Genome and transcriptome assembly: state of the art and best practices

30.3.2021

Monika Čechová

@biomonika
biomonika.org

Presentations

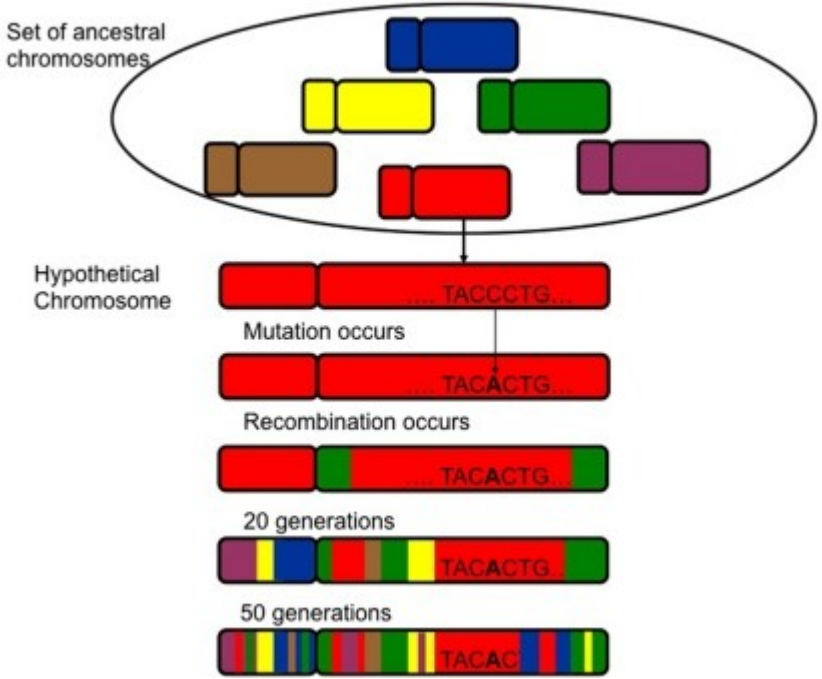
Variant calling

Variant detection

Variants in the Illumina.bam file:

- SNP 1bp
- ins 1bp
- del 1bp
- ins 50bp
- del 50bp
- ins 250bp
- translocation 1000bp

Linkage disequilibrium



In population genetics, **linkage disequilibrium** (**LD**) is the non-random association of alleles at different loci in a given population. Loci are said to be in linkage disequilibrium when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124487/

