



The Genomic HyperBrowser

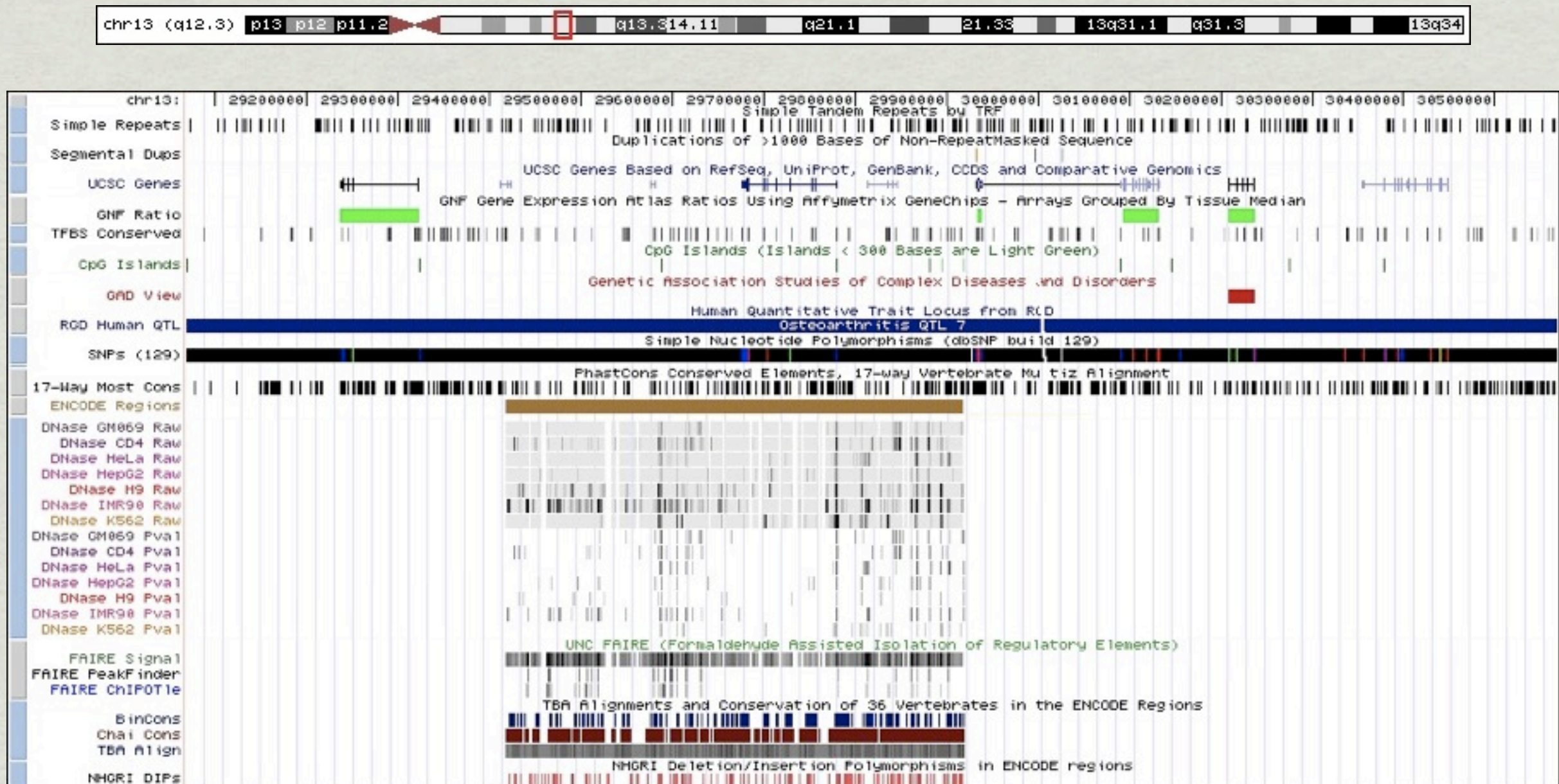
A tutorial

Sveinung Gundersen, PhD, Oslo University Hospital

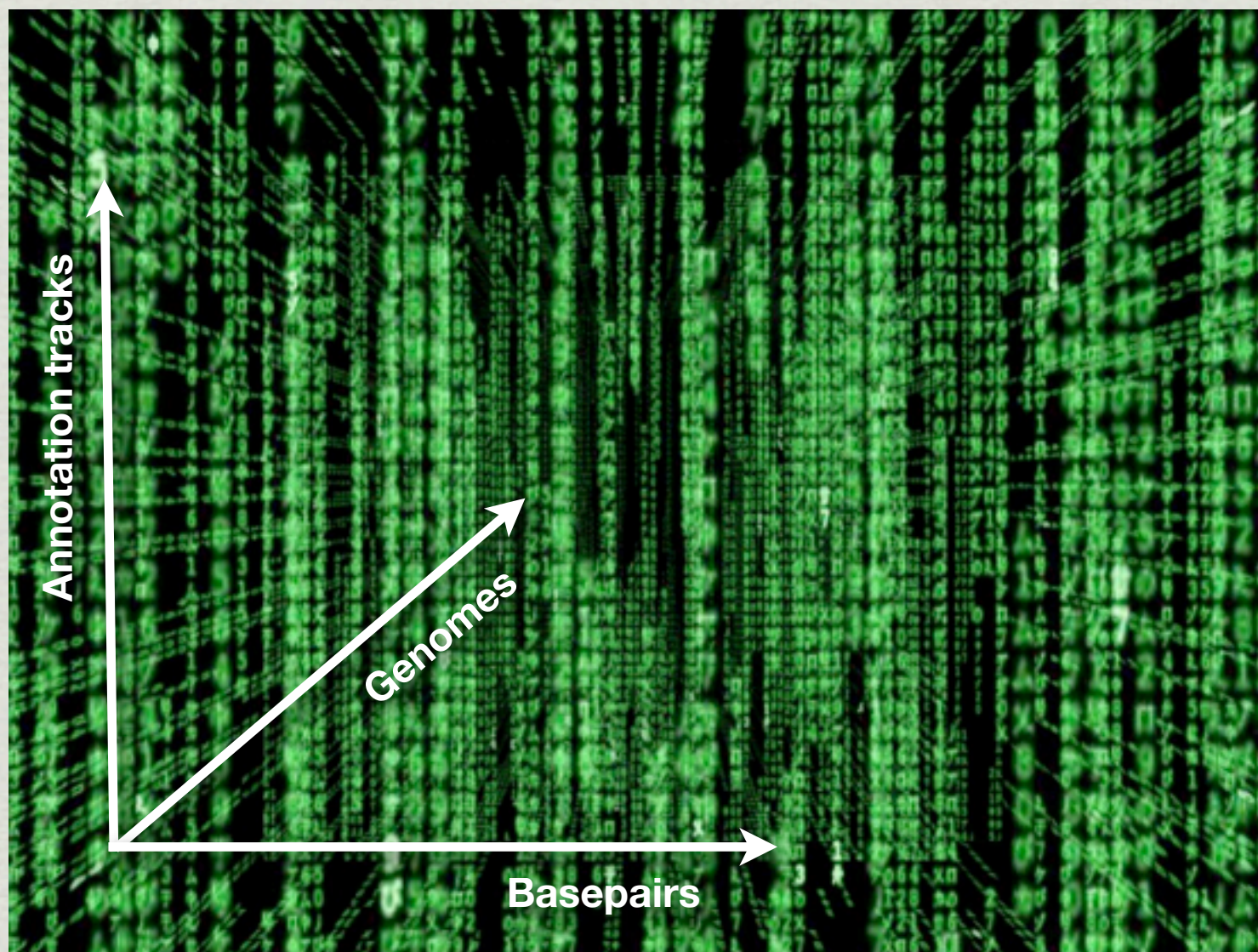
Overview

- Introduction
- Tracks and descriptive analysis
- 1. Demo
- Hypothesis testing in the real world
- 2. Demo
- Notes on hypothesis testing
- Exercises

