Statistical genomics

INF-BIO5121/INF-BIO9121
October 22. 2015, Oslo

Boris Simovski and Sveinung Gundersen
BMI/Genomic HyperBrowser team
Department of Informatics, UiO

Overview of session

09:00-10:15 Introduction. Tracks and track types

10:30-12:00 Hypothesis testing

12:00-13:00 Lunch

13:00-13:50 Hypothesis testing (cont.)

14:00-14:50 Statistical details

15:00-15:50 Data upload. Implementation. Reproducibility

16:00-17:00 Course wrap-up and exam handout

Introduction

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

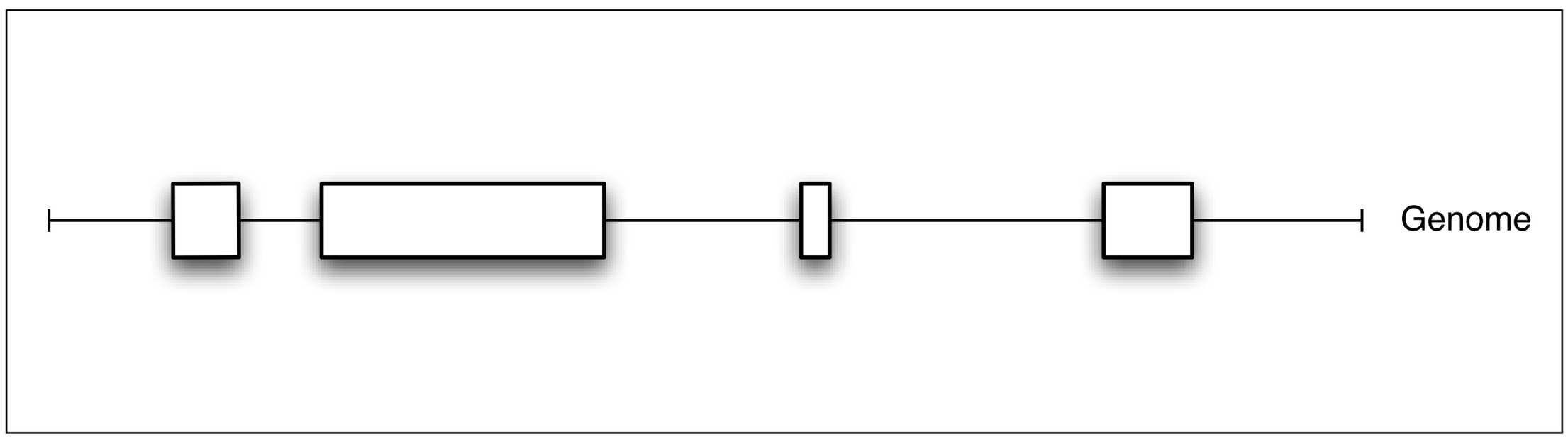
About the Genomic HyperBrowser

- 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:

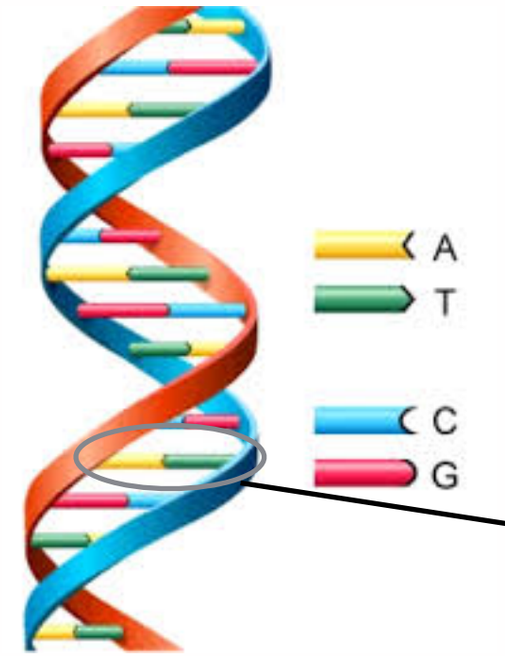
  The course is about statistical genomics. The concepts are the same if you use other tools!

What are genes?

This! :



Genome as a line



ATCTGTGACCTGA

base-pair

1 2 3 4 5 6 7 8 9 10 11 12 13

How to represent genes on the line coordinate?

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

- Poorly defined term
- 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 (epi)genomics

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

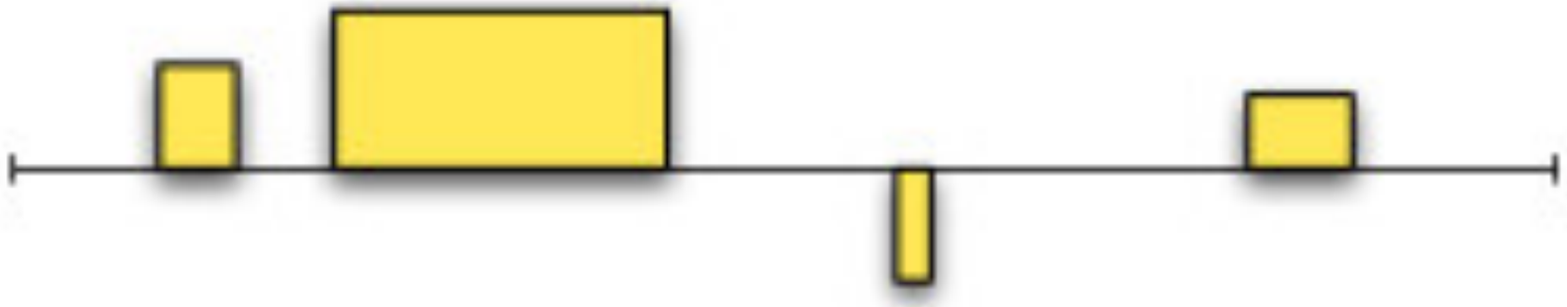
- However:
  - Many of the concepts are central statistical concepts that can be used for other types of analyses

Tracks and track types

Representation of genes

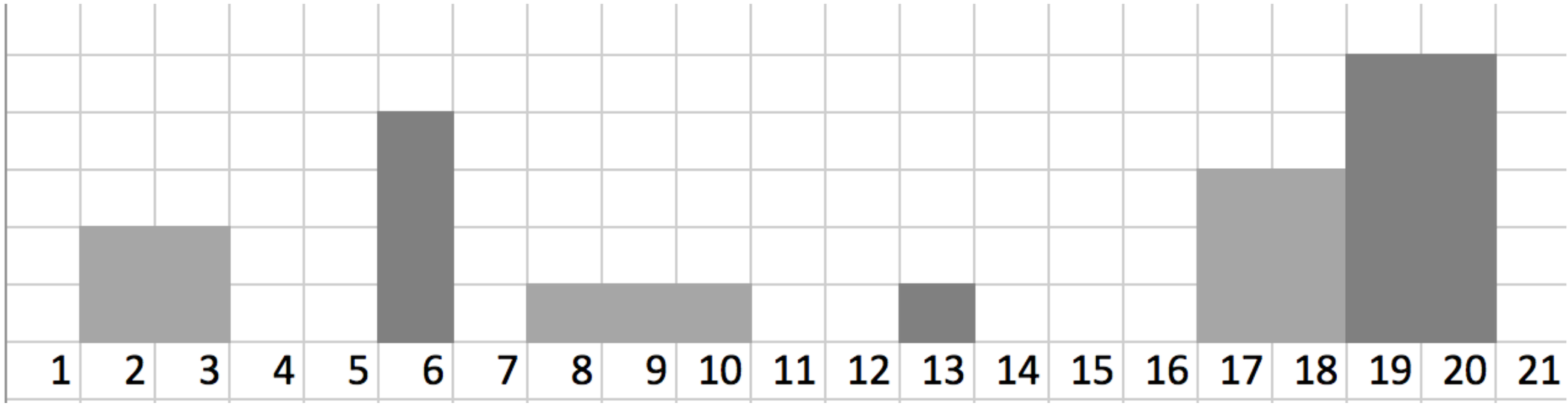
chr7	127471196	127472363
chr7	127472388	127473530
chr7	127473555	127474697
chr7	127474701	127475864
chr7	127475893	127477031
chr7	127477121	127478198
chr7	127478300	127479365
chr7	127479375	127480532
chr7	127480538	127481699

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

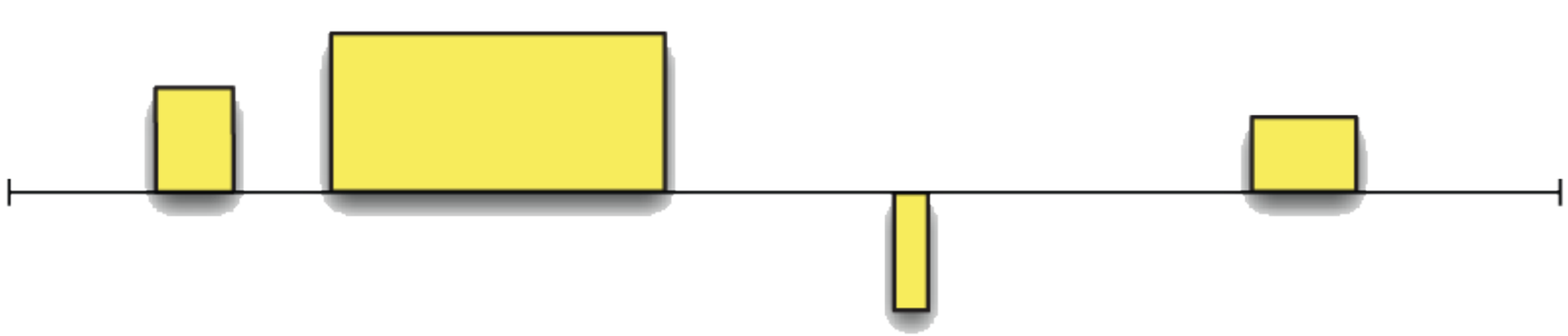
Exercise 1



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

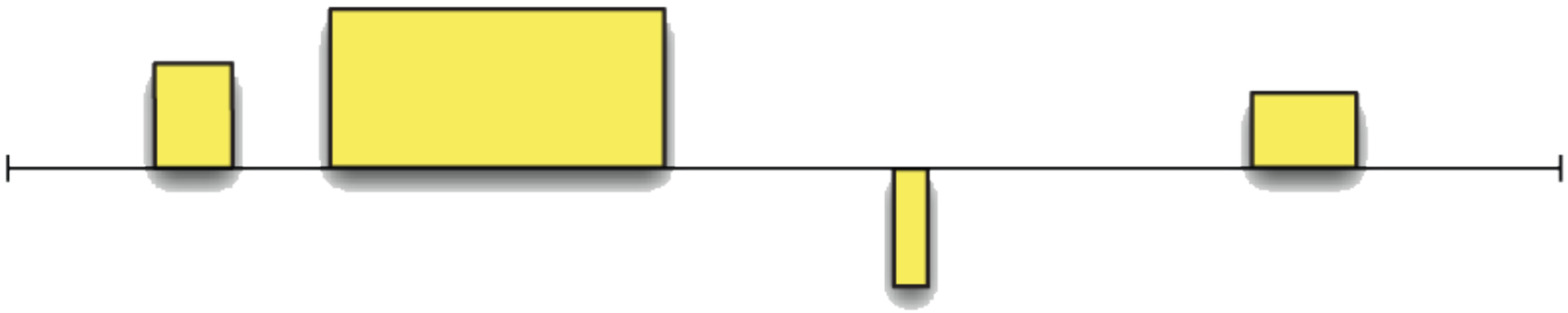
- In the last example, we showed genes as segments on the genome line, with attached RNA-seq read count values
- This track is of a **track type** we call "valued segments"



Valued Segments (VS)

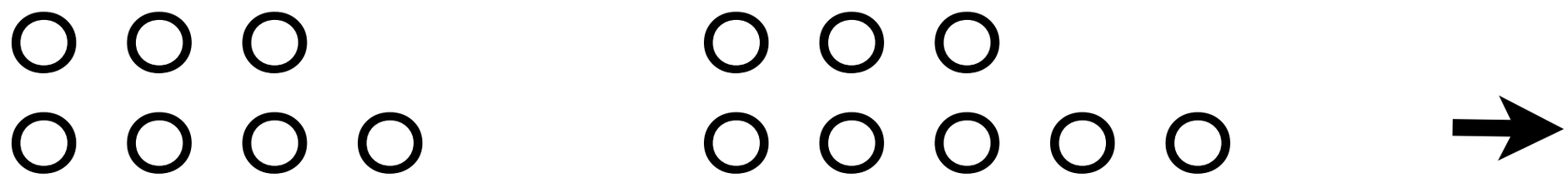
- Track types are mathematical / conceptual models used to categorize track according to their main characteristics

Exercise 2

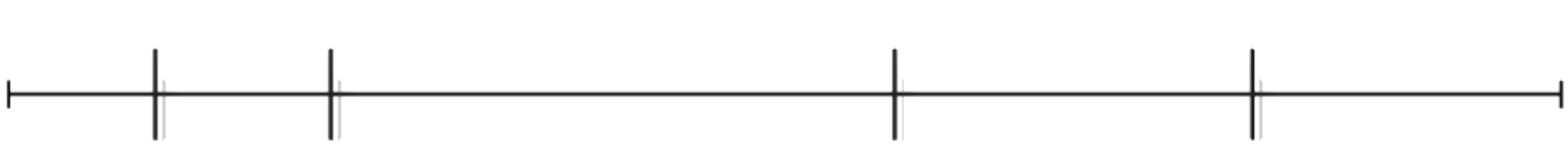


Valued Segments (VS)

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

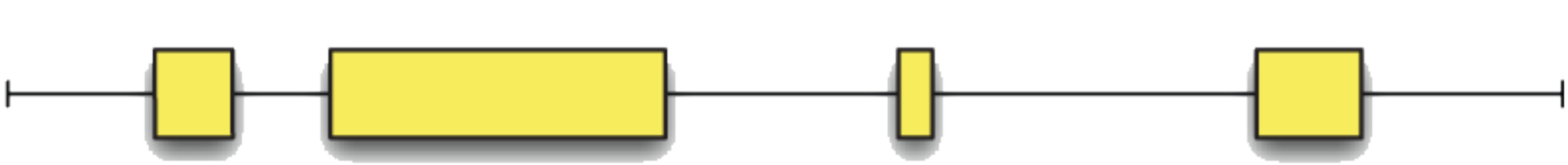


Points



Points (P)

Segments



Segments (S)

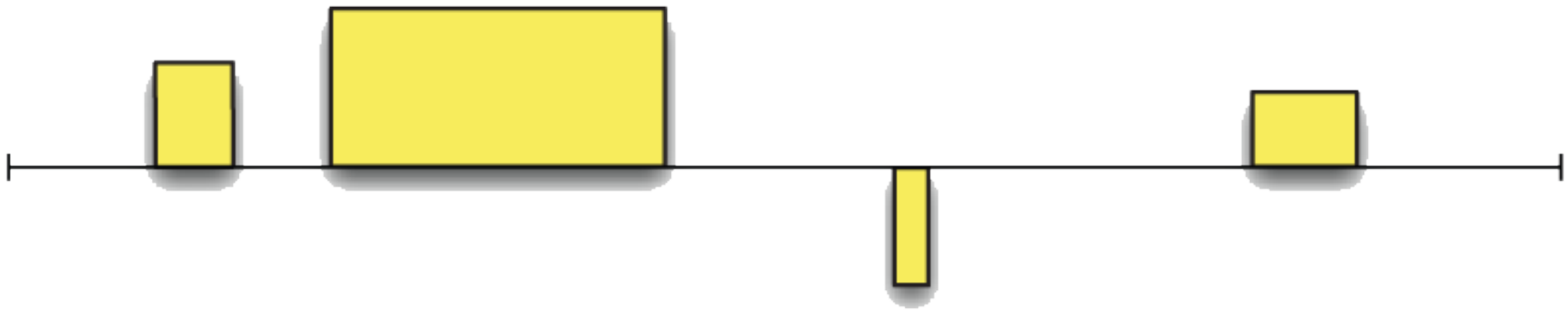
Genome Partition

Genome Partition (GP)

Valued Points

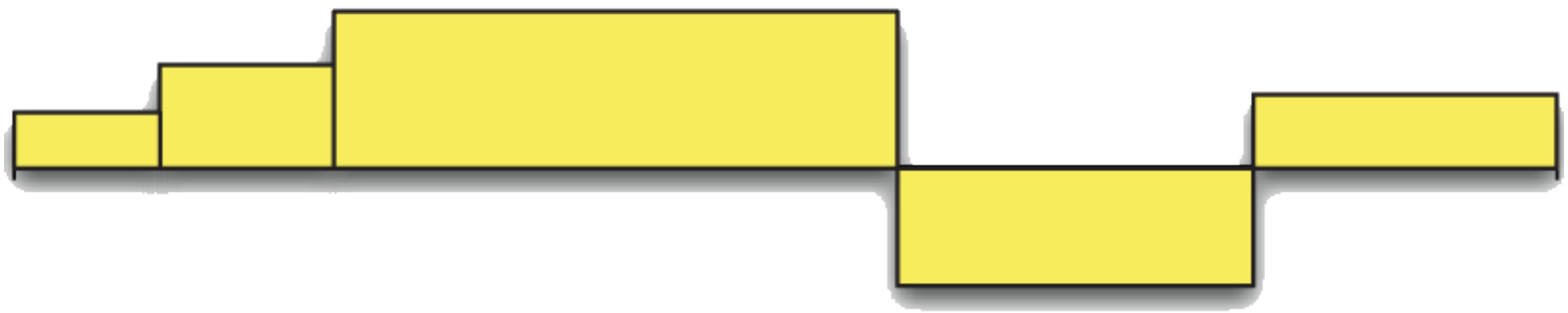
Valued Points (VP)

