Bioinformatics carpentry - Transcriptomics

Markus Wolfien

markus.wolfien@uni-rostock.de

Galaxy Training – 14th April 2021

www.sbi.uni-rostock.de

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

Our schedule for today

- Q & A from Yesterday
- Introduction
  - General introduction to transcriptomics
  - Choosing the correct technology
  - Basic data analysis principles
  - Trainee-specific requests
- Hands-on (joint)
  - Quality control of fastq files
  - RNA-Seq mapping algorithms
  - Quantification of alignment files
- Hands-on (individual)
  - Further Hands on time (individual)

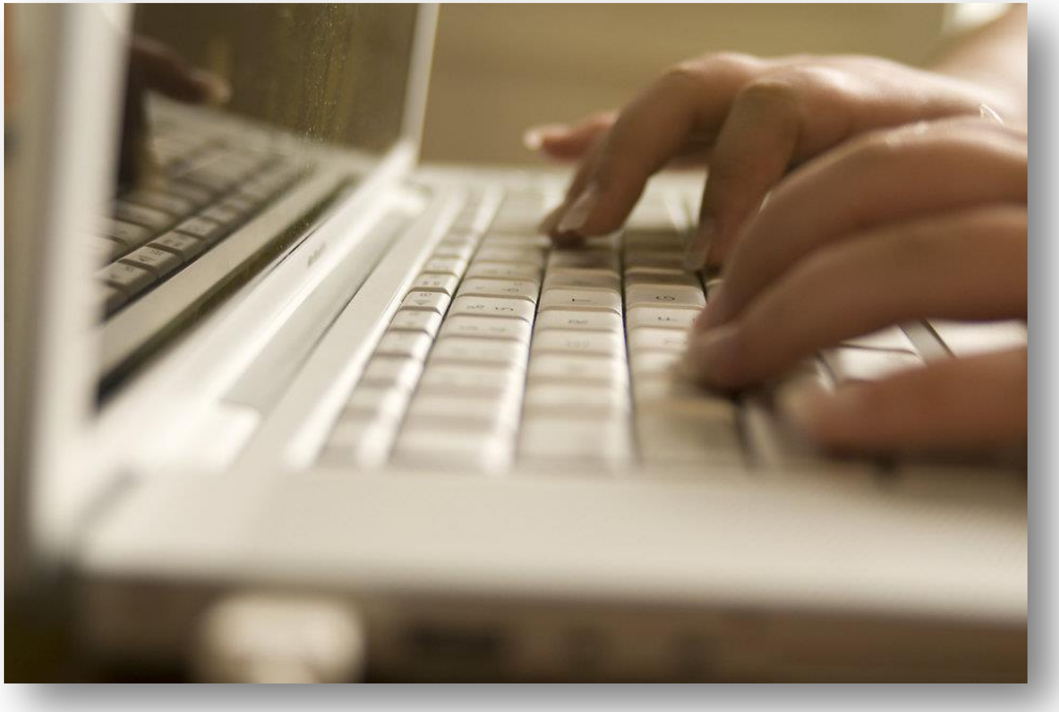
Gene expression theory

Data analysis procedures

Analyzing RNA-Seq data with Galaxy

Go and get the slides!

Schedule & Slides at:
https://github.com/destairdenbi/trainings



Transcription from genes to a functional gene products

The Role of Genetic Expression

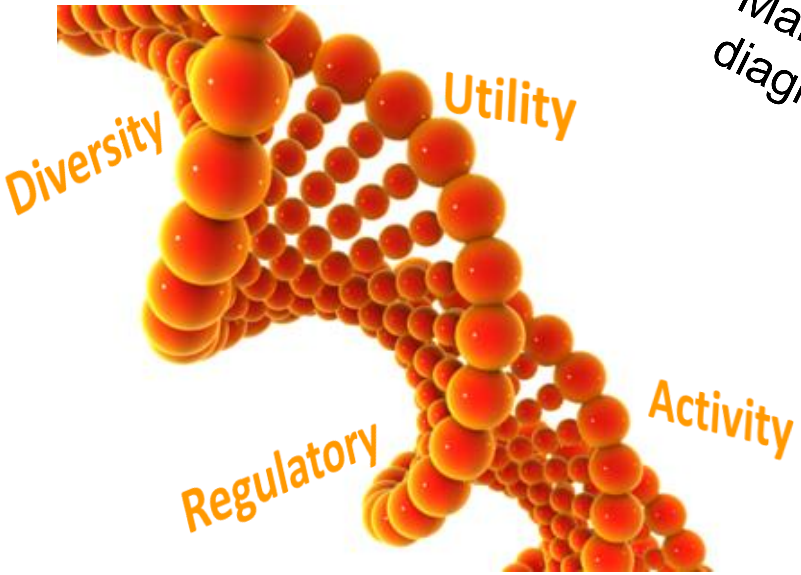
Many different variations and subtypes

Information about regulatory mechanisms

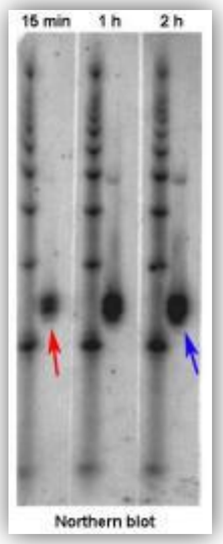
Active and measurable state of the cell …

… ,but only a snapshot

Many different therapeutical and diagnostical approaches

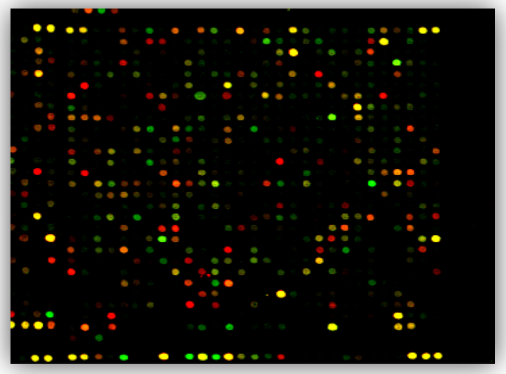


Measureing gene expression



Northern Blot

Reverse Transcription PCR



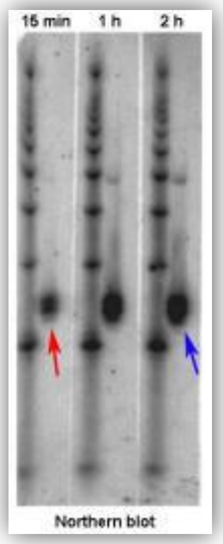
Microarrays

Microarrys are still good and useful, e.g.,

- Quantify known transcripts, isoforms
- Investigate pathway activity (small assay‘s less than 150 USD)
- Less sample amount needed
- Less data intense and computational resource intense

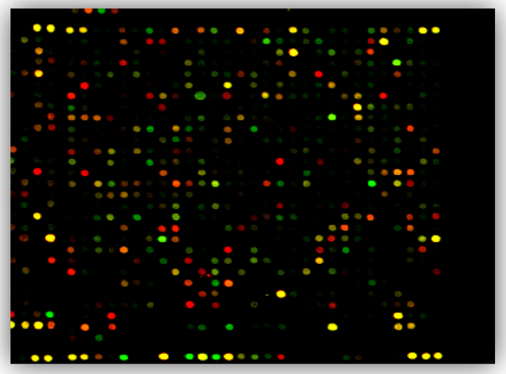
… but

Measureing gene expression

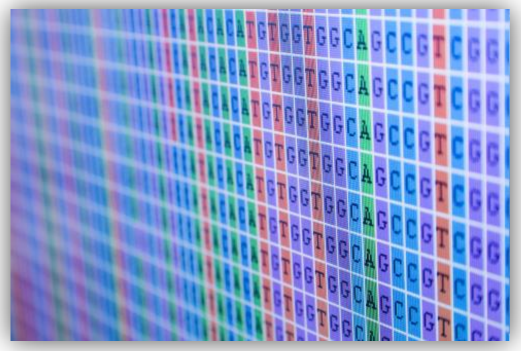


Northern Blot

Reverse Transcription PCR

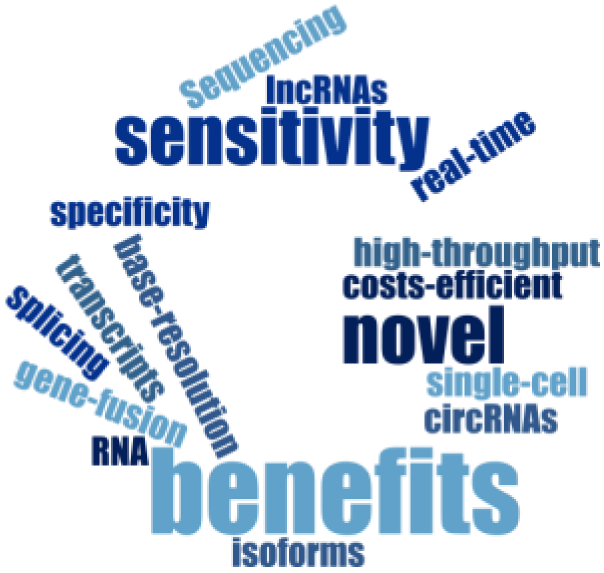


Microarrays



NGS

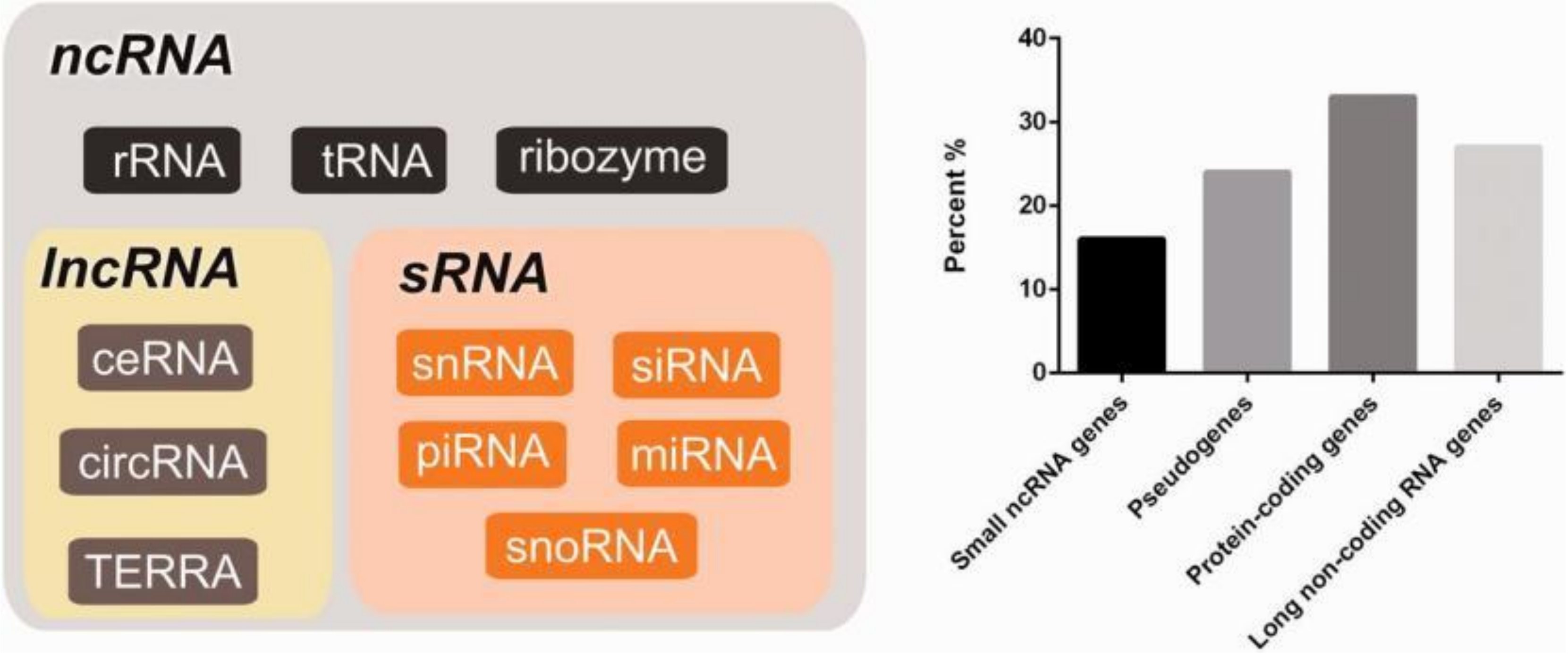
Why do we need NGS?



