Genome analysis and statistical testing

Sveinung Gundersen
PhD, Radiumhospitalet/IFI

What you will learn

09.00 - 12.00:

- Genome analysis
    - Statistical testing

13.00 - 17.00:

- Different user interfaces for bioinformatics
- Overview of Galaxy
- Reproducibility

The form of these sessions

- We briefly introduce a topic
- You do a short hands-on
- We explain the topic in more detail

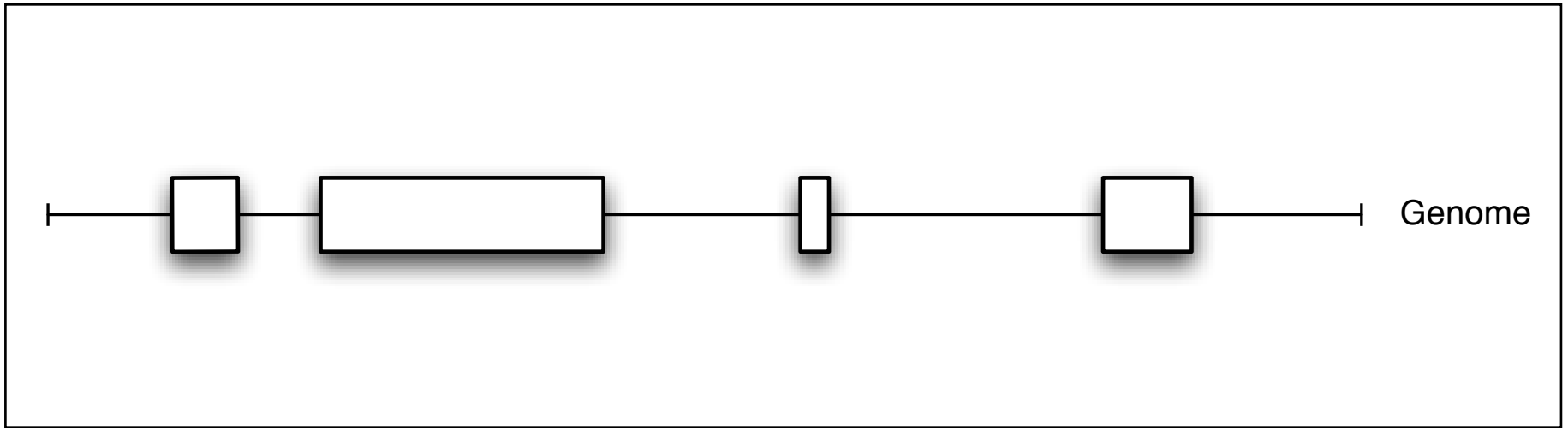
- ... we repeat this for a sequence of increasingly advanced/detailed topics

Biological cases, but not depth

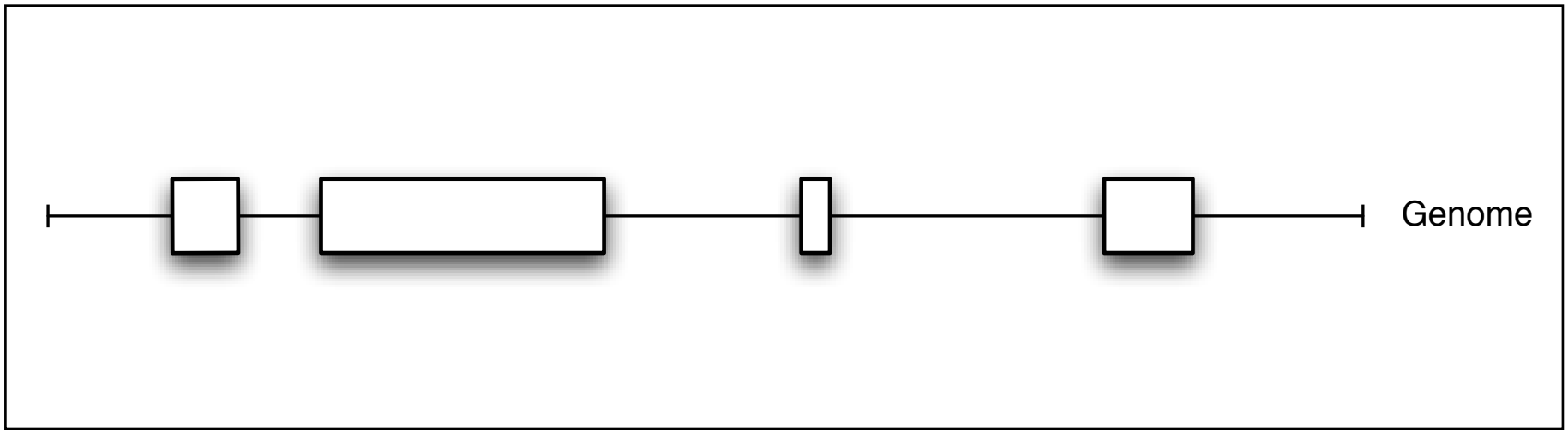
- We will use biological cases, but not focus on biological interpretation:
  - You are the experts in biology, not us
  - Our message is the methodology and its generic (statistical) interpretations

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

- Google: "UCSC"
- URL: http://genome.ucsc.edu/cgi-bin/hgGateway

- Try:
  - Small region and whole chromosome
  - Add/remove tracks, change their appearance

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

- A relation between methylation patterns and repeating elements? (Genome Res. 2009 19: 221-233)
- Distinct methylation for tissue-specific genes?(Genome Res. 2010 20: 1493-1502)
- Cooperative histone modifications? (Nat Genet 2008 40:897-903)