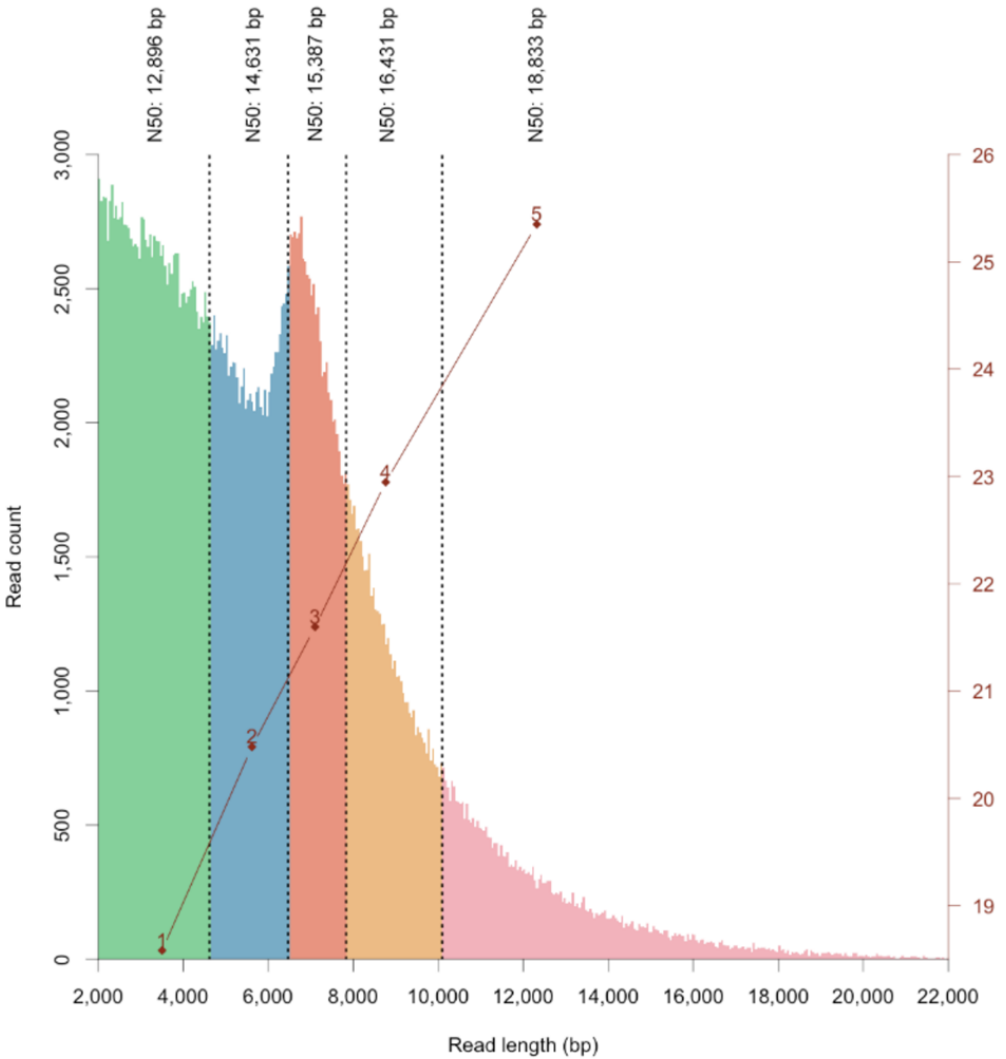
https://en.wikipedia.org/wiki/Linkage_disequilibrium

