The Genomic HyperBrowser

Statistical genome analysis
made transparent and accessible

Eivind Hovig and Geir Kjetil Sandve

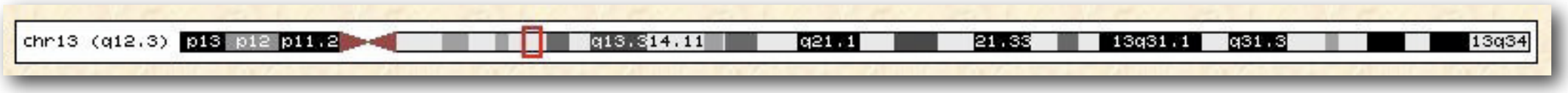
Oslo University Hospital

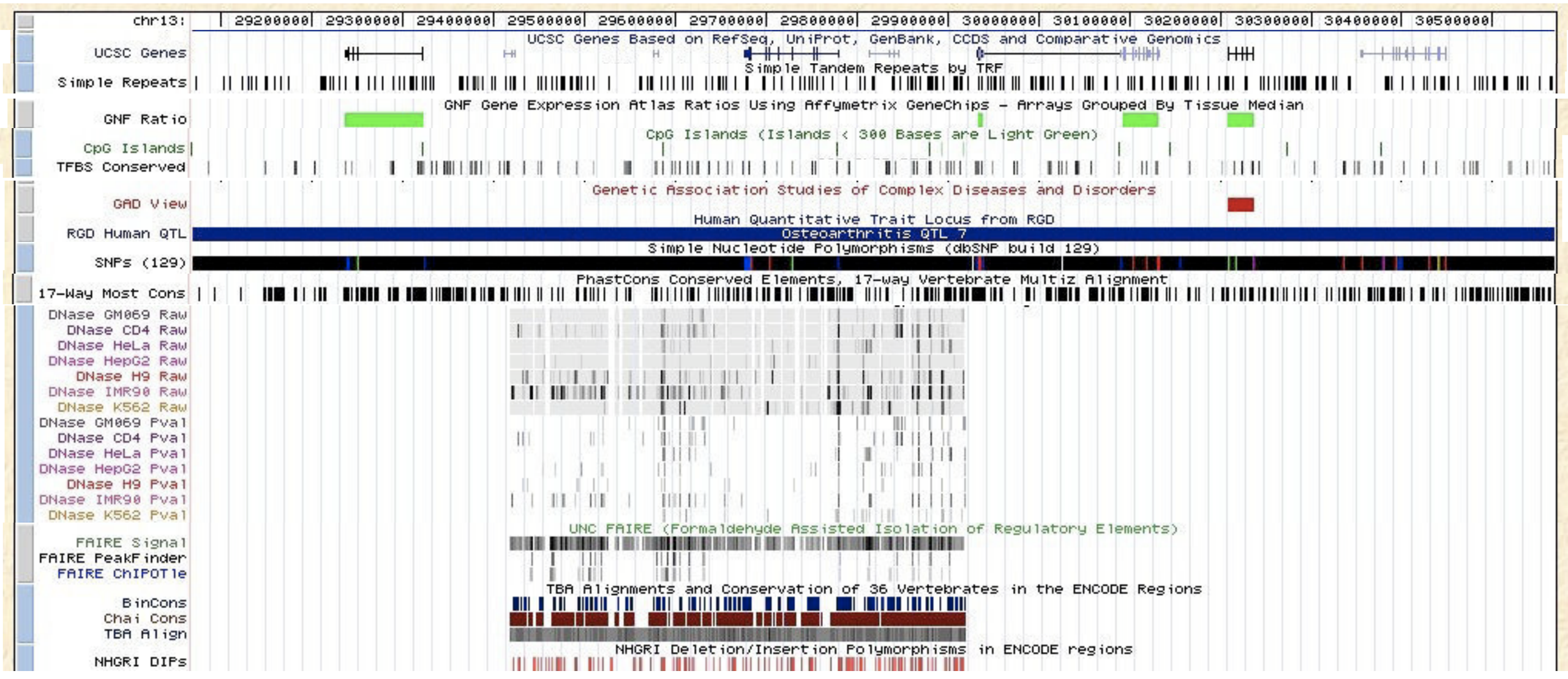
UiO : University of Oslo

Challenges and opportunities in the post genomic era

- I will next year generate more data than all up to now
- But asking the crucial questions is still difficult!
- Need analysis methodology to match the power of data generation systems