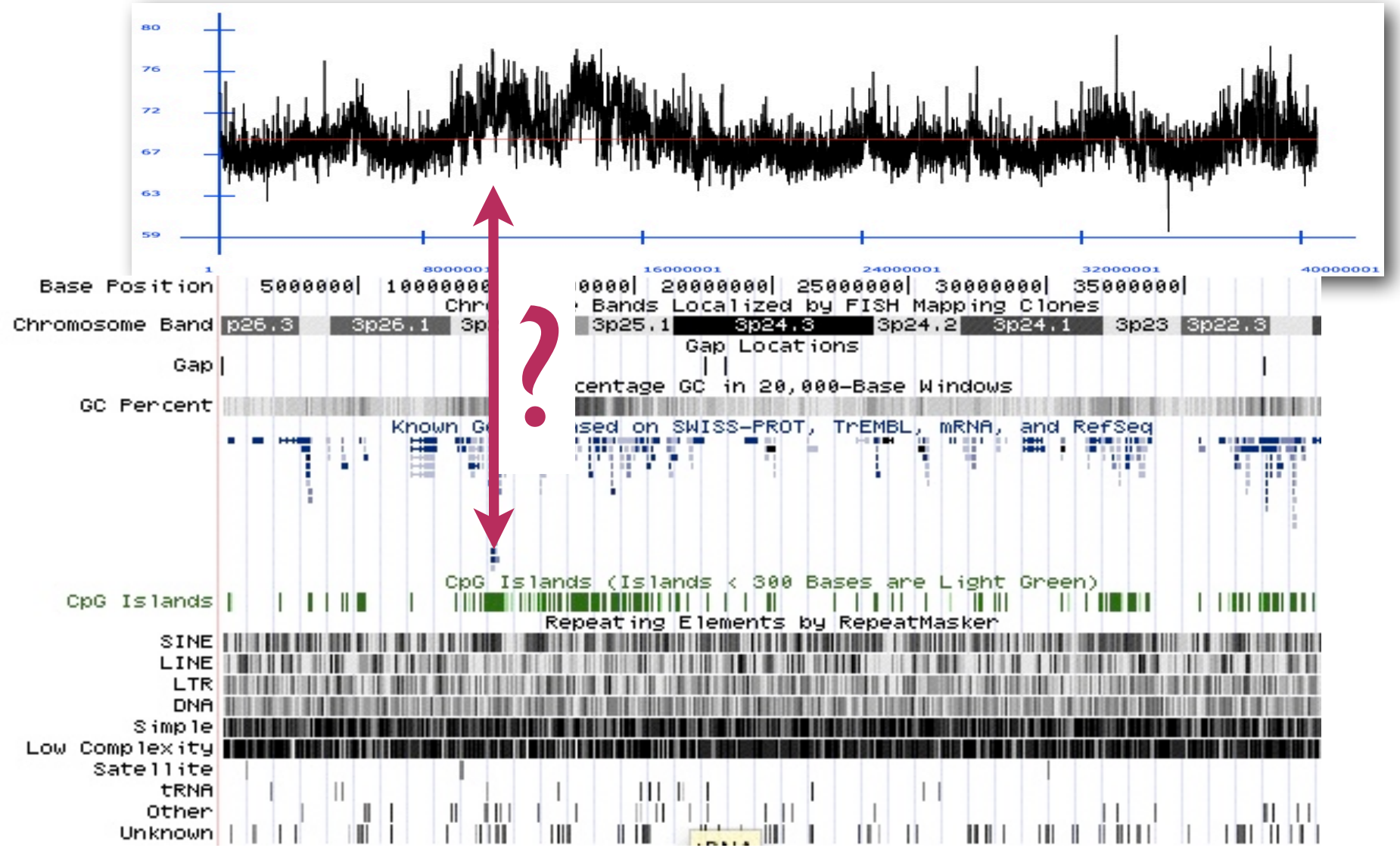
Example analyses (cont.)

- Fragile sites, breakpoints and repeats? (Genome Biology 2006 7:R115)
- Copy number variation, repeats, duplications and genes? (Genome Res. 2009 19: 1682-1690)
- Methylation and active genes at T-Cell G0->G1 (Genome Res. 2009 19: 1325-1337)

Example analyses (cont.)

- Virus integration vs genes, CpG, GC-content
  (Journal of Virology 2007 6731–6741)
- Methylation patterns in embryonic cells
  (PNAS 2010 107:10783–10790)

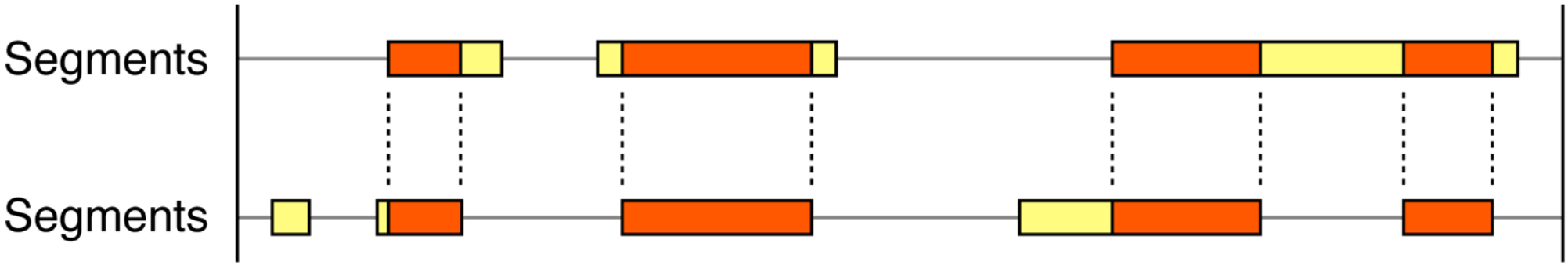
This can't be it?!



Co-location of genomic features

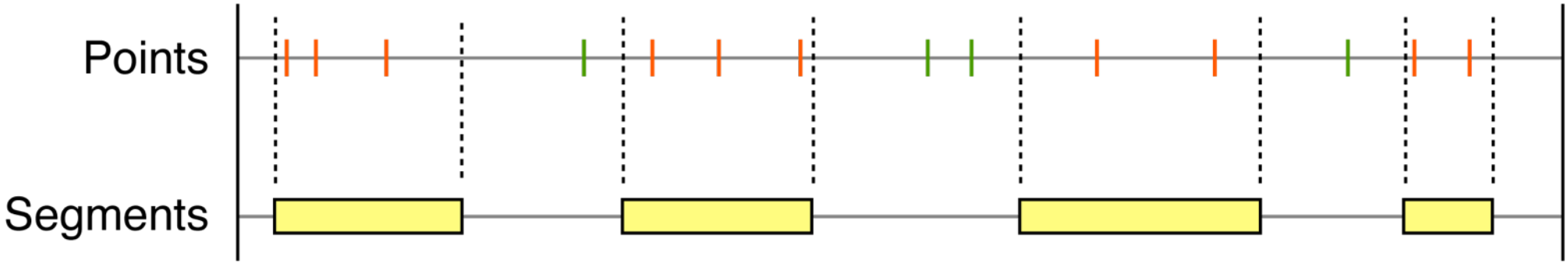
- Common question:
  do genomic feature X and Y occur
  (more than expected)
  at the same locations in the genome?
- Used to discover novel relations
- May indicate a direct causal relation, or hint to indirect association.

How does this look at the drawing board?



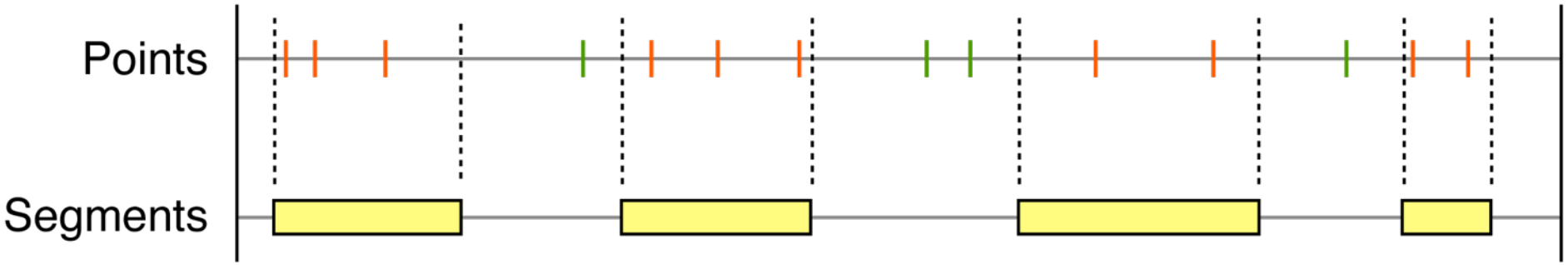
overlap > expected?

How does this look at the drawing board?



inside > outside ?

How does this look at the drawing board?