Valued Segments



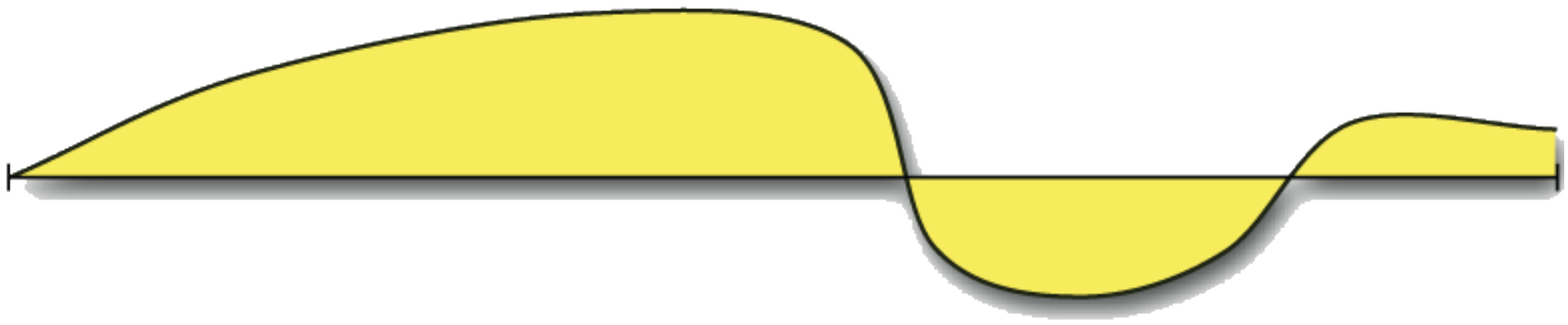
Valued Segments (VS)

Step Function



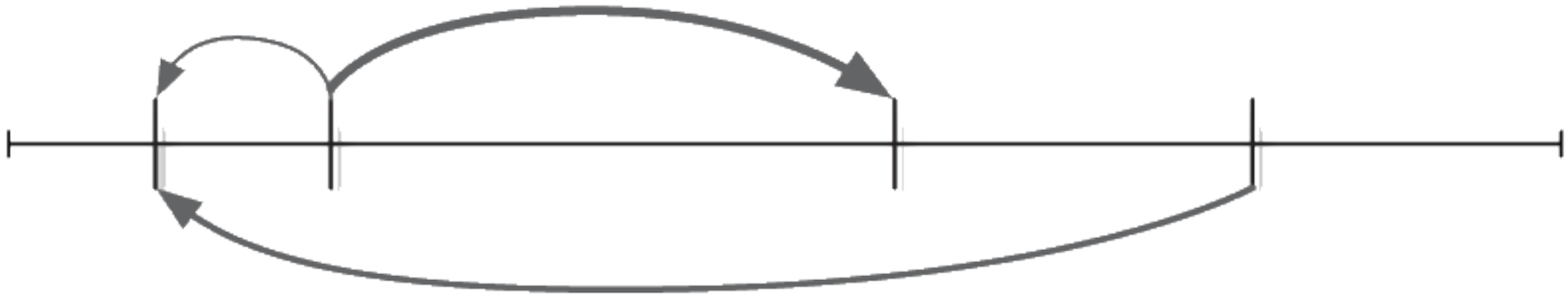
Step Function (SF)

Function



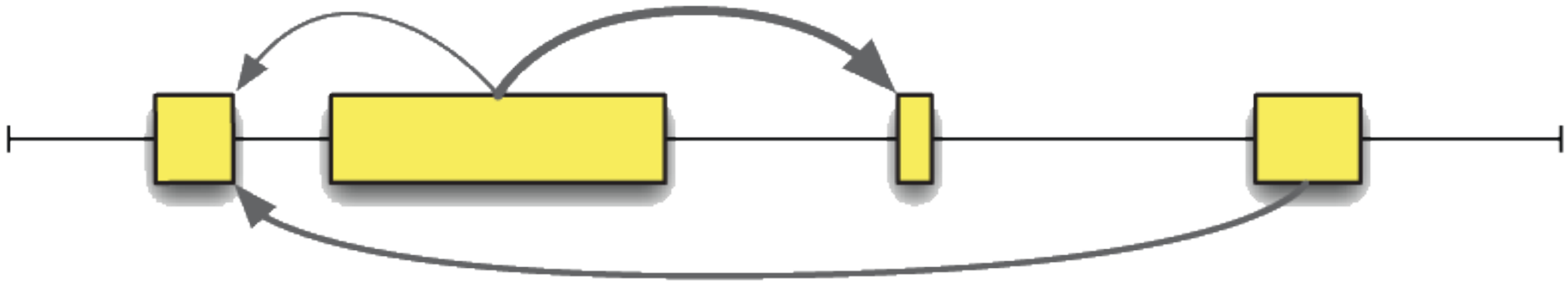
Function (F)

Linked Points



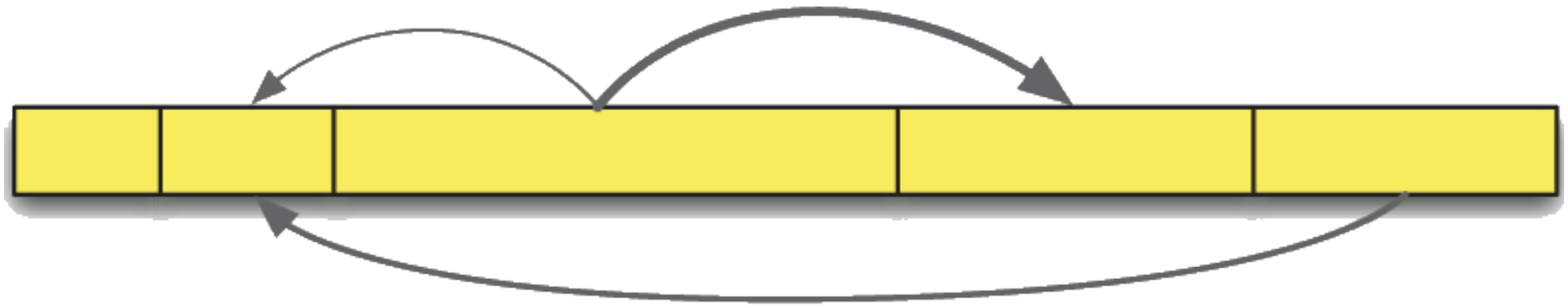
Linked Points (LP)

Linked Segments



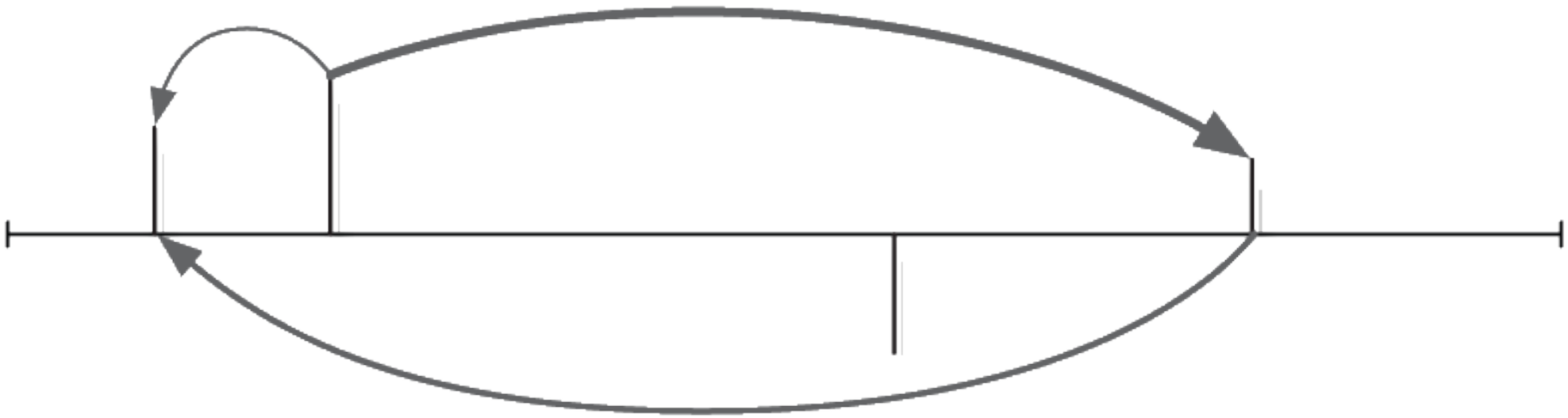
Linked Segments (LS)

Linked Genome Partition



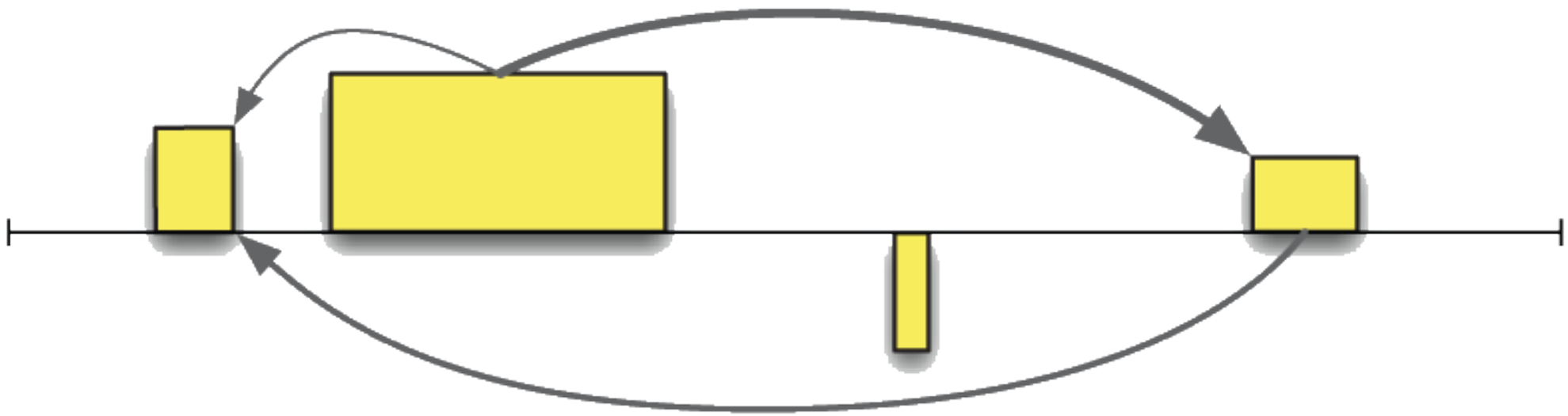
Linked Genome Partition (LGP)

Linked Valued Points



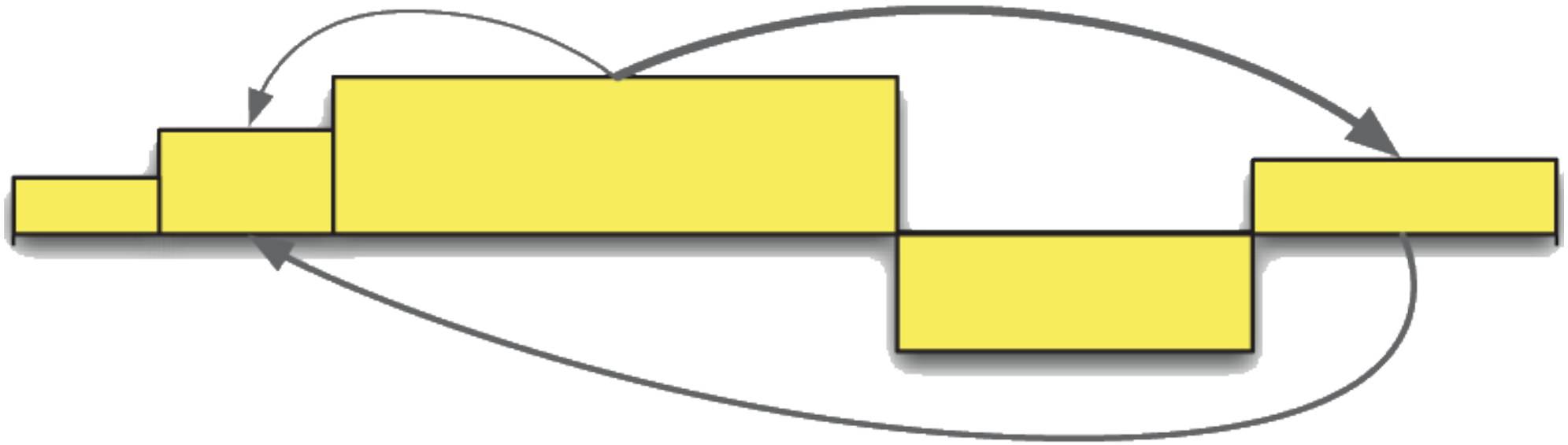
Linked Valued Points (LVP)

Linked Valued Segments



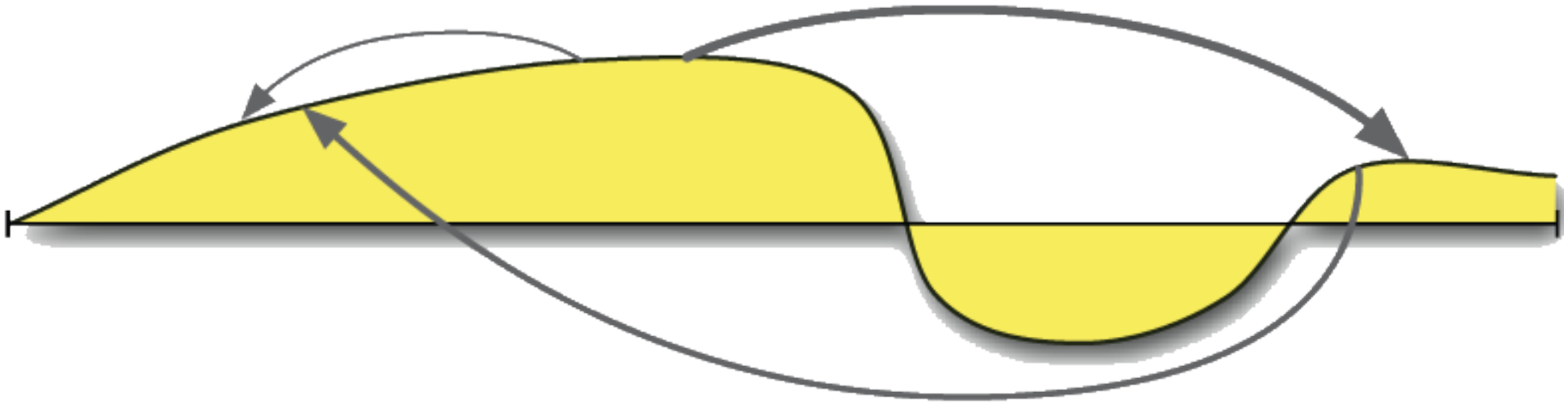
Linked Valued Segments (LVS)

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?

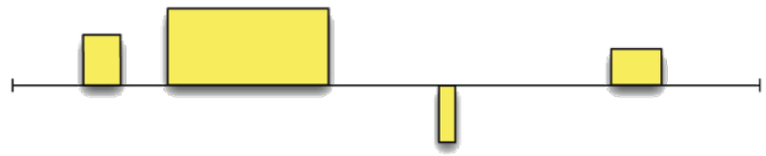
- For each track, which **track type** should be used to represent the data?

