Introduction to RNA-Seq data analysis with Galaxy

Markus Wolfien and Andrea Bagnacani

de.NBI Training – 7th March 2018 Kiel

www.sbi.uni-rostock.de

Making sense out of data – providing meaning to models

SYSTEMS BIOLOGY BIOINFORMATICS ROSTOCK

Omic analyses and data integration

Data management & standardisation (Fairdom partner)

eHealth iOS Application

SYSTEMS BIOLOGY BIOINFORMATICS ROSTOCK

Computational modeling & machine learning

Summer schools and workshops

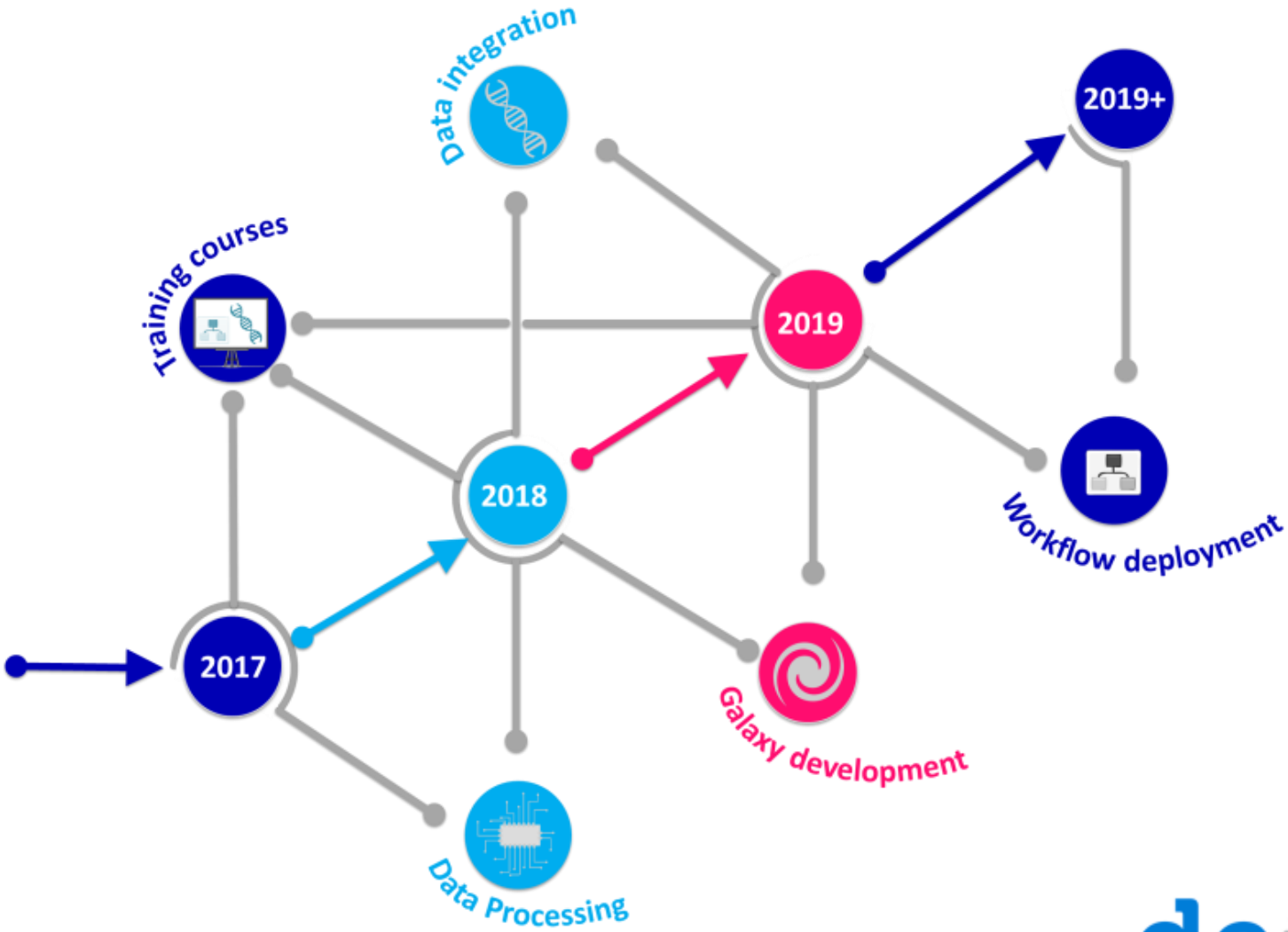
Systems Medicine Pre-clinical trials

sbi.uni-rostock.de/destair

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

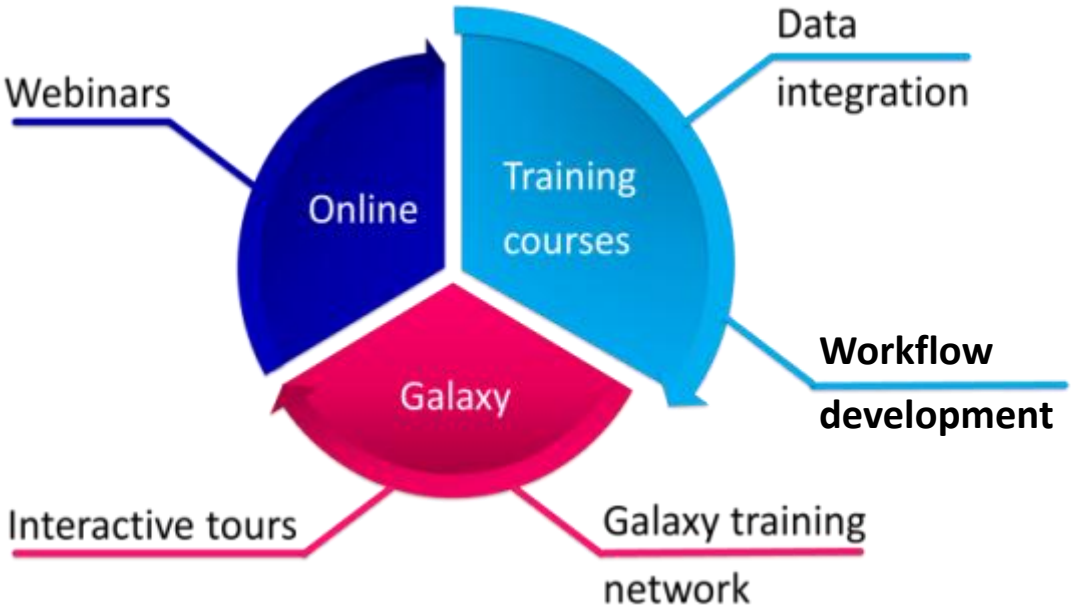
Our aim is to enable a comprehensive **analysis of RNA-Seq experiments as a service.** To enable maximum usefulness, interconnectivity, and accessibility for the developed approaches and services, we will provide dedicated **workshops**, **training programs and screen casts** for bioinformaticians and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

de.STAIR – RNA-Seq analysis and integration



GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

de.STAIR – RNA-Seq analysis and integration



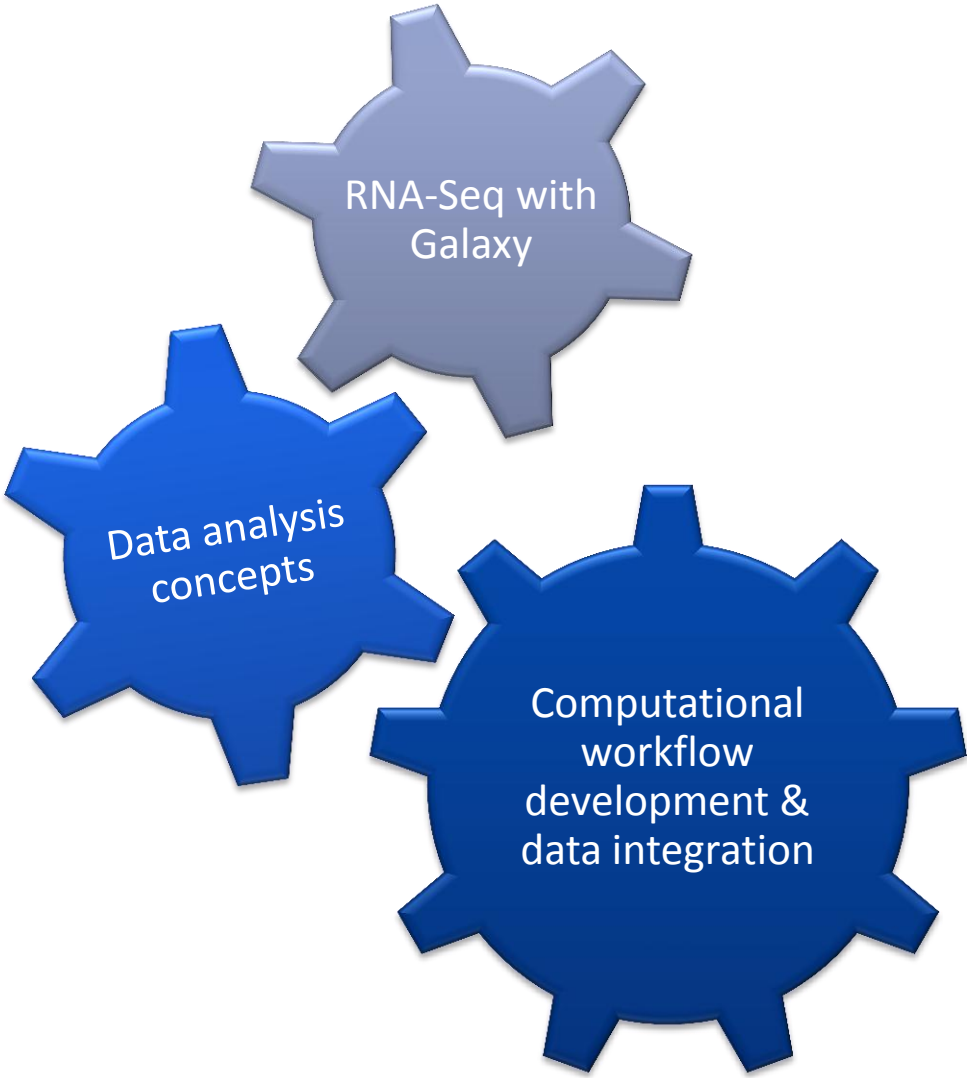
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

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 a use case in the medical field



What is big data?

"The basic idea behind the phrase "Big Data" is that everything we do is increasingly leaving a digital trace (or data), which we (and others) can use and analyze. Big data therefore refers to our ability to make use of the ever-increasing volumes of data."