Genomic datasets: More than just gene lists

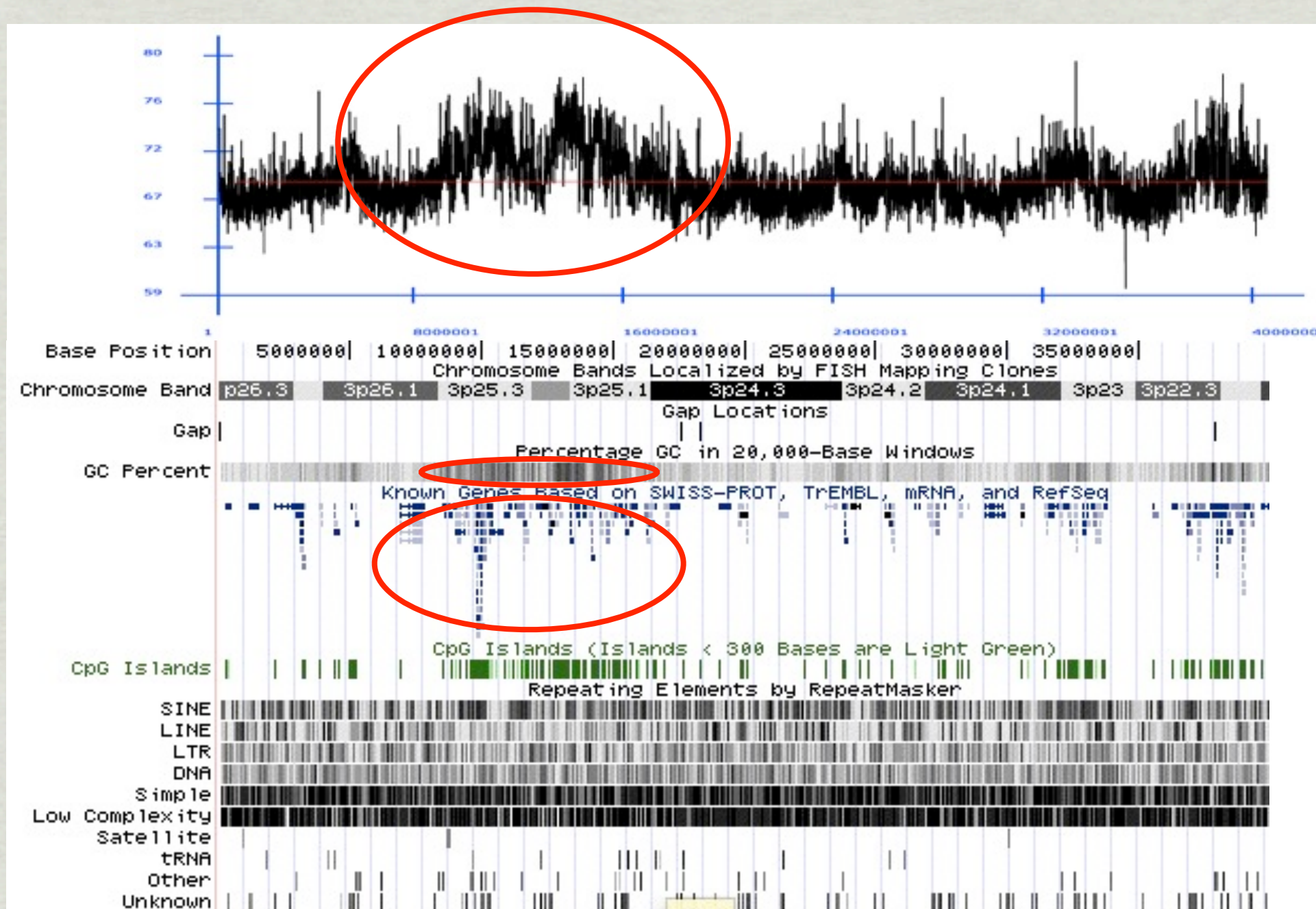


The Matrix - Reloaded again



Billions of basepairs
x 1000s of features
x 1000s of individuals
x 100s of cell types
x 100s of genomes
...

Chromosome 3p

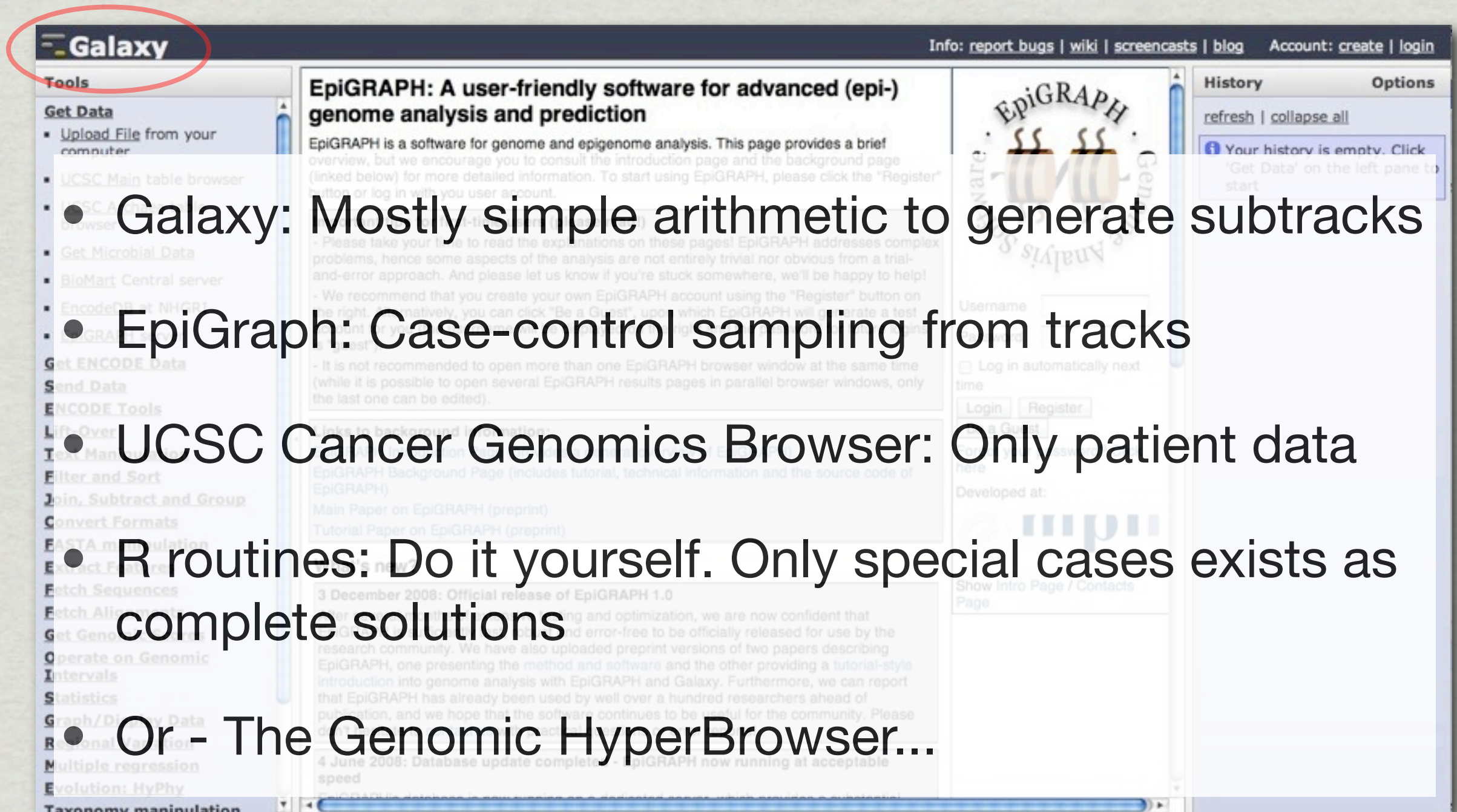


Melting temperature

GC %

Genes

Existing resources for statistical analysis



The screenshot shows the Galaxy web interface. The 'Galaxy' logo is circled in red in the top left corner. The main content area displays the 'EpiGRAPH: A user-friendly software for advanced (epi-) genome analysis and prediction' page. The left sidebar contains a 'Tools' section with various options like 'Get Data', 'Send Data', and 'Statistics'. The right sidebar shows a 'History' section with a message: 'Your history is empty. Click "Get Data" on the left pane to start.'

- Galaxy: Mostly simple arithmetic to generate subtracks
- EpiGraph: Case-control sampling from tracks
- UCSC Cancer Genomics Browser: Only patient data
- R routines: Do it yourself. Only special cases exists as complete solutions
- Or - The Genomic HyperBrowser...

Basic Features

- Massive statistical analysis of genomic data
- Not just a framework, but a guided approach to practical statistics
- Built-in statistical and biological knowledge
- Supports a variety of data types and file formats
- Includes a variety of standard and custom-made datasets (tracks)
- Includes a variety of statistical tests
- Operates on up to two genomic tracks at a time
- Optimized for large-scale, genome-wide analyses

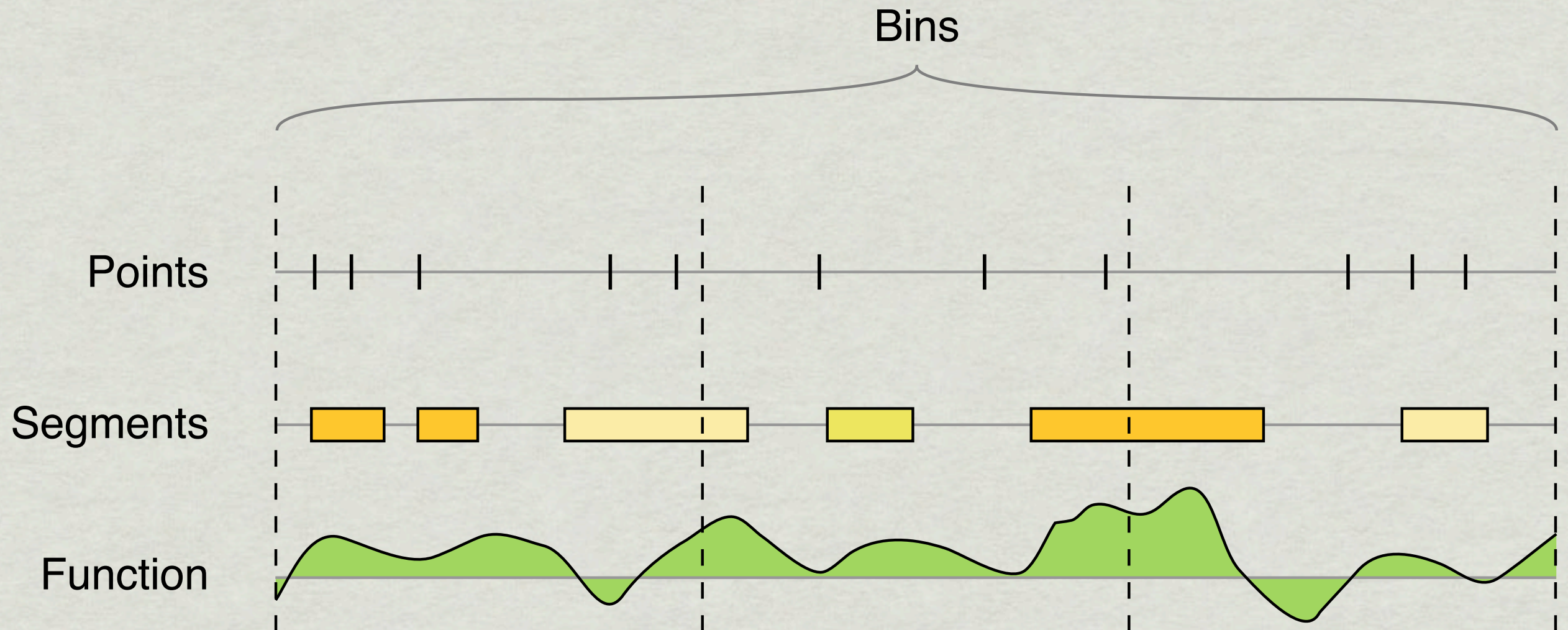
User interface

- Simple
- Questions, not statistics
- Only relevant choices at all times
- Every choice documented for verification by a statistician!

Available tracks and track types

- All standard tracks (UCSC and BioMart)
- Custom tracks of your making (e.g. expression data)
- A multitude of DNA structural tracks
 - Melting
 - Bubbles
 - Curvature
 - Bending
 - Quadruplex G
- Chromatin tracks
 - Nucleosome prediction
 - Histone methylation
 - Histone acetylation
 - SATB1 prediction
 - Lamina domains
- Binding prediction tracks
- arrayCGH
- Viral insertions
- Literature-derived tracks (182484)
- Oligonucleotide tracks (1364)
- ...