Figure from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3098715/

Long reads

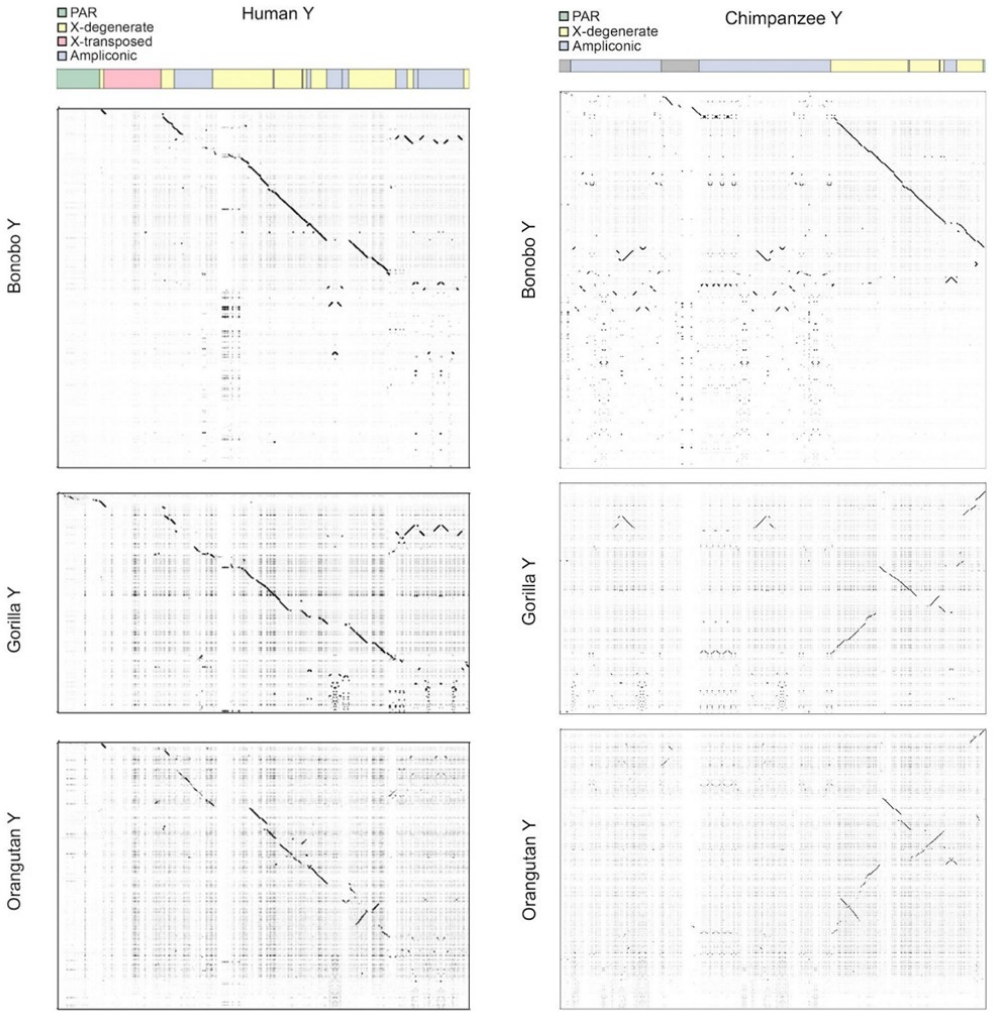
Lessons learned

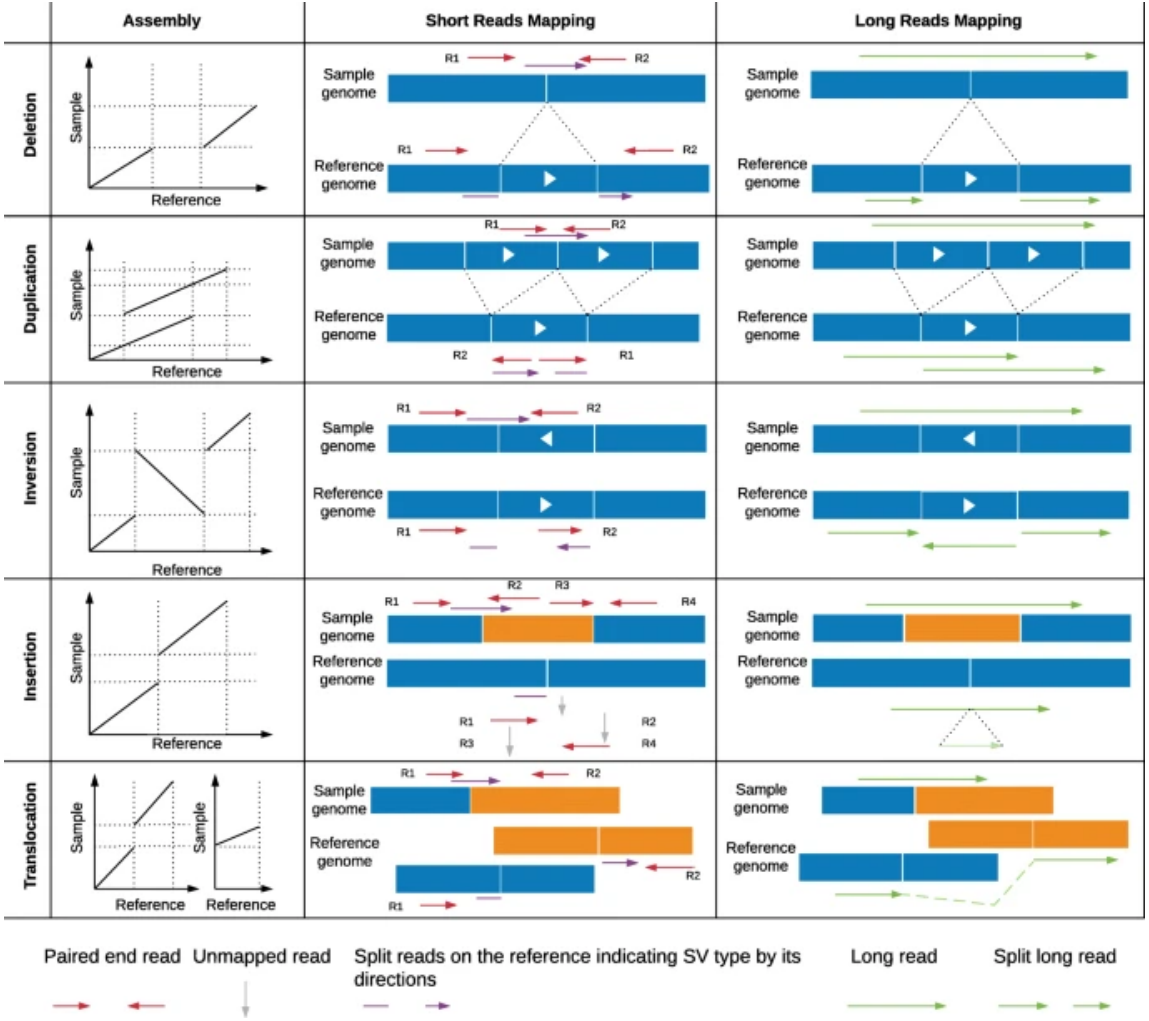
- Long reads do the heavy-lifting
- Combining orthogonal sequencing technologies is essential
- Evaluation strategy is important
  - Total length, Ns, number of scaffolds, N50 (quast)
  - Gene content
  - Transposable element content
  - Dotplot analysis in relation to a closely related species
- Trial and error is a necessary strategy for a fast-paced field