Bernhard Marr, 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

Transcription from genes to a functional gene products

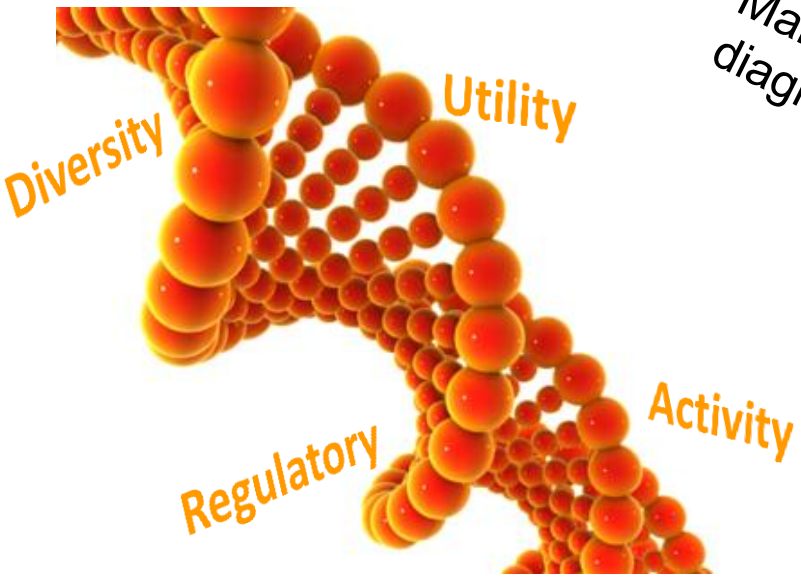
Many different variations and subtypes

Information about regulatory mechanisms

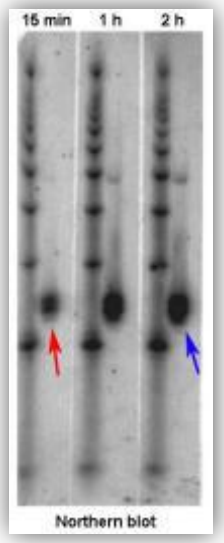
Active and measurable state of the cell …