Exercise 3

Points (P)

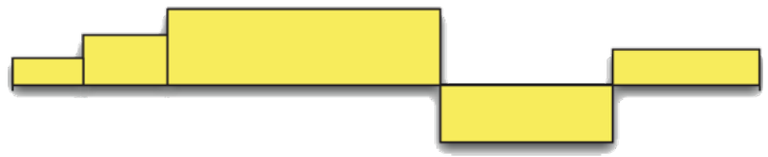
Valued Points (VP)

Segments (S)

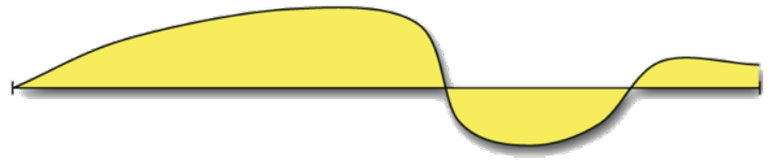


Valued Segments (VS)

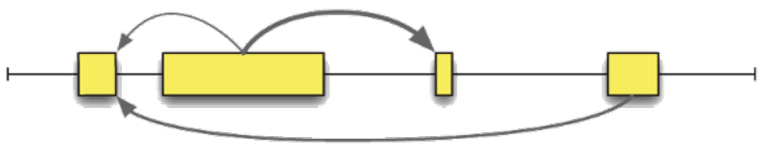
Genome Partition (GP)



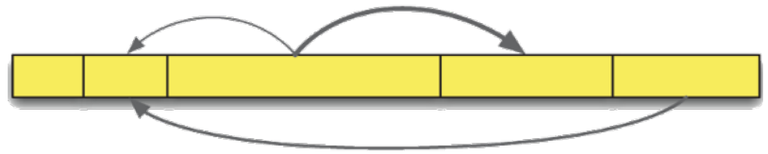
Step Function (SF)



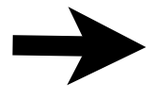
Function (F)



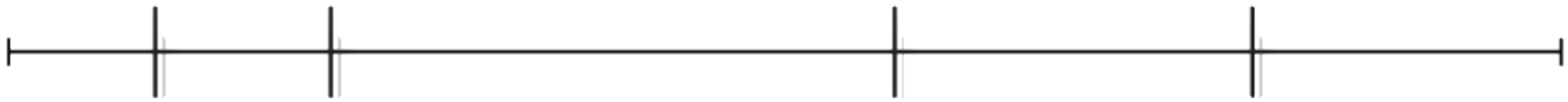
Linked Segments (LS)



Linked Genome Partition (LGP)



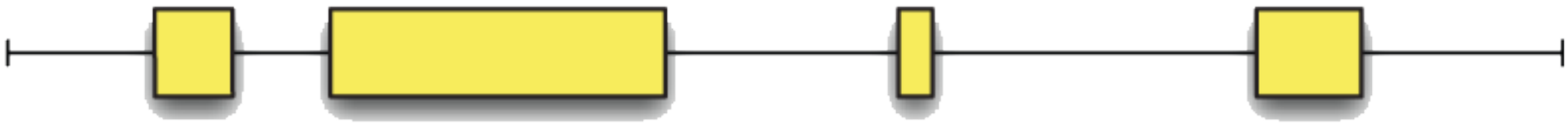
Points



Example tracks:

- SNPs, GWAS
-

Segments



Example tracks:

- Transcripts
- DNA methylation
- ChIP-seq peaks

↵

Genome Partition

Example tracks:

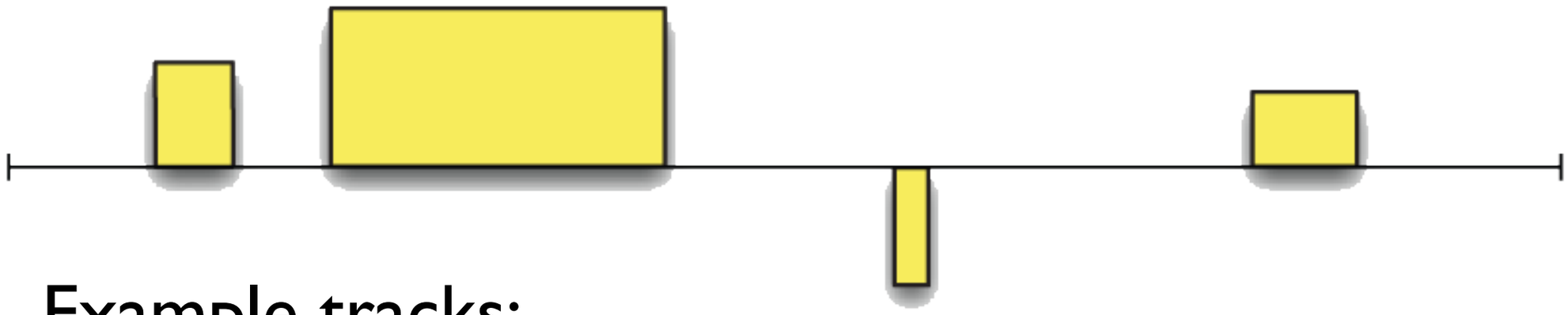
- Chromosomes
- Chromatin state segmentation
-

Valued Points

Example tracks:

- SNPs with allele frequency
-
-
-

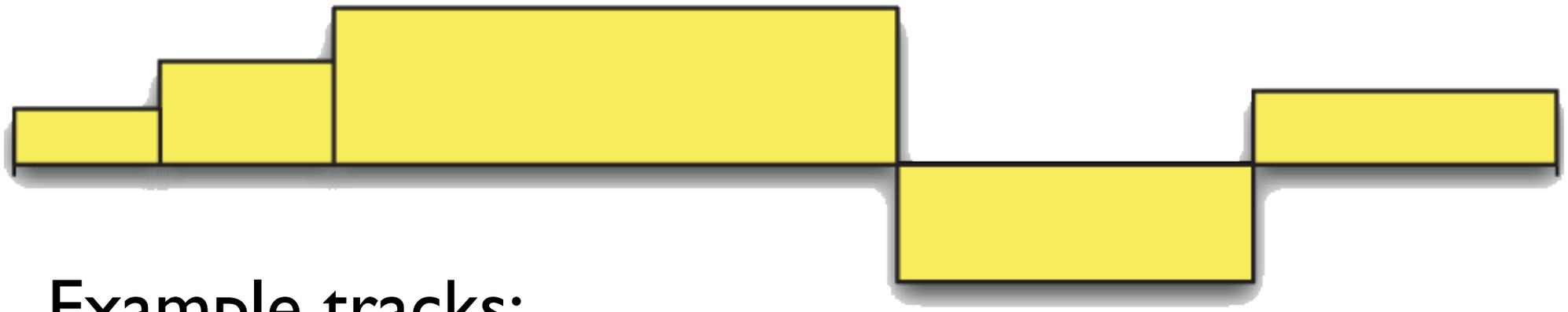
Valued Segments



Example tracks:

- Genes with expression values
-
-
-

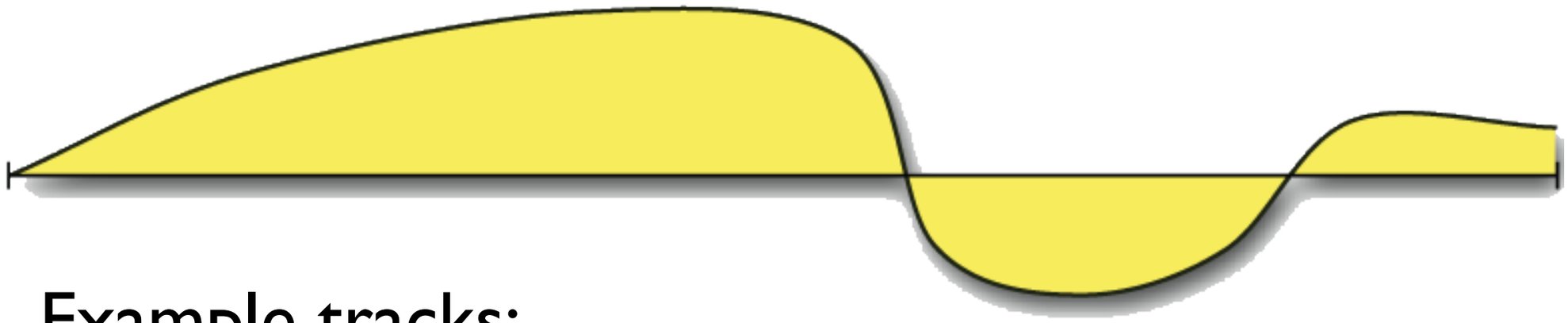
Step Function



Example tracks:

- RNA-seq signal track (equal read count in a box)
- Copy number variation
-
-

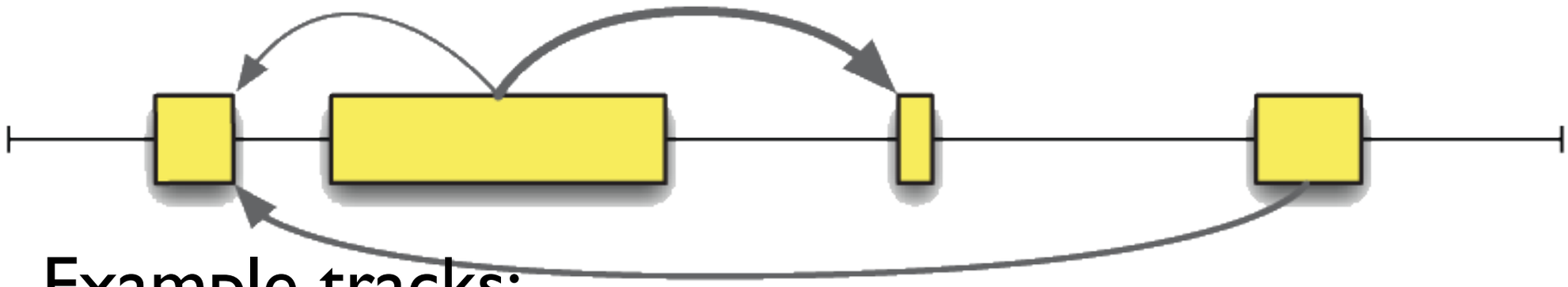
Function



Example tracks:

- RNA-seq signal track (one value per base pair)
- Quality scores
- ChiP-seq signal track
-

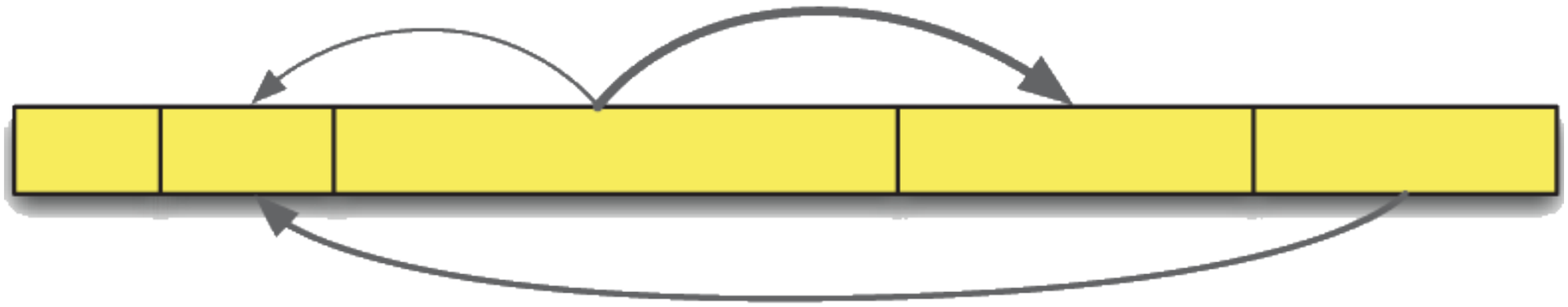
Linked Segments



Example tracks:

- Exons linked as transcripts
- Gene networks
- ChIA-PET
- Transposons

Linked Genome Partition



Example tracks:

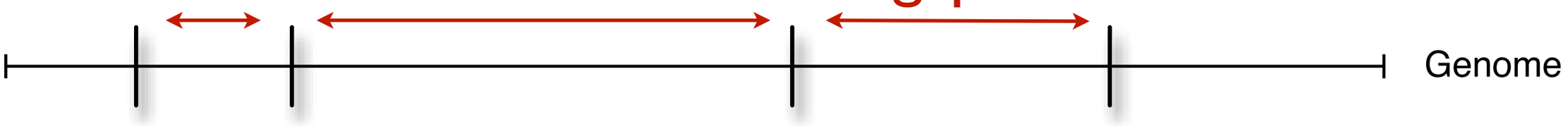
- Hi-C
-
-

↵

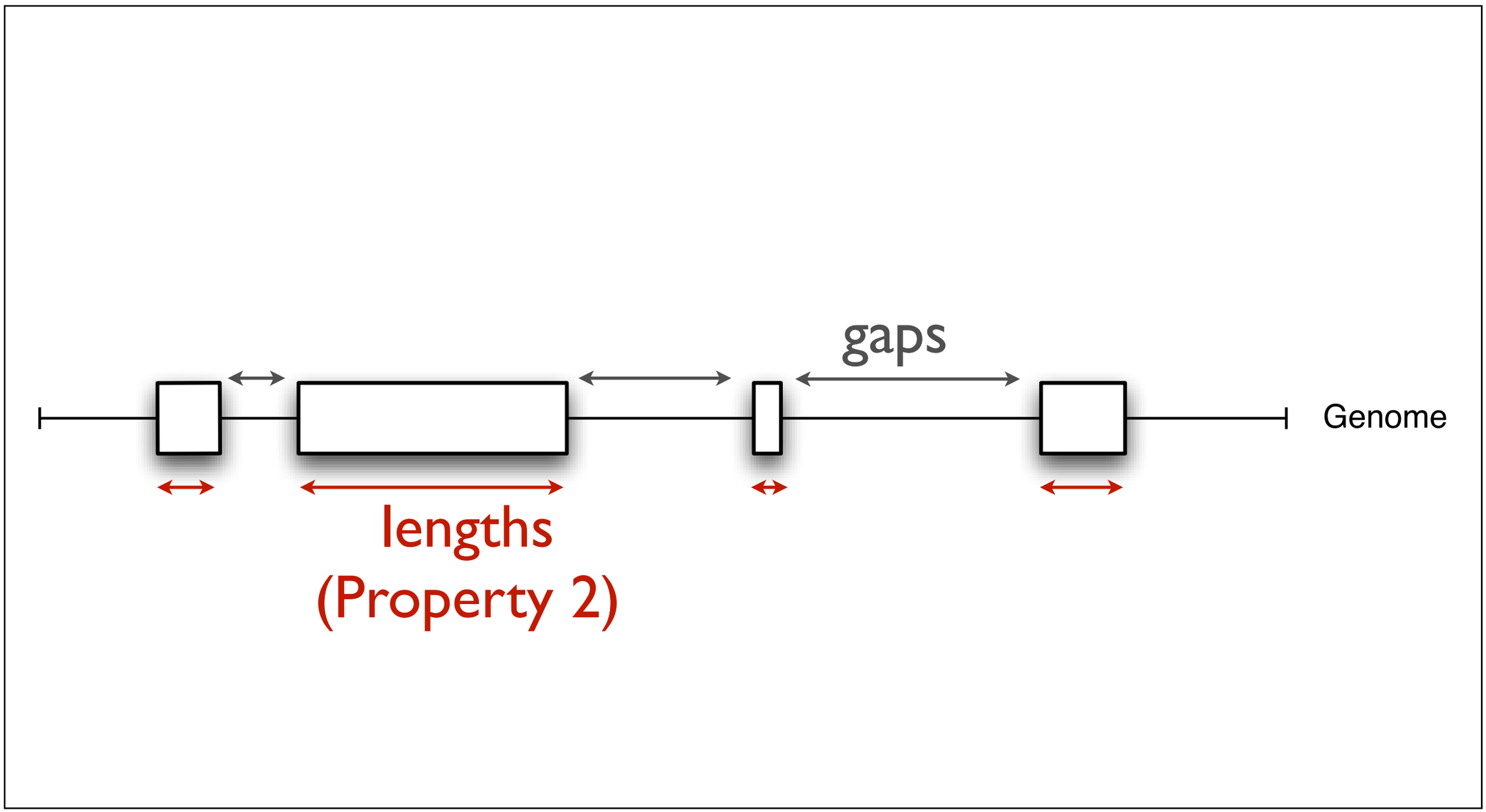
Core properties of tracks

Property 1:

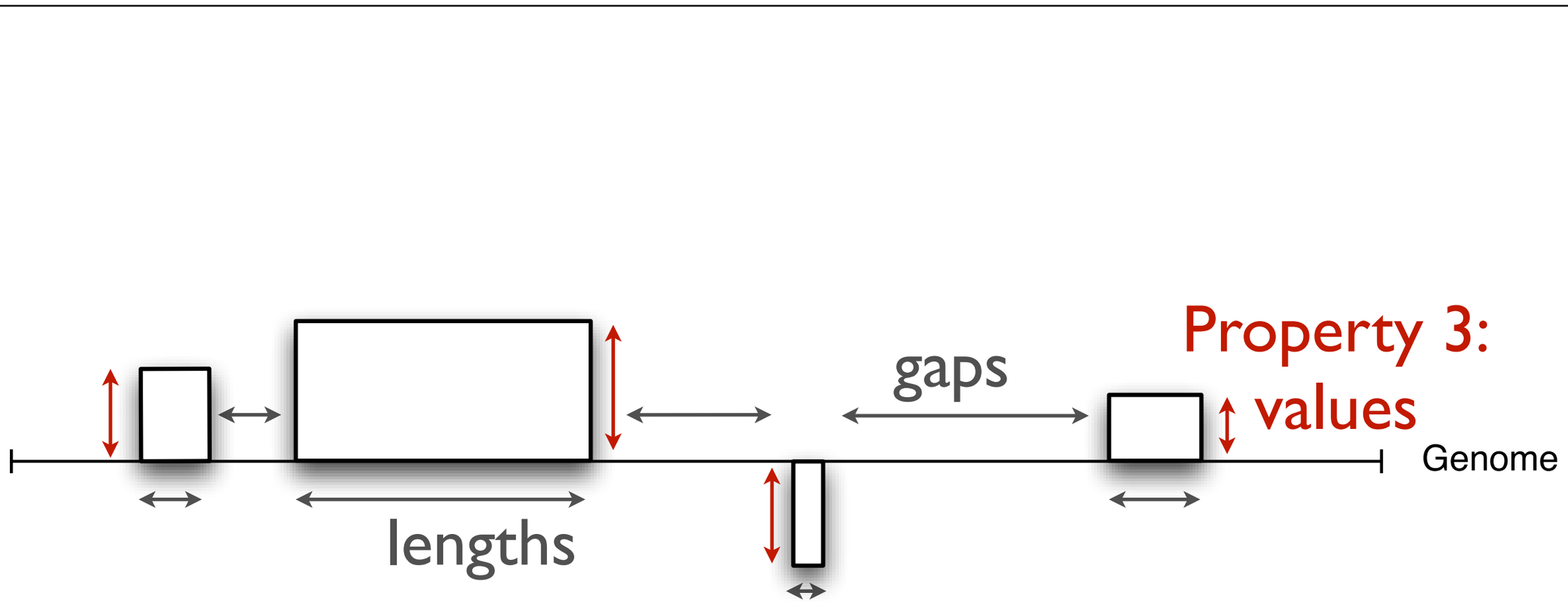
gaps



Core properties of tracks

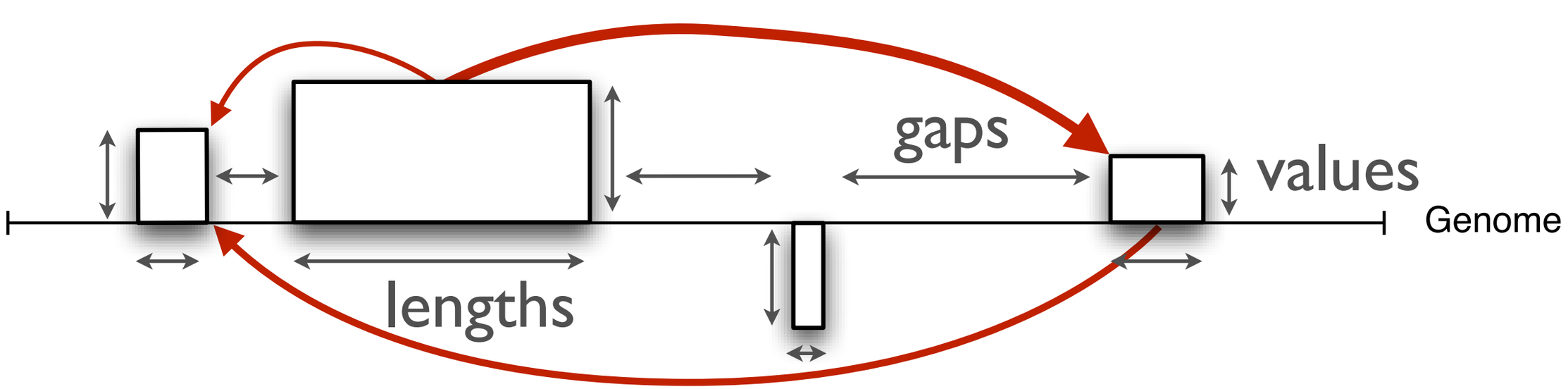


Core properties of tracks



Core properties of tracks

Property 4: interconnections



Exercise 4:
tracks in the real world

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

- Try out:
  - Zoom to small region and whole chromosome
  - Add/remove some tracks
  - Change the appearance of some tracks

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)

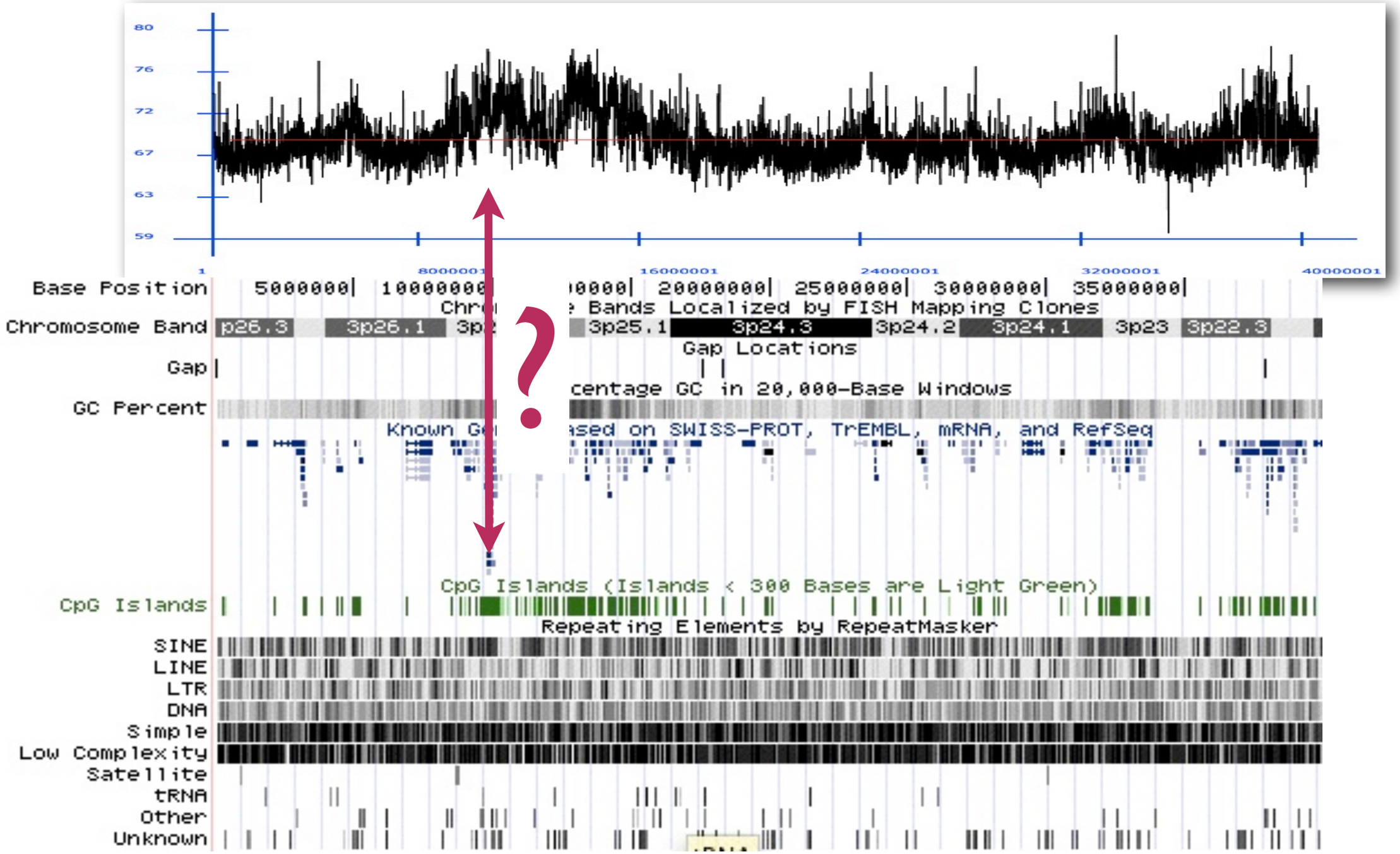
- Copy number variation, repeats, duplications and genes? (Genome Res. 2009 19: 1682-1690)

- 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

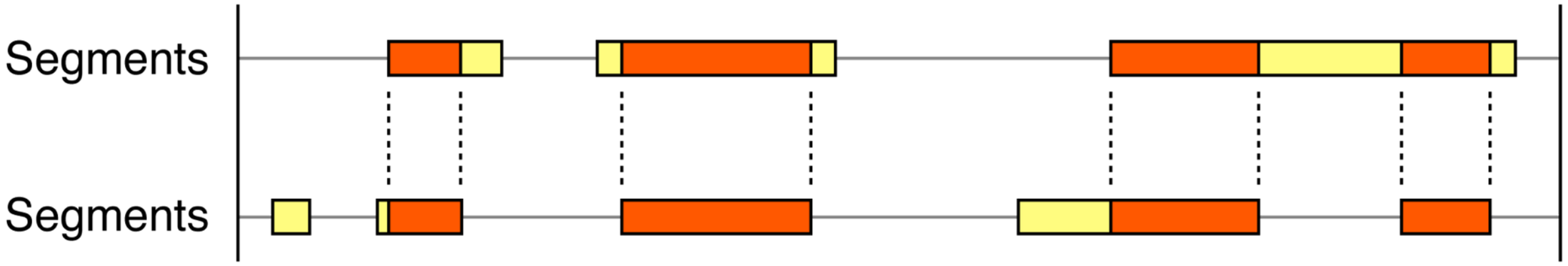
- Typical question:

> *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?

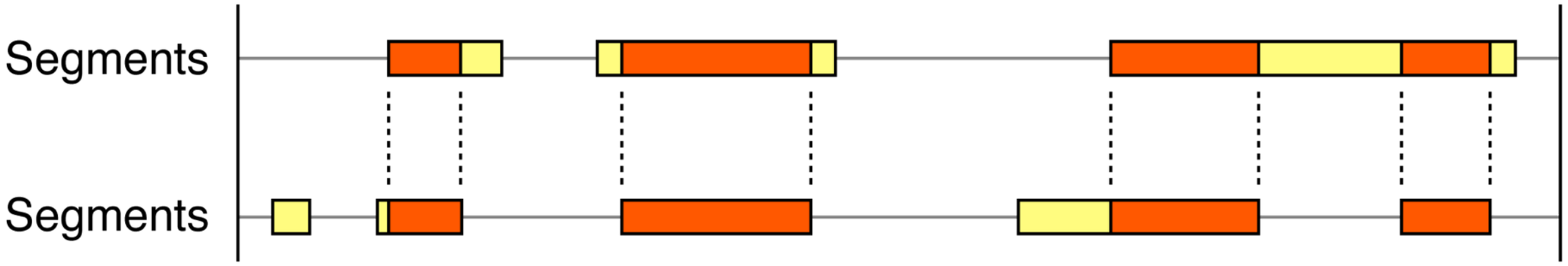
How does this look at the whiteboard?



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

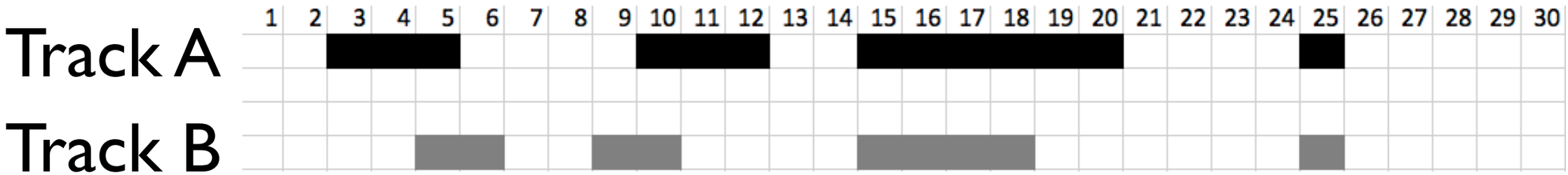
How does this look at the whiteboard?



overlap > expected?

What now?

Exercise 5

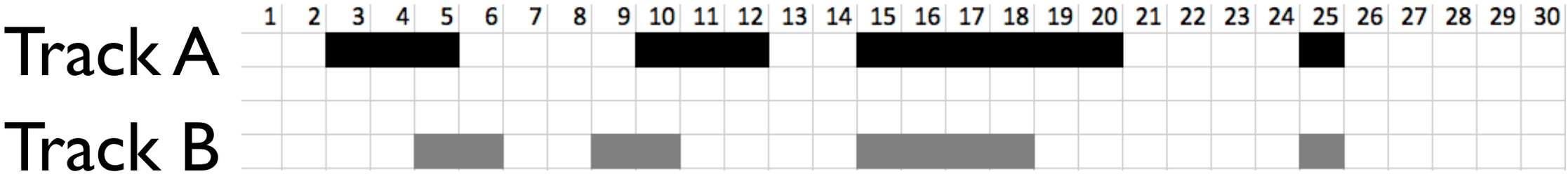


Calculate:

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

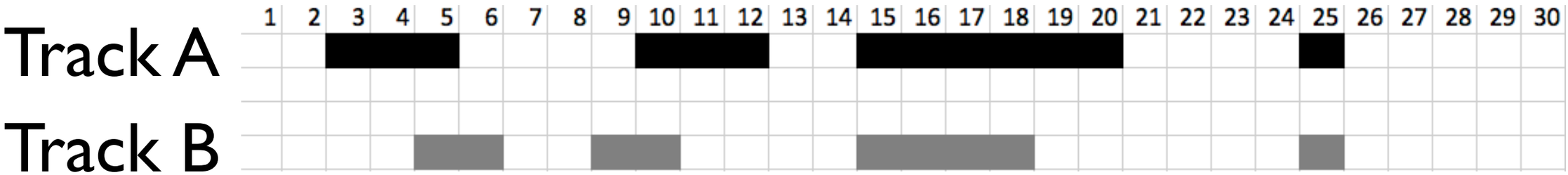
Hypothesis testing

- Alternative hypothesis (H1)
  - What you really want to show (e.g. relation between tracks A and B)
- Null hypothesis (H0)
  - A neutral baseline (e.g. no relation exists between tracks A and B)
- Test statistic
  - A measure of the aspect of interest (e.g. base pair overlap)
- P-value
  - How likely is the observation (or more extreme), given H0
  - Observation is unlikely (e.g. p-value less than 0.05) -> reject H0, data supports H1 (=significant results)
  - The p-value "cut-off" (0.05) is called significance level (usually denoted by α)
  - Not rejecting the null hypothesis does not mean that it is true. It could for instance be that you lack data
- Two-tailed vs. right-tailed vs. left-tailed

Null models (1/2)

- A model from which the null hypothesis arises
- Mathematical computation of the null model is usually out of reach
- Simulation by Monte Carlo is often the solution (you already did this)
- How to randomize the data?
  - Preservation of 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

Exercise 6b



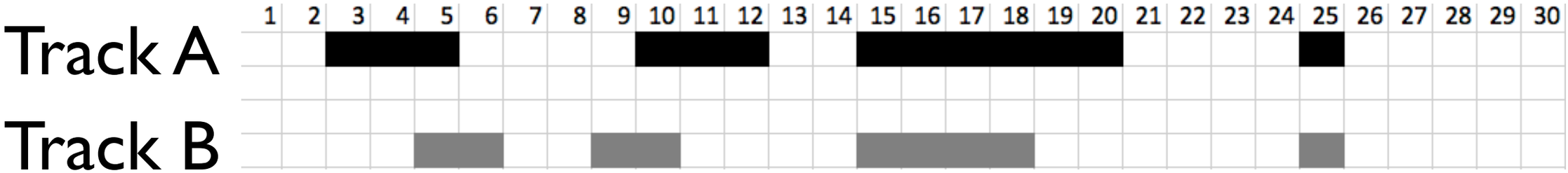
Create a random control track for track B, by

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 (2/2)

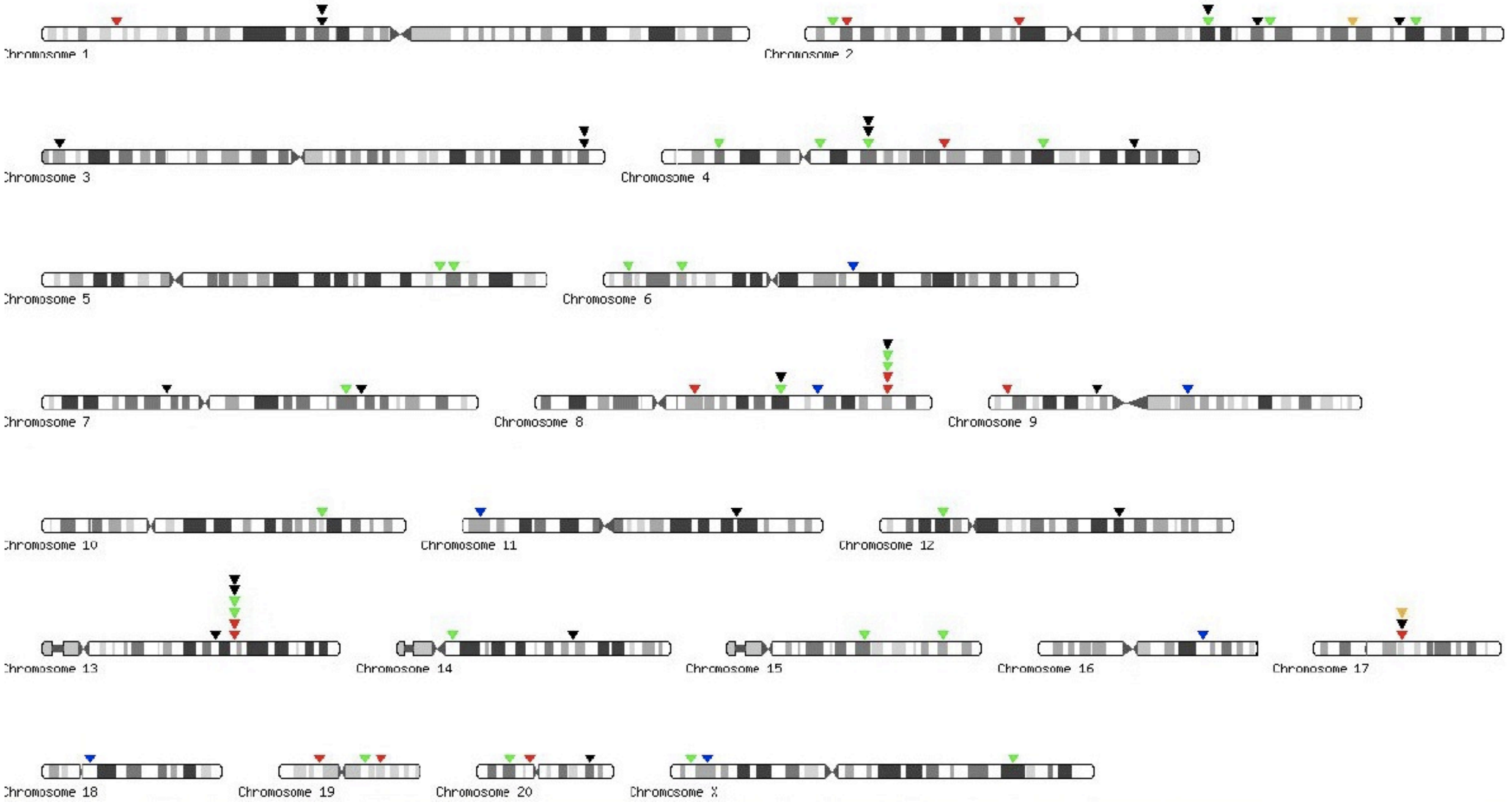
- Examples of preservation strategies
  - Preserve segment length (already seen this)
  - Preserve segment and gap length (this too)
- For points (segments with length 1)
  - Preserve point count
  - Preserve inter-point distance
- For all this cases we randomize the position of the track elements.