Track Formats



- 5 Main data types (UP, MP, US, MS, F)
- Mark can be a number, a character, a category, a vector
- The system can convert from segments to points when needed (start, mid or end)

Descriptive statistics

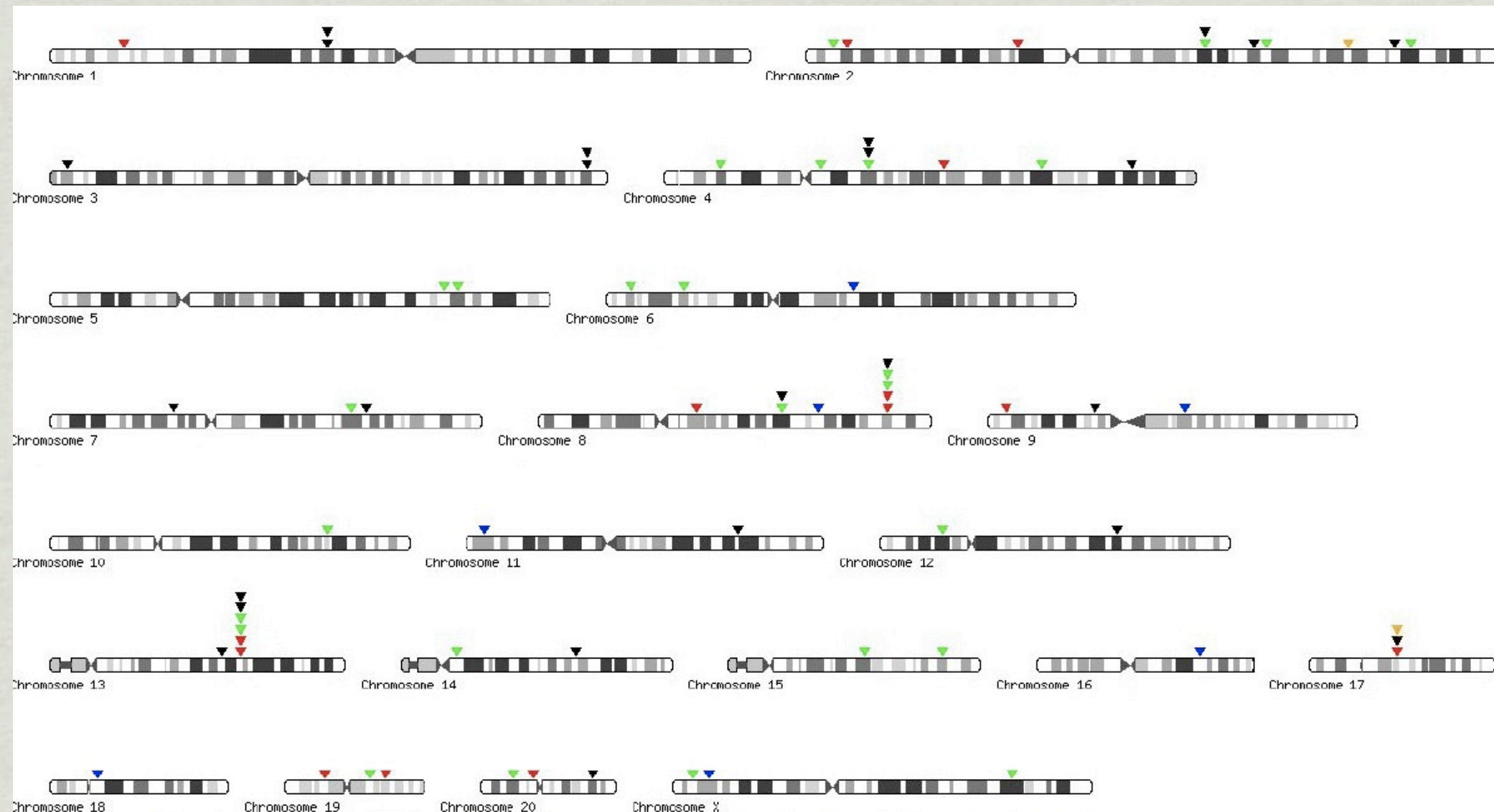
(for one or two tracks)

- Basic:
 - Counts
 - Coverage
 - Overlap
 - Enrichment
 - Sum, Mean, Variance
 - Inside vs outside versions
 - Correlations
- Distributions:
 - Lengths
 - Marks
 - Distances
- Plots:
 - Scatter plot
 - Bin-scaled plot
 - Histogram
 - Genomewide plot

1. Demo

- Dataset from Barski et. al. 2007: ChIP-seq data on histone modifications in human T-cells
- How is the frequency of nucleosomes with histone modification H3K4me3 around Transcription Start Sites of genes?

Hypothesis testing in the real world



Are HPV close to genes?

Hypothesis testing in the real world

**Now,
what do you do?**

First get data

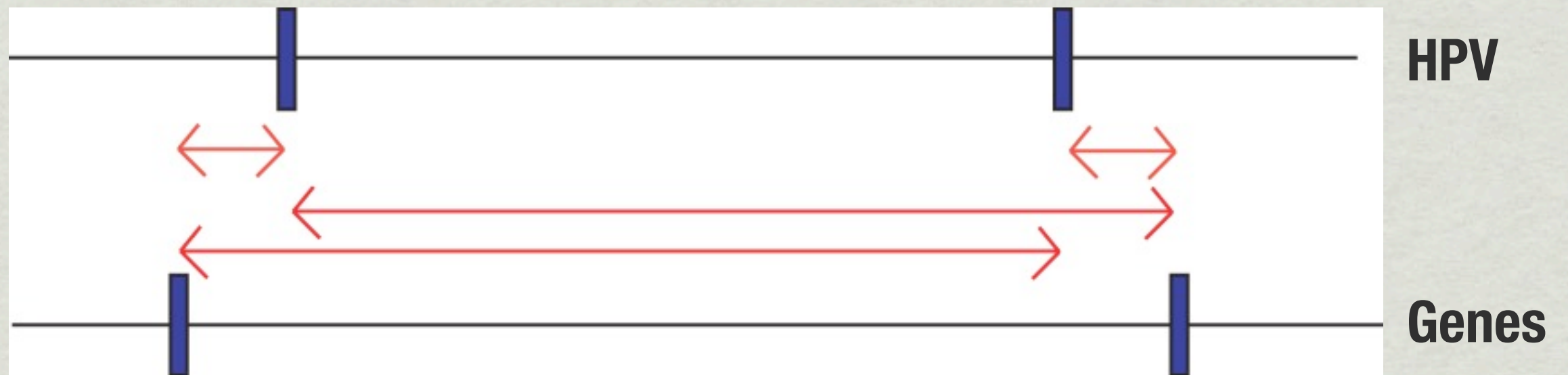
- HPV in text-file..
- Download genes..

Now what?

Specify hypothesis



Specify hypothesis



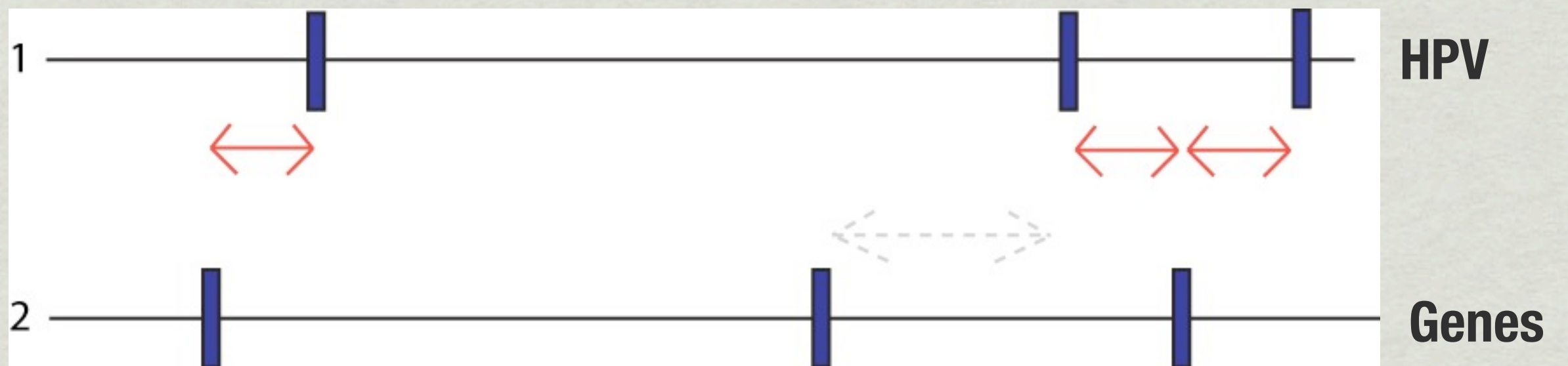
- Which distances?