Or take the enormous increase in public data





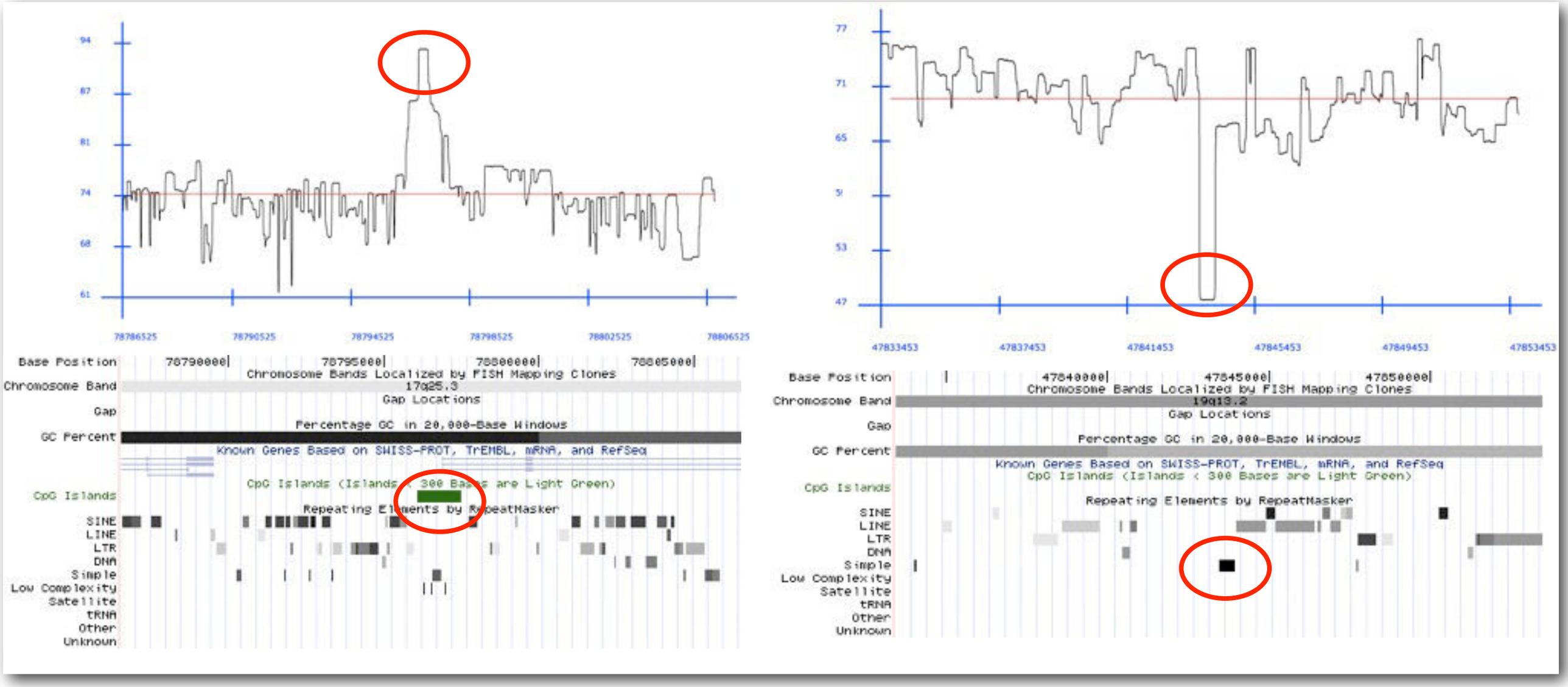
ENCODE, FANTOM, GEO, Roadmap Epigenomics ...