… ,but only a snapshot

Many different therapeutical and diagnostical approaches

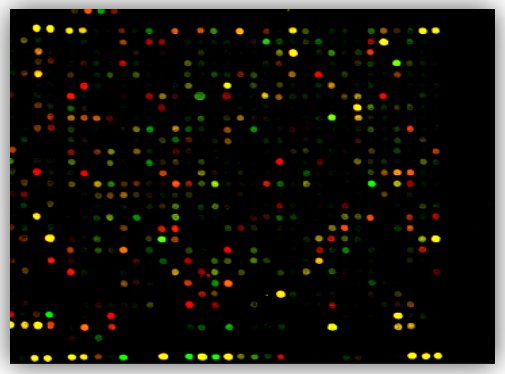


Measureing gene expression

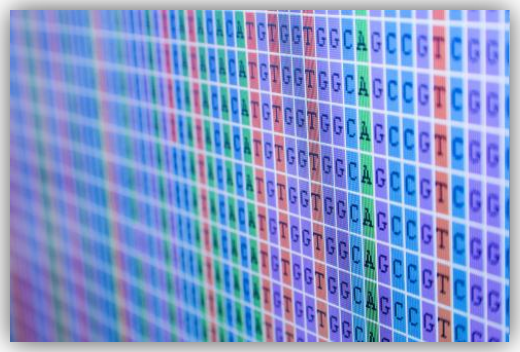


Northern Blot

Reverse Transcription PCR

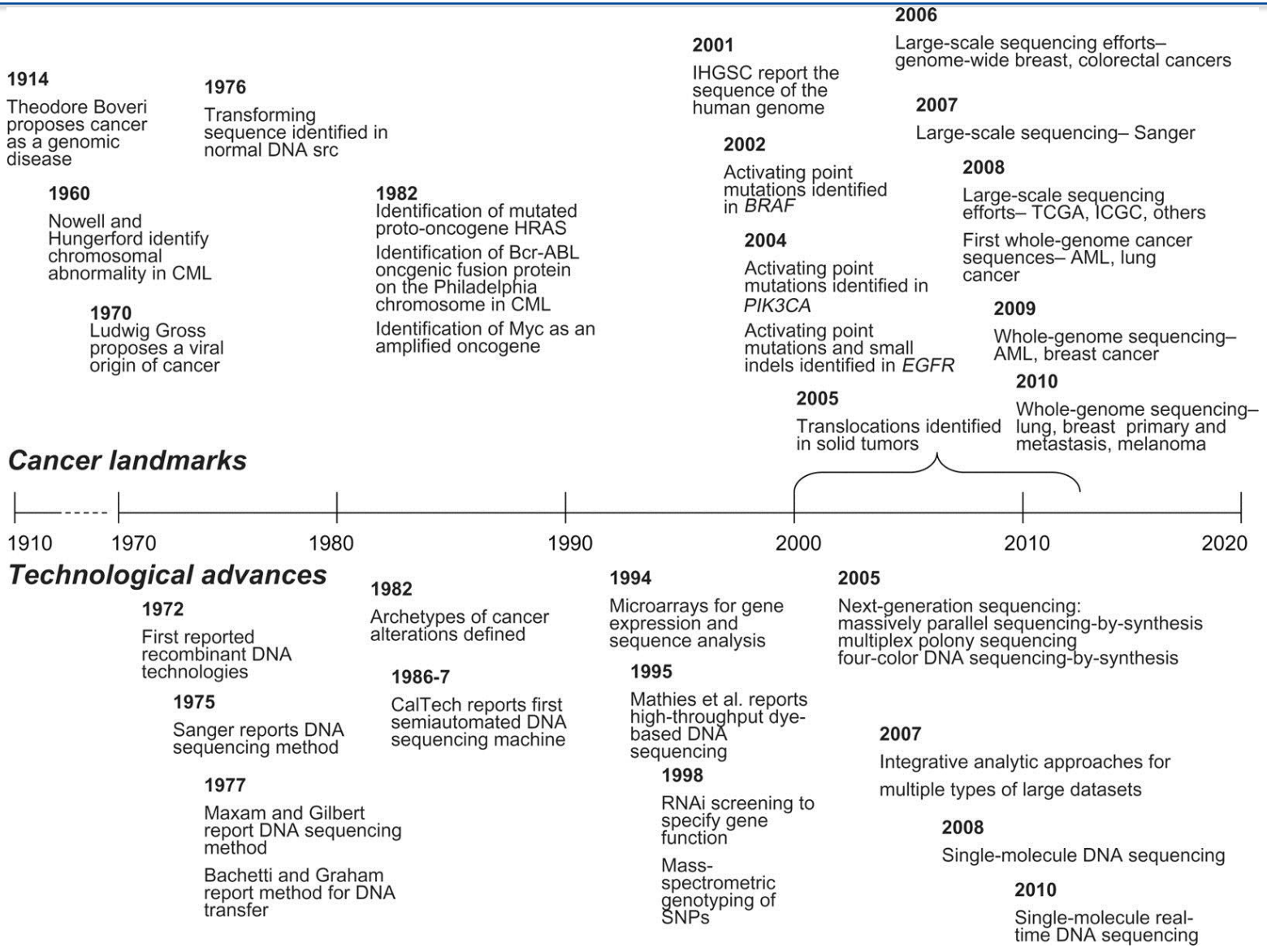


Microarrays

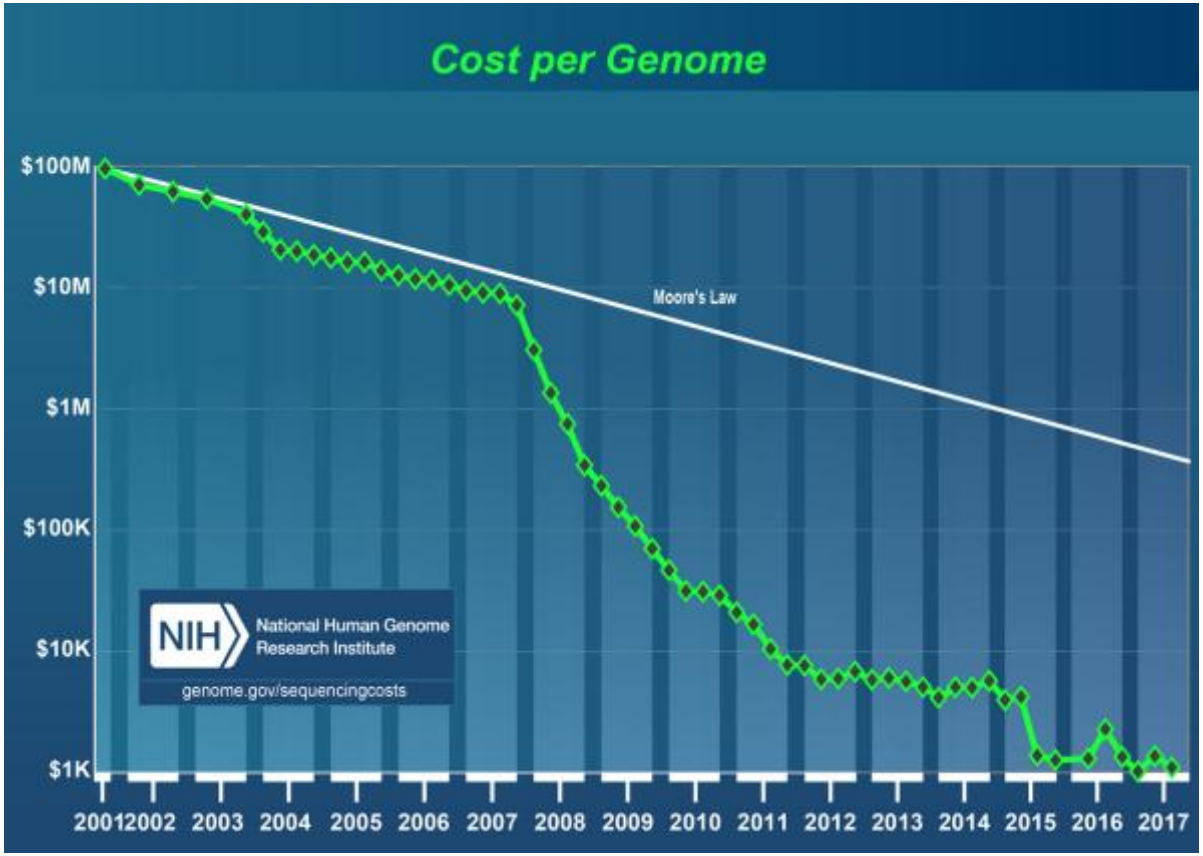


NGS

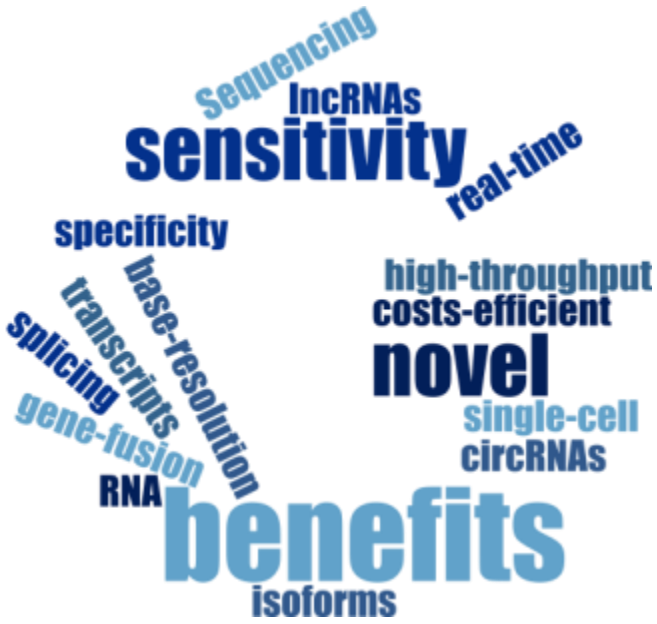
Evolution of Sequencing technologies



Technical advances lead the way

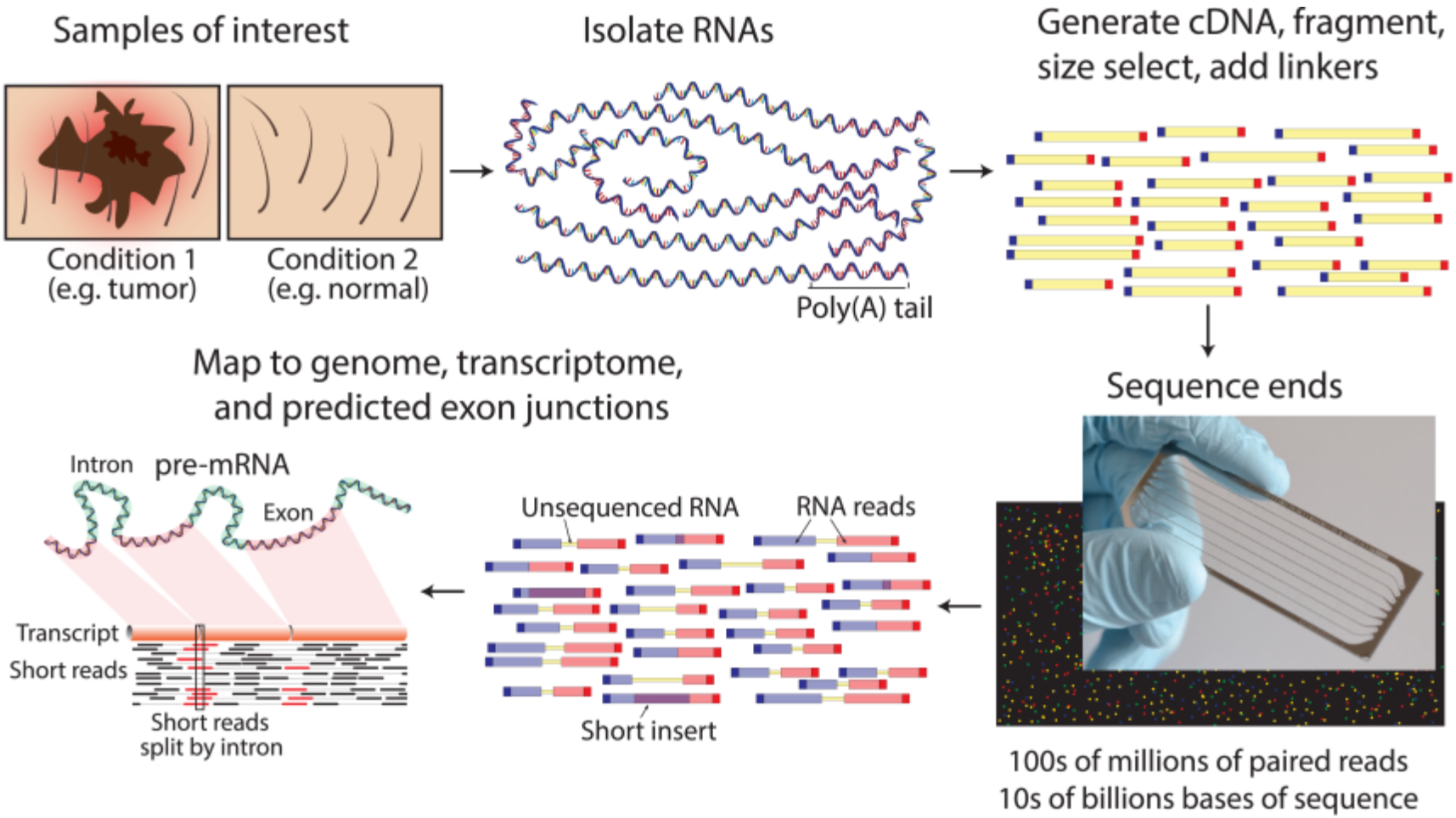


Why do we need NGS?



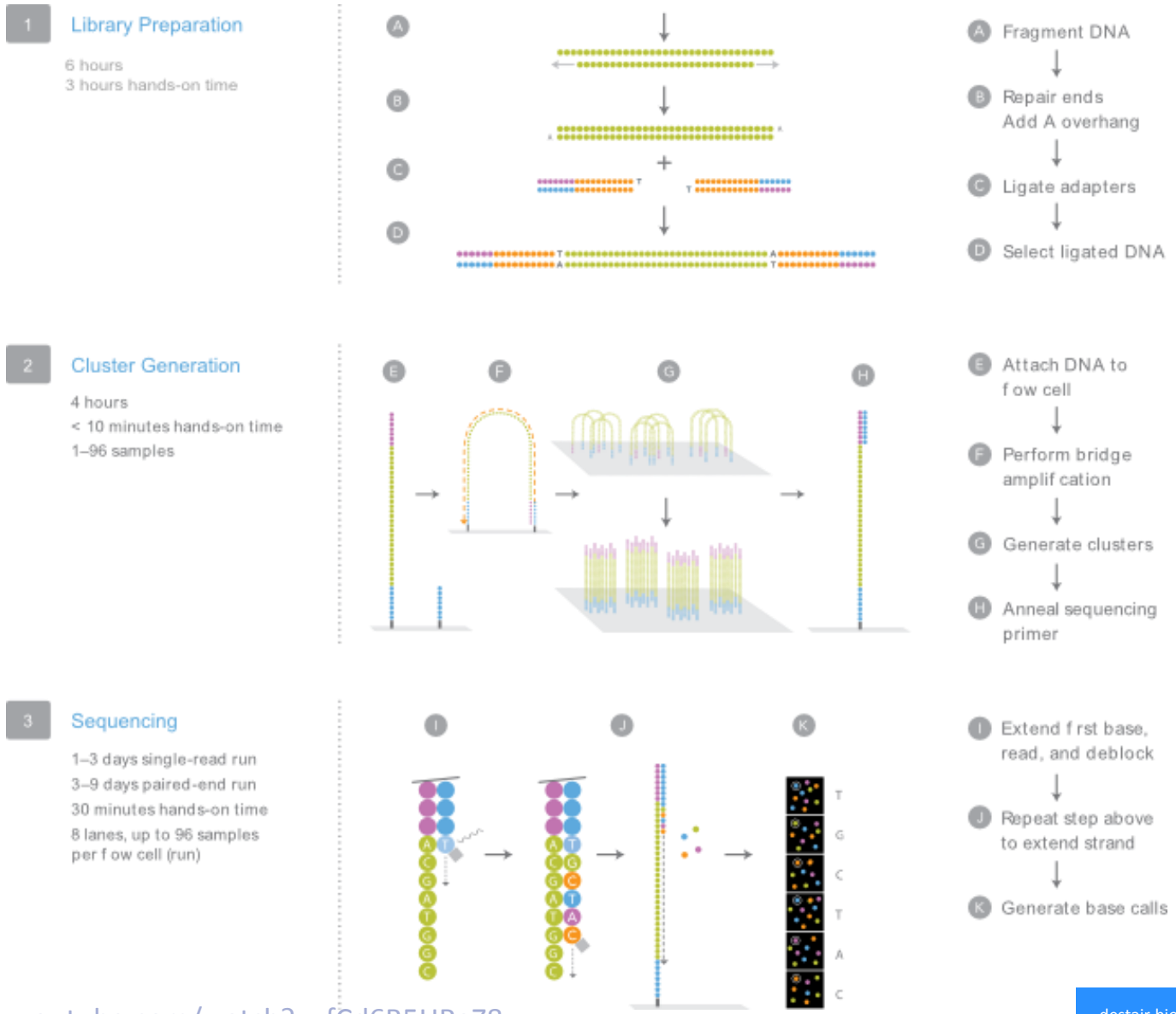
"RNA-Seq is able to identify thousands of differentially expressed genes, tens of thousands of differentially expressed gene isoforms and can detect mutations and germline variations for hundreds to thousands of expressed genetic variants, as well as detecting chimeric gene fusions, transcript isoforms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