Tomaszkiewicz, Rangavittal, Cechova et al., 2016

Species	Assembly length (Mb)	NG50 (in bp, using G=8.5 Mb)	Number of scaffolds	Ns (in Mb)
Orangutan	17.4	1,388,499	1,178	1.3
Gorilla	14.3	150,017	268	0.008
Bonobo	23.4	153,556	3,590	0.8
Chimpanzee	26.4	-	-	1.1
Human	57.2	-	-	33.6





https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1828-7

Assembly strategies

Trio binning

Homozygous versus heterozygous genome assembly.

De novo assembly of haplotype-resolved genomes with trio binning

Sergey Koren, Arang Rhie, Brian P Walenz, Alexander T Dilthey, Derek M Bickhart, Sarah B Kingan, Stefan Hiendleder, John L Williams, Timothy P L Smith ✉ & Adam M Phillippy ✉

Nature Biotechnology **36**, 1174–1182(2018) | Cite this article

This will be followed by in-class discussion.

Trio binning

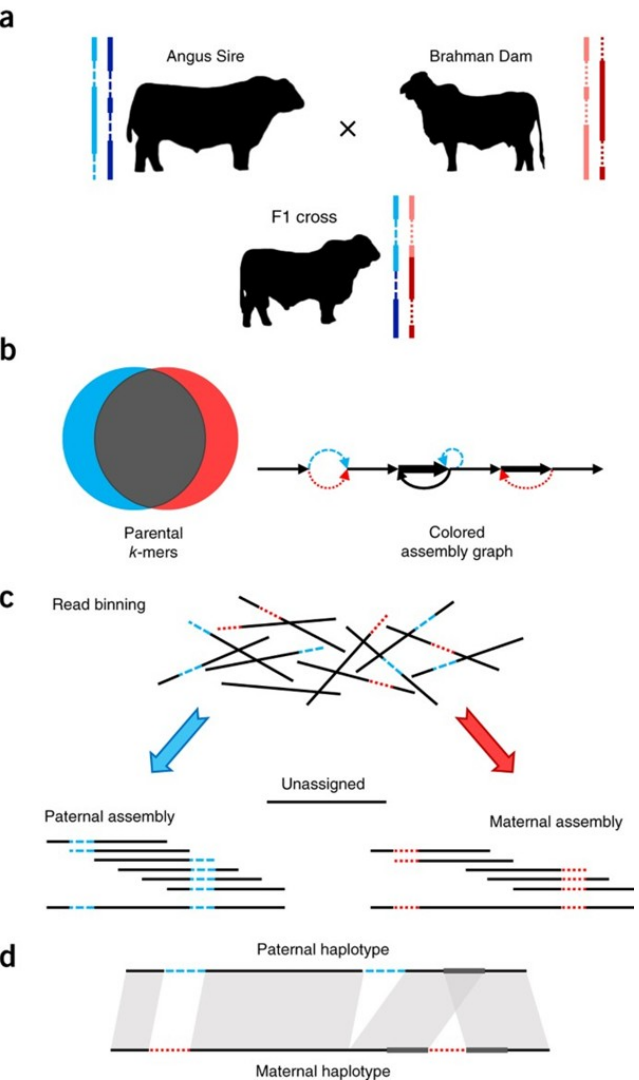
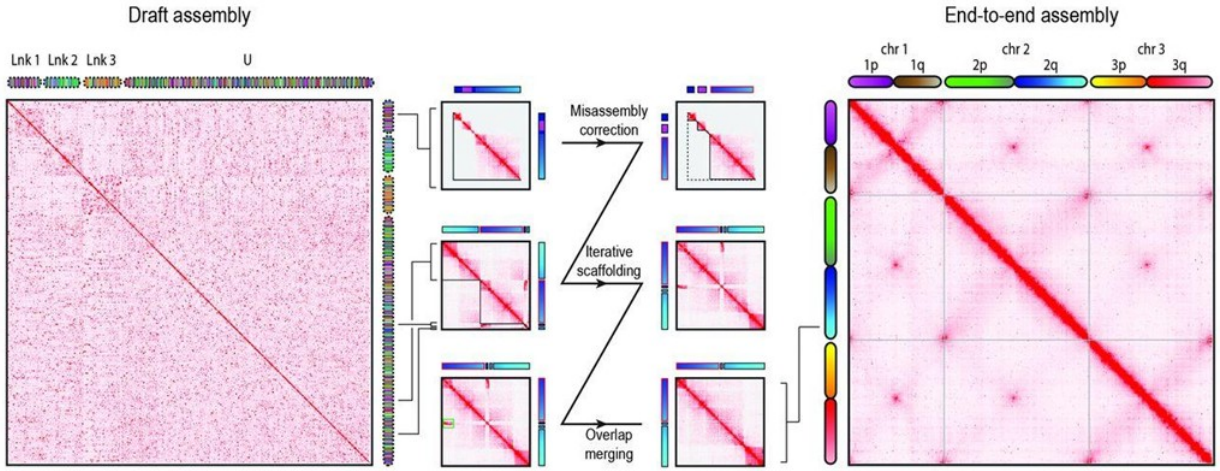


Fig. 1 Starting with a draft assembly, we used Hi-C data to correct misjoins, scaffold, and merge overlaps, thereby generating an assembly of the Ae. aegypti mosquito genome with chromosome-length scaffolds.