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

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

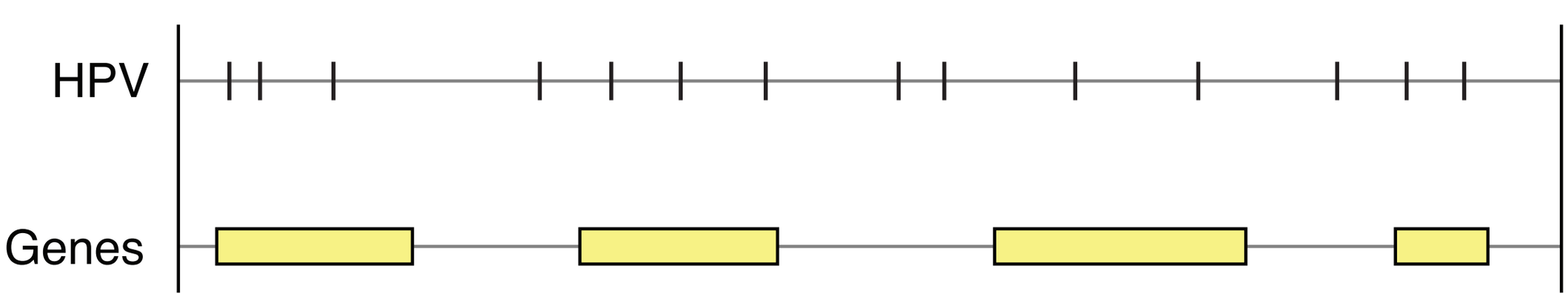
"Viruses might be expected to integrate near genes. Our results confirm such preferential localization inside genes for HPV [Human Papillomavirus], where 75 out of 119 determined integration sites (attached) fall inside genes."

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?

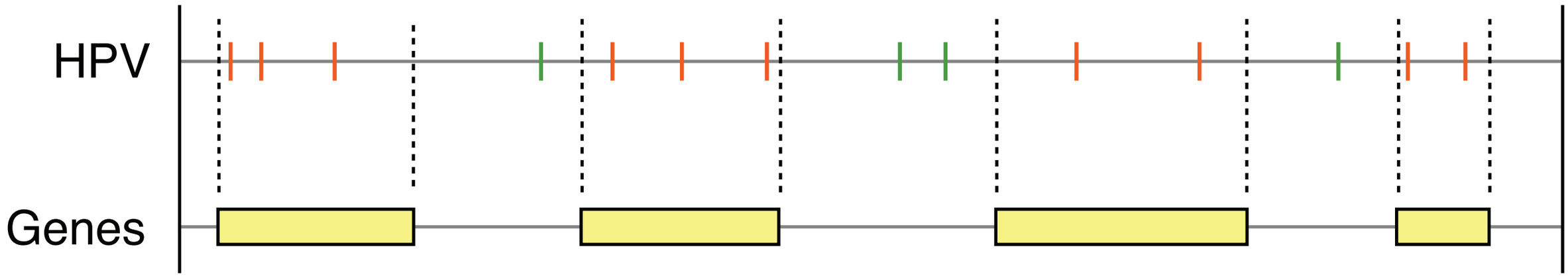
Exercise 7: HPV and genes

Student answers:

1. Which test statistic would you choose?

Number of HPV that fall inside genes	3	2
Observed / expected of number inside	15	17
Proportion of HPV points falling inside genes	4	4
Average proximity to gene	1	2
I do not know	1	3

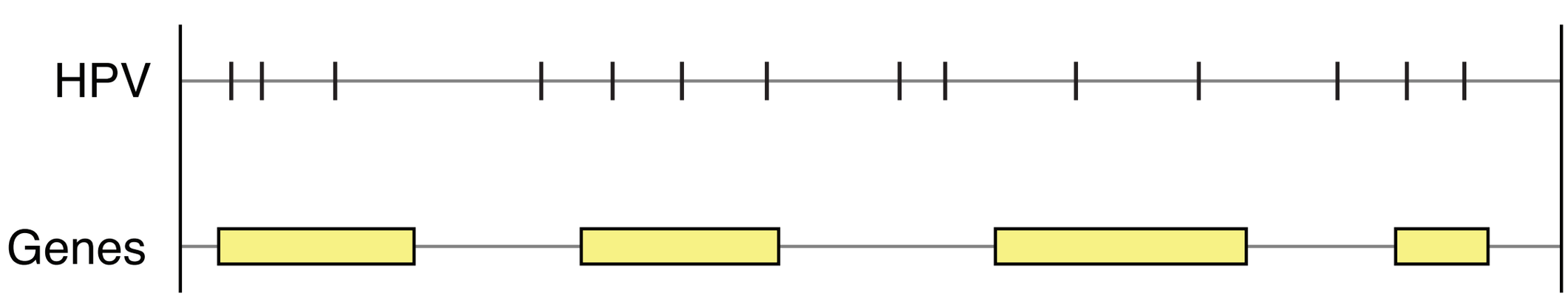
A possible test statistic



- 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

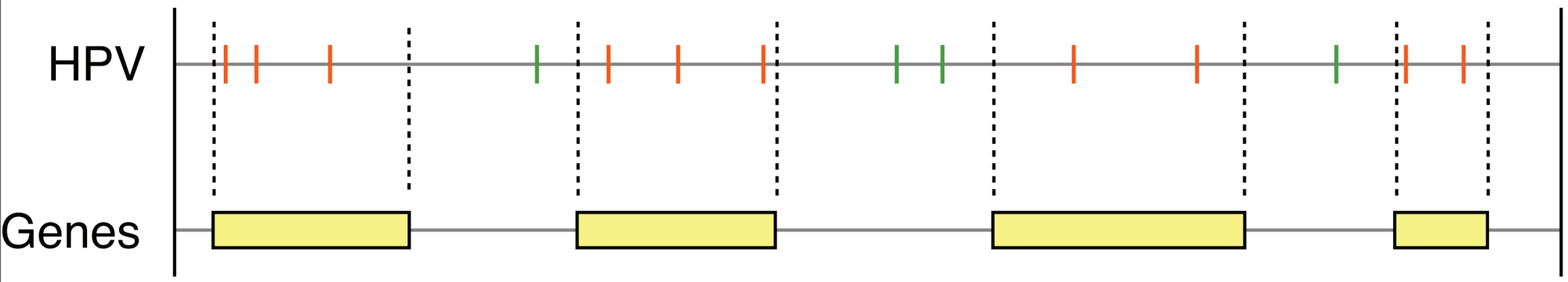
Exercise 8: HPV and genes

Student answers:

2. Which null model would you choose?

Randomize HPV positions, preserve number of points and length of gaps	1	1
Randomize HPV numbers and positions, preserve length of gaps	1	1
Randomize HPV positions, preserve number of points	22	20
Randomize gene positions, preserve segment and gap length	2	2
Randomize gene positions, preserve segment	0	1
Bootstrapping		
i don't know	3	1

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 HPV, gene lenght, randomize gene position	Refseq	0,667
Preserve genes, number of HPV, randomize position	Ensembl	0,013
Preserve HPV, gene and gap lenght , randomize gene position	Refseq	0,49
Preserve HPV, gene lenght, randomize gene position	Refseq	0,45

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

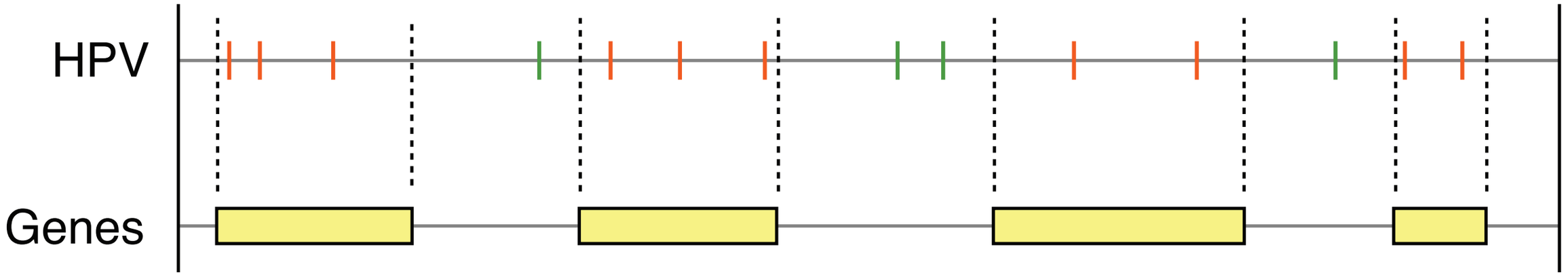
- Duan, [...] and Noble (Nature, 2011):
  - Extensive significant 3D co-localization of functional elements, assessed by hypergeometric distribution
- Witten and Noble (NAR, 2012):
  - Hypergeometric had unrealistic implications. Telomeres and breakpoints may not be co-located after all.. (cancelled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

Further into statistical details

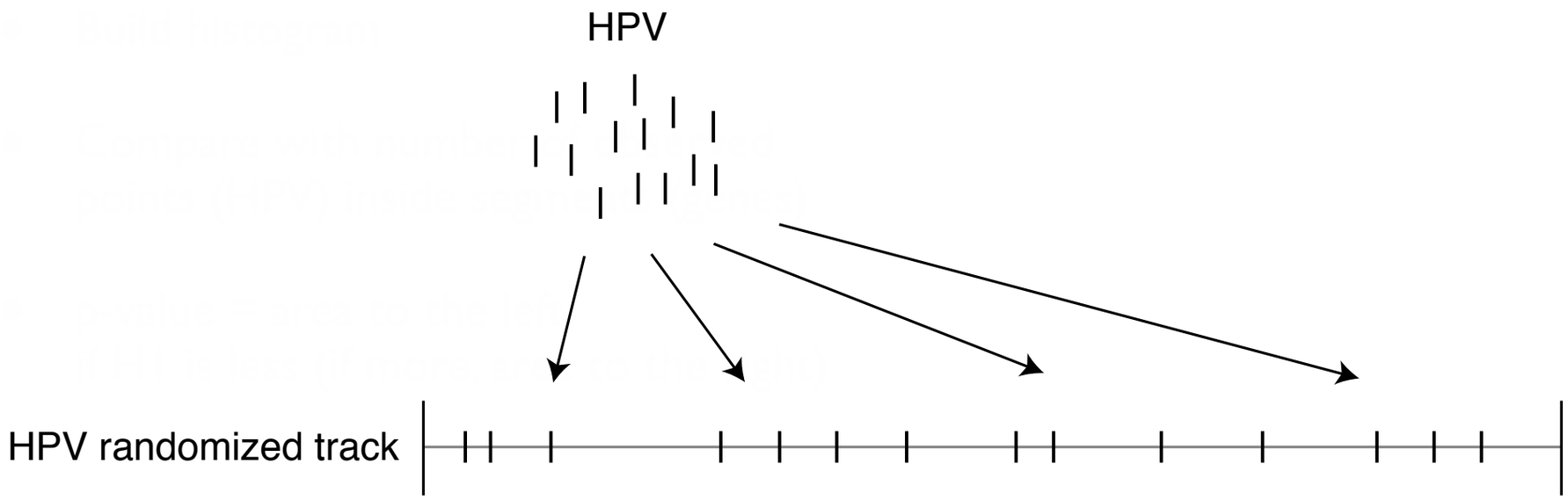
Further into statistical details: distributions



- 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"

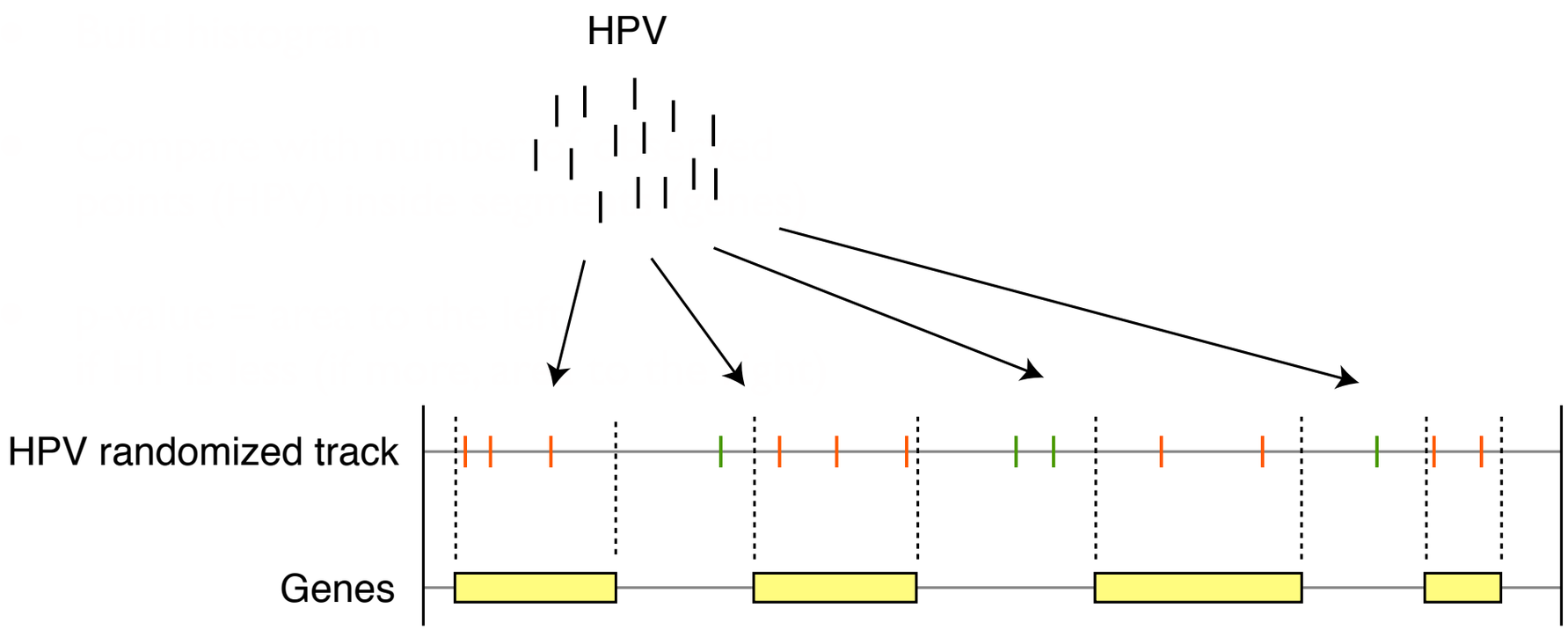
- Randomize point (HPV) locations (null model)



Monte Carlo test on "points inside segments"

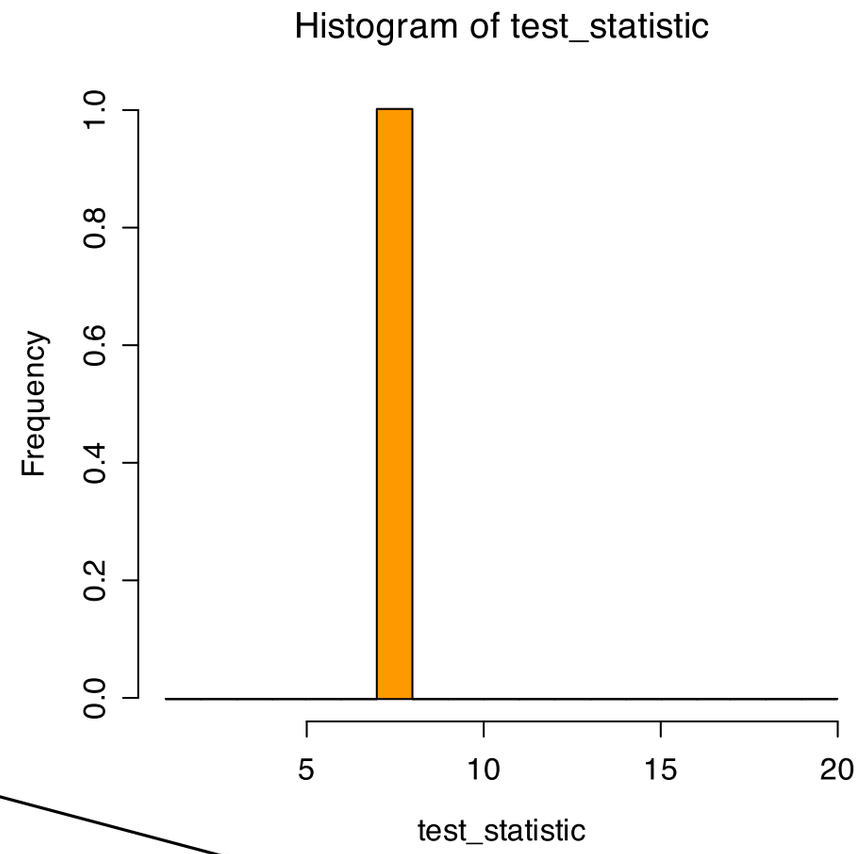
- Randomize point (HPV) locations (null model)

