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

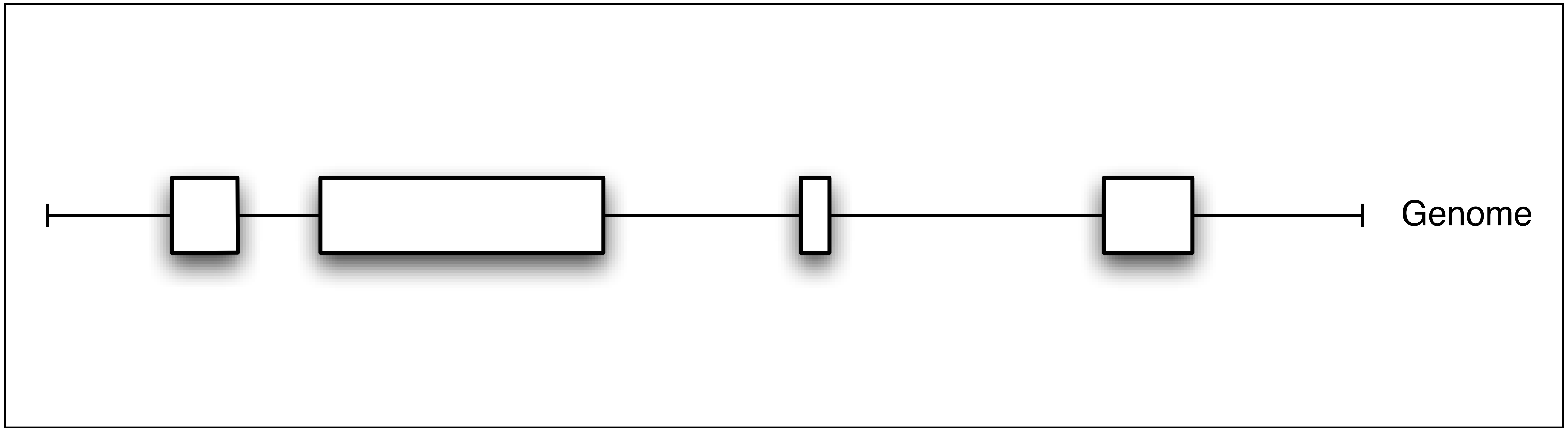
- We will make use of the GSuite HyperBrowser in this session
- The HyperBrowser is a software system for statistical analysis, developed locally at UiO
- However:

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

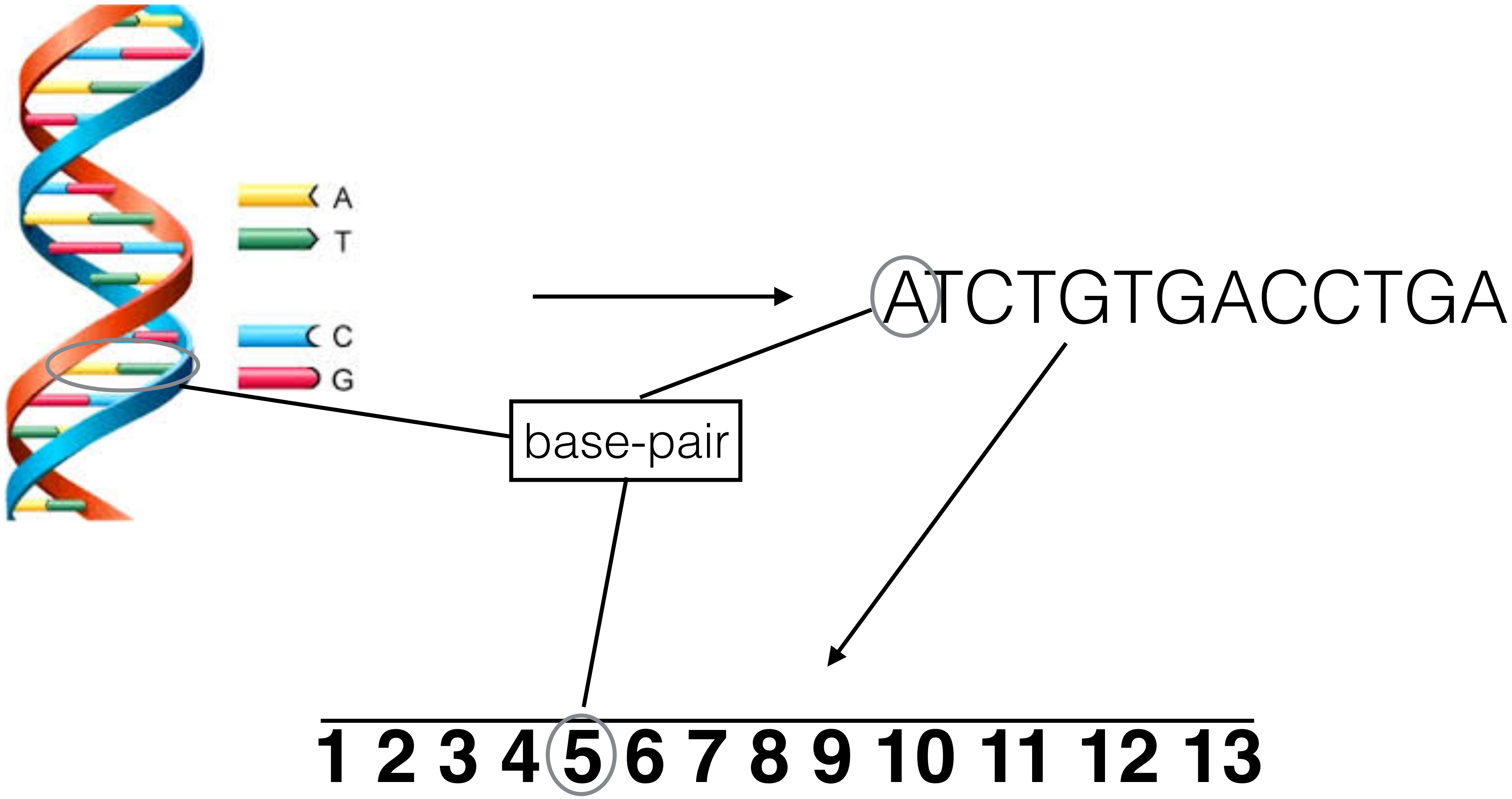
Introduction

What are genes?

This! :



Genome as a line

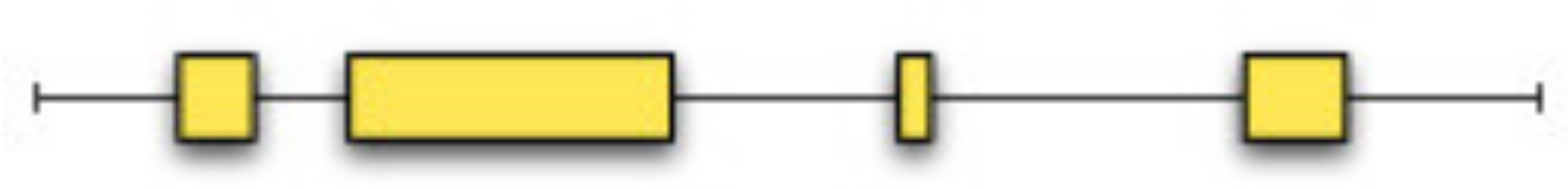


ATCTGTGACCTGA

base-pair

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

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



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

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

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

Statistical genomics

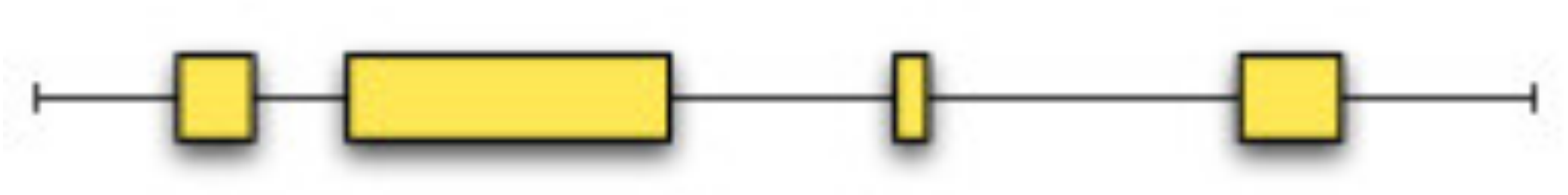
- Often used for statistical analysis of:
  - Gene lists (e.g. Gene set enrichment analysis, GSEA)
  - Gene expression (Differential expression)
  - SNPs (e.g. Genome-wide association studies, GWAS)
  - etc..
- We are not going to do any of the above

Statistical genomics

- Statistical analysis of genomic tracks
  - Tracks: genome-wide datasets than can be positioned along a reference genome (DNA)
- However:
  - Many of the concepts are central statistical concepts that can be used for other types of analyses

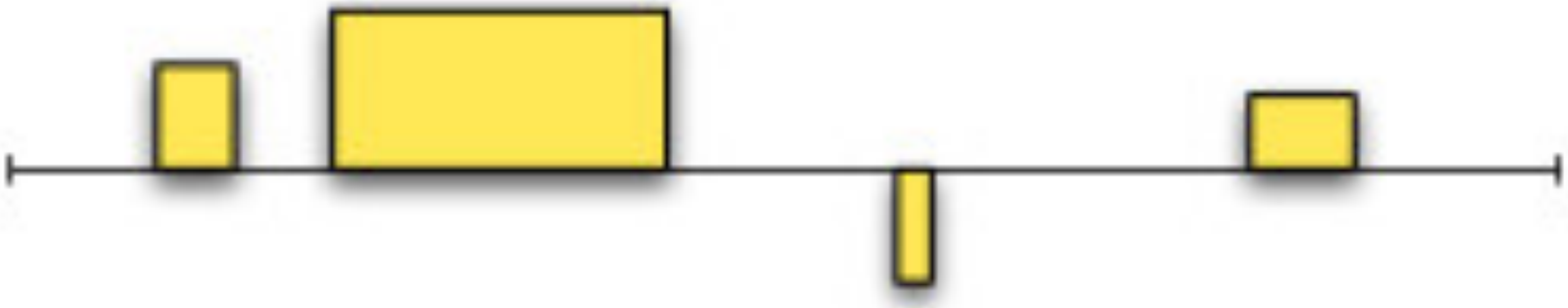
Tracks and track types

Representation of genes



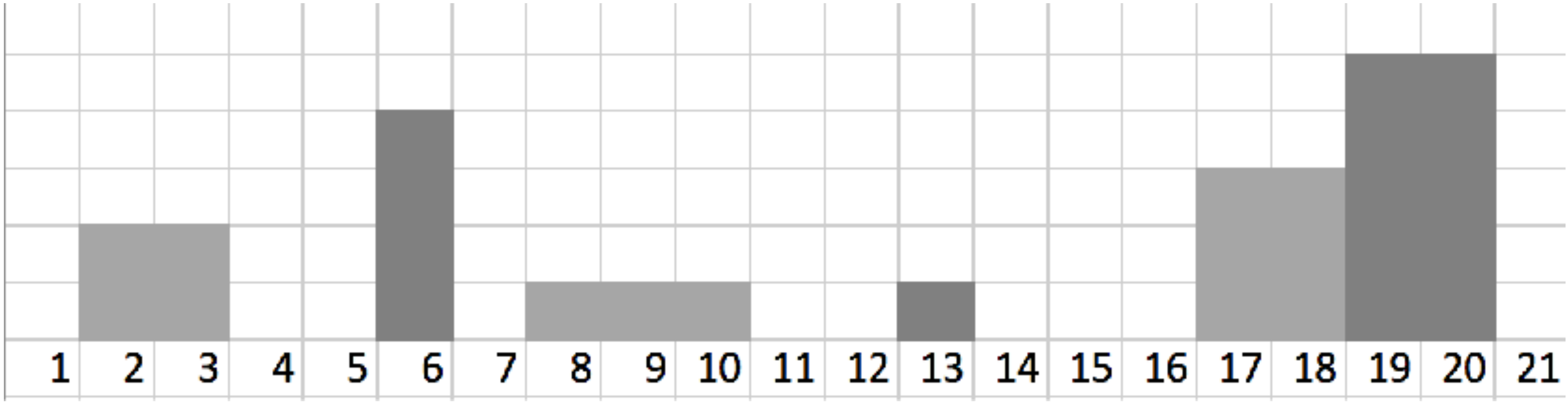
chr7	127471196	127472363
chr7	127472388	127473530
chr7	127473555	127474697
chr7	127474701	127475864
chr7	127475893	127477031
chr7	127477121	127478198
chr7	127478300	127479365
chr7	127479375	127480532
chr7	127480538	127481699

How about gene expression data (RNA-seq)?