High melting temperature

Chr. 17: 93.25

Low melting temperature

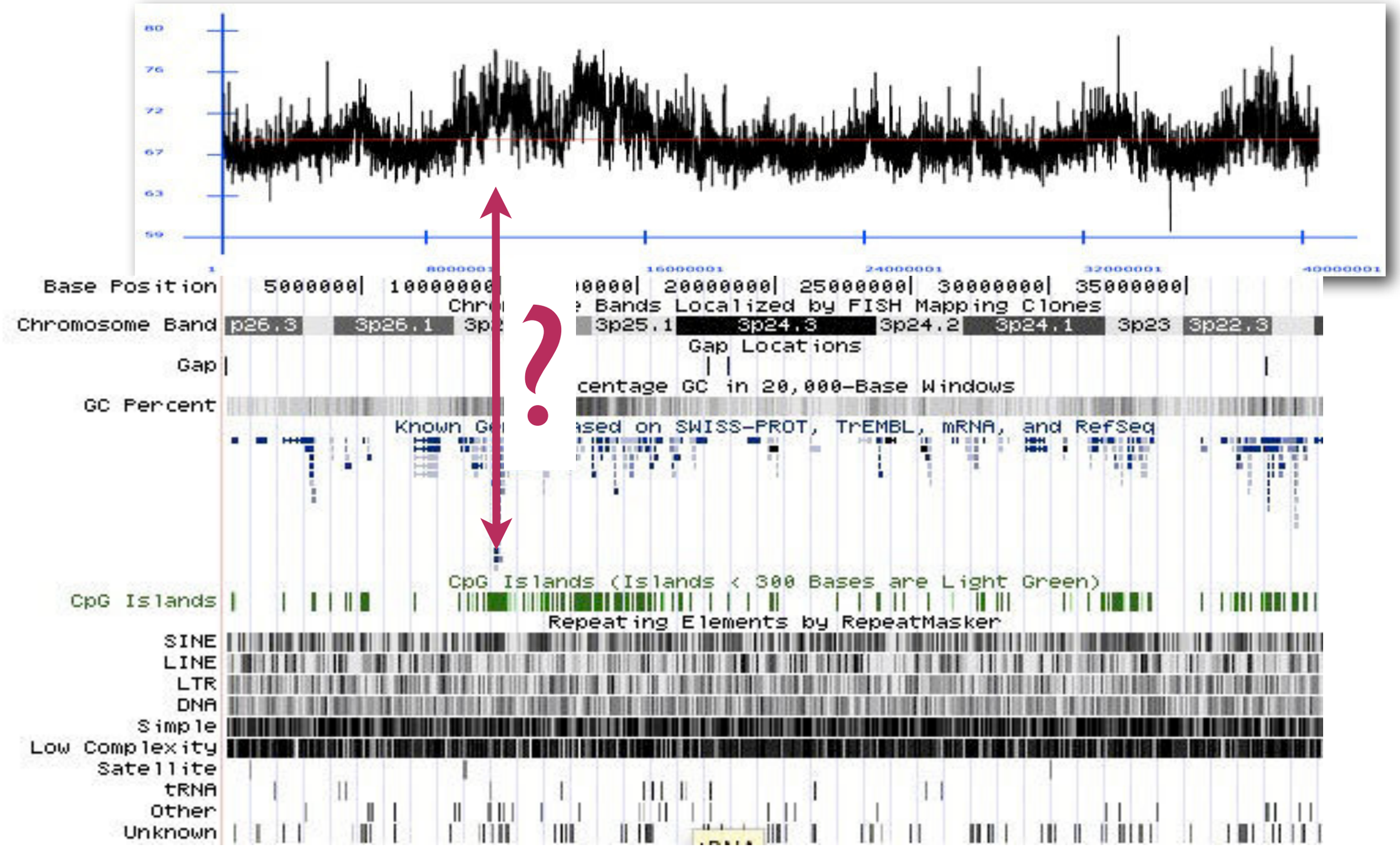
Chr. 19: 47.75



CpG Islands

AT simple repeats

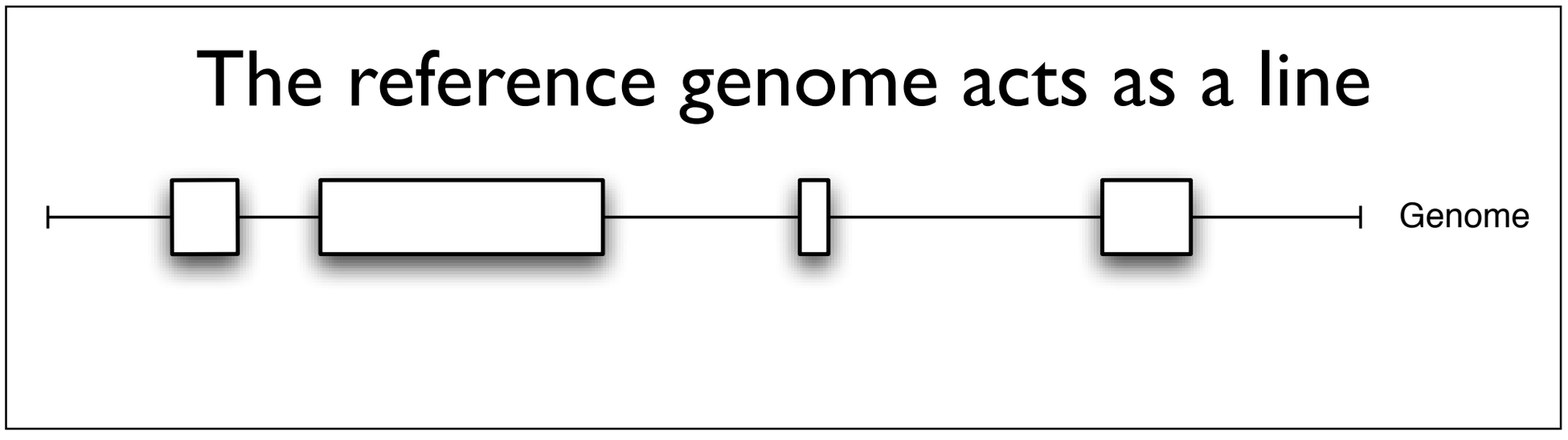
This can't be it?!



