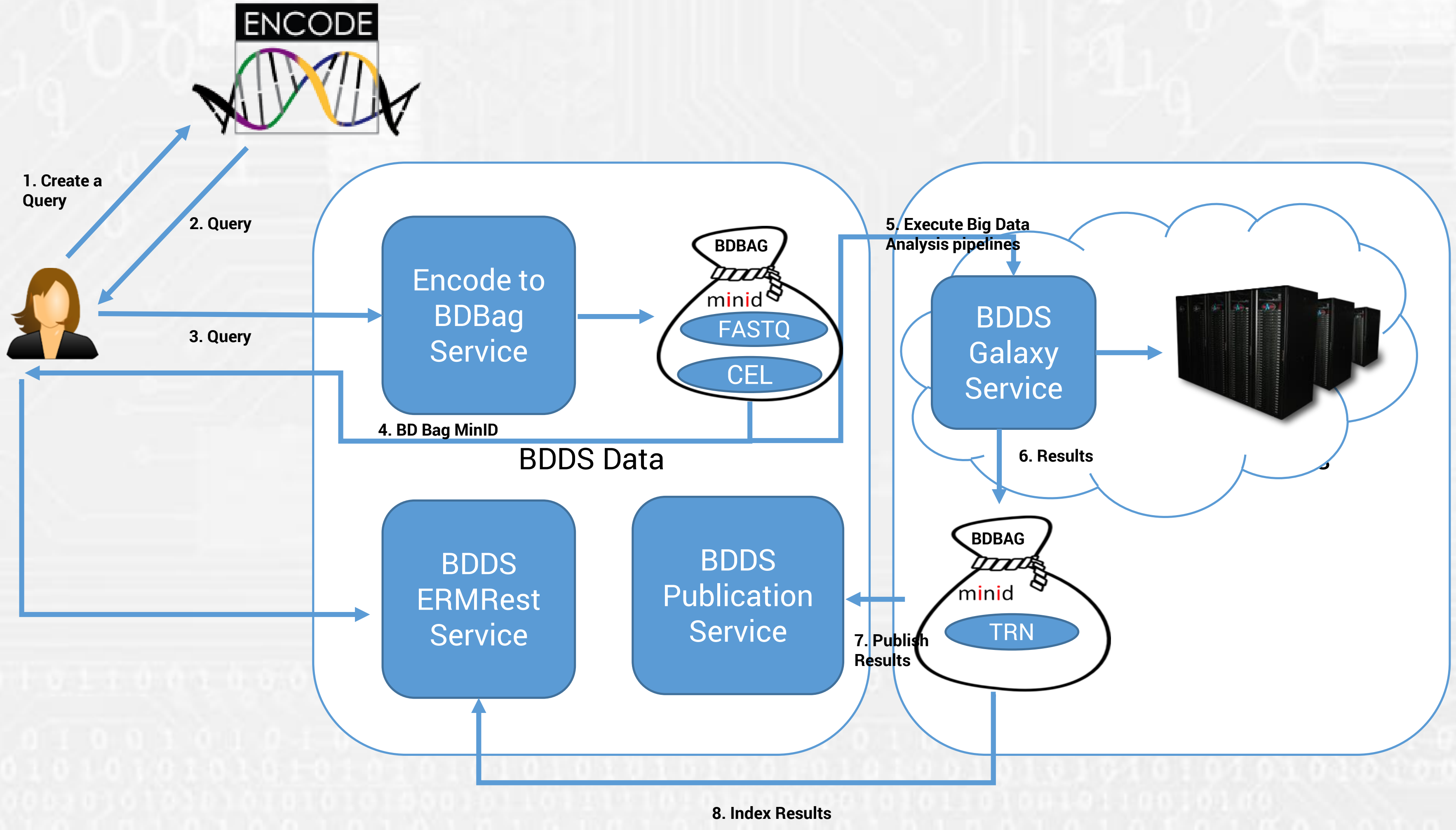
https://bdds.globusgenomics.org/

BDDS Solutions:
Enabling TReNA

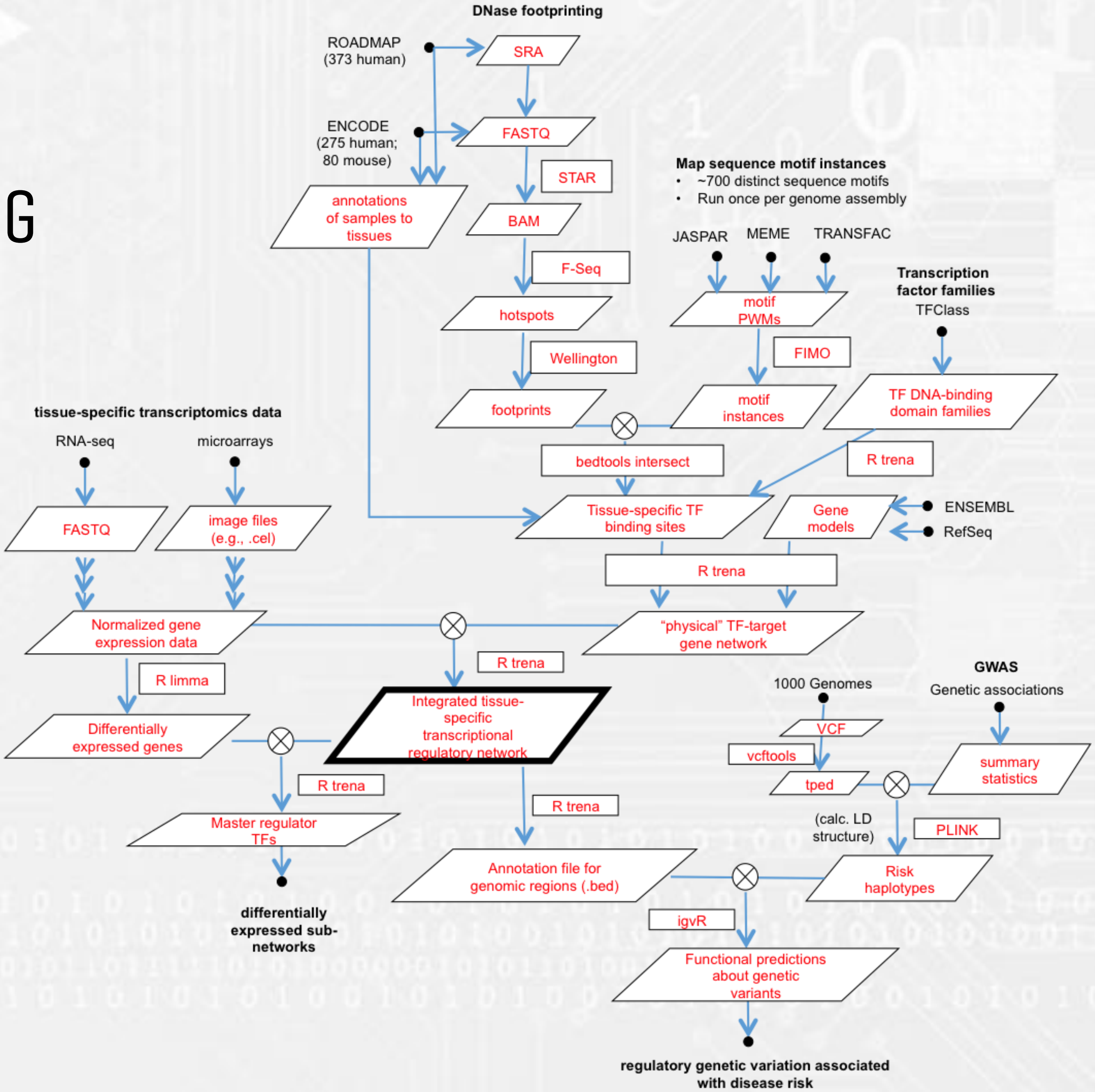
- BDBags and BDBag Tooling
- ENCODE to BDBag Web Service
- minids and minid Tooling
- Cloud based Data Transfers
- BDDS Galaxy Analytics Platform

Implementing the DNase Hypersensitivity Workflow on the BDDS Galaxy Platform

Plan overview



DNase Footprinting Data Processing Workflow



DNase footprinting

ROADMAP (373 human)

SRA

ENCODE (275 human; 80 mouse)

FASTQ

STAR

annotations of samples to tissues

BAM

F-Seq

hotspots

Wellington

footprints

Map sequence motif instances

- ~700 distinct sequence motifs
- Run once per genome assembly

JASPAR

MEME

TRANSFAC

motif PWMs

FIMO

motif instances

Transcription factor families

TFClass

TF DNA-binding domain families

R trena

bedtools intersect

Tissue-specific TF binding sites

Gene models

ENSEMBL

RefSeq

R trena

"physical" TF-target gene network

tissue-specific transcriptomics data

RNA-seq

microarrays

FASTQ

image files (e.g., .cel)

