

Project 2025

BINF-F401

Vincent Detours

Overview of course evaluation

Written exam

- Questions assess your understanding of the theory
- It's an open book exam or 2 hours (but don't expect to study the course during these two hours!)
- You've got a personal grade

Project

- You apply ideas and tools of the course on real data
- It's a mini research project evaluated from a written report
- It's done by groups of three students, so grade is collective

Final grade is the harmonic mean of exam and project (see https://en.wikipedia.org/wiki/Harmonic_mean). Thus, you need reasonable scores for both exam and project to pass.

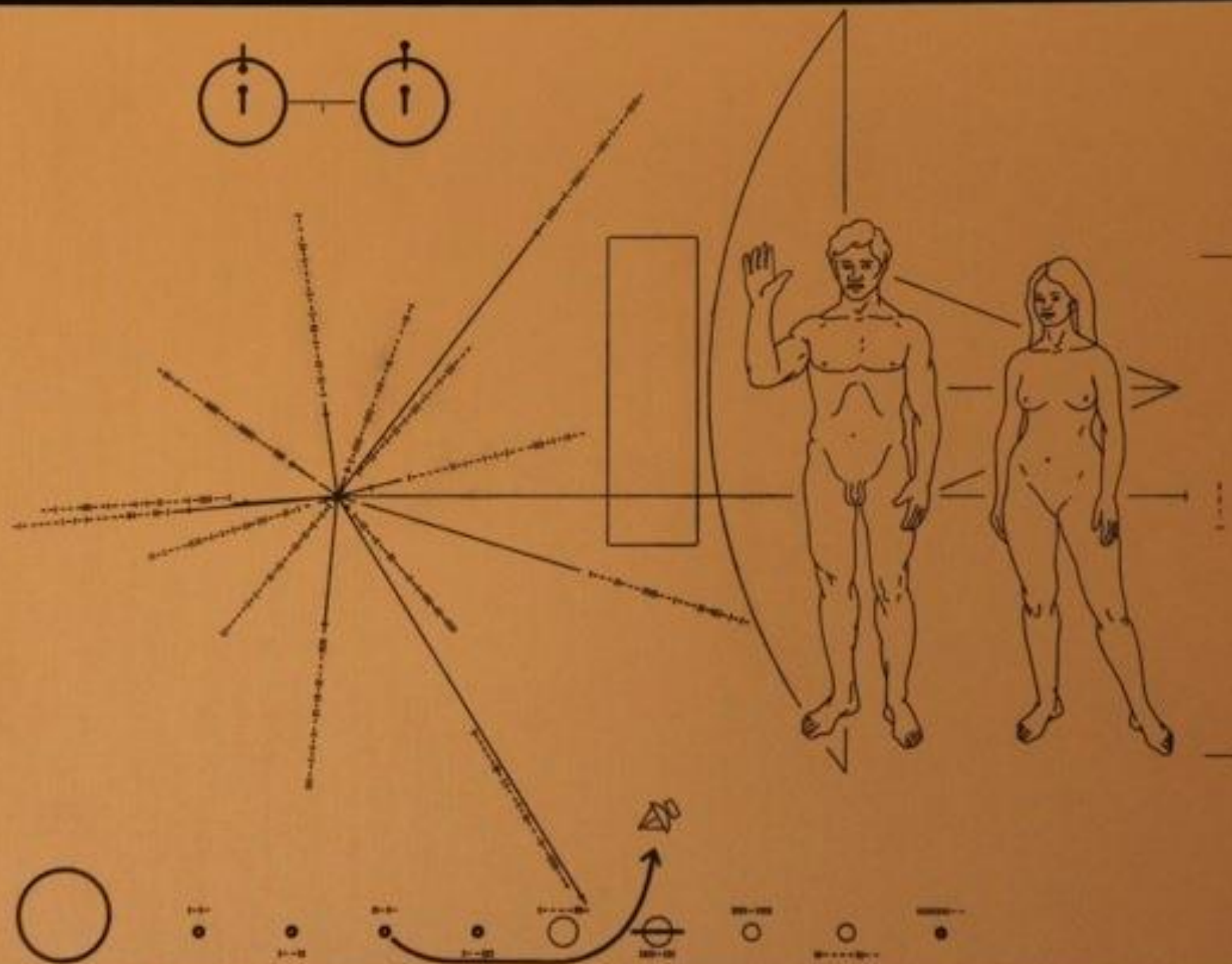
Scientific background

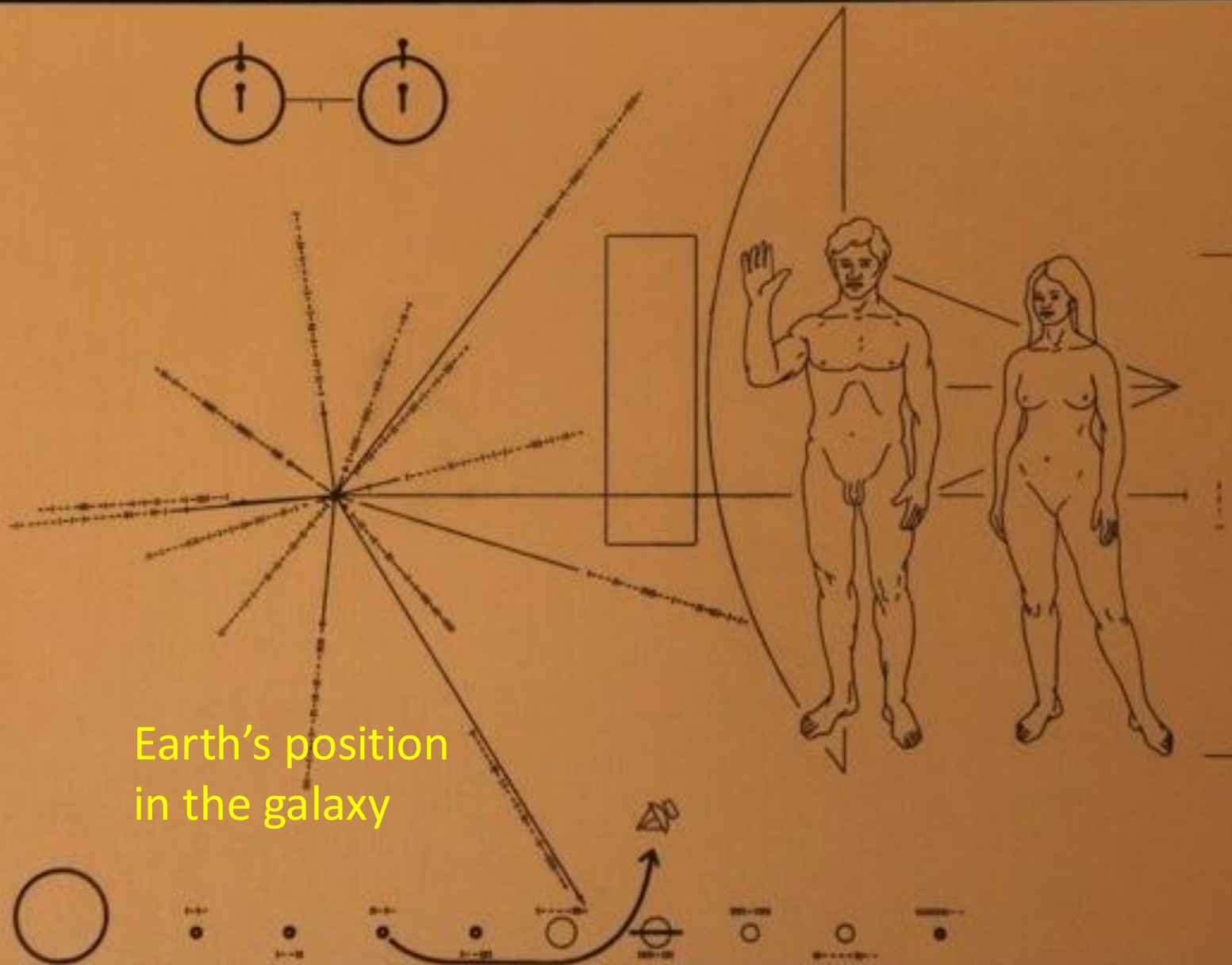
Quantifying human sexual dimorphism with the dimorphism index



