Genome analysis and reproducibility

What you will learn

- Genome analysis
- Statistical testing
- Reproducibility

The form of this session

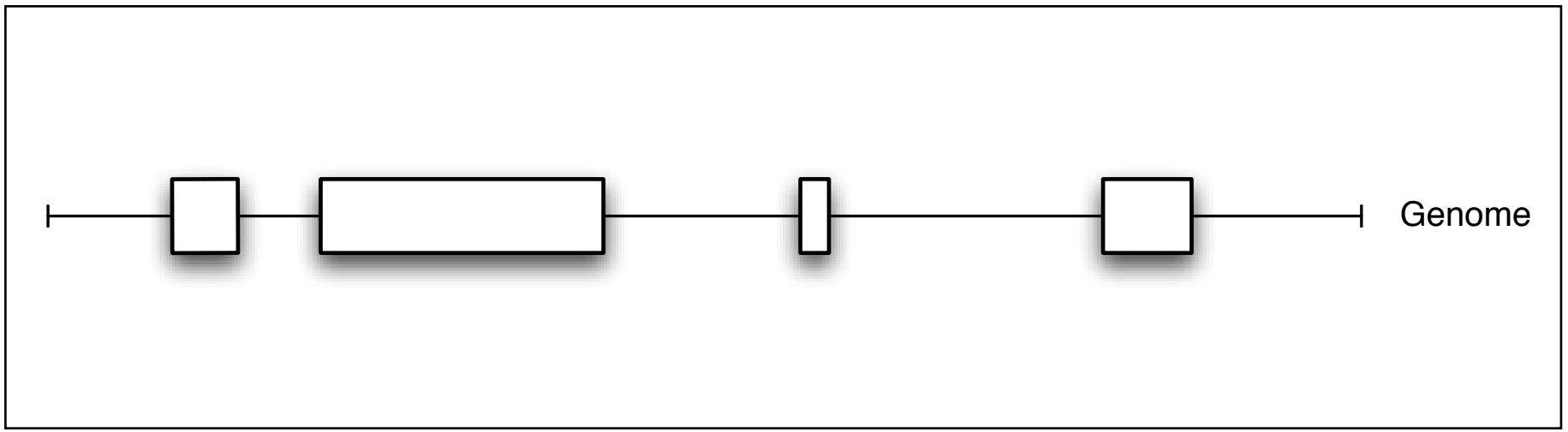
- I briefly introduce a topic
- You do a short hands-on
- I explain the topic in more detail
- ... we repeat this for a sequence of increasingly advanced/detailed topics
- And, I won't tell you exactly what to do in hands-ons

Biological cases, but not depth

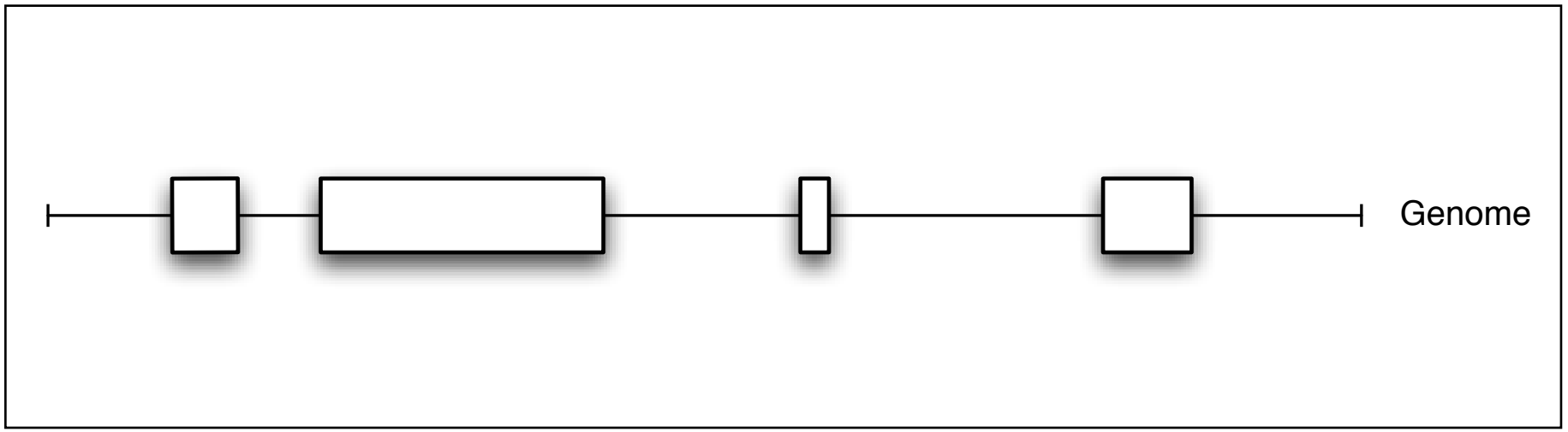
- We will use biological cases, but not focus on biological interpretation:
  - You are the experts in biology, not me
  - My message is the methodology and its generic (statistical) interpretations

What are genes?

This! :



What are genes?



Reference genome acts like coordinate system for genomic data

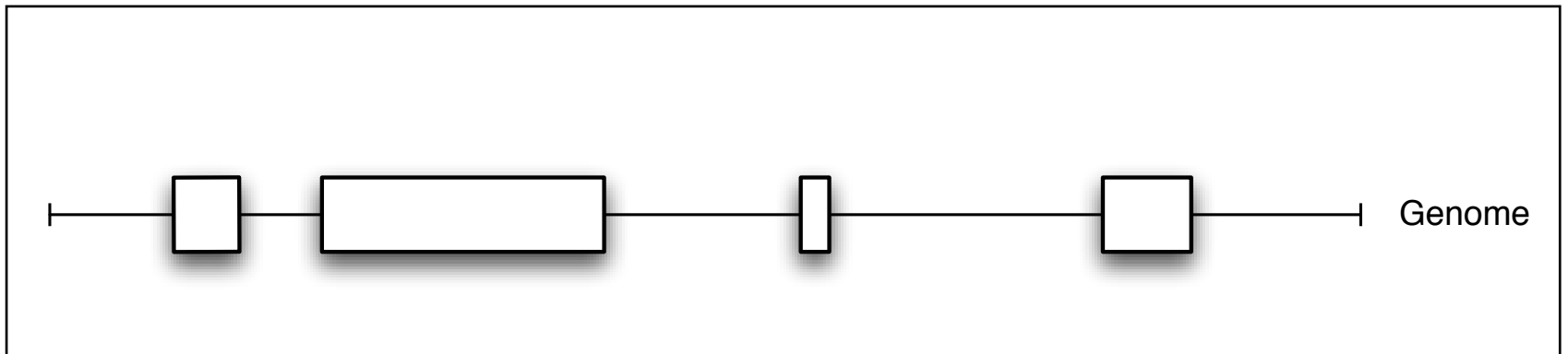
```
chr21 10079666 10120808 NM_001187
chr21 13332357 13412442 NR_026916
chr21 13700575 13700652 NR_036164
chr21 13904368 13935777 NM_174981
chr21 14137324 14142556 NR_026755
```

The UCSC genome browser

- You have previously in the course seen how genomic tracks can be inspected at UCSC

Examples of genomic data

- Genes locations, gene expression
- Repeating elements
- Evolutionary conserved regions
- DNA methylation, histone modifications
- SNPs, copy-number variations
- Disease-associated regions



So, what about analysis?