Olga Dudchenko et al. Science 2017;356:92-95

https://www.dnazoo.org/assemblies

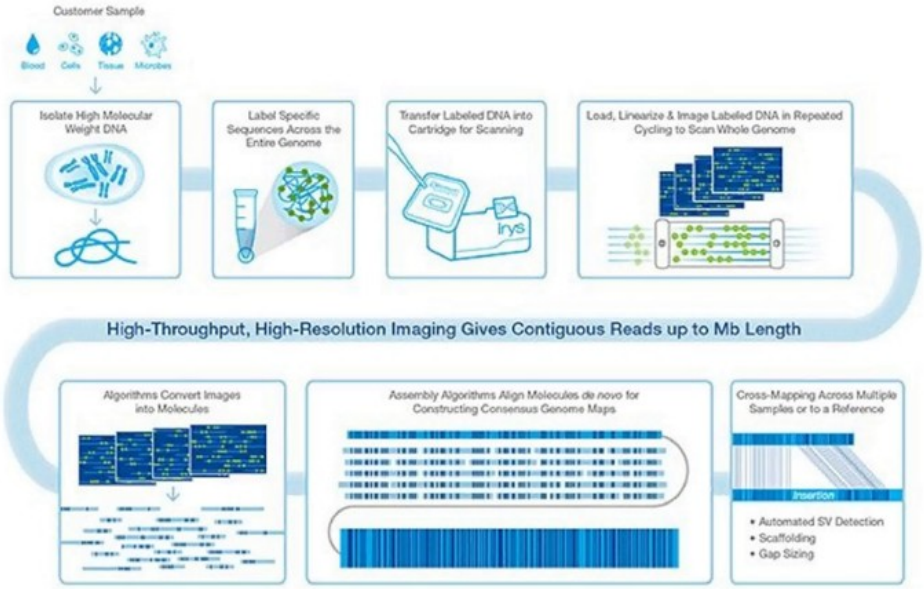
Science AAAS

Bionano Genomics

BioNano Genomics Irys Workflow

High-throughput workflow gives:

- Automated SV Detection
- Scaffolding
- Gap Sizing



The use of optical maps for scaffolding and large-scale structural variants.

T2T assembly

Read methods of the following paper to in order to understand the methods authors used for the successful assembly of the human chromosome 8.

Discuss these methods in pairs (10 minutes).

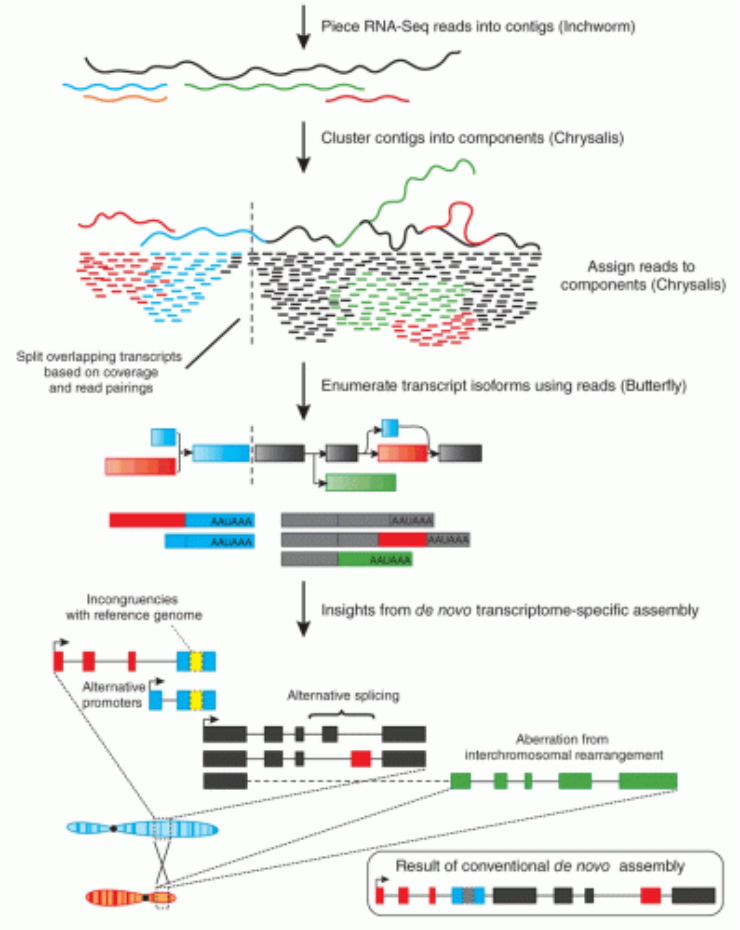
The breakout session will be followed by an in-class discussion (10 minutes).