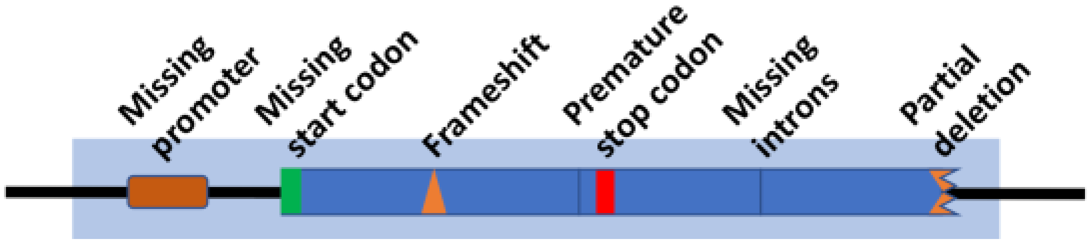
"RNA-Seq is able to identify thousands of differentially expressed genes, tens of thousands of differentially expressed gene isoforms, and can detect mutations and germline variations for hundreds to thousands of expressed genetic variants, as well as detecting chimeric gene fusions, transcript isoforms,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

Wang, *Nat Rev. Genet*., 2009

Classes of ncRNAs

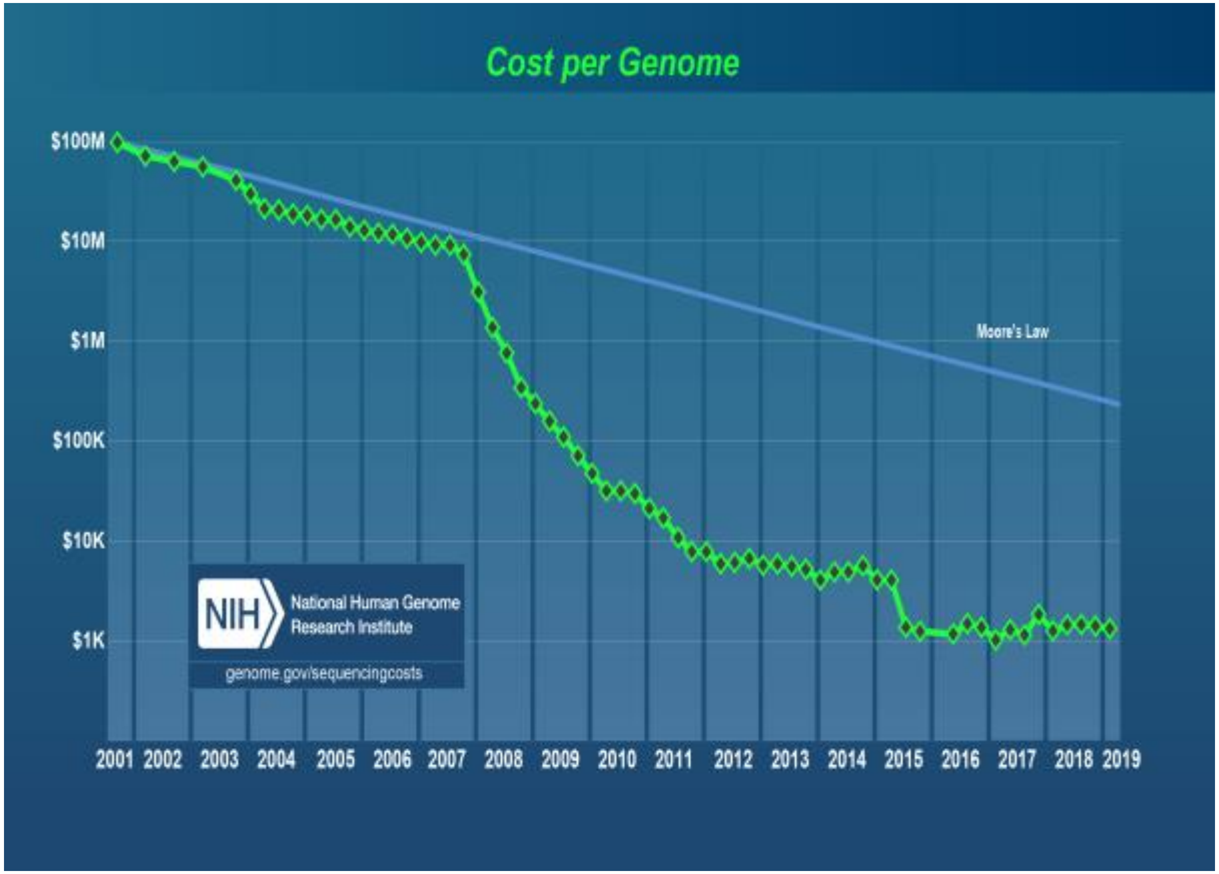


Common defects of pseudogenes:

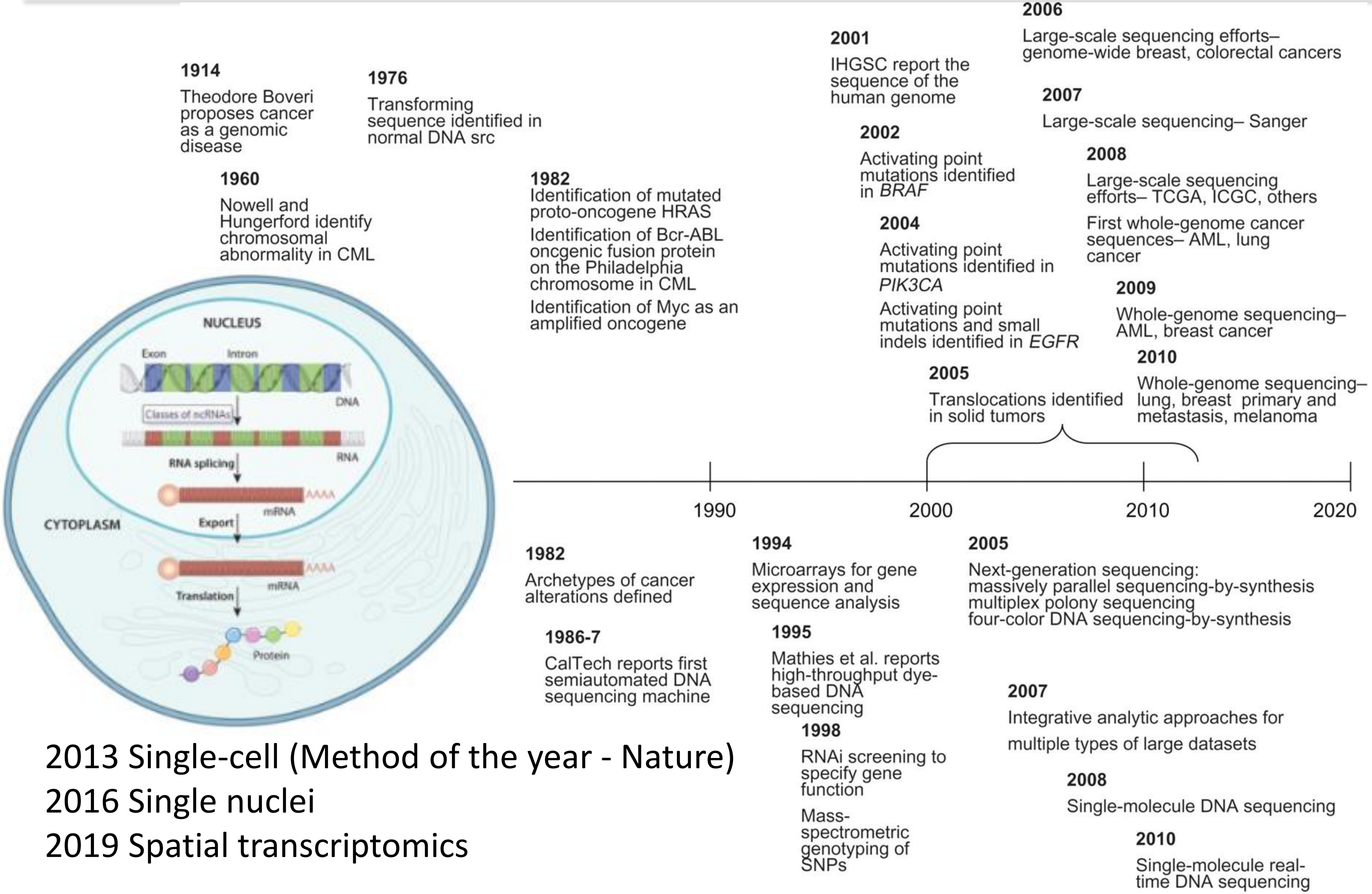


Schmitz, *Brief Bioinform*. 2016

Technical advances lead the way



Evolution of sequencing technologies



1914
Theodore Boveri proposes cancer as a genomic disease

1960
Nowell and Hungerford identify chromosomal abnormality in CML

1976
Transforming sequence identified in normal DNA src

1982
Identification of mutated proto-oncogene HRAS

Identification of Bcr-ABL oncgenic fusion protein on the Philadelphia chromosome in CML

Identification of Myc as an amplified oncogene

2001
IHGSC report the sequence of the human genome

2002
Activating point mutations identified in *BRAF*

2004
Activating point mutations identified in *PIK3CA*

Activating point mutations and small indels identified in *EGFR*

2005
Translocations identified in solid tumors

2006
Large-scale sequencing efforts– genome-wide breast, colorectal cancers

2007
Large-scale sequencing– Sanger

2008
Large-scale sequencing efforts– TCGA, ICGC, others

First whole-genome cancer sequences– AML, lung cancer

2009
Whole-genome sequencing– AML, breast cancer

2010
Whole-genome sequencing– lung, breast primary and metastasis, melanoma

1982
Archetypes of cancer alterations defined

1986-7
CalTech reports first semiautomated DNA sequencing machine

1994
Microarrays for gene expression and sequence analysis

1995
Mathies et al. reports high-throughput dye-based DNA sequencing

1998
RNAi screening to specify gene function

Mass-spectrometric genotyping of SNPs

2005
Next-generation sequencing:
massively parallel sequencing-by-synthesis
multiplex polony sequencing
four-color DNA sequencing-by-synthesis