From sample to readout



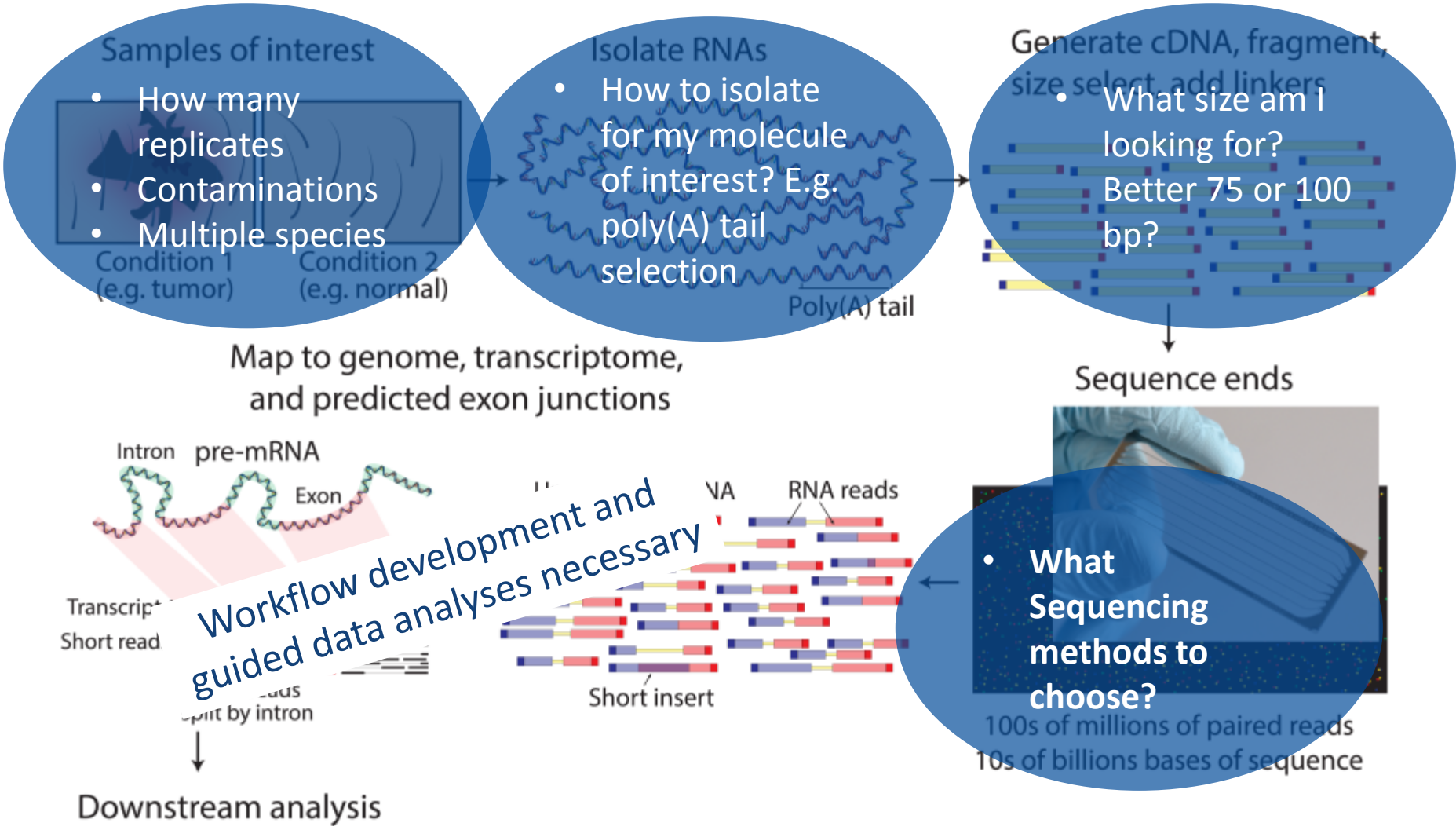
Griffith, Plos Comp. Biol., 2015

How do I get my NGS data?