Example analyses

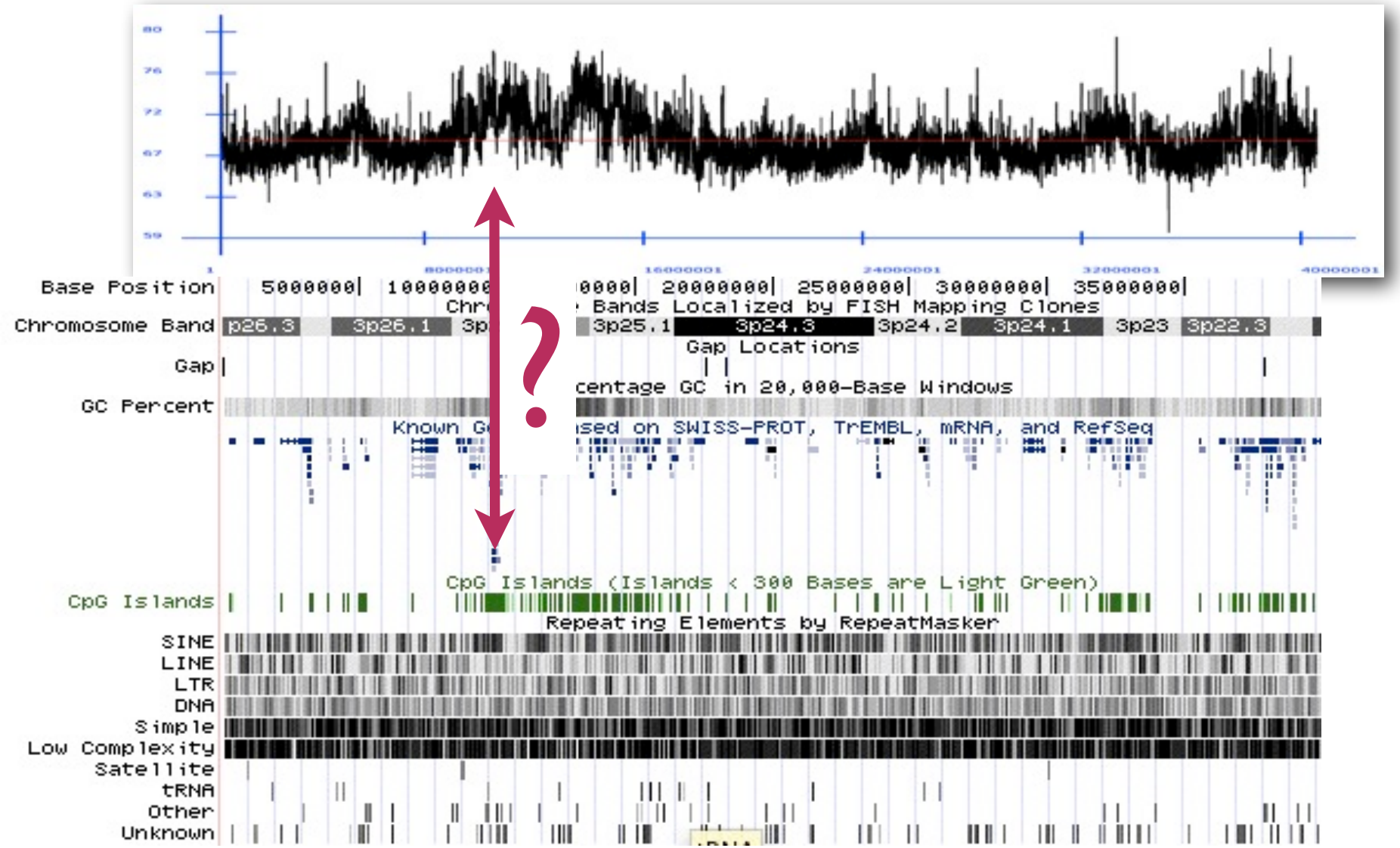
- Distinct methylation for tissue-specific genes? (Genome Res. 2010 20: 1493-1502)
- Methylation patterns in embryonic cells (PNAS 2010 107:10783–10790)
- A relation between methylation patterns and repeating elements? (Genome Res. 2009 19: 221-233)
- Methylation and active genes at T-Cell G0->G1 (Genome Res. 2009 19: 1325-1337)

Example analyses (cont.)

- Cooperative histone modifications?
  (Nat Genet 2008 40:897-903)
- Fragile sites, breakpoints and repeats?
  (Genome Biology 2006 7:R115)
- Copy number variation, repeats, duplications and genes? (Genome Res. 2009 19: 1682-1690)
- Virus integration vs genes, CpG, GC-content
  (Journal of Virology 2007 6731–6741)

So, how to do analysis?

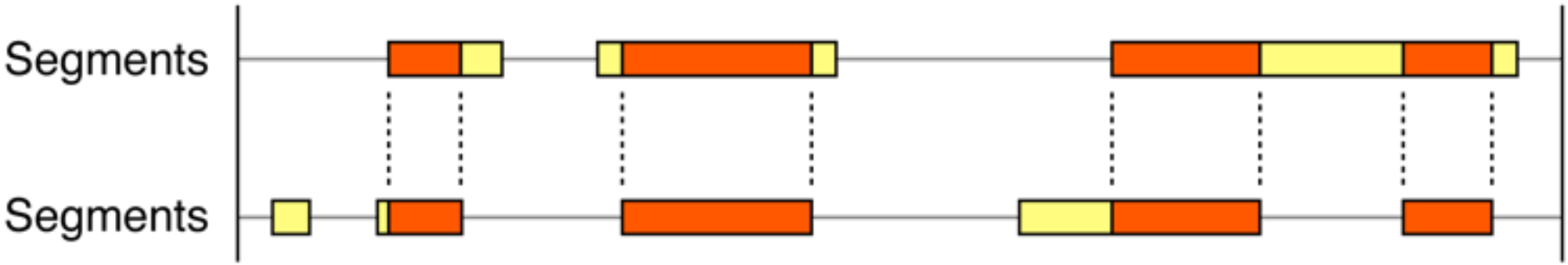
This can't be it?!