Normalized gene expression data

R limma

Differentially expressed genes

R trena

Integrated tissue-specific transcriptional regulatory network

R trena

Master regulator TFs

differentially expressed sub-networks

R trena

Annotation file for genomic regions (.bed)

1000 Genomes

VCF

vcftools

tped

(calc. LD structure)

GWAS

Genetic associations

summary statistics

PLINK

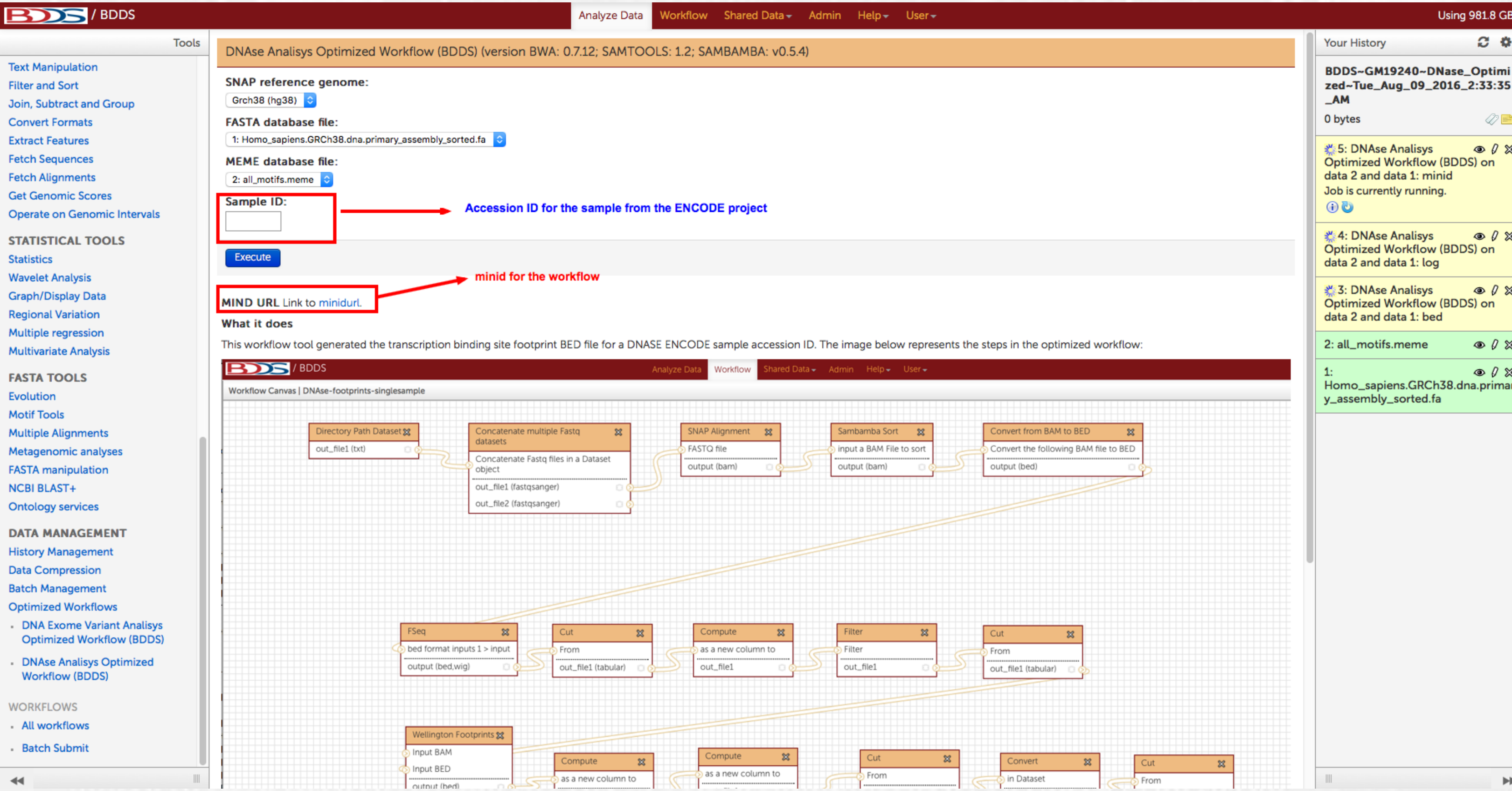
Risk haplotypes

igvR

Functional predictions about genetic variants

regulatory genetic variation associated with disease risk

BDDS Galaxy Implementation



Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals

STATISTICAL TOOLS
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis

FASTA TOOLS
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NCBI BLAST+
Ontology services

DATA MANAGEMENT
History Management
Data Compression
Batch Management
Optimized Workflows
- DNA Exome Variant Analisys Optimized Workflow (BDDS)
- DNAse Analisys Optimized Workflow (BDDS)

WORKFLOWS
- All workflows
- Batch Submit

DNAse Analisys Optimized Workflow (BDDS) (version BWA: 0.7.12; SAMTOOLS: 1.2; SAMBAMBA: v0.5.4)

SNAP reference genome: Grch38 (hg38)

FASTA database file: 1: Homo_sapiens.GRCh38.dna.primary_assembly_sorted.fa

MEME database file: 2: all_motifs.meme

Sample ID:

Accession ID for the sample from the ENCODE project

Execute

minid for the workflow

MIND URL Link to minidurl.

What it does

This workflow tool generated the transcription binding site footprint BED file for a DNASE ENCODE sample accession ID. The image below represents the steps in the optimized workflow:

BDDS~GM19240~DNase_Optimized~Tue_Aug_09_2016_2:33:35_AM

0 bytes

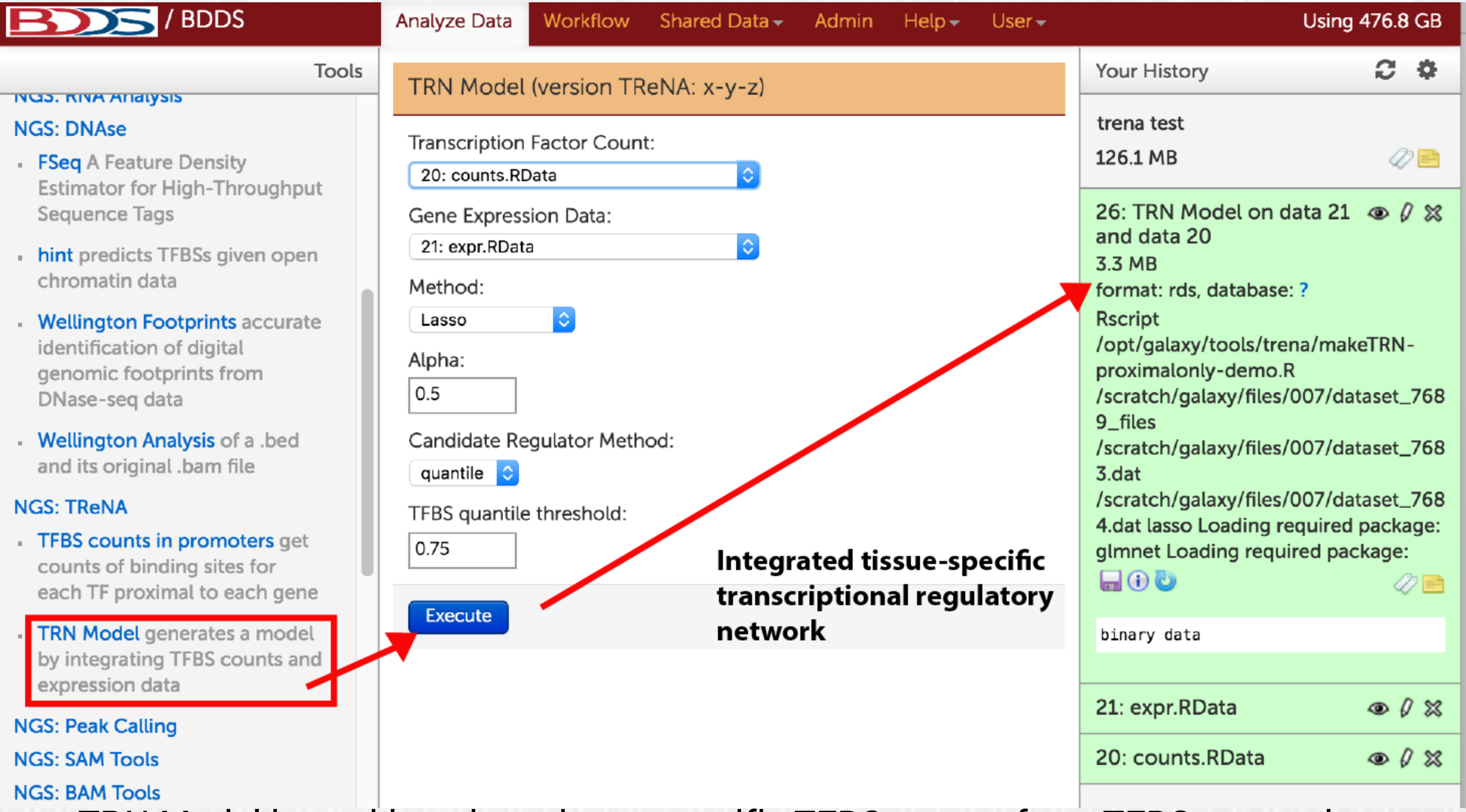
5: DNAse Analisys Optimized Workflow (BDDS) on data 2 and data 1: minid
Job is currently running.