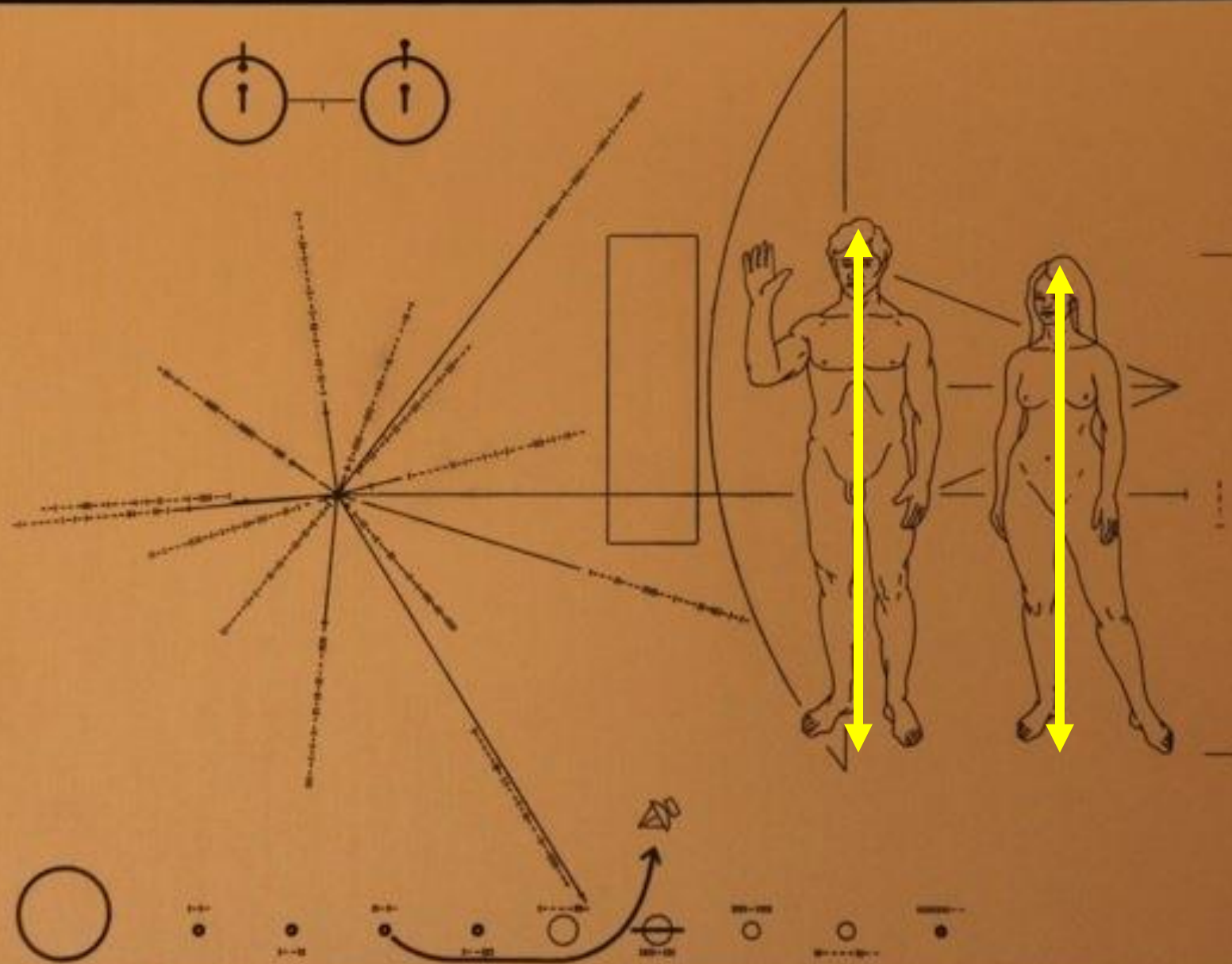


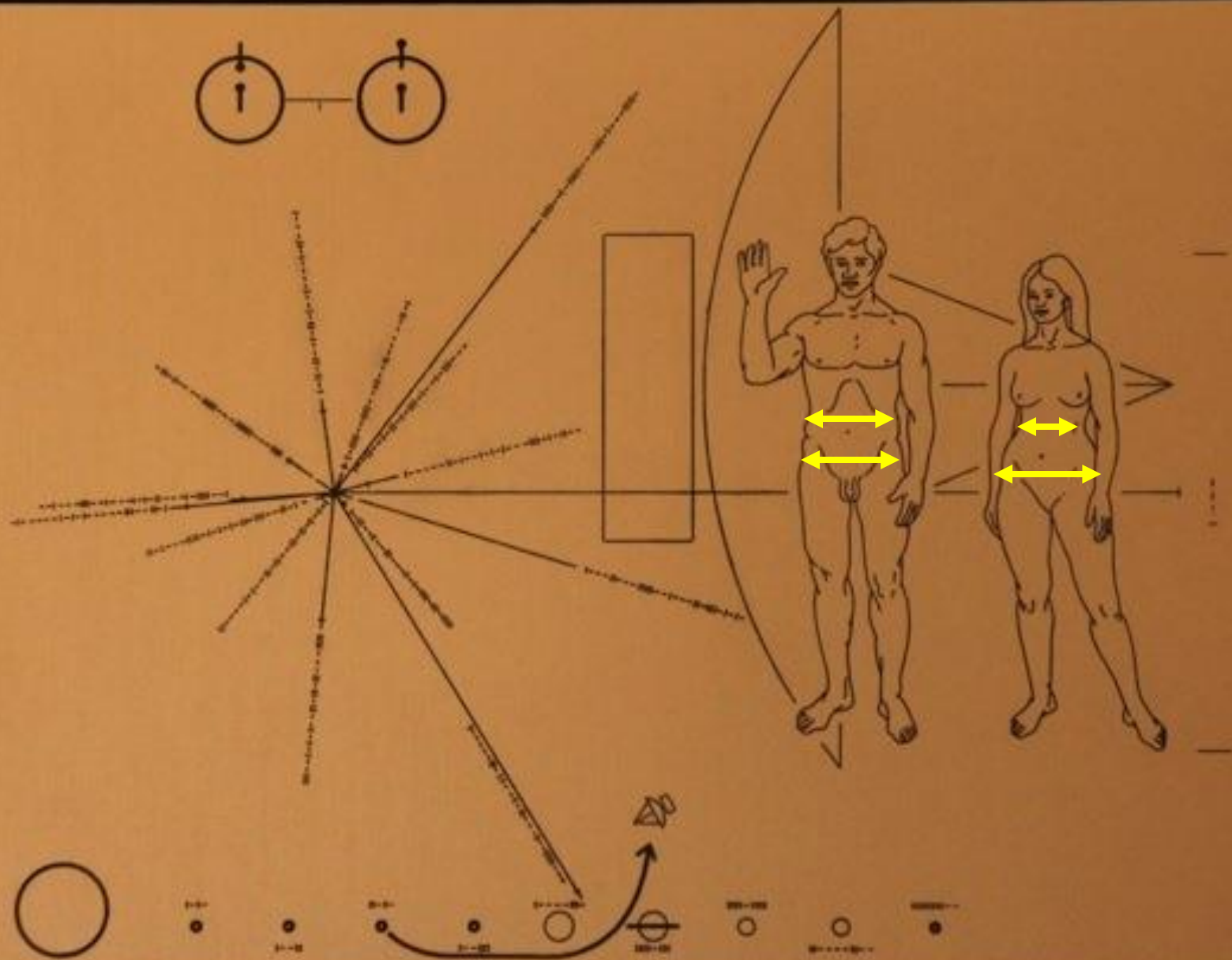




