Big Data in Medicine

Thomas Mohr

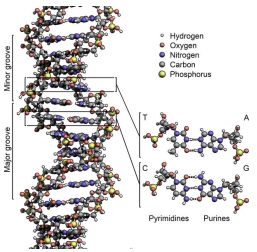
Institute for Analytical Chemistry - University of Vienna, Vienna.
Center of Cancer Research - Medical University of Vienna, Vienna.
ScienceConsult - DI Thomas Mohr KG, Guntramsdorf.

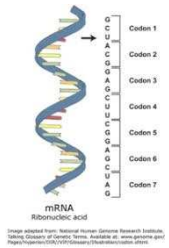
SS2024

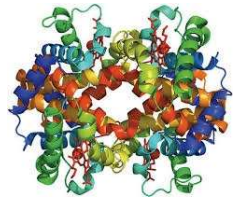
Second Lecture

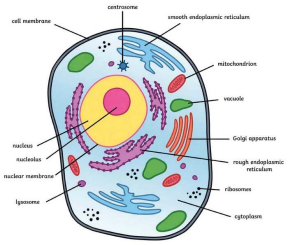
The biological foundations of -omics data.

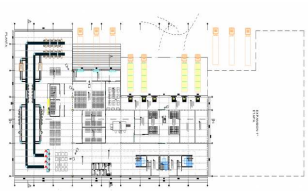
How does the plan of life function?

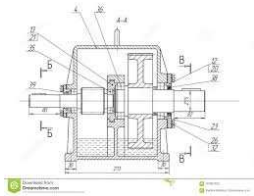


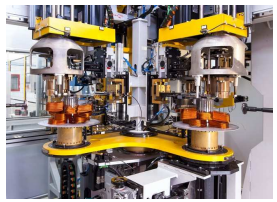


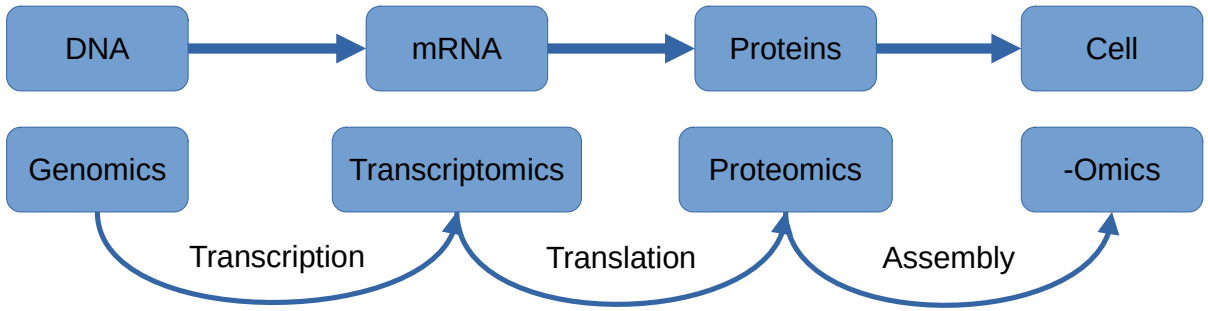












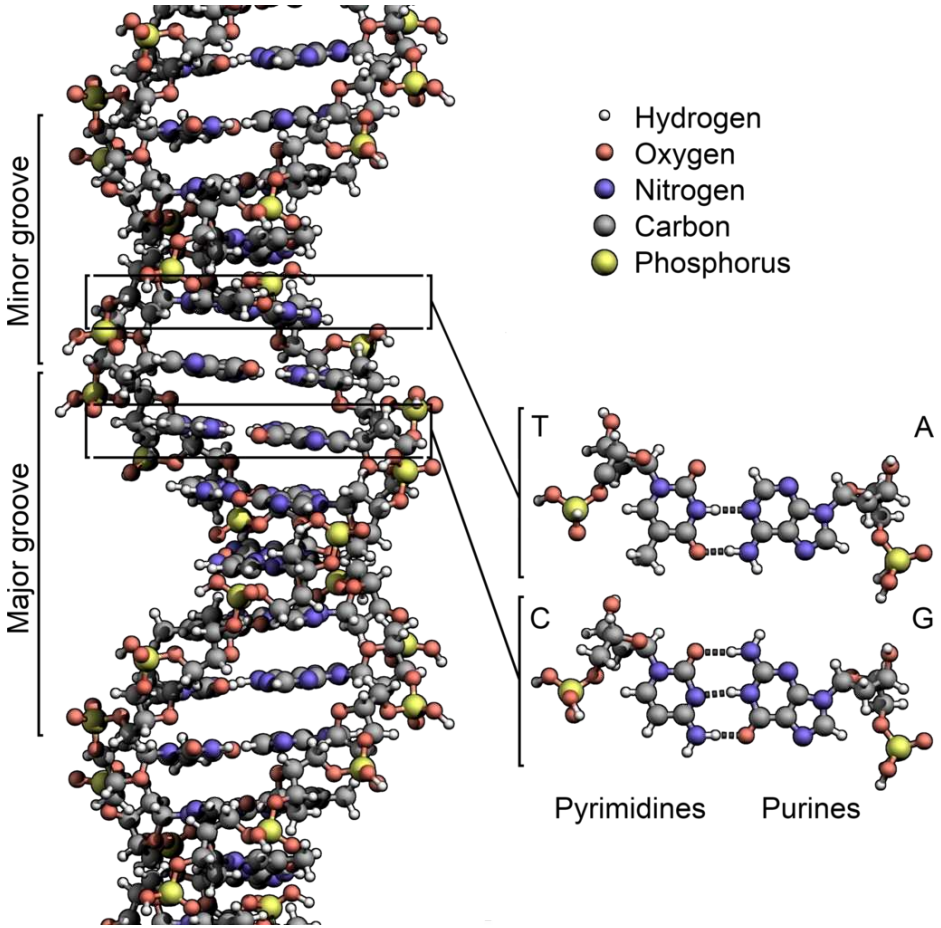
DNA (desoxyribonucleic acid) is the master plan of life.

DNA is a molecule composed of two so-called *polynucleotide* strands that form a double helix and contain the genetic information of an organism.

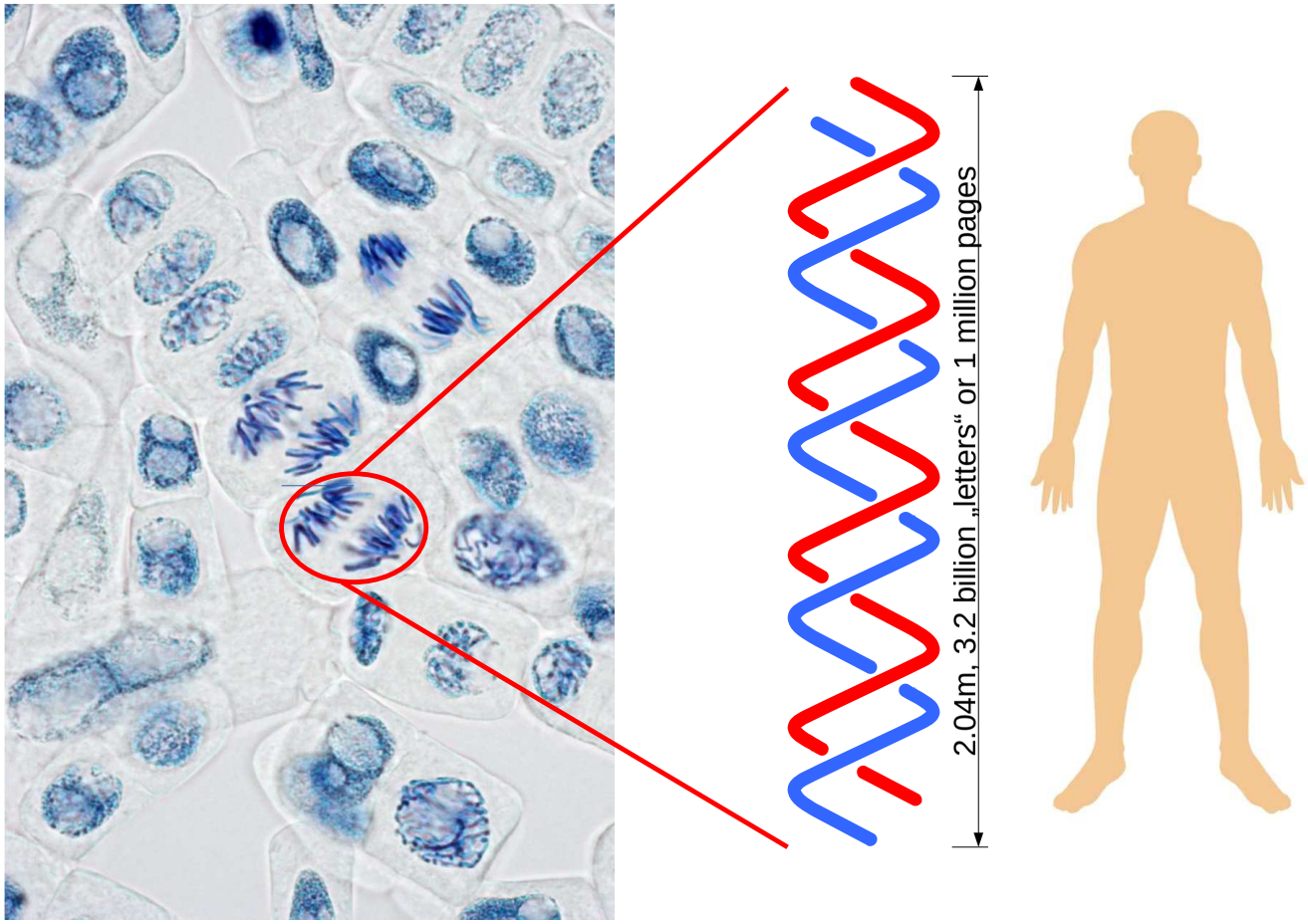
It is composed of the subunits pyrimidine bases cytosine (C) and thymidine (T) and the purine bases adenine (A) and guanosine (G).

These four bases, C, T, A and G, stand for the letters of life and enable a DNA to encode all relevant information to build a cell.

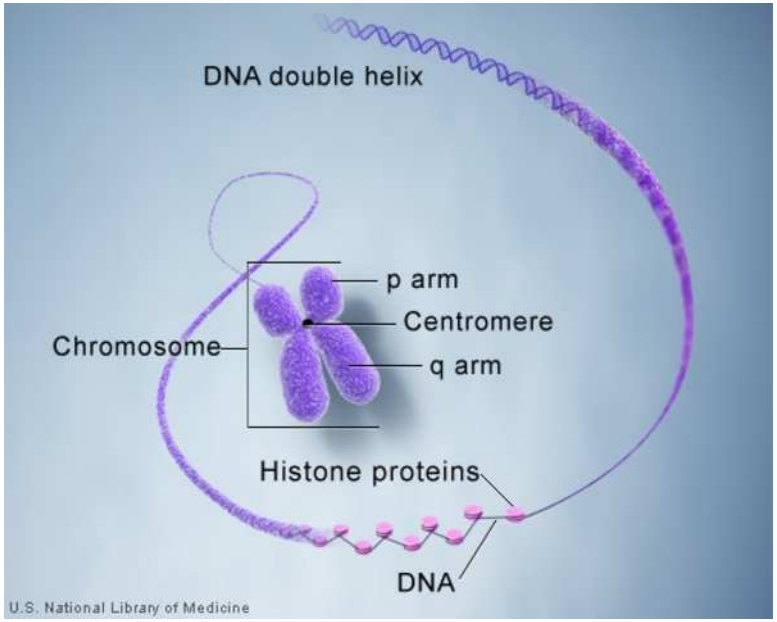
How does DNA look like?

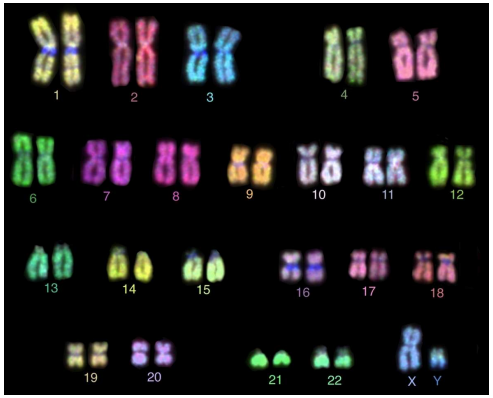


DNA in Numbers



How is DNA organised?

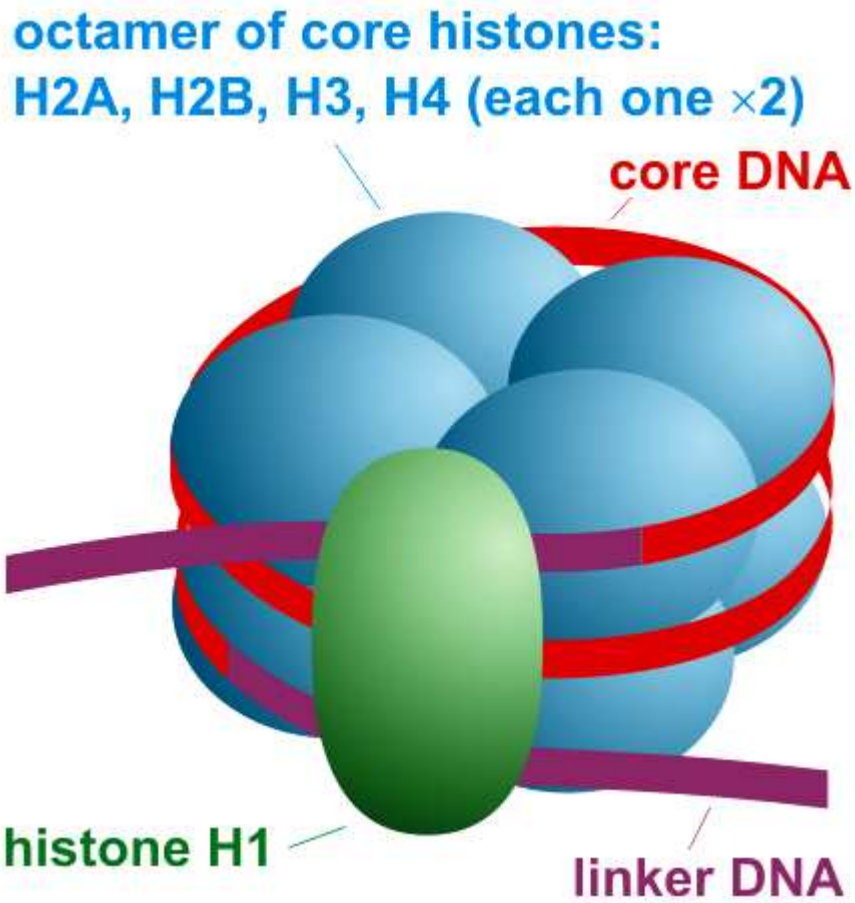




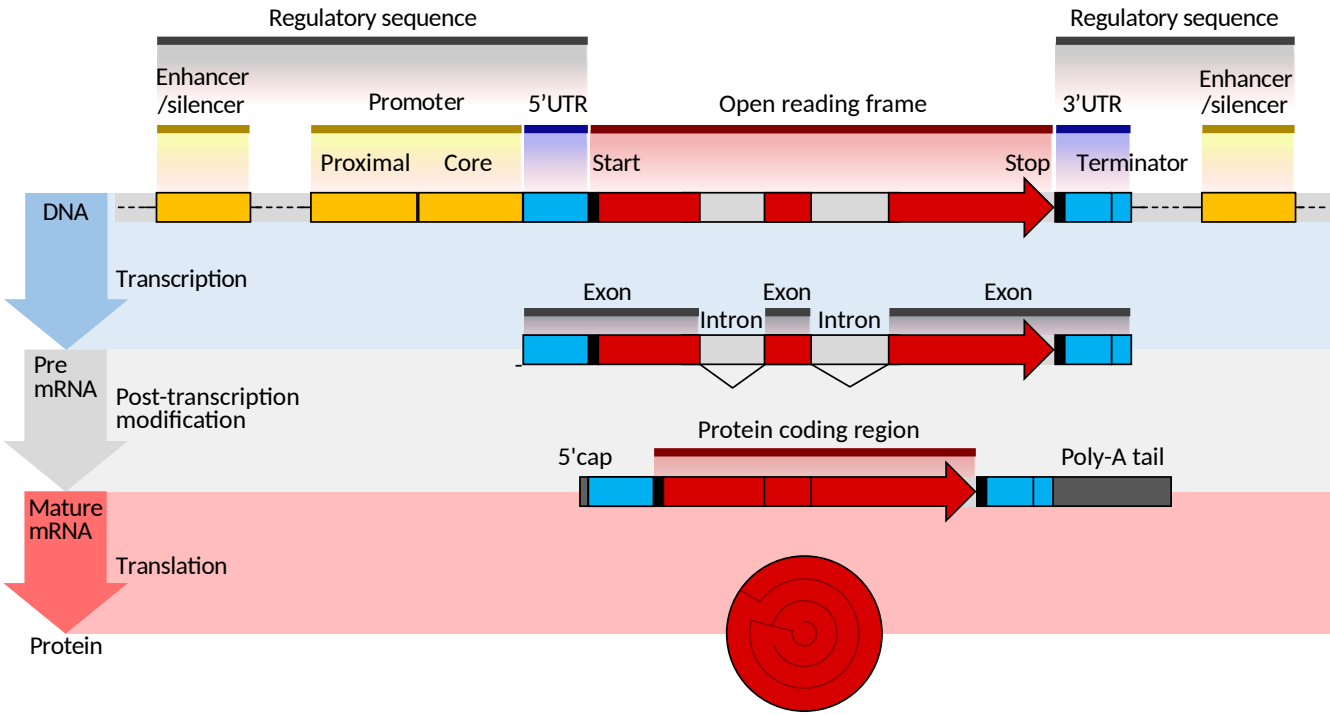
MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US); [updated Jun 24; cited 2020 Jul 1]. Available from: https://medlineplus.gov/.

Genes & Health: https://www.genesandhealth.org/genes-your-health/46-%E2%80%93-magical-number

How is DNA organised - the role of histones.



The paragraphs of life - the genes.



Shafee T, Lowe R (2017). "Eukaryotic and prokaryotic gene structure". WikiJournal of Medicine 4 (1). DOI:10.15347/wjm/2017.002. ISSN 20024436.

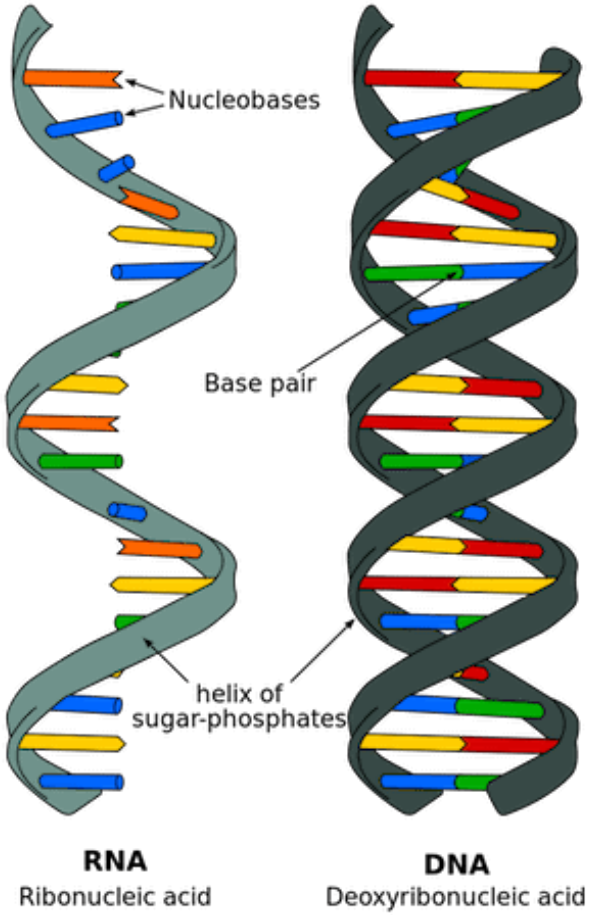
What is mRNA (messenger Ribonucleic acid)?

mRNA is a single-stranded polynucleotide strand consisting of a chain of cytosine, thymidine, guanosine, and uracil bases linked by a sugar backbone.

RNA differs from DNA in two respects: First, the usage of uracil instead of thymidine, and second, the use of ribose instead of desoxyribose as a backbone.

These differences result in less chemical stability, thus facilitating gene expression regulation by mRNA degradation.

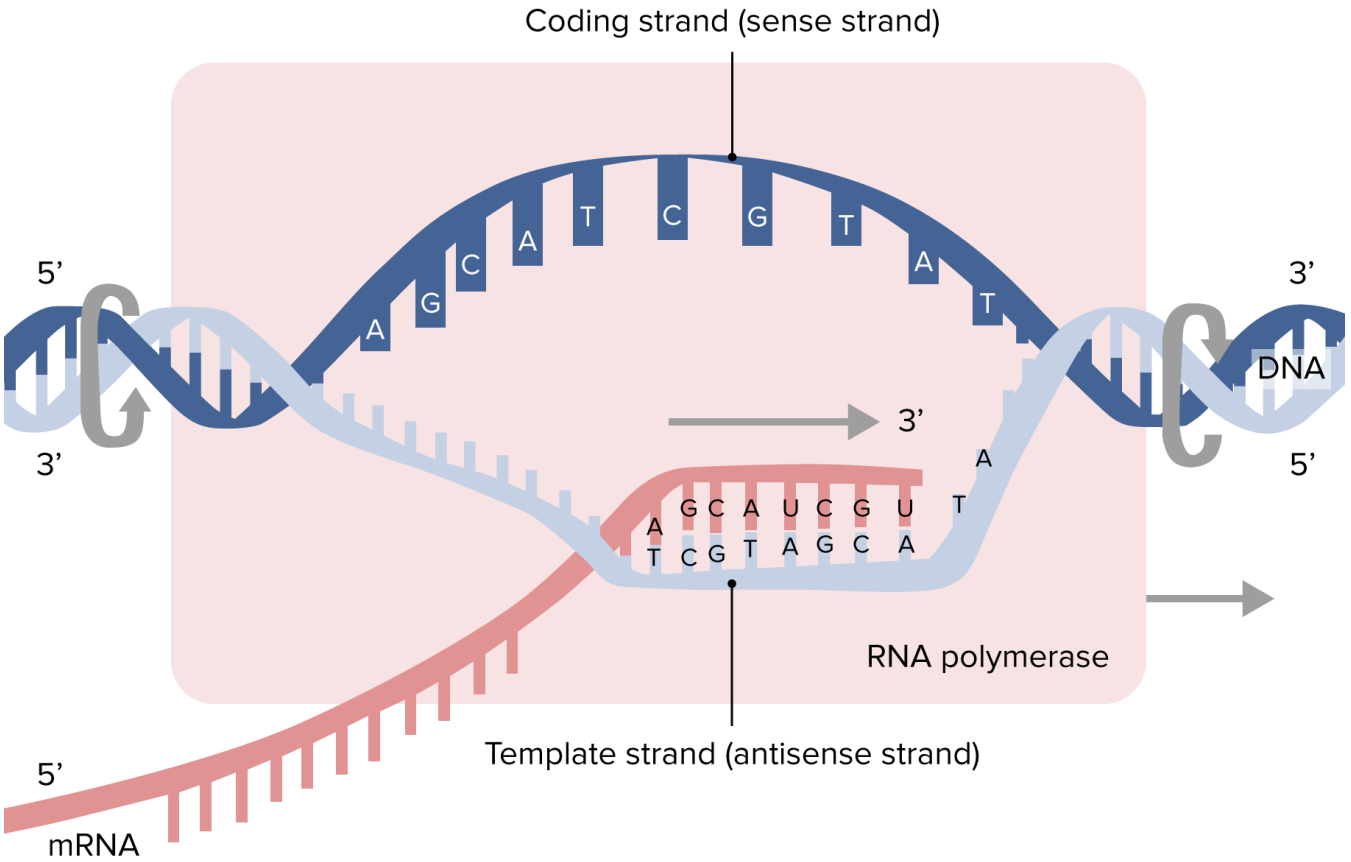
How does mRNA look like?



How is DNA transcribed into mRNA?

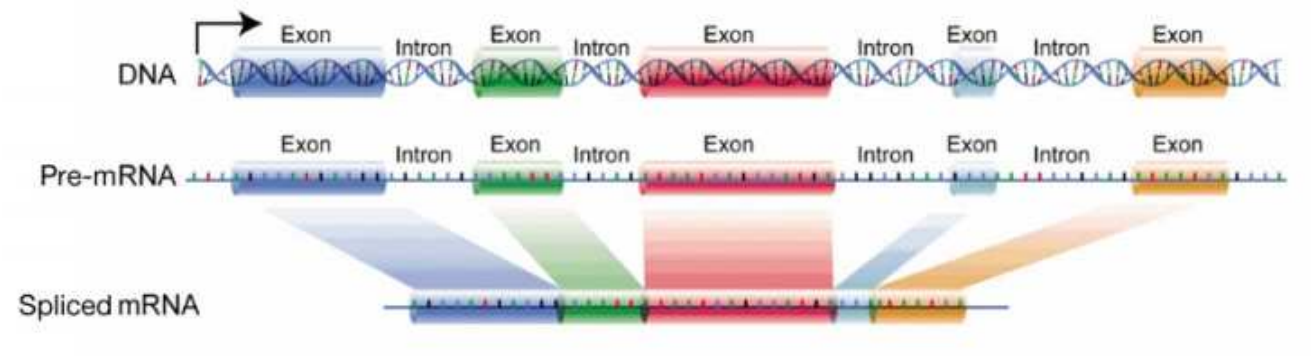
1. A so-called transcription factor binds to a gene's promoter region and initiates the RNA polymerase's binding.
2. RNA polymerase detwists the DNA double-helix and separates the two strands.
3. RNA polymerase adds RNA nucleotides (complementary to one DNA strand's nucleotides).
4. RNA sugar-phosphate backbone forms with assistance from RNA polymerase to form an RNA strand.
5. Hydrogen bonds of the RNA–DNA helix break, freeing the newly synthesised RNA strand.

How is DNA transcribed into mRNA?



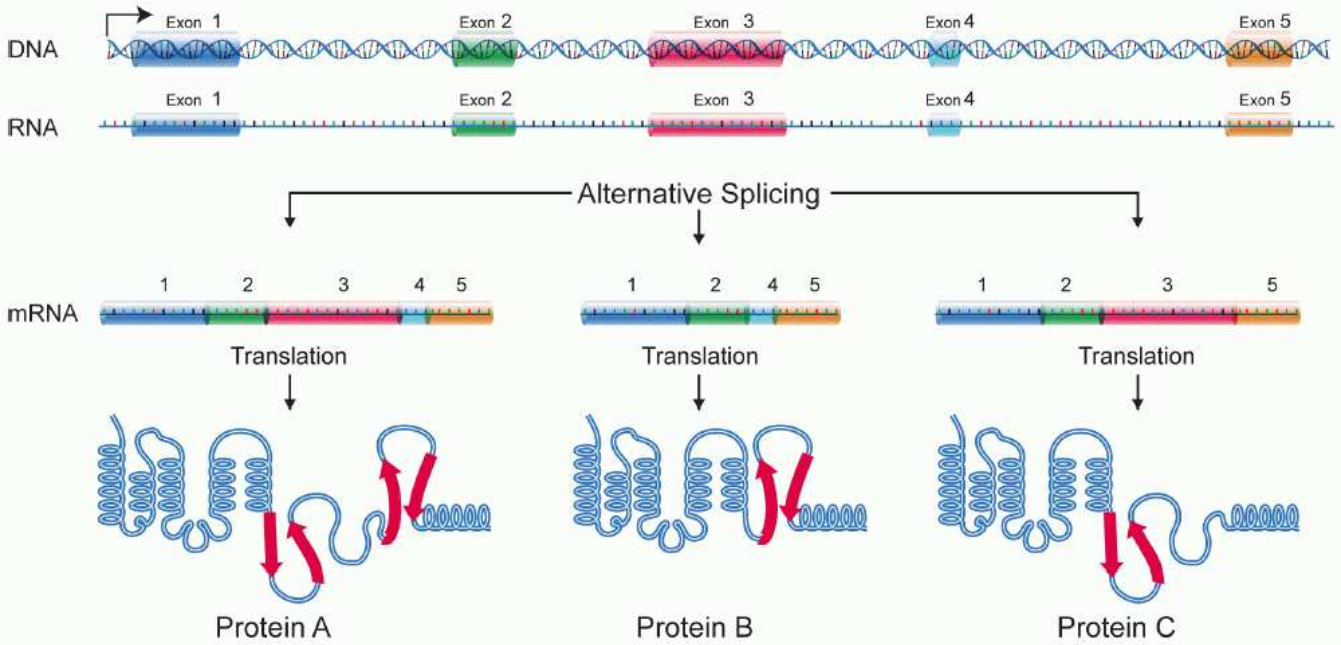
https://www.lecturio.com/

How to prepare a functional mRNA - exons and introns.



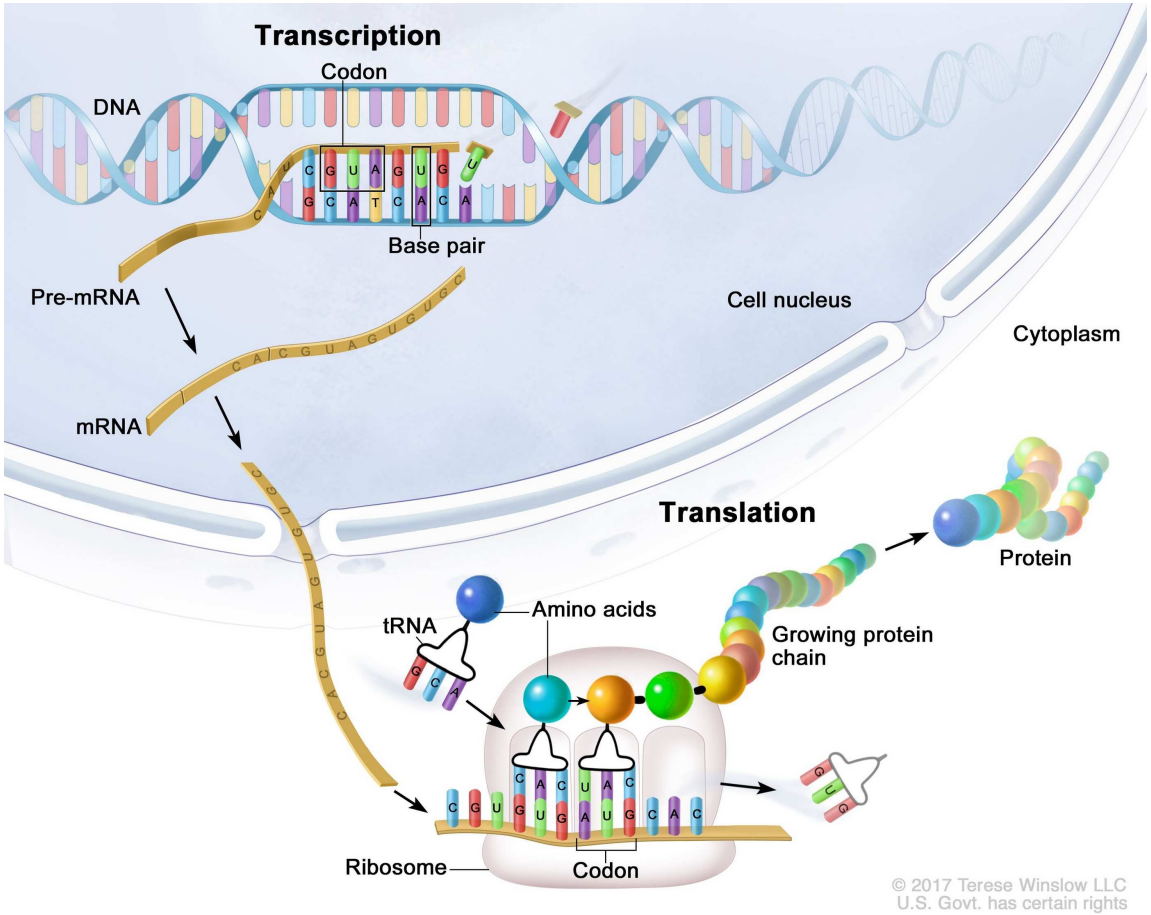
Chemistry Libre Texts, University of Arkansas

Exons and Introns - the key to even more diversity.



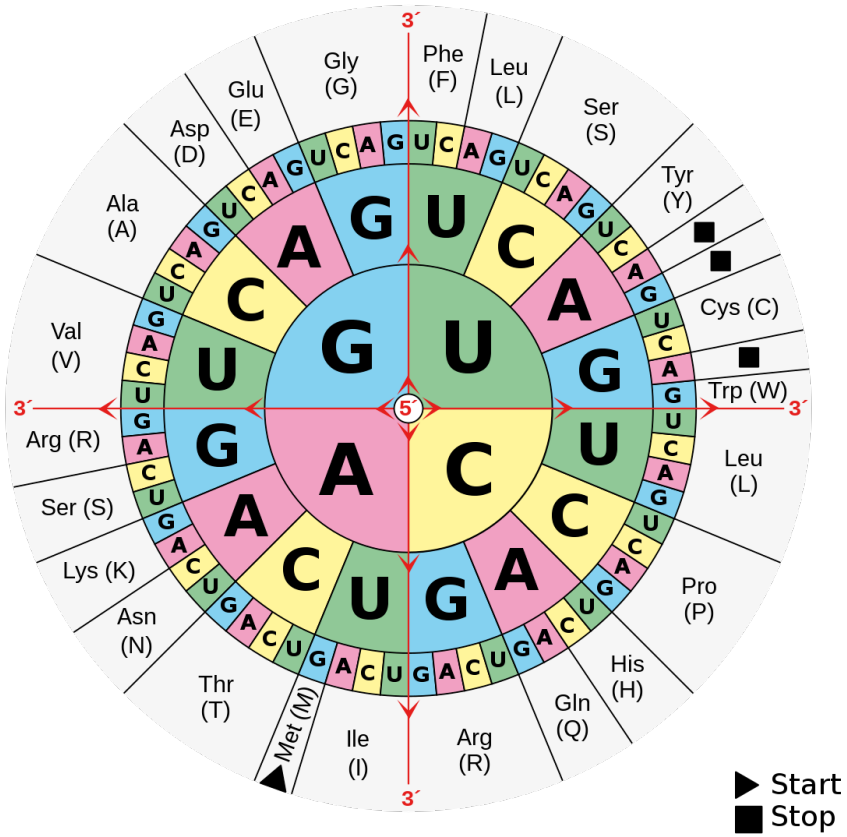
National Human Genome Research Institute

How do we get proteins? The process of translation.



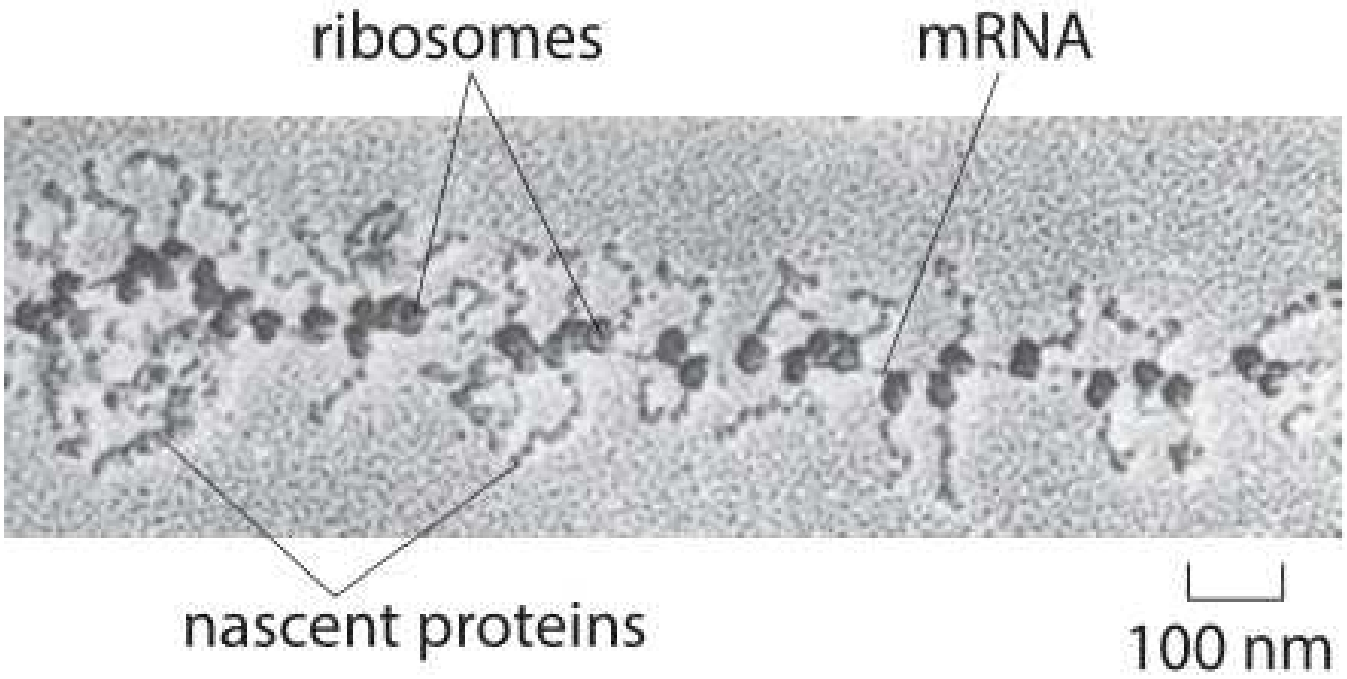
Teresa Winston LLC

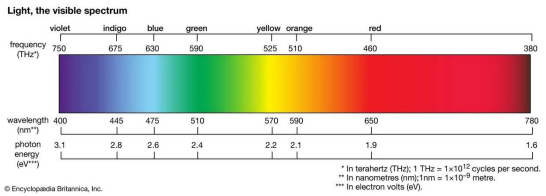
The genetic Code - the letters of life



https://de.wikipedia.org/wiki/Genetischer_Code#/media/Datei:Aminoacids_table.svg

How does the translation process look in the microscope?

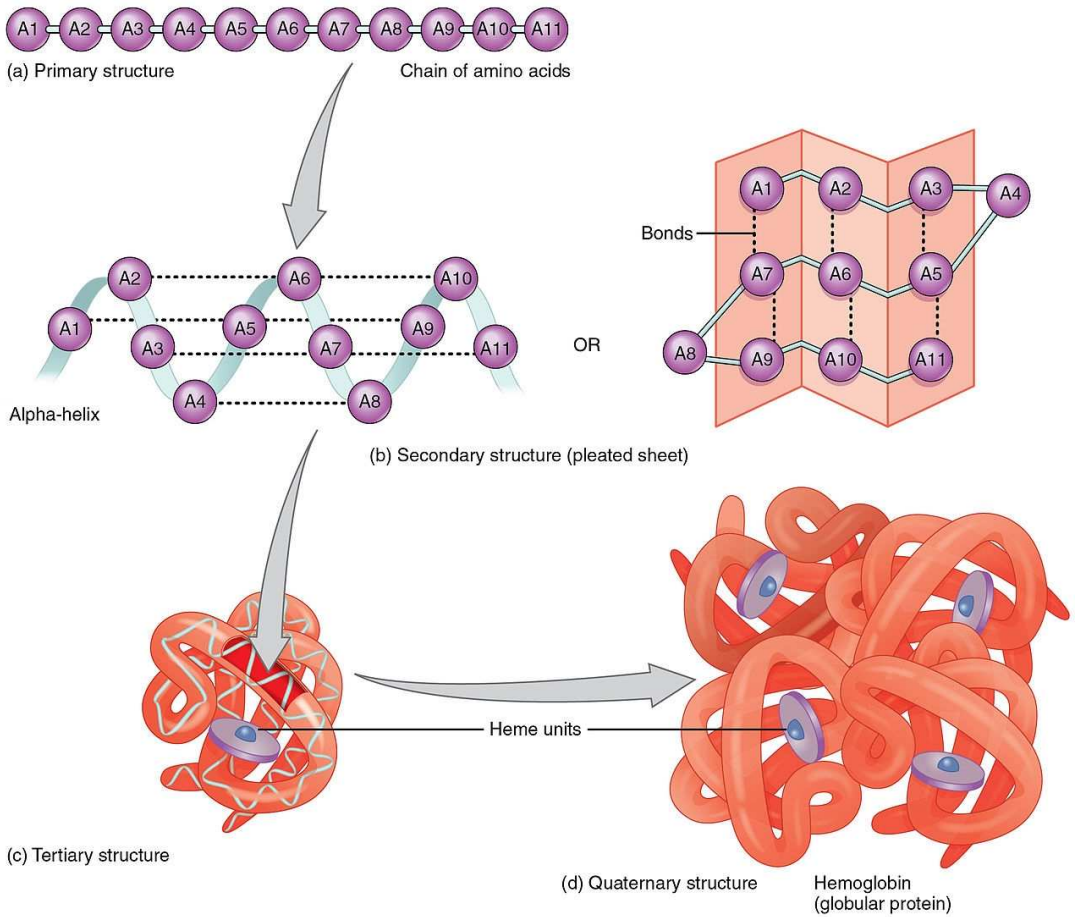




Milo, R., & Phillips, R. (2015). Cell biology by the numbers. CRC Press,

Encyclopedia Britannica

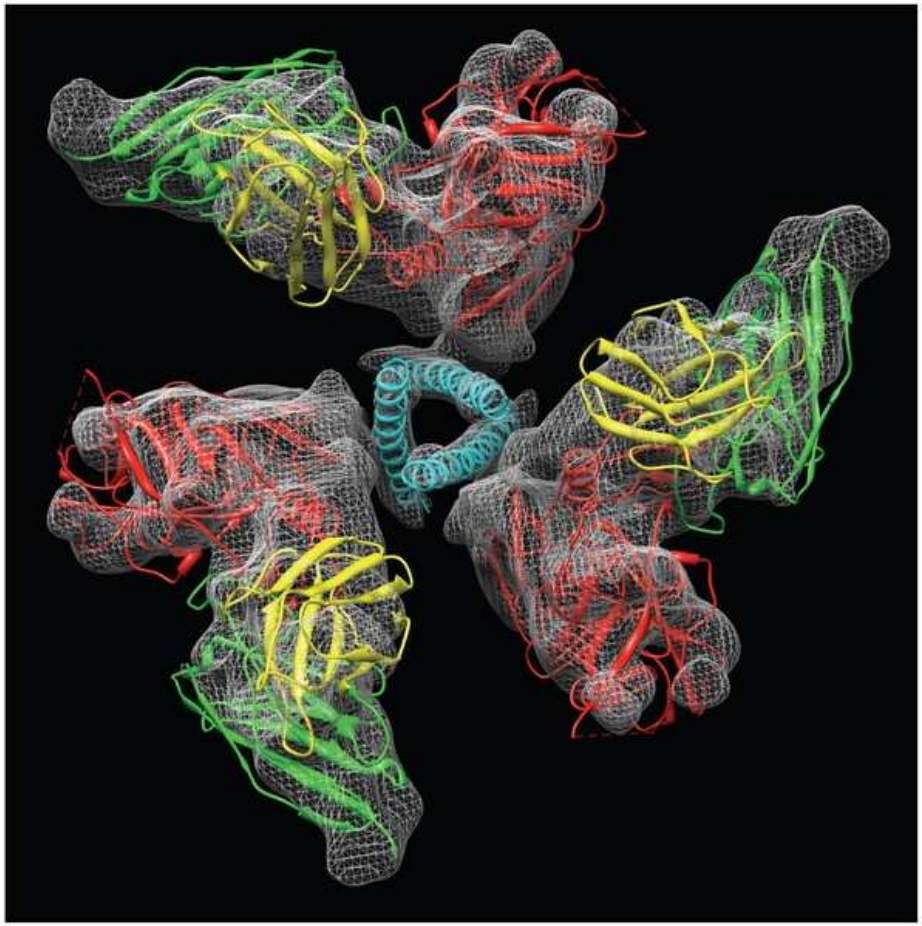
What are the structures of proteins?



Milo, R., & Phillips, R. (2015). Cell biology by the numbers. CRC Press,

Encyclopedia Britannica

How does that look like in the microscope?



National Institute of Health

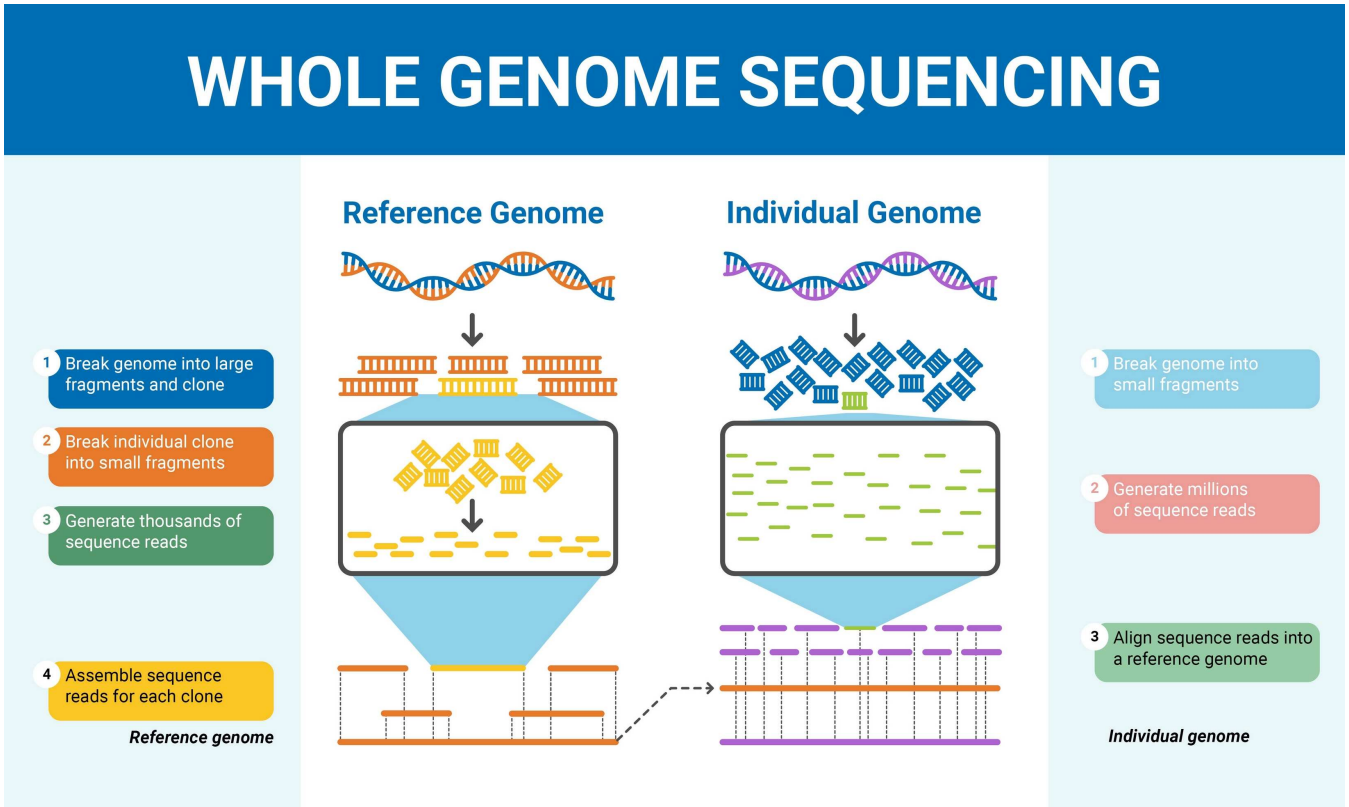
The wet-lab equipment behind your data

The lab itself - movie vs reality

The labs - what does the work look like?

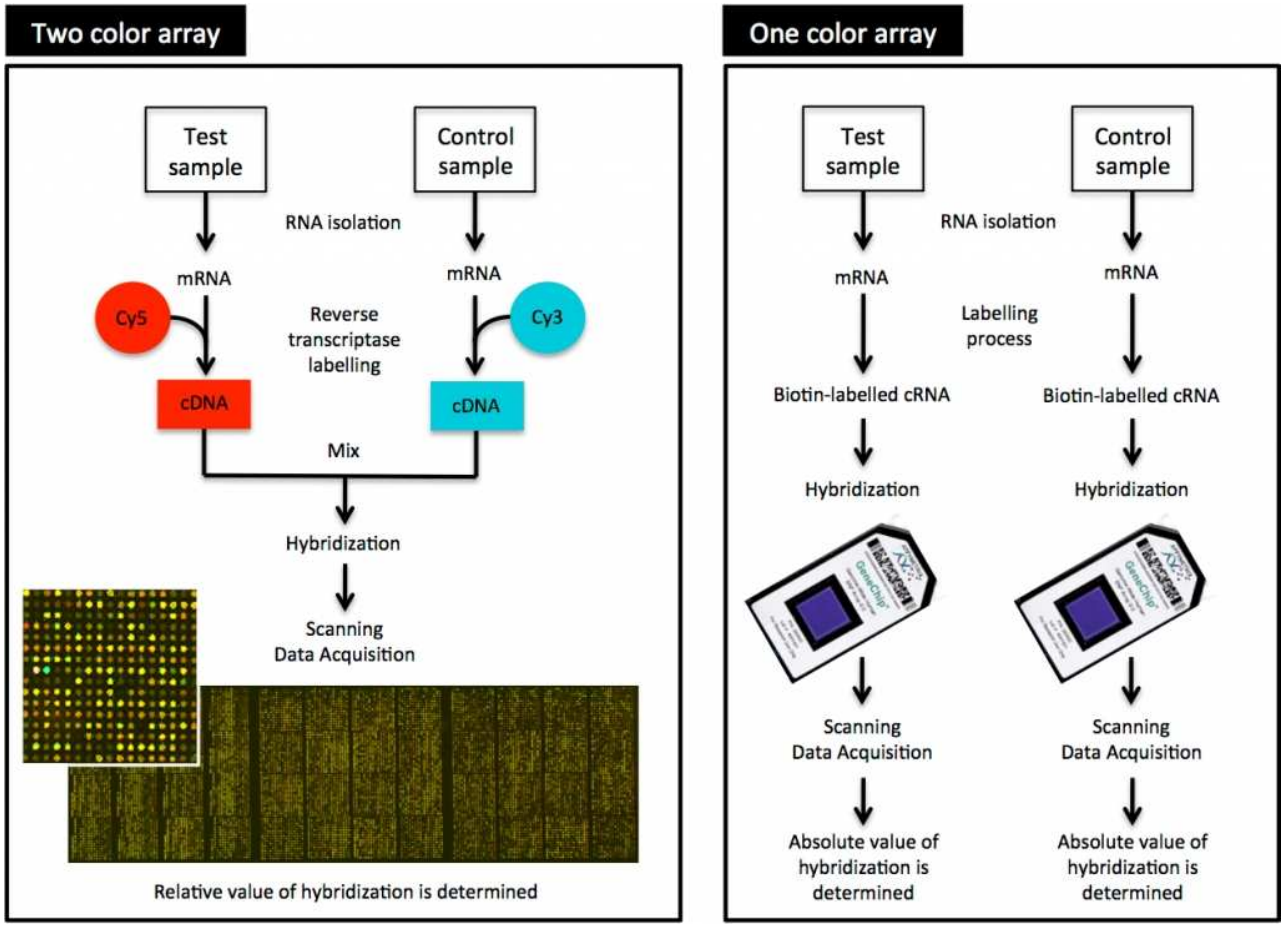
Weill Cornell Graduate School of Medical Sciences.

Starting with DNA - whole genome sequencing



https://sequencing.com/education-center/whole-genome-sequencing

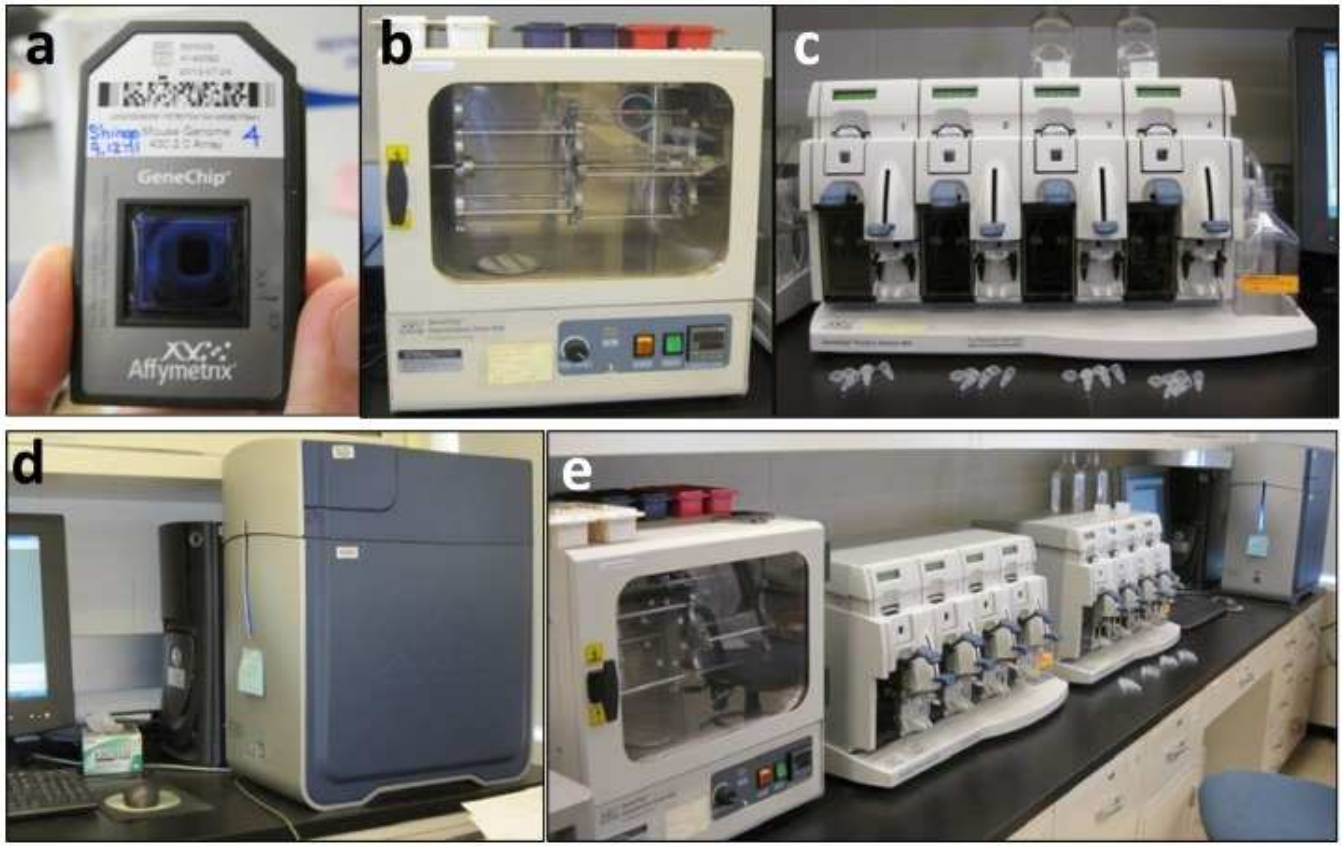
The oldest -omics technology - DNA Microarrays



How does this look in the lab?

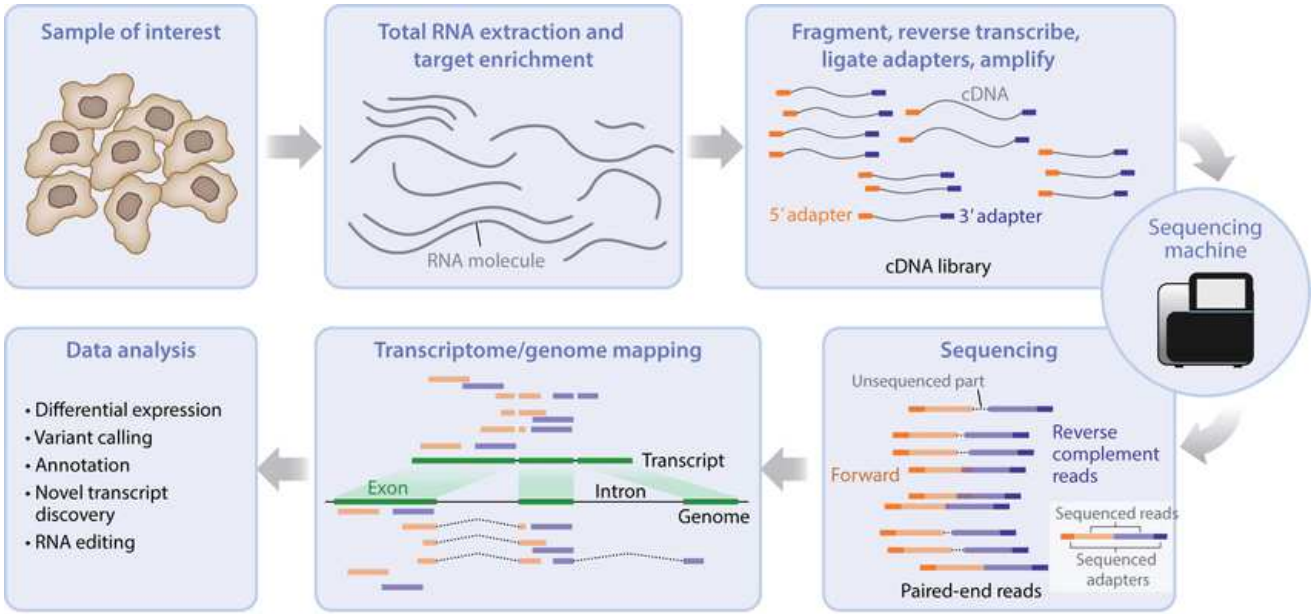
University of Berkeley, PhyloChip: DNA Microarray for Rapid Profiling of Microbial Populations IB-2229

How does that look like in the lab?



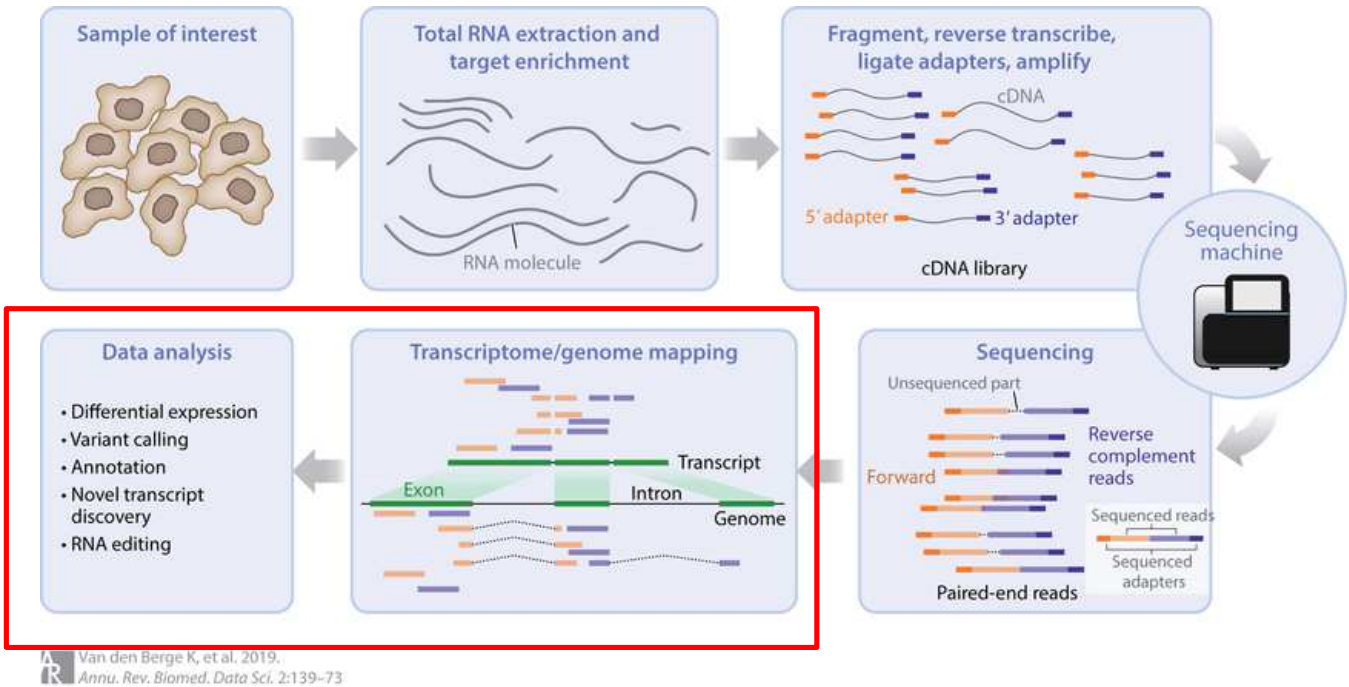
Blumenberg, Miroslav. (2012). SKINOMICS: Transcriptional profiling in dermatology and skin biology. Current genomics. 13. 363-368. 10.2174/138920212801619241.

The next step - RNA Sequencing



Van den Berge K, et al. 2019.
Annu. Rev. Biomed. Data Sci. 2:139–73

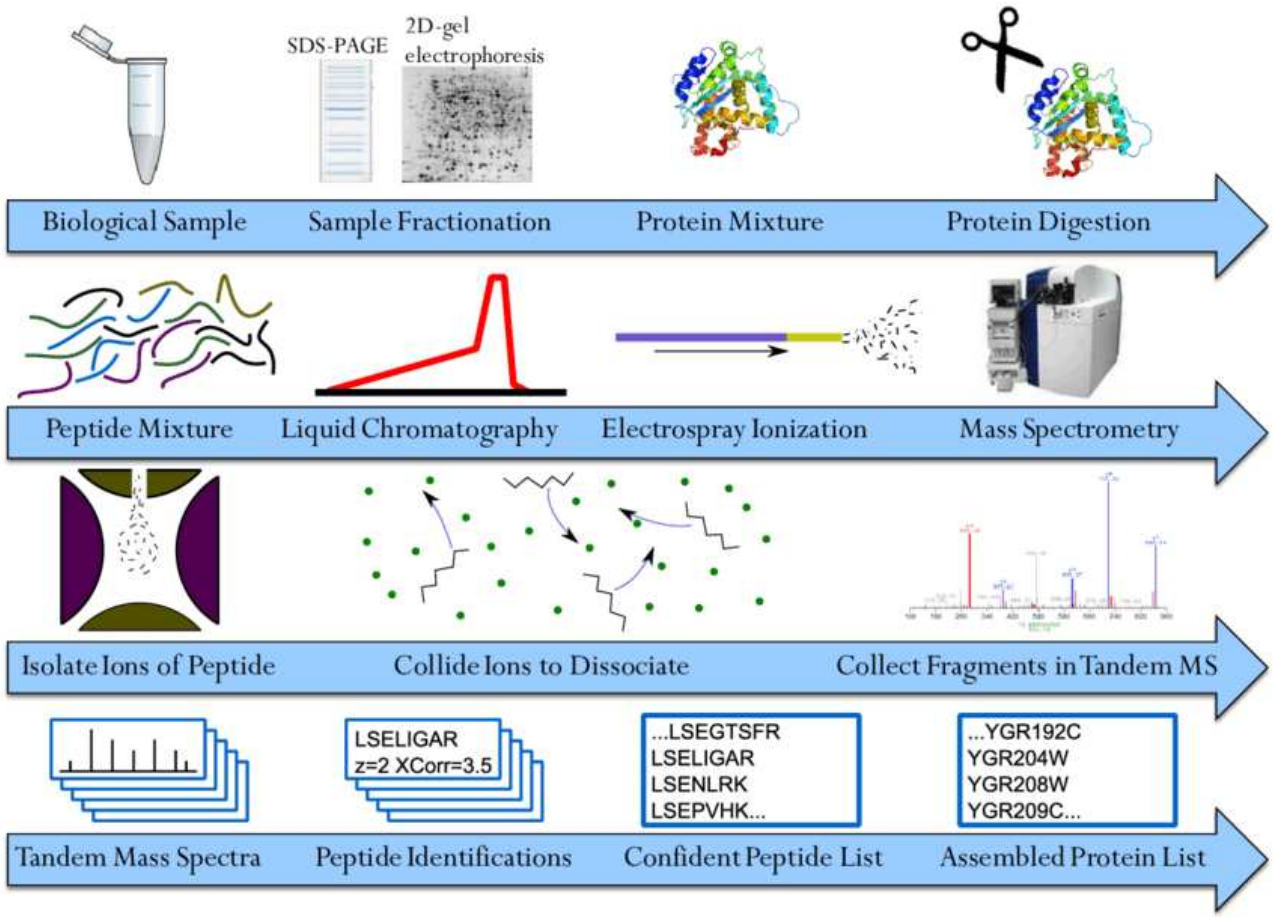
The next step - RNA Sequencing - where do bioinformaticians come into play?



How does RNA Sequencing look in the lab?

https://omegabioservices.com

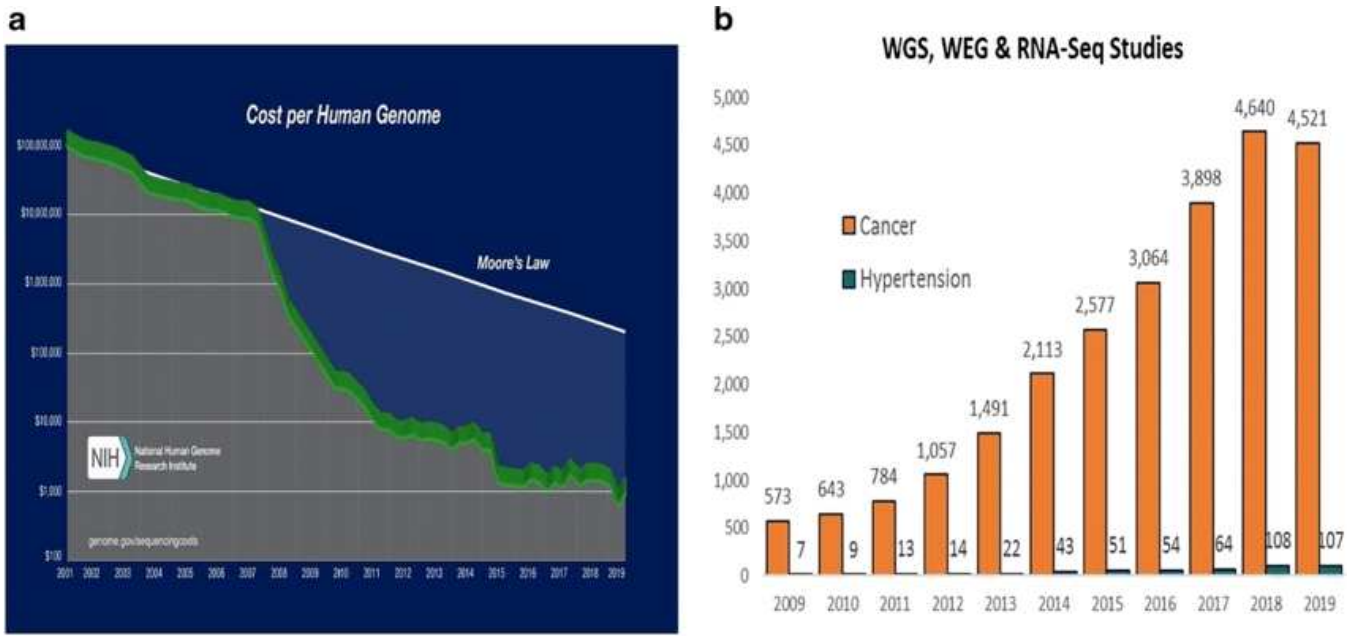
Moving it one level higher - Mass Spectrometry



Chen, Yao-Yi & Dasari, Surendra & Ma, Ze-Qiang & Vega-Montoto, Lorenzo & Li, Ming & Tabb, David. (2012). Refining comparative proteomics by spectral counting to account for shared peptides and multiple search engines. Analytical and bioanalytical chemistry. 404. 1115-25. 10.1007/s00216-012-6011-x.

Mass Spectrometry - seen in the lab

How much does that cost?



Mueller, Franco. (2020). AI (Artificial Intelligence) and Hypertension Research. Current Hypertension Reports. 22. 10.1007/s11906-020-01068-8.

What is the output?

The output of these experiments are

- An matrix with m features $\times$ n samples
- A matrix with n samples $\times$ o traits
- A matrix with m features $\times$ p feature annotations.

It is now **your** task to put context behind the data.

E-learning assignment

Read:

Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, and Smyth, GK (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7),

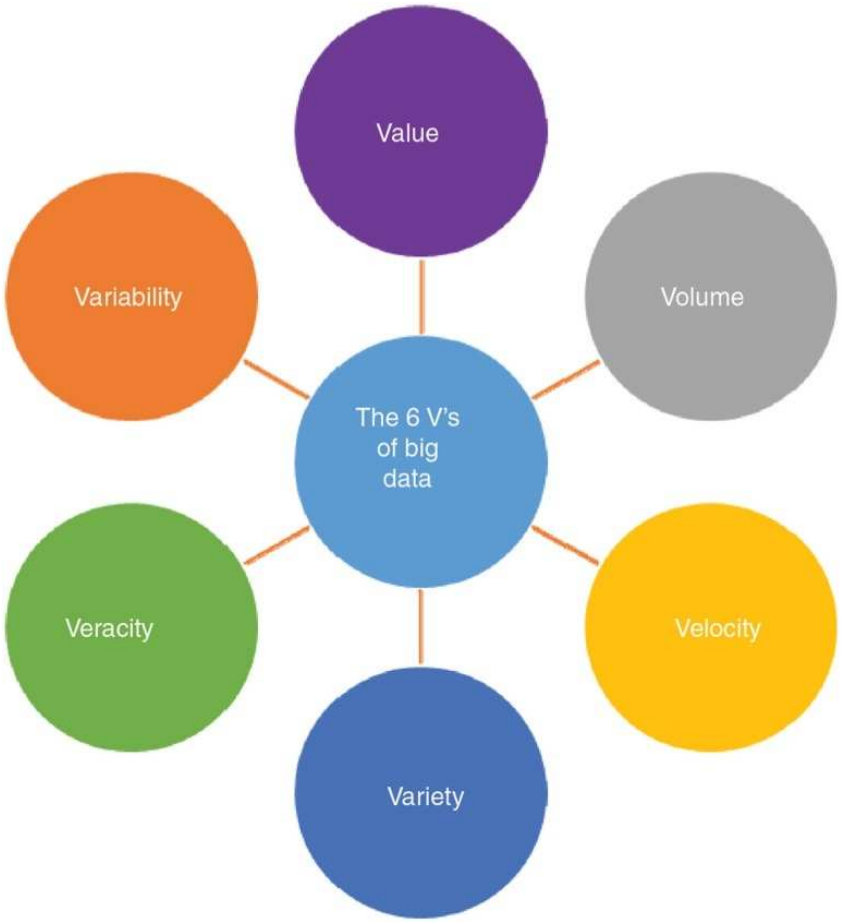
and

Phipson, B, Lee, S, Majewski, IJ, Alexander, WS, and Smyth, GK (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. Annals of Applied Statistics 10(2), 946–963

Furthermore, prepare the linear models approach.

Third lecture

The six V's of Big Data



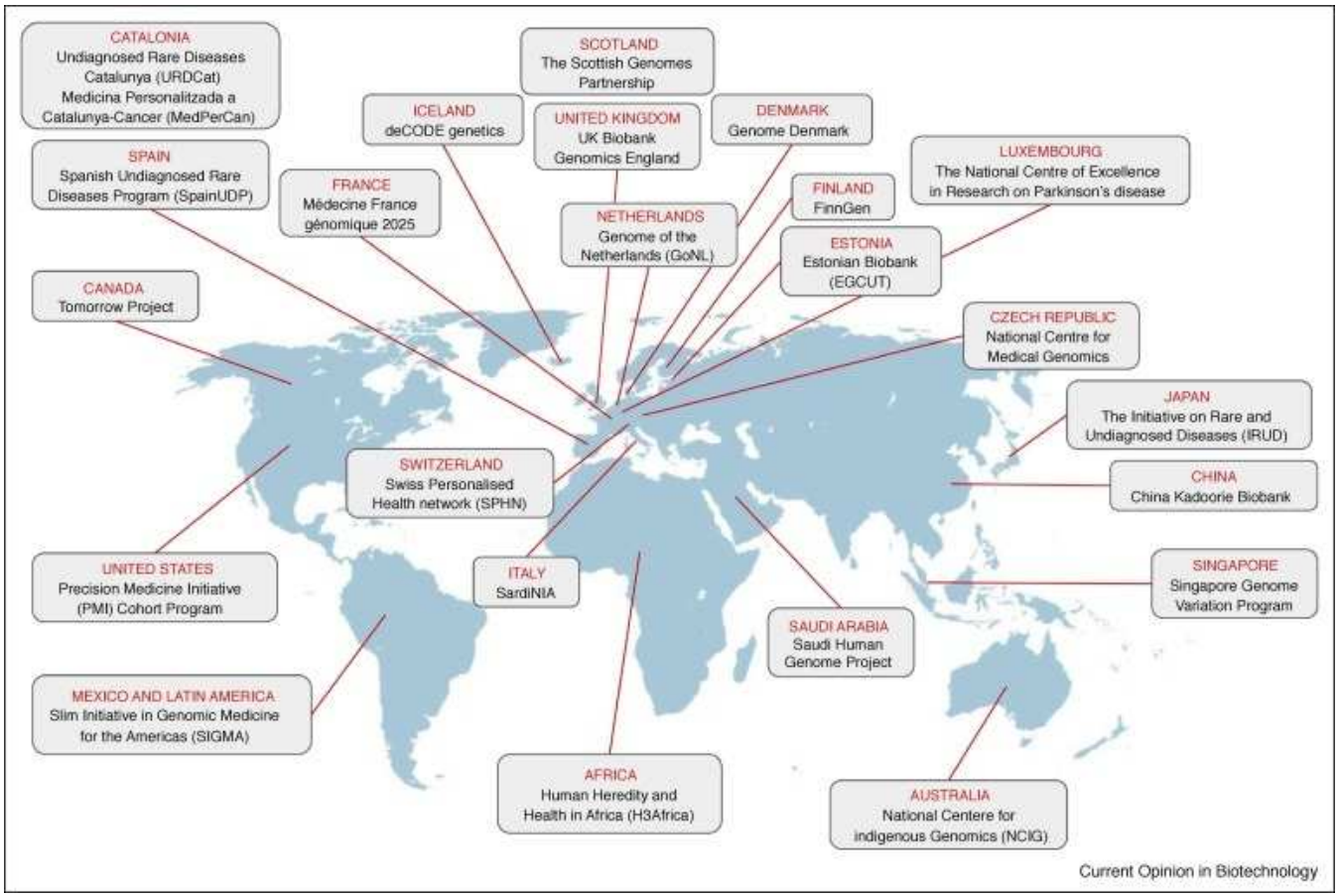
Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. J Integr Bioinform. 2018 May 10;15(3):20170030. doi: 10.1515/jib-2017-0030. PMID: 29746254; PMCID: PMC6340124.

Technologies used in big data analysis

Technology	Percentage
Statistical analysis	47.6%
Data mining	39.0%
Data visualization	34.1%
Structured Query Language	28.0%
Parallel processing	7.3%

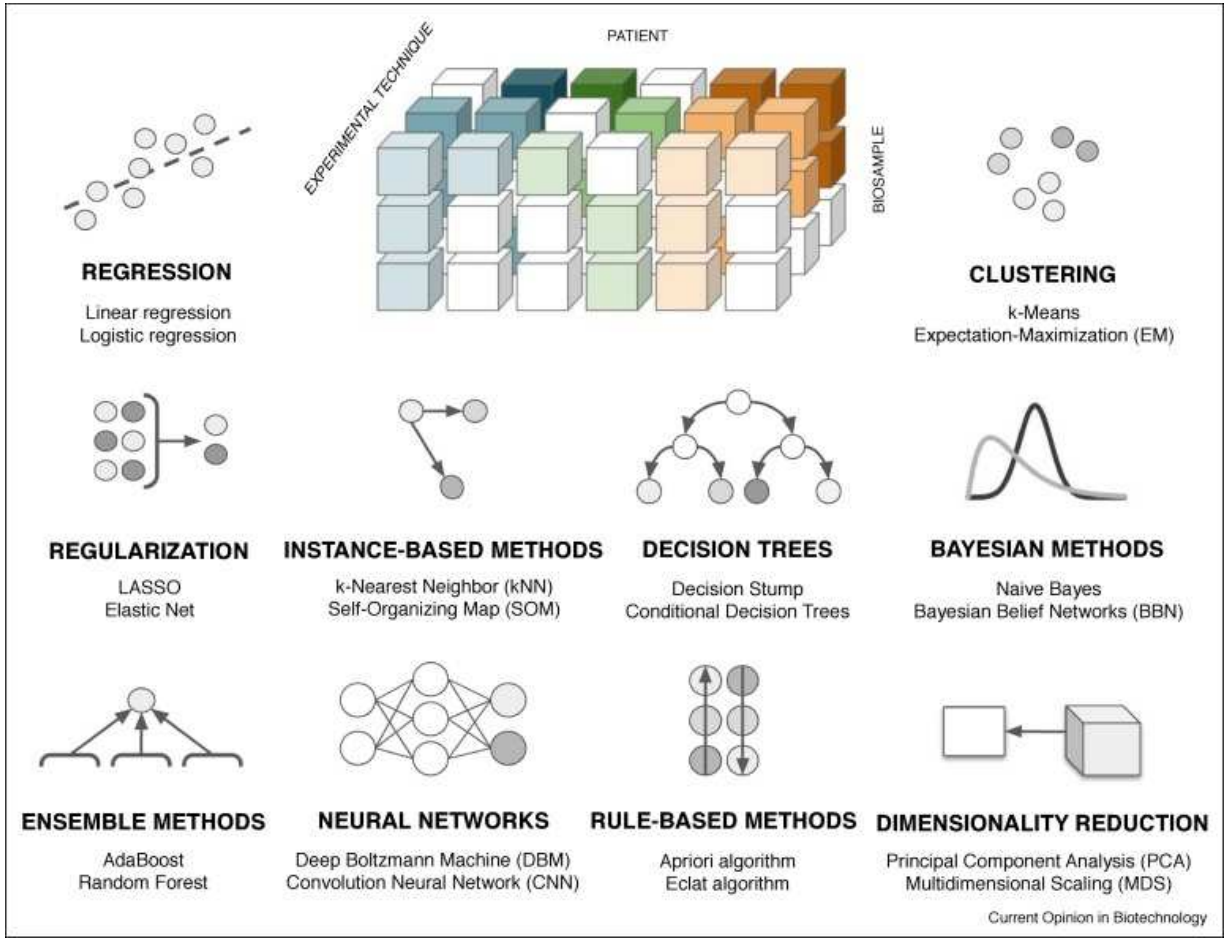
Dolezel D, McLeod A. Big Data Analytics in Healthcare: Investigating the Diffusion of Innovation. Perspect Health Inf Manag. 2019 Jul 1;16(Summer):1a. PMID: 31423120; PMCID: PMC6669368

Example: Geographic scope of ongoing population-scale sequencing initiatives for Personalized Medicine



Davide Cirillo, Alfonso Valencia, Big data analytics for personalized medicine, Current Opinion in Biotechnology,Volume 58, 2019, Pages 161-167, ISSN 0958-1669

Machine learning algorithms for multi-view data analysis.



Davide Cirillo, Alfonso Valencia, Big data analytics for personalized medicine, Current Opinion in Biotechnology,Volume 58, 2019, Pages 161-167, ISSN 0958-1669,

How does the amount of data change during the analysis?

Omics data are HUGE initially, but shrink dramatically during analysis

Example:

Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations (EGAS00001001563)

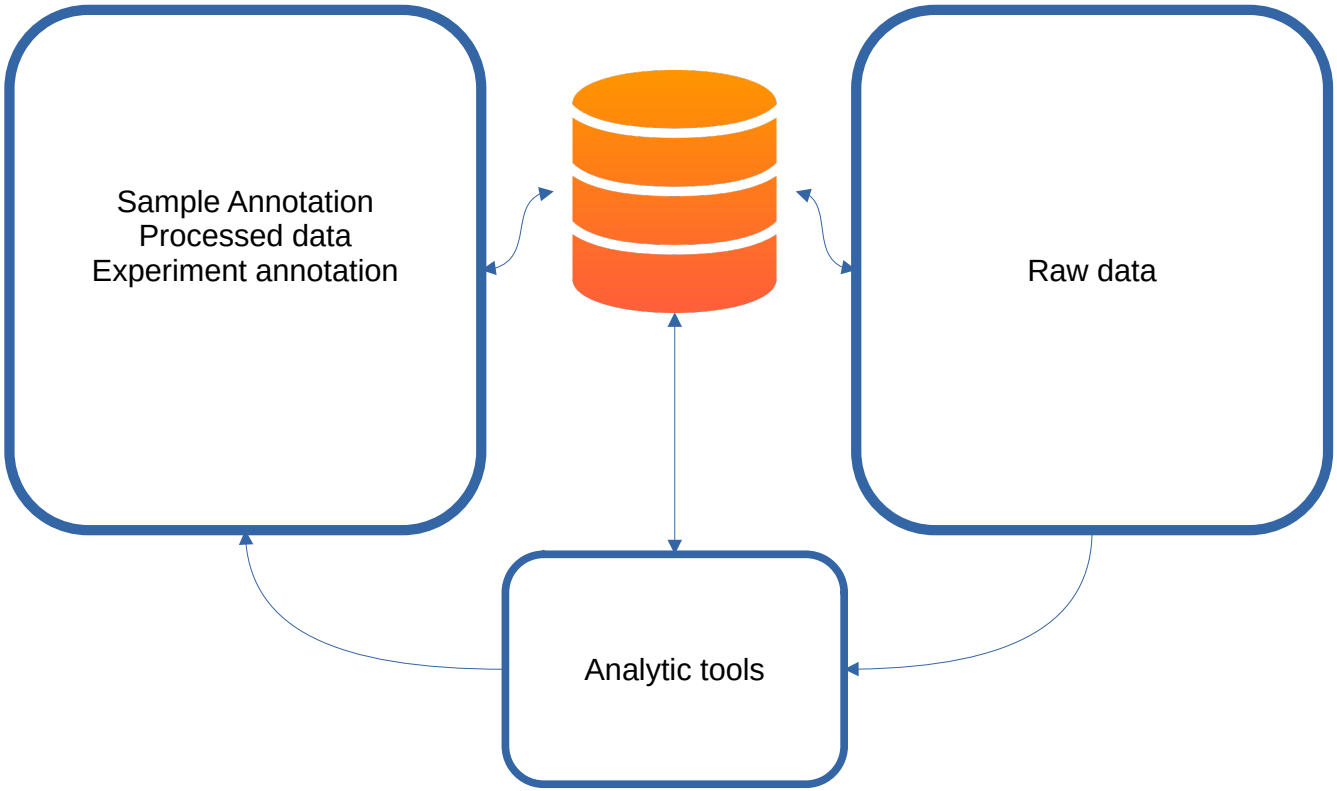
216 malignant pleural mesothelioma (MPM) tumors:

- Transcriptomes (n = 211),
- whole exomes (n = 99)
- targeted exomes (n = 103)

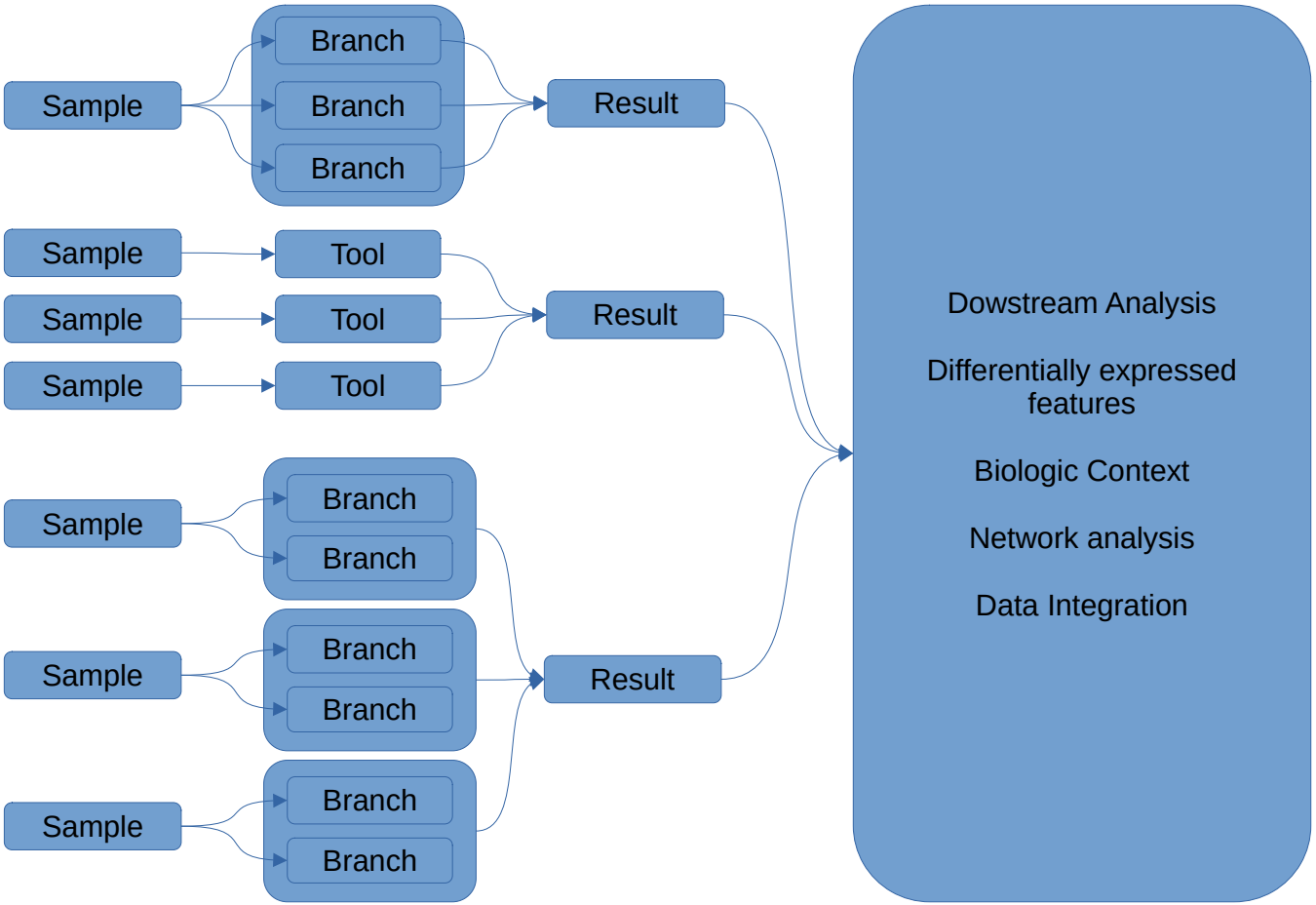
The raw data are ~ 3TB $\rightarrow$ has to be processed by a server

The processed data are ~200M $\rightarrow$ can be processed by a Desktop/Laptop

How does the amount of data change during the analysis?



Parallelization - the key concept in initial -omics data analysis



Getting R ready for parallel computing

R per se does not do parallel computing

- Install special algebraic libraries (openblas, ATLAS)
- install special packages (e.g. furrr or pdbR)
- Design of the workflows

Doing reproducible science

The goal of reproducible research is to tie specific instructions to data analysis and experimental data so that an analysis can be recreated, better understood and verified.

There are following fields:

- Literate programming
- pipeline toolkits
- package reproducibility
- project workflows
- code/data formatting tools
- format converters
- object caching.

A very good documentation can be found here:

https://cran.r-project.org/web/views/ReproducibleResearch.html

Version Control

All major workflows should be under version control using either

- git
- mercurial
- svn.

Why? Version control helps maintain reproducibility and documentation.

Task 1: Install the version control system of your choice on your laptop and link RStudio to the repository

Task 2: Download E-GEOD-51401 as ZIP from BioStudies and extract the archive

Task 3: Download pdata.xlsx from the eDesktop

Getting the data into R

How do we get the data into the machine? Getting data into R is highly data source dependent. As a general rule:

- Affymetrix chips: R-packages *affy* (older chips) or *oligo* (newer ones) $\Rightarrow$ AffyBatch or FeatureSet objects
- Agilent chips: R-package *limma*, input is the output of the program *Featureextract*
- RNASeq data: Count matrices resp. files from the alignment $\Rightarrow$ package-specific constructor.
- Proteomics data: Intensity tables for *MSStats* $\Rightarrow$ MSStatsobject.

Task 4: use the affy package to read the data from E-GEOD-51401 into R

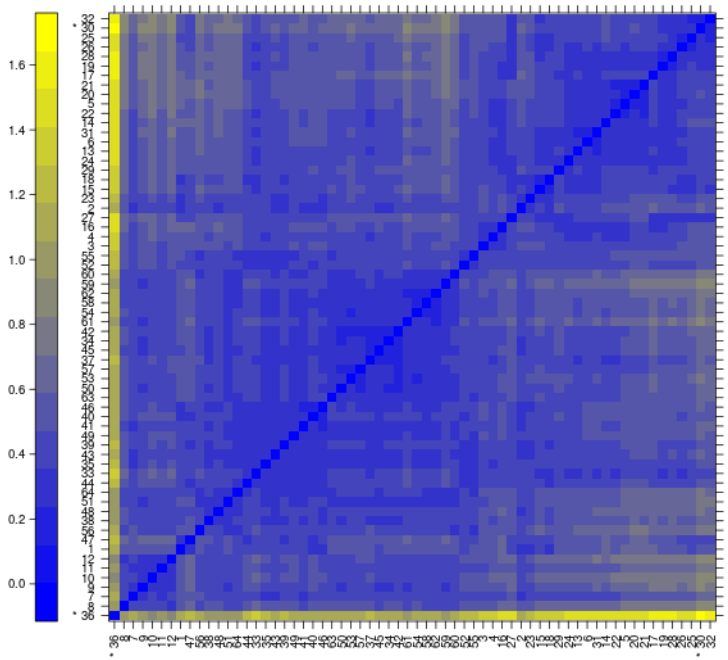
Task 5: use the import function from the rio package to import

Quality Control - underestimated but extremely important

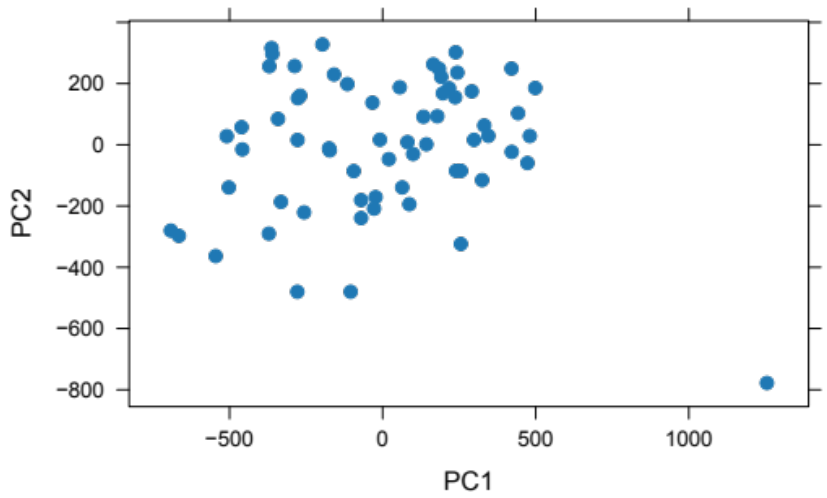
Quality control answers following questions:

- Are the data sufficiently good to warrant processing?
- Are there technical fails?
- Are there batches?
- Are there outliers?

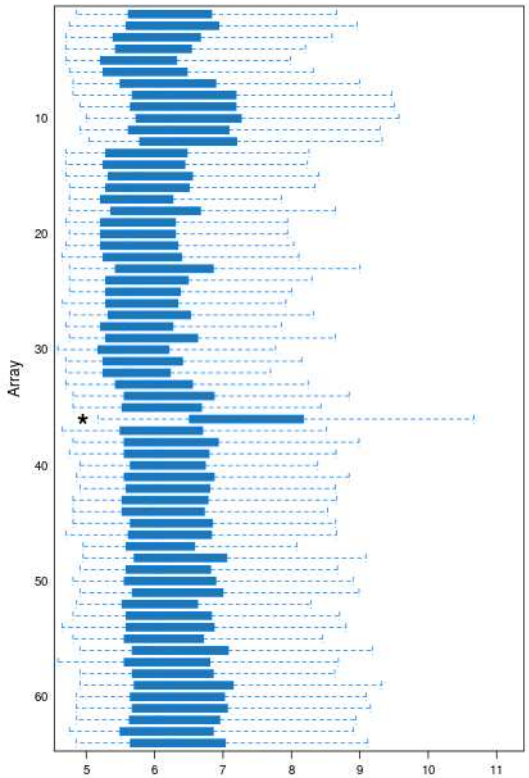
Quality Control - Distance between arrays and principal component analysis

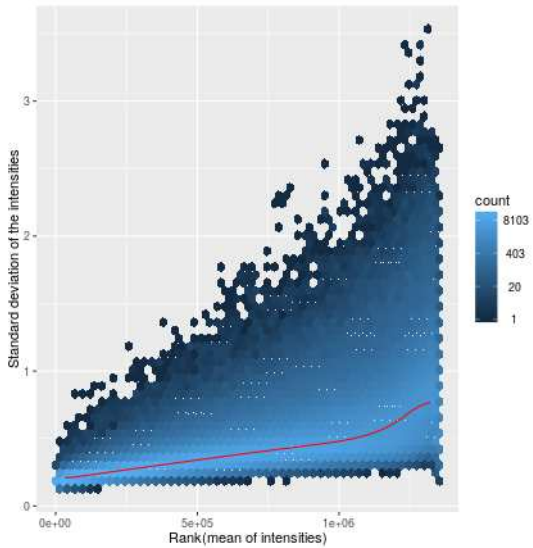


Quality Control - Distance between arrays and principal component analysis

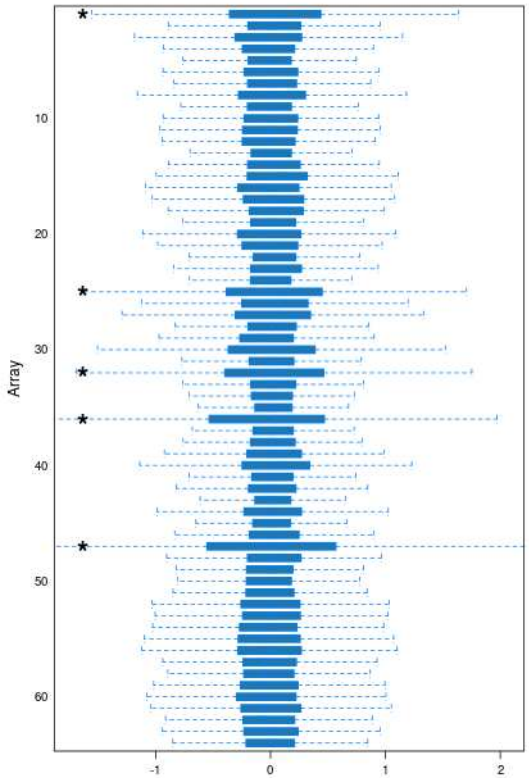


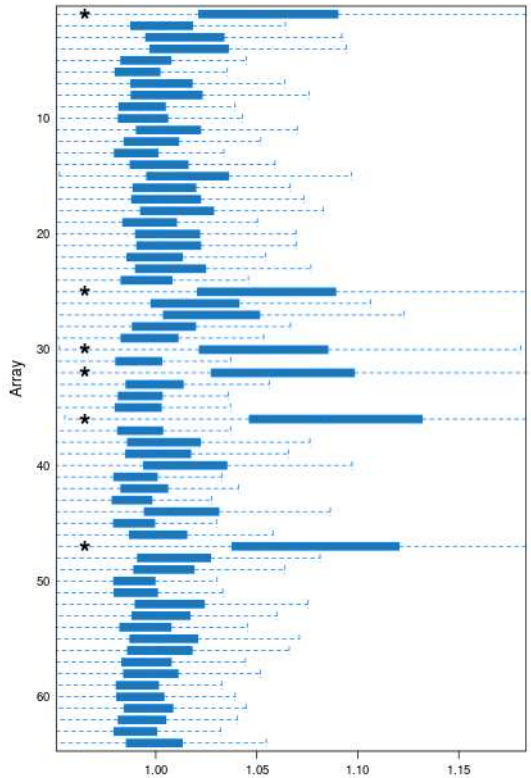
Quality Control - Distribution of the data and variance independence



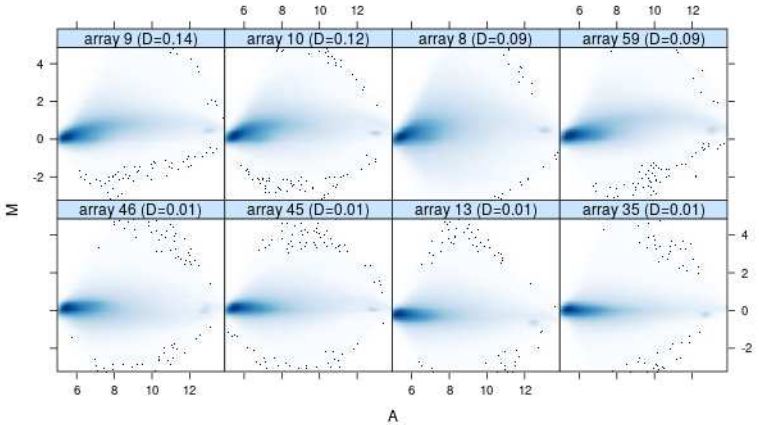


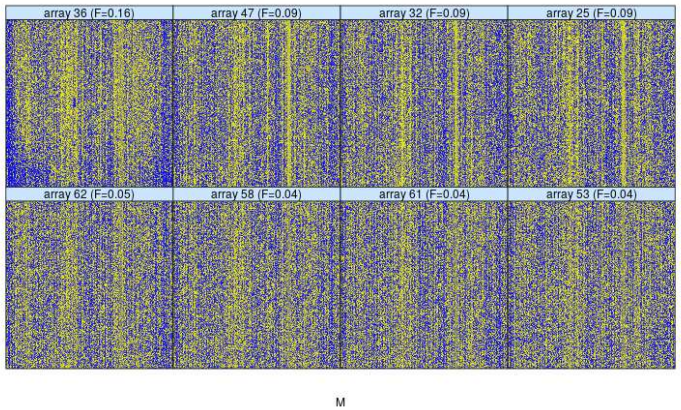
Quality Control - Relative Log Expression and Normalized Unscaled Standard Error





Quality Control - MA Plots and Spatial Distribution





Task 6: Use the arrayQualityMetrics function to conduct a quality control

Fourth lecture

Linear Models, the mainstay of initial -omics data analysis

What is a Linear Model?

- This is simply an extension of multiple regression
- Or Multiple Regression is just a simple form of the General Linear Model
- Multiple Regression only looks at ONE dependent (Y) variable
- Whereas GLM allows analysing of several dependent, Y, variables in a linear combination, i.e. multiple regression is a GLM with only one Y variable
- ANOVA, t-test, F-test, etc., are also forms of the GLM

What is a Linear Model?

Mathematically speaking . . .

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

in simple form

$$Y = X_1\beta_1 + X_2\beta_2 + \ldots + X_j\beta_k + \epsilon_1$$

How do we tell the algorithm the experimental design?

Defining the experiment design is done via the **design matrix**

Design matrices represent the linear model:

condition 1, *timepoint* 1, *replicate* 1		1	0	0	0
condition 1, *timepoint* 1, *replicate* 2		1	0	0	0
condition 2, *timepoint* 1, *replicate* 1		0	1	0	0
condition 2, *timepoint* 1, *replicate* 2	=	0	1	0	0
condition 1, *timepoint* 2, *replicate* 1		0	0	1	0
condition 1, *timepoint* 2, *replicate* 2		0	0	1	0
condition 2, *timepoint* 2, *replicate* 1		0	0	0	1
condition 2, *timepoint* 2, *replicate* 2		0	0	0	1

How are linear models written in R?

Linear models in R are defined as:

- model <- make.model(~factor1 + factor2 + . . . + factorn)
- this leads to a model matrix where the first level of factor1 is the control to which all other levels of factor1, factor2, etc. are compared.

The model matrix might be a bit counterintuitive.

How are linear models written in R?

A more intuitive approach, especially when working with contrasts is following:

factor <- paste(factor1, factor2, . . . , factorn, sep = “_“) %>%
factor(levels = c(”level1”, “level2”, etc))

The model is then defined as:

model <- make.model(~0+factor)

A trick to make things more human readable

To make things easier for contrasting, we set the names of the columns to the levels of the factor:

colnames(model) <- levels(factor)

For our matrix above we would have:

factor <- paste(condition, timepoint, sep = “_“) %>% (factor, levels = c(”condition1_timepoint1”, “condition2_timepoint2”, <etc>.))

of course, one has to replace <condition1> with the name, for instance “control”, and <timepoint1> with for example “0hrs”, resulting in “control_0hrs”.

How do we tell the algortihm which conditions to compare - the concept of contrasts.

Comparisons are defined via a **contrast matrix**

The contrast matrix represents the contrasts one wants to examine. A contrast is written as $k_1 \cdot condition_1 + k_2 \cdot condition_2$ with $k_1 = 1$ and $k_2 = -1$

Contrasts have to be **orthogonal**. That means the sum of the coefficients has to be zero. If that is not the case -> something is wrong.

What to do in case of unbalanced designs - weighting.

Unbalanced designs with combinations of conditions require **weighted** contrast matrices

- for example, condition 2 has three samples, and condition 3 has six samples
- we compare condition 1 versus condition 2 and 3,
- the contrast looks like this:
  $condition_1 - (\frac{6}{9} condtion_2 + \frac{3}{9} condition_3)$
- and the column entries in the contrast matrix would be:
  $1, -\frac{6}{9}, -\frac{3}{9}$

How are contrasts represented?

Contrasts are represented in a so-called **contrasts matrices**:

$$
\begin{array}{l}
\textit{condition}\ 1,\ \textit{timepoint}\ 1, \\
\textit{condition}\ 2,\ \textit{timepoint}\ 1, \\
\textit{condition}\ 1,\ \textit{timepoint}\ 2, \\
\textit{condition}\ 2,\ \textit{timepoint}\ 2,
\end{array}
=
\begin{bmatrix}
-1 & 0 & -1 & 0 \\
1 & 0 & 0 & -1 \\
0 & -1 & 1 & 0 \\
0 & 1 & 0 & 1
\end{bmatrix}
$$

What are the results of such an analysis?

The results are typically presented as:

- A feature annotation to allow the translation of featureIDs into something people can work with
- A log fold change
- A standard error
- A p-value
- An adjusted p-value
- A log odds ratio

What is the log fold change?

- Most bioinformatics tasks encompass the detection of changes. In -omics this is typically the change in the analyte.
- The log fold change is calculated as $log_2 \frac{condition_1}{condition_2} = log_2(condition_1) - log_2(condition_2)$

Why the base 2?

chosing a log base of 2 allows to immediately see how much the expression changes:

- $condition_1 = 2 \cdot condition_2 \rightarrow log_2 \frac{2 \cdot condition_2}{condition_2} \rightarrow log_2 \frac{2}{1} \rightarrow log_2 2 = 1$
- $condition_1 = \frac{1}{2} \cdot condition_2 \rightarrow log_2 \frac{condition_2}{2 \cdot condition_2} \rightarrow log_2 \frac{1}{2} \rightarrow log_2 0.5 = -1$

If $log_2 FC = 1, 2, 3 \Rightarrow$ the abundance of the analyte is doubled, fourfold, eightfold.

If $log_2 FC = -1, -2, -3 \Rightarrow$ the abundance of the analyte is halved, one fourth, one eighth.

Why do we adjust the p-value?

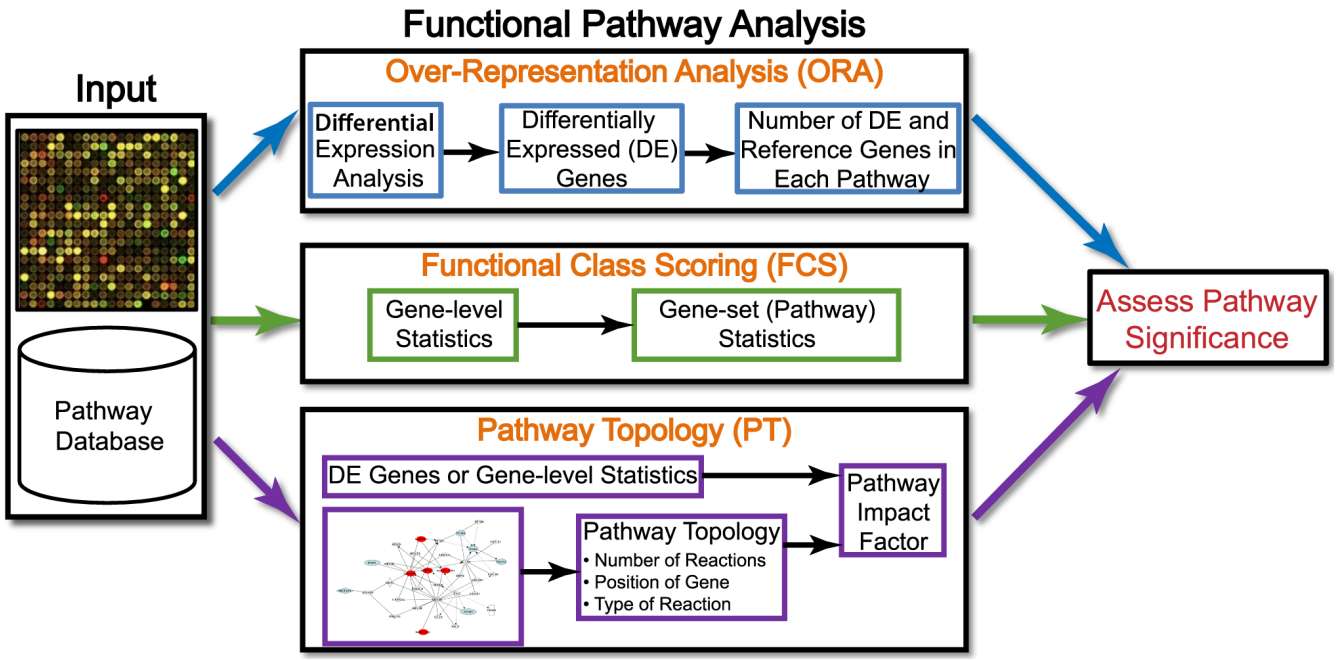
- In -omics data, we do not only perform one test, we perfom n tests.
- We are *not* interested in either one of the tests, but in both combined, therefore p-values are combined
- Probability combination is defined as:
  $Pr(test_1 \cup test_2 \cup .... \cup test_n) = 1 - \prod_{i=1}^{n}(1 - Pr(test_i)$
- The combined p-value rises with the number of tests.
- To keep the overall p-value below our threshold, p-values have to be adjusted.

P-value adjustment methods

- Family Wise Error Rate (FWER)
  - Gives the probability that at least one discovery among our discoveries is a false positive
  - Holm - Sidak
  - Bonferroni
- False Discovery Rate (FDR)
  - Gives the rate of false positive amoung our discoveries
  - Benjamini-Hochberg
  - Benjamini-Yekuteli
  - Storey's q

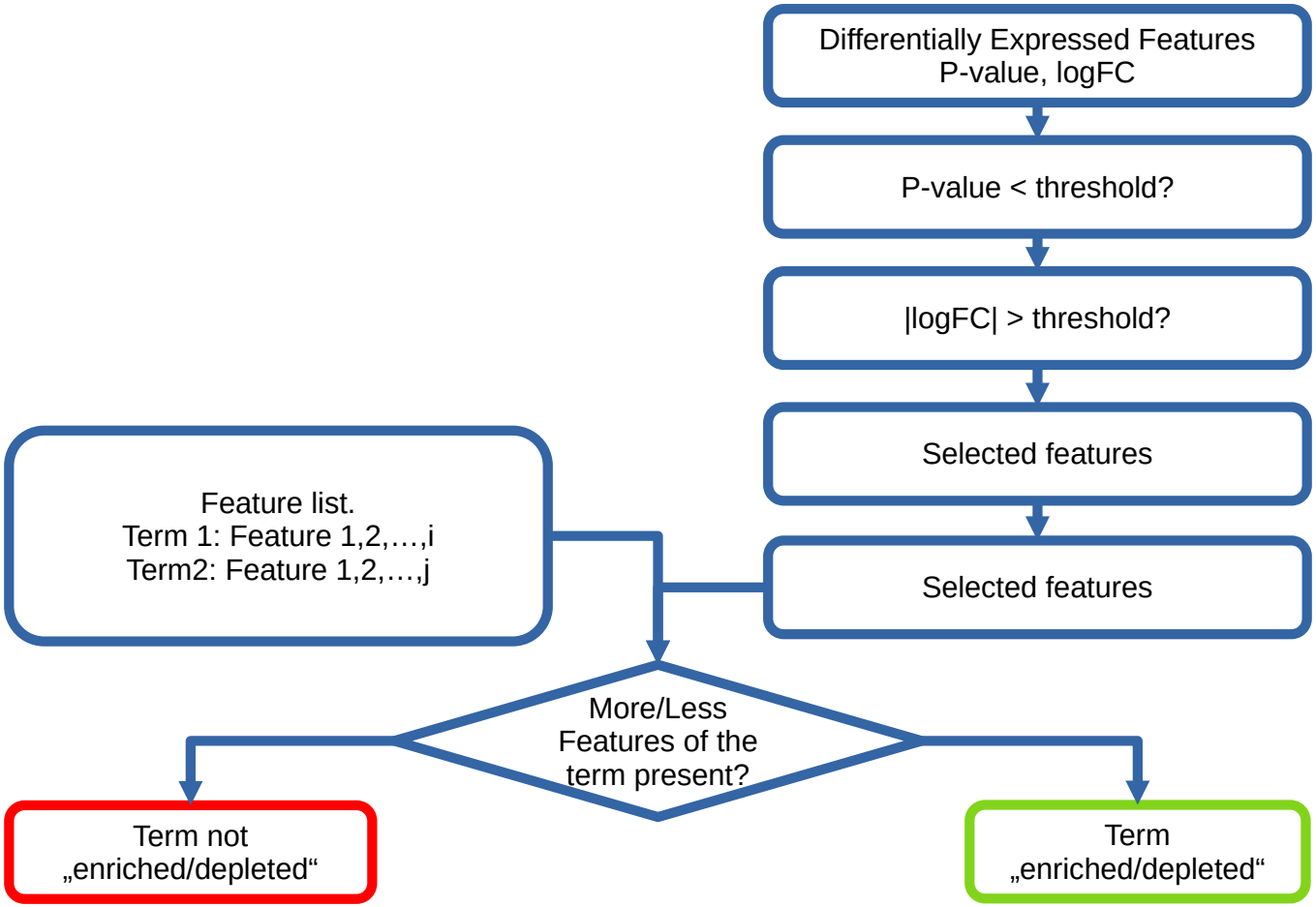
Putting things in perspective - Biologic context analysis

Generations of analysis methods

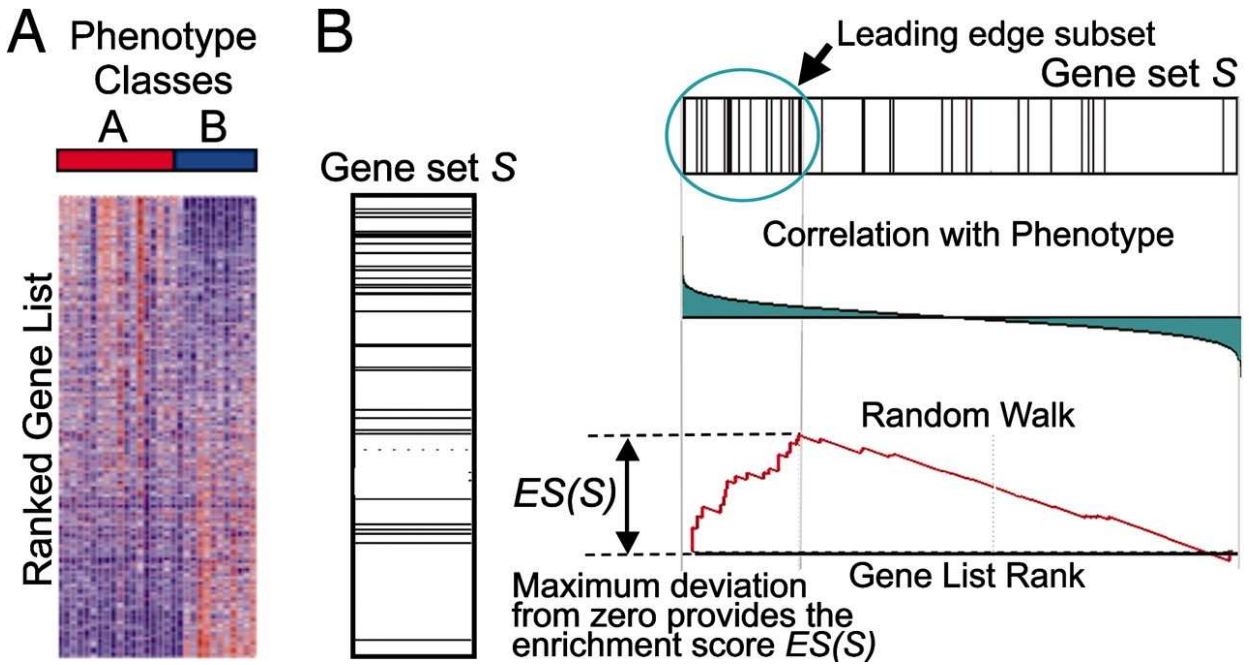


Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 8(2): e1002375. https://doi.org/10.1371/journal.pcbi.1002375

Generation one: Guilt by association - Term Enrichment analyses

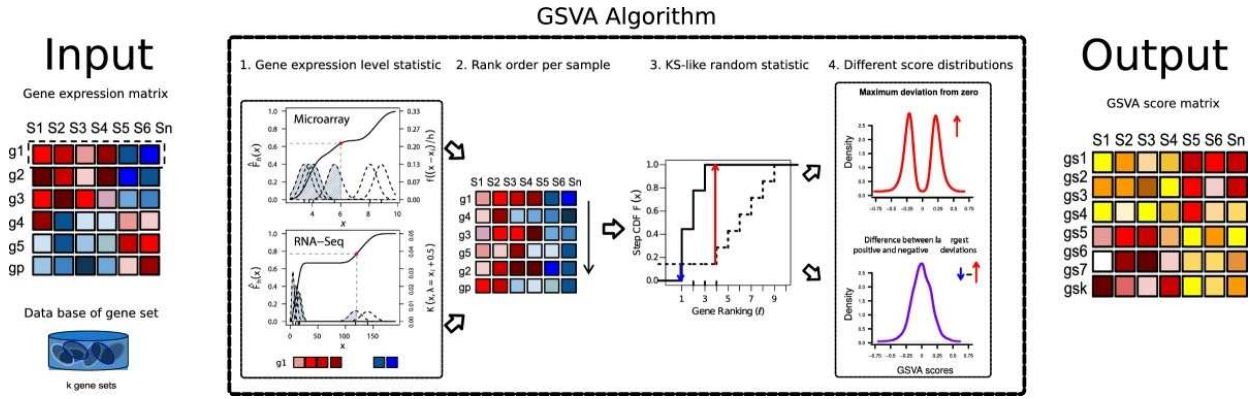


Generation two: Guilt by accumulation - Set Enrichment/Variation Analyses



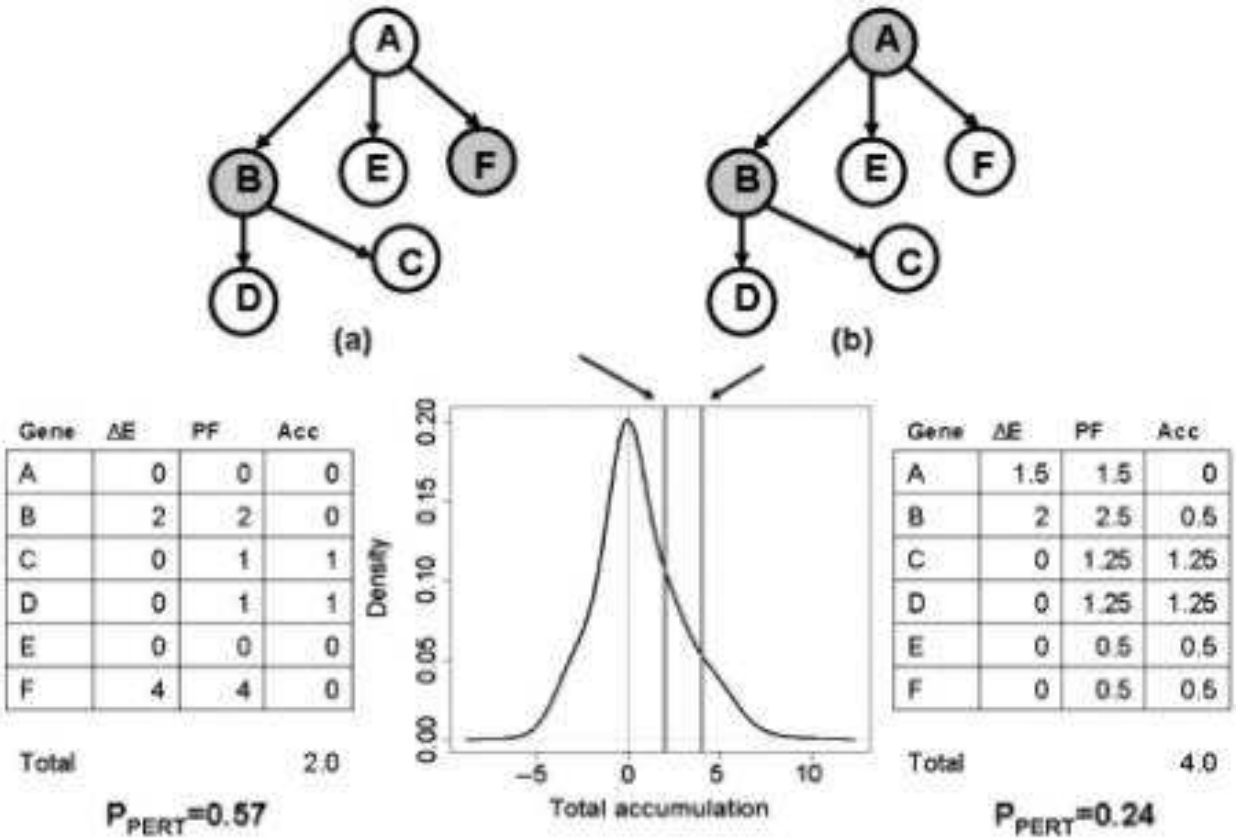
Subramanian A, Tamayo P, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50.

Generation two: Guilt by accumulation - Set Enrichment/Variation Analyses



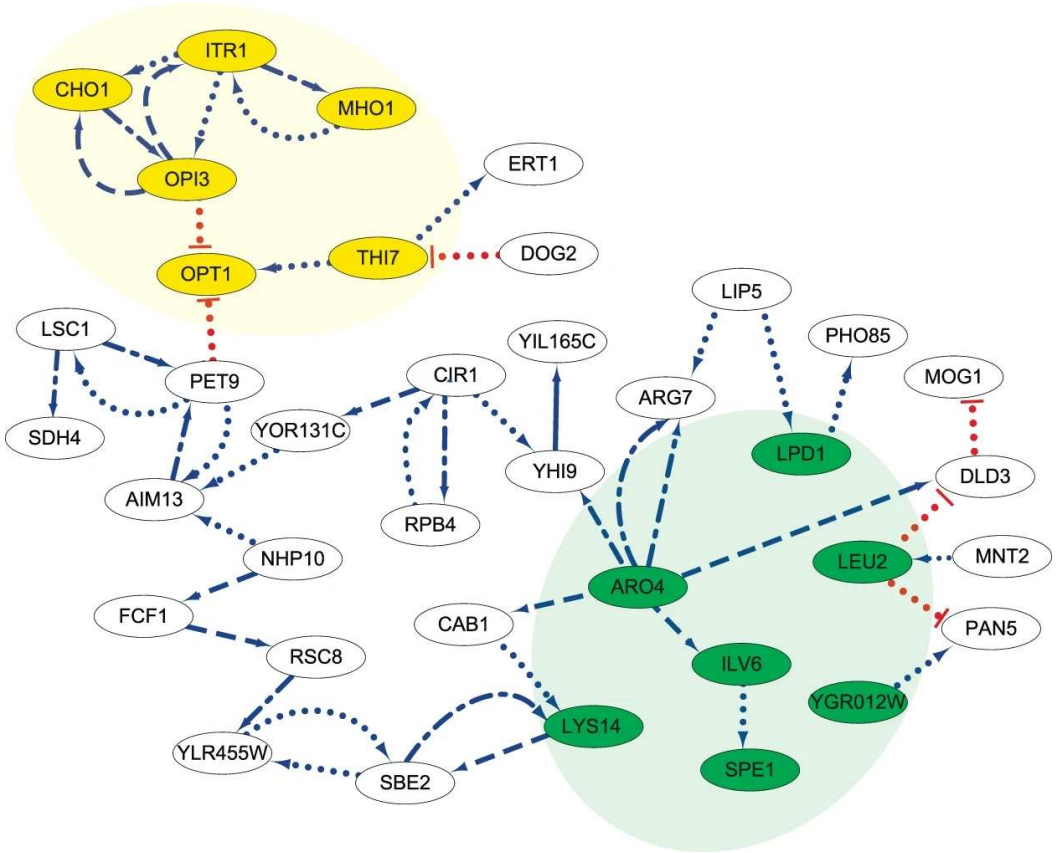
Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 14, 7 (2013). https://doi.org/10.1186/1471-2105-14-7

Generation three: taking into account pathway topology - Signalling Pathway Impact Analyses



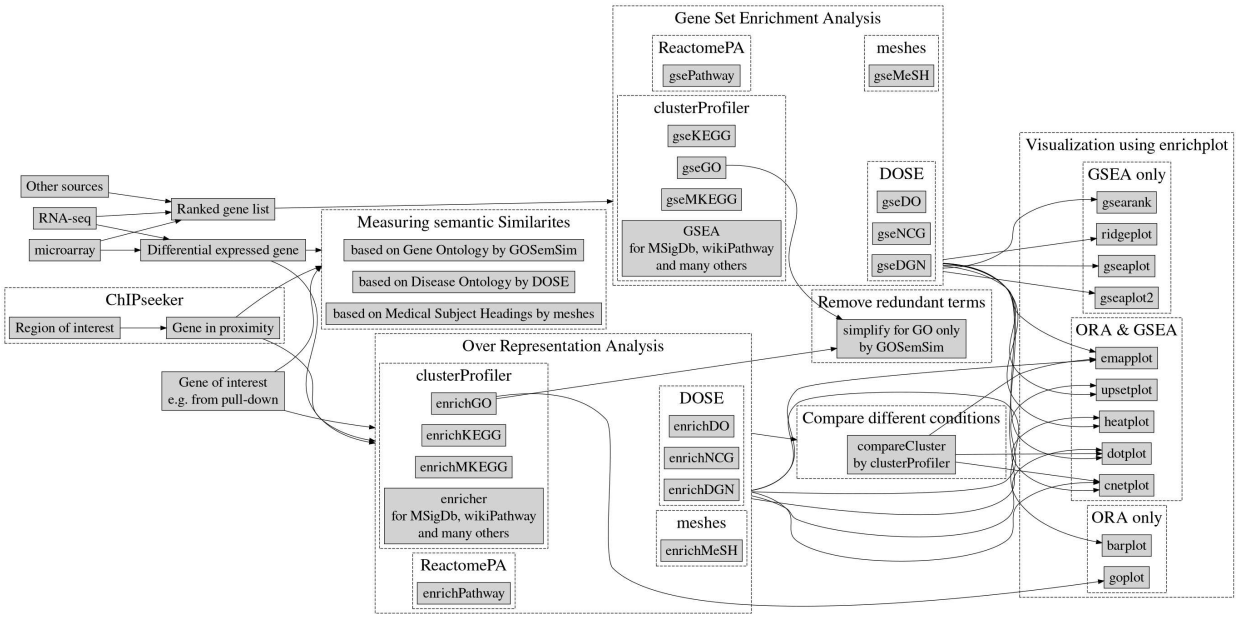
Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis. Bioinformatics. 2009 Jan 1;25(1):75-82. doi: 10.1093/bioinformatics/btn577. Epub 2008 Nov 5. PMID: 18990722; PMCID: PMC2732297.

Generation four: modeling reality - Network based analyses



Chen, C., Zhang, D., Hazbun, T.R. et al. Inferring Gene Regulatory Networks from a Population of Yeast Segregants. Sci Rep 9, 1197 (2019). https://doi.org/10.1038/s41598-018-37667-4

clusterProfiler - a versatile tool to put data into biologic context.



G Yu, LG Wang, Y Han, QY He. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology 2012, 16(5):284-287.

Tasks for you to accomplish

- Task 1: Install clusterProfiler and GSVA packages
- Task 2: Carry out a Term Enrichment Analysis and a Gene Set Variation Analysis on your data
- Task 3: Look at the results, and try to deduce what is going on.

Fifth lecture

Question and Answers, Coding