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

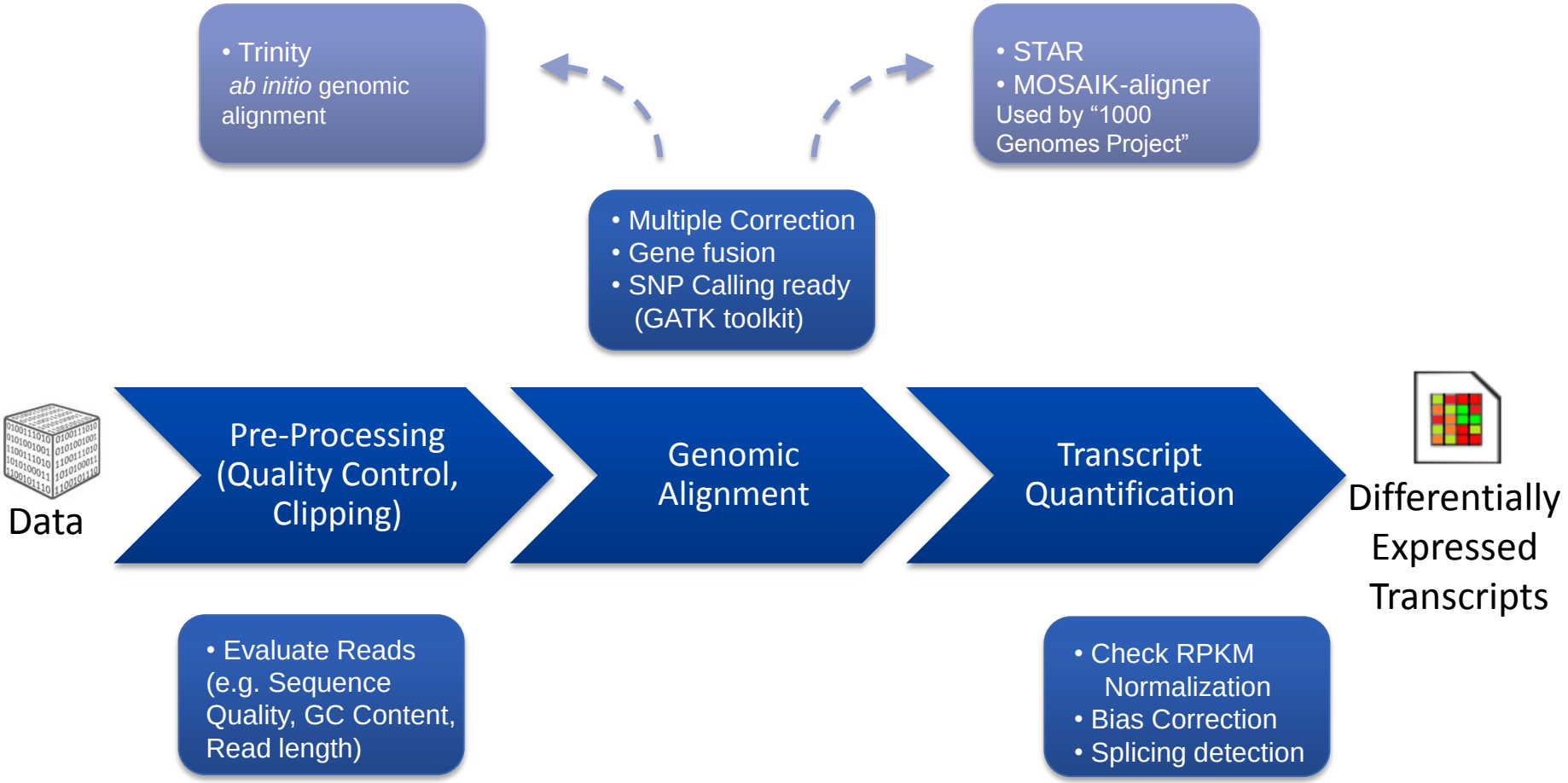
From sample to readout



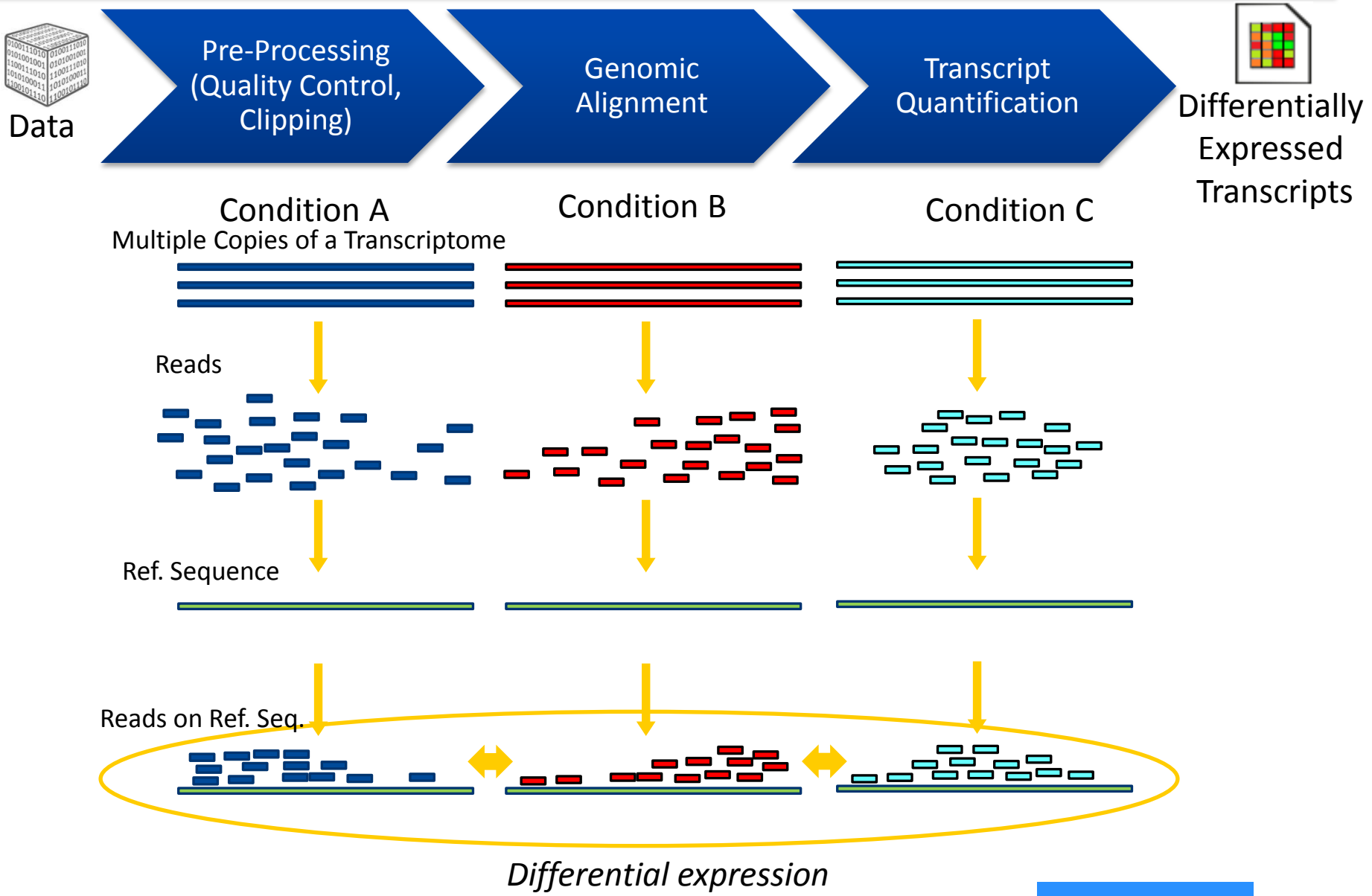
Griffith, Plos Comp. Biol., 2015

Data analysis

Basic workflow for data processing



Differential expression



Big Data and the need for new analyses

broadinstitute.org

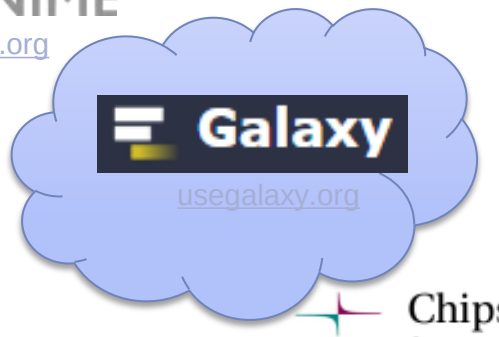
geneprof.org

big.crg.cat/services/grape

knime.org

python.org

mapman.gabipd.org



usegalaxy.org

r-project.org/

chipster.csc.fi

gene-talk.de

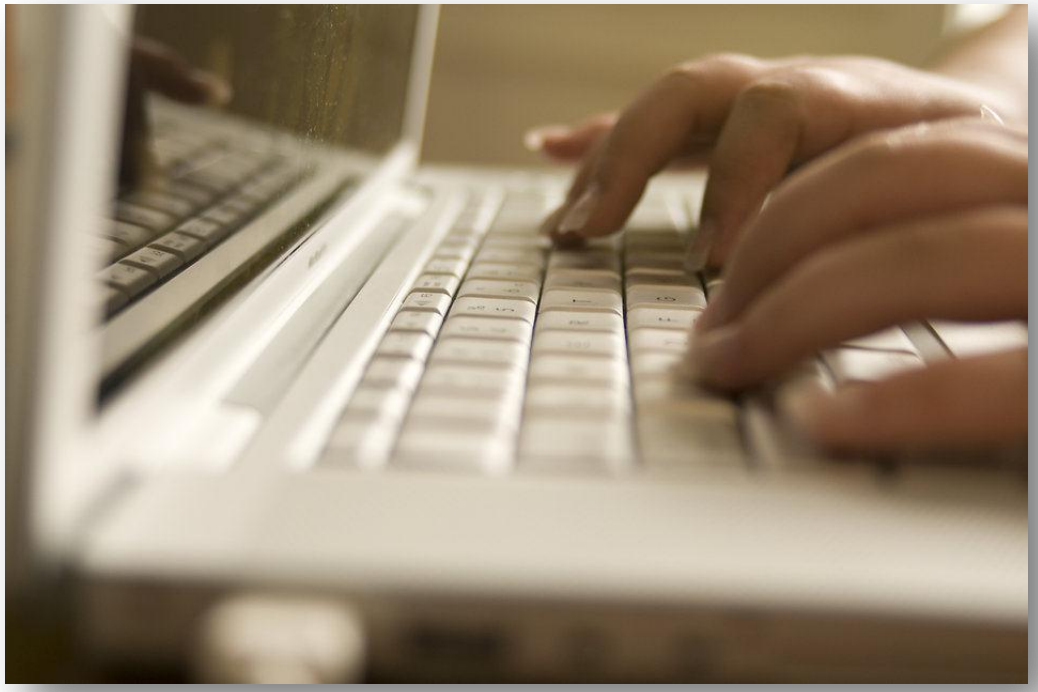
bioconductor.org

illumina.com

Hands on part 1:

“Register and login to Galaxy. Explore and get familiar with your Galaxy.“

Please visit and explore Galaxy
usegalaxy.org



Galaxy training material

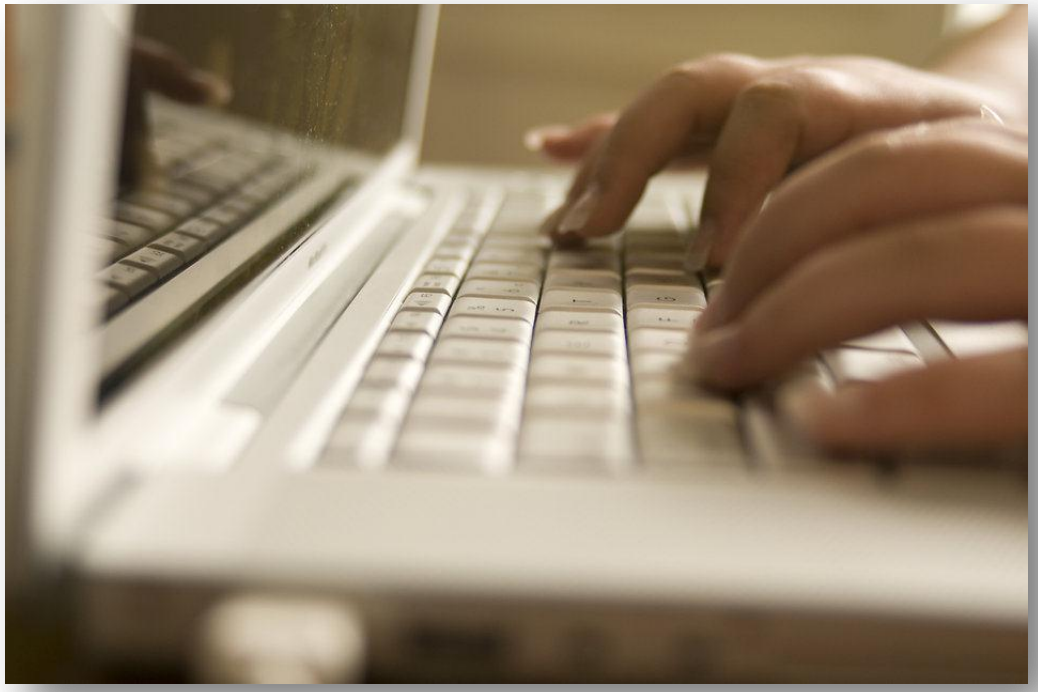
Training material: http://galaxyproject.github.io/training-material/

Manuscript: https://www.biorxiv.org/content/early/2017/11/29/225680

Hands on part 2 9:00 – 10:00

“RNA-Seq data preprocessing and quality control“

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

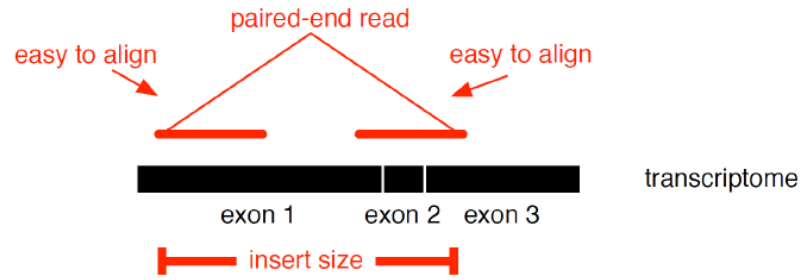


Hands on part 3 10:15 – 11:45

“Application of different read mapping approaches for genomic alignment“

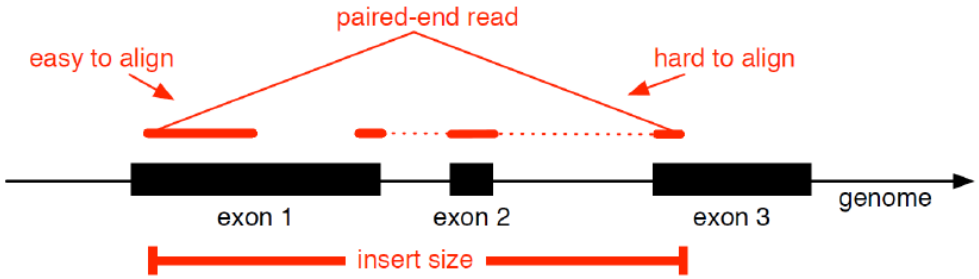
Genomic alignment - Mapping to a reference genome

Transcriptome alignment



- reliable gene models required
- no detection of novel genes

Genome alignment (splice-aware read alignment)

