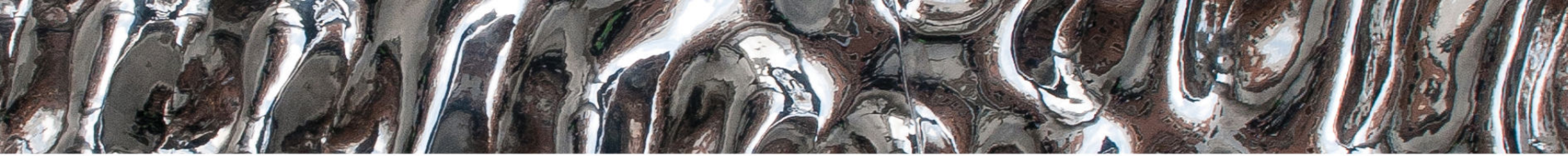


UiO : Department of Informatics
University of Oslo

GSuite Tools – efficiently manage and analyze collections of genomic data

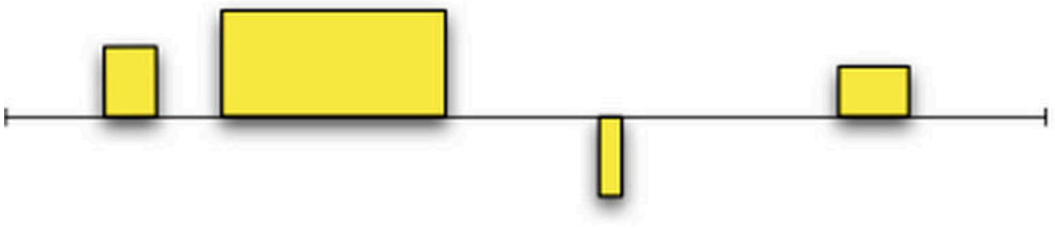
Boris Simovski, Sveinung Gundersen, Abdulrahman Azab, Diana Domanska, Eivind Hovig, Geir Kjetil Sandve

Outline

- Genomic tracks, HB
- Why collections?
- GSuite format
- GSuite Tools
- Demo

Genomic tracks

- Collection of objects of a specific genomic feature with base-pair-locations relative to reference genome assemblies
- Powerful way of representing genome-scale data
- “Identifying elemental genomic track types and representing them uniformly” – S. Gundersen et al (2011)