Co-location of genomic features

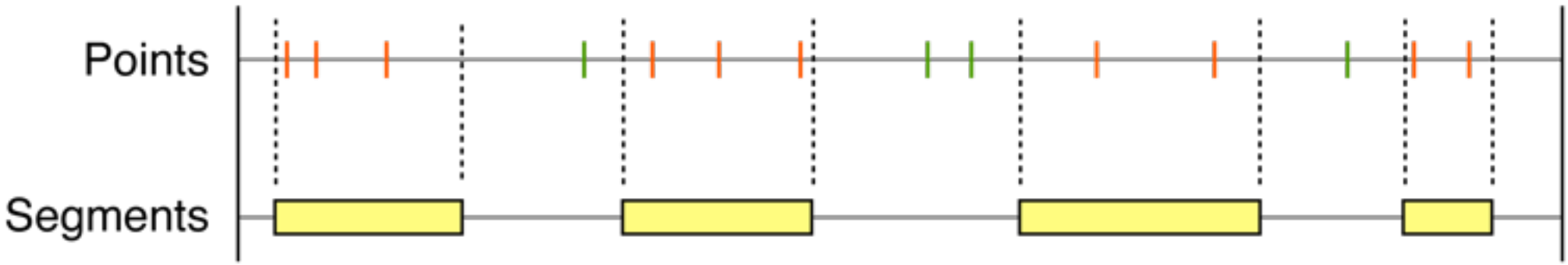
- Common question:
  do genomic feature X and Y occur
  (more than expected)
  at the same locations in the genome?
- Used to discover novel relations
- May indicate a direct causal relation, or hint to indirect association.

How does this look at the drawing board?



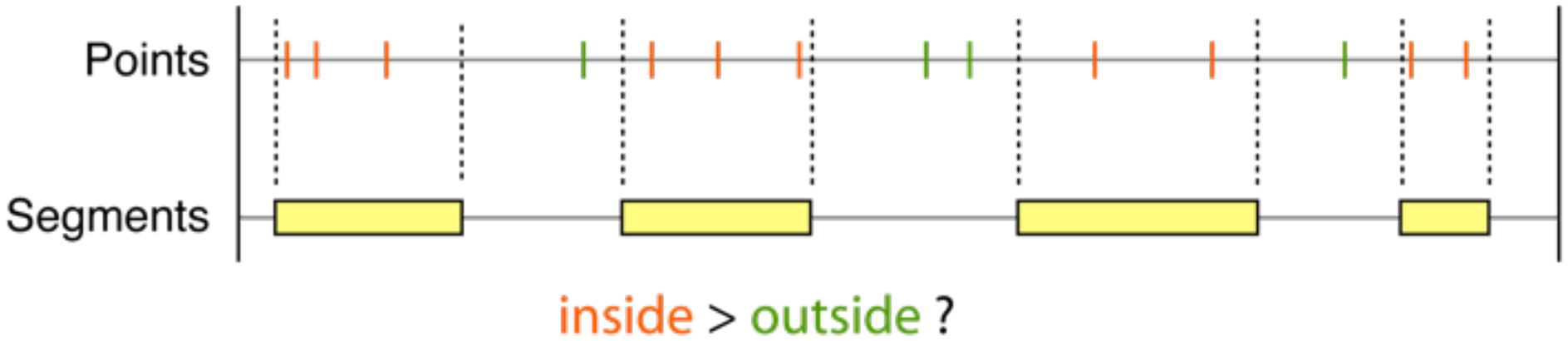
overlap > expected?

How does this look at the drawing board?



inside > outside ?

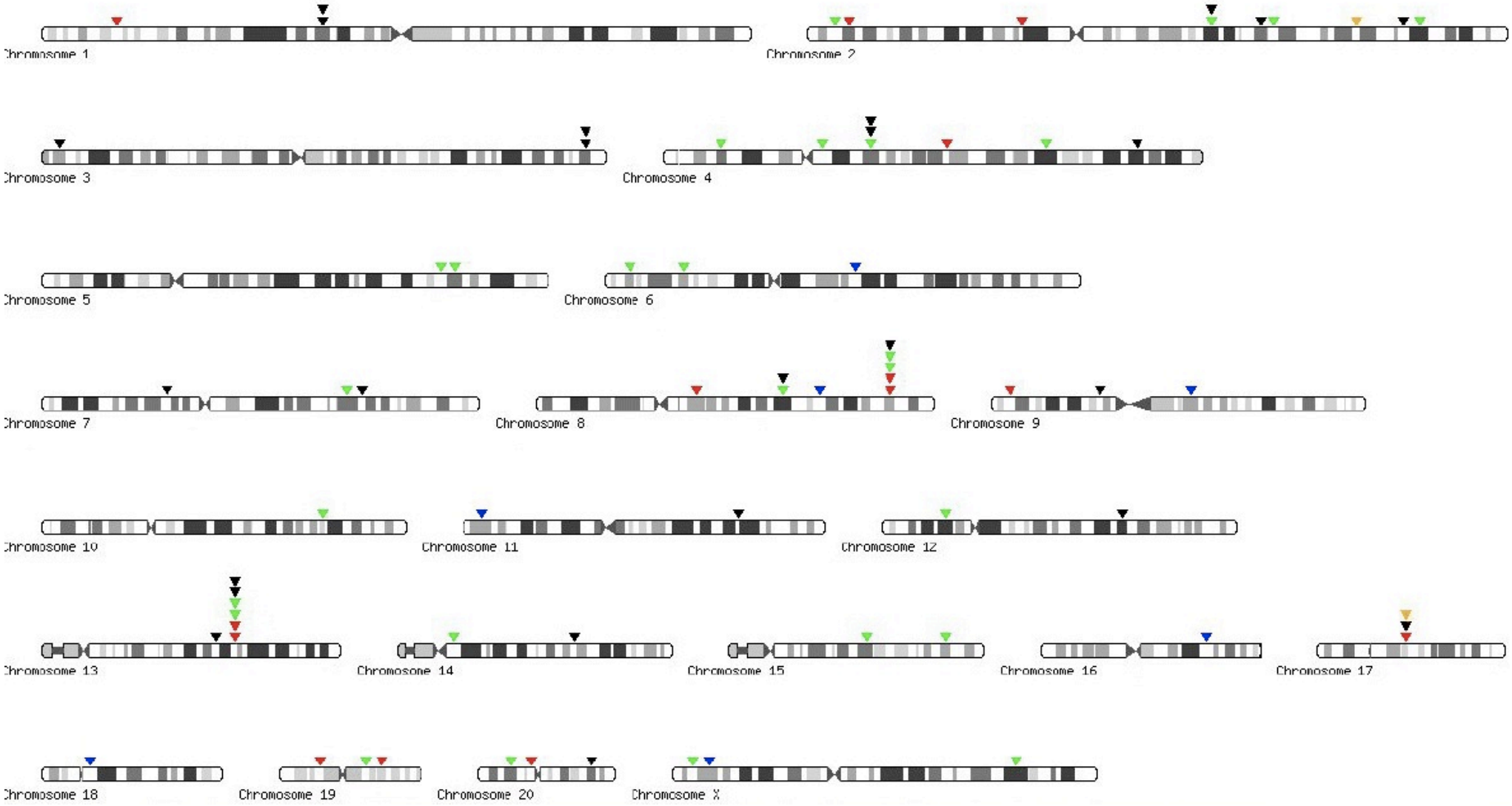
How does this look at the drawing board?



- Issues in practice:
  - How to collect and represent data
  - How to count points inside
  - How to conclude on relation or not

Interpreting a claim

"Viruses might be expected to integrate near genes. Our results confirm such preferential localization inside genes for HPV, where 75 out of 119 determined integration sites (attached) fall inside genes."



