Statistical (epi)genomics

IN-BIOS9000

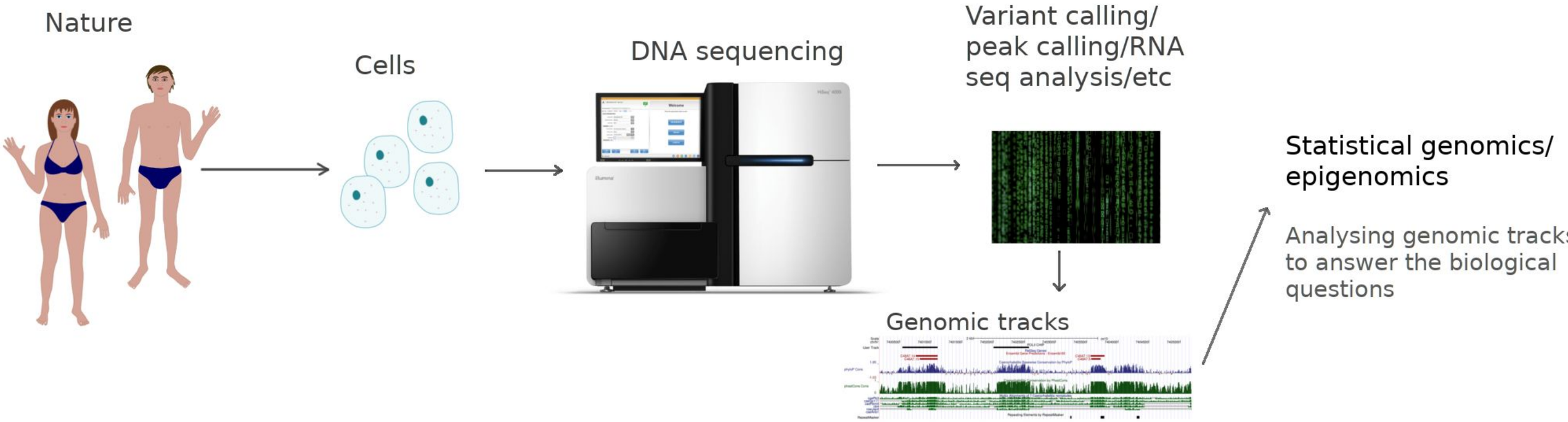
November 1 and 2, 2018

Ivar Grytten

Biomedical Informatics Research Group

Department of Informatics, UiO

What is statistical genomics/epigenomics?



Using **statistics** to answer biological questions by investigating the relationship between **genomic and epigenomic data sets**

What type of statistics will we use and what kind of data?

- Quite simple statistics, such as computing number of base pairs covered by two datasets
- Basic hypothesis testing
- We will use different kinds of genomic and epigenomic data, such as position of genes, open chromatin, transcription factors binding sites, genomic variants etc.

Learning outcomes

1. Know the principles required to do analysis of genomic/epigenomic datasets, including hypothesis testing and Monte Carlo simulation
2. Be able to analyse the relationship between genomic and epigenomic dataset (in order to answer biological questions)
3. Be able to make reasoned choices about null models, test statistics, parameter choices and other important details when doing such analyses, and know how these choices might affect results.
4. Descriptive statistics/investigating data sets
5. Reproducibility. Why it is important and simple ways to improve reproducibility in bioinformatics.

Format of these sessions

- Briefly introduce a topic
- Short exercise
- Topic is explained more in detail
- ... we repeat this for a sequence of increasingly advanced/detailed topics

Approximate schedule day 1

09:00-10:45 Introduction and methodology (reference genomes, tracks, basic analysis)

11:00 - 11:30 Hypothesis testing and introduction to Monte Carlo Simulation

11:30-12:30 Lunch

12:30-13:45 Doing a real analysis (investigating HPV integration sites and genes)

14:00-14:30 Descriptive statistics using BEDtools/Hyperbrowser

14:40-15:20 More statistics, deeper into null models and test statistics

Approximate schedule day 2

09:00-09:15 Summary of day 1

09:15-11:30 Analysing more than two tracks (investigating Multiple Sclerosis)

11:30-12:00 Lunch

12:00-13:00 Reproducibility

13:00-14:00 Summary of everything, questions, quiz

14:30-15:00 Home exams handed out

Tools we will use

- Bedtools
  - Easy-to-use command line tool for simple analysis and operations on genomic data sets

- The Genomic HyperBrowser
  - Web tool for statistical analysis of genomic data sets

We will mostly focus on the concepts, not too much on the tools. The concepts are the same if you use other tools.

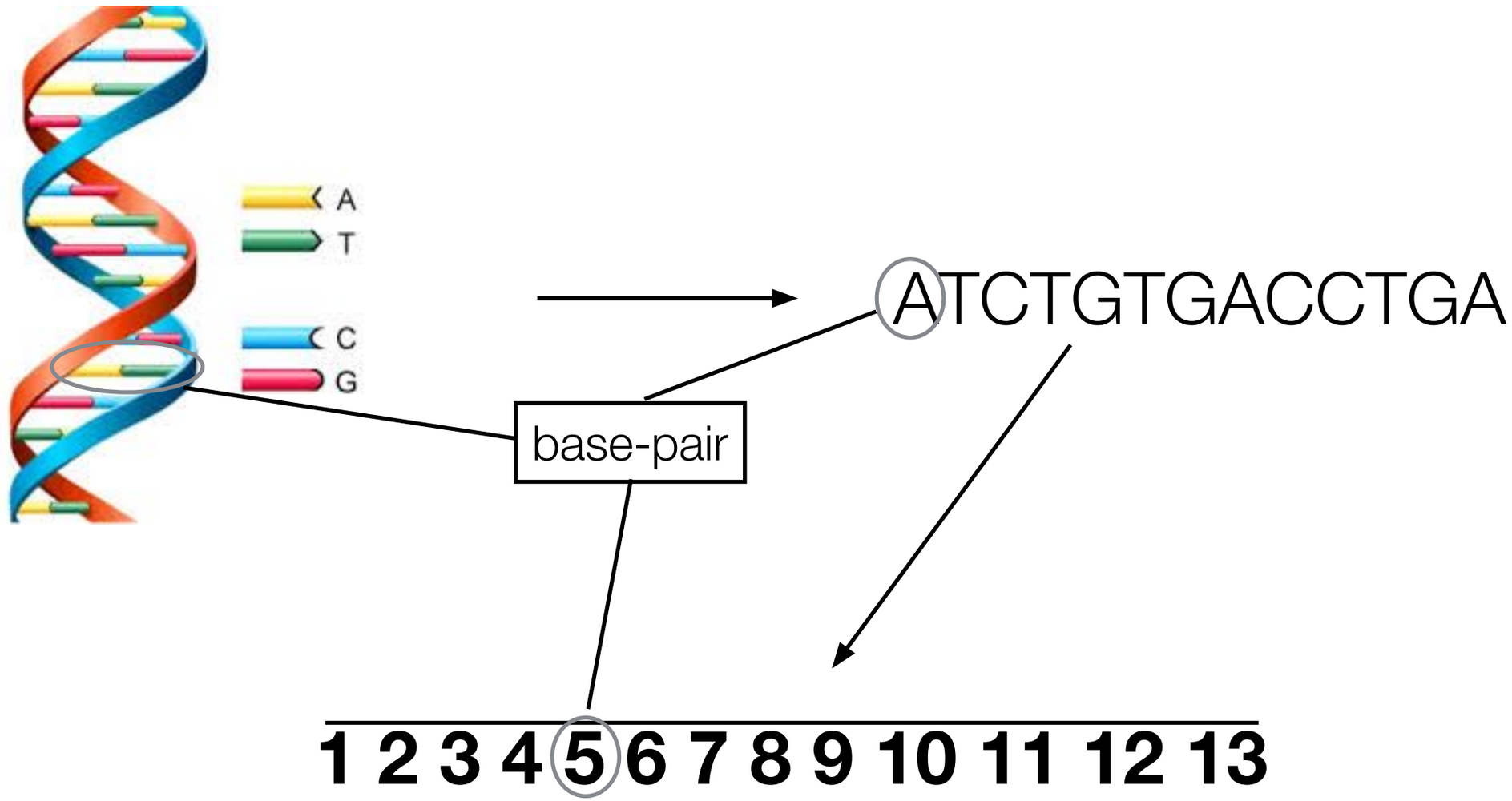
Models and data representation

Reference genomes, genomic tracks

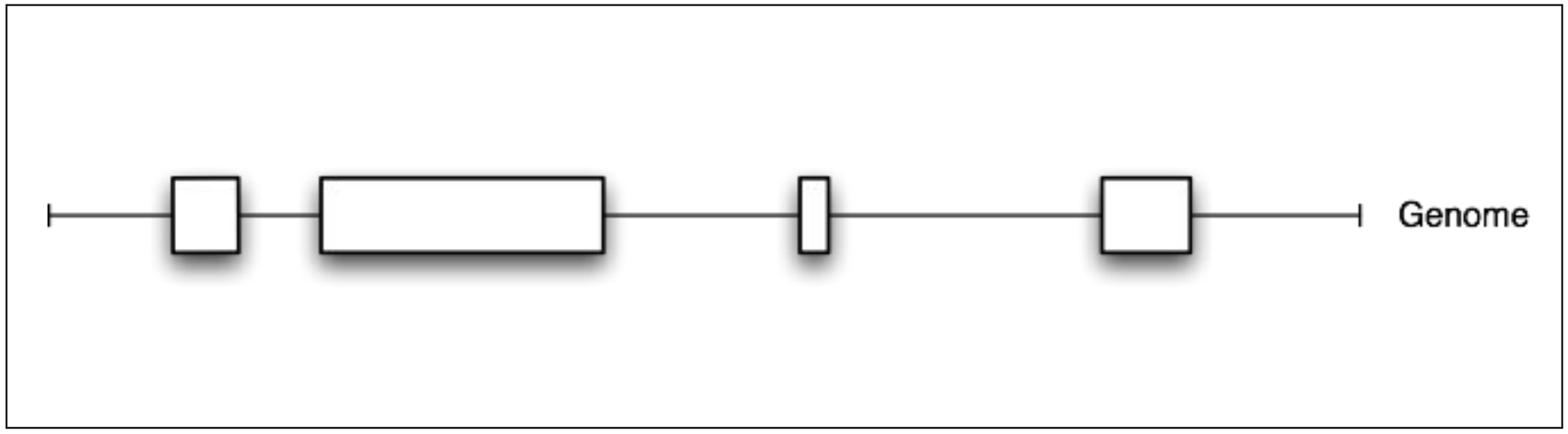
What are reference genomes?

What are reference genomes?

Genome as a line (coordinate system)



How to represent genes on reference genomes?



How to represent genes on reference genomes?

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

Why represent genes and other genomic features as elements on a reference genome?

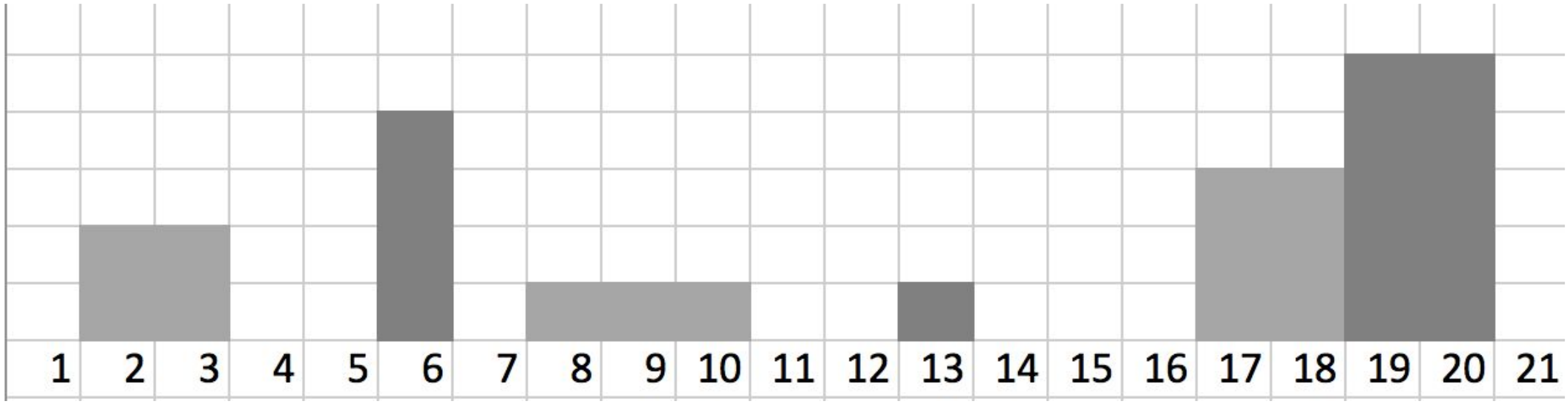
- Main reason: It makes it easy to compare different datasets (e.g., whether they have elements on the same position)
- It is a simple and compact way of representing elements.
  - You only need the reference genome coordinates, and then you know the sequence within the elements.
- Makes it simple to compute the distance between genomic elements

Tracks and track types

Representing genomic elements on the reference genome

- Genomic tracks are genomic datasets represented on a reference genome
- As there are many different types of genomic datasets, there are many different track types

Exercise 1

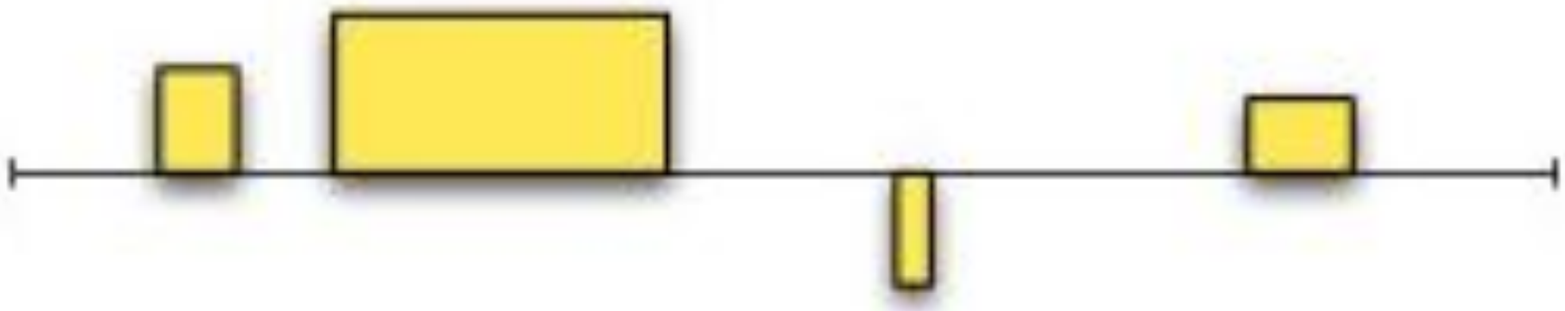


a) Base-pairs covered (coverage) 11
b) Proportion of genome covered 0.52
c) Average segment length 1.83
d) Average gap length 1.43
e) Average value 1.33 per bp
2.54 per bp (only segments)
2.67 per segment

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 (found by doing RNA-seq)?