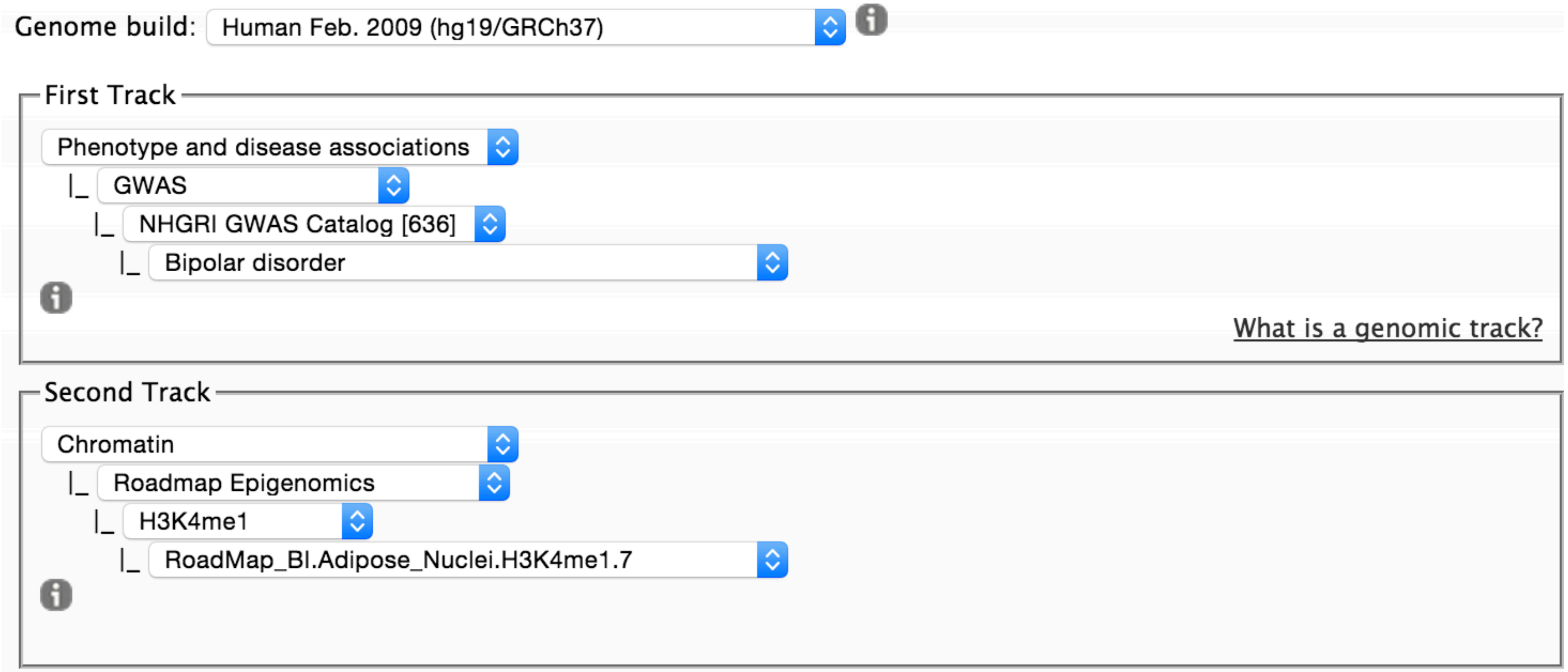


Valued Segments (VS)

The Genomic HyperBrowser

- Open-ended web server system for processing and statistical analysis of genomic tracks
- Offers a set of statistical analyses
  - Descriptive statistics
  - Hypothesis testing
  - Single track or a pair of tracks

Instead of demo… (1/3)



Genome build: Human Feb. 2009 (hg19/GRCh37)

First Track

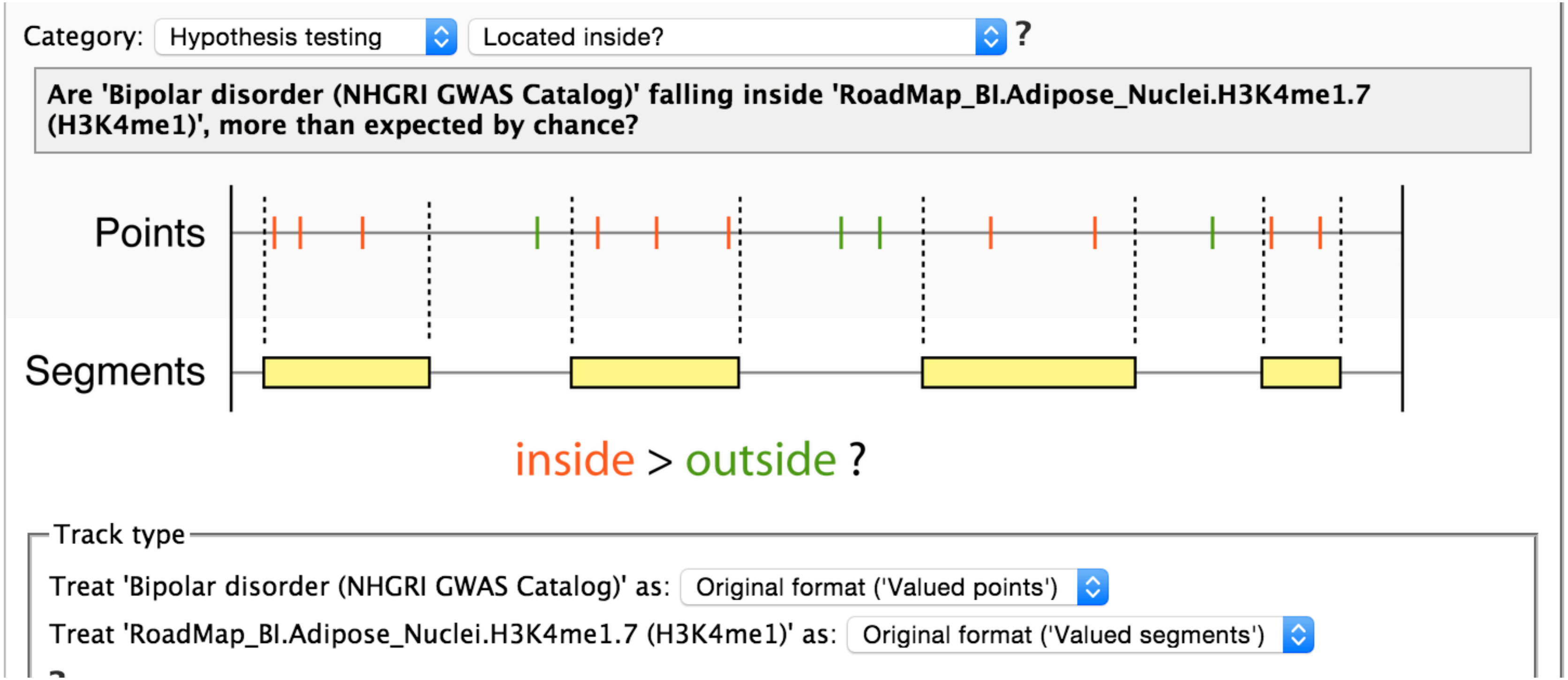
Phenotype and disease associations
|_ GWAS
|_ NHGRI GWAS Catalog [636]
|_ Bipolar disorder