HPV integration sites

Interpreting a claim

"Viruses might be expected to integrate near genes. Our results confirm such preferential localization inside genes for HPV, where 75 out of 119 determined integration sites (attached) fall inside genes."

How would you go forth in reproducing such a claim?

(bioinformatically, i.e. trusting the supplied genomic locations)

Down to the ground

- Exploring HPV data in the Genomic HyperBrowser

Now you try!

- Everybody find:
  - whether HPV is preferentially located inside genes
  - proportion of genome covered by genes
  - number of HPV sites inside genes

  - http://hyperbrowser.uio.no
  - The Genomic HyperBrowser->Perform analysis, in left-hand menu
  - (HPV: hg19 - Phenotype and disease associations:
    Assorted experiments:Virus integration, HPV specific..)

Making justified choices is indeed hard!

- The choice of data may influence results
  - Both source and exact version of genes might matter
  - Can sometimes justify, e.g. based on sensitivity/ specificity trade-off
  - Should ideally show how results vary with choice of data
  - Should at least be very precise in what was done (accessibility, transparency, reproducibility)

Making justified choices is indeed hard (2)

- There is usually more than one possible test for a given biological question
  - The choice has to be made,
    and can't be resolved automatically
  - Statistical and biological implications play together to determine what may be reasonable
  - Should at least expose the different possibilities

Hypothesis testing

- Alternative hypothesis (H1)
  - What you really want to show (more HPV in genes)
- Null hypothesis (H0)
  - A neutral baseline (HPV equally inside/outside)
- P-value
  - How likely is observation (or more extreme), given H0
  - Observation unlikely -> reject H0, left with H1

Hypothesis testing: the challenges

- Alternative hypothesis (H1)
  - What you really want to show (more HPV in genes)
- Null hypothesis (H0)
  - A neutral baseline (HPV equally inside/outside)
- P-value
  - How likely is observation (or more extreme), given H0
  - Observation unlikely -> reject H0, left with H1

Mathematically imprecise?

Is it easy to define?

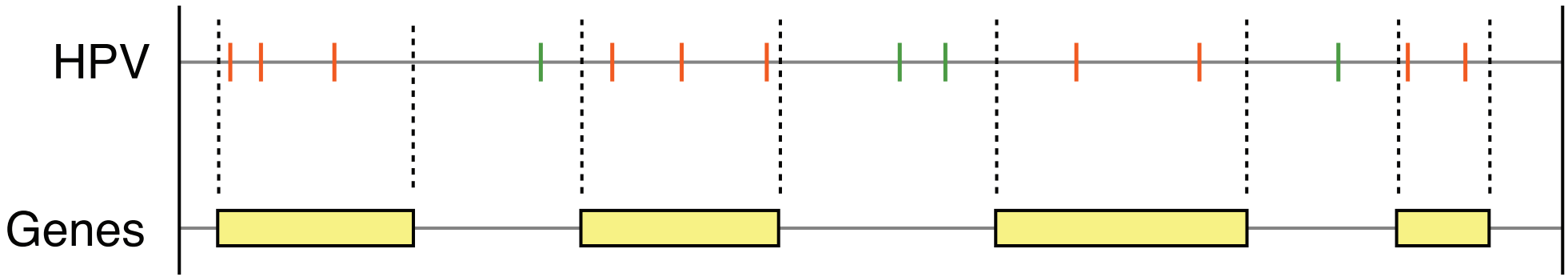
How to compute?

Or maybe unlikely for other reason?

How to compute p-value?

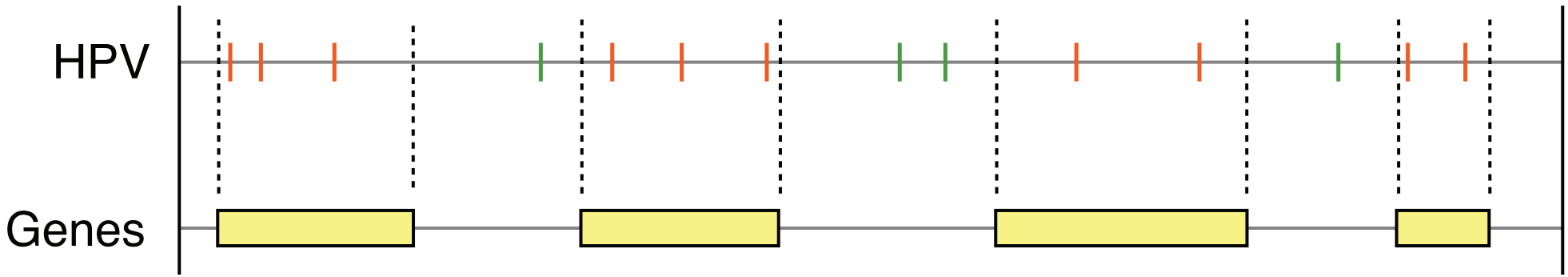
- Look at normal distribution table? Run a t-test?
  - But where to put in HPV and genes?