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

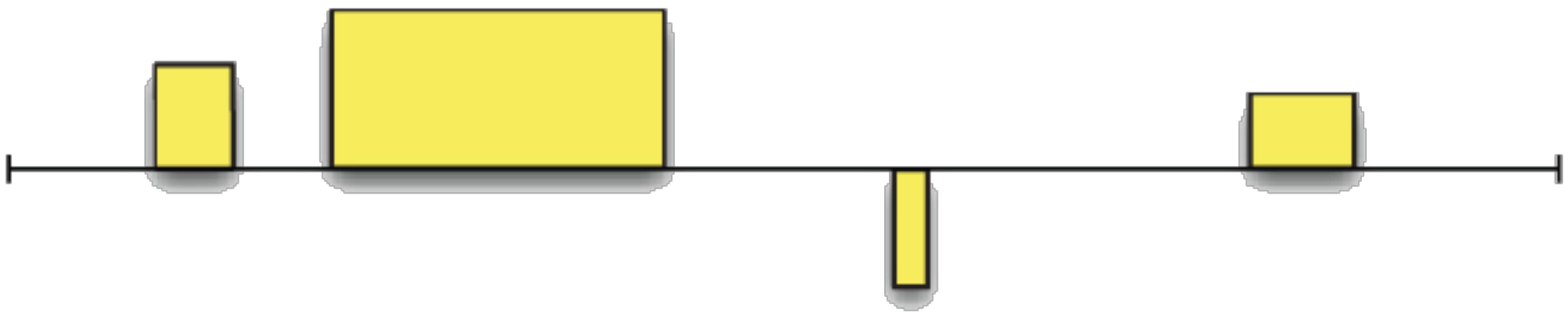
Exercise 1



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

Track types

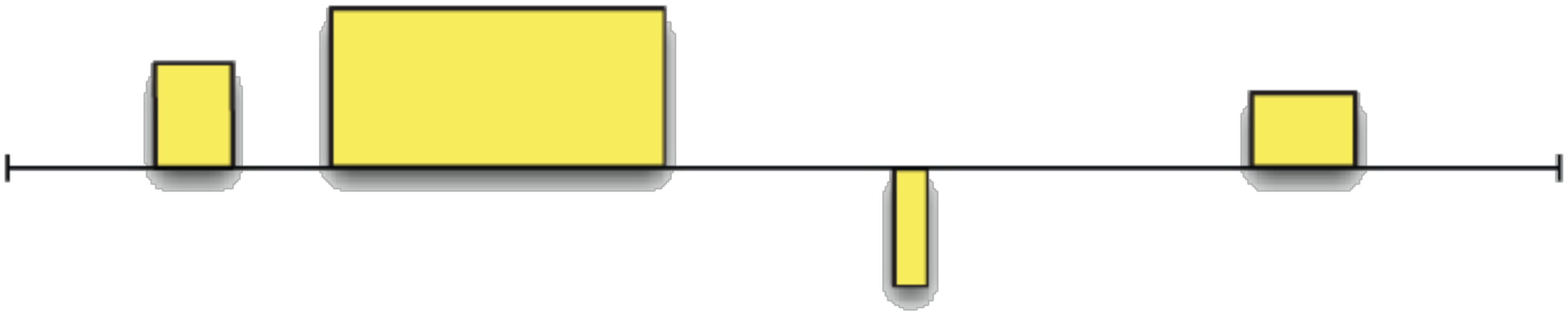
- In the last example, we showed genes as segments on the genome line, with attached RNA-seq read count values
- This track is of a **track type** we call "valued segments"



Valued Segments (VS)

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

Exercise 2



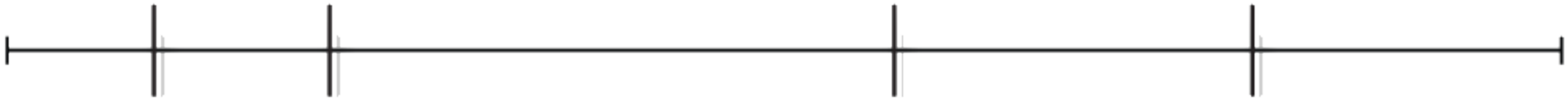
Valued Segments (VS)

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

○ ○ ○ ○ ○ ○

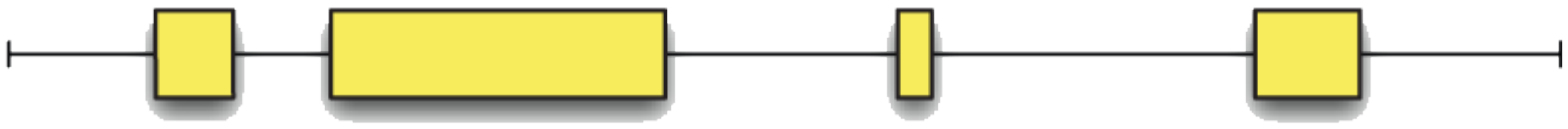
○ ○ ○ ○ ○ ○ ○ ○ ○ →

Points



Points (P)

Segments



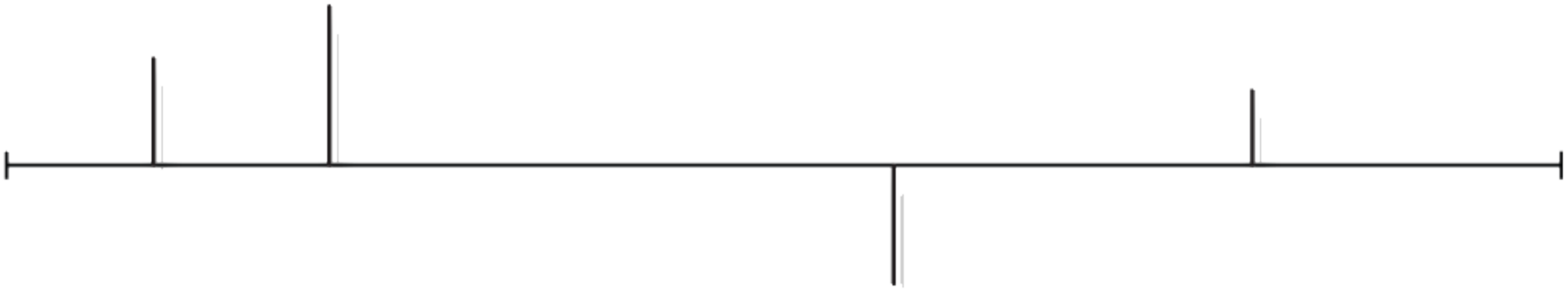
Segments (S)

↵

Genome Partition

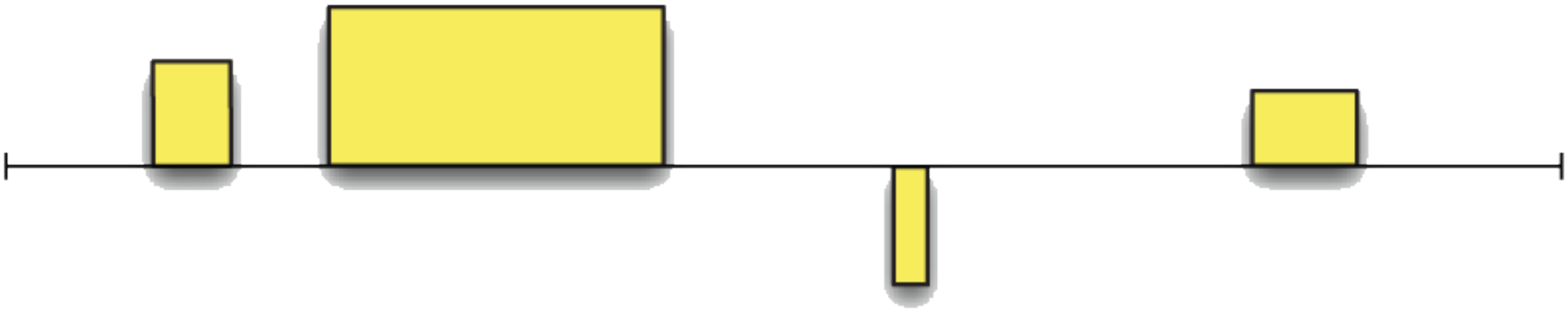
Genome Partition (GP)

Valued Points



Valued Points (VP)

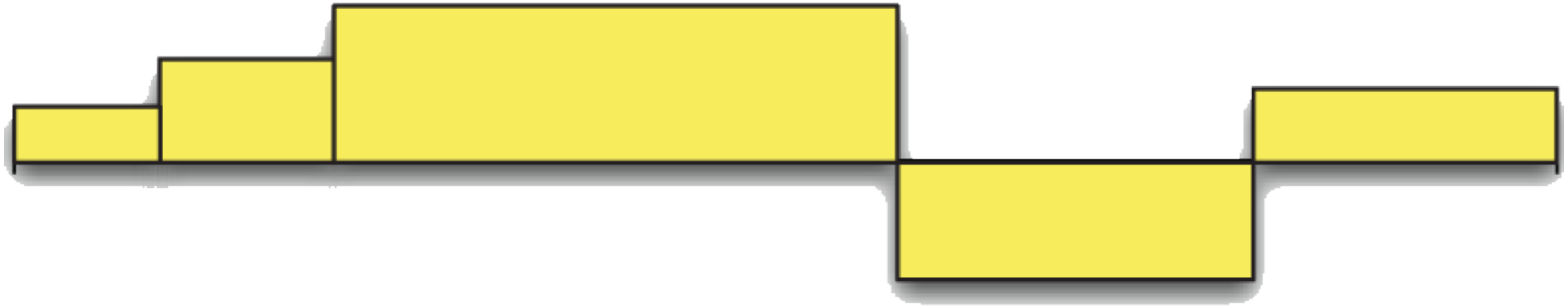
Valued Segments



Valued Segments (VS)

↵

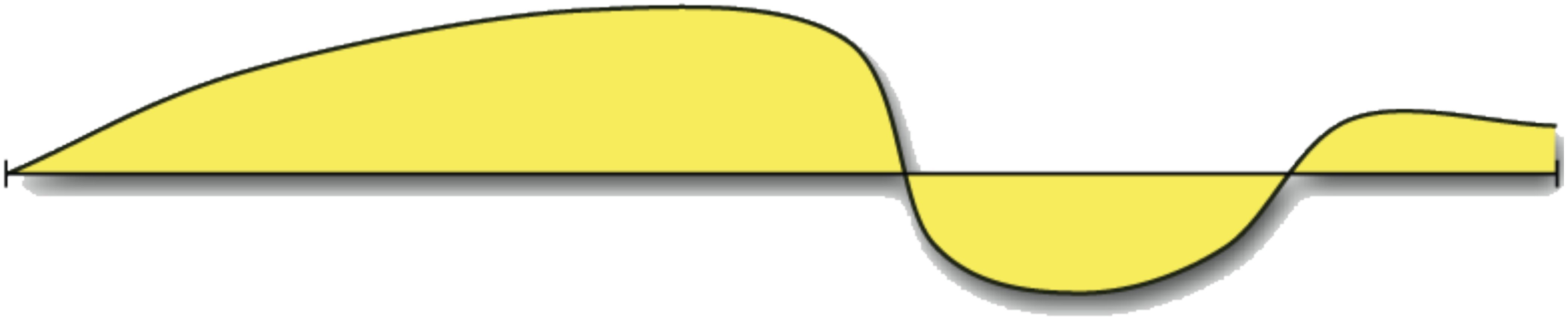
Step Function



Step Function (SF)