4: DNAse Analisys Optimized Workflow (BDDS) on data 2 and data 1: log

3: DNAse Analisys Optimized Workflow (BDDS) on data 2 and data 1: bed

2: all_motifs.meme

1: Homo_sapiens.GRCh38.dna.primary_assembly_sorted.fa



- FSeq A Feature Density Estimator for High-Throughput Sequence Tags
- hint predicts TFBSs given open chromatin data
- Wellington Footprints accurate identification of digital genomic footprints from DNase-seq data
- Wellington Analysis of a .bed and its original .bam file

- TFBS counts in promoters get counts of binding sites for each TF proximal to each gene
- TRN Model generates a model by integrating TFBS counts and expression data

TRN Model (version TReNA: x-y-z)

Transcription Factor Count: 20: counts.RData

Gene Expression Data: 21: expr.RData

Method: Lasso

Alpha: 0.5

Candidate Regulator Method: quantile

TFBS quantile threshold: 0.75

trena test
126.1 MB

26: TRN Model on data 21 and data 20
3.3 MB
format: rds, database: ?

```
Rscript
/opt/galaxy/tools/trena/makeTRN-proximalonly-demo.R
/scratch/galaxy/files/007/dataset_7689_files
/scratch/galaxy/files/007/dataset_7683.dat
/scratch/galaxy/files/007/dataset_7684.dat lasso Loading required package:
glmnet Loading required package:
```

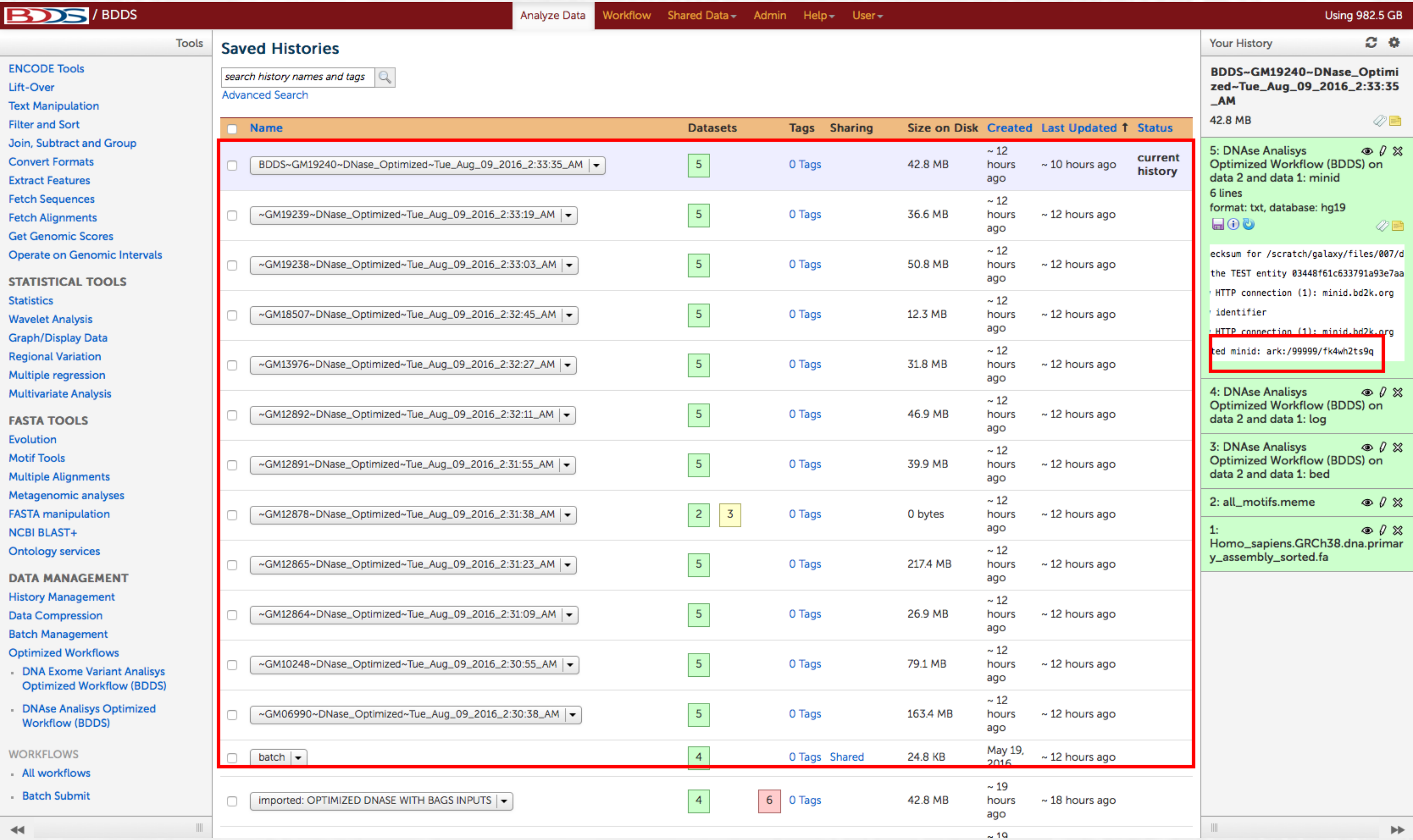
```
binary data
```

21: expr.RData

20: counts.RData

Integrated tissue-specific transcriptional regulatory network

TRN Model is working given tissue-specific TFBS counts from TFBS counts in promoters and corresponding gene expression data.



Saved Histories

☐	Name	Datasets		Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
☐	BDDS~GM19240~DNase_Optimized~Tue_Aug_09_2016_2:33:35_AM	5		0 Tags		42.8 MB	~ 12 hours ago	~ 10 hours ago	current history
☐	~GM19239~DNase_Optimized~Tue_Aug_09_2016_2:33:19_AM	5		0 Tags		36.6 MB	~ 12 hours ago	~ 12 hours ago	
☐	~GM19238~DNase_Optimized~Tue_Aug_09_2016_2:33:03_AM	5		0 Tags		50.8 MB	~ 12 hours ago	~ 12 hours ago	
☐	~GM18507~DNase_Optimized~Tue_Aug_09_2016_2:32:45_AM	5		0 Tags		12.3 MB	~ 12 hours ago	~ 12 hours ago	
☐	~GM13976~DNase_Optimized~Tue_Aug_09_2016_2:32:27_AM	5		0 Tags		31.8 MB	~ 12 hours ago	~ 12 hours ago	
☐	~GM12892~DNase_Optimized~Tue_Aug_09_2016_2:32:11_AM	5		0 Tags		46.9 MB	~ 12 hours ago	~ 12 hours ago	
☐	~GM12891~DNase_Optimized~Tue_Aug_09_2016_2:31:55_AM	5		0 Tags		39.9 MB	~ 12 hours ago	~ 12 hours ago	
☐	~GM12878~DNase_Optimized~Tue_Aug_09_2016_2:31:38_AM	2	3	0 Tags		0 bytes	~ 12 hours ago	~ 12 hours ago	
☐	~GM12865~DNase_Optimized~Tue_Aug_09_2016_2:31:23_AM	5		0 Tags		217.4 MB	~ 12 hours ago	~ 12 hours ago	
☐	~GM12864~DNase_Optimized~Tue_Aug_09_2016_2:31:09_AM	5		0 Tags		26.9 MB	~ 12 hours ago	~ 12 hours ago	
☐	~GM10248~DNase_Optimized~Tue_Aug_09_2016_2:30:55_AM	5		0 Tags		79.1 MB	~ 12 hours ago	~ 12 hours ago	
☐	~GM06990~DNase_Optimized~Tue_Aug_09_2016_2:30:38_AM	5		0 Tags		163.4 MB	~ 12 hours ago	~ 12 hours ago	
☐	batch	4		0 Tags	Shared	24.8 KB	May 19, 2016	~ 12 hours ago	
☐	imported: OPTIMIZED DNASE WITH BAGS INPUTS	4	6	0 Tags		42.8 MB	~ 19 hours ago	~ 18 hours ago	
							~ 19		