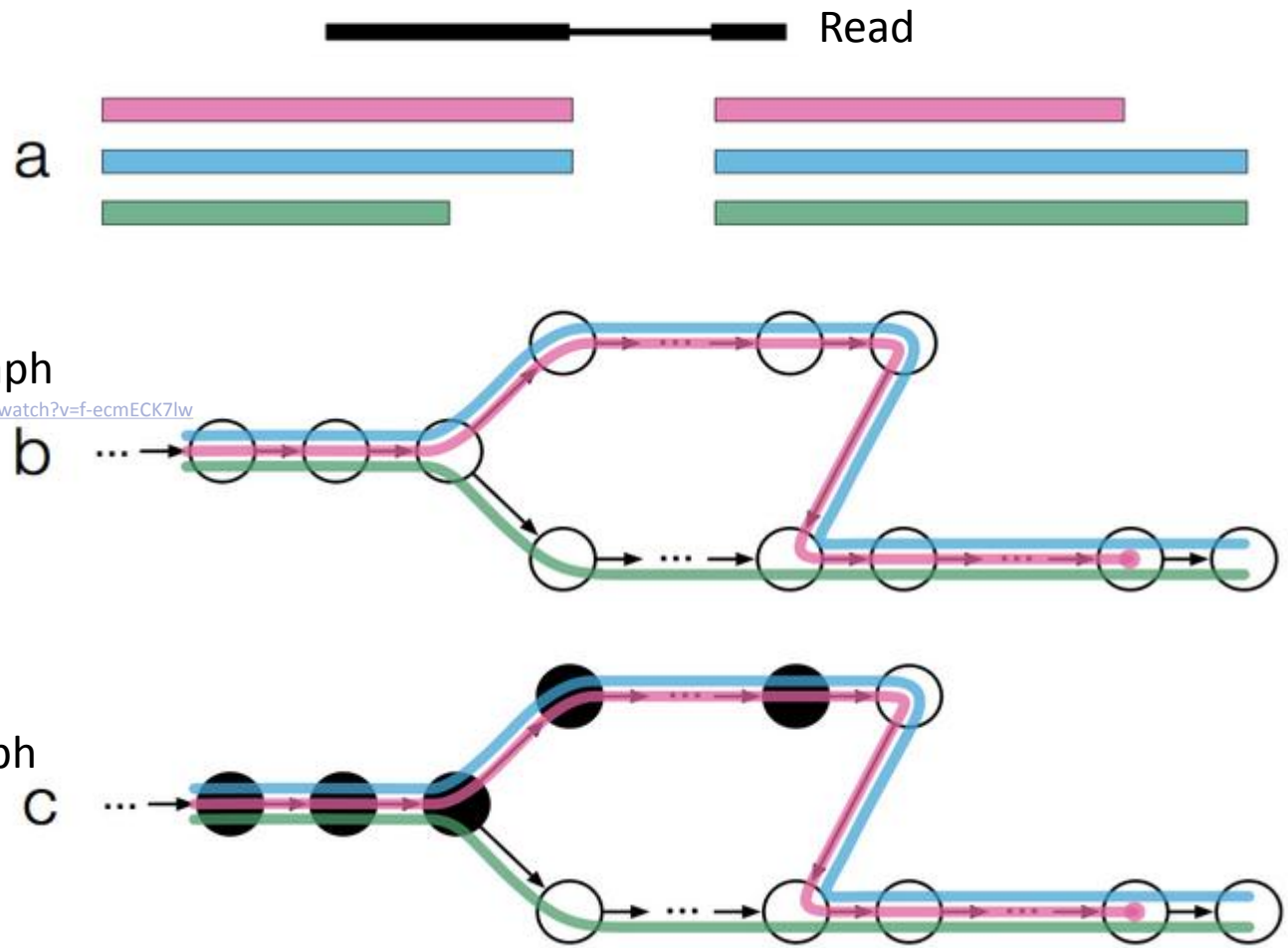


+ detection of novel genes and isoforms

Turro, EMBO, 2012

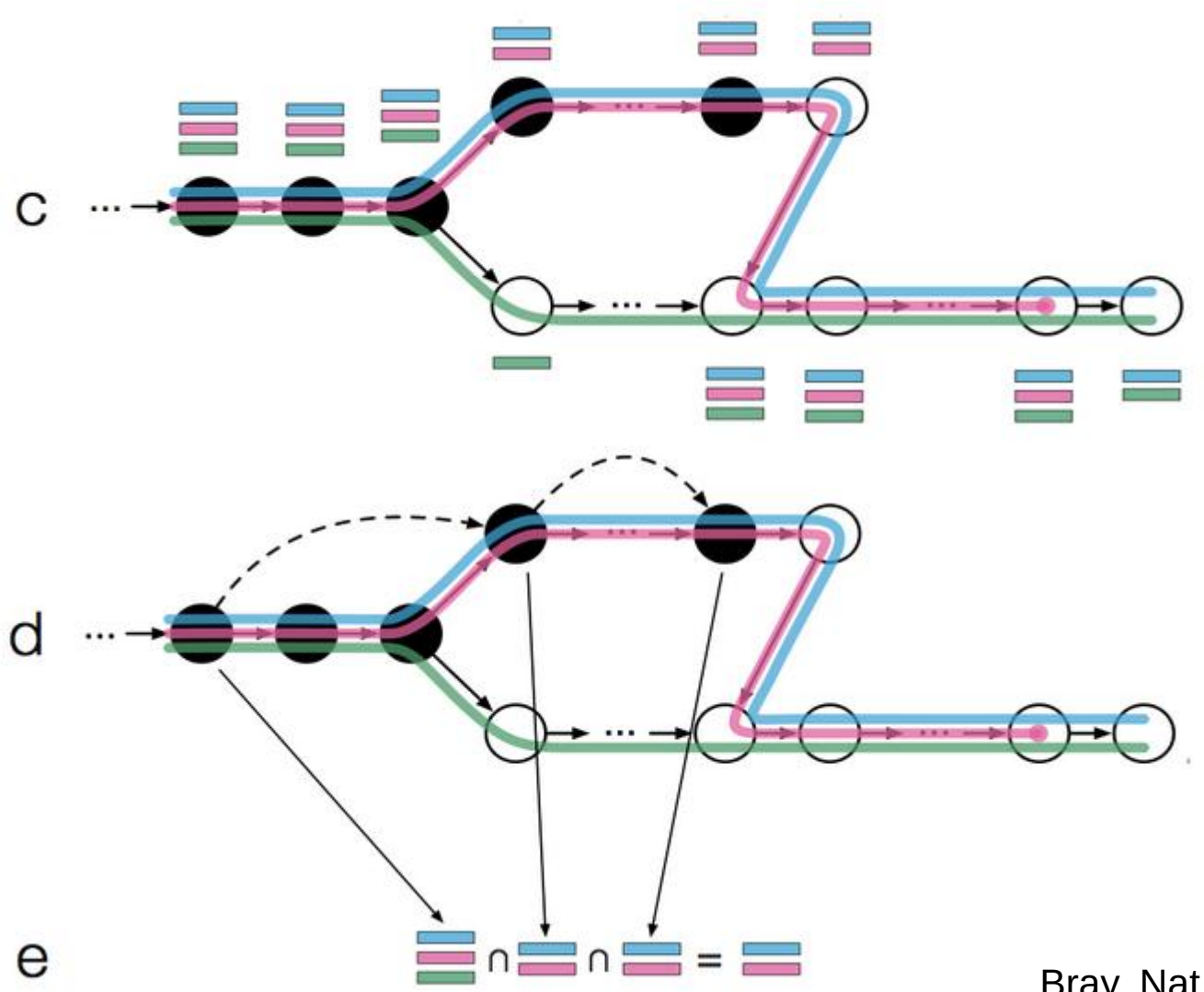
For clinical usage combination of different algorithms possible:

Genomic alignment - pseudoalignment



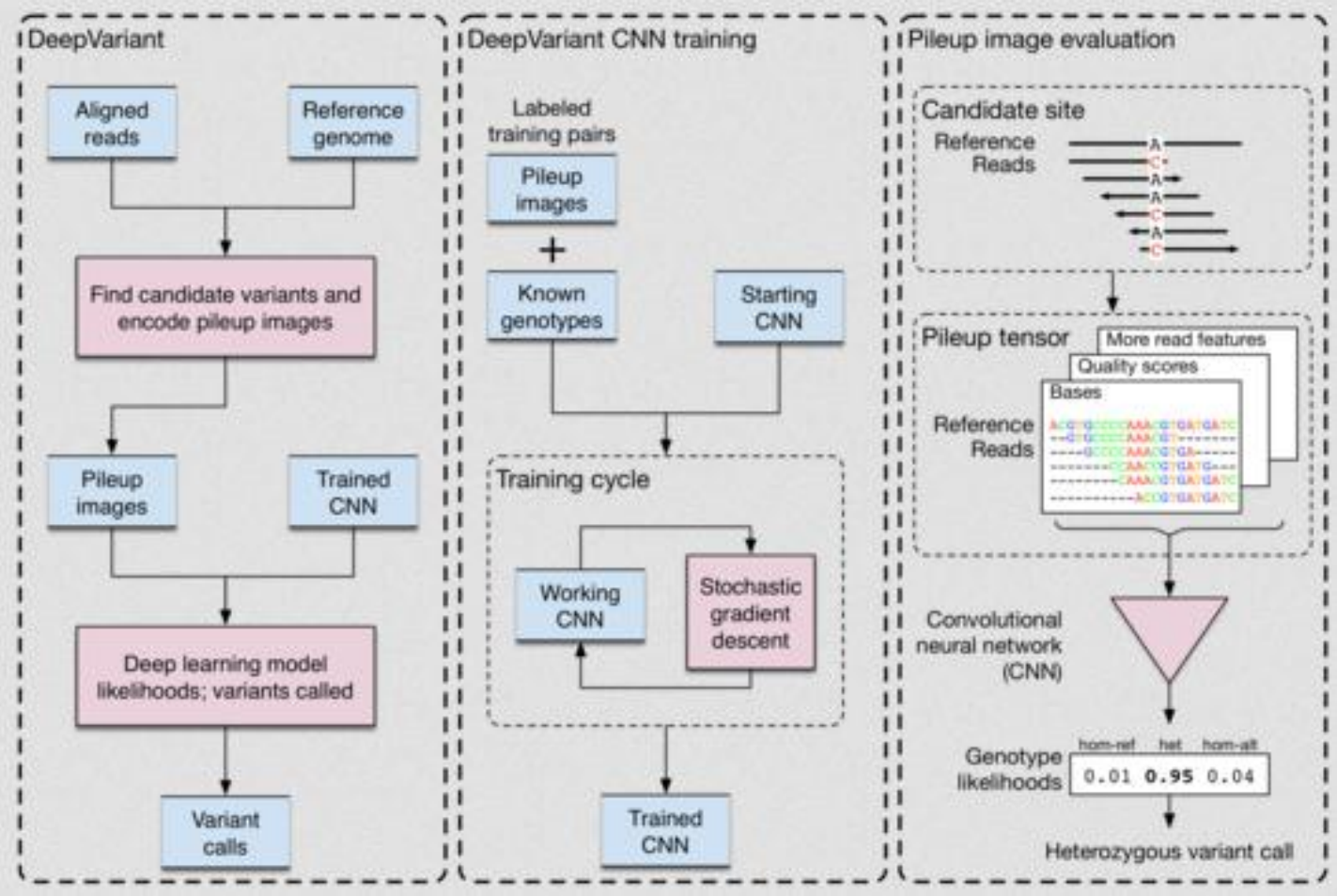
Bray, Nat Biotech, 2016

Genomic alignment - pseudoalignment



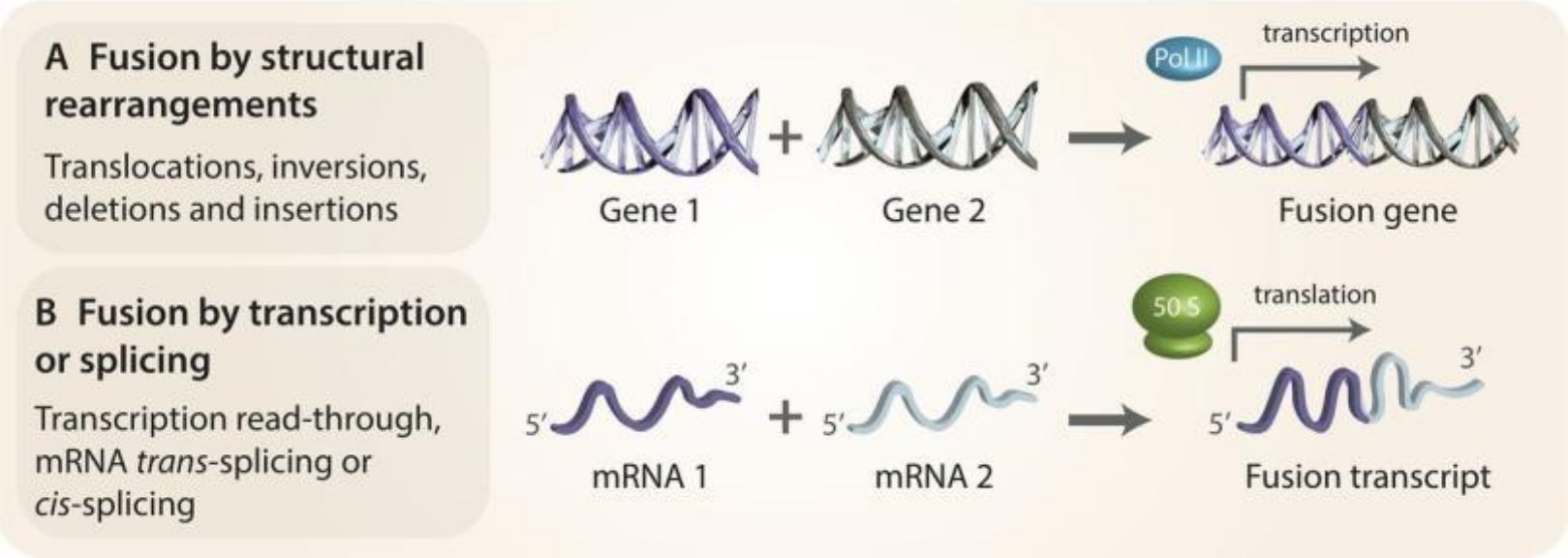
Bray, Nat Biotech, 2016

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

Transcript quantification

- 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, isoforms, ...)
  - RNA libraries (samples)