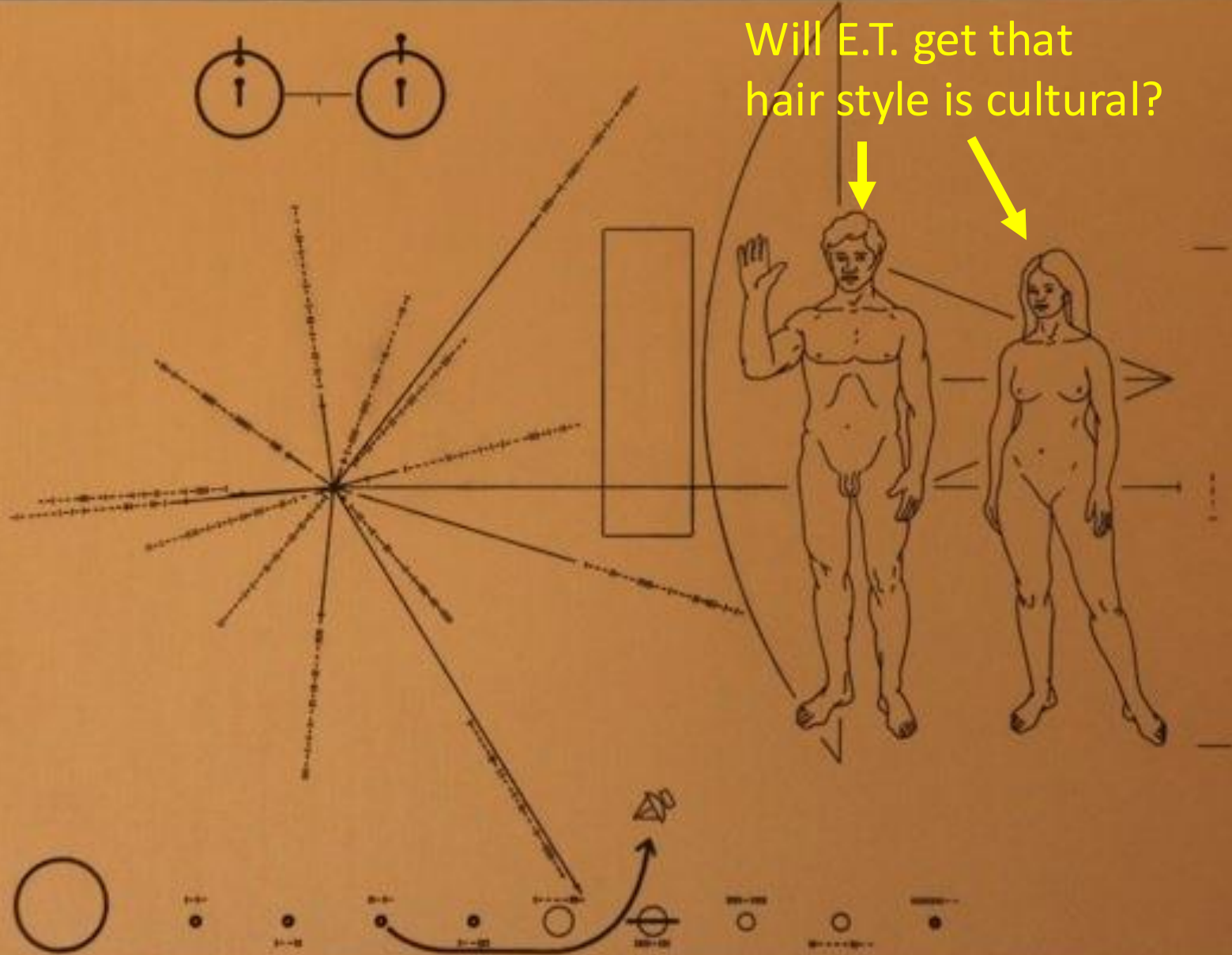


Earth's position
in the galaxy





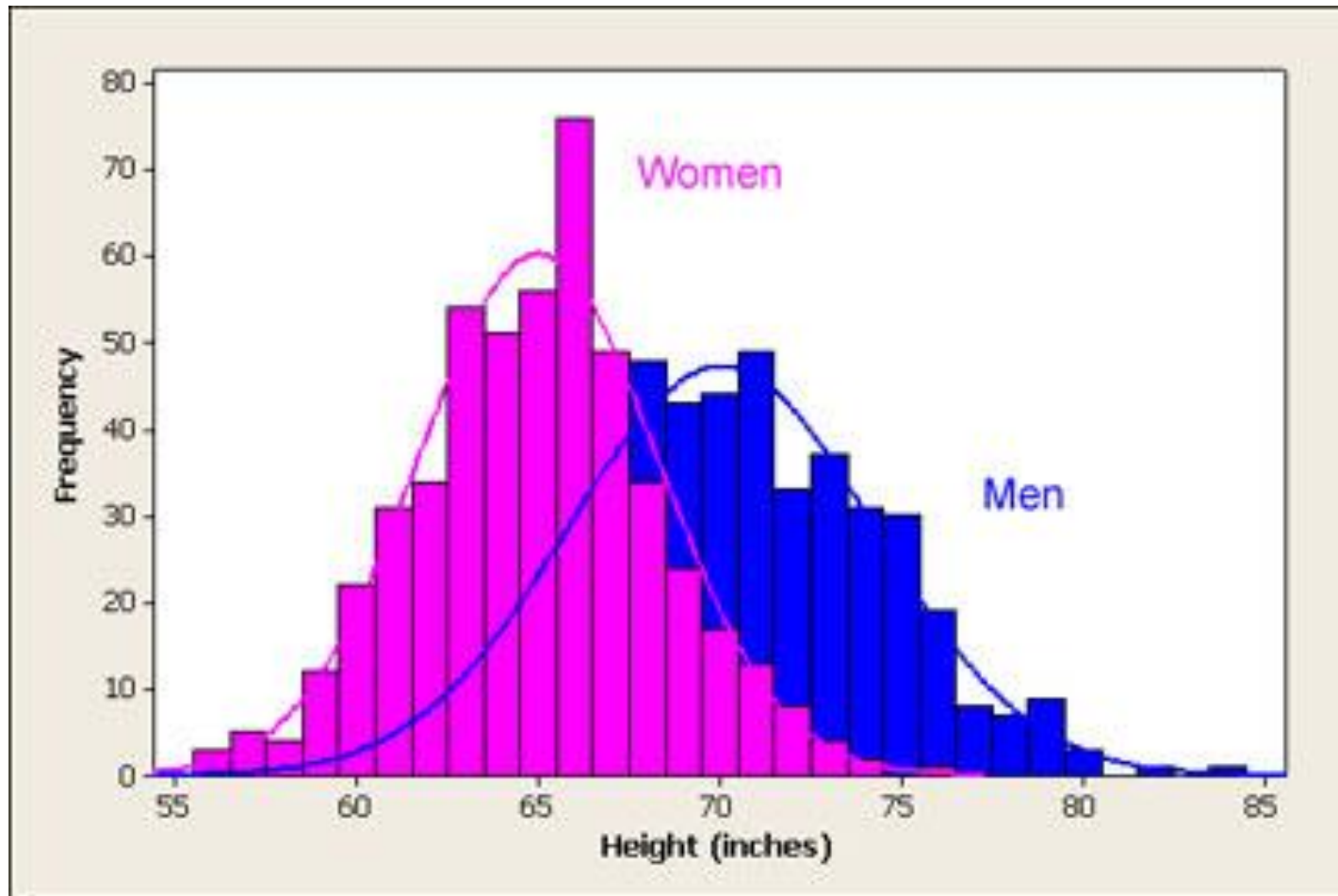
Will E.T. get that
hair style is cultural?



Human sexually dimorphic traits

- Height
- Breast
- Body hairs
- Muscular mass
- Hip-to-waiste ratio
- Disease incidence (not on the Pioneer plate)
- Etc.

Human sexually dimorphic traits lie in overlapping continuums



A loaded question...

A classic example:

- Broca reported in 1861 larger brain size in male than female...
- ...but did not control for body size.
- Research on male/female brain variations (beyond the genitalia) has been controversial, perhaps tainted with cultural biases

The dimorphism index

We propose a **dimorphism index**, which is

- **Unbiased**, i.e.
 - It does not rest on any *a priori* regarding what makes male and female different,
 - It does not even assume that dimorphism is present
 - It only assumes that individual belong to one genetic sex, XX or XY
- **Quantitative**, we approach sexual dimorphism as a continuum
- **Universal**, i.e. it is applicable over the entire phenotypic space, all scales of biology and all human organs
- **Meaningful to individuals**, setting the work apart from population-level investigations

A clinically relevant topic



Perspective

Sex contextualism in laboratory research: Enhancing rigor and precision in the study of sex-related variables



This paper presents examples of limits of binary sex stratification in clinical research

The dimorphism index may provide a principled, context-specific, adjustment variable

Cell, 2024

The dimorphism index

Intuitively, the dimorphic ratio of Mister Smith's liver is akin to the probability that it is more similar to the liver of another man than to the liver of a female.

So the dimorphic ratio is defined *separately*

- For each organ
- For male and female

The dimorphism index

- The similarity of any pair of genome-wide expression profiles can be defined as their **Spearman's correlation across all genes**
- Given the matrix of similarities between livers from all individuals in the population the dimorphism ratio of Mister Smith's liver is defined as **the median across all pairs men**, [Mister Smith, Mister X], of

proportion of **males** with a liver more similar to the
liver of Mister Smith than mister X is

proportion of **females** with a liver more similar to
the liver of Mister Smith than mister X is

The higher Mister Smith liver dimorphism ratio, the more masculine his liver is

The dimorphism index



These are the livers of individuals in the gene expression space, projected on the 2D screen (using PCA)

The closer they are, the more similar are their liver transcriptomes

The dimorphism index

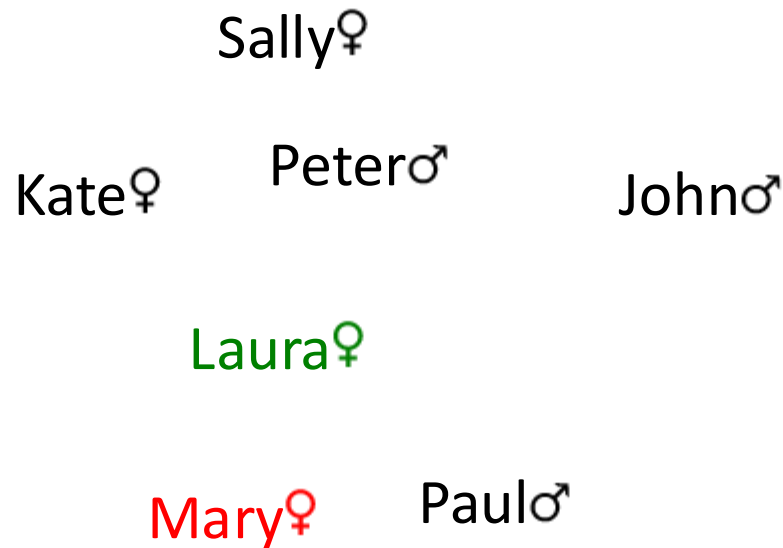


These are, say, the livers of individuals in the gene expression space, projected on the 2D screen (using PCA)

The closer they are, the more similar are their liver transcriptomes

Let's compute the dimorphism ratio of **Mary**...

The dimorphism index



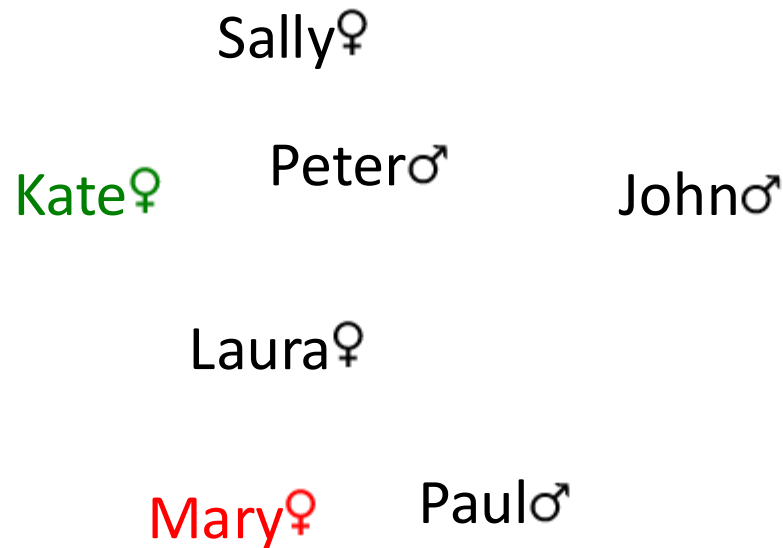
First we use **Laura** as female reference

Laura is the woman the closest to Mary

Mary is closer to Paul than Laura

So the ratio is $(1/3)/(1/3)=1$,
3 is the number of females
(excluding Mary!) and males

The dimorphism index



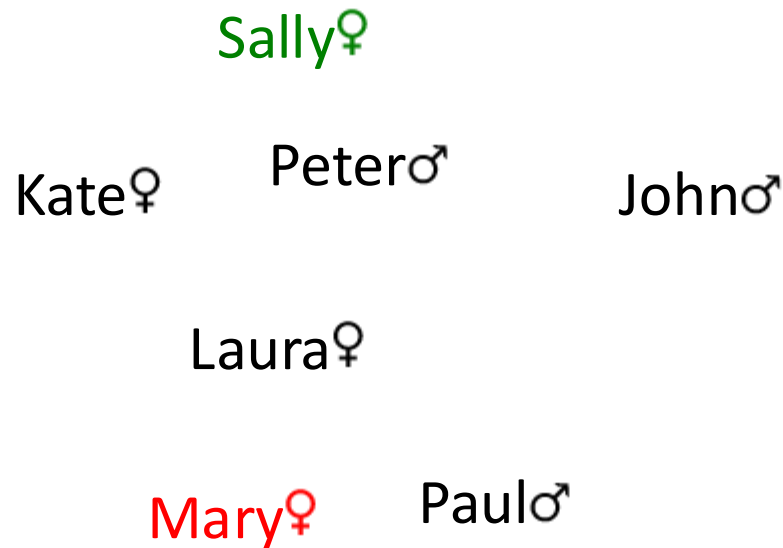
Now we use **Kate** as female reference

Laura is closer to Mary than Kate -> $2/3$

Paul is closer to Mary than Kate -> $1/3$

So the ratio is $(2/3)/(1/3)=2$

The dimorphism index



Finally, we use Sally as female reference

Laura and Kate are closer to Mary than Kate -> 3/3

Paul and Peter are closer to Mary than Kate -> 2/3

So the ratio is $(3/3)/(2/3)=3/2$

The dimorphism index

Sally♀
Kate♀ Peter♂ John♂
Laura♀
Mary♀ Paulo♂

Overall, the dimorphism ratio of Mary is the *median* of the ratios obtained with all the female references, here Laura, Kate and Sally:

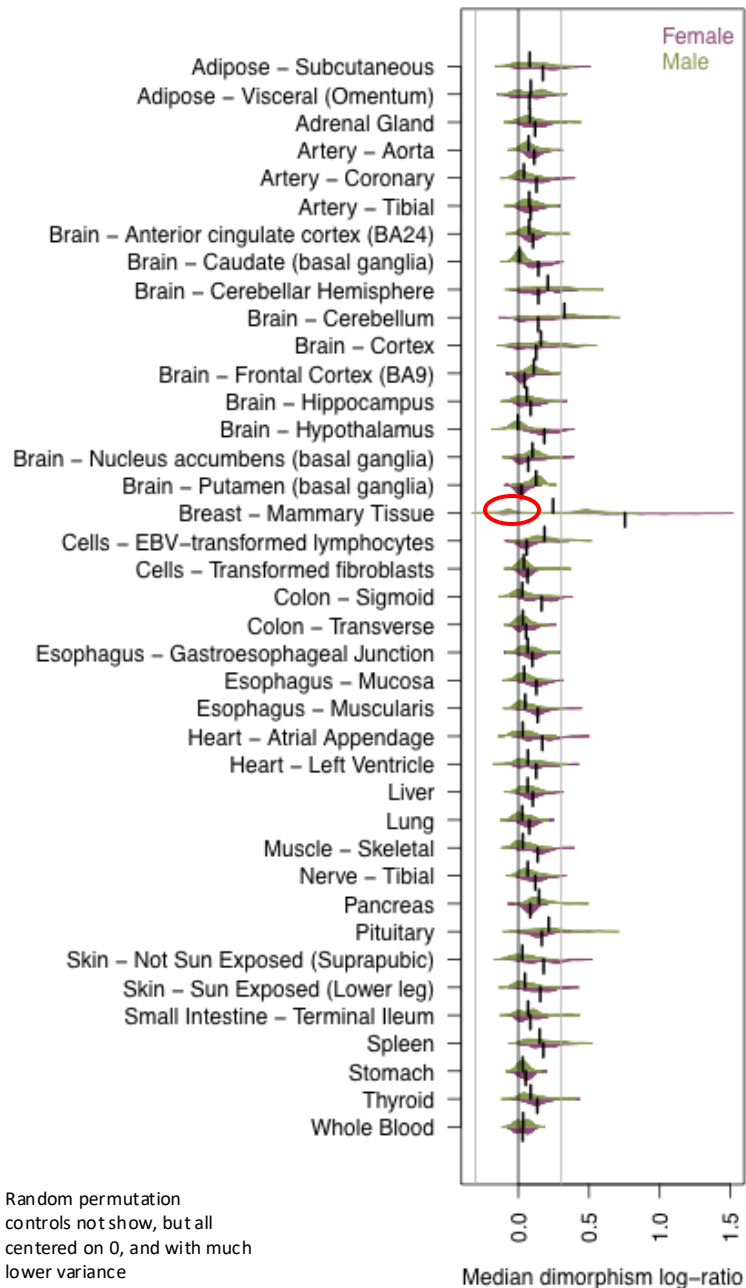
$$\text{median}(1, 2, 3/2) = 1.5$$

Exercise: who has a higher liver dimorphism ratio, John or Peter?

(In practice I average the entire calculation across 100 bootstrap samples for better stability and to balance male and female)

The dimorphism index

- Is defined with respect to a population (here the GETx subjects)
- Takes into account the entire molecular phenotype of organs
- Is applicable independently to all organs
- Is computed separately for males and female, i.e. it does not assume a male/female antagonism
- It *does* assume a male/female dichotomy (controversial in some circles), but assume nothing regarding the traits underlying this dichotomy
- Is applicable to any high dimension data set phenotype.



The dimorphism index applied to organ transcriptomes

- It breast is the most dimorphic (non genital) organ
 - Yet, 11 females and 49 males, have **reversely** dimorphic breast...
 - GTEx pathology notes report gynecomasty in 63% of reversely dimorphic vs. 8% of dimorphic male breasts ($p=5 \times 10^{-11}$)
 - Another factor in male is 'fibrosis' ($p=0.01$)
- Thus,
- the dimorphic index match the intuitive idea that breast is the most dimorphic organ
 - the dimorphic index detected gynecomasty, although I didn't even know this word before I ran this computation!

Technical aspects

The data in *your* hands

Overview of GTEx

- 54 organs from 948 donors, most with
 - High-throughput genotypes
 - Genome-wide gene expression (RNA-seq and small RNA sequencing)
 - Histology images (20X scans)
 - donor level clinical annotations
- GTEx samples were collected postmortem
- For an overview of the GTEx v8 release, see <https://www.gtexportal.org/home/tissueSummaryPage>

Overview of GTEx

- 54 organs from 948 donors, most with
 - High-throughput genotypes
 - Genome-wide gene expression (**RNA-seq** and small RNA sequencing)
 - **Histology images** (20X scans)
 - **donor level clinical annotations**
- GTEx samples were collected postmortem
- For an overview of the GTEx v8 release, see <https://www.gtexportal.org/home/tissueSummaryPage>

Data in your hands

Each project group will be provided a data from a specific organ:

1. Covariates, including clinical data, technical data and pre-computed dimorphism indices
2. A RNA read count matrix
3. Miscellaneous resources

for about 250-600 samples of the organ assigned to your group.

Let me know if data is too big to run the analysis on your computer. We can decide to subsample the population.

Covariates: demographics/health

The clinical data matrix (covariates.csv) includes,

AGE_DECADE: age at death (10 yrs brackets)

SEX: 1=male, 2=female

PATHOLOGY_NOTES: Note from the pathologists who examined histology slides

PATHOLOGY_CATEGORIES: medical conditions diagnosed post-mortem from histology slides (extracted from PATHOLOGY_NOTES)

Covariates : technical

The clinical data matrix (covariates.csv) also includes,

ISCHEMIC_TIME: It's the number of minutes elapsed between the presumed donor death and tissue collection.

RNA_INTEGRITY_NUMBER: a measure of RNA degradation

Details about clinical annotations here:

https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs000424/phs000424.v8.p2/pheno_variable_summaries/phs000424.v8.pht002742.v8.GTEX_Subject_Phenotypes.data_dict.xml

Covariates : clinical/technical

The clinical data matrix (metadata.csv) also includes,

HARDY_SCALE: A number from 0 to 4 summarizing the circumstances of death

0=Ventilator Case

1=Violent and fast death

2=Fast death of natural causes

3=Intermediate death

4=Slow death

Covariates: dimorphism indices

You will have:

RNA_DI: Transcriptional dimorphism index computed from whole RNA-seq transcriptomes for men and for women

HISTO_DI: I also provide the female and male indices computed from AI-based histology profiles

Importantly,

- The higher the female dimorphism index of a woman is the more feminine she is
- The higher the male dimorphism index of a man is the more virile he is
- So, a high index value means different things for females and males
- The female and male populations should be treated separately

Covariates : miscellaneous

File covariates.csv also includes

SUBJID: the GTEx ID of the subject, e.g. GTEX-111CU

These are also used to index the columns of the gene expression matrix

Gene expression matrix

- The gene expression matrix is in `RNA_counts_adjusted.csv`
- Columns stand for subjects, except the first one which stands for gene names
- Rows stand for genes, except the first one which stands for SUBJIDs
- Entries are adjusted sequencing read counts: We tried as much as possible to normalize out the effect of technical covariates from the expression data


Your project step-by-step

Questions to be addressed

Overview of the project

In a nutshell, the overall goal is to (try to) formulate hypotheses about the biology behind organ-specific human sexual dimorphism

Q1: explore clinical and technical variables, and dimorphism indices

1. How are they distributed?
2. Are the variables correlated with one another?
3. Point of emphasis: Are the dimorphism indices associated with one another (RNA vs. histo indices)? With other variables ?
4. Point of emphasis: are some technical variables possibly confounding dimorphism indices and clinical variables ? 