2007
Integrative analytic approaches for multiple types of large datasets

2008
Single-molecule DNA sequencing

2010
Single-molecule real-time DNA sequencing

2013 Single-cell (Method of the year - Nature)
2016 Single nuclei
2019 Spatial transcriptomics

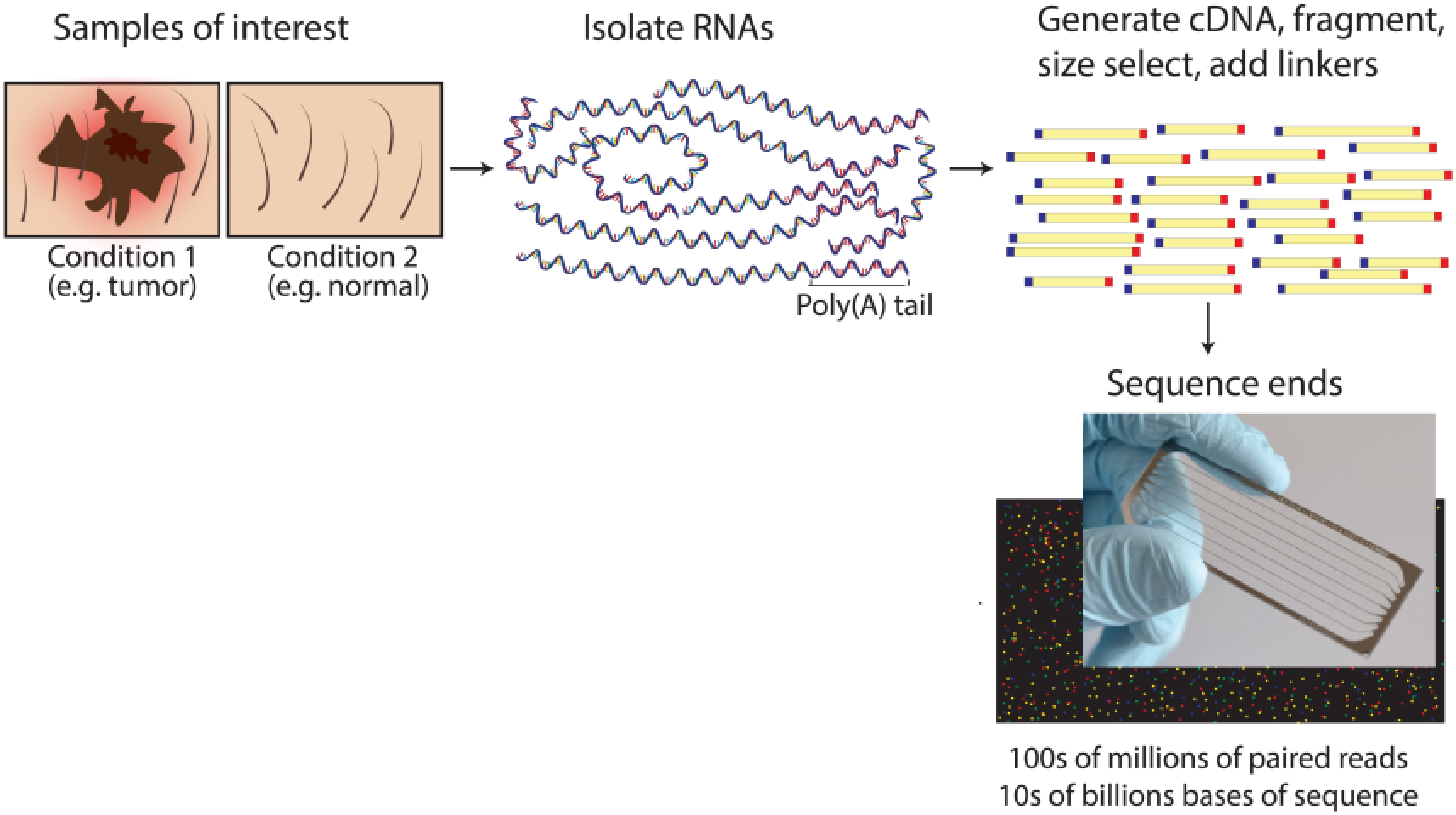
Sequencing type comparison

Common
“Bulk“ RNA-sequencing

Single-cell RNA-sequencing

Spatial transcriptomics

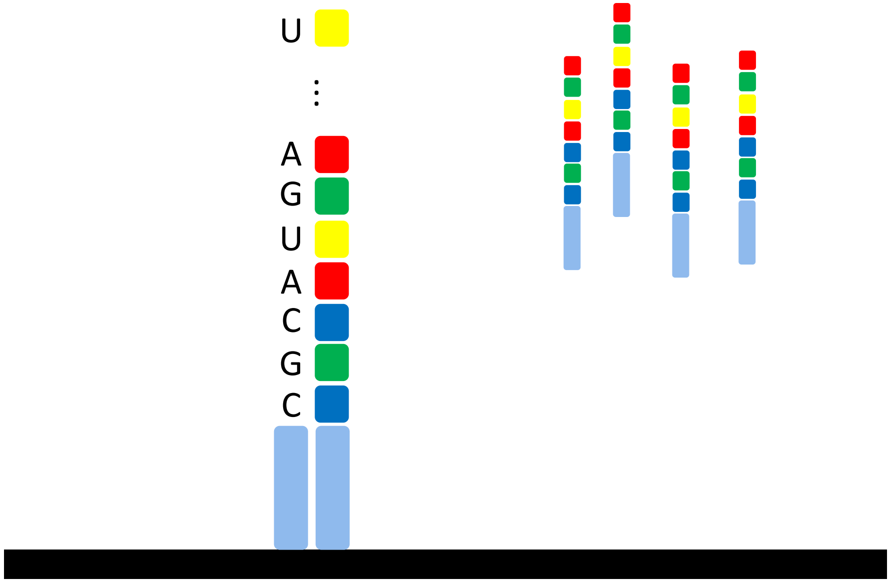
From sample to readout



Griffith, Plos Comp. Biol., 2015

How NGS works - brief

Example: Sequencing



>25 *10^6 Sequences

RNA-Sequencing

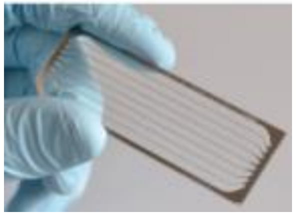
RNA sample of the patient

Adapter

Flow cell

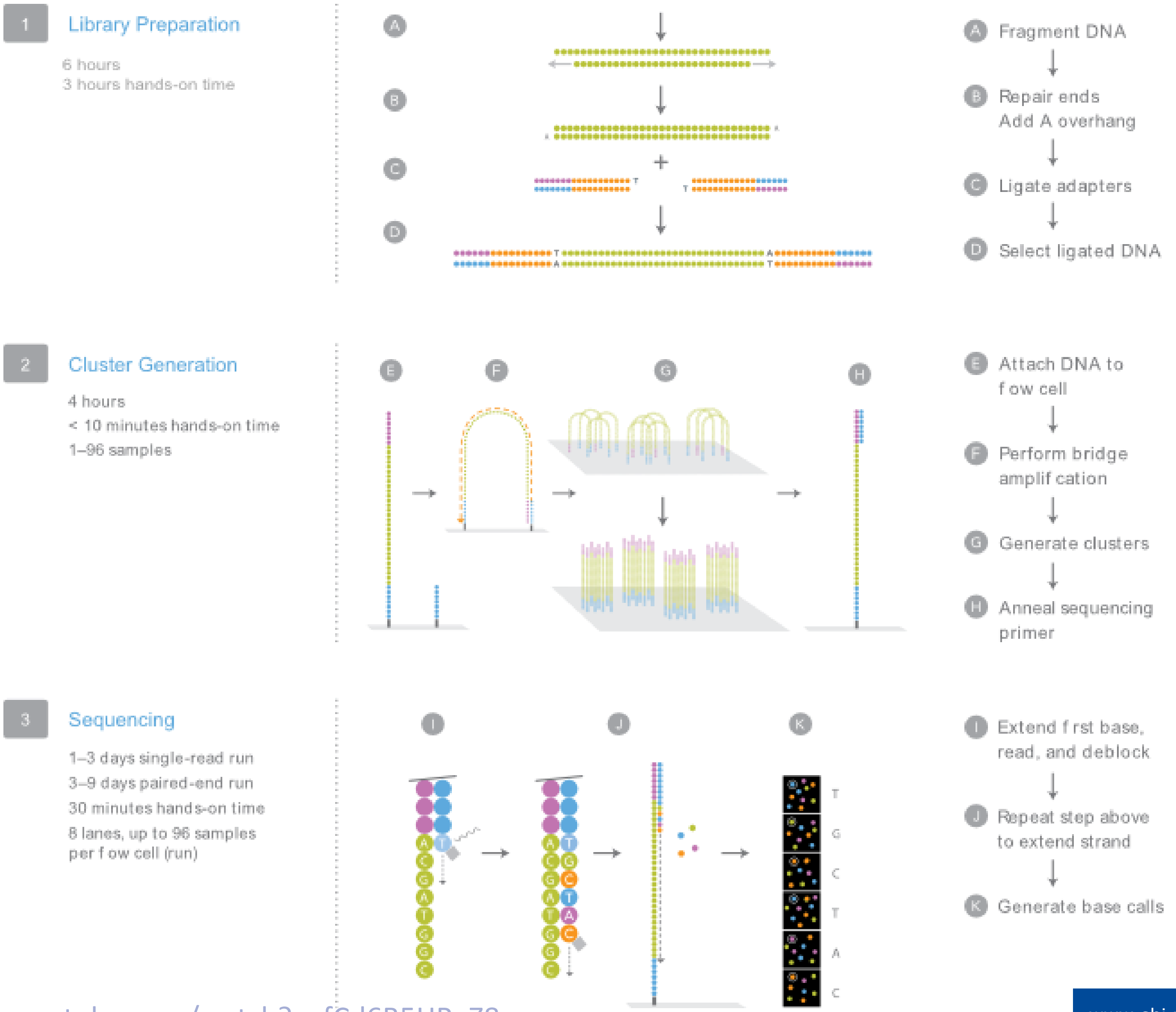
A – U

G – C



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 flow cell
- F Perform bridge amplification
- G Generate clusters
- H Anneal sequencing primer

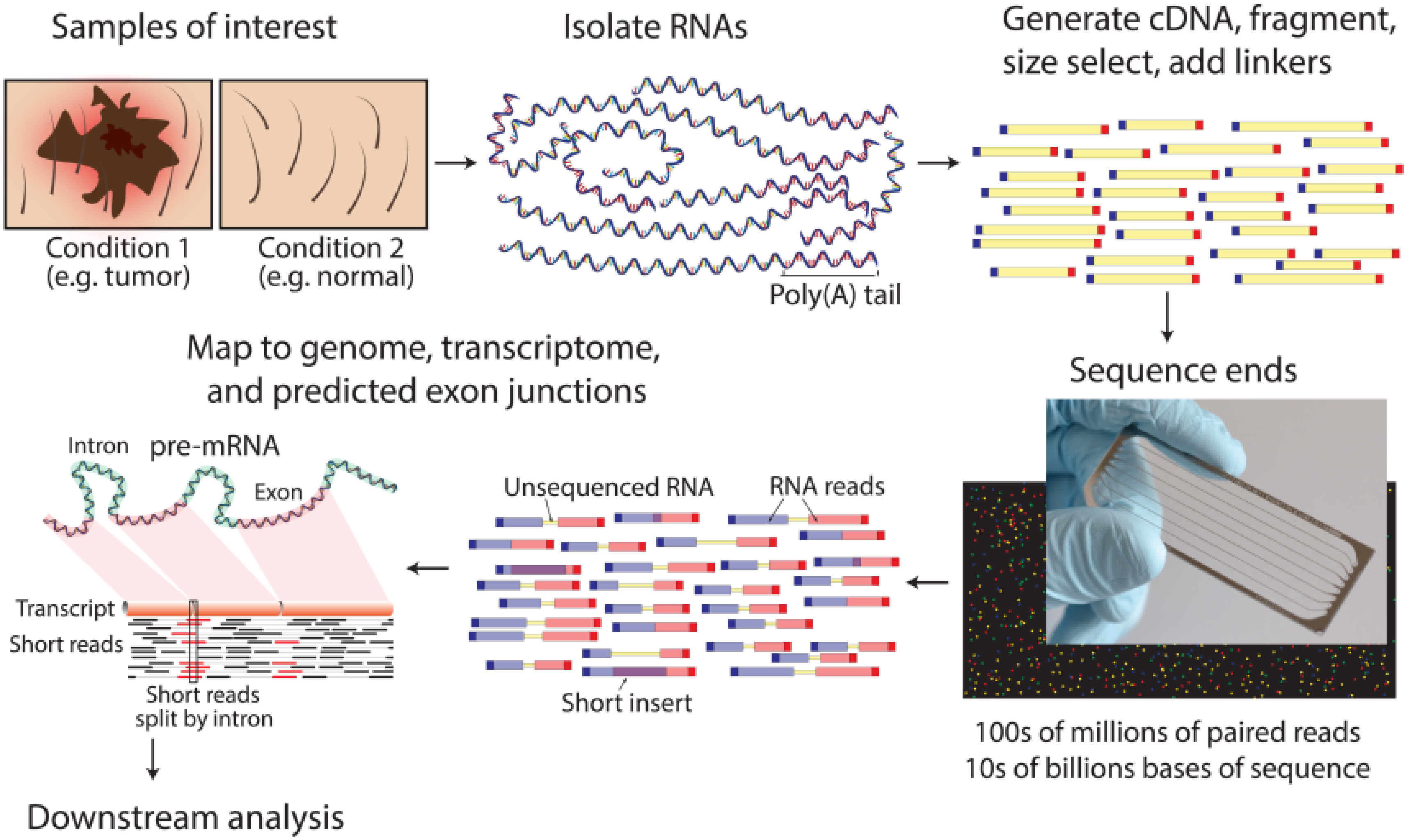
3 Sequencing

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

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

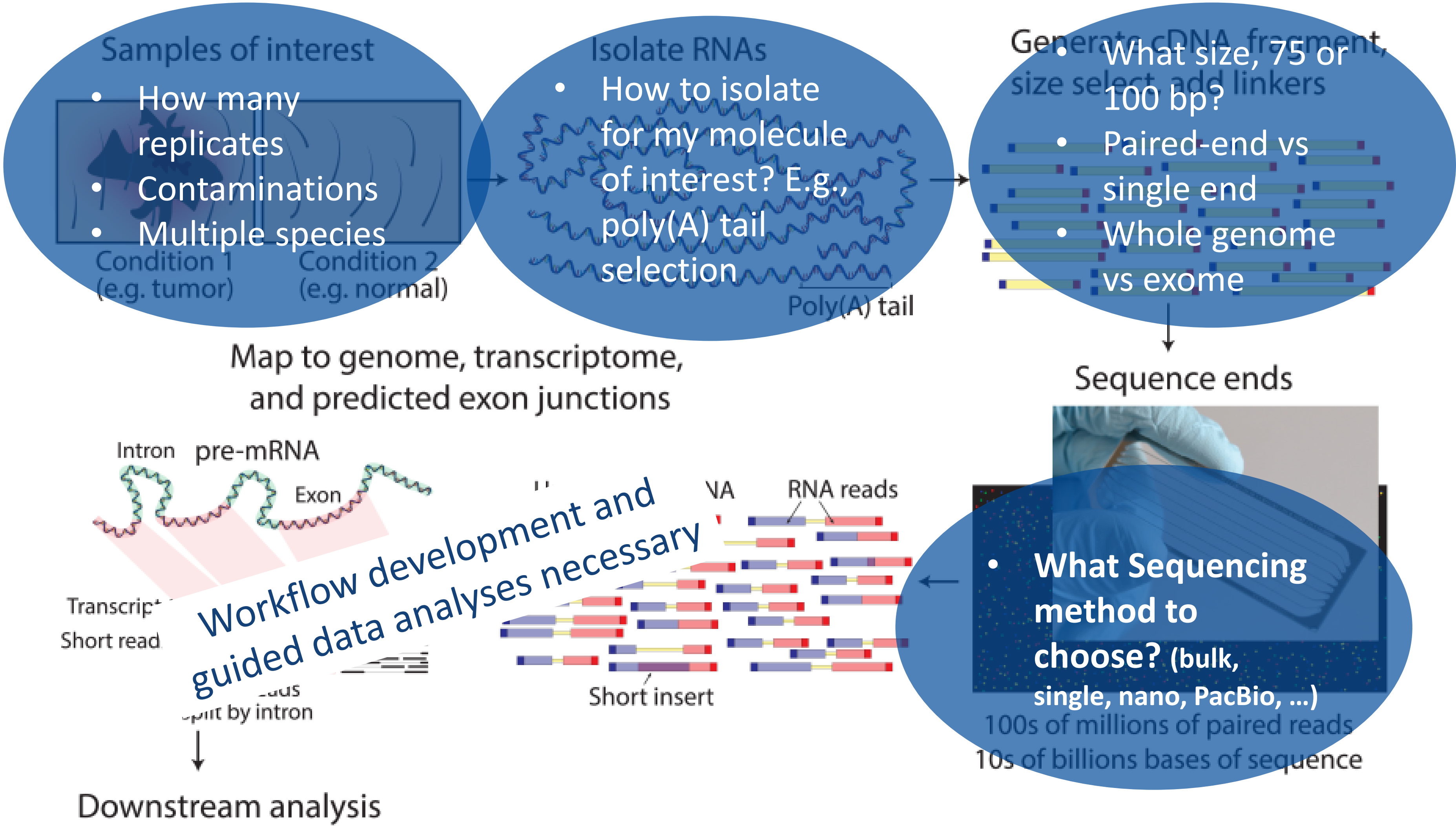
https://www.youtube.com/watch?v=fCd6B5HRaZ8

From sample to readout


Samples of interest
Condition 1
(e.g. tumor)
Condition 2
(e.g. normal)
Isolate RNAs
Poly(A) tail
Generate cDNA, fragment, size select, add linkers
Sequence ends
100s of millions of paired reads
10s of billions bases of sequence
Unsequenced RNA
RNA reads
Short insert
Map to genome, transcriptome, and predicted exon junctions
Intron
pre-mRNA
Exon
Transcript
Short reads
Short reads split by intron
Downstream analysis

Griffith, Plos Comp. Biol., 2015

From sample to readout



- How many replicates
- Contaminations
- Multiple species

- How to isolate for my molecule of interest? E.g., poly(A) tail selection

- What size, 75 or 100 bp?
- Paired-end vs single end
- Whole genome vs exome

- **What Sequencing method to choose? (bulk, single, nano, PacBio, …)**

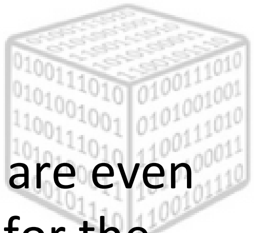
Workflow development and guided data analyses necessary

Griffith, Plos Comp. Biol., 2015

Data analysis

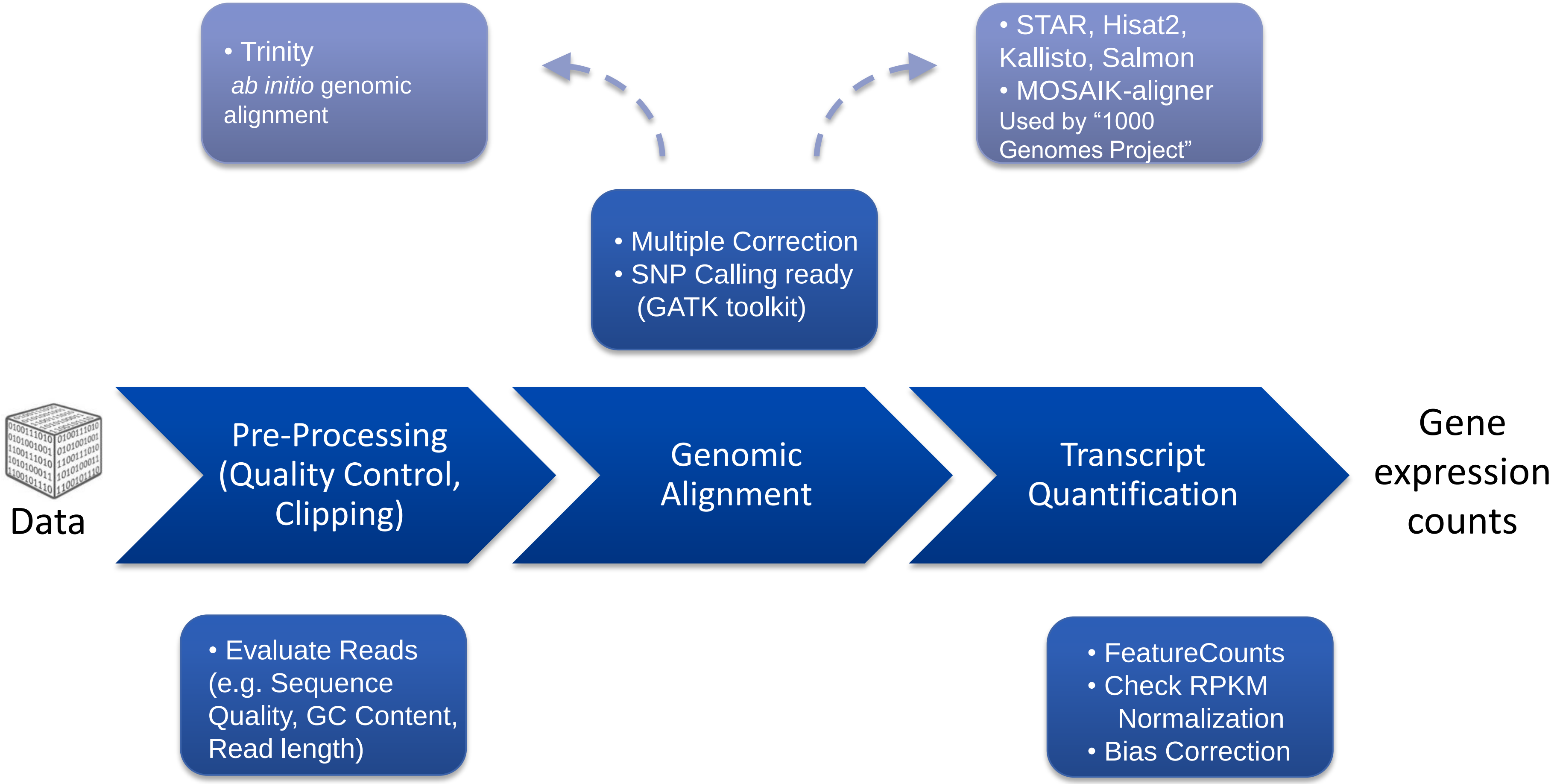
Where do I get NGS data?

- Databases, 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.
  - Galaxy histories, e.g., covid19 specific RNA-Seq data - https://covid19.usegalaxy.eu/u/nekrut/h/rnaseq



- 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)

Basic workflow for differential expression analysis



- Trinity

ab initio genomic alignment

- STAR, Hisat2, Kallisto, Salmon
- MOSAIK-aligner

Used by “1000 Genomes Project”

- Multiple Correction
- SNP Calling ready (GATK toolkit)

Data

Pre-Processing (Quality Control, Clipping)

Genomic Alignment

Transcript Quantification

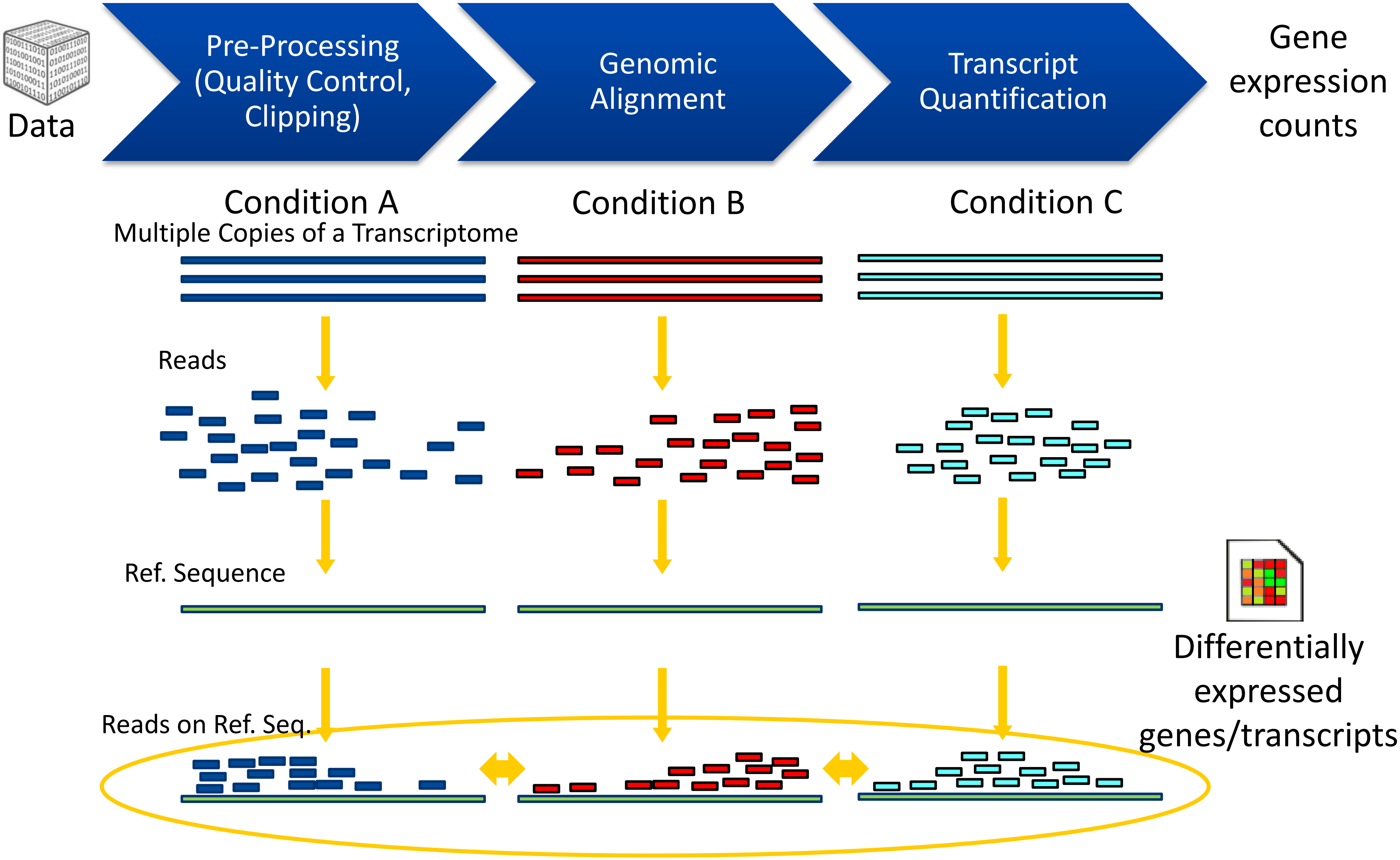
Gene expression counts

- Evaluate Reads (e.g. Sequence Quality, GC Content, Read length)

- FeatureCounts
- Check RPKM Normalization
- Bias Correction

Basic workflow for differential expression analysis

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK



Data

Pre-Processing (Quality Control, Clipping)

Genomic Alignment

Transcript Quantification

Gene expression counts

Condition A

Multiple Copies of a Transcriptome

Condition B

Condition C

Reads

Ref. Sequence

Reads on Ref. Seq.

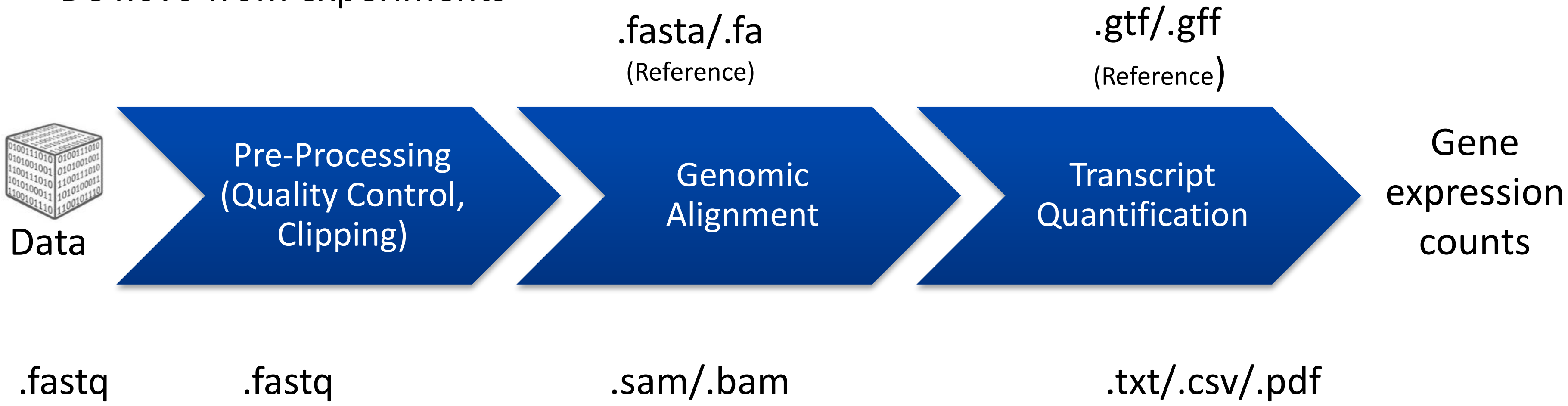
Differentially expressed genes/transcripts

Differential expression (DESeq2, Sleuth)

Sequencing data formats

Obtain references from:

- Galaxy (build-in)
- Ensembl
- UCSC
- NCBI
- *De novo* from experiments



.fasta/.fa
(Reference)

.gtf/.gff
(Reference)

Data

Pre-Processing (Quality Control, Clipping)

Genomic Alignment

Transcript Quantification

Gene expression counts

.fastq

.fastq

.sam/.bam

.txt/.csv/.pdf

Big Data and the need for new analyses

GenePattern

broadinstitute.org

GeneProf

geneprof.org

Grape

big.crg.cat/services/grape

Open for Innovation KNIME

knime.org

mapman.gabipd.org

Galaxy

usegalaxy.org

python™

python.org

R

r-project.org/

Chipster

Open source platform for data analysis

chipster.csc.fi

GeneTalk

gene-talk.de

BaseSpace®

illumina.com

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

bioconductor.org

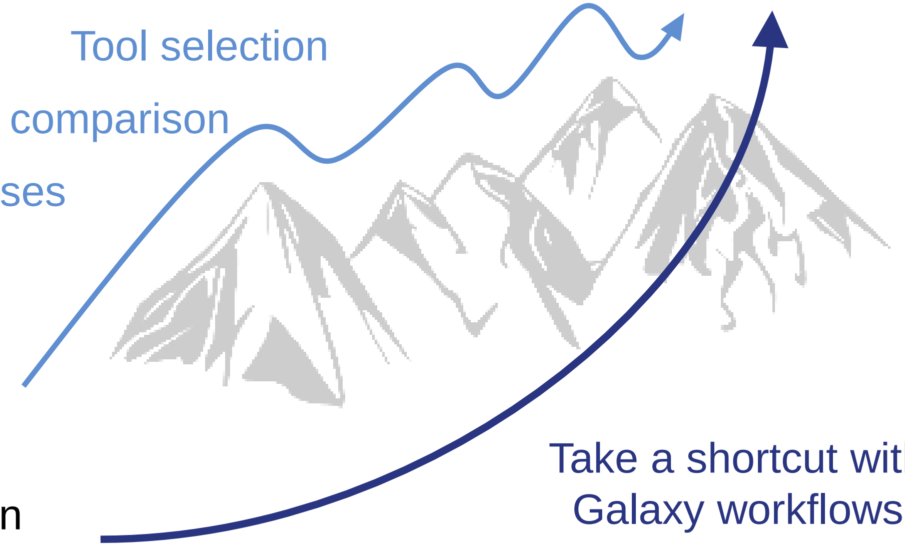
… more than a hundred available

Different Galaxy servers around me

- Main galaxy (US): https://usegalaxy.org/
- European Galaxy (de.NBI support): https://usegalaxy.eu/
- More than 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 - (our module on Friday) https://github.com/destairdenbi/galaxy-modular-workflow-generator

Why using workflows for data analysis?

Programming language?
Analysis approach?



Solution
Data output

Implementation & testing

Tool selection

Tool comparison

Type of analyses

Research question
Data input

Take a shortcut with
Galaxy workflows!

- 👍 Applicable
- 🔍 Transparent
- ⏩ Modular

Key challenges:

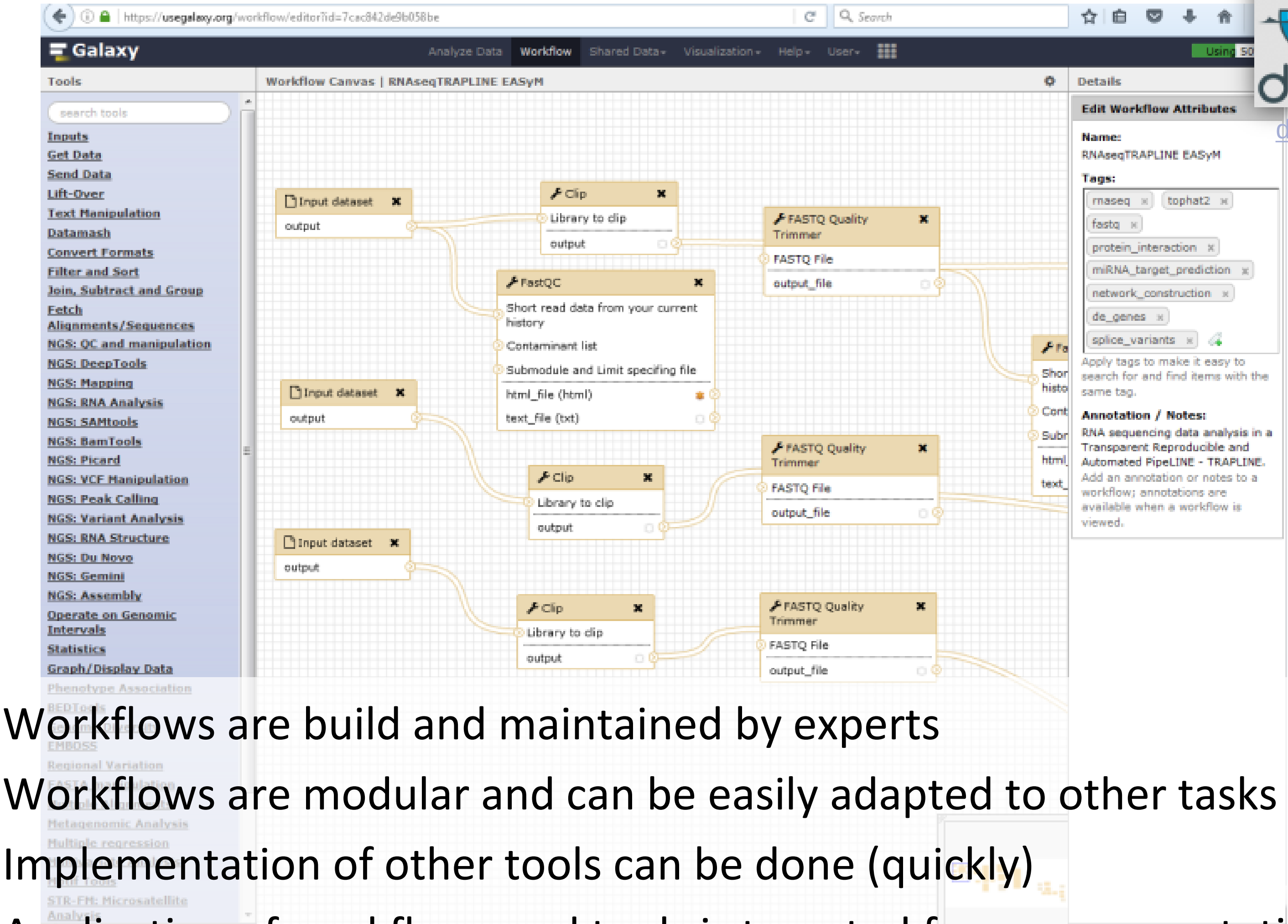
- High data heterogenity
- Large number of tools
- Interdisciplinarity

→ Integration & analysis
of different data
is essential

Lott *et al.* 2017, Journal of Biotechnology

Using workflow development

- Key performance of Galaxy: usegalaxy.eu



Edit Workflow Attributes

Name:
RNAseqTRAPLINE EASyM

Tags:
rnaseq, tophat2, fastq, protein_interaction, miRNA_target_prediction, network_construction, de_genes, splice_variants

Apply tags to make it easy to search for and find items with the same tag.

Annotation / Notes:
RNA sequencing data analysis in a Transparent Reproducible and Automated PipeLINE - TRAPLINE.

Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

- 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

docker.com

biocontainers.pro

bioconda.github.io

elixir-europe.org

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

denbi.de

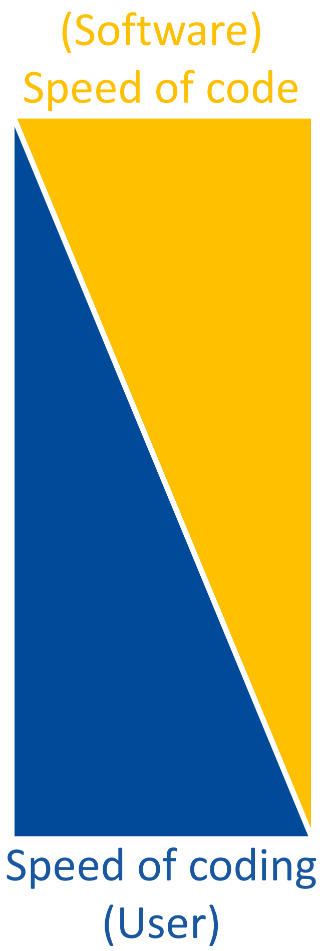
Interactive environments

Python - iJupyter

- Freely available
- Python is a general purpose language, great for data structures and programming in general, it has a vast collection of libraries that one can use

R-Studio

- Freely available
- Oriented to statistical analysis and data processing in a smaller scale. It has a very huge collection of packages to do almost anything one might imagine with data and they are easy to install

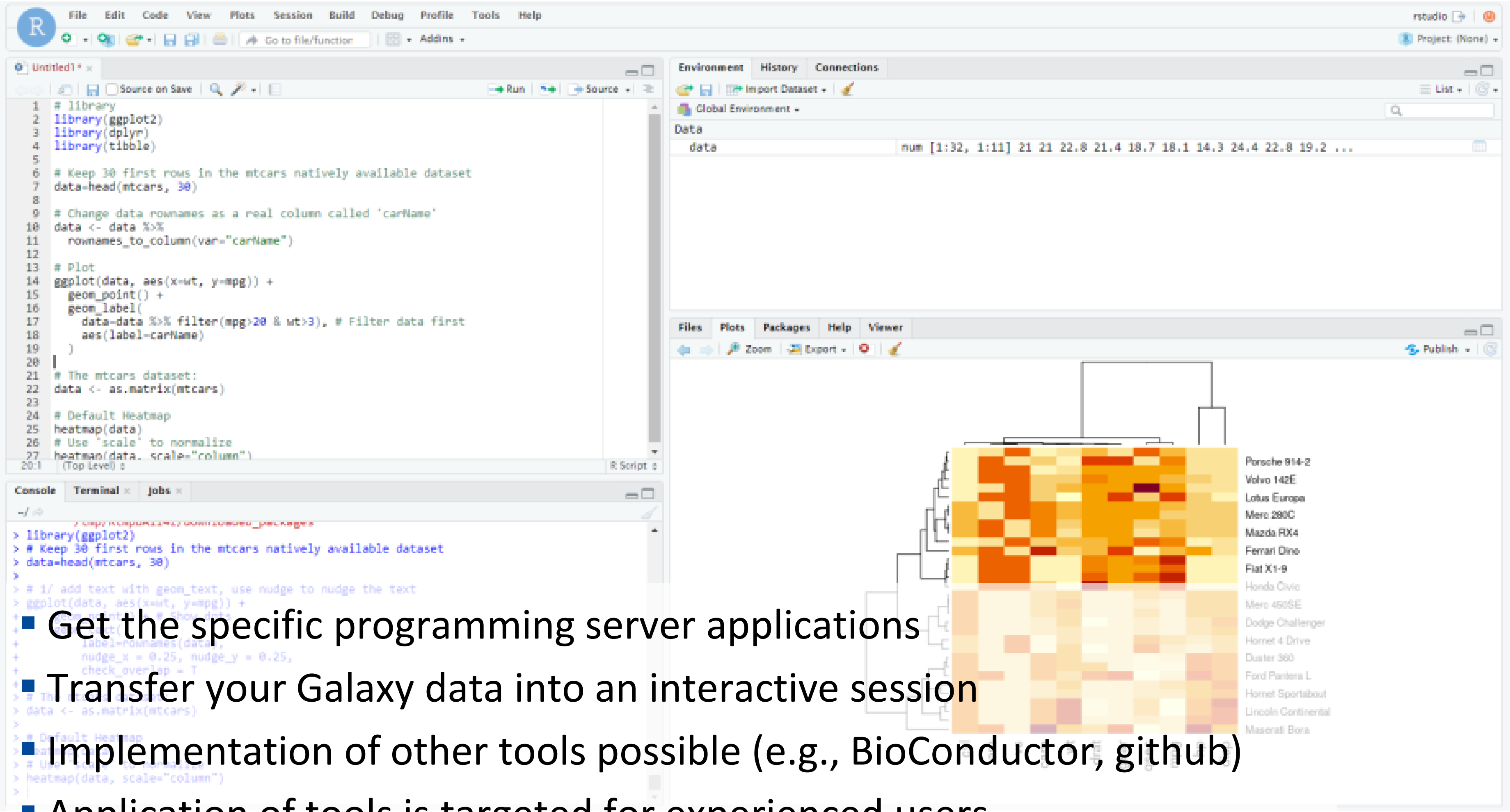


(Software)
Speed of code

Speed of coding
(User)

Interactive environments

- Key performance of Galaxy: usegalaxy.eu



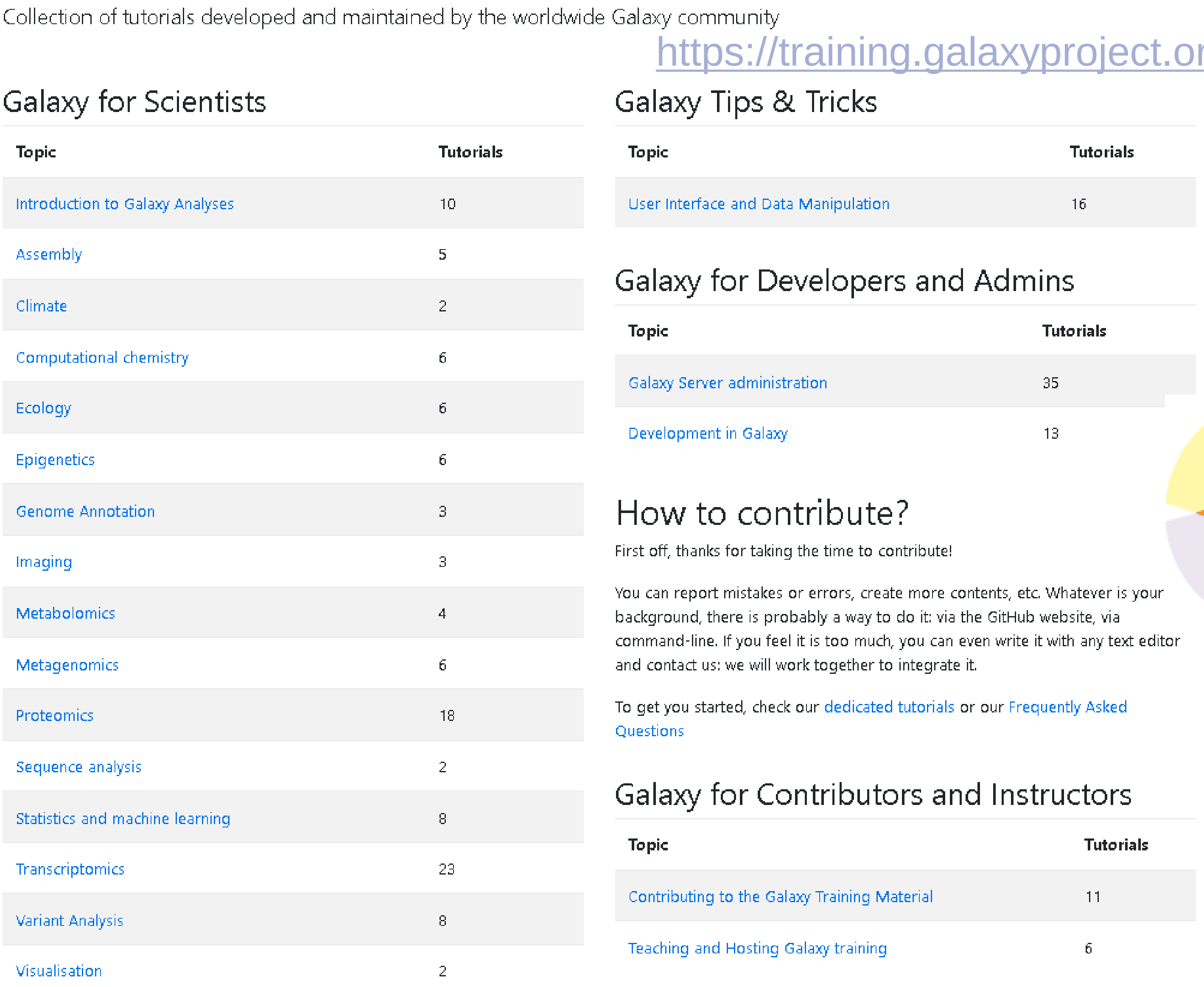
```
# library
library(ggplot2)
library(dplyr)
library(tibble)

# Keep 30 first rows in the mtcars natively available dataset
data=head(mtcars, 30)

# Change data rownames as a real column called 'carName'
data <- data %>%
  rownames_to_column(var="carName")

# Plot
ggplot(data, aes(x=wt, y=mpg)) +
  geom_point() +
  geom_label(
    data=data %>% filter(mpg>20 & wt>3), # Filter data first
    aes(label=carName)
  )

# The mtcars dataset:
data <- as.matrix(mtcars)

# Default Heatmap
heatmap(data)
# Use 'scale' to normalize
heatmap(data, scale="column")
```

```
> library(ggplot2)
> # Keep 30 first rows in the mtcars natively available dataset
> data=head(mtcars, 30)
>
> # 1/ add text with geom_text, use nudge to nudge the text
> ggplot(data, aes(x=wt, y=mpg)) +
+   geom_point() + # Show dots
+   geom_text(
+     label=rownames(data),
+     nudge_x = 0.25, nudge_y = 0.25,
+     check_overlap = T
+ )
> # The mtcars dataset:
> data <- as.matrix(mtcars)
>
> # Default Heatmap
> heatmap(data)
> # Use 'scale' to normalize
> heatmap(data, scale="column")
>
```

```
Data
  data    num [1:32, 1:11] 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
```

- Get the specific programming server applications
- Transfer your Galaxy data into an interactive session
- Implementation of other tools possible (e.g., BioConductor, github)
- Application of tools is targeted for experienced users

Welcome to the Galaxy training network



Collection of tutorials developed and maintained by the worldwide Galaxy community

https://training.galaxyproject.org/training-material/

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	10
Assembly	5
Climate	2
Computational chemistry	6
Ecology	6
Epigenetics	6
Genome Annotation	3
Imaging	3
Metabolomics	4
Metagenomics	6
Proteomics	18
Sequence analysis	2
Statistics and machine learning	8
Transcriptomics	23
Variant Analysis	8
Visualisation	2

Galaxy Tips & Tricks

Topic	Tutorials
User Interface and Data Manipulation	16

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	35
Development in Galaxy	13

How to contribute?

First off, thanks for taking the time to contribute!

You can report mistakes or errors, create more contents, etc. Whatever is your background, there is probably a way to do it: via the GitHub website, via command-line. If you feel it is too much, you can even write it with any text editor and contact us: we will work together to integrate it.

To get you started, check our dedicated tutorials or our Frequently Asked Questions

Galaxy for Contributors and Instructors

Topic	Tutorials
Contributing to the Galaxy Training Material	11
Teaching and Hosting Galaxy training	6

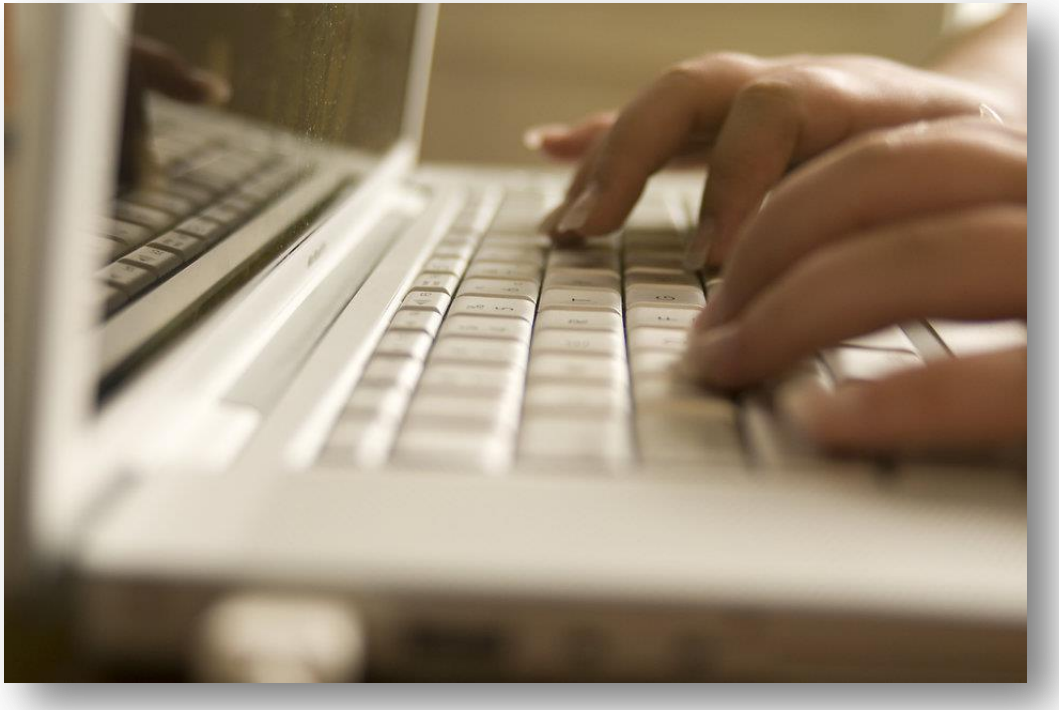
Batut *et al.* 2018, Cell Systems

Hands on part:

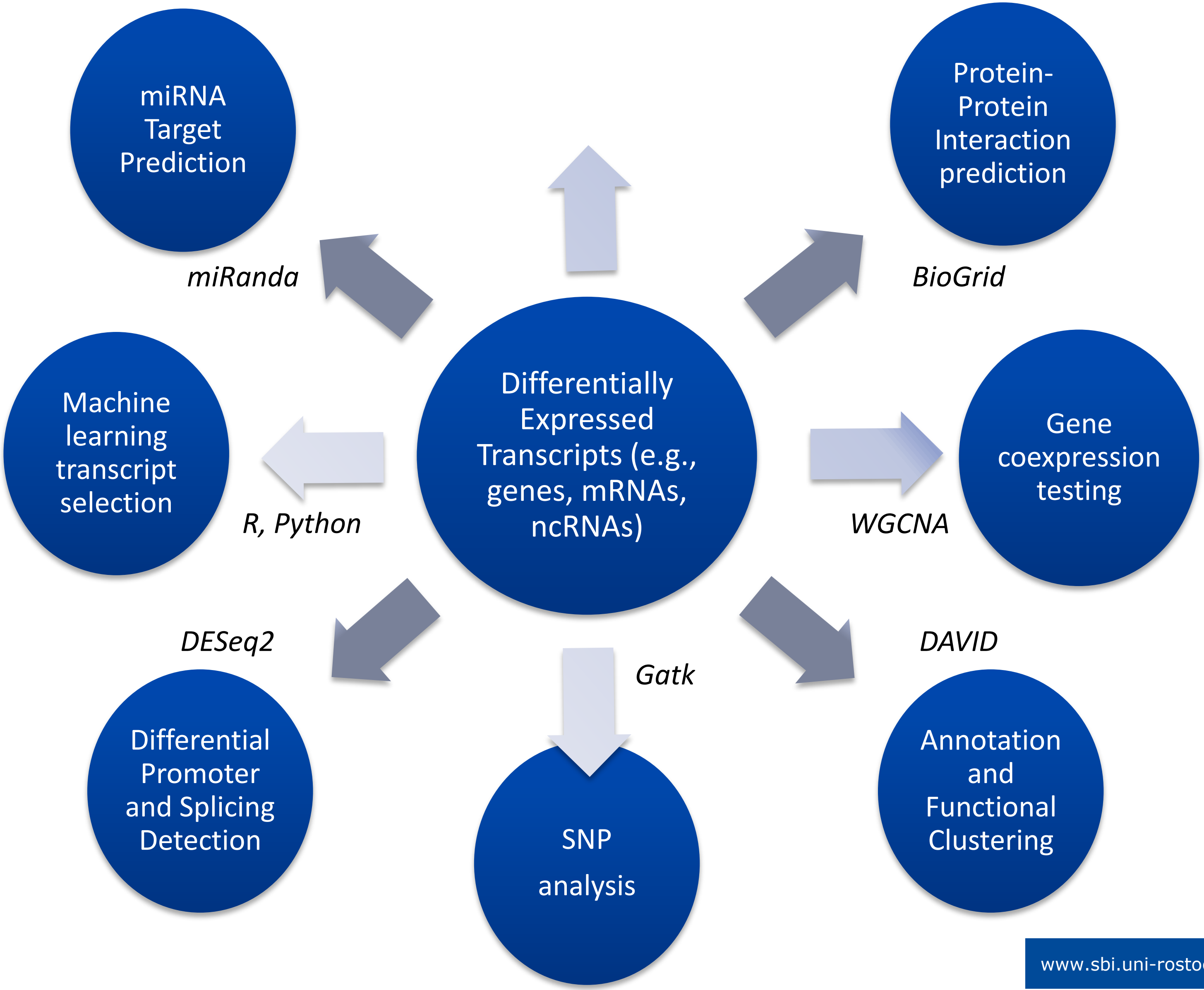
“RNA-Seq data processing and interpretation“

Material: https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html

Please visit and explore Galaxy
usegalaxy.eu



Linking and integrating data



Differentially Expressed Transcripts (e.g., genes, mRNAs, ncRNAs)

miRNA Target Prediction

miRanda

Protein-Protein Interaction prediction

BioGrid

Machine learning transcript selection

R, Python

Gene coexpression testing

WGCNA

Differential Promoter and Splicing Detection

DESeq2

SNP analysis

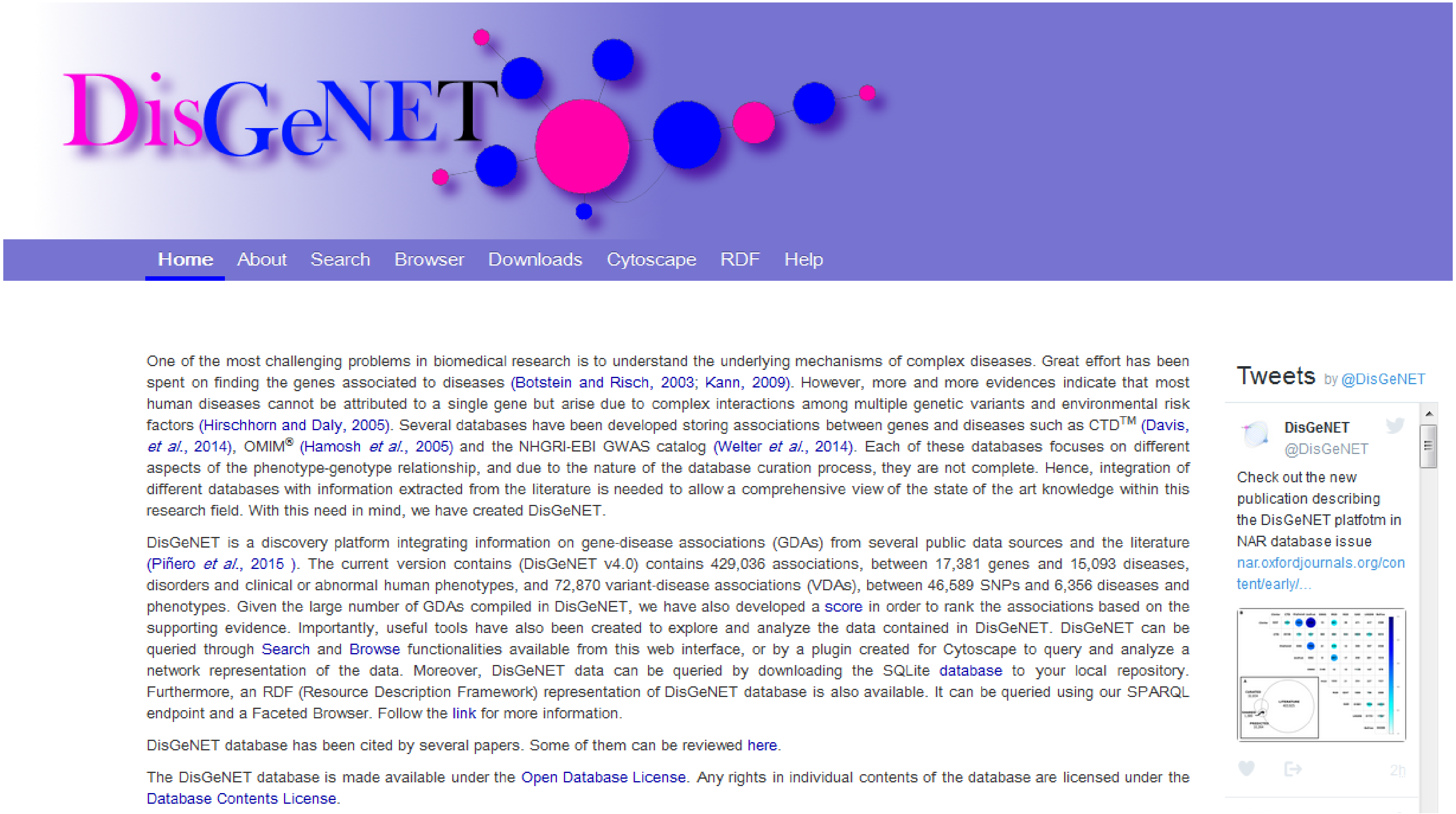
Gatk

Annotation and Functional Clustering

DAVID

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^{TM} (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 diseases and phenotypes. Given the large number of GDAs compiled in DisGeNET, we have also developed a score in order to rank the associations based on the supporting evidence. Importantly, useful tools have also been created to explore and analyze the data contained in DisGeNET. DisGeNET can be queried through Search and Browse functionalities available from this web interface, or by a plugin created for Cytoscape to query and analyze a network representation of the data. Moreover, DisGeNET data can be queried by downloading the SQLite database to your local repository. Furthermore, an RDF (Resource Description Framework) representation of DisGeNET database is also available. It can be queried using our SPARQL 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 licensed under the Database Contents License.

Tweets by @DisGeNET

DisGeNET
@DisGeNET

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

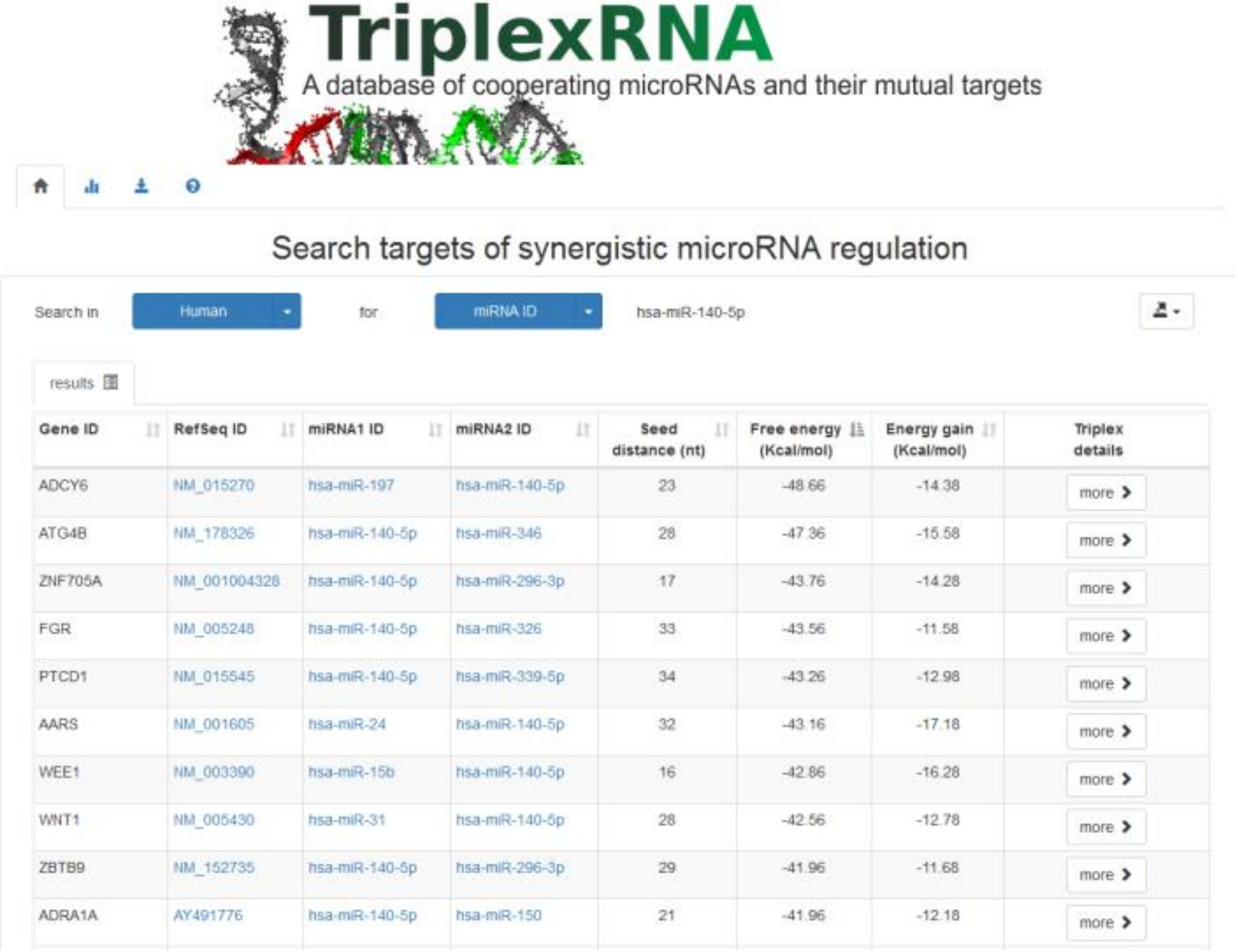
2h

Linking and integrating data

- Gene seq enrichment analysis (GSEA) – by means of Gene Ontology and Pathway information (e.g., WikiPathways, KEGG, Reactome)
  - Cytoscape (http://www.cytoscape.org/)
    - ClueGo/Cluepedia (http://apps.cytoscape.org/apps/cluego)
    - BiNGO (http://apps.cytoscape.org/apps/bingo)
  - David (https://david.ncifcrf.gov/summary.jsp)
  - Enrichr (http://amp.pharm.mssm.edu/Enrichr/)
  - gProfiler (https://biit.cs.ut.ee/gprofiler/gost)
    - Available in Galaxy (gProfilerGOSt)

Explore miRNA cooperativity for miRNA-mRNA pairings

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



TriplexRNA

A database of cooperating microRNAs and their mutual targets

Search targets of synergistic microRNA regulation

Search in Human for miRNA ID hsa-miR-140-5p

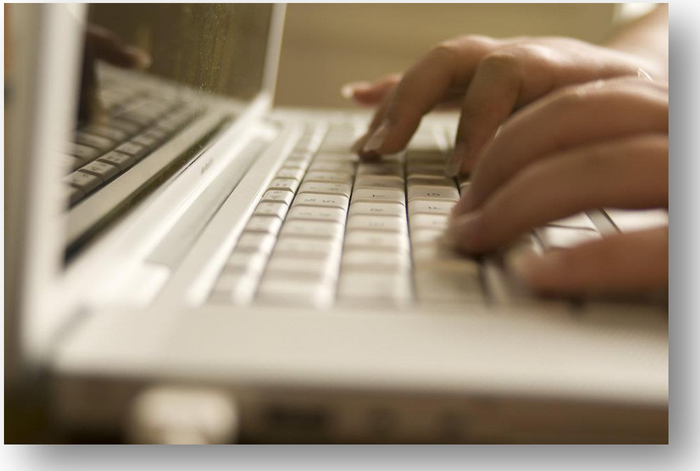
results

Gene ID	RefSeq ID	miRNA1 ID	miRNA2 ID	Seed distance (nt)	Free energy (Kcal/mol)	Energy gain (Kcal/mol)	Triplex details
ADCY6	NM_015270	hsa-miR-197	hsa-miR-140-5p	23	-48.66	-14.38	more ❯
ATG4B	NM_178326	hsa-miR-140-5p	hsa-miR-346	28	-47.36	-15.58	more ❯
ZNF705A	NM_001004328	hsa-miR-140-5p	hsa-miR-296-3p	17	-43.76	-14.28	more ❯
FGR	NM_005248	hsa-miR-140-5p	hsa-miR-326	33	-43.56	-11.58	more ❯
PTCD1	NM_015545	hsa-miR-140-5p	hsa-miR-339-5p	34	-43.26	-12.98	more ❯
AARS	NM_001605	hsa-miR-24	hsa-miR-140-5p	32	-43.16	-17.18	more ❯
WEE1	NM_003390	hsa-miR-15b	hsa-miR-140-5p	16	-42.86	-16.28	more ❯
WNT1	NM_005430	hsa-miR-31	hsa-miR-140-5p	28	-42.56	-12.78	more ❯
ZBTB9	NM_152735	hsa-miR-140-5p	hsa-miR-296-3p	29	-41.96	-11.68	more ❯
ADRA1A	AY491776	hsa-miR-140-5p	hsa-miR-150	21	-41.96	-12.18	more ❯

Individual - Hands on parts:

Please visit and explore the Galaxy Training Material, material includes different topics such as:

- Nanopore assembly
- De novo transcriptome reconstruction from RNA-Seq
- Visualization: Volcano plot
- Visualization: Heatmap
- RNA-Seq from genes to pathways
- GO enrichment analysis
- Single cell RNA-Seq
- Variant calling (from DNA)



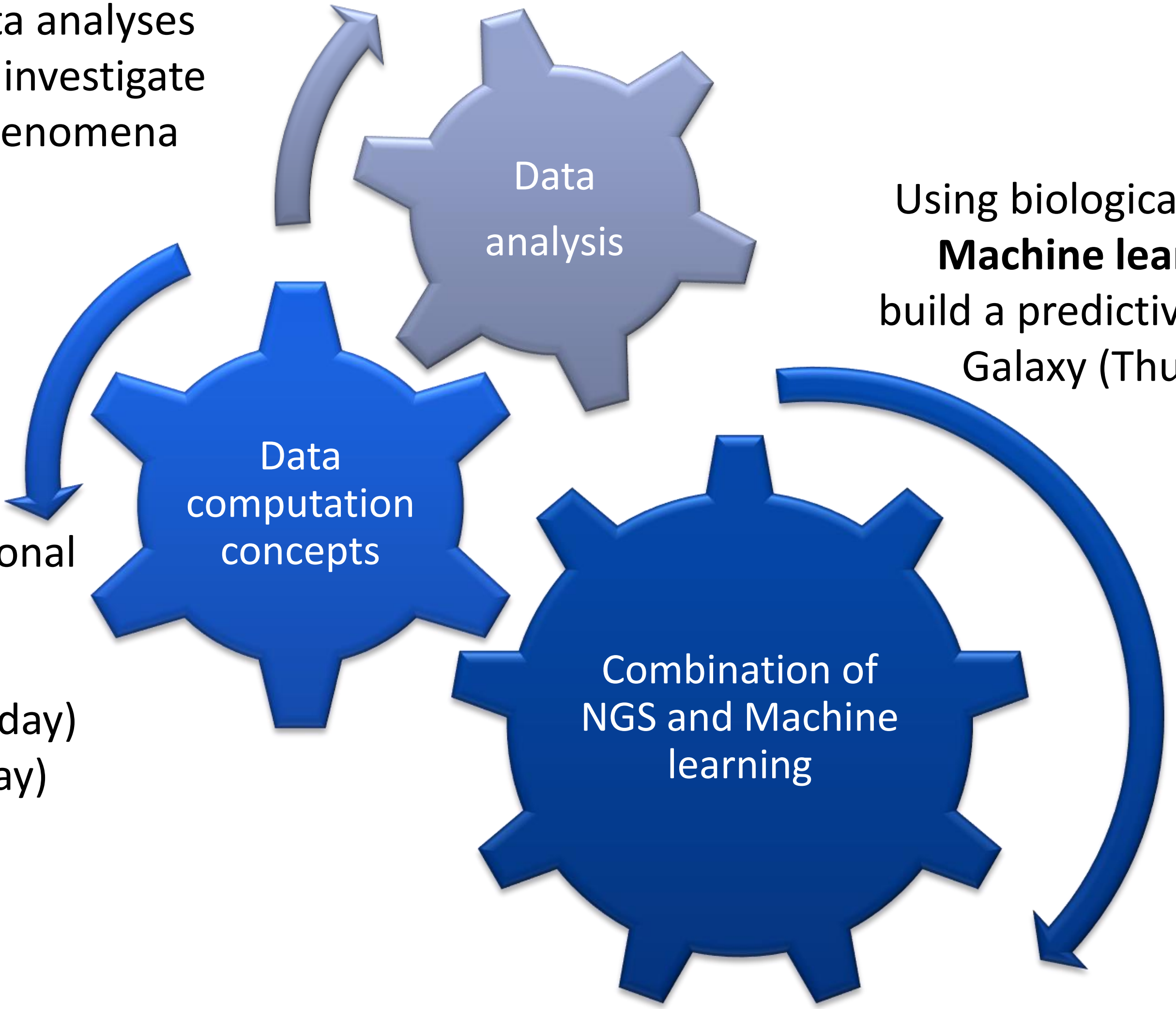
Literature for best practices in RNA-Seq

- 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

What else will we learn?


Data analysis
Data computation concepts
Combination of NGS and Machine learning

Using NGS data analyses approaches to investigate biological phenomena

Using biological data and **Machine learning** to build a predictive model in Galaxy (Thursday)

Utilizing computational technologies,
like **Galaxy**
(Monday - Wednesday)
and **Docker** (Friday)

Acknowledgements

Wolfgang Hess (University of Freiburg)
Steffen Lott (University of Freiburg)
Steve Hoffmann (FLI Jena)
Konstantin Riege (FLI Jena)

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

Supported by:

deNBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
denbi.de

elixir
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).

We hope you enjoyed the training!

Universität Rostock

Traditio et Innovatio

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

deSTAIR

Galaxy Training Network

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE