

Statistical epigenomics

INF-BIO 5121/9121

October 09-11 2017, Oslo

Boris Simovski and Ivar Grytten

BMI/Genomic HyperBrowser team

Department of Informatics, UiO

Where are you now (in this course)?

- You have some genomic feature datasets, e.g.:
 - SNPs datasets found by doing variant calling
 - Expressed genes datasets discovered by RNA-Seq
 - .. or any other genomic features (e.g. position of transcription factor binding sites)
- What is this module about?
 - You will do statistical analyses on such datasets, e.g.:
 - Learn how you can find the relationship between e. g. expressed genes and SNPs

What will you learn?

- To investigate the relationships between genomic features, by doing statistical testing
- The underlying principles and models behind such analysis (tracks, track types)
- How to create a suitable model when doing such analysis (including null models and test statistics)
- Which errors people typically do
- Learn to use the Genomic Hyperbrowser, which will make you able to do this kind of analysis on huge

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 Example analysis. The Genomic HyperBrowser.

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

12:30-13:00 Binary similarity measures.

13:00-15:00 Analysis of track collections. The GSuite
HyperBrowser.

Overview of session

Day 3:

09:00-09:45 Recap of days I and II.

10:00-12:00 Reproducibility

12:00-12:15 Home exam

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 GSuite HyperBrowser

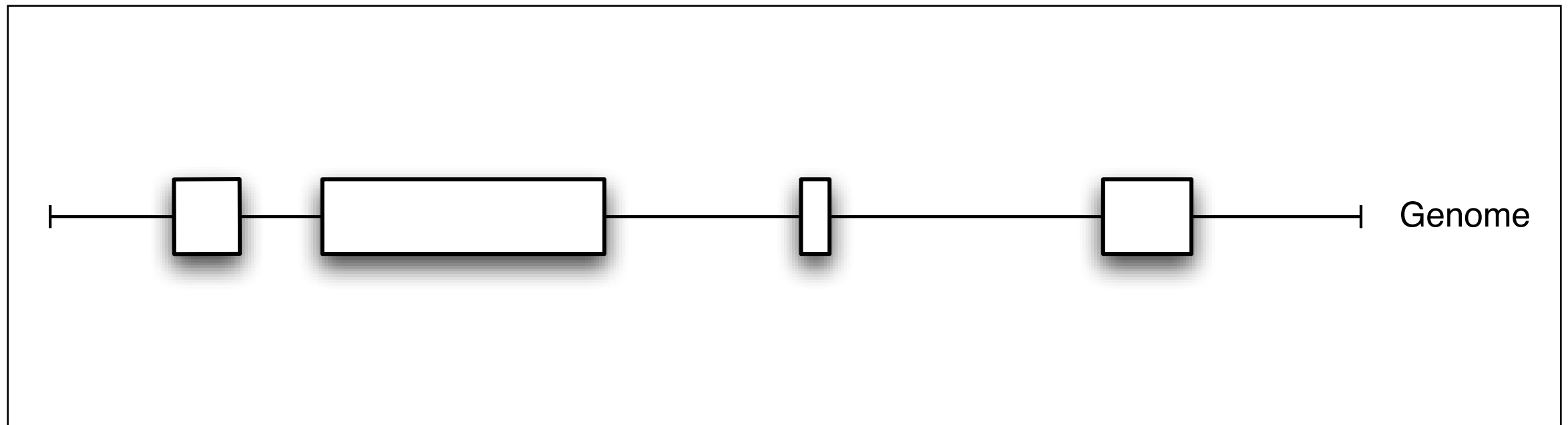
- We will make use of the GSuite HyperBrowser in this session
- The HyperBrowser is a software system for statistical analysis, developed locally at UiO
- However:

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

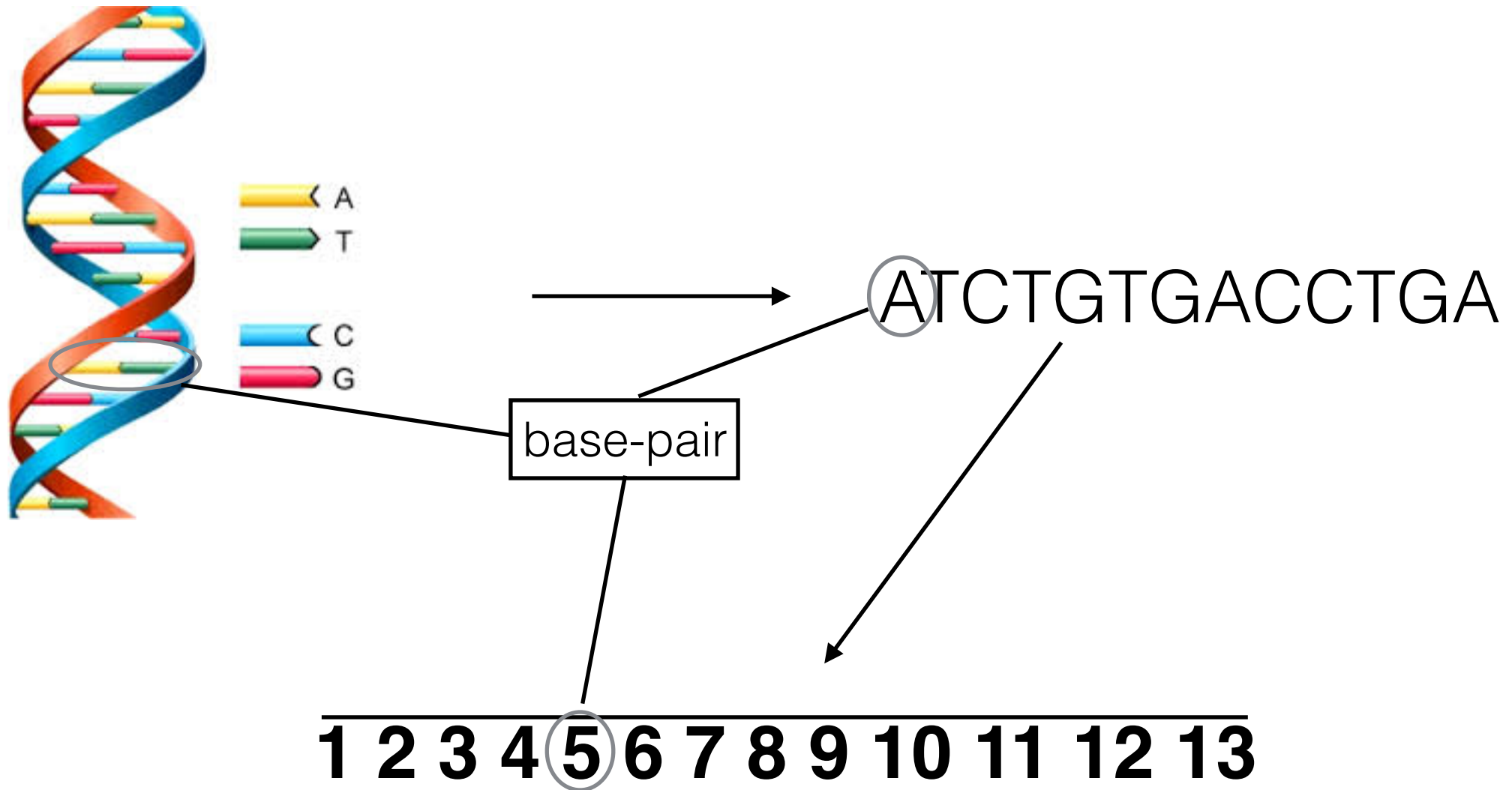
Introduction

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 that 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

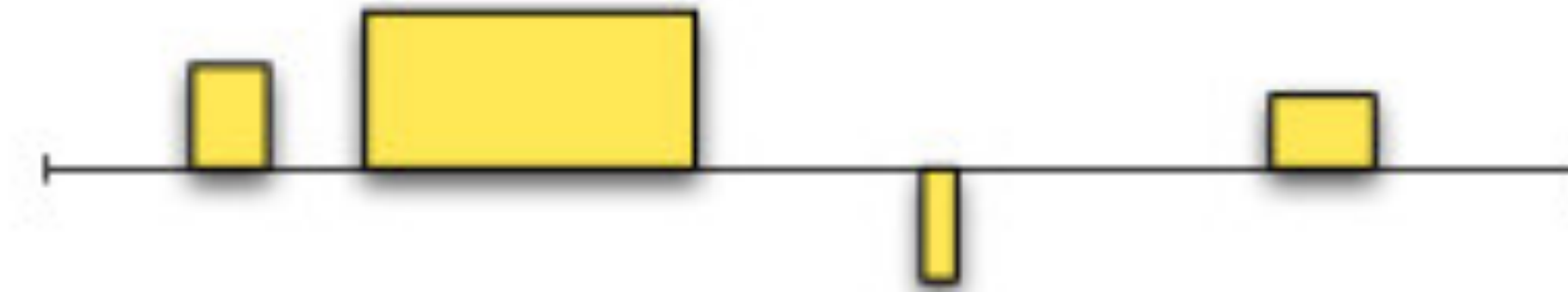
Tracks and track types

Representation of genes



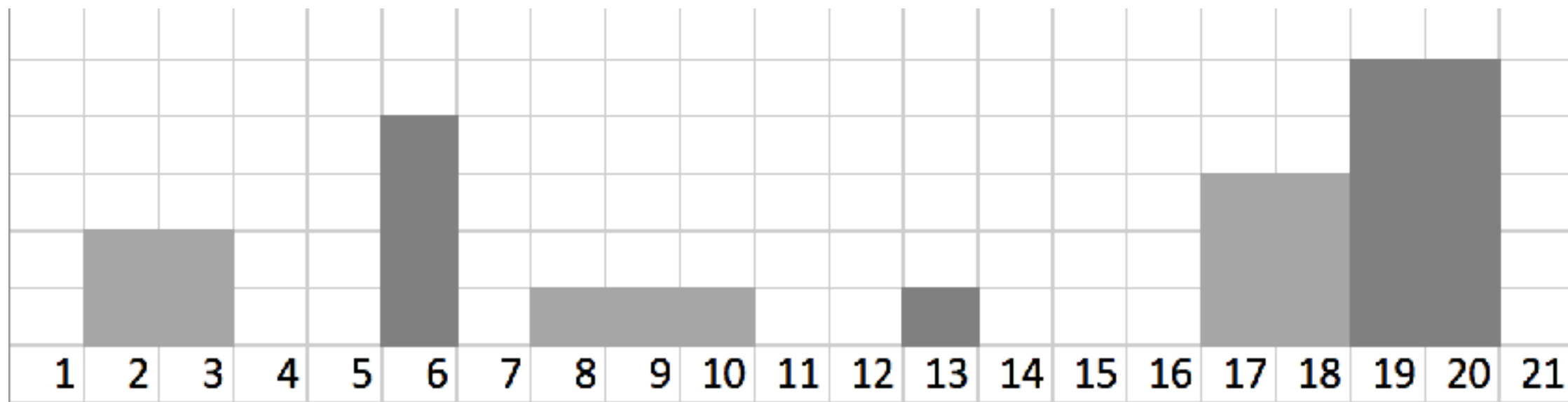
chr7	127471196	127472363
chr7	127472388	127473530
chr7	127473555	127474697
chr7	127474701	127475864
chr7	127475893	127477031
chr7	127477121	127478198
chr7	127478300	127479365
chr7	127479375	127480532
chr7	127480538	127481699

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

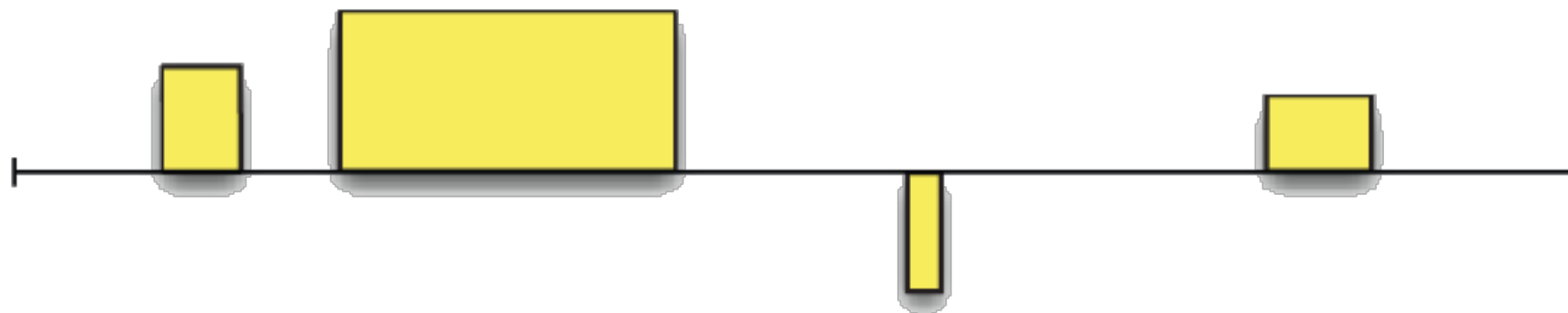
Exercise I



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

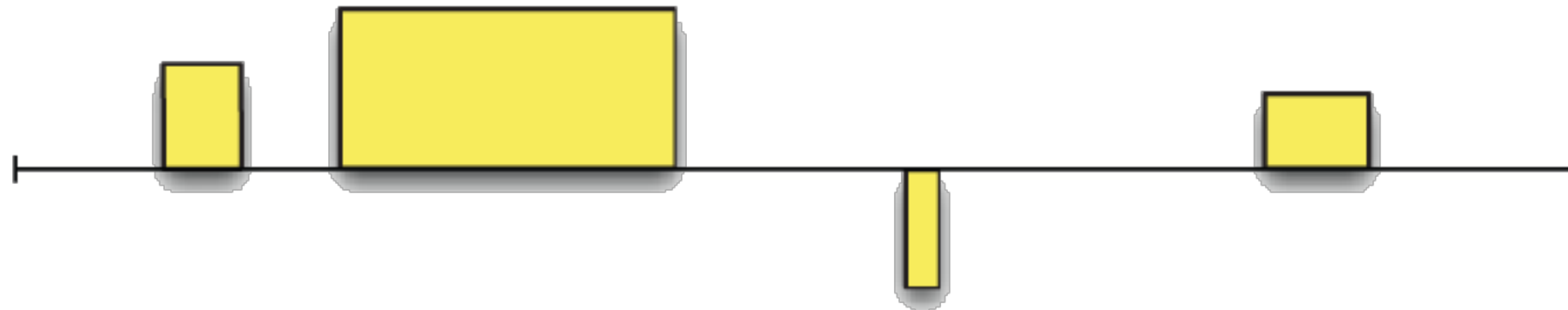
- In the last example, we showed genes as segments on the genome line, with attached RNA-seq read count values
- This track is of a **track type** we call “valued segments”



Valued Segments (VS)

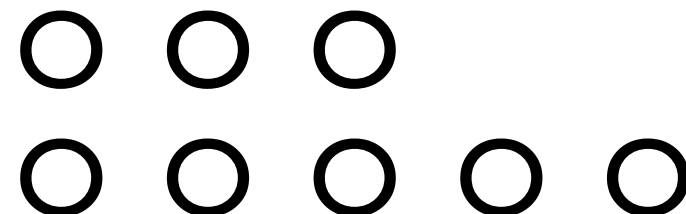
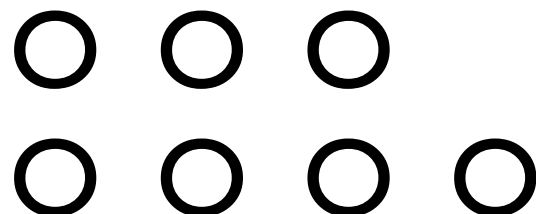
- Track types are mathematical / conceptual models used to categorize tracks according to their main characteristics

Exercise 2

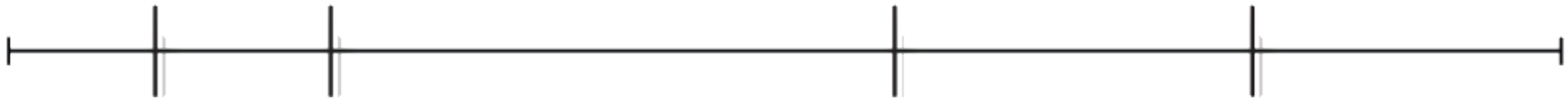


Valued Segments (VS)

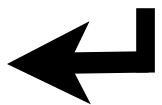
- What other **track types** can you think of?
 - Discuss with your neighbour (2-3 min)
 - Classroom discussion



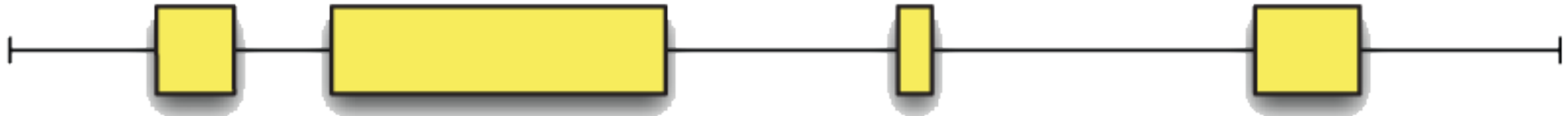
Points



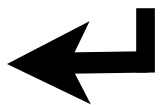
Points (P)



Segments



Segments (S)



Genome Partition



Genome Partition (GP)



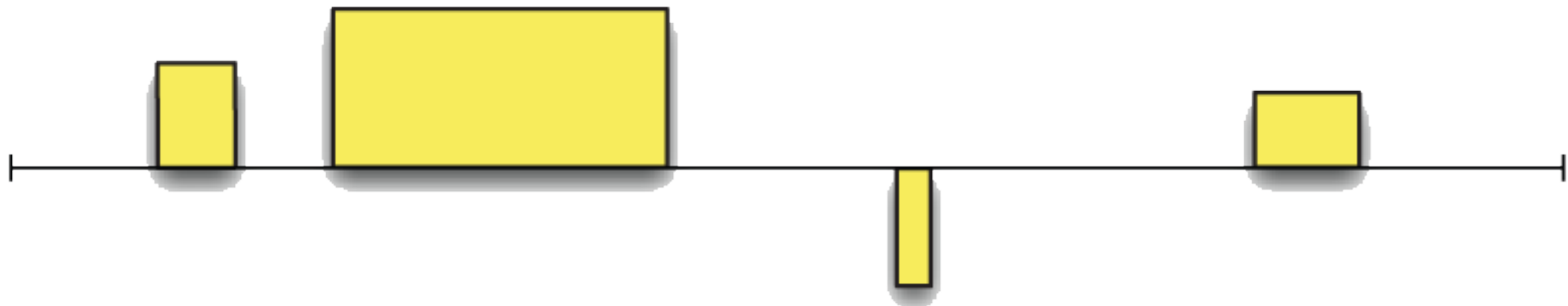
Valued Points



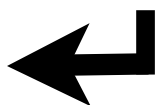
Valued Points (VP)



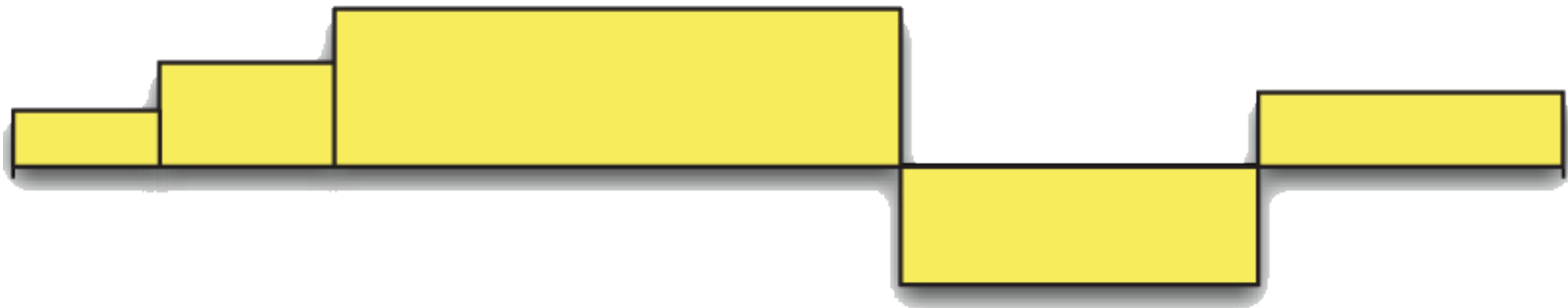
Valued Segments



Valued Segments (VS)



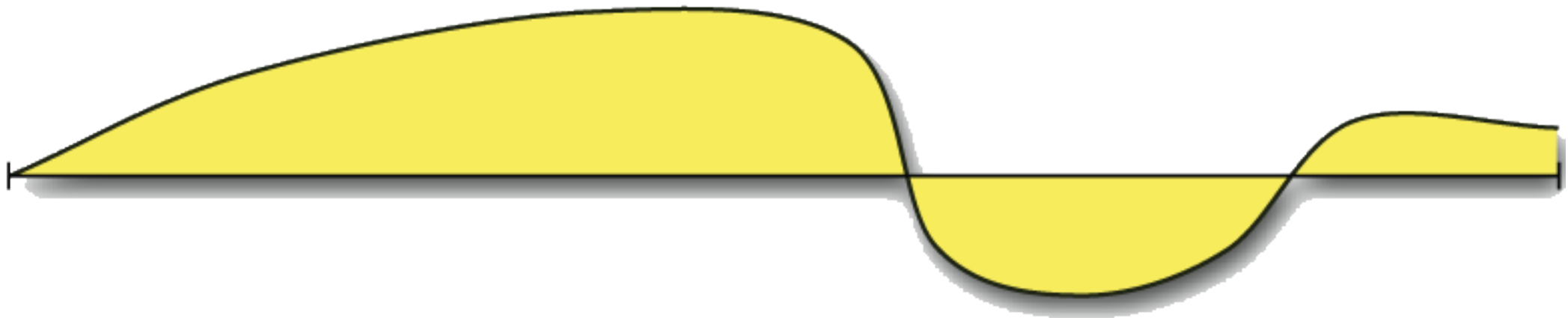
Step Function



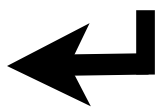
Step Function (SF)



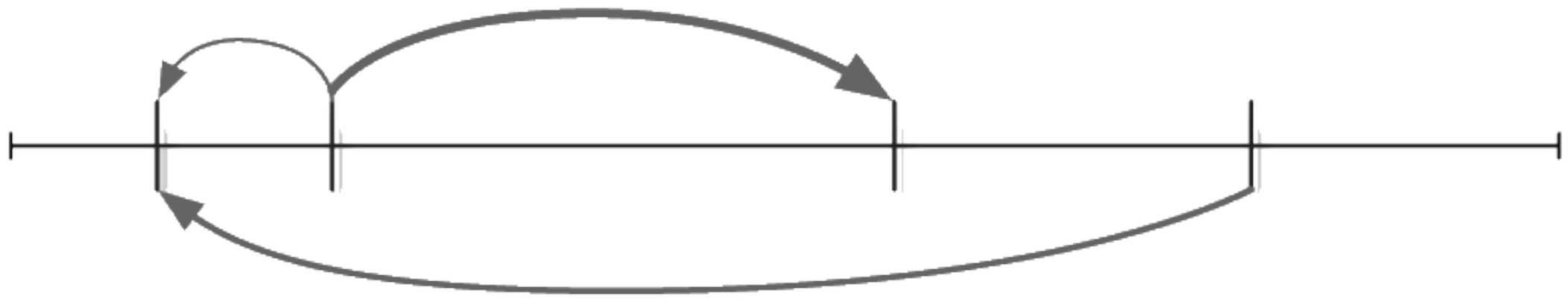
Function



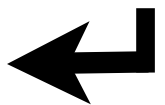
Function (F)



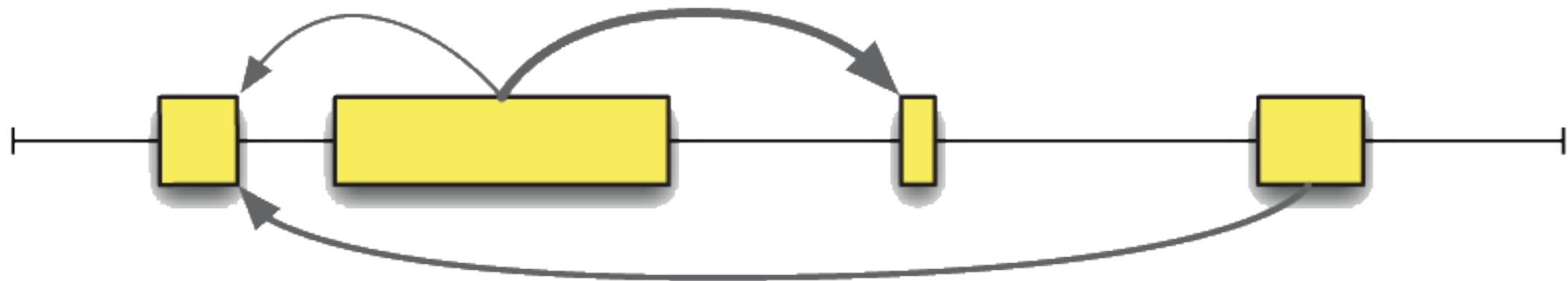
Linked Points



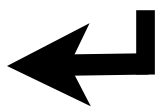
Linked Points (LP)



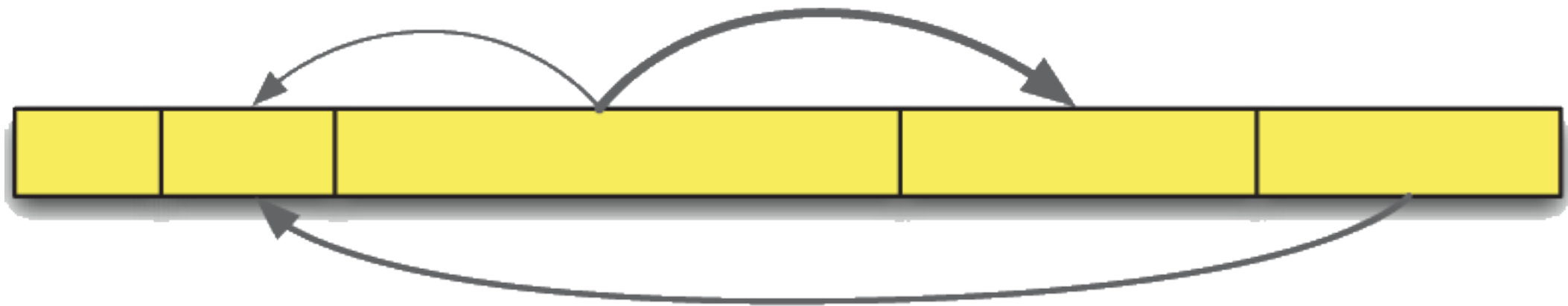
Linked Segments



Linked Segments (LS)



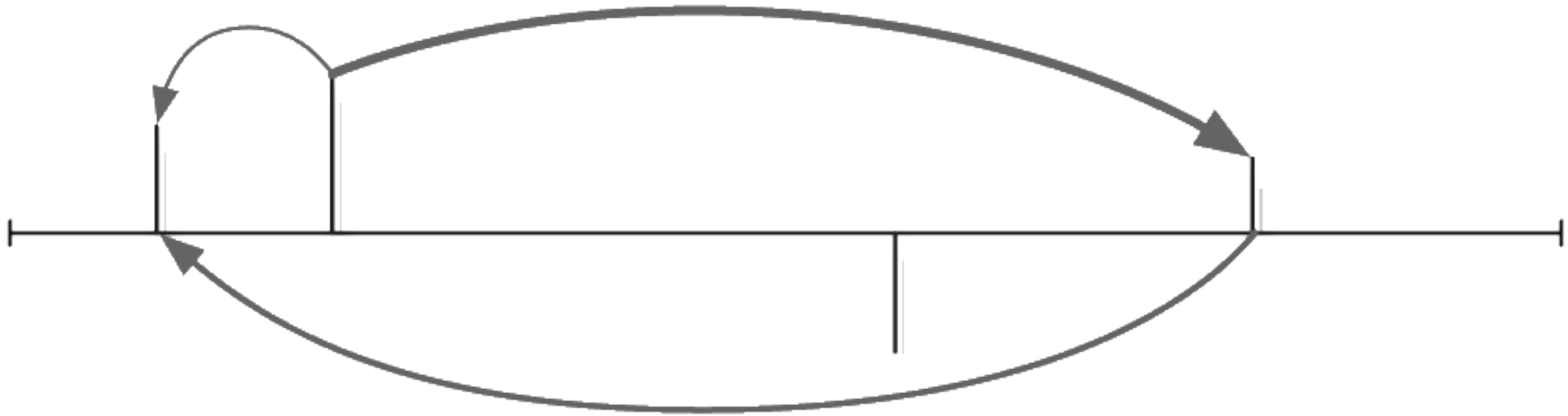
Linked Genome Partition



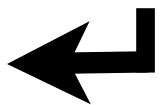
Linked Genome Partition (LGP)



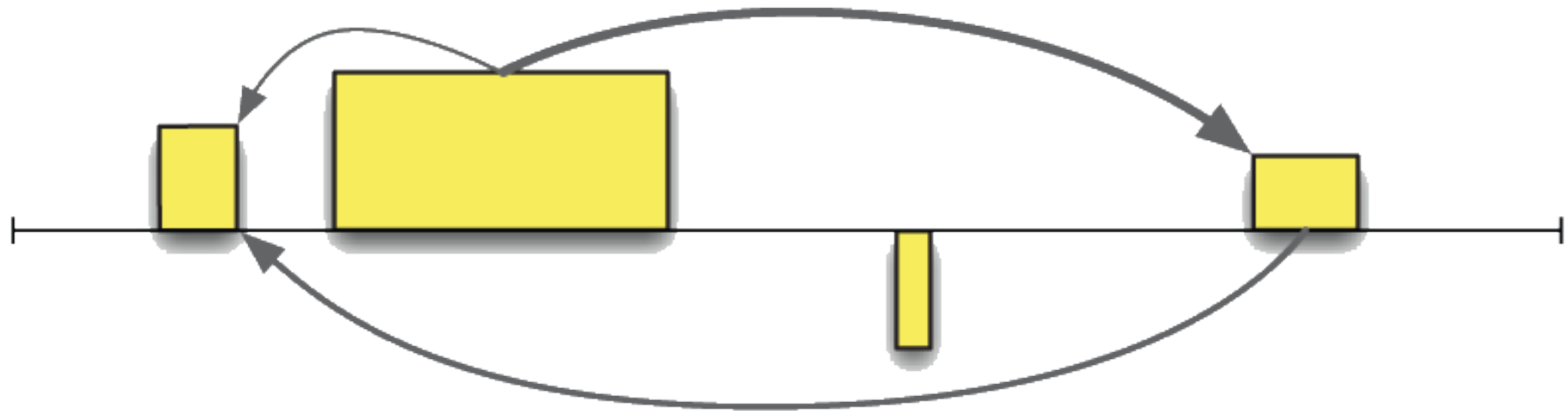
Linked Valued Points



Linked Valued Points (LVP)



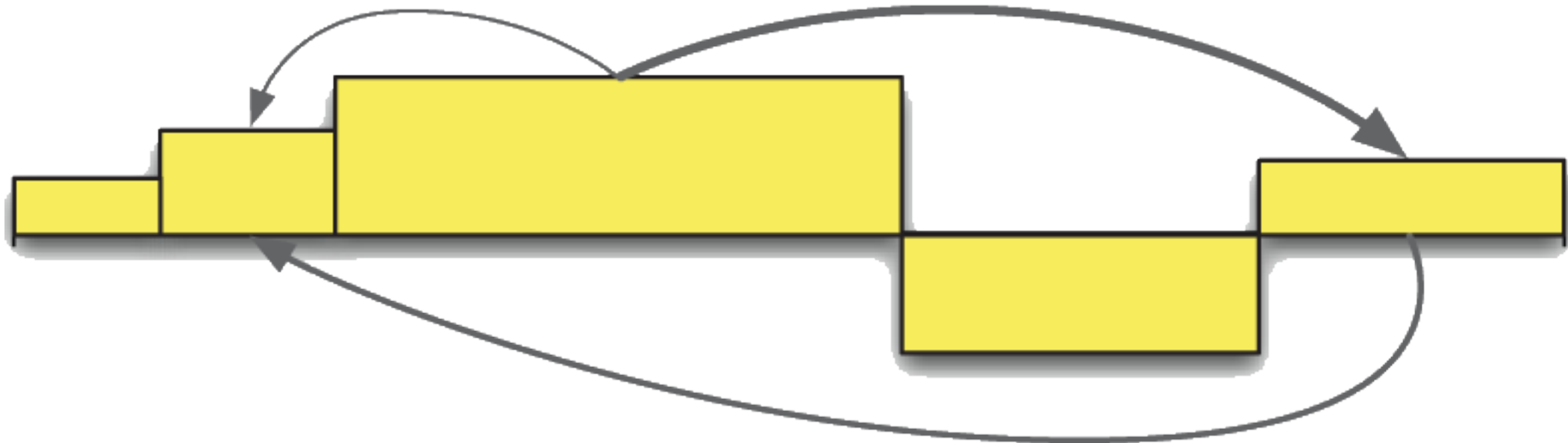
Linked Valued Segments



Linked Valued Segments (LVS)



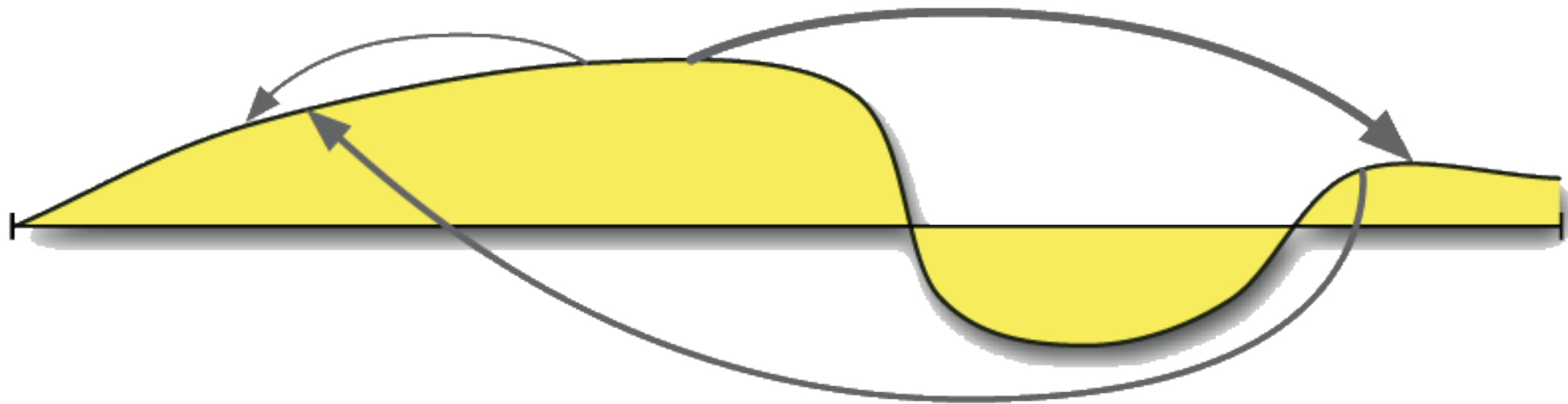
Linked Step Function



Linked Step Function (LSF)



Linked Function



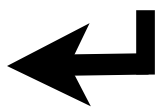
Linked Function (LF)



Linked Base Pairs



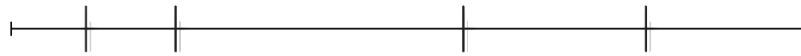
Linked Base Pairs (LBP)



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?

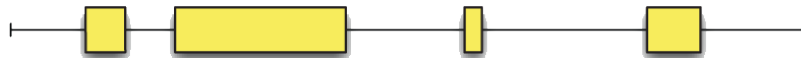
Exercise 3



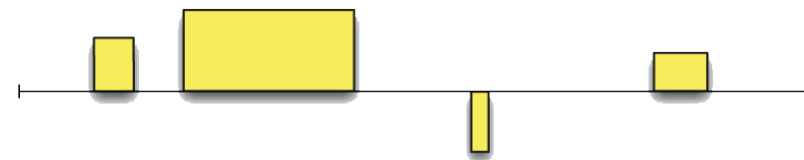
Points (P)



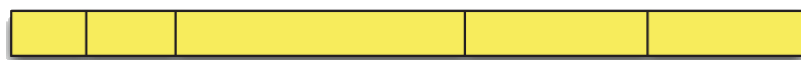
Valued Points (VP)



Segments (S)



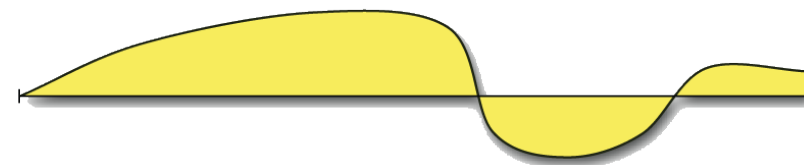
Valued Segments (VS)



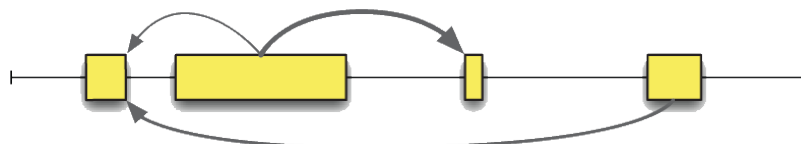
Genome Partition (GP)



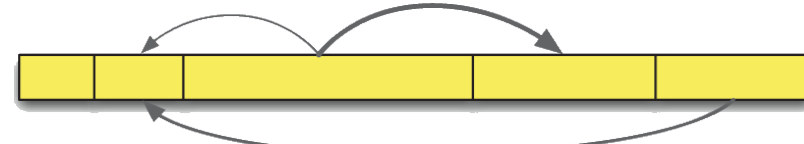
Step Function (SF)



Function (F)



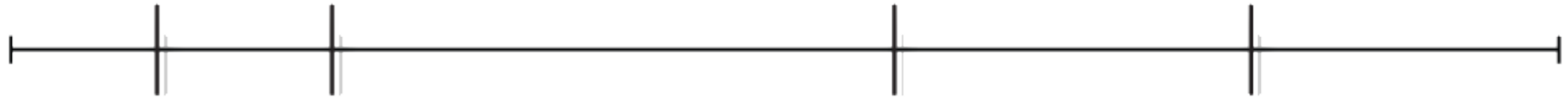
Linked Segments (LS)



Linked Genome Partition (LGP)



Points

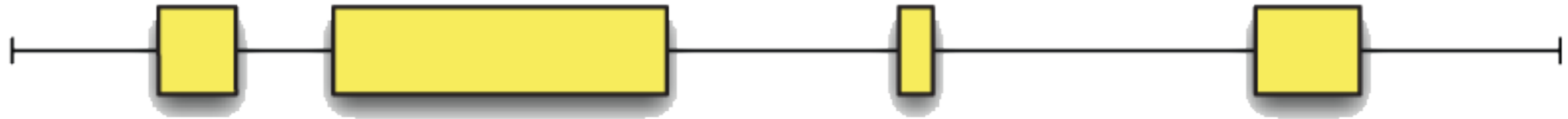


Example tracks:

- SNPs



Segments



Example tracks:

- Genes
- Transcription factor binding sites



Genome Partition



Example tracks:

- Chromosomes
- Chromosome arms
- Chromatin state segmentation



Valued Points

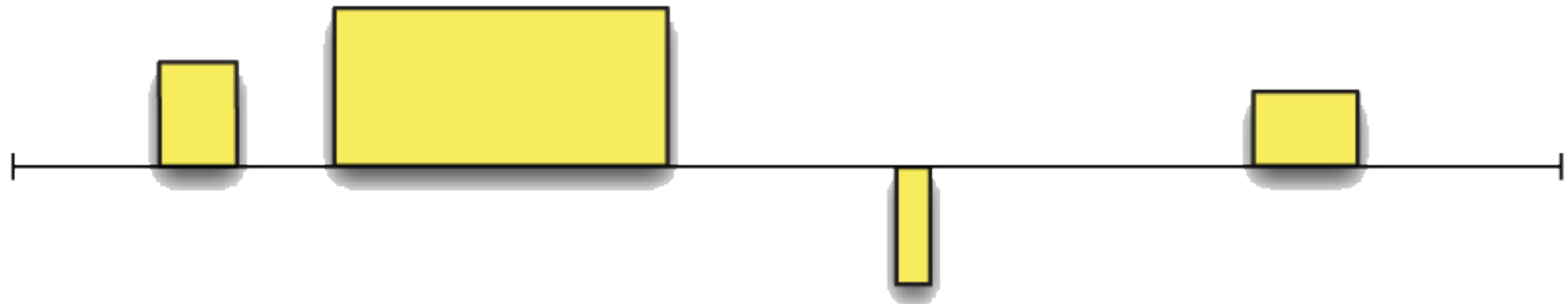


Example tracks:

- SNPs with allele frequency
- SNPs with quality



Valued Segments

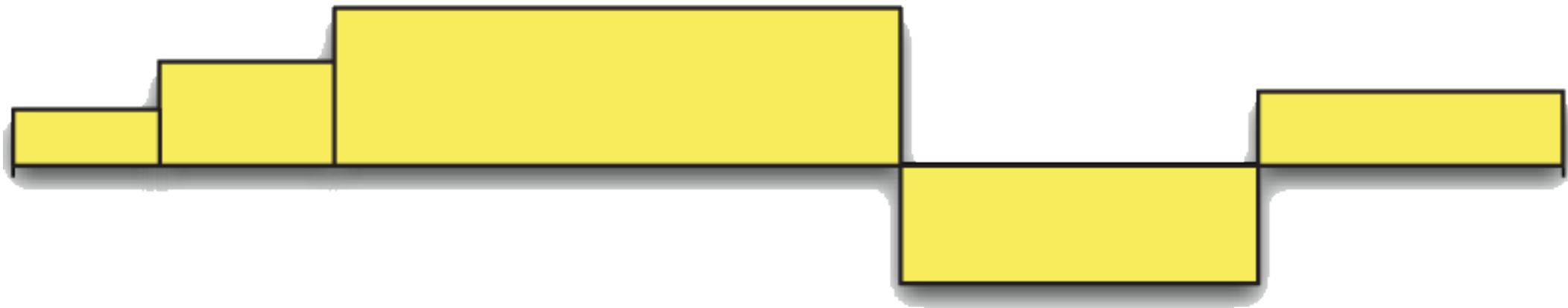


Example tracks:

- Genes with expression values
-



Step Function

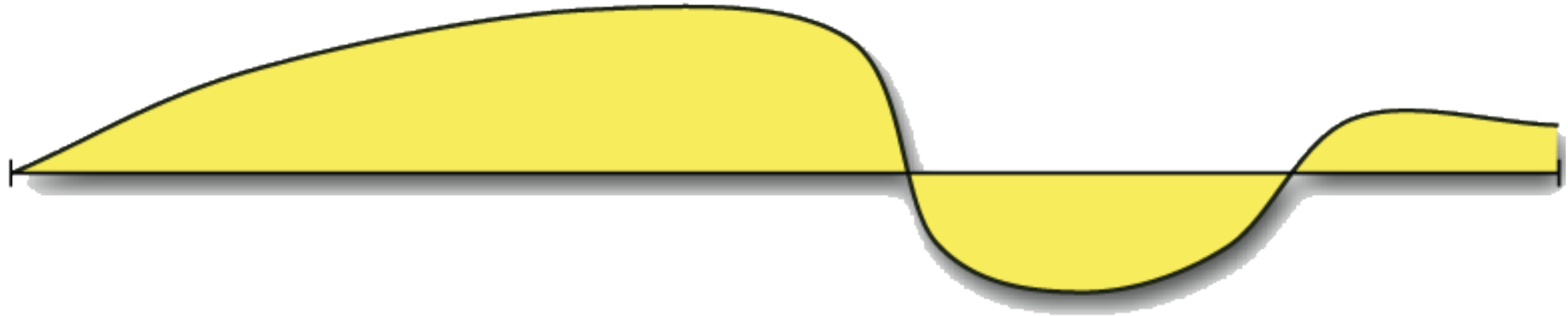


Example tracks:

- GC content (per partition)



Function

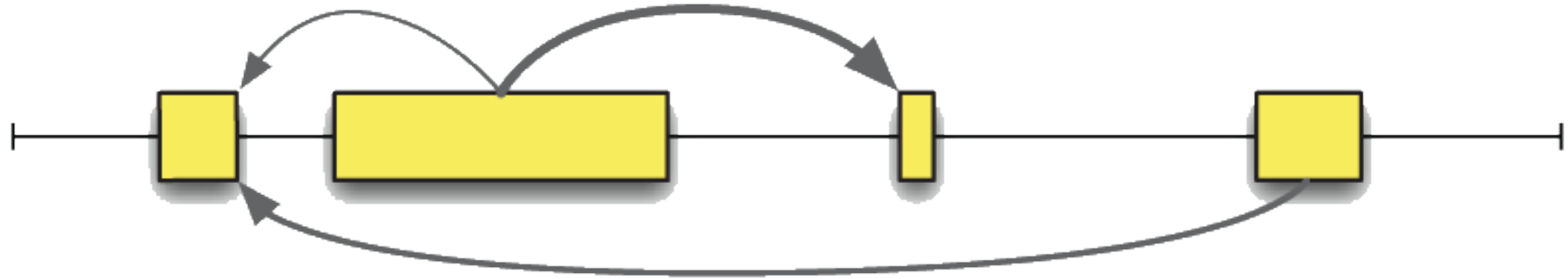


Example tracks:

- DNA melting temperature
- Coverage (RNA-seq)



Linked Segments

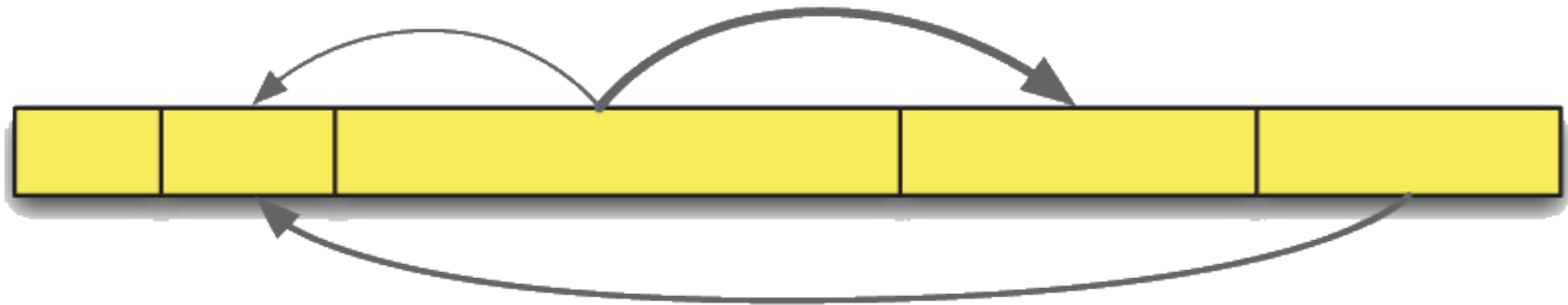


Example tracks:

- ChIA-PET
- Co-expressed genes



Linked Genome Partition

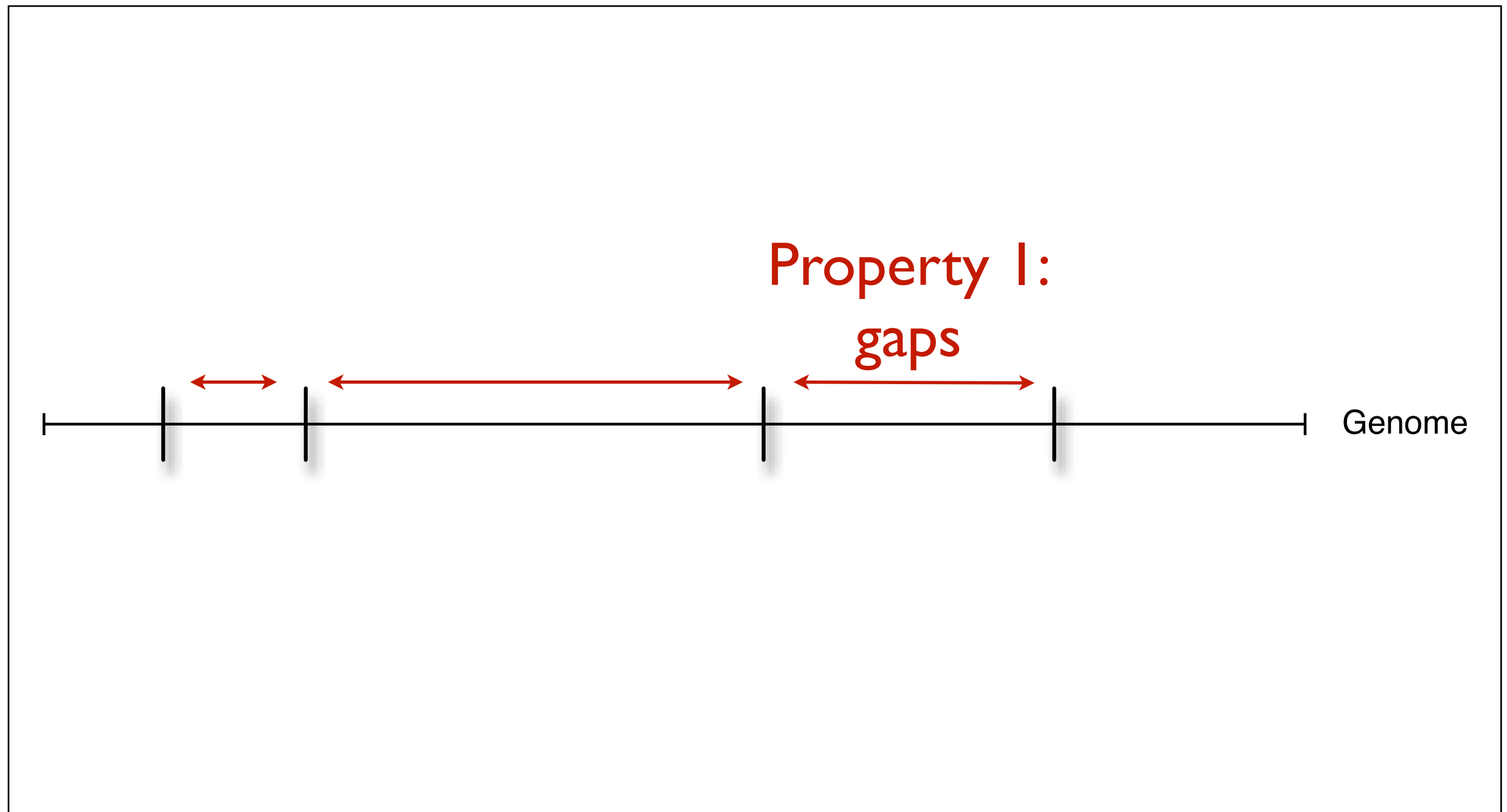


Example tracks:

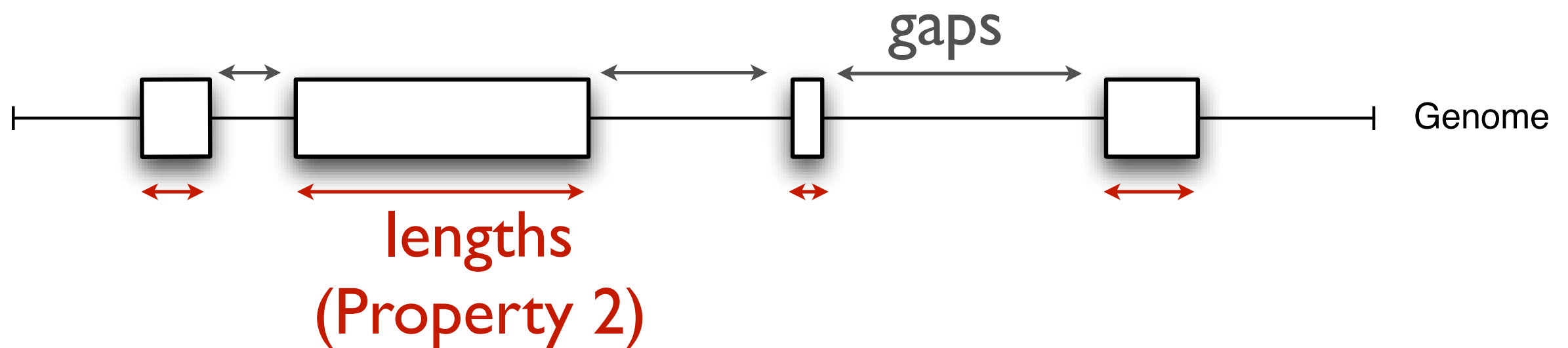
- Hi-C (3D chromatin conformation)



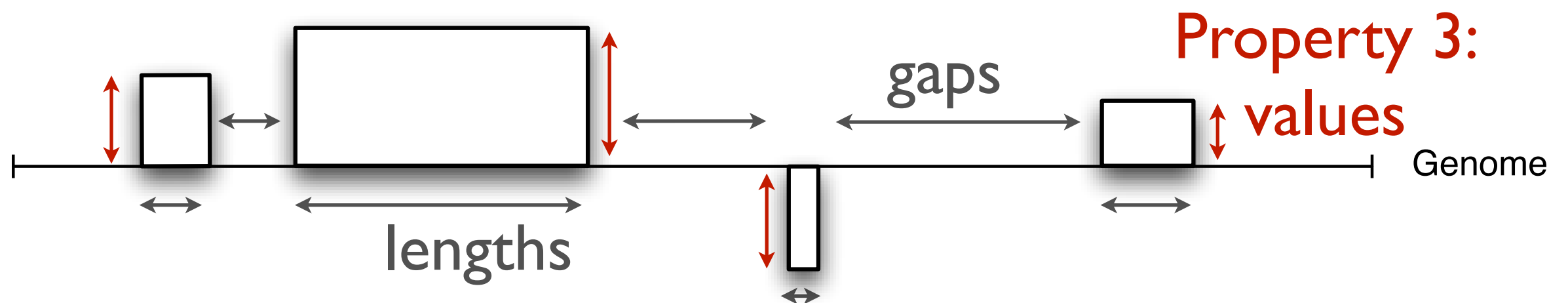
Core properties of tracks



Core properties of tracks

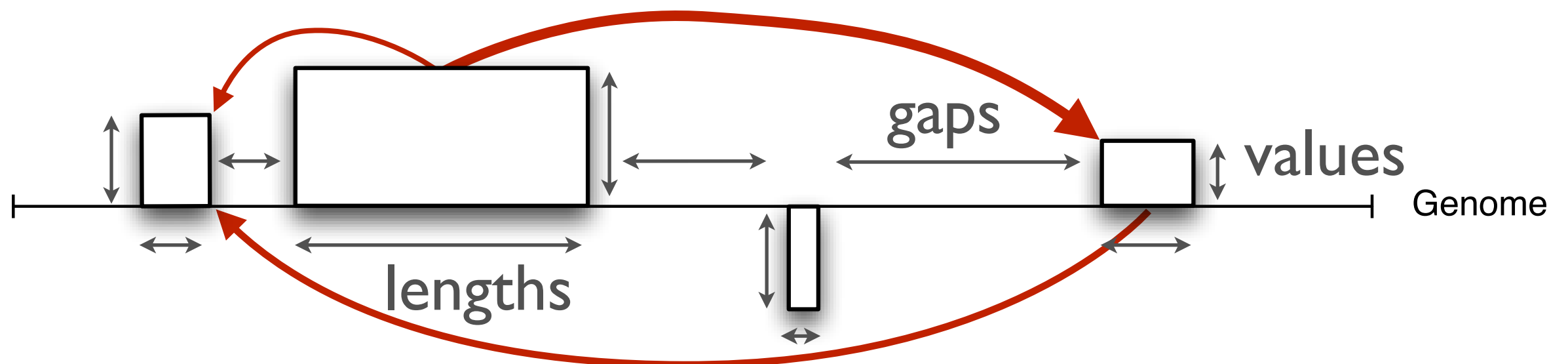


Core properties of tracks



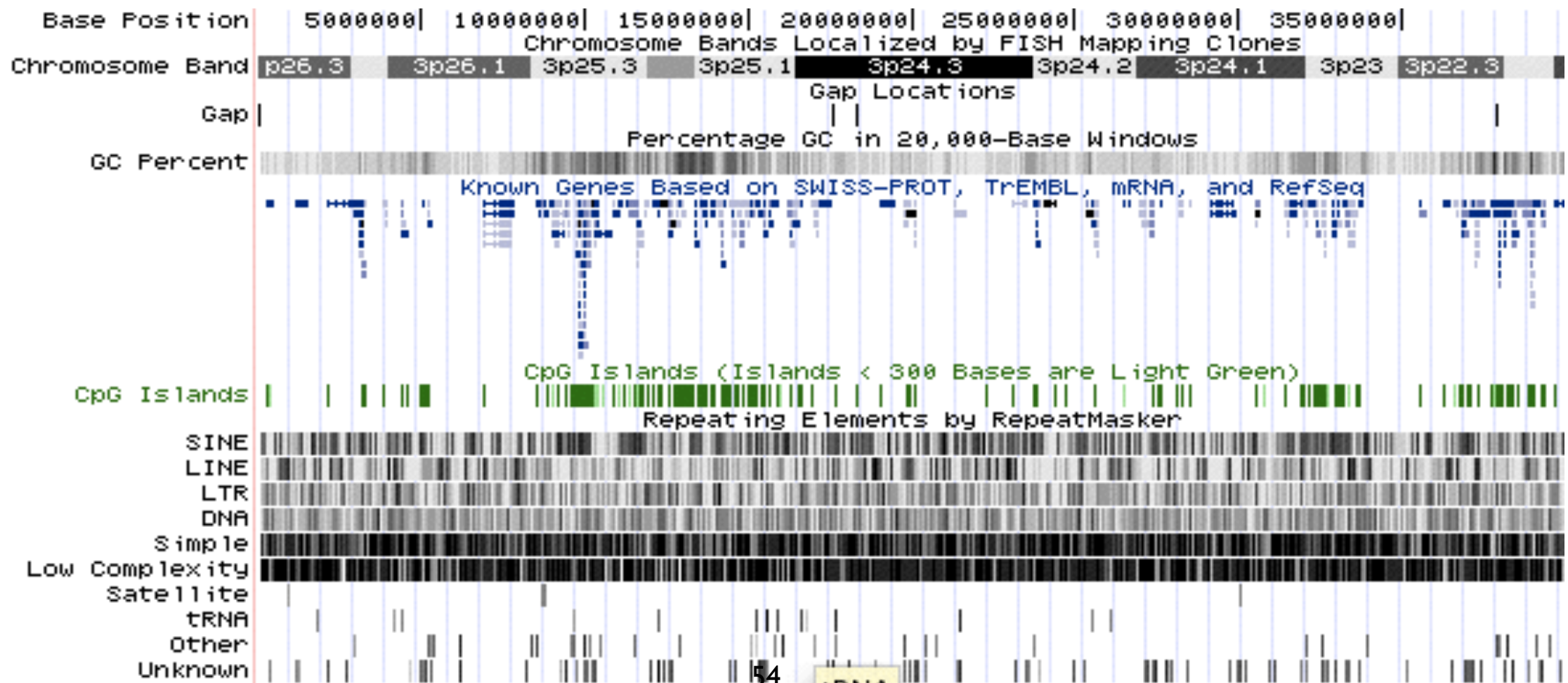
Core properties of tracks

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



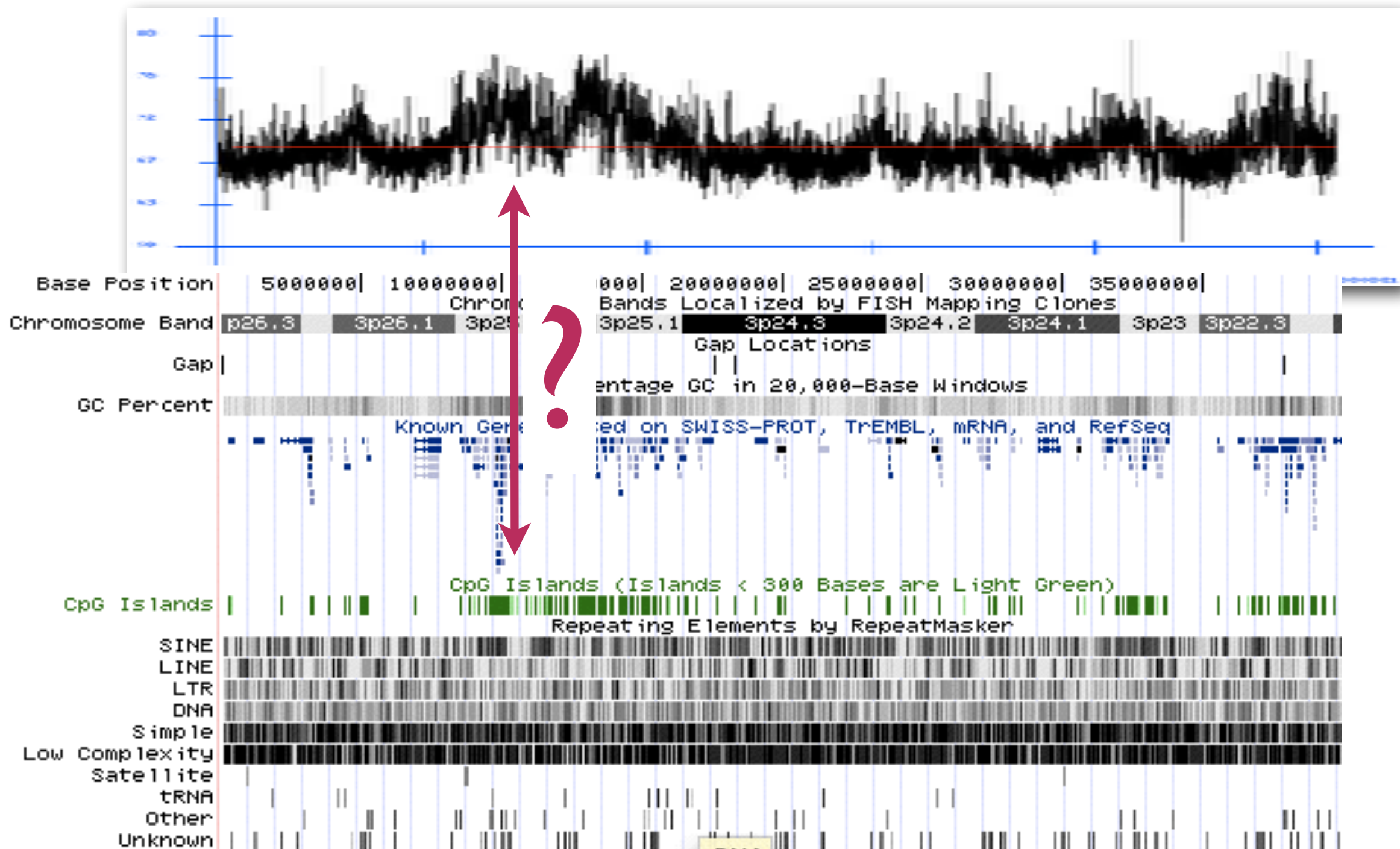
So, what about analysis?

Example analyses

- Age-associated hyper-methylated regions in the human brain overlap with bivalent chromatin domains (Watson et al. 2012)
- Genomic regions associated with multiple sclerosis are active in B cells (Disanto et al. 2012)
- DNase hypersensitive sites and association with multiple sclerosis (Sandve et al. 2012)

Example analyses (cont.)

- Vitamin D receptor binding, chromatin states and association with multiple sclerosis (Sandve et al. 2012)
- DNase hypersensitive sites and association with multiple sclerosis (Disanto et al. 2013)



This can't be it!!

Co-occurrence of genomic features

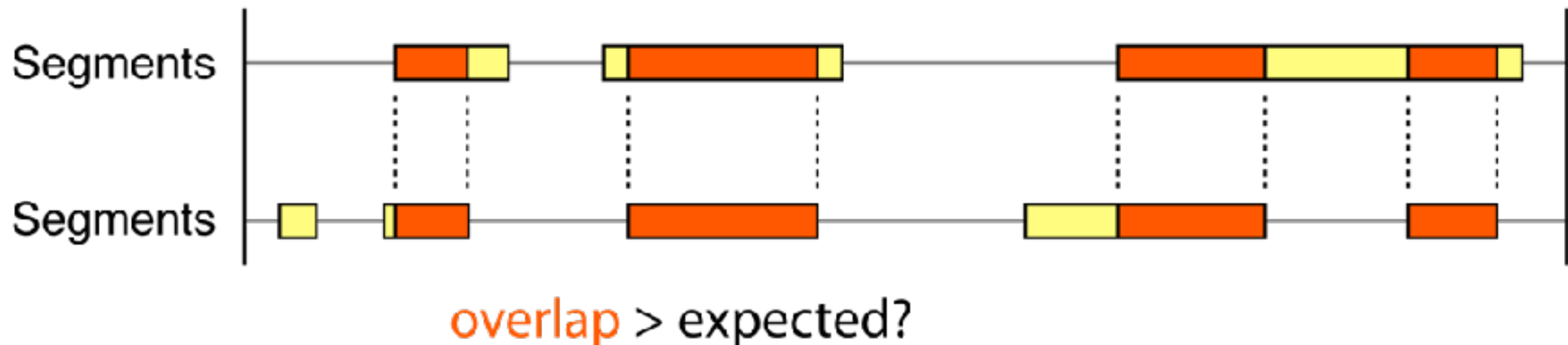
- Typical question:

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

Co-occurrence of genomic features

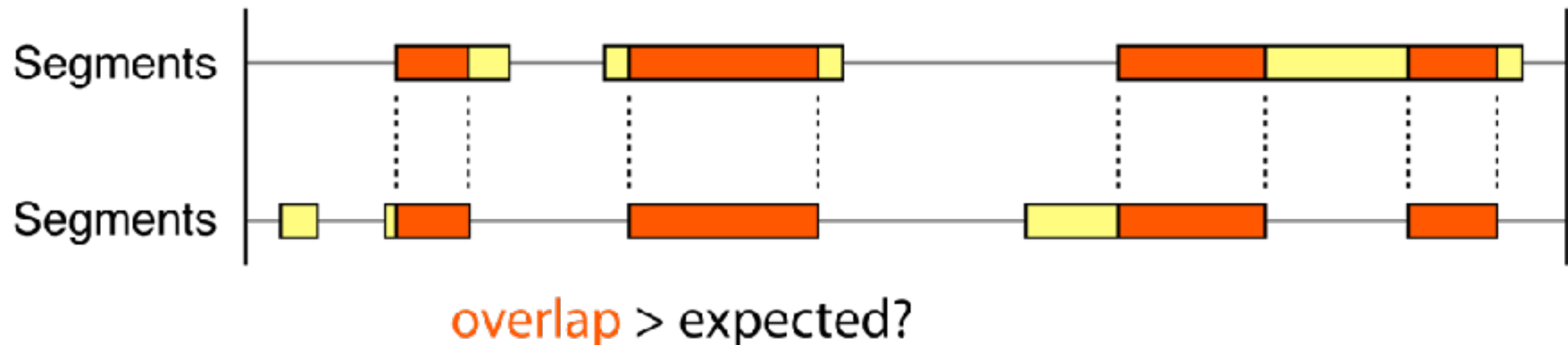
- What can such analyses be used for?
- Discover novel relations between tracks (can be done just by simply using public datasets):
 - May e.g. suggest that the biological features represented by the tracks are involved in the same cellular mechanism
- Relate experimental dataset to existing biological 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?



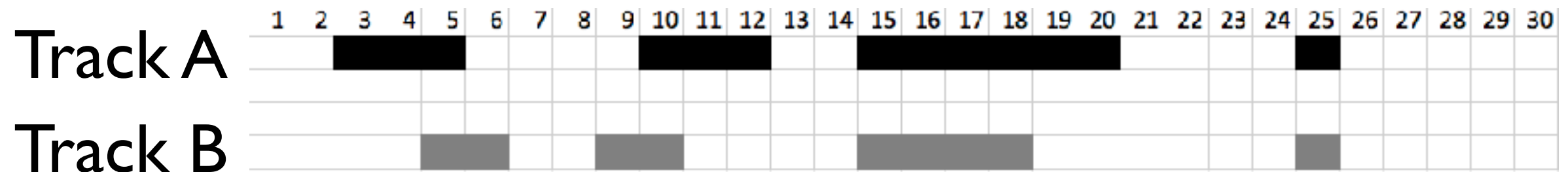
- As evident, this analysis makes sense when you have two tracks of type “segments”
- Generally, the type of analysis is dependent of the track types:
 - Each single track type defines a set of analyses appropriate for that track type (e.g. counting, coverage)
 - Each pair of track types defines another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

How does this look at the whiteboard?



What now?

Exercise 5

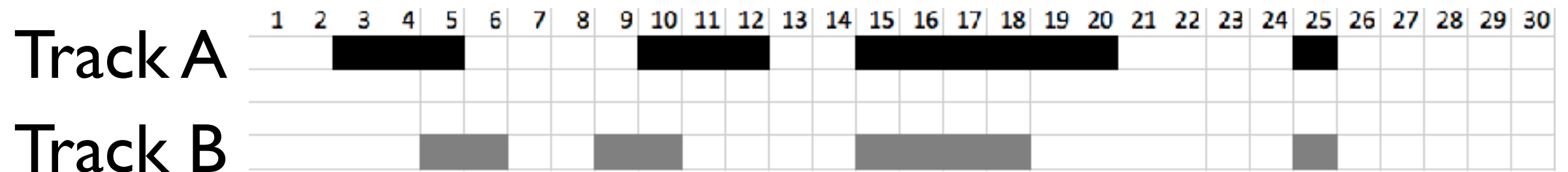


Calculate:

- the number of overlapping base-pairs between tracks A and B 7
- the proportion of overlapping base-pairs (in respect to the genome) 23.3%
- the expected number of overlapping base-pairs (assuming independent tracks) 3.9
- 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

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

Help: http://46.101.93.163/monte_carlo/

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? -
Count the proportion of boxes that are more extreme

Hypothesis testing

Statistical methods

- Basic idea behind statistical methods:
 - Consider the data are generated by a probabilistic model
 - Evaluate variability of the observed data in relation to what is expected to be generated by the assumed probabilistic model

Probabilistic model

- The data are generated non-deterministically - for a single event (e.g. measurement) instead of a single outcome, the probabilistic model describes a probability distribution, assigning a probability to each possible outcome.
- Parametric - the complexity of the model is bounded by its finite set of parameters (e.g. Normal distribution with given mean and variance).
- Non-parametric - the set of parameters is not finite, it depends on the current state of the observed data.

Intuitive example

- Someone claims that they can guess the outcome (head or tail) when a fair coin is flipped.
- Do an experiment to investigate
- You throw the coin 5 times, and the person guesses correctly every time.
- What is the probability of the claim being false?

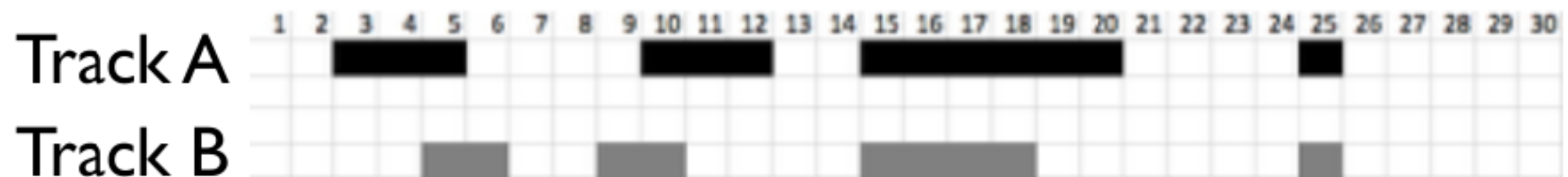
General setup

- Alternative hypothesis (H_1) - the claim you wish to test (e.g. person can guess coin flip)
- Null hypothesis (H_0) - a neutral baseline that can be reasonably assumed to be true (e.g. person can't guess better than an random guesser)
- Test statistic - measurement of the observed data that best captures the aspect of interest (e.g. nr of guessed coin flips, 5/5)

- **P-value** - given the assumption that H_0 is true, what is the probability to observe a value equal, or more extreme, of the observed ($p=0.5^5 = 0.031$)
- Significance level α - the cut-off under which the p-value is considered significant (often 0.05 or 0.01)
- If $p < \alpha$, then H_0 is rejected, meaning the evidence supports H_1 (e.g. the person is psychic?)
- Two-tailed vs. right-tailed vs. left-tailed

More realistic example

- **Claim:** The two genomic tracks, A and B, co-occur (more than expected by random chance)
- What is the null hypothesis?
- How can we compute the p-value in this case?



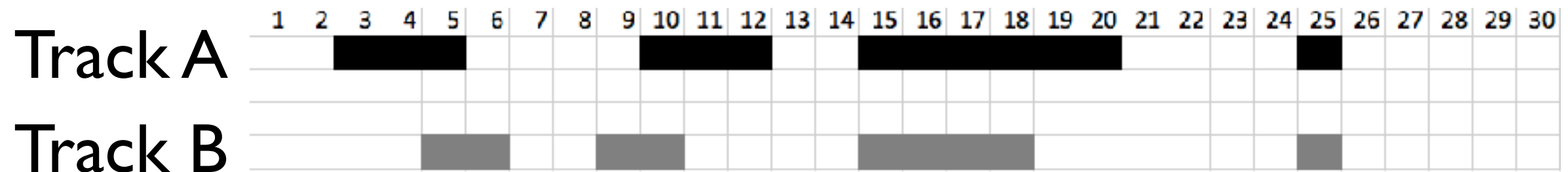
Null models

- A model from which the null hypothesis arises
- In genomics, mathematical computation of the null model is usually out of reach
- Simulation by Monte Carlo is often the solution (you already did this)
 - Permutation testing, but enumerating all possible permutations is not possible
 - For each randomization (of the track elements) calculate the value of the test statistic

- How to randomize the data?
 - Preservation of the structure in data
 - 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
 - Randomize one or both of the tracks

- 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 these cases we randomize the position of the track elements.

Exercise 6b



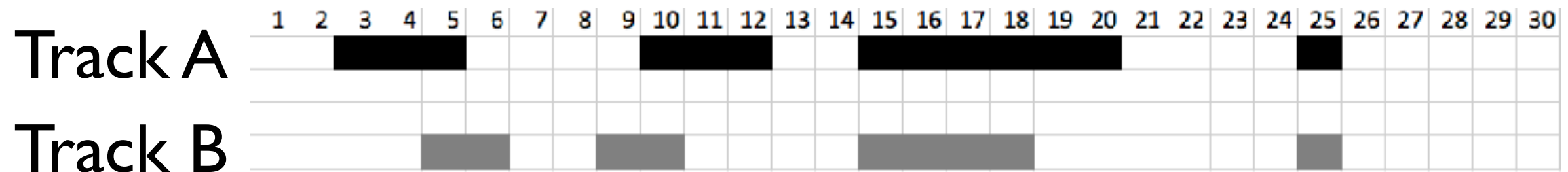
Create a random control track for track B, by

- Take each (grey) base pair and move it to a random location (do not keep existing segments)
- Take each segment and move it to a random location (preserving segment lengths)**
- 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:

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

What conclusion can you draw from the results?

P-value considerations and pitfalls

- ASA: <http://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf>
- Statement on statistical significance and p-values
 1. P-values can indicate how incompatible the data are with a specified statistical model.
 2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

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

Hypothesis testing errors

- When running a hypothesis test there's the possibility to make one of two types of errors
- Type I error: Rejecting H_0 when it is true
- Type II error: Not rejecting H_0 when it is false

	H₀ accepted	H₀ rejected
H₀ true	TN	FD
H₀ false	FN	TD

FD - False Discovery (Type I error)

FN - False Non-Discovery (Type II error)

Multiple testing

- When testing one hypothesis, with α set to 0.05, we accept the chance to make a false discovery (Type I error) 5% of the time.
- It is not uncommon, in an experiment, to test several hypotheses simultaneously.
- In genomics in particular, the number of independent tests can be in range of 10 000.

- In such cases, for a significance level 0.05, we expect around 500 false discoveries
- Even when the number of tests is relatively small ($m=10$), the probability of making at least one false discovery is high

$$1 - P(\text{no false discoveries}) =$$
$$1 - (1 - \alpha)^{10} = 1 - (1 - 0.05)^{10} = 0.4$$

	H₀ accepted	H₀ rejected	Total
H₀ true	TN	FD	T0
H₀ false	FN	TD	T1
Total	N	D	m

Controlling the errors

- Controlling Per-Comparison Type I Error (PCER) - uncorrected, $P(FD_i) < \alpha$ for all m tests
- Controlling Family-wise Type I Error (FWER) - e.g. Bonferroni, $P(FD_i) < \alpha/m$, $P(FD > 0) < \alpha$
- Controlling the False Discovery Rate (FDR) - $FDR = E(FD/D) < \alpha$

Bonferroni

- For m tests, the significance level is set to α/m ; the adjusted p-values are $P_i^{\text{adj}} = \min(m * P_i, 1)$
- The Bonferroni method for multiple test correction assumes all tests are independent of each other
- It is very conservative for large m , and it will rule out potentially interesting discoveries

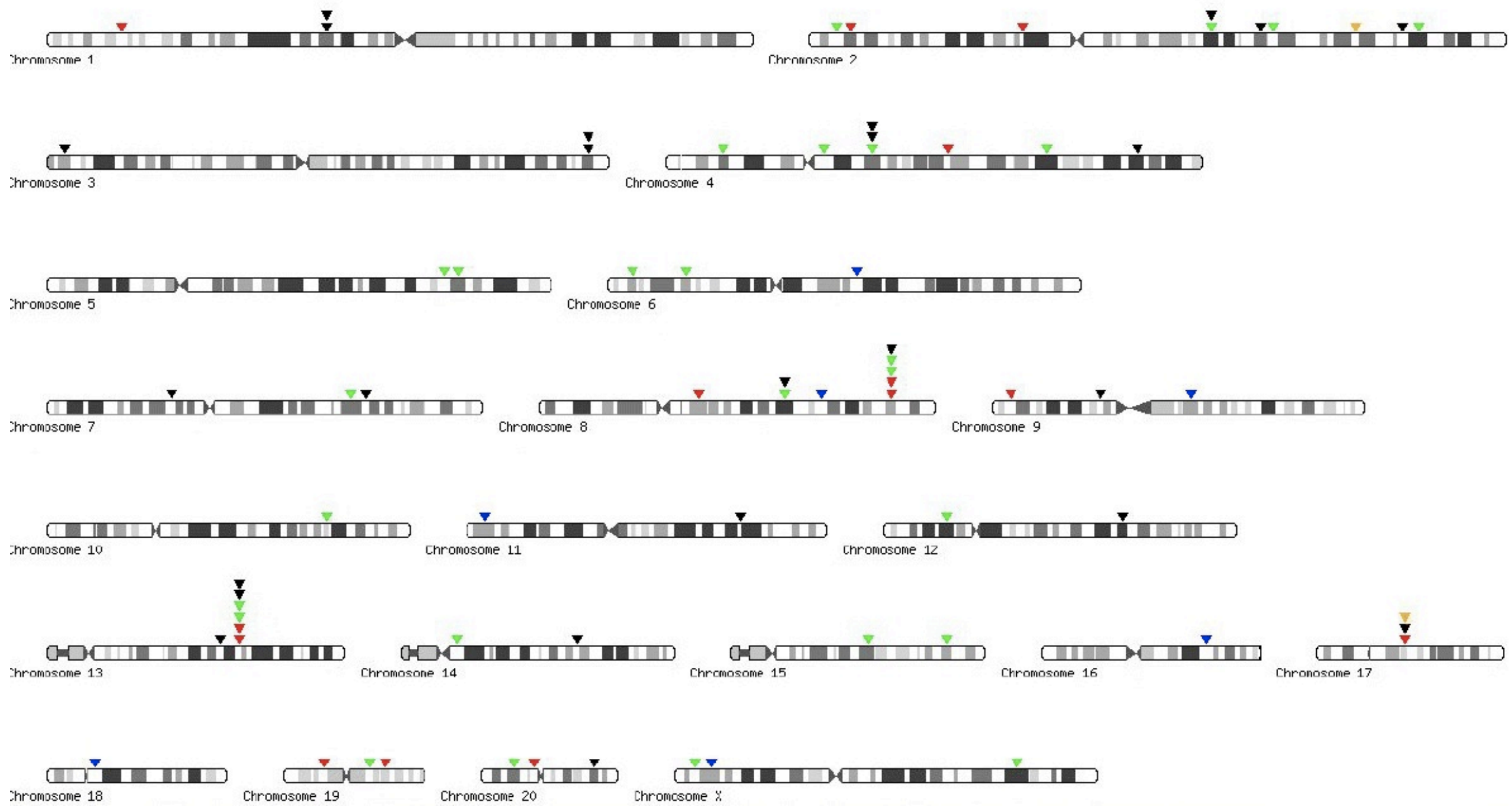
FDR - the Benjamini & Hochberg method

- Controls the expected proportion of false discoveries
 1. Select Q , the false discovery rate (e.g 0.1)
 2. Sort the original p-values $p_1, p_2, p_3 \dots$
 3. Compare each p_i to its corresponding BH critical value $q_i = (i/m) * Q$
 4. The largest $p_i > q_i$ is considered significant, as well as all the other smaller p-values.

A real example

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

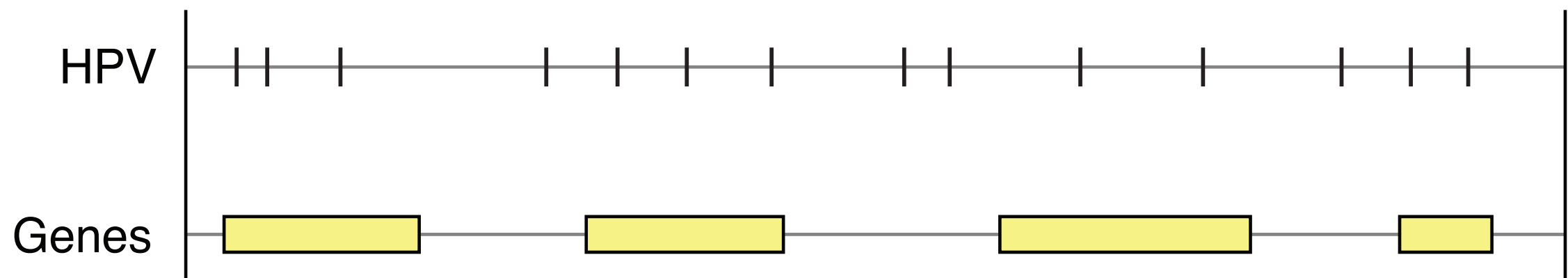
"Viruses might be expected to integrate near genes. Our results confirm such preferential localization inside genes for HPV [Human Papillomavirus], where 75 out of 119 determined integration sites (attached) fall inside genes."

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?

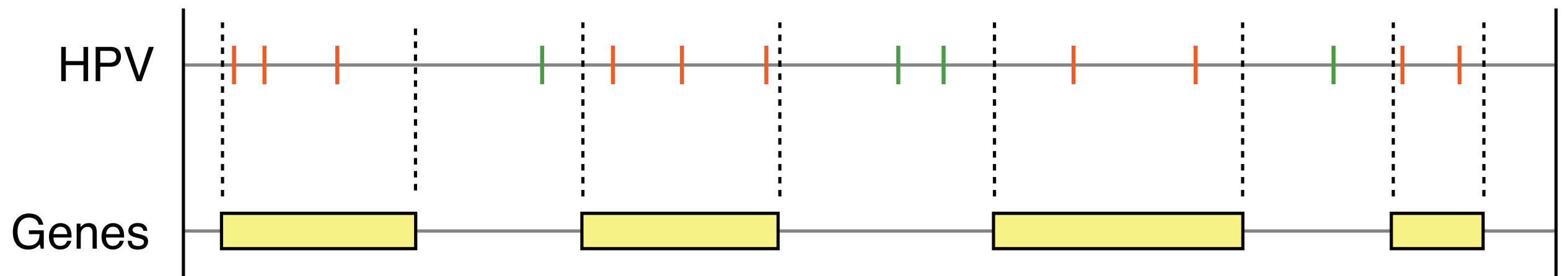
Exercise 7: HPV and genes

Student answers:

I. Which test statistic would you choose?

Observed vs Expected overlap	3	
Nr HPV sites outside genes	0	
Nr HPV sites inside genes	5	
Nr HPV sites near genes	2	
Proportion of HPV sites inside genes	12	

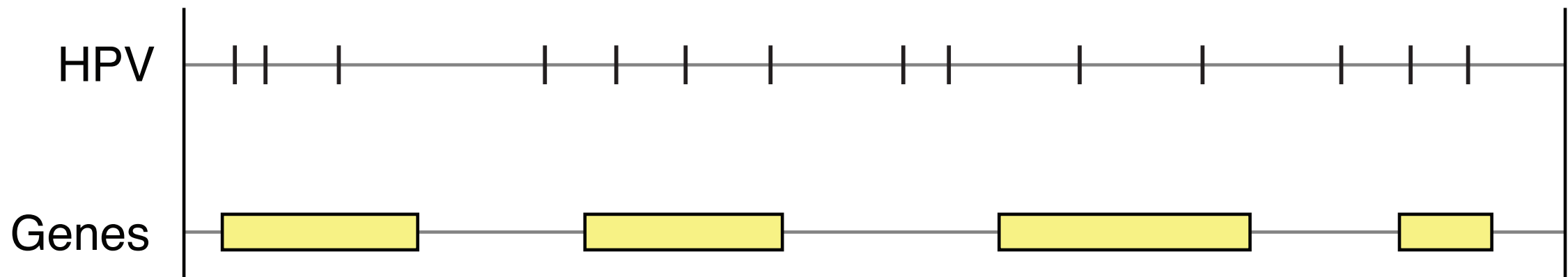
A possible test statistic



- 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

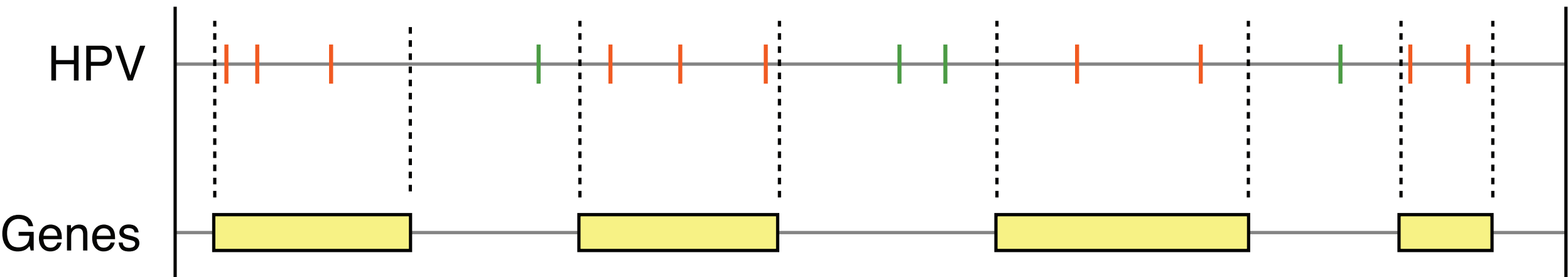
Exercise 8: HPV and genes

Student answers:

2. Which null model would you choose?

Randomize T1, preserve gaps T1 and nr of points	2	2
Randomize T1, keep nr of points	2	
Randomize T1, keep groups together	4	7
Randomize T2	0	
Randomize T1 and T2	0	
Randomize T2, preserve lengths and gaps	1	3

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 T2, Preserve nr of points T1, Randomize T1	Ensembl	0.0063
preserve T2, Preserve gaps T1, Randomize T1	Ensembl	0.0196
preserve T2, Preserve nr of points T1, Randomize T1	Ensembl	0.03
preserve T2, Preserve nr of points T1, Randomize T1	RefSeq	0.31
preserve T2, Preserve nr of points T1, Randomize T1	RefSeq	0.5071
preserve T2, Preserve nr of points T1, Randomize T1	Ensembl	0.05

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

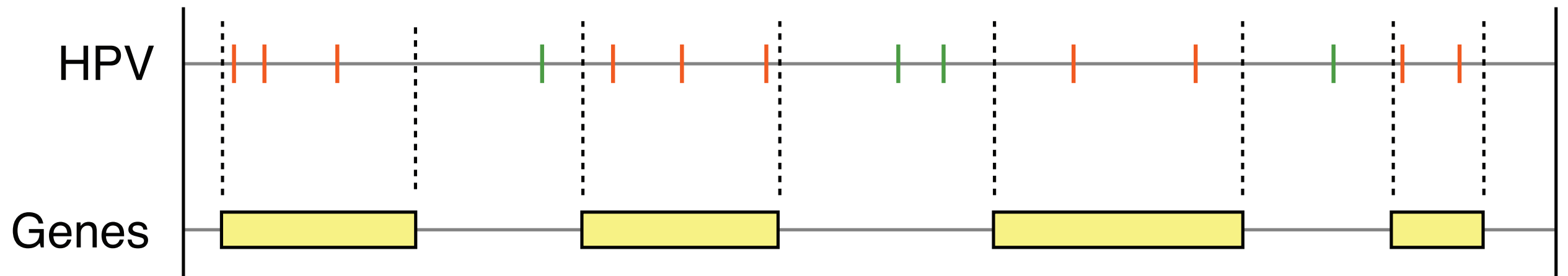
Further into statistical details: the test-statistic

- 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

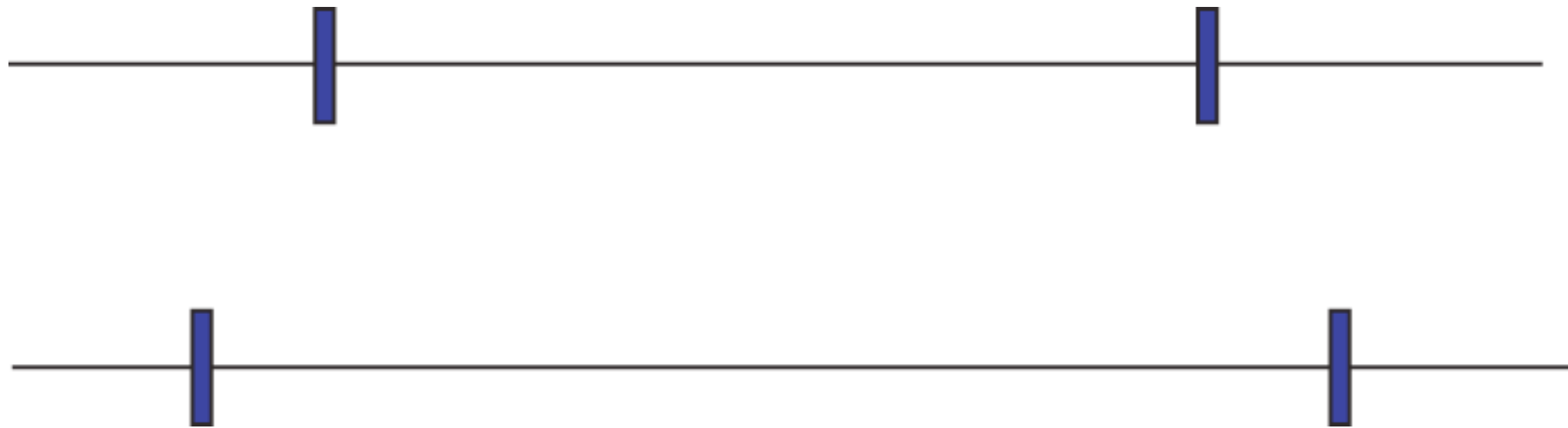
Back to the whiteboard: the test-statistic



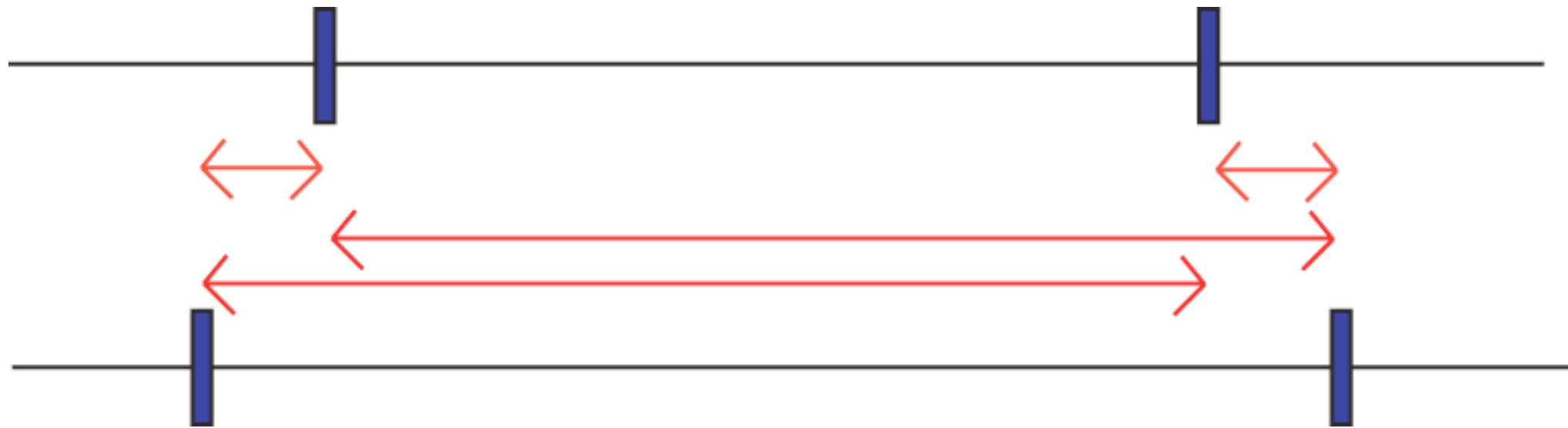
- For “located inside”:
 - Could simply count the number of HPV sites falling inside genes

Back to the whiteboard:

Must quantify “close”

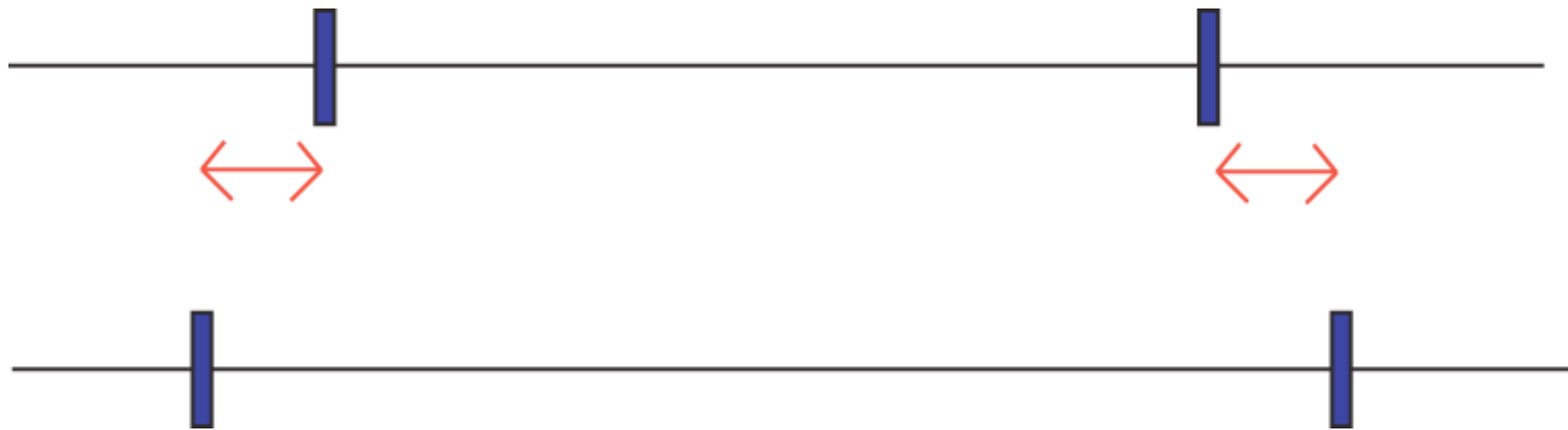


But that's trivial, sure: Just count bp distance!?



- 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!?



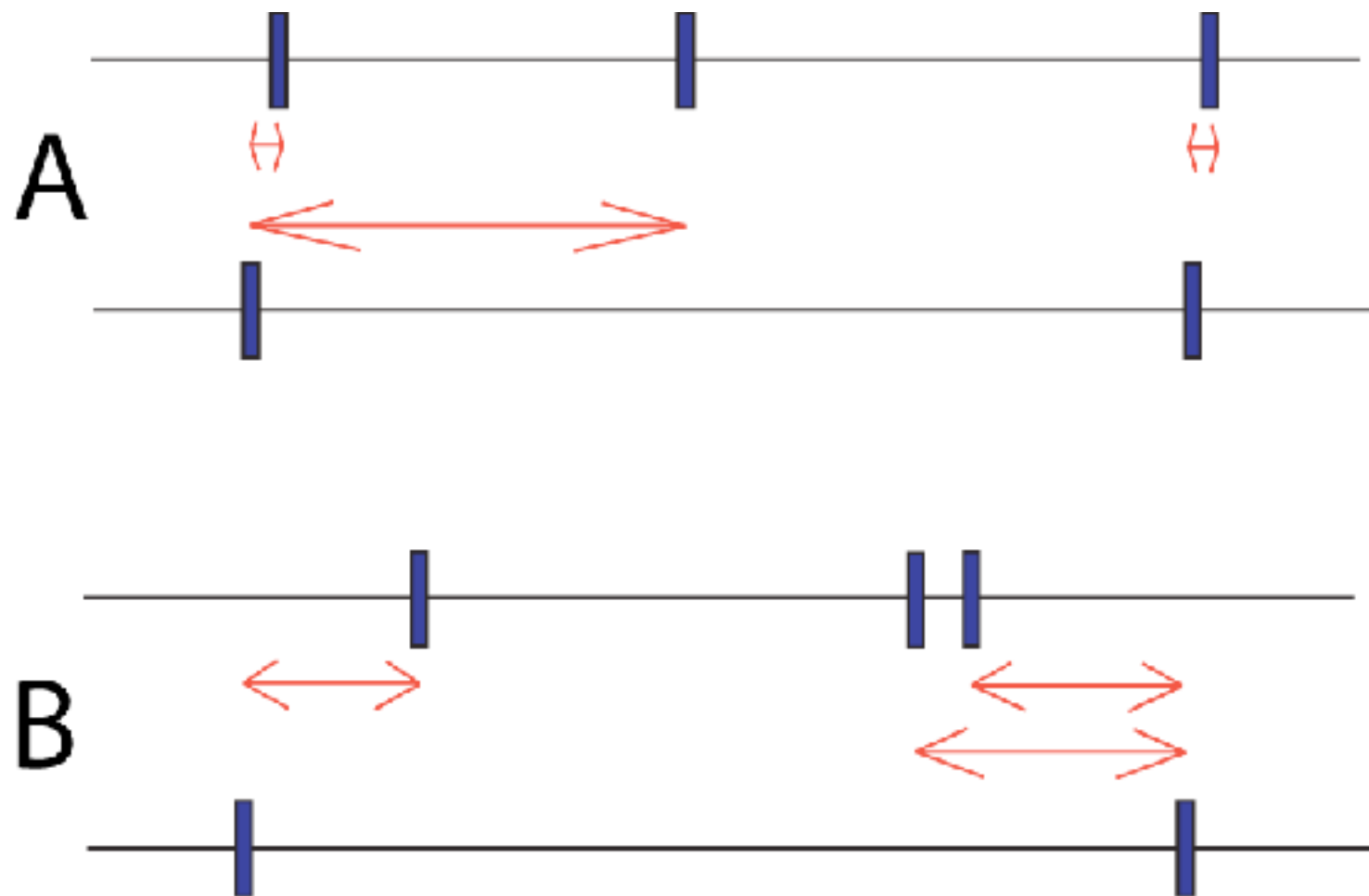
- But which distances - not all vs all!!
 - Only shortest! From 1 to 2! But MC needs a single number..

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



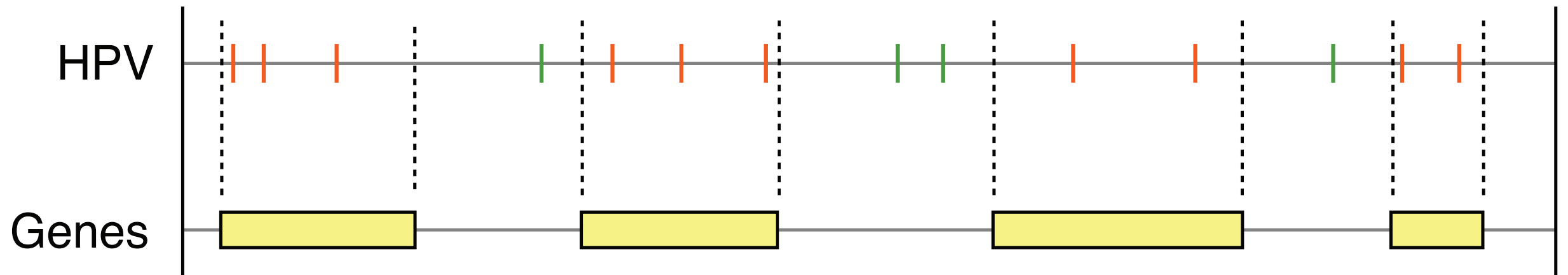
- But which distances - not all vs all!!
 - Only shortest! From 1 to 2! But MC needs a single number..
 - Can use sum/average of distances

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

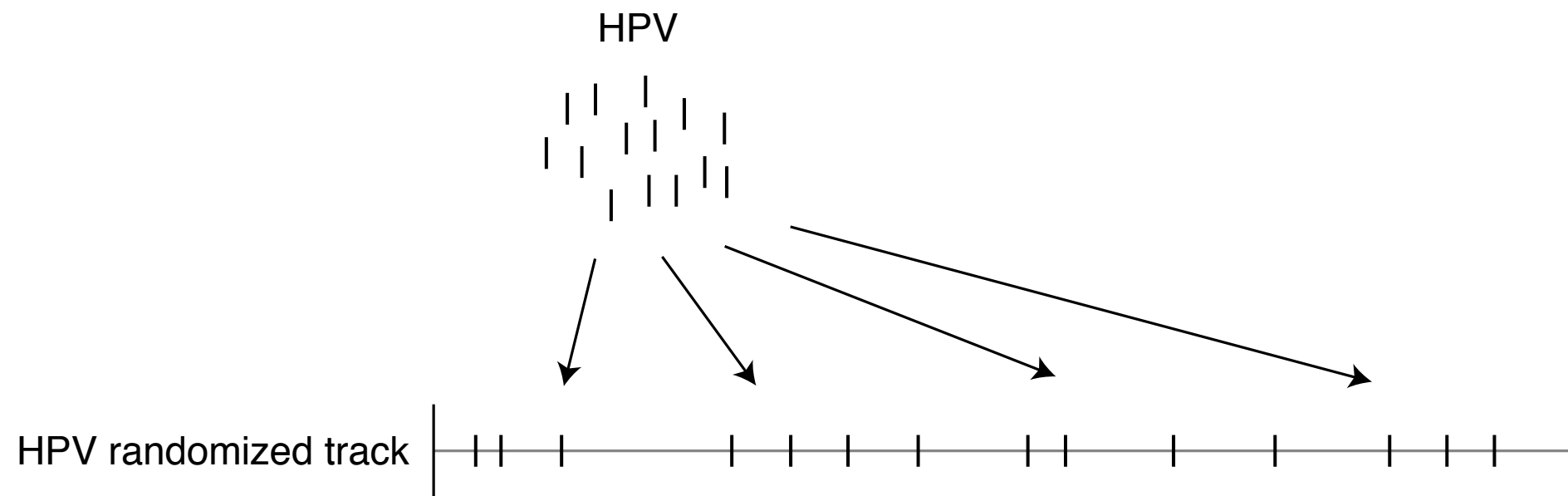
Further into statistical details: distributions



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

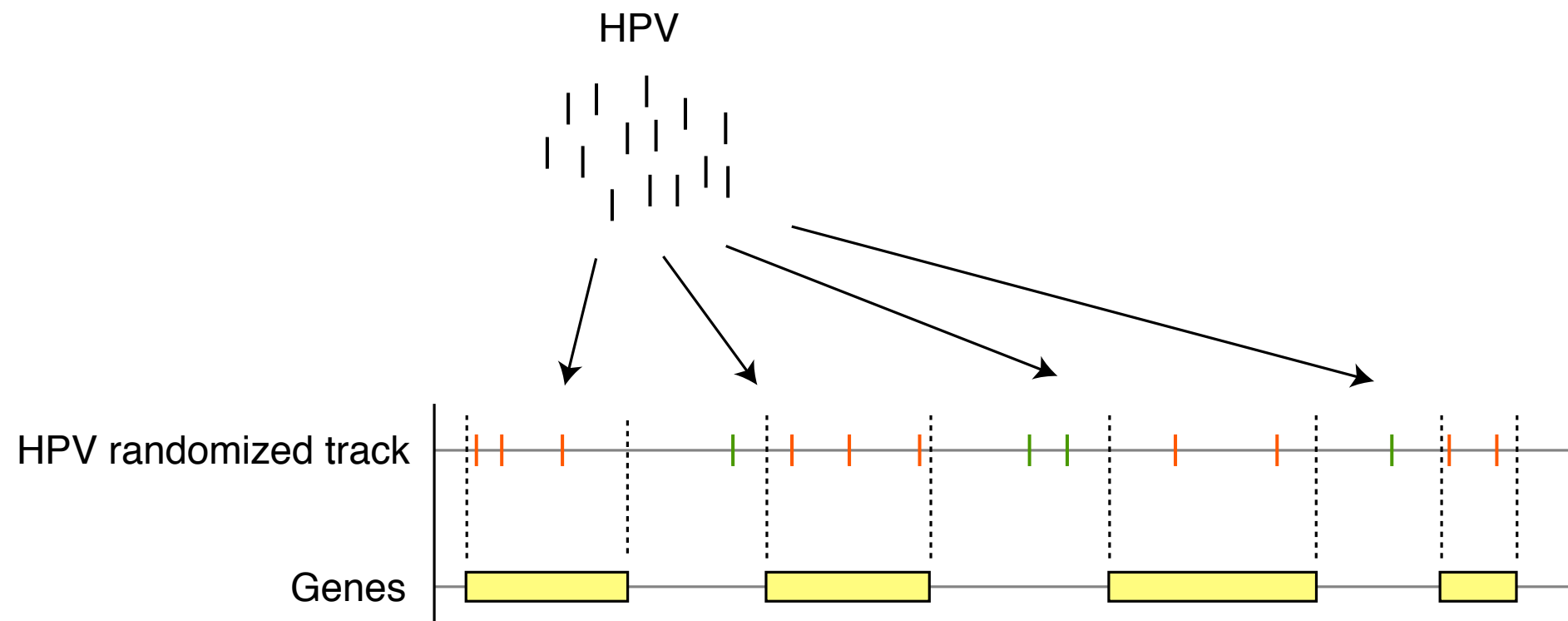
Monte Carlo test on “points inside segments”

- Randomize point (HPV) locations
(null model)



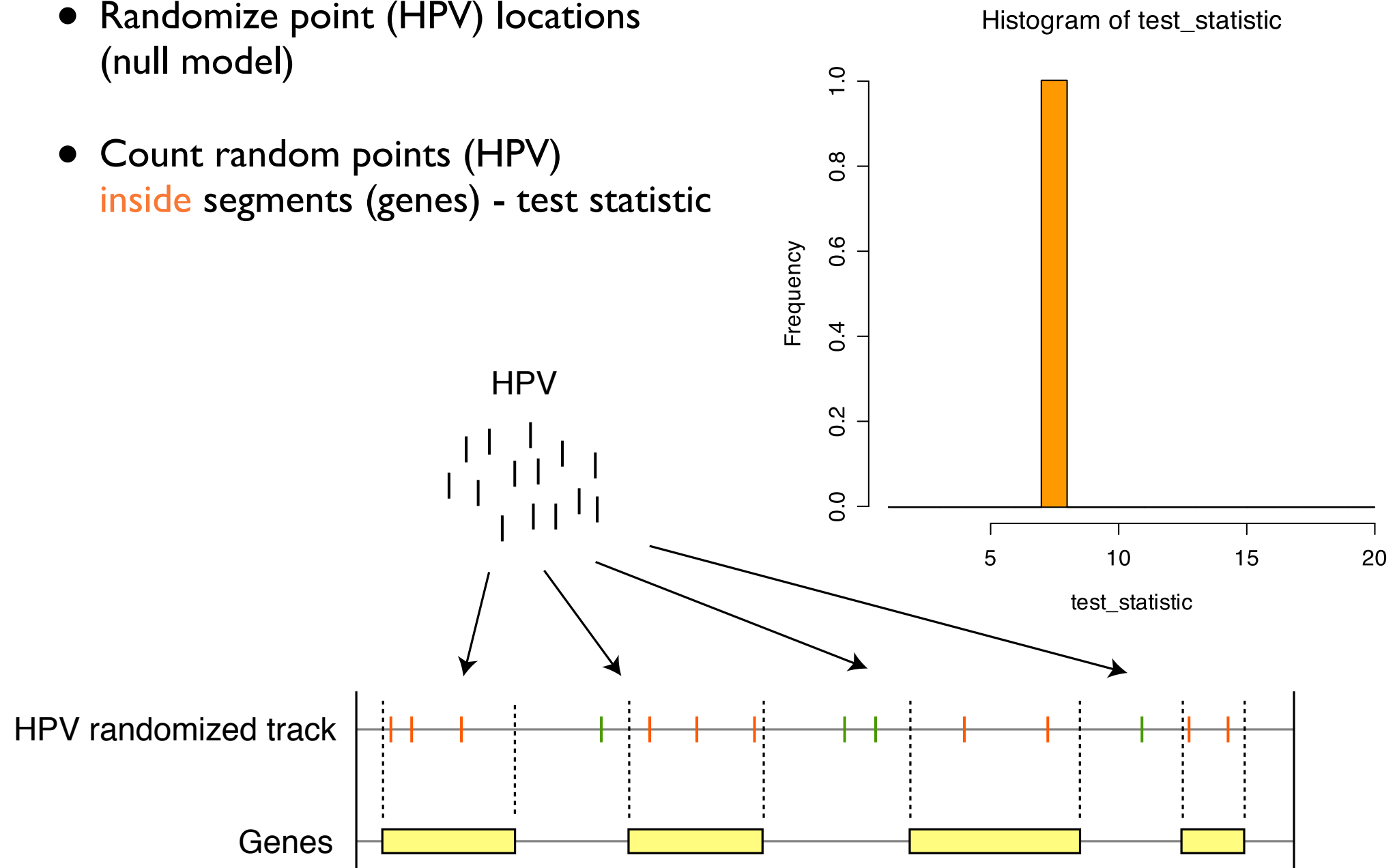
Monte Carlo test on “points inside segments”

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



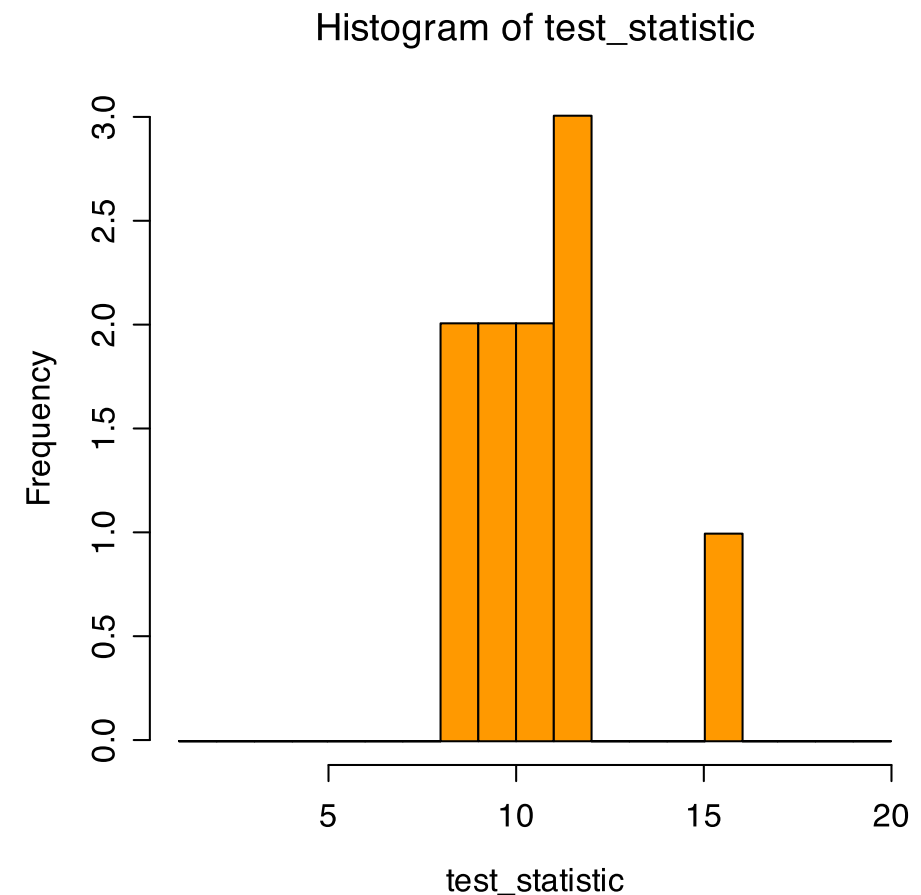
Monte Carlo test on “points inside segments”

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



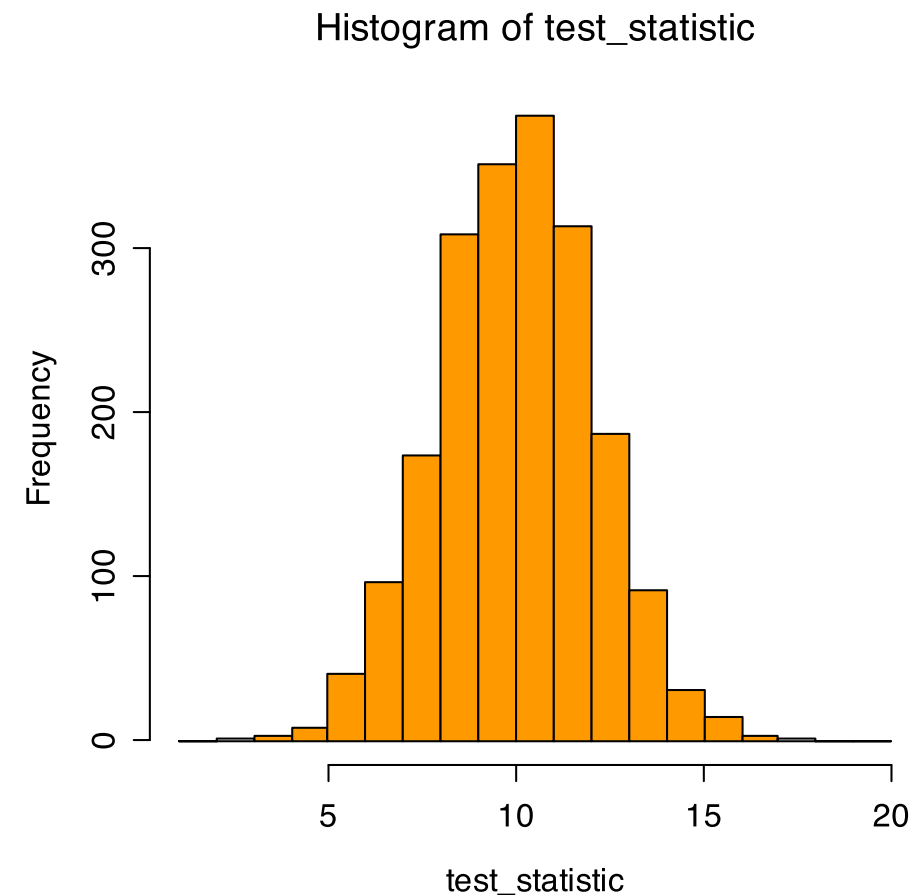
Monte Carlo test on “points inside segments”

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



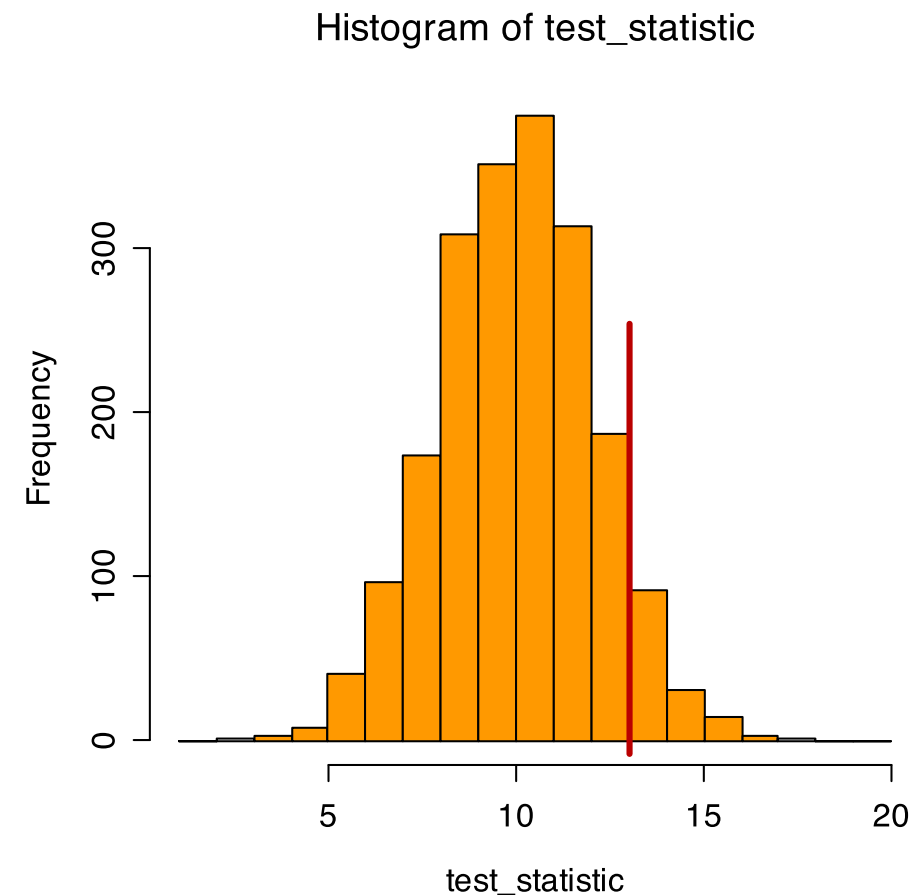
Monte Carlo test on “points inside segments”

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



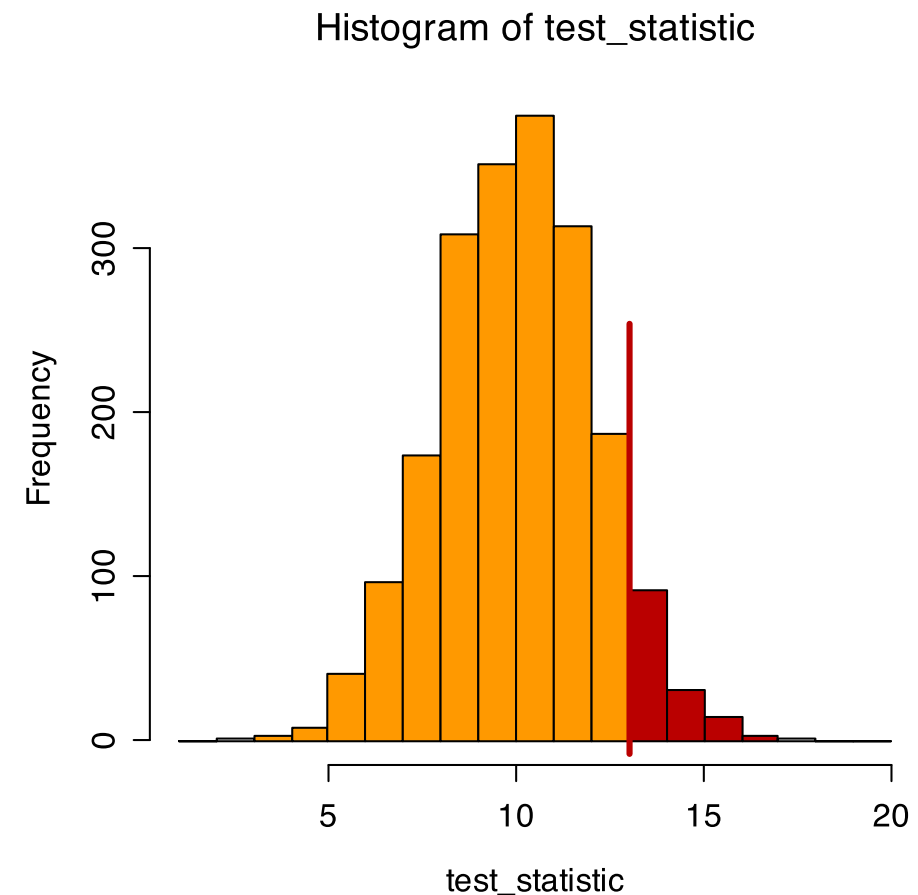
Monte Carlo test on “points inside segments”

- 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”

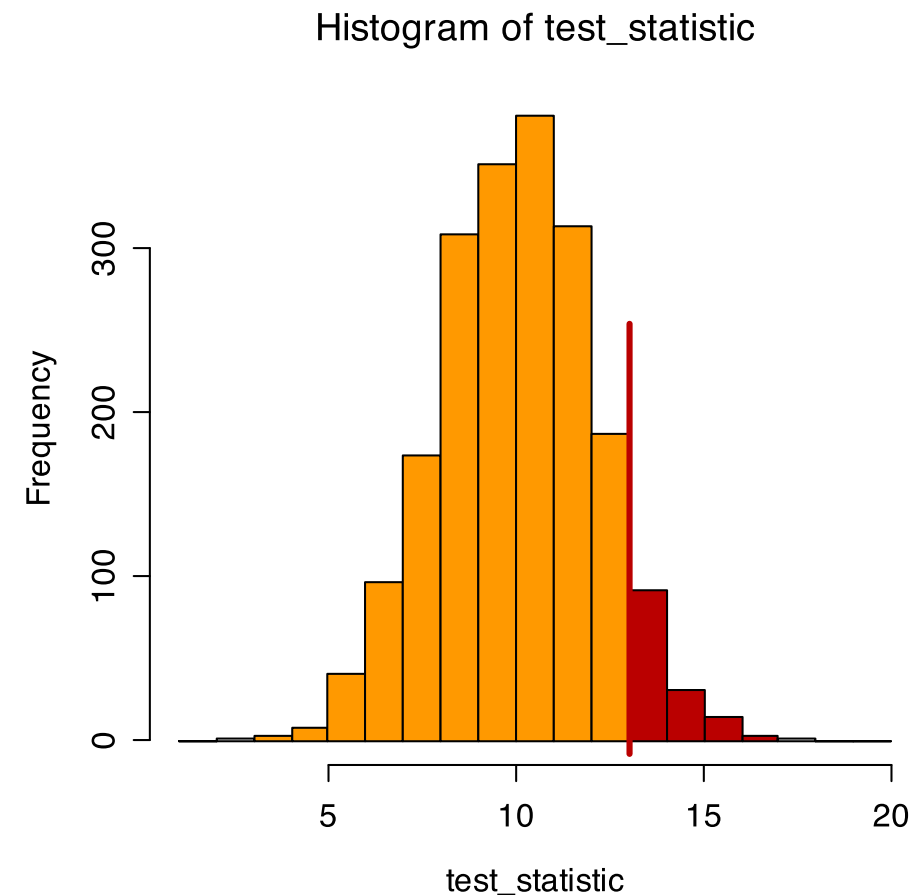
- Randomize point (HPV) locations (null model)
- Count random points (HPV) **inside** segments (genes) - test statistic
- Repeat a number of times
- Build histogram
- Compare with number of observed points (HPV) inside segments (genes)
- p-value = area to the right if alt hypothesis is “more” (if “less”,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



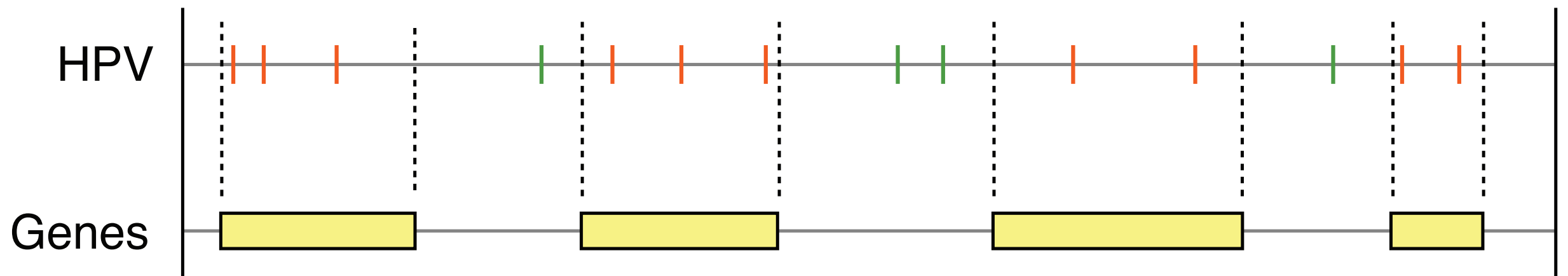
p-value = 0.08

Monte Carlo: distribution

- What we have done now is to build a random discrete distribution (with discrete meaning that it 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 H_0)
- A statistician may answer: “yes, a binomial distribution”

Binomial distribution

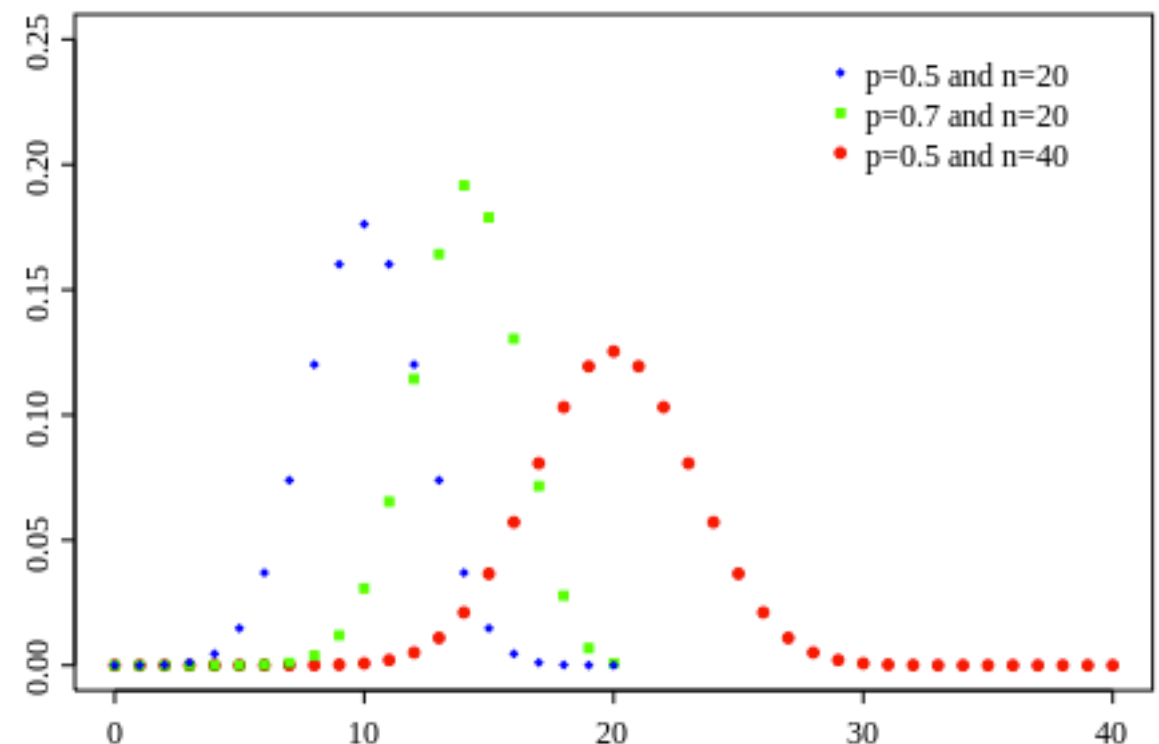
- Flip a coin ***n*** number of times
 - Two outcomes: heads or tails
- But: one side may be heavier than another

- E.g. the probability of tails:

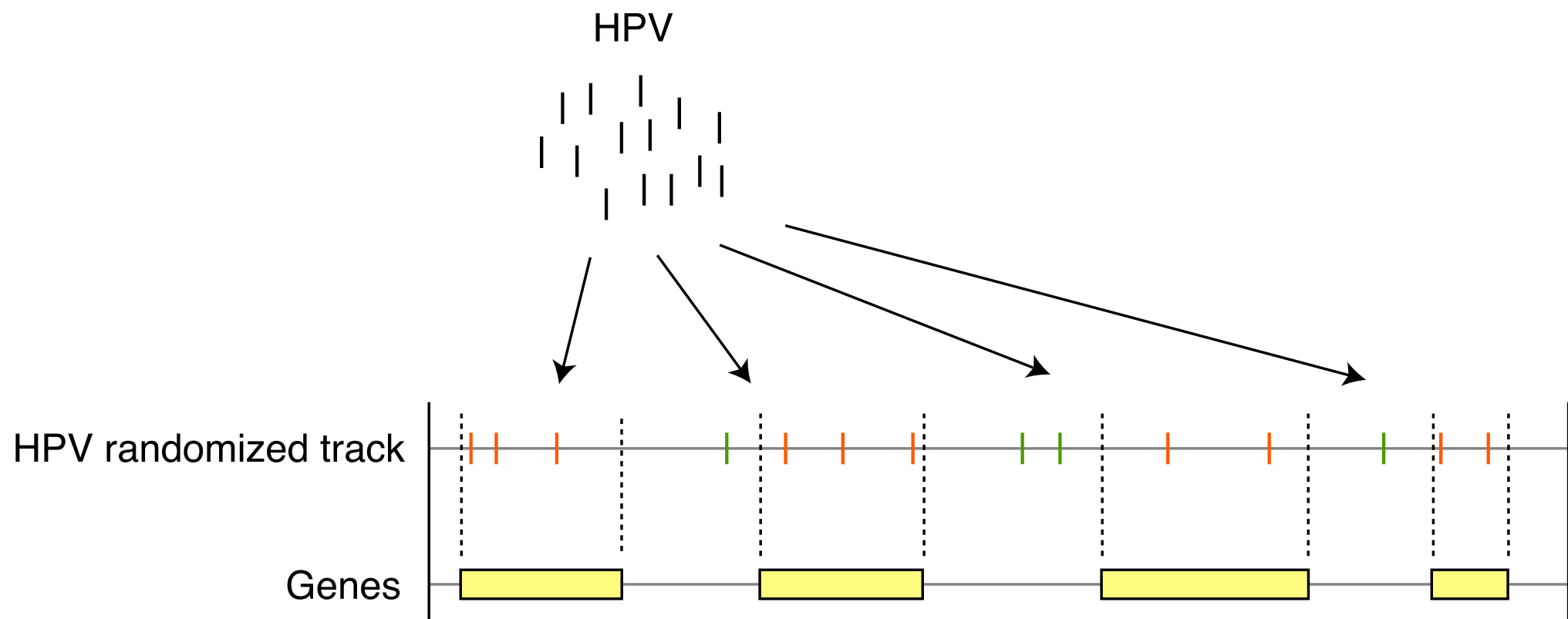
$$P(\text{tails}) = p = 0.6$$

$$P(\text{heads}) = 1-p = 0.4$$

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

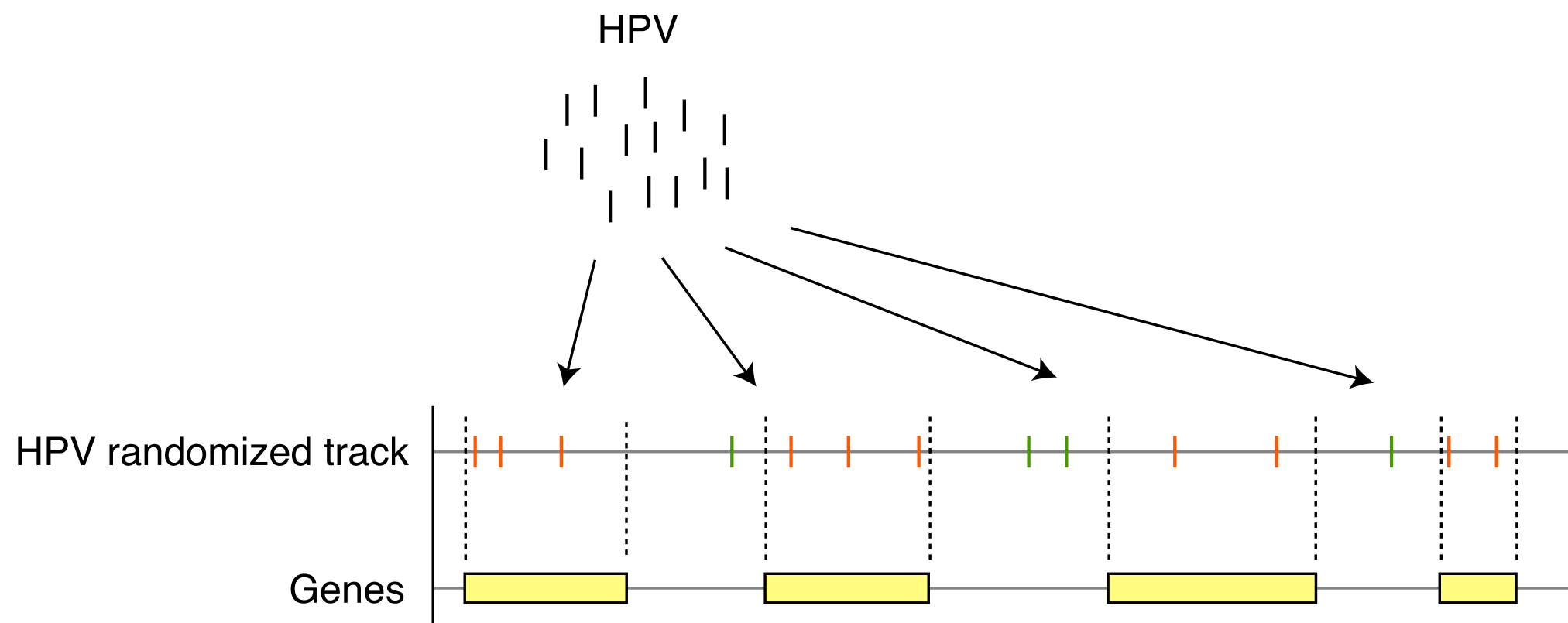


Binomial distribution



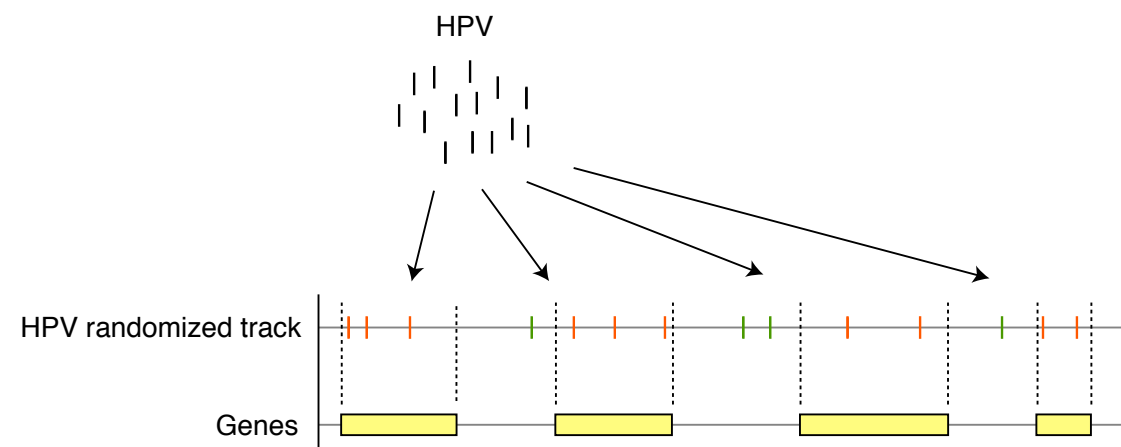
- In this case, each HPV is a coin, and it can either fall into a gene or not, depending on how much of the genome that is covered by genes
- n = number of HPV
- p = proportional coverage of genes

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

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 II

- Goal: Get familiar with the GSuite HyperBrowser
<https://hyperbrowser.uio.no>
- 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

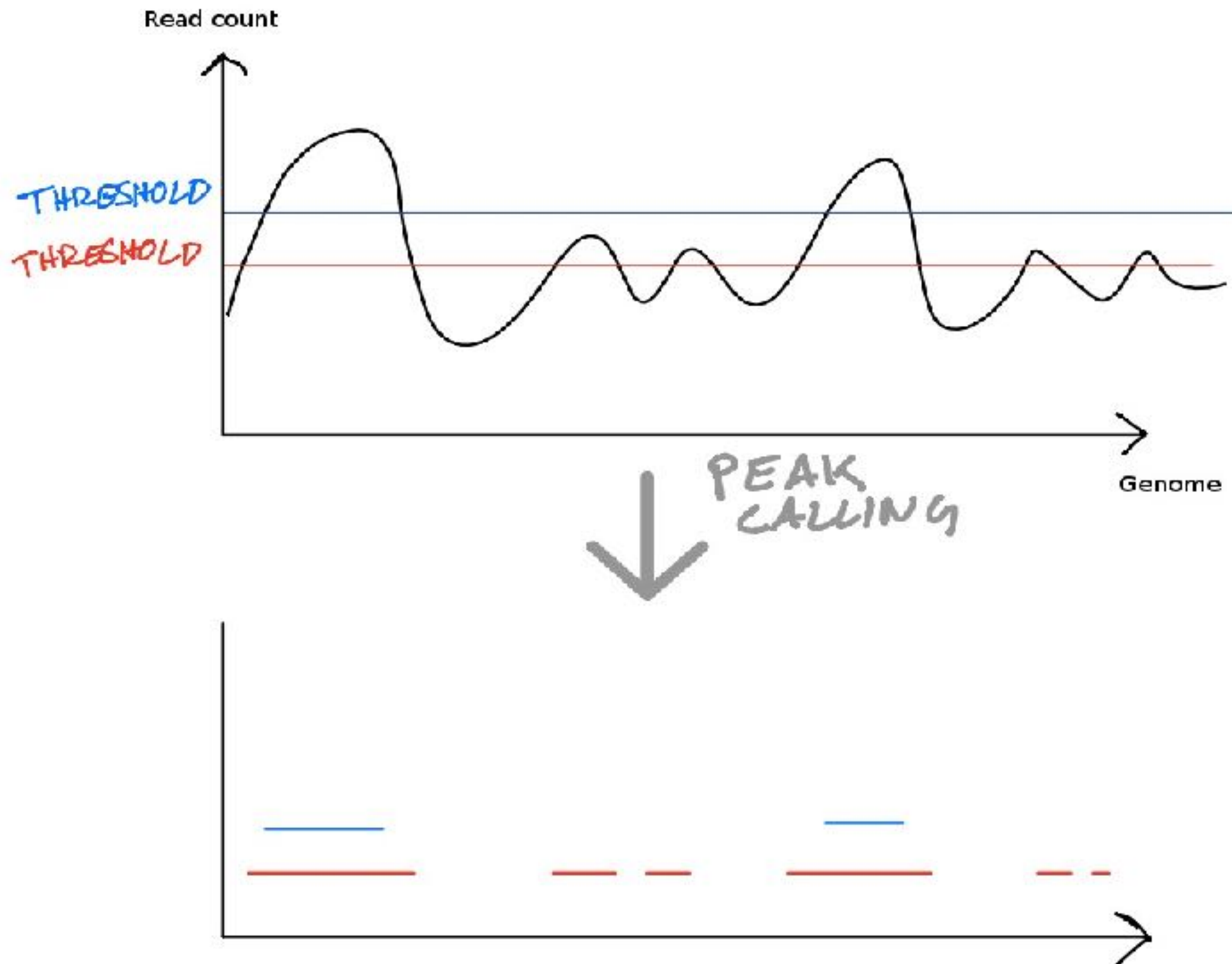
... and the following question:

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

Discussion on similarity measures (cont.)

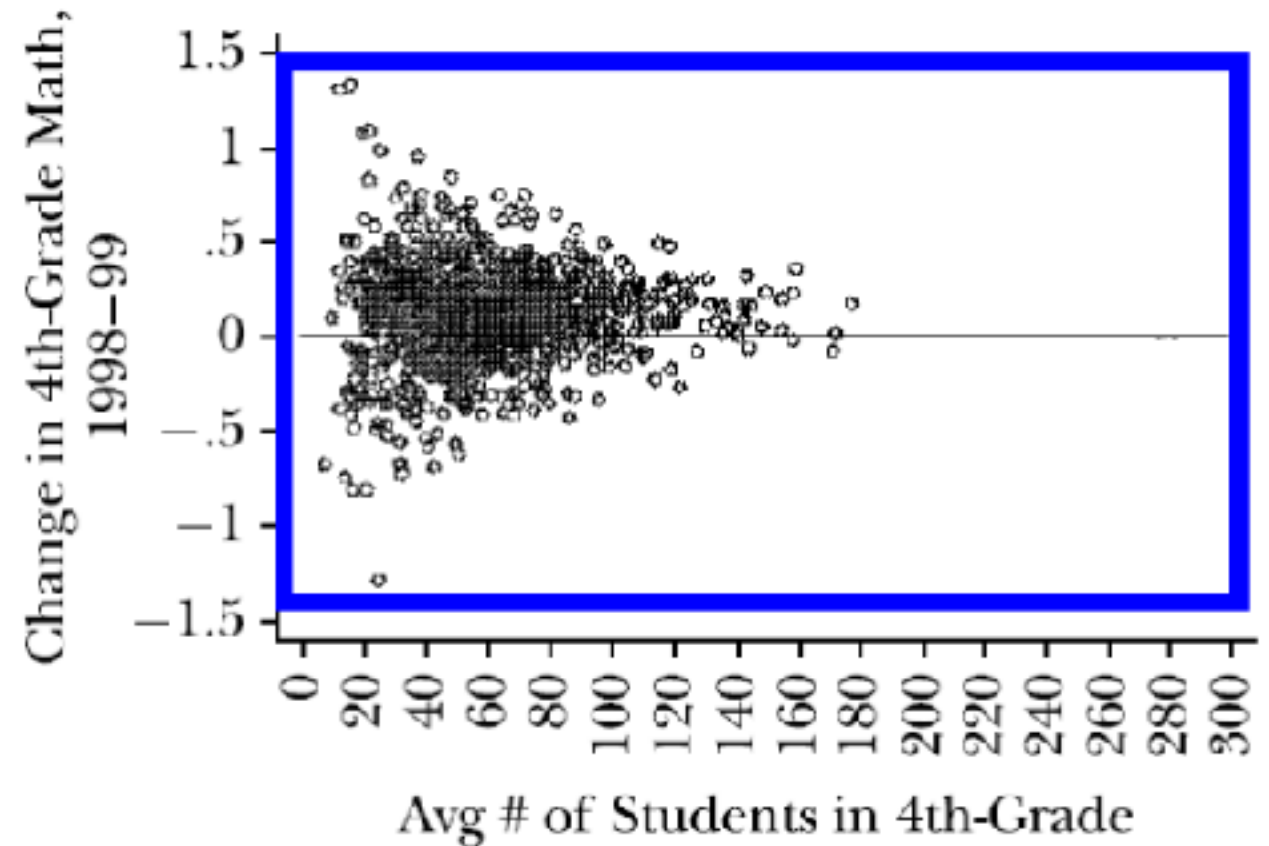
- Challenge:
 - The different TF tracks are of different “size”, meaning that the number of base pairs covered by each track varies a lot.
 - Not because of less or more binding
 - Can be because some data sets lack data, or are produced in a more conservative way (e.g. stricter threshold in peak calling)

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/>

Binary similarity measures

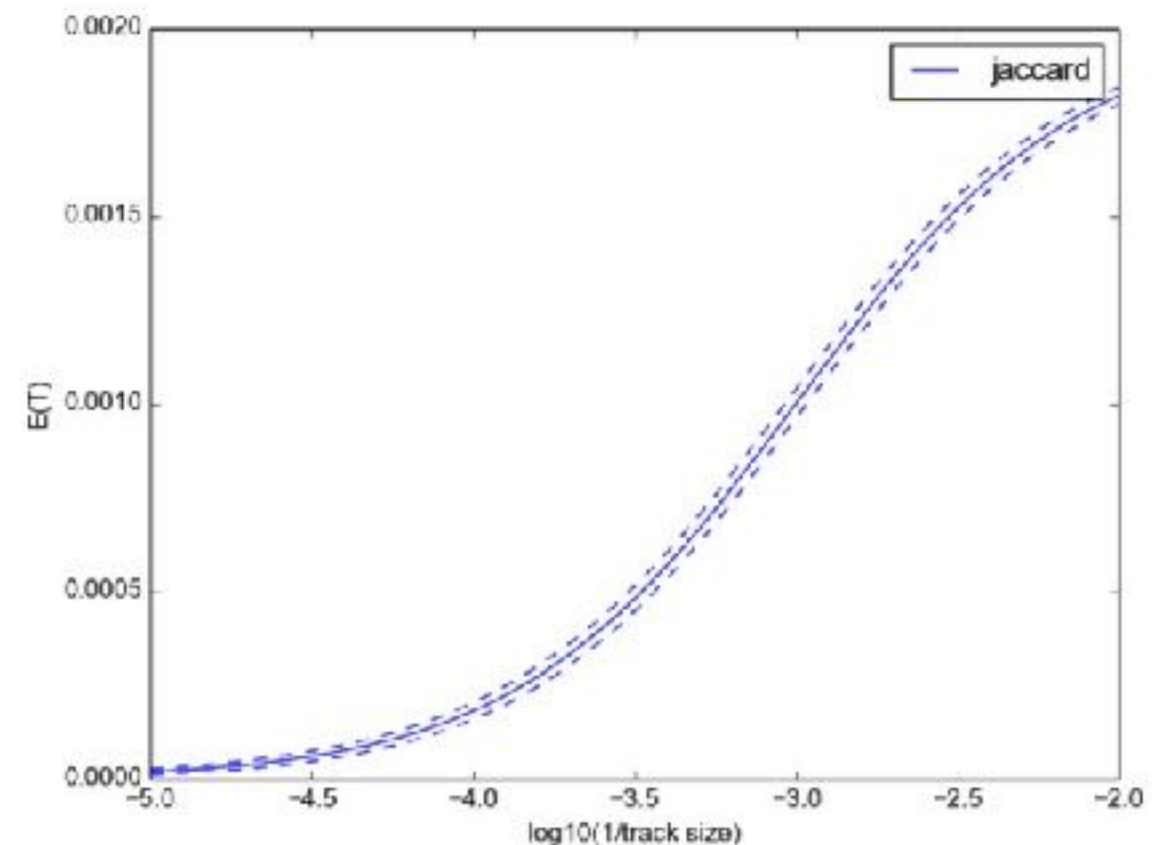
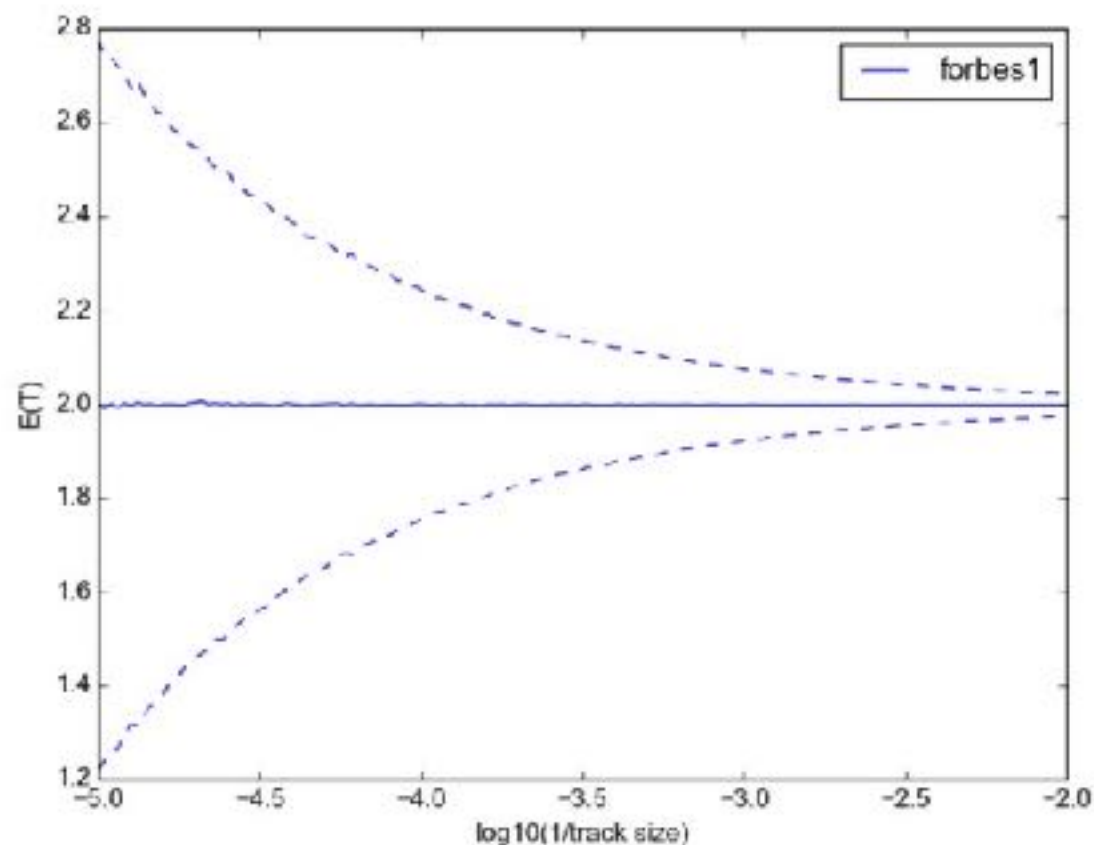
- A binary similarity measure is a function $S(\text{trackA}, \text{trackB})$ that takes two tracks as input and returns the “similarity” between these two tracks
- Called “binary” because originally not used for genomic tracks, but to compute the similarity between two binary vectors
- The most famous 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 pairs¹⁴⁶ covered by both tracks)

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 sizes
- 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](http://www.iiisci.org/journal/CV$/sci/pdfs/GS315JG.pdf)
- In the Genomic HyperBrowser (Gsuite), you can select among different binary similarity measures






Exercise 12a

- 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

Exercise 12b

- For the same datasets, rerun the analysis, this time also generating p-values for each of the histone modification tracks
- NB: Select “moderate resolution of global p-value”
- Use Bonferroni and FDR multiple testing correction to determine which of the results are statistically significant

Exercise 12 results

Rank	Track title	Similarity to query track 	P-value 	Overlap between query and reference track (bps) 	Genome coverage of track (bps) 	Nr. of reference track elements 	mark
1	E047-H3K4me1.broadPeak	2.44188781332	0.00398406374502	337813	375646878	113865	H3K4me1
2	E047-H3K9ac.broadPeak	1.35101124563	0.00398406374502	387124	778072857	1033352	H3K9ac
3	E047-H3K4me3.broadPeak	1.48153128847	0.00796812749004	231744	424743561	402716	H3K4me3
4	E047-H3K27ac.broadPeak	1.56420675445	0.0159362549801	213764	371081705	302688	H3K27ac
5	E047-H3K36me3.broadPeak	1.29301029128	0.286852589641	294972	619452406	410731	H3K36me3
6	E047-H3K27me3.broadPeak	0.965243769222	0.892430278884	303457	853668558	577299	H3K27me3
7	E047-H3K9me3.broadPeak	0.450777677098	1.0	116667	702772505	273731	H3K9me3

Summary of statistical epigenomics

Data

- High-throughput sequencing
 - RNA-Seq (position of expressed genes)
 - Variant-calling (position of SNPs or other variants)
 - ChIP-seq (position of e.g. transcription factor binding sites)
- Typical formats you will be using in real analysis:
 - VCF
 - Bigbed, bed
 - Any files containing the position of genomic elements

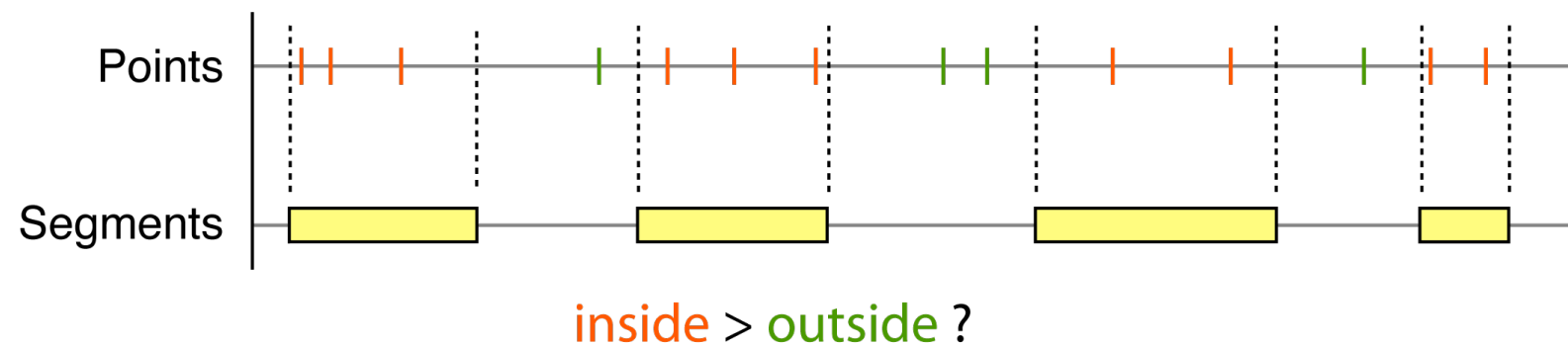
Data representation

- *Track*: A set of genomic features
 - A dataset that can be positioned along the reference genome
- Tracks are represented by different *track types*, which are models that makes it easy to represent the track on a computer (e.g. in a text file)
 - *Examples*: Segments, valued points, genome partition



Analysis

- Typical question: Do genomic feature A and B co-occur more than expected by chance?
 - We answer this question using a *Hypothesis test*
 -

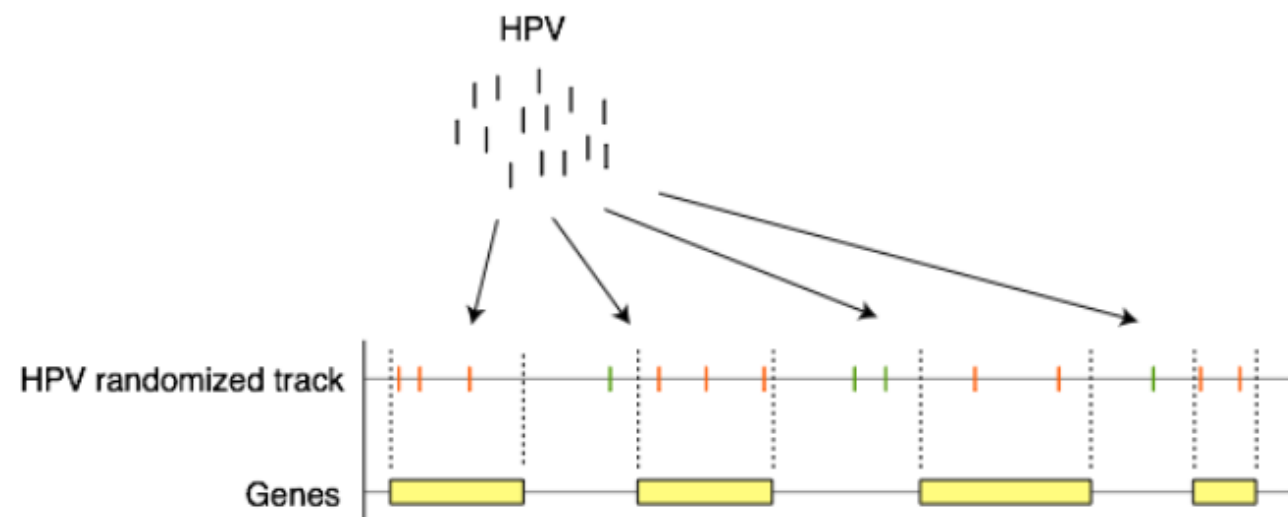


Analysis

- Co-occurrence is measured by a test statistic
 - E.g. the number of base pairs overlapping between two tracks
- We “compare” the computed test statistic to what we get when there is no association
 - Either analytically or by doing Monte Carlo Simulation
 - This requires a null model

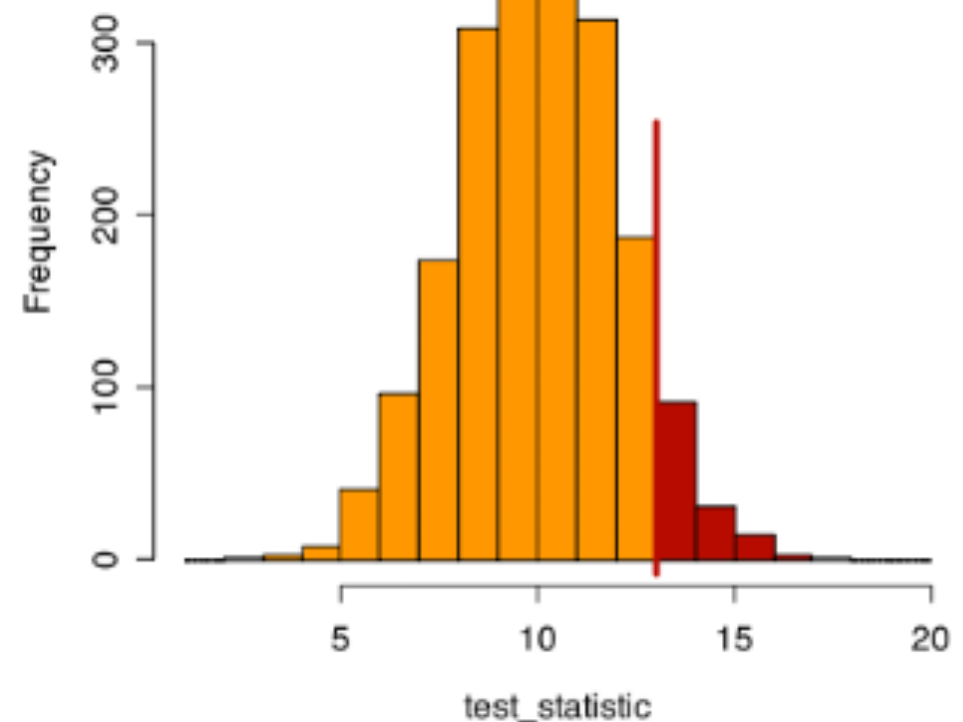
Analysis

- An example of a null model:
 - We assume that SNPs are distributed uniformly across the genome when there is no association
- Preservation strategies makes the null model more realistic:
 - We can for instance preserve the inter-point/segment distances.



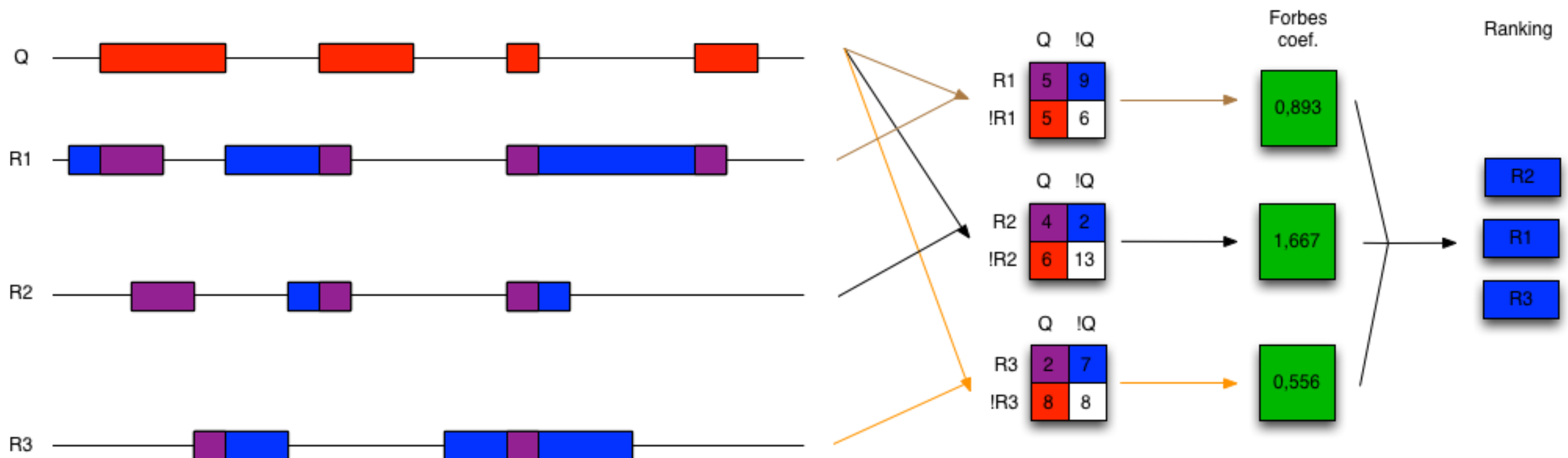
Finding the p-value

- Can be found either *analytically* or by doing *Monte Carlo simulation*
 - *Analytically*: We assume a distribution of the test statistic
 - *Monte Carlo*: We simulate the distribution by computing the test statistic for random samples. We compare our observed test statistic with those simulated.



Analysis of track collections

- Typical question:
 - Which reference tracks is most similar to a query track
 - We rank the reference tracks by similarity
- Different similarity measures will give different results:



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)

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