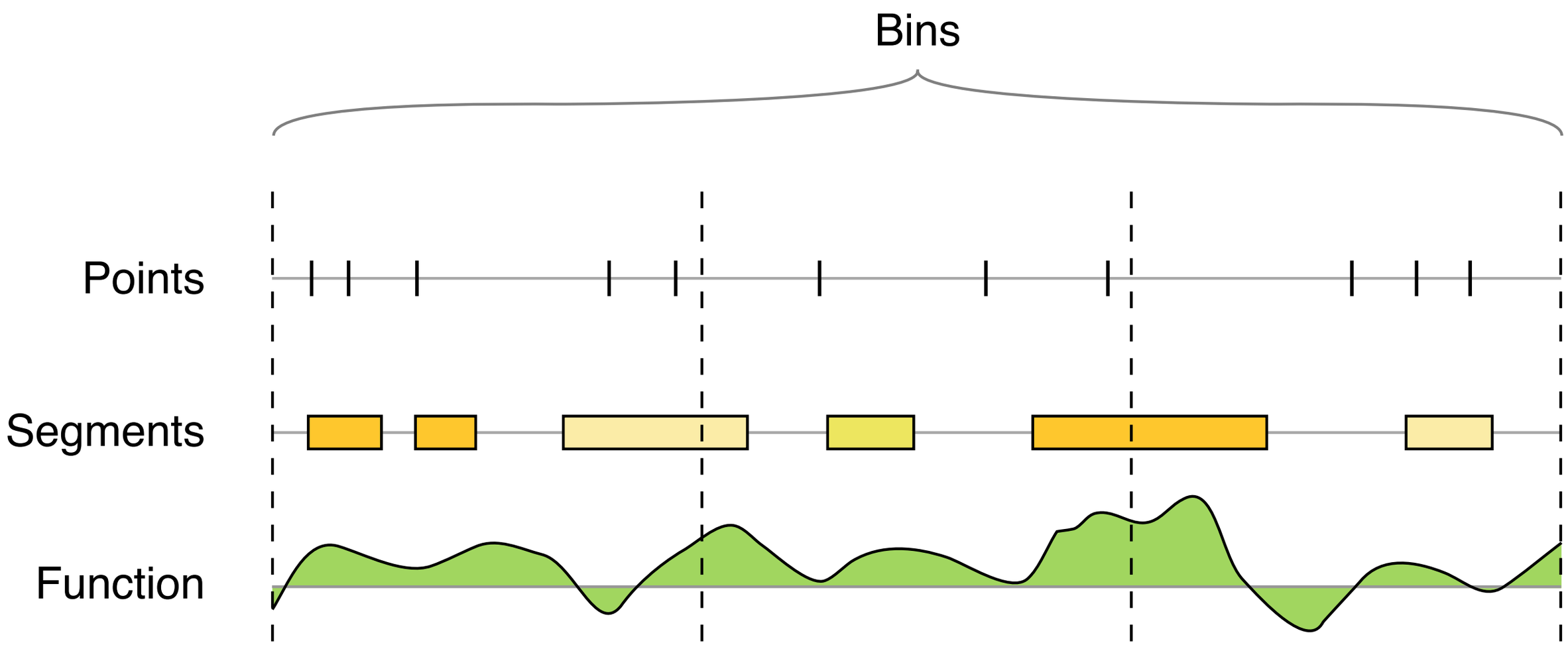
The whiteboard view of genome data

Reference genome acts like coordinate system for genomic data

```
chr21 10079666 10120808 NM_001187
chr21 13332357 13412442 NR_026916
chr21 13700575 13700652 NR_036164
chr21 13904368 13935777 NM_174981
chr21 14137324 14142556 NR_026755
```



Delineating basic types of genomic tracks



Cataloguing generic analyses of relations between the formal data types

P	P	Different frequencies?
P	P	Located nearby?
P	S	Located inside?
P	S	Located nonuniformly inside?
P	S	Located nearby?
S	S	Similar segments?
S	S	Overlap?
S	S	Located nearby?
F	F	Correlated?
P	F	Higher values at locations?
S	F	Higher values inside?
P	VS	Located in segments with high values?
S	VP	Higher values inside segments?
VP	VP	Nearby values similar?
P	VS (c/c)	Located in case segments
VS (c/c)	S	Preferential overlap?
VP (cat)	VS (cat)	Category pairs differentially co-located?
LGP	P	Colocalized in 3D?

Biological example

- B-cells important for the pathology of multiple sclerosis?

Genome build: Human Mar. 2006 (hg18/NCBI36)

First Track

-- From history (bed, wig, ...) --

1: imported: MS regions [hg18]

What is a genomic track?

Second Track

Chromatin [221]

|_ Chromatin state segmentation [144]

|_ wgEncodeBroadHmmGm12878HMM [16]

|_ 1 Active Promoter

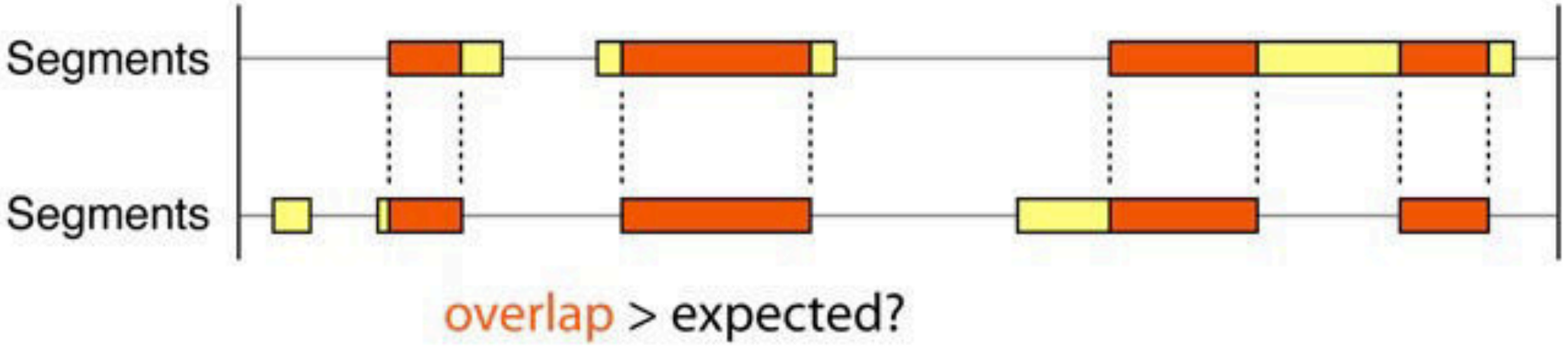
Analysis

Category: Hypothesis testing

Overlap?

?

Are 'imported. MS regions (1)' overlapping '1 Active Promoter (wgEncodeBroadHmmGm12878HMM)', more than expected by chance?



You asked:

Are 'imported: MS regions' overlapping '1 Active Promoter (wgEncodeBroadHmmGm12878HMM)', more than expected by chance?

Simplistic answer:

Yes - the data suggests this (p-value: 0.004975)

Precise answer:

The p-value is 0.004975 for the test

H0: The segments of track 1 are located independently of the segments of track 2 with respect to overlap

vs

H1: The segments of track 1 tend to overlap the segments of track 2

Low p-values are evidence against H0.

The test was also performed for each bin separately, resulting in 12 significant bins out of 26, at 10% FDR* (17 bins excluded from FDR-analysis due to lacking p-values).

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%.

```
@REGION=__chrArms__

@BINNING=*

@TN1=Chromatin:Chromatin%20state
%20segmentation:wgEncodeBroadHmmGm12878HMM:1%20Active
%20Promoter

@TN2=Phenotype and disease associations:Assorted
experiments:Multiple Sclerosis, Sawcer et al. (2011)

@ANALYSIS=RandomizationManagerStat(tf1=TrivialFormatConv
erter,tail=more,assumptions=PermutedSegsAndSampledInters
egsTrack_,rawStatistic=TpRawSegsOverlapStat,maxSamples=1
00,numResamplings=100,tf2=TrivialFormatConverter)

hg18|@REGION|@BINNING|@TN1|@TN2|@ANALYSIS
```

@REGION=__chrArms__

@BINNING=*

@TN1=Chromatin:Chromatin%20state
%20segmentation:wgEncodeBroadHmmGm12878HMM:* 1%20Active
%20Promoter

@TN2=Phenotype and disease associations:Assorted
experiments:Multiple Sclerosis, Sawcer et al. (2011)

@ANALYSIS=RandomizationManagerStat(tf1=TrivialFormatConv
erter,tail=more,assumptions=PermutedSegsAndSampledInters
egsTrack_,rawStatistic=TpRawSegsOverlapStat,maxSamples=1
00,numResamplings=100,tf2=TrivialFormatConverter)

hg18|@REGION|@BINNING|@TN1|@TN2|@ANALYSIS

B-cells important for MS?

- MS-associated regions (GWAS)
- Active regions in B-cells (chromatin state AP)
- Do MS overlap unexpectedly with B-cell AP?
  - But: They may also overlap other-cell AP..
  - Must use case-control analysis

Combine two BED files into single case-control track

Genome build: Human Mar. 2006 (hg18/NCBI36)

Select track to be used as case: 8 - Extract track from HyperBrowser repc

Select track to be used as control: 9 - Extract track from HyperBrowser repc

Shared regions should be: removed

Execute

Corresponding batch command line:

The Genomic HyperBrowser (v1.6)

Genome build: Human Mar. 2006 (hg18/NCBI36)

First Track

-- From history (bed, wig, ...) -- | 1: imported: MS regions [hg18]

What is a genomic track?

Second Track

-- From history (bed, wig, ...) -- | 10: Combine two BED files into single cas

Analysis

Category: Hypothesis testing | Preferential overlap? ?

Are 'Combine two BED files into single case-control track (10)' marked as case overlapping unexpectedly more with 'imported. MS regions (1)' than 'Combine two BED files into single case-control track (10)' marked as control?

Track type

Treat 'Combine two BED files into single case-control track (10)' as:
Original format (")

Treat 'imported. MS regions (1)' as: Original format (")

?

Options

Alternative hypothesis: more

Null model: Preserve segments of both tracks; permu

Inspect parameters of the analysis | Toggle debug

You asked:

Are 'Combine two BED files into single case-control track' marked as case overlapping unexpectedly more with 'imported: MS regions' than 'Combine two BED files into single case-control track' marked as control?

Simplistic answer:

Yes - the data suggests this (p-value: 0.004975)

Precise answer:

The p-value is 0.004975.

Low p-values are evidence against H0.

The test was also performed for each bin separately, resulting in 4 significant bins out of 26, at 10% FDR* (17 bins excluded from FDR-analysis due to lacking p-values).

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

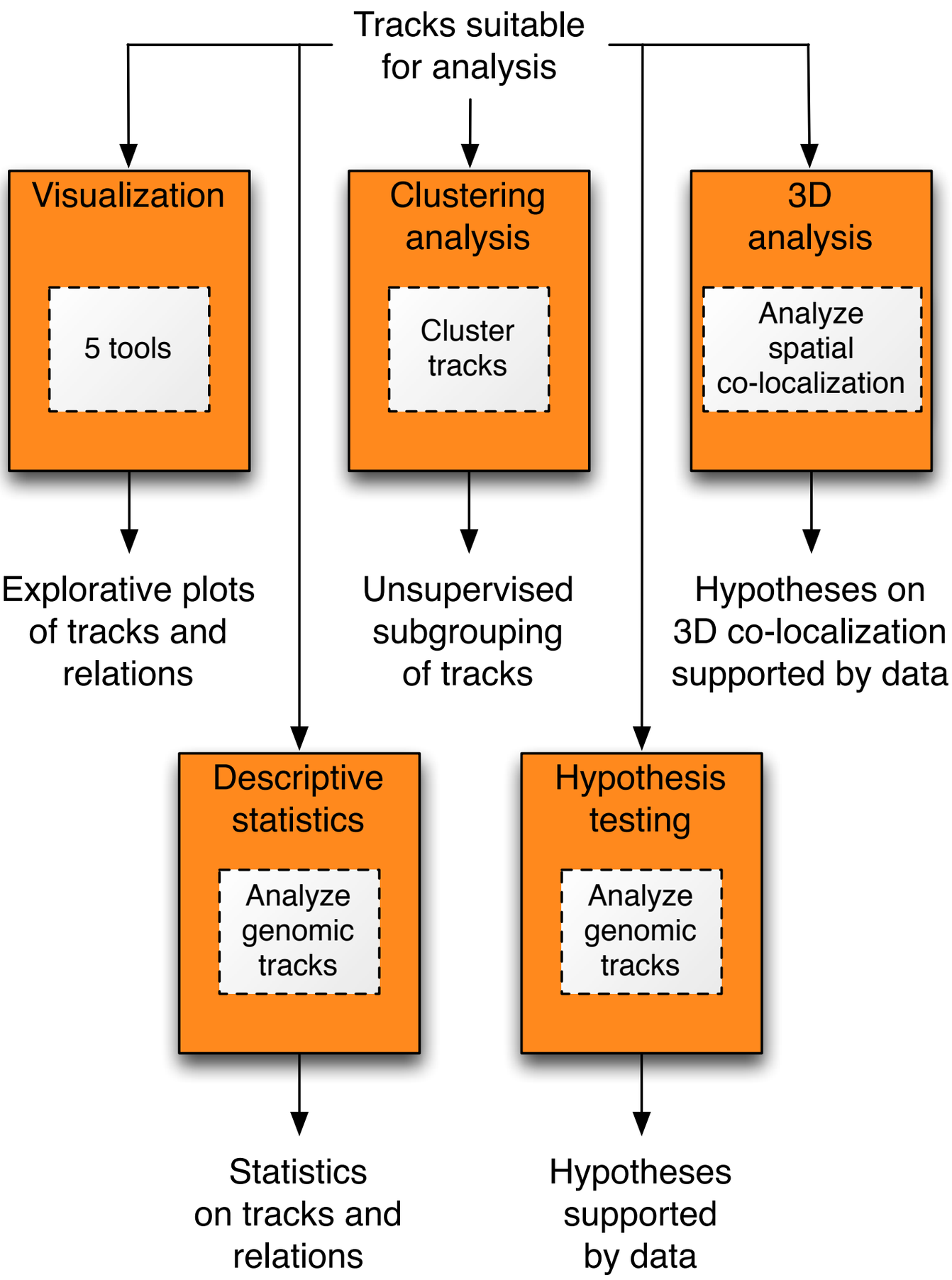
* False Discovery Rate: The expected proportion of false positive results among the significant bins is no more than 10%.