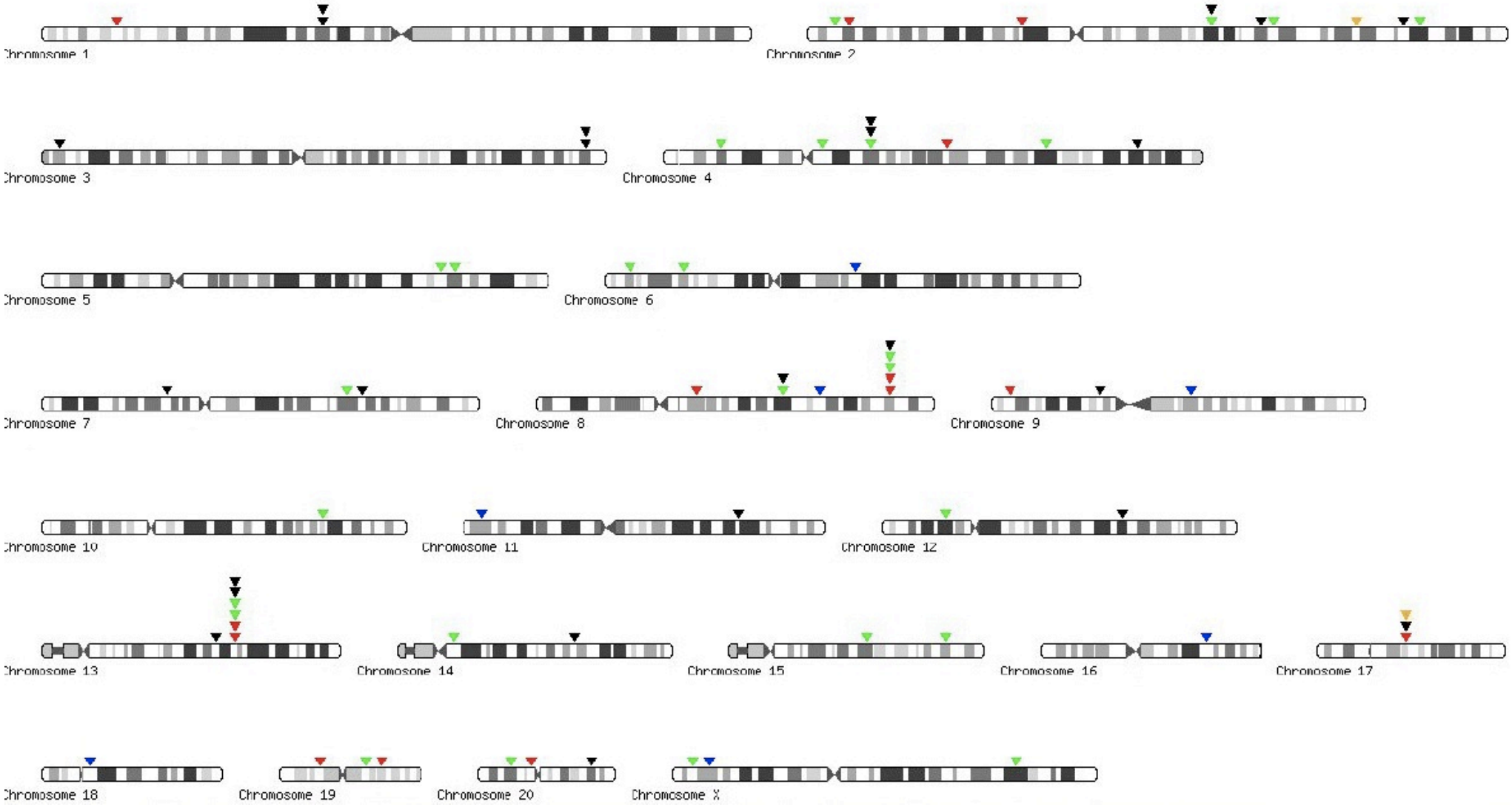


inside > outside ?

- 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?

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/test
  - 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?

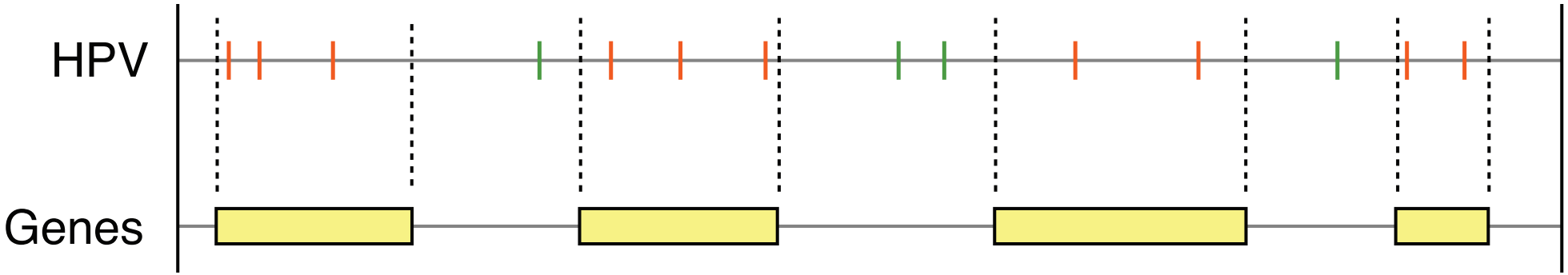
How to compute?

Or maybe unlikely for other reason?

How to compute p-value?

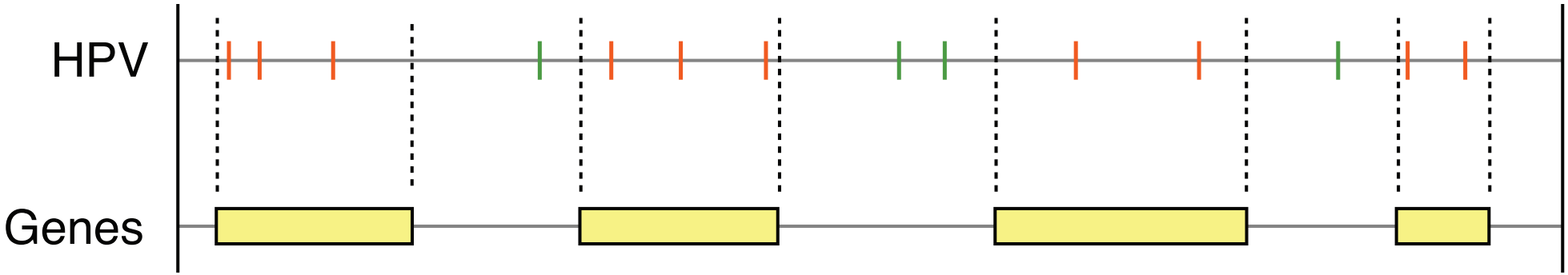
- Look at normal distribution table? Run a t-test?
  - But where to put in HPV and genes?