The quest for a distribution

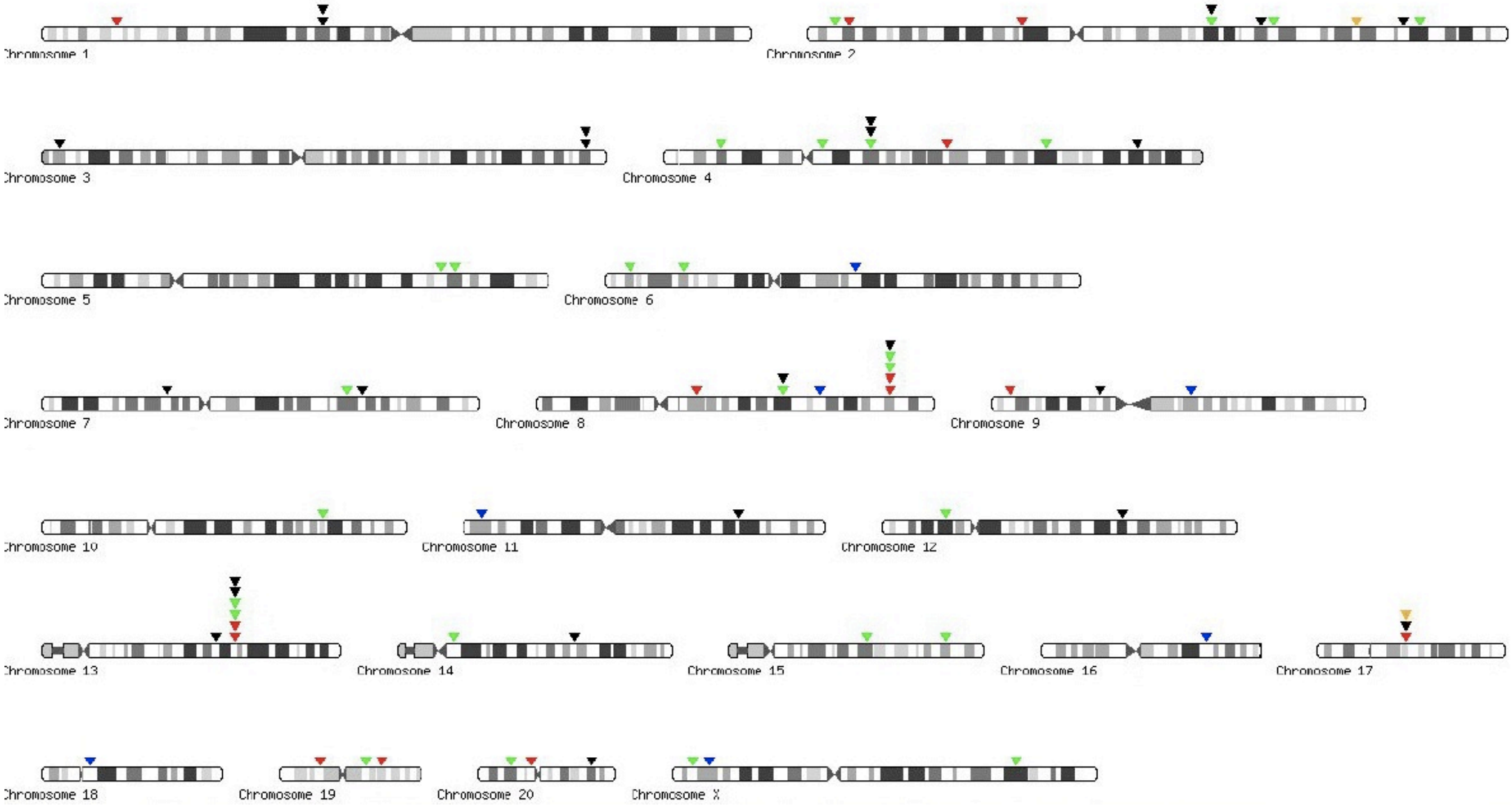


- Can we find a suited distribution?
  (for number of HPV sites inside genes under H0)
  - Statistician may find that "yes: a binomial distribution"
  - Would you be comfortable assuming a binomial distribution?
    Or better: Would you have any clue on the implications?

The quest for a distribution



- The implication of using a binomial distribution
  - What is binomially distributed - HPV or genes?
  - Neither! This only applies to the measure.
  - Instead, HPV assumed independently and uniformly distributed
  - Not trivial to see, and if found: is this acceptable?



HPV integration sites

How to compute p-value?

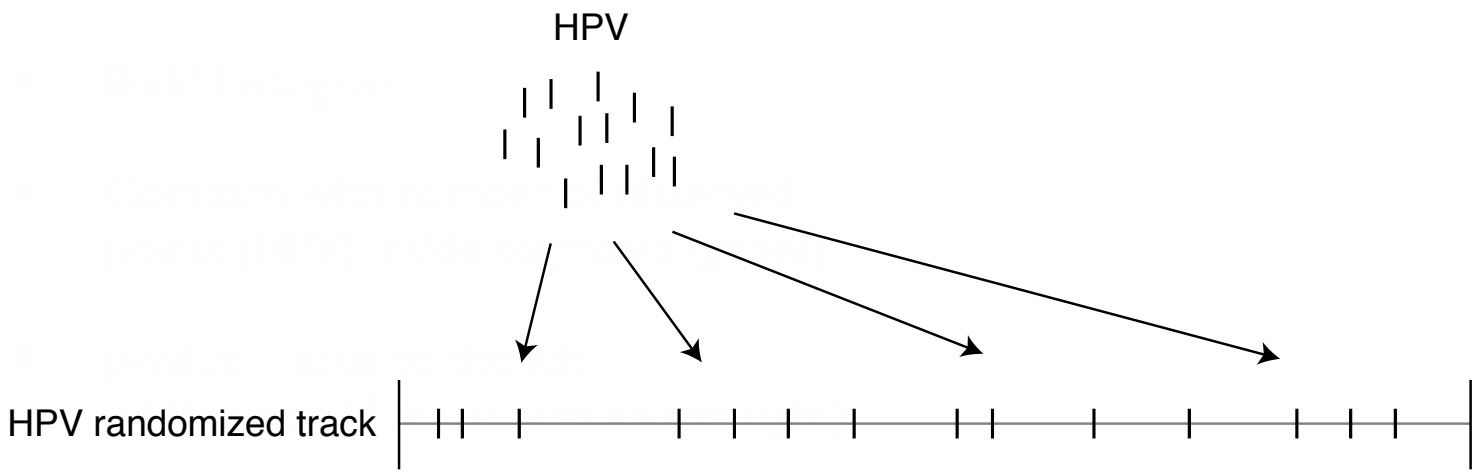
- Look at normal distribution table? Run a t-test?
  - But where to put in HPV and genes?
- Turns out that thinking about standard tests and distributions becomes awkward
  - Instead, do it the modern way..

Meet Monte Carlo

- Null model:
  - How to randomize data (precise rendition of H0)
  - Where could HPV be located under H0..
- Test-statistic:
  - How to measure aspect of interest
  - Number of HPV sites located inside genes
- P-value:
  - How often is ***test-statistic*** from ***null model*** more extreme than for observation?
  - How often are 65 or more random HPV inside genes?

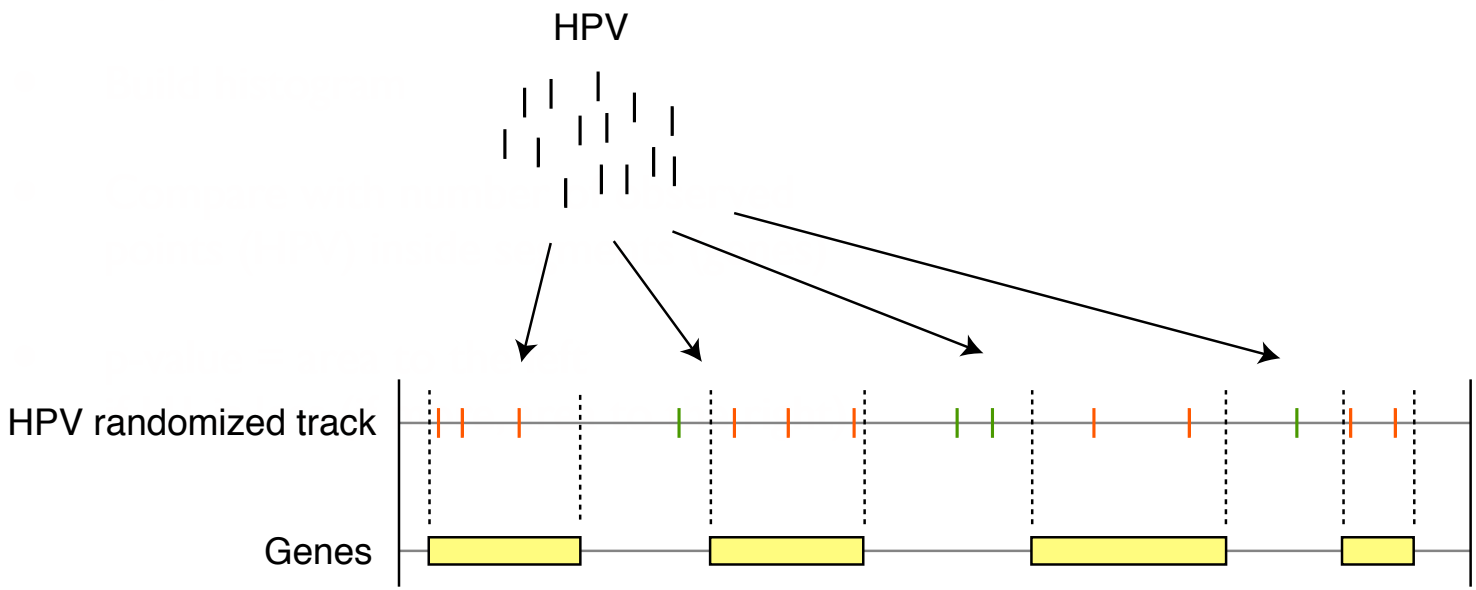
Monte Carlo test on "points inside segments"

