IN-BIOS5000/9000

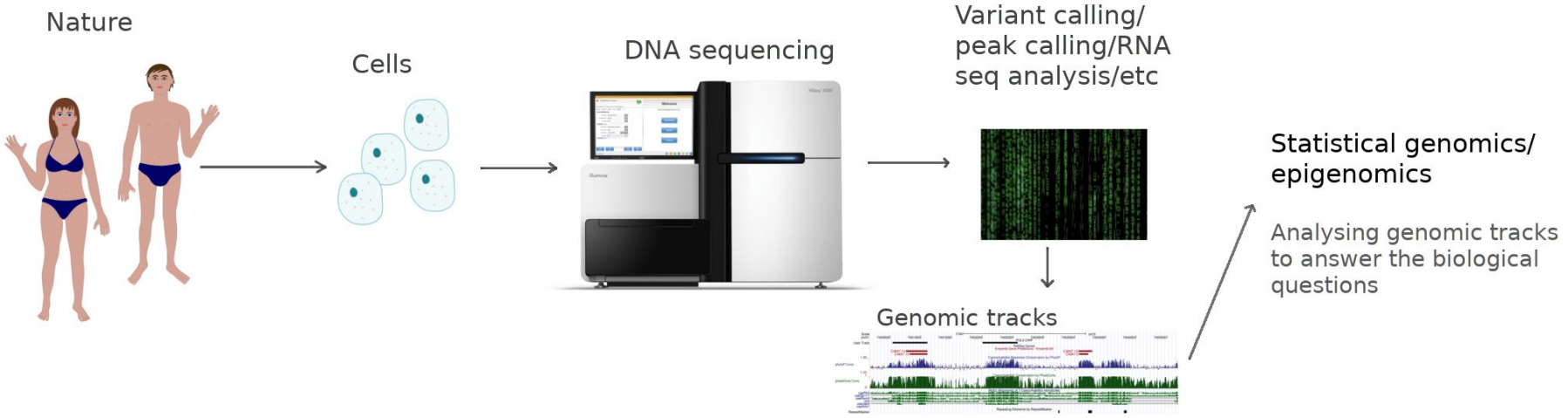
Statistical (epi)genomics

Oktober 27-28 2020, Ivar Grytten

Welcome!

- Feel free to ask questions any time in the chat! I'll check it now and then.
- Use Mattermost (especially if you have questions when I'm not on Zoom)
- The Zoom lecture will be recorded, but I will edit out any students speaking before uploading

What is statistical genomics/epigenomics?



Using **statistics** to answer biological questions by analysing **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.

What kind of tools will we use?

- Some of you have experience with programming and some have not
- In order to try to make this fun for everyone, there will be an option on many exercies:
  - Either use Python
  - .. or use tools that require no programming skills
  - .. or try both
- The idea is that those of you who want to learn more programming will be able to do that, and that those who are already experienced can get some fun challenges and not be bored
- However: Learning the exact tools and/or programming techniques is not the essential part of this module, I mostly want you to learn the underlying principles

How will we do this?

- I will present the topics
  - There will be a few interruptions with small exercises inbetween
- You will do some bigger exercises using what you have learned from the topics presented
- We repeat this on new topics
  - The topics build on each other

What do we mean with the word “statistic”?

- The same as Wikipedia:

  “A statistic (singular) or sample statistic is any quantity computed from values in a sample that is used for a statistical purpose”

- The number of genes in a dataset is a statistic. Average gene length is another

- Sometimes we also talk about “statistics” (the field)

Learning outcomes

In general:

- Become better at bioinformatics “problem solving”
  - dataset, hypothesis, ideas -> reasoning, critical thinking -> applying methods and tools -> discussing the findings
- Get better at statistical thinking
- Be able to apply the methods and tools you learn to solve real problems

More specific, you will learn about:

- Genomic tracks, how we represent genomic/epigenomic datasets on the computer
- Descriptive statistics for analysing such tracks
- Hypothesis testing (on genomic tracks)
- Learn to use some tools: Galaxy, Bedtools, The Genomic Hyperbrowser, Python

Approximate schedule

Day 1:

- **09:15-10:45:** Introduction and methodology (reference genomes, tracks, basic analysis, with some short exercises inbetween)
- **11:00-11:15:** Introduction to the tools we will be using
- **11:15-13:30:** Exercises (you work individually, and take a lunch break when you want)
- **13:30-13:45:** We go through the exercises together
- **13:45-14:30:** Hypothesis testing
- **14:30-15:30** Exercises (you work individually)

Day 2:

- **09:15-10:00:** We go through the exercise from the day before, and recap day 1
- **10:00-10:30:** More theory (statistics, monte carlo simulation)
- **10:30-11:00:** Multi-track analysis
- **11:00-13:00:** Exercices (you work individually, and take lunch break when you want)
- **13:00-14:00:** We go through the exercices togheter
- **14:15-15:30:** Summary/discussion

About the exercises

- You will be working on these individually and I will be available on Zoom (as if you were in a class room)
- It is important that you ask questions whenever you need help (on Zoom, email or chat)
  - I can help students individually and will probably explain stuff on Zoom while you are working
- The exercises these two days will follow the same “case”
  - Investigating SNPs known to be associated with the disease Multiple Sclerosis

Before we start, just a few poll questions to get to know you

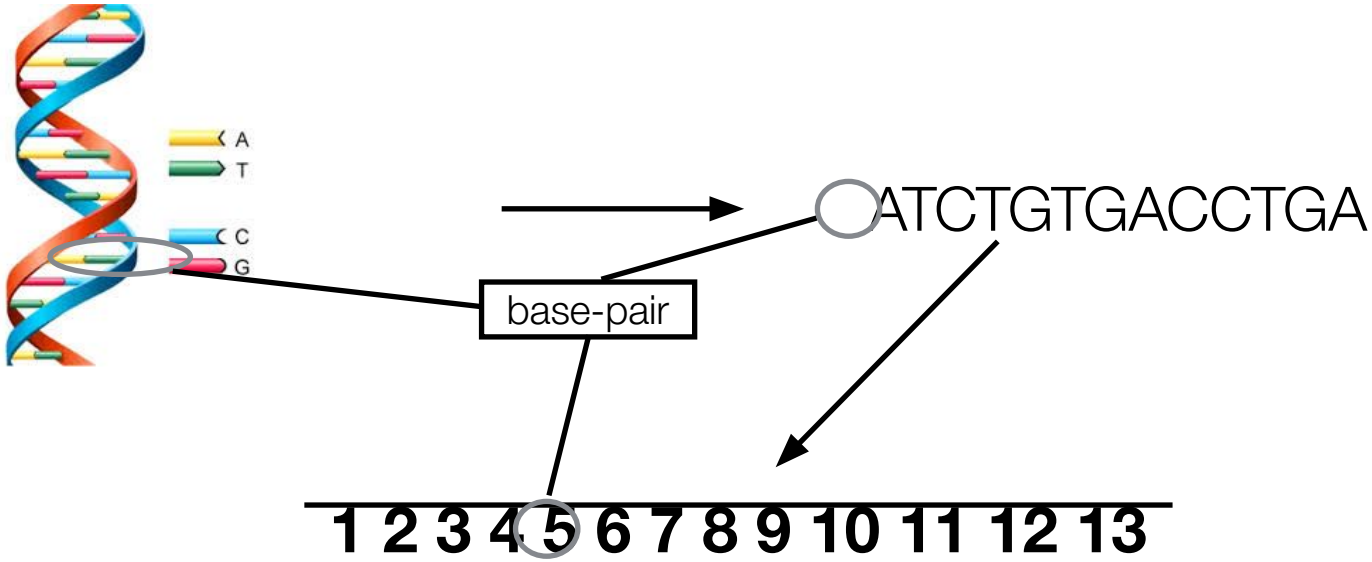
Part 1

Models and data, 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?

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)

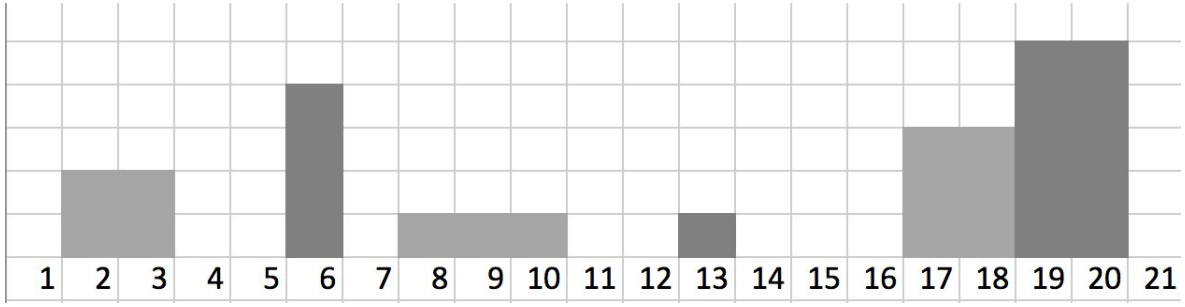
- It is a simple and compact way of representing elements.
  - You only need the reference genome coordinates, and then you know the sequence within the elements.
- Makes it quite easy to compute the distance between genomic elements

Genomic tracks

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



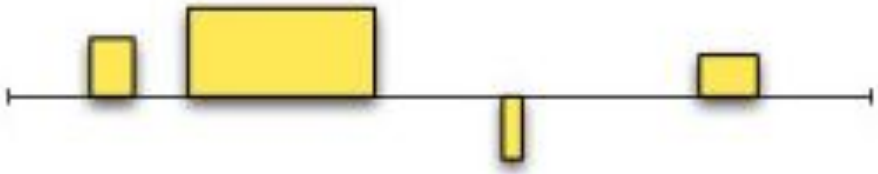
Compute the following statistics:

a) Number of base-pairs covered — 11
b) Proportion of genome covered — 11/21 or 0.52 or 52%
c) Average segment length — 1.83
d) Average gap length — 1.43
e) Average value — 1.33 per bp (or 1.6); 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

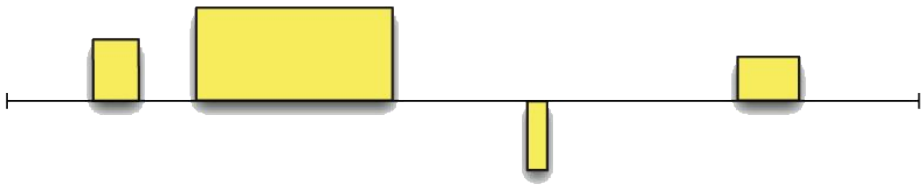
How about gene expression data (found by doing RNA-seq)?



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

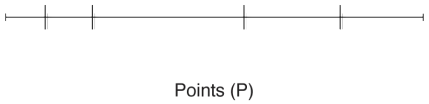
- In the last example, we showed genes as segments on the genome line, with attached RNA-seq read count values
- This track is of a **track type** we call “valued segments”



Valued Segments (VS)

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

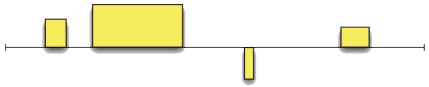
Track types



Points (P)

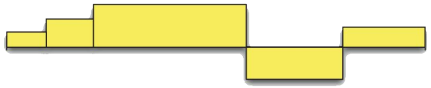
Valued Points (VP)

Segments (S)

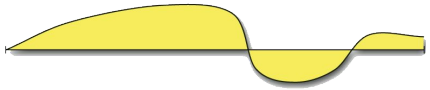


Valued Segments (VS)

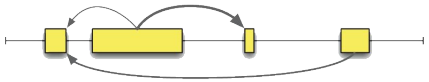
Genome Partition (GP)



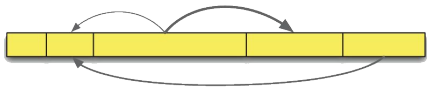
Step Function (SF)



Function (F)



Linked Segments (LS)

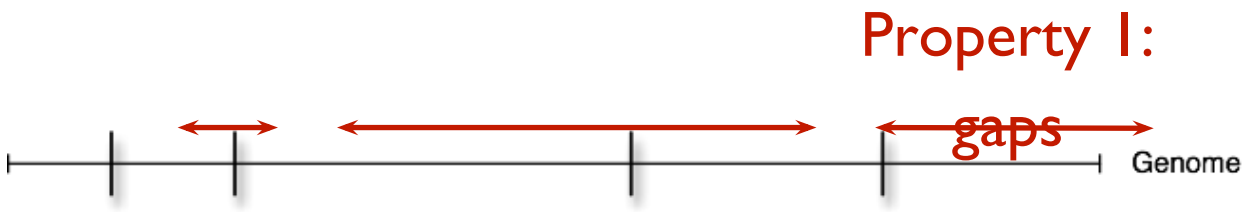


Linked Genome Partition (LGP)

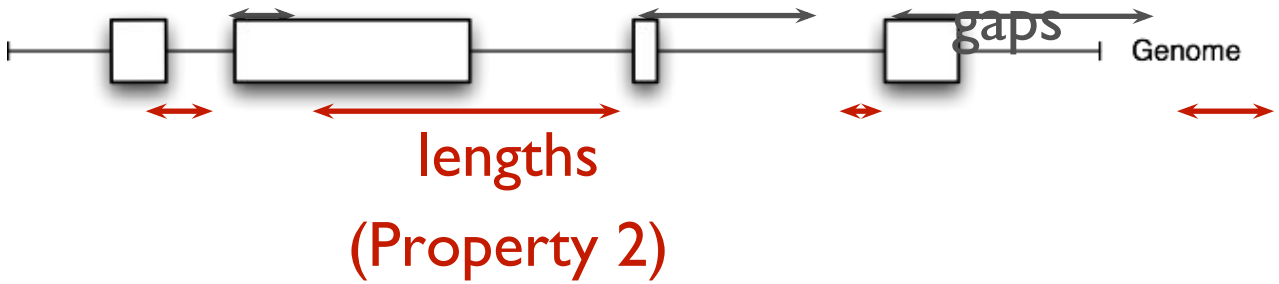
Typical genomic tracks

- 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...

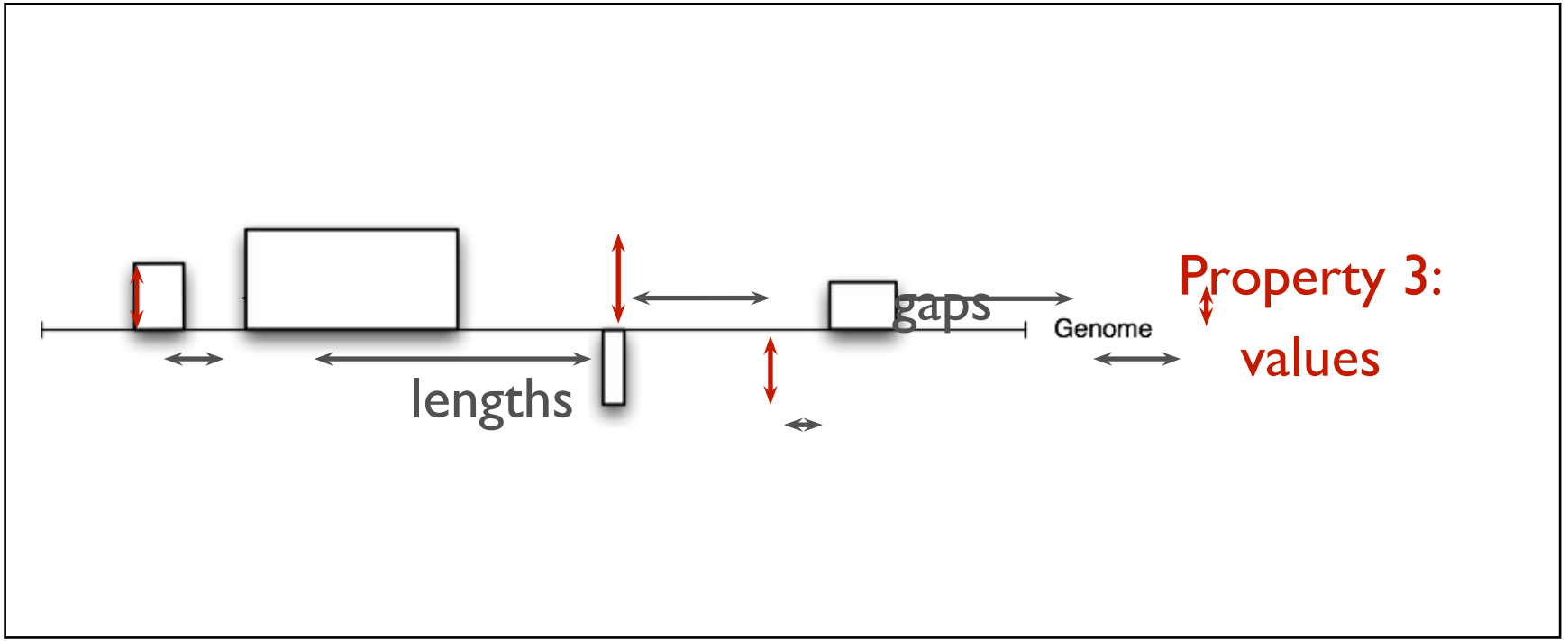
Core properties of tracks



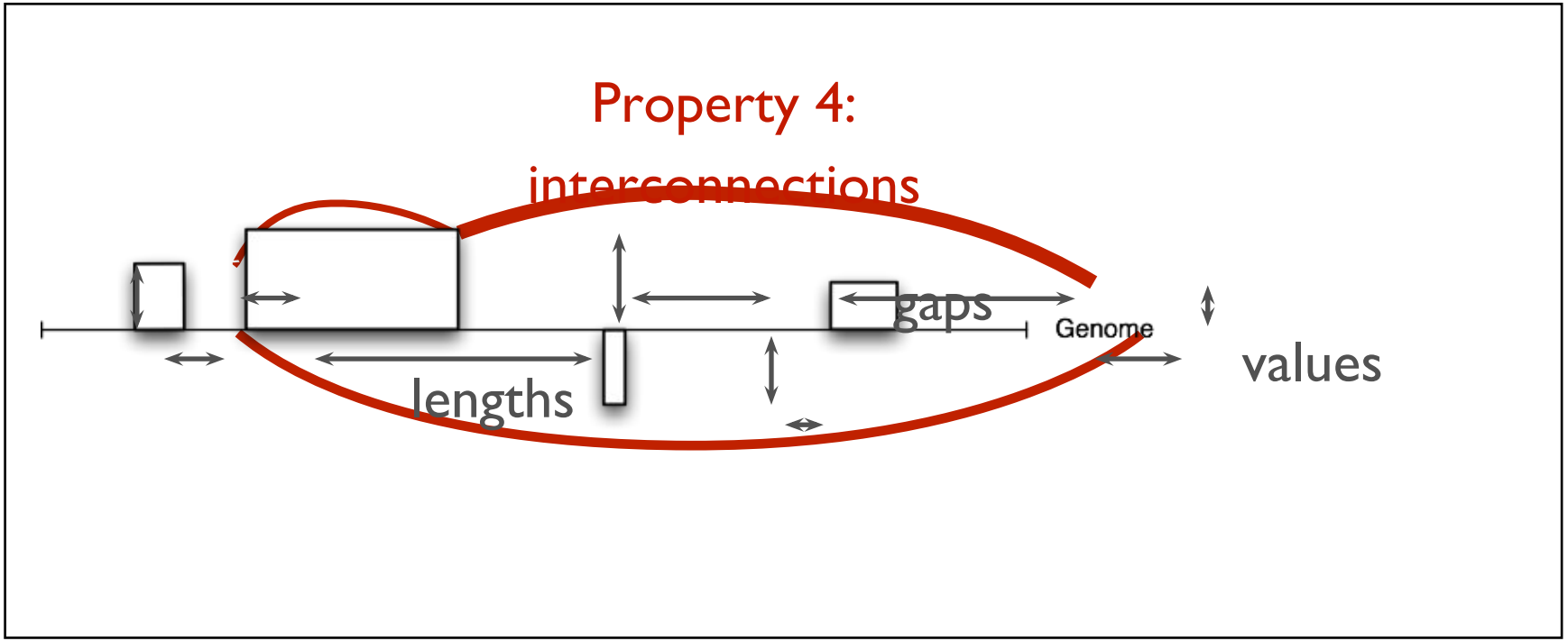
Core properties of tracks



Core properties of tracks

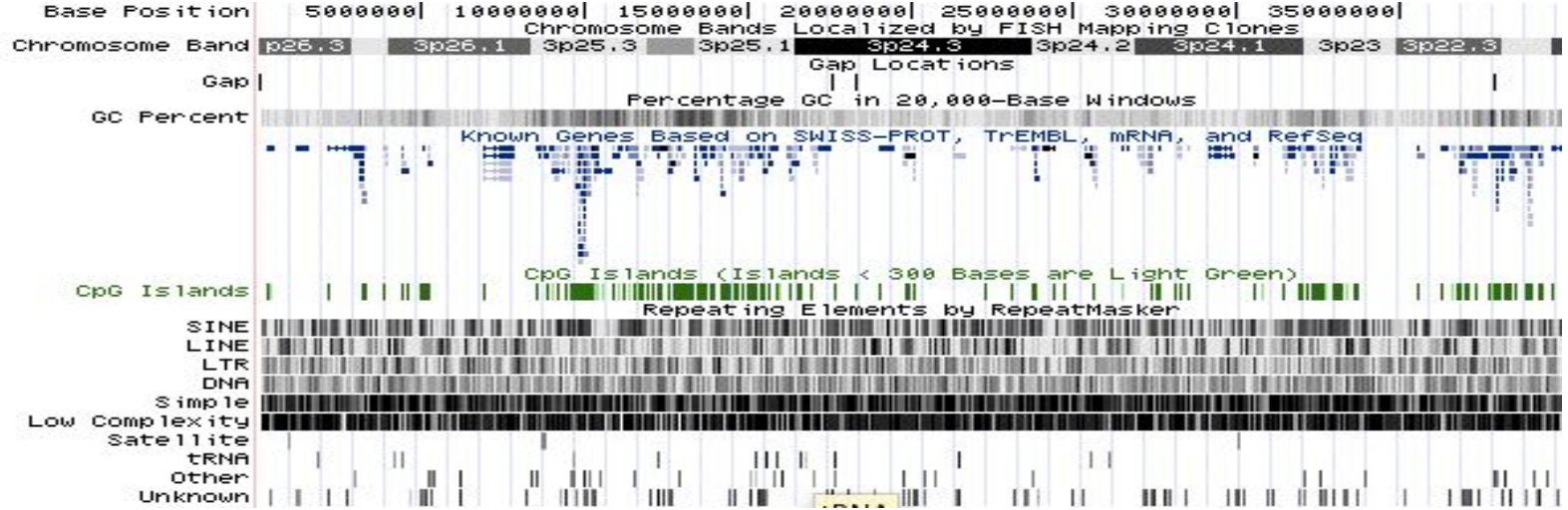


Core properties of tracks



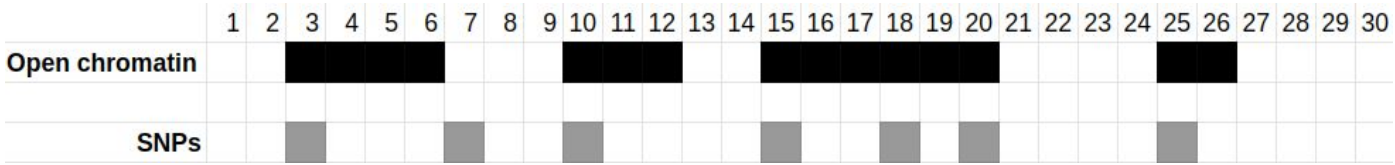
Tracks in the real world

- UCSC Genome Browser is built around the concept of genomic tracks
- Each row is a track, and many track types are supported

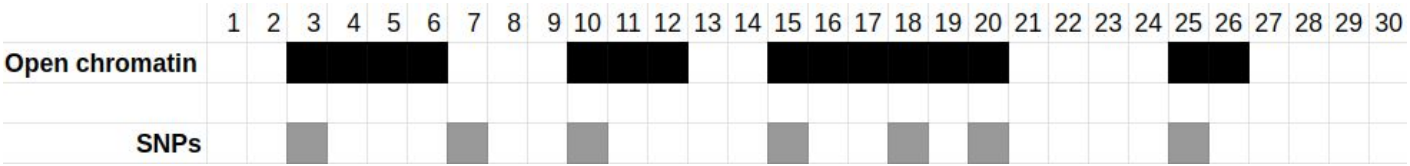


Hands-on exercise with two tracks

- We suspect that SNPs associated with Multiple Sclerosis often occur in regions of the genome that have open chromatin (for a given cell).
- We perform an ATAC-seq experiment on this cell type and get a **segment track.**



Exercise 2



Calculate:

a. Number of SNPs inside regions with open chromatin 6

b. Ratio of SNPs that are inside regions with open chromatin

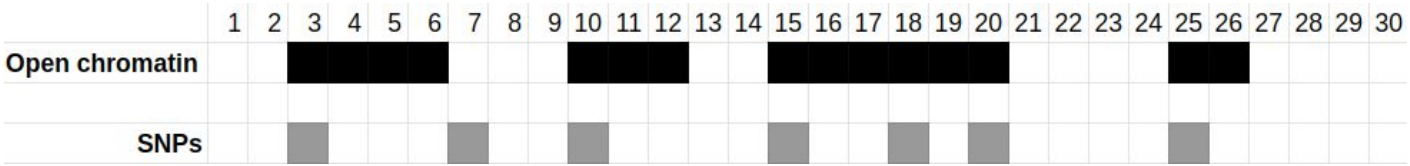
6/7 or 85.7%

What conclusions can we draw from the results?

- **86.7% of all SNPs are inside regions with open chromatin**

- 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 2



Calculate:

a. Number of SNPs inside regions with open chromatin — 6

b. Ratio of SNPs that are inside regions with open chromatin — 85.8%

Try to find:

c. the proportion of the genome that has open chromatin — 50%

d. the expected number of SNPs that would be in open chromatin if all 7 SNPs were positioned randomly — 3.5

e. the proportion of observed to expected SNPs inside open chromatin — 1.7

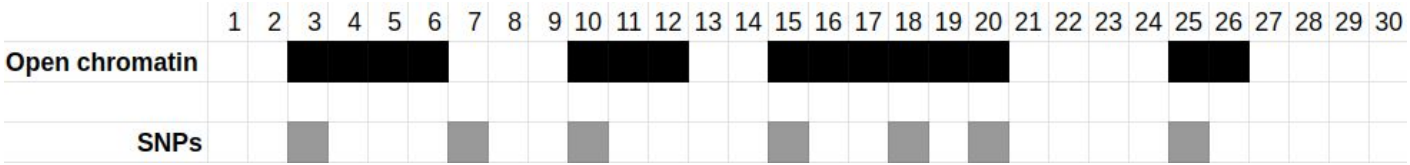
What we did in the exercise

- In order to assess whether 6/7 SNPs being inside open chromatin is 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 1.7 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 (two SNPs are allowed to be next to each other)

b) Compute number of “random” SNPs within open chromatin segments

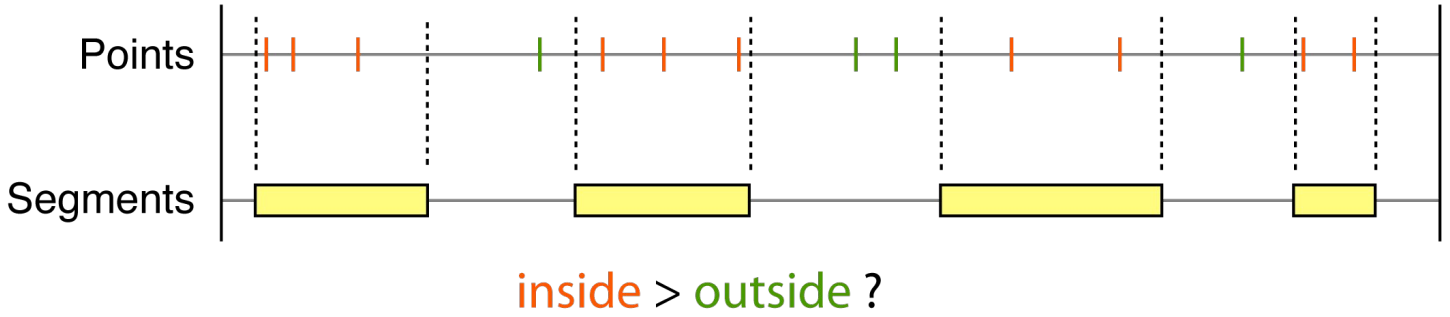
You can find the tracks here:
https://docs.google.com/spreadsheets/d/1-zWvGKx0kMfkvtPCoF8C4fmpOi_3FPcSzcaa5GFFIIs/edit?usp=sharing

What conclusions can we draw from this histogram?

- We drew a histogram of all the counts from a random SNP track on the whiteboard
- We can compare the observed count with these simulated counts (which the histogram shows)
- This shows us whether the observed count (6) is high compared to what we would expect by chance
- You now have a feeling of the data and some methods.
  - We will now start working with real data
  - .. but first some theory

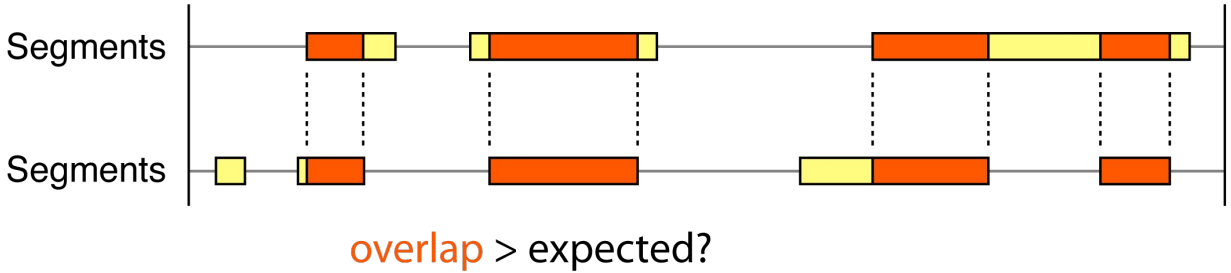
Typical analysis questions

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



Typical analysis questions

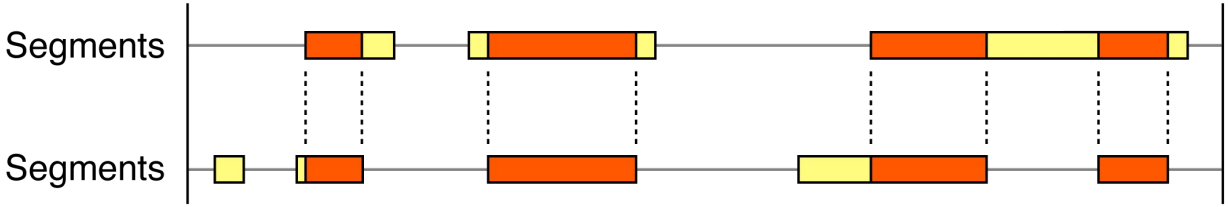
Do genomic feature X and Y overlap more than expected by chance?



Typical analysis questions

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

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?

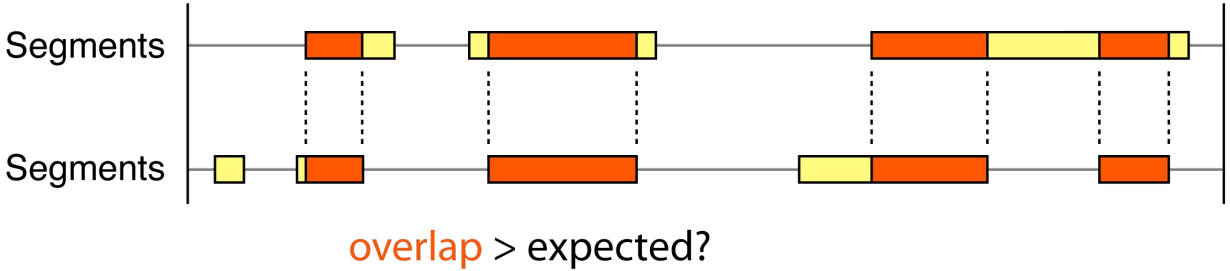
Some examples of this type of analysis

- 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 …… cells (Disanto et al. 2012)
- DNase hypersensitive sites and association with multiple sclerosis (Sandve et al. 2012)

Some examples of this type of analysis

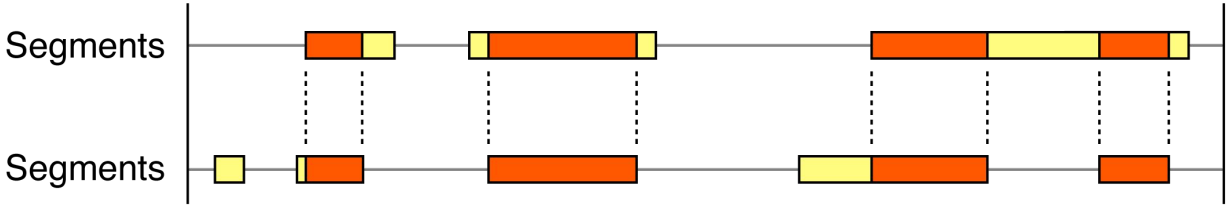
- Vitamin D receptor binding, chromatin states and association with multiple sclerosis (Sandve et al. 2012)
- DNase hypersensitive sites and association with multiple sclerosis (Disanto et al. 2013)

How does this look at the whiteboard?



- 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?

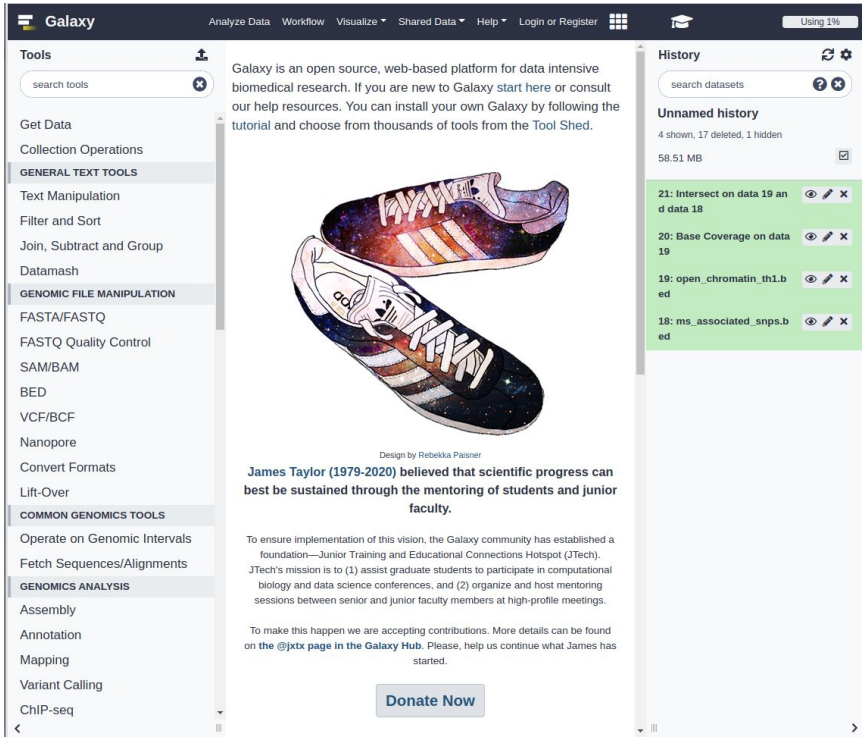
How to assess whether overlap > expected?

Part2

Introduction to the tools we will be using:
Galaxy, Python and Bedtools

Galaxy

- Galaxy is a web platform for computational research (mainly within bioinformatics)
- Galaxy includes a lot of other tools, such as Bedtools, SPAdes
- We will learn:
  - How to upload a genomic track
  - Do simple operations on that genomic track



Bedtools

- A command line tool
- Has a very good documentation and is quite easy to use: https://bedtools.readthedocs.io/en/latest/
- Usually operates on either one or two files (typically bed-files)
- I'll show how you run bedtools live ...

The Genomic Hyperbrowser

- It is “identical” to Galaxy, but has some additional tools for doing statistics, and the user interface can be a bit more messy to use

Python

- A very useful and powerful programming language for solving bioinformatics problems
- Many of you are probably familiar with Python
- Using Python on the exercises will be optional
- If you do want to use Python, you can run Python scripts on the server, or locally on your computer

Part 3

Exercises (descriptive statistics)

Exercises

- Back to the Multiple Sclerosis case
- We have a set of SNPs that we know are associated with Multiple Sclerosis
- We have a theory about these SNPs occuring in regions of open chromatin in a given cell type more than is common
- We do an ATAC-seq experiment and optain a genomic track with regions having open chromatin

Exercise is at https://github.com/uio-bmi/statistical_genomics_exercises

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

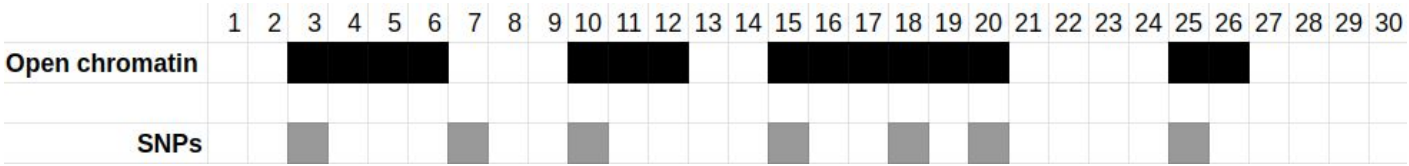
- A hypothesis test quantifies the certainty of concluding a hypothesis (p-value)
  - For some cases, a very small p-value might be requested, e.g when concluding on the effect of a drug

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.

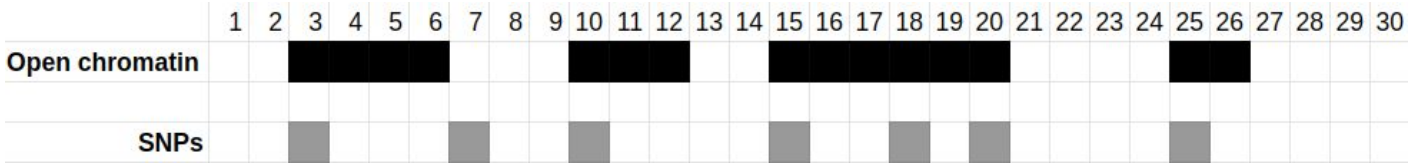
Back to our case

- **Claim:** SNPs are often in regions with open chromatin (more than expected by chance)
- What is the null model? (How do we assume these tracks behave when there is no association)



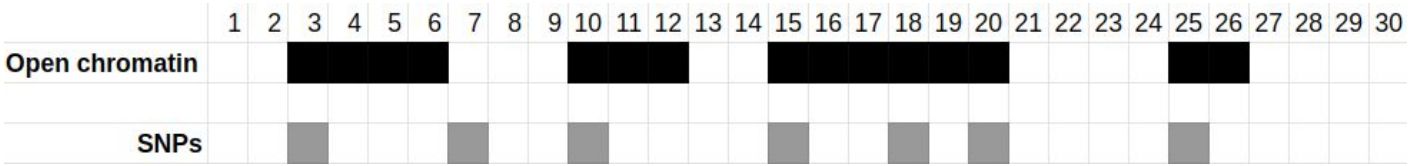
How to make random samples in this case?

- 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?

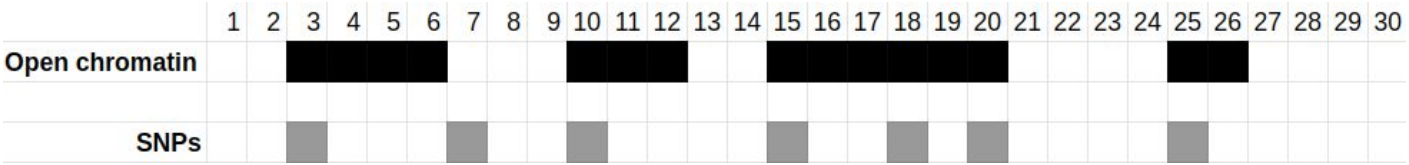


Monte Carlo

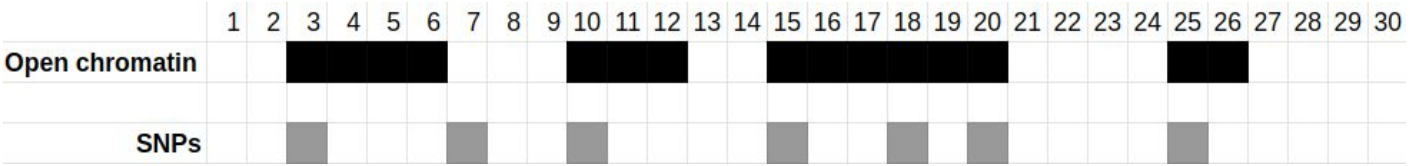
- Simulate many samples from the null model
  - E.g. many pairs of tracks following the same properties
- For each simulation, compute the co-occurrence
  - E.g. the number of base pairs overlap
- Compute how often the co-occurrence found **using the null model** was as extreme or more extreme than the co-occurrence found in **our observation**
  - If this happened rarely (e.g. < 1 % of the times), we conclude there is an association (with significance level 0.01)



- **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 (two SNPs are allowed to be next to each other)

b) Take each (grey) base pair and move it to a random location, but keep distances between SNPs
   - 2, 3, 2, 4, 2, 1, 4, 5
   - 7 snps
   - Do this exercise twice, so you get two numbers

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

Brief about test-statistics

- A test statistic measures an effect
- If we want to measure how much two sets of genes overlap, a possible test statistic is the number of base pairs covered by both genes
- Test statistics are necessary when doing hypothesis testing
  - But deciding on a test-statistic is often tricky

Analysis part 2

Exercises

Investigating Multiple Sclerosis SNPs and open chromatin

You will find this exercise on https://github.com/uio-bmi/statistical_genomics_exercises

Exercise 9

Student answers:

What can we learn from this analysis?

- Assumptions (null model) matter
  - p-value and conclusion differs depending on null model
  - How do we know what null model we should use?
- But what have we really found here?
  - That SNPs associated with M.S. in this given cell are more often inside regions with open chromatin than random SNPs
  - Are these random SNPs realistic?
- Using this kind of analysis (comparing against random data) has limitations
  - Is there an alternative approach that better would answer the question of whether this association is something unique/special?
  - How about comparing against open chromatin from other cells?
  - Let’s look into this! (but first some theory...)

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 SNPs inside open chromatin regions” or “ratio of SNPs within regions” 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: “SNPs are distributed randomly across the genome in no 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 open chromatin regions with at least one SNP in them
  - Number of SNPs that are either inside or close (< 10k bp) to a gene

How to select a good null model?

- Choose a null model that is as realistic as possible for the case where no effect/association is expected
- Example of some null models for our case (some are better than others):
  - X% of the base pairs in the genome are independently covered by open chromatin, there are Y SNPs sites on random locations.
  - All open chromatin regions are kept, but moved to random locations. SNPs are moved to random locations.
  - Genes are kept where they are, SNPs 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
- Your job is to “translate” a biological question into a “bioinformatics” problem definition (test statistic, model choice, etc)

Making justified choices can be hard

- In addition to model selection, the choice of data can also influence results
- Sometimes one can easily justify what data to 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 SNPs since we assume that in the null model, SNPs has no preference to be in 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

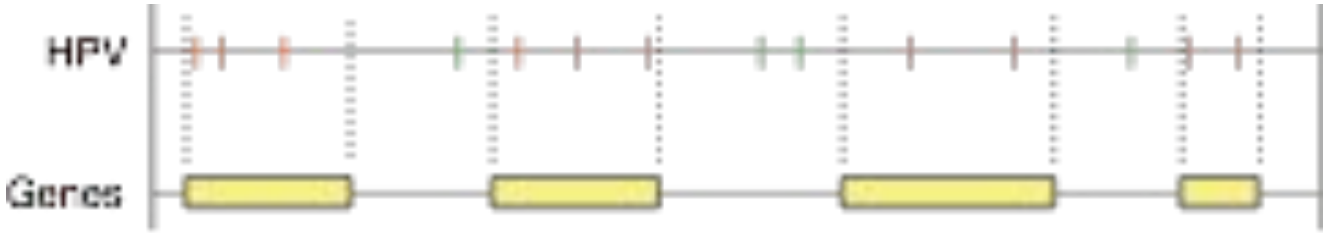
- Generally:
  - Use test-statistic that gives best (lowest) p-value
  - Use null model that gives worst (highest) p-value
- Reasoning:
  - Use measure that best catches relation of interest
  - Use the most realistic model of nature (null model)

Defining a test statistic is not always easy

- Consider the following claim:

"HPV virus is expected to integrate in the genome near genes.”

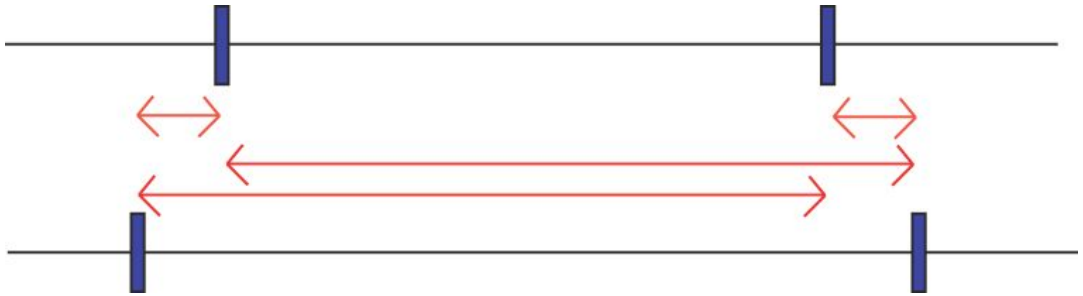
Measuring closeness is not trivial



- 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.

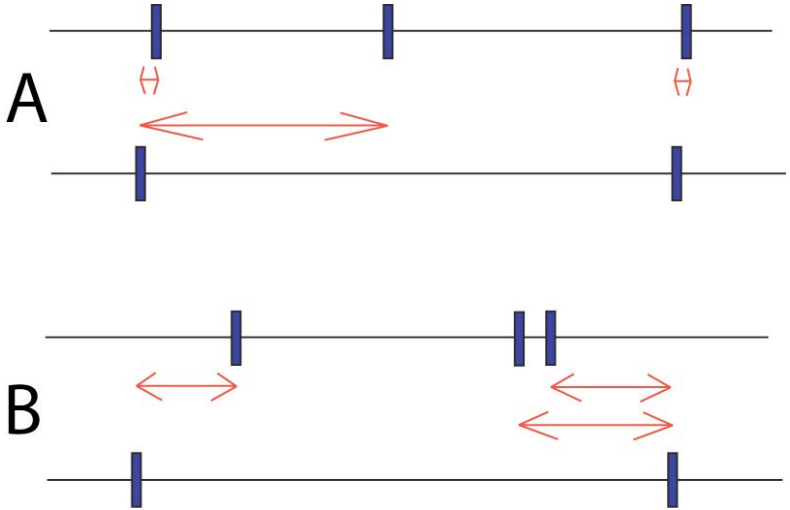
We can calculate distance between pairs of elements

- 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

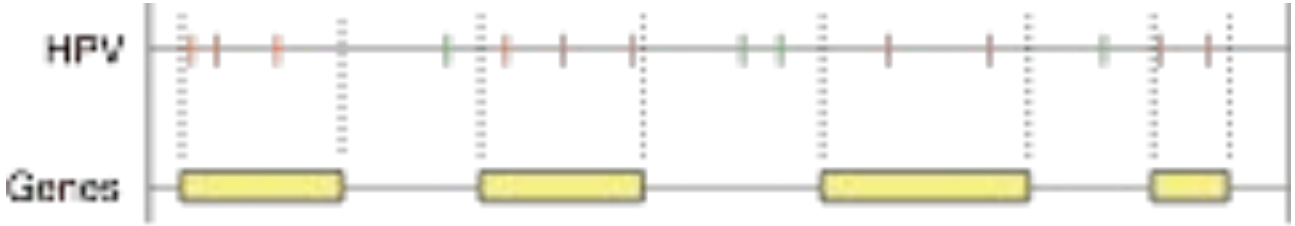
- Use sum or average of of distances?

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

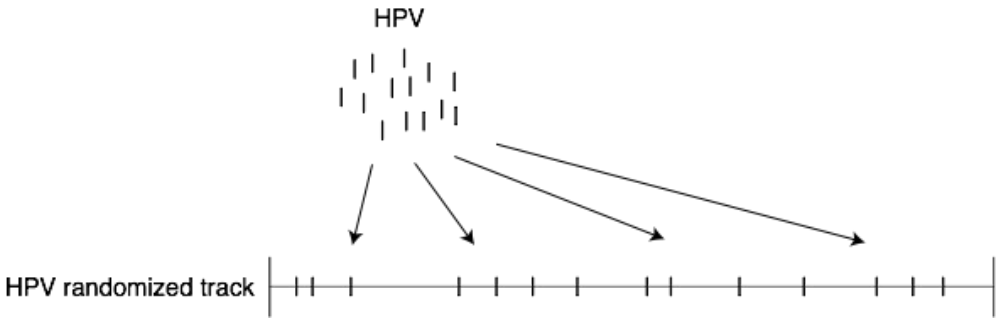
Monte Carlo

- 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

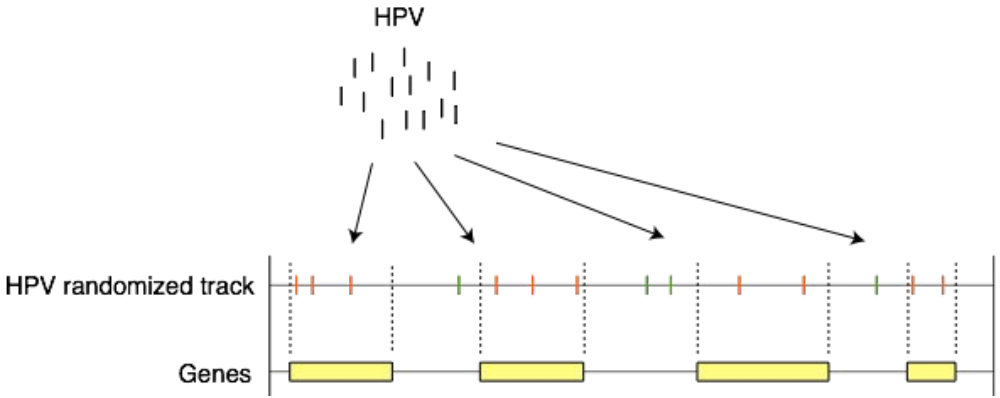
- Why is Monte Carlo simulation so powerful? Let's see how it can be used on points and segments.

Monte Carlo test on "points inside segments"

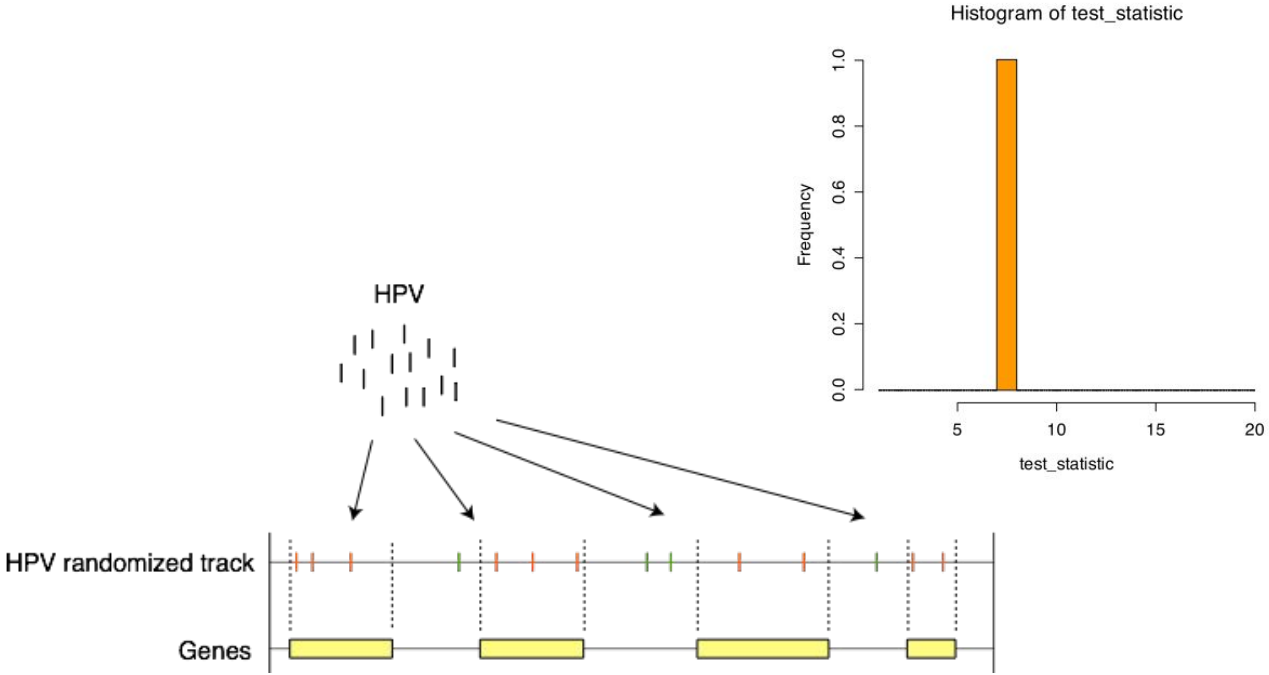
- Randomize point (HPV) locations (null model)



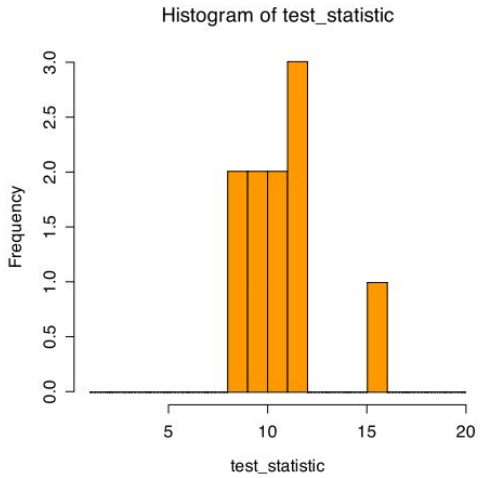
Monte Carlo test on
“points inside segments”



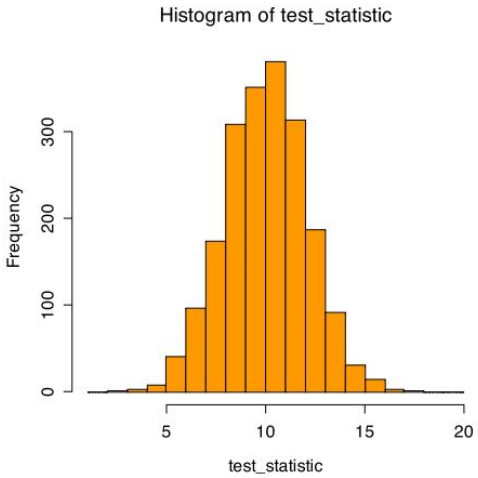
Monte Carlo test on “points inside segments”



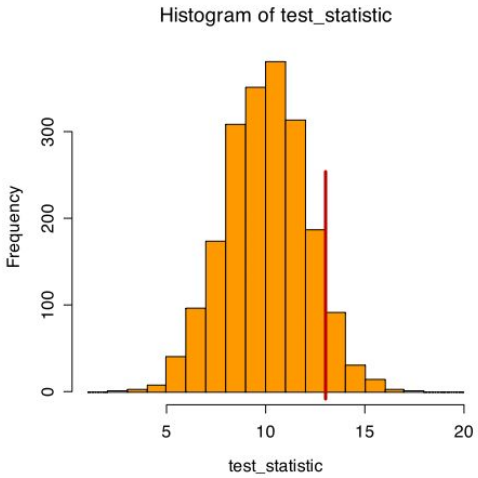
Monte Carlo test on “points inside segments”



Monte Carlo test on "points inside segments"

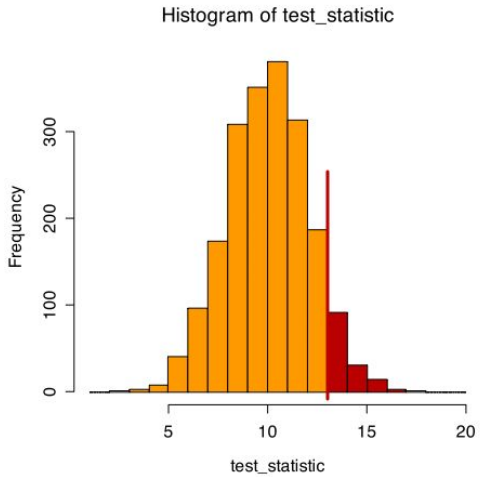


Monte Carlo test on "points inside segments"



Monte Carlo test on "points inside segments"

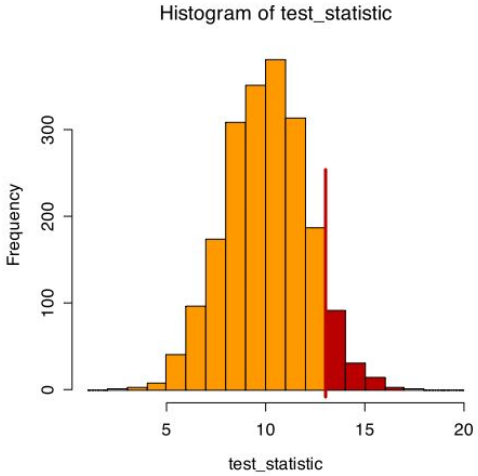
- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times
- Build histogram
- Compare with number of observed points (HPV) inside segments (genes)
- p-value = area to the right if alt hypothesis is "more" (if "less",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



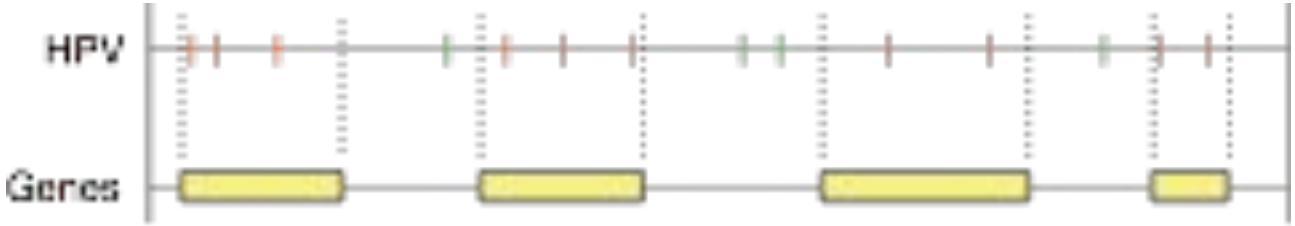
p-value = 0.08

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)

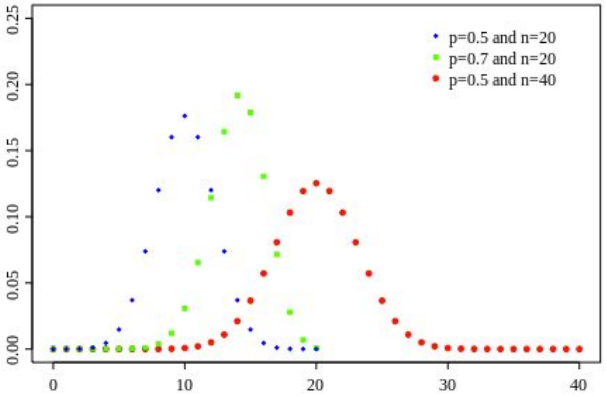
Binomial distribution

- 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:

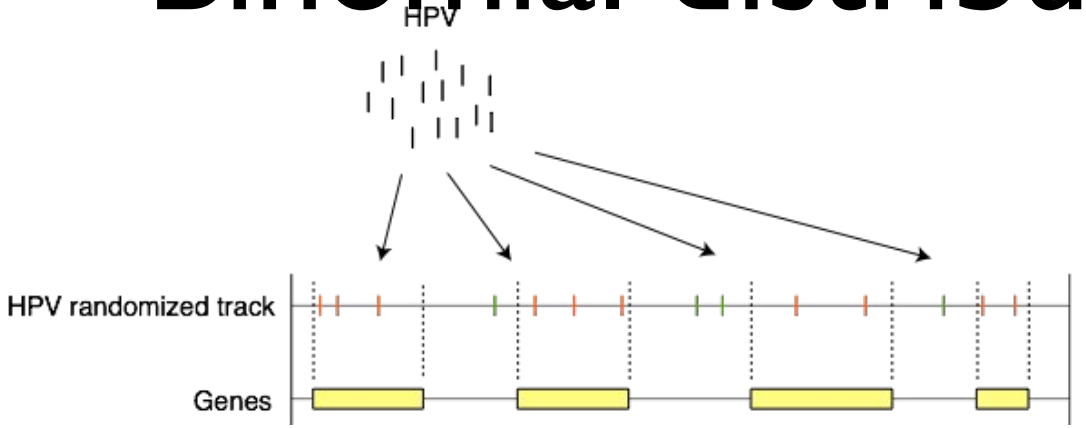
```
P(tails) = p = 0.6

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

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

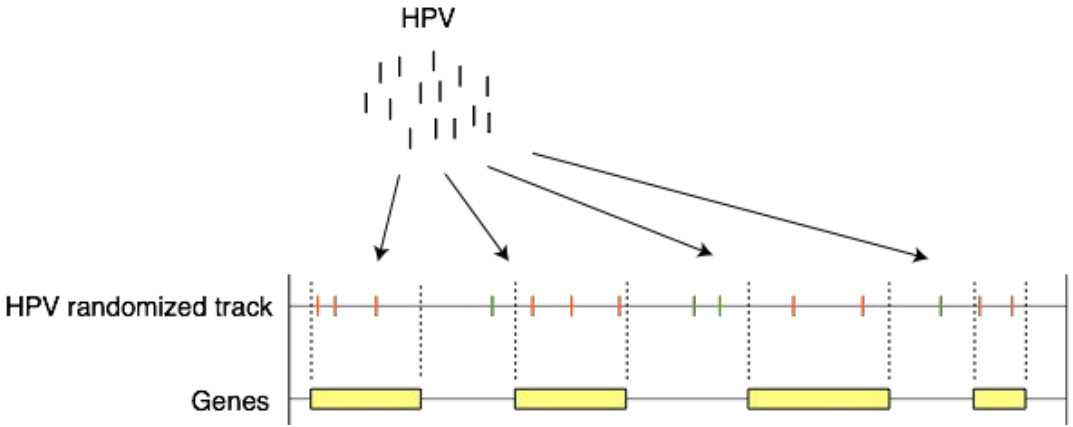


Binomial distribution



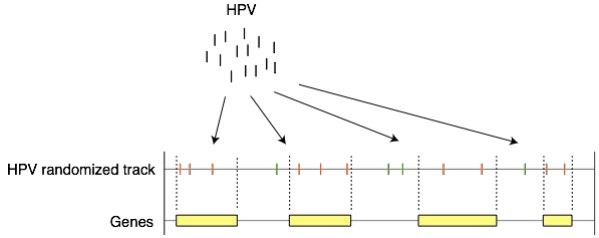
- In this case, each HPV is a coin, and it can either fall into a gene or not, depending on how much of the genome that is covered by genes
- n = number of HPV
- p = proportional coverage of genes

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 SNPs and open chromatin 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 are we expected to 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

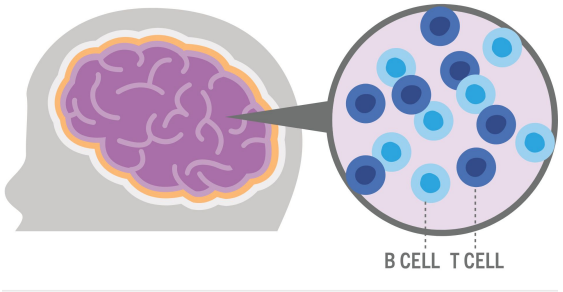
- Association: A & B are related, show up together.
- Causation: A causes B
- Using statistical testing for the co-localization of two tracks, we can only find whether there is an association
- Causation often requires speculation, biological understanding, experimentally determined mechanisms

Back to Multiple Sclerosis, but this time with multiple 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?



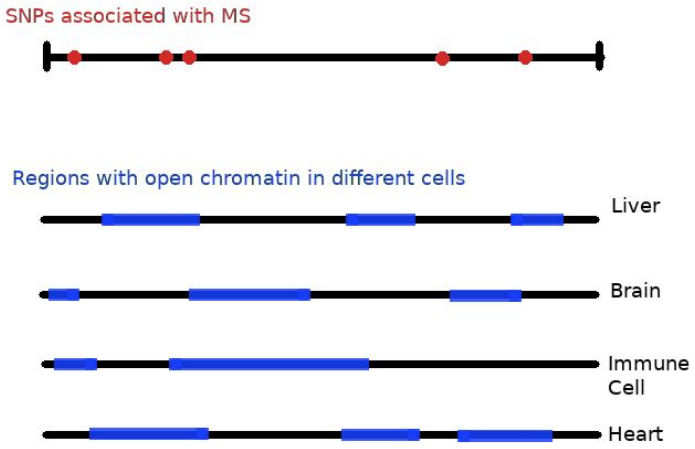
Exercise: Investigating Multiple Sclerosis

We have the following:

A track with position of variants (SNPs) we know are associated with Multiple Sclerosis

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.

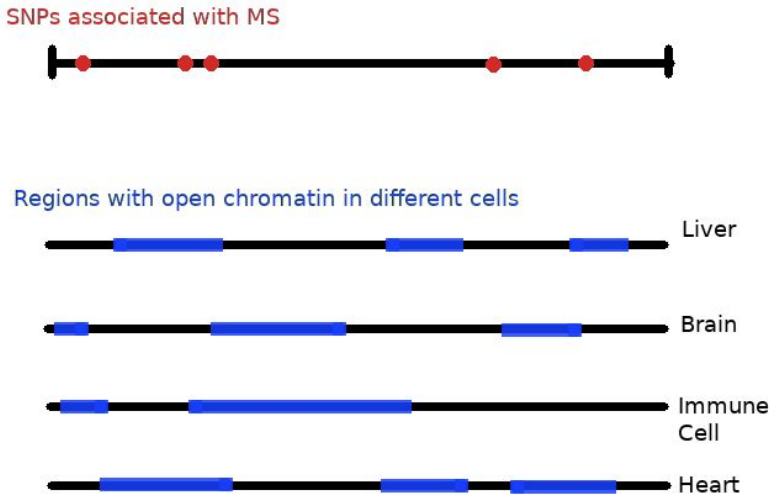
We have tracks of regions containing open chromatin for many cell types



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

Exercise: Investigating Multiple Sclerosis in multiple cells

You'll find this Exercise in the Github repository



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

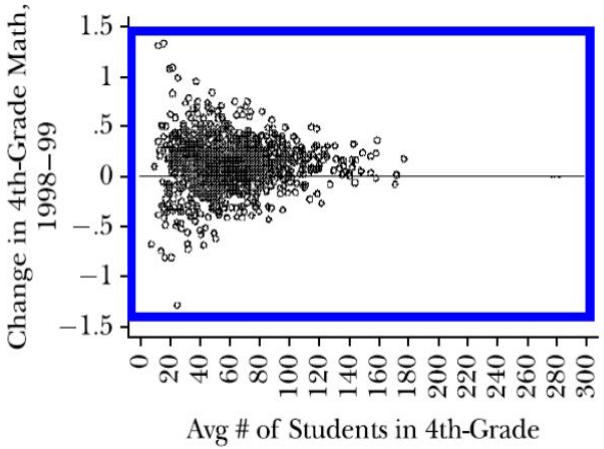
Jaccard is most common to use, but it is not a good measure for similarity of genomic tracks, since it is biased by giving higher similarity for tracks that cover many base pairs.

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.



Difference between similarity measure and test statistic

- A test statistic computes a number that is compared to a probability distribution of that test statistic under the null momdel
  - Thus, it is “scale invariant” (doesn’t matter if we take the mean or sum for instance)
- A similarity measure is computed for different data sets, and then these are compared.
  - It needs to handle datasets of different sizes.
  - Often it finds the proportion of something in regards to something else.

Typical multitrack analysis questions

- Which tracks in a collection are most representative or most atypical?
- Which tracks in a collection coincide most strongly with a target track?
- Are certain tracks of one collection coincide particularly strongly with certain tracks of another collection?
- Which genomic regions are mostly enriched with the segments of tracks in a collection?
- In which genomic regions are tracks of a collection coinciding the most?

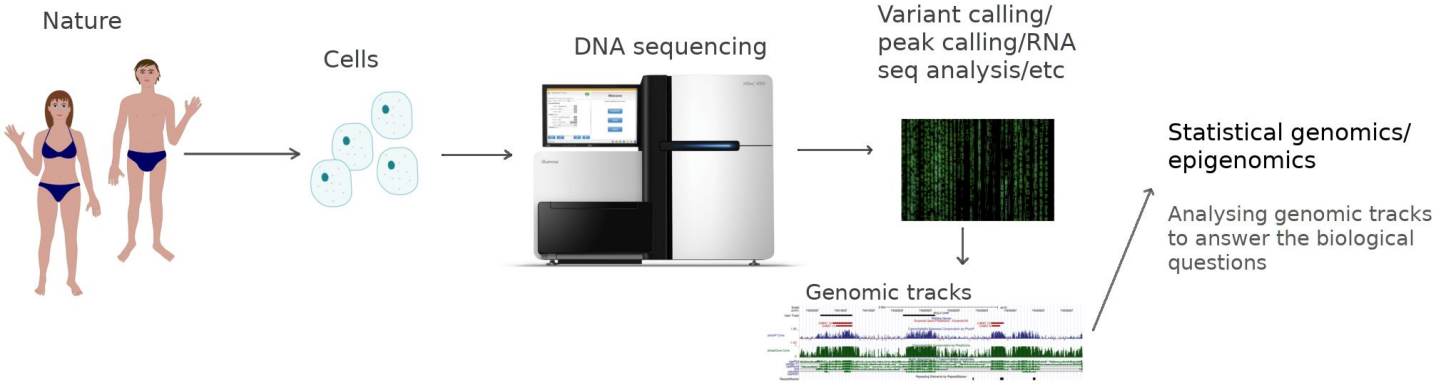
The basic mode of the Hyperbrowser has a lot of cool example analyses using these questions

Summary

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

What is statistical genomics/epigenomics?



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

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)

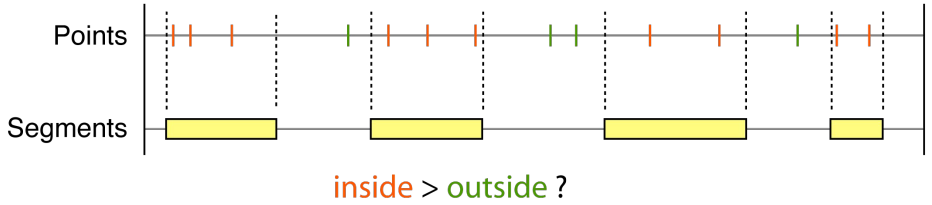
- Typical formats you will be using in real analysis:
  - VCF
  - Bigbed, bed
  - Any files containing the position of genomic elements

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

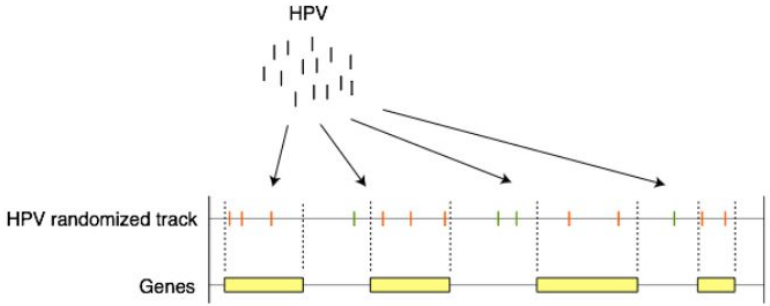
- Co-occurence is measured by a **test statistic**
  - E.g. the number of base pairs overlapping between two tracks

- 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

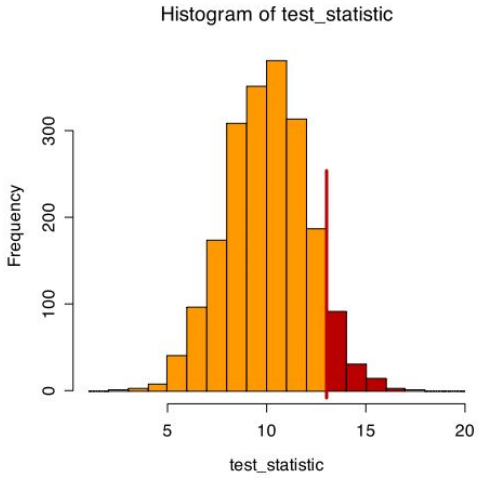
- An example of a null model:
  - We assume that SNPs are distributed uniformly across the genome when there is no association

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



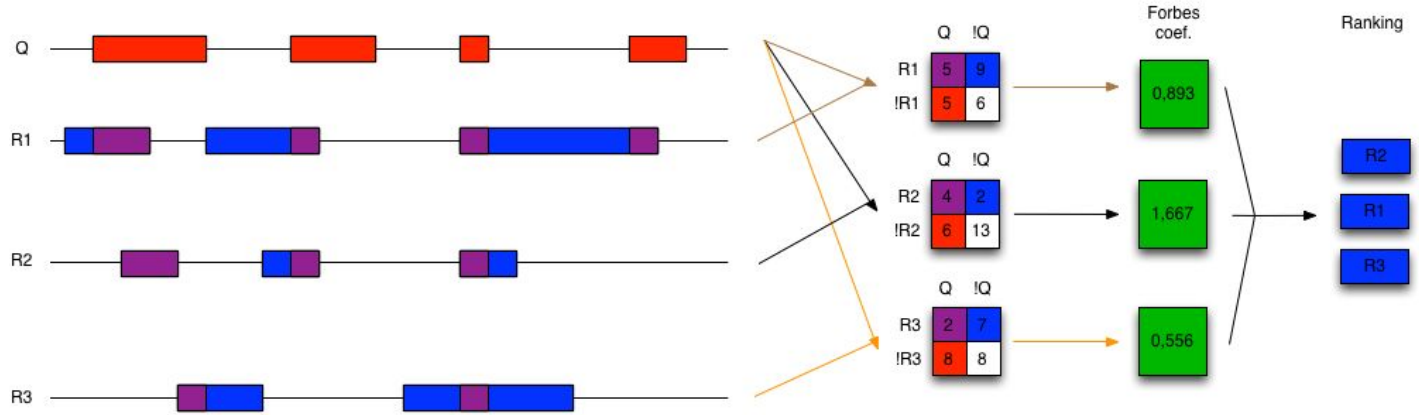
Finding the p-value

- 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:
  - Forbes is usually a good choice



results

t the bottom)

Questions?

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