Your History

BDDS~GM19240~DNase_Optimized~Tue_Aug_09_2016_2:33:35_AM

42.8 MB

5: DNAse Analisys Optimized Workflow (BDDS) on data 2 and data 1: minid
6 lines
format: txt, database: hg19

```
ecksum for /scratch/galaxy/files/007/d
the TEST entity 03448f61c633791a93e7aa
 HTTP connection (1): minid.bd2k.org
 identifier
 HTTP connection (1): minid.bd2k.org
ted minid: ark:/99999/fk4wh2ts9q
```

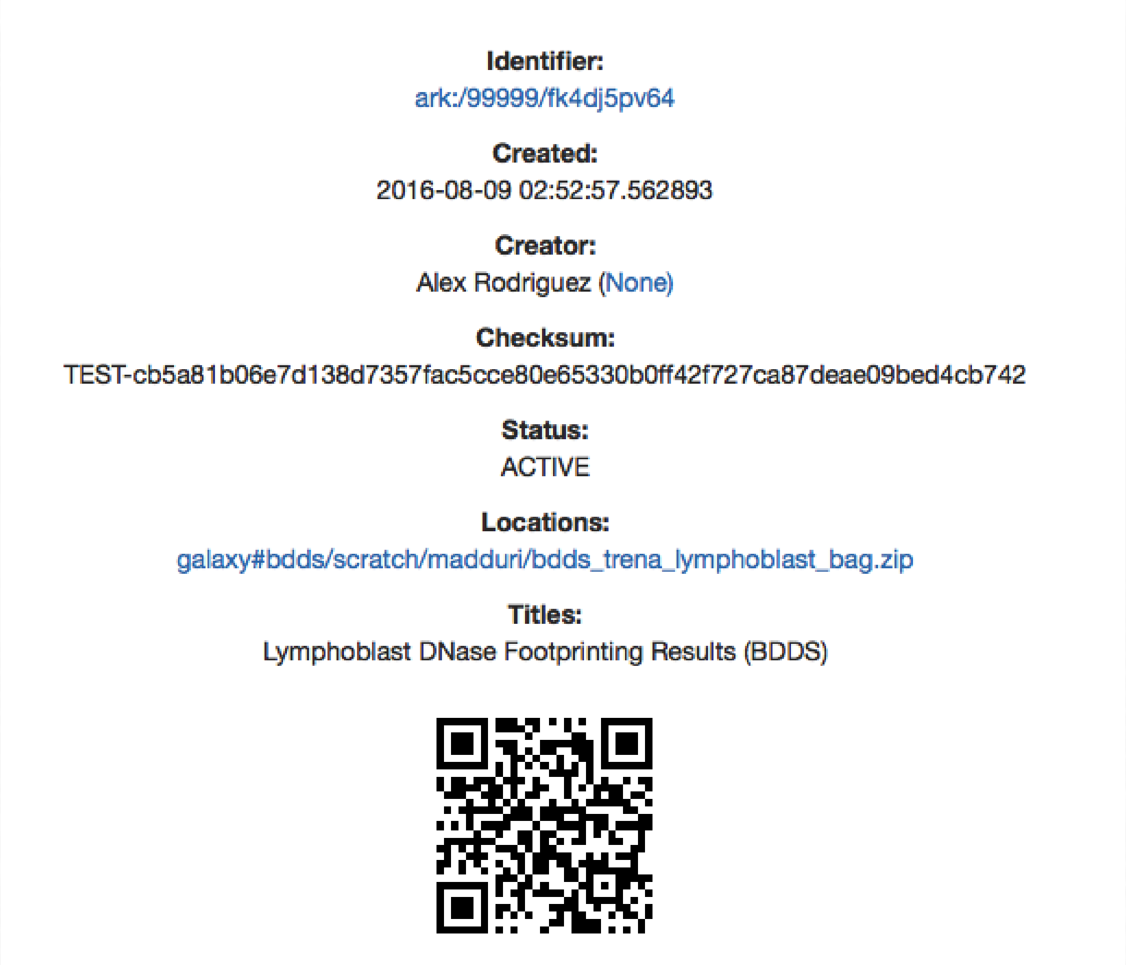
4: DNAse Analisys Optimized Workflow (BDDS) on data 2 and data 1: log

3: DNAse Analisys Optimized Workflow (BDDS) on data 2 and data 1: bed

2: all_motifs.meme

1: Homo_sapiens.GRCh38.dna.primary_assembly_sorted.fa

Publishing Results



Identifier:

ark:/99999/fk4dj5pv64

Created:

2016-08-09 02:52:57.562893

Creator:

Alex Rodriguez (None)

Checksum:

TEST-cb5a81b06e7d138d7357fac5cce80e65330b0ff42f727ca87deae09bed4cb742

Status:

ACTIVE

Locations:

galaxy#bdds/scratch/madduri/bdds_trena_lymphoblast_bag.zip

Titles:

Lymphoblast DNase Footprinting Results (BDDS)

What can we do today that we couldn't do before?

- Generate BDBags containing researcher-defined subsets of ENCODE data
- Uniquely identify ENCODE data sets using minids
- Copy data directly from ENCODE to BDDS Galaxy platform by specifying minid
- Run complex DNase footprinting analysis workflow on Galaxy platform with full provenance
- Uniquely identify workflow instantiation with minid
- Return analysis results in BDBag with minid

BDDS Tools used in the TReNA Approach

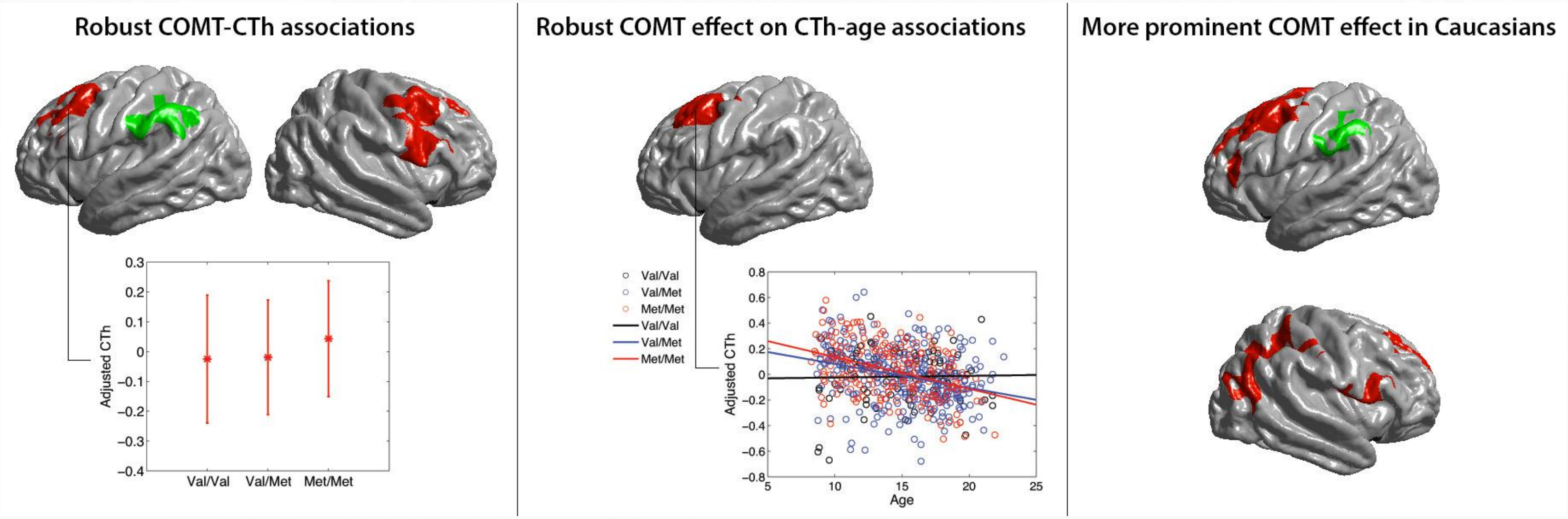
- Minids and Minid Tooling
- BDBags and BDBag Tooling
- ENCODE to BDBag Web Service
- Cloud based Data Transfers between ENCODE and BDDS Galaxy
- BDDS Galaxy Cloud-based Analytics Platform

A Platform for Phenome Wide Association Studies (PheWAS)