- Randomize point (HPV) locations (null model)



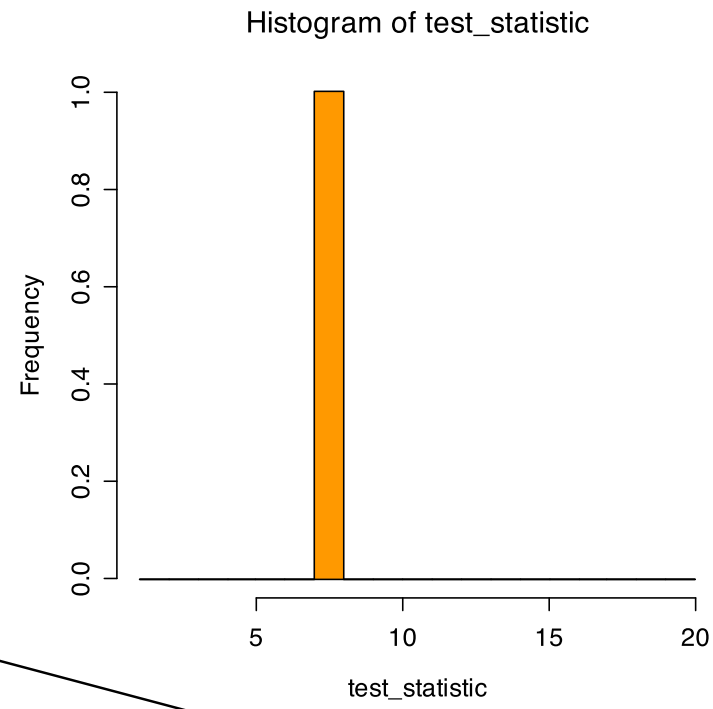
Monte Carlo test on "points inside segments"

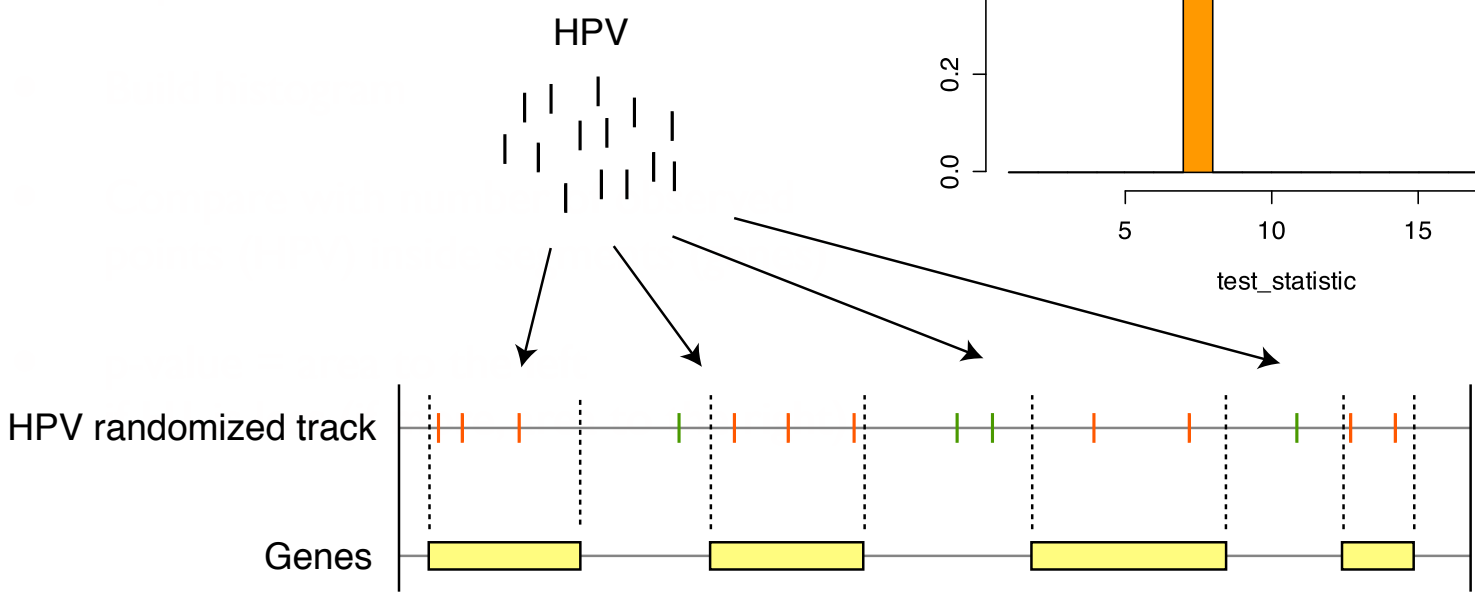
- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic



Monte Carlo test on "points inside segments"

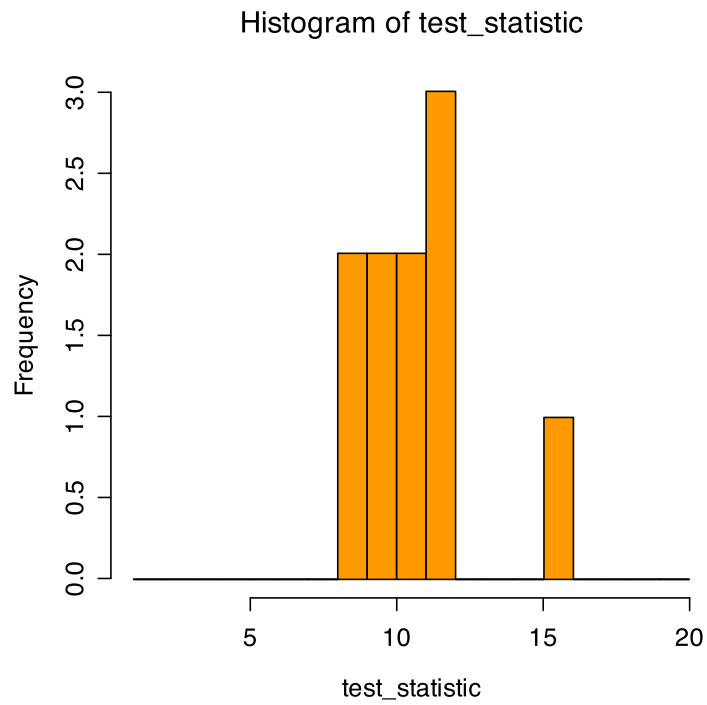
- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic





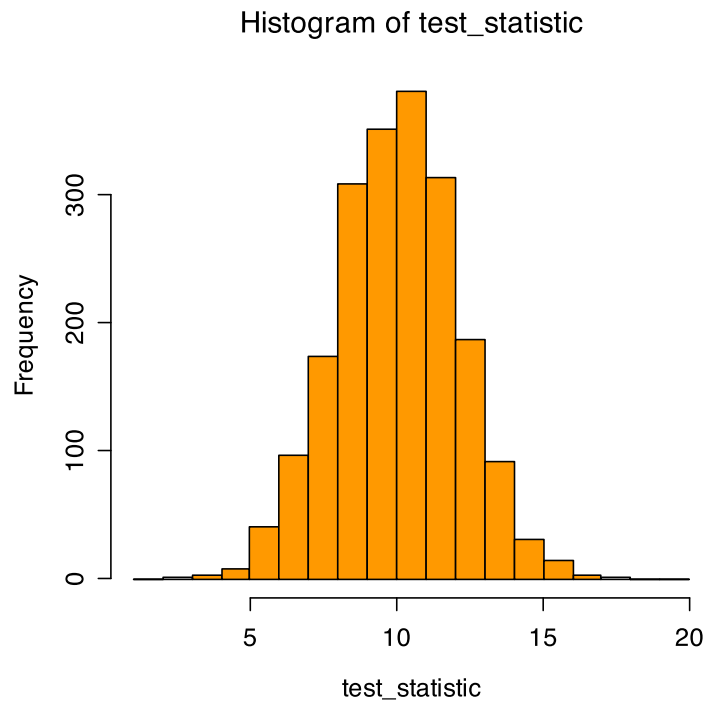
Monte Carlo test on "points inside segments"

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times



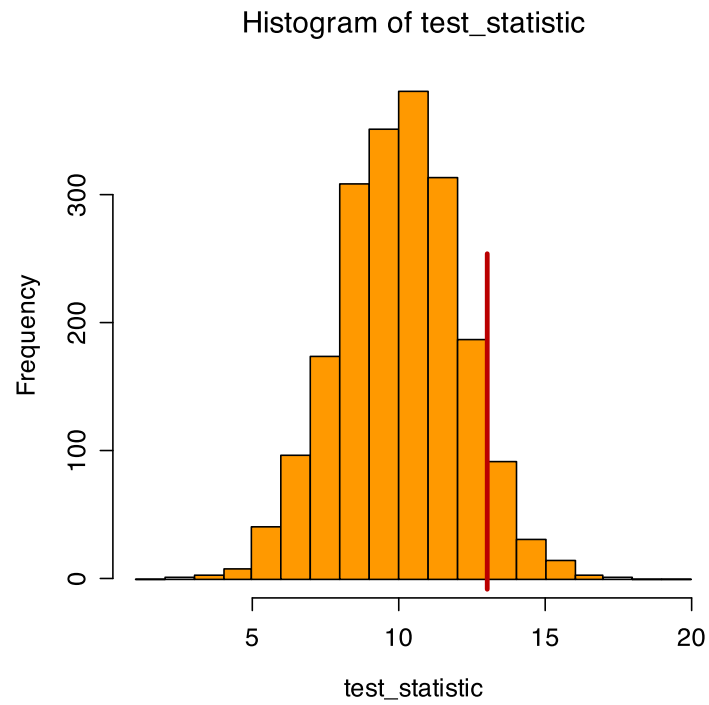
Monte Carlo test on "points inside segments"

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times
- Build histogram



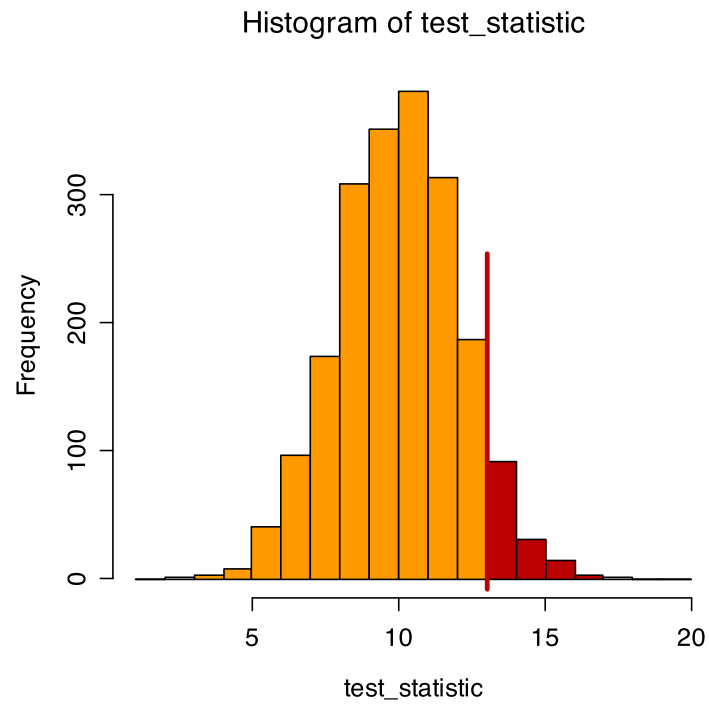
Monte Carlo test on "points inside segments"

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times
- Build histogram
- Compare with number of observed points (HPV) inside segments (genes)



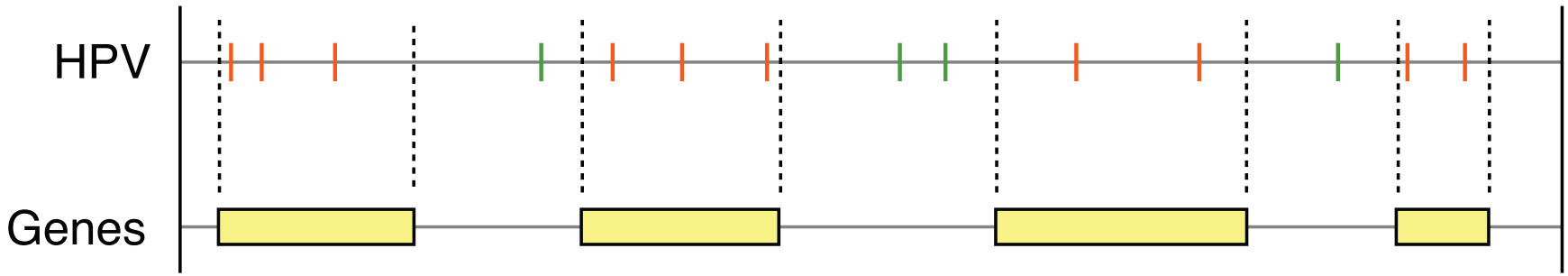
Monte Carlo test on "points inside segments"

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times
- Build histogram
- Compare with number of observed points (HPV) inside segments (genes)
- p-value = area to the right if H1 is more (if less, area to the left)



p-value = 0.08

Back to HPV and genes



- Didn't like implications of binomial distribution?
- With Monte Carlo, you can shuffle how you like
  - Throw HPV around uniformly and independently (like binomial)
  - Keep clustering tendency of HPV (shuffle HPV spacings)
  - Keep HPV as is, only shuffle genes (in various ways)

You try!

