Introduction to sequencing data analysis

Markus Wolfien, Andrea Bagnacani, and Olaf Wolkenhauer

de.NBI Training – 6th March 2019 Rostock

www.sbi.uni-rostock.de

EUROPÄISCHE UNION
Europäischer Sozialfonds

Europäische Fonds EFRE, ESF und ELER
in Mecklenburg-Vorpommern 2014-2020

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

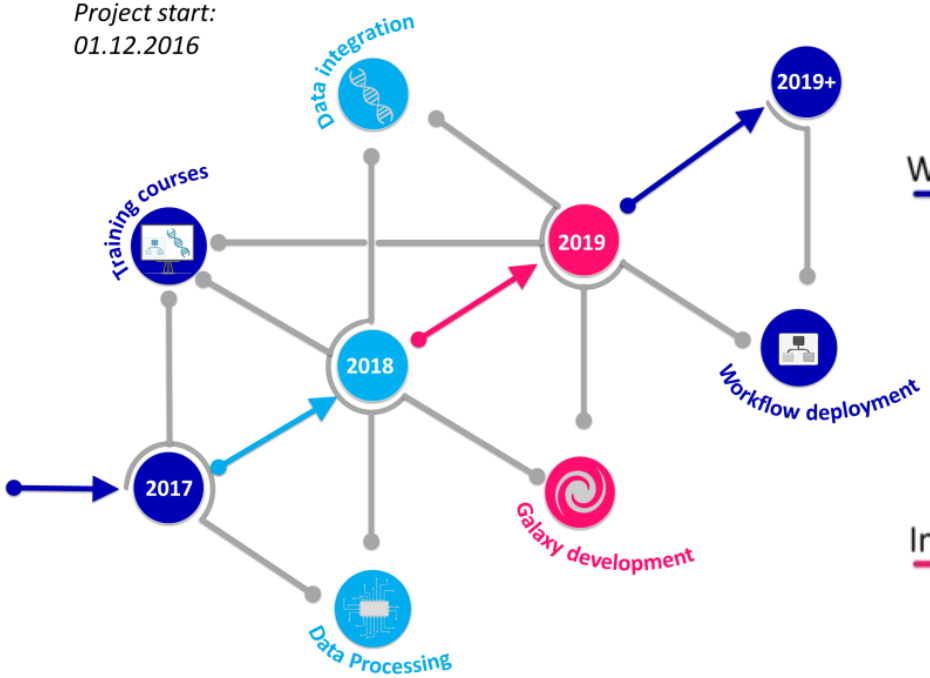
sbi.uni-rostock.de/destair

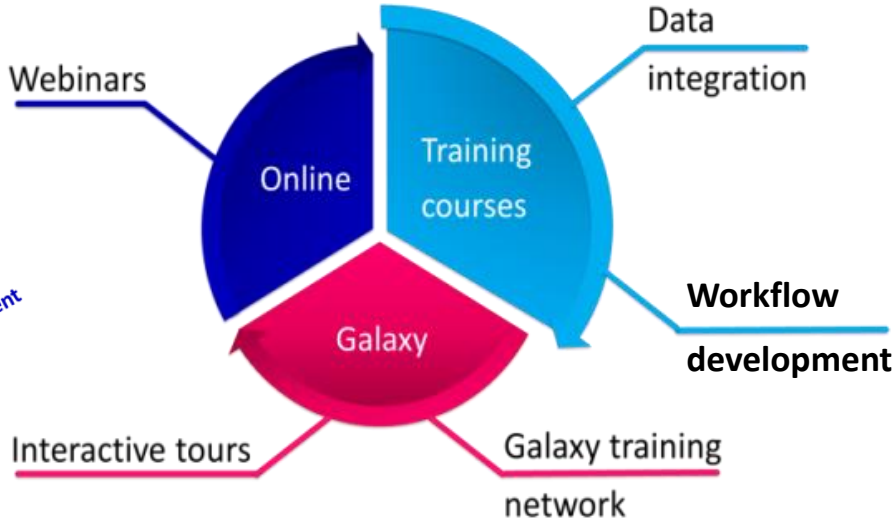
Structured Analysis and Integration of RNA-Seq experiments (de.STAIR)

Our aim is to enable a comprehensive **analysis of RNA-Seq experiments as a service.** To enable maximum usefulness, interconnectivity, and accessibility for the developed approaches and services, we will provide dedicated **workshops**, **training programs and screen casts** for bioinformaticians and other life scientists.

de.STAIR – RNA-Seq analysis and integration

Project start:
01.12.2016





GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

Objectives for the training today

- What is medical Big data?
- Why using NGS?
- The Galaxy around me

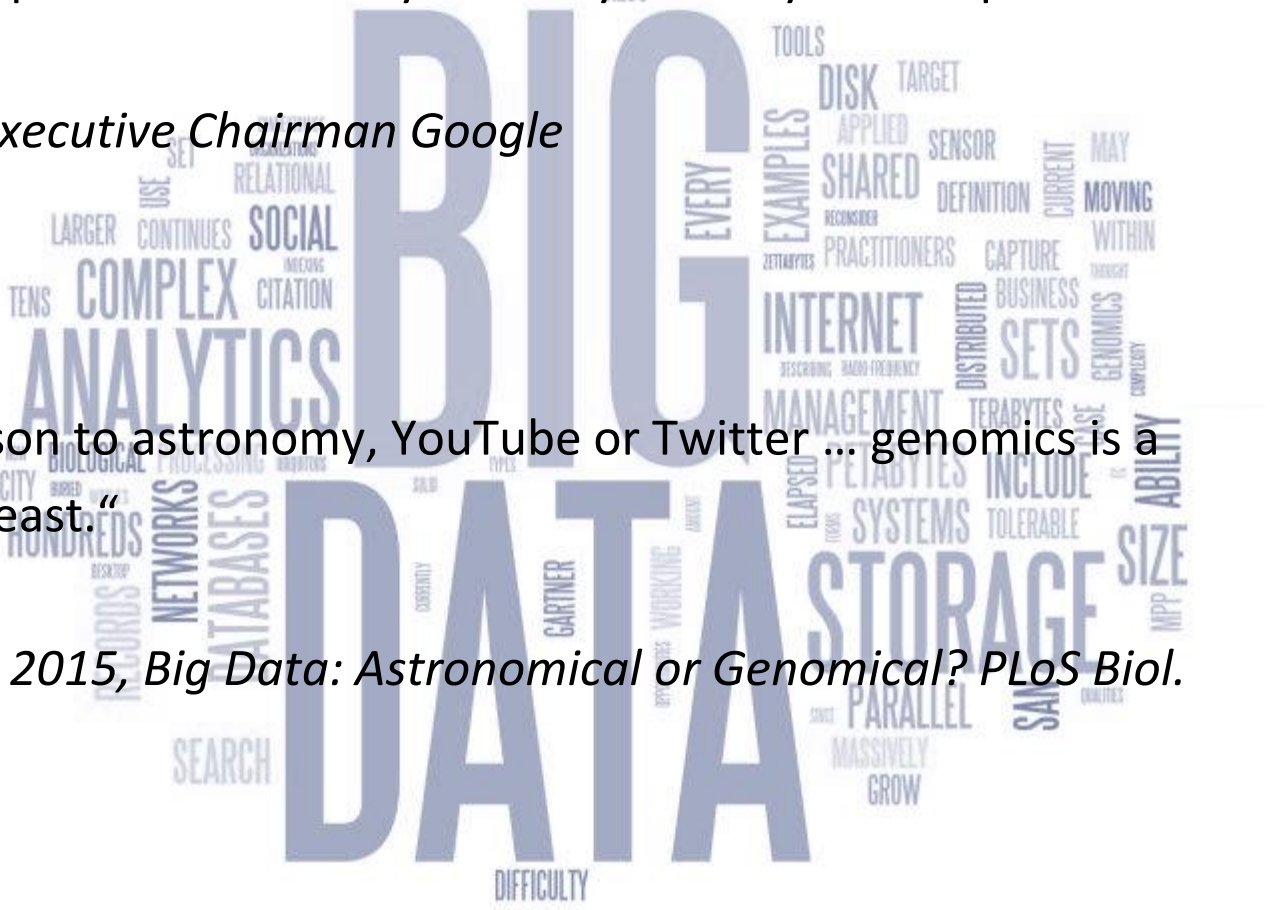
- How do we analyze NGS data?
- Are there best practices?
- Are there automation strategies?

- Explore RNA-Seq data analysis on use cases

What is big data?

"From the dawn of civilization until 2003, humankind generated five exabytes of data. NOW we produce five exabytes every two days … and pace is accelarating."

Eric Schmidt, Executive Chairman Google



"… In comparison to astronomy, YouTube or Twitter … genomics is a four-headed beast."

Stephens et al. 2015, Big Data: Astronomical or Genomical? PLoS Biol.

Transcription from genes to a functional gene products

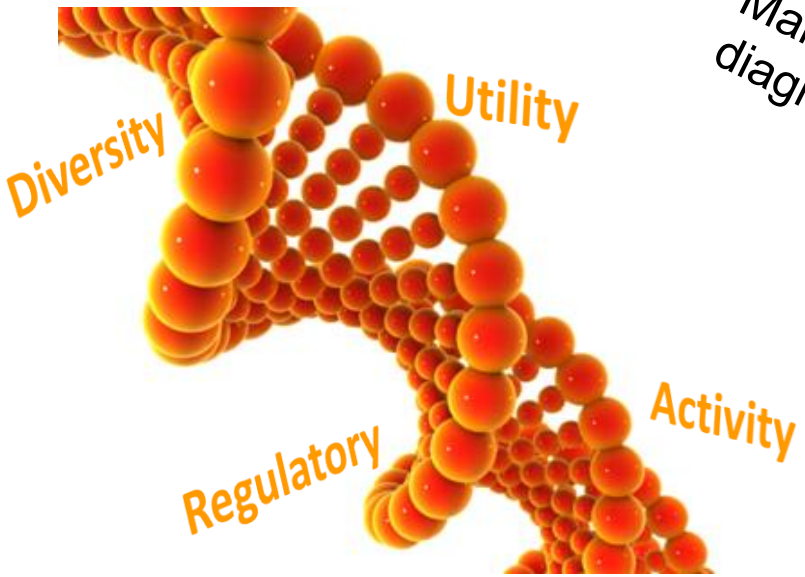
Many different variations and subtypes

Information about regulatory mechanisms

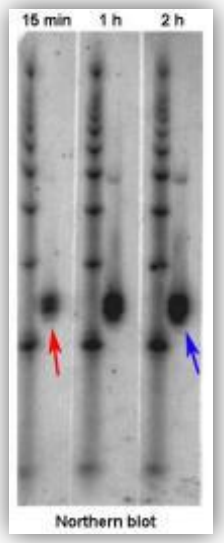
Active and measurable state of the cell …

… ,but only a snapshot

Many different therapeutical and diagnostical approaches

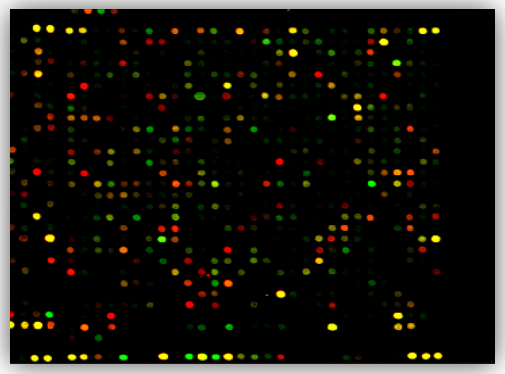


Measureing gene expression

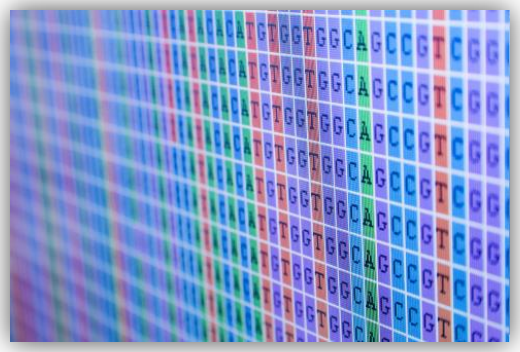


Northern Blot

Reverse Transcription PCR

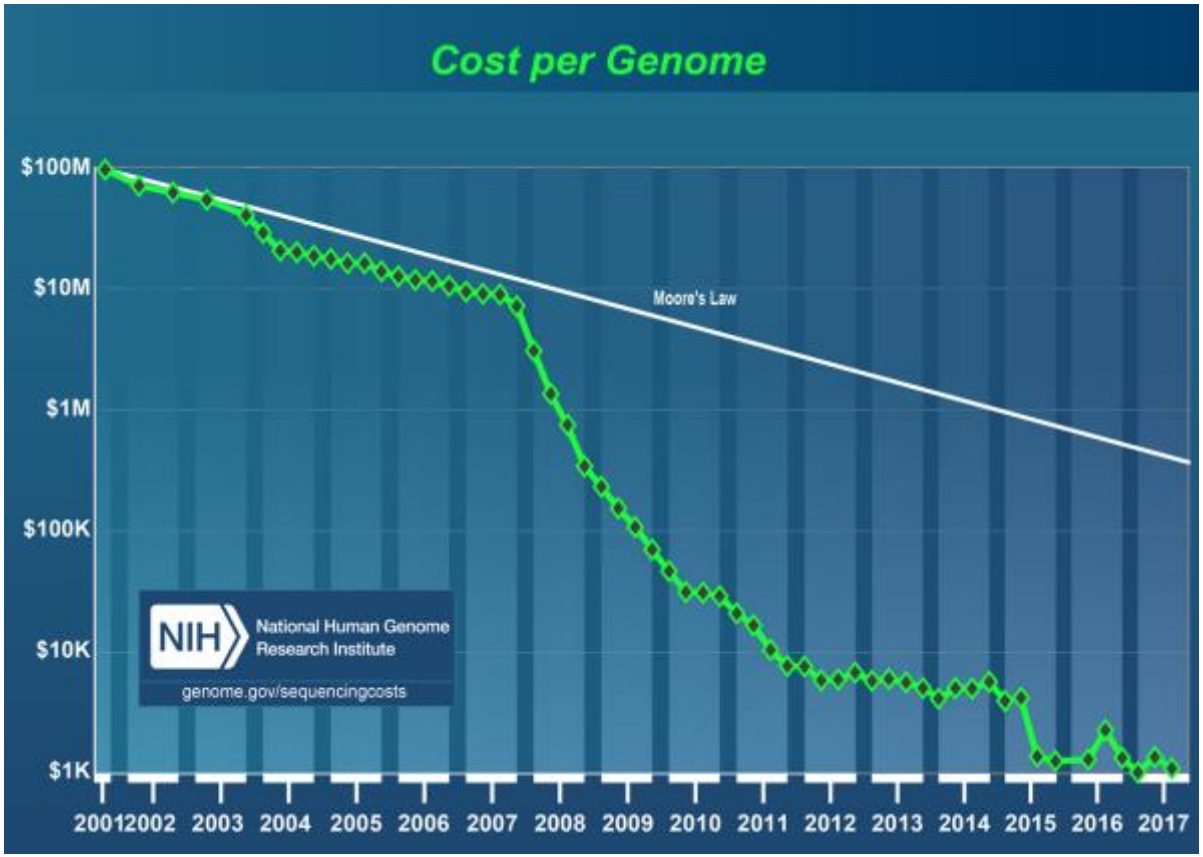


Microarrays

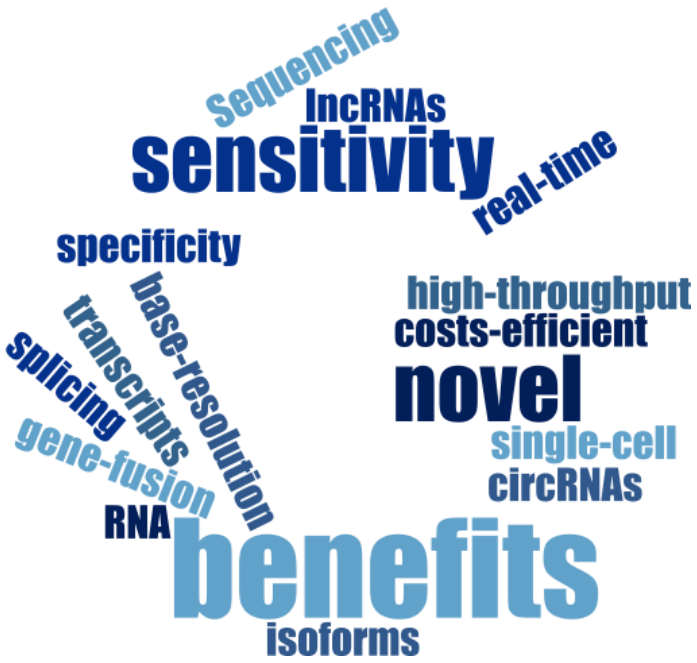


NGS

Technical advances lead the way

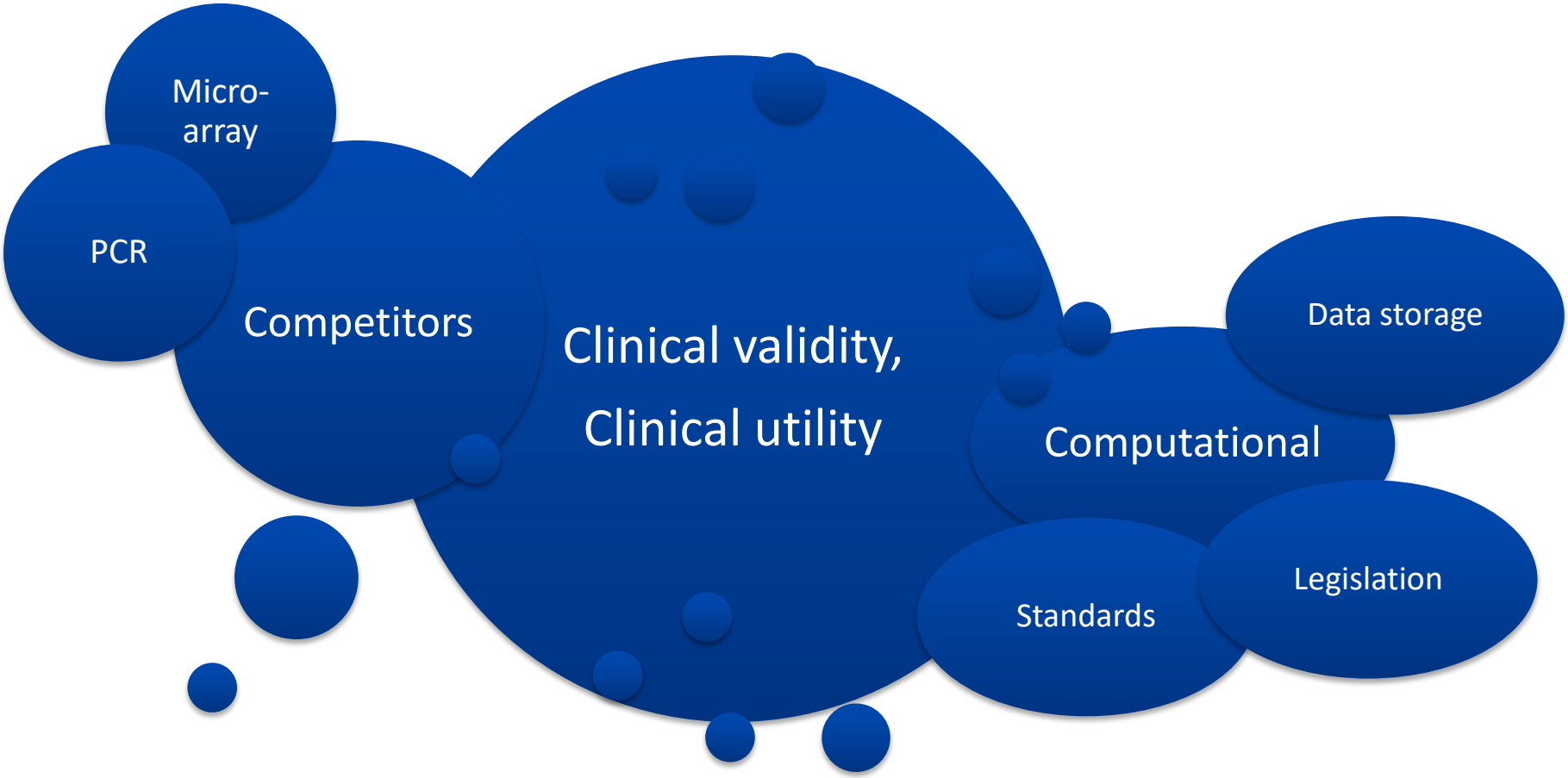


Why do we need NGS?