Neuroimaging PheWAS

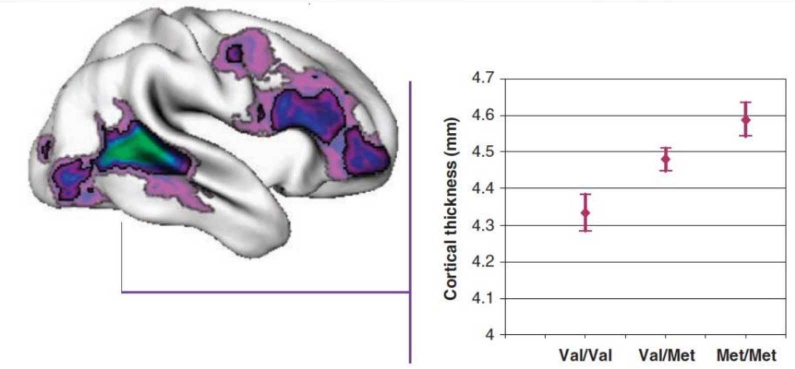
- What is PheWAS?
  - One SNP -> a wide variety of neuroimaging phenotypes (inverse of GWAS)
- Why PheWAS?
  - Unbiasedly validates GWAS/single-phenotype studies findings and explores new system-level genetic associations.
- Challenges
  - Complexity, heterogeneity, and volume of the data
  - Complex and sophisticated brain image processing
  - Multiple-comparison correction
  - Result visualization

- PheWAS findings (Zhao,…, Toga, Nat Neurosci, submitted)



Shaw, Molecular Psychiatry (2009) 14, 348–355

Raznahan, Neuroimage (2011) 57, 1517-23



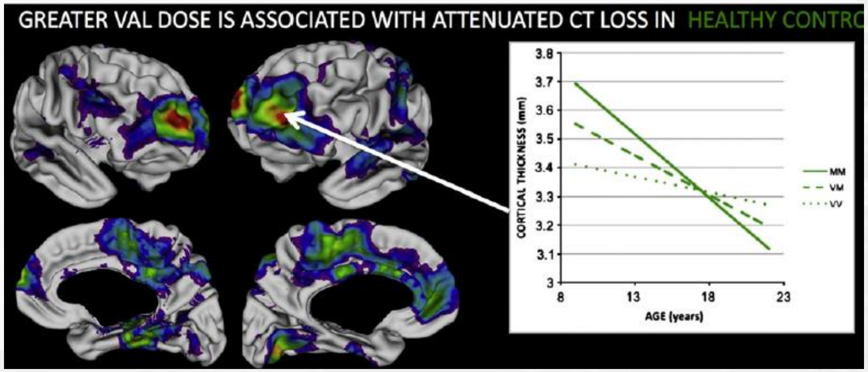
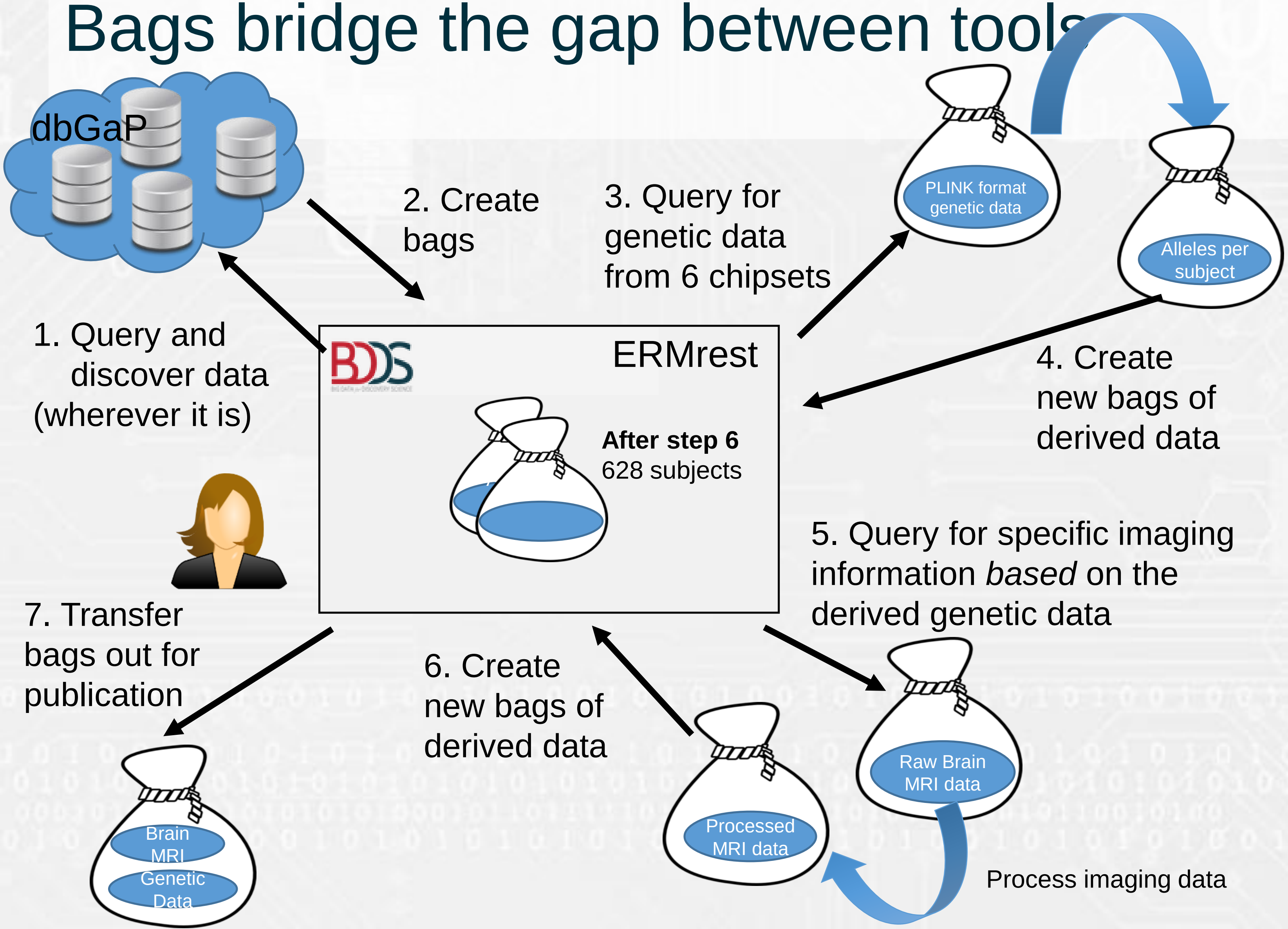


Image PheWAS

1. Assemble Data Collections
2. Identify subjects with images and extract images
3. Compute image phenotypes
   - Use Freesurfer with different atlases and computed measures
4. Associate Freesurfer results with each subject.
5. Quality control on derived data. Rerun on bad results
6. Identify subset of subjects that have variant of interest in SNP being considered
7. Collect up all phenotype data associated with identified subset
8. Do correlation analysis of phenotypes for the SNP to look for predictive correlations.

