Statistical epigenomics

MBV-INF x410
November 25 2016, Oslo

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

Overview of session

Day 1:

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

10:30-11:15 Analysis of tracks.

11:30-12:00 Hypothesis testing I

12:00-13:00 Lunch

13:00-14:00 Hypothesis testing II

14:15-16:00 Reproducibility

About this module

The form of these sessions

- We briefly introduce a topic
- You do a short exercise
- We explain the topic in more detail
- ... we repeat this for a sequence of increasingly advanced/detailed topics

Biological cases, but not depth

- We will use biological cases, but not focus on biological interpretation:
  - You are the experts in biology, not us
  - Our message is the methodology and its generic (statistical) interpretations
  - Feel free to correct us if we say something wrong

About the Genomic HyperBrowser

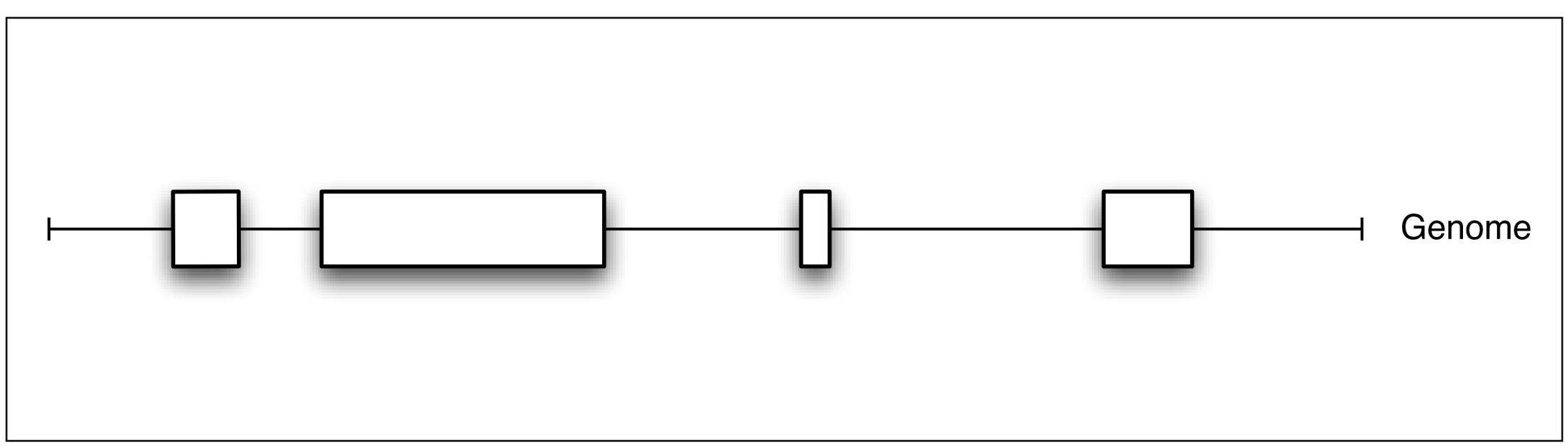
- We will make use of the Genomic HyperBrowser in this session
- The HyperBrowser is a software system for statistical analysis, developed locally at UiO
- However:

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

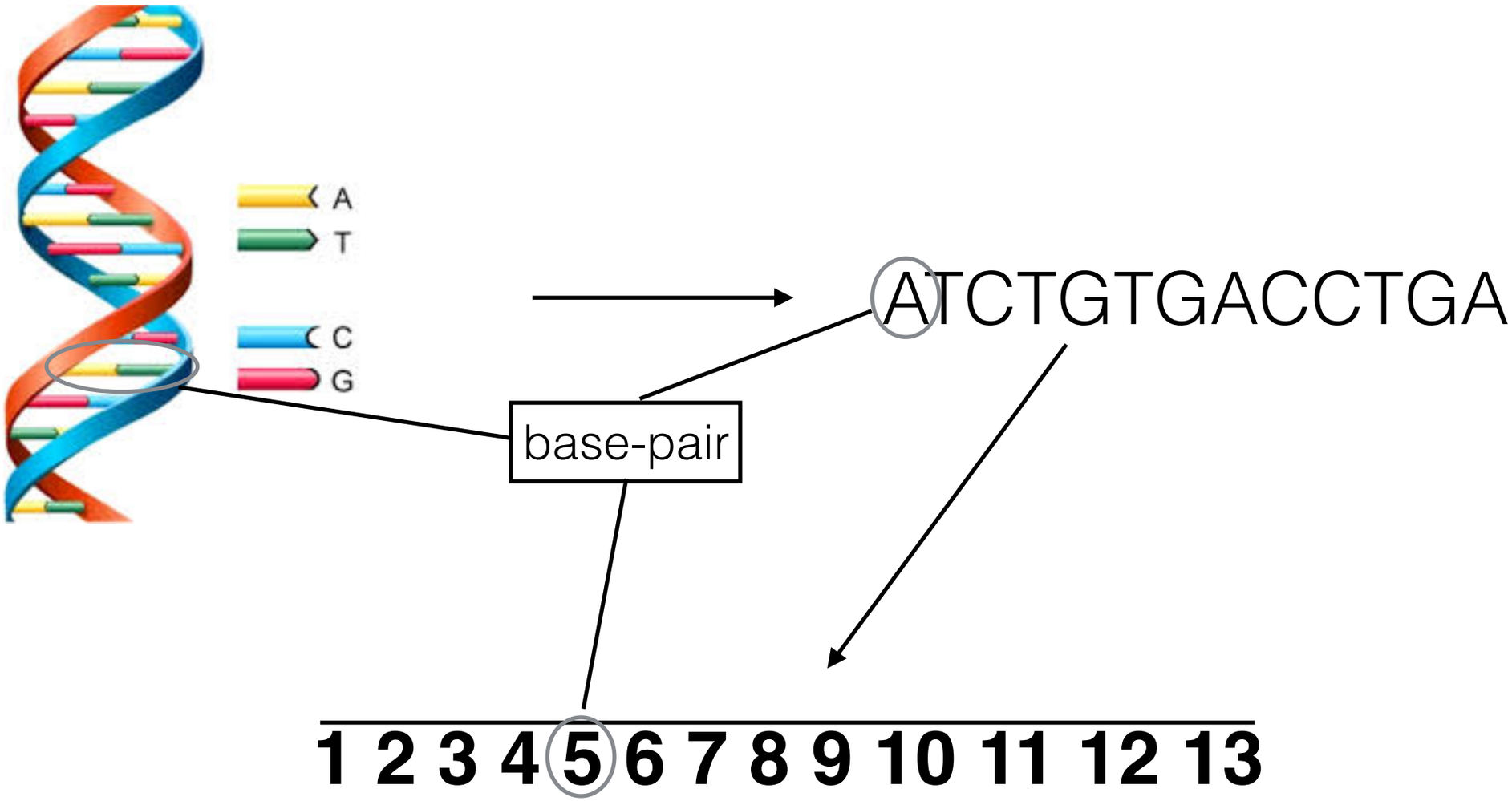
Introduction

What are genes?

This! :



Genome as a line



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

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

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

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

Statistical genomics

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

Statistical epigenomics

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

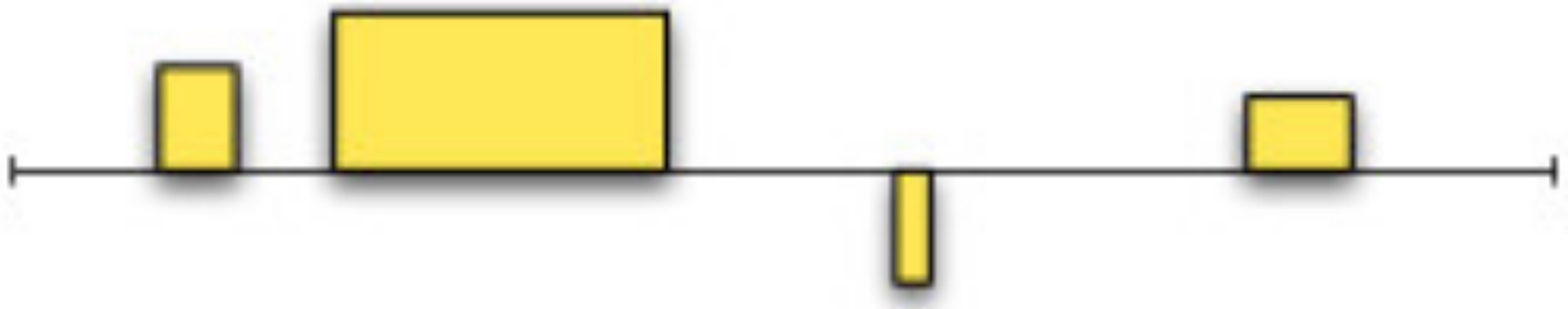
- However:
  - Many of the concepts are central statistical concepts that can be used for other types of analyses

Tracks and track types

Representation of genes

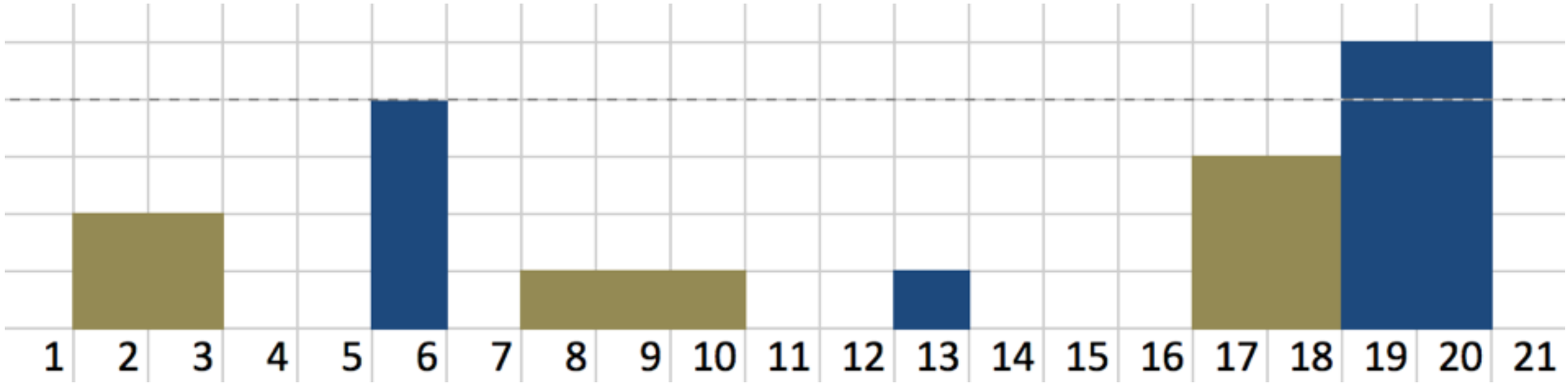
chr7	127471196	127472363
chr7	127472388	127473530
chr7	127473555	127474697
chr7	127474701	127475864
chr7	127475893	127477031
chr7	127477121	127478198
chr7	127478300	127479365
chr7	127479375	127480532
chr7	127480538	127481699

How about gene expression data (RNA-seq)?



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

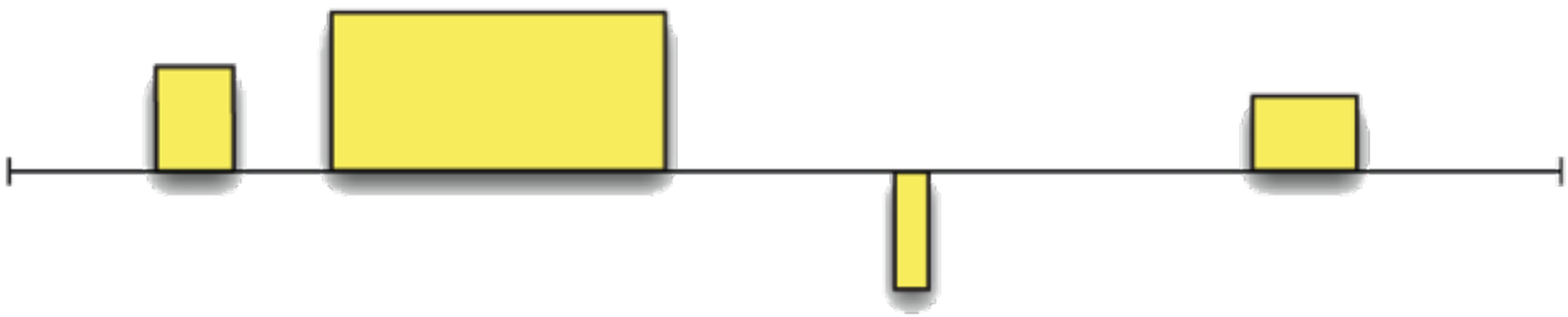
Exercise 1



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

Track types

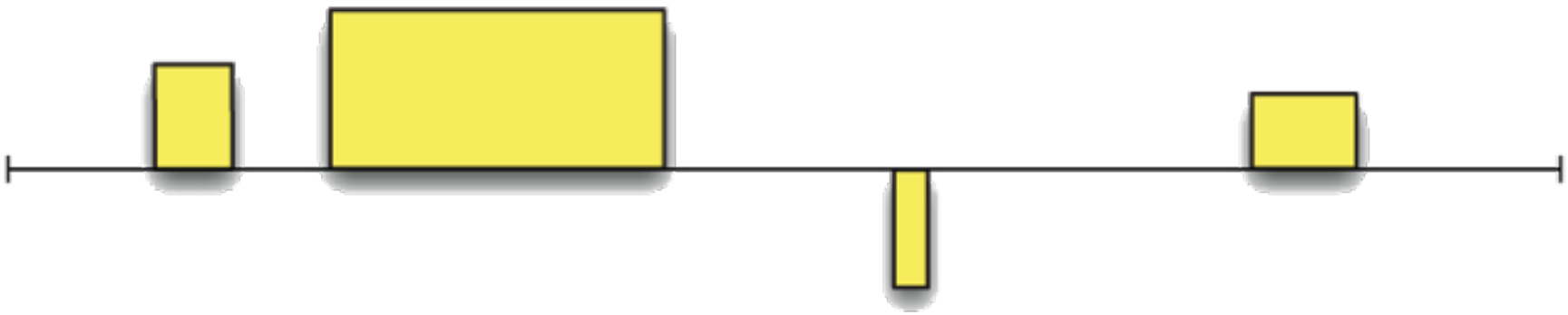
- In the last example, we showed genes as segments on the genome line, with attached RNA-seq read count values
- This track is of a **track type** we call "valued segments"



Valued Segments (VS)

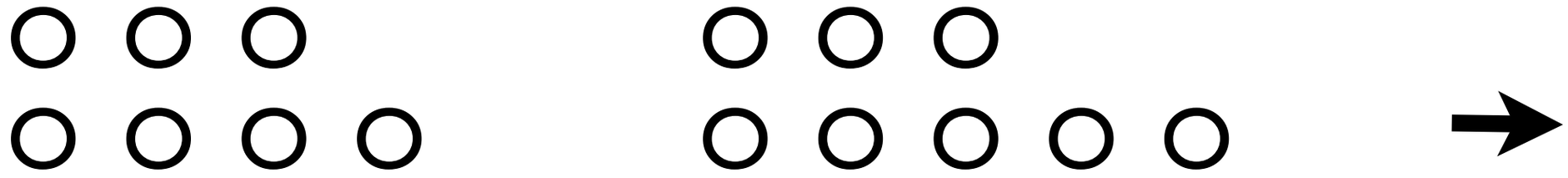
- Track types are mathematical / conceptual models used to categorize track according to their main characteristics

Exercise 2

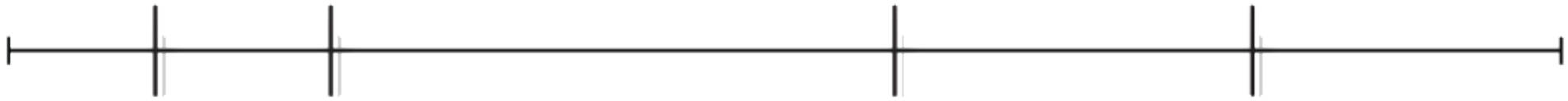


Valued Segments (VS)

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

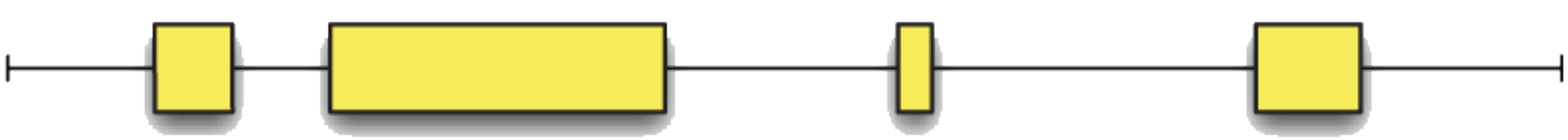


Points



Points (P)

Segments



Segments (S)

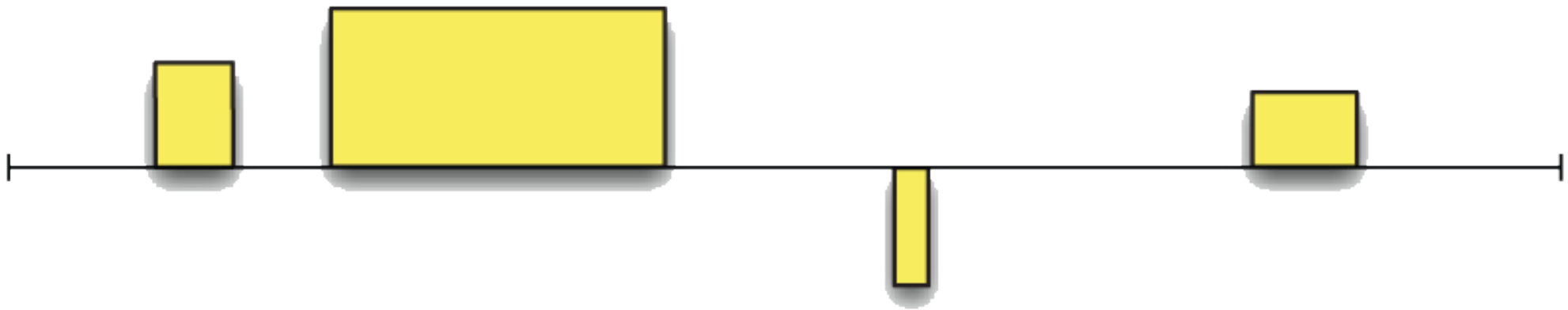
Genome Partition

Genome Partition (GP)

Valued Points

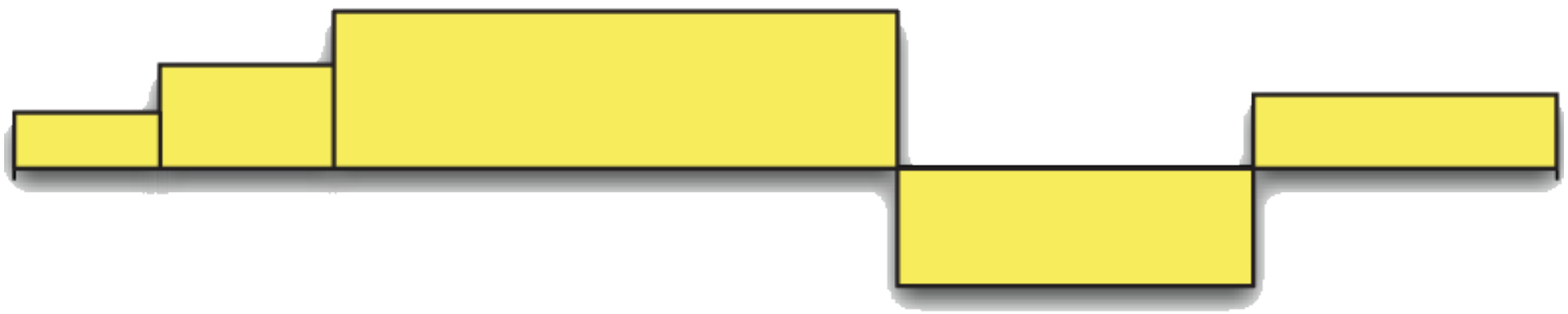
Valued Points (VP)

Valued Segments



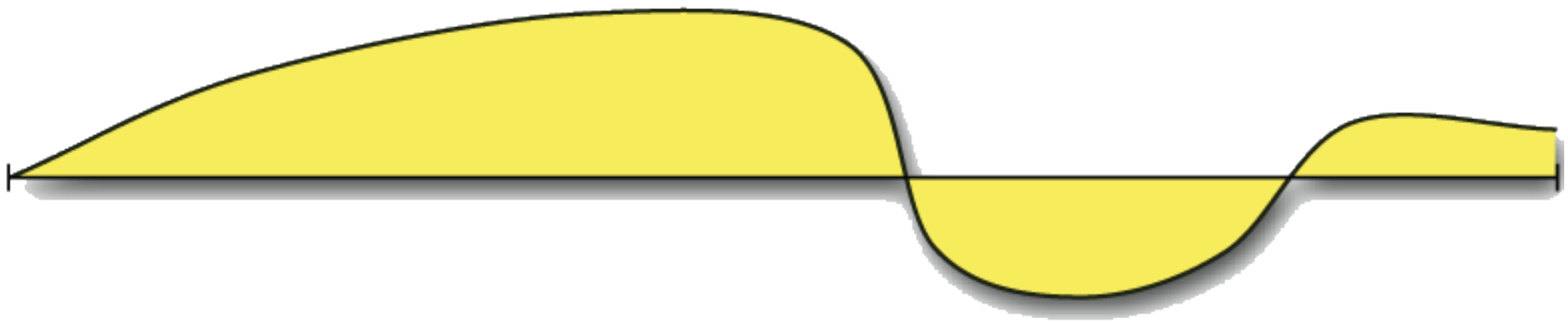
Valued Segments (VS)

Step Function



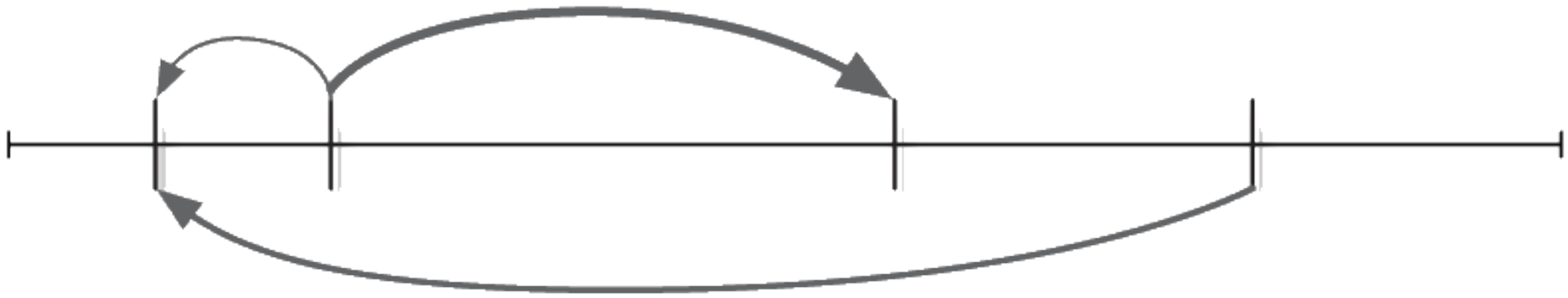
Step Function (SF)

Function



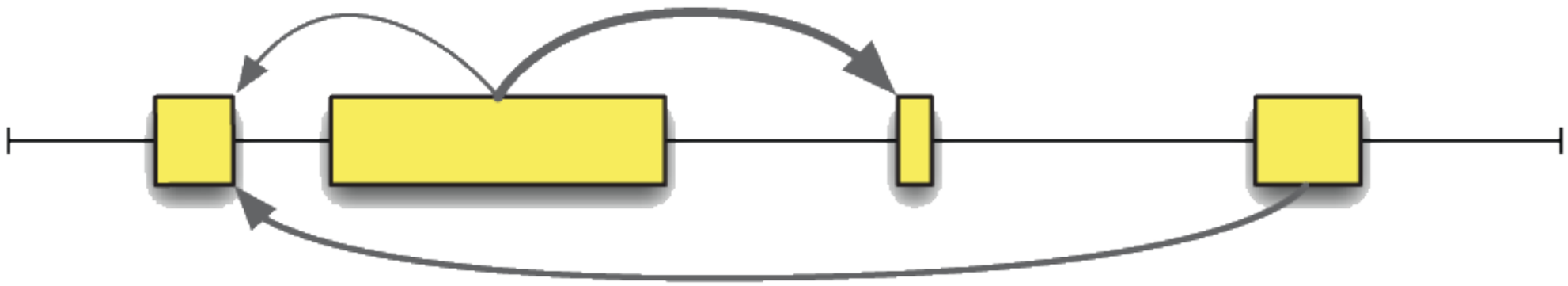
Function (F)