What is a genomic track?

Second Track

Chromatin
|_ Roadmap Epigenomics
|_ H3K4me1
|_ RoadMap_BI.Adipose_Nuclei.H3K4me1.7

Instead of demo… (2/3)



Category: Hypothesis testing | Located inside? ?

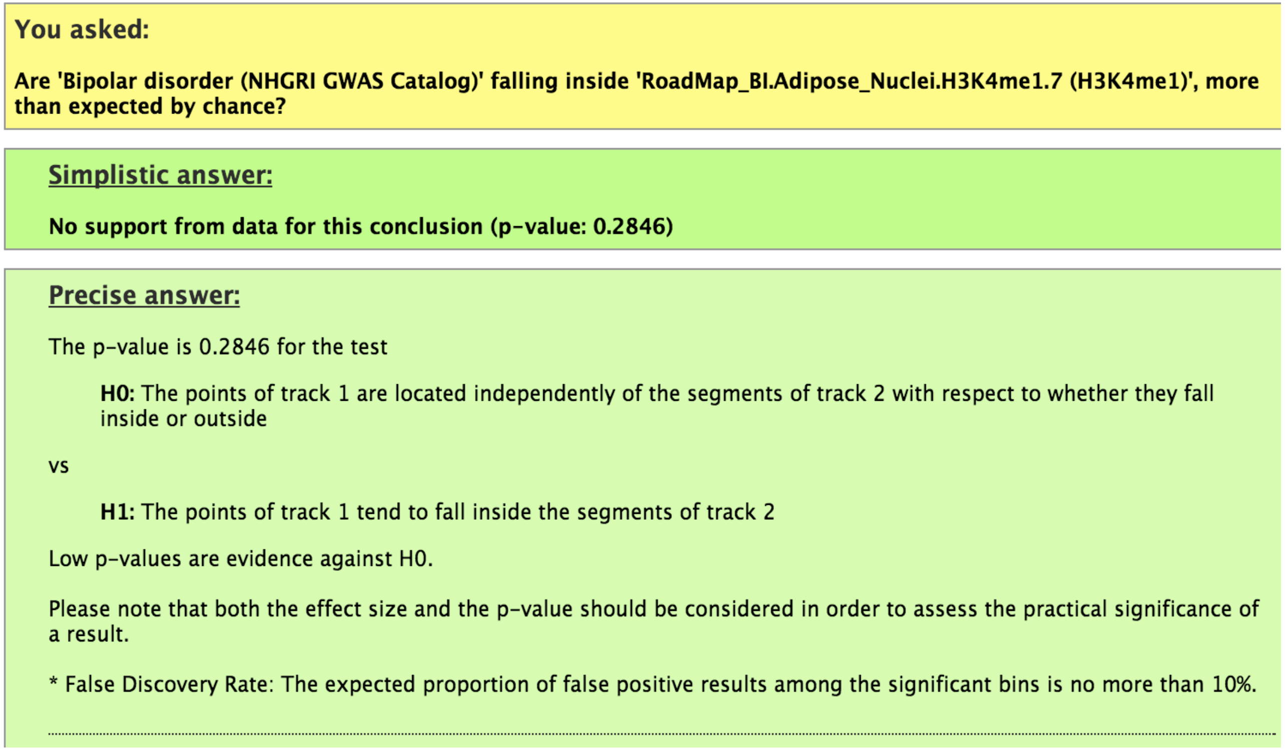
Are 'Bipolar disorder (NHGRI GWAS Catalog)' falling inside 'RoadMap_BI.Adipose_Nuclei.H3K4me1.7 (H3K4me1)', more than expected by chance?