- Try different gene data sources and assumptions (null models) on HPV-gene relation
- Use redo functionality ()
- Who get's the best p-value;)

Data and assumptions matter!

- HPV inside Ensembl genes? *(default assumptions)*
  - Yes, but a bit weak evidence *(p-value=0.013)*
- HPV inside Refseq genes? *(default assumptions)*
  - No! *(p-value=0.4512)*
- Inside Ensembl (v2)? *(Preserve inter-HPV distances)*
  - Yes! (p-value=0.007)
- Inside Ensembl (v3)? *(Randomize genes)*
  - Maybe.. (p-value=0.027)

An example of alternative assumptions

- Duan, [...] and Noble (Nature, 2011):
  - Extensive significant 3D co-localization of functional elements, assessed by hypergeometric distribution
- Witten and Noble (NAR, 2012):
  - Hypergeometric had unrealistic implications. Telomeres and breakpoints may not be co-located after all.. (cancelled 4 of 11 findings)

In silico analysis and reproducibility

- Bioinformaticians gets surprised every time they need to redo/modify previous analyses
- But you bench biologists already know the importance of reproducibility!
- You also know that even with a detailed lab journal, reproduction is a challenge
- The question is then how this manifests itself when doing analysis on a computer

What is in silico reproducibility?

- Basically the same issues as at the bench:
  - Materials -> Data sources
  - Experiment conditions -> Analysis parameters
  - Equipment (and models) -> Programs (and versions)
- And the same challenges:
  - Are all relevant conditions described accurately?
  - Will the same materials and equipment be available?

What is the current status of reproducibility?

- Less than half of selected microarray experiments published in Nature Genetics could be reproduced
(Ioannidis et al., Nat Genet 2009)

- More than half [of surveyed papers] do not provide primary data and list neither the version nor the parameters used [for read mapping]
(Nekrutenko and Taylor., Nat Rev Genet 2012)

Why should you care?

(about making your analyses reproducible)

- Because it's the right thing to do!
- .. and the one that's struggling with its reproduction is often the future you
- Journals are becoming aware of the issues
- Reviewers may value it
- Anyway, it's the same as at the bench..

How could we document the HPV analysis?

- We should note, archive and distribute all data sources and parameters used in the analysis
  - But it shouldn't be necessary to duplicate this information, also with the risk of introducing errors..
- HyperBrowser automatically compiles a report
  - Includes all inputs, parameters and computation details
- Galaxy even allows direct storing and sharing of the whole analysis, including re-run possibility
  - A Galaxy Page, linking to shared history, is all we need

You try!

- Create a Galaxy Page
  - User->Saved Pages (you will have to register a user)
  - “Add new page”
  - Click chosen name under “Title” and “edit content”
  - “Embed Galaxy object”->history
  - For now, just write very brief explanatory text
  - “Save”, “Close” and “Share or Publish” via Link..

Some simple rules for reproducibility

- Whenever making a claim, note a reference to supportive data
  - ".. MS occur preferentially inside AP in B-cells [hist:HbLecture-8] .."
- For every result of interest, keep track of how it was produced
  - Solved automatically by redo-functionality if using Galaxy
- Provide public access to scripts, runs and results
  - Provide link to Galaxy Page that embed histories with all runs and results

Some simple rules for reproducibility (cont.)

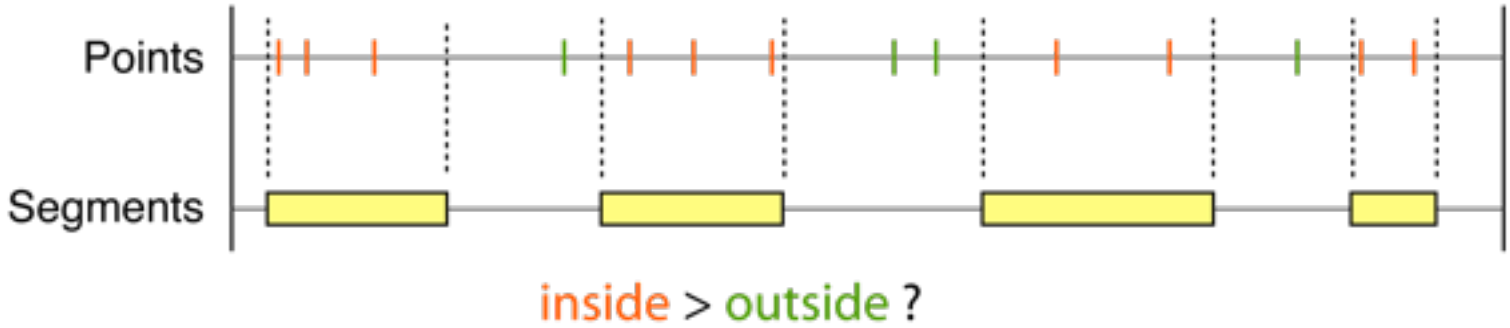
- Use executable documentation and verification
  - Galaxy histories document analysis and are executable
- Generate hierarchical analysis output, allowing layers of increasing detail to be inspected
  - HyperBrowser provides conclusion, full table and local results
- Always store raw data behind plots
  - Result plots of HyperBrowser analyses come with underlying numbers

Summary

- Genomic data can be visualized in tools like *"UCSC genome browser"* and analyzed in tools like *"The Genomic HyperBrowser"*
- Monte Carlo is a powerful, flexible and transparent method for hypothesis testing
- Although tools may offer simple user interfaces, they can't make all choices for you
- Integrated analysis frameworks makes it easy to provide precise documentation of analyses

The Genomic HyperBrowser

Main aim: Analyze any data in the form of genomic tracks



Services:

- "Out of the box" web-based system
- Custom-tailored development of functionality
- Statistical guidance for genome analysis

Builds upon:

- Large and well-tested code base
- Hundreds of statistical analyses and tools
- Cross-disciplinary competence in genome analysis
- Experience with GWAS, TFBS, 3D, chromatin...

Our vision

- Be able to answer any question involving data that can be represented as tracks!
  - Whether the question regards
    the DNase accessibility in
    multiple sclerosis associated regions
    or
    the number of camels in Afghanistan in 2005

What have we been working on?

- Methods in genomics:
  - Statistical testing, clustering, large-scale analyses, multi-track comparisons, cross-species analysis ..
- Applications of genome analysis:
  - Chromatin, 3D genome organization, mechanistic studies of disease, gene regulation, viral integration ..

Publications

- Methodology described in 2010 Genome Biology paper
- Three further papers on methodology, three on applications
- Eight papers in preparation or review