Bags bridge the gap between tools



1. Query and discover data (wherever it is)

2. Create bags

3. Query for genetic data from 6 chipsets

4. Create new bags of derived data

5. Query for specific imaging information *based* on the derived genetic data

6. Create new bags of derived data

7. Transfer bags out for publication

Process imaging data

After step 6
628 subjects

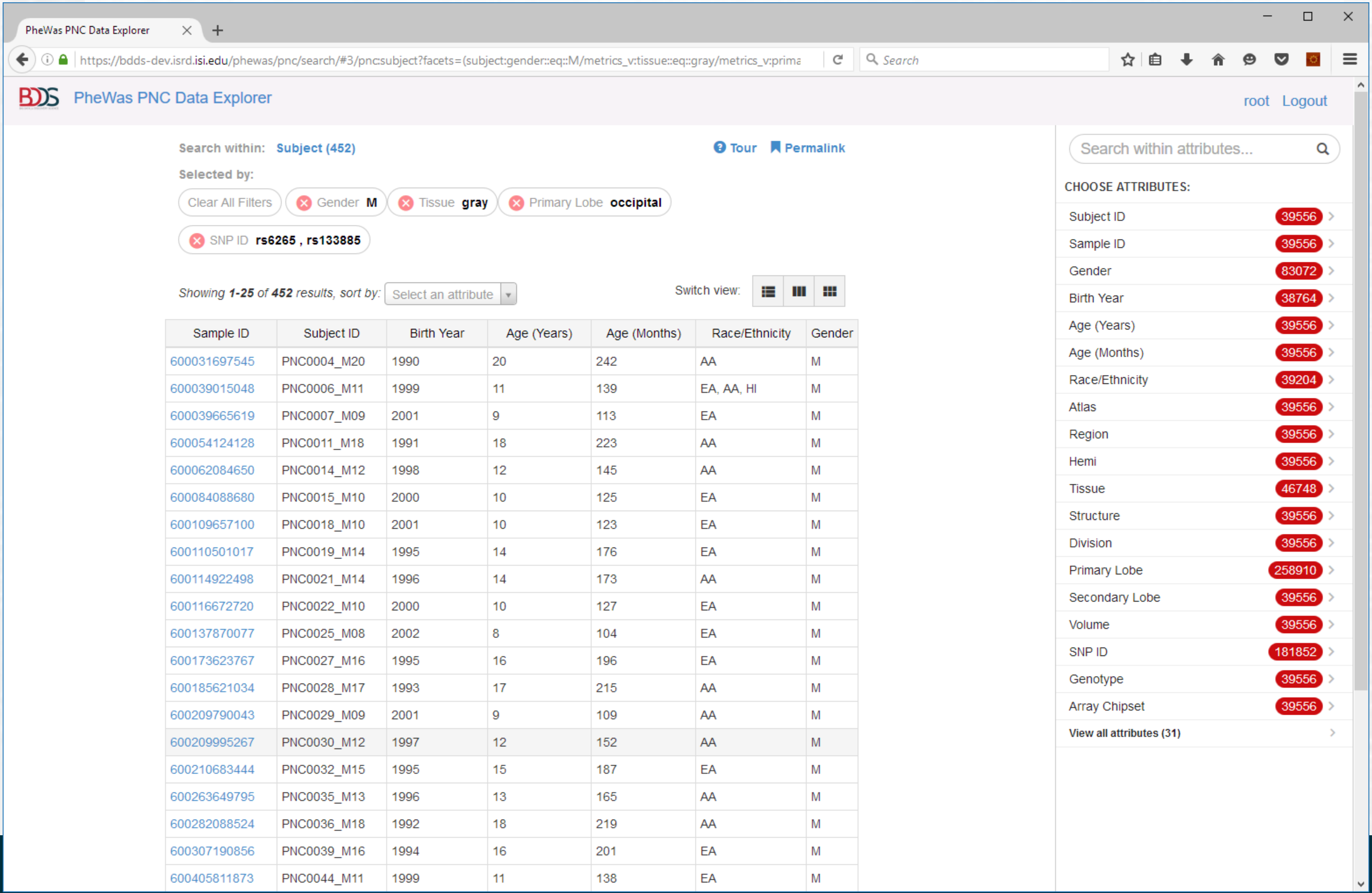
Assemble Data Collections

- …create bag with references to dbGap
- Log into dbGap and resolve references
- Assemble results in local directory
- Validate bag
- Ingest bag into catalog

Philadelphia Neurodevelopmental Consortium

- 8719 subjects in study
  - Baseline clinical elements
- 6 different SNP array chipsets resulting in a combined set of 1,873,486 distinct SNPs (out of a possible 85 million in the human genome).
  - The total combinatorial space of the genomic data is 5,435,533,460 (SNP, subject, allele) tuples across the 8719 subjects
- 997 of the subjects have MRI imaging data

Managing data collections



https://bdds-dev.isrd.isi.edu/phewas/pnc/search/#3/pnc:subject?facets=(subject:gender::eq::M/metrics_v:tissue::eq::gray/metrics_v:prima

PheWas PNC Data Explorer

Search within: Subject (452)

Selected by:

Clear All Filters | Gender **M** | Tissue **gray** | Primary Lobe **occipital** | SNP ID **rs6265 , rs133885**

*Showing **1-25** of **452** results, sort by:* Select an attribute

Sample ID	Subject ID	Birth Year	Age (Years)	Age (Months)	Race/Ethnicity	Gender
600031697545	PNC0004_M20	1990	20	242	AA	M
600039015048	PNC0006_M11	1999	11	139	EA, AA, HI	M
600039665619	PNC0007_M09	2001	9	113	EA	M
600054124128	PNC0011_M18	1991	18	223	AA	M
600062084650	PNC0014_M12	1998	12	145	AA	M
600084088680	PNC0015_M10	2000	10	125	EA	M
600109657100	PNC0018_M10	2001	10	123	EA	M
600110501017	PNC0019_M14	1995	14	176	EA	M
600114922498	PNC0021_M14	1996	14	173	AA	M
600116672720	PNC0022_M10	2000	10	127	EA	M
600137870077	PNC0025_M08	2002	8	104	EA	M
600173623767	PNC0027_M16	1995	16	196	EA	M
600185621034	PNC0028_M17	1993	17	215	AA	M
600209790043	PNC0029_M09	2001	9	109	AA	M
600209995267	PNC0030_M12	1997	12	152	AA	M
600210683444	PNC0032_M15	1995	15	187	EA	M
600263649795	PNC0035_M13	1996	13	165	AA	M
600282088524	PNC0036_M18	1992	18	219	AA	M
600307190856	PNC0039_M16	1994	16	201	EA	M
600405811873	PNC0044_M11	1999	11	138	EA	M

CHOOSE ATTRIBUTES:

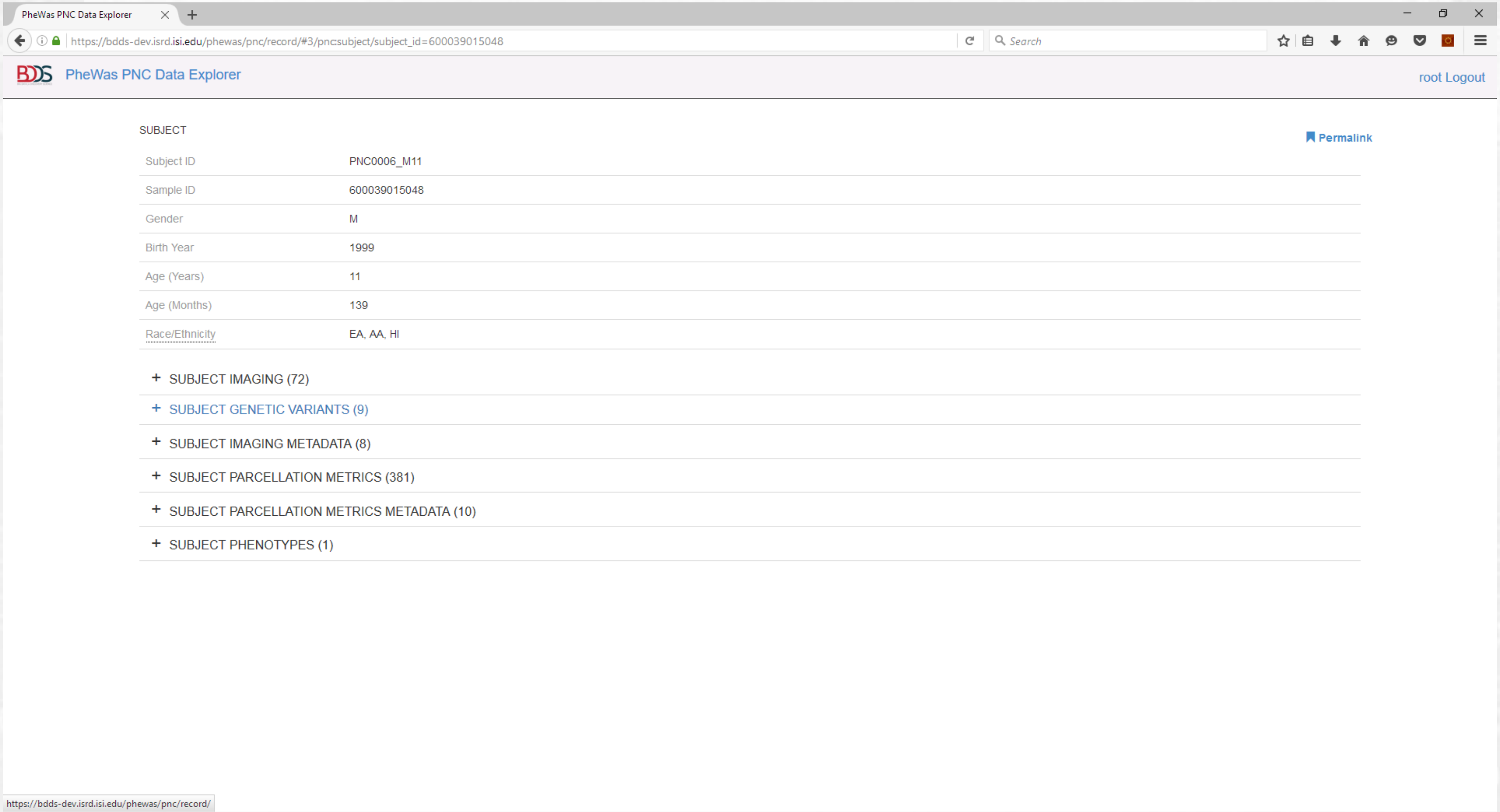
Subject ID	39556
Sample ID	39556
Gender	83072
Birth Year	38764
Age (Years)	39556
Age (Months)	39556
Race/Ethnicity	39204
Atlas	39556
Region	39556
Hemi	39556
Tissue	46748
Structure	39556
Division	39556
Primary Lobe	258910
Secondary Lobe	39556
Volume	39556
SNP ID	181852
Genotype	39556
Array Chipset	39556

