Statistical epigenomics

INF-BIO 5121/9121
October 11 2016, Oslo

Boris Simovski and Ivar Grytten
BMI/Genomic HyperBrowser team
Department of Informatics, UiO

Overview of session

Day 1:

09:00-10:30 Introduction. Tracks and track types.

10:45-11:30 Analysis of tracks.

11:30-12:30 Lunch

12:30-13:45 Hypothesis testing.

14:00-16:00 The Genomic Hyperbrowser. Examples and exercises.

Overview of session

Day 2:

09:00-09:15 Recap of day 1.

09:15-10:15 Descriptive statistics.

10:30-11:30 Further into statistical details.

11:30-12:30 Lunch

12:30-14:00 Reproducible research.

14:00-16:00 Analysis of track collections. The GSuite HyperBrowser.

About this module

The form of these sessions

- We briefly introduce a topic
- You do a short exercise
- 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
  - Feel free to correct us if we say something wrong

About the Genomic HyperBrowser

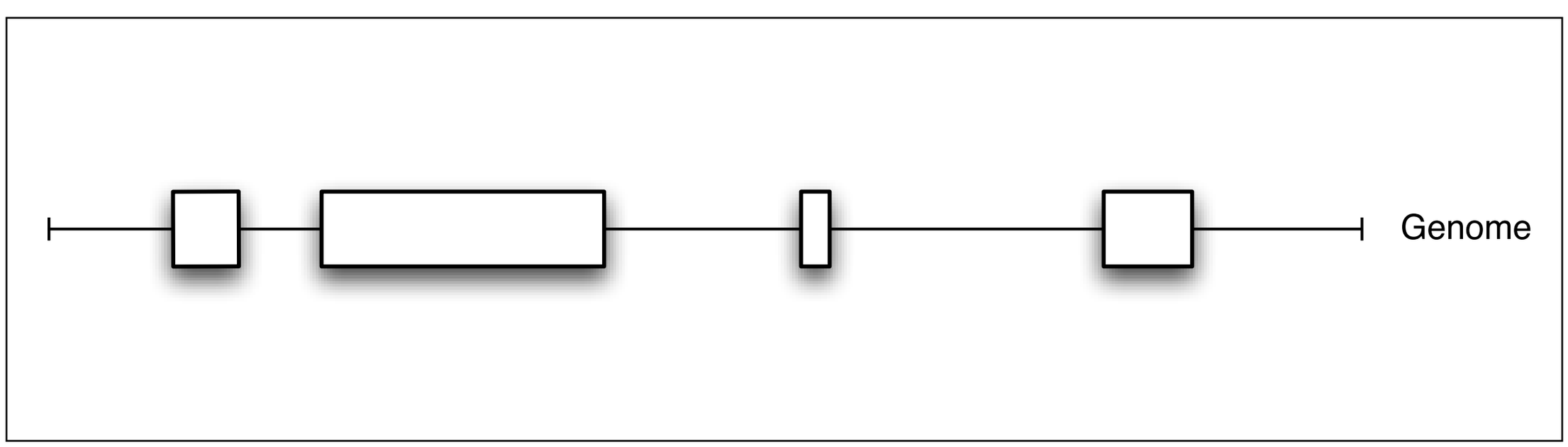
- We will make use of the Genomic HyperBrowser in this session
- The HyperBrowser is a software system for statistical analysis, developed locally at UiO
- However:

  The course is about statistical genomics. The concepts are the same if you use other tools!

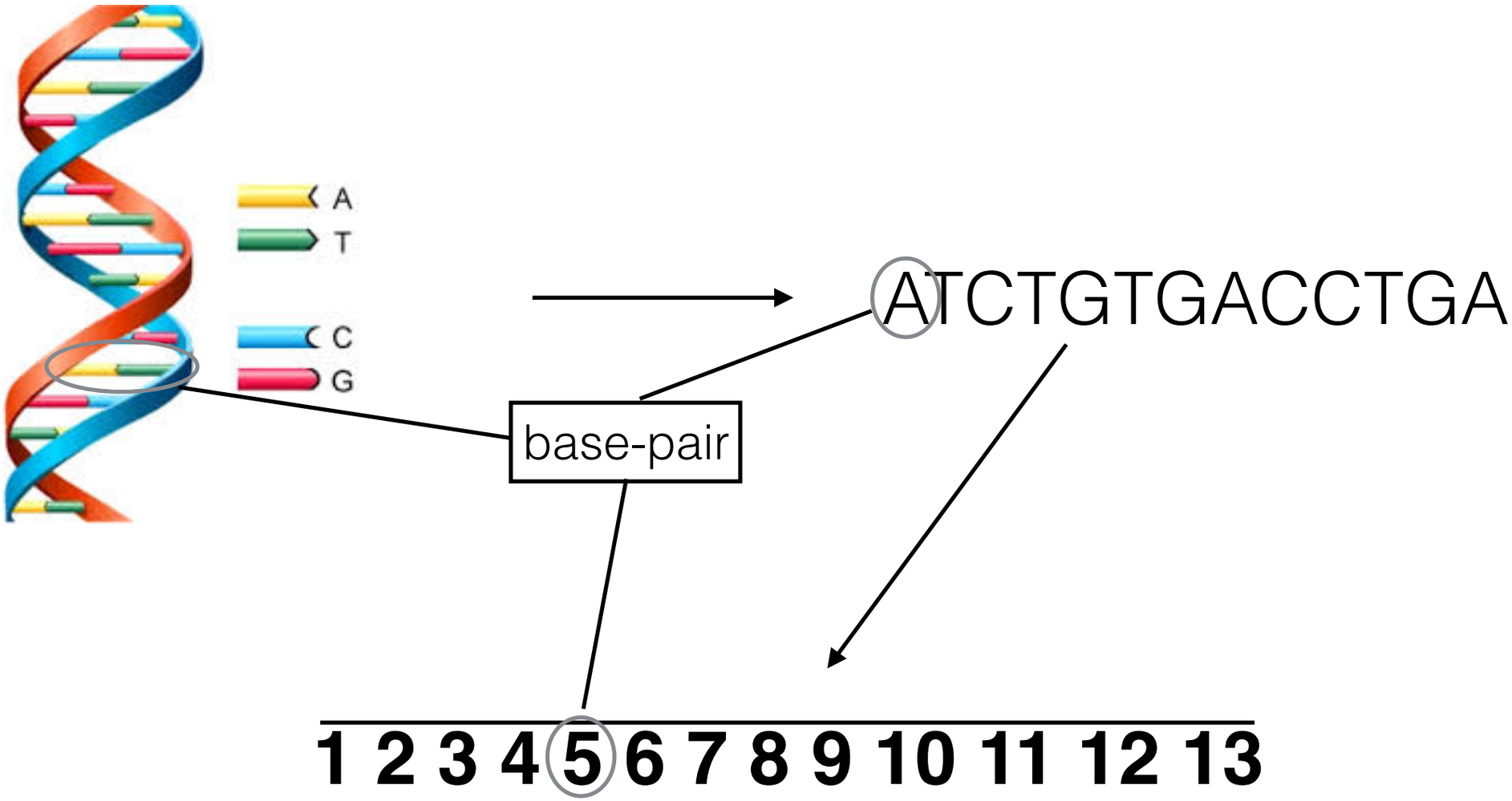
Introduction

What are genes?

This! :



Genome as a line



How to represent genes on the ‘genome as a line’?

chr7	127471196	127472363
chr7	127472388	127473530
chr7	127473555	127474697
chr7	127474701	127475864
chr7	127475893	127477031
chr7	127477121	127478198
chr7	127478300	127479365
chr7	127479375	127480532
chr7	127480538	127481699

What are genes not (in this part of the course)?

- A sequence of base pairs (e.g. ACGTGTC)
  - We only care about start and end positions...
- An identifier (e.g. *BRCA2*), or a list of these
  - We need some positional information
- Pathway nodes (gene -> mRNA -> protein)
  - We only look at what is happening relative to the reference genome as a line

Statistical genomics

- Often used for statistical analysis of:
  - Gene lists (e.g. Gene set enrichment analysis, GSEA)
  - Gene expression (Differential expression)
  - SNPs (e.g. Genome-wide association studies, GWAS)
  - etc..
- We are not going to do any of the above

Statistical genomics

- Statistical analysis of genomic tracks
  - Tracks: genome-wide datasets than can be positioned along a reference genome (DNA)

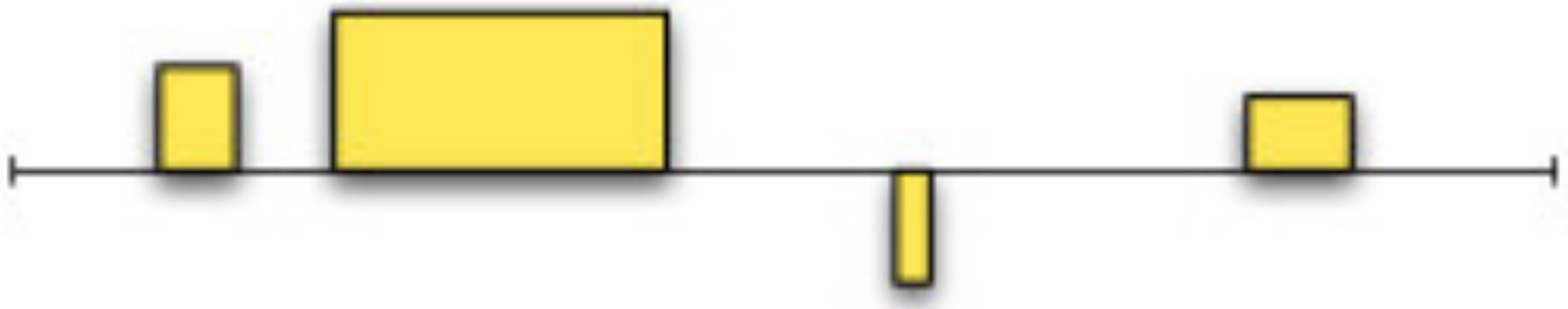
- However:
  - Many of the concepts are central statistical concepts that can be used for other types of analyses

Tracks and track types

Representation of genes

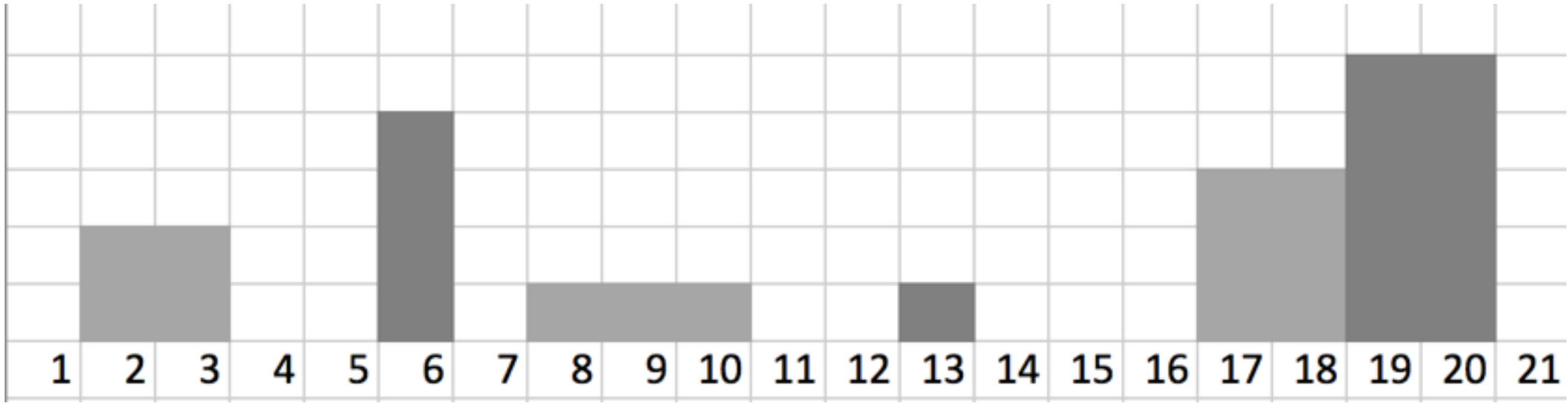
chr7	127471196	127472363
chr7	127472388	127473530
chr7	127473555	127474697
chr7	127474701	127475864
chr7	127475893	127477031
chr7	127477121	127478198
chr7	127478300	127479365
chr7	127479375	127480532
chr7	127480538	127481699

How about gene expression data (RNA-seq)?



chr7	127471196	127472363	17
chr7	127472388	127473530	31
chr7	127473555	127474697	73
chr7	127474701	127475864	13
chr7	127475893	127477031	83
chr7	127477121	127478198	93
chr7	127478300	127479365	29
chr7	127479375	127480532	59
chr7	127480538	127481699	63

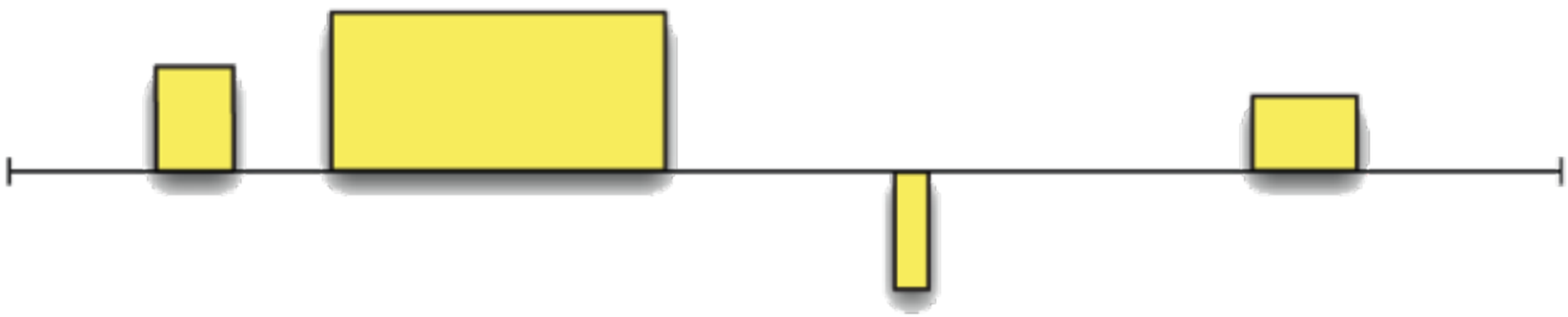
Exercise 1



a) Base-pair count (coverage) 11
b) Coverage proportion 0.52
c) Average segment length 1.83
d) Average gap length 1.43
e) Average value 1.33 per bp
2.54 per bp (only segments)
2.67 per segment

Track types

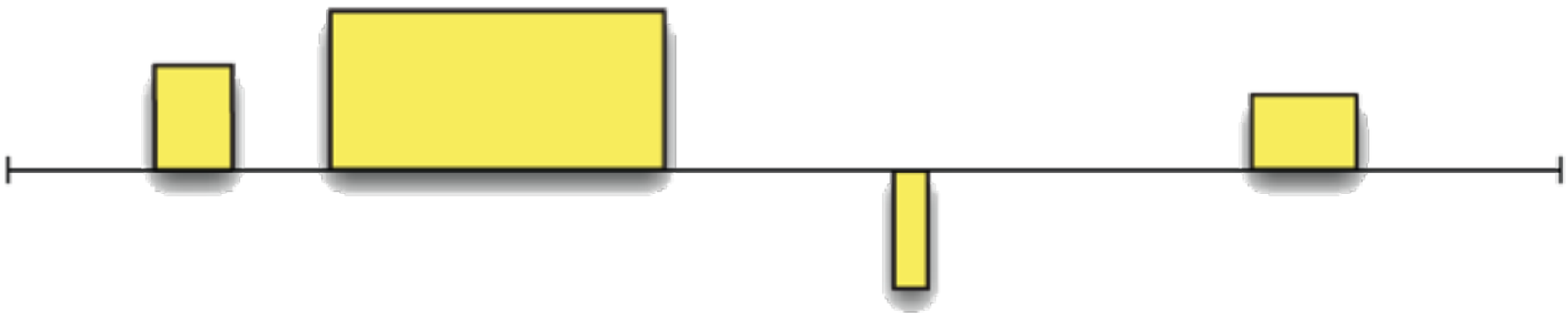
- In the last example, we showed genes as segments on the genome line, with attached RNA-seq read count values
- This track is of a **track type** we call "valued segments"



Valued Segments (VS)

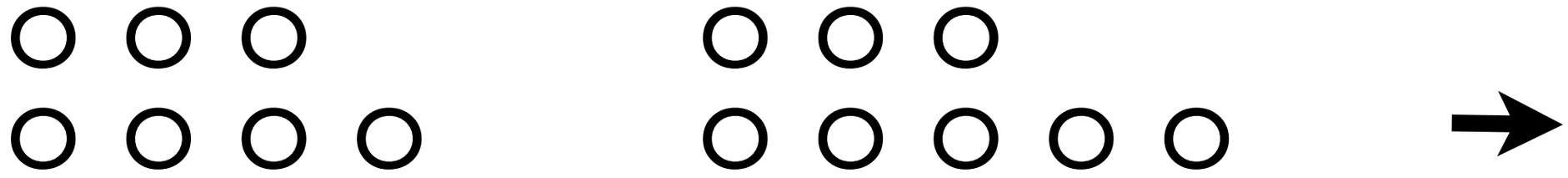
- Track types are mathematical / conceptual models used to categorize track according to their main characteristics

Exercise 2

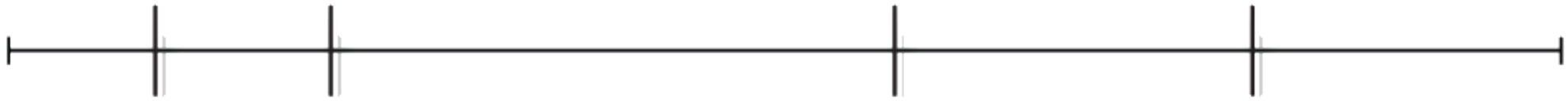


Valued Segments (VS)

- What other **track types** can you think of?
  - Discuss with your neighbour (2-3 min)
  - Classroom discussion

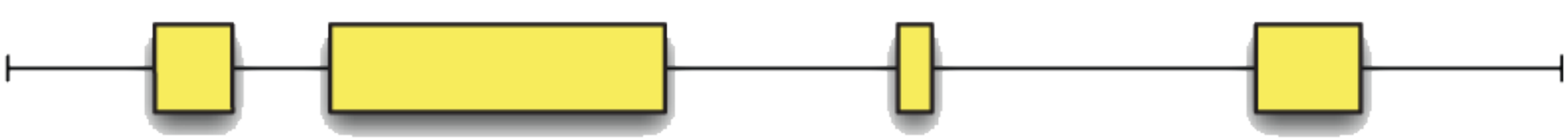


Points



Points (P)

Segments



Segments (S)

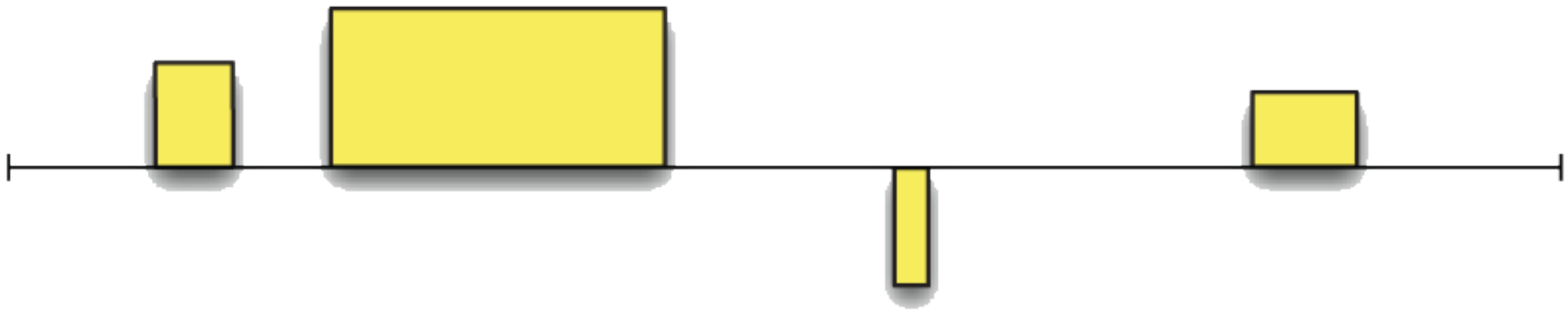
Genome Partition

Genome Partition (GP)