↵

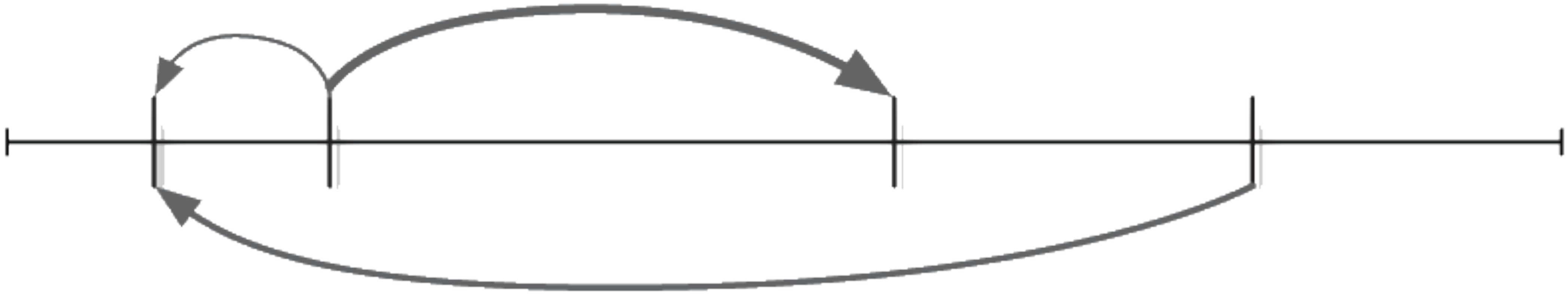
Function



Function (F)

↵

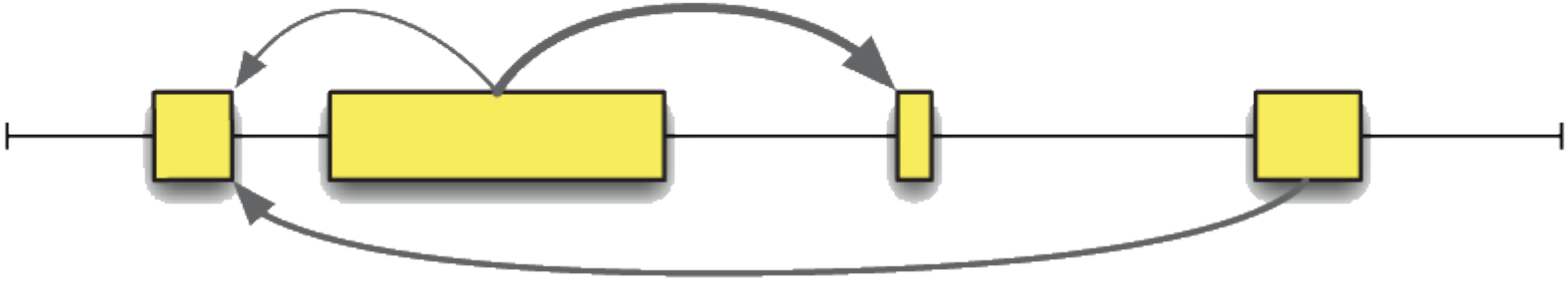
Linked Points



Linked Points (LP)

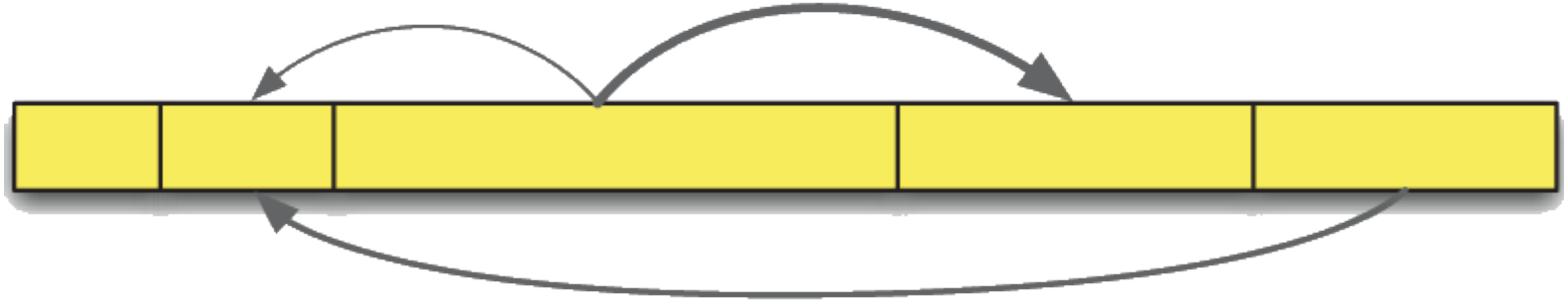
↵

Linked Segments



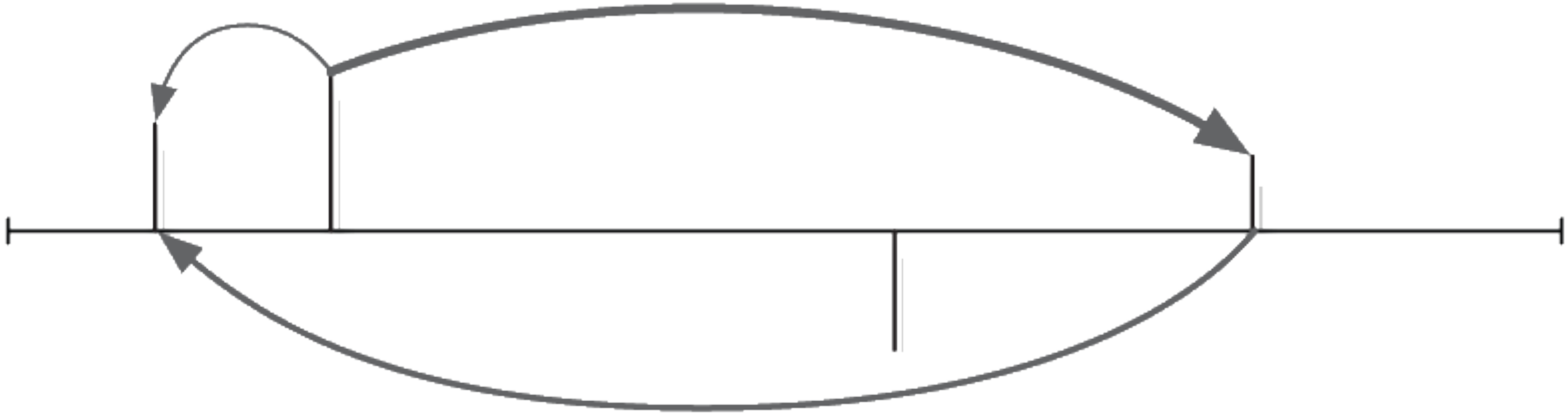
Linked Segments (LS)

Linked Genome Partition



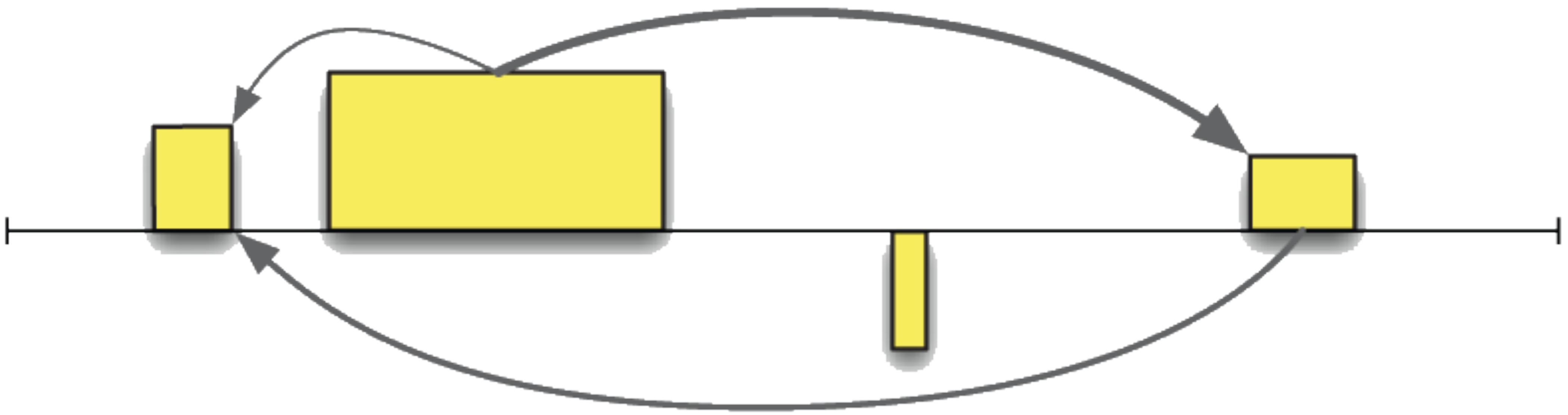
Linked Genome Partition (LGP)

Linked Valued Points



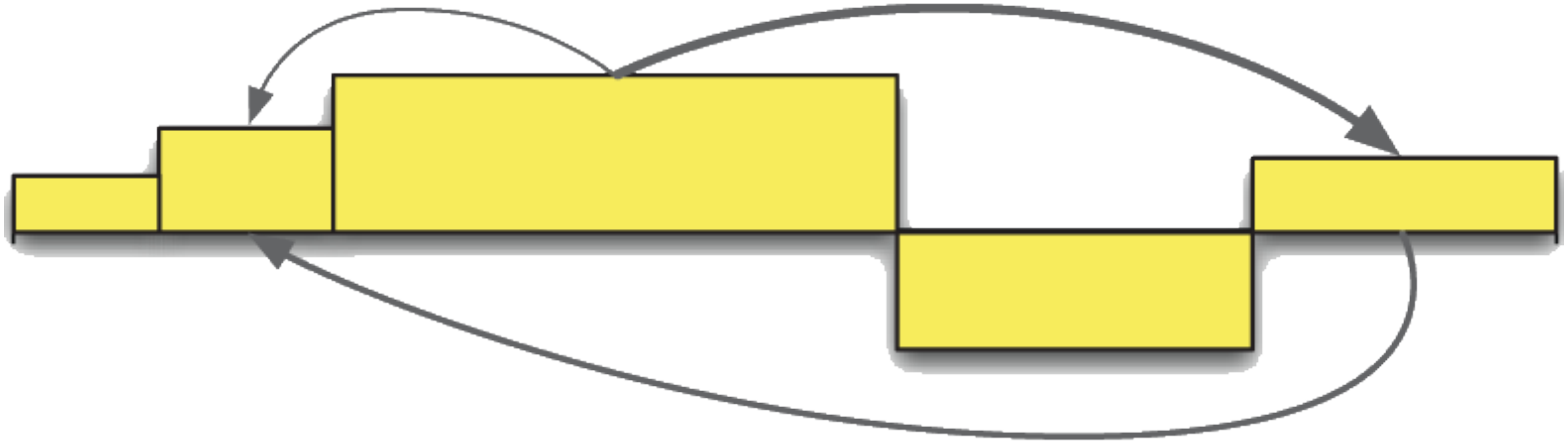
Linked Valued Points (LVP)