Track type

Treat 'Bipolar disorder (NHGRI GWAS Catalog)' as: Original format ('Valued points')

Treat 'RoadMap_BI.Adipose_Nuclei.H3K4me1.7 (H3K4me1)' as: Original format ('Valued segments')

Instead of demo… (2/3)



You asked:

Are 'Bipolar disorder (NHGRI GWAS Catalog)' falling inside 'RoadMap_BI.Adipose_Nuclei.H3K4me1.7 (H3K4me1)', more than expected by chance?

Simplistic answer:

No support from data for this conclusion (p-value: 0.2846)

Precise answer:

The p-value is 0.2846 for the test

> **H0:** The points of track 1 are located independently of the segments of track 2 with respect to whether they fall inside or outside

vs

> **H1:** The points of track 1 tend to fall inside the segments of track 2

Low p-values are evidence against H0.

Please note that both the effect size and the p-value should be considered in order to assess the practical significance of a result.

* False Discovery Rate: The expected proportion of false positive results among the significant bins is no more than 10%.

Why dataset collections?

- Even more genome-wide data is now publicly available
- Multiple track analysis is the natural next step
- Analyze a collection of tracks of a specific genomic feature for different cell lines (e.g. H3K4me3 for cell lines from different tissue)
- Analyze a collection of tracks of genomic features for a specific cell line (e.g. several histone modifications for a liver tissue cell)