Valued Points

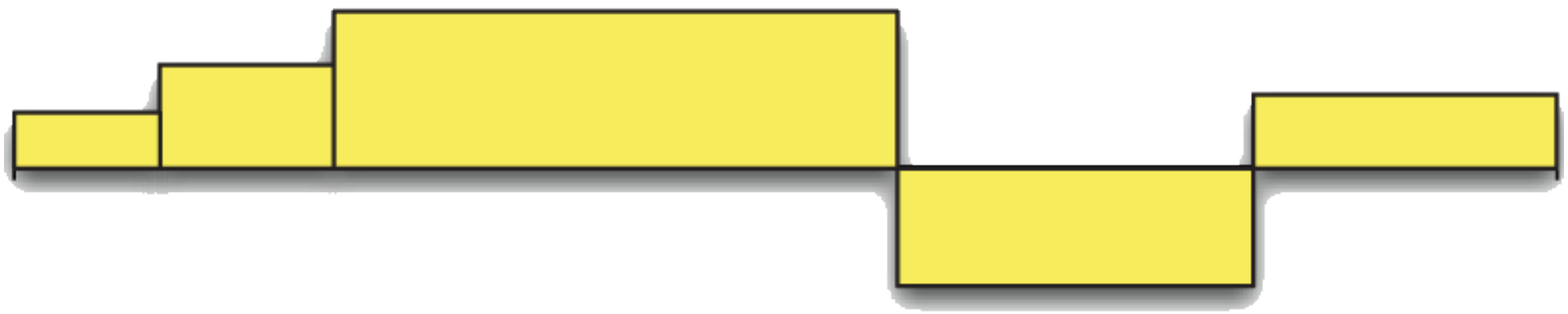
Valued Points (VP)

Valued Segments



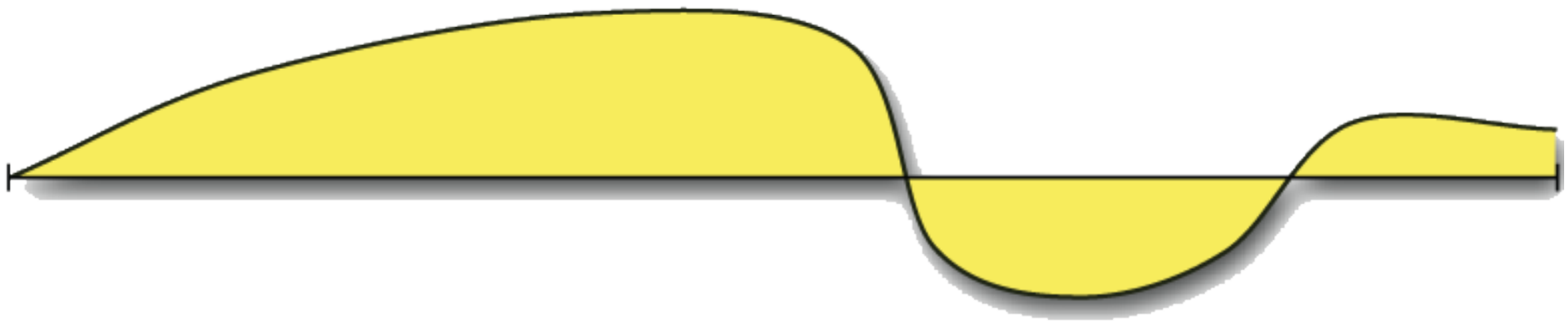
Valued Segments (VS)

Step Function



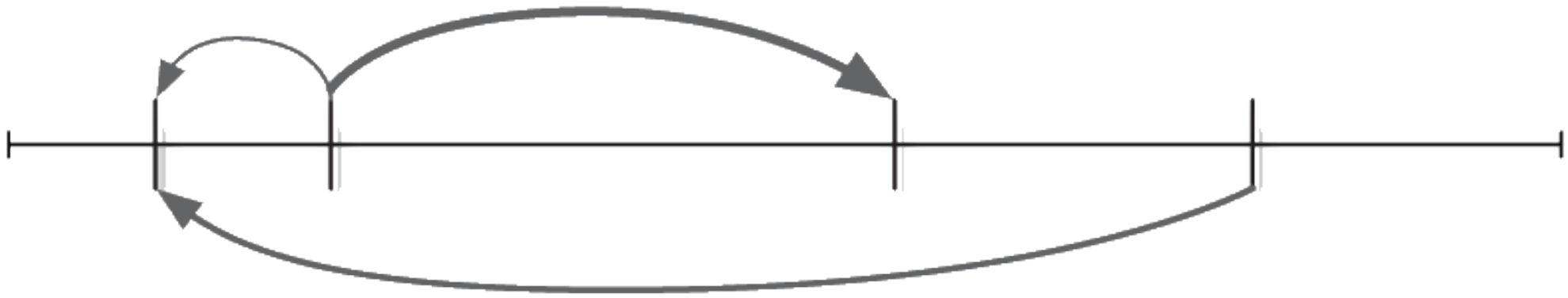
Step Function (SF)

Function



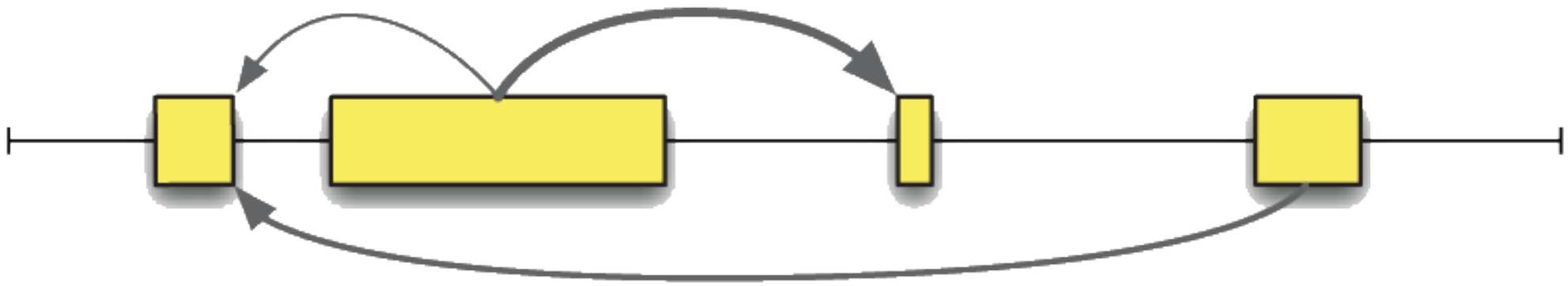
Function (F)

Linked Points



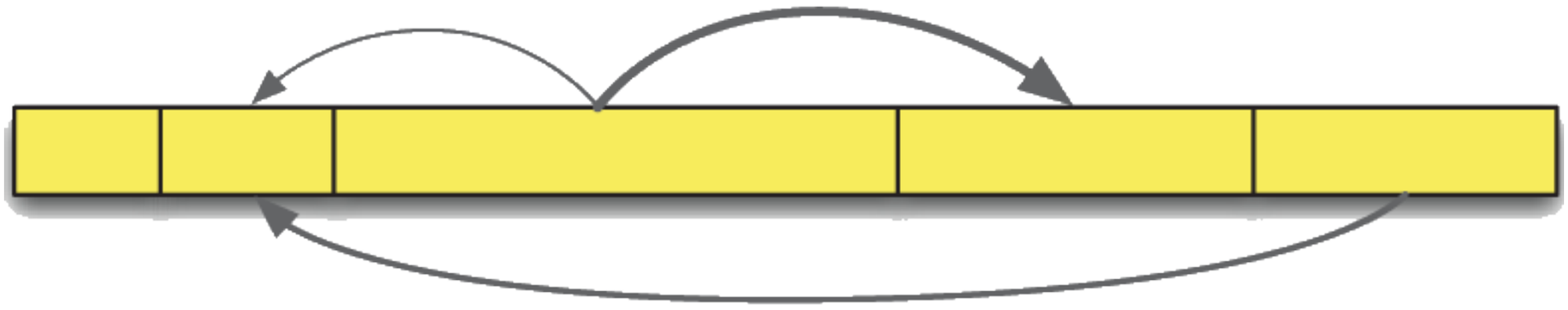
Linked Points (LP)

Linked Segments



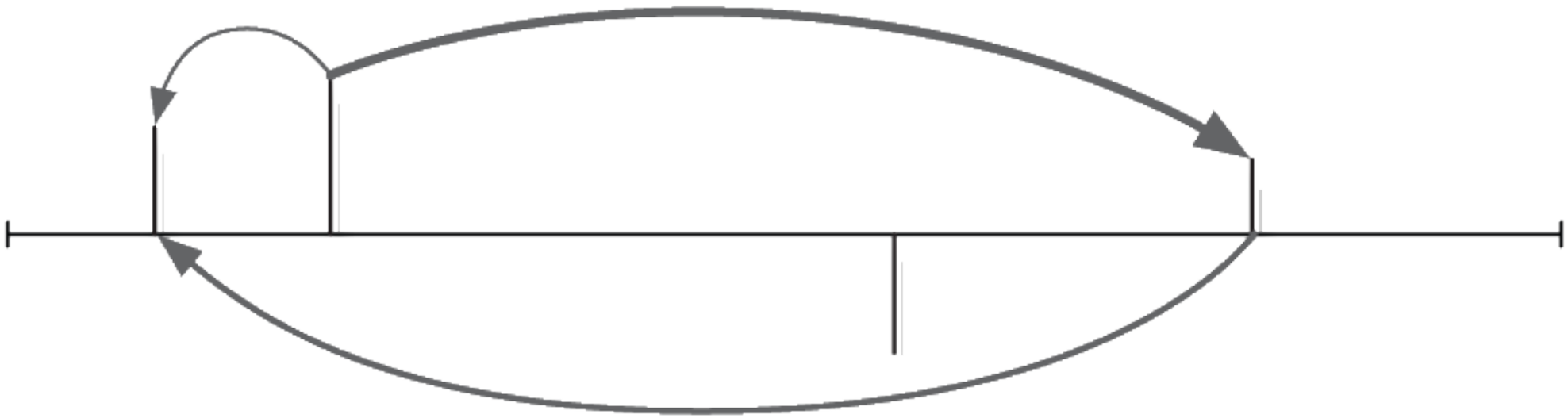
Linked Segments (LS)

Linked Genome Partition



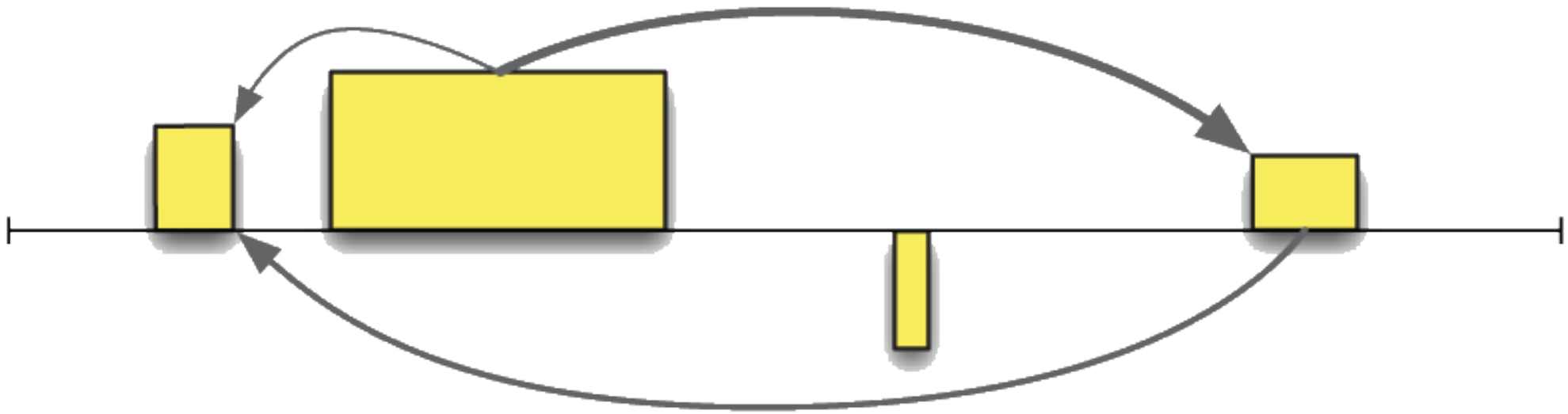
Linked Genome Partition (LGP)

Linked Valued Points



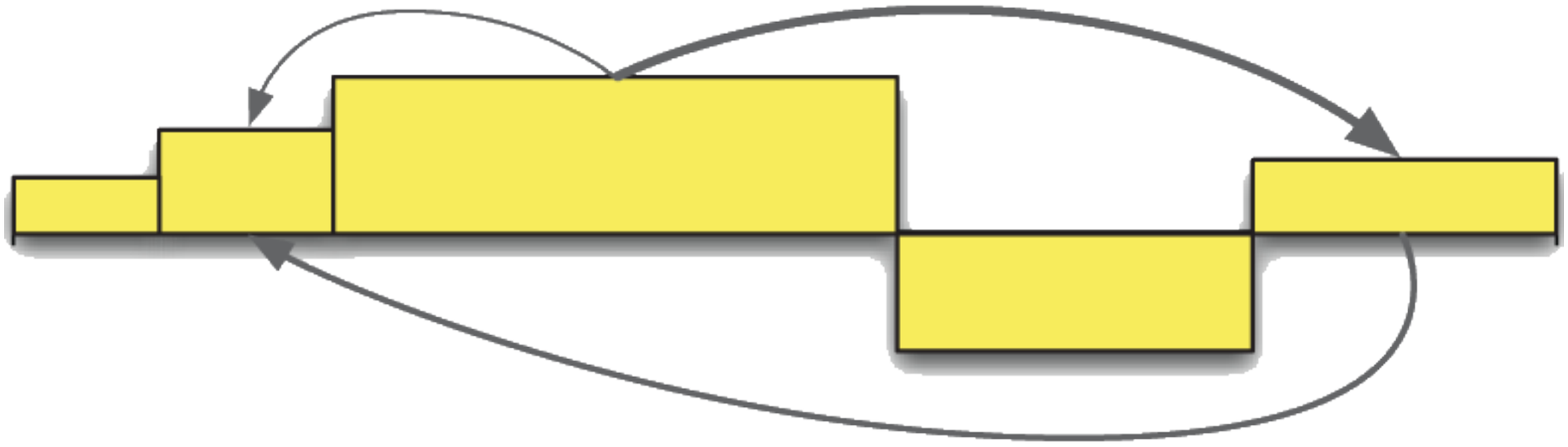
Linked Valued Points (LVP)

Linked Valued Segments



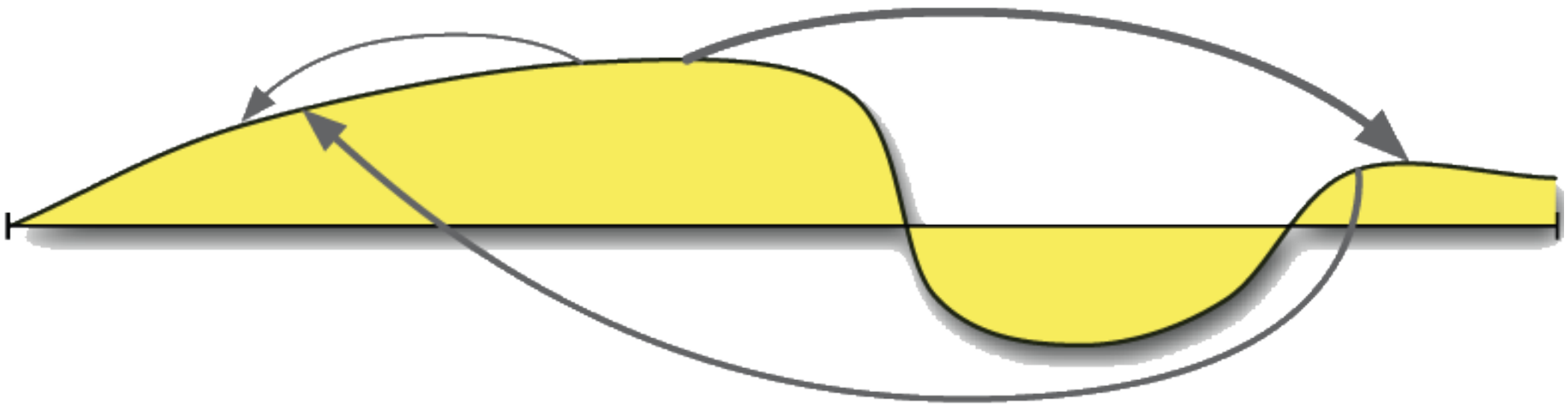
Linked Valued Segments (LVS)

Linked Step Function



Linked Step Function (LSF)

Linked Function



Linked Function (LF)

Linked Base Pairs

Linked Base Pairs (LBP)

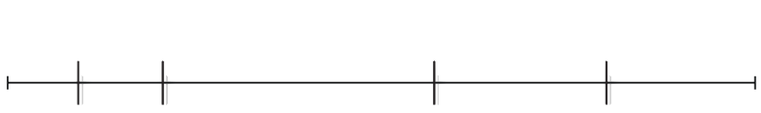
Exercise 3

- Tracks: genome-wide datasets than can be positioned along the a reference genome (DNA)

- Brainstorm: which **tracks** can you think of?

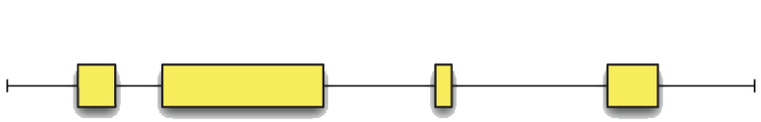
- For each track, which **track type** should be used to represent the data?

Exercise 3

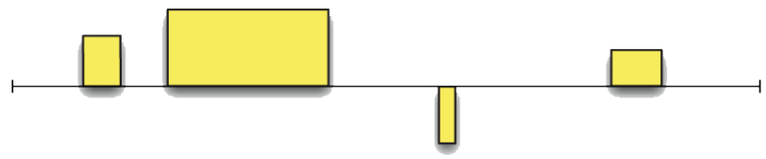


Points (P)

Valued Points (VP)

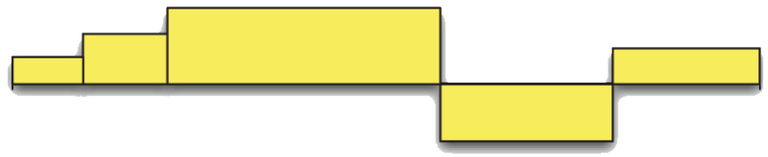


Segments (S)

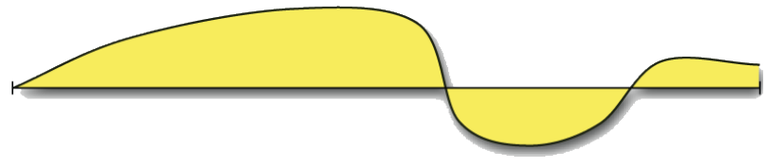


Valued Segments (VS)

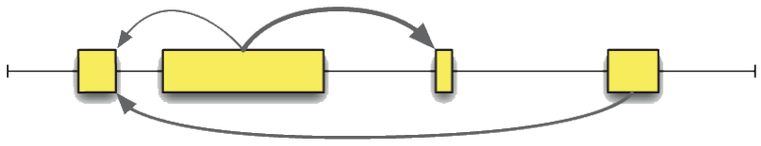
Genome Partition (GP)



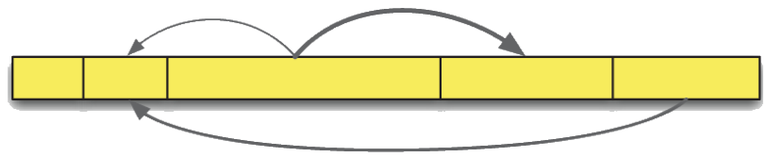
Step Function (SF)



Function (F)

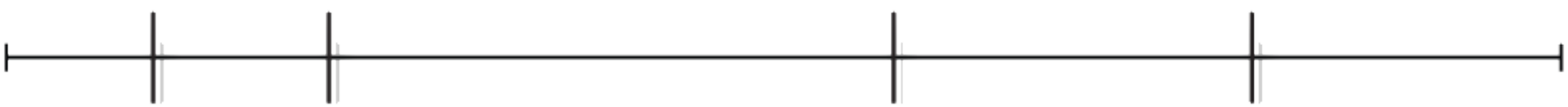


Linked Segments (LS)



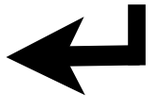
Linked Genome Partition (LGP)

Points

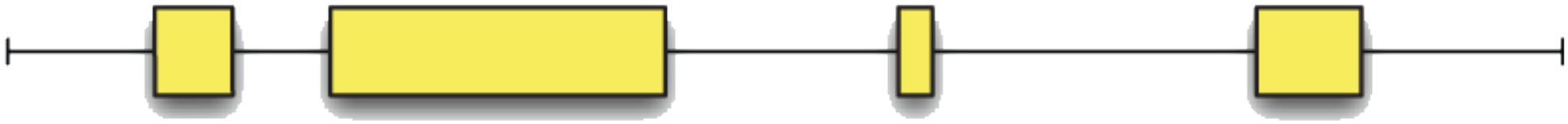


Example tracks:

- SNPs



Segments



Example tracks:

- Splice variants

↵

Genome Partition

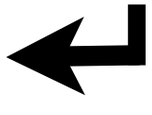
Example tracks:

- Chromosomes
- Euchromatin/Heterochromatin

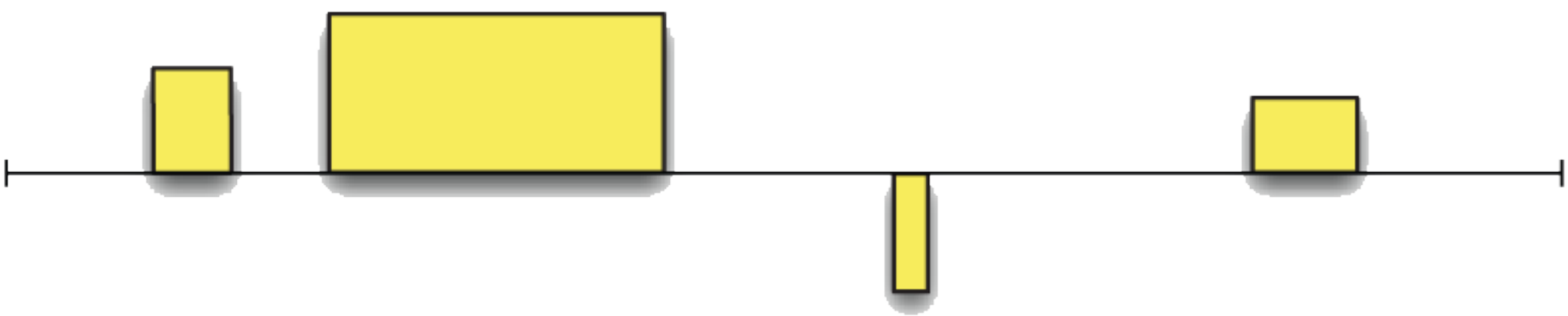
Valued Points

Example tracks:

- SNPs with quality
- SNPs with allele freq



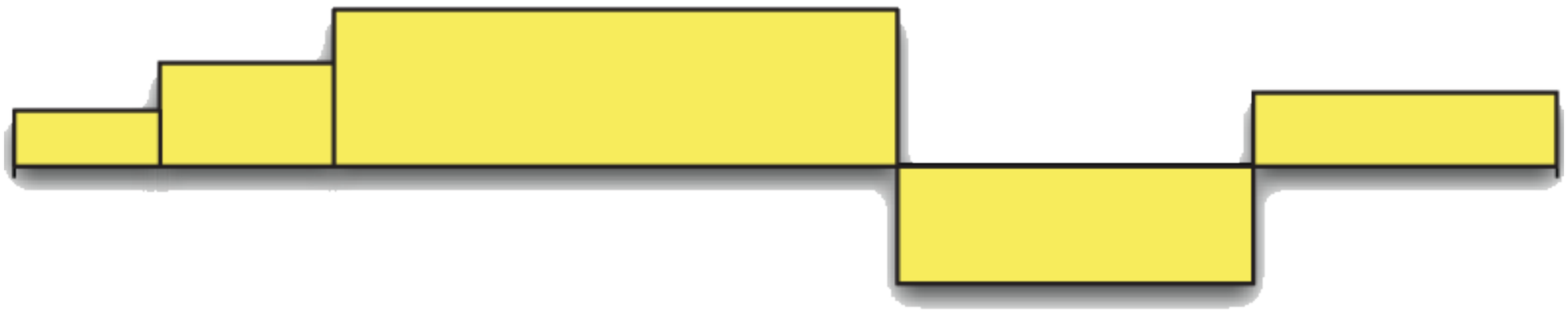
Valued Segments



Example tracks:

- GC content
-

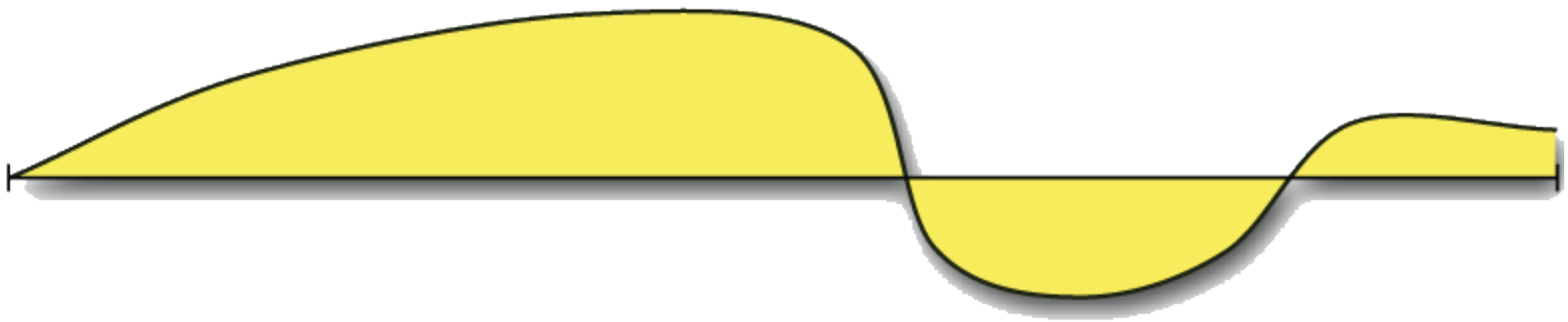
Step Function



Example tracks:

- CNV

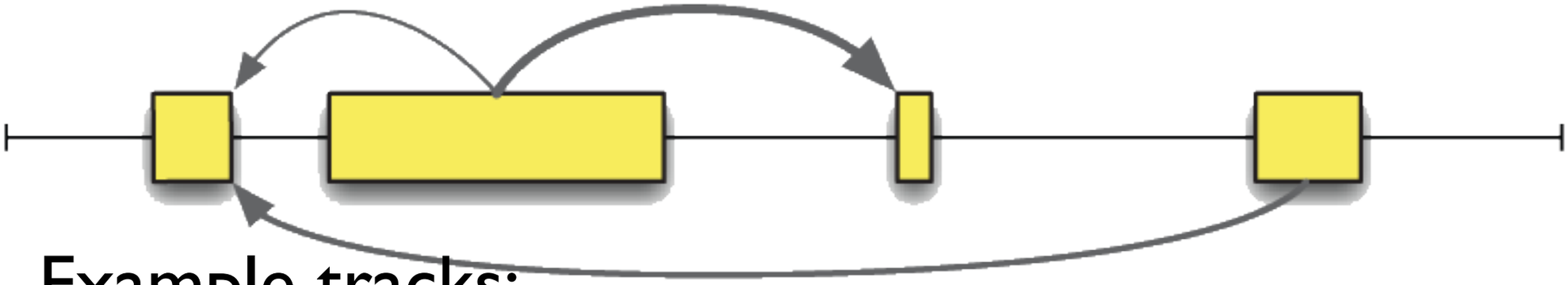
Function



Example tracks:

- DNA melting temperature

Linked Segments



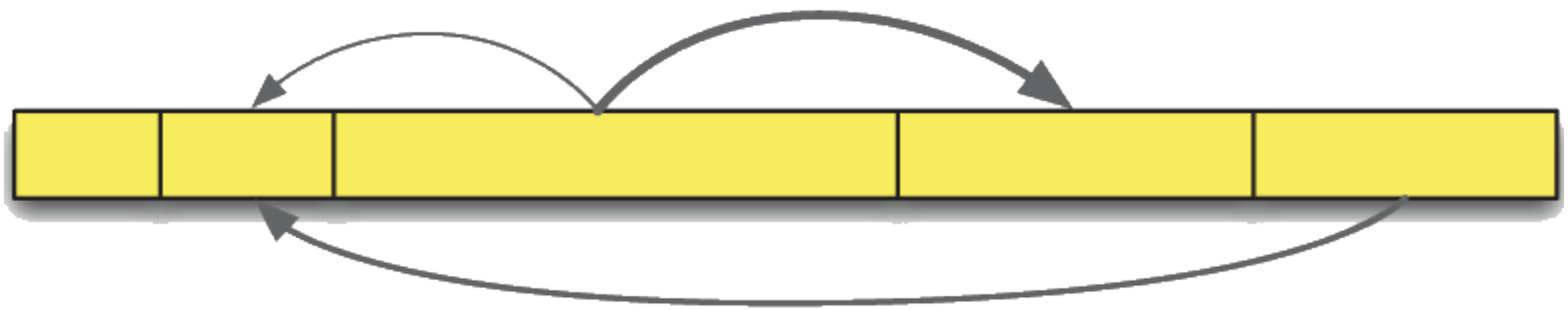
Example tracks:

Splice variants

Gene regulation

Fusion Genes

Linked Genome Partition



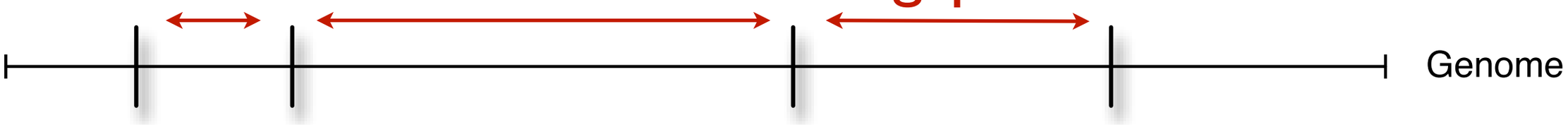
Example tracks:

- Hi-C
-
-

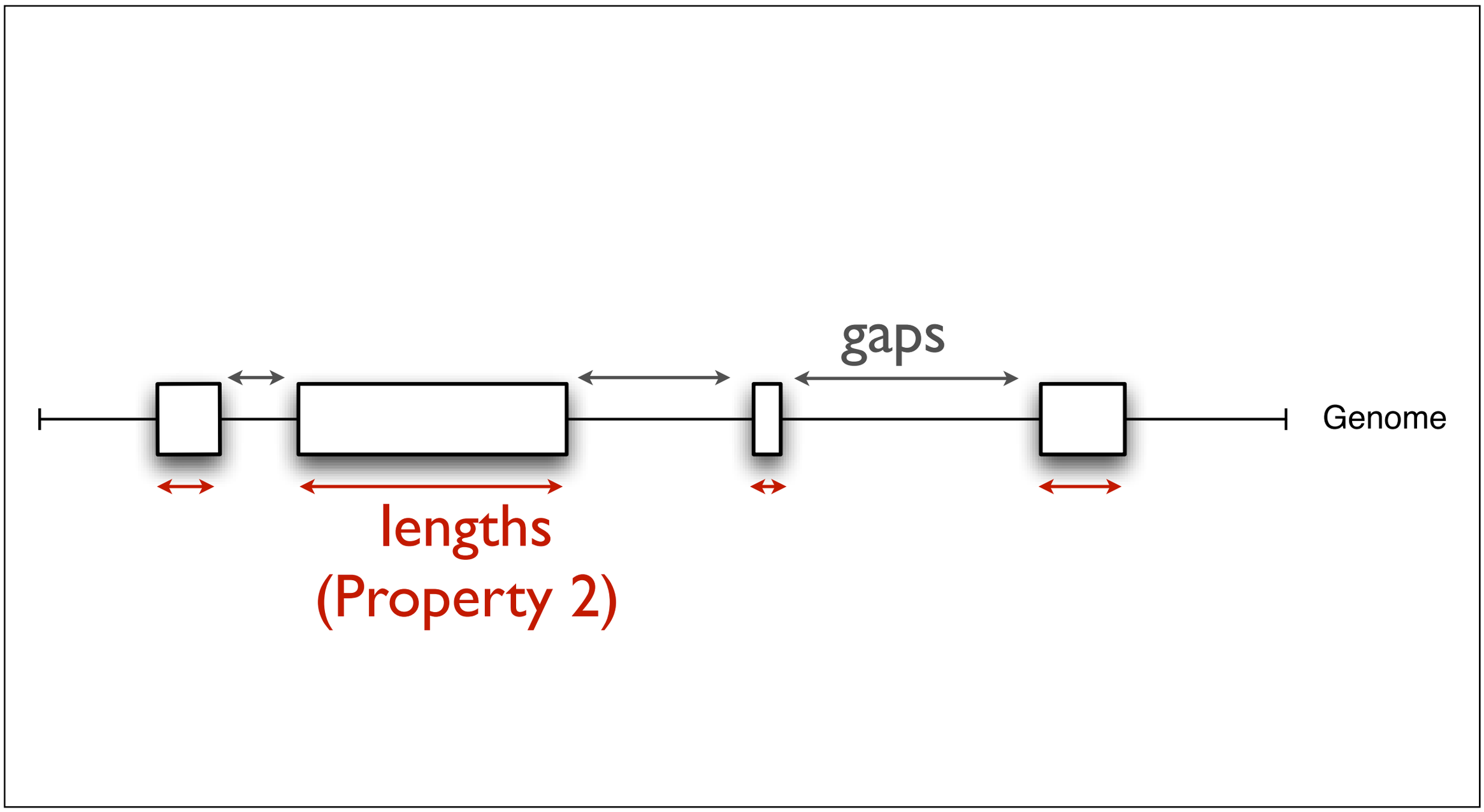
Core properties of tracks

Property 1:

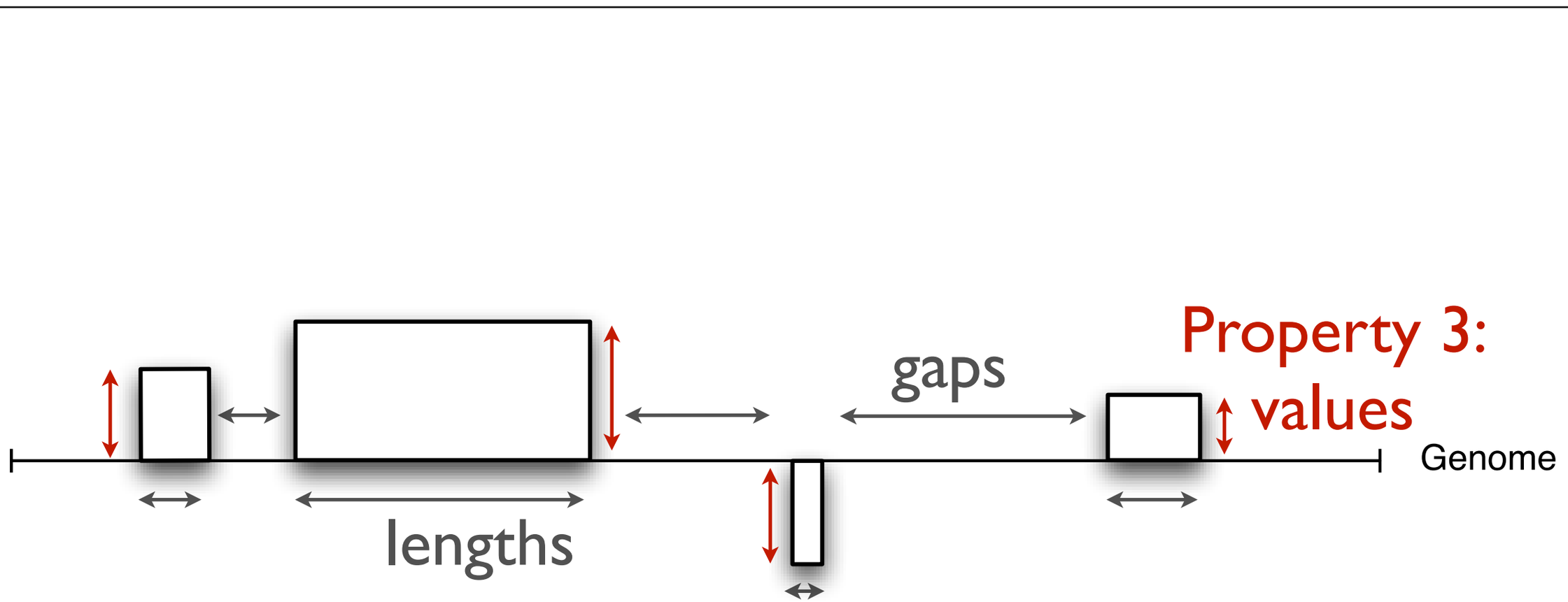
gaps



Core properties of tracks

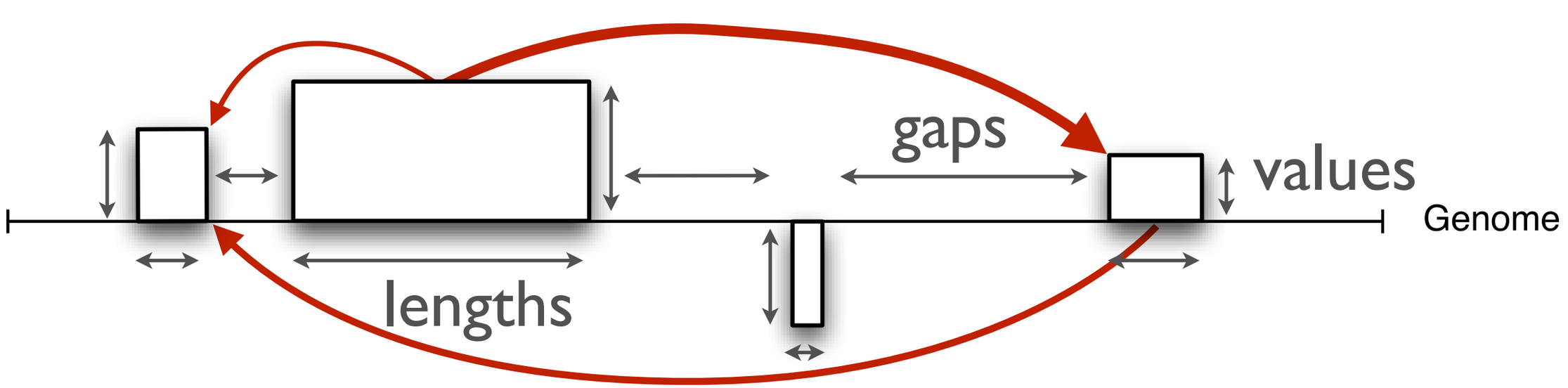


Core properties of tracks



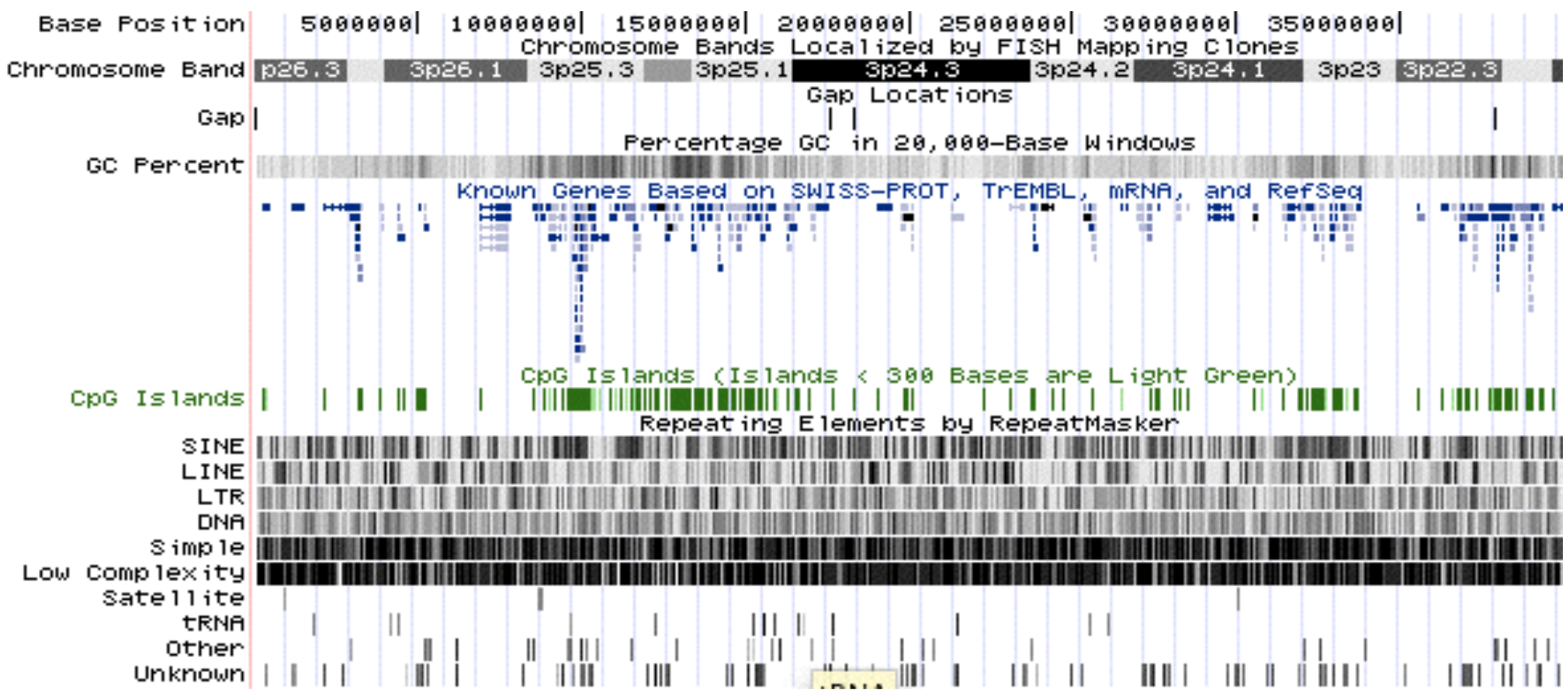
Core properties of tracks

Property 4: interconnections



Tracks in the real world

- Remember the UCSC Genome Browser?
- Each row is a track, and many of the track types are supported



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)

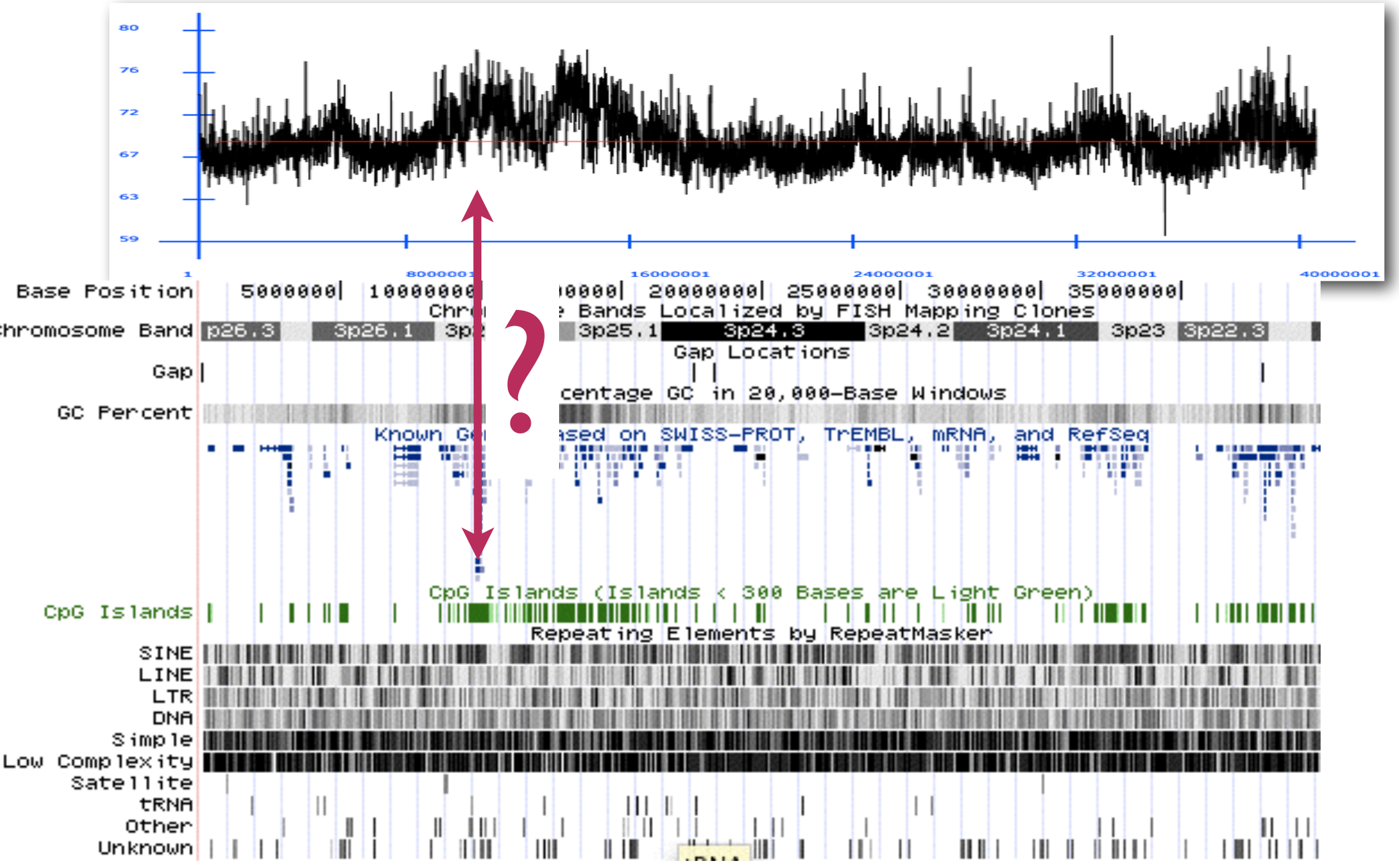
Example analyses (cont.)

- 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-occurrence of genomic features

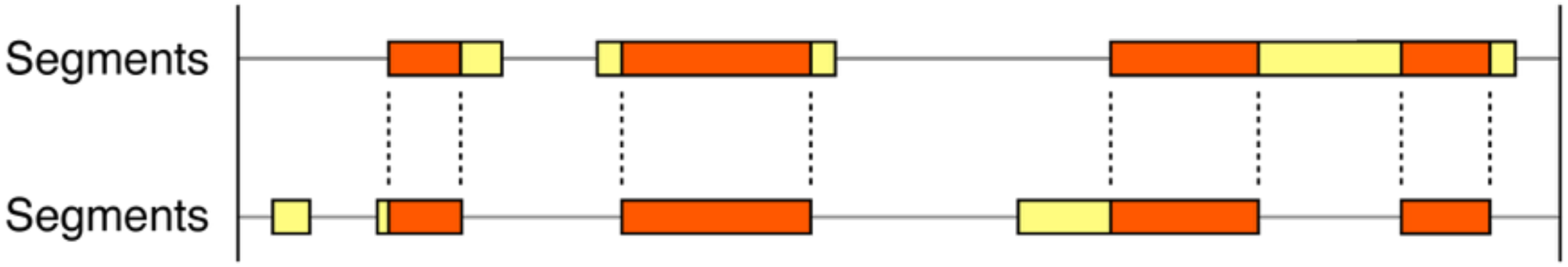
- Typical question:

> *do genomic feature X and Y occur (more than expected) at the same locations in the genome?*

Co-occurrence of genomic features

- What can such analyses be used for?
  - Discover novel relations between tracks (can be done with only public datasets):
    - May e.g. suggest that the biological features represented by the tracks are involved in the same cellular mechanism
  - Relate experimental dataset to existing biological features
    - Compare experimental data with chromatin tracks from different cell/tissue types:
      - In which cell/tissue types does the mechanism in question happen?

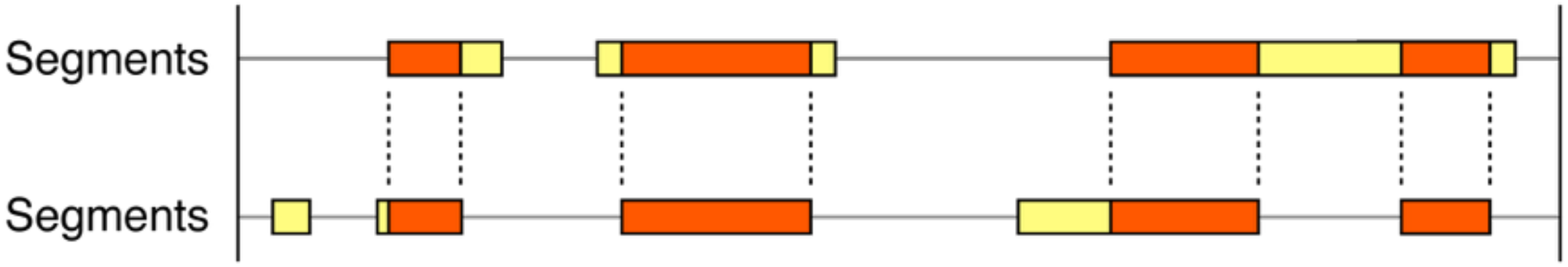
How does this look at the whiteboard?



overlap > expected?

- As evident, this analysis makes sense when you have two tracks of type “segments”
- Generally, the type of analysis is dependent of the track types:
  - Each single track type defines a set of analyses appropriate for that track type (e.g. counting, coverage)
  - Each pair of track types defines another set of relational analyses (e.g. overlap, correlation...) specific to that combination

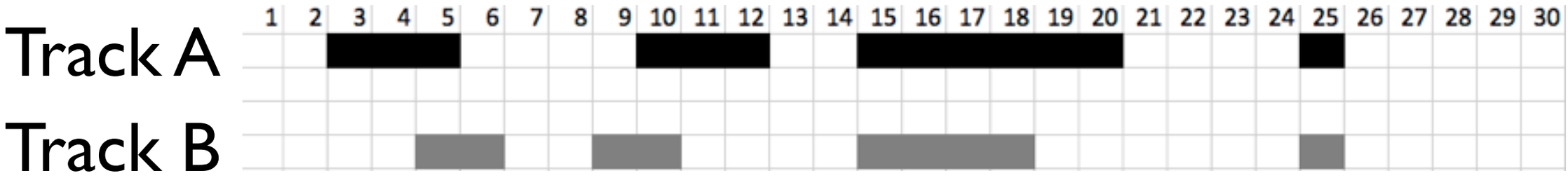
How does this look at the whiteboard?



overlap > expected?

What now?

Exercise 5

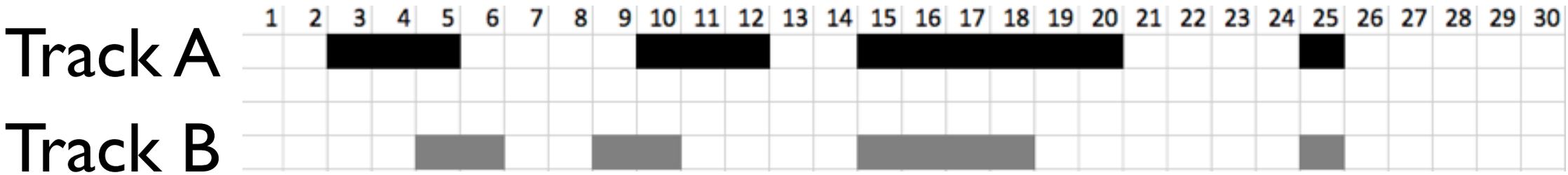


Calculate:

a.	the number of overlapping base-pairs between tracks A and B	7
b.	the proportion of overlapping base-pairs (in respect to the genome)	23.3%
c.	the expected number of overlapping base-pairs (assuming independent tracks)	3.9
d.	the proportion of observed to expected overlap (= a type of enrichment)	1.8

What conclusion can you draw from the results?

Exercise 6a



Create a random control track for track B, by

a) Take each (grey) base pair and move it to a random location (do not keep existing segments)

Exercise 6a

- What is the overlap between the original track A with your random control B track?
- Let's build a histogram of your results
- How extreme is the original observation?
- If we count the proportion of boxes that are more extreme, we have the p-value

Hypothesis testing

What you will learn:

- What hypothesis tests are and why they are useful
- How to perform a simple hypothesis test
- How you can “investigate” a biological question by using a hypothesis test

What is a hypothesis test?

- From Wikipedia:
  - *An hypothesis test is a statistical test that is used to determine whether there **is enough evidence in a sample of data** to infer that a certain condition is **true for the entire population**.*
- Based on some data, infer whether a condition is true
- Typically consider the probability of the condition NOT being true based on the data (this is a p-value)

What is a hypothesis test? (cont.)

- Example:
  - Someone claims that they can guess the outcome (head or tail) when a fair coin is flipped.
  - Do some trials, investigate whether this is true
  - You throw the coin 5 times, and the person guesses correct every time.
    - What is the probability of the claim being false?

Example, more formally

- H_0: **Null hypothesis**. This is what we believe until proven wrong
  - *The person is not able to predict tail or head*
- H_1: **Alternative hypothesis**, that we want to investigate
  - *The person is able to predict tail or head for this coin*
- **Test statistic**: *Measure that best captures the phenomena you are investigating (e.g. overlap)*
- ***P-value**: Given H_0 being true, it is the probability to actually observe the observed data. If this probability is small enough, we reject the null hypothesis, and conclude H_1.*
  - *In our example: p-value = 0.5^5 = 0.031*

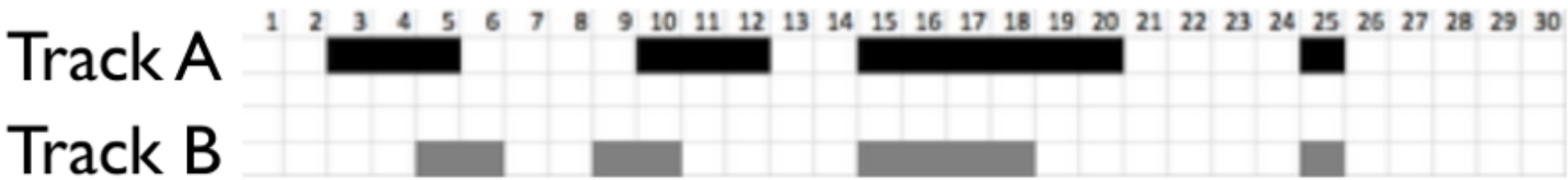
Why use hypothesis tests?

- Sometimes hard or impossible to make conclusions without.
  - What if the person guessed correct 520 out of 1000 times?
  - Even harder when working with biological data

- A hypothesis test quantifies the certainty of concluding a hypothesis (p-value)
  - For some cases, a very small p-value might be requested, e.g when concluding on the effect of a drug

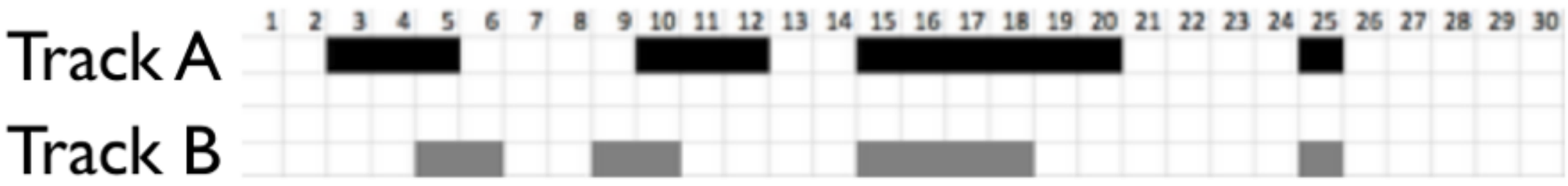
Null model

- A null model is the model in which the null hypothesis arises from
  - The “base case” where we assume the condition in the null hypothesis is true.
- In the case of the coin flips, the null model is simple:
  - In the null model we assume it is not possible to predict outcome, so guessing correct result has probability 0.5
- A claim with a less simple null model:
  - **Claim**: A genomic track co-occurs (more than expected by chance/coincidence) with another genomic track



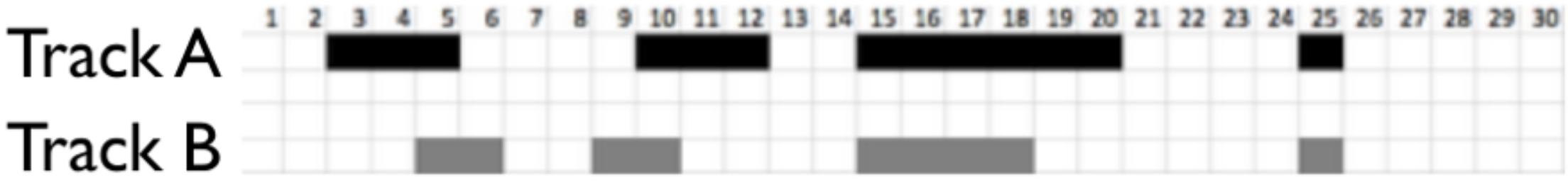
A more “complicated” claim

- **Claim:** The two genomic tracks (in the picture) co-occur (more than one should expect by random tracks to do)
  - What is the null hypothesis?
  - What is the null model?
- How can we compute the p-value in this case?
  - We can measure the co-occurrence and find the probability that this co-occurrence would be found in the null-model (where there is no association)



Monte Carlo

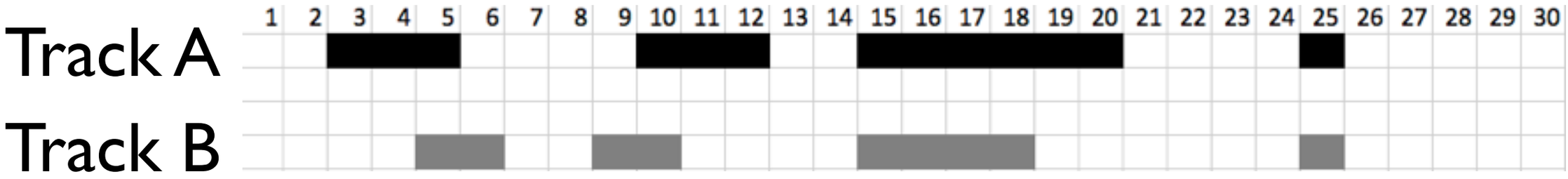
- Simulate many samples from the null model
  - E.g. many pairs of tracks following the same properties
- For each simulation, compute the co-occurrence
  - E.g. the number of base pairs overlap
- Compute how often the co-occurrence found using the null model was as extreme or more extreme than the co-occurrence found in our observation
  - If this happened rarely (e.g. < 0.5 % of the times), we conclude there is an association (with significance level 0.005)



How to make random samples in this case?

- Preservation of structure in data
  - Should reflect the combination of stochastic and selective events that constitutes the evolution behind the observed genomic feature
  - Reflect biological realism, but also allow sufficient variation to permit the construction of tests

Exercise 6b



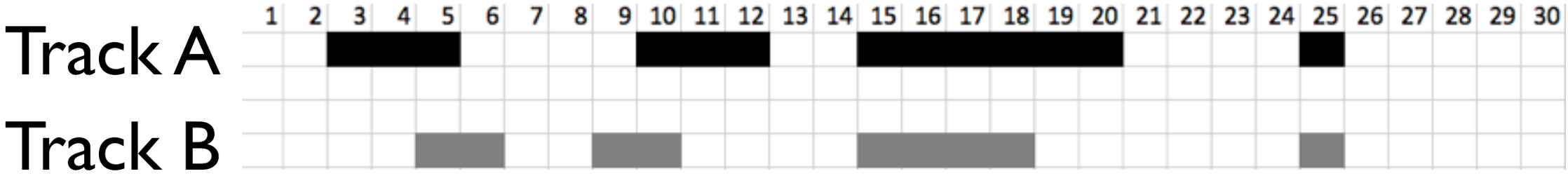
Create a random control track for track B, by

a) Take each (grey) base pair and move it to a random location (do not keep existing segments)
b) **Take each segment and move it to a random location (preserving segment lengths)**
c) **Preserve segment and gap (inter-segment) lengths, randomize order**

Exercise 6b

- What is the overlap between the original track A with your random control B track?
- Let's build a histogram of your results
- How extreme is the original observation?
- If we count the proportion of boxes that are more extreme, we have the p-value

Remember this?



Calculate:

a.	the number of overlapping base-pairs	7
b.	the proportion of overlapping base-pairs (in respect to the genome)	23.3%
c.	the expected number of overlapping base-pairs (assuming independent tracks)	3.9
d.	the proportion of observed to expected overlap (= a type of enrichment)	1.8

What conclusion can you draw from the results?

Null models

- Examples of preservation strategies
  - Preserve segment length (already seen this)
  - Preserve segment and gap length (this too)
  - For points (segments with length 1)
    - Preserve point count
    - Preserve inter-point distance
  - For all this cases we randomize the position of the track elements.

Association vs. causation

- Association: A & B are related, show up together.
- Causation: A causes B
- Using statistical testing, we can only find whether there is an association
- Causation requires speculation, biological understanding, experimentally determined mechanisms

Multiple testing

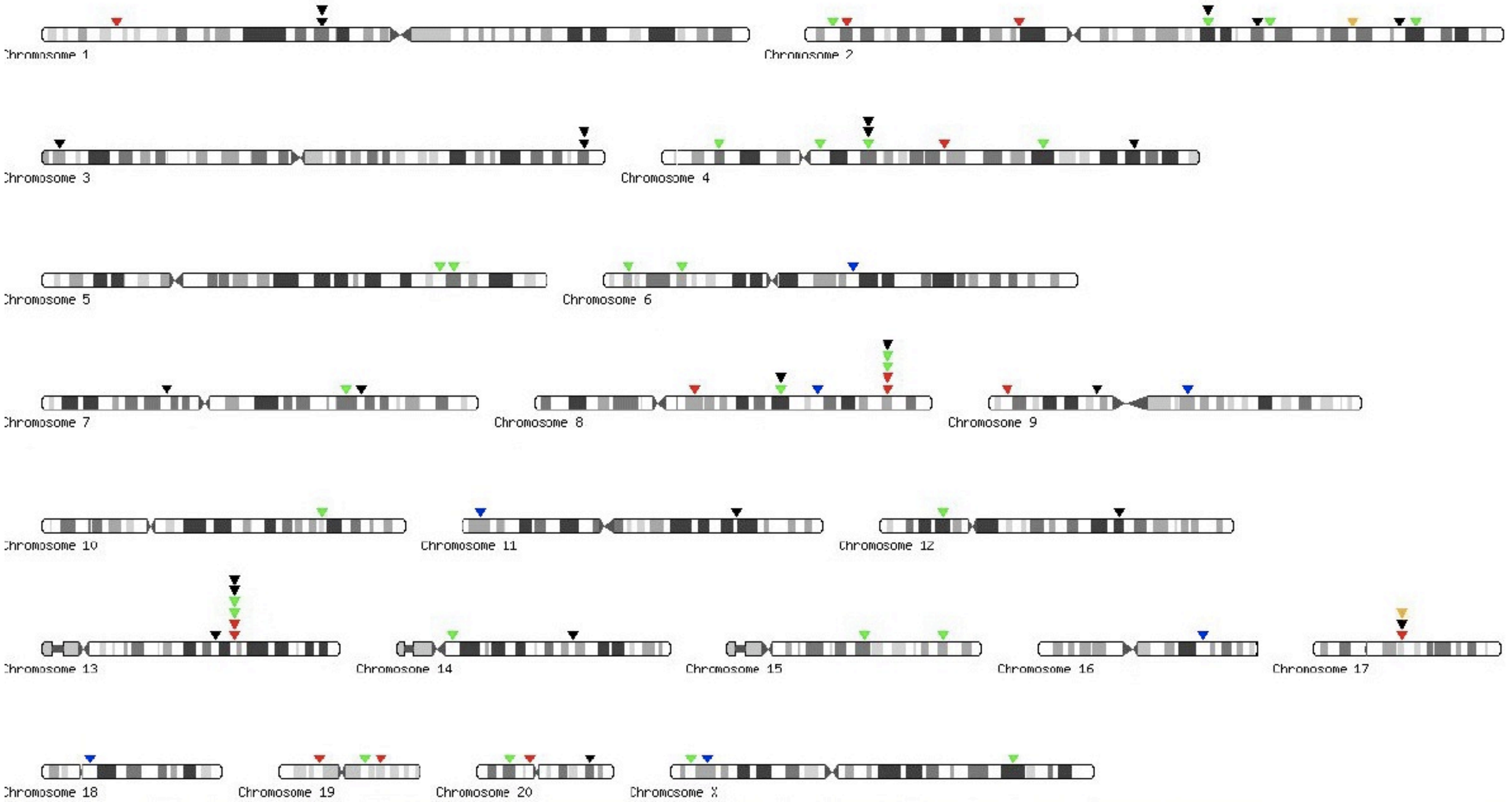
- Bonferroni correction
  - Assume all tests are independent
  - Too conservative (less power)
  - For m tests, multiply each p-value by m

Multiple testing (FDR)

- False discovery rate
  - Control the proportion (δ) of false positives among the set of rejected hypotheses
  - Order the unadjusted m p-values
  - Find the test with the highest rank, j, for which the p-value, p , is less than or equal to (j/m) x δ

Interpreting a claim

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



HPV integration sites

Interpreting a claim

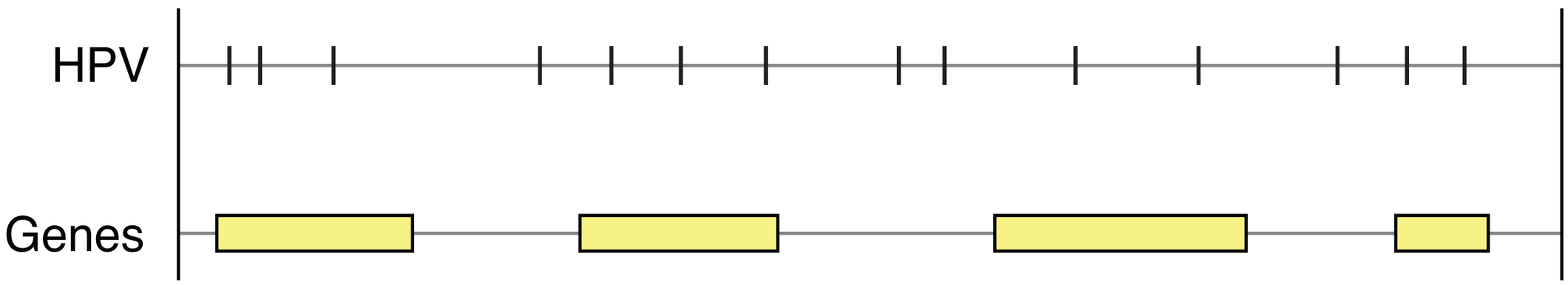
"Viruses might be expected to integrate near genes. Our results confirm such preferential localization inside genes for HPV [Human Papillomavirus], where 75 out of 119 determined integration sites (attached) fall inside genes."

How would you go forth in reproducing such a claim?

Which tracks do we have? What are their track types?

Exercise 7: HPV and genes

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



Note down (in silence):

1. Which test statistic would you choose?

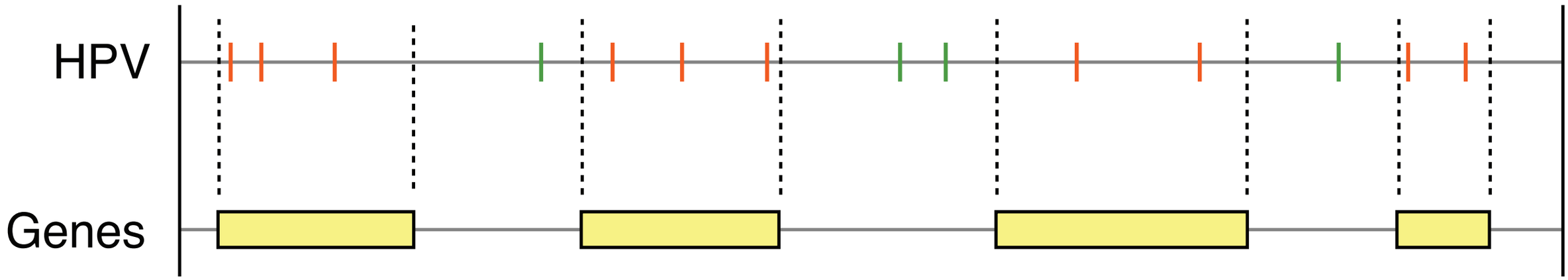
Exercise 7: HPV and genes

Student answers:

1. Which test statistic would you choose?

The overlap between HPV sites and expanded genes	5	
The overlap between expanded HPV and genes	0	
Proportion of HPV sites within genes	2	
The number of HPV sites within genes	0	
Proportion of HPV within vs outside	9	

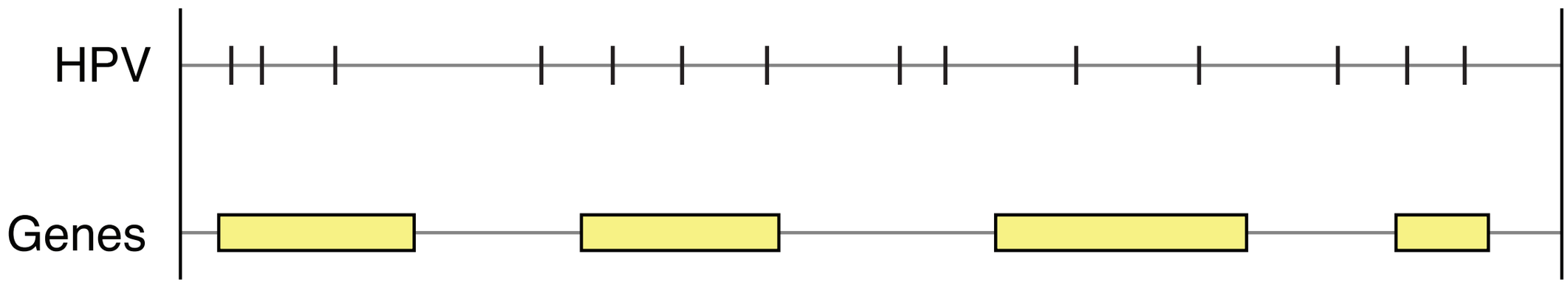
A possible test statistic



- Count number of points of track 1 (HPV) that are located inside segments of track 2 (Genes)

Exercise 8: HPV and genes

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



Note down (in silence):

2. Which null model would you choose?
   a) Which track to randomize?
   b) What to preserve / randomize?

Null models for segments:
- Preserve segment length
- Preserve segment and gap length

For points:
- Preserve point count
- Preserve inter-point distance

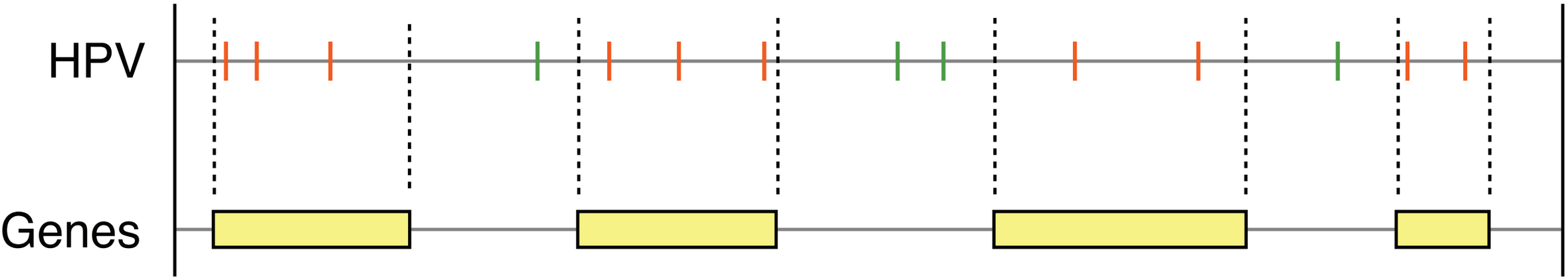
Exercise 8: HPV and genes

Student answers:

2. Which null model would you choose?

Randomize HPV points	10	1
Randomize HPV, but nothing before the first or after the last		
Randomize HPV, keep interpoint distances	2	4
Keep genes, shuffle HPV (keep number of HPV)		
Shuffle genes (keep lengths and inter lengths), shuffle HPV (keep number)		
Randomize genes, preserve lengths	1	2

Exercise 9: HPV and genes



Test statistic: Count number of points of track 1 (HPV) that are located inside segments of track 2 (Genes)

- Go to the Genomic HyperBrowser (https://hyperbrowser.uio.no), using Firefox
- Register a new user (User->Register, top right corner)
- Go to Statistical analysis of tracks -> Analyze genomic track, in the left hand menu
- Genome: hg19
- Track 1 (HPV): Phenotype and disease associations:
  Assorted experiments:Virus integration, HPV specific..
- Track 2 (Genes): Find yourself
- Figure out the rest yourself
- **NB:** Set random seed to 0 (so that you can compare results)
- **NB2:** MC stands for Monte Carlo. Use a Monte Carlo null model and set the sampling depth to "Quick and rough"

Exercise 9: HPV and genes

Student answers:

Which p-values did you get? Which null model did you use?

Preserve seg, number of points, randomize points	Ensembl	0,006
Preserve points, seg length, gap length, randomize segs	Ensembl	0,015
Preserve seg, number of points, randomize points	Refseq	0,507
Preserve seg, number of points, randomize points (MC)	Ensembl	0,02
Preserve seg, number of points, randomize points (MC)	REfseq	0,445
Preserve seg, number of points, randomize points (MC)	Ensembl	0,049
Preserve seg, preserve number points, randomize (MC)	Refseq	0,09

Recap from day 1

- Genomic tracks. Track types.
- Analysis. Co-occurrence of genomic features
- Hypothesis testing. Monte Carlo. Null models.

How much of the human genome is covered by genes?

Exercise 10: descriptive statistics

- Use HyperBrowser again
- What is the coverage (base-pair count) of the different **gene** tracks?

  RefSeq: 1 216 642 705

  Ensembl: 1 539 666 812
- What proportion of the genome do they cover?

  RefSeq: 0.4254

  Ensembl: 0.5383
- What is the number of mutual base-pairs of the different **gene** tracks?

  1 196 508 344 (41.84%)

Descriptive statistics

- Now you actually carried out the analysis in the opposite order than what is recommended
- You should first use descriptive statistics to get to know the datasets before defining and testing your hypothesis
- Visualizing your data in different ways is often very helpful for understanding it

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. how strict definition of a gene one should use
  - One 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

Making justified choices is indeed hard (3)

- Selecting a null model is a very important step, that often has large consequences for the results
  - You always assume a null model when doing hypothesis tests, for instance “assuming a normal distribution”
  - In bioinformatics articles, it is an often overlooked step
  - At the minimum, it should be possible to infer the null model from e.g. the type of test, but it is always better to state it explicitly
  - Much better is actually discussing the assumptions of the hypothesis tests from biological and statistical points of view

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)

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)
- Always:
  - Double-check with a statistician (and a biologist, if you are not one)

Further into statistical details

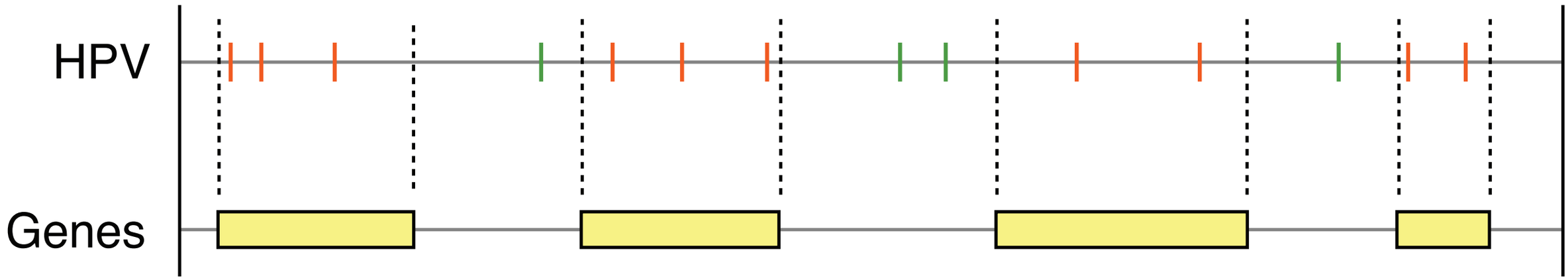
Further into statistical details: the test-statistic

- 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

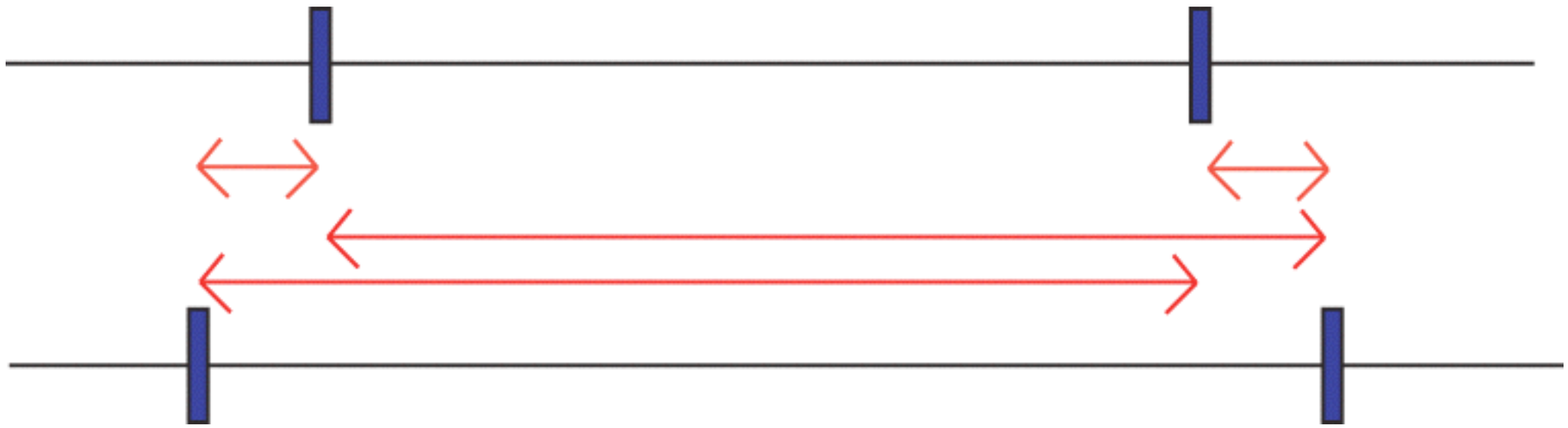
Back to the whiteboard: the test-statistic



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

Back to the whiteboard: Must quantify “close”

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

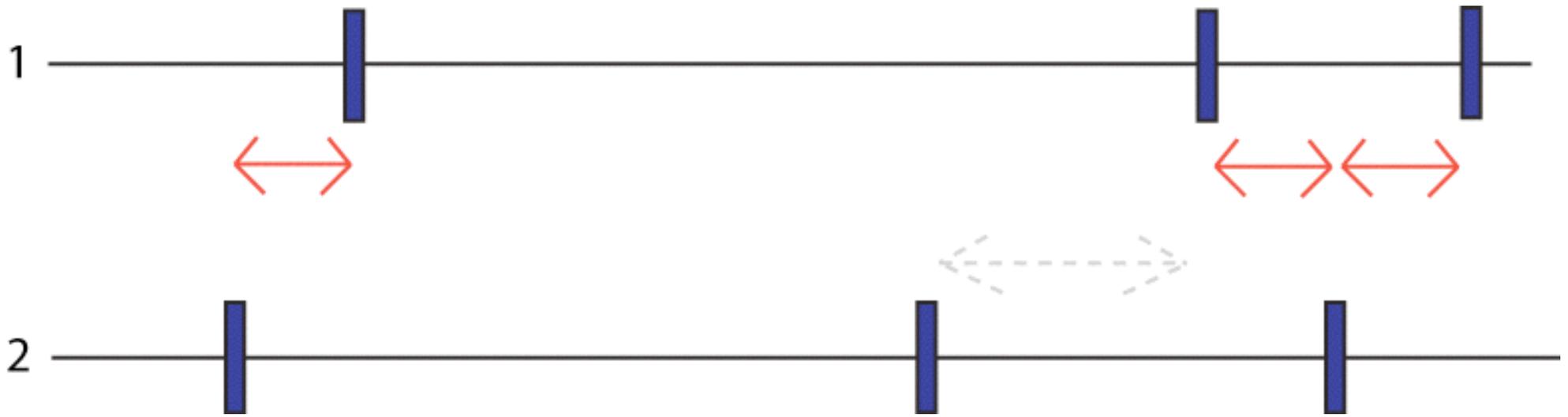


- But which distances - not all vs all?!