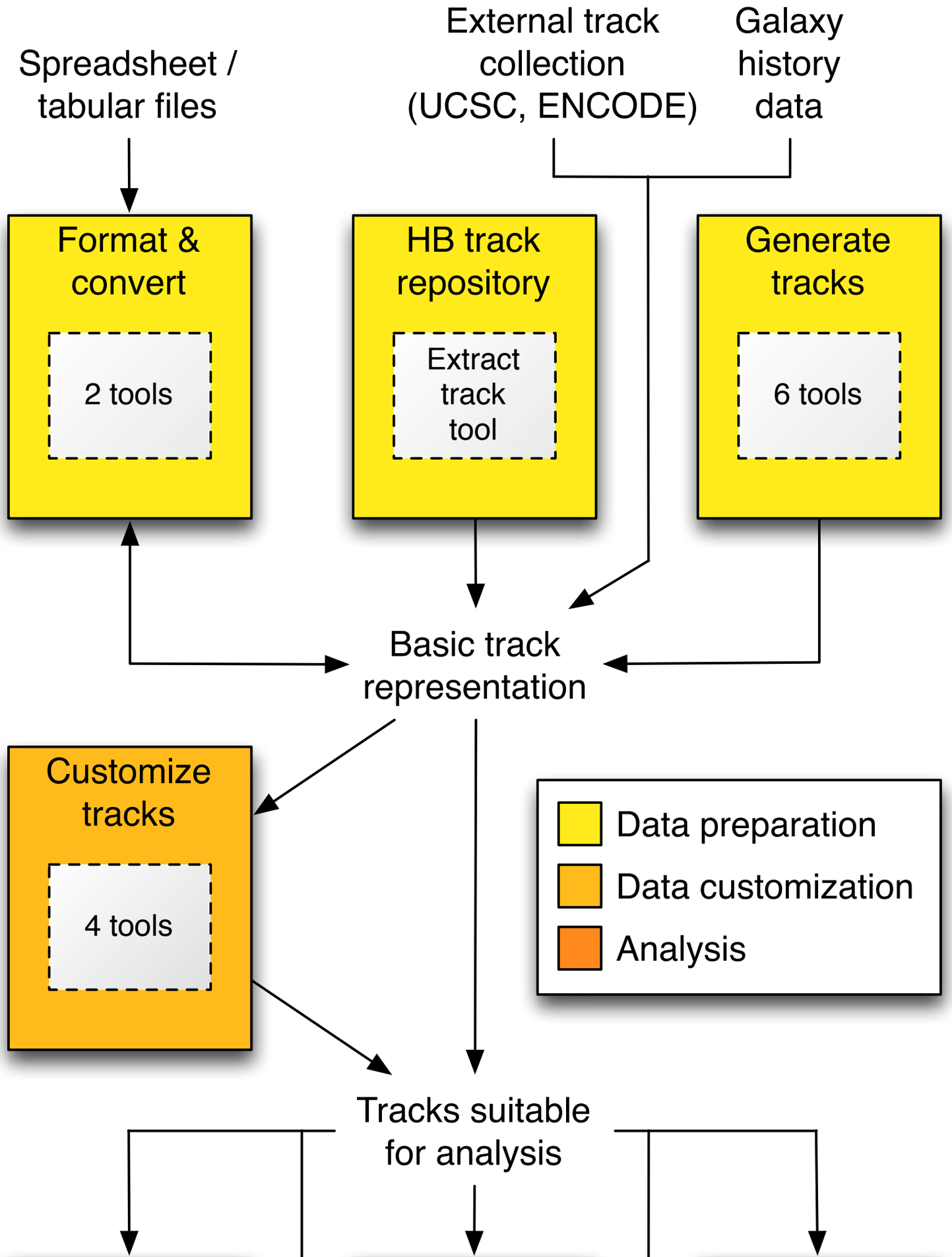
P-values were computed under the **null model** defined by the following preservation and randomization rules:

Preserve segments of both tracks; permute case and control assignment of T1-segments

The Genomic HyperBrowser

- Robust statistical treatment
- Dynamic Galaxy web interface
    - Determines meaningful analyses and options from data
- 76 statistical analyses
- 42 analysis-centric tools


Tracks suitable for analysis
Visualization
5 tools
Clustering analysis
Cluster tracks
3D analysis
Analyze spatial co-localization
Explorative plots of tracks and relations
Unsupervised subgrouping of tracks
Hypotheses on 3D co-localization supported by data
Descriptive statistics
Analyze genomic tracks
Hypothesis testing
Analyze genomic tracks
Statistics on tracks and relations
Hypotheses supported by data


Spreadsheet / tabular files
External track collection (UCSC, ENCODE)
Galaxy history data
Format & convert
2 tools
HB track repository
Extract track tool
Generate tracks
6 tools
Basic track representation
Customize tracks
4 tools
Data preparation
Data customization
Analysis
Tracks suitable for analysis

(sfi)² Statistics for Innovation

The Research Council of Norway

UiO : University of Oslo

Oslo University Hospital

If you have a genomic track, we can analyze it!

If you have a generic question for which we have no answer, we will develop it!

- Google "HyperBrowser" and try out the web system
- PubMed "HyperBrowser" and skim through our 2013 NAR article