Linked Points



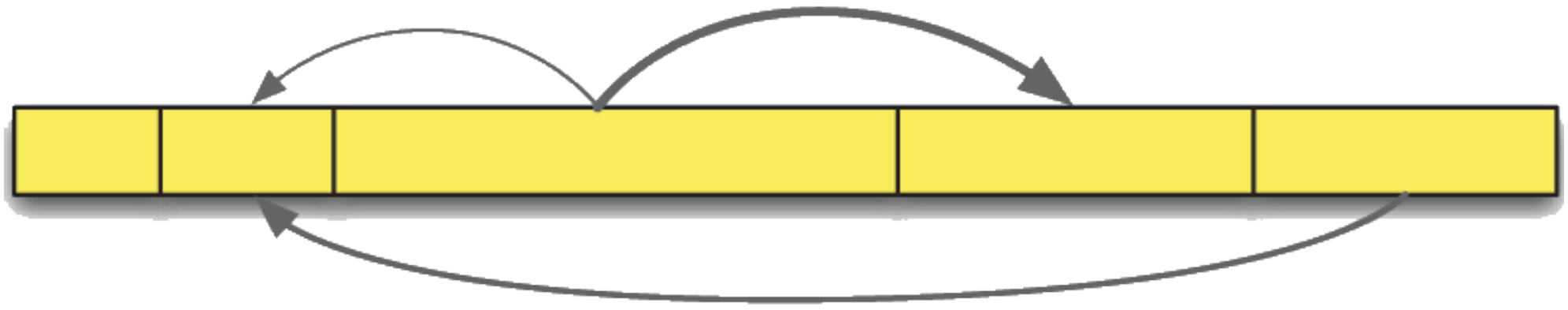
Linked Points (LP)

Linked Segments



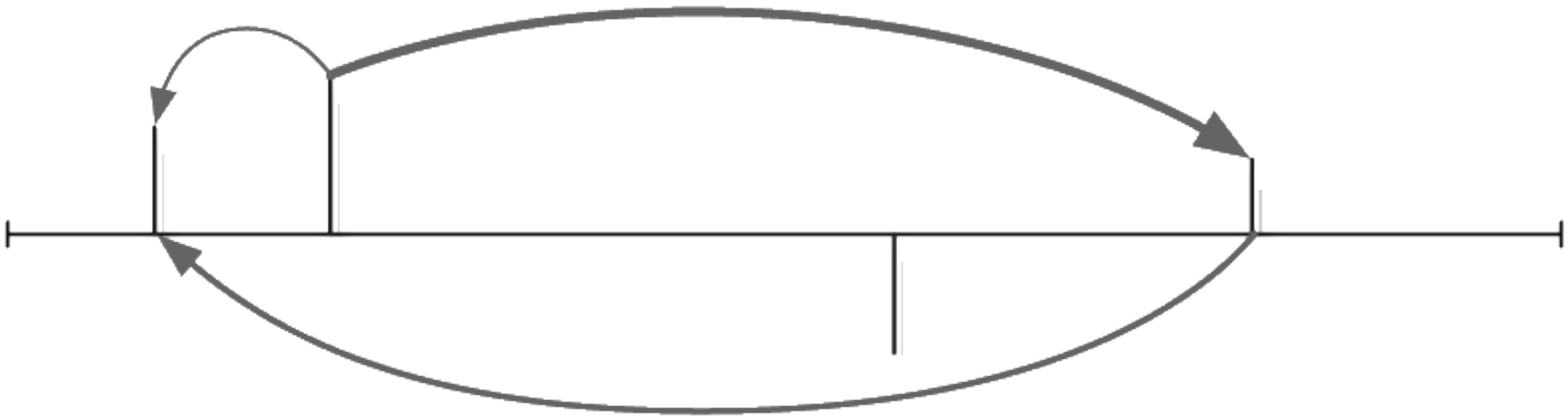
Linked Segments (LS)

Linked Genome Partition



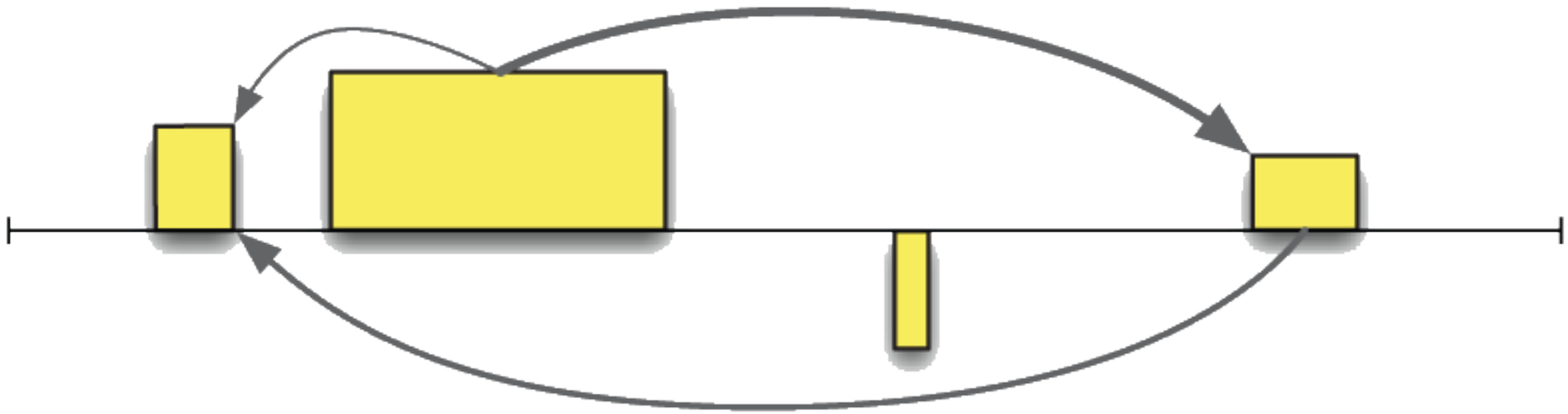
Linked Genome Partition (LGP)

Linked Valued Points



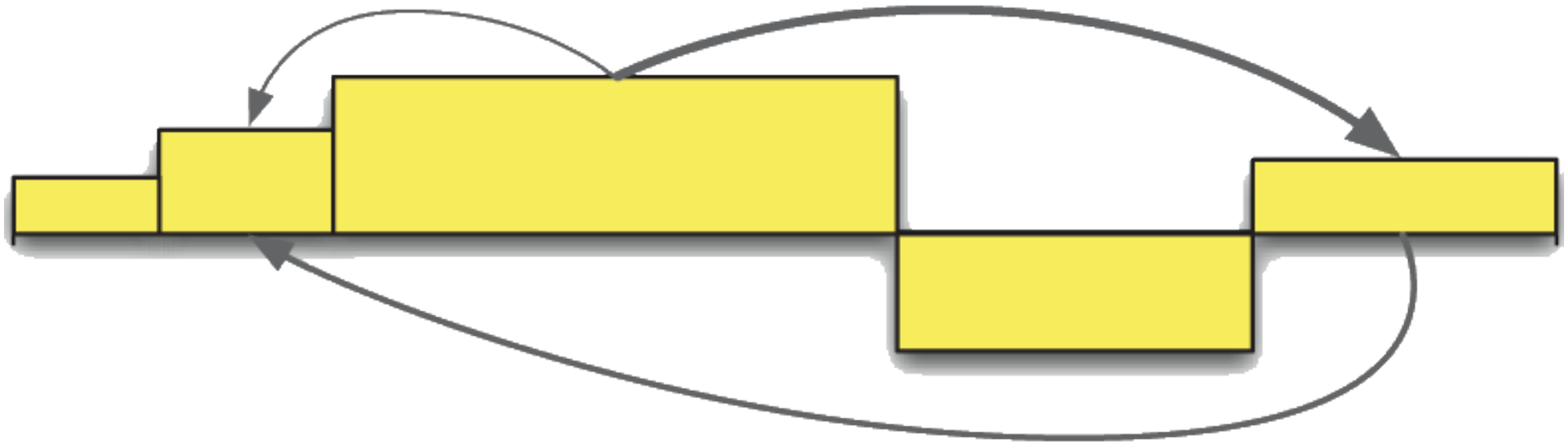
Linked Valued Points (LVP)

Linked Valued Segments



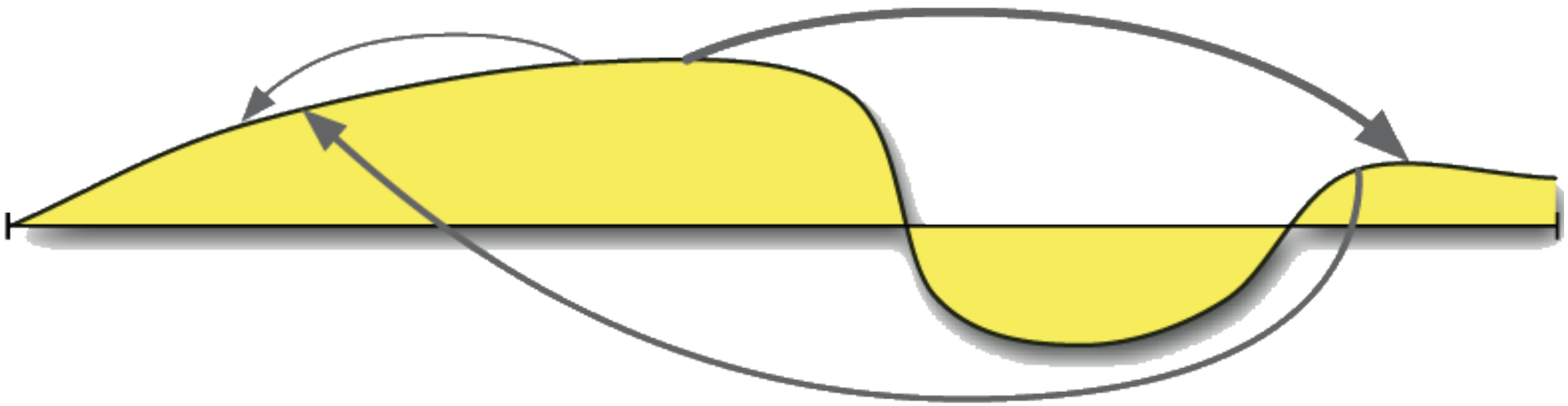
Linked Valued Segments (LVS)

Linked Step Function



Linked Step Function (LSF)

Linked Function



Linked Function (LF)

Linked Base Pairs

Linked Base Pairs (LBP)

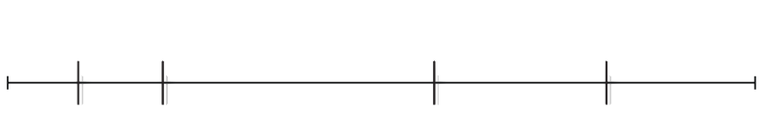
Exercise 3

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

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

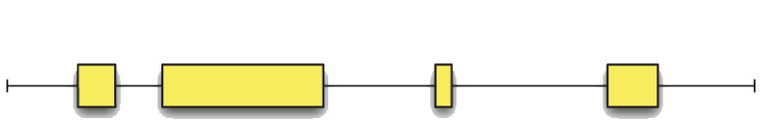
- For each track, which **track type** should be used to represent the data?