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!

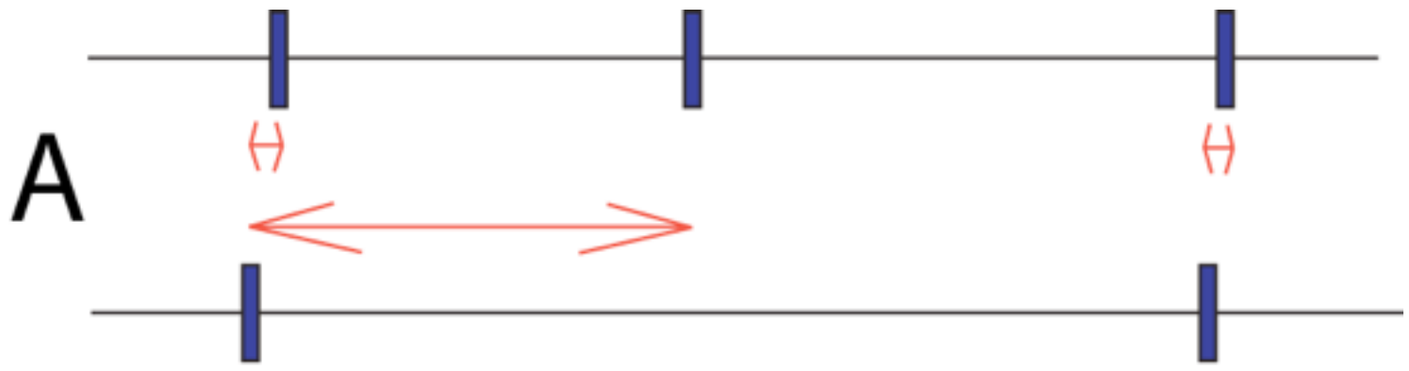
But that's trivial, sure: Just count bp distance!?

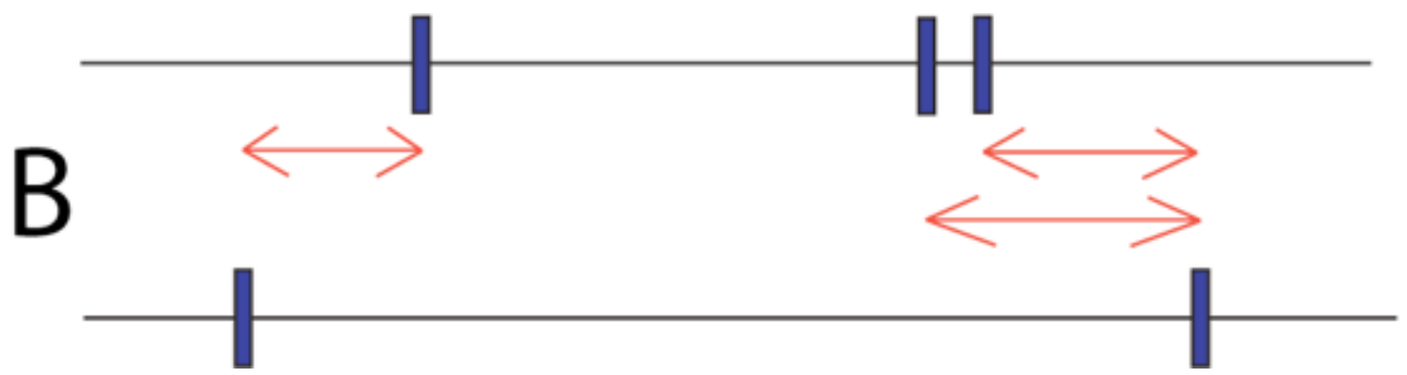
- But which distances - not all vs all?!
  - Only shortest! From 1 to 2! But MC needs a single number..

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!?

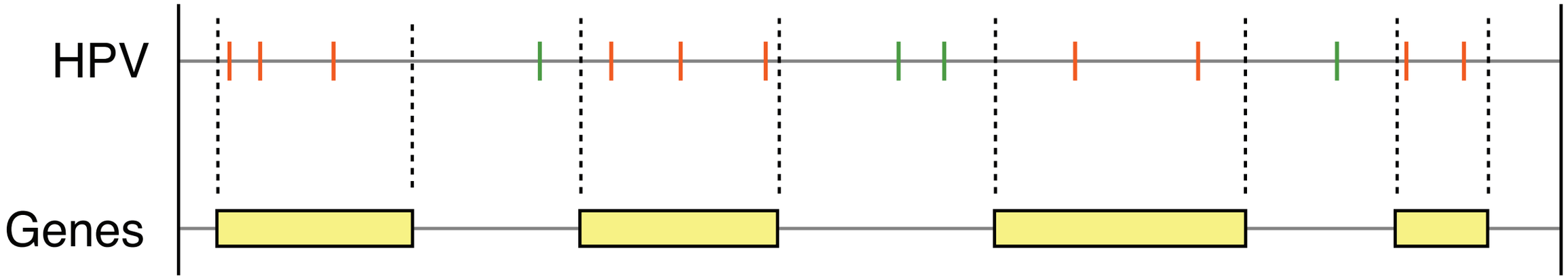
Same degree of closeness?!





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

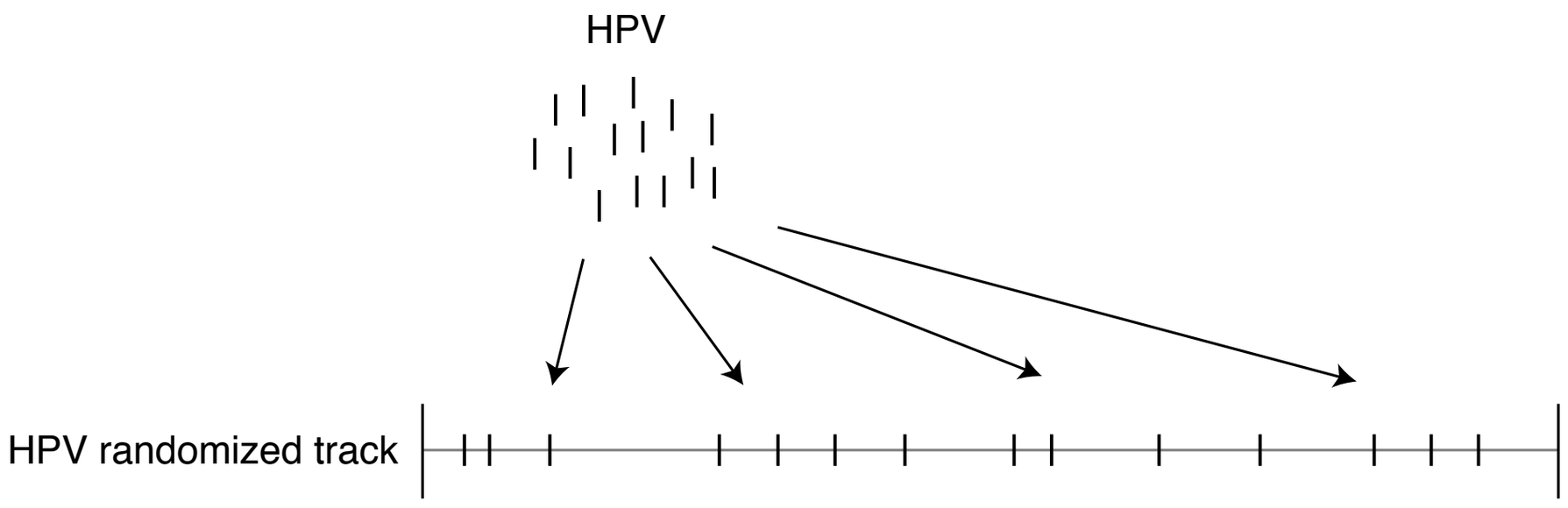
Further into statistical details: distributions



- You have probably read many times: "We assume XYZ is normally distributed"
- How is this related to Monte Carlo?
- Let us recap

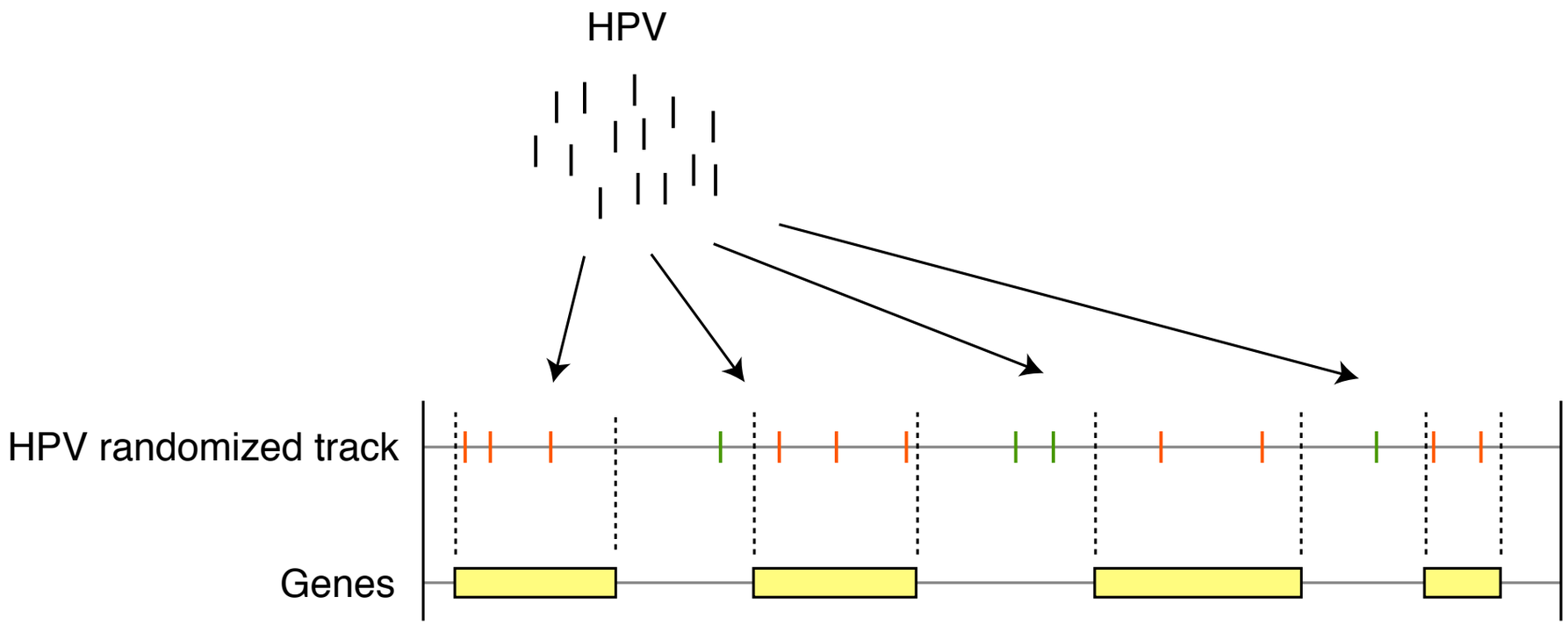
Monte Carlo test on "points inside segments"

- Randomize point (HPV) locations (null model)



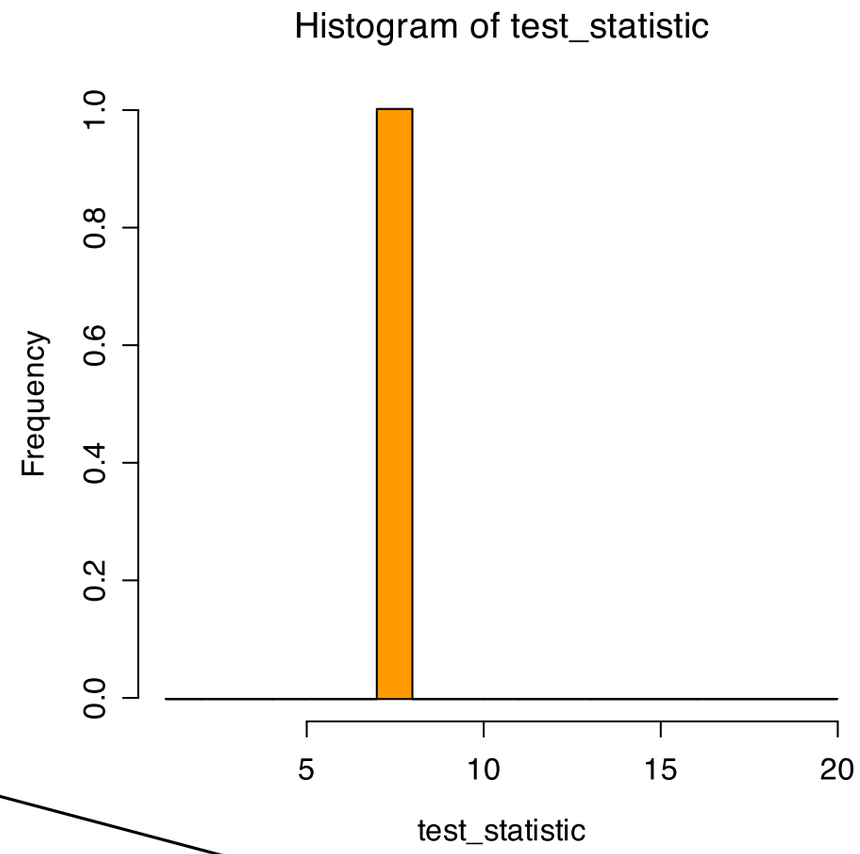
Monte Carlo test on "points inside segments"

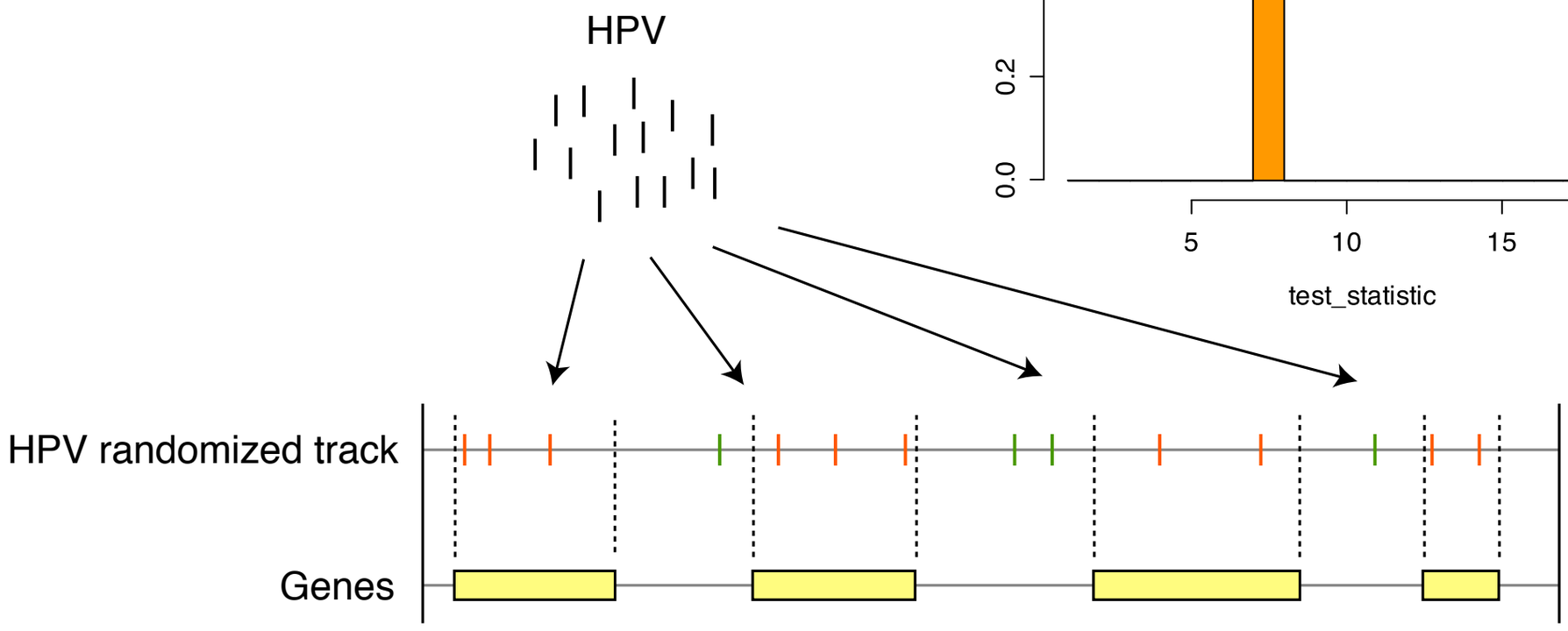
- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic



Monte Carlo test on "points inside segments"

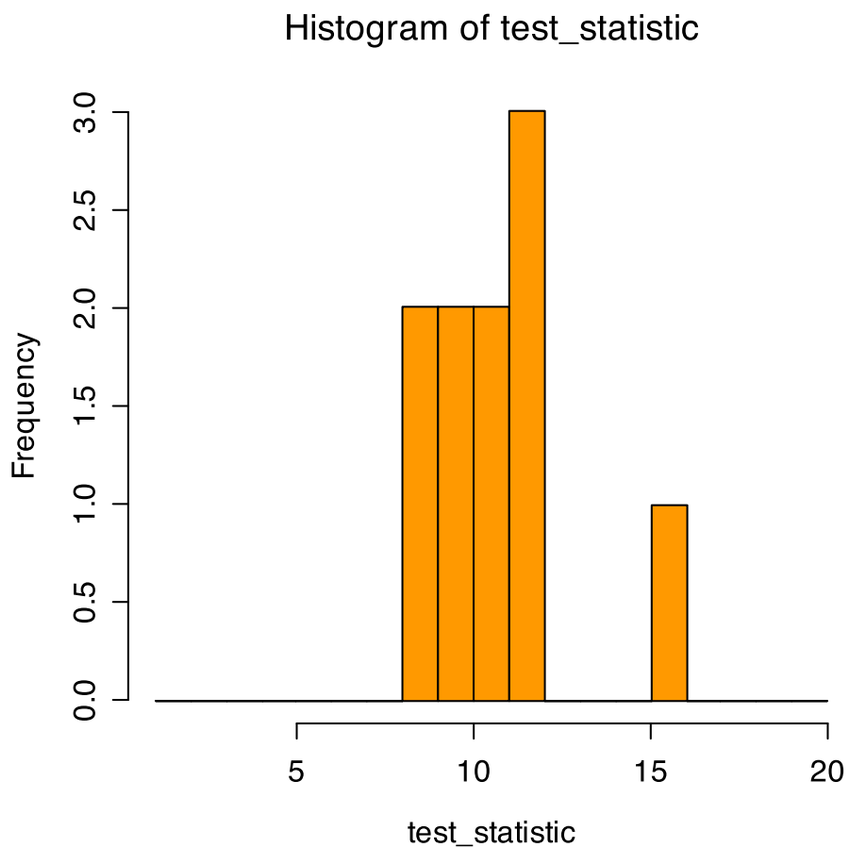
- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic





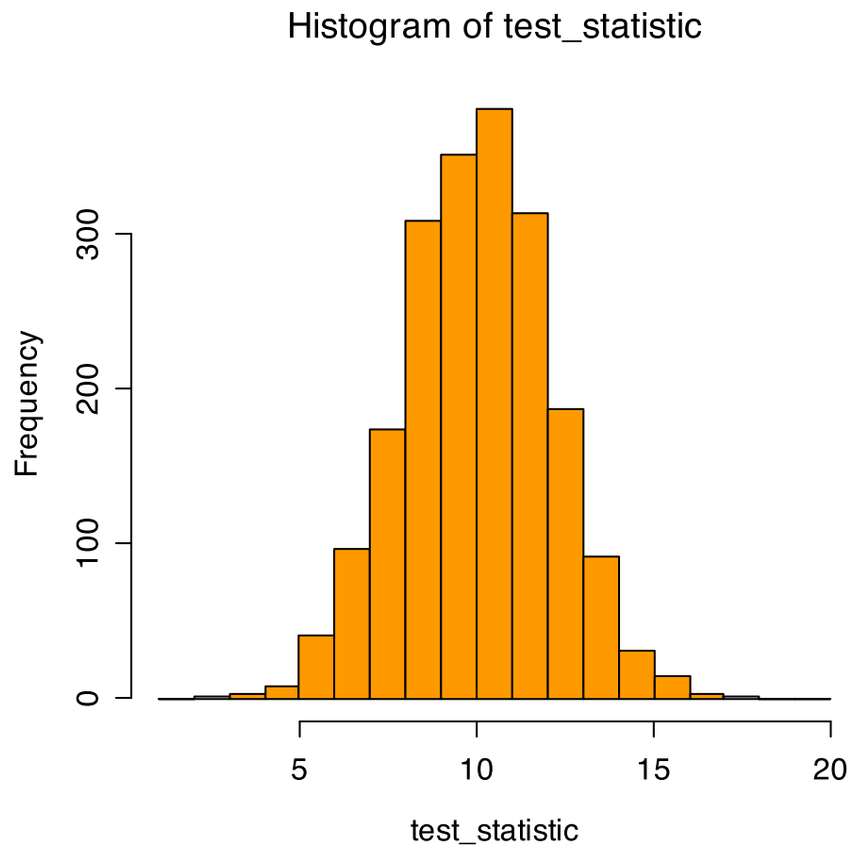
Monte Carlo test on "points inside segments"

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



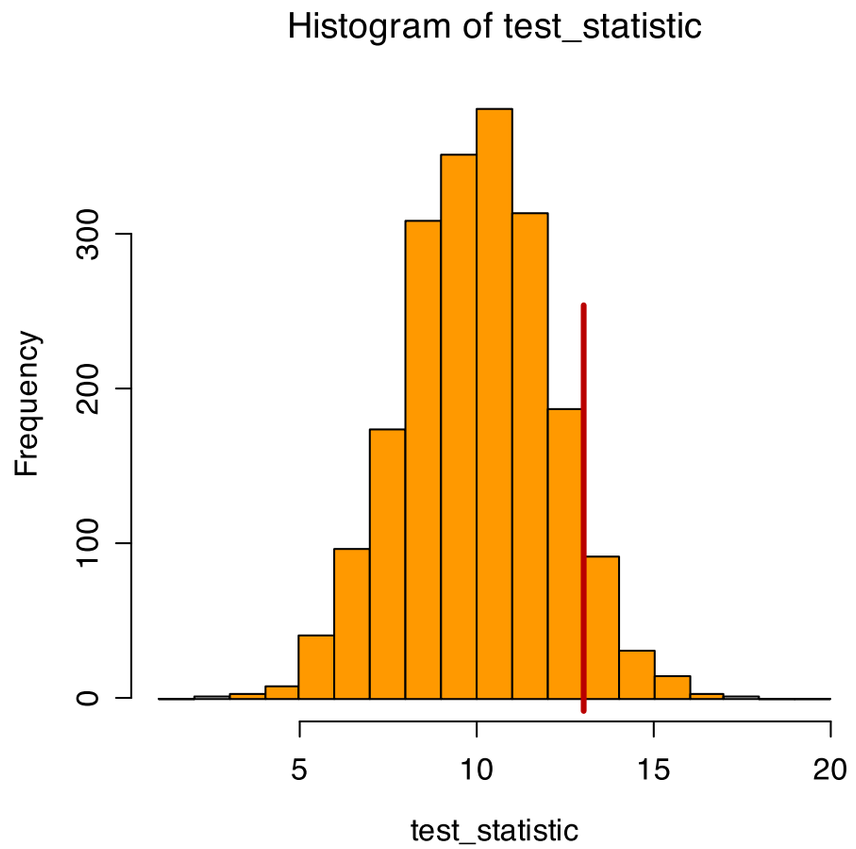
Monte Carlo test on "points inside segments"

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



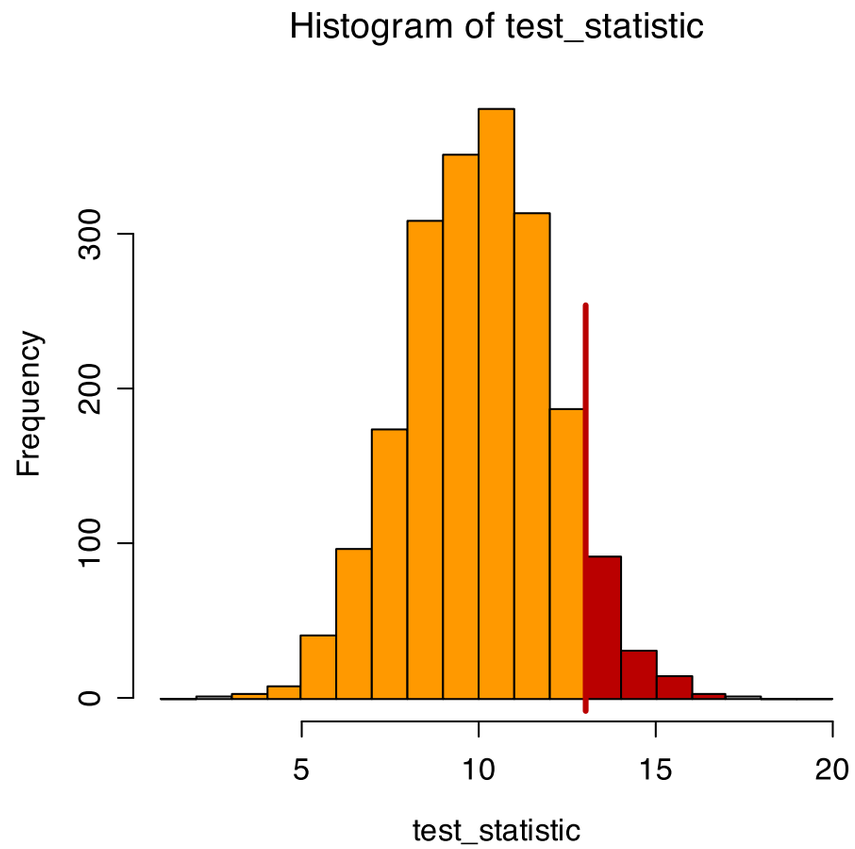
Monte Carlo test on "points inside segments"

- 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 alt hypothesis is "more" (if "less", area to the left)



p-value = 0.08

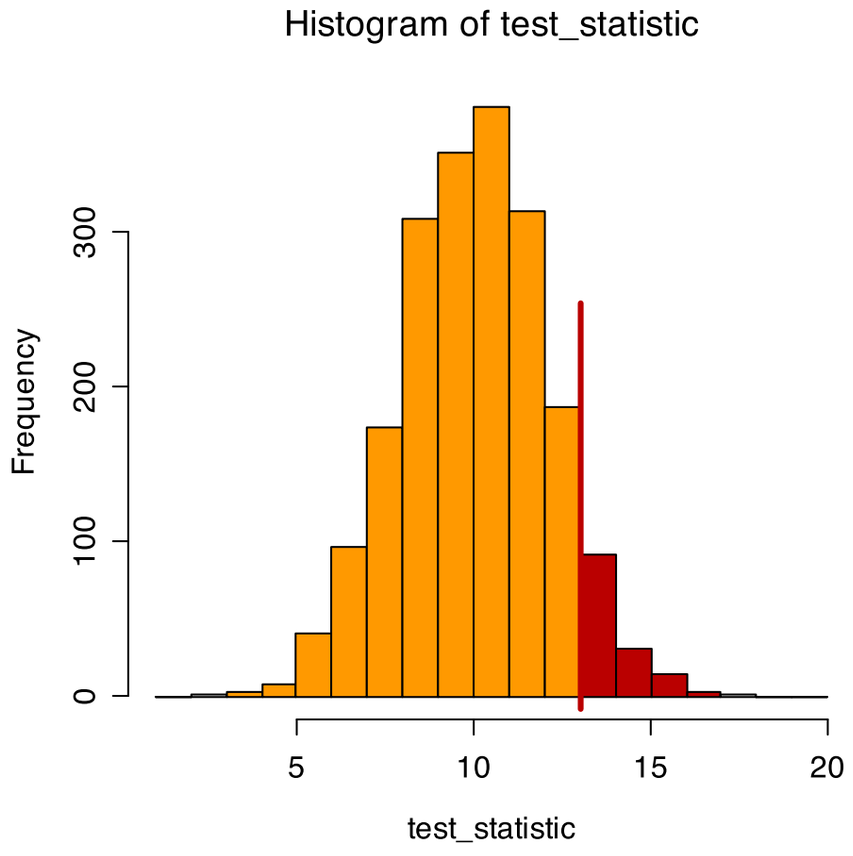
Monte Carlo: distribution

- What we have done now is to build a random discrete distribution (with discrete meaning that is is not smooth)

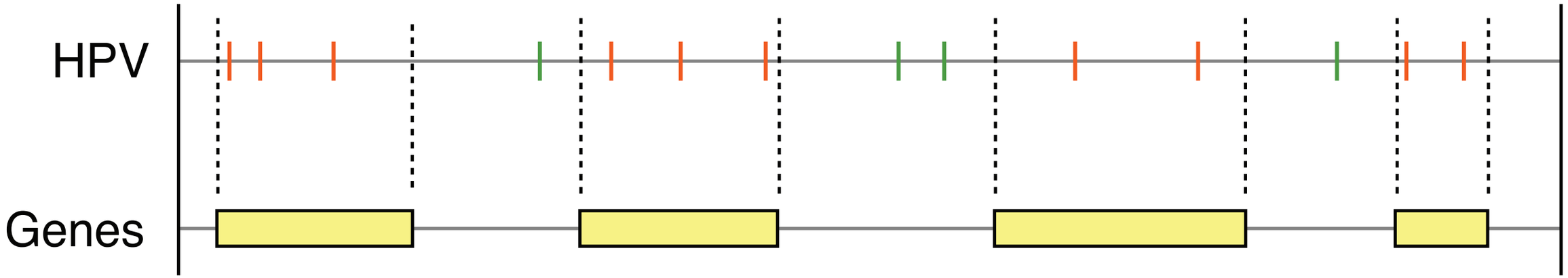
- We do this using Monte Carlo (which is slow) because we have no reason to assume a standard analytical distribution (such as the normal distribution)

  - (By analytical distribution we mean a distribution that can be described by mathematical formulas)

- In some cases, however, one can actually assume such distributions...



Further into statistical details: distributions



- Can we find a suited analytical distribution? (for number of HPV sites inside genes under H0)
- A statistician may answer: “yes, a binomial distribution”

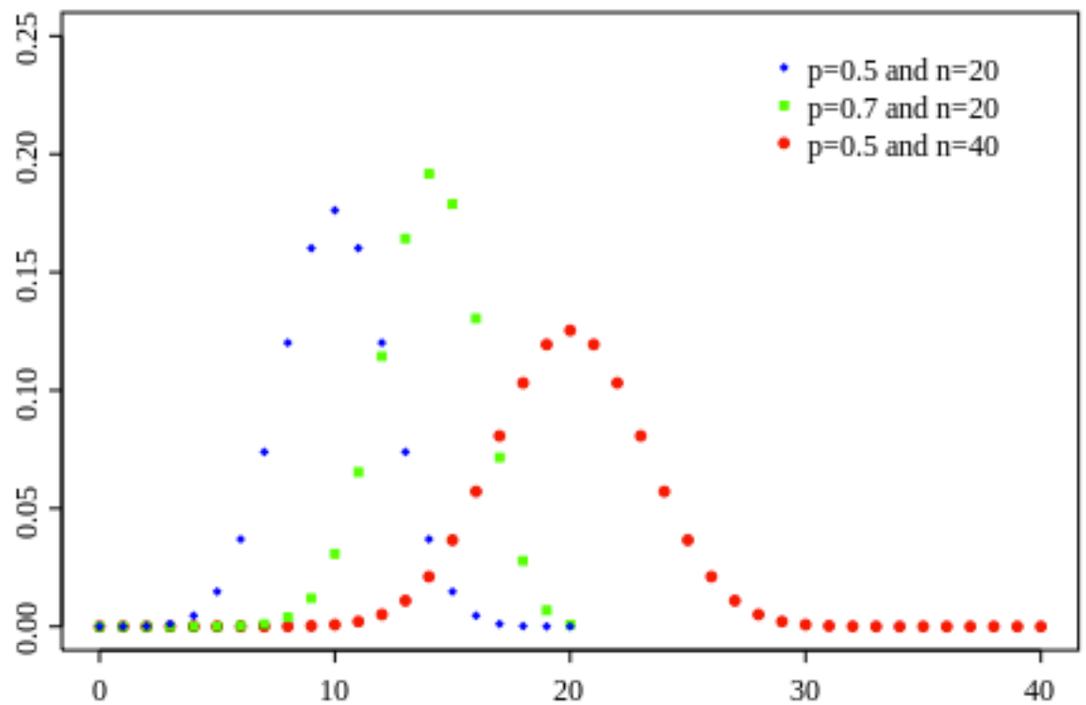
Binomial distribution

- Flip a coin ***n*** number of times
  - Two outcomes: heads or tails
- But: one side may be heavier than another
  - E.g. the probability of tails:

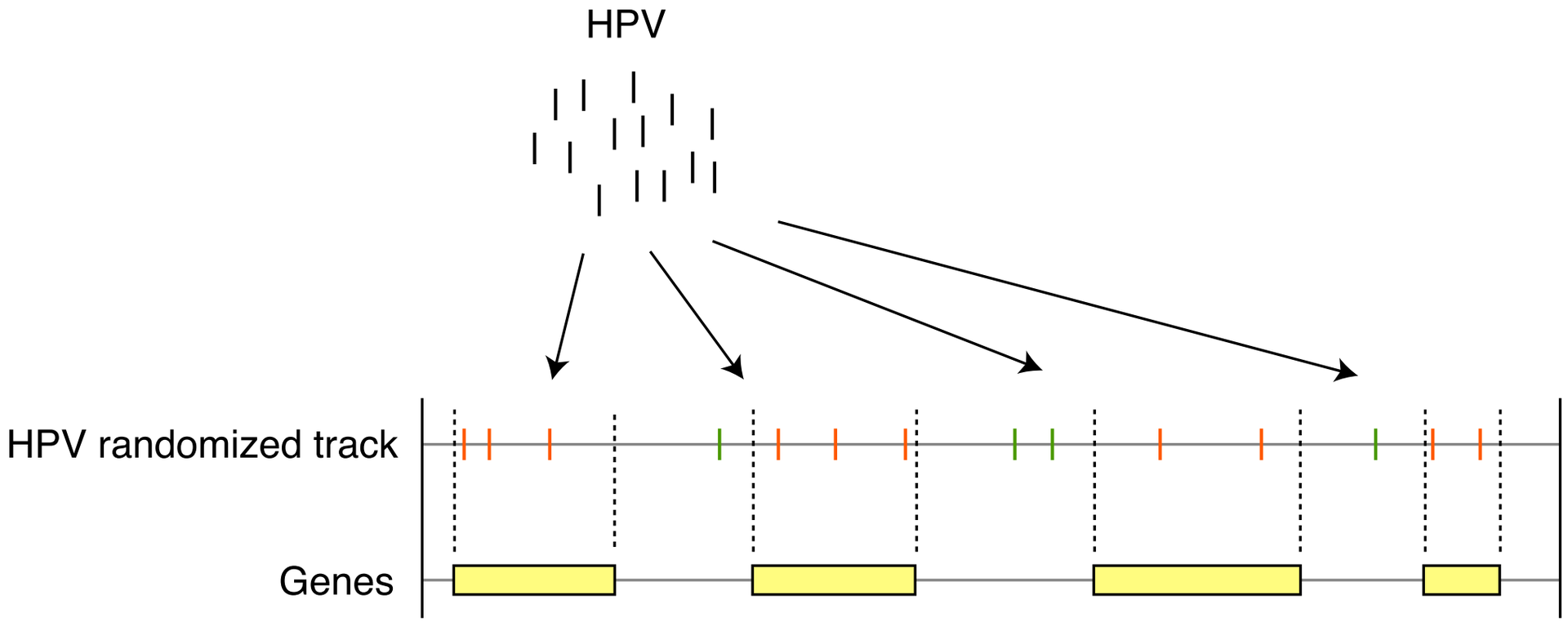
```
P(tails) = p = 0.6

P(heads) = 1-p = 0.4
```

- The distribution is dependent on **p** and **n**

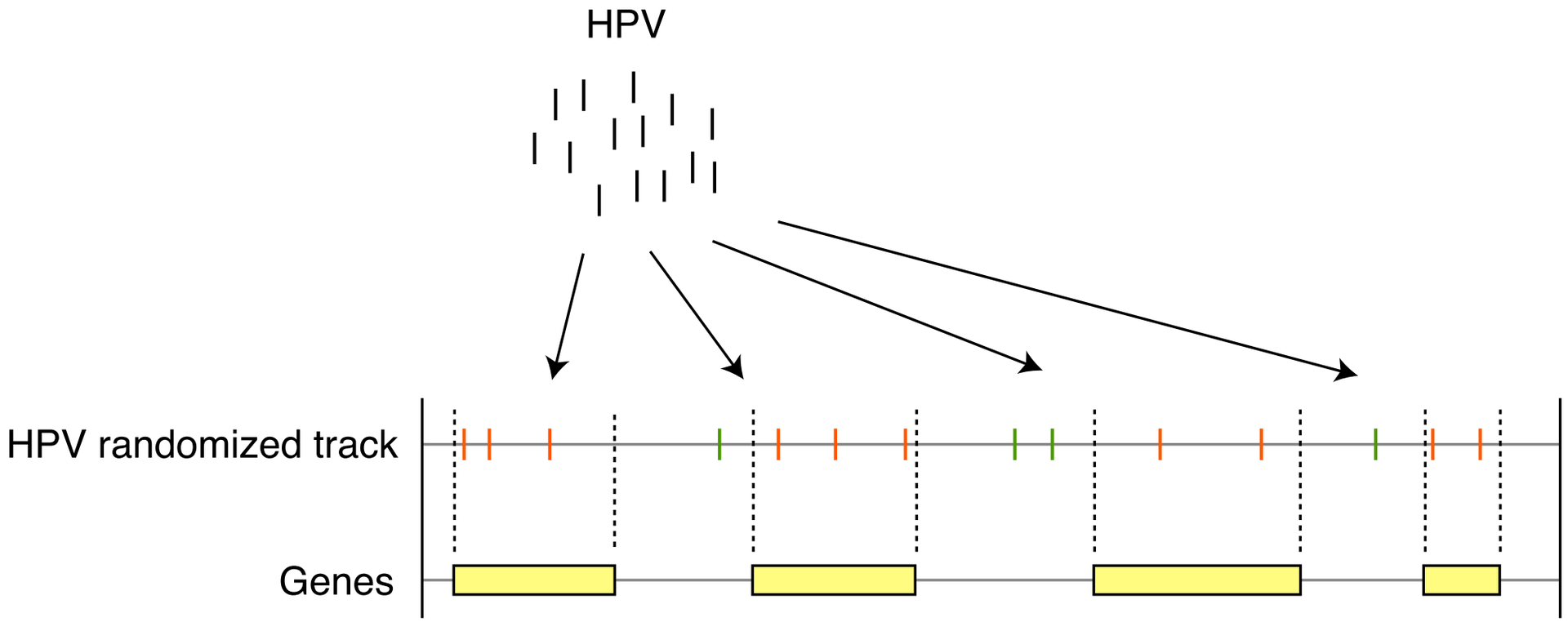


Binomial distribution



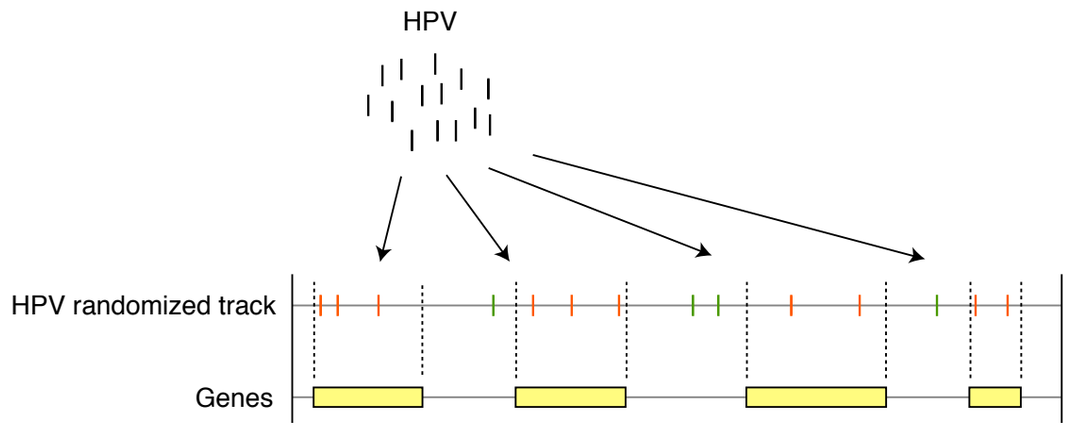
- In this case, each HPV is a coin, and it can either fall into a gene or not, depending on how much of the genome that is covered by genes
- n = number of HPV
- p = proportional coverage of genes

Binomial distribution



- Would you be comfortable assuming a binomial distribution? Or better: Would you have any clue on the implications?

Binomial 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
    - Same as MC null model: Preserve point count, randomize position (In the HyperBrowser, the binomial distribution is the null model without "MC")
  - Not trivial to see, and if found: is this acceptable?
  - If not acceptable, one can use Monte Carlo to randomize however one wants

Reproducibility

Reproducibility

- The advantages of making your research reproducible have been discussed in previous sessions
- The Genomic HyperBrowser is built on top of Galaxy, and thus keeps all its functionality for reproducible research
- In this part, you will carry out an exercise to test out reproducibility in practice

Exercise 13

- You will receive a document describing an analysis, which will be different from the one of your neighbor
- Carry out the analysis in a new history
- Make sure that the names of the history and elements are understandable
- Create a Galaxy page with your results (explained in the document)
- When finished, share your Galaxy Page with your neighbor
- The neighbor should rerun the analysis with another null model
- Discuss among yourself whether it was easy to understand and redo the analysis

Ten 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
- Record all intermediate results, when possible in readable formats
  - Intermediate steps of creating case-control are stored as history elements
- Provide public access to scripts, runs and results
  - Provide link to Galaxy Page that embed histories with all runs and results

Ten 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

Ten simple rules for reproducibility (cont.)

- Archive all external programs and custom scripts, in the versions that were used
  - Galaxy provides this publicly and explicitly. HyperBrowser is version controlled and can be contacted.
- Avoid manual, non-trackable procedures
  - We have performed all analysis steps in the Galaxy system
- For analyses including randomness, note underlying random seeds
  - HyperBrowser allows a particular random seed to be set (results are then deterministic, like a frozen snapshot of randomness)

Analysis of track collections

Why multitrack analysis?

- So far, we did statistical analysis on a pair of datasets
- Recent improvements in sequencing technologies allow genome-wide profiles for a variety of biological features to be systematically generated for a wide range of cell types (e.g. H3K27ac for different cell types, or all histone modifications in liver cells).
- One should take advantage of all the available data

Representing track collections

- The GSuite format
  - Simple tabular format
    - One line per track
    - Allows metadata (per collection and per track)
- Remote vs Local tracks

Multitrack analysis questions

- Which tracks in a collection are most representative or most atypical?
- Which tracks in a collection coincide most strongly with a target track?
- Are certain tracks of one collection coincide particularly strongly with certain tracks of another collection?
- Which genomic regions are mostly enriched with the segments of tracks in a collection?
- In which genomic regions are tracks of a collection coinciding the most?

The GSuite HyperBrowser

- A comprehensive solution for the analysis of track collections (GSuites) across the genome and epigenome
- Provides tools for
  - Acquisition
  - Customization
  - Analysis

Exercise 14

- Goal: Get familiar with the GSuite HyperBrowser https://hyperbrowser.uio.no/gsuite
- Basic vs Advanced user mode
- By navigating through the basic mode execute an analysis of your choice from start to end

Acquisition of track collections

- From public repositories
- From local datasets
- From the HyperBrowser repository

Customization of track collections

- Downloading and preprocessing
- Modifying the collections
- Modifying the datasets themselves

Statistical analysis of track collections

- Determine representative and atypical tracks in a GSuite
- Determine GSuite tracks coinciding with a target track
- Determine coinciding track combinations from two suites
- Determine regions where GSuite tracks are enriched
- Determine regions where GSuite tracks co-occur more strongly

Discussion on similarity measures

Assume the following data

- 300 tracks representing TF binding sites for different TFs
- 1 other track representing TF binding sites for one specific TF

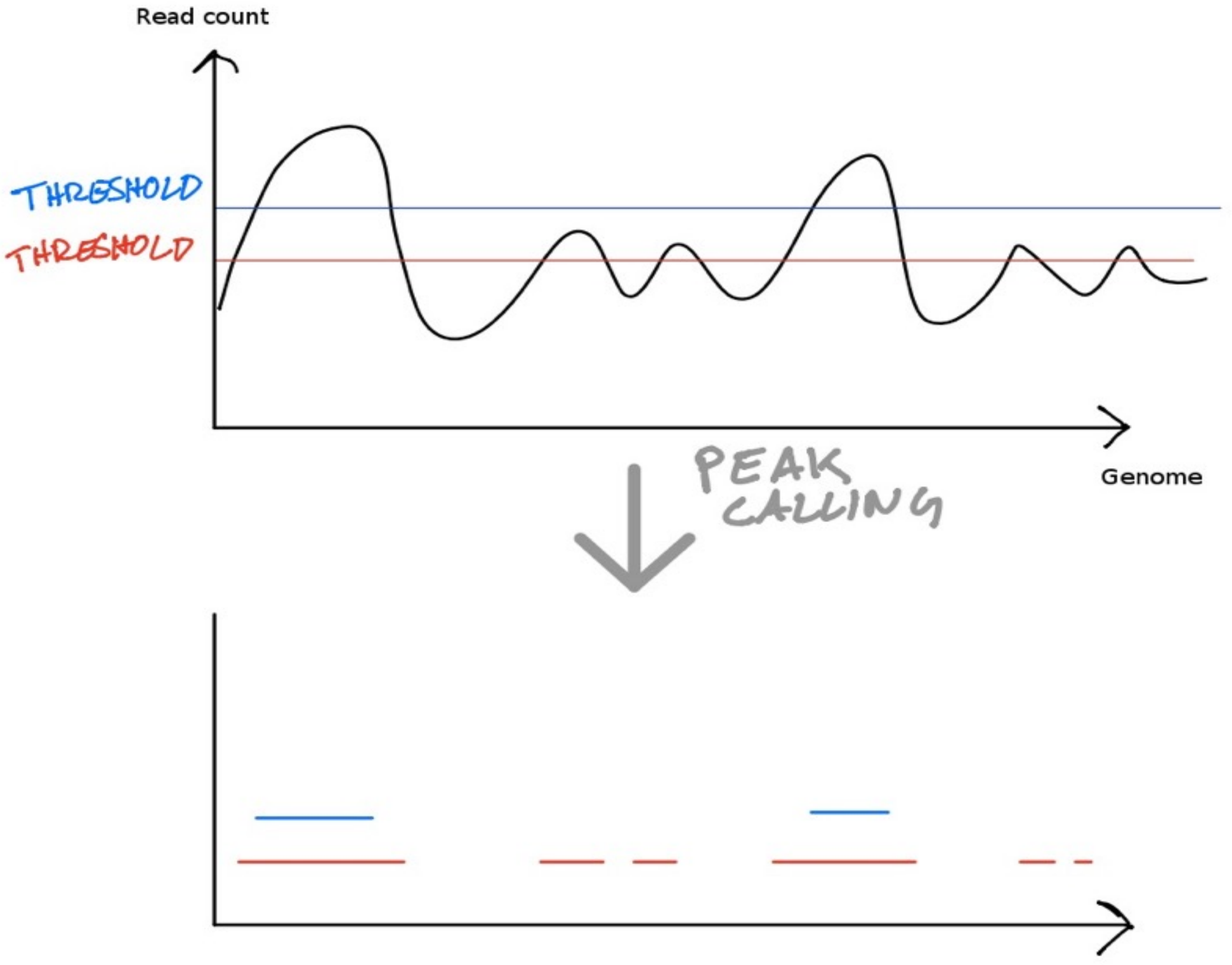
… and the following question:

- What TF track among the 300 is most similar to the separate TF?

Discussion on similarity measures (cont.)

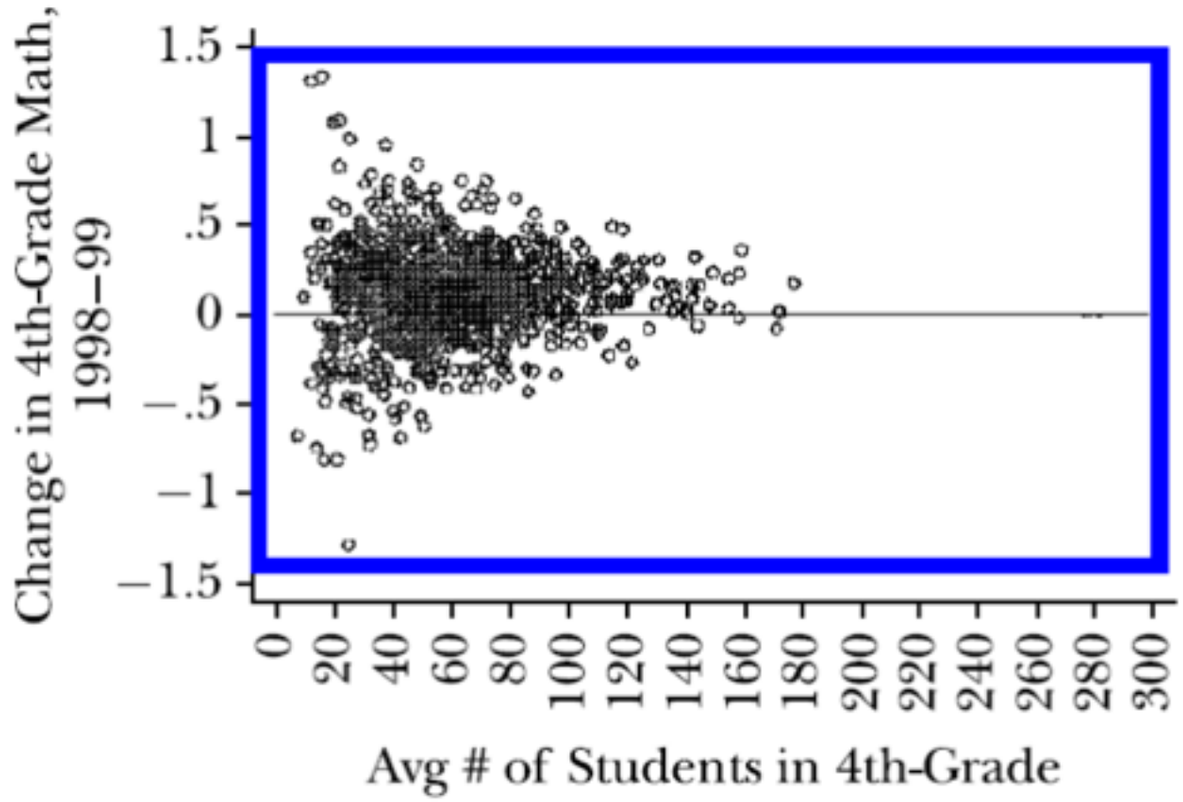
- Challenge:
  - The different TF tracks are of different “size”, meaning that the number of base pairs covered by each track varies a lot.
  - Not because of less or more binding
  - Can be because some data sets lack data, or are produced in a more conservative way (e.g. stricter threshold in peak calling)

How two TF tracks from the same TF can be of different “size”



A digression: The Small Schools Myth

- The Bill gates foundation spent $2 billion in funding small schools, after research showed that small schools performed better
- Later admitted they were wrong
- What was the research? Small schools always rank top.



Source: http://marginalrevolution.com/

Statistics on ranking tracks based on similarity to a reference track

- Each track is a sample of the actual genomic feature

- We want to rank the tracks based on the proportion of the underlying probability of each base pair falling inside the *reference track* and falling outside the reference track

- Need a measure that is an estimator of this

Binary similarity measures

- A binary similarity measure is a function S(trackA, trackB) that takes two tracks as input and returns the “similarity” between these two tracks
- Called “binary” because originally not used for genomic tracks, but to compute the similarity between two binary vectors
- The most famous and most used measure is the Jaccard similarity measure, first used by Jaccard to cluster ecological species in 1901
  - Today often used in scientific publications to measure similarity between genomic tracks

- Jaccard = intersection / union

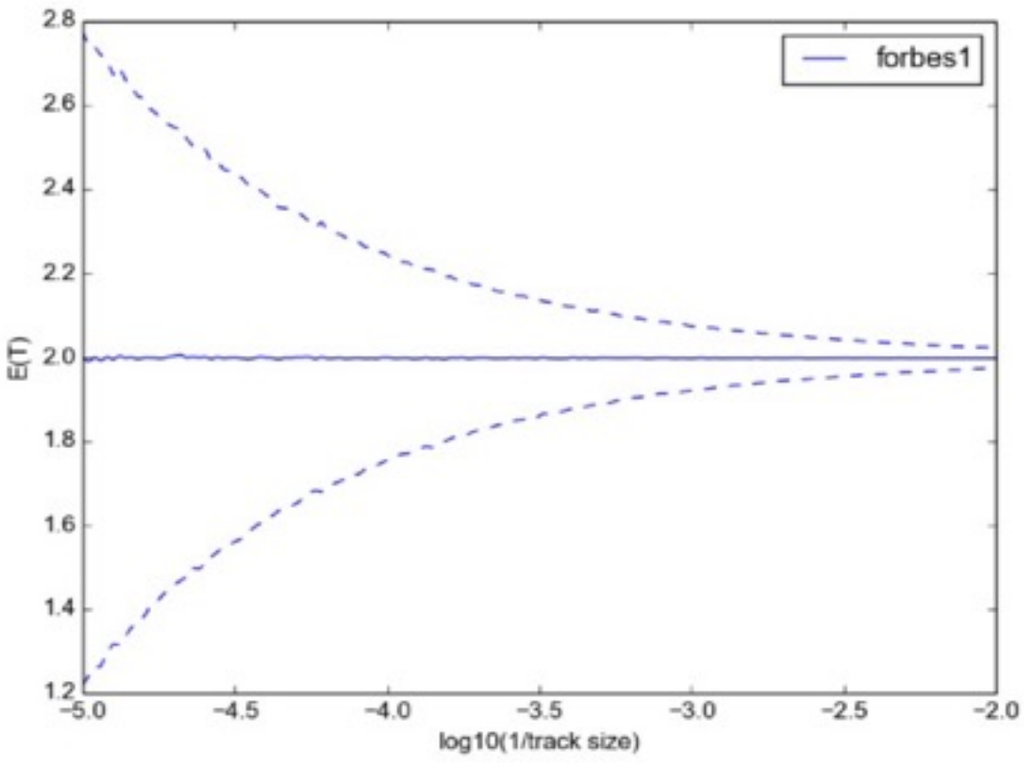
  (Number of base pairs covered by both tracks divided by number of base pairs covered by at least one track)

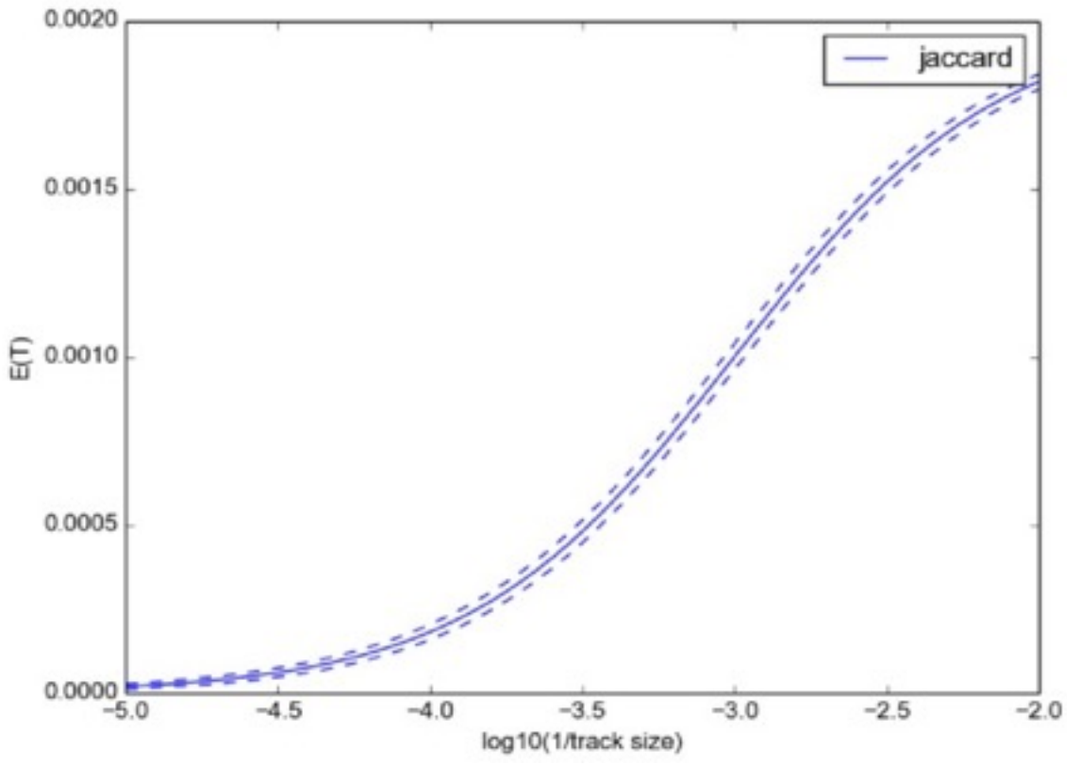
- Forbes = observed / expected

  (Number of base pairs

Biases in binary similarity measures

- We will not explain the statistical details here, but:
  - Jaccard favours large tracks
  - Forbes has no bias, but has high variance for small tracks (winners curse)





Simulation of the Jaccard similarity and Forbes measure of two tracks:
- Jaccard increases with track size
- Forbes is unbiased, but has high variance for small tracks

Discussion on similarity measures Take home message

- Be aware of winners curse
- The similarity measure you choose will either have variance or bias (in expected value) that is depending on the track size
- Forbes is unbiased, but has high variance for small track s
- Jaccard has low variance, but favours large tracks
- There are at least 76 different binary similarity measures. Try different, investigate their properties: http://www.iiisci.org/journal/CV$/sci/pdfs/GS315JG.pdf
- In the GSuite HyperBrowser, you can choose among different binary similarity measures

Exercise 15

- Using the GSuite HyperBrowser find out which histone modification broad-peak datasets from peripheral blood primary T CD8+ naive cells show the highest association to multiplex sclerosis associated regions
- NB: Use the demo dataset for MS
- NB2: Use the curated catalog to acquire the appropriate track collection

Conclusion

Main conclusions

- Tracks and track types are useful concepts for representing genome-wide positional data
- Monte Carlo is a powerful, flexible and transparent method for hypothesis testing
- Choice of data, test statistic, null model and implementation details are all difficult, and have consequences for the results
- You should be aware of the choices you make. The software cannot make all the choices for you
- The more realistic assumptions you make, the less publishable your results will typically be! :-) (but they will be more correct...)
- It is important to do your analyses in a reproducible way (by e.g. using Galaxy or the Genomic HyperBrowser)

The basic skills we want you to learn

- Quality control (both reads and analysis results)
- Study design (e.g. replicates)
- Principles of mapping
- Principles of assembly
- Statistics, hypothesis testing
- Summary statistics and visualisation
- Sanity checking/validation of results
- Model system versus non-model system organisms
- Reproducibility
- Finding data, and munging it

Any questions?

- Feel free to contact us:
  - borissim@ifi.uio.no
  - ivargry@ifi.uio.no