View all attributes (31)

Export Bags for Phenotype calculation

- Bag structure used to collect data sets, along with description of what should be computing
- Collect results of calculation into bag and reingest
- Parcellation process generates 381 distinct regional measurements per subject, for a total of 370,641 regional neuro-anatomical measurements
- Surface-based model generates > 2 millions local measurements per subject, for a total of > 2 billions local neuro-anatomical measurements
- Image data for the 997 subjects consists of 70930 files (including derived images) @ 666GB

Details on one data element



PheWas PNC Data Explorer
https://bdds-dev.isrd.isi.edu/phewas/pnc/record/#3/pnc:subject/subject_id=600039015048
Search
PheWas PNC Data Explorer
root Logout
Permalink

SUBJECT

Subject ID	PNC0006_M11
Sample ID	600039015048
Gender	M
Birth Year	1999
Age (Years)	11
Age (Months)	139
Race/Ethnicity	EA, AA, HI

+ SUBJECT IMAGING (72)

+ SUBJECT GENETIC VARIANTS (9)

+ SUBJECT IMAGING METADATA (8)

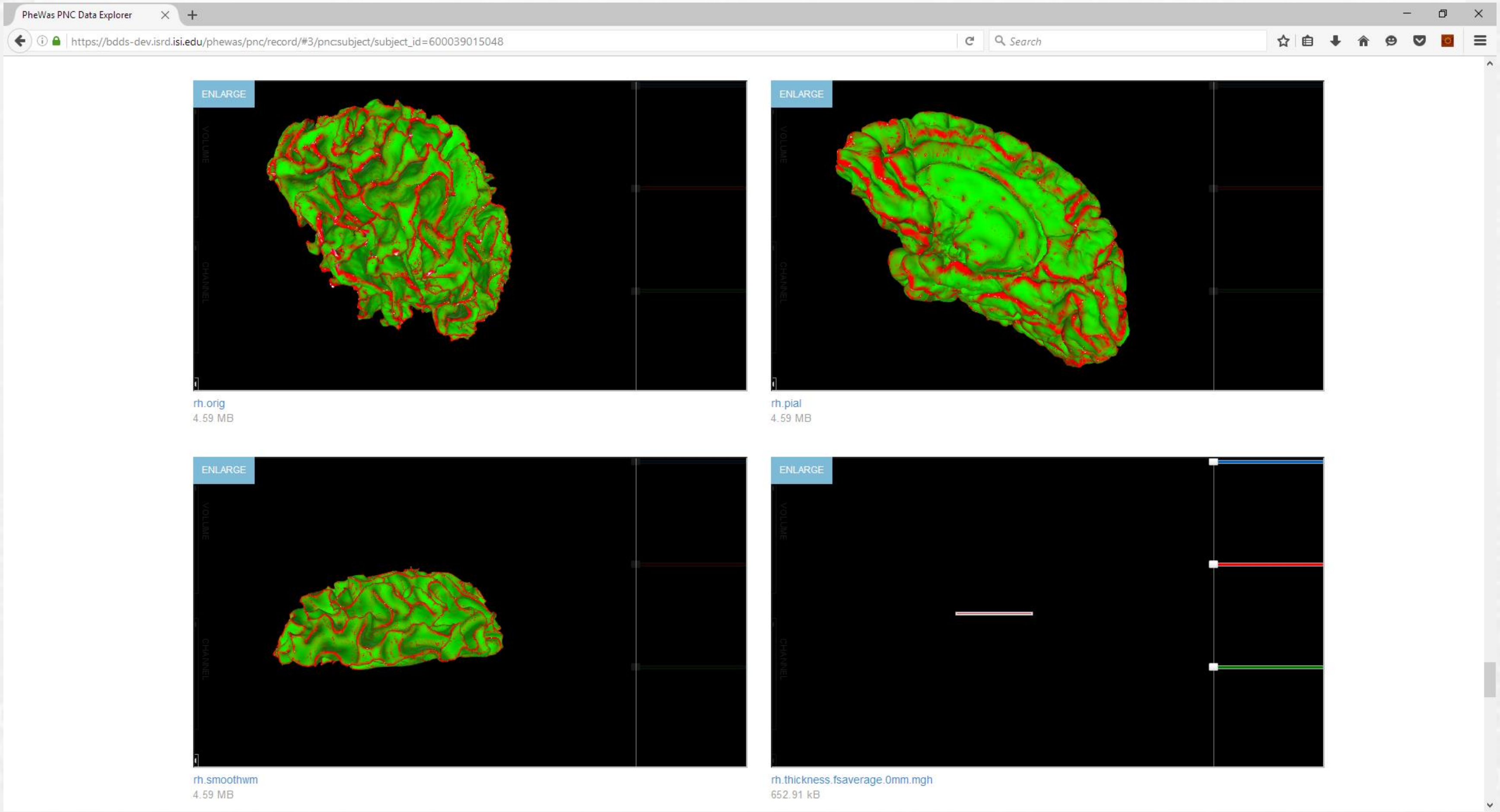
+ SUBJECT PARCELLATION METRICS (381)

+ SUBJECT PARCELLATION METRICS METADATA (10)

+ SUBJECT PHENOTYPES (1)

https://bdds-dev.isrd.isi.edu/phewas/pnc/record/

QC on derived data


PheWas PNC Data Explorer
https://bdds-dev.isrd.isi.edu/phewas/pnc/record/#3/pnc:subject/subject_id=600039015048
Search
ENLARGE
VOLUME
CHANNEL

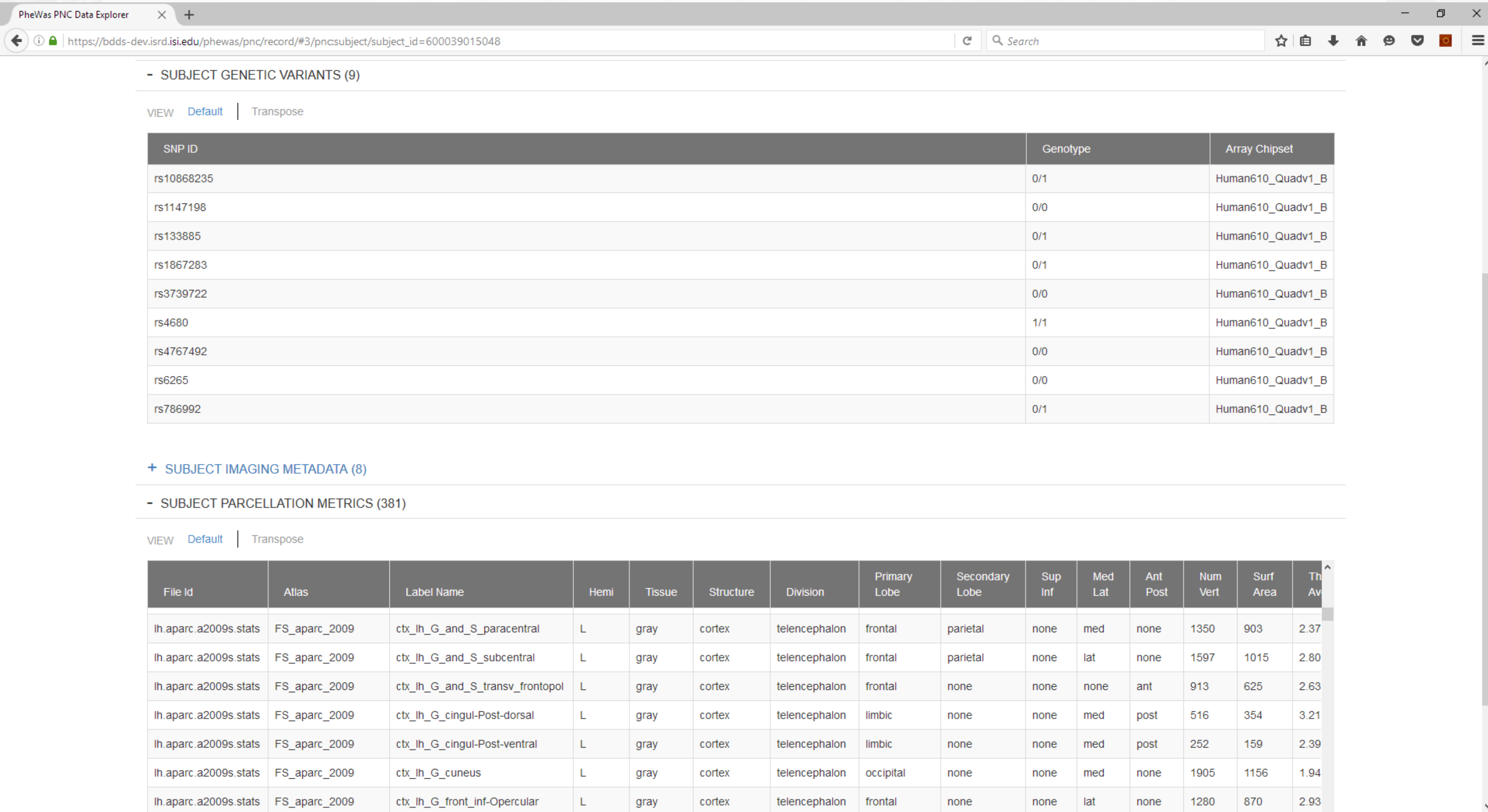
rh.orig
4.59 MB

rh.pial
4.59 MB

rh.smoothwm
4.59 MB

rh.thickness.fsaverage.0mm.mgh
652.91 kB

Complex data relationships…



https://bdds-dev.isrd.isi.edu/phewas/pnc/record/#3/pnc:subject/subject_id=600039015048

- SUBJECT GENETIC VARIANTS (9)

VIEW Default | Transpose

SNP ID	Genotype	Array Chipset
rs10868235	0/1	Human610_Quadv1_B
rs1147198	0/0	Human610_Quadv1_B
rs133885	0/1	Human610_Quadv1_B
rs1867283	0/1	Human610_Quadv1_B
rs3739722	0/0	Human610_Quadv1_B
rs4680	1/1	Human610_Quadv1_B
rs4767492	0/0	Human610_Quadv1_B
rs6265	0/0	Human610_Quadv1_B
rs786992	0/1	Human610_Quadv1_B

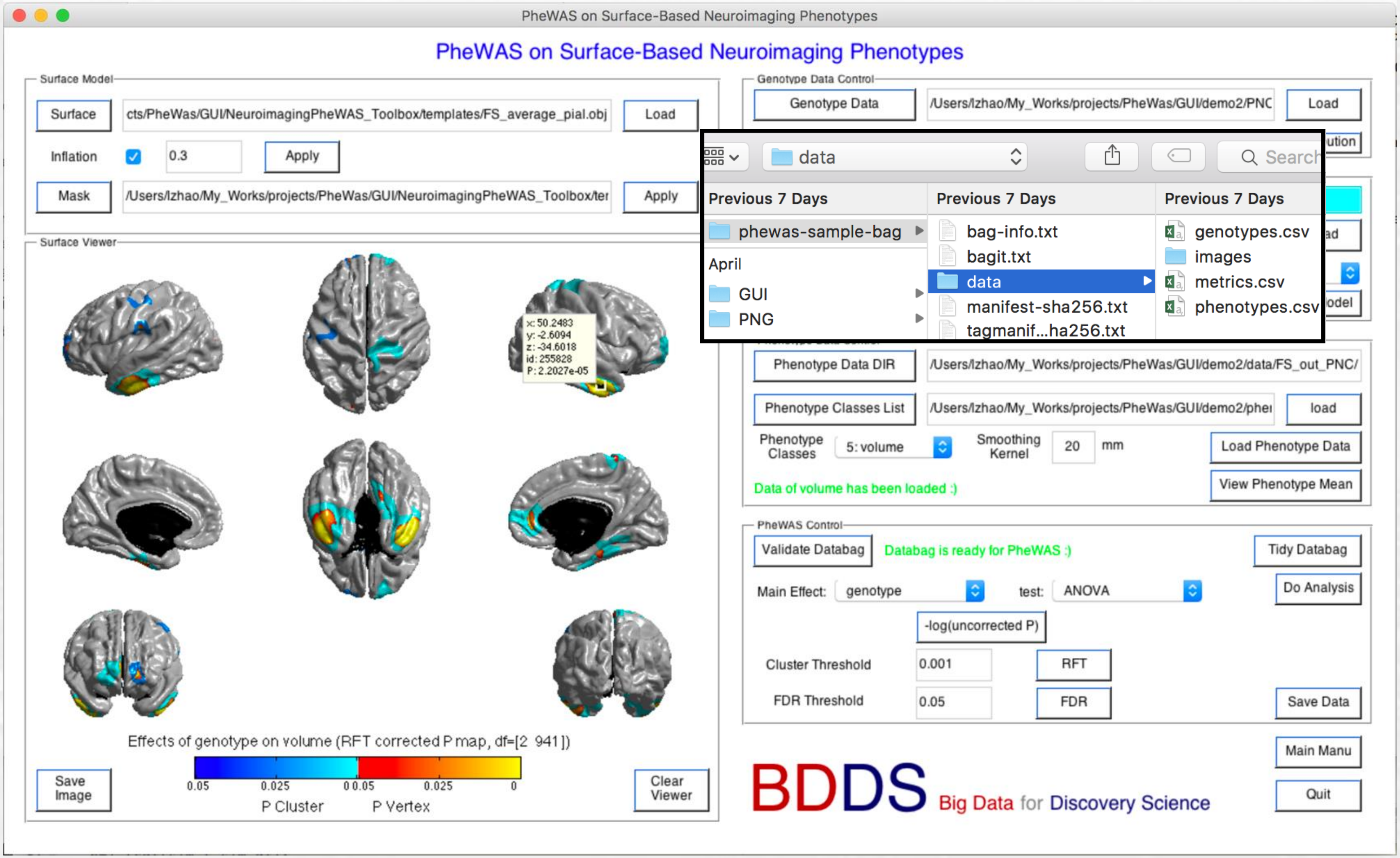
+ SUBJECT IMAGING METADATA (8)

- SUBJECT PARCELLATION METRICS (381)

VIEW Default | Transpose

File Id	Atlas	Label Name	Hemi	Tissue	Structure	Division	Primary Lobe	Secondary Lobe	Sup Inf	Med Lat	Ant Post	Num Vert	Surf Area	Th Av
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_and_S_paracentral	L	gray	cortex	telencephalon	frontal	parietal	none	med	none	1350	903	2.37
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_and_S_subcentral	L	gray	cortex	telencephalon	frontal	parietal	none	lat	none	1597	1015	2.80
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_and_S_transv_frontopol	L	gray	cortex	telencephalon	frontal	none	none	none	ant	913	625	2.63
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_cingul-Post-dorsal	L	gray	cortex	telencephalon	limbic	none	none	med	post	516	354	3.21
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_cingul-Post-ventral	L	gray	cortex	telencephalon	limbic	none	none	med	post	252	159	2.39
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_cuneus	L	gray	cortex	telencephalon	occipital	none	none	med	none	1905	1156	1.94
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_front_inf-Opercular	L	gray	cortex	telencephalon	frontal	none	none	lat	none	1280	870	2.93

NeuroimagingPheWAS Toolbox



PheWAS on Surface-Based Neuroimaging Phenotypes

Surface: cts/PheWas/GUI/NeuroimagingPheWAS_Toolbox/templates/FS_average_pial.obj

Inflation ☑ 0.3

Mask: /Users/lzhao/My_Works/projects/PheWas/GUI/NeuroimagingPheWAS_Toolbox/ter

Genotype Data: /Users/lzhao/My_Works/projects/PheWas/GUI/demo2/PNC

Phenotype Data DIR: /Users/lzhao/My_Works/projects/PheWas/GUI/demo2/data/FS_out_PNC/

Phenotype Classes List: /Users/lzhao/My_Works/projects/PheWas/GUI/demo2/phe

Phenotype Classes: 5: volume

Smoothing Kernel: 20 mm

Data of volume has been loaded :)

Databag is ready for PheWAS :)

Main Effect: genotype

test: ANOVA

Cluster Threshold: 0.001

FDR Threshold: 0.05

x: 50.2483
y: -2.6094
z: -34.6018
id: 255828
P: 2.2027e-05

Effects of genotype on volume (RFT corrected P map, df=[2 941])

What can we do now we couldn't do before?

- Broad survey for true system-level genetic associations across the whole population
  - All kinds of imaging genome data and processing
  - Not just for PNC, PING
  - Extensible to other phenotypes, not just FreeSurfer
- Build more complex studies the previously possible
  - e.g. TRENA + PheWAS
- Reproducible, complex, multistep big-data analysis

BDDS Demos

This afternoon

- TReNA
  - Ravi Madduri, Ben Heavner
- PheWAS
  - Carl Kesselman, Mike D’Arcy, Kristi Clark, Lu Zhao
- Panther
  - Huaiyu Mi, Anushya Muruganujan
- Data Publication
  - Ian Foster

  - Dry Creek Valley I room