The structure, function, and evolution of a complete human chromosome 8

Glennis A. Logsdon, Mitchell R. Vollger, PingHsun Hsieh, Yafei Mao, Mikhail A. Liskovykh, Sergey Koren, Sergey Nurk, Ludovica Mercuri, Philip C. Dishuck, Arang Rhie, Leonardo G. de Lima, David Porubsky, Andrey V. Bzikadze, Milinn Kremitzki, Tina A. Graves-Lindsay, Chirag Jain, Kendra Hoekzema, Shwetha C. Murali, Katherine M. Munson, Carl Baker, Melanie Sorensen, Alexandra M. Lewis, Urvashi Surti, Jennifer L. Gerton, Vladimir Larionov, Mario Ventura, Karen H. Miga, Adam M. Phillippy, Evan E. Eichler

doi: https://doi.org/10.1101/2020.09.08.285395

Transcriptome assembly



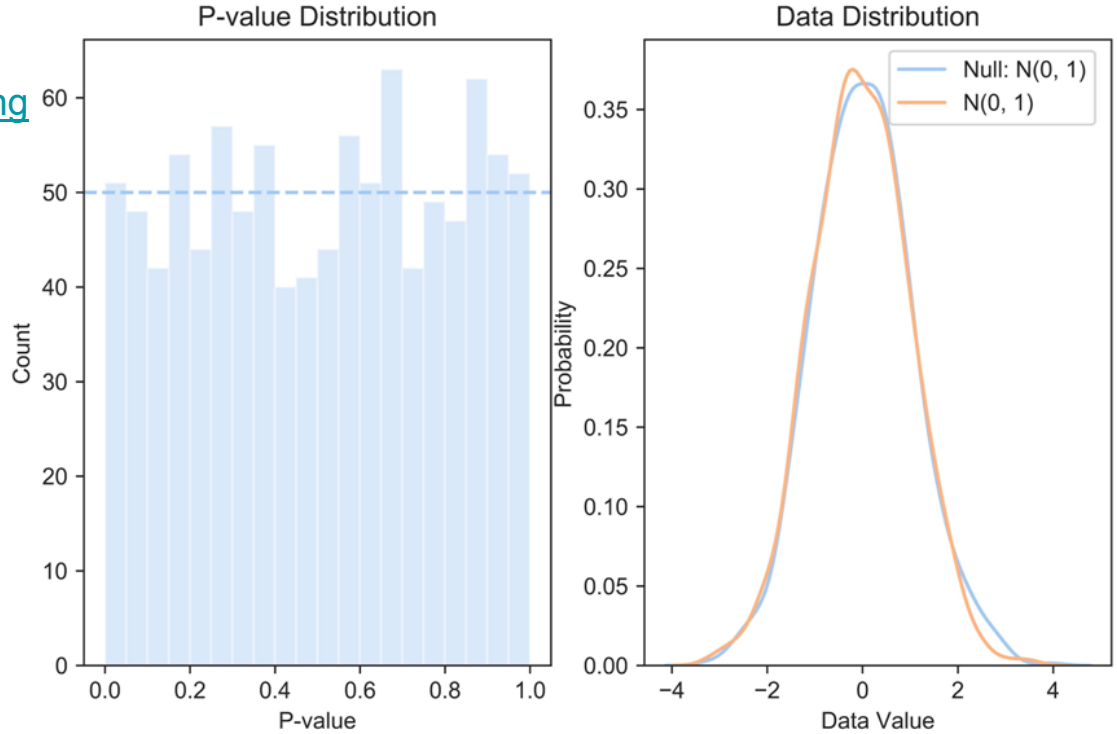
https://evomics.org/learning/genomics/trinity/