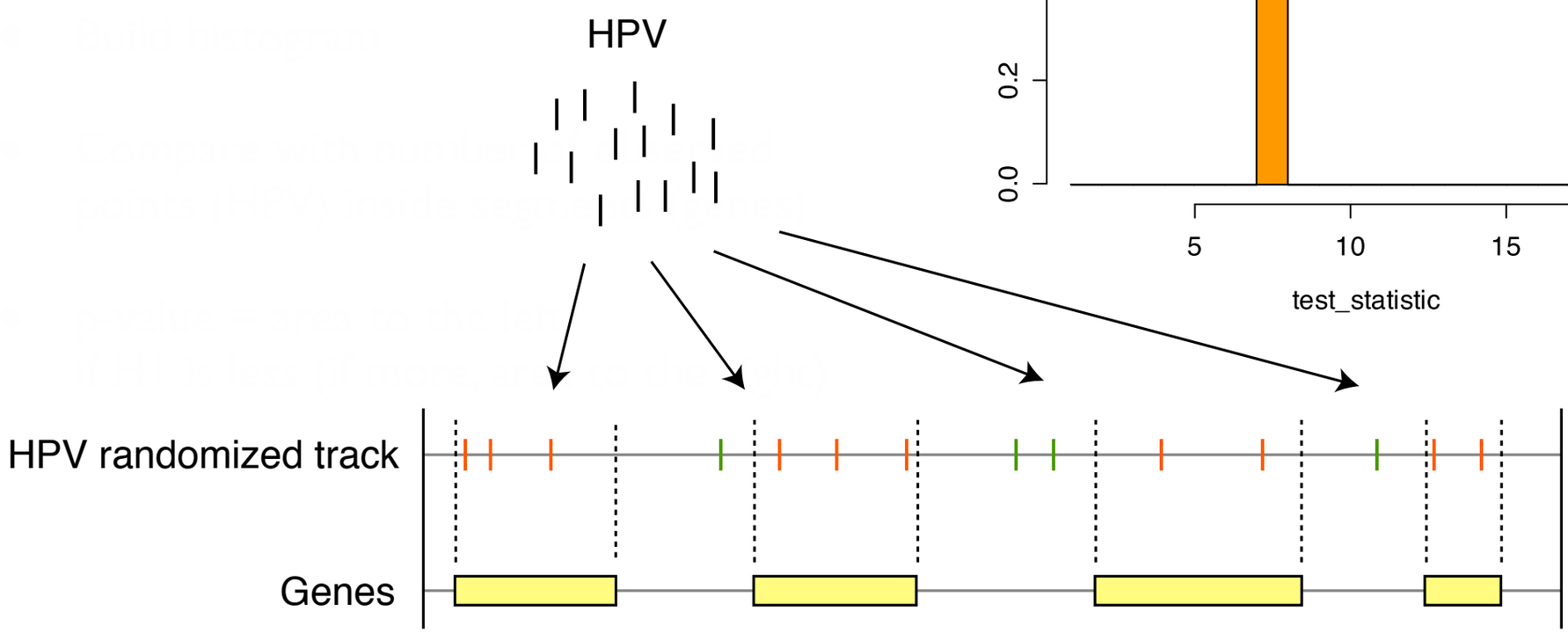
- Count random points (HPV) inside segments (genes) - test statistic



Monte Carlo test on "points inside segments"

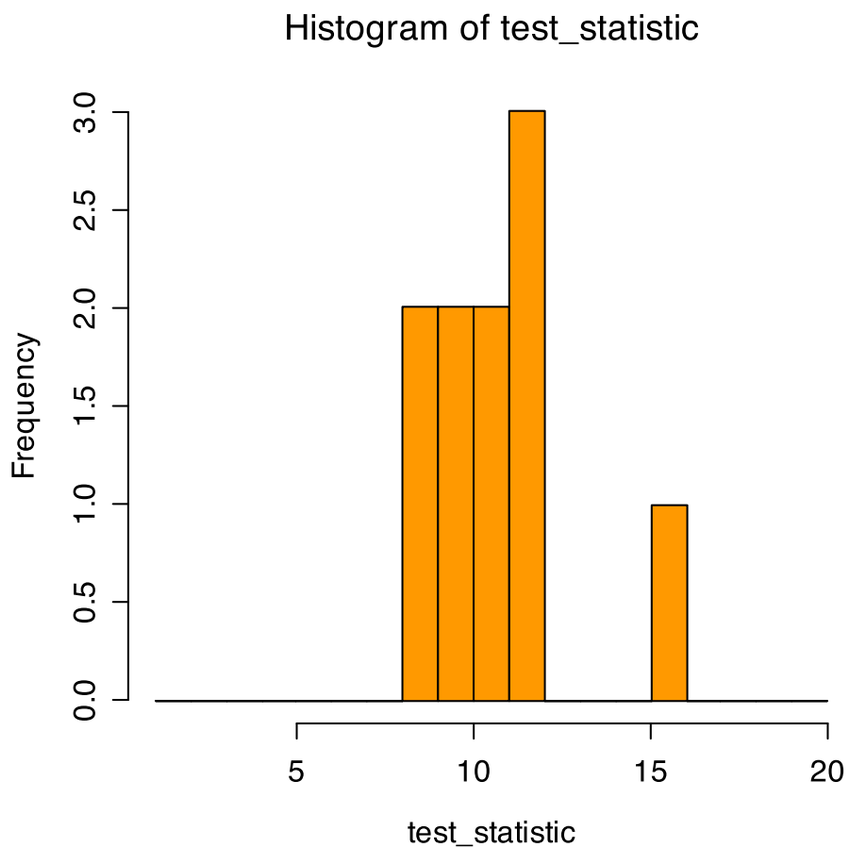
- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic





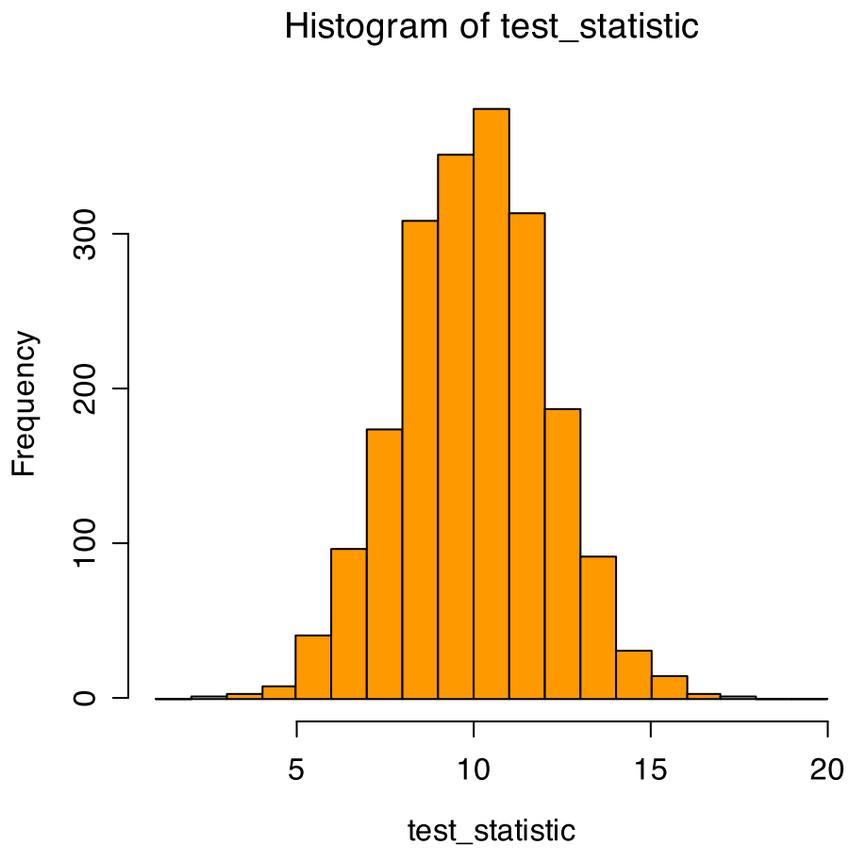
Monte Carlo test on "points inside segments"

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times
- Build Histogram
- Compare with number of observed points (HPV) in the segments (genes)
- p-value = area to the left (right) of more extreme test statistic



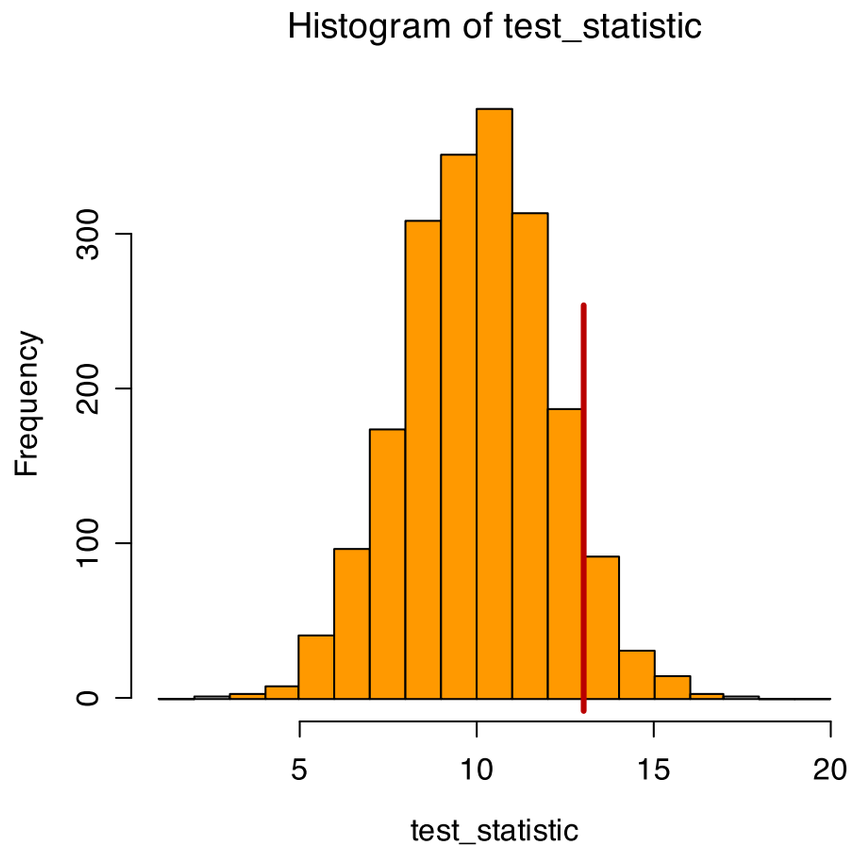
Monte Carlo test on "points inside segments"

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



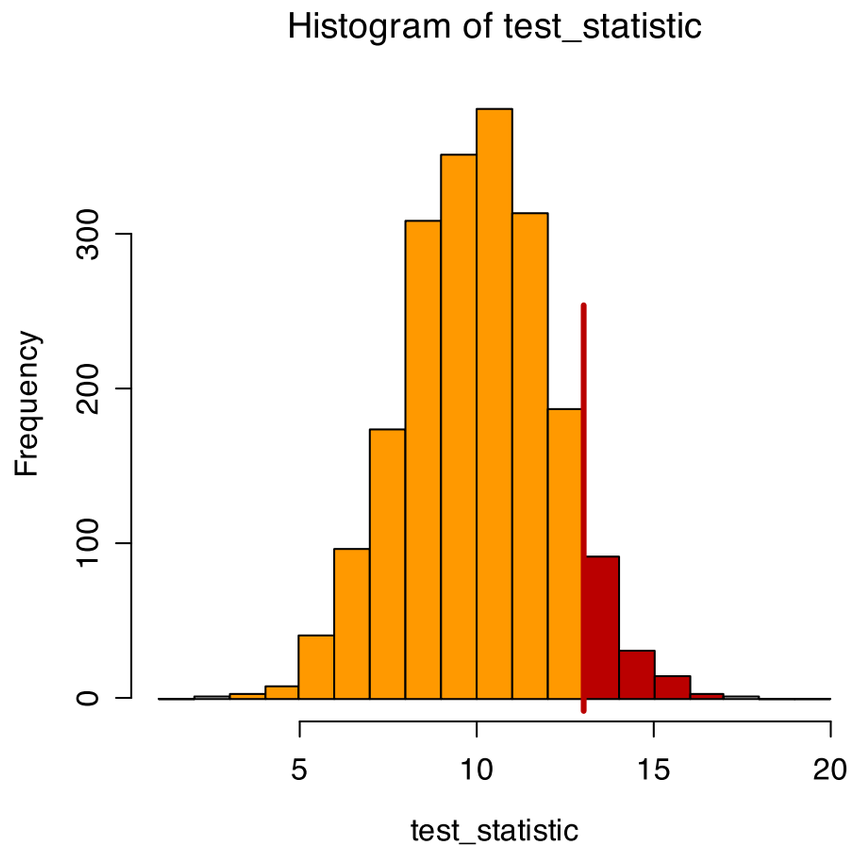
Monte Carlo test on "points inside segments"

- 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

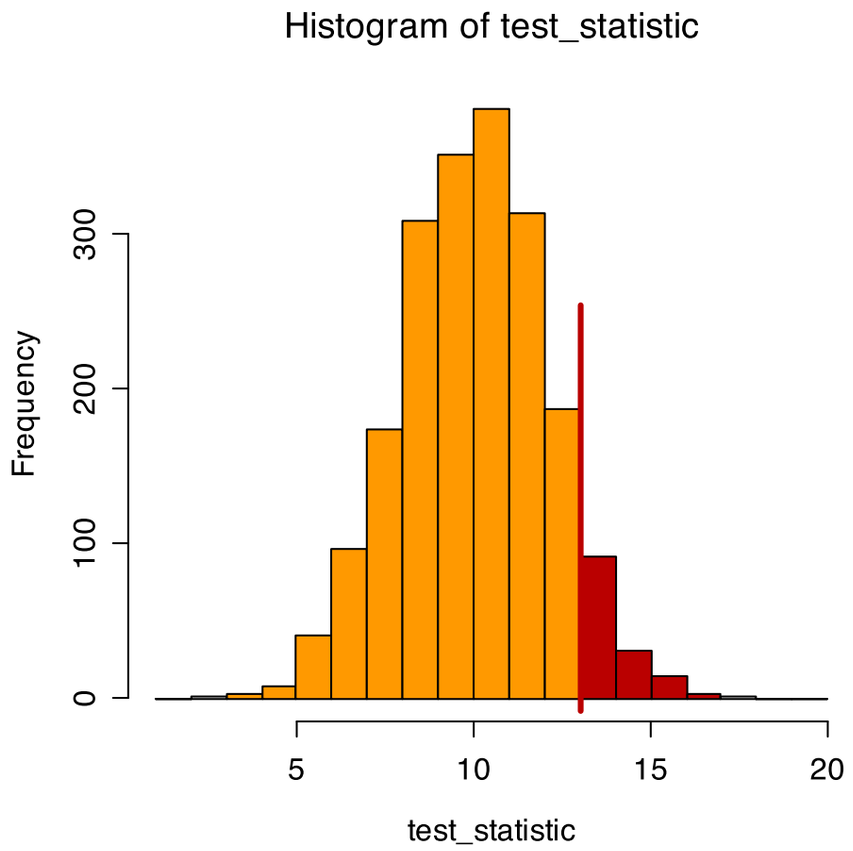
Monte Carlo: distribution

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

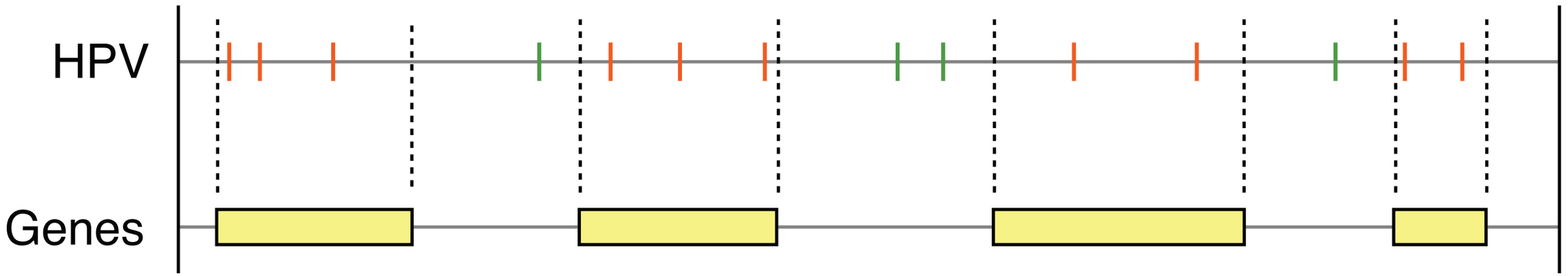
- We do this using Monte Carlo (which is slow) because we have no reason to assume a standard analytical distribution (such as the normal distribution)