Exercise 3

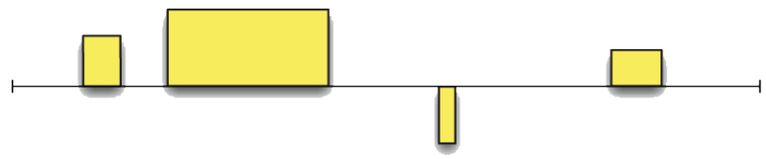


Points (P)

Valued Points (VP)

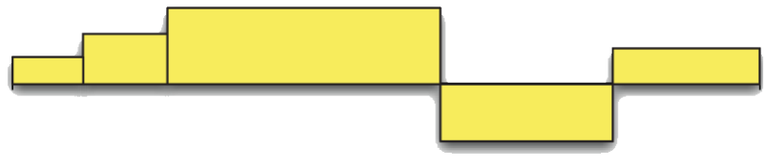


Segments (S)

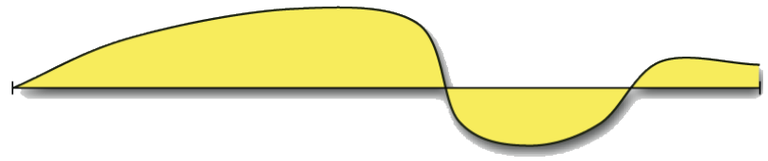


Valued Segments (VS)

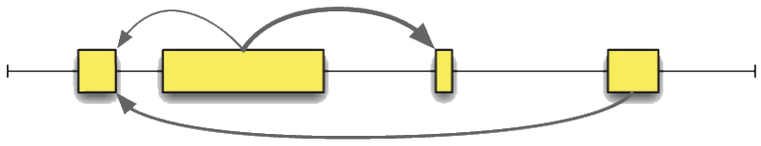
Genome Partition (GP)



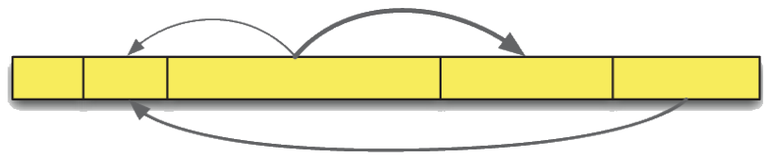
Step Function (SF)



Function (F)

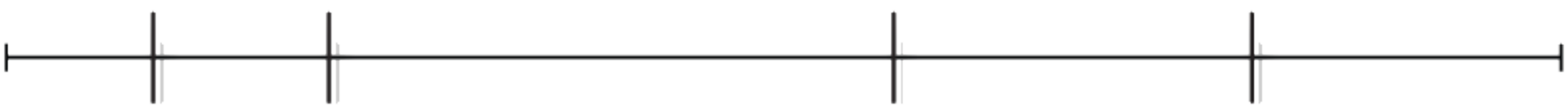


Linked Segments (LS)



Linked Genome Partition (LGP)

Points



Example tracks:

- SNP

Segments

Example tracks:

Genome Partition

Example tracks:

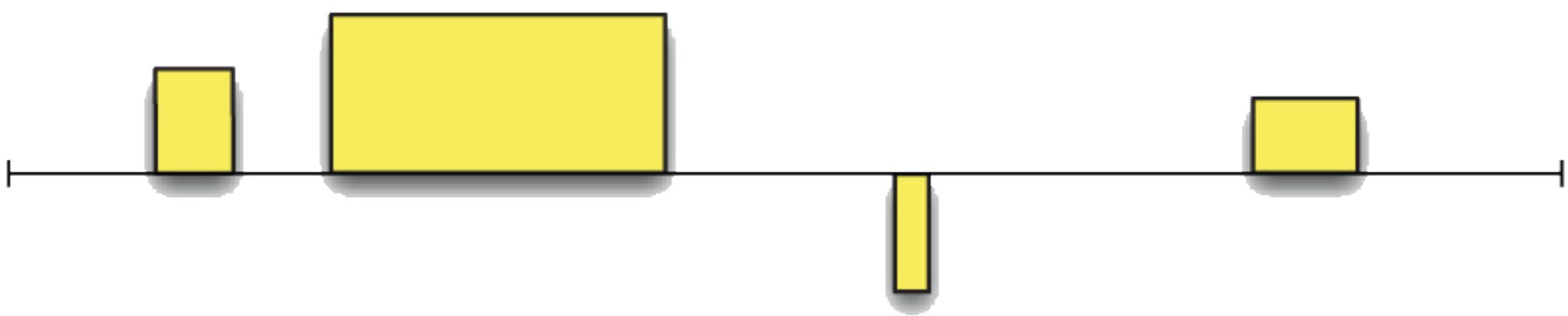
- Chromosomes
- Heterochromatin/Euchromatin

Valued Points

Example tracks:

- SNPs with freq
-

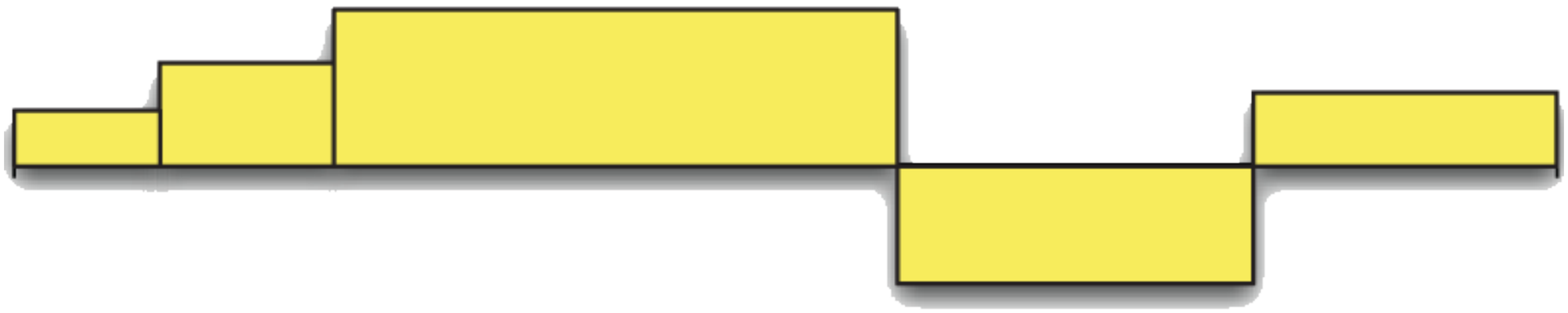
Valued Segments



Example tracks:

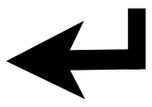
-
-

Step Function

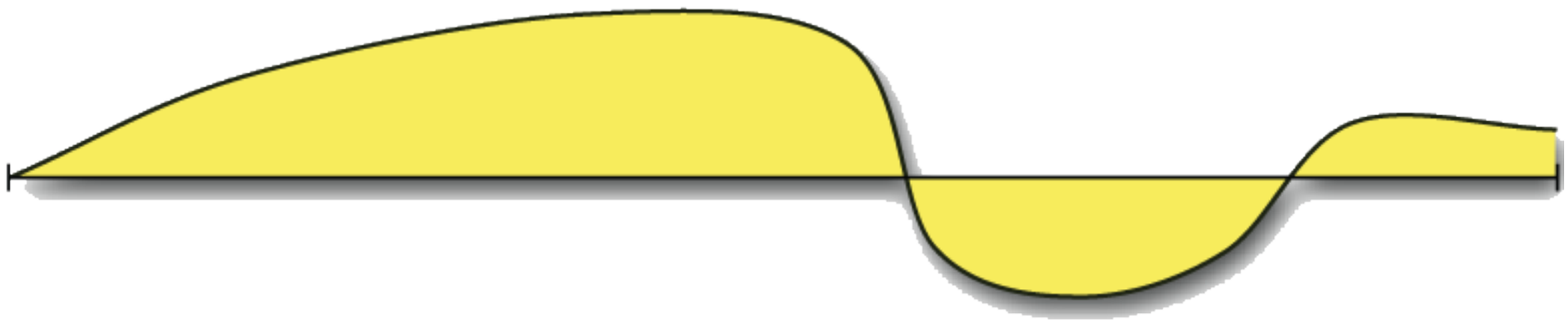


Example tracks:

- GC content



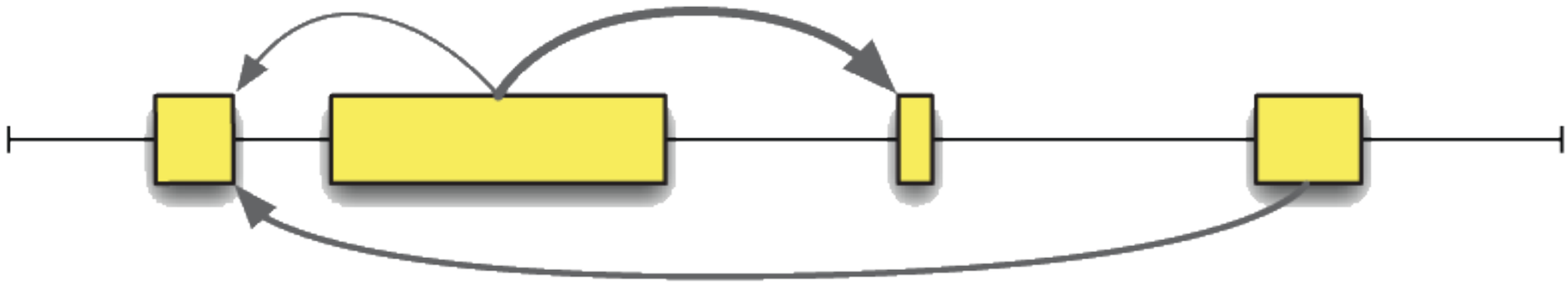
Function



Example tracks:

- DNA melting temp.

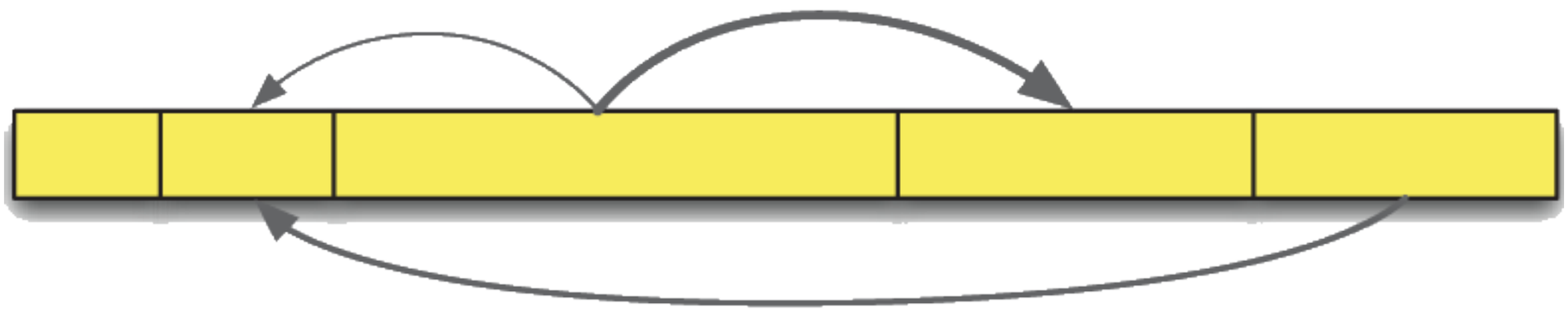
Linked Segments



Example tracks:

- Exons (linked)
- TFs

Linked Genome Partition



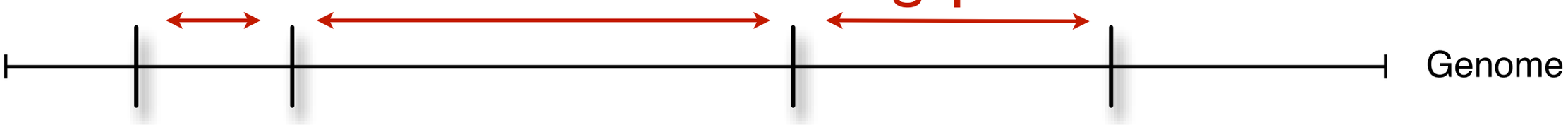
Example tracks:

- HiC data
-
-

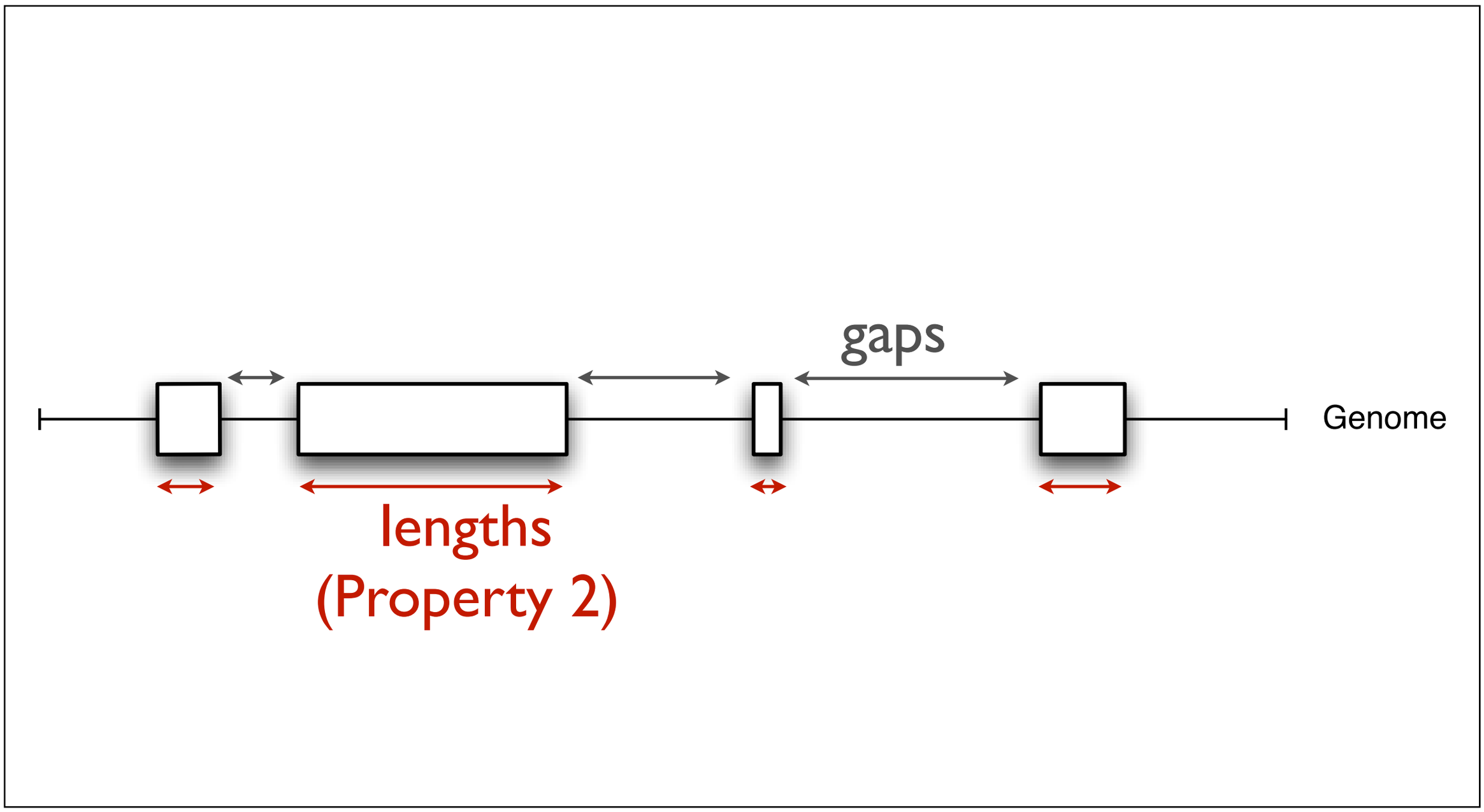
Core properties of tracks

Property 1:

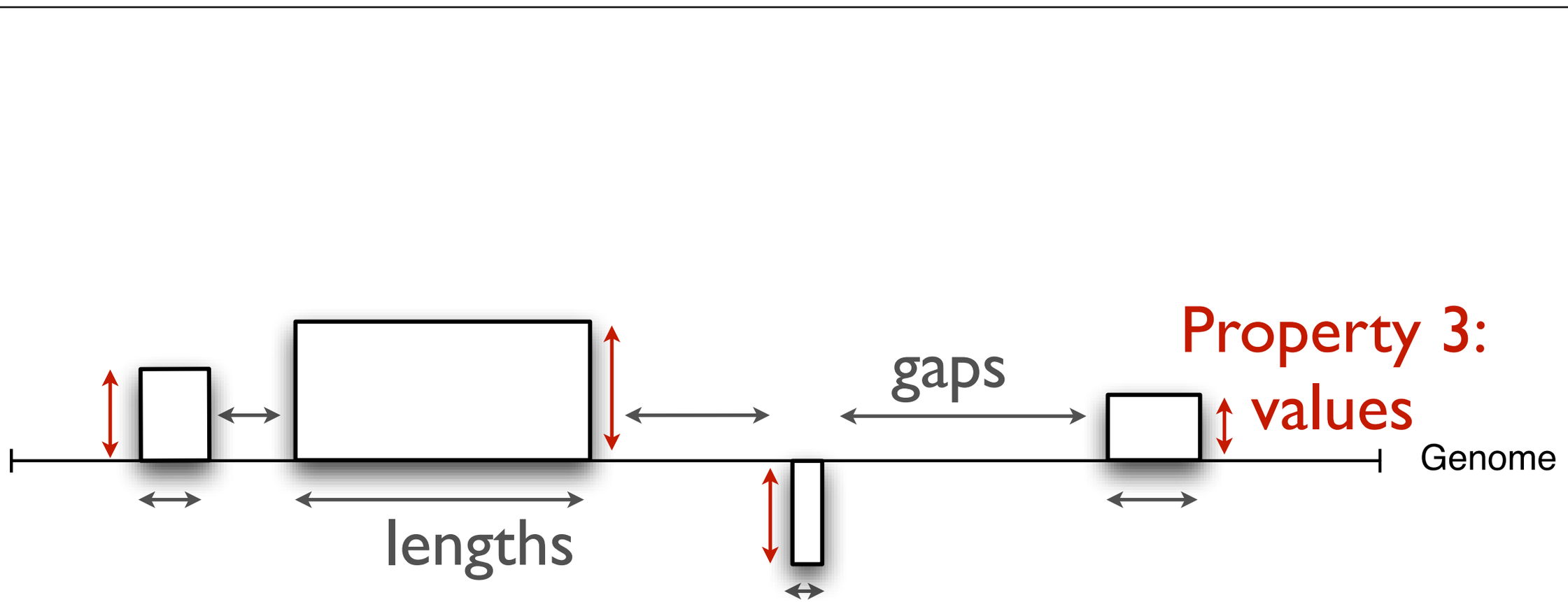
gaps



Core properties of tracks

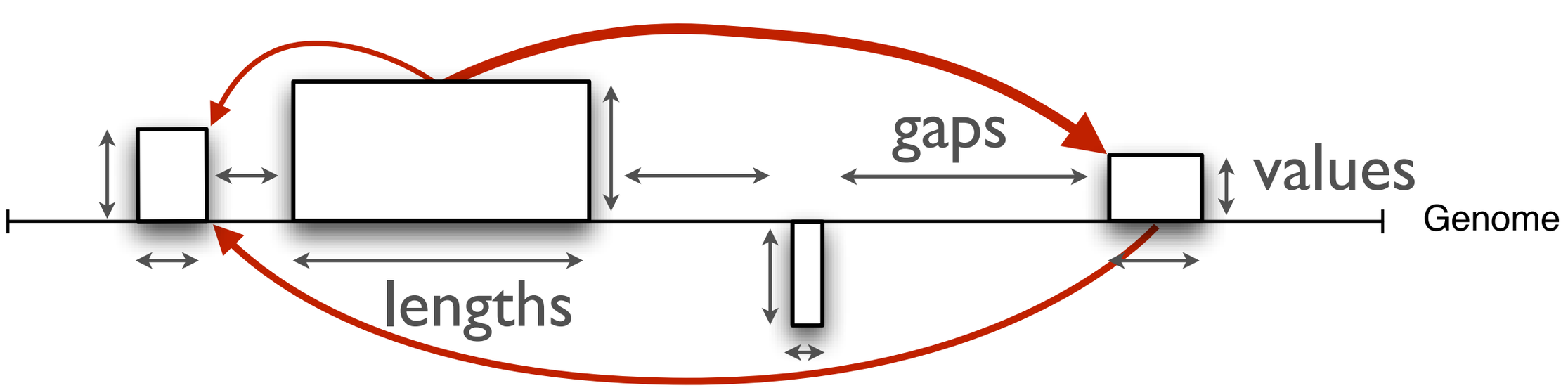


Core properties of tracks



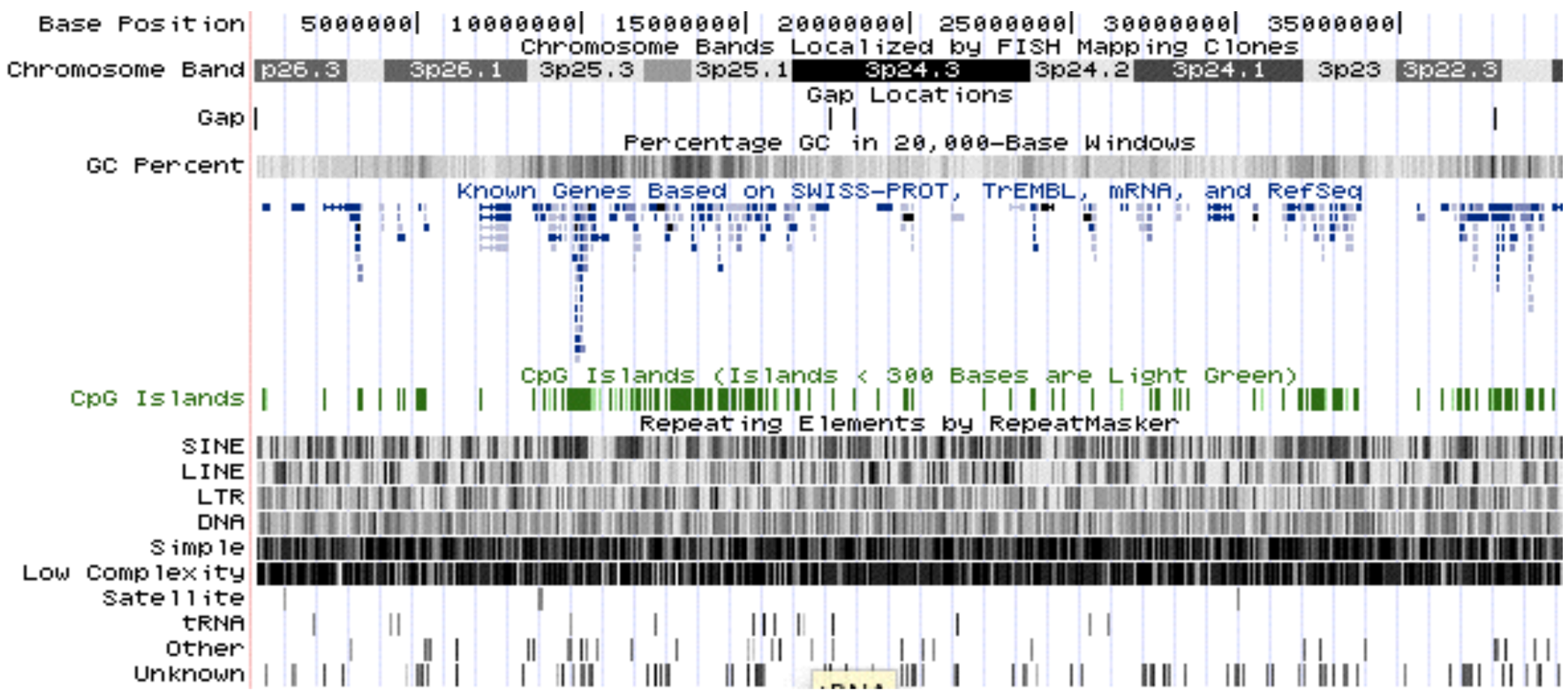
Core properties of tracks

Property 4: interconnections



Tracks in the real world

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



So, what about analysis?

Example analyses

- A relation between methylation patterns and repeating elements? (Genome Res. 2009 19: 221-233)
- Distinct methylation for tissue-specific genes?(Genome Res. 2010 20: 1493-1502)
- Cooperative histone modifications? (Nat Genet 2008 40:897-903)

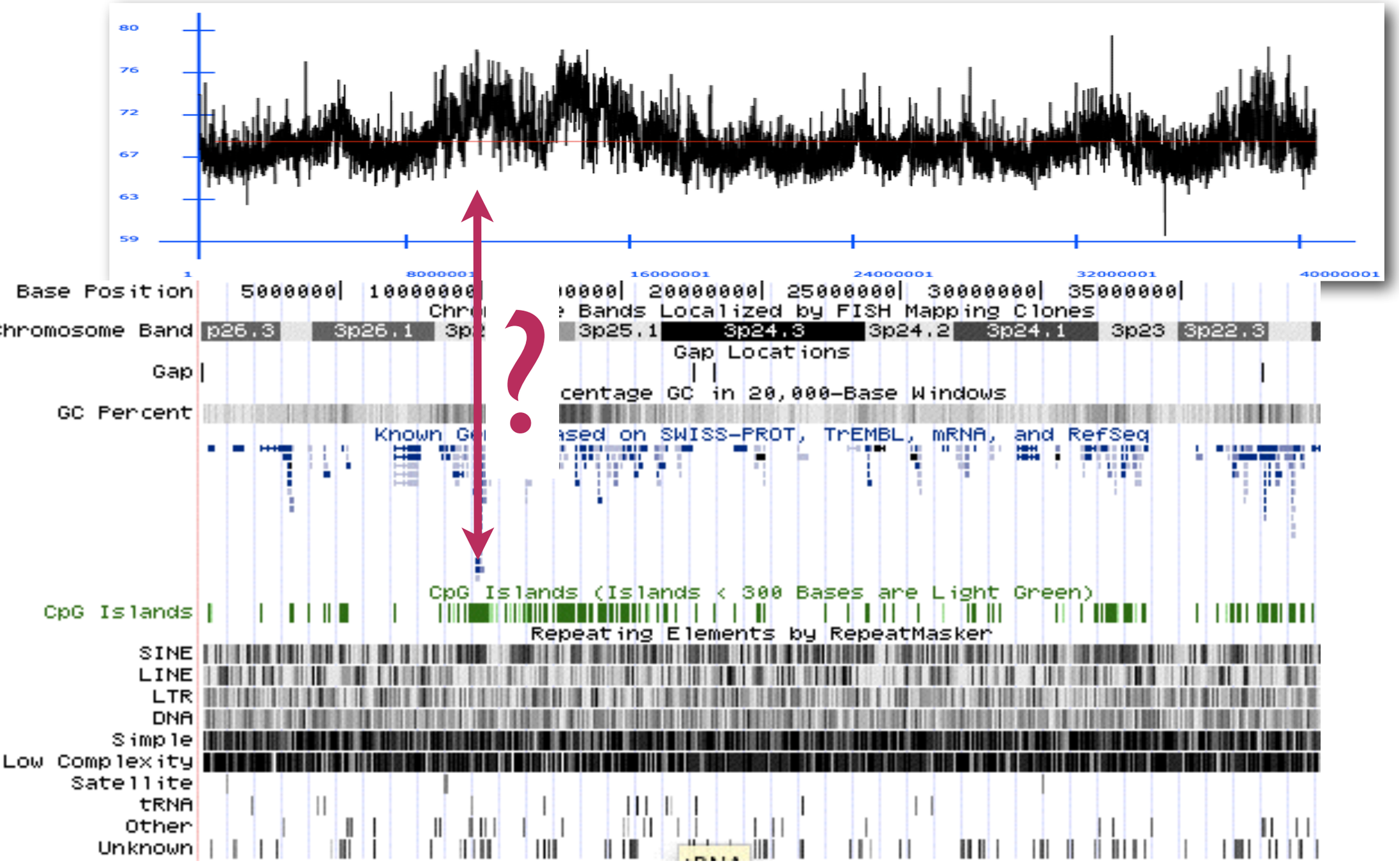
Example analyses (cont.)

- Fragile sites, breakpoints and repeats?
  (Genome Biology 2006 7:R115)
- Copy number variation, repeats, duplications and genes? (Genome Res. 2009 19: 1682-1690)
- Methylation and active genes at T-Cell G0->G1 (Genome Res. 2009 19: 1325-1337)

Example analyses (cont.)

- Virus integration vs genes, CpG, GC-content
  (Journal of Virology 2007 6731–6741)
- Methylation patterns in embryonic cells
  (PNAS 2010 107:10783–10790)

This can't be it?!



Co-occurrence of genomic features

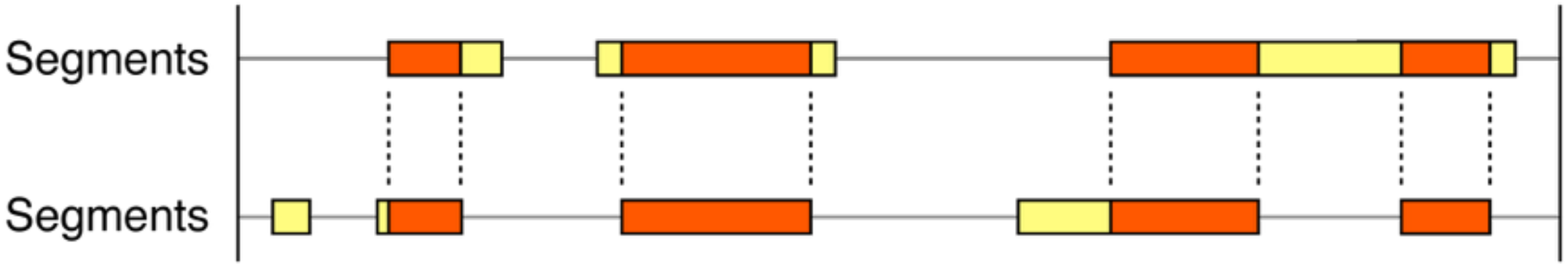
- Typical question:

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

Co-occurrence of genomic features

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

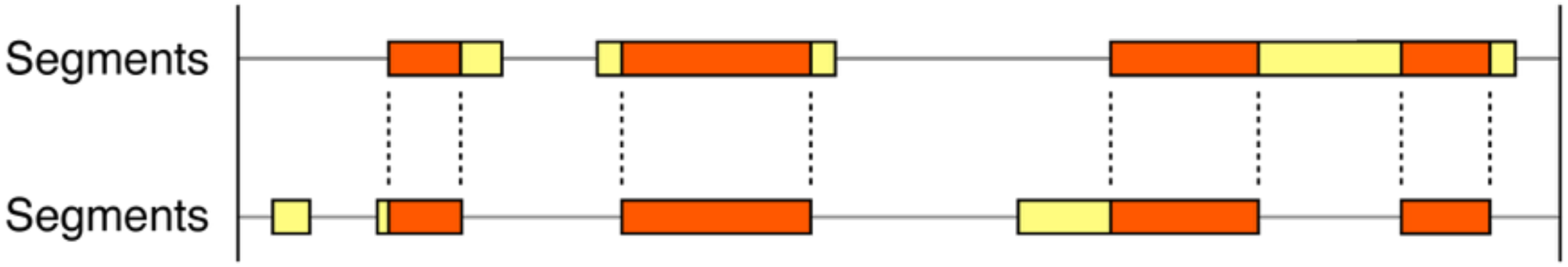
How does this look at the whiteboard?



overlap > expected?

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

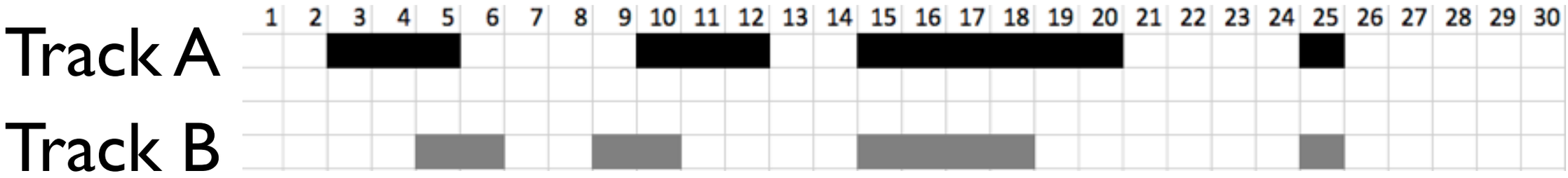
How does this look at the whiteboard?



overlap > expected?

What now?

Exercise 5

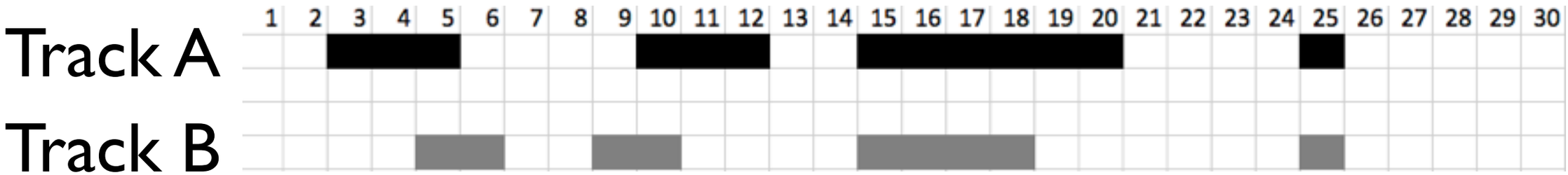


Calculate:

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

What conclusion can you draw from the results?

Exercise 6a



Create a random control track for track B, by

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

Exercise 6a

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

Hypothesis testing

What you will learn:

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

What is a hypothesis test?

- From Wikipedia:
  - *An hypothesis test is a statistical test that is used to determine whether there **is enough evidence in a sample of data** to infer that a certain condition is **true for the entire population**.*

- Based on some data, infer whether a condition is true

- Typically do this by computing the probability of the given data being observed if the condition is false (p-value)

What is a hypothesis test? (cont.)

- Example:
  - Someone claims that they can guess the outcome (head or tail) when a fair coin is flipped.

  - Do some trials, investigate whether this is true
  - You throw the coin 5 times, and the person guesses correct every time.
    - What is the probability of the claim being false?

Example, more formally

- H_0: **Null hypothesis**. This is what we believe until proven wrong
  - *The person is not able to predict tail or head*
- H_1: **Alternative hypothesis**, that we want to investigate
  - *The person is able to predict tail or head for this coin*
- ***P-value**: Given H_0 being true, it is the probability of observing the observed data. If this probability is small enough, we reject the null hypothesis, and conclude H_1.*
  - *In our example: p-value = 0.5^5 = 0.031*

Example, more formally

- **Significance level:** The probability of rejecting the null hypothesis when it is true
  -Choose a significance level before doing the hypothesis test
- ***Test statistic:*** *A value computed based on the data, measuring the wanted effect.*
  -E.g. mean, a count, number of base pairs overlap

Why use hypothesis tests?

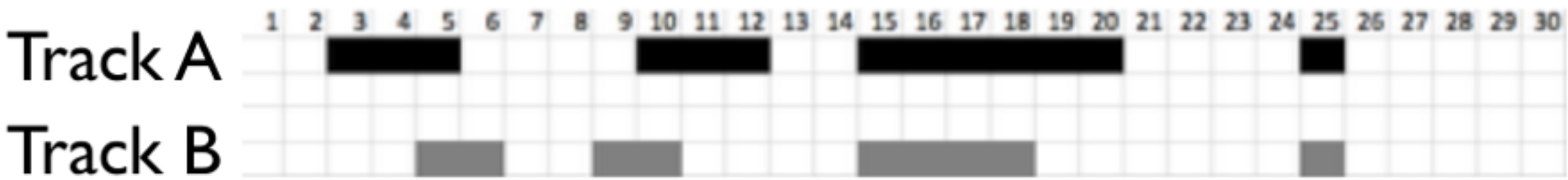
- Sometimes hard or impossible to make conclusions without.
  - What if the person guessed correct 520 out of 1000 times?
  - Even harder when working with biological data
- A hypothesis test quantifies the certainty of concluding a hypothesis (p-value)
  - For some cases, a very small p-value might be requested, e.g when concluding on the effect of a drug

Hypothesis test – general procedure

1) Define your null hypothesis and alternative hypothesis
2) Find a good test statistic that captures what you want to investigate
3) Select a significance level
4) Find the probability of observing the given data given that the null hypothesis is true (this requires a null model)

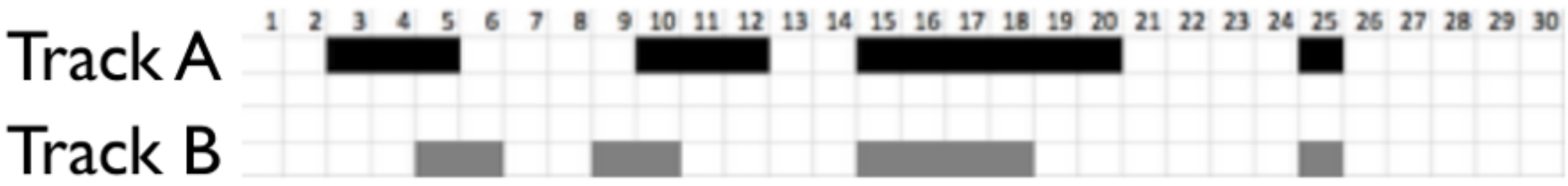
Null model

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



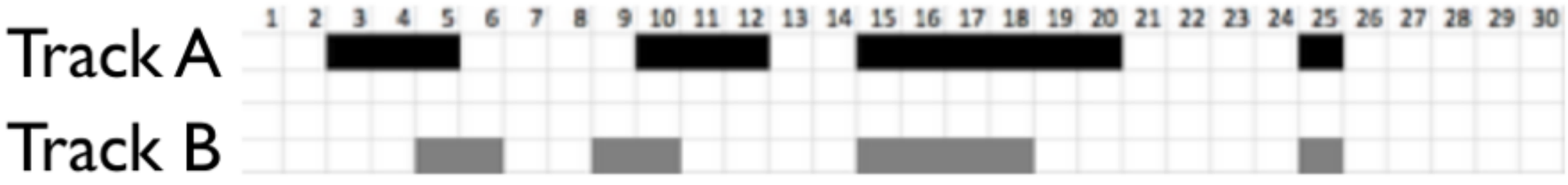
A more “complicated” claim

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



Monte Carlo

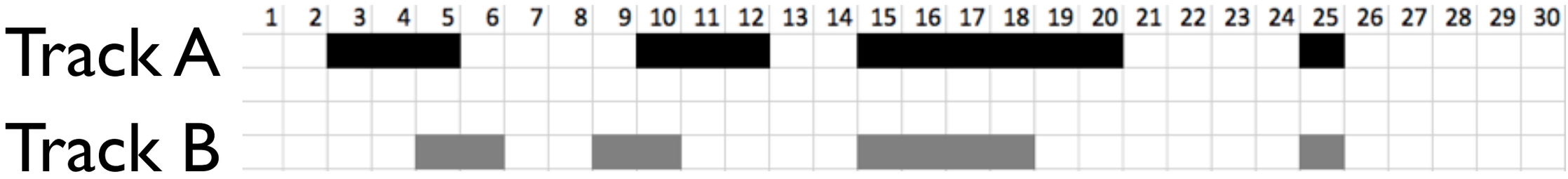
- Simulate many samples from the null model
  - E.g. many pairs of tracks following the same properties
- For each simulation, compute the co-occurrence
  - E.g. the number of base pairs overlap
- Compute how often the co-occurrence found using the null model was as extreme or more extreme than the co-occurrence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



How to make random samples in this case?

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

Exercise 6b



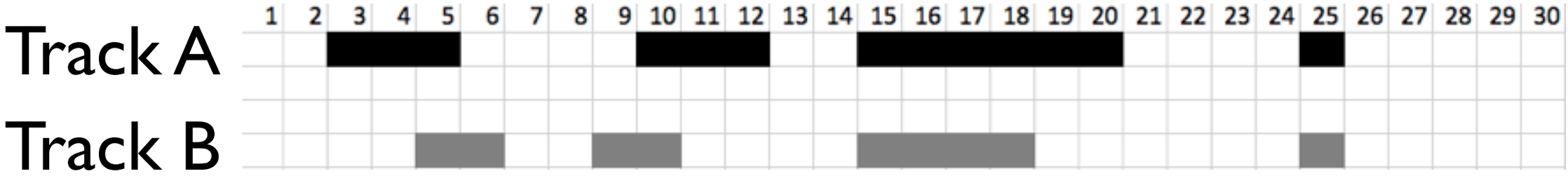
Create a random control track for track B, by

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

Exercise 6b

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

Remember this?



Calculate:

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

What conclusion can you draw from the results?

Null models (2/2)

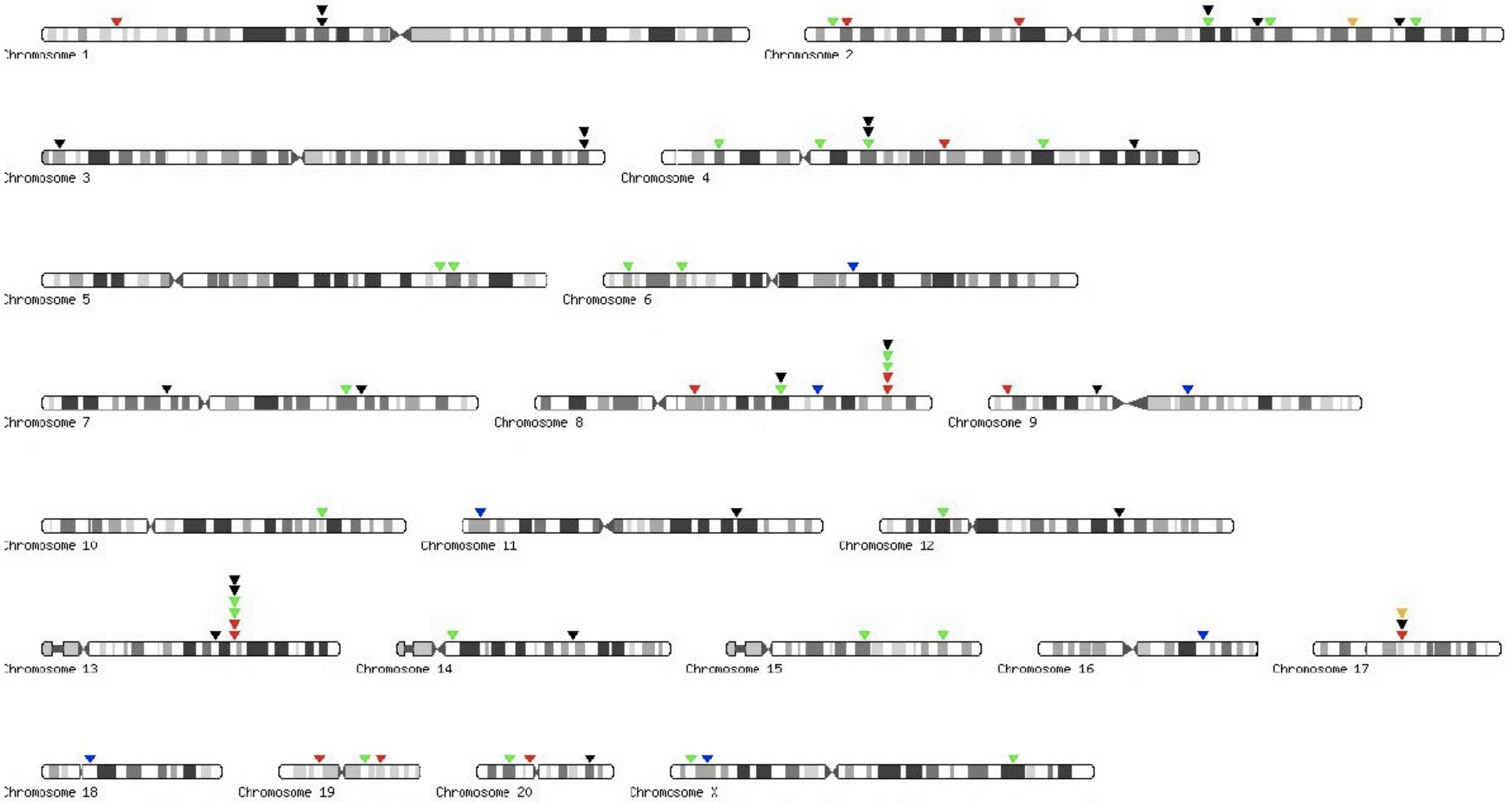
- Examples of preservation strategies
  - Preserve segment length (already seen this)
  - Preserve segment and gap length (this too)
  - For points (segments with length 1)
    - Preserve point count
    - Preserve inter-point distance
  - For all this cases we randomize the position of the track elements.

Association vs. causation

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

Interpreting a claim

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



HPV integration sites

Interpreting a claim

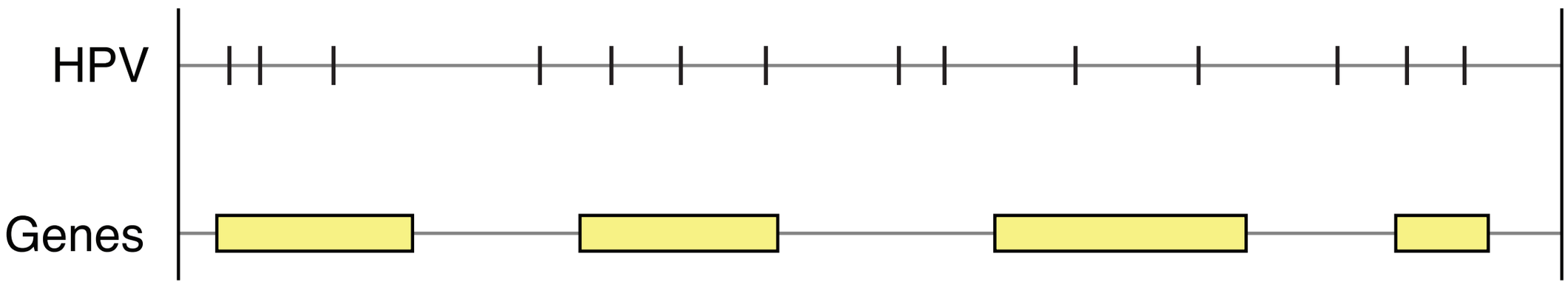
"Viruses might be expected to integrate near genes. Our results confirm such preferential localization inside genes for HPV [Human Papillomavirus], where 75 out of 119 determined integration sites (attached) fall inside genes."

How would you go forth in reproducing such a claim?

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

Exercise 7: HPV and genes

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



Note down (in silence):

1. Which test statistic would you choose?

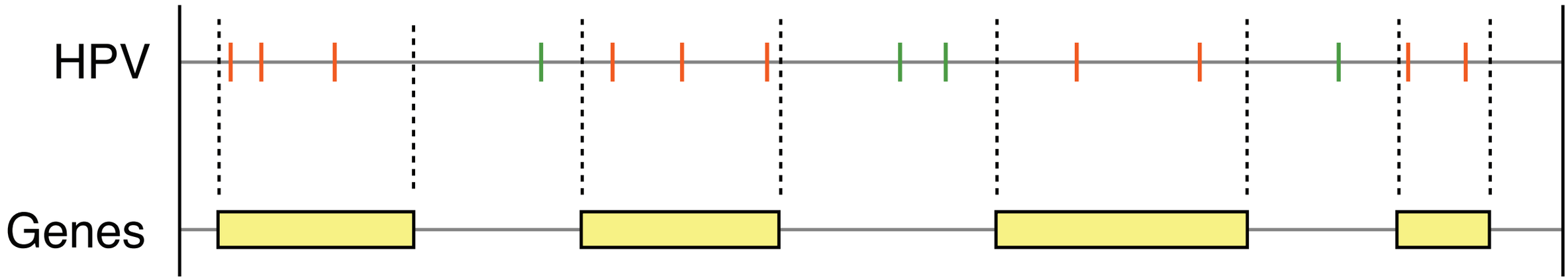
Exercise 7: HPV and genes

Student answers:

1. Which test statistic would you choose?

Proportion of HPV in Genes	6	
Count HPV divided to Gene coverage proportion	3	
Count HPV inside Genes	5	
Count Genes containing HPV sites	2	
Observed vs expected		

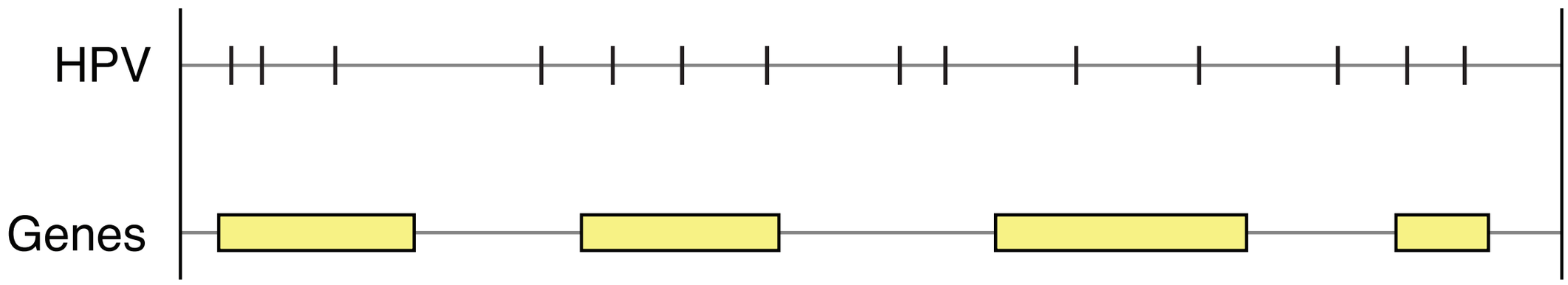
A possible test statistic



- 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

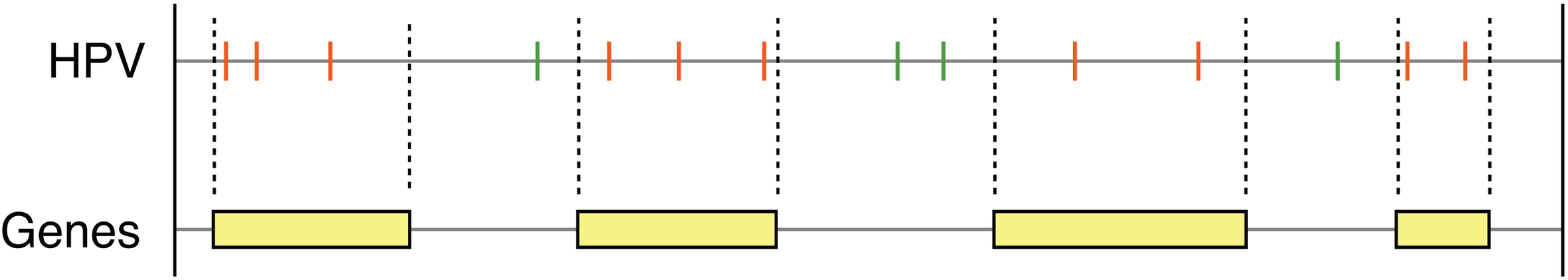
Exercise 8: HPV and genes

Student answers:

2. Which null model would you choose?

Preserve Genes and randomise HPV	7	
Preserve Genes and inter-point distance randomize HPV	4	
Preserve HPV point randomize Genes	0	
Preserve Genes and minimum interpoint distance	1	
none	9	

Exercise 9: HPV and genes



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

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

Exercise 9: HPV and genes

Student answers:

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

Preserve segments and nr of point, randomize points	0.5	Refseq
Preserve segments and nr of point, randomize points (MC)	0.45	Refseq
Preserve segments and nr of point, randomize points (MC)	0,006	Ensem
Preserve segments and nr of point, randomize points (MC)	0,049	Ensem
Preserve segment and interpoint distance, randomize points (MC)	0.47	Refseq
Preserve segment and interpoint distance, randomize points (MC)	0,02	Ensem
Preserve segment and interpoint distance, randomize points (MC)	0,42	Refseq

How much of the human genome is covered by genes?

Exercise 10: descriptive statistics

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

  RefSeq: 1 216 642 705

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

  RefSeq: 0.4254

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

  1 196 508 344 (41.84%)

Descriptive statistics

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

Making justified choices is indeed hard!

- The choice of data may influence results
  - Both source and exact version of genes might matter
  - Can sometimes justify e.g. how strict definition of a gene one should use
  - One should ideally show how results vary with choice of data
  - Should at least be very precise in what was done (accessibility, transparency, reproducibility)

Making justified choices is indeed hard (2)

- There is usually more than one possible test for a given biological question
- The choice has to be made,
  and can't be resolved automatically
- Statistical and biological implications play together to determine what may be reasonable
- Should at least expose the different possibilities

Making justified choices is indeed hard (3)

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

An example of alternative assumptions

- Duan, [...] and Noble (Nature, 2011):
  - Extensive significant 3D co-localization of functional elements, assessed by hypergeometric distribution
- Witten and Noble (NAR, 2012):
  - Hypergeometric had unrealistic implications. Telomeres and breakpoints may not be co-located after all.. (cancelled 4 of 11 findings)

Any rules of thumb?

(for the statistical testing)

- Maybe:
  - Use test-statistic that gives best (lowest) p-value
  - Use null model that gives worst (highest) p-value
- Reasoning:
  - Use measure that best catches relation of interest
  - Use the most realistic model of nature (null model)
- Always:
  - Double-check with a statistician (and a biologist, if you are not one)

Reproducibility

My claim:

Bioinformaticians

(esp. those with a biology background)

are too fond of the command line!

The command-line approach to bioinformatics

- We want to run a tool, say Bowtie
- Try: "bowtie"
- "module load bowtie"
- Try: "bowtie" (Yes, it's there)
- What were the options?
- "bowtie -h"

bowtie -h

```
Usage:
  bowtie [options]* <ebwt> {-1 <m1> -2 <m2> | --12 <r> | <s>} [<hit>]

  <m1>    Comma-separated list of files containing upstream mates (or the
          sequences themselves, if -c is set) paired with mates in <m2>
  <m2>    Comma-separated list of files containing downstream mates (or the
          sequences themselves if -c is set) paired with mates in <m1>
  <r>     Comma-separated list of files containing Crossbow-style reads.  Can be
          a mixture of paired and unpaired.  Specify "-" for stdin.
  <s>     Comma-separated list of files containing unpaired reads, or the
          sequences themselves, if -c is set.  Specify "-" for stdin.
  <hit>   File to write hits to (default: stdout)
Input:
  -q                 query input files are FASTQ .fq/.fastq (default)
  -f                 query input files are (multi-)FASTA .fa/.mfa
  -r                 query input files are raw one-sequence-per-line
  -c                 query sequences given on cmd line (as <mates>, <singles>)
  -C                 reads and index are in colorspace
  -Q/--quals <file>  QV file(s) corresponding to CSFASTA inputs; use with -f -C
  --Q1/--Q2 <file>   same as -Q, but for mate files 1 and 2 respectively
  -s/--skip <int>    skip the first <int> reads/pairs in the input
  -u/--qupto <int>   stop after first <int> reads/pairs (excl. skipped reads)
  -5/--trim5 <int>   trim <int> bases from 5' (left) end of reads
  -3/--trim3 <int>   trim <int> bases from 3' (right) end of reads
  --phred33-quals    input quals are Phred+33 (default)
  --phred64-quals    input quals are Phred+64 (same as --solexa1.3-quals)
  --solexa-quals     input quals are from GA Pipeline ver. < 1.3
  --solexa1.3-quals  input quals are from GA Pipeline ver. >= 1.3
  --integer-quals    qualities are given as space-separated integers (not ASCII)
```

bowtie -h

```
Alignment:
  -v <int>           report end-to-end hits w/ <=v mismatches; ignore qualities
    or
  -n/--seedmms <int> max mismatches in seed (can be 0-3, default: -n 2)
  -e/--maqerr <int>  max sum of mismatch quals across alignment for -n (def: 70)
  -l/--seedlen <int> seed length for -n (default: 28)
  --nomaqround       disable Maq-like quality rounding for -n (nearest 10 <= 30)
  -I/--minins <int>  minimum insert size for paired-end alignment (default: 0)
  -X/--maxins <int>  maximum insert size for paired-end alignment (default: 250)
  --fr/--rf/--ff     -1, -2 mates align fw/rev, rev/fw, fw/fw (default: --fr)
  --nofw/--norc      do not align to forward/reverse-complement reference strand
  --maxbts <int>     max # backtracks for -n 2/3 (default: 125, 800 for --best)
  --pairtries <int>  max # attempts to find mate for anchor hit (default: 100)
  -y/--tryhard       try hard to find valid alignments, at the expense of speed
  --chunkmbs <int>   max megabytes of RAM for best-first search frames (def: 64)
Reporting:
  -k <int>           report up to <int> good alignments per read (default: 1)
  -a/--all           report all alignments per read (much slower than low -k)
  -m <int>           suppress all alignments if > <int> exist (def: no limit)
  -M <int>           like -m, but reports 1 random hit (MAPQ=0); requires --best
  --best             hits guaranteed best stratum; ties broken by quality
  --strata           hits in sub-optimal strata aren't reported (requires --best)
Output:
  -t/--time          print wall-clock time taken by search phases
  -B/--offbase <int> leftmost ref offset = <int> in bowtie output (default: 0)
  --quiet            print nothing but the alignments
  --refout           write alignments to files refXXXXX.map, 1 map per reference
  --refidx           refer to ref. seqs by 0-based index rather than name
```

bowtie -h

```
  --al <fname>       write aligned reads/pairs to file(s) <fname>
  --un <fname>       write unaligned reads/pairs to file(s) <fname>
  --max <fname>      write reads/pairs over -m limit to file(s) <fname>
  --suppress <cols>  suppresses given columns (comma-delim'ed) in default output
  --fullref          write entire ref name (default: only up to 1st space)
Colorspace:
  --snpphred <int>   Phred penalty for SNP when decoding colorspace (def: 30)
     or
  --snpfrac <dec>    approx. fraction of SNP bases (e.g. 0.001); sets --snpphred
  --col-cseq         print aligned colorspace seqs as colors, not decoded bases
  --col-cqual        print original colorspace quals, not decoded quals
  --col-keepends     keep nucleotides at extreme ends of decoded alignment
SAM:
  -S/--sam           write hits in SAM format
  --mapq <int>       default mapping quality (MAPQ) to print for SAM alignments
  --sam-nohead       supppress header lines (starting with @) for SAM output
  --sam-nosq         supppress @SQ header lines for SAM output
  --sam-RG <text>    add <text> (usually "lab=value") to @RG line of SAM header
Performance:
  -o/--offrate <int> override offrate of index; must be >= index's offrate
  -p/--threads <int> number of alignment threads to launch (default: 1)
  --mm               use memory-mapped I/O for index; many 'bowtie's can share
  --shmem            use shared mem for index; many 'bowtie's can share
Other:
  --seed <int>       seed for random number generator
  --verbose          verbose output (for debugging)
  --version          print version information and quit
  -h/--help          print this usage message
```

At last....

- Call “bowtie /path/input.fastq ...(a bunch and of options and some some, and even more options)... > /path/to/bowtieLog.txt 2>&1 &”
- We get back to it next morning

Now isn't this good enough ?!

Log in to server. Profile OK?

Confusing and error-prone

- Call "bowtie /path/input.fastq ...(a bunch and of options and some some, and even more options)... > /path/to/bowtieLog.txt 2>&1 &"
- We get back to it next morning

How was this again?

How to keep this running when I log off? nohup? screen?

Will I remember? Will it be ready then?

Where did I log to this time?

But I wanted to run it on the cluster!!

- How were those SLURM things again?...

Reproducibility

- Bioinformaticians gets surprised every time they need to redo/modify previous analyses
- But lab biologists already know the importance of reproducibility!
- They also know that even with a detailed lab journal, reproduction is a challenge
- The question is then how this manifests itself when doing analysis on a computer

What is *in silico* reproducibility?

- Basically the same issues as in the lab:
  - Materials -> Data sources
  - Experiment conditions -> Analysis parameters
  - Equipment (with model number) -> Programs (with version number)
- And the same challenges:
  - Are all relevant conditions described accurately?
  - Will the same materials and equipment be available?

What is the status of reproducibility in the literature?

- Less than half of selected microarray experiments published in Nature Genetics could be reproduced
  (Ioannidis et al., Nat Genet 2009)

- More than half [of surveyed papers] do not provide primary data and list neither the version nor the parameters used [for read mapping]
  (Nekrutenko and Taylor., Nat Rev Genet 2012)

Why should you care?
(about making your analyses reproducible)

- Because it's the right thing to do!
- Journals are becoming aware of the issues
- Reviewers will value it
- But the main argument:
  - The one that's struggling with reproducing your analysis is often the future *you*

Galaxy supports reproducibility

- Automatically tracks *metadata* at every step
  - Which are the datasets?
  - What are the parameters?
  - Which tools, and which version of the tool?
  - What are the outputs
- Users can annotate the steps to capture the *intent* of the analysis!

Galaxy supports reproducibility

- All jobs can be rerun later, by independent scientists
- Galaxy automatically captures data inputs, analysis parameters, program versions, and intermediate data
- Workflows capture common analysis sequences, *i.e.* typical experimental setups. Can be reused for other datasets and experiments
- The Genomic HyperBrowser is built on top of Galaxy, and thus keeps all its functionality for reproducible research

Galaxy pages

- A Galaxy page is a solution in Galaxy to support sharing and publication of results
- A Galaxy page is somewhat like an ordinary web page (and can be accessed from a URL), with formatting:
    - Headers
    - Paragraphs of texts
    - Images...
- In addition, you can add Galaxy datasets, histories and workflows.
    - In this way, you can share your results and the parameters used. Other researcher can easily rerun your jobs with the same or other options.

Exercise 13

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

Ten simple rules for reproducibility

- For more reading on reproducibility (tips and concepts), read:

"Ten simple rules for reproducible computational research." Sandve, Geir Kjetil, et al., PLOS Computational Biology (2013): e1003285.

Conclusion

Main conclusions

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

Any questions?

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