The quest for a distribution

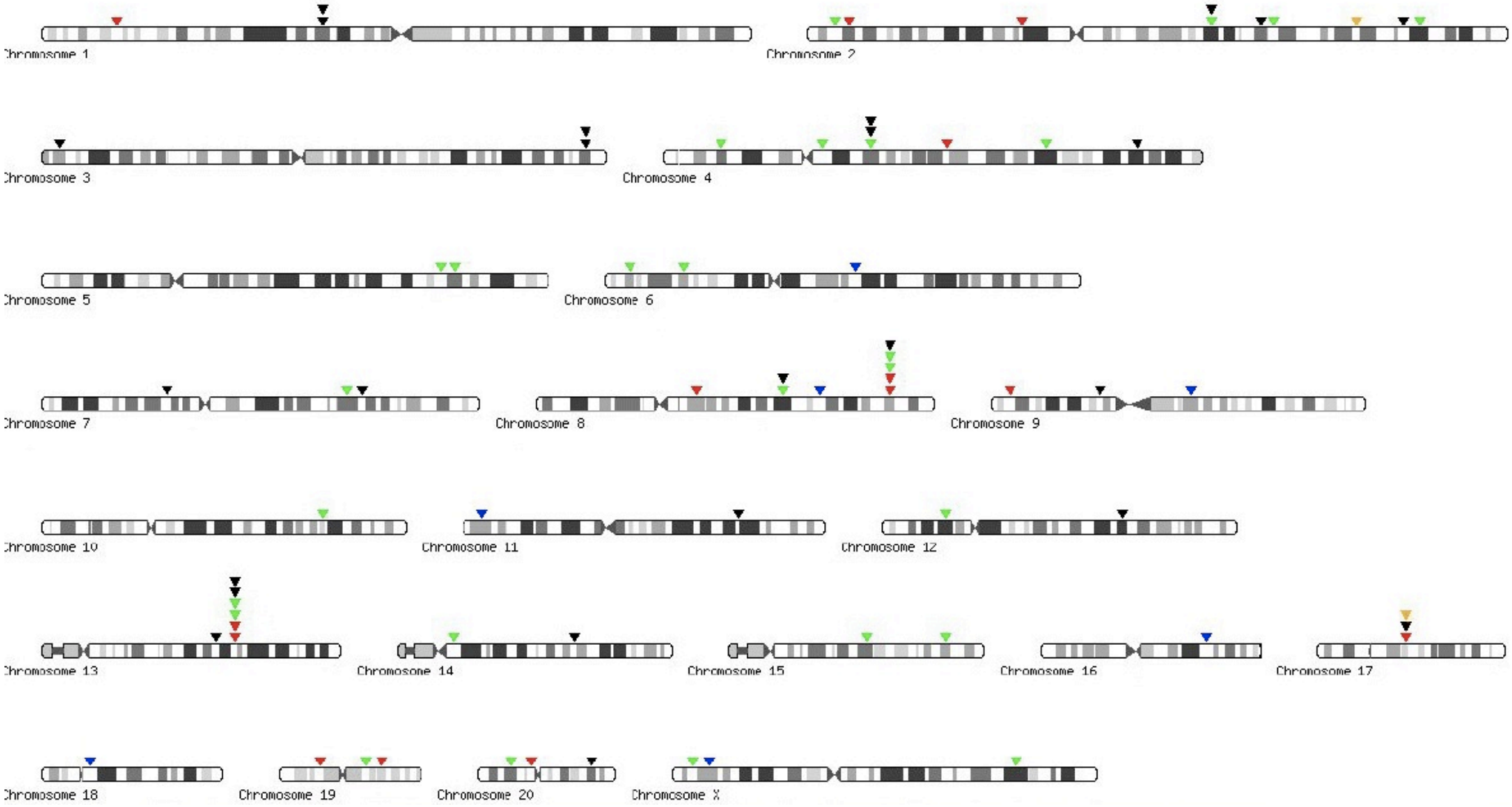


- Can we find a suited distribution?
  (for number of HPV sites inside genes under H0)
  - Statistician may find that "yes: a binomial distribution"
  - Would you be comfortable assuming a binomial distribution?
    Or better: Would you have any clue on the implications?

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?

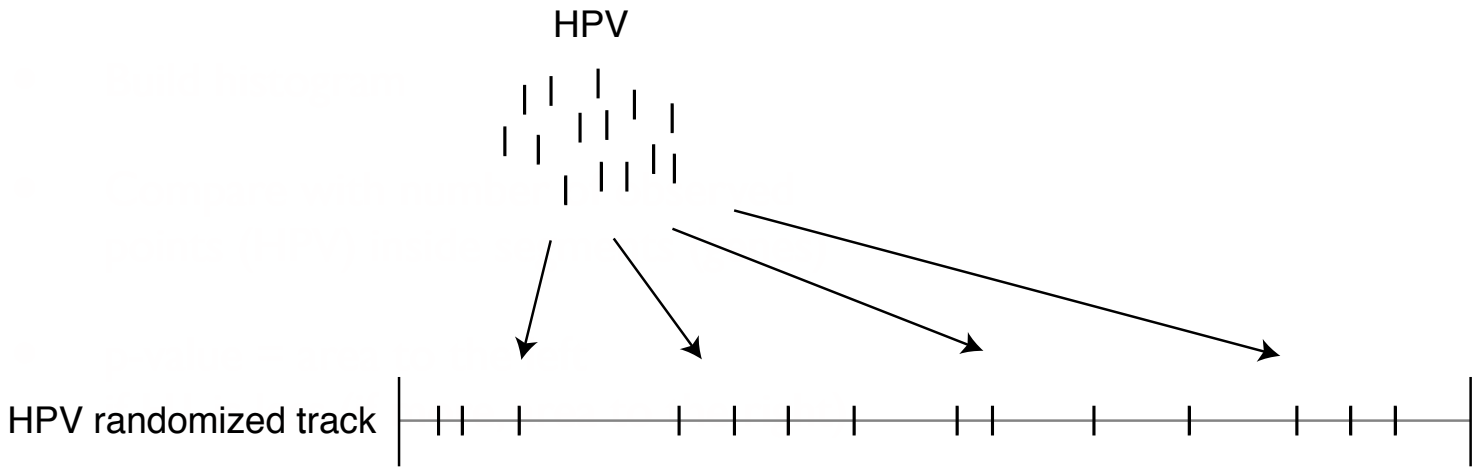
- Look at normal distribution table? Run a t-test?
  - But where to put in HPV and genes?
- Turns out that thinking about standard tests and distributions becomes awkward
  - Instead, do it the modern way..

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?

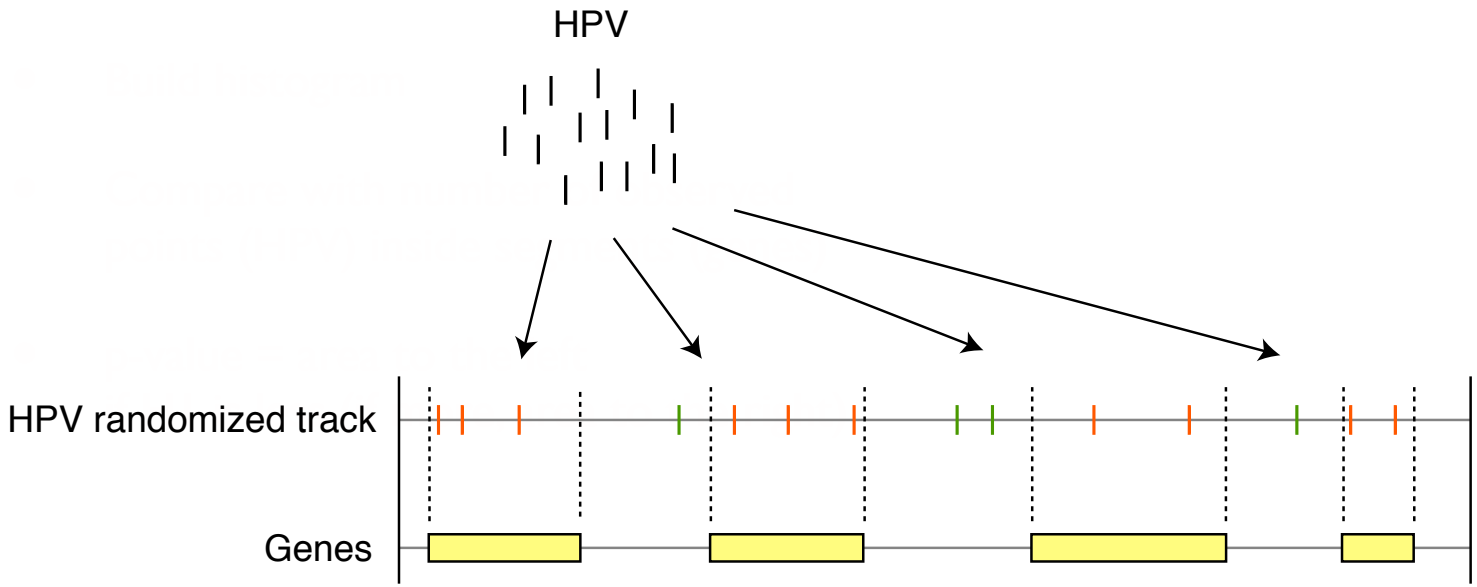
Monte Carlo test on "points inside segments"

- Randomize point (HPV) locations (null model)



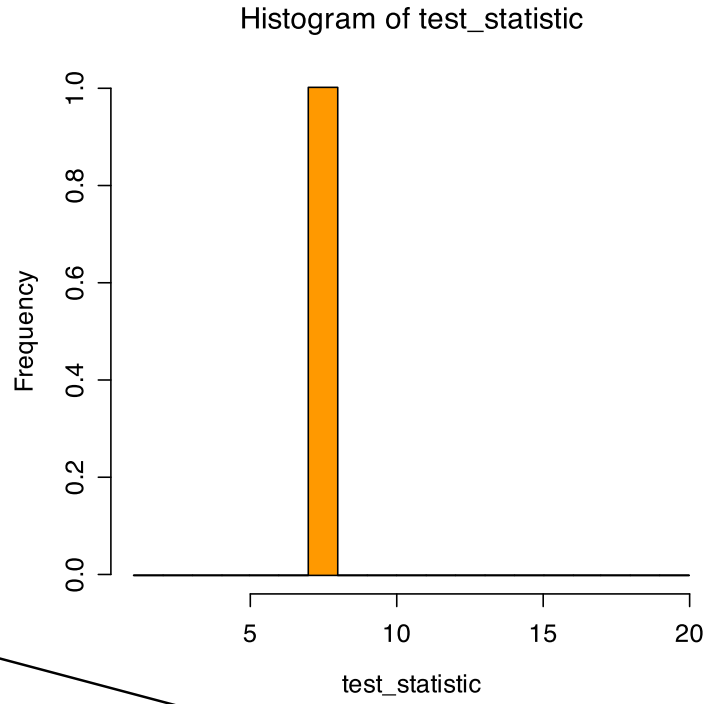
Monte Carlo test on "points inside segments"

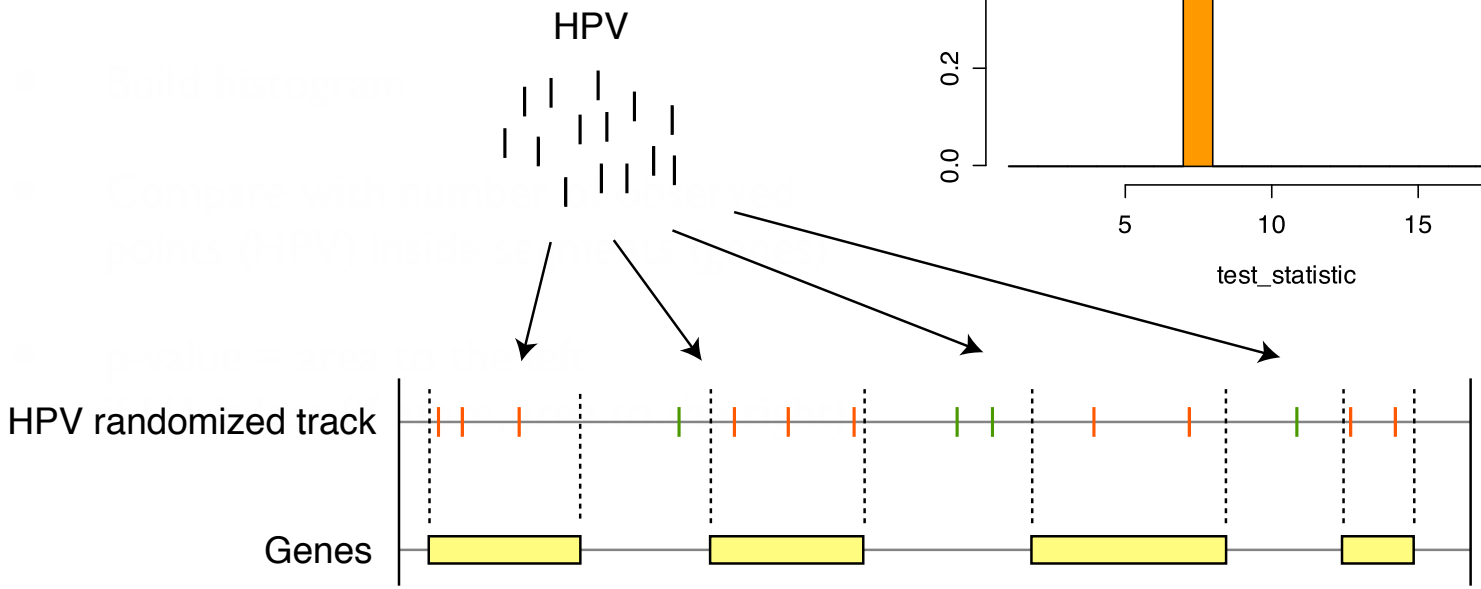
- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic



Monte Carlo test on "points inside segments"

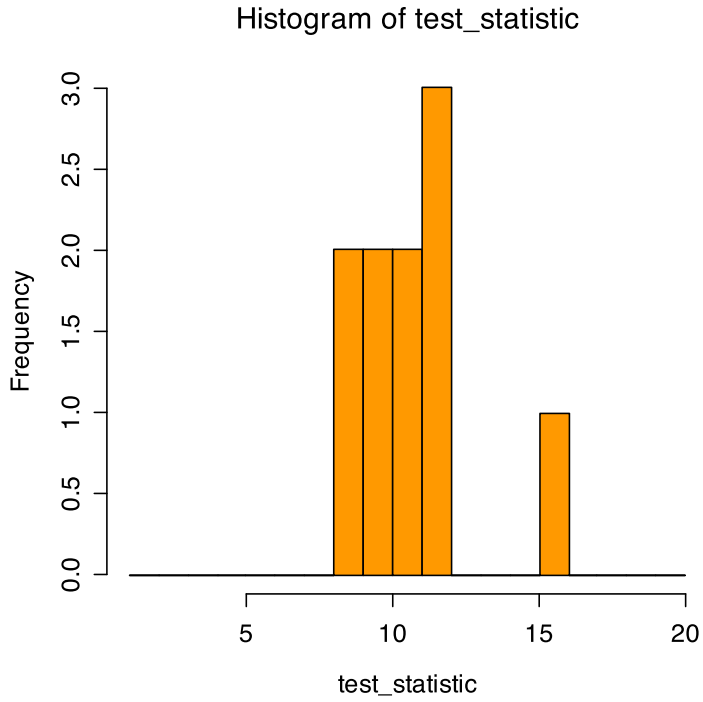
- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic





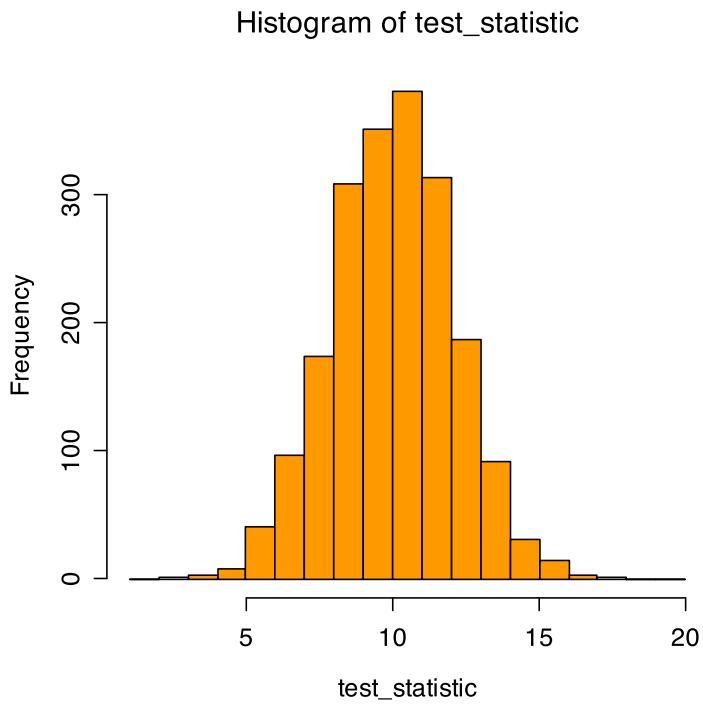
Monte Carlo test on "points inside segments"

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



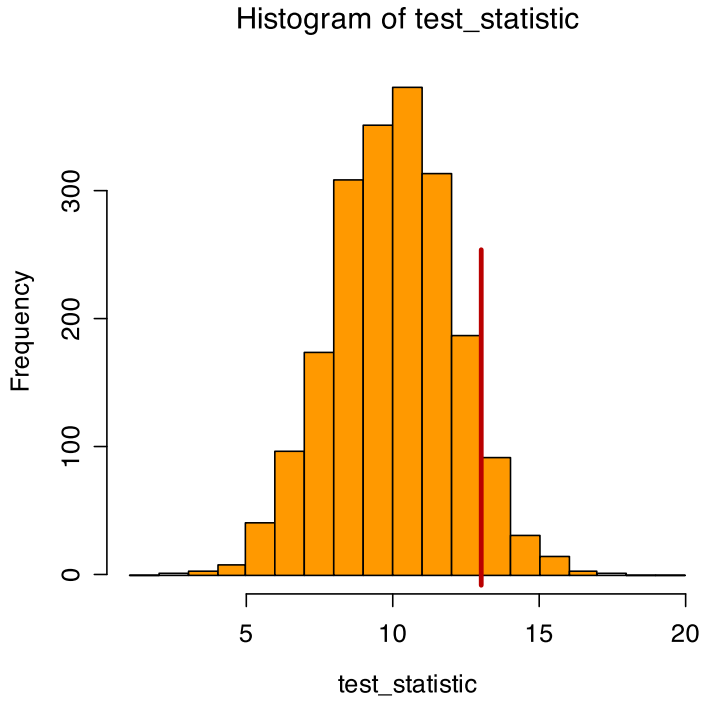
Monte Carlo test on "points inside segments"

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



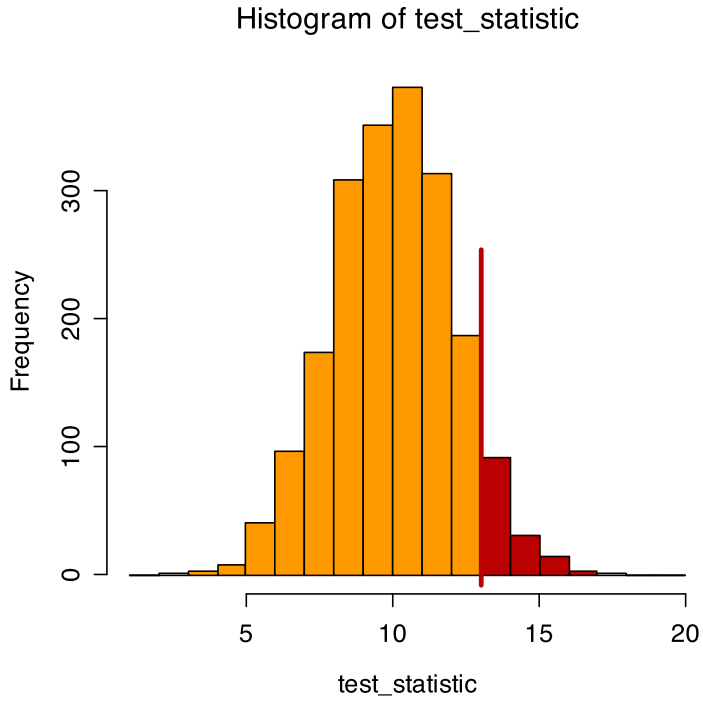
Monte Carlo test on "points inside segments"

- 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)



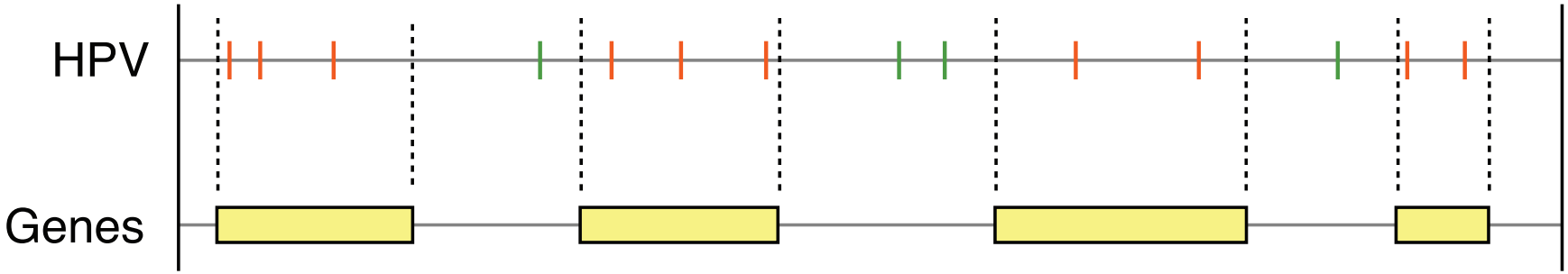
Monte Carlo test on "points inside segments"

- 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 less (if less, area to the left)



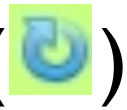
p-value = 0.08

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)

Further into statistical details: the test-statistic

- Original claim:

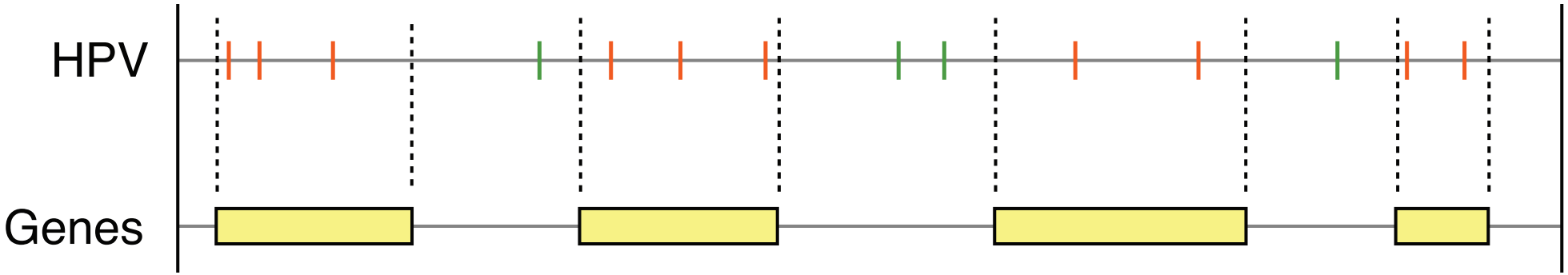
  "Viruses might be expected to integrate near genes. Our results confirm such preferential localization inside genes for HPV, where 75 out of 119 determined integration sites (attached) fall inside genes."

- Let's instead analyze distance to TSS

You try!

- “Perform analysis” to ask whether HPV is located nearby upstream end points of genes (TSS)
- Use redo - only slight changes are needed..

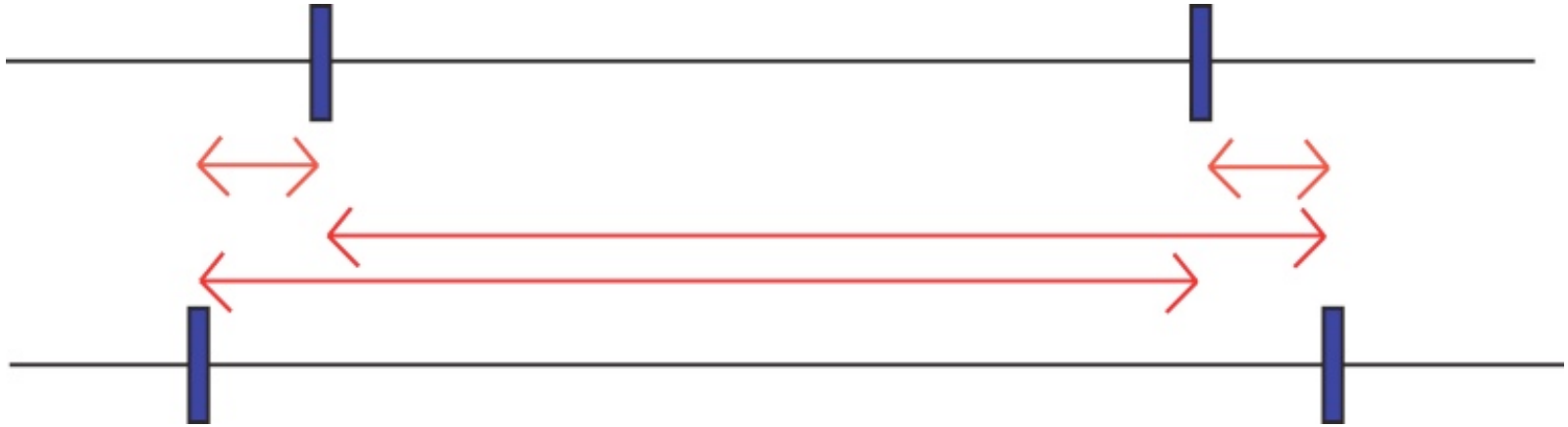
Back to drawing board: the test-statistic



- For "located inside":
  - Could simply count the number of HPV sites falling inside genes

Back to drawing board: Must quantify “close”

But that's trivial, sure: Just count bp distance!?

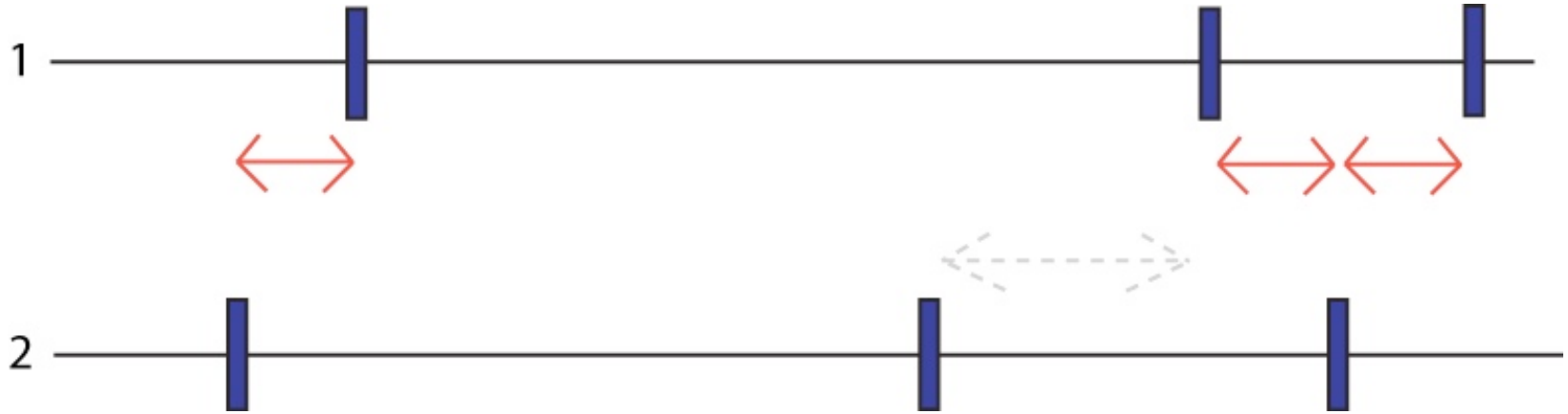


- But which distances - not all vs all?!

But that's trivial, sure: Just count bp distance!?

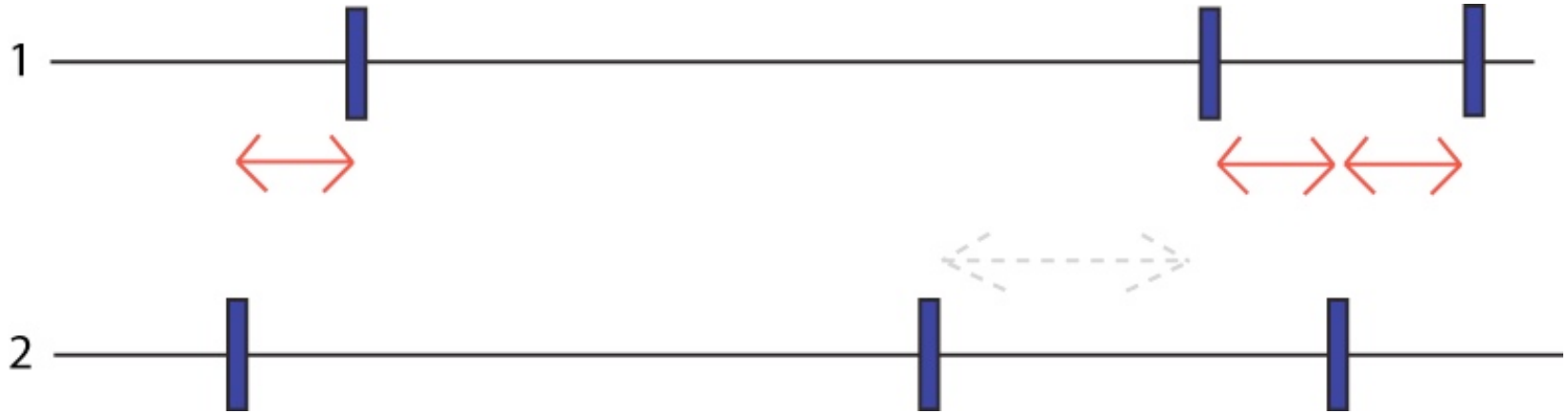
- But which distances - not all vs all?!
  - Only shortest!

But that's trivial, sure: Just count bp distance!?



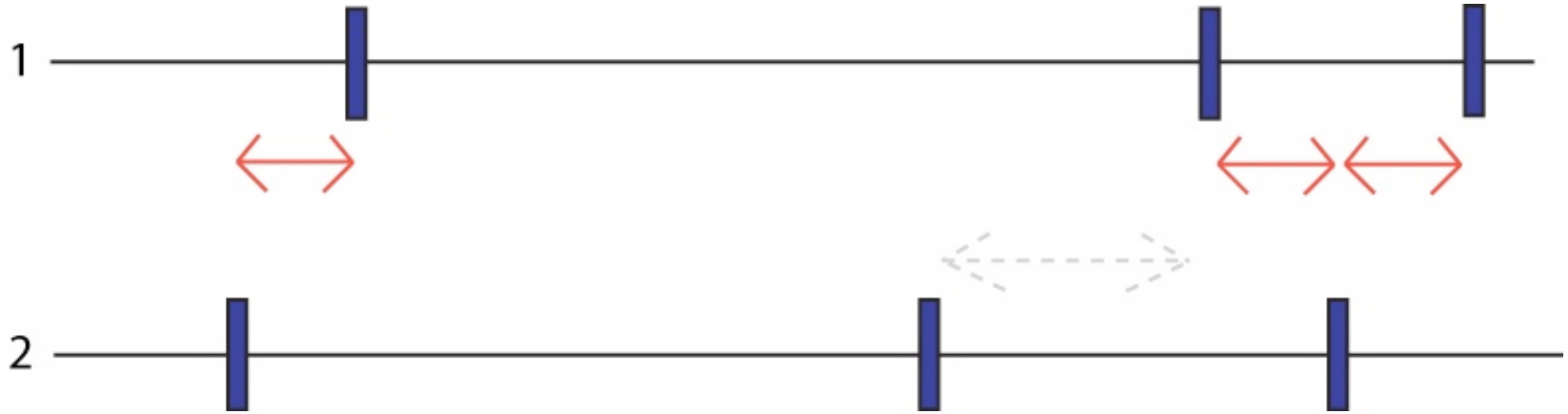
- But which distances - not all vs all?!
  - Only shortest! From 1 to 2!

But that's trivial, sure: Just count bp distance!?



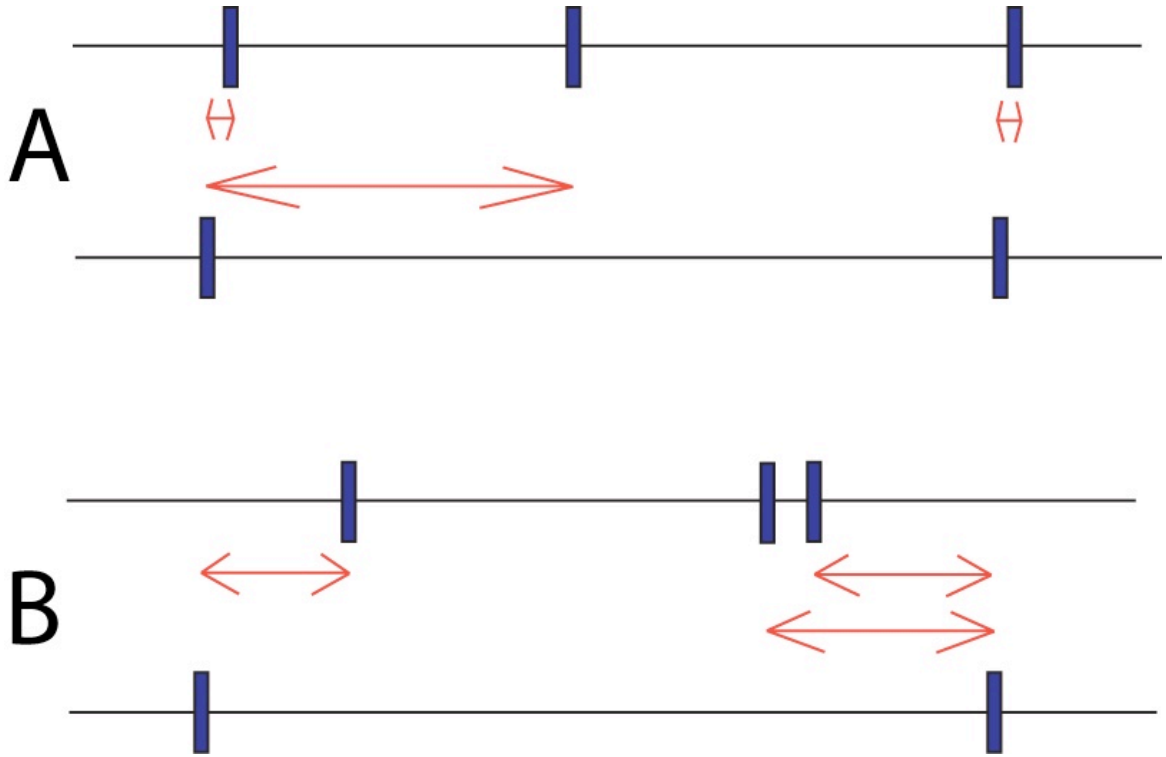
- But which distances - not all vs all?!
  - Only shortest! From 1 to 2! But MC needs a single number..

But that's trivial, sure: Just count bp distance!?



- But which distances - not all vs all?!
  - Only shortest! From 1 to 2! But MC needs a single number..
  - Just use sum/average of distances!?

Same degree of close?!



- Two scenarios with same (arithmetic) average..
  - Scenario A indicates relation, but not B !?
  - If so, can be captured by instead using geometric average

You try!

- Can you find a significant HPV-gene relation?
- Would you be comfortable reporting (publishing) this relation?
- If so, what would be an acceptable way to report it?

Any rules of thumb?

(for the statistical testing)

- Maybe:
  - Use test-statistic that gives best (lowest) p-value
  - Use null model that gives worst (highest) p-value
- Reasoning:
  - Use measure that best catches relation of interest
  - Use the most realistic model of nature (null model)

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