Specify hypothesis



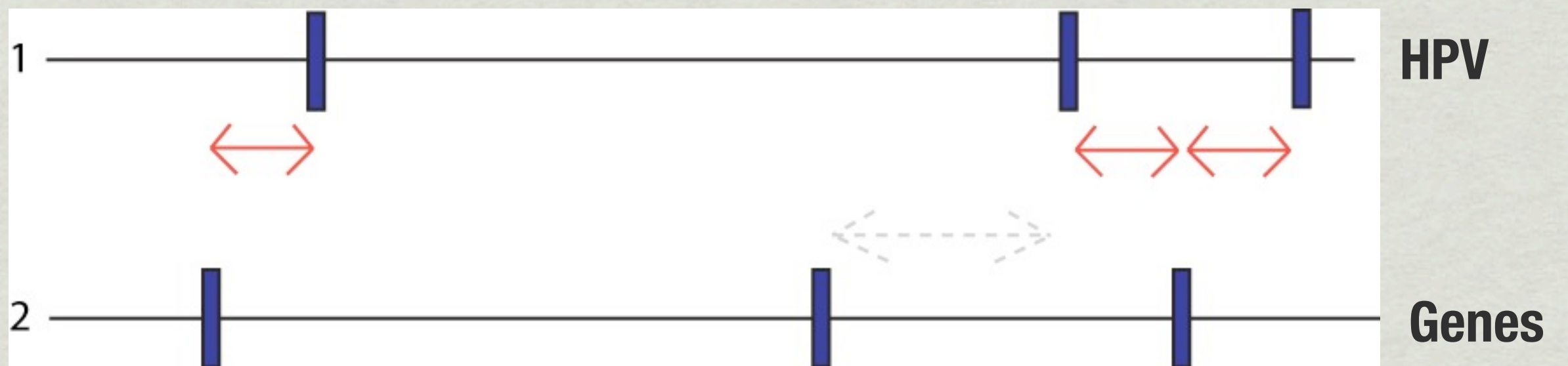
- Which distances: Only shortest

Specify hypothesis



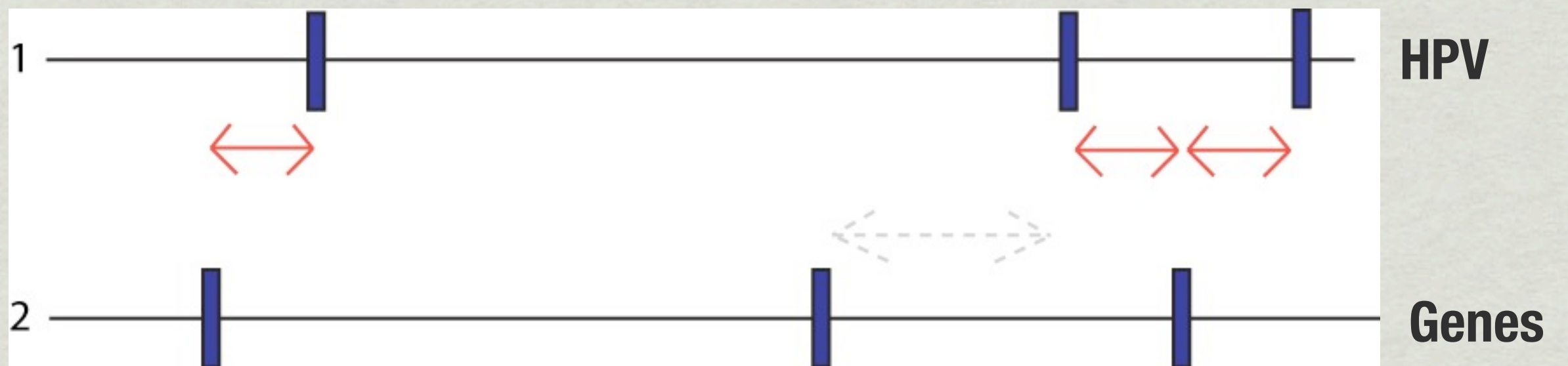
- Which distances: Only shortest, from 1 to 2

Specify hypothesis



- Significantly close?

Specify hypothesis

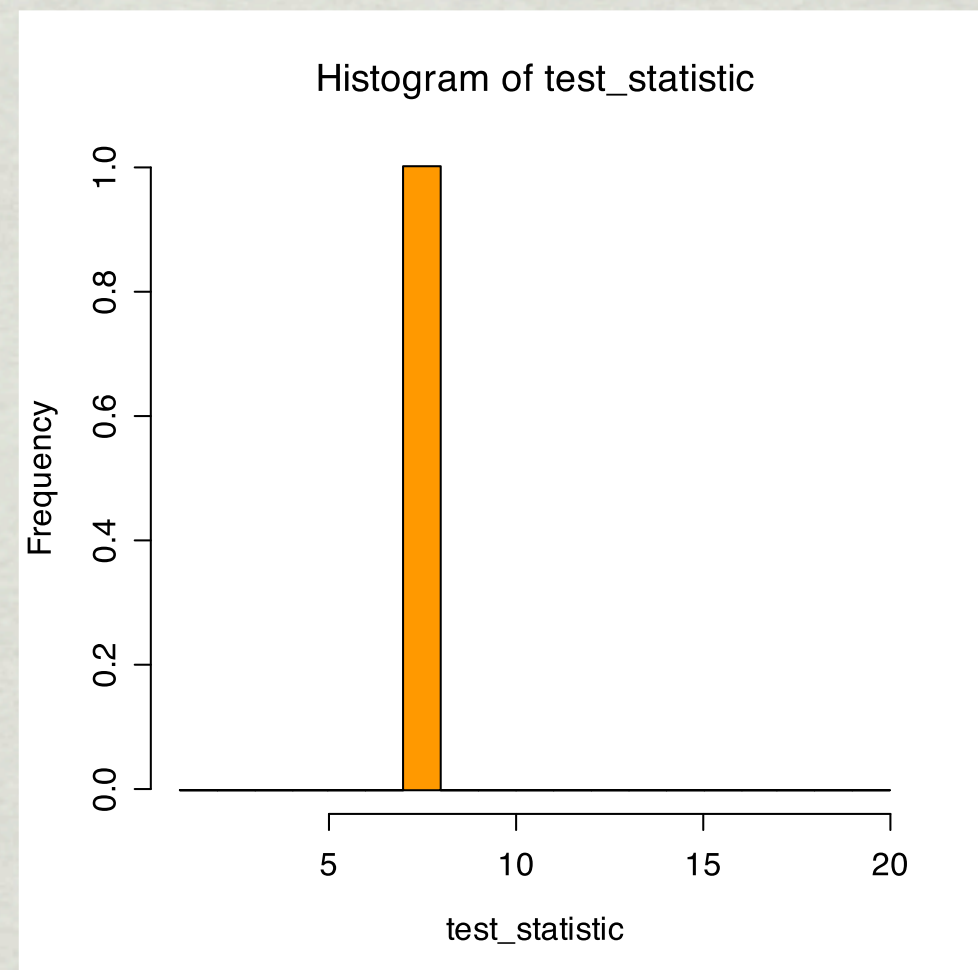


- Significantly close? Use Monte Carlo..

Detour:

What is Monte Carlo?

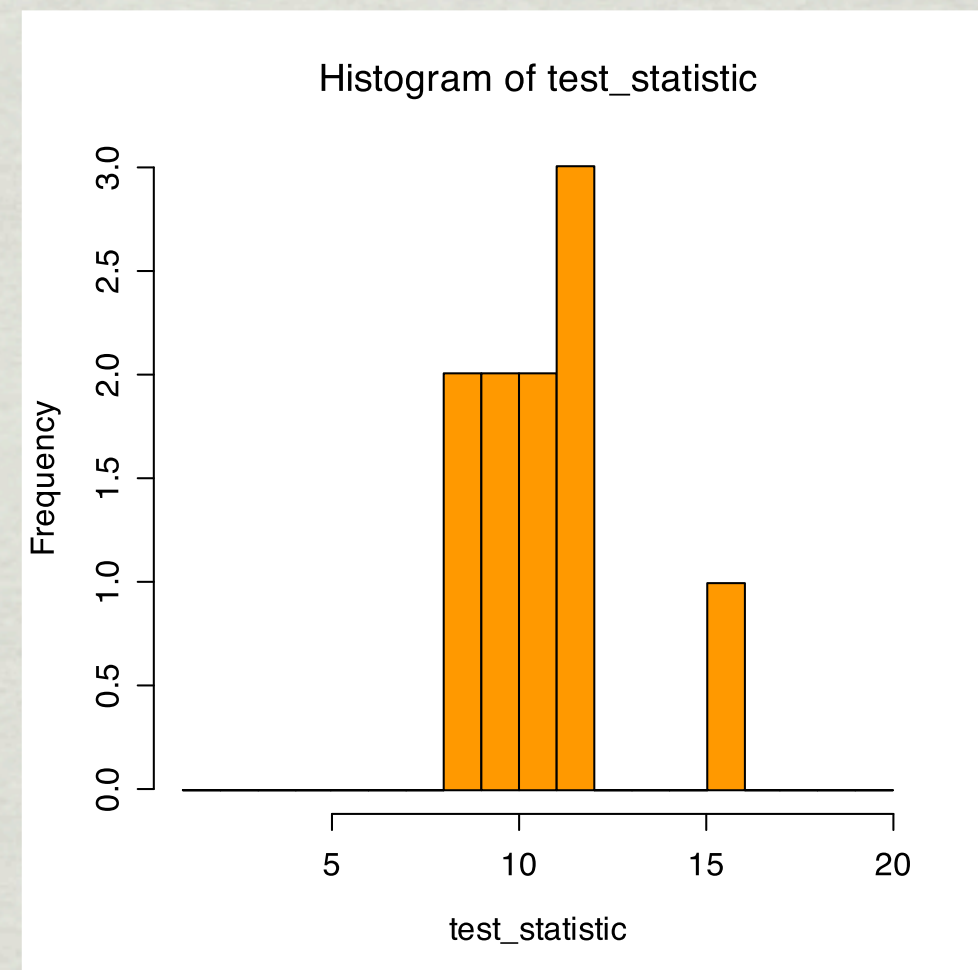
- Randomize test statistic



Detour:

What is Monte Carlo?