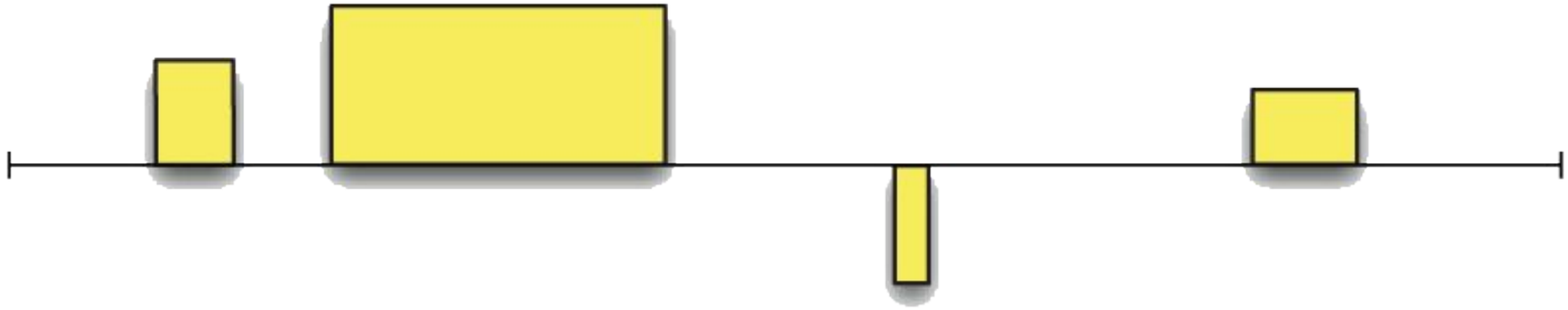


chr7	127471196	127472363	17
chr7	127472388	127473530	31
chr7	127473555	127474697	73
chr7	127474701	127475864	13
chr7	127475893	127477031	83
chr7	127477121	127478198	93
chr7	127478300	127479365	29
chr7	127479375	127480532	59
chr7	127480538	127481699	63

Track types

- In the last example, we showed genes as segments on the genome line, with attached RNA-seq read count values

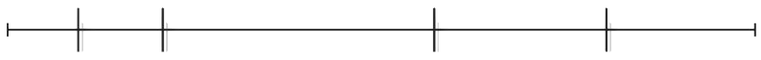
- This track is of a **track type** we call “valued segments”



Valued Segments (VS)

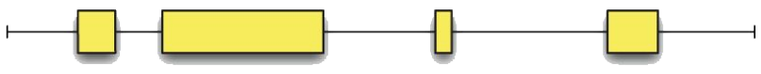
- Track types are models used to categorize tracks according to their main characteristics

Track types

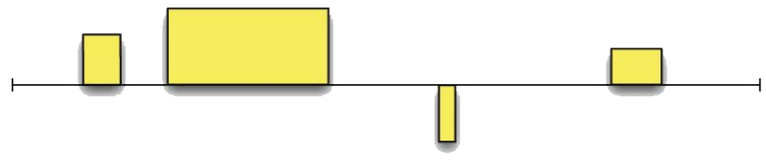


Points (P)

Valued Points (VP)

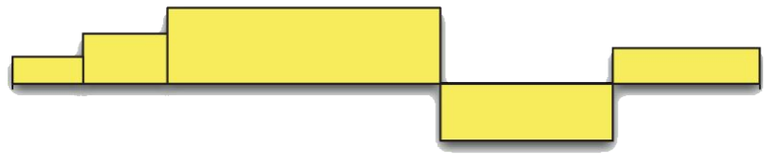


Segments (S)

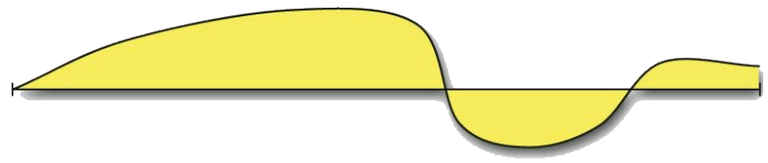


Valued Segments (VS)

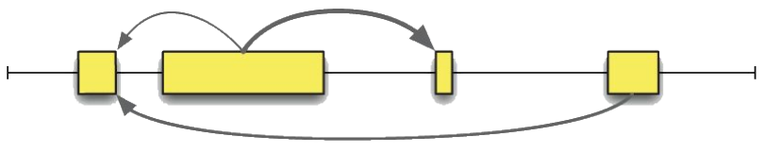
Genome Partition (GP)



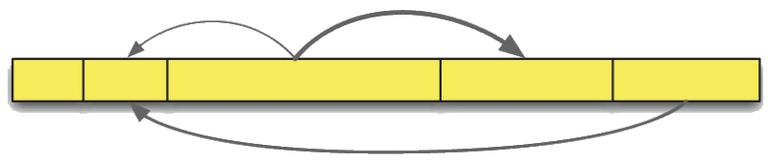
Step Function (SF)



Function (F)



Linked Segments (LS)



Linked Genome Partition (LGP)

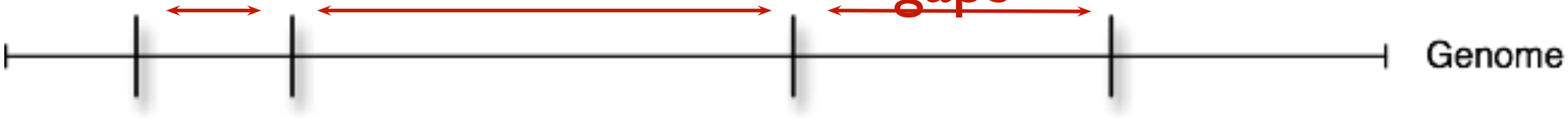
Typical genomic tracks

- Open chromatin (represented as segments)
- SNPs (represented as points, or valued points, e.g. with frequencies)
- Transcription factor binding sites (“peaks”) (represented as segments)
- Genes (represented as segments, valued segments with expression as values)
- etc...

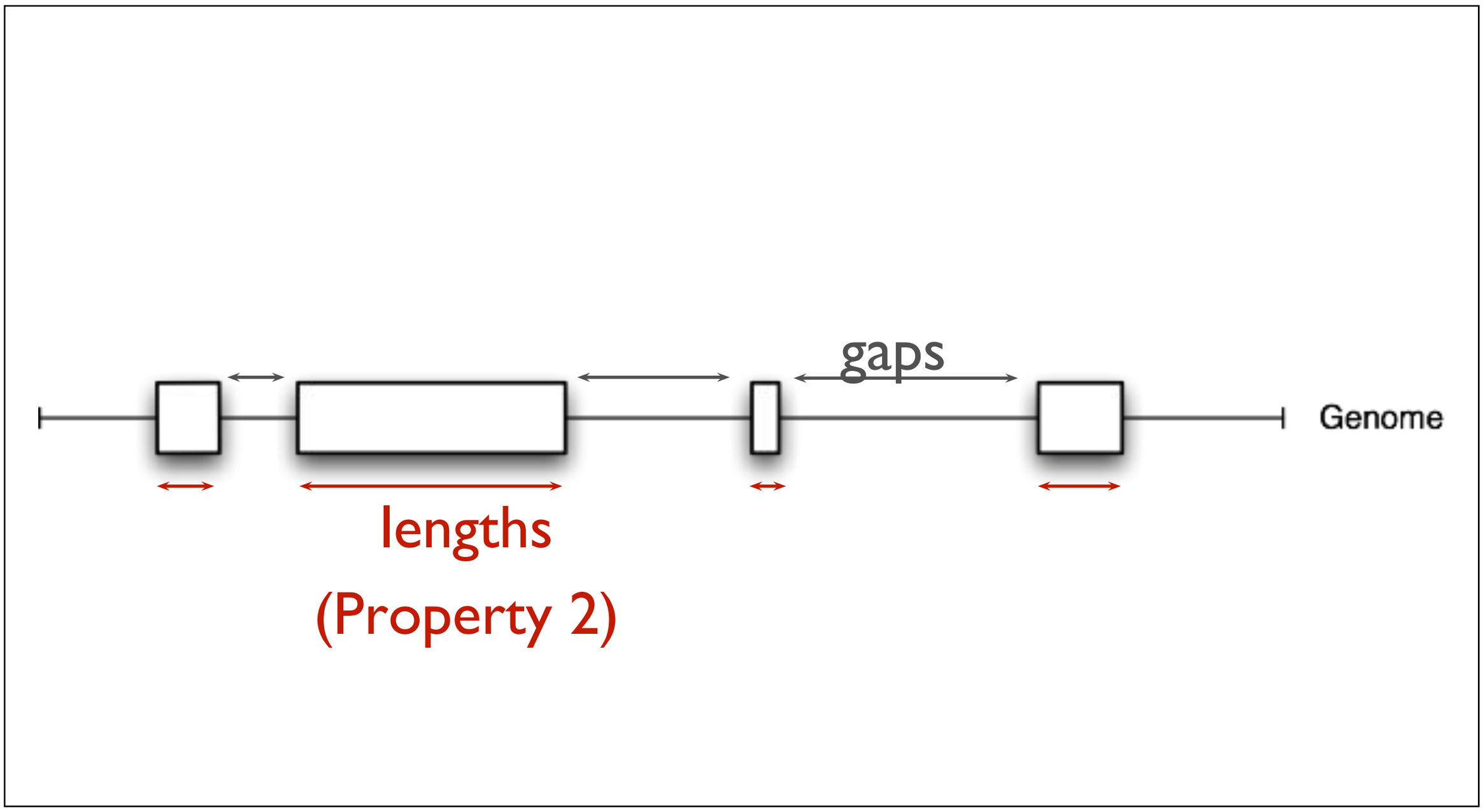
Core properties of tracks

Property 1:

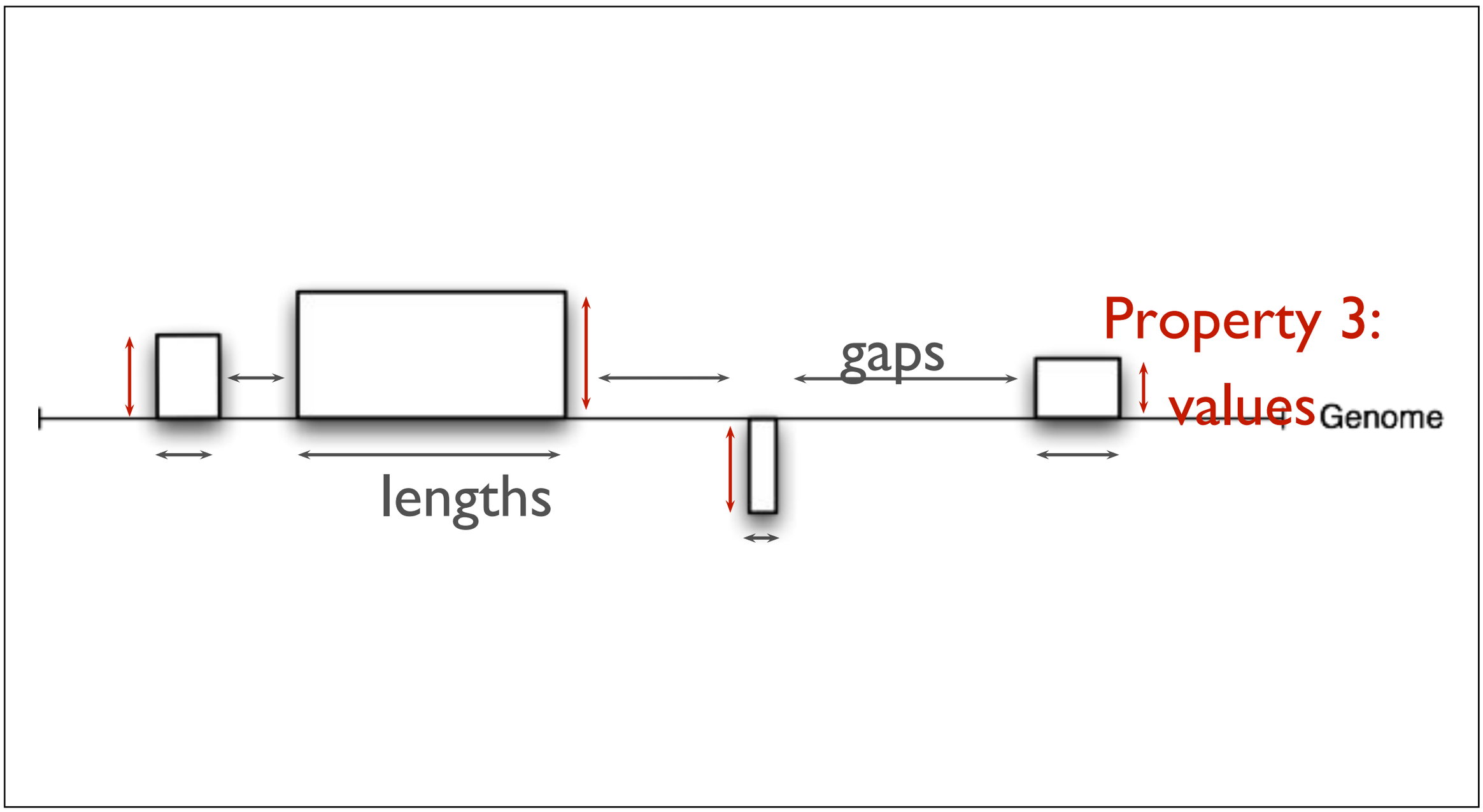
gaps



Core properties of tracks



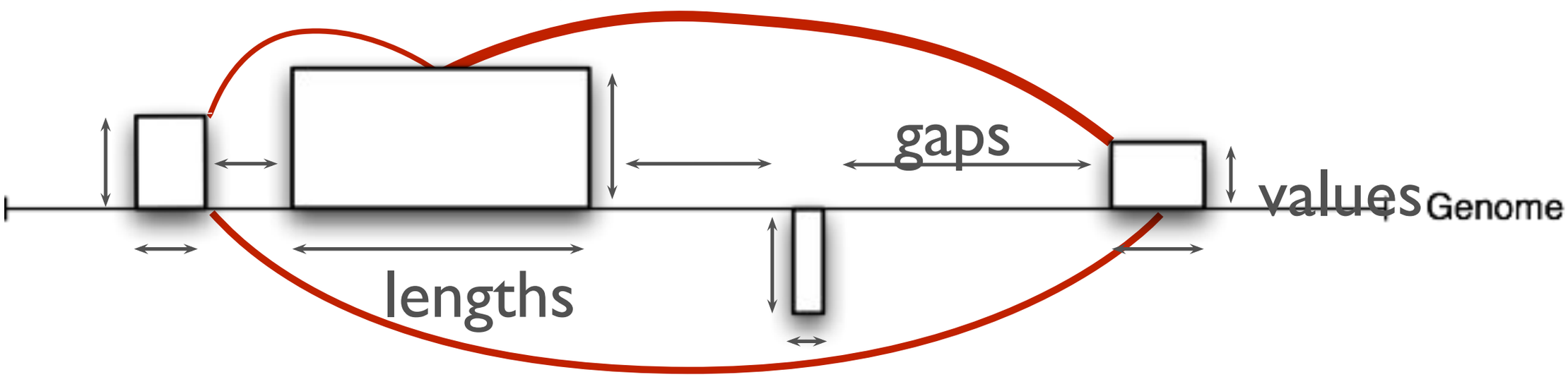
Core properties of tracks



Core properties of tracks

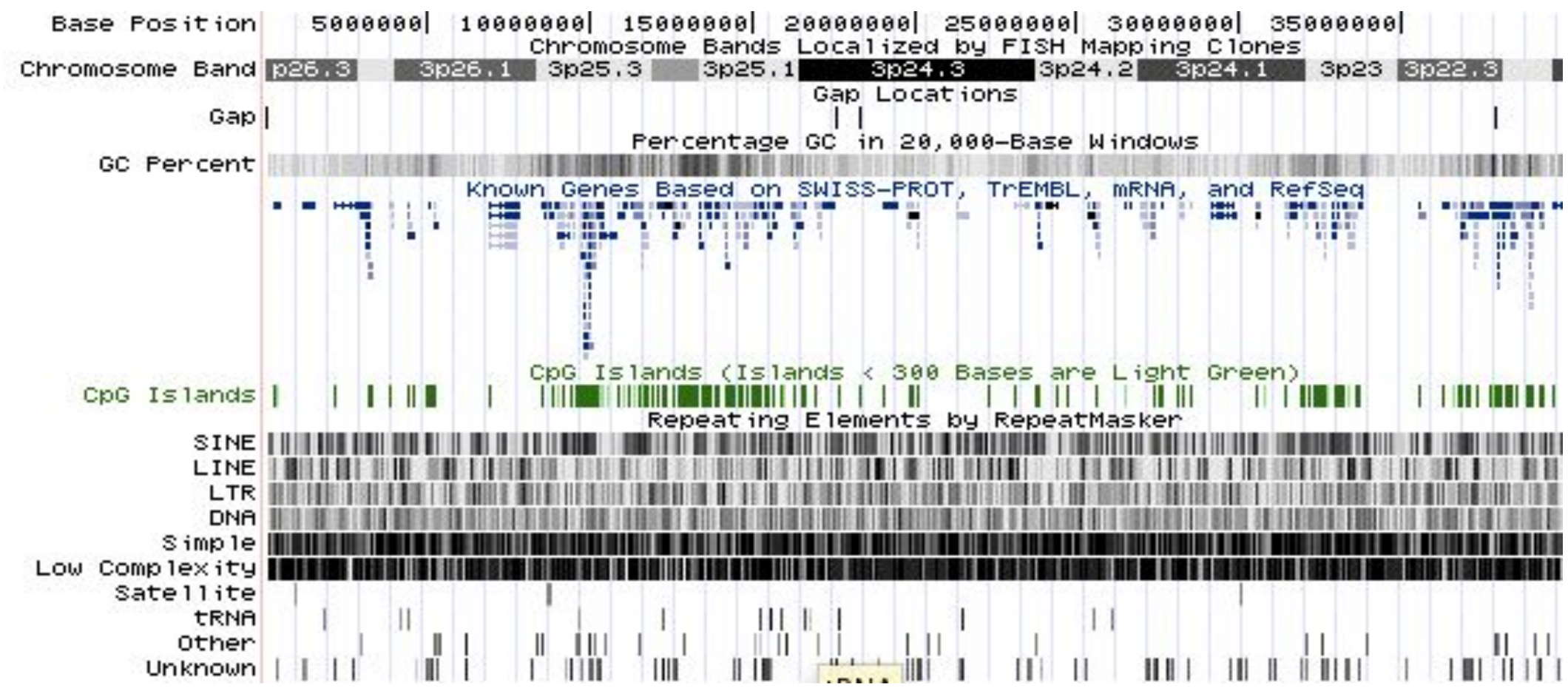
Property 4:

interconnections



Tracks in the real world

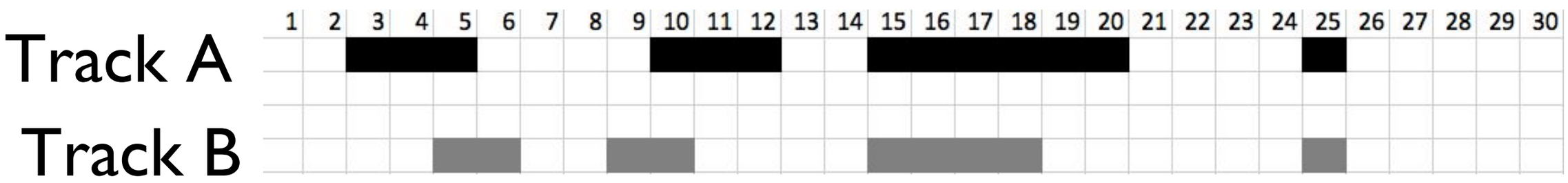
- UCSC Genome Browser
- Each row is a track, and many of the track types are supported



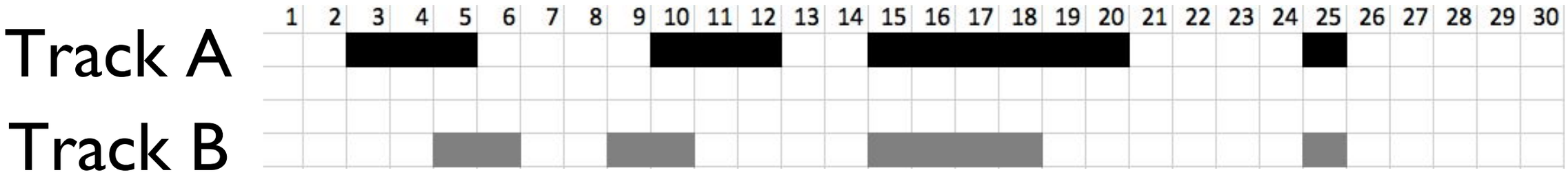
Hands on exercise with two tracks

- A researcher suspects that two transcription factors (TFs) tend to bind to the same locations in the genome

- Thus, he performs a ChIP-seq experiment for the two TFs, performs peak calling and ends up with two **tracks** on a reference genome (saying where the transcription factors bind):



Exercise 2a



Calculate:

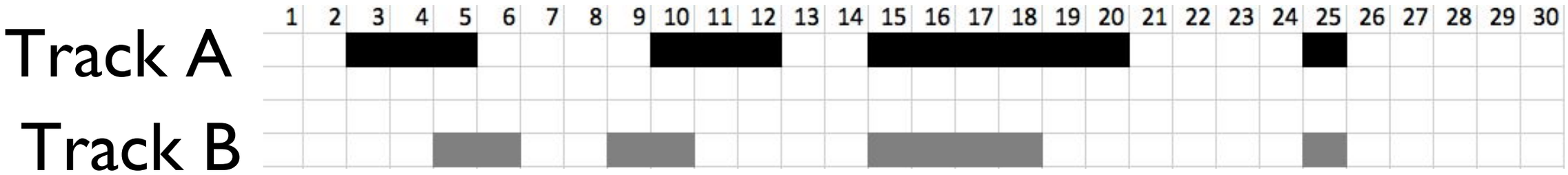
a. the number of overlapping base-pairs between tracks A and B (base pairs covered by both tracks) — 7

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

What conclusions can we draw from the results?