GSuite format

- A tabular format
- Represent dataset collections and some basic metadata
- One genomic track per line
- Easy to create
- Flexible
- Easy to process by analysis tools
- Specification: https://hyperbrowser.uio.no/gsuite/static/hyperbrowser/gsuite/GSuite_specification.txt?x=x

```
##location: unknown
##file type: unknown
##track type: unknown
##genome: unknown
```

```
##location: remote
##file type: unknown
##track type: unknown
##genome: unknown
###uri       title
http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/E115-
H2A.Z.broadPeak.gz    0_E115-H2A.Z.broadPeak.gz
http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/E115-
H3K27ac.broadPeak.gz                1_E115-H3K27ac.broadPeak.gz
http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/E115-
H3K27me3.broadPeak.gz               2_E115-H3K27me3.broadPeak.gz
http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/E115-
H3K36me3.broadPeak.gz               3_E115-H3K36me3.broadPeak.gz
http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/E115-
H3K4me1.broadPeak.gz                4_E115-H3K4me1.broadPeak.gz
```

```
##location: local
##file type: text
##track type: unknown
##genome: unknown
###uri        title
galaxy:/e4efca/E115-H2A.Z.broadPeak;broadPeak        0_E115-H2A.Z.broadPeak
galaxy:/e4efca/E115-H3K27ac.broadPeak;broadPeak      1_E115-H3K27ac.broadPeak
galaxy:/e4efca/E115-H3K27me3.broadPeak;broadPeak     2_E115-H3K27me3.broadPeak
galaxy:/e4efca/E115-H3K36me3.broadPeak;broadPeak     3_E115-H3K36me3.broadPeak
galaxy:/e4efca/E115-H3K4me1.broadPeak;broadPeak      4_E115-H3K4me1.broadPeak
```

```
##location: local
##file type: binary
##track type: valued segments
##genome: hg19
###uri         title
hb:/ext/gsuite/006/6522/0_E115-H2A.Z.broadPeak         0_E115-H2A.Z.broadPeak
hb:/ext/gsuite/006/6522/1_E115-H3K27ac.broadPeak       1_E115-H3K27ac.broadPeak
hb:/ext/gsuite/006/6522/2_E115-H3K27me3.broadPeak    2_E115-H3K27me3.broadPeak
hb:/ext/gsuite/006/6522/3_E115-H3K36me3.broadPeak    3_E115-H3K36me3.broadPeak
hb:/ext/gsuite/006/6522/4_E115-H3K4me1.broadPeak      4_E115-H3K4me1.broadPeak
```

```
##location: local
##file type: binary
##track type: segments
##genome: multiple
###uri        title           track_type  genome
hb:/ext/gsuite/006/6522/0_E115-H2A.Z.broadPeak          0_E115-H2A.Z.broadPeak           valued
segments   hg19
hb:/ext/gsuite/006/6522/1_E115-H3K27ac.broadPeak        1_E115-H3K27ac.broadPeak         valued
segments   hg19
hb:/ext/gsuite/006/6522/2_E115-H3K27me3.broadPeak       2_E115-H3K27me3.broadPeak        valued
segments   hg19
hb:/ext/gsuite/006/6522/3_E115-H3K36me3.broadPeak       3_E115-H3K36me3.broadPeak        valued
segments   hg19
hb:/ext/gsuite/006/6522/4_E115-H3K4me1.broadPeak        4_E115-H3K4me1.broadPeak         valued
segments   hg19
hb:/Genes and gene subsets/Genes/CCDS     CCDS (Genes)          valued segments        hg18
hb:/Genes and gene subsets/Genes/Ensembl  Ensembl (Genes)       valued segments        hg18
hb:/Genes and gene subsets/Genes/GeneID   GeneID (Genes)        valued segments        hg18
hb:/Genes and gene subsets/Genes/Hinxton Coverage       Hinxton Coverage (Genes)
             segments   hg18
```

GSuite Tools

- 1. Compile GSuite – locate and fetch tracks.
- 2. Customize GSuite – manipulate rows and columns.
- 3. Analyze GSuite – several multitrack scenarios.

Compile GSuite

- From a remote source
  - Currently supported public database:
    - ENCODE, Roadmap Epigenomics, Cancer Genome Atlas, FANTOM 5, ICGC Data Portal, BLUEPRINT project hub, NHGRI-EBI GWAS Catalog
  - Supported protocols
    - http(s), ftp, rsync
- From a catalog of chromatin tracks
- From datasets in history
- From HyperBrowser repository
- From an archive (gsuite.tar, gsuite.zip)

Customize GSuite

- Select subset of metadata columns
- Select subset of tracks (rows) in GSuite
- Combine several GSuites
- Manipulate textual datasets referred in GSuite
- Preprocess for analysis

Analyze GSuite

- Analyze relations of tracks in GSuite.
- Screen a track against a collection.
- Screen two GSuits against each other.
- Visualize analysis results
  - Charts, heatmaps, Venn diagram
- Few domain-specific analysis tools

Demo.

Questions?

Useful links

- GSuite Tools
  - https://hyperbrowser.uio.no/gsuite/
- GSuite format specification
  - https://hyperbrowser.uio.no/gsuite/static/hyperbrowser/gsuite/GSuite_specification.txt?x=x
- Publication on genomic track types
  - http://www.biomedcentral.com/1471-2105/12/494/
- Publication on the Genomic HyperBrowser
  - http://www.ncbi.nlm.nih.gov/pubmed/23632163