    - (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



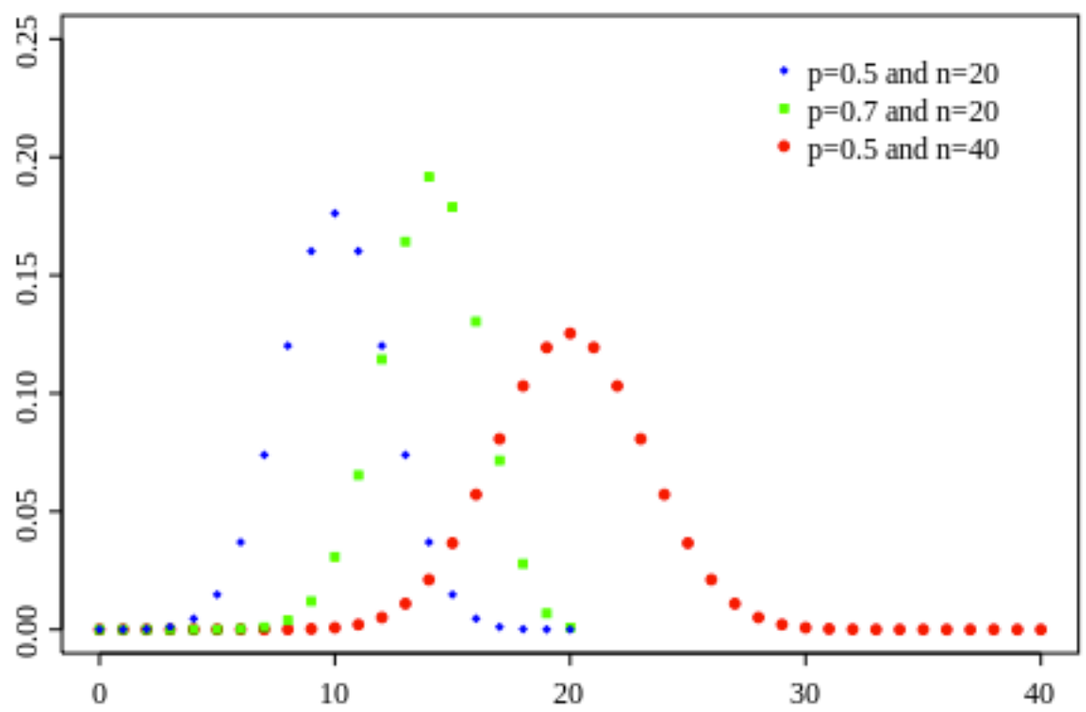
- Can we find a suited analytical distribution? (for number of HPV sites inside genes under H0)
- A statistician may answer: “yes, a binomial distribution”

Binomial distribution

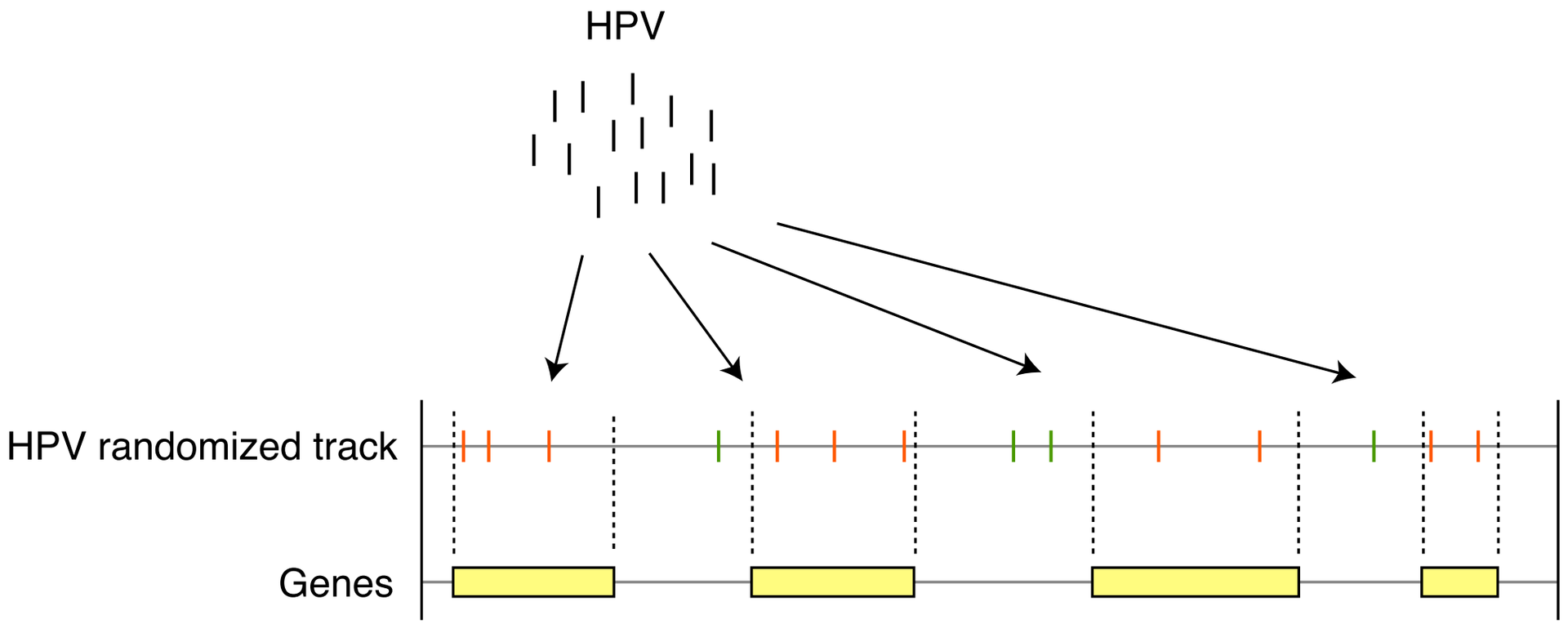
- Flip a coin ***n*** number of times
  - Two outcomes: heads or tails
- But: one side may be heavier than another
  - E.g. the probability of tails:

    `P(tails) = p = 0.6`

    `P(heads) = 1-p = 0.4`
- The distribution is dependent on **p** and **n**

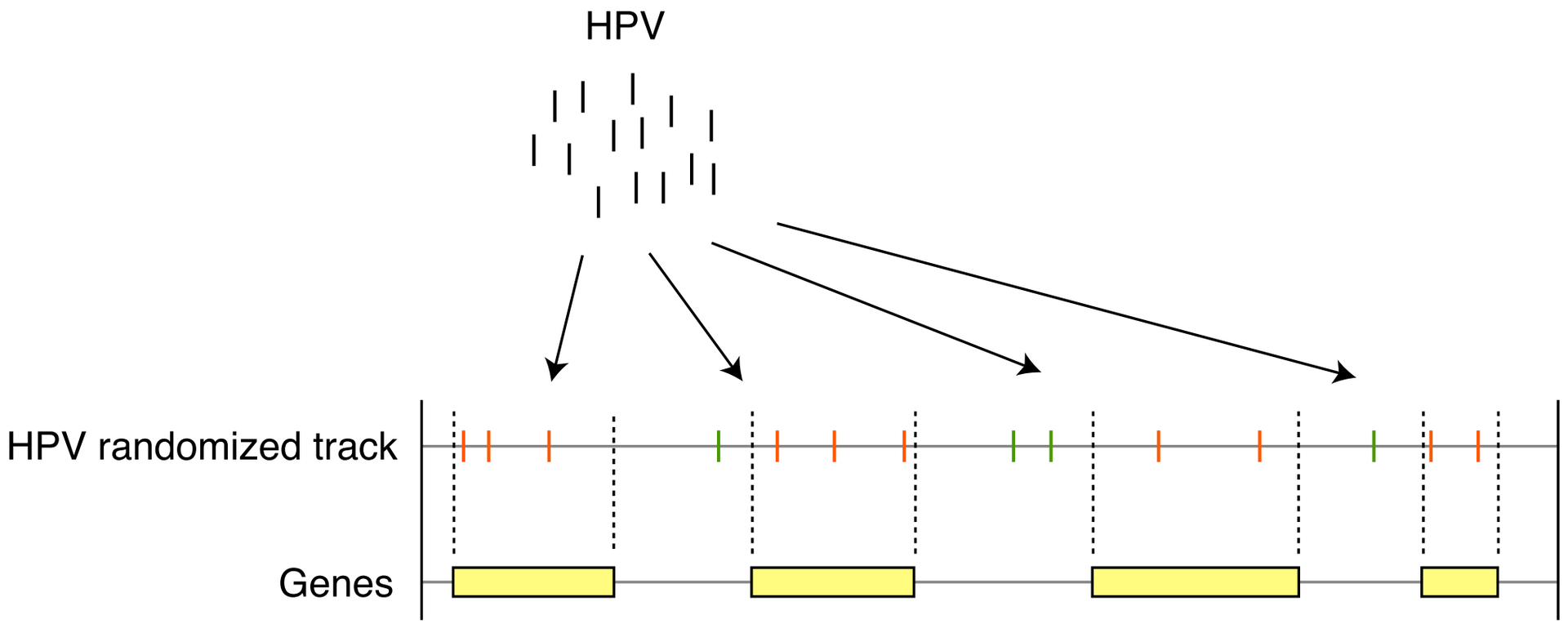


Binomial distribution



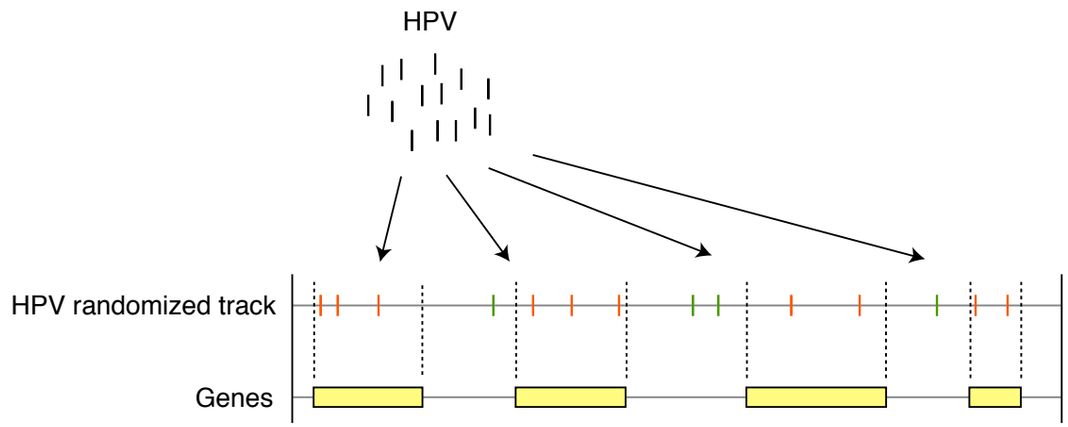
- In this case, each HPV is a coin, and it can either fall into a gene or not, depending on how much of the genome that is covered by genes
- n = number of HPV
- p = proportional coverage of genes

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

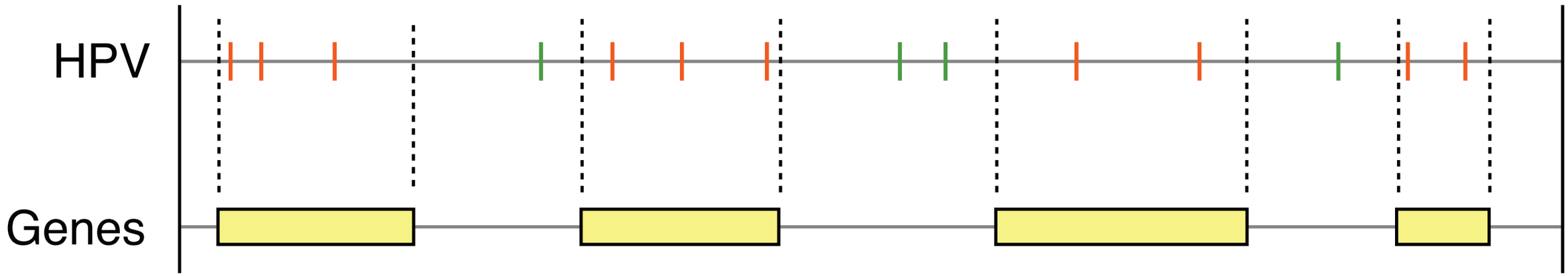
Further into statistical details: the test-statistic

- 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."

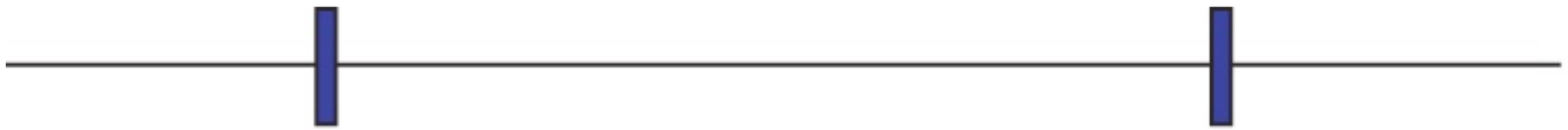
- Let's instead analyze distance to TSS

Back to the whiteboard: the test-statistic

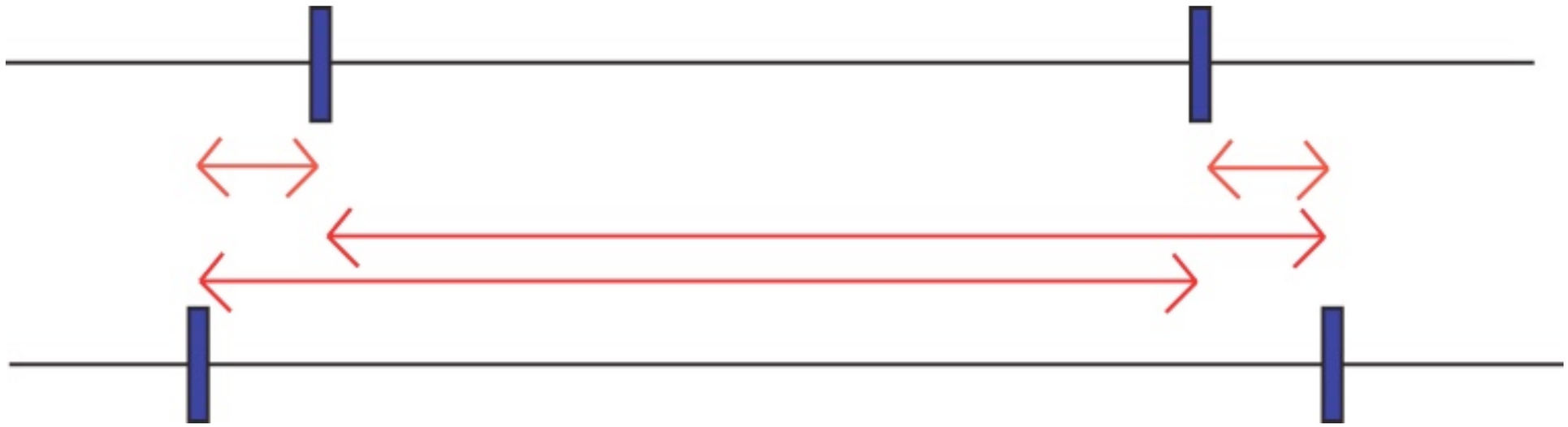


- For “located inside”:
  - Could simply count the number of HPV sites falling inside genes

Back to the whiteboard: Must quantify “close”

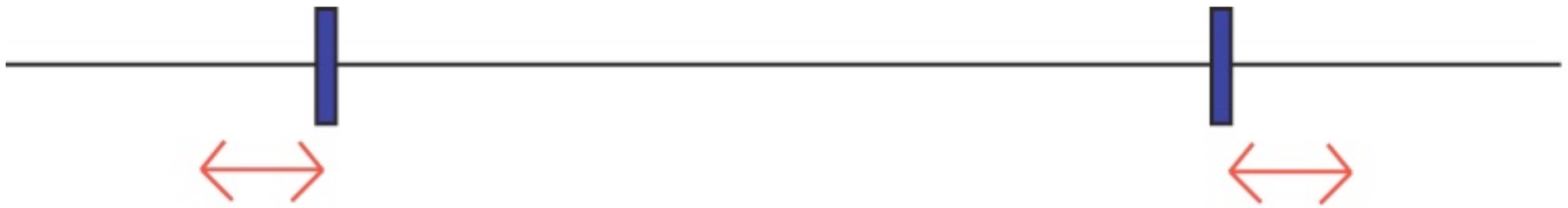


But that's trivial, sure: Just count bp distance!?