- **23** % of the genome is covered by both tracks
- How to know if that number is high or low?
  - Could we compute the number for other pairs of such data sets and compare?
  - How many base pairs would be covered by the two data sets if they both covered random base pairs?

Exercise 2b



Calculate:

a. the number of overlapping base-pairs between tracks A and B (base pairs covered by both tracks) — 7

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

Try to find:

c. the expected number of overlapping base-pairs (how many base pairs would be covered by both tracks if they covered random base pairs) — 3.9

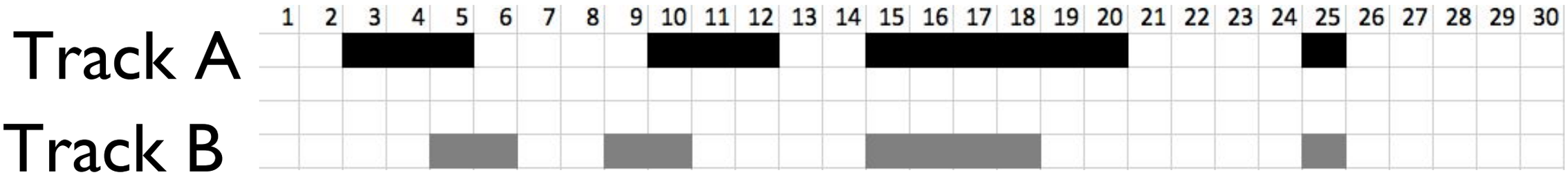
d. the proportion of observed to expected overlap — 1.8

What we did in the exercise

- In order to assess whether 7 base pairs overlap was much, we compared this number to the expected overlap (what the overlap would be on average if the two tracks were random).
- But how do we know that 1.8 times the expected overlap is high (could this happen easily by chance)?

- Let's investigate this!

Exercise 3a



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) Compute the overlap between your random track and Track A

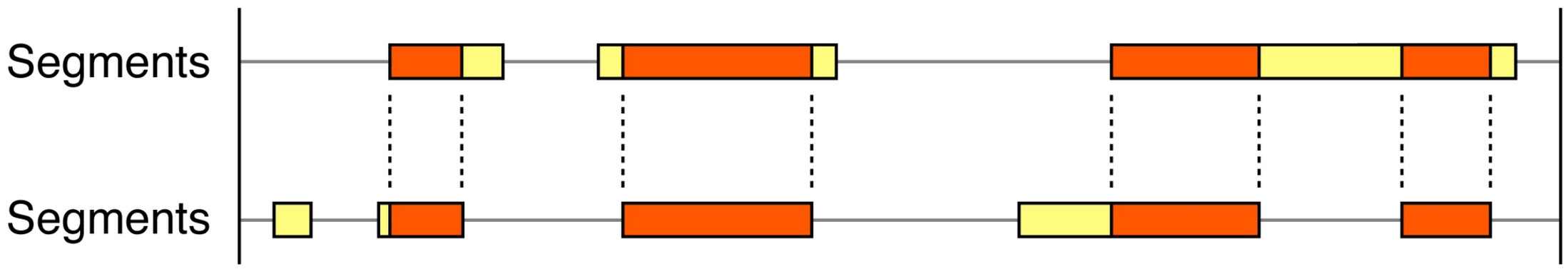
(You can also play around with this tool: http://46.101.93.163/monte_carlo/)

What conclusions can we draw from this histogram?

- We drew a histogram of all overlaps with random track on the blackboard
- We can compare the overlap between A and B with all the overlaps between A and random track (which the histogram shows)
- This shows us whether the overlap between A and B is extreme or typical.
- We will come back to this, but first some theory...

Typical analysis questions

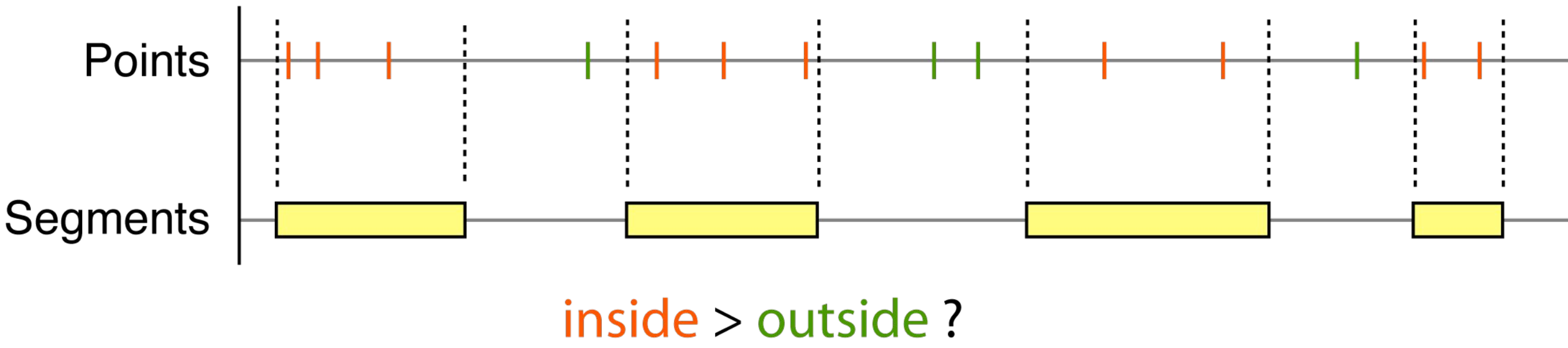
Do genomic feature X and Y overlap more than expected by chance?



overlap > expected?

Typical analysis questions

Do genomic feature X (points) fall inside Y more than expected by chance?



Typical analysis questions

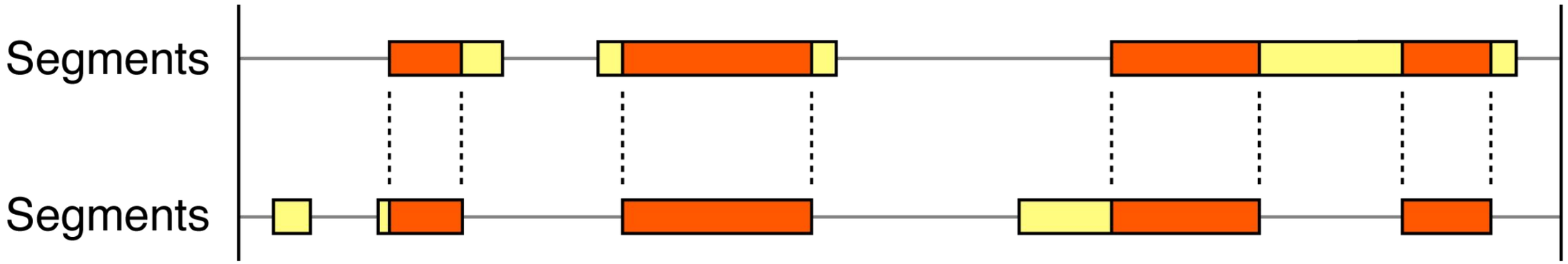
Are points in track X closer to elements in track Y than expected by chance?

Typical analysis questions

- This is the main class of questions we will be working with (with some variations):

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

Co-occurrence of genomic features



overlap > expected?

What can such analyses be used for?

- Discover novel relations between tracks:
  - 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?

Some examples of this type of analysis

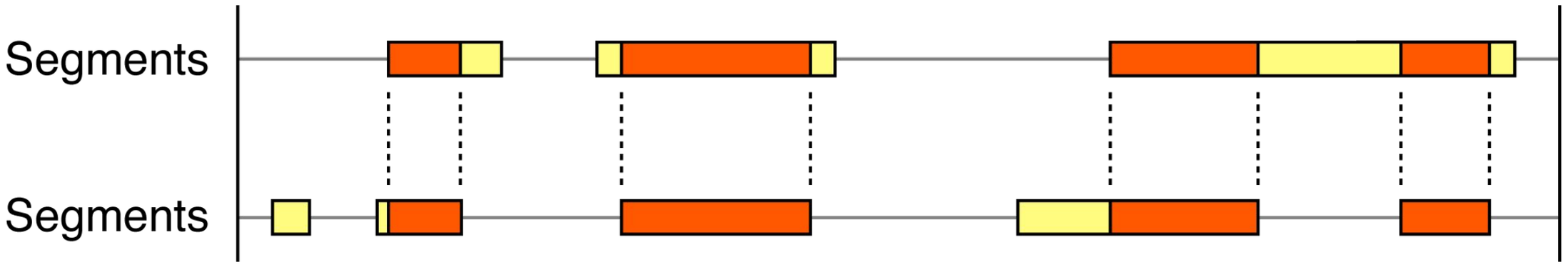
- Age-associated hyper-methylated regions in the human brain overlap with bivalent chromatin domains (Watson et al. 2012)
- Genomic regions associated with multiple sclerosis are active in B cells (Disanto et al. 2012)
- DNase hypersensitive sites and association with multiple sclerosis (Sandve et al.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

Some examples of this type of analysis

- 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)

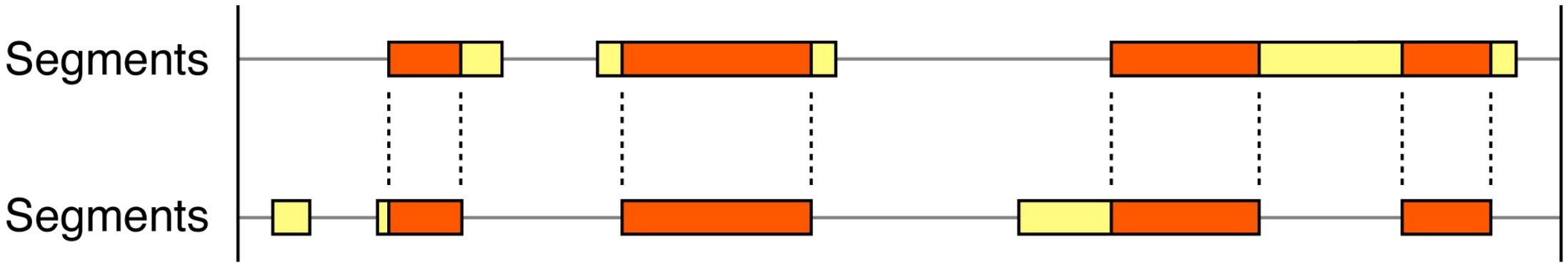
How does this look at the whiteboard?



overlap > expected?

- This analysis only 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?

How to assess whether overlap > expected?

Hypothesis testing

Example

Someone claims that they are able to taste whether **tea** or **milk** was added to a cup first.

You want to test whether they are able to taste the difference or not.

Example

We create an experiment. We give a person 4 cups (you toss a coin for every cup to decide whether to add milk first or after)

Example

- Assume the person is correct in 4 out of the 4 cups. How can we assess whether that person is able to tell the difference or not?
- If she is guessing, what is the probability of getting 4 out of 4 correct?
- 0.5^4 = 6.25%
- The probality of guessing correct 4 times is quite low, so we might believe here

Example

- We want to be more certain, so we give her 50 cups in a row (each time we throw a coin to decide whether to have milk in first or not)
- She guesses correct 34 of the times
- How certain are we that she can tell the difference?
- Probability of having 34 correct out of 50 by blindly guessing is (by binomial distribution) only **0.8%**
- We can quite confidently conclude that she is able to tell the difference

The example as a hypothesis test

- Until otherwise proven, we assume she cannot tell the difference. This is our **null hypothesis**.
- We want to investigate whether an **alternative hypothesis** might be true: She can tell the difference.
- We make some observations (give her tea and let her guess). We compute the probability of the null hypothesis being true based on the observations. This is a **p-value**.
- If the **p-value** is low, we reject the null hypothesis and conclude on the alternative hypothesis.

Example

- Someone claims that gene A is generally more expressed than gene B in the population (more than expected by chance)
- Do an experiment to investigate
- You check 5 people
- What is the probability of the claim being false?

Example

Assume there is no preference for any gene:

- It is “50/50” whether gene A or gene B is the most expressed gene
- You check 5 people, and gene A is always expressed more
- What is the probability that this happened by chance?

More formally

- **Null hypothesis** (H_0) **-** a neutral baseline that can be reasonably assumed to be true:
  She cannot taste the difference

- **Alternative hypothesis** (H_1) **-** the claim you wish to test:
  She can taste the difference

- **Test statistic -** measurement of the observed data that captures the aspect of interest:
  E.*g. number of times she correctly tasted the difference*

- **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

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

- If **p** < **α**, then the null hypothesis is rejected, meaning the evidence supports the alternative hypothesis

Why use hypothesis tests?

- Sometimes hard or impossible to make conclusions without.
  - What if you observed that she was able to taste the difference **540 out of 1000** times?
  - Even harder when working with biological data where numbers are less intuitive

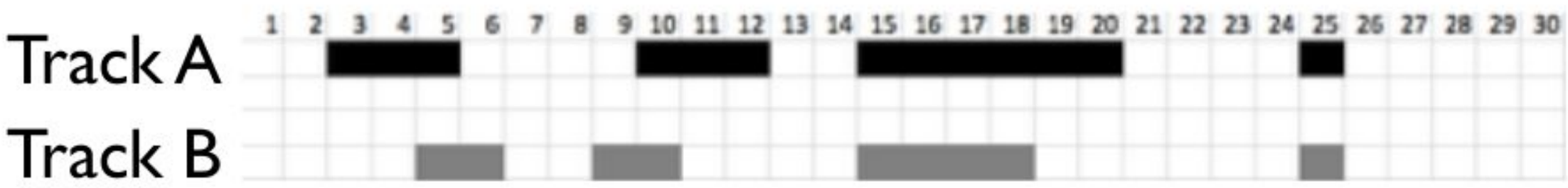
- A hypothesis test quantifies the certainty of concluding a hypothesis (p-value)
  - For some cases, a very small p-value might be requested, e.g when concluding on the effect of a drug

Null model

- A null model is the model in which the null hypothesis arises from
  - The “base case” where we assume the condition in the null hypothesis is true.

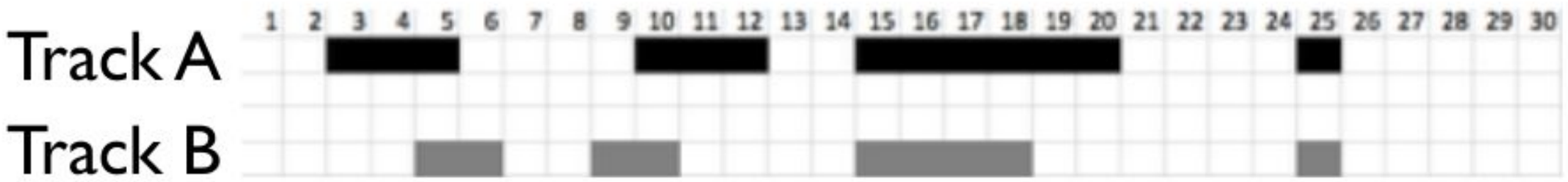
A claim with a less simple null model:

- **Claim**: A genomic track co-occurs (more than expected by chance) with another genomic track



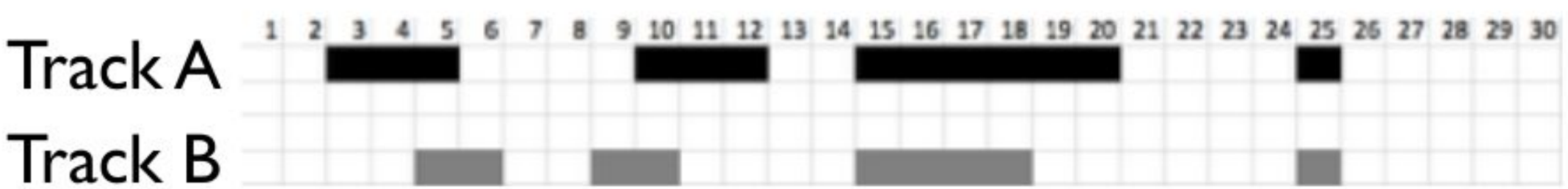
Back to our case

- **Claim:** The two genomic tracks, A and B, co-occur (more than expected by chance)
- What is the null model? (How do we assume these tracks behave when there is no association)



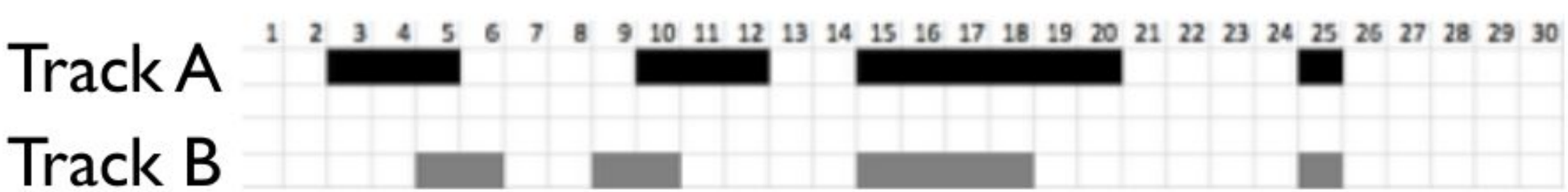
How to make random samples in this case?

- Preservation of structure in data
  - Should be realistic
  - Reflect biological realism
  - Many ways to to this, not trivial
- Then, given a null model, how do we find the p-value?

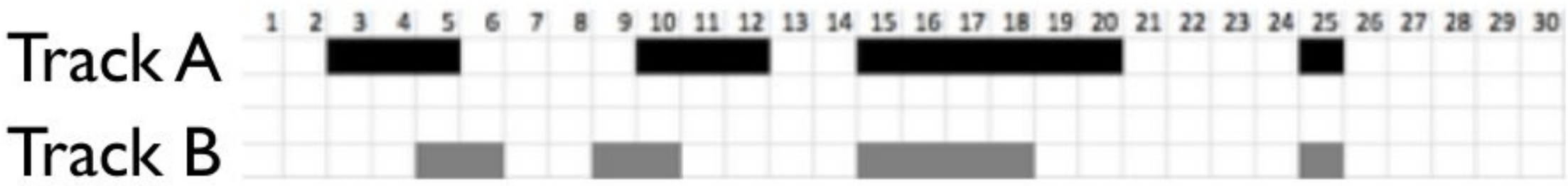


Monte Carlo

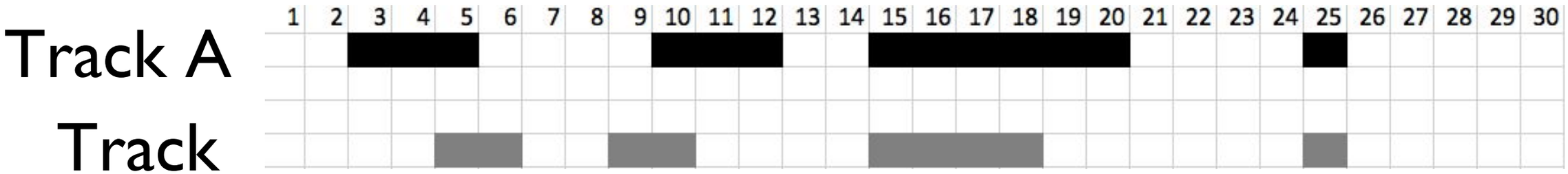
- Simulate many samples from the null model
  - E.g. many pairs of tracks following the same properties
- For each simulation, compute the co-occurrence
  - E.g. the number of base pairs overlap
- Compute how often the co-occurrence found **using the null model** was as extreme or more extreme than the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



- **Examples of preservation strategies**
  - Preserve segment length
  - Preserve segment and gap length
- 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 3b



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 3b

- 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

Analysis part 2

We will investigate this 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

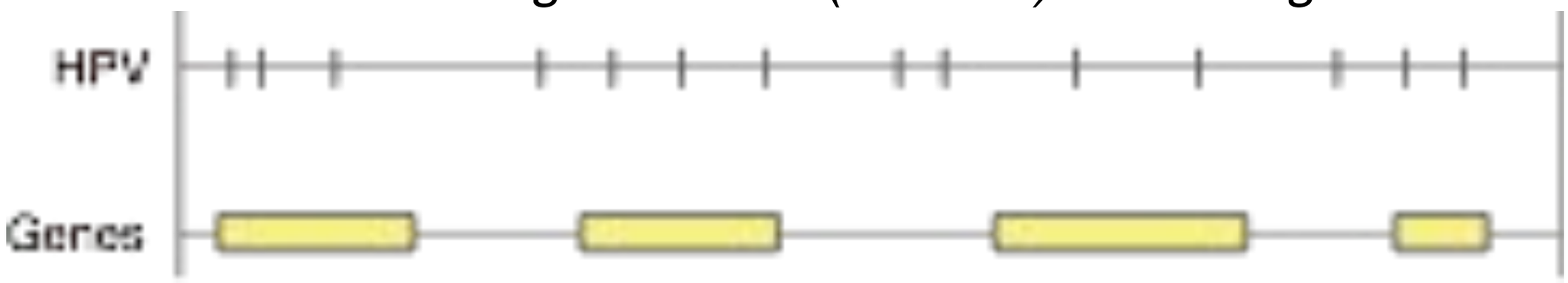
"Viruses might be expected to integrate near genes. Our results confirm such preferential localization inside genes for HPV [Human Papillomavirus], where 75 out of 119 determined integration sites (attached) fall inside genes."

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?

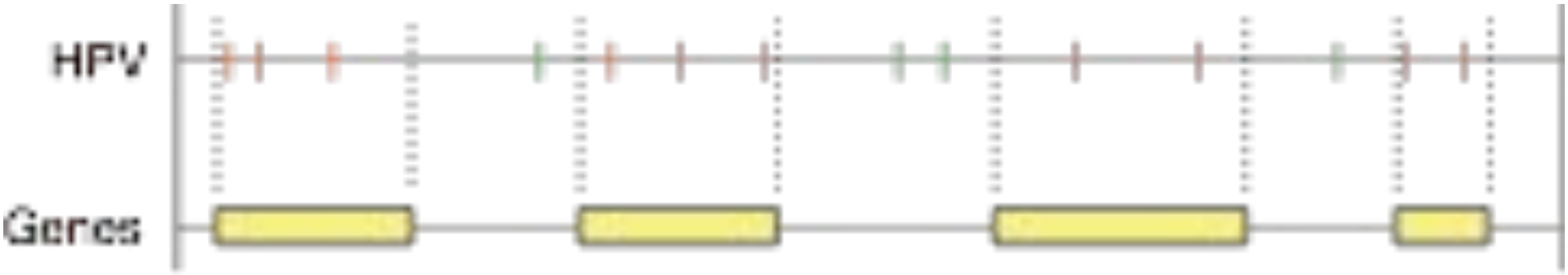
Exercise 7: HPV and genes

Student answers:

1. Which test statistic would you choose?

Ratio of HPV sites falling inside genes		
Average of distances from each HPV site to its closest gene	4	
Take median or average of log of distances instead of average		
Count number of genes with HPV site less than X bp away		
Count number of HPV sites with gene less than X bp away	7	

A possible test statistic



Compute the average distance from each HPV site (track 1) to its nearest gene (track 2)

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?

Some possible null models for segments:

- Preserve segment length
- Preserve segment and gap length

For points:

- Preserve point count
- Preserve inter-point distance

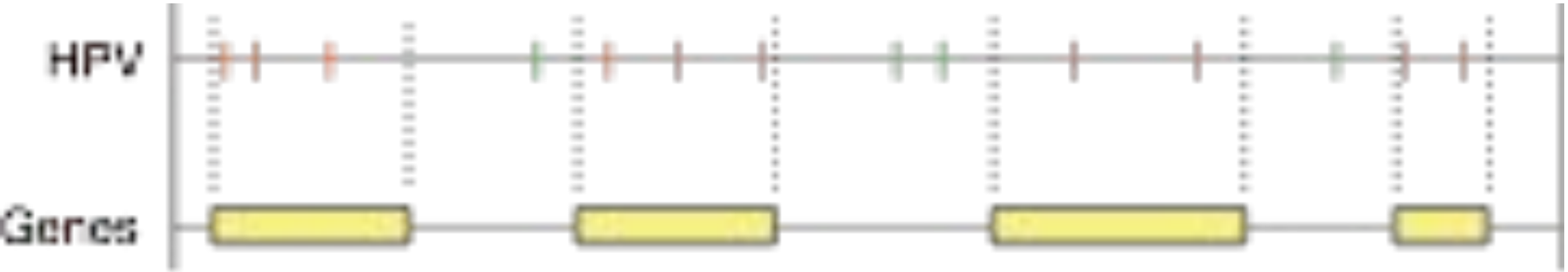
Exercise 8: HPV and genes

Student answers:

2. Which null model would you choose?

Move HPV sites to random positions but keep number of sites		
Keep HPV where they are, move genes randomly around (keep segments)		
Move HPV sites to random positions but keep number of sites (keep inter-point distances)		

Exercise 9: HPV and genes



Check whether HPV sites are integrated closer to genes than expected by chance, using the Genomic Hyperbrowser. Follow these steps, and make your own choices when needed.

1. Go to the Genomic HyperBrowser (https://hyperbrowser.uio.no). Register a new user if you want to keep what you do.
2. Go to Statistical analysis of tracks -> Analyze genomic track, in the left hand menu
3. Genome: hg19
4. Track 1 (HPV): Phenotype and disease associations:
   Assorted experiments:Virus integration, HPV specific..
5. Track 2 (Genes): Find a suitable gene dataset yourself
6. Figure out the rest yourself (but select to analyse "**P-S** Located nearby?" (P means we treat first track as points and S means we treat second track as segments)
7. **NB:** Set random seed to 0 (so that you can compare results) and MCFDR sampling depth to "Quick and rough indication"
8. **NB2:** Choose to compare in chromosome arms.
9. If you have time: Try using the test statistic "number of HPV sites inside genes" instead of average distance to gene

Exercise 9: HPV and genes

Student answers:

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

Average log distance points to segments. Randomize point positions. REFSEQ	0.29	
Same as above with ENSEMBL. Kept interpoint distances.	0.0196	
Number of points inside genes. Randomize points (preserve segments) ENSEMBL	0.0068	
Same as above with REFSEQ	0.5071	
Average log distance points to segments. Randomize point positions. ENSEMBL	0.039	
Average distance points to segments. Randomize point positions.		

Data and assumptions matter

- **Using Refseq genes:**
  - Null model: Preserve genes, randomize HPV sites:
    - Test statistic “Average log distance”: *No support from data for this conclusion (p-value: **0.2941**)*
    - Test statistic “Average distance”: *Maybe - weak evidence (p-value: **0.01961**)*

- **Using ENSEMBL genes:**
  - Null model: Preserve genes, randomize HPV sites:
    - Test statistic “Average log distance”: *Maybe - weak evidence (**p-value: 0.03922**)*
    - Test statistic “Average distance”: *Maybe - weak evidence (**p-value: 0.01961**)*

Using test statistic “count number of HPV sites inside genes”:

- Using Refseq genes, p-value: 0.4512
- Using ENSEMBL genes, p-value: 0.007
- Using exons: 0.000947

Get to know the data

Descriptive statistics

Exercise

1. How many HPV sites do we have in the dataset?
2. How much of the genome is covered by genes?
3. How many HPV sites are inside genes?

We will answer these questions using both BEDtools and the Genomic HyperBrowser.

1. How many HPV sites do we have in the dataset?
2. How much of the genome is covered by genes?
3. How many HPV sites are inside genes?

Using BEDtools

1) Install BEDtools (https://bedtools.readthedocs.io/en/latest/content/installation.html) if you want it locally or just run it from the VM
2) Type **bedtools** on your command line to check that it is installed properly
3) Download the data files
   wget http://folk.uio.no/ivargry/norbis/ensembl.bed
   wget http://folk.uio.no/ivargry/norbis/refseq.bed
   wget http://folk.uio.no/ivargry/norbis/hpv_sites.bed
4) BEDtools has no direct way to only get sum of elements. But we can run jaccard with the same file twice to get the intersection between that file and it self, which gives the number of elements:
   bedtools jaccard -a refseq.bed -b refseq.bed
5) Try to use bedtools intersect to get number of HPV sites inside genes (see documentation for intersect). Hint: Count number of files in output file with wc -l filename.bed

Using the Genomic Hyperbrowser

1) Go to the tool **Analyze genomic tracks**
2) Select hg19 as the reference genome
3) Choose the track you want to analyse (e.g. the HPV sites track or one of the genes tracks). We only want to analyse one track if checking e.g. coverage.
4) Try to find out the rest yourself.

Hint: When we want to find number of elements or base pairs covered, we want to compute a descriptive statistics. Choose descriptive statistics under analysis, and find the descriptive statistics that answers your question.

Bonus question: How many base pairs are covered by both Ensembl genes and Refseq genes?

Exercise 10: descriptive statistics

- How many base pairs are covered by the genes?

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 are useful

- It is a good idea to compute descriptive statistics before doing an analysis (get to know your data)
- Often this leads to surprising findings (e.g. that the gene list we use cover much more of the genome than we thought)

More about test statistics and null models

How to choose good null models and test statistics, and why it is tricky

Quick recap of what test statistics and null models are

- A **test statistic** is used to quantify the effect we want to measure
  - E.g. the test statistic “number of HPV sites closer than 10 000 base pairs away from a gene” might be used to measure whether HPV tend to integrate close to genes

- A **null model** is a model of the “world” where we assume no effect (or association)
  - An example of a null model: “HPV integrates randomly across the genome, with no preference to integrate near genes or in any other systematic way”

How to select a good test statistic?

- *Choose a test statistic that **best** captures/measures the effect you are investigating, with as little noise as possible.*
- Example of some possible test statistics (some are bad):
  - Number of chromosomes that have at least one gene with an HPV site close
  - Number of genes with at least one HPV sites close to itself
  - Number of HPV sites close to a gene

How to select a good null model?

- Choose a null model that is as realistic as possible if no effect/association is expected
- Example of some null models for the HPV case (some are better than others):
  - 37% of the base pairs in the genome are independently covered by genes, there are 119 HPV sites on random locations.
  - All the genes are kept, but on random locations. HPV sites are at random locations
  - Genes are kept where they are, HPV sites are moved to random locations

Making justified choices can be hard

- 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

Making justified choices can be hard

- The choice of data may influence results
  - E.g. 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 can be hard

- 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, this is an often overlooked step
- Much better is actually discussing the assumptions of the hypothesis tests from biological and statistical points of view
    - E.g. “we randomize HPV integration sites since we assume that in the null model, the virus has no preference to integrate to any specific locations”

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)

Rules of thumb

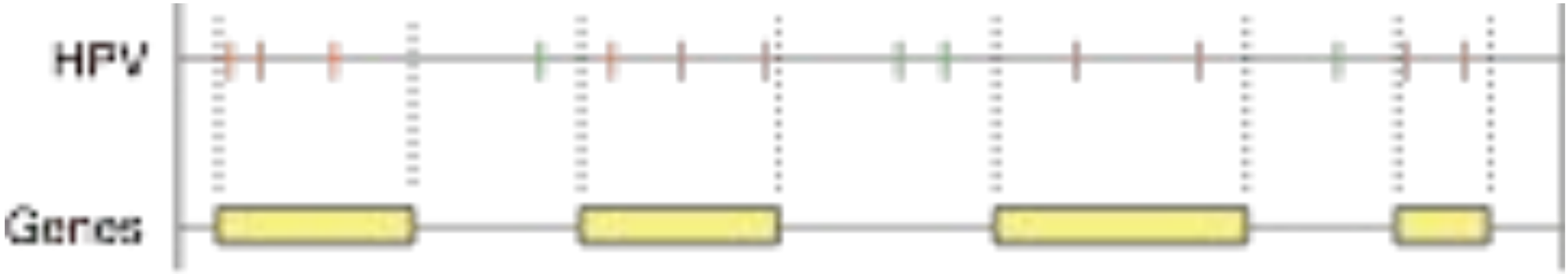
- Generally:
    - Use test-statistic that gives best (lowest) p-value
    - Use null model that gives worst (highest) p-value
- Reasoning:
    - Use measure that best catches relation of interest
    - Use the most realistic model of nature (null model)

Defining a test statistic is not always easy

- 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."

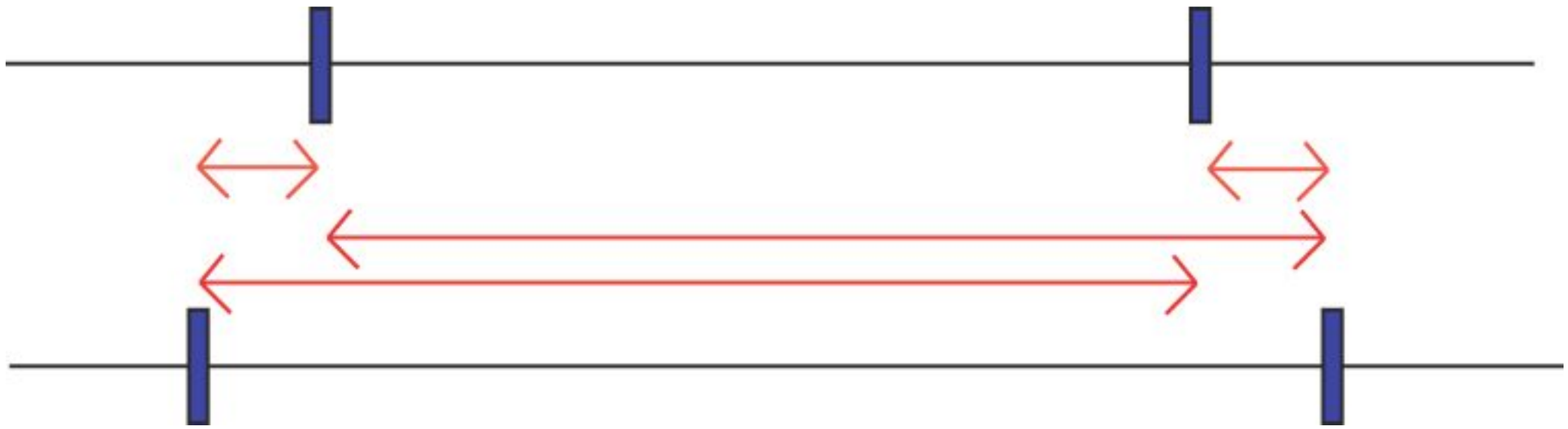
Measuring closeness is not trivial



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

How to quantify close?

We can calculate distance between pairs of elements

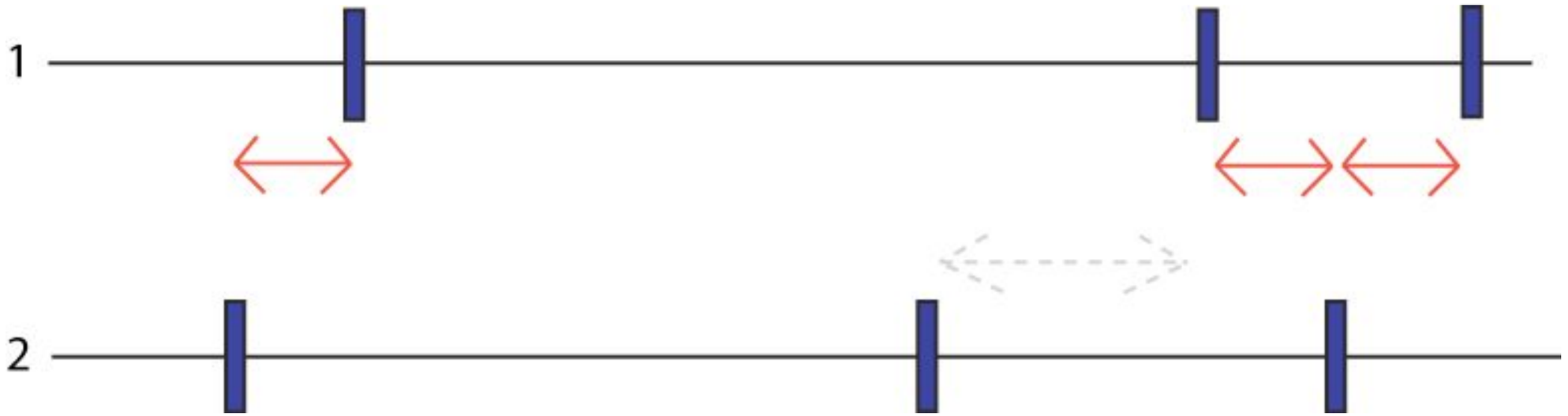


- But which pairs of elements to use - not all vs all?

We can calculate distance between pairs of elements

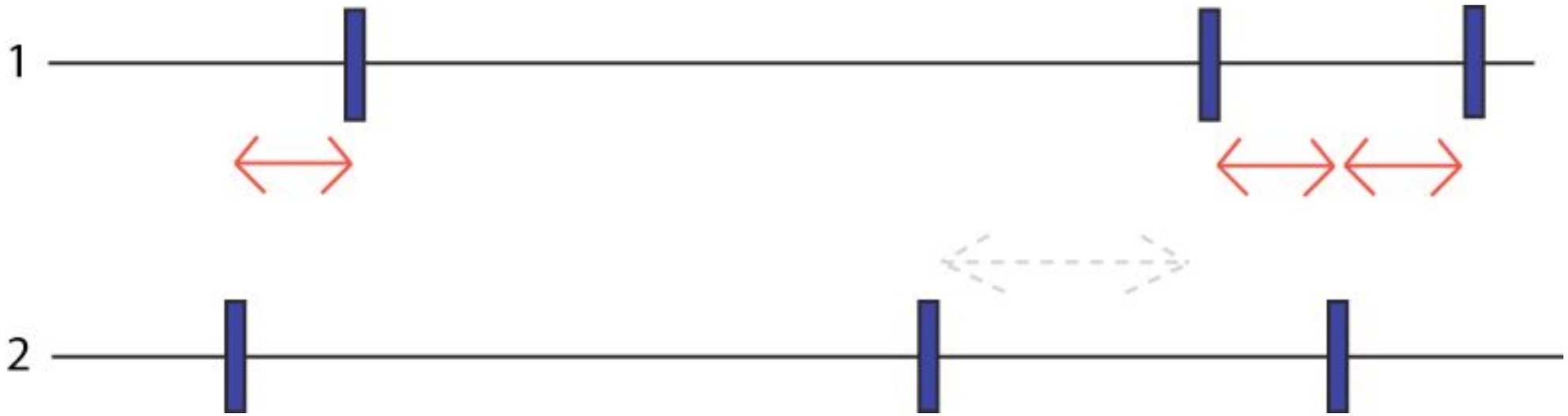
- But which distances - not all vs all?
  - We can match each gene boundary to the nearest HPV site

We can calculate distance between pairs of elements



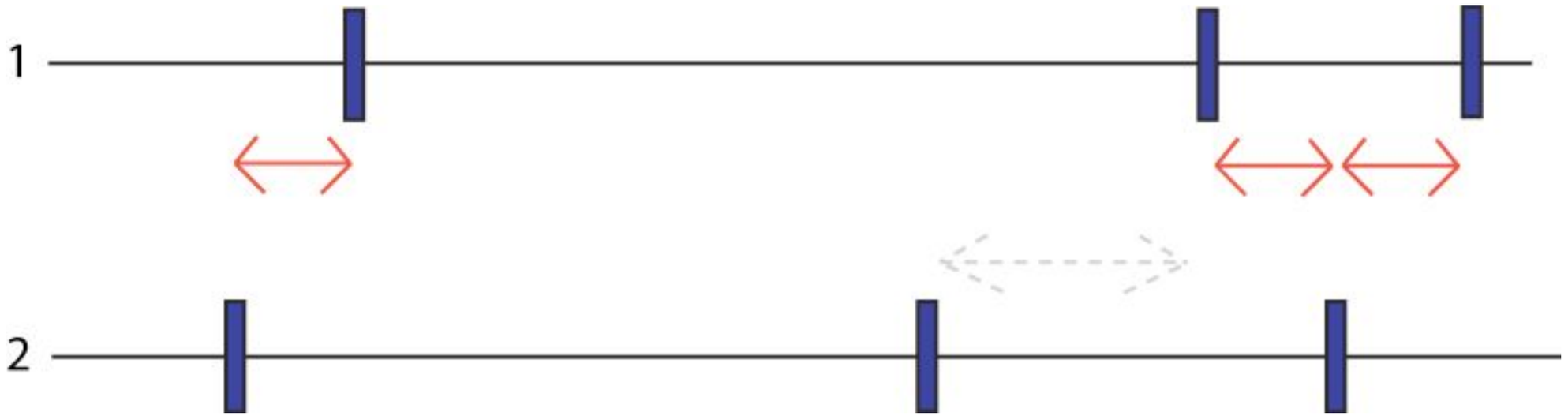
- But which distances - not all vs all?
  - This is not a symmetric measure. Not the same to match 1 against 2 as 2 against 1.

We can calculate distance between pairs of elements



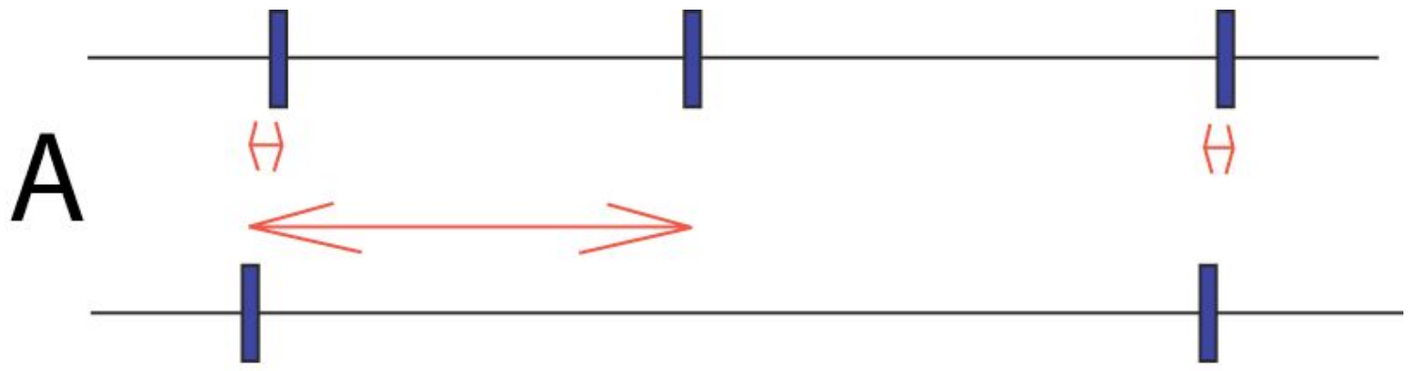
- But which distances - not all vs all?
  - If we decide on one of them, we still need a single number as our test statistic.

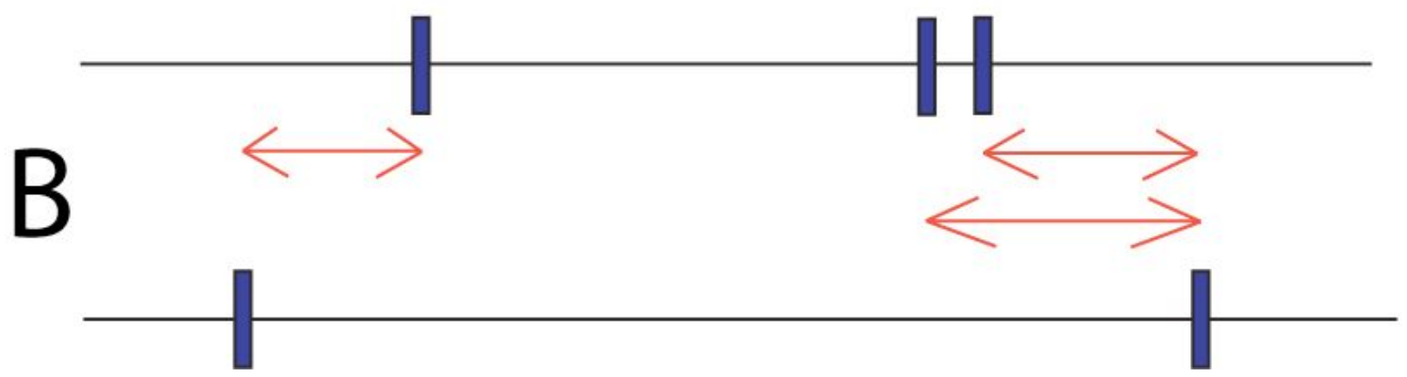
We can calculate distance between pairs of elements



- Use sum or average of of distances?

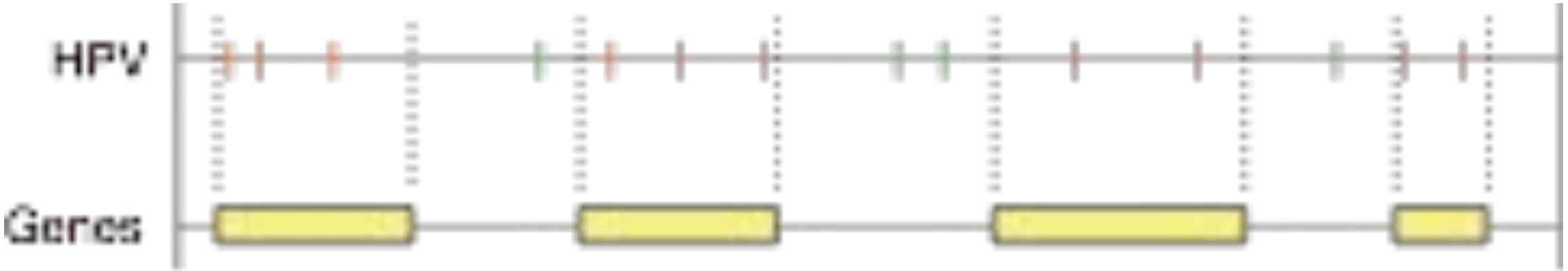
Same degree of closeness?





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

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

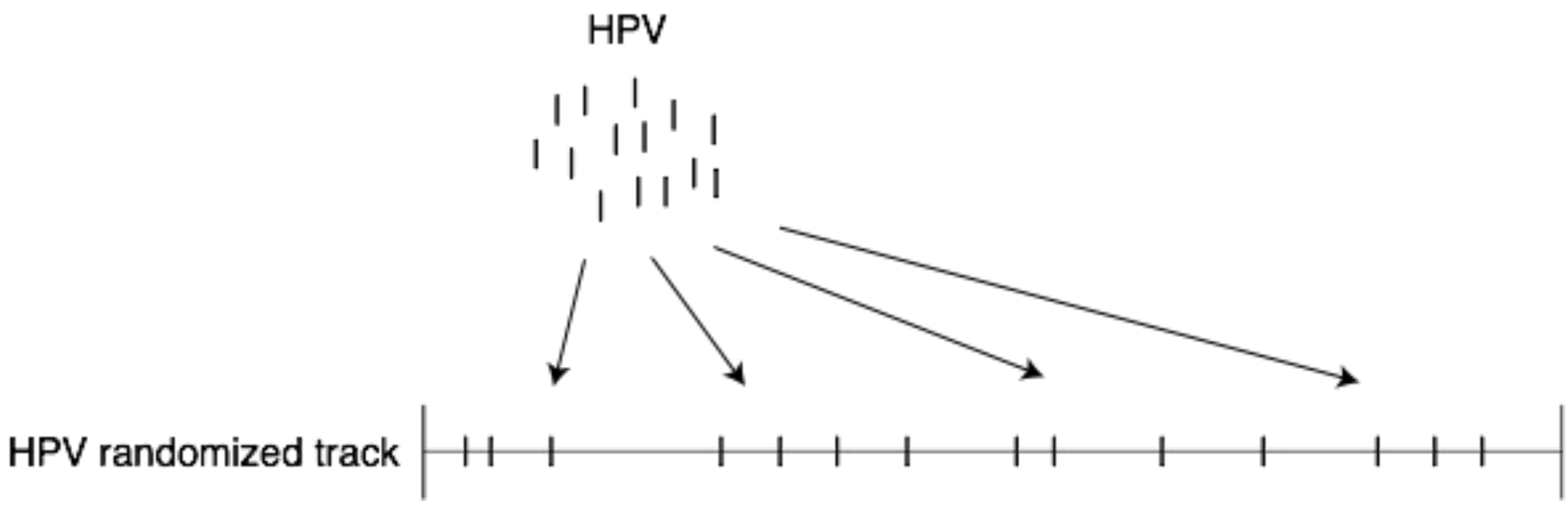
Monte Carlo

- We used Monte Carlo simulation in the first exercises when each student randomized a track and computed a test statistic using track A and the randomized track

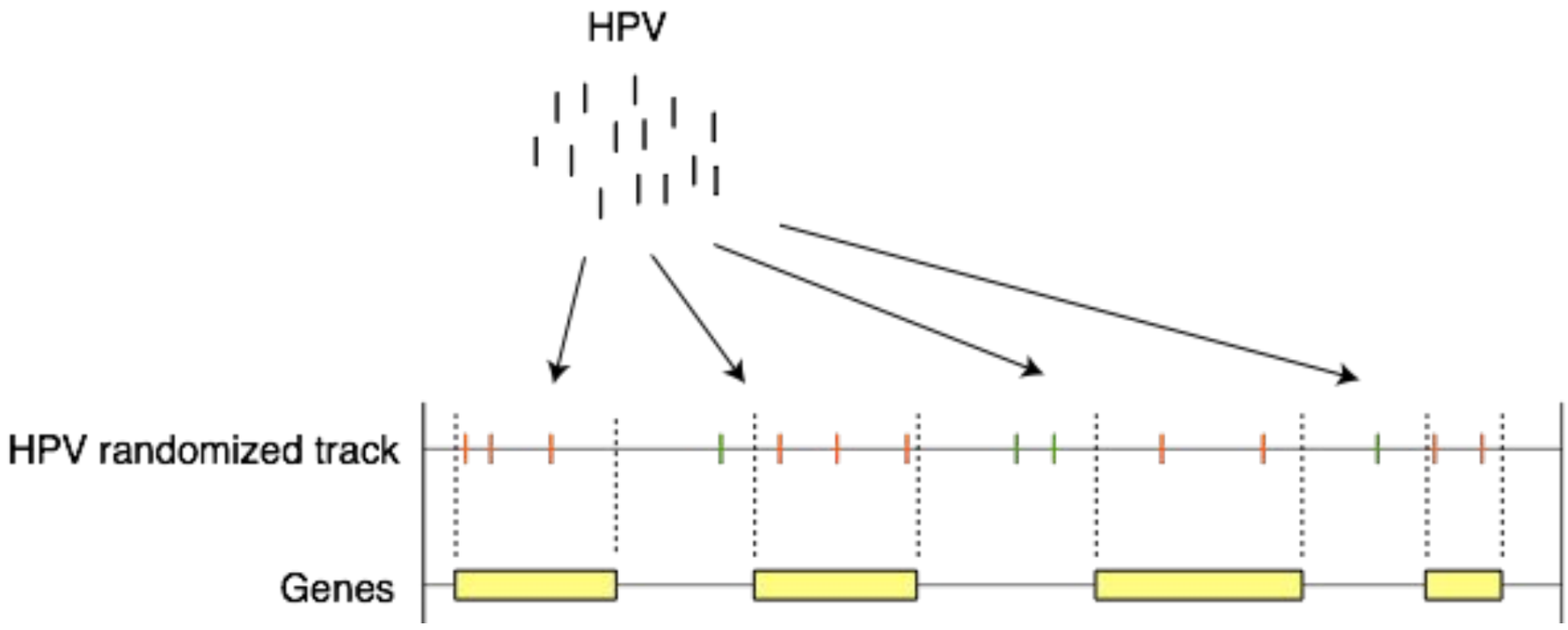
- Why is Monte Carlo simulation so powerful? Let’s see how it can be used on points and segments.

Monte Carlo test on "points inside segments"

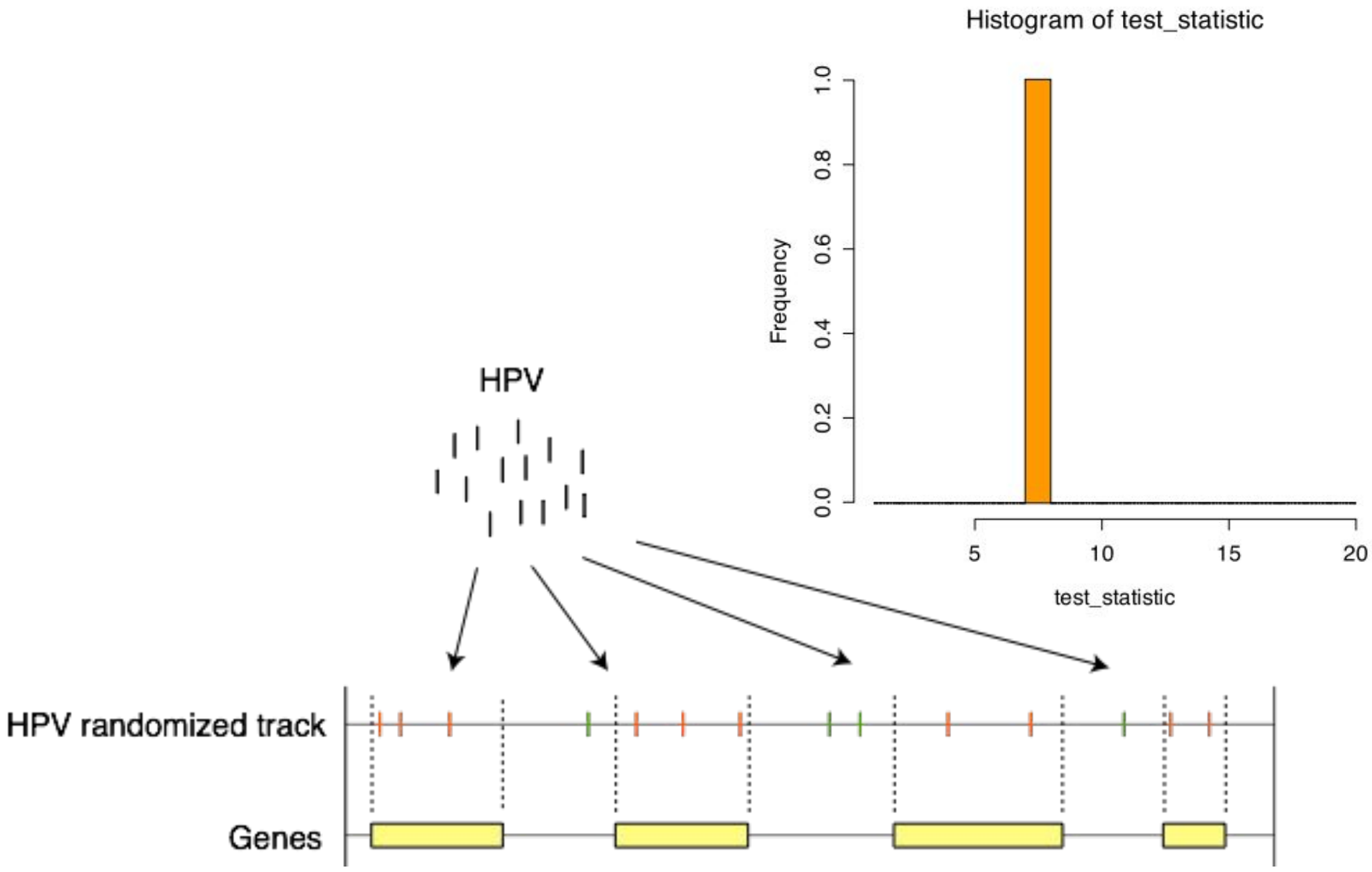
- Randomize point (HPV) locations (null model)



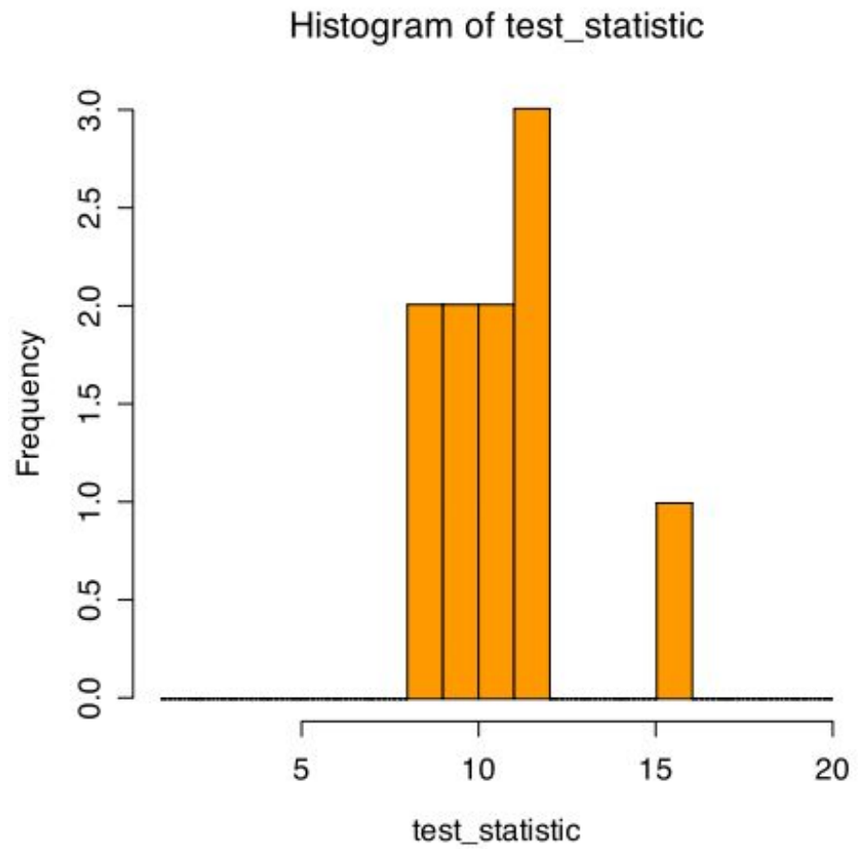
Monte Carlo test on “points inside segments”



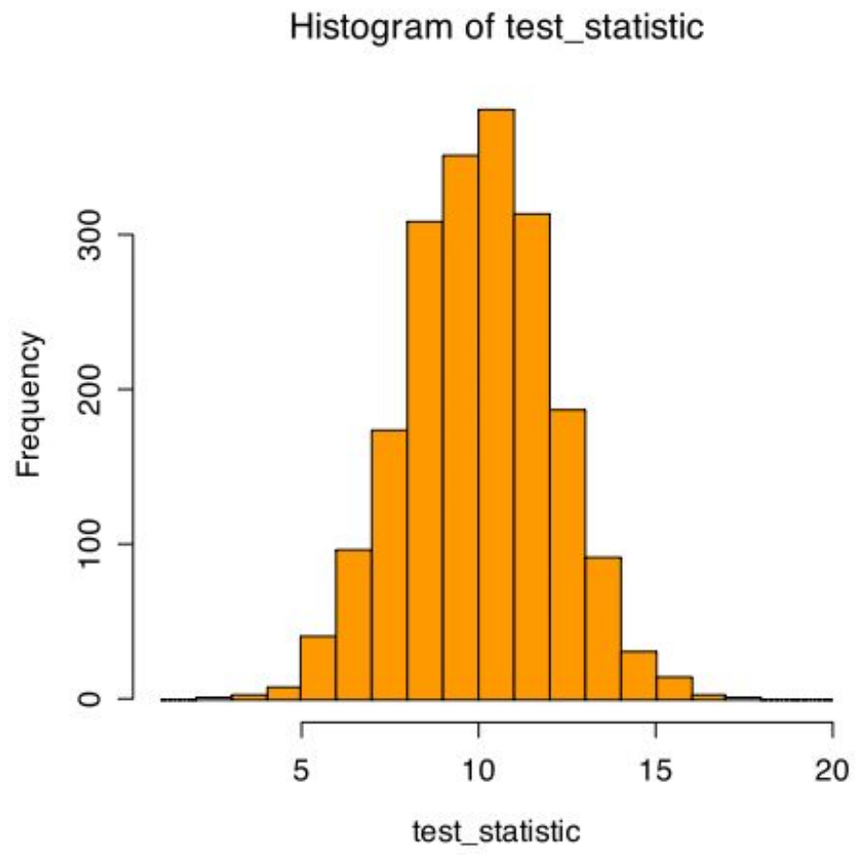
Monte Carlo test on "points inside segments"



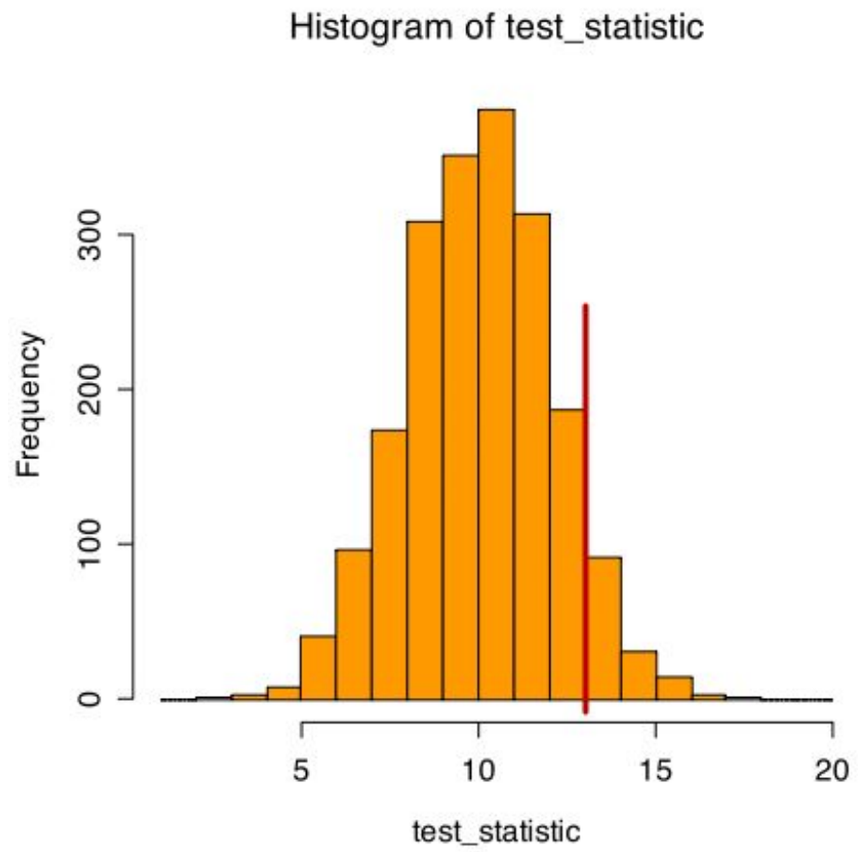
Monte Carlo test on "points inside segments"



Monte Carlo test on "points inside segments"

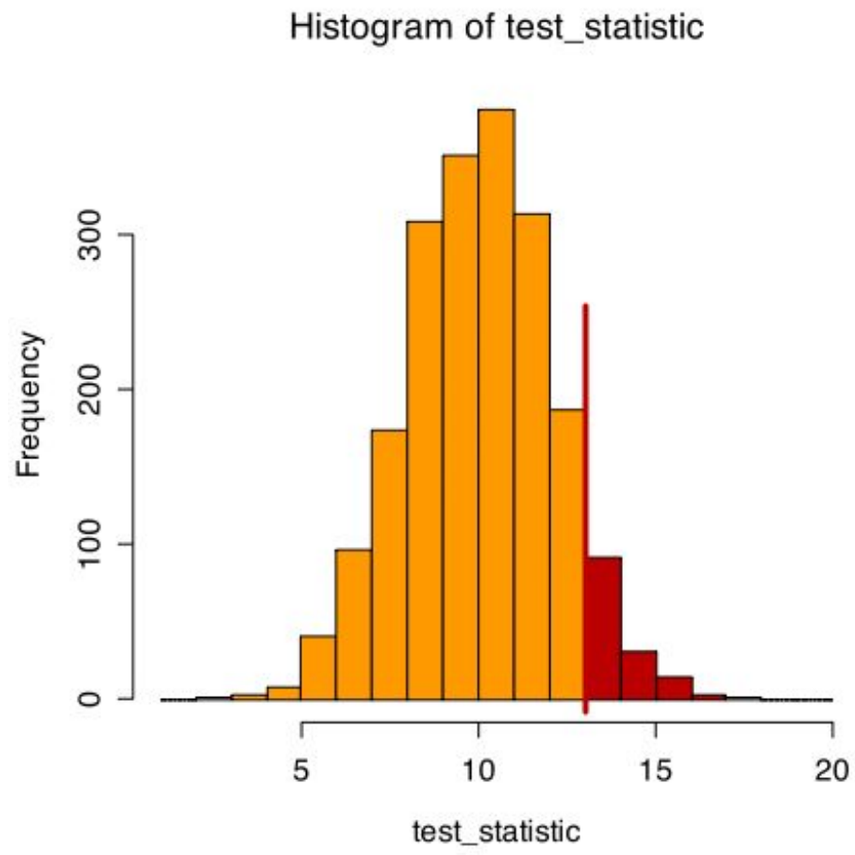


Monte Carlo test on "points inside segments"



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

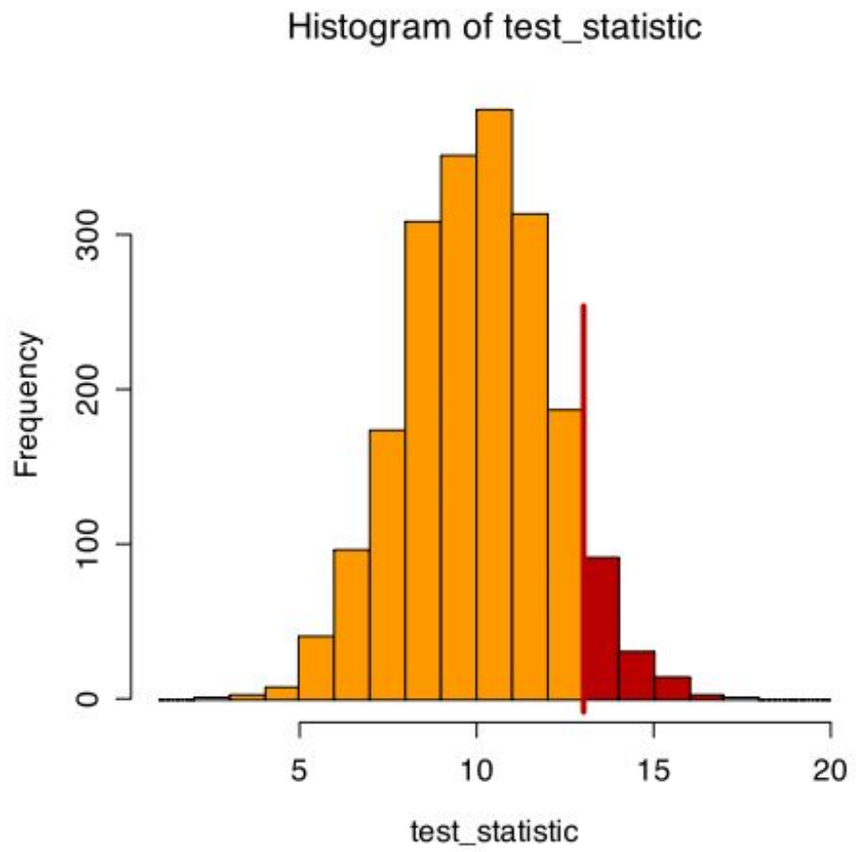
Monte Carlo: distribution

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

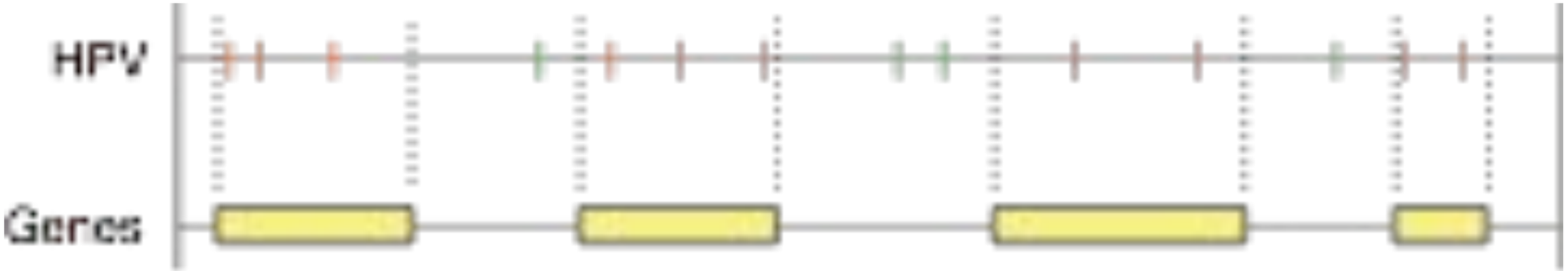
- We do this using Monte Carlo (which is slow) because we have no reason to assume a standard analytical distribution (such as the normal distribution)

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

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



Is there a faster alternative method to Monte Carlo simulation?