"RNA-Seq is able to identify thousands of differentially expressed genes, tens of thousands of differentially expressed gene isoforms and can detect mutations and germline variations for hundreds to thousands of expressed genetic variants, as well as detecting chimeric gene fusions, transcript isoforms and splice variants."
Wang, *Nat Rev. Genet*., 2009

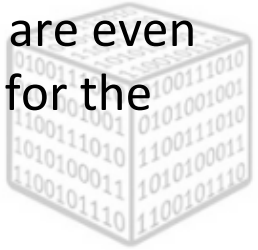
What are current challenges of NGS, e.g. in the clinic



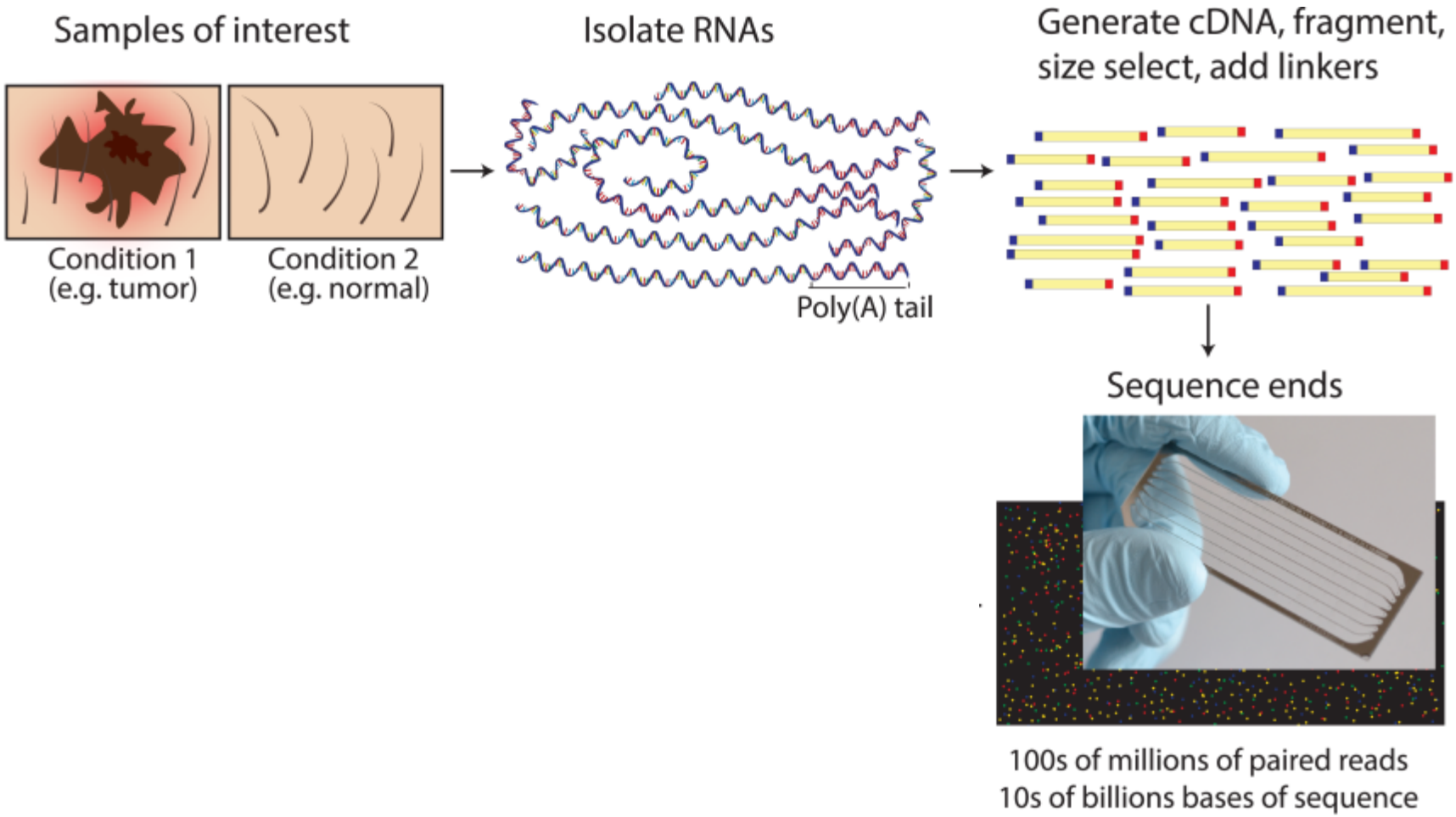
Where do I get NGS data?

- Databases, two popular examples
  - Sequence Read Archive (SRA) - https://www.ncbi.nlm.nih.gov/sra
    - Makes biological raw sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets (including Roche 454 GS System, Illumina Genome Analyzer, Applied Biosystems SOLiD System, Helicos Heliscope, Complete Genomics, and Pacific Biosciences SMRT).
  - The Cancer Genome Atlas (TCGA) - https://portal.gdc.cancer.gov/
    - Publishing the Pan-Cancer Atlas : a collection of cross-cancer analyses delving into overarching themes on cancer, including cell-of-origin patterns, oncogenic processes and signaling pathways.

- Experimental partners who have a wet lab for sample preparation or are even equipped with a sequencing device (there are also lots of companies for the sequencing procedure with the latest machines)



From sample to readout



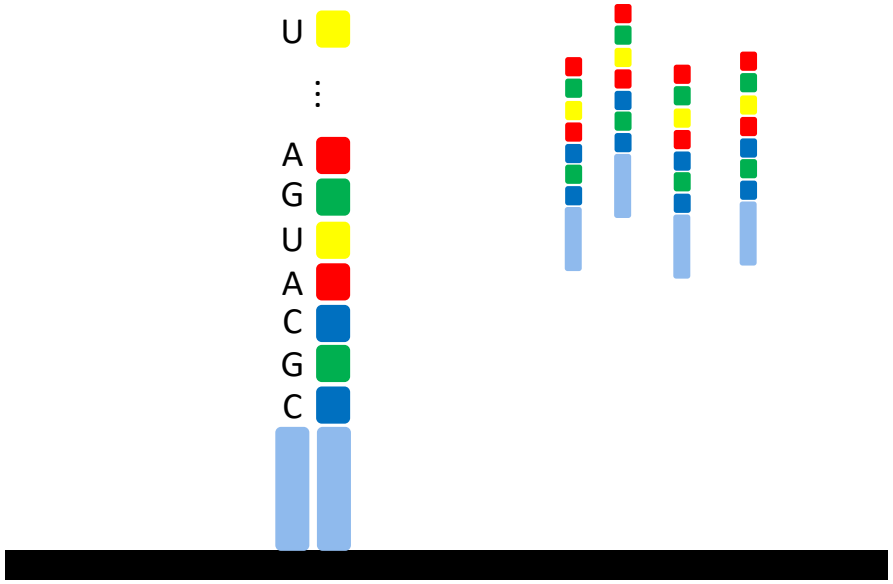
Griffith, Plos Comp. Biol., 2015

More accurate with RNA-Sequencing?

Sequencing procedure:

A – U

G – C

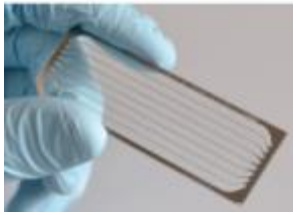


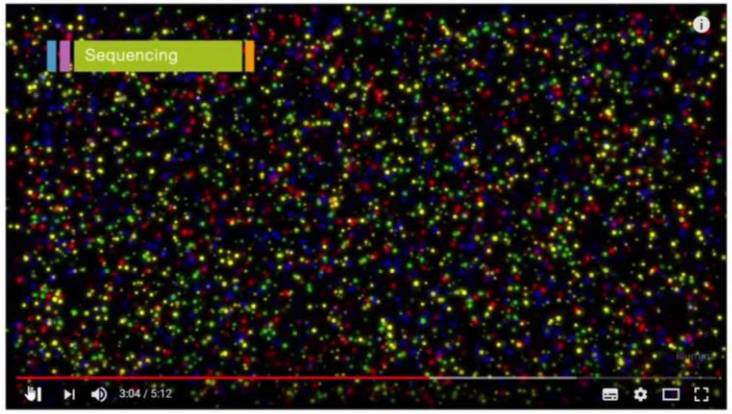
>10^7 Sequences

Sequencing Sample

Adapter

Flow cell



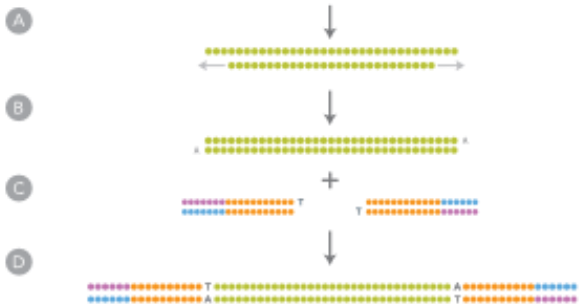


https://www.illumina.com

How do I get my NGS data - detailed?

1 Library Preparation

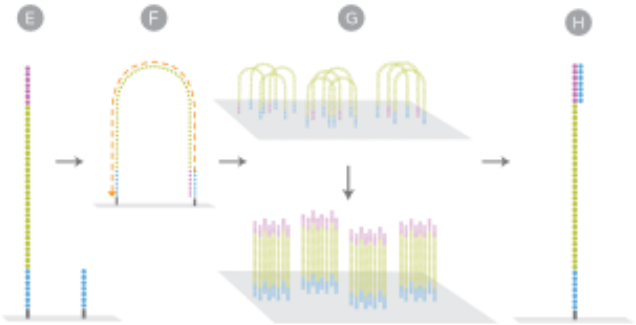
6 hours
3 hours hands-on time



A Fragment DNA
B Repair ends
Add A overhang
C Ligate adapters
D Select ligated DNA

2 Cluster Generation

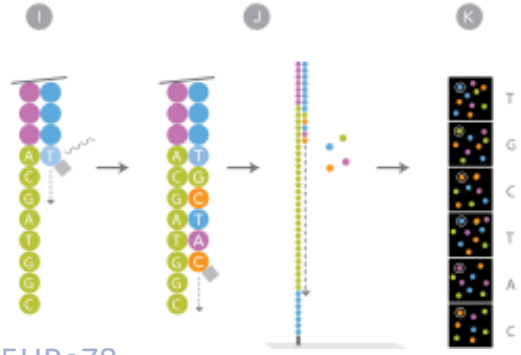
4 hours
< 10 minutes hands-on time
1–96 samples



E Attach DNA to
f ow cell
F Perform bridge
amplif cation
G Generate clusters
H Anneal sequencing
primer

3 Sequencing

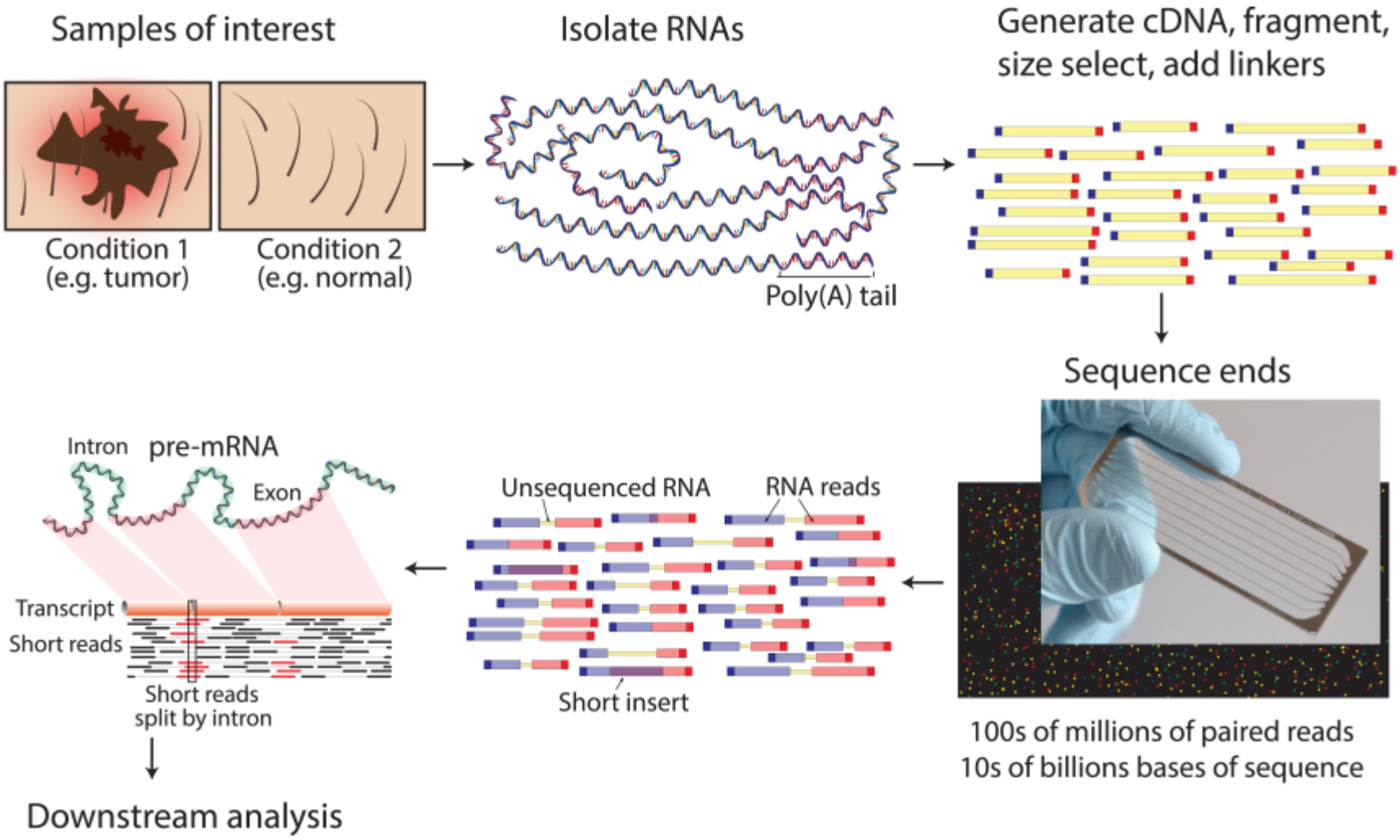
1–3 days single-read run
3–9 days paired-end run
30 minutes hands-on time
8 lanes, up to 96 samples
per f ow cell (run)



I Extend f rst base,
read, and deblock
J Repeat step above
to extend strand
K Generate base calls

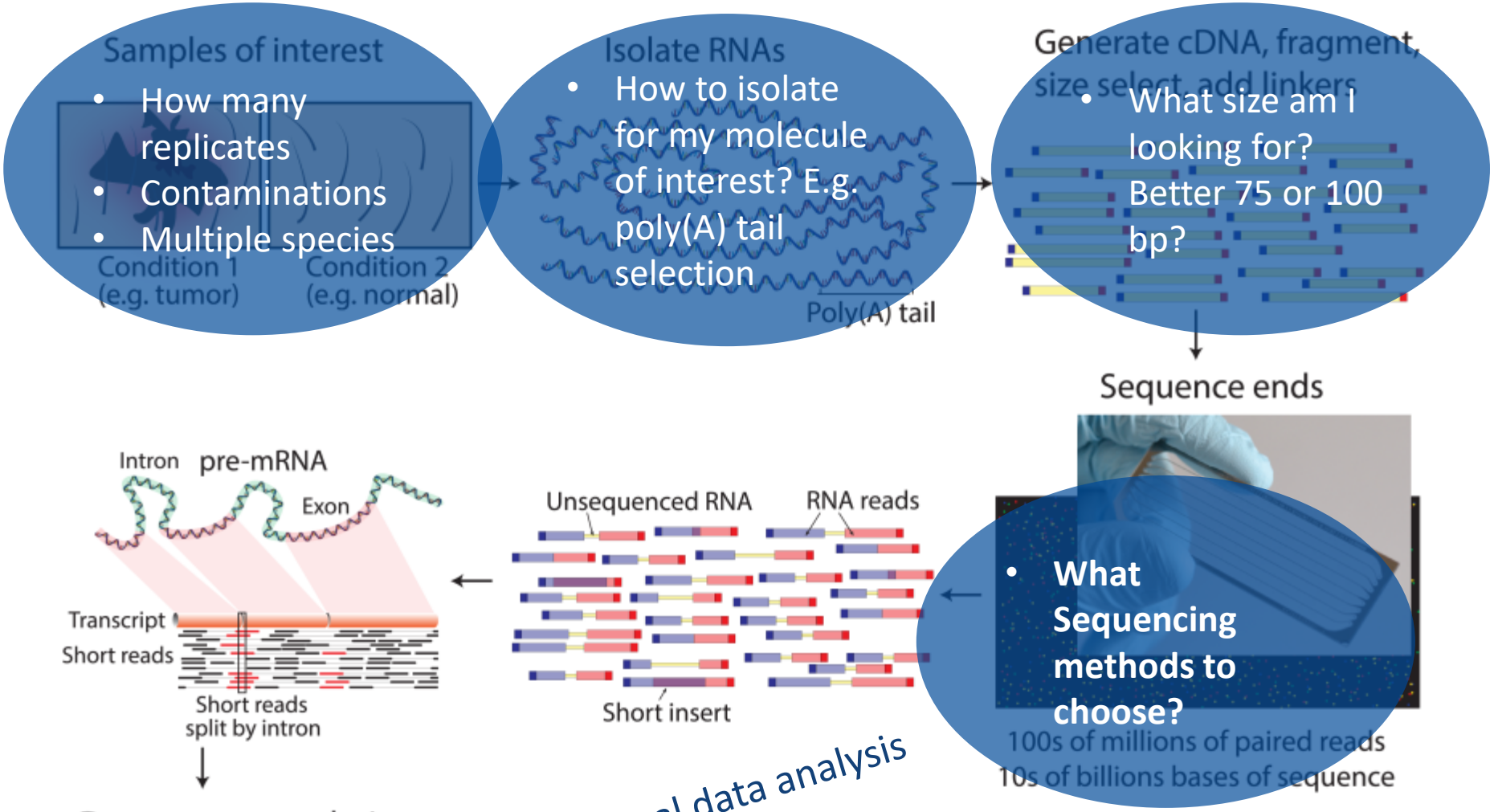
https://www.youtube.com/watch?v=fCd6B5HRaZ8

From sample to readout



Griffith, Plos Comp. Biol., 2015

From sample to readout



Computational data analysis is inevitable

Griffith, Plos Comp. Biol., 2015

Big Data and the need for new analyses

broadinstitute.org

geneprof.org

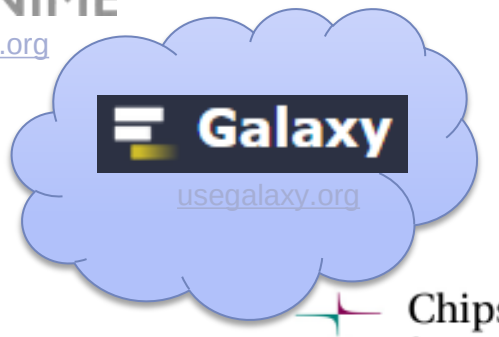
Grape

big.crg.cat/services/grape

knime.org

python.org

mapman.gabipd.org



usegalaxy.org

r-project.org/

snakemake

chipster.csc.fi

gene-talk.de

bioconductor.org

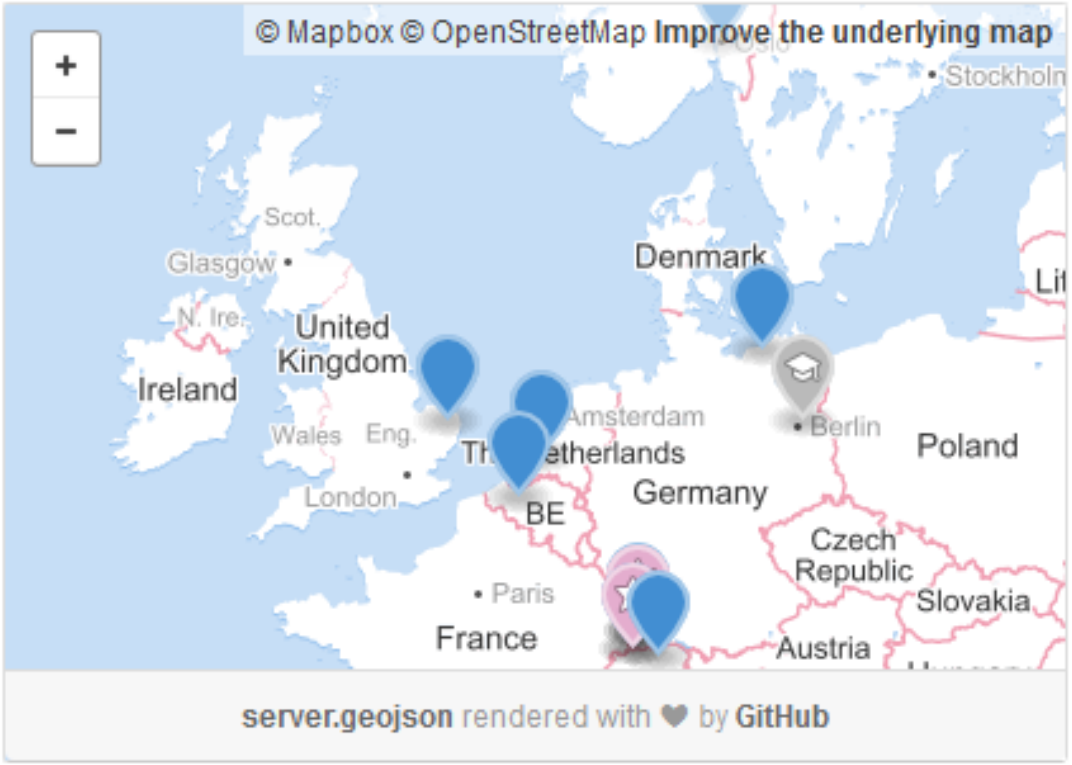
illumina.com

Supporting new data analysis approaches

- Key performance of Galaxy
  - Accessibilty
  - Reproducibility
  - Transparency

jupyter

jupyter.org/



Galaxy

usegalaxy.org

elixir

elixir-europe.org

de.NBI

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

denbi.de

Different Galaxy servers around me

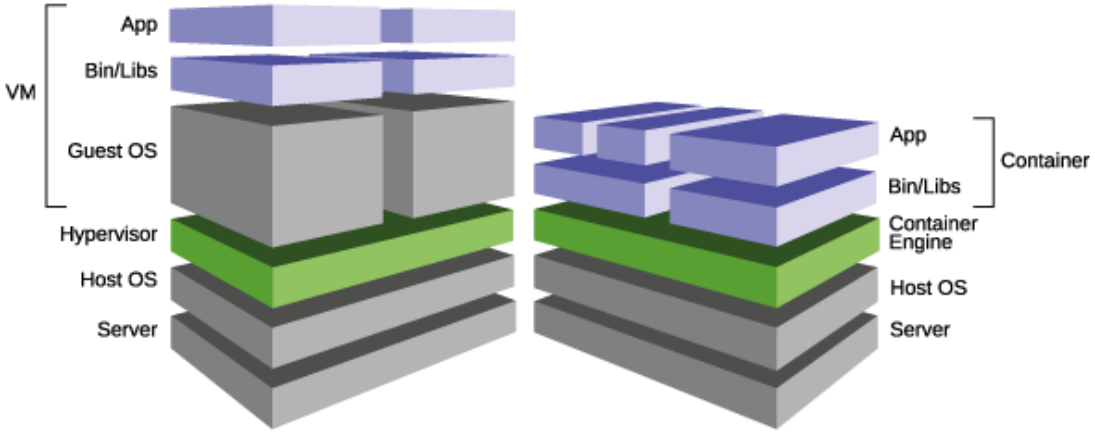
- Main galaxy (US): https://usegalaxy.org/
- European Galaxy (de.NBI support): https://usegalaxy.eu/
- More thhan 125 dedicated servers about every kind of scientific research https://galaxyproject.org/use/
- Have your own Galaxy with Docker!
  - RNA-Workbench - https://github.com/bgruening/galaxy-rna-workbench
  - Galaxy Modular Workflow Generator - https://github.com/destairdenbi/galaxy-modular-workflow-generator

What is Docker?

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

Docker is used to run software packages called "**containers**". Containers are isolated from each other and bundle their own application, tools, libraries and configuration files; they can communicate with each other through well-defined channels.

- Build for scale
- Extensible and flexible
- E.g. used by Ebay, GE, illumina, Spotify



One tool to share them all

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

Galaxy + docker = Symbiosis!

- Tailor-made, user specific and integration into a general framework to develop workflows adressing the users need and facilitating a reuse

- Stand-alone Docker container which “conserves“ your whole tool compilation (for an easy use – one command line or kitematic.com click!)

```
docker run -p 8080:80 bgruening/galaxy-rna-workbench
```

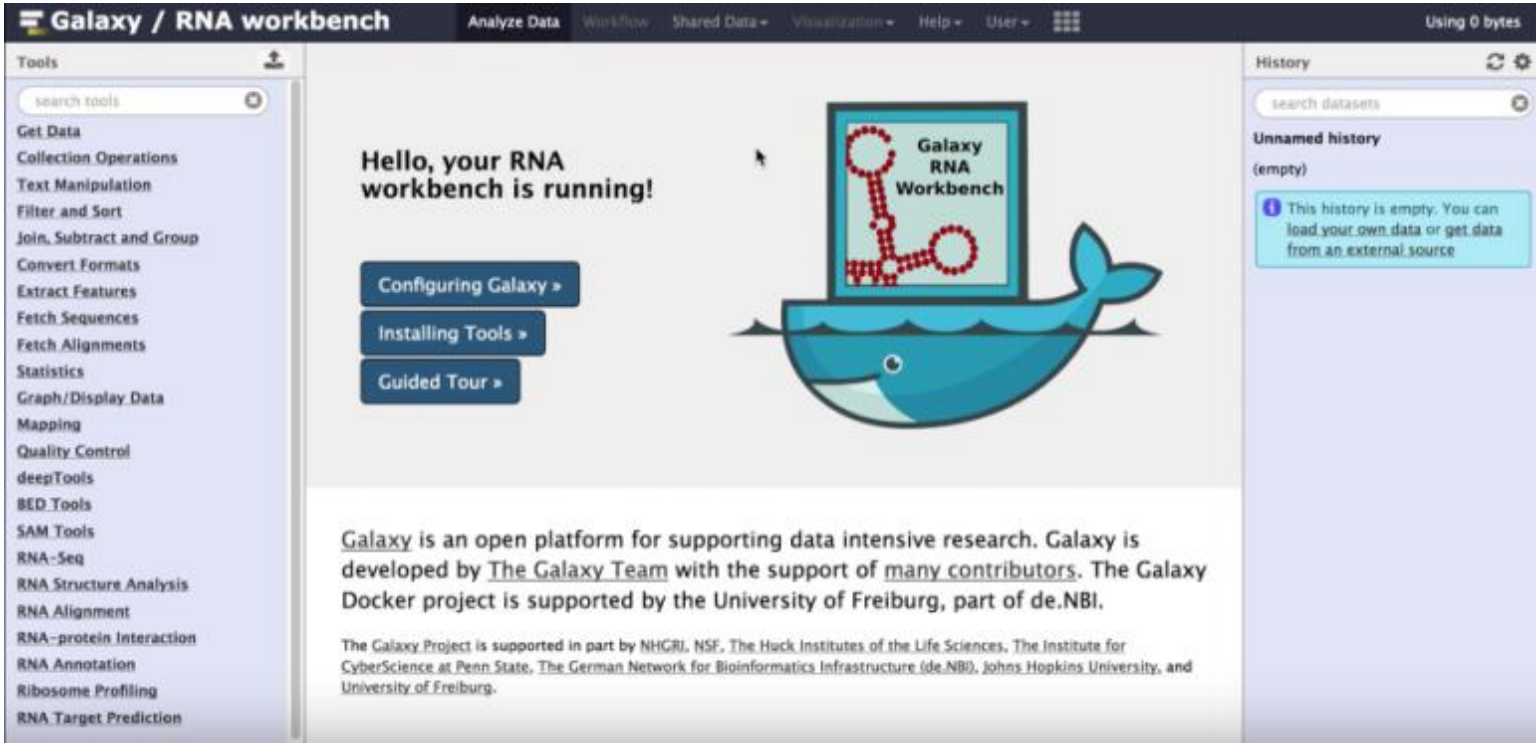
- Get a single minimized Docker container for every tool and obtain a maximum flexibility (more advanced)

Supporting the RNA Galaxy-workbench

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

- Specialized Galaxy instance for RNA analyses povided by the RBC
- Contains +50 tools for structure analyses, annotation, alignment and many more

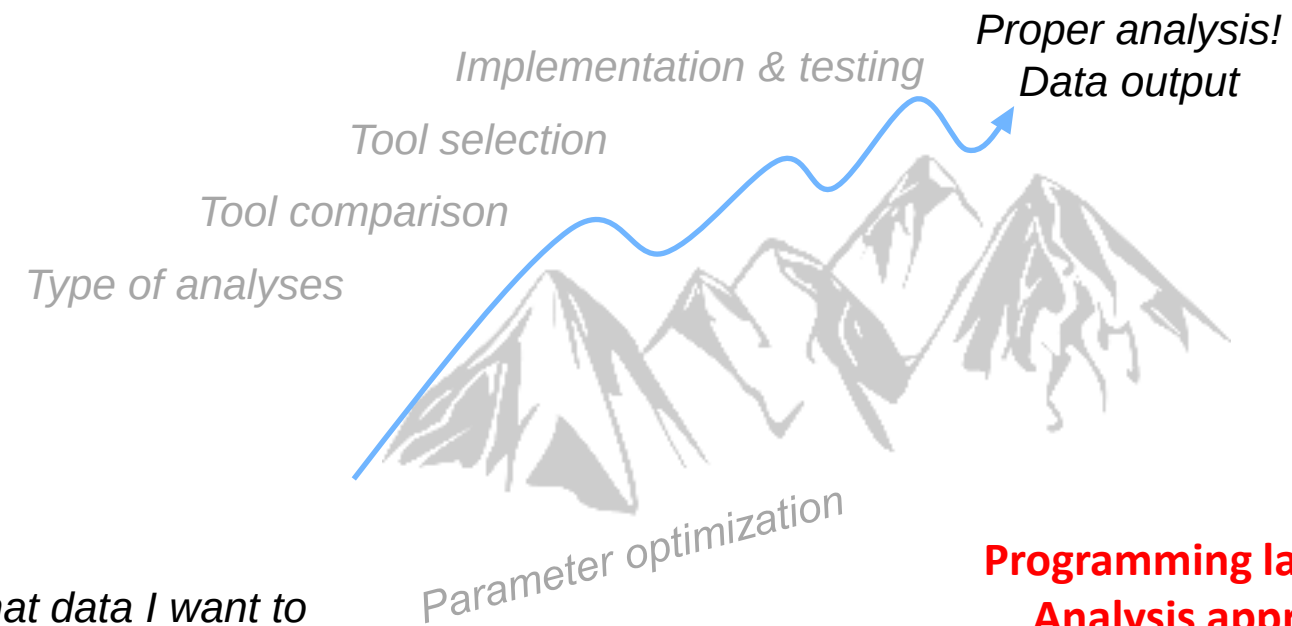
github.com/bgruening/galaxy-rna-workbench



Gruening *et al.*, *NRA*, *2017*

The struggle for the right approaches

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK



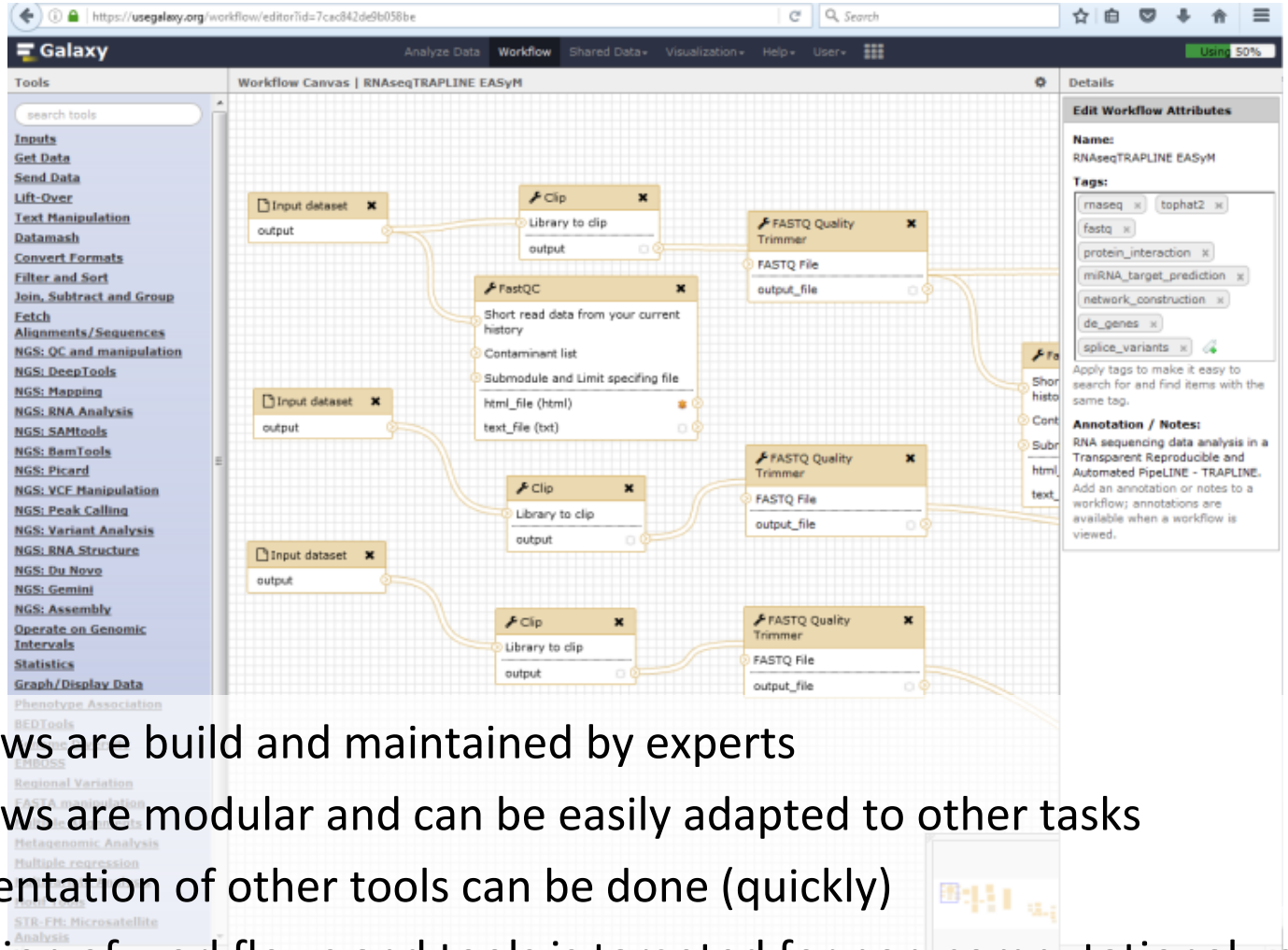
What data I want to analyze?
Data input

Programming language?
Analysis approach?

Lott, Wolfien, Riege, Bagnacani, et al., *J.Biotech*, 2017

Using workflow development

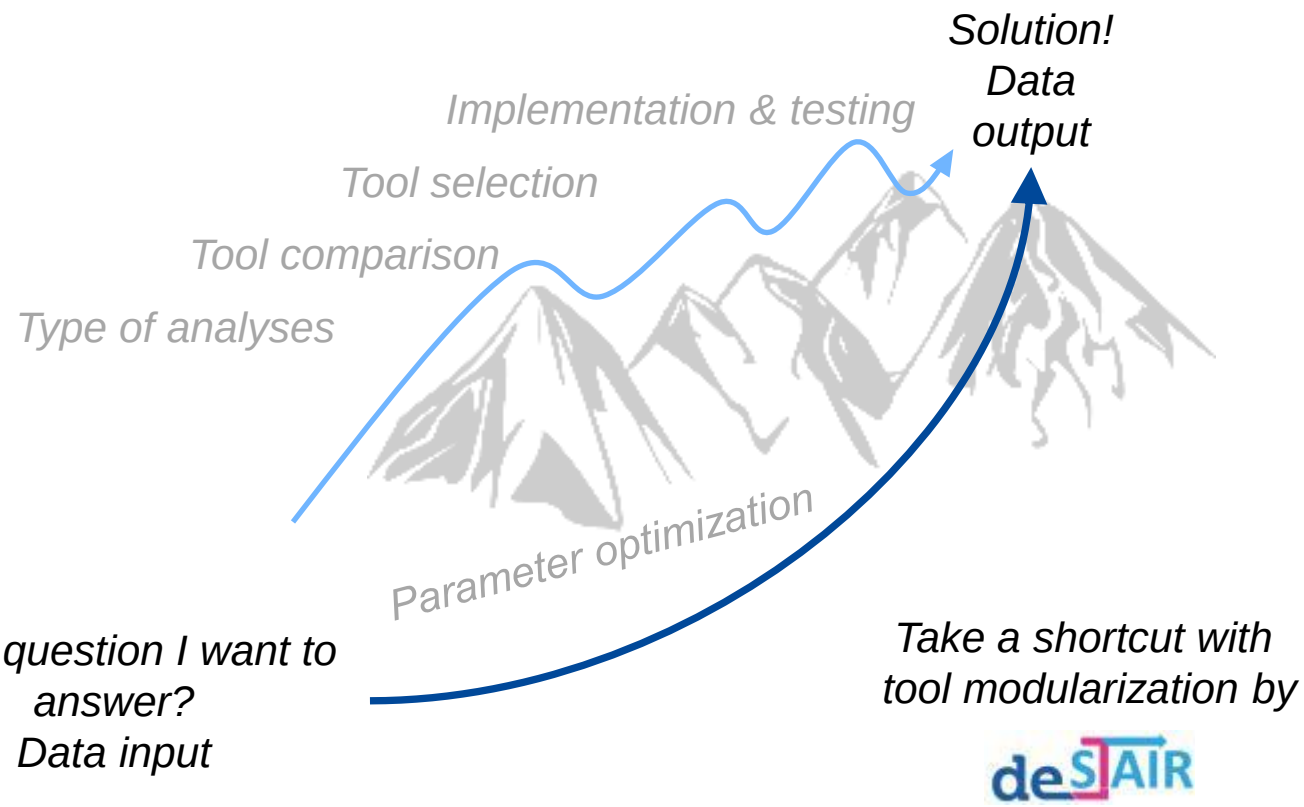
SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK



- Workflows are build and maintained by experts
- Workflows are modular and can be easily adapted to other tasks
- Implementation of other tools can be done (quickly)
- Application of workflows and tools is targeted for non-computational users

Why using workflows?

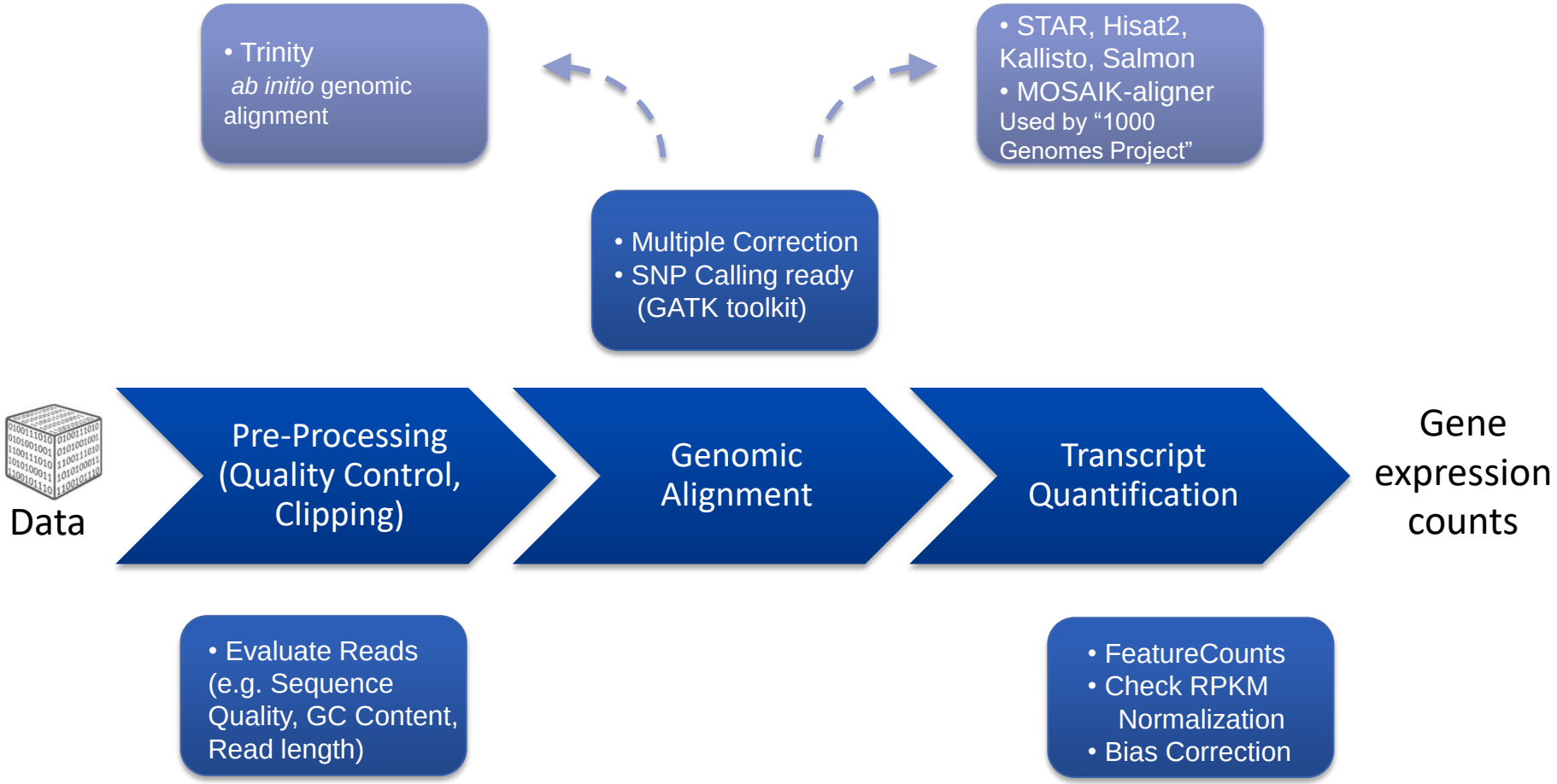
SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK



Lott, Wolfien, Riege, Bagnacani, *et al*., *J.Biotech*, 2017

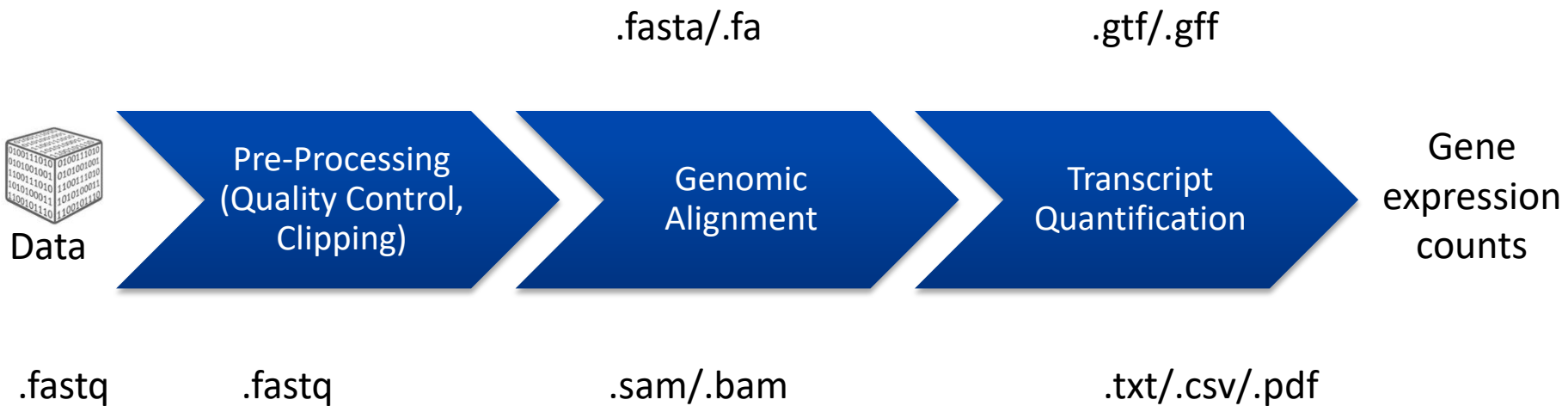
Basic workflow for differential expression analysis

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK



Sequencing data formats

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK



Basic workflow for differential expression analysis

Data

Pre-Processing (Quality Control, Clipping)

Genomic Alignment

Transcript Quantification

Gene expression counts

Condition A

Condition B

Condition C

Multiple Copies of a Transcriptome

Reads

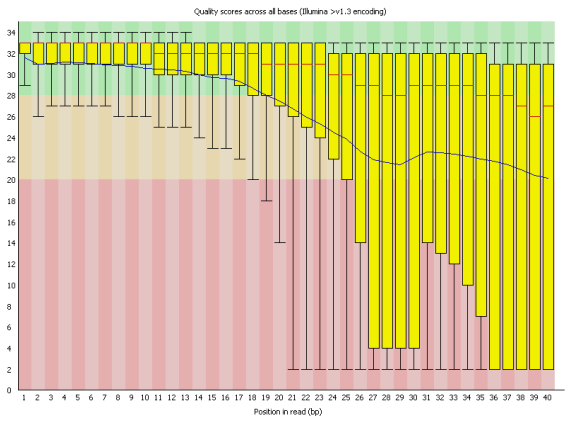
Ref. Sequence

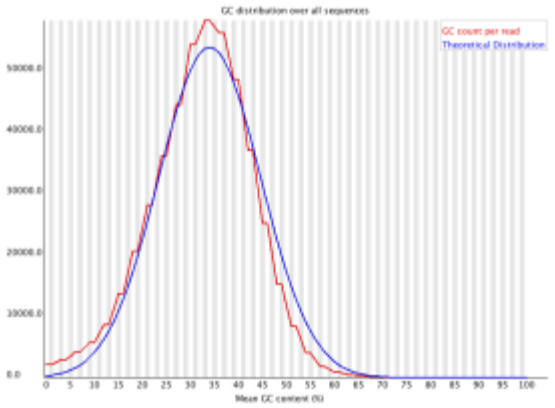
Reads on Ref. Seq.

Differentially expressed genes/transcripts

Differential expression (DESeq2, Sleuth)

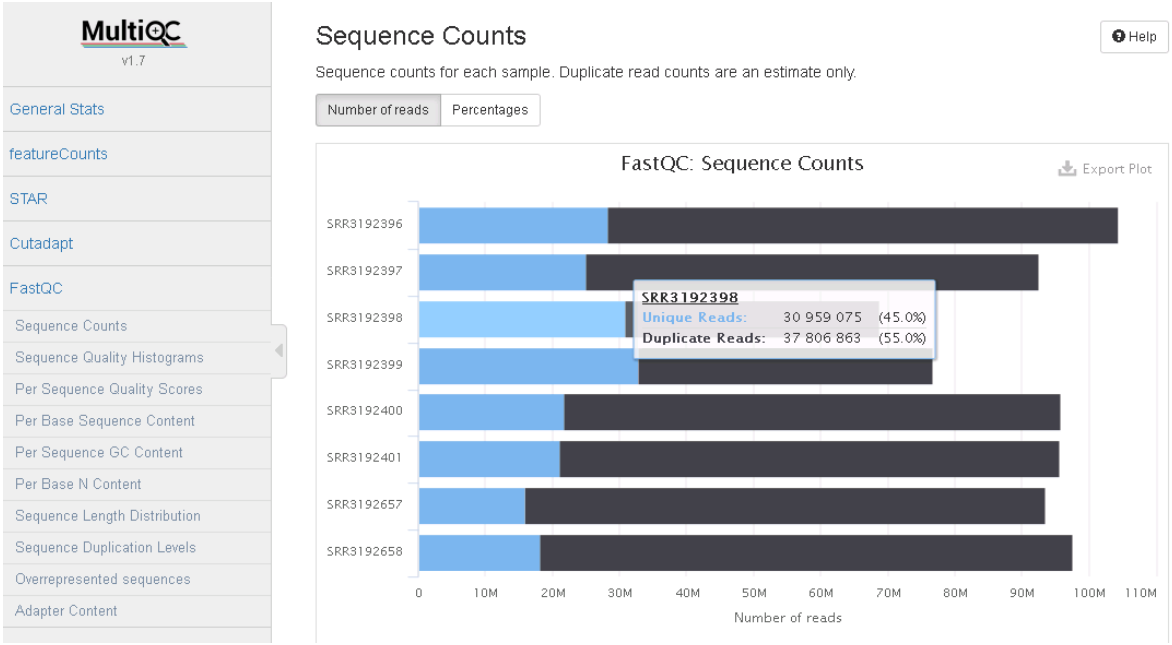
Visualization of pre-processing results



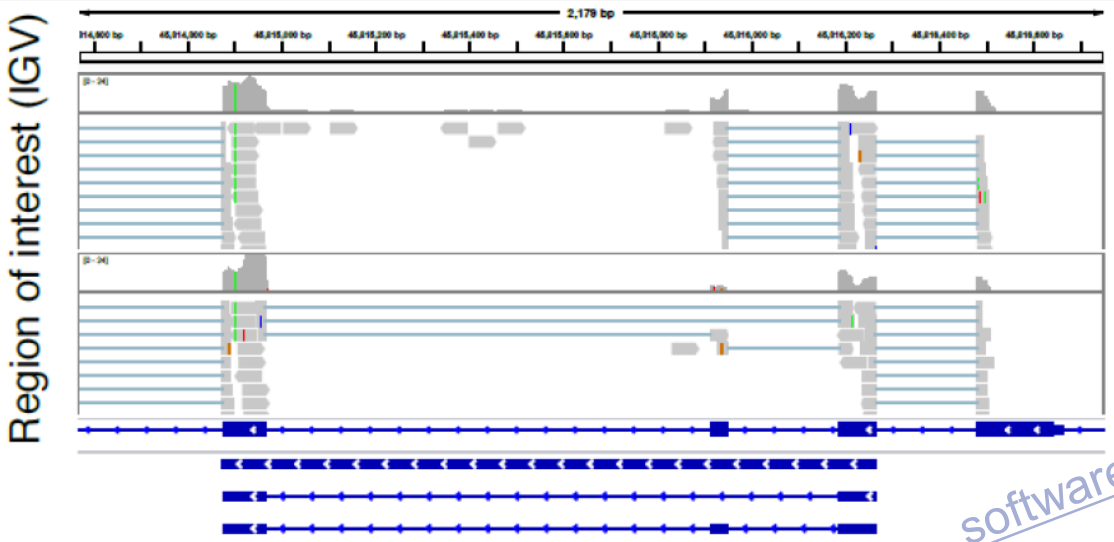


FastQC - https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

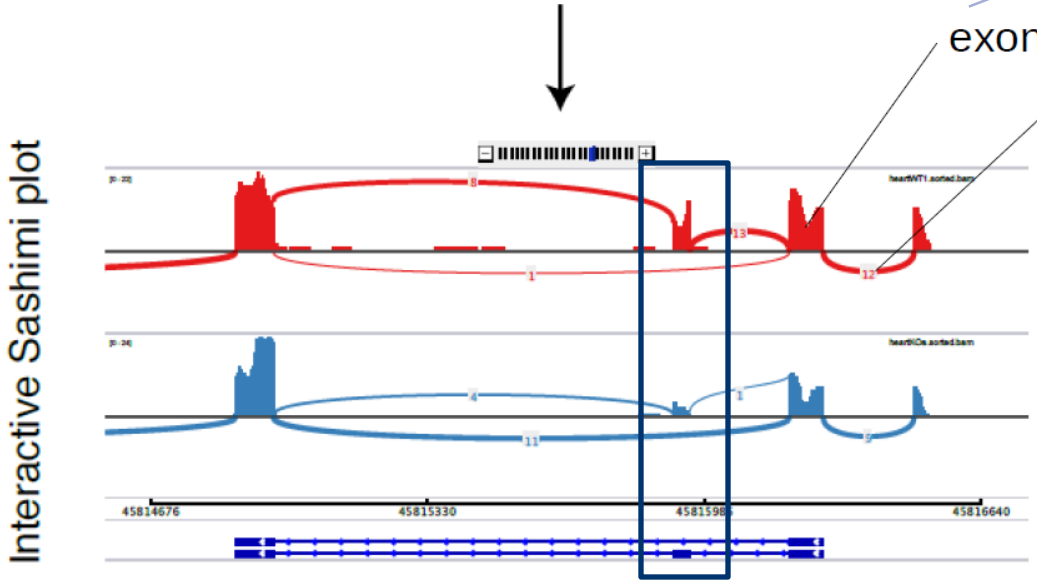
MultiQC - https://multiqc.info/



Visualization of mapping results



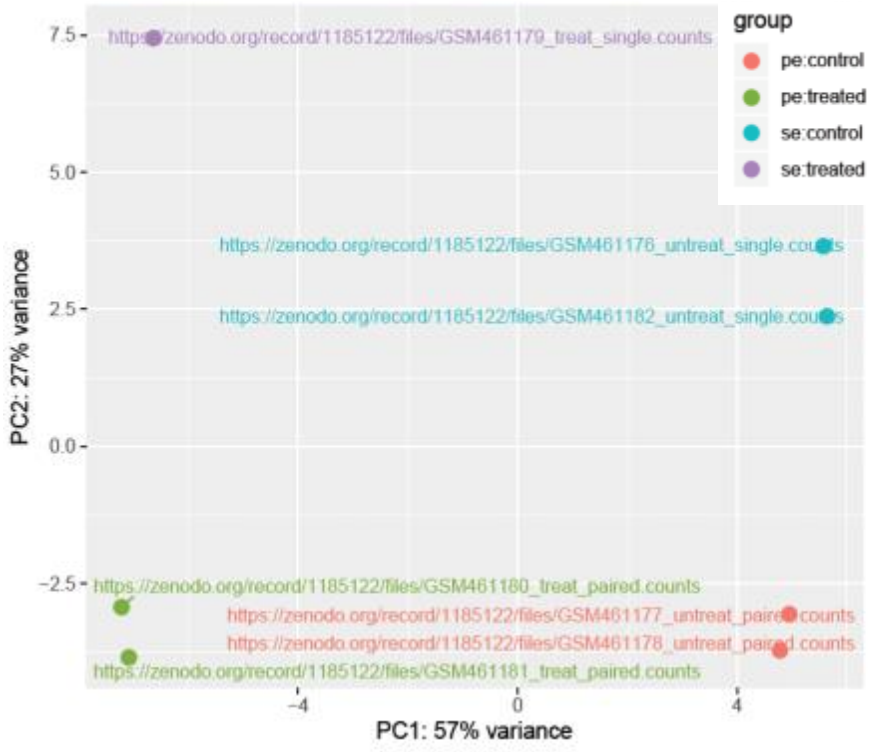
software.broadinstitute.org/software/igv/

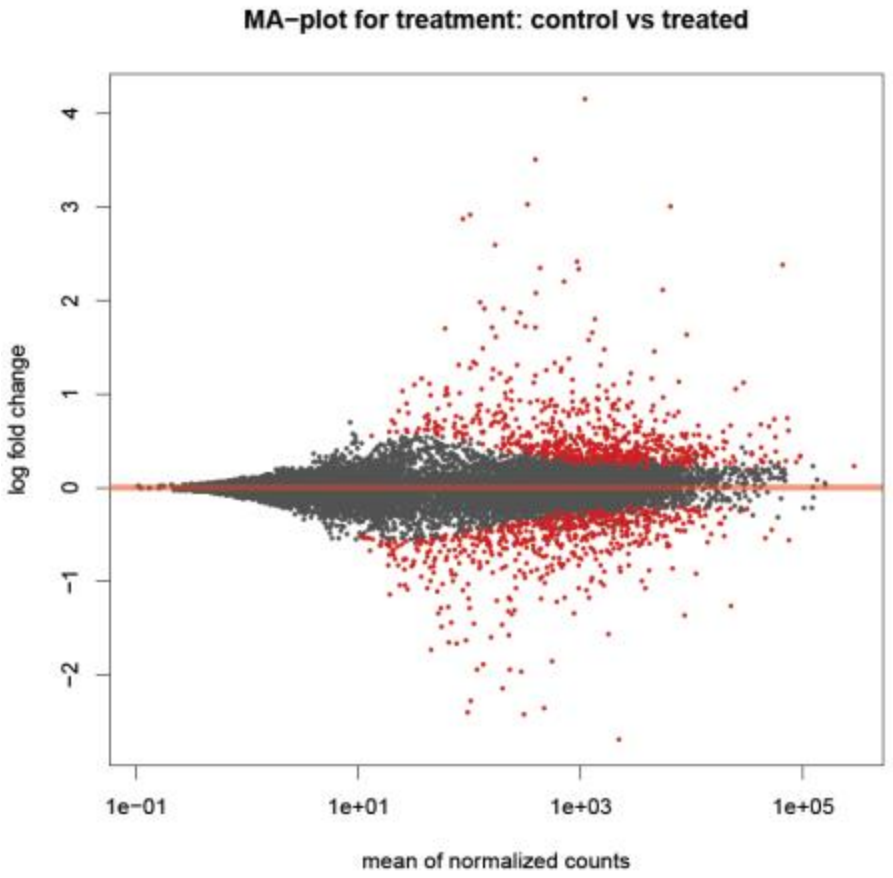


Sashimi plots:
Quantitative visualization of read coverage along exons and splice junctions

Katz, *EMBO*, 2015

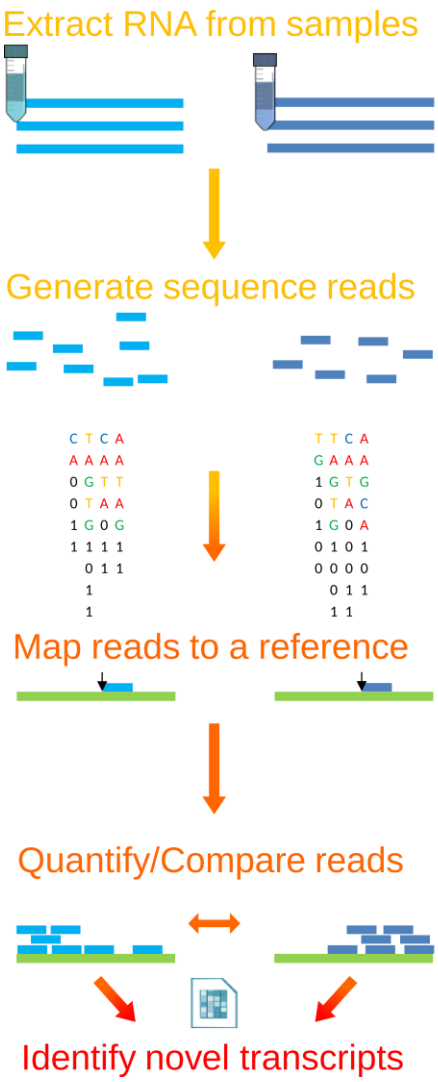
Visualization of quantification results





DESeq2- Love et al. 2014
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302049/

Sequencing experimentation and data analysis summary



Experiment

Sample preparation
- ☐ Amount of replicates (n ≤ 3) for the desired experiment
- ☐ No contaminations that effect the RNA level
- ☐ Multiple species within a sample (e.g. metatranscriptomics)
- ☐ Cell culture routine with/out antibiotics

RNA isolation
- ☐ Choice of isolation (e.g. Poly(A)-selection)
- ☐ Use latest purification kits
- ☐ Work RNAase free
- ☐ Low number of cell passage
- ☐ Use globin depletion kit for whole blood

Library generation
- ☐ Selection of suitable polymerases, thermocycling times and buffer conditions
- ☐ Quickly transfer RNA into cDNA
- ☐ Choice for desired amount and size of reads
- ☐ Choice for suitable sequencing method

Data analysis

Preprocessing
- ☐ Demultiplex data and remove barcodes
- ☐ Do quality control (check for GC content, read size, sequencing errors in lanes & cycles)
- ☐ Perform adapter clipping and quality trimming
- ☐ Check again for the quality

Genomic alignment
- ☐ Use the latest reference genome (e.g. UCSC, Ensembl)
- ☐ Choose correct aligner for your task (e.g. RNA quantification, *de novo* alignment, gene-fusion detection, etc.)

Quantification
- ☐ Use the latest annotation set
- ☐ Normalize your data (e.g. RPKM, TPM)
- ☐ Use different statistics to identify significantly differentially expressed transcripts
- ☐ Fold-change > 2
- ☐ q-value < 0.05

Wolfien *et al*. 2019

Literature for best practices

- Lott SC, Wolfien M, Riege K, Bagnacani A, Wolkenhauer O, Hoffmann S, et al. Customized workflow development and data modularization concepts for RNA-Sequencing and metatranscriptome experiments. *J Biotechnol*. 2017 Jul; Available from: http://linkinghub.elsevier.com/retrieve/pii/S0168165617314992

- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016. Available from: http://genomebiology.com/2016/17/1/13

- Wolfien M, Brauer DL, Bagnacani A, Wolkenhauer O. Workflow Development for the Functional Characterization of ncRNAs. In *Springer Nature*, New York, NY; 2019. Available from: http://link.springer.com/10.1007/978-1-4939-8982-9_5

Do you want more NGS? - Further training courses

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	7
Assembly	4
ChIP-Seq data analysis	3
Epigenetics	3
Genome Annotation	3
Metabolomics	2
Metagenomics	2
Proteomics	12
Sequence analysis	5
Statistics and machine learning	3
Transcriptomics	17
Variant Analysis	6

Galaxy Tips & Tricks

Topic	Tutorials
Data Manipulation	4
User Interface and Features	3

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	34
Development in Galaxy	13

http://galaxyproject.github.io/training-material/

de.STAIR & Galaxy training material

Course material: https://github.com/destairdenbi/trainings

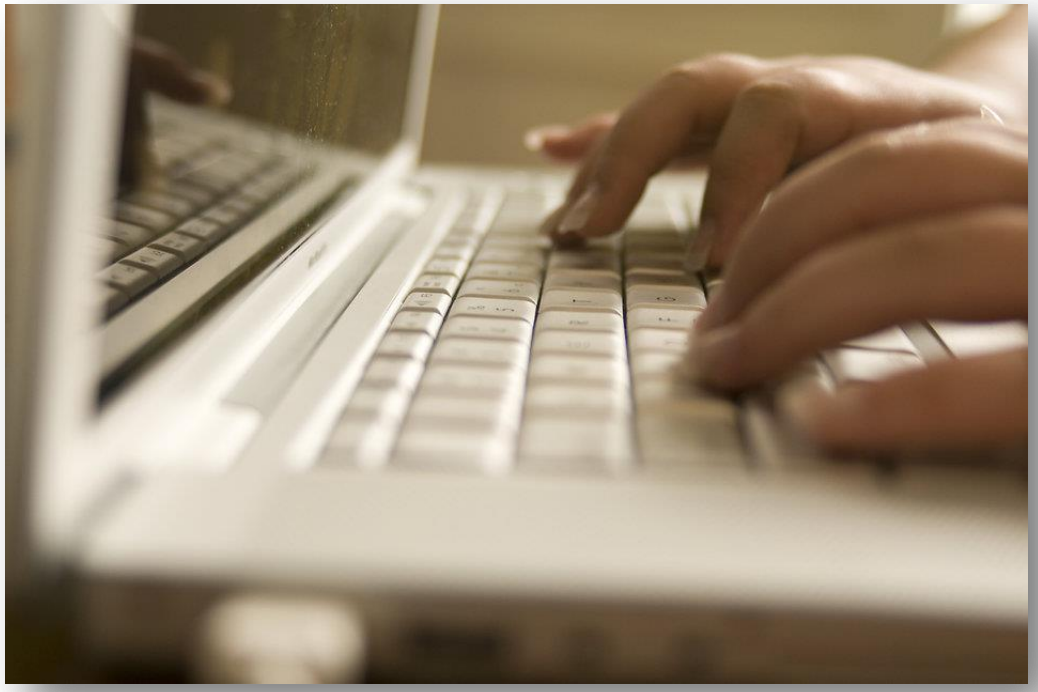
Training material: http://galaxyproject.github.io/training-material/

Manuscript: https://www.cell.com/cell-systems/fulltext/S2405-4712(18)30230-8

Hands-on part 1 11:15 – 12:00

“Introduction to Galaxy“

Material: https://galaxyproject.github.io/training-material/topics/introduction/



Time for a break …

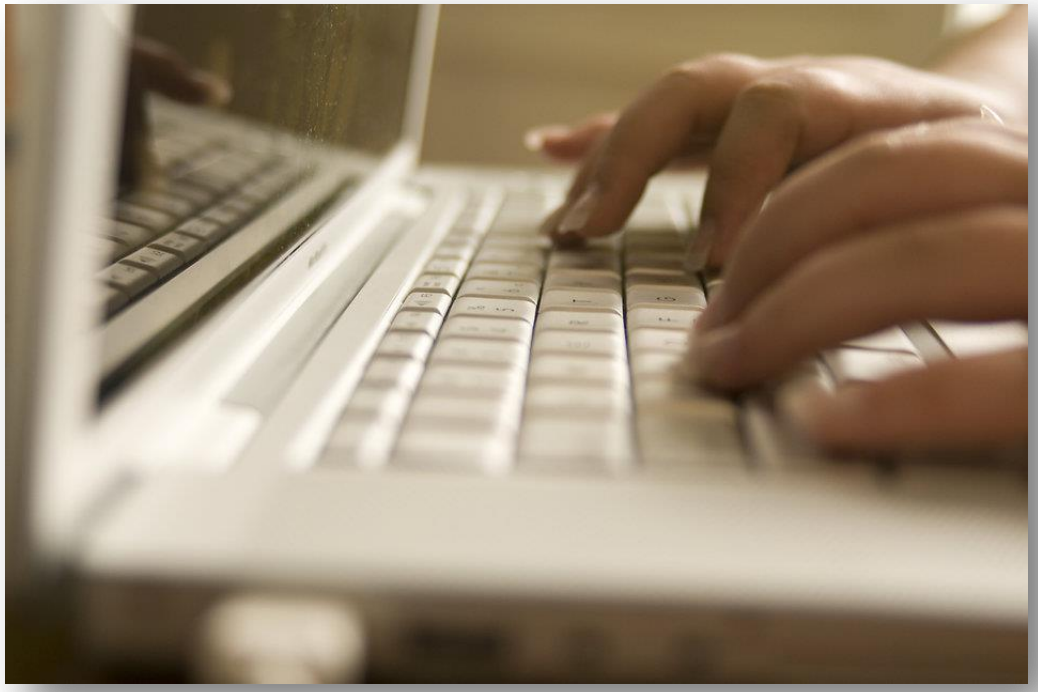
SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

12:00 – 13:45

Hands-on part 2 13:45 – 14:30

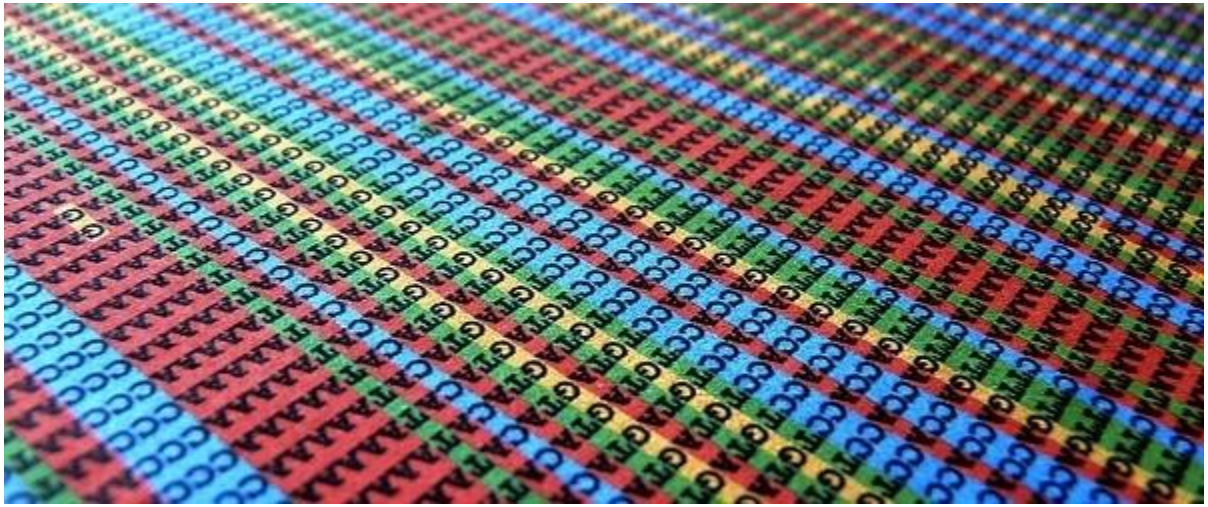
“RNA-Seq data preprocessing and quality control“

Material: http://galaxyproject.github.io/training-material/topics/sequence-analysis/



Hands-on part 3 14:30 – 15:15

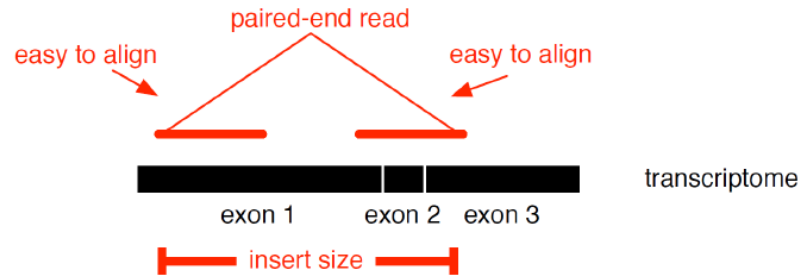
"Application of different read mapping approaches for genomic alignment"



Genomic alignment - Mapping to a reference genome

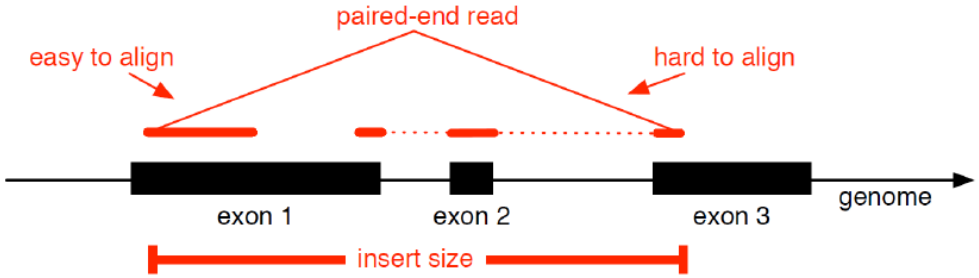
SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

Transcriptome alignment



- reliable gene models required
- no detection of novel genes

Genome alignment (splice-aware read alignment)

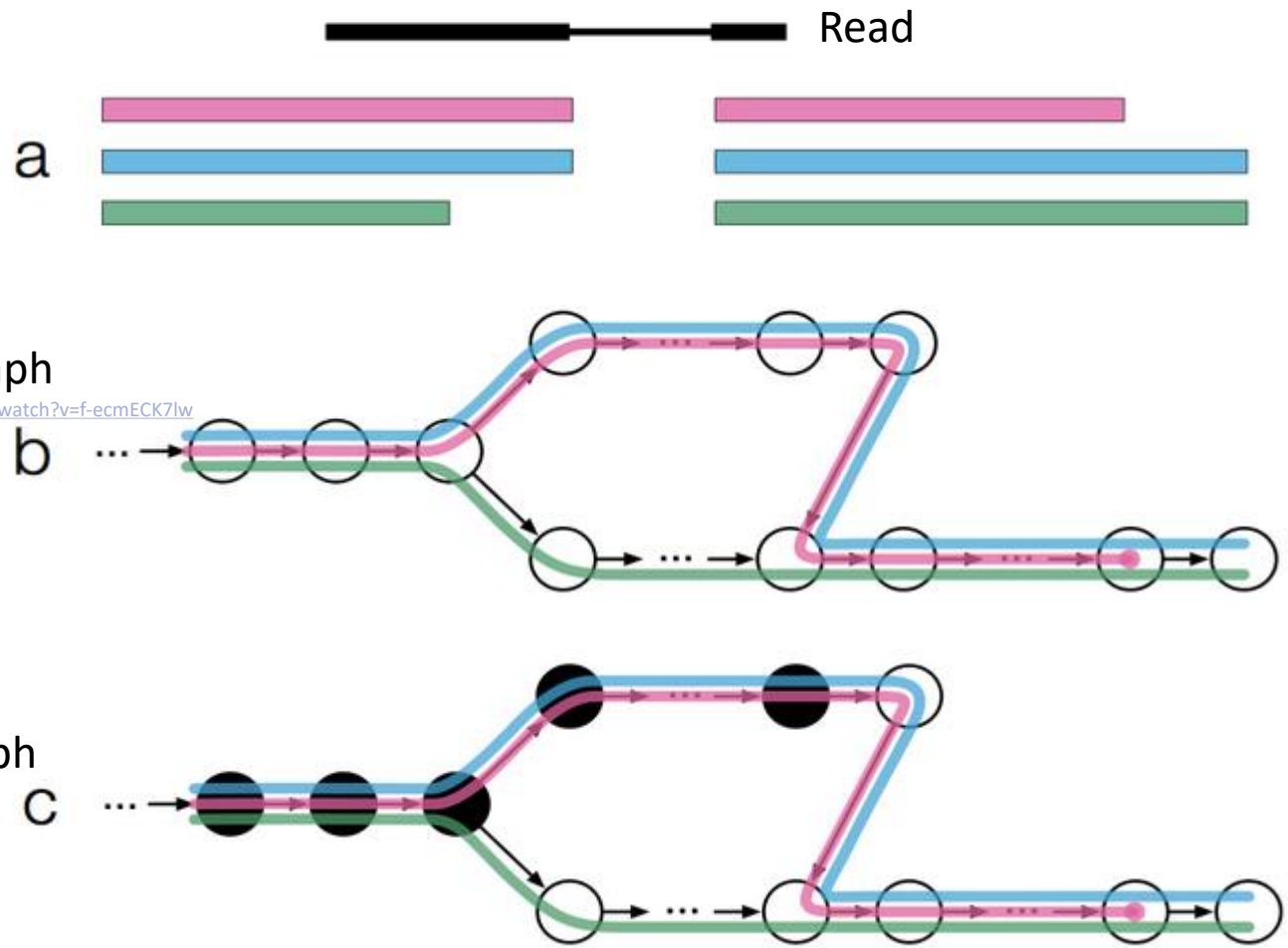


+ detection of novel genes and isoforms

Turro, EMBO, 2012

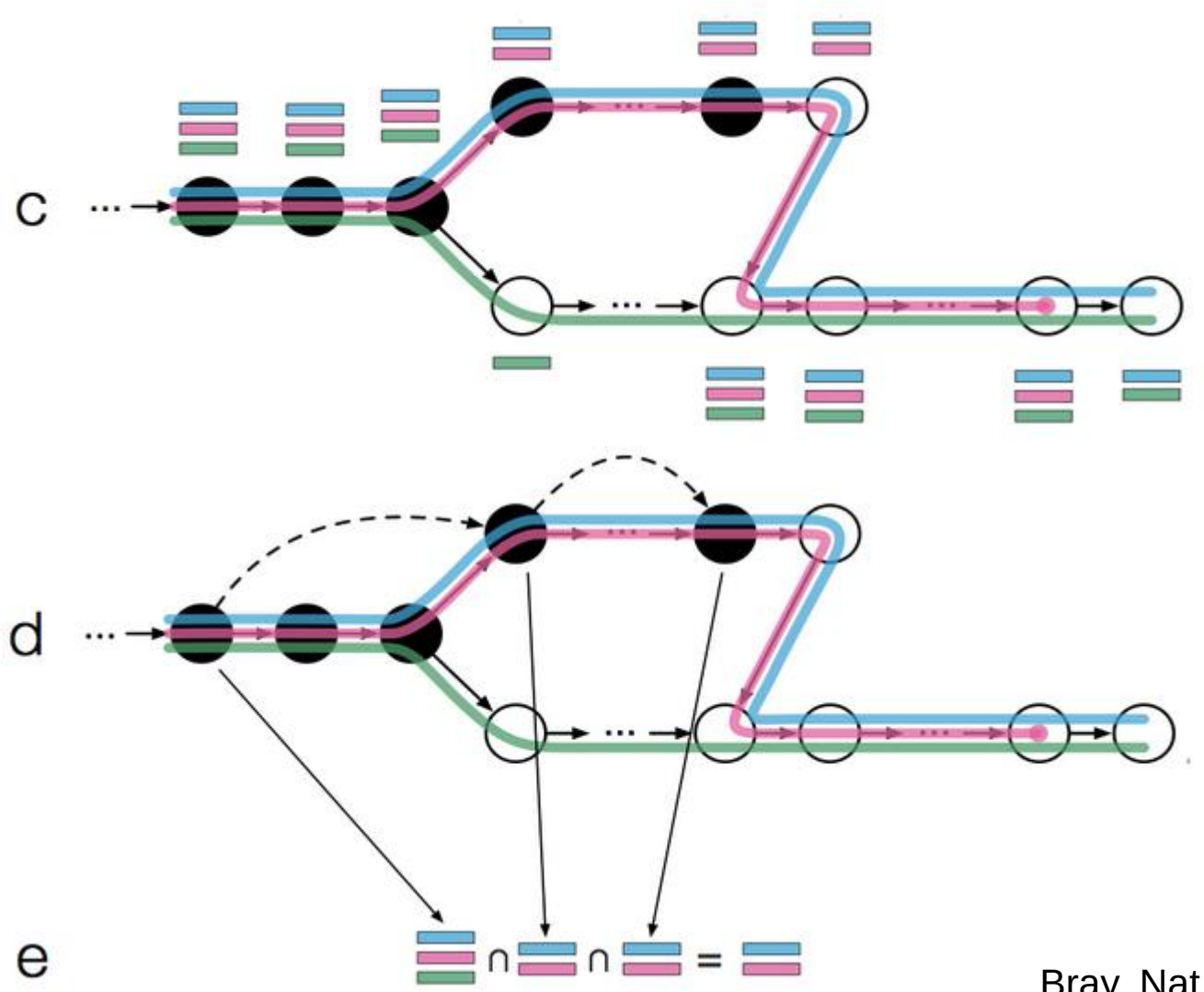
For clinical usage combination of different algorithms possible:

Genomic alignment - pseudoalignment



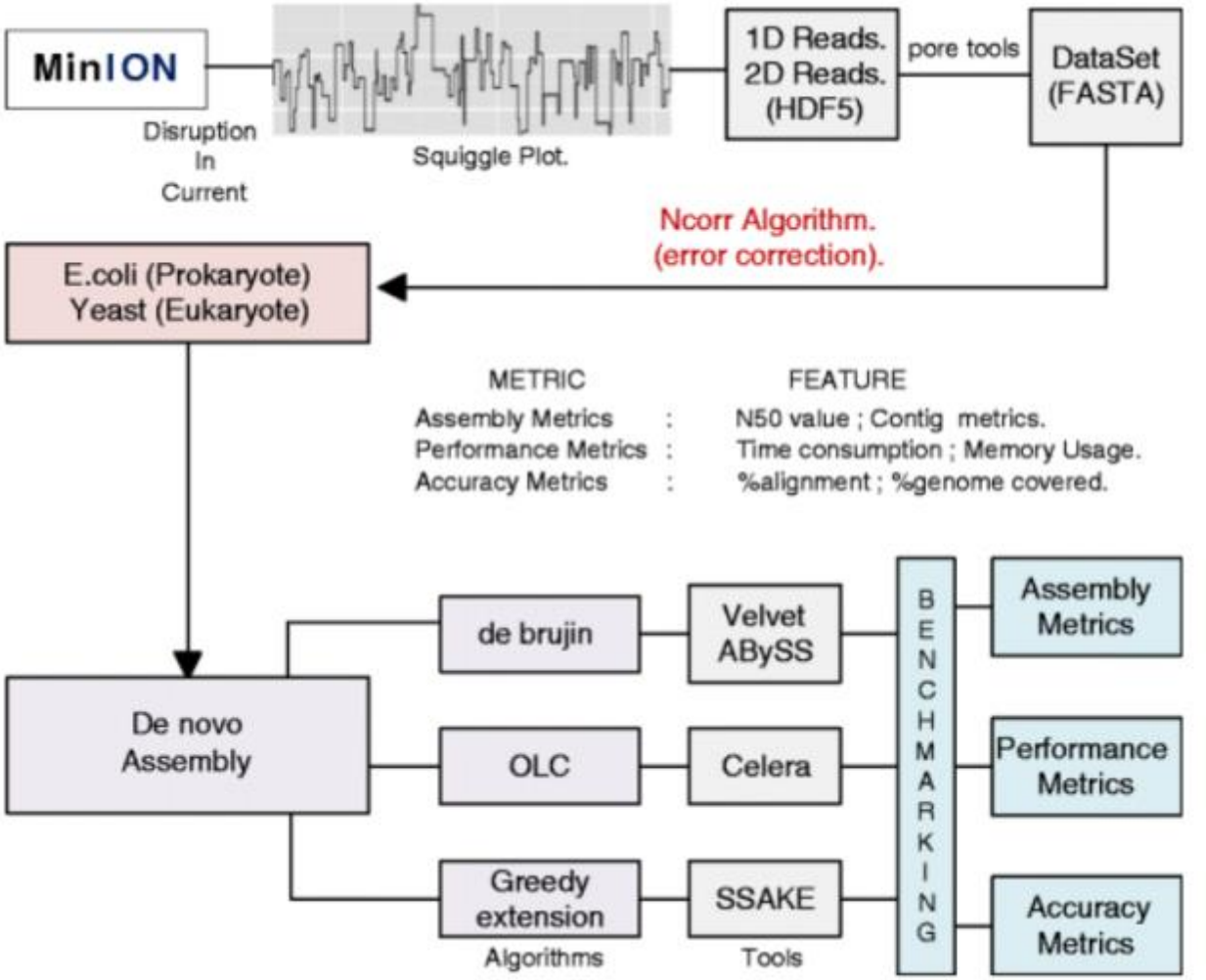
Bray, Nat Biotech, 2016

Genomic alignment - pseudoalignment



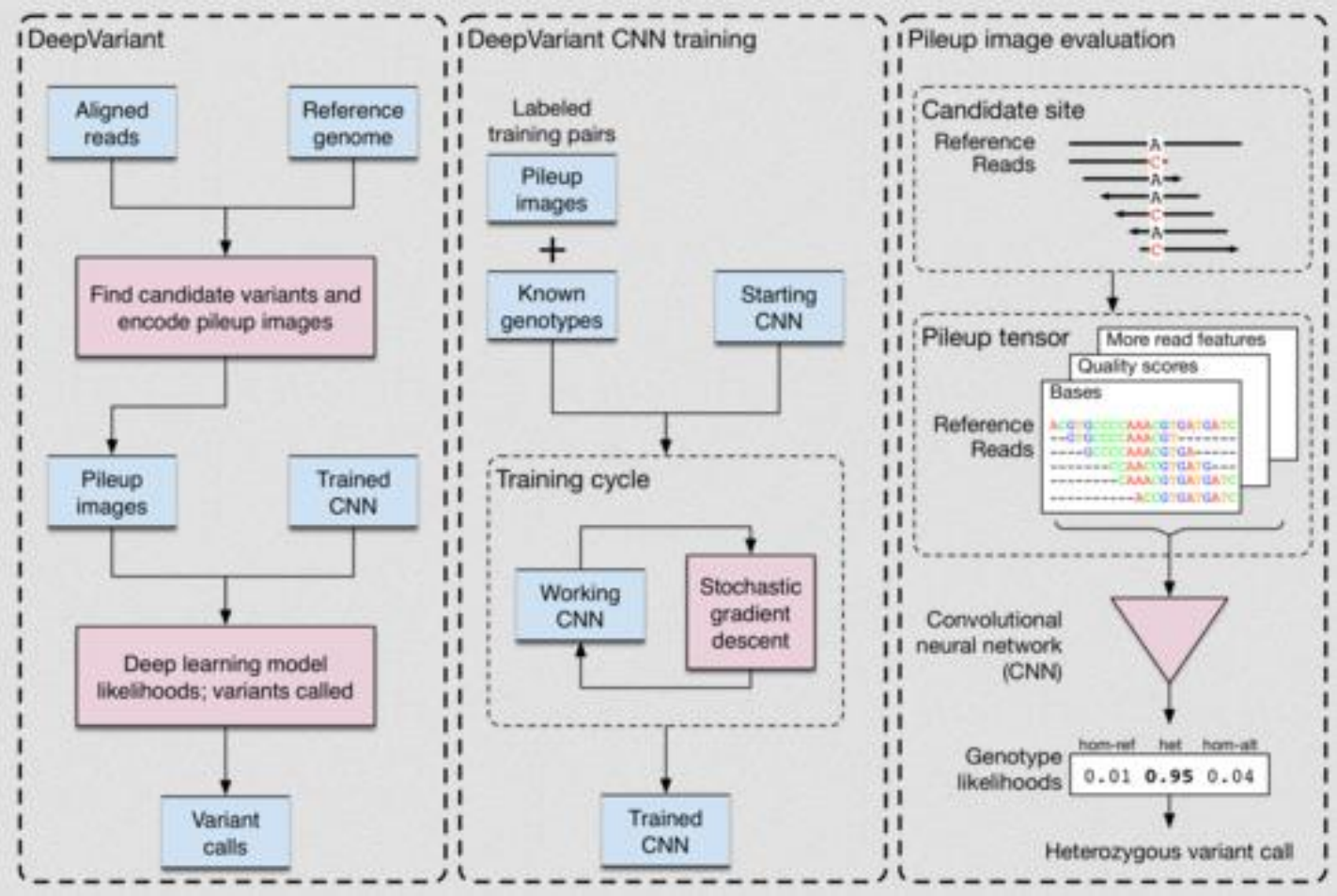
Bray, Nat Biotech, 2016

De novo alignment for nanopore seq



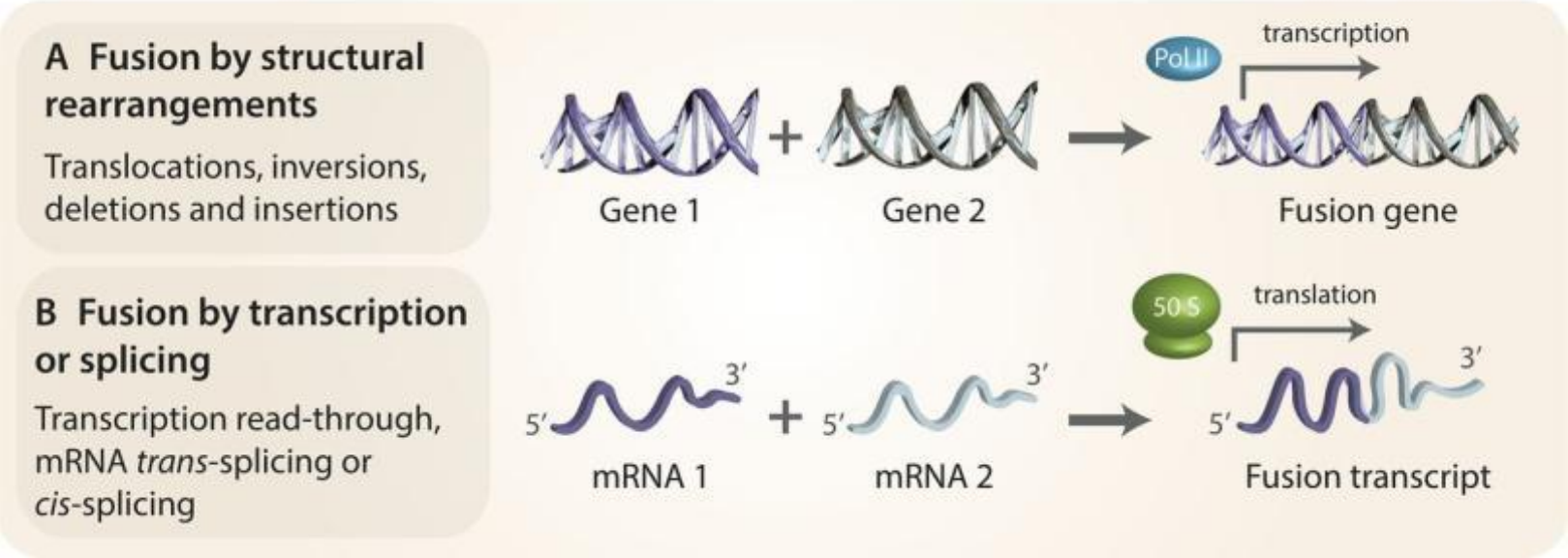
Cherkuri, BMC Genomics, 2016

Genomic alignment – deep neural network



https://github.com/google/deepvariant

Gene fusion detection



Natasha, *Nucl. Acids Res.*, 2016

" Gene fusions are associated with oncogenic properties, and often act as driver mutations in a wide array of cancer types."

- Deregulating one of the involved genes
- Forming a fusion protein with oncogenic functionality
- Inducing a loss of function

Yoshihara, *Oncogene*, 2015

Visualization of .bam files

http://software.broadinstitute.org/software/igv/

The James Hutton Institute

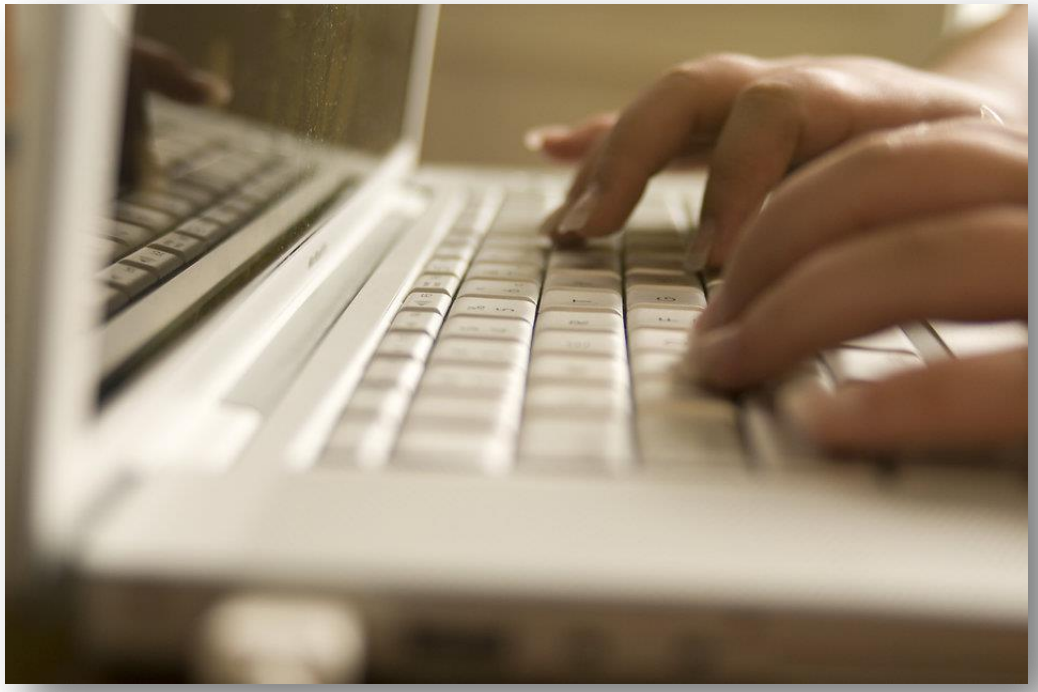
Tablet

https://ics.hutton.ac.uk/tablet/

Hands-on part 3 14:30 – 15:15

“Application of different read mapping approaches for genomic alignment“

Material: http://galaxyproject.github.io/training-material/topics/sequence-analysis/



Time for a break …

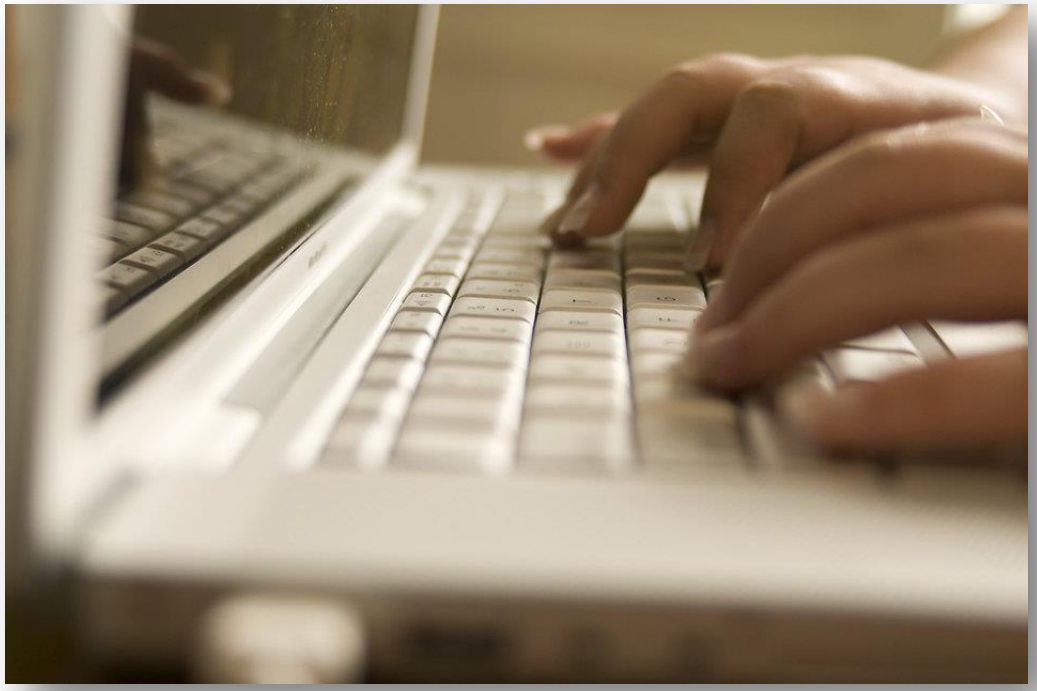
SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

15:15 – 15:45

Hands-on part 4 15:45 – 16:30

“Quantification and DE analysis of RNA-Seq alignments“

Material: Part of https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html#analysis-of-the-differential-gene-expression



Transcript quantification

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

- Read counts
  - Count the reads per feature
    - relatively easy: count the number of reads per gene, exon, …
  - How to handle multi-mapping reads (i.e. reads with multiple alignments)?

- Normalization - aims to make expression levels comparable across:
  - Features (genes, transcripts, isoforms, ...)
  - RNA libraries (samples)
  - Batch (different technologies, vendors, locations)

- Normalization methods:
  - TPM (featureCounts, htseqCount, Kallisto)
  - RPKM / FPKM (Cufflinks /Cuffdiff) (Mortazavi, Nat Meth, 2008)
  - TMM (edgeR) (Robinson & Oshlack, Genome Biol, 2010)
  - DESeq2 (DESeq2) (Love et al., Genome Biol, 2014)

TPM normalization

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

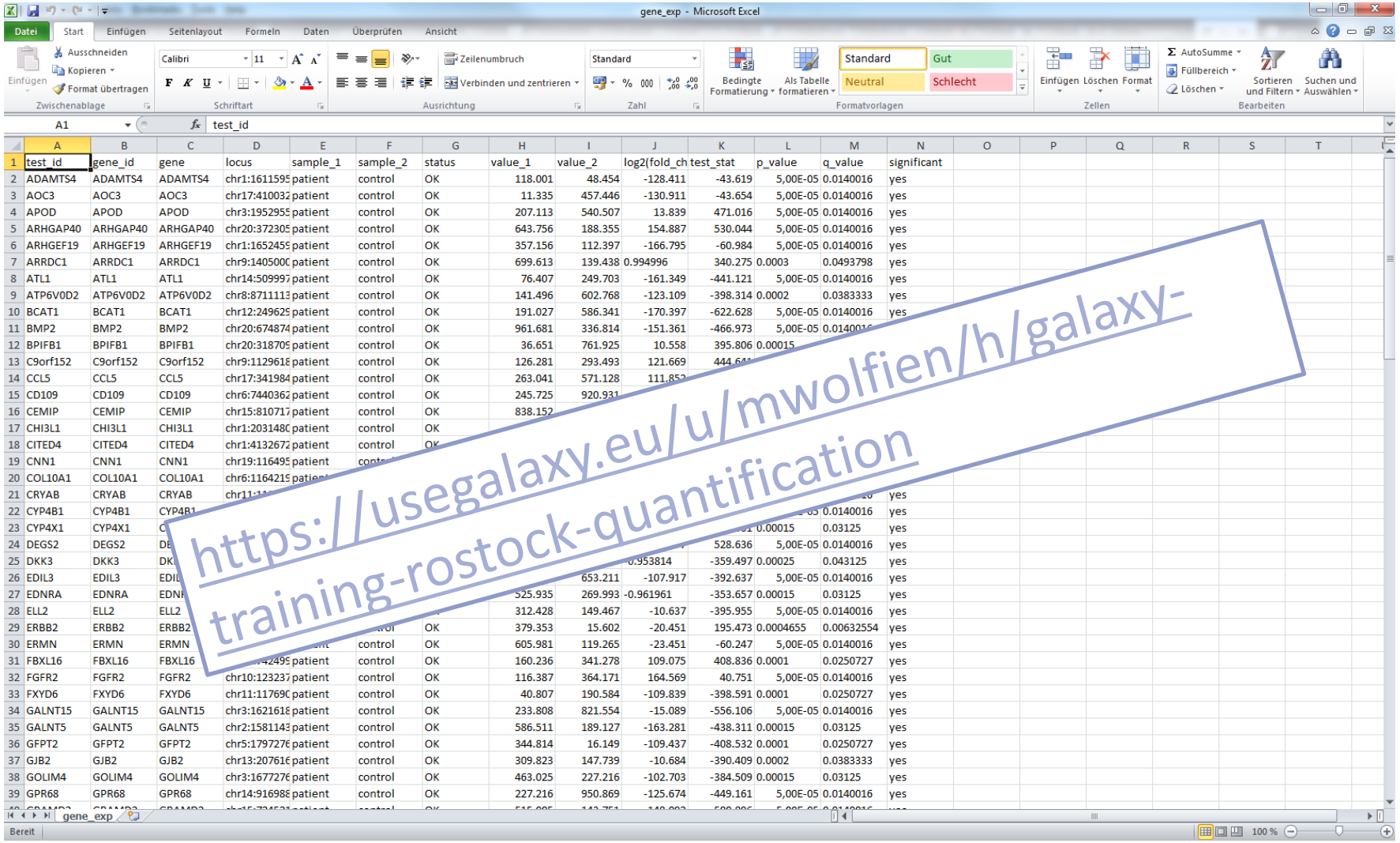
- **T**ranscripts **P**er **M**illion (TPM) is a normalization method for RNA-seq, should be read as "for every 1,000,000 RNA molecules in the RNA-seq sample, *x* came from this gene/transcript."

  1. For each transcript in the gene model, the number (raw count) of reads mapped is divided by the transcript's length, giving a normalized transcript-level expression.
  2. The sum of ALL normalized transcript expression values is divided by 1,000,000, to create a scaling factor.
  3. Each transcript's normalized expression is divided by the scaling factor, which results in the TPM value.
  4. Gene-level TPM's are calculated by summing up the transcript-level TPM for each gene.

- In this scaling, the sum of all TPMs (transcript-level or gene-level) should always equal 1,000,000.
- For a given sample, TPM values will linearly scale with FPKM values for genes or transcripts, but FPKM will not add up to 1,000,000 (http://www.arrayserver.com/wiki/index.php?title=TPM_and_FPKM)
- Differences with and without normalization and differences among them stated at https://www.youtube.com/watch?v=TTUrtCY2k-w

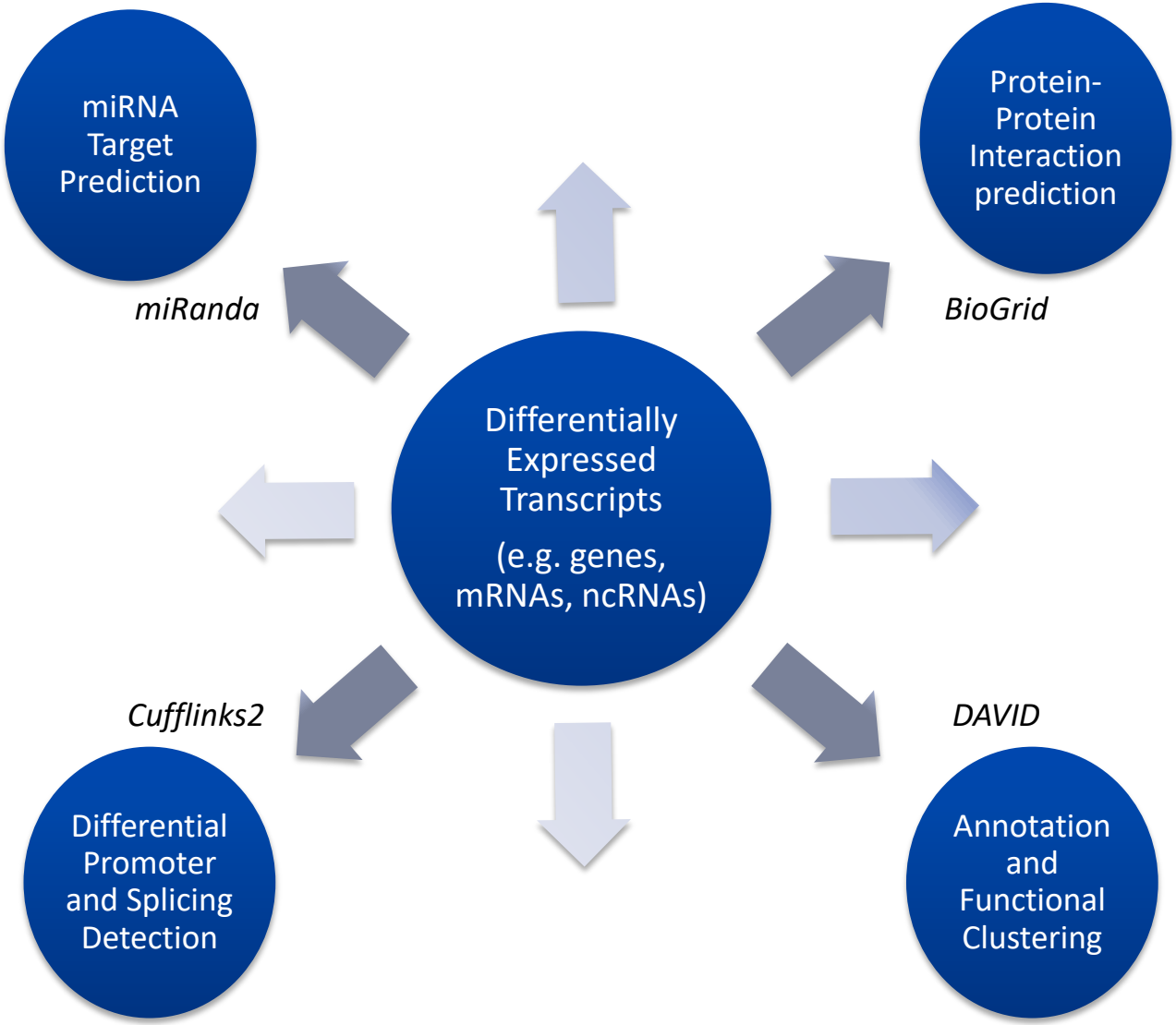
Get your data set!

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK



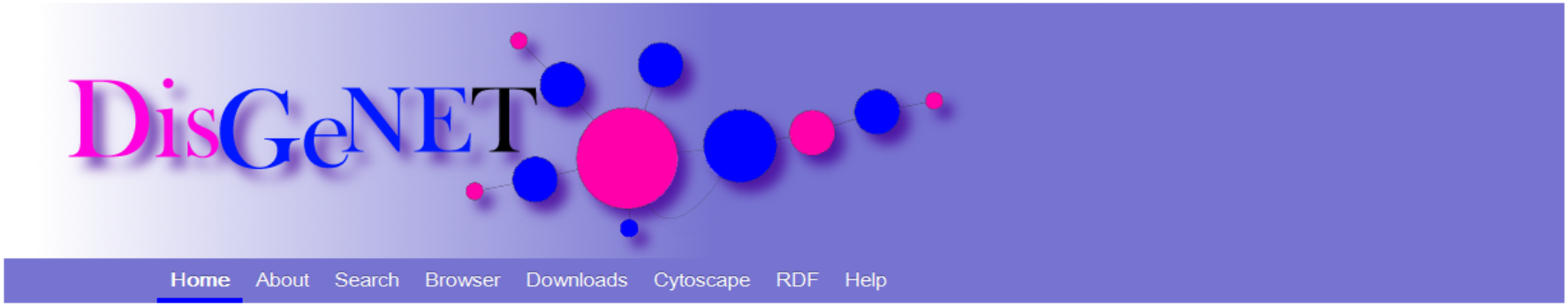
test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_ch	test_stat	p_value	q_value	significant
ADAMTS4	ADAMTS4	ADAMTS4	chr1:1611595	patient	control	OK	118.001	48.454	-128.411	-43.619	5,00E-05	0.0140016	yes
AOC3	AOC3	AOC3	chr17:410032	patient	control	OK	11.335	457.446	-130.911	-43.654	5,00E-05	0.0140016	yes
APOD	APOD	APOD	chr3:1952955	patient	control	OK	207.113	540.507	13.839	471.016	5,00E-05	0.0140016	yes
ARHGAP40	ARHGAP40	ARHGAP40	chr20:372305	patient	control	OK	643.756	188.355	154.887	530.044	5,00E-05	0.0140016	yes
ARHGEF19	ARHGEF19	ARHGEF19	chr1:1652455	patient	control	OK	357.156	112.397	-166.795	-60.984	5,00E-05	0.0140016	yes
ARRDC1	ARRDC1	ARRDC1	chr9:140500C	patient	control	OK	699.613	139.438	0.994996	340.275	0.0003	0.0493798	yes
ATL1	ATL1	ATL1	chr14:509997	patient	control	OK	76.407	249.703	-161.349	-441.121	5,00E-05	0.0140016	yes
ATP6V0D2	ATP6V0D2	ATP6V0D2	chr8:8711113	patient	control	OK	141.496	602.768	-123.109	-398.314	0.0002	0.0383333	yes
BCAT1	BCAT1	BCAT1	chr12:249625	patient	control	OK	191.027	586.341	-170.397	-622.628	5,00E-05	0.0140016	yes
BMP2	BMP2	BMP2	chr20:674874	patient	control	OK	961.681	336.814	-151.361	-466.973	5,00E-05	0.0140016	yes
BPIFB1	BPIFB1	BPIFB1	chr20:318705	patient	control	OK	36.651	761.925	10.558	395.806	0.00015		
C9orf152	C9orf152	C9orf152	chr9:1129618	patient	control	OK	126.281	293.493	121.669	444.641			
CCL5	CCL5	CCL5	chr17:341984	patient	control	OK	263.041	571.128	111.853				
CD109	CD109	CD109	chr6:7440362	patient	control	OK	245.725	920.931					
CEMIP	CEMIP	CEMIP	chr15:810717	patient	control	OK	838.152						
CHI3L1	CHI3L1	CHI3L1	chr1:203148C	patient	control	OK							
CITED4	CITED4	CITED4	chr1:4132672	patient	control	OK							
CNN1	CNN1	CNN1	chr19:116495	patient	control	OK							
COL10A1	COL10A1	COL10A1	chr6:1164215	patient	control	OK							
CRYAB	CRYAB	CRYAB	chr11:11	patient	control	OK							yes
CYP4B1	CYP4B1	CYP4B1										0.0140016	yes
CYP4X1	CYP4X1	C									0.00015	0.03125	yes
DEGS2	DEGS2	DE								528.636	5,00E-05	0.0140016	yes
DKK3	DKK3	DK							0.953814	-359.497	0.00025	0.043125	yes
EDIL3	EDIL3	EDIL						653.211	-107.917	-392.637	5,00E-05	0.0140016	yes
EDNRA	EDNRA	EDNI					525.935	269.993	-0.961961	-353.657	0.00015	0.03125	yes
ELL2	ELL2	ELL2					312.428	149.467	-10.637	-395.955	5,00E-05	0.0140016	yes
ERBB2	ERBB2	ERBB2				OK	379.353	15.602	-20.451	195.473	0.0004655	0.00632554	yes
ERMN	ERMN	ERMN			control	OK	605.981	119.265	-23.451	-60.247	5,00E-05	0.0140016	yes
FBXL16	FBXL16	FBXL16	.42495	patient	control	OK	160.236	341.278	109.075	408.836	0.0001	0.0250727	yes
FGFR2	FGFR2	FGFR2	chr10:123237	patient	control	OK	116.387	364.171	164.569	40.751	5,00E-05	0.0140016	yes
FXYD6	FXYD6	FXYD6	chr11:11769C	patient	control	OK	40.807	190.584	-109.839	-398.591	0.0001	0.0250727	yes
GALNT15	GALNT15	GALNT15	chr3:1621618	patient	control	OK	233.808	821.554	-15.089	-556.106	5,00E-05	0.0140016	yes
GALNT5	GALNT5	GALNT5	chr2:1581143	patient	control	OK	586.511	189.127	-163.281	-438.311	0.00015	0.03125	yes
GFPT2	GFPT2	GFPT2	chr5:179727E	patient	control	OK	344.814	16.149	-109.437	-408.532	0.0001	0.0250727	yes
GJB2	GJB2	GJB2	chr13:207616	patient	control	OK	309.823	147.739	-10.684	-390.409	0.0002	0.0383333	yes
GOLIM4	GOLIM4	GOLIM4	chr3:1677276	patient	control	OK	463.025	227.216	-102.703	-384.509	0.00015	0.03125	yes
GPR68	GPR68	GPR68	chr14:916988	patient	control	OK	227.216	950.869	-125.674	-449.161	5,00E-05	0.0140016	yes

Interconnection of RNA-Seq data



Linking and integrating data

- DisGeNET (http://www.disgenet.org/)



One of the most challenging problems in biomedical research is to understand the underlying mechanisms of complex diseases. Great effort has been spent on finding the genes associated to diseases (Botstein and Risch, 2003; Kann, 2009). However, more and more evidences indicate that most human diseases cannot be attributed to a single gene but arise due to complex interactions among multiple genetic variants and environmental risk factors (Hirschhorn and Daly, 2005). Several databases have been developed storing associations between genes and diseases such as CTD™ (Davis, *et al.*, 2014), OMIM® (Hamosh *et al.*, 2005) and the NHGRI-EBI GWAS catalog (Welter *et al.*, 2014). Each of these databases focuses on different aspects of the phenotype-genotype relationship, and due to the nature of the database curation process, they are not complete. Hence, integration of different databases with information extracted from the literature is needed to allow a comprehensive view of the state of the art knowledge within this research field. With this need in mind, we have created DisGeNET.

DisGeNET is a discovery platform integrating information on gene-disease associations (GDAs) from several public data sources and the literature (Piñero *et al.*, 2015). The current version contains (DisGeNET v4.0) contains 429,036 associations, between 17,381 genes and 15,093 diseases, disorders and clinical or abnormal human phenotypes, and 72,870 variant-disease associations (VDAs), between 46,589 SNPs and 6,356 phenotypes. Given the large number of GDAs compiled in DisGeNET, we have also developed a score in order to rank the associations supporting evidence. Importantly, useful tools have also been created to explore and analyze the data contained in DisGeNET. DisG queried through Search and Browse functionalities available from this web interface, or by a plugin created for Cytoscape to query a network representation of the data. Moreover, DisGeNET data can be queried by downloading the SQLite database to your loc Furthermore, an RDF (Resource Description Framework) representation of DisGeNET database is also available. It can be queried using endpoint and a Faceted Browser. Follow the link for more information.

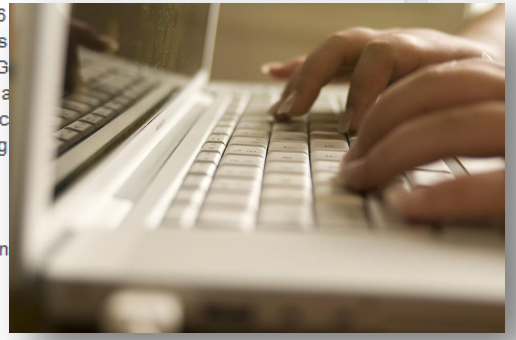
DisGeNET database has been cited by several papers. Some of them can be reviewed here.

The DisGeNET database is made available under the Open Database License. Any rights in individual contents of the database are licen Database Contents License.

Tweets by @DisGeNET

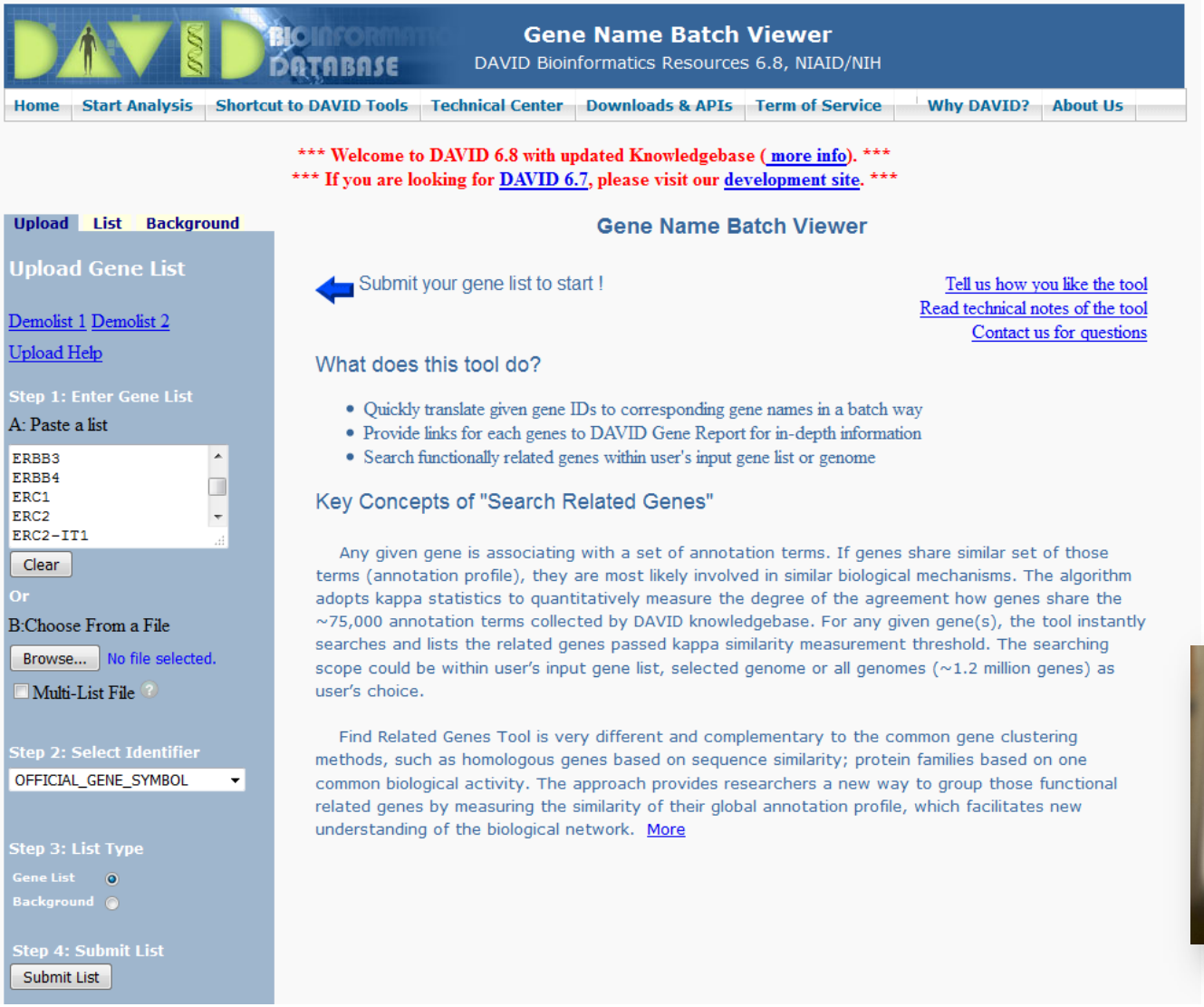
DisGeNET
@DisGeNET

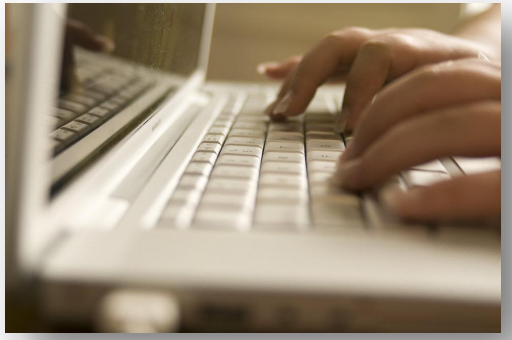
Check out the new publication describing the DisGeNET platfotm in NAR database issue nar.oxfordjournals.org/con



Linking and integrating data

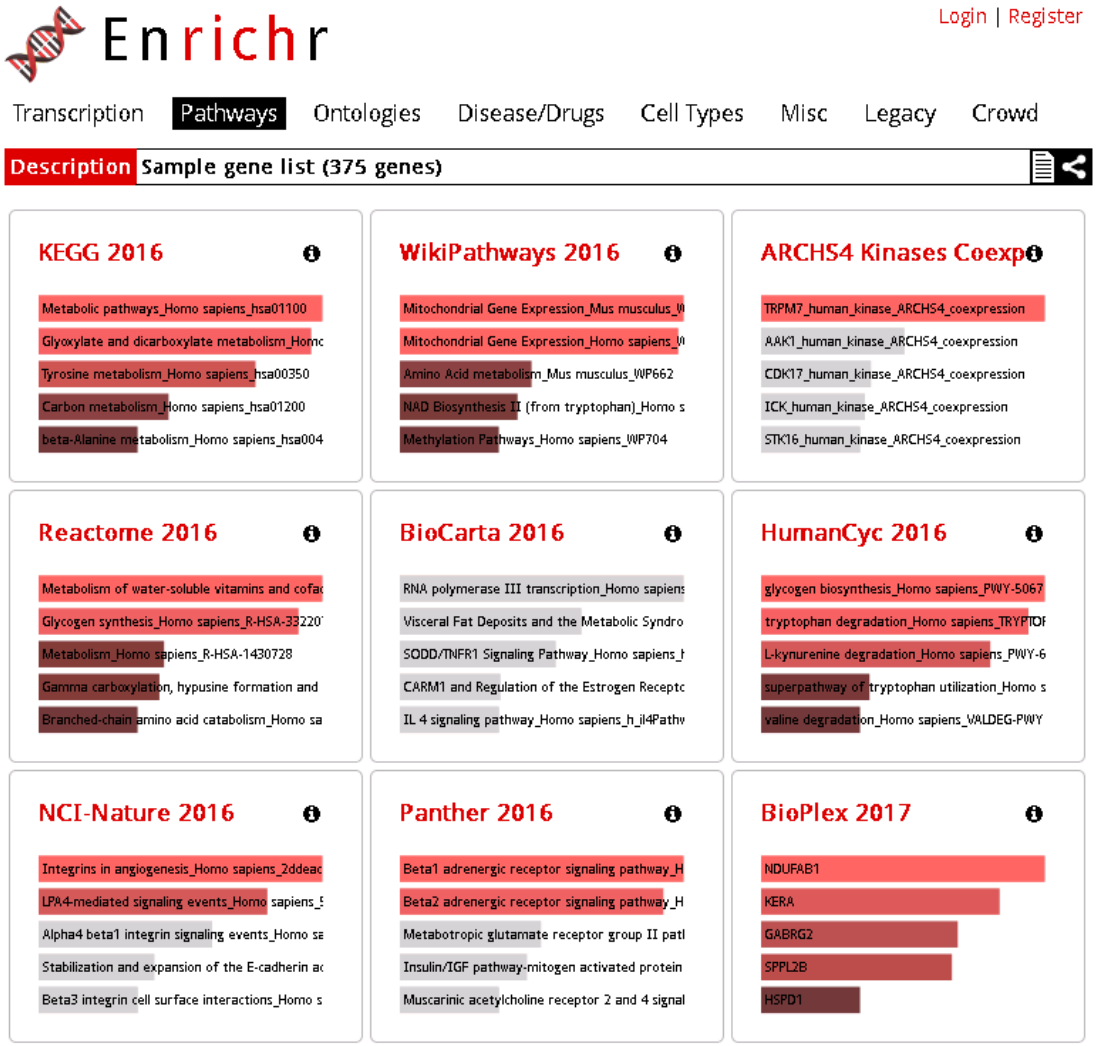
- David (https://david.ncifcrf.gov/list.jsp)

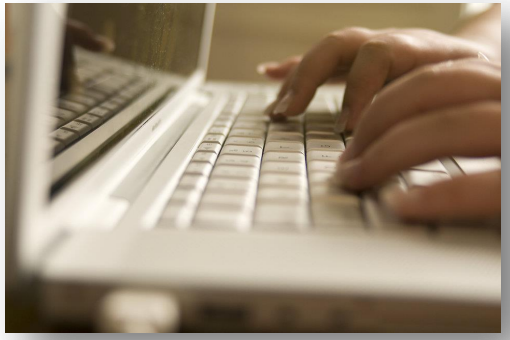




Linking and integrating data

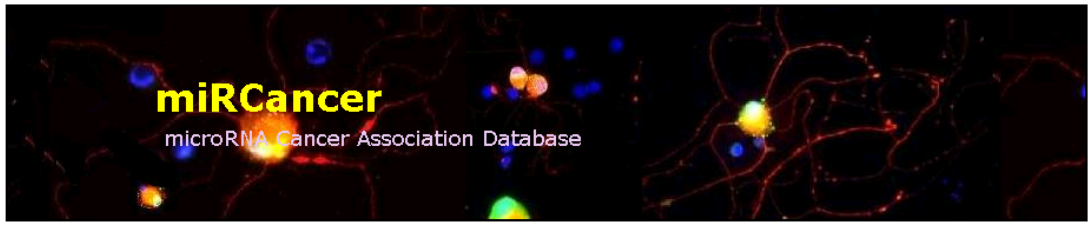
- Enrichr (http://amp.pharm.mssm.edu/Enrichr/)





Linking and integrating data

- miRCancer db - Find up regulated miRNAs (http://mircancer.ecu.edu/index.jsp)



Search miRCancer | Browse miRCancer | Sequence Analysis | Download | Help | About Us

Search

Search for miRNA and cancer associations

miRNA name
mir-145
Example: mir-145, hsa-mir-21, let-7

Or And

cancer name
Example: lung, breast cancer

Search

Mailing List

To get annoucements about miRCancer database update, new avaliable functionality and more, please subscribe to our mailing list.

Links

miRCancer : microRNA Cancer Association Database

miRCancer provides comprehensive collection of microRNA (miRNA) expression profiles in various human cancers which are automatically extracted from published literatures in PubMed. It utilizes text mining techniques for information collection. Manual revision is applied after auto-extraction to provide 100% precision.

User can search the database by miRNA and/or cancer names in the miRCancer Search page. Our website also provides two sequence analysis tools: clustering and chi-square analysis which can perform analysis on all or selected pool of miRNA sequences.

Reference

If you make use of the information presented here, please cite the following references:

miRCancer: a microRNA-cancer association database constructed by text mining on literature
Boya Xie; Qin Ding; Hongjin Han; Di Wu
Bioinformatics, Vol. 29, Issue 5, pp.638-644, 2013

Text Mining on Big and Complex Biomedical Literature, Big Data Analytics in Bioinformatics and Healthcare, IGI Global, 2014.
Boya Xie; Qin Ding; Di Wu

MIRSAT & MIRCDB: An Integrated microRNA Sequence Analysis Tool and a Cancer-associated microRNA Database
Boya Xie; Robert Hochberg; Qin Ding; Di Wu
International Conference on Bioinformatics and Computational Biology, Honolulu, Hawaii, 2010, pp. 159-164.

Latest Counts

miRNA: 57984
Cancer: 196
miR-Cancer: 7325
Paper: 5723

Update: February 18, 2019

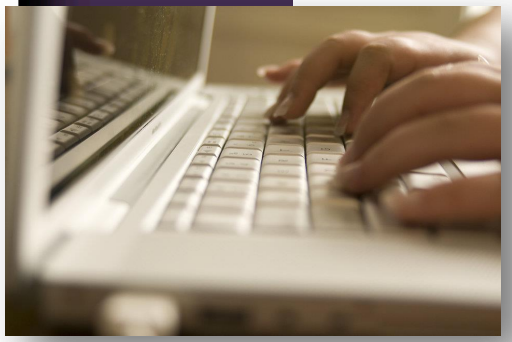
News

According to Google Scholar, By 2019 February, our journal paper has been cited 242 times.

The latest miRCancer data are updated with:

Literatures from PubMed (queried on June 30, 2018)

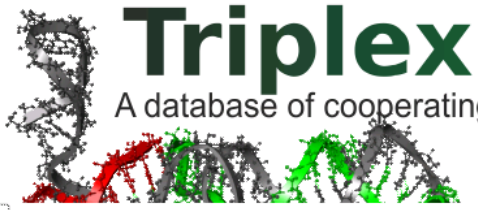
miRNA from miRBase (Release 22)



Explore miRNA cooperativity to justify diagnosis

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

- TriplexRNA database (https://www.sbi.uni-rostock.de/triplexrna/)



TriplexRNA

A database of cooperating microRNAs and their mutual targets

What is the TriplexRNA database?

Description

The triplexRNA database contains predicted RNA triplexes composed of two cooperatively acting microRNAs (miRNAs) and their mutual target mRNAs.
The phenomenon of miRNAs that cooperatively repress target gene transcription is illustrated below and elucidated in detail in our article Schmitz et al. (2014).

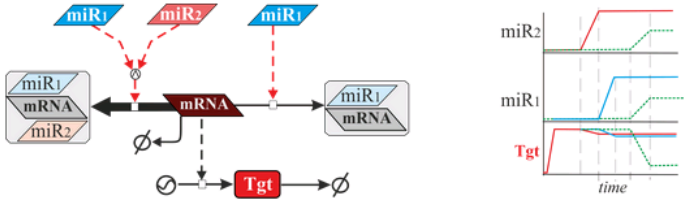
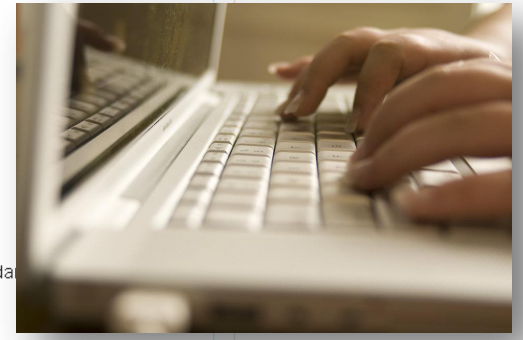


Figure 1 General principle of cooperative target regulations by pairs of miRNAs
The illustration on the left shows how a target mRNA can be repressed by either a single miRNA or by a pair of cooperating miRNAs. While in the first case miRNA and target form a duplex structure, the second case leads to the formation of a RNA triplex. On the right side we illustrated the repressive effect on the target that is induced either by a single miRNA (red and blue lines) or by two cooperating miRNAs (green dashed line). Even if the expression of the cooperating miRNAs is only mildly up-regulated an enhanced repressive effect can be observed as compared to the cases where single miRNAs are highly up-regulated.

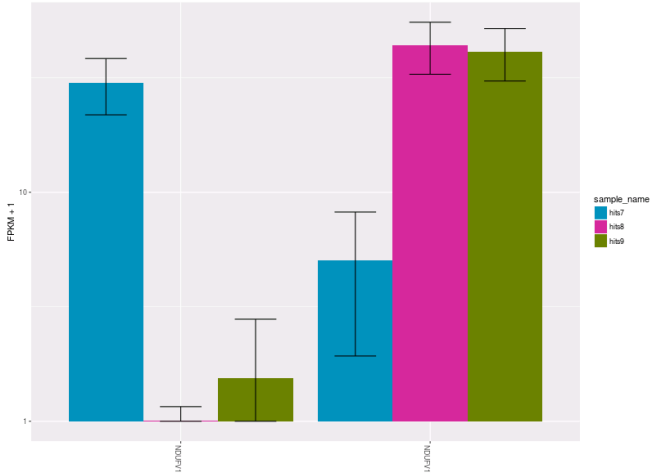
The following information about predicted RNA triplexes can be retrieved from the TriplexRNA database: triplex free energies, experimental evidences, seconda
simulations of target gene repression and more...

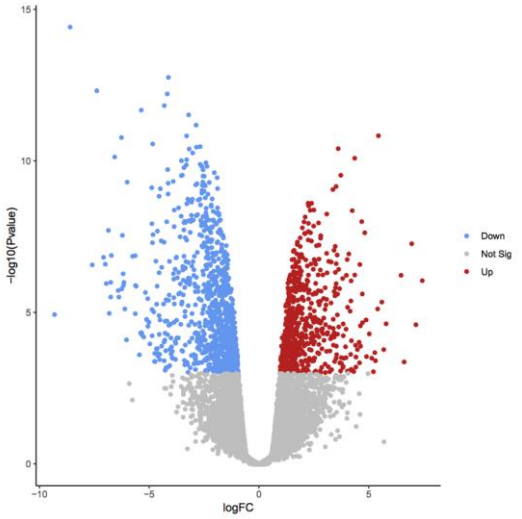


Further visualizations available in Galaxy

- Available at: https://galaxyproject.github.io/training-material/topics/transcriptomics/

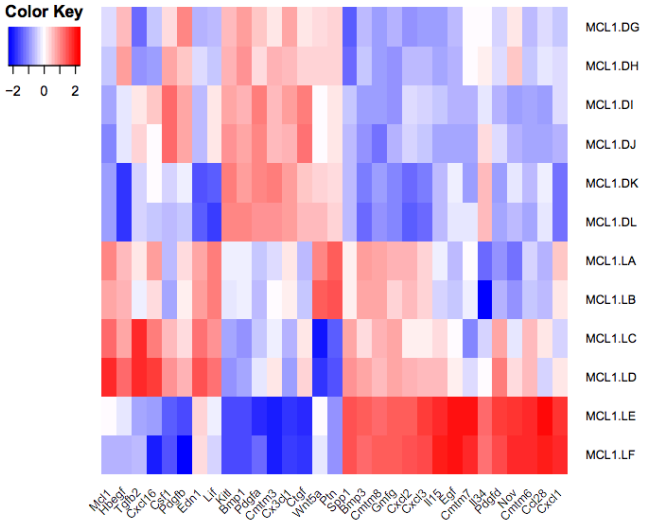
CummeRbund





Volcano plots

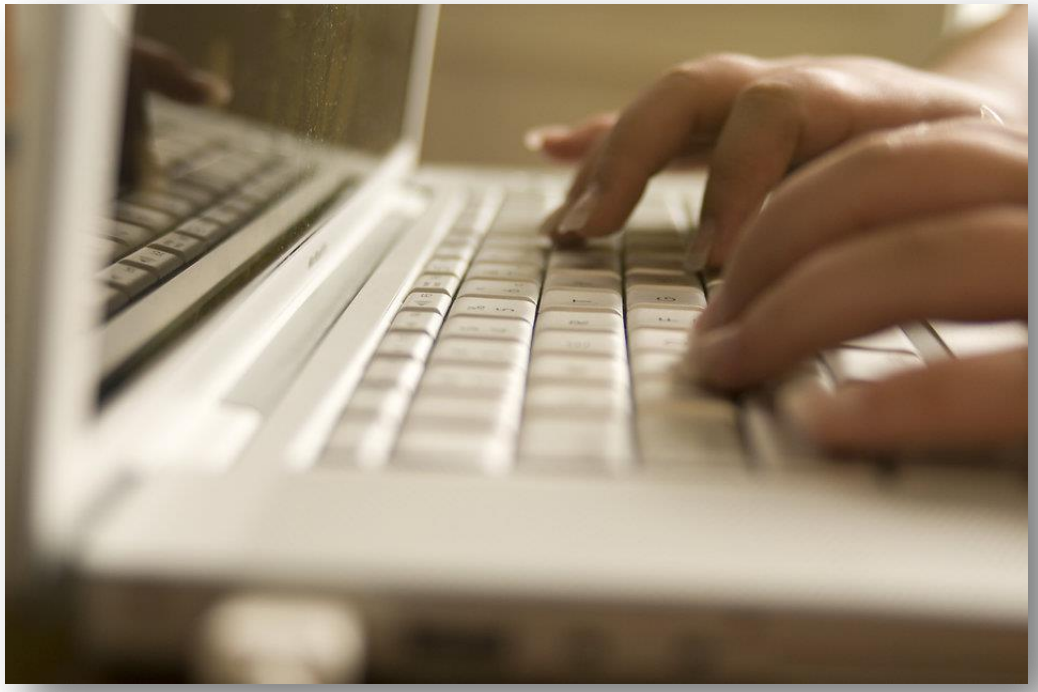
Heatmaps



Hands-on part 6 16:30 – 17:15

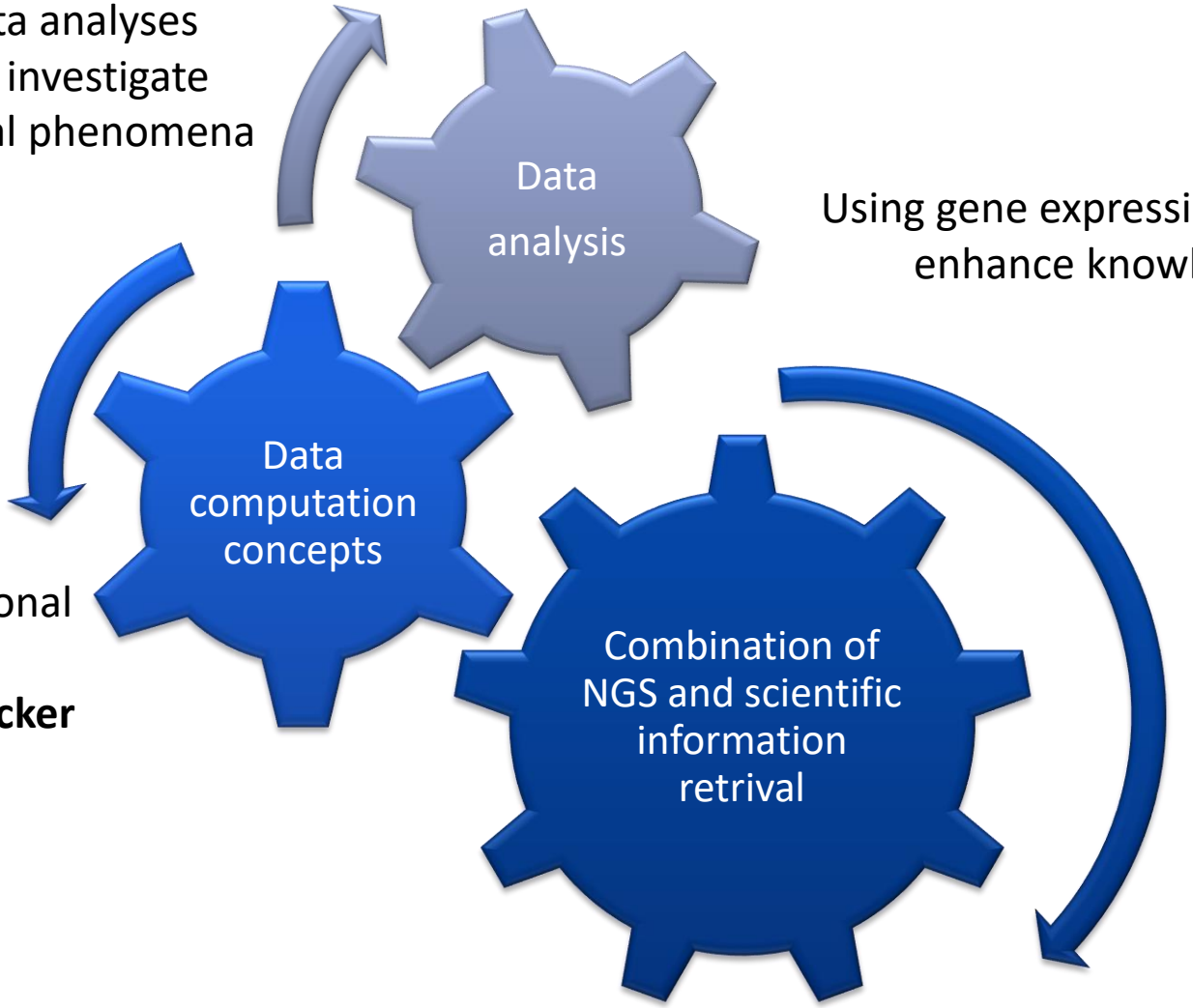
“Galaxy modular workflow generator“

Material: https://github.com/destairdenbi/galaxy-modular-workflow-generator



What did we learn so far?

Using NGS data analyses approaches to investigate various biological phenomena



Using gene expression data to enhance knowledge

Utilizing computational technologies, like **Galaxy** and **Docker**

Acknowledgements

Olaf Wolkenhauer (University of Rostock)
Wolfgang Hess (University of Freiburg)
Steve Hoffmann (University of Leipzig)

Rolf Backofen (University of Freiburg)
Björn Grüning (University of Freiburg)

Supported by:

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
denbi.de

elixir GERMANY
elixir-europe.org

Bundesministerium für Bildung und Forschung
bmbf.de

EUROPÄISCHE UNION
Europäischer Sozialfonds
Europäische Fonds EFRE, ESF und ELER in Mecklenburg-Vorpommern 2014-2020

European Social Fund (ESF) program of the European Union (ESF/14-BM-A55-0027).