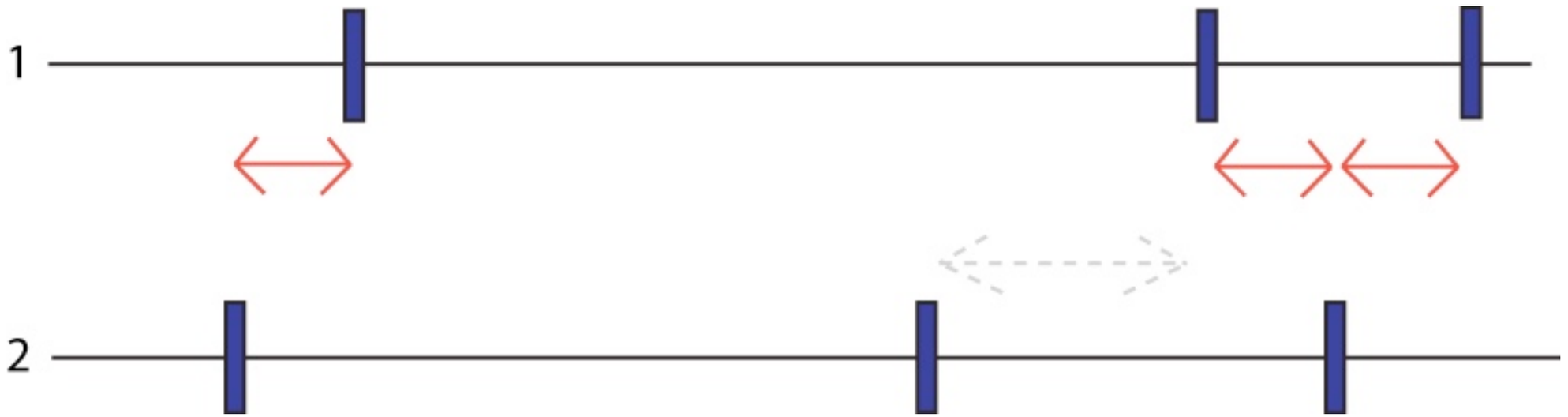
- But which distances - not all vs all?!

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



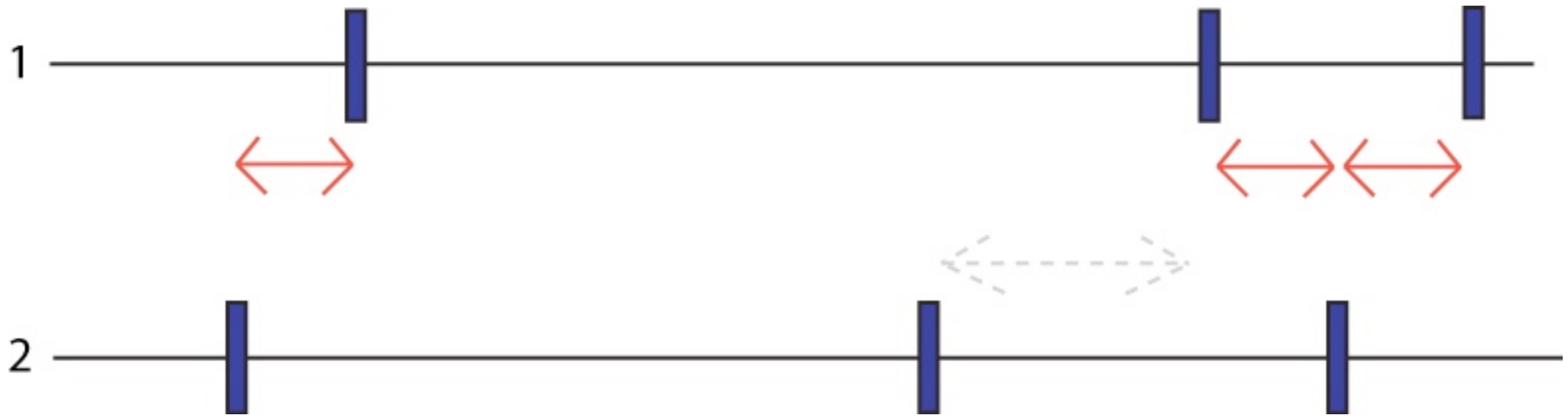
- But which distances - not all vs all?!
  - Only shortest!

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



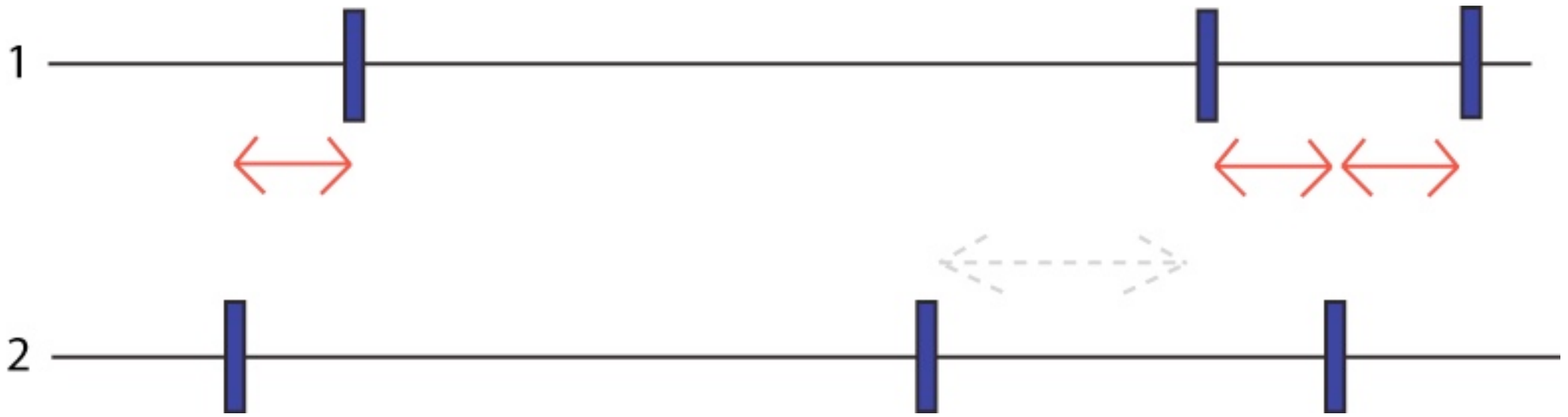
- But which distances - not all vs all?!
  - Only shortest! From 1 to 2!

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



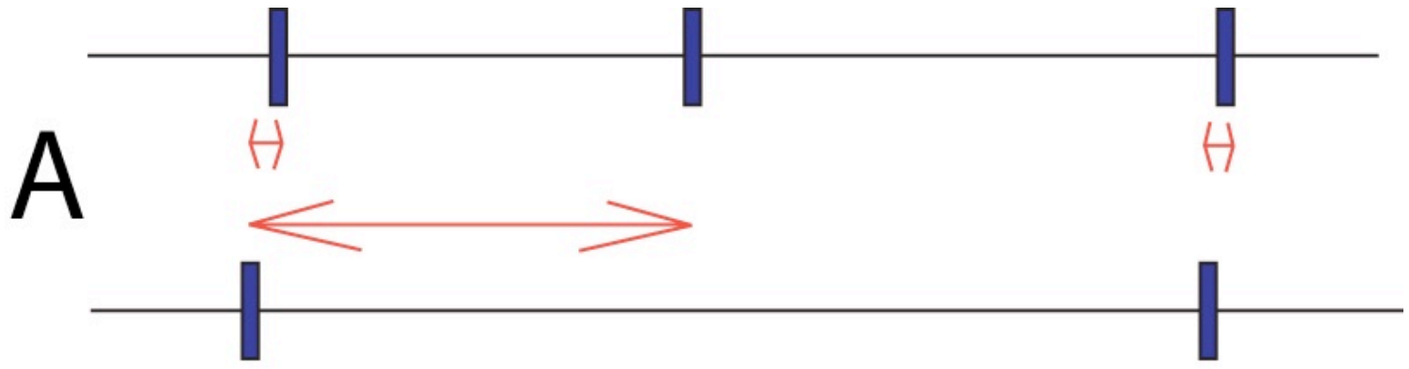
- But which distances - not all vs all?!
  - Only shortest! From 1 to 2! But MC needs a single number..

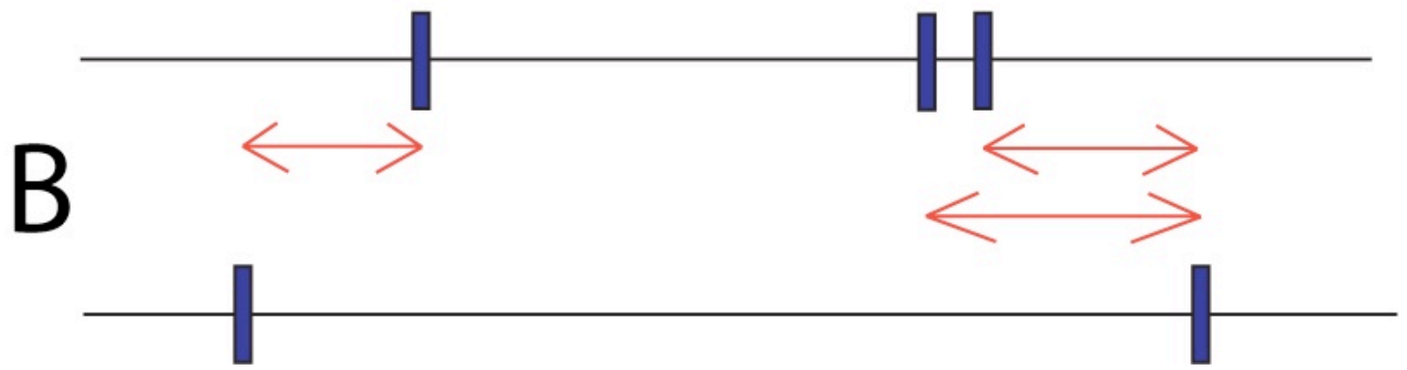
But that's trivial, sure: Just count bp distance!?



- 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

Exercise 11

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

Exercise 11

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

Multiple hypotheses testing

- The setup:
  - A set of n hypotheses that you test simultaneously (e.g. n=20)
  - Significance level α of 0.05
- When n tests are considered, even if all of the null hypotheses are true, there's a certain probability that at least one of them will be rejected by pure chance.
- This probability increases when n increases.

Multiple hypotheses testing

- The Bonferroni correction
  - The significance cut-off is set to α/n (for n=20, the p-value needs to be less than 0.0025)
  - The problem: we assume all tests are independent of each other, which is rarely the case in reality. This makes the process highly conservative, leading to actually missing out on significant findings
- Controlling for the False Discovery Rate (FDR)
  - The proportion of false significant findings, i.e. tests that rejected the null hypothesis by chance when it was actually true
  - This approach is less conservative than the Bonferroni correction

Implementation

Implementation

- Until now, the treatment has been mostly theoretical, not really considering the problem of software
- As a bioinformatician you will often have to choose between:
  - Using software developed by someone else
  - Developing you own scripts/software

Things to take care of (when developing your own software)

- Parse data
- Indexing to easily find data for a region
- Develop test statistic (e.g. how many points are inside segments?)
- Create algorithms that don't take forever to complete
- Implement Monte Carlo correctly
- If very large data:
  Split, intermediate computation, combine

Things to take care of (when developing your own software)

- Must check for bugs!
- Double-check with test cases
- Any silly bugs?
- Formats understood correctly?
- Remembered strand?

Finally..

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

Using existing software

- In the Genomic HyperBrowser, we have spent 8 years of development to create a feature-rich and well tested system
- There are other alternatives to the HyperBrowser, for instance BEDtools (command line) or Bioconductor (in R), that could also be used
- But remember that all shuffling assumes a null model.
  - E.g. in BEDtools:
    - Uniform shuffling with possible overlaps
    - Centromeres not explicitly handled
    - Both of the above gives better (more significant) results than with more realistic assumptions
  - In the HyperBrowser
    - User choice of Null models, all with overlap handling

      (NB: error in article in curriculum!)
    - Centromeres are handled

And the moral is..

- Before you start spending (way too much) time on developing from scratch,
- .. be sure to first spend enough time looking for an existing solution
- .. and also be aware of assumptions in software that will have implications for your results
  - (at least try alternatives and report your choice)

Data upload

Data upload (in the real world)

HPV data_GeirKjetil.xls

New Open Save Print Import Copy Paste Format Undo Redo AutoSum Sort A-Z Sort Z-A Gallery Toolbox 100% Zoom Help

Sheets Charts SmartArt Graphics WordArt

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Sample**	**Chromosome**	**Strand**	**Coordinates**								
2	T183	1	+	24601512								
3	T627	1	+	94760681								
4	D3571	1	-	98596236								
5	T1981	2	-	10121207								
6	T174	2	-	15985150								
7	T644	2	+	74886368								
8	D3918	2	+	141574075								
9	T654	2	+	141790172								
10	F251	2	+	157178230								
11	T4177	2	+	160743472								
12	D3931	2	+	191141041								
13	D3994	2	+	207736951								
14	D4034	2	+	213747033								
15	D3622	3	+	4202544								
16	D3829	3	-	191175170								
17	D4023	3	-	190530615								
18	T2341	4	+	20558738								
19	D3899	4	+	55177868								
20	D3421	4	-	74906729								
21	D3458	4	-	74841098								
22	D4049	4	-	74862313								
23	T1051	4	-	100606970								
24	D3545	4	-	139323911								
25	D3826	4	+	169238923								

Commonly used format: Excel!

What we want

```
##gtrack version: 1.0
##track type: points
##1-indexed: true
##end inclusive: true
###seqid  start      strand  id
chr1      24601512   +       T183
chr1      94760681   +       T627
chr1      98596236   -       D3571
chr2      10121207   -       T1981
chr2      15985150   -       T174
chr2      74886368   +       T644
chr2      141574075  +       D3918
chr2      141790172  +       T654
chr2      157178230  +       F251
chr2      160743472  +       T4177
```

GTrack format

GTrack

- File format developed for the HyperBrowser (but can be used by other software)
- There are more common file formats you can use, e.g.
  - BED, WIG, GFF...
  - Main problem: does not support all track types

Which track types do existing formats support?

BED

Format	Ref.	Data	Repr.	P	S	VP	VS	GP	SF	F	L	Strand	#Cols	Value type
GFF3/GTF	[2]	General	Tab.	√(1)	√	√(1)	√				(2)	√	9	Float (3)
BED/bigBed	[4]	General	Tab./ Bin.	√(1)	√	√(1)	√				(2)	√	3-12	Int(0-1000) /string (4)
BED15	[4]	Microarray	Tab.			√(1)	√				(2)	√	15	List of floats (5)
bedGraph	[4]	General	Tab.			√(1)	√						4	Float
WIG/bigWig (fixedStep)	[8]	General	Tab./ Bin.			√	√		√	√			1	Float
WIG/bigWig (variableStep)	[8]	General	Tab./ Bin.			√	√						2	Float
CNT	[36]	Copy number	Tab.			√							4	Float
Personal Genome SNP	[4]	Variation	Tab.			√(1)	√						7	String (6)
VCF	[37]	Variation	Tab.			√	√						≥ 8	String (6) (3)
GVF	[6]	General/ Variation	Tab.	√(1)	√	√(1)	√				(2)	√	9	Float (3)
PSL	[4]	Alignment	Tab.		√		√					√	21	Int (7)
SAM/BAM	[38]	Alignment	Tab./ Bin.		√		√					√	11	Int /string (8)
BioHDF	[39]	Alignment	Bin.		√		√					√	11	Int /string (8)
MAF	[4]	Multiple Alignment	Tab.		√		√				(9)	√	2-7	Float /string (8)
FASTA	[40]	Sequence	Text							√			N/A	Char
DAS XML	[12]	General	XML	√(1)	√	√(1)	√				(2)	√	N/A	Float
BioXSD 1.0	[16]	General	XML	√(10)	√(10)	√(10)	√(10)				√(11)	√	N/A	Float (12)
USeq	[19]	General	Bin.	√	√	√	√					√	N/A	Int/float/string
Genomedata	[41]	General	Bin.			√	√		√	√			N/A	Int/float/char

GTrack format

- All track types are supported by GTrack (may replace most of the formats of the last slide)
- Supports a variable number of columns
- Fully supported by the Genomic HyperBrowser
- HyperBrowser includes 7 specific GTrack tools, including a tool for converting between GTrack and other file formats
- http://www.gtrack.no

What does the contents mean?

Track type
(you know this...)

```
##gtrack version: 1.0
##track type: points
##1-indexed: true
##end inclusive: true
###seqid  start     strand  id
chr1      24601512  +       T183
chr1      94760681  +       T627
chr1      98596236  -       D3571
chr2      10121207  -       T1981
chr2      15985150  -       T174
chr2      74886368  +       T644
chr2      141574075 +       D3918
chr2      141790172 +       T654
chr2      157178230 +       F251
chr2      160743472 +       T4177
```

Columns are:
- chromosome
- (start) position
- DNA strand
- ID of element

Biologists start counting at 1, and a segment from [1,10] includes base pairs 1 through 10

Computers start counting at 0, and a segment from [0,10] includes base pairs 0 through 9

Exercise 12

- Aim: Convert from spreadsheet document to GTrack for analysis
- "*Create GTrack file from unstructured tabular data*" tool (under "GTrack tools" header:
  - Select: Tabular file from input box
  - Copy contents from spreadsheet, paste into box
  - Skip 1 line
  - Select column names: id, seqid, strand, start
  - Select specific genome: Yes: hg18
  - Select: 1-indexed, end inclusive
  - Select: Auto-correct to the best match in the genome
  - Click Execute
- Use the "Analyze genomic tracks" tool. You can now select the track "from history". Try to find the number of SNPs in each chromosome.

Data upload in Galaxy

- In the exercise, you just copied and pasted some data
- For larger datasets, Galaxy has an upload tool, under “Get Data”=>”Upload File from your computer”
- Here, you can select files from your local computer, or you can paste URLs for data located on the web

Reproducibility

Reproducibility

- The advantages of making your research reproducible have been discussed in previous sessions (among others, the "introduction to Galaxy" session)
- 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
  - sveinung.gundersen@medisin.uio.no