- Can we find a suited analytical distribution?
  (for number of HPV sites inside genes under H0)

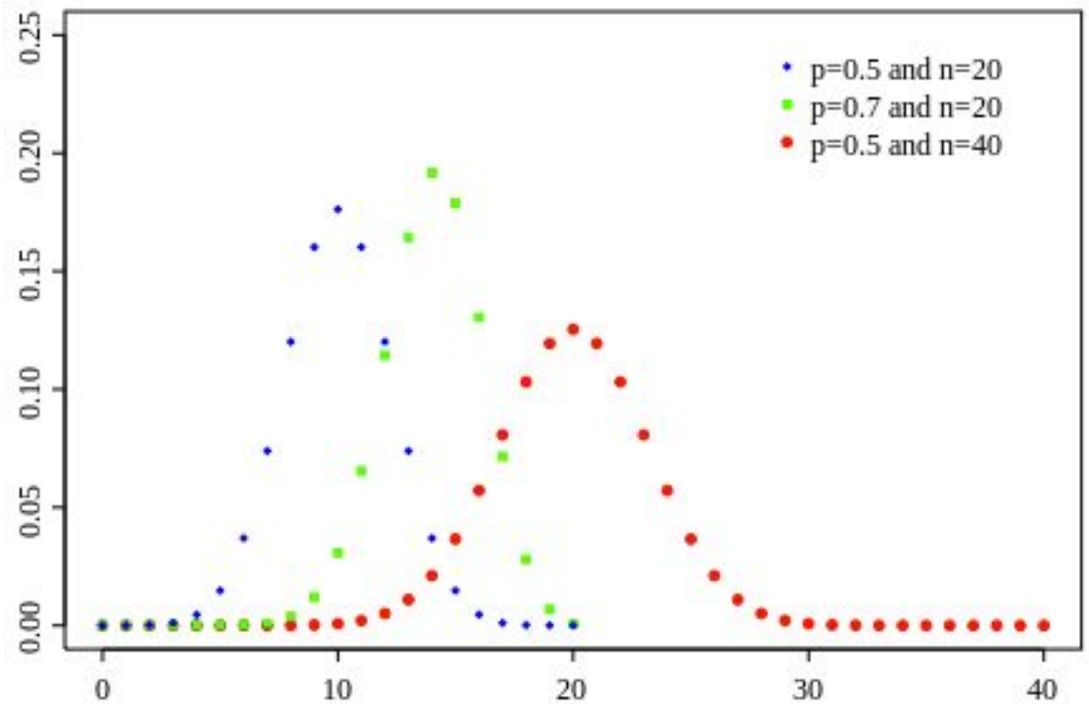
Binomial distribution

- 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:

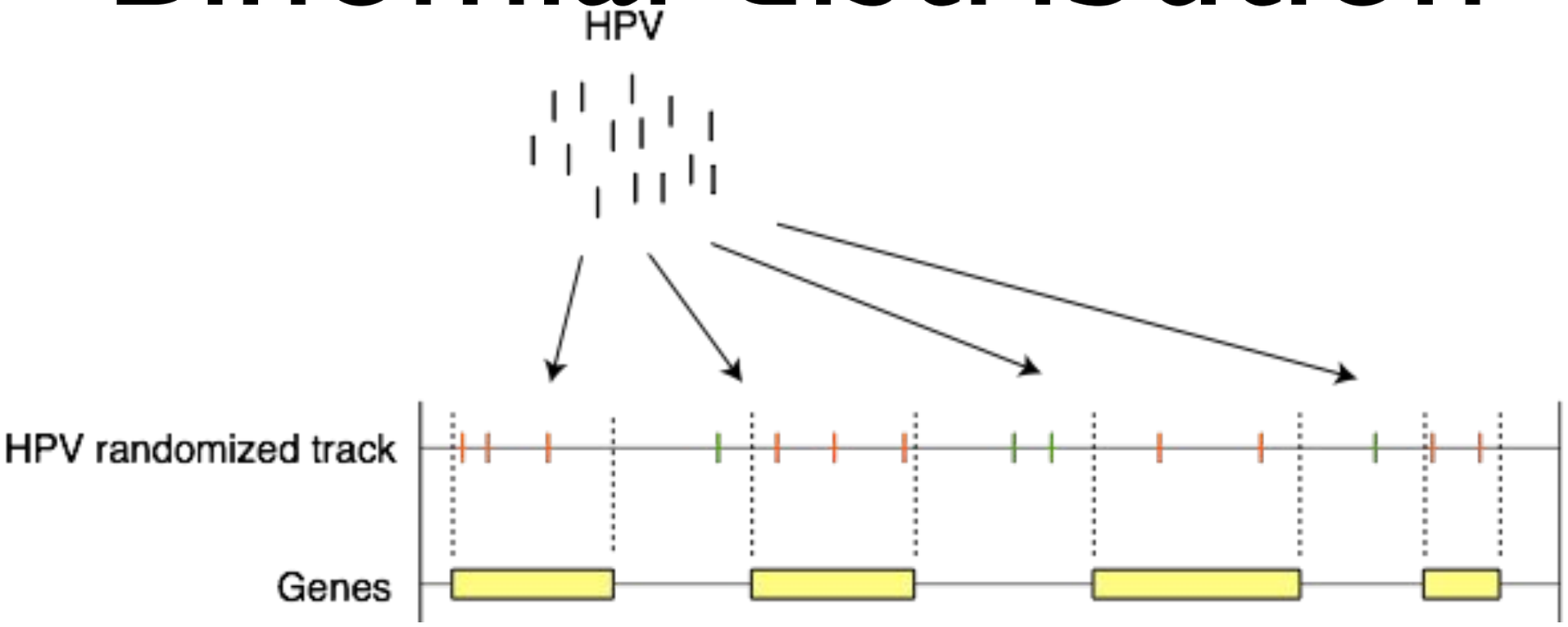
```
P(tails) = p = 0.6

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

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

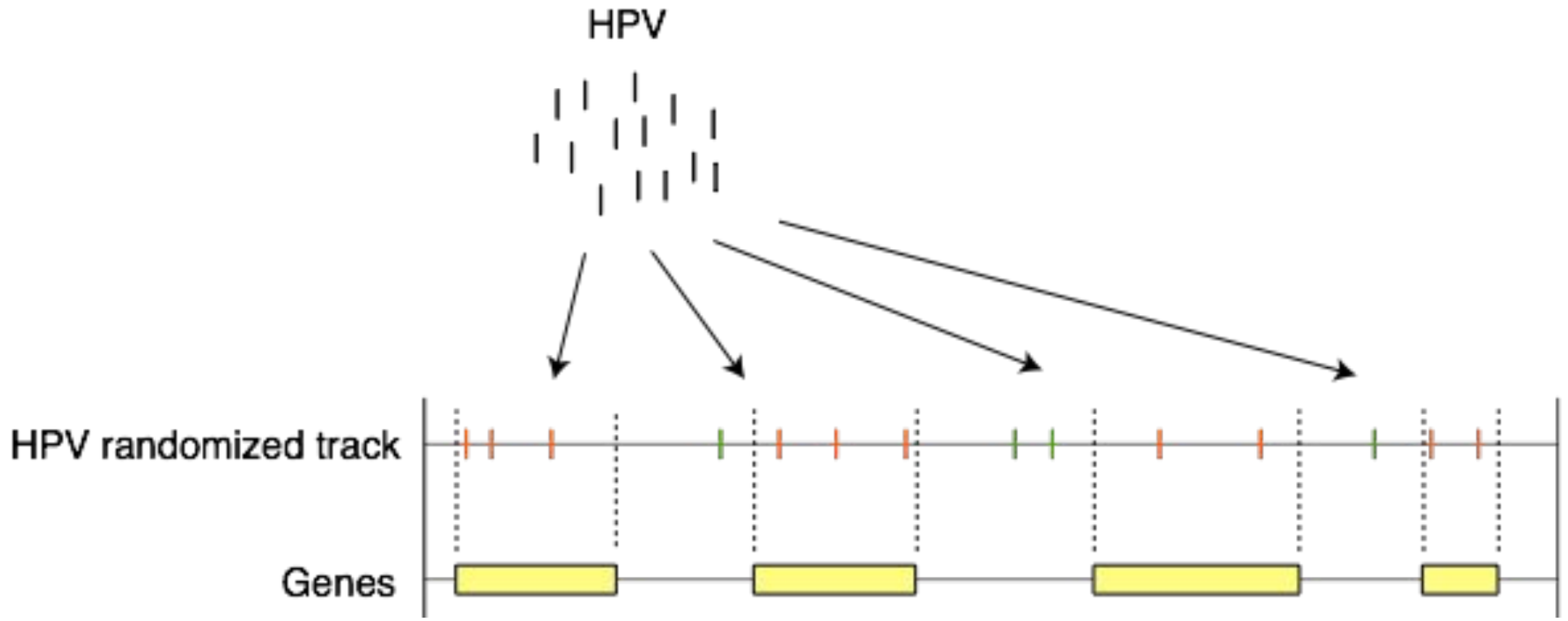


Binomial distribution



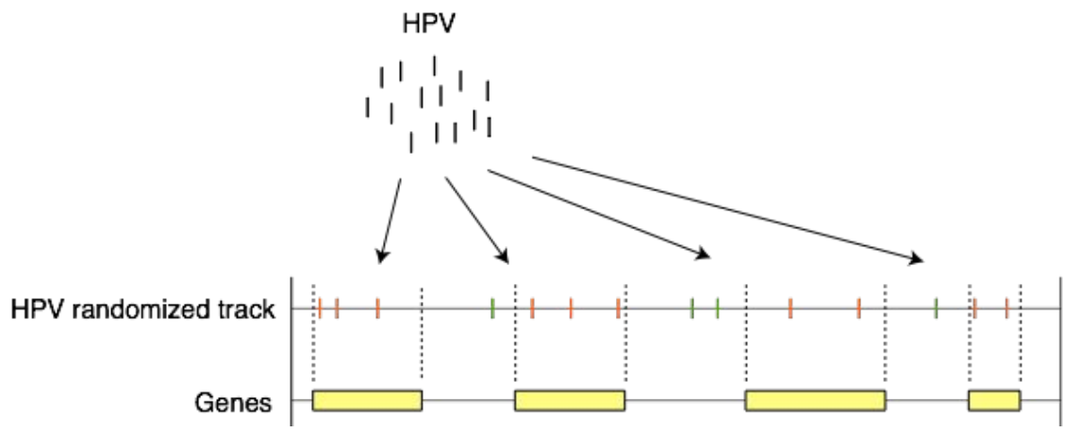
- In this case, each HPV is a coin, and it can either fall into a gene or not, depending on how much of the genome that is covered by genes
- n = number of HPV
- p = proportional coverage of genes

Binomial distribution



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

Binomial distribution



- What is binomially distributed - HPV or genes? The count of HPV within genes.
- 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")
- It seems that we can find an analytical distribution when genes are fixed and HPV sites are randomized.
  - However: For most null models, an analytical distribution is hard to find

Multiple testing

- Assume we are doing 1000 co-localization analyses like the one we did with HPV integration sites and genes, in order to find possible associations
- For each test, we accept a probability of 5% of rejecting H0 even though H0 is true (we accept H1 if the probability of H0 being true is 5% based on the data)
- If we do 1000 tests, how many false positive results will we then have?

Multiple testing

- The expected number of false findings will be 50
- There are several methods for controlling for multiple testing
- Most important, you should keep in mind that checking multiple hypothesis increases the chance of false positive findings (p-value hacking)

Bonferroni

- For m tests, the significance level is set to α/m (we require a much lower p-value)
- 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

Association vs. causation

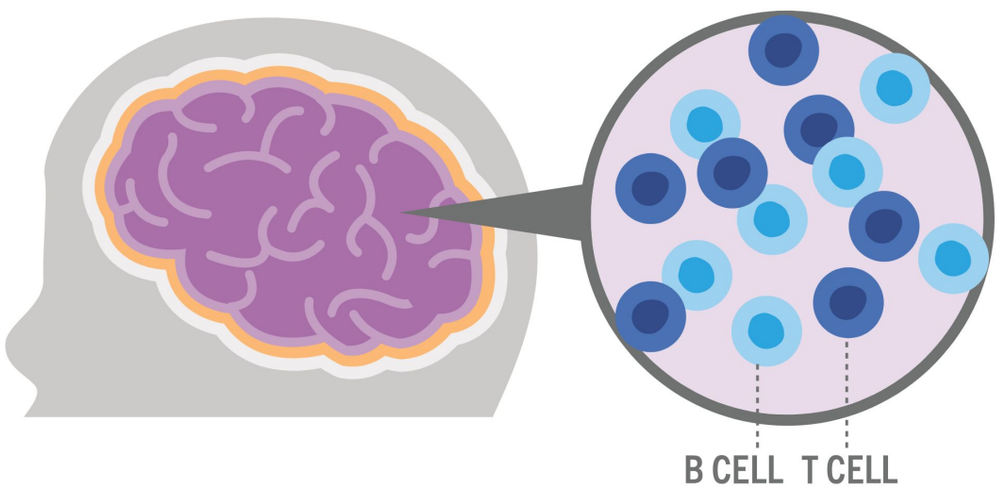
- Association: A & B are related, show up together.
- Causation: A causes B
- Using statistical testing for the co-localization of two tracks, we can only find whether there is an association
- Causation often requires speculation, biological understanding, experimentally determined mechanisms

Analysis of more than two tracks

Investigating Multiple Sclerosis

- Multiple Sclerosis (MS) is a disease in which the nervous system in the brain gradually gets damaged.
- A set of heritable genomic variants (SNPs) are found to be associated with MS.

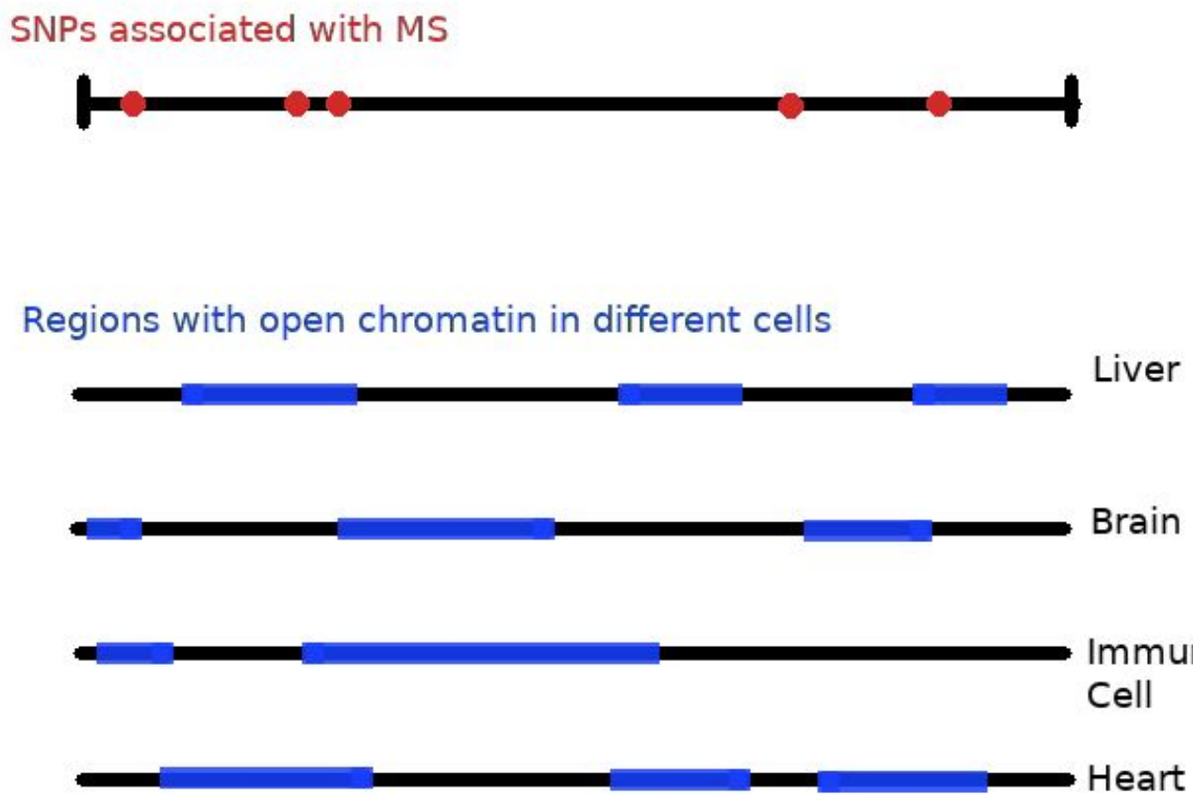
Our task: Find the cell in which the disease is active (where the SNPs might play a role). Is it the brain, or is it somewhere else?



Exercise: Investigating Multiple Sclerosis

We have the following:

- A track with position of variants (SNPs) we know are associated with Schizophrenia
- We suspect that these SNPs are able to affect gene regulation when they are inside or close to open chromatin. Open/closed chromatin varies between cell types.
- We have tracks of regions containing open chromatin for many cell types

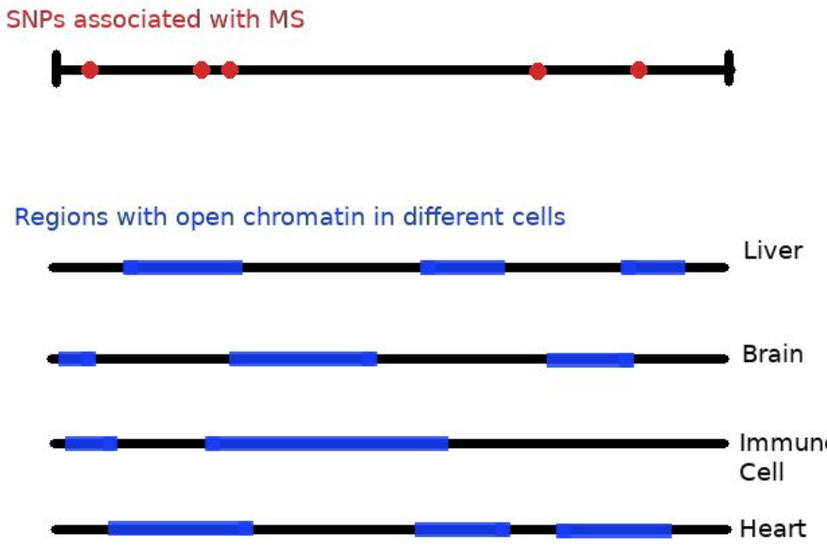


How can we find out in which cell types Multiple Sclerosis might be active?

Exercise: Investigating Multiple Sclerosis

1. Get a track with variants associated with Multiple Sclerosis on hg19 using the tool "**Extract track from HyperBrowser repository**" (try to navigate the track menu to find the right track)
2. We want to expand these SNPs to larger segments, since we want to compare the area around the SNPs to open chromatin. Use the tool "**Expand or contract points/segments**". Expand 5k bp in each direction.
3. Create a GSuite file with all open chromatin tracks using the tool "**Create a GSuite from an integrated catalog of genomic datasets**". Choose *DNA footprinting* as method. You should get 57 tracks (choose broad peaks), keep all of them.
4. Use the tool "**Determine GSuite tracks coinciding with a target track**". Choose the details yourself, but avoid computing any p-values in the choice of analysis questions (since that will be slow).

   Tips: Choose to include "cell/tissue type" in the results table



Two common similarity measures

- **Jaccard = intersection / union**

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

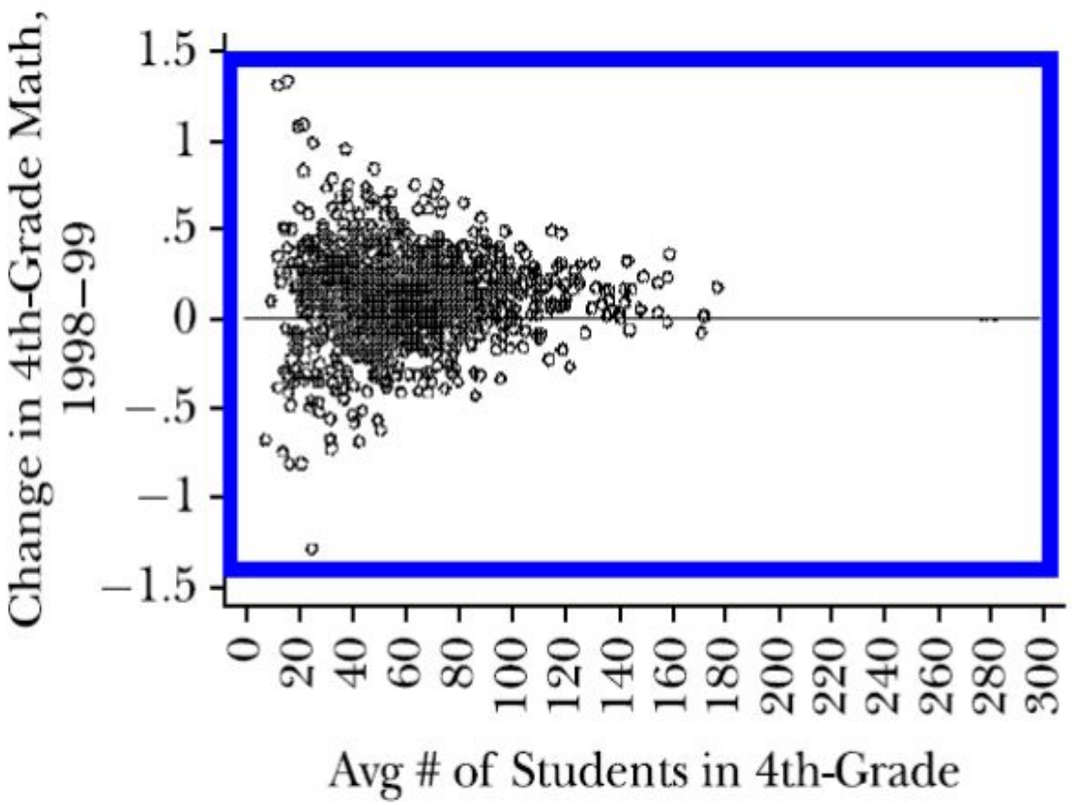
- **Forbes = observed / expected**

  Number of base pairs covered by both tracks divided by expected number of base pairs covered

Jaccard is most common to use, but it is not a good measure for similarity of genomic tracks, since it is biased by giving higher similarity for tracks that cover many base pairs.

Different similarity measures might give different results

- The commonly used similarity measure the Jaccard Index favours tracks with high coverage (intersection divided by union of tracks)



- Forbes might be better to use, but be aware that tracks with little data might be ranked high (and low) by chance (the “small schools myth”)

- You should use different similarity measures, and see if results are consistent.

Doing this exercise with BEDtools

- You will need to download all open chromatin tracks and run “BEDtools jaccard” with them.
- BEDtools does not support any other similarity measures than Jaccard and cannot compute p-values

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?

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

Exercise 13

Make an analysis reproducible.

Choose either to do the HPV integration sites vs. genes analysis or the descriptive statistics analysis reproducible.

1. Carry out the analysis in a new history
2. Make sure that the names of the history and elements are understandable
3. Create a Galaxy page with your results
4. When finished, share your Galaxy Page with your neighbor
5. The neighbor should check that he/she is able to rerun the analysis (also by using e.g. a different null model)
6. Discuss among yourself whether it was easy to understand and redo the analysis

Ten simple rules for reproducibility

1: Whenever making a claim, note a reference to supportive data

- *".. MS occur preferentially inside AP in B-cells [hist:HbLecture-8] .."*

2: For every result of interest, keep track of how it was produced

- *Solved automatically by redo-functionality if using Galaxy*

3: Record all intermediate results, when possible in readable formats

- *Intermediate steps of creating case-control are stored as history elements*

4: 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.)

5: Use executable documentation and verification

- Galaxy histories document analysis and are executable

6: Generate hierarchical analysis output, allowing layers of increasing detail to be inspected

- HyperBrowser provides conclusion, full table and local results

7: Always store raw data behind plots

- Result plots of HyperBrowser analyses come with underlying numbers

Ten simple rules for reproducibility (cont.)

8: Archive all external programs and custom scripts, in the versions that were used

- Galaxy provides this publicly and explicitly. HyperBrowser is version controlled.

9: Avoid manual, non-trackable procedures

- We have performed all analysis steps in the Galaxy system

10: 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)

Summary

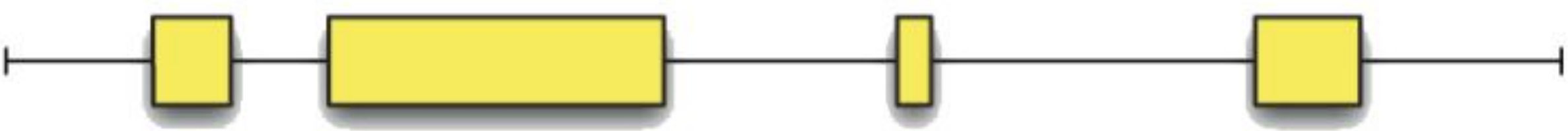
Data

- 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

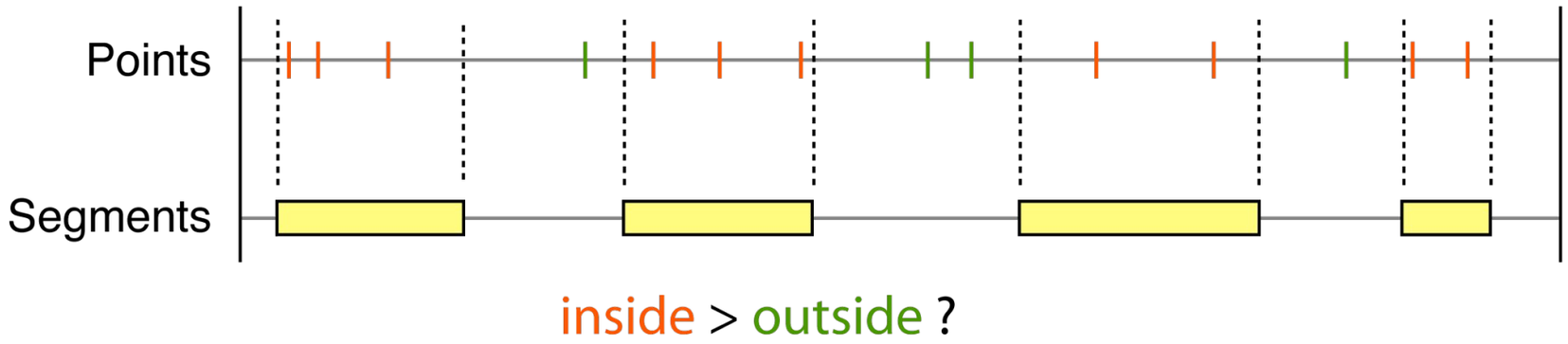
Data representation

- *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*



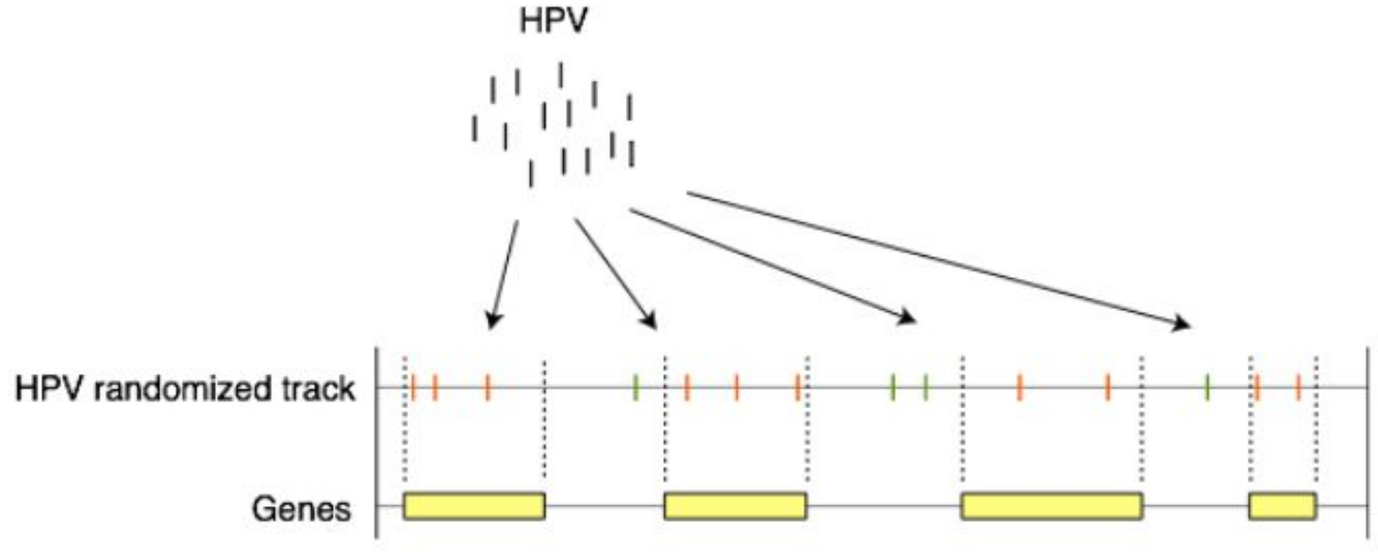
Analysis

- Co-occurence 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

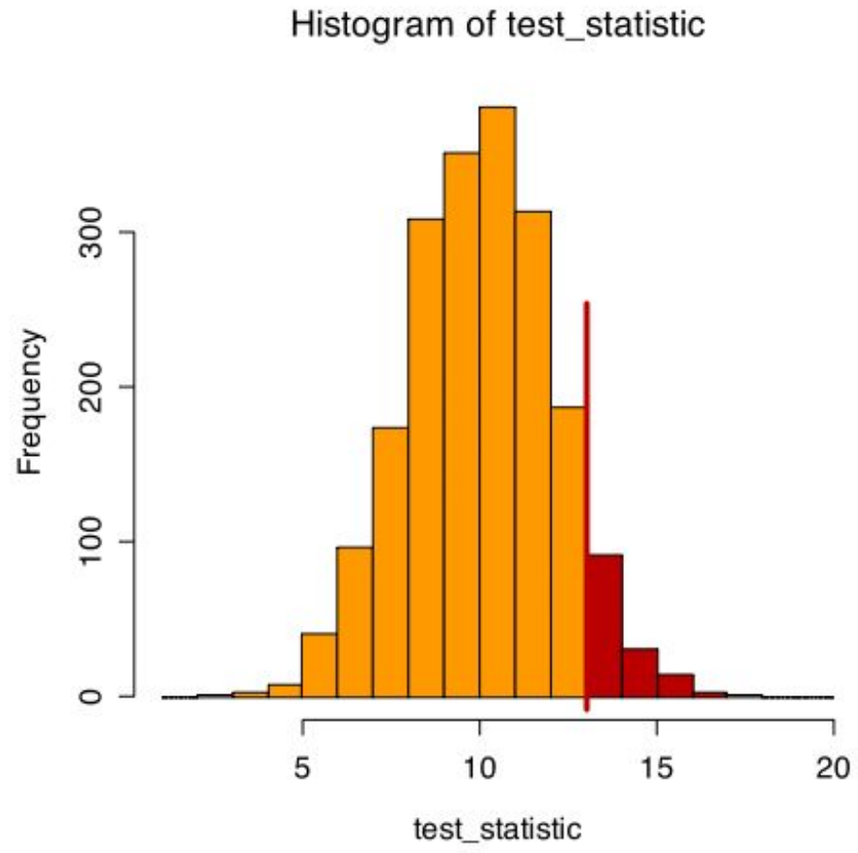
- An example of a null model:
  - We assume that SNPs are distributed uniformly across the genome when there is no association

- Preservation strategies makes the null model more realistic:
  - We can for instance preserve the inter-point/segment distances.



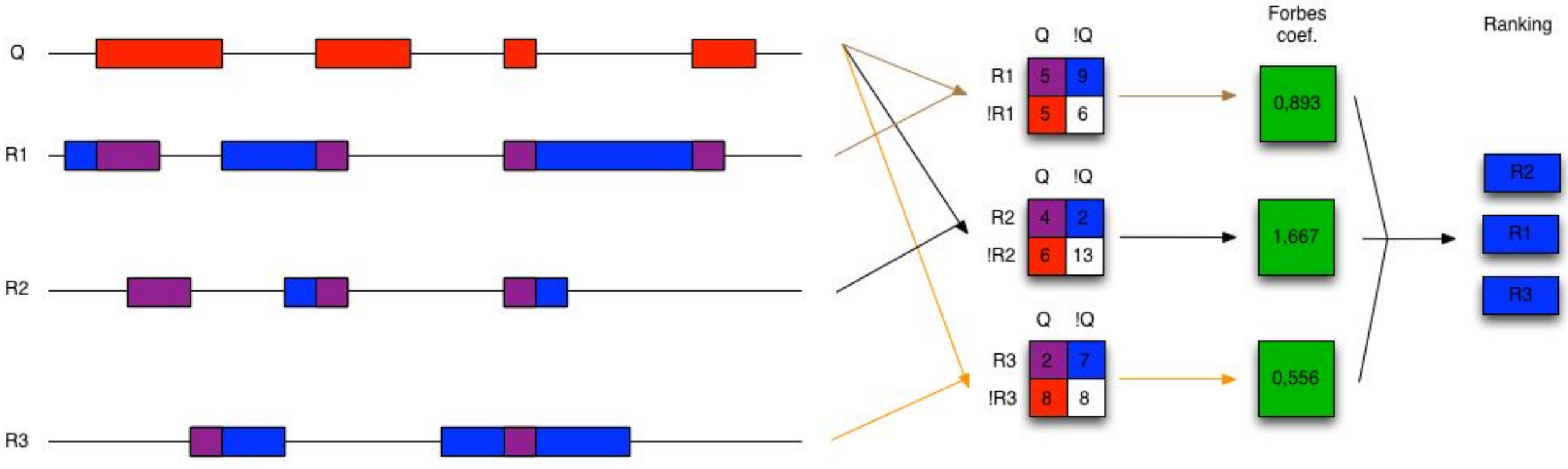
Finding the p-value

- 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:



Questions?

- Feel free to reach out if you have questions after the course ivargry@ifi.uio.no