How to interpret a p-value histogram

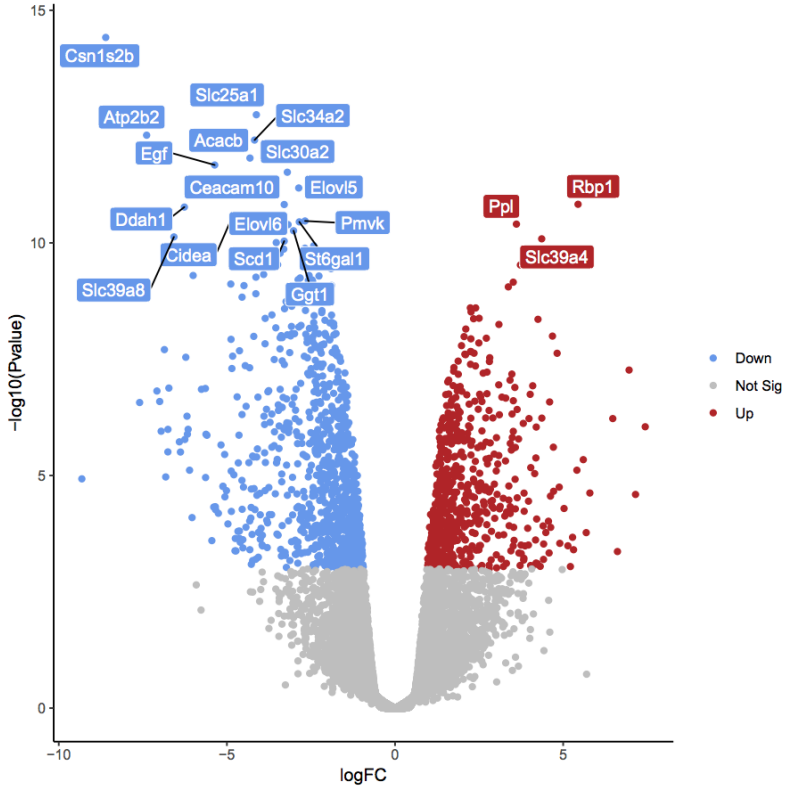
http://varianceexplained.org/statistics/interpreting-pvalue-histogram/

https://towardsdatascience.com/how-to-test-your-hypothesis-using-p-value-uniformity-test-e3a43fc9d1b6

https://imgs.xkcd.com/comics/significant.png



Differential gene expression



https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-volcanoplot/tutorial.html

Galaxy

Galaxy is an open source, web-based platform for data intensive biomedical research.

Galaxy

https://usegalaxy.org/ or https://usegalaxy.eu/

Tutorials:

https://training.galaxyproject.org/

Czech community:

https://lists.galaxyproject.org/lists/galaxy-czech.lists.galaxyproject.org/

Homework (6 points)

Remember there is no “one right way” to do an analysis. Choose parameters that you think are the most suitable for your goal.

- Create an account at https://usegalaxy.eu/
- Load following fastq files as a Collection (List of Pairs):

  https://zenodo.org/record/3541678/files/A1_left.fq.gz
  https://zenodo.org/record/3541678/files/A1_right.fq.gz
  https://zenodo.org/record/3541678/files/A2_left.fq.gz
  https://zenodo.org/record/3541678/files/A2_right.fq.gz
  https://zenodo.org/record/3541678/files/A3_left.fq.gz
  https://zenodo.org/record/3541678/files/A3_right.fq.gz
  https://zenodo.org/record/3541678/files/B1_left.fq.gz
  https://zenodo.org/record/3541678/files/B1_right.fq.gz
  https://zenodo.org/record/3541678/files/B2_left.fq.gz
  https://zenodo.org/record/3541678/files/B2_right.fq.gz
  https://zenodo.org/record/3541678/files/B3_left.fq.gz
  https://zenodo.org/record/3541678/files/B3_right.fq.gz

- Run FastQC before and after trimming reads with Trimmomatic. Trim for quality and consider whether the adaptor removal should be performed.

Homework (6 points)

Remember there is no “one right way” to do an analysis. Choose parameters that you think are the most suitable for your goal.

- Assemble the trimmed reads with Trinity. Trinity will output both gene and isoform files. Focus on the isoforms.
- Align trimmed reads to this de-novo reference assembly and estimate read abundance per isoform (Align reads and estimate abundance on a de novo assembly of RNA-Seq data). Use salmon as **Abundance estimation method.**
- Rename the datasets: A1_raw, A2_raw, A3_raw, B1_raw, B2_raw, B3_raw
- Build expression matrix for your de novo assembly of RNA-Seq data by Trinity (this is the first step in the differential gene expression pipeline)
- Share your history with the user cechova.biomonika@gmail.com
- Export your history to a file and upload your .tar.gz to the Odevzdávarna by April 13th, 2021

This exercise is inspired by the following draft tutorial:

> https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/full-de-novo/tutorial.html#merge-the-mapping-tables-and-compute-a-tmm-normalization