Hint: Tools to address this may include PCA analysis, showing a variable's correlation matrix, specific plots,...

Hint: DTHHRDY is more akin to a categorical variable.

Hint: although AGE is discretized by decade, you may still treat it a numerical variable

Hint: Pay particular attention to disease states, these are relevant to clinical applications of the indices

Q2: the scale of DI-associated gene expression

1. Count the number of genes associated with the male and female dimorphism indices. 
2. Do the same with other clinical and technical variables.
3. What does it tell you about the magnitude of the dimorphism-associated gene expression signal? In particular, how does it compare to the signal associated with SEX and AGE?

Q2: the scale of DI-associated gene expression

Hint: You may filter transcripts beforehand. You don't need transcripts that are not/little expressed in your organ. It's also a good idea to focus on transcripts showing high variability (use the median average deviation, a.k.a. `mad()` in R). Don't hesitate to be drastic in your filtering (e.g. it's OK to remove >50% of all transcripts)

Hint: Use dedicated RNA-seq tools that take count data as input (e.g. DESeq2, edgeR, etc.)

Hint: These tools implements multivariate model formulas that will enable the treatment of potential confounders. For example, formula '`~AGE+X+Y`' will compute the multivariate association of morphological clusters with variables AGE, X and Y. So, if you look at AGE from this model it will be adjusted for the variations of X and Y.

Thus, DESeq2 and edgeR not only automatically handles normalization for total count and other intensity biases of count data, but it also enables you to cancel out your variables of choice in the analysis.

Hint: Could statistical power complicate your interpretation?



Q3: interpreting the DI-associated transcriptome

1. Which hormone receptors are associated with female and male dimorphism indices? With SEX?
2. Which pathways are associated with female and male dimorphism indices? With SEX?

Q3: interpreting the DI-associated transcriptome



Hint: You choose the hormone receptors, remember that we want to be as unbiased as possible with respect to what is or is not dimorphic.

Hint: The direction of the gene expression variation matters.

Q3: interpreting the DI-associated transcriptome

Hint: For computational ease, you may use the fgsea package, which takes as input a gene ranking (i.e. provided by the differential analyses of Q3 output)

(Yes, I know, it's not using the the gold standard sample permutation of GSEA. But I tried it for you and it's reasonable in this context where the transcriptional signal is quite massive)

Hint: The REACTOME gene sets are provided in file c2.cp.reactome.v7.5.1.symbols.gmt in the resource folder. The fgsea package provides a function to read *.gmt files

Your project step-by-step

Handing out your report

Project report

You are asked to describe your work in a written report

It includes :

- An short introduction to the topic
- A careful description of methods and results
- A discussion of your results for each questions

Dumping code and results is *not* good enough, you must show me you understand what you are doing

Project report

- The report is handed out as a **PDF** file or a **PDF-converted Jupyter notebook**
- You may put the larger outputs as appendices to the report, or hide them in your notebook, and discuss the most salient findings in the main text
- I am not defining a specific report length. Good reports are neither too short, nor too long. 15-20 pages (appendix not included) seems reasonable.

Project report

- This report is a good exercise for your master thesis and for scientific writing in general
- Try to be specific, precise and quantitatively accurate. For example, don't write '*many genes are associated with the index*', but '*N genes are associated with the index*', where N is the number you have computed

Grading of your projects

- Questions Q1-Q3: **15pt**
- Quality of your PDF report (it needs to be concise, clear, precise, well presented), **3pt**
- Reproducibility: all your analyses must be reproducible, i.e. all the steps are carefully documented, including parameter settings and software versions, **2pt**
- ...and **2pt** to be gained beyond 20 for the motivated ones who go off the beaten track with original analyses, etc.

Handing out your project report

- Deadline is **June 22nd 23:59:59**.
- Send me the report by e-mail (Vincent.Detours@ulb.be), **including 'BINF-F401' in the subject line**
- Get back to me if I don't acknowledge receipt of your report within 3 days.

Your project step-by-step

Rules of the game

Groups

- There will be 10 groups of 2 students
- Each group is assigned a specific organ
- You will be informed by mail about which group you've been assigned to if you haven't chosen one by March 20th, when each group will be informed about where their data is

Rules of the game

Choose your own weapons. I do advise R, but I'll accept alternatives.

You are encouraged to communicate with one another, but

- Remember, each group has its own unique dataset
- Beware of herd effects, the majority and/or the leading figures can be dead wrong
- Plagiarism is not acceptable
- ChatGPT and other AI text generators are unreliable for scientific facts, they easily spit-out falshood with an impeccable style, including made-up citations of papers that don't exist! **It's up to you to use or not AI text generators as long as 1- you clearly insert AI generated text in quote marks and 2- say which AI wrote it and provide the prompt used to generate the text.** Of course you will be held responsible for all the text in your report.

Rules of the game

I'll address your questions by e-mail, if

- You use your brain and do your home work before asking, e.g. I won't reply if the answer is in the R documentation or can be obtained from a basic web search. (I basically request the same etiquette as in any technical forum on the Internet.)
- The subject line of all your correspondance with me must start with 'BINF-F401: '

Some tips from past experience...

- This is a science project, your technical choices must be argued and the calculations must be reproducible, i.e. all the information needed to rerun the analyses must be provided. *In science details do matter.*
- Dumping graphics and stats in a document is not enough. The students who got the best grades examined critically their results and demonstrated that they understood the limits and biological significance of what they did.
- Technical problems are hard to anticipate, so don't wait for the last minute to discover them, and then be left with no time to overcome them.