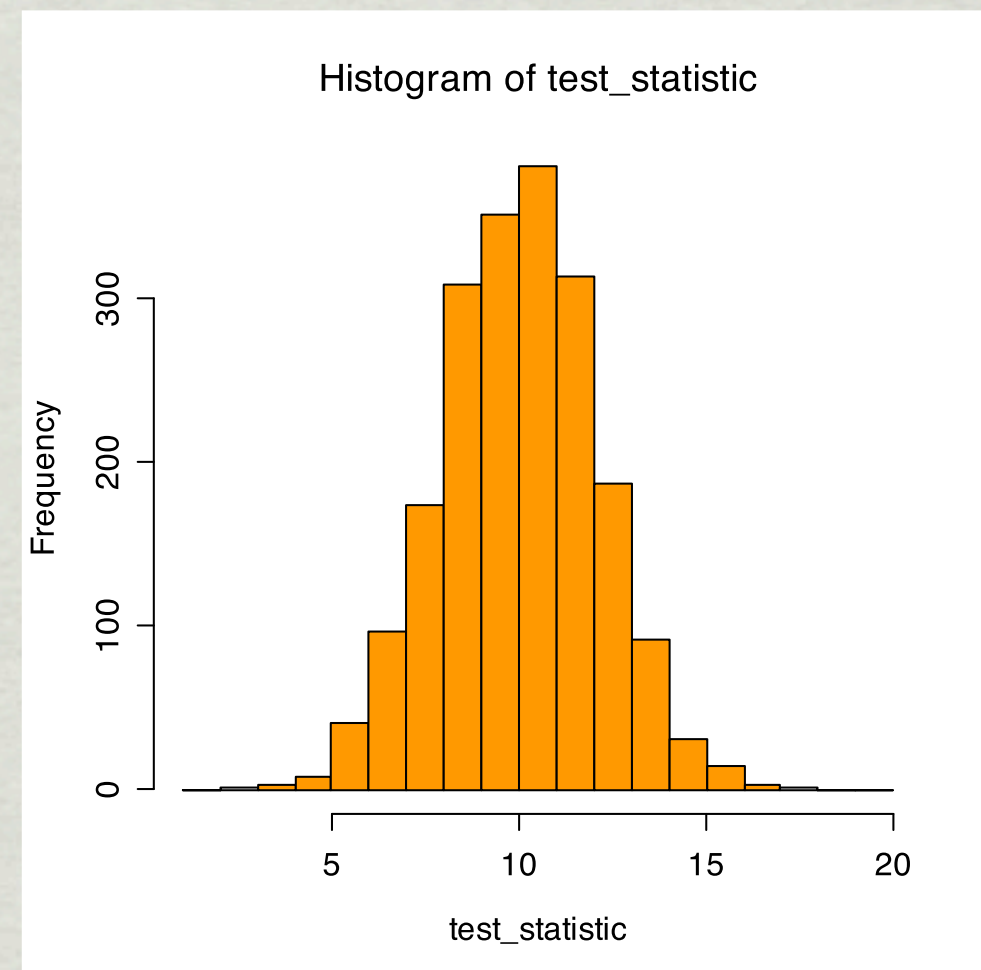
- Randomize test statistic
- Repeat a number of times



Detour:

What is Monte Carlo?

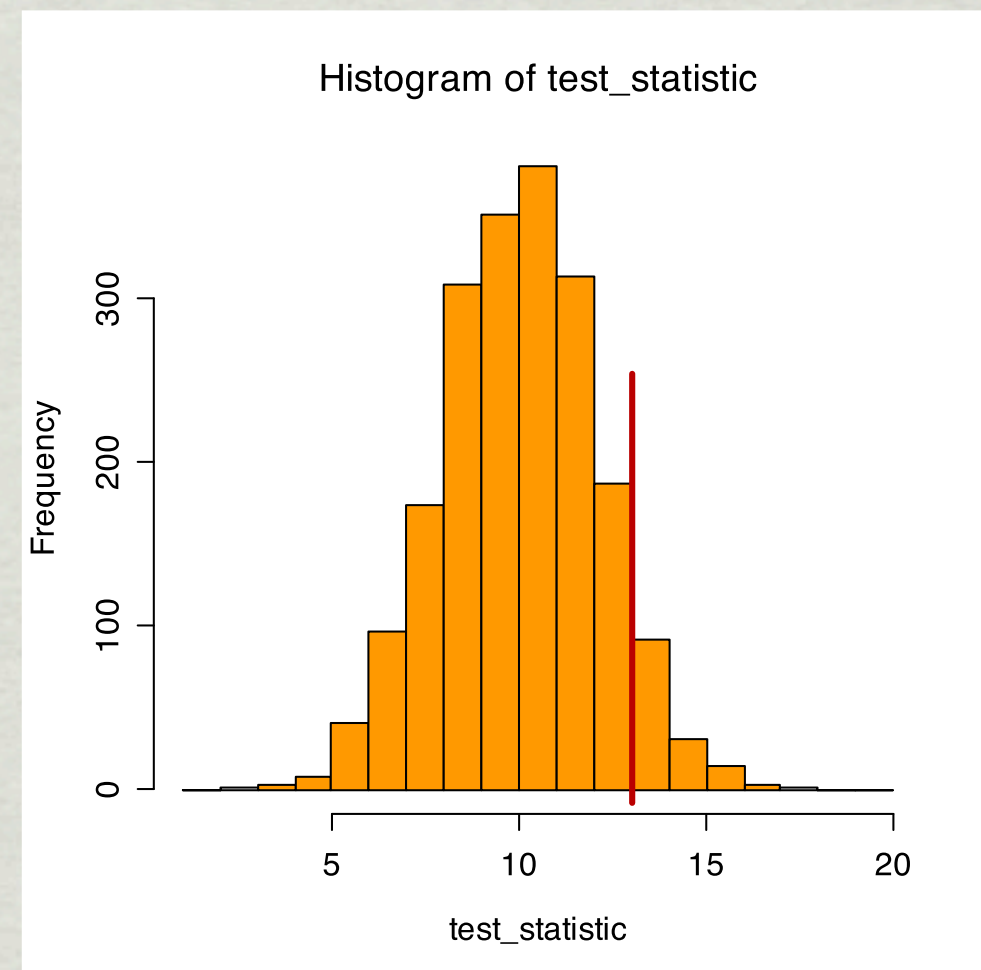
- Randomize test statistic
- Repeat a number of times
- Build histogram



Detour:

What is Monte Carlo?

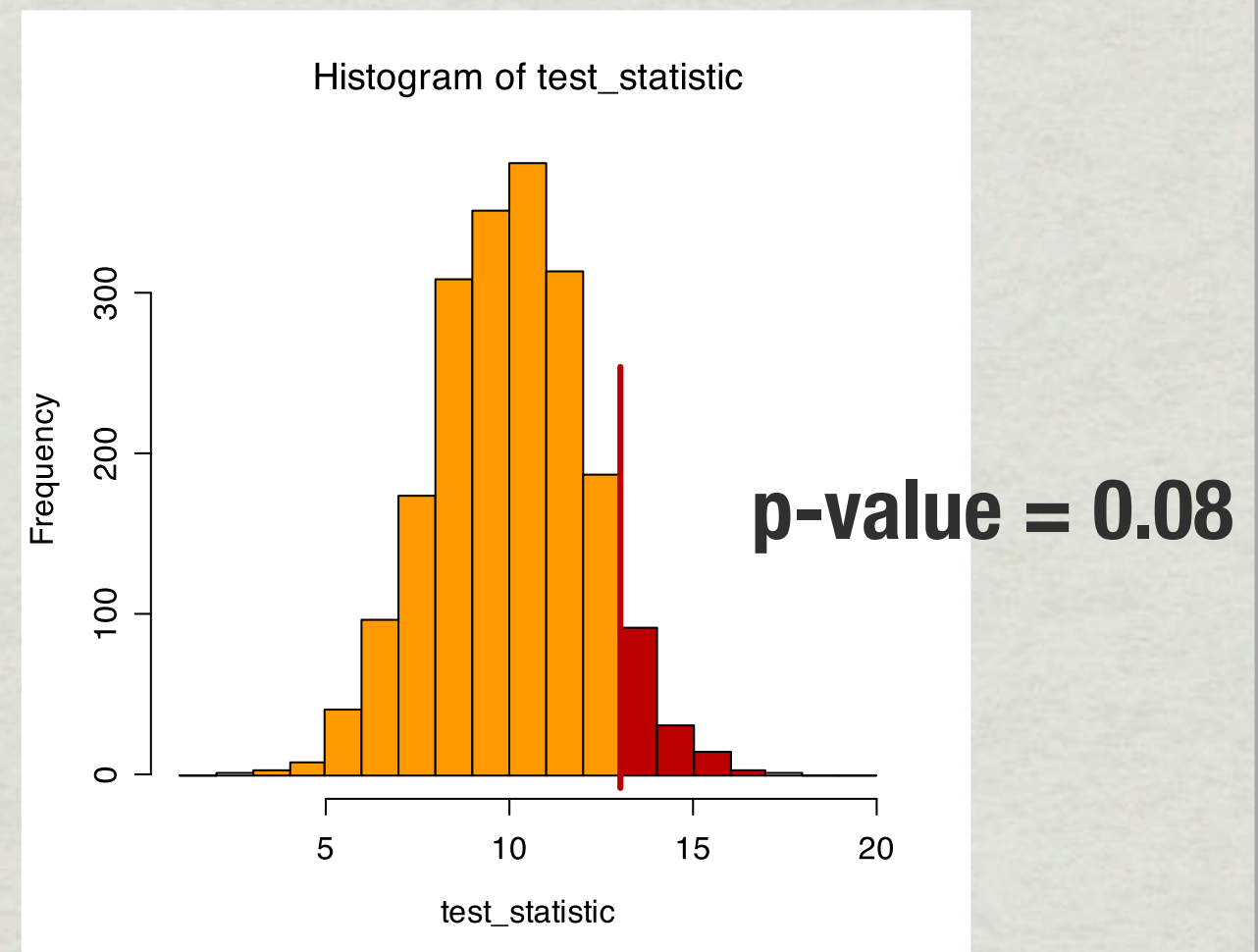
- Randomize test statistic
- Repeat a number of times
- Build histogram
- Compare with observed value



Detour:

What is Monte Carlo?