Linked Valued Segments



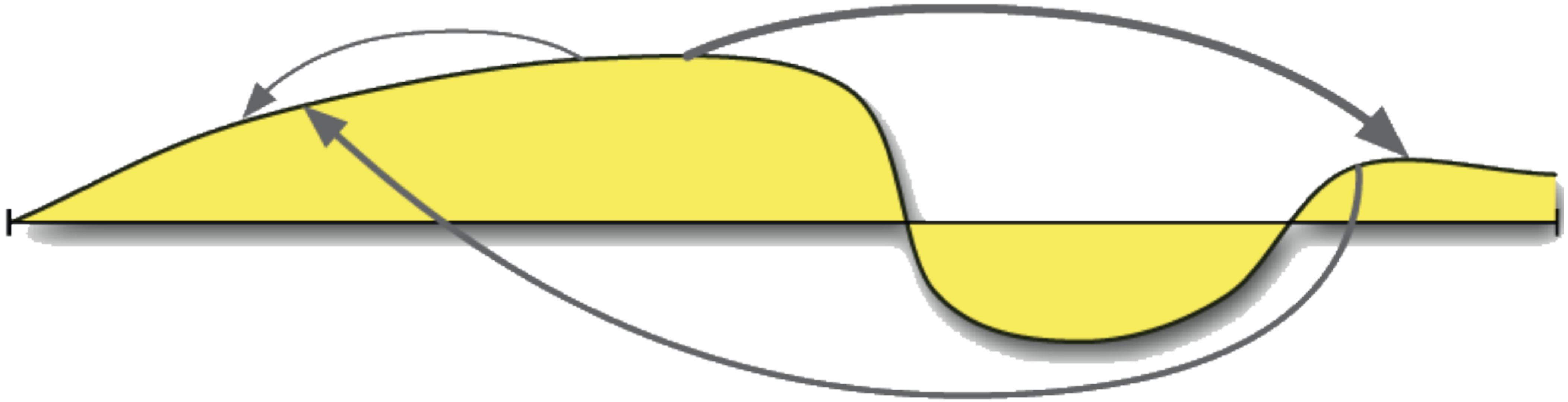
Linked Valued Segments (LVS)

Linked Step Function



Linked Step Function (LSF)

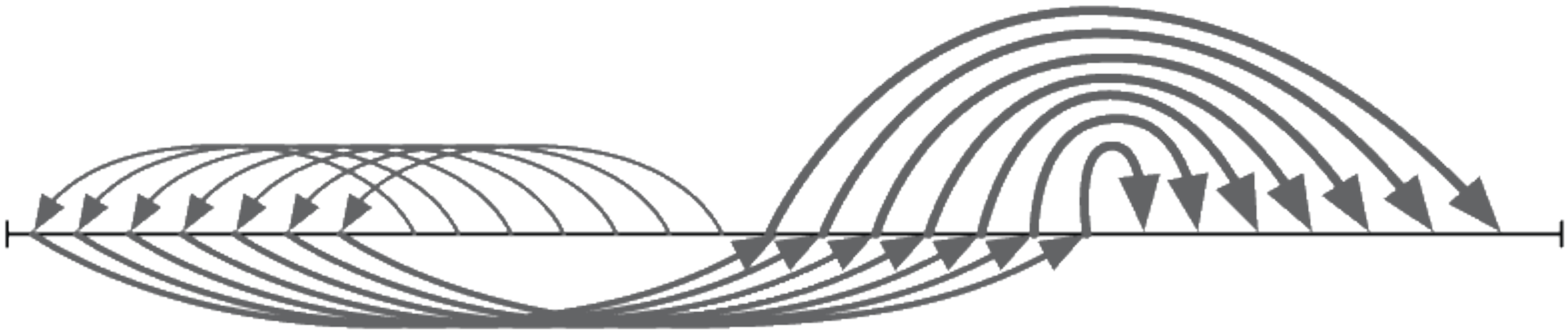
Linked Function



Linked Function (LF)

↵

Linked Base Pairs

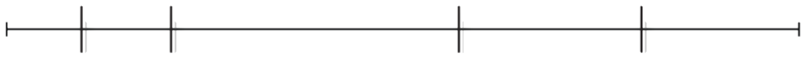


Linked Base Pairs (LBP)

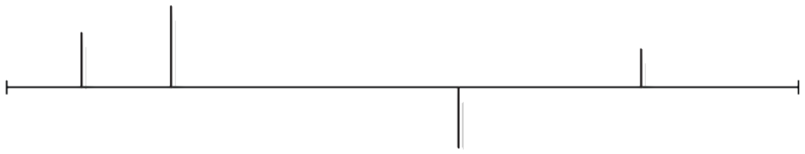
Exercise 3

- 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?

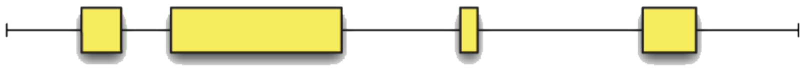
Exercise 3



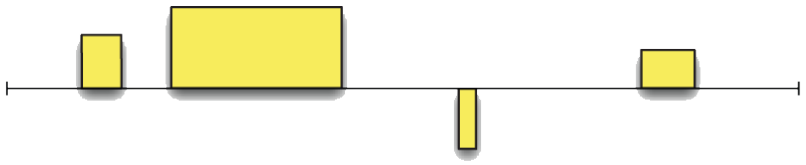
Points (P)



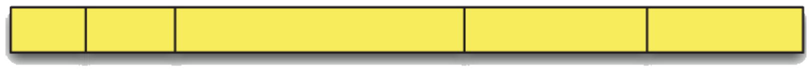
Valued Points (VP)



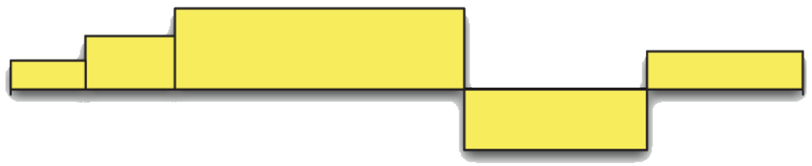
Segments (S)



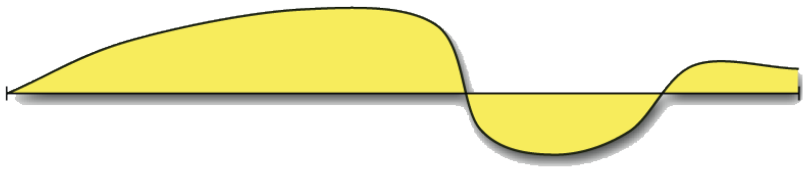
Valued Segments (VS)



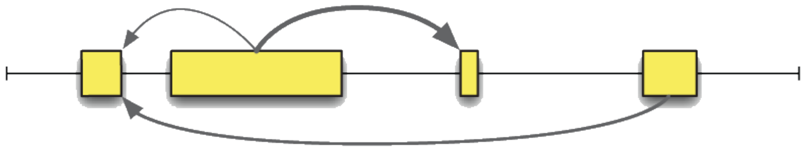
Genome Partition (GP)



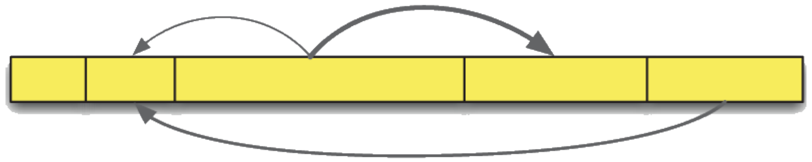
Step Function (SF)



Function (F)

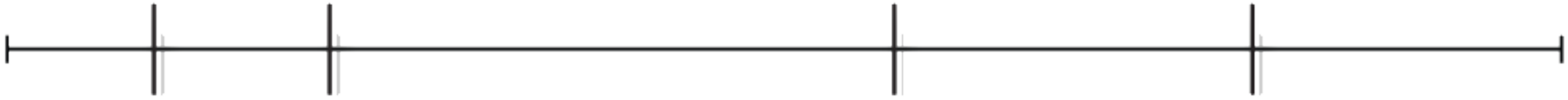


Linked Segments (LS)



Linked Genome Partition (LGP)

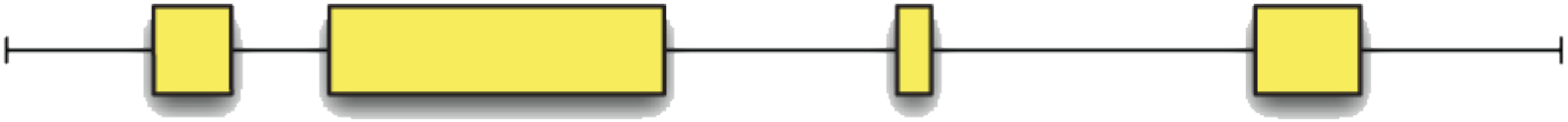
Points



Example tracks:

- SNPs

Segments

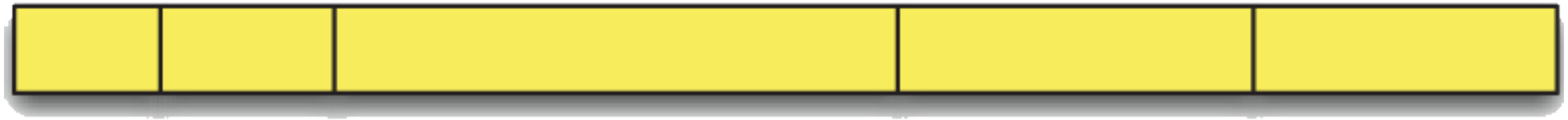


Example tracks:

- Genes
- Transcription factor binding sites

↵

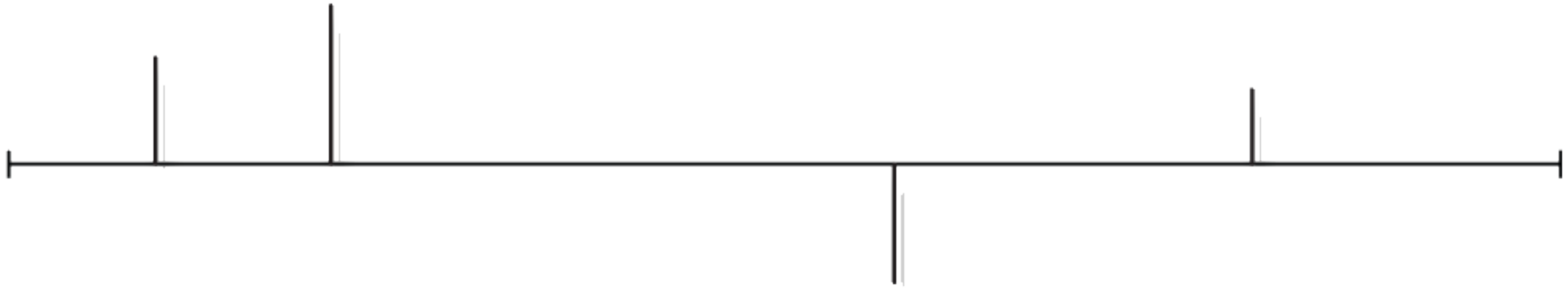
Genome Partition



Example tracks:

- Chromosomes
- Chromosome arms
- Chromatin state segmentation

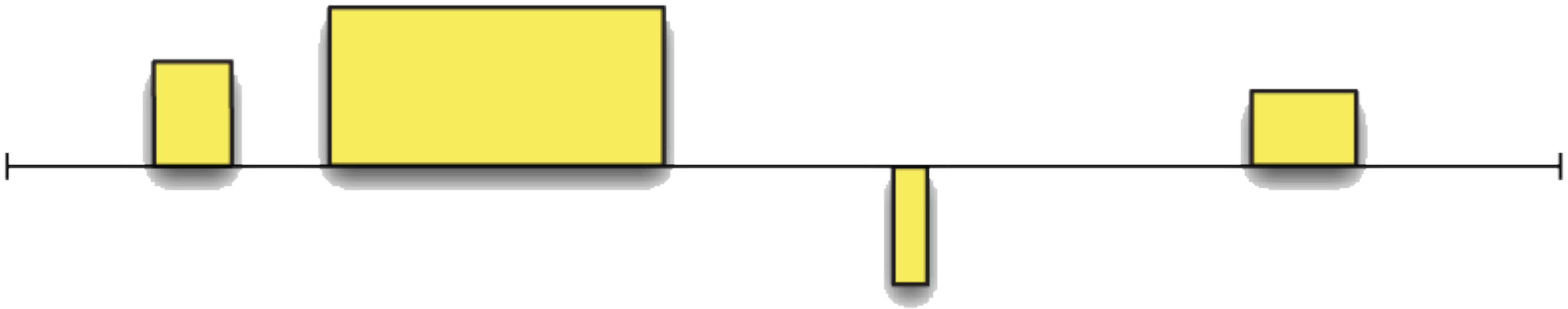
Valued Points



Example tracks:

- SNPs with allele frequency
- SNPs with quality

Valued Segments

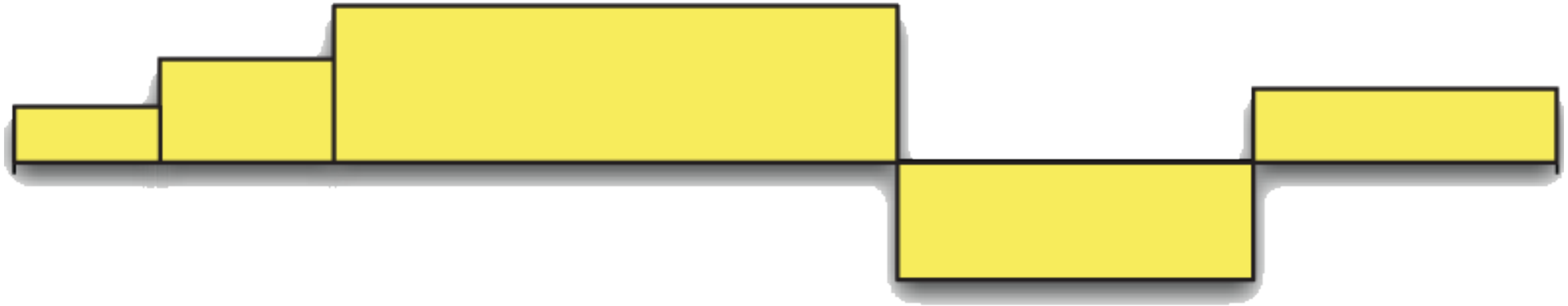


Example tracks:

- Genes with expression values
-

↵

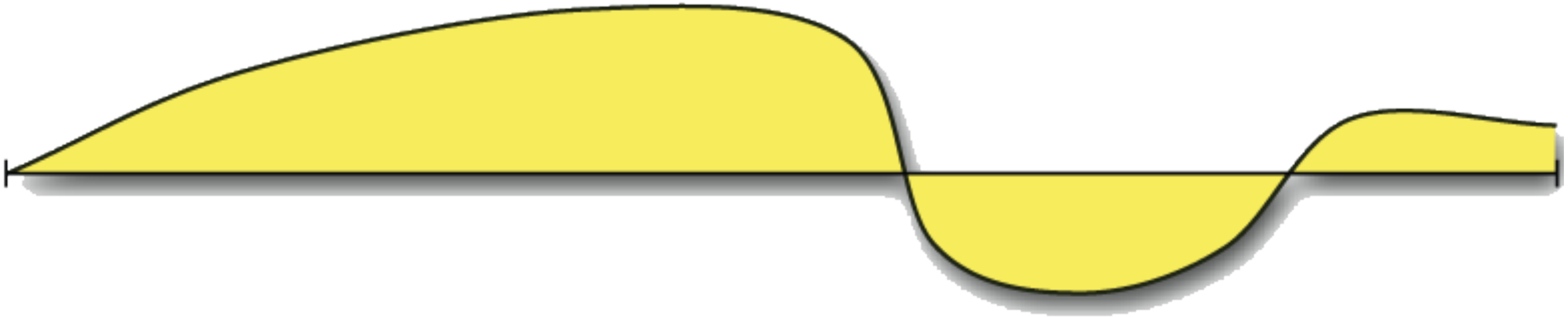
Step Function



Example tracks:

- GC content (per partition)

Function

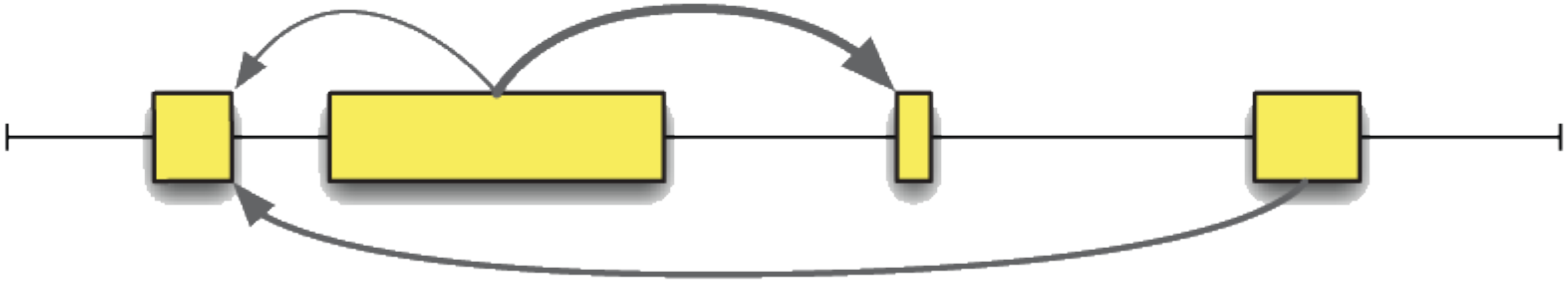


Example tracks:

- DNA melting temperature
- Coverage (RNA-seq)

↵

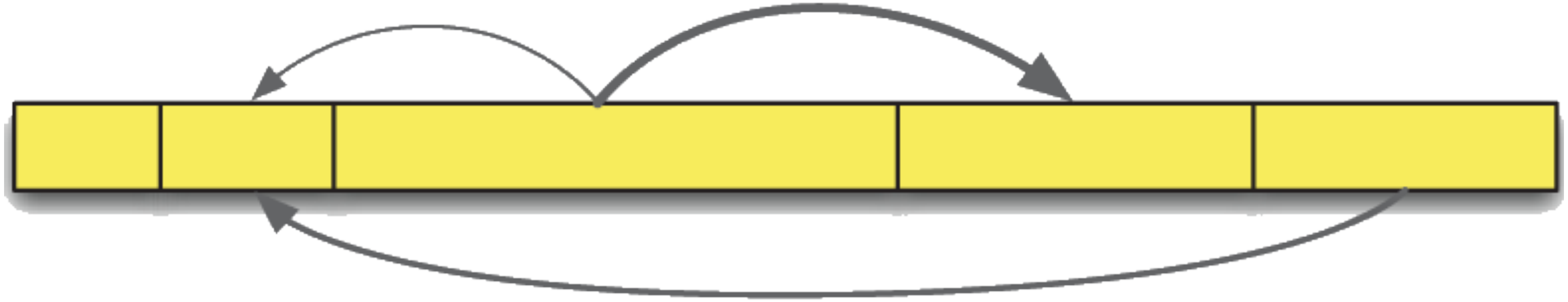
Linked Segments



Example tracks:

- ChIA-PET
- Co-expressed genes

Linked Genome Partition

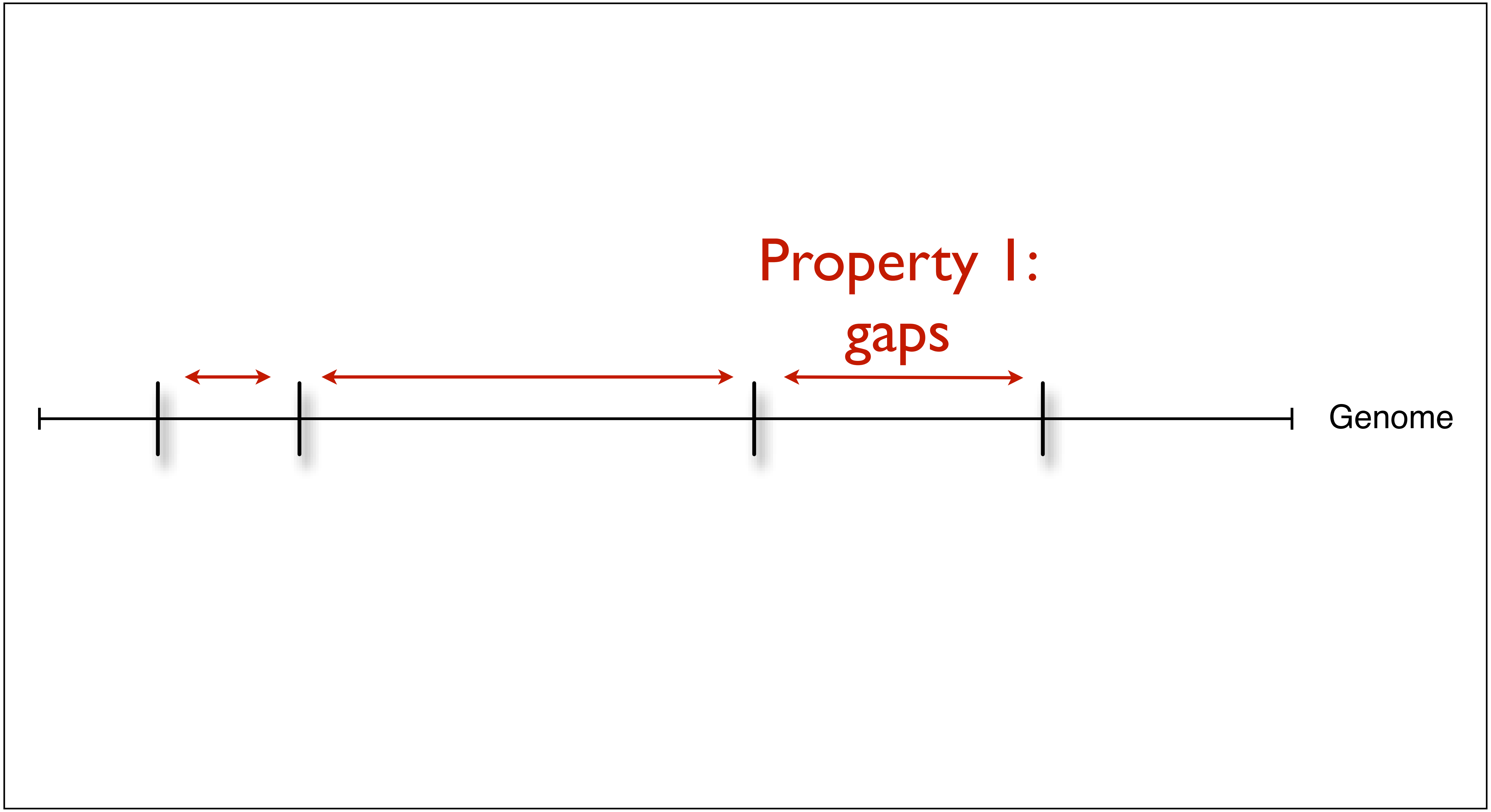


Example tracks:

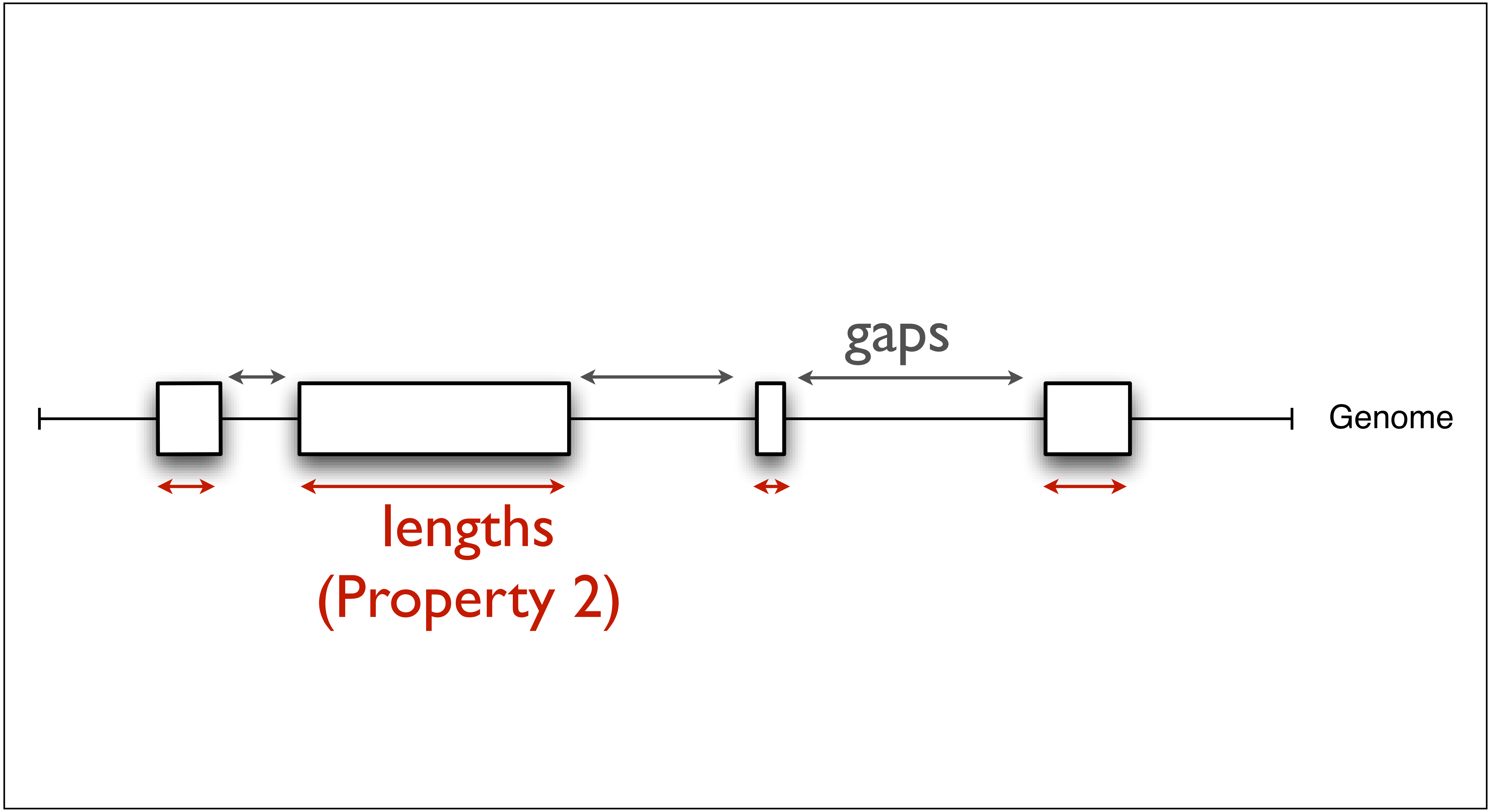
- Hi-C (3D chromatin conformation)

↵

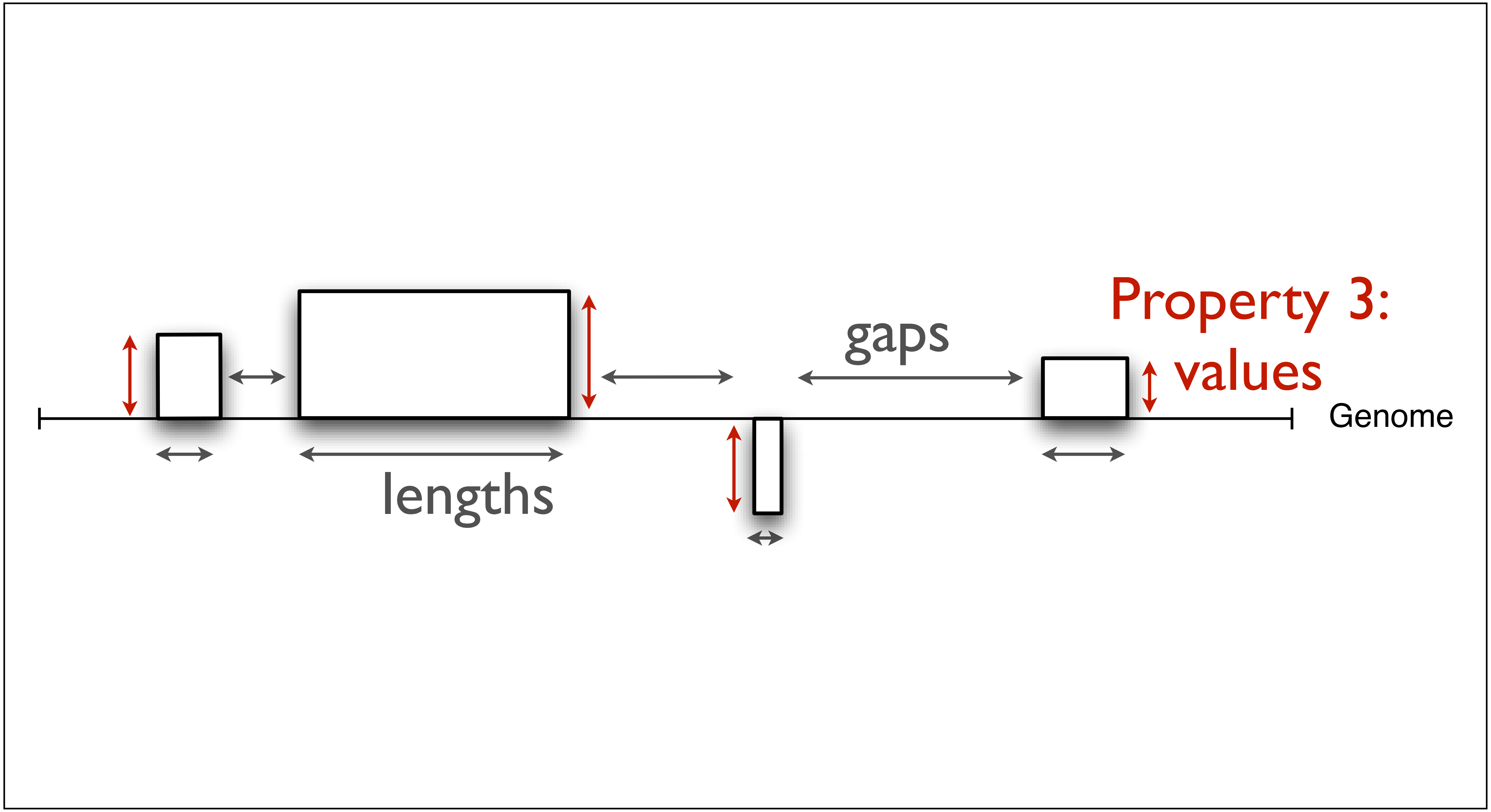
Core properties of tracks



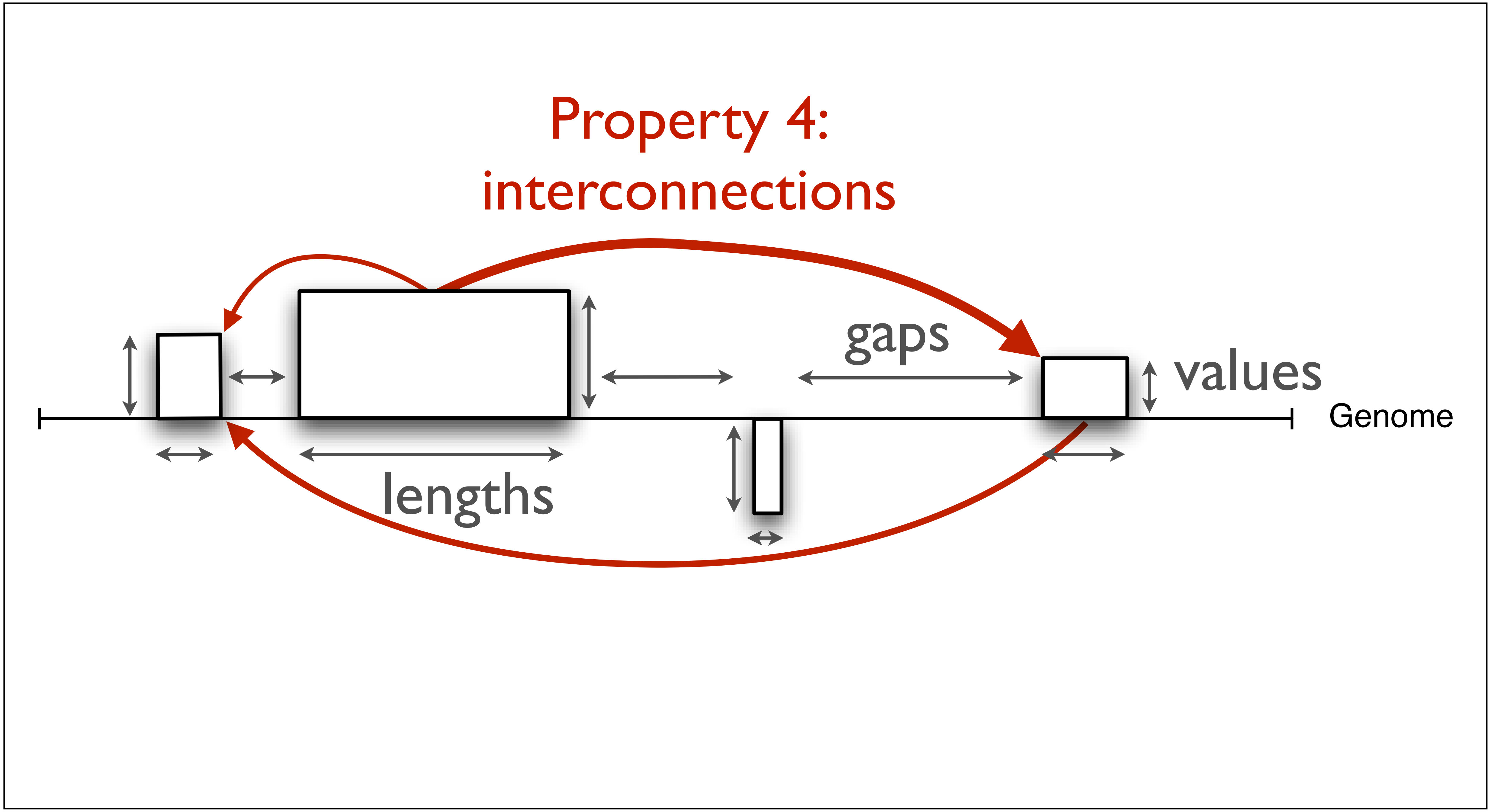
Core properties of tracks



Core properties of tracks

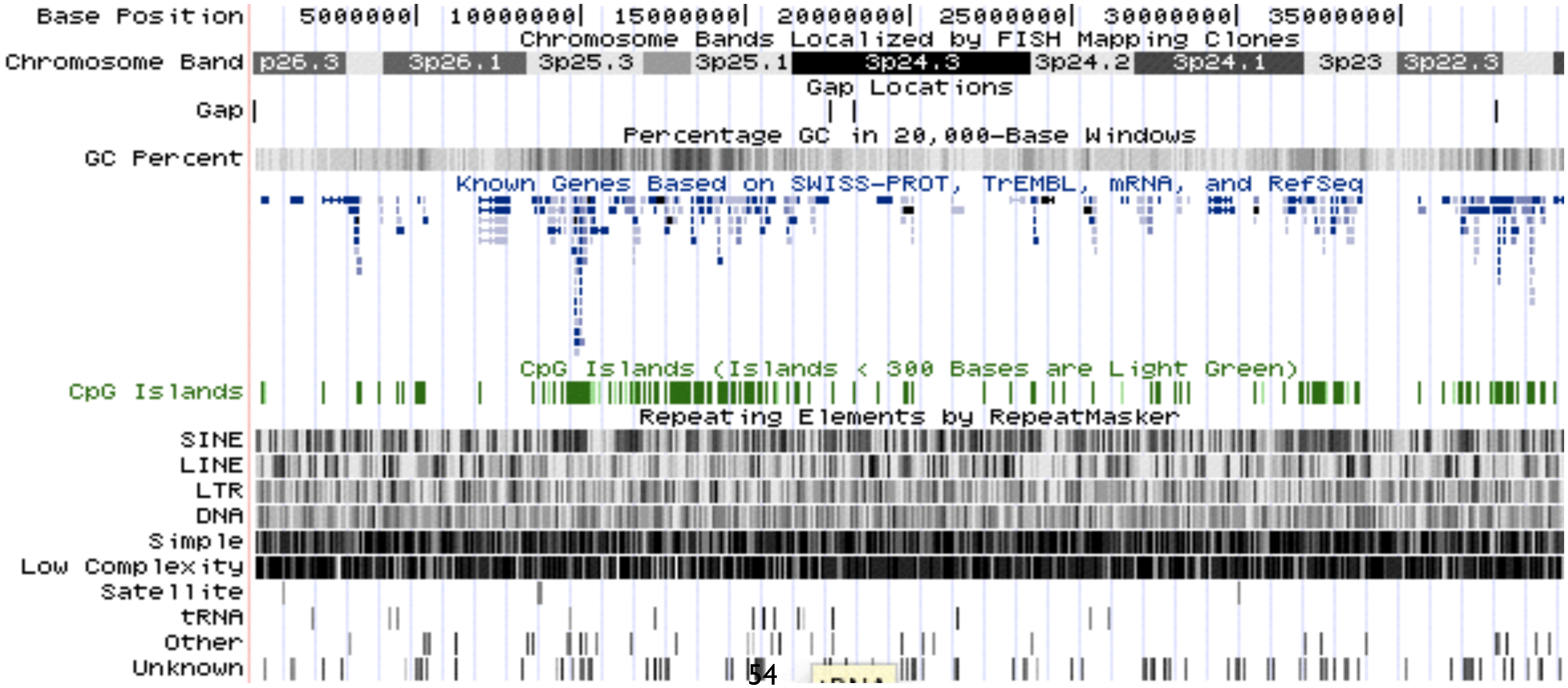


Core properties of tracks



Tracks in the real world

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



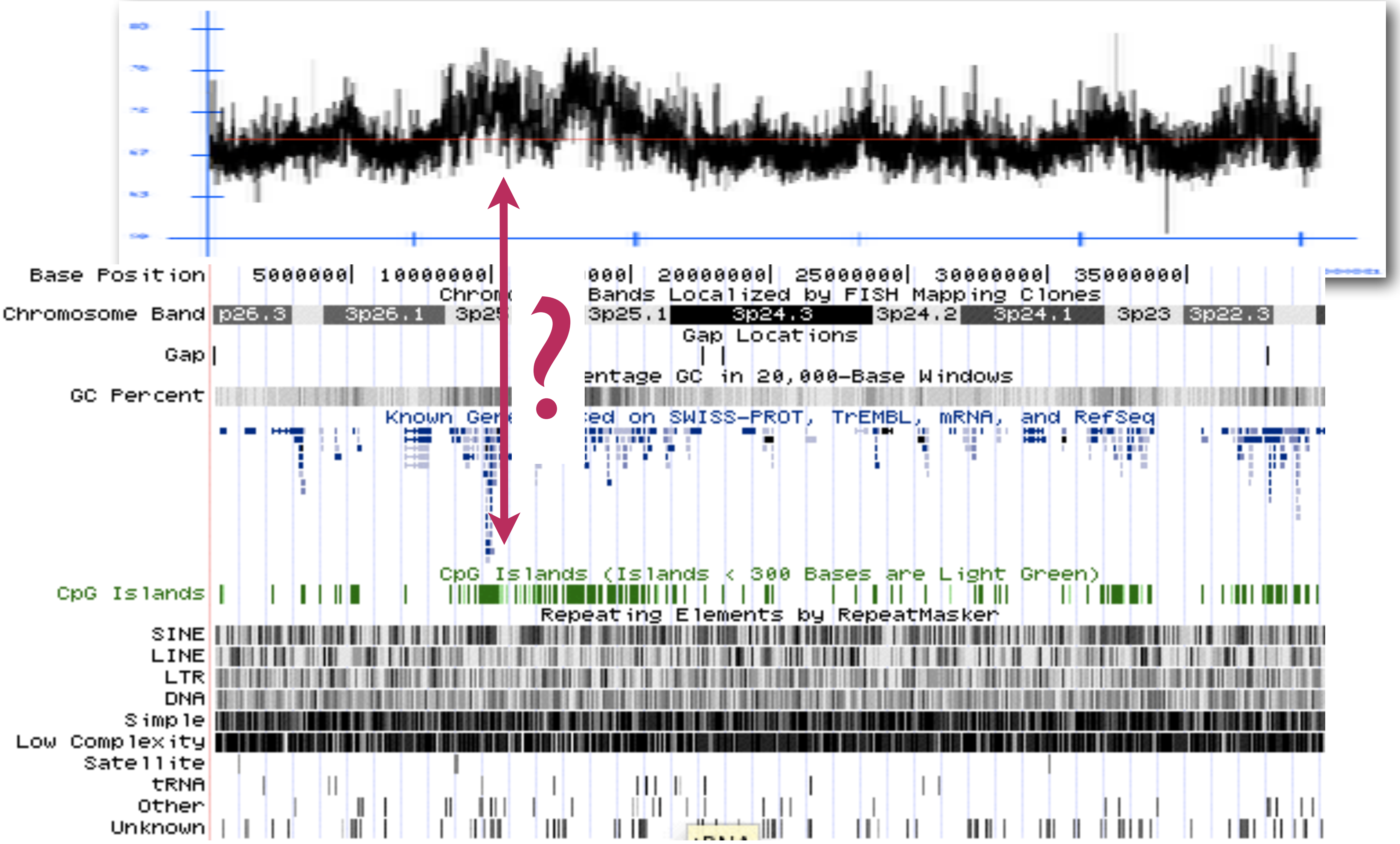
So, what about analysis?

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

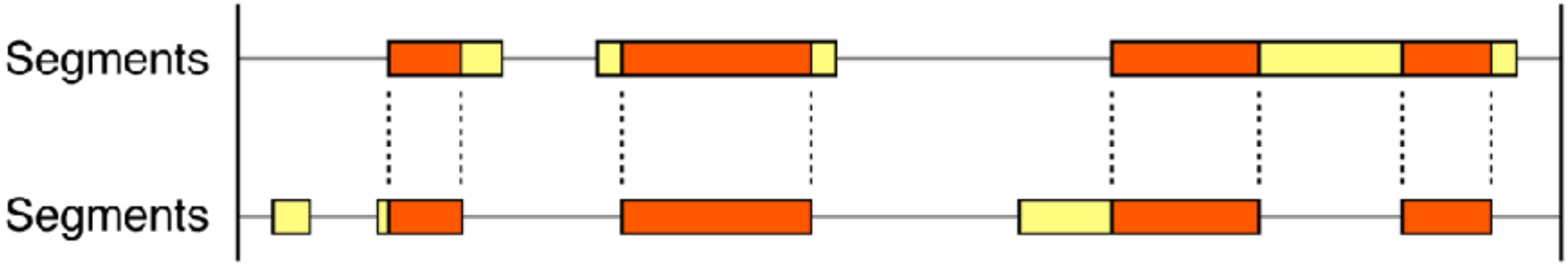
- Typical question:

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?

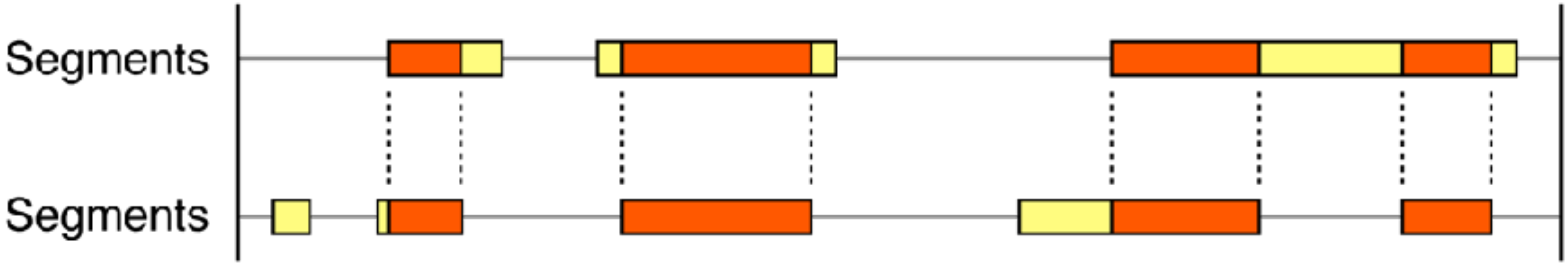
How does this look at the whiteboard?



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

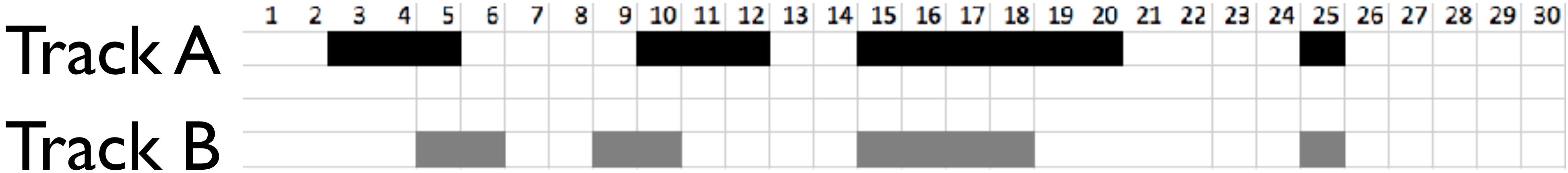
How does this look at the whiteboard?



overlap > expected?

What now?

Exercise 5



Calculate:

a. the number of overlapping base-pairs between tracks A and B — 7

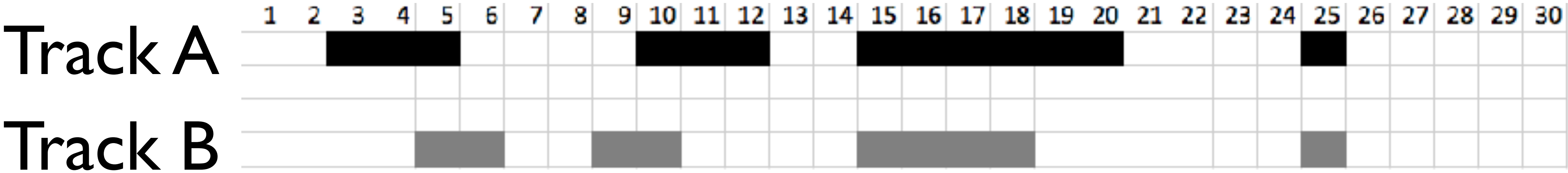
b. the proportion of overlapping base-pairs (in respect to the genome) — 23.3%

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)

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 it's finite set of parameters (e.g. Normal distribution with given mean and variance).
- Non-parametric - the set of parameters is not finite, it depends of 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)

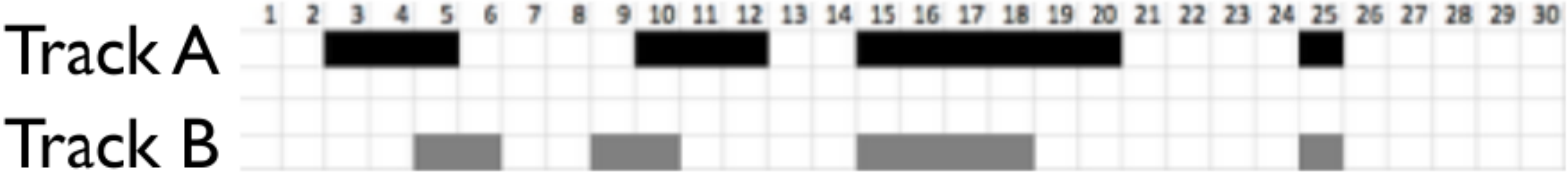
- Significance level **α** - the cut-off under which the p-value is considered significant (often 0.05 or 0.01)

- If **p** < **α**, 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?



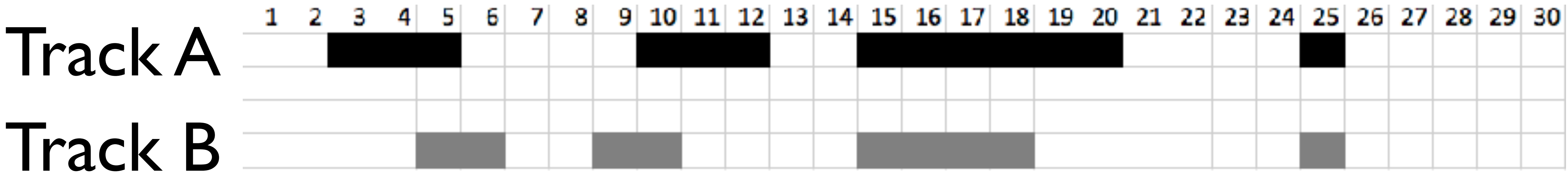
Null models

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

Exercise 6b



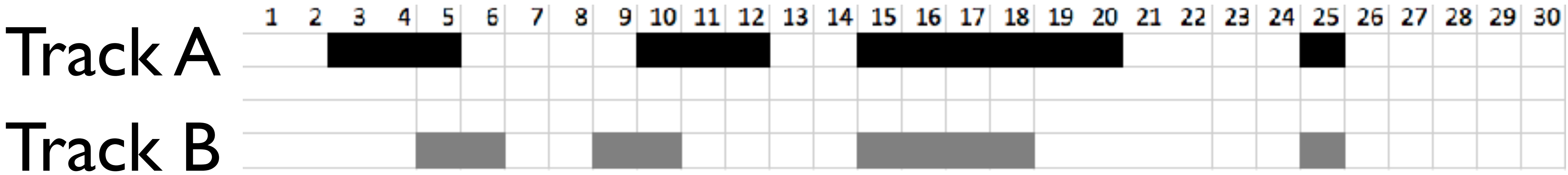
Create a random control track for track B, by

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?

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 H0 when it is true
- Type II error: Not rejecting H0 when it is false

	H_0 accepted	H_0 rejected
H_0 true	TN	FD
H_0 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

$$1 - P(\text{no false discoveries}) =$$

$$1-(1-\alpha)^{10}=1-(1-0.05)^{10}=0.4$$

	H_0 accepted	H_0 rejected	Total
H_0 true	TN	FD	T0
H_0 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 $P_i^{adj}=\min(m*Pi,1)$
- 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

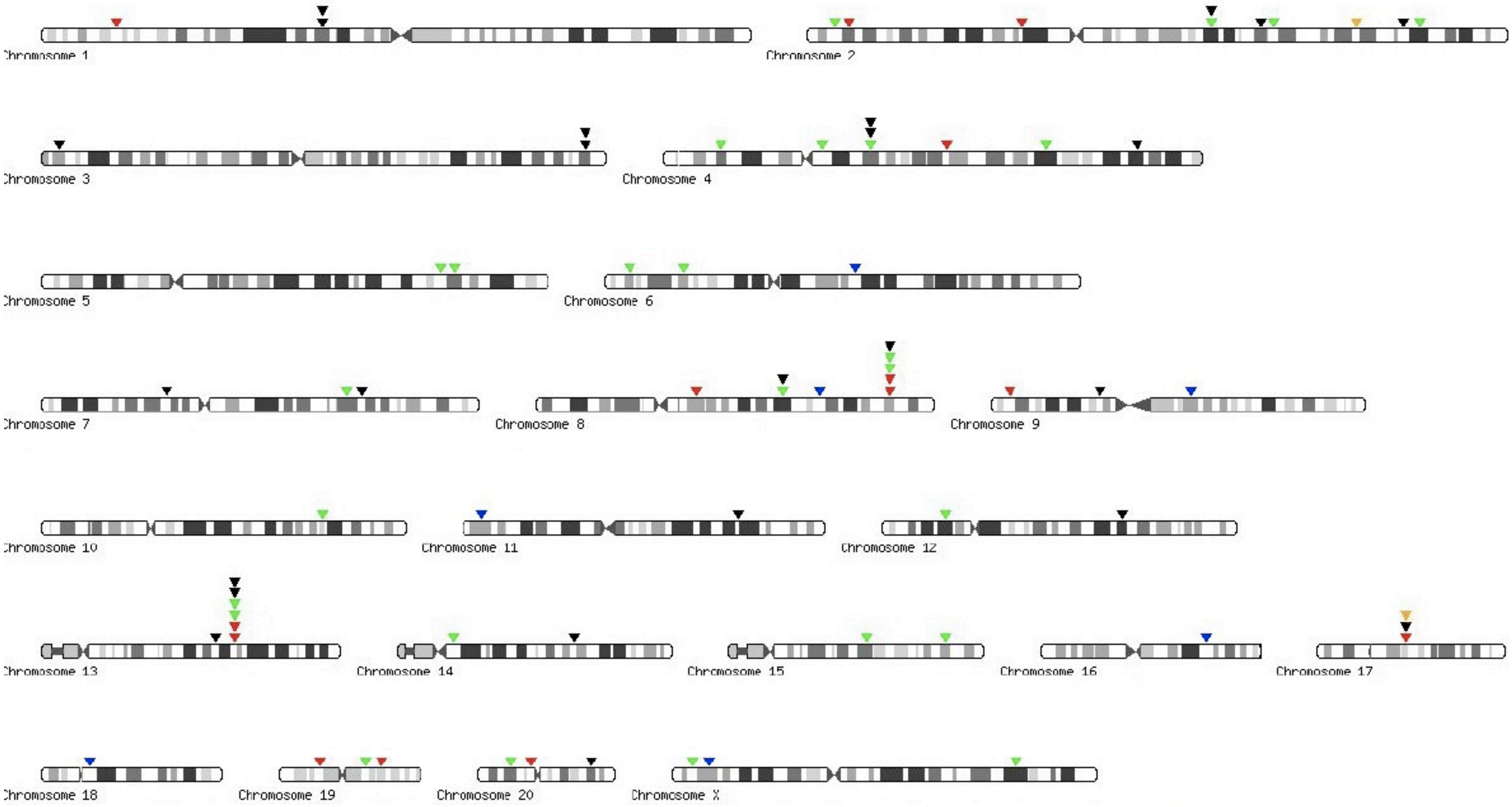
- Controls the expected proportion of false discoveries

1. Select Q, the false discovery rate (e.g 0.1)
2. Sort the original p-values p_1, p_2, p_3…
3. Compare each p_i to it's 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

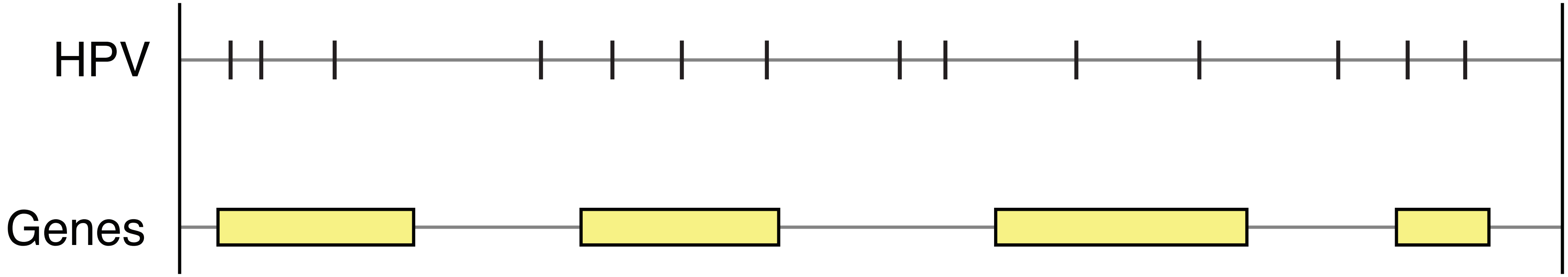
"Viruses might be expected to integrate near genes. Our results confirm such preferential localization inside genes for HPV [Human Papillomavirus], where 75 out of 119 determined integration sites (attached) fall inside genes.”

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?

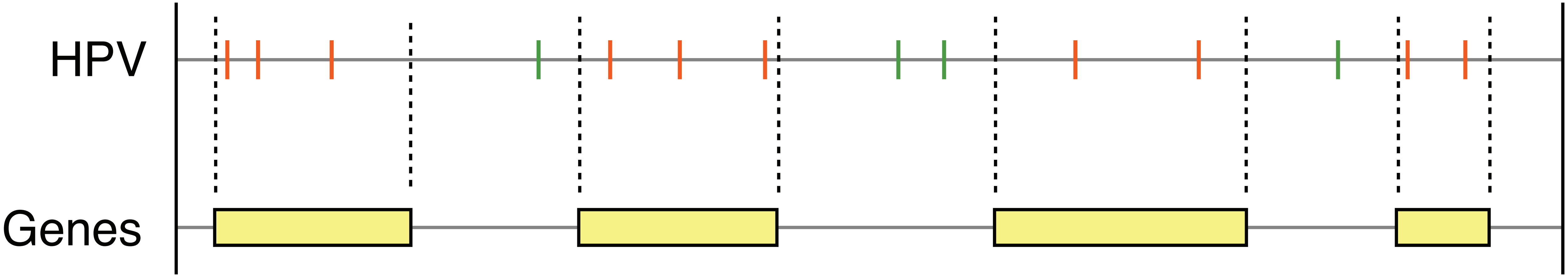
Exercise 7: HPV and genes

Student answers:

1. 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	

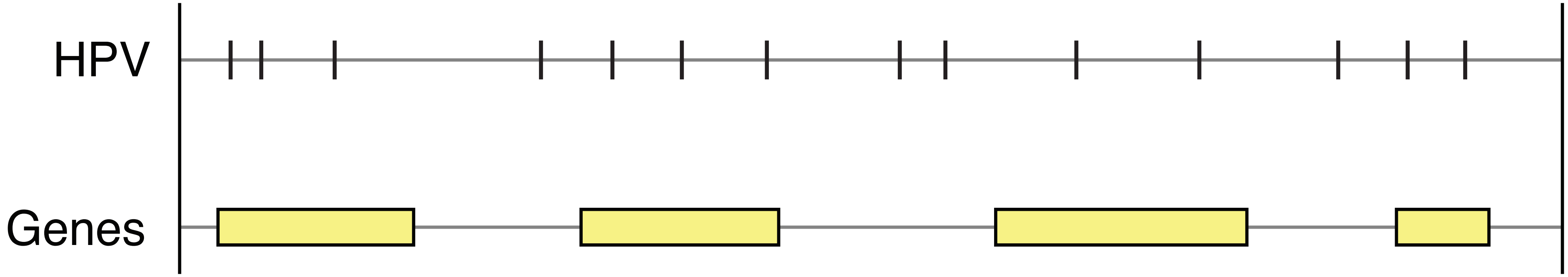
A possible test statistic



- 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

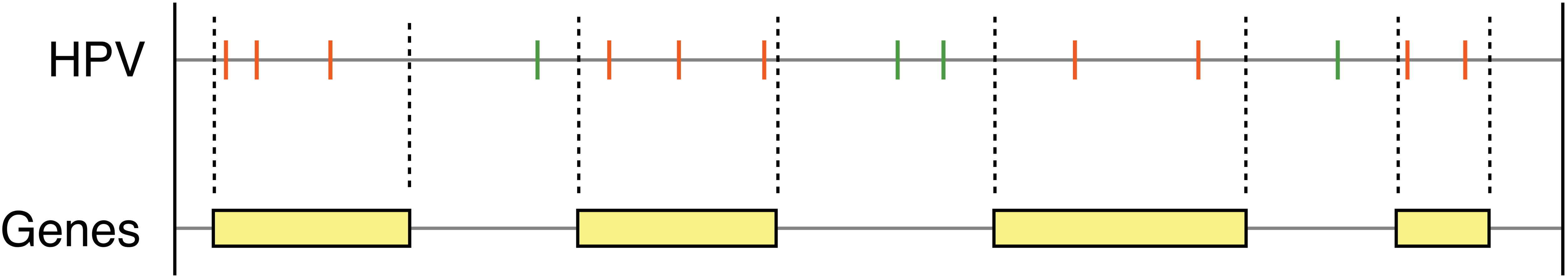
Exercise 8: HPV and genes

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?

Any questions?

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