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

RPKM normalization

- **R**eads **P**er **K**ilobase per **M**illion

- $$RPKM = \frac{Raw\ number\ of\ reads}{Exon\ length} * \frac{1.000.000}{Number\ of\ reads\ mapped\ in\ the\ sample}$$

- In RNA-seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from itHowever:
  - The total number of fragments is biased towards larger genes
  - Total number of fragments is related to total library depth

- Differences with and without normalization and differences among them stated at https://www.youtube.com/watch?v=TTUrtCY2k-w

Visualization of .bam files

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

The James Hutton Institute

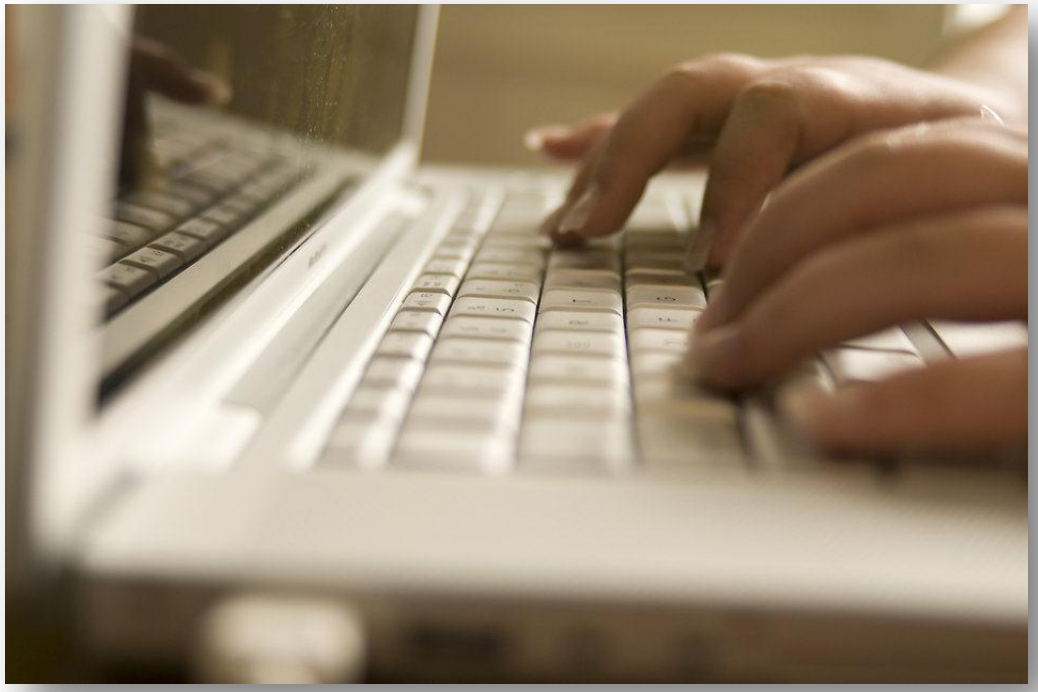
Tablet

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

Hands on part 3 10:15 – 11:45

“Application of different read mapping approaches for genomic alignment“

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



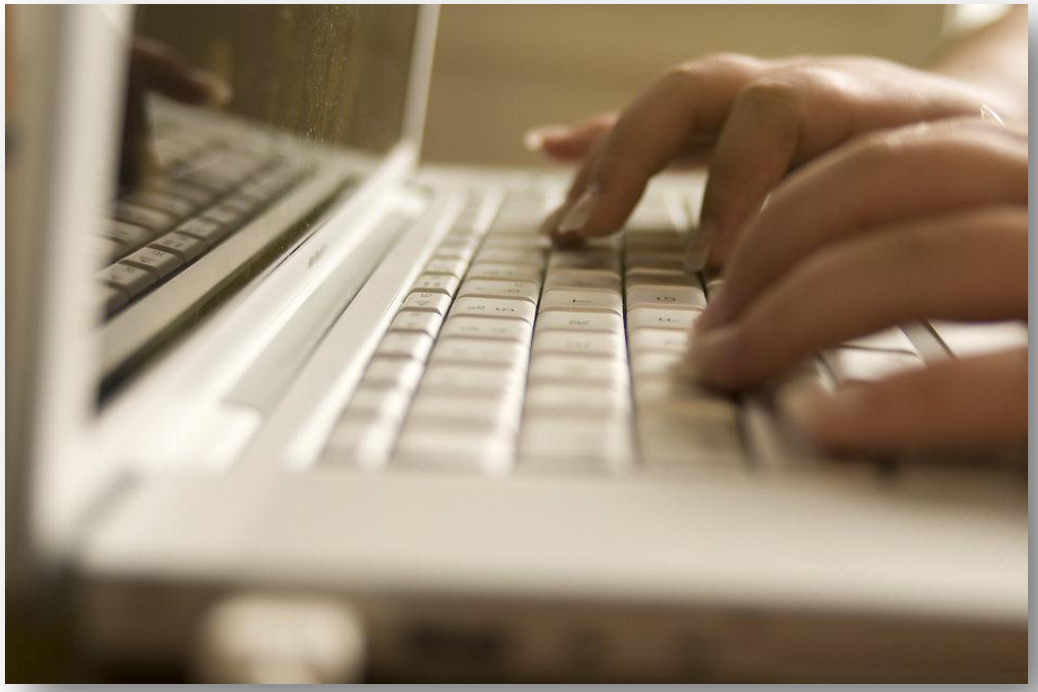
Time for a break …

11:45 – 12:30

Hands on part 4 12:30 – 13:30

“Visualizations of RNA-Seq experiments with Galaxy“

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



Setting up your Galaxy with Docker & Kitematic

Part 5 13:30 – 14:00

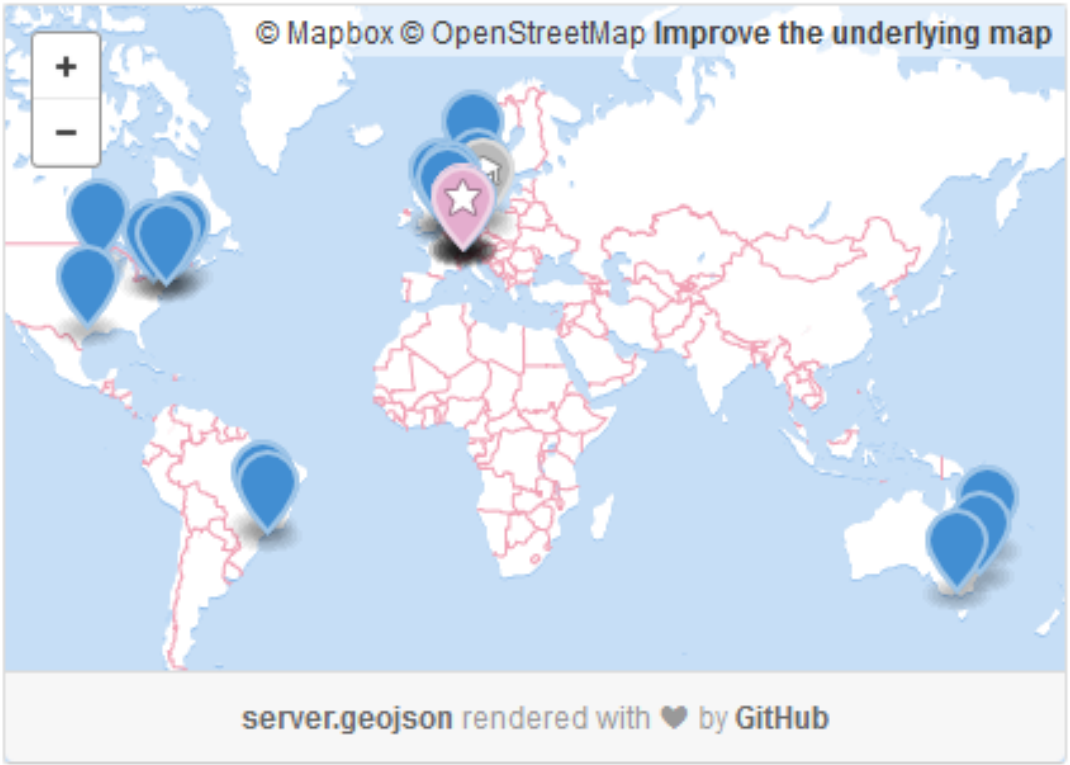
“Introduction into the underlying technology of Galaxy“

Supporting new data analysis approaches

- Key performance of Galaxy
  - Accessibilty
  - Reproducibility
  - Transparency

jupyter

jupyter.org/



docker

Galaxy

usegalaxy.org

elixir

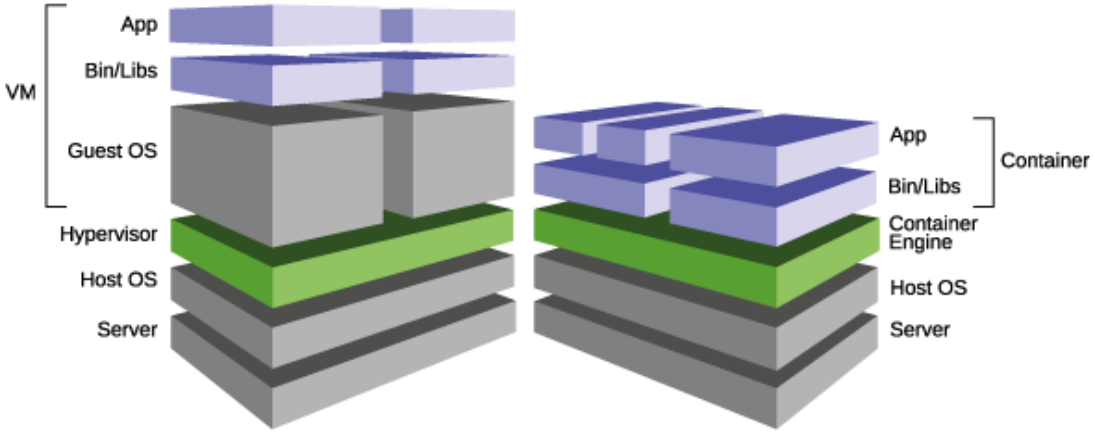
elixir-europe.org

de.NBI

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

denbi.de

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



One tool to share them all

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)

Example Dockerfile: TRAPLINE + Docker

```
FROM bgruening/galaxy-stable

MAINTAINER Markus Wolfien, markus.wolfien@gmail.com

ENV GALAXY_CONFIG_BRAND "TRAPLINE_160801"

WORKDIR /galaxy-central

RUN install-repository \
    "--url https://toolshed.g2.bx.psu.edu/ -o devteam --name fastq_groomer" \
    "--url https://toolshed.g2.bx.psu.edu/ -o devteam --name fastq_trimmer_by_quality" \
    "--url https://toolshed.g2.bx.psu.edu/ -o devteam --name fastx_clipper" \
    "--url https://toolshed.g2.bx.psu.edu/ -o devteam --name tophat_fusion_post" \
    "--url https://toolshed.g2.bx.psu.edu/ -o scottx611x --name tophat2_with_gene_annotations" \
    "--url https://toolshed.g2.bx.psu.edu/ -o devteam --name cufflinks" \
    "--url https://toolshed.g2.bx.psu.edu/ -o devteam --name cuffmerge" \
    "--url https://toolshed.g2.bx.psu.edu/ -o devteam --name cuffcompare" \

VOLUME ["/export/", "/data/", "/var/lib/docker"]

EXPOSE :80
EXPOSE :21
EXPOSE :8080

CMD ["/usr/bin/startup"]
```

Source Container

Tools to be added to the new Container

Galaxy out of the box (container)

Short *Kitematic* and *Docker-Galaxy* showcase

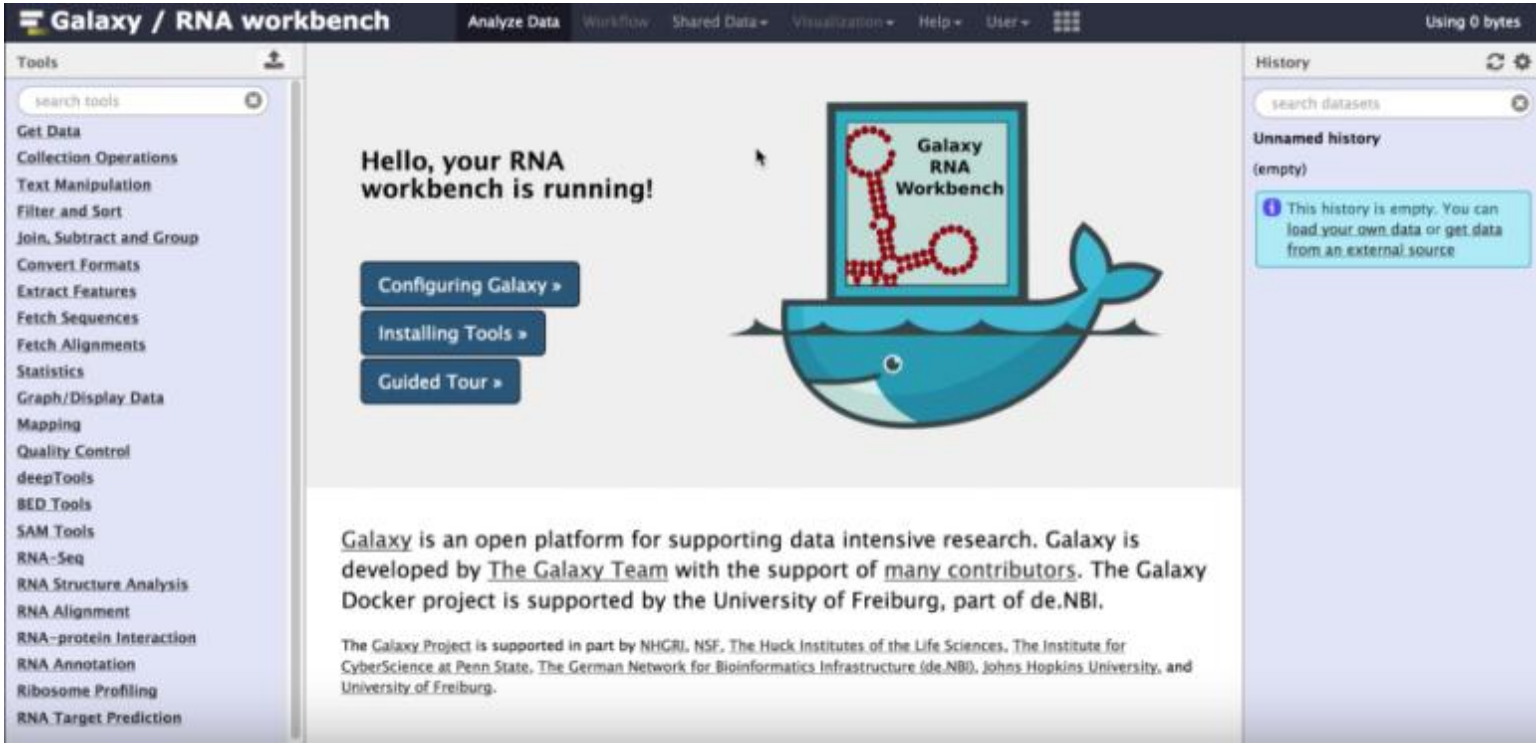
https://kitematic.com/

Containerization!
New workflow technologies eliminate
“works on my machine” problems

Supporting the RNA Galaxy-workbench

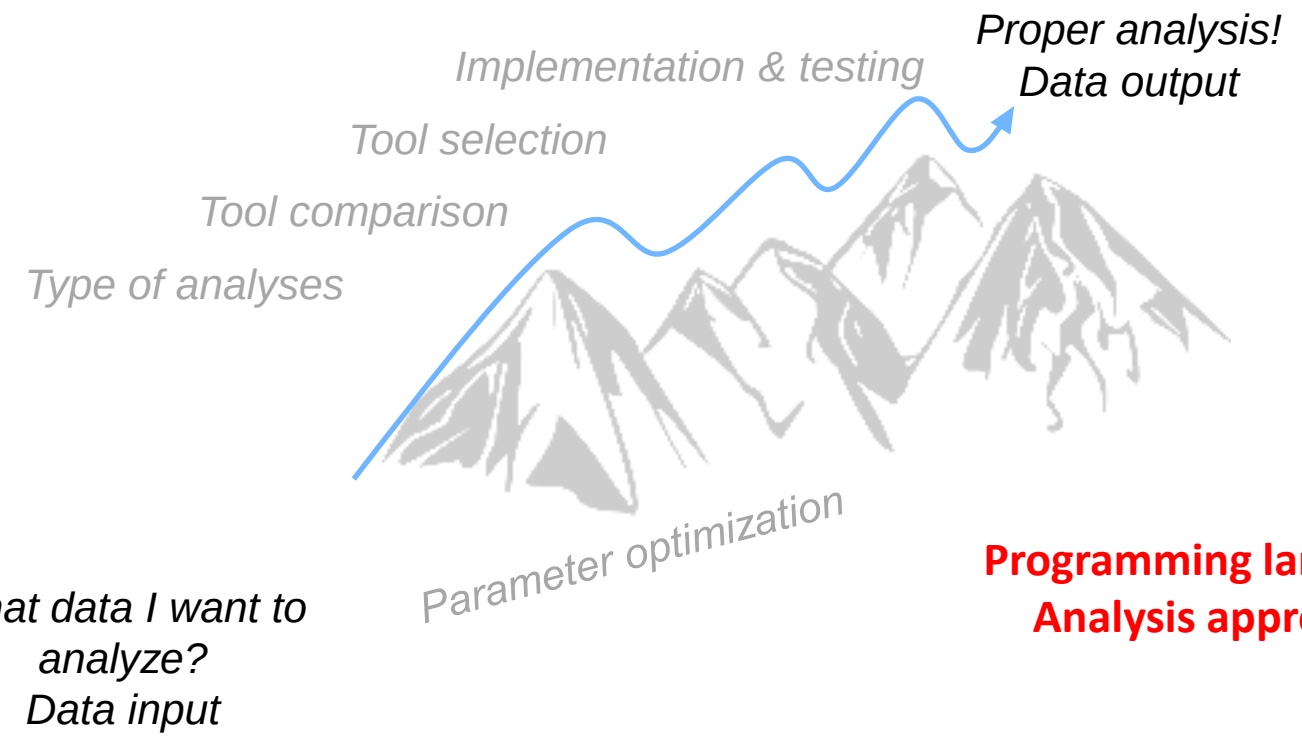
- Specialized Galaxy instance for RNA analyses povided by the RBC
- Contains +50 tools for structure analyses, annotation, alignment and many more

github.com/bgruening/galaxy-rna-workbench



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

The struggle for the right approaches



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

Transparent Reproducible Automated PipeLINE - TRAPLINE

Galaxy Modules implemented:

Data

↓

Pre-Processing

↓

Alignment

↓

Differential Expression (DE)

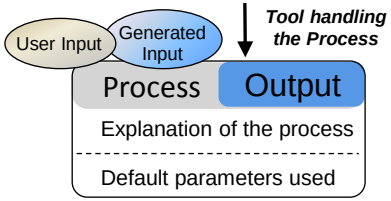
↓

Transcriptomic Analysis

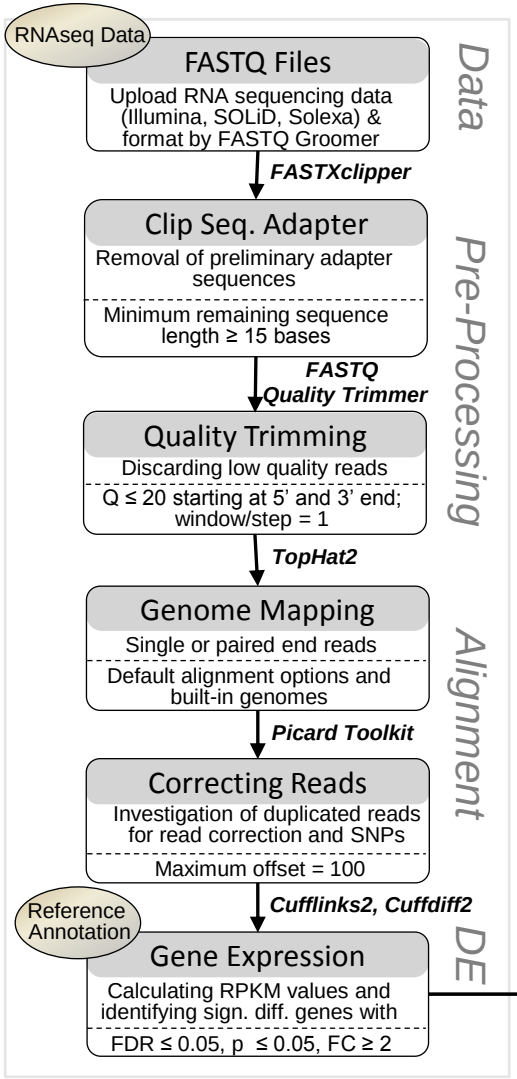