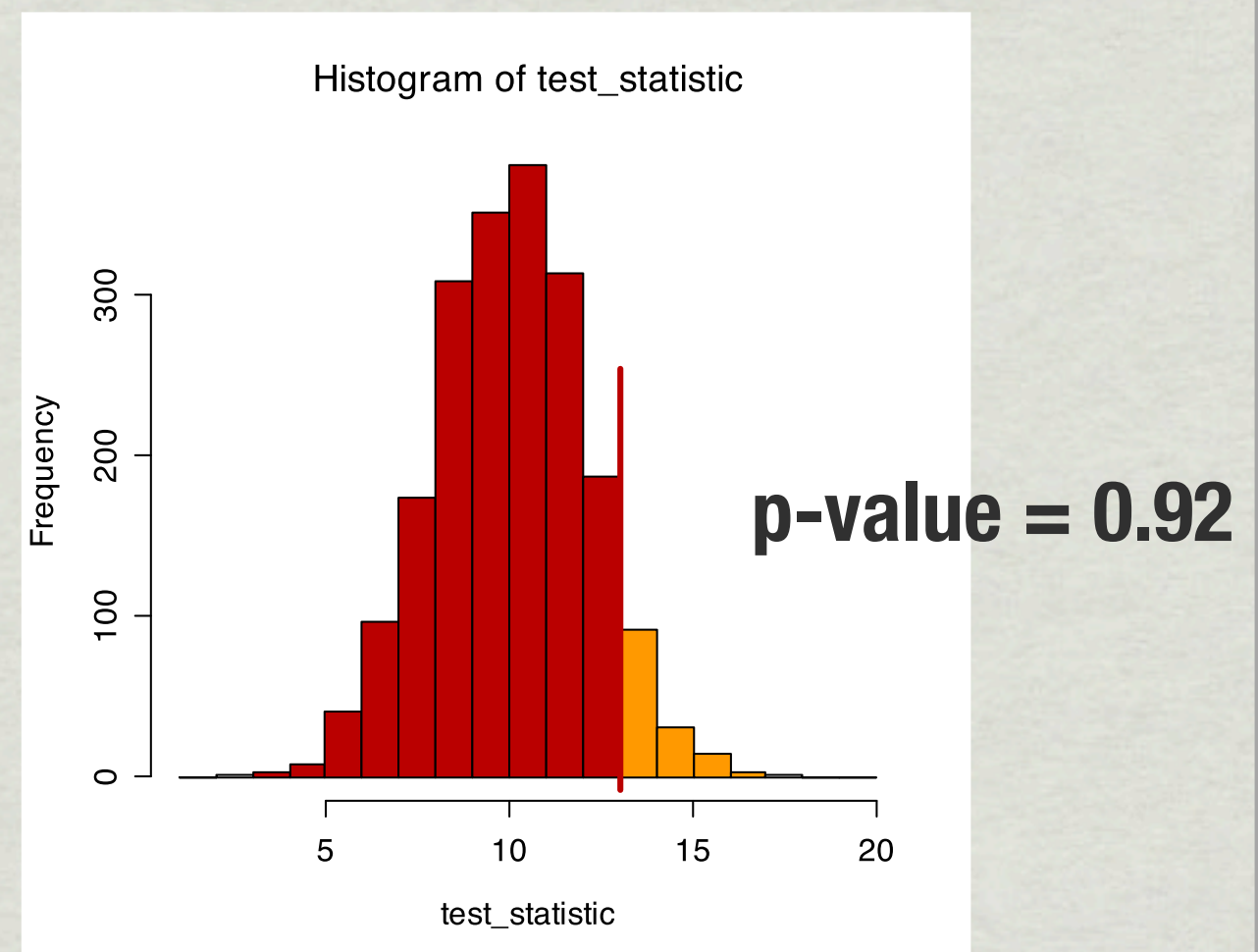
- Randomize test statistic
- Repeat a number of times
- Build histogram
- Compare with observed value
- p-value = Area to the right (right-tailed) when total area sums to 1
(< 0.05 is usually significant)



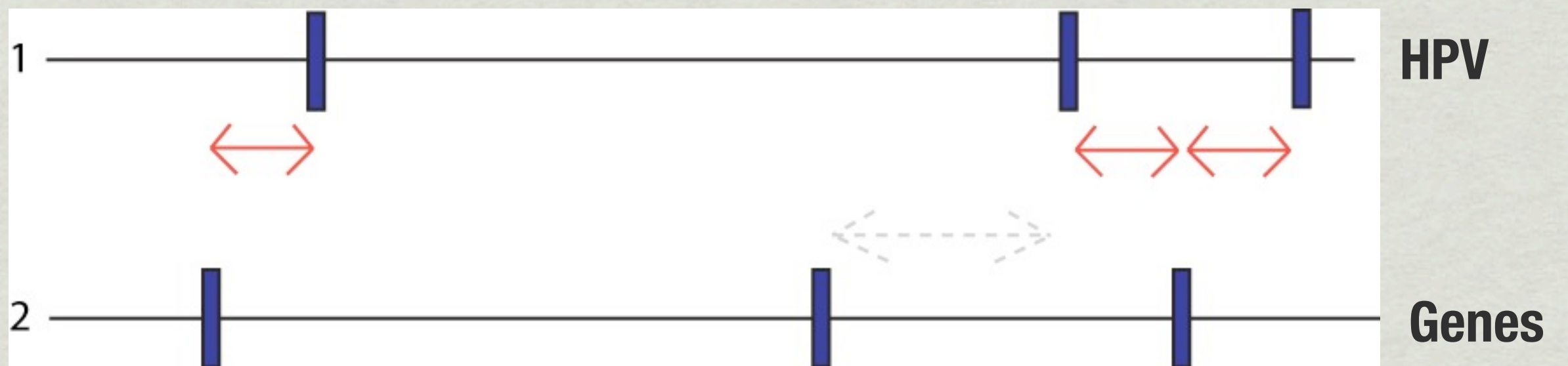
Detour:

What is Monte Carlo?

- Randomize test statistic
- Repeat a number of times
- Build histogram
- Compare with observed value
- p-value = Area to the right (right-tailed) when total area sums to 1 (< 0.05 is usually significant)
- Can also be left- or two-tailed

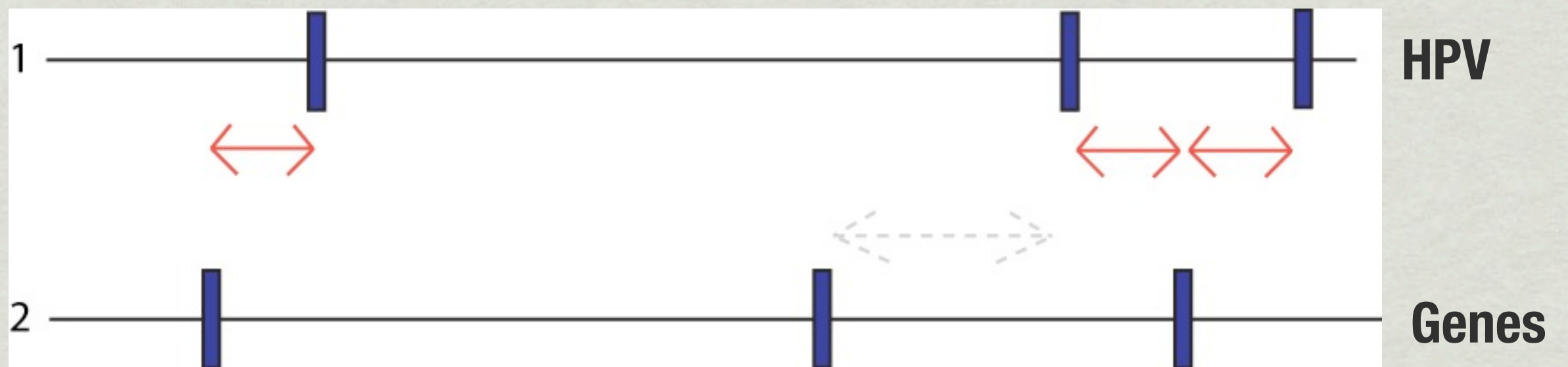


Specify hypothesis



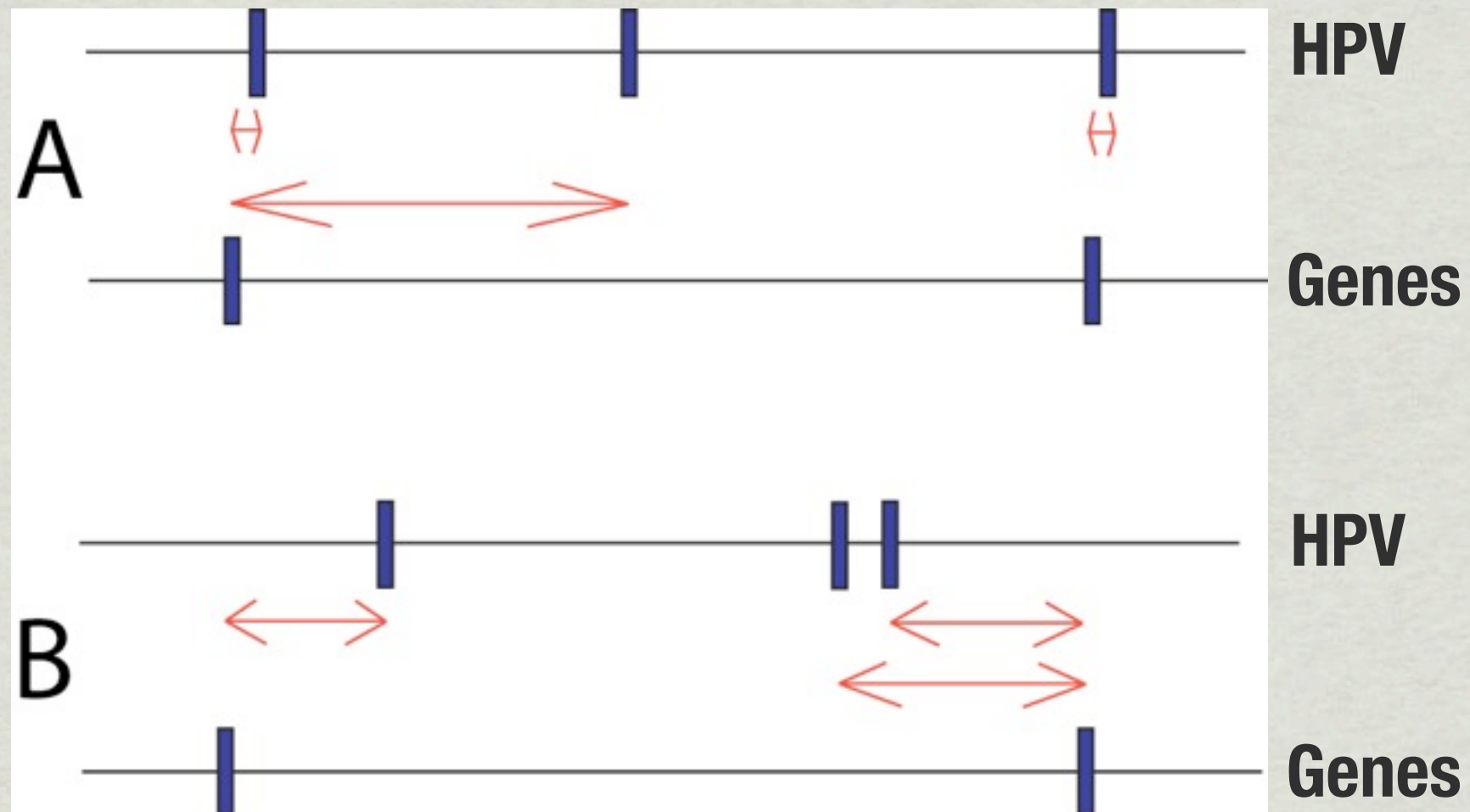
- Significantly close? Use Monte Carlo..
But many dists!

Specify hypothesis



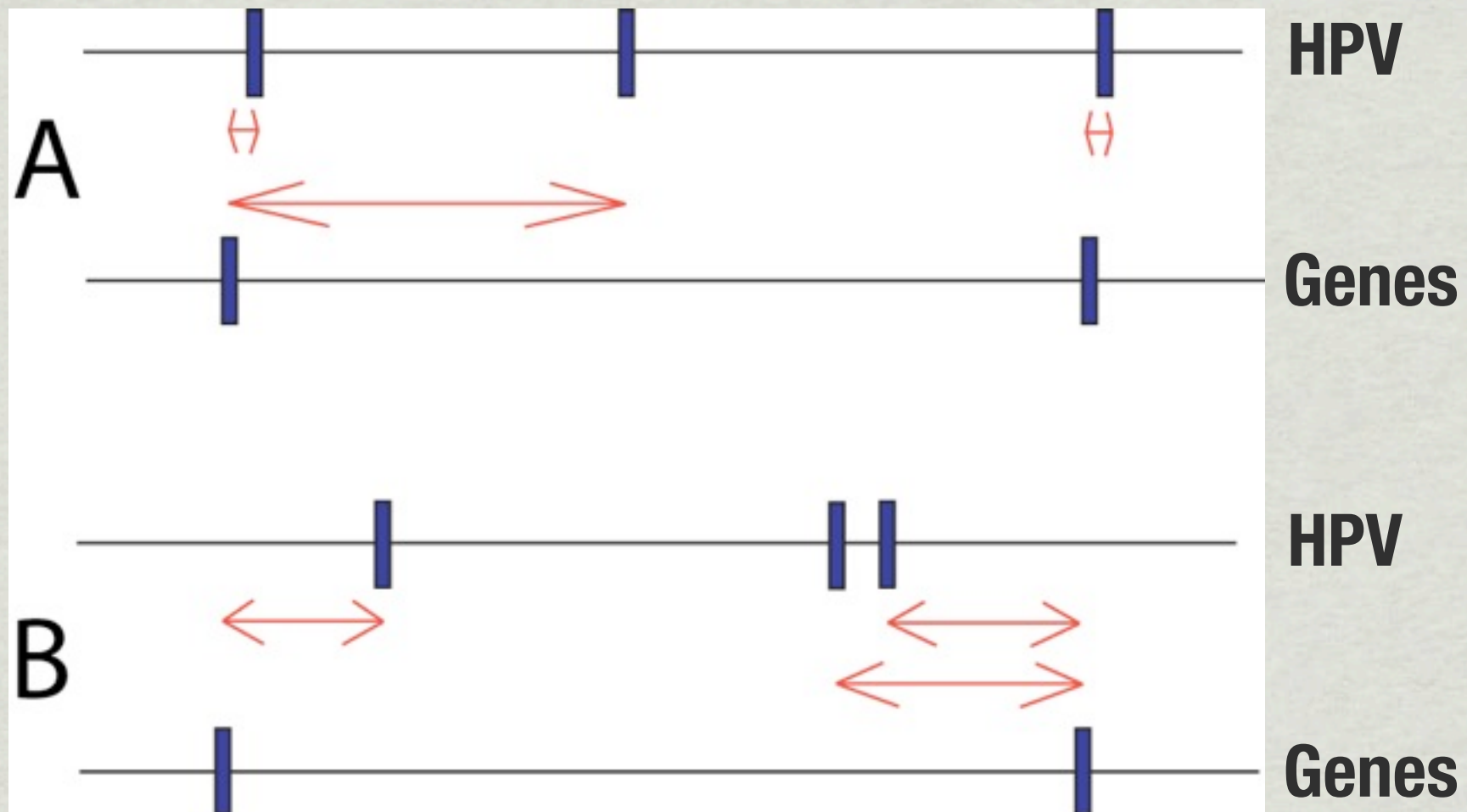
- Significantly close? Use Monte Carlo..
Average of dists?!

Specify hypothesis



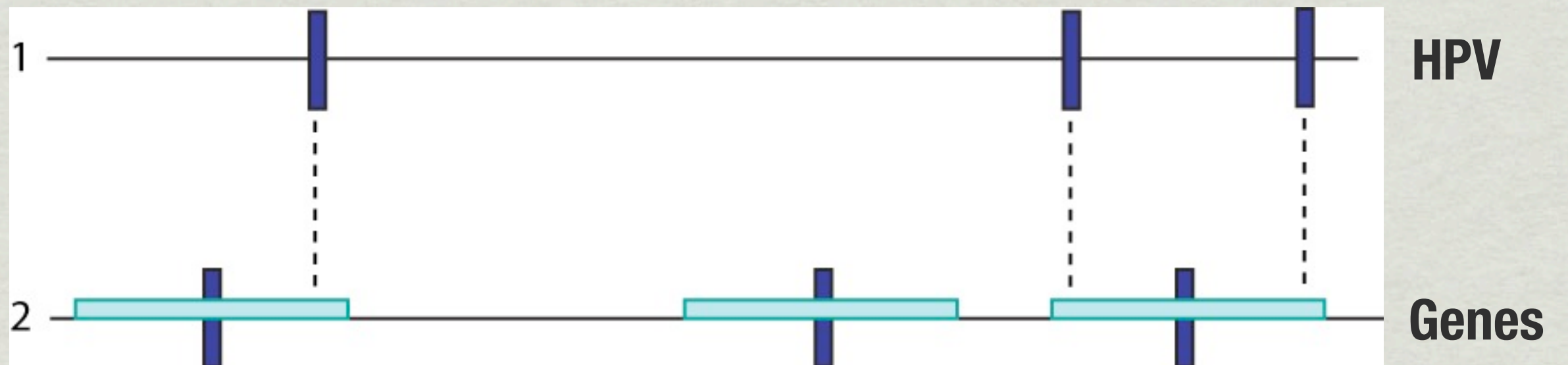
- Significantly close? Use Monte Carlo..
Average of dists?!

Specify hypothesis



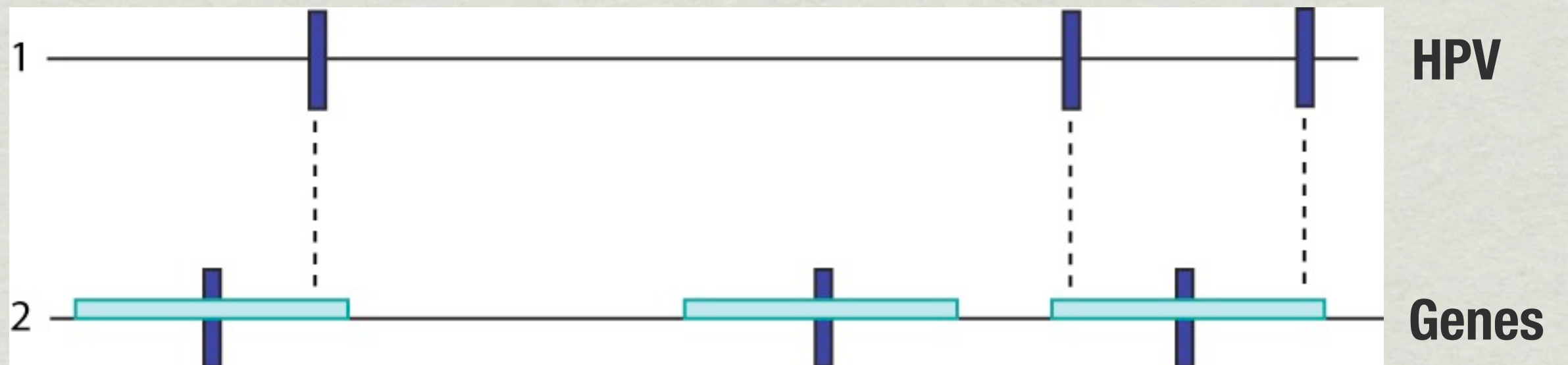
- Significantly close? Use Monte Carlo..
Geometric avg?!

Or, for something (entirely) different..



- How many HPV-sites in regions around genes?
(or HPV in exon upstreams?)

And significance..



- Trivial with Monte Carlo..
- Can it be found analytically?
 - Binomial distribution!

Almost there..

- Must first double-check with (another) statistician..!
- And then - how to implement?

Implementation

- Parse data
- Take upstreams
- Determine if points inside any segments
- Binomial test
- If large data:
Split, intermediate computation, combine

Still not there!

- Must check for bugs!
 - Any silly bugs?
 - Formats understood correctly?
 - Remembered strand?
- Double-check which points declared inside and outside..

Finally..

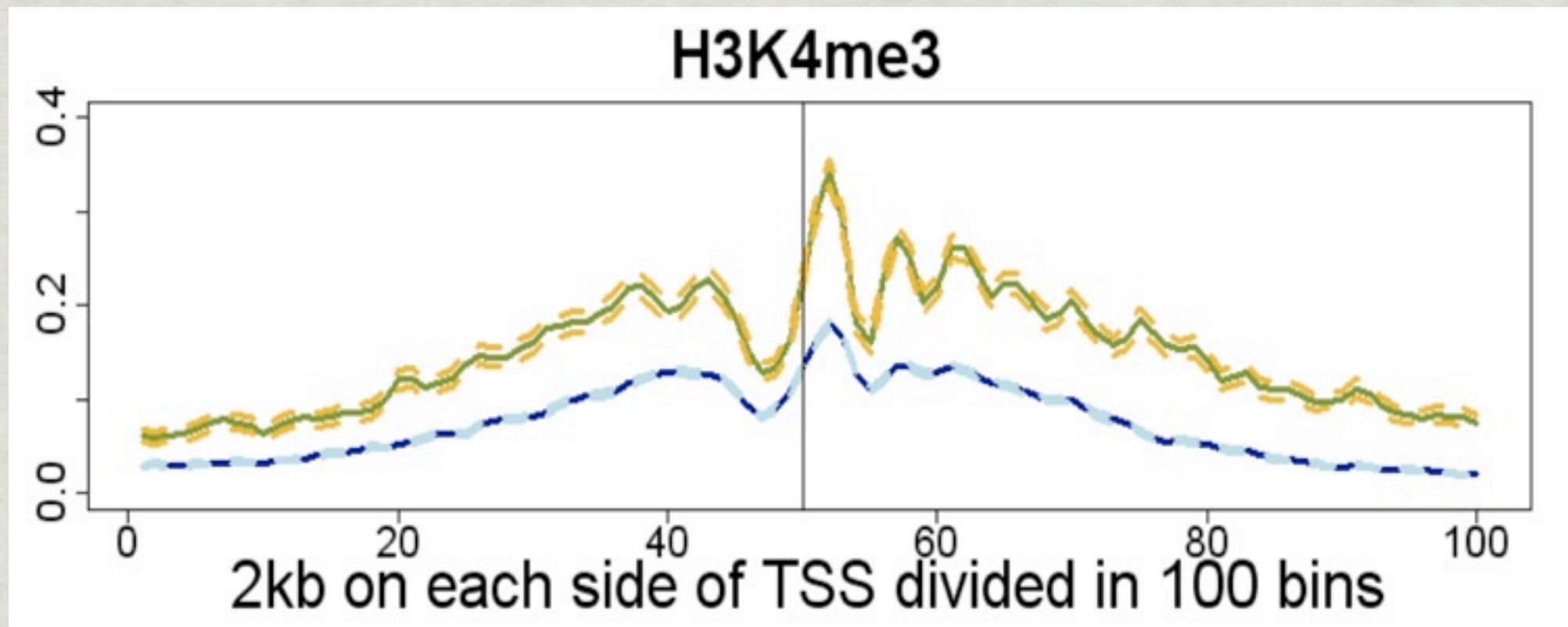
- We can now dump the code and never have to use or look at it anymore (hopefully..)

Or...

We could use the Genomic HyperBrowser

(2. Demo)

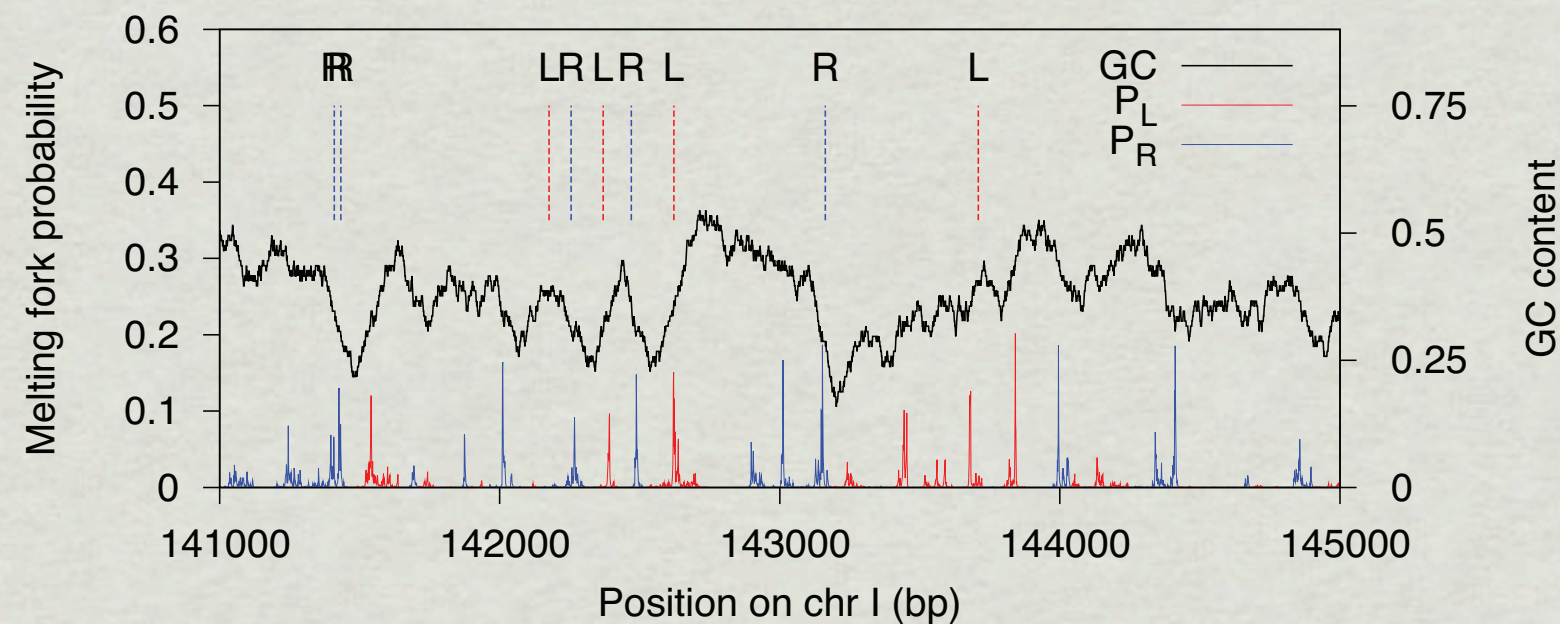
Example 2: Microarray vs. ChIP-seq



- Do histone modification H3K4me3 contribute to expression, more than expected by chance? How much?
- Recreate result from Barski et. al. 2007, but using hypothesis testing

(Example: Halfdan Rydbeck)

Example 3: Melting forks vs. exons

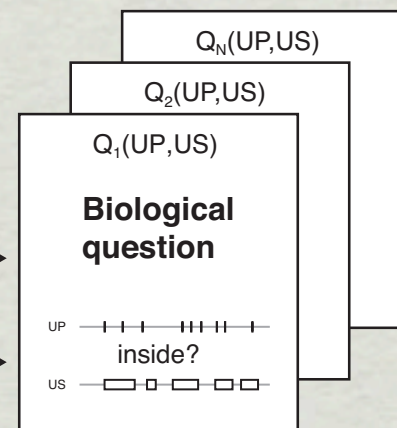
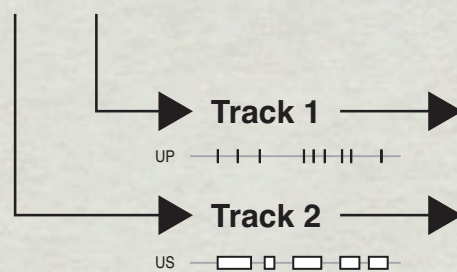
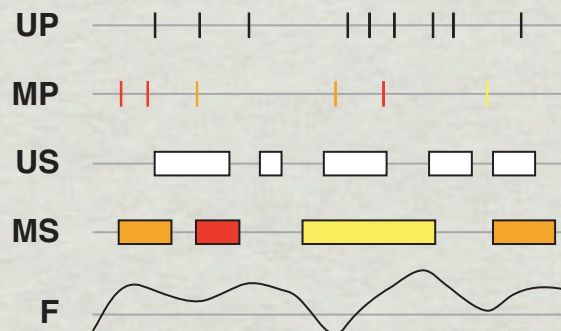


- Are the melting fork probabilities higher at exon ends than expected by chance?
- What about the GC content?

(Example: Eivind Tøstesen)

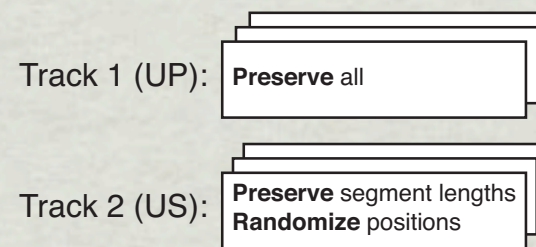
The HyperBrowser approach

Data

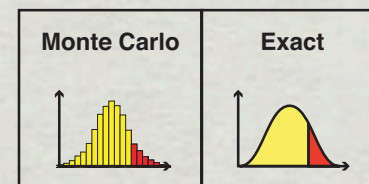


Analysis

Null hypothesis

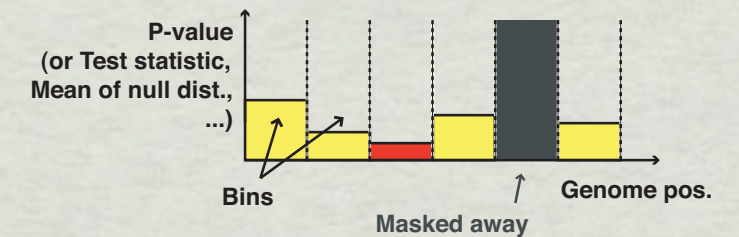
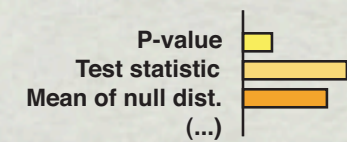


Statistical test



Results

Global results



Local results

The Null Hypothesis

- The null hypothesis (or null model) is defined by a distribution (Monte Carlo or analytical)
- Is determined by Preservation and Randomization
- Choices of these should reflect biological knowledge. Very hard. Should in principle model 3-4 billion years of the random process that is called evolution.
- The alternative hypothesis is usually one of “less than, more than, different”

Local results

- A separate test is carried out for every bin
- We have a multiple test problem: On average, 1/20 of the bins may give significant results, even though nothing is significant
- FDR (False Discovery Rate) is used. At default 10% of the significant bins are accepted as false positives
- FDR-values in the tables are the proportion of false positives we must accept if we hold the results for true
- Local significance may show a real difference in the data, but it is not immediately clear whether this difference in data actually corresponds to a biological phenomena
- Significance could be because of few data points

Limitations

- Only cis-type questions possible at this time (but the future is 3D!)
- Only one genome at a time
- Hardware (especially for large simulations)
- Need for much statistical effort to address all relevant questions

Planned extensions

- More tracks and track types
- More statistics
- New genomes
- Cell type specificity
- Meta-analysis
- Better and more graphical output
- ...and much more

Publications

(Currently under review..)

The team



Support



Web-site

Official site:

<http://hyperbrowser.uio.no>

For exercises today:

<http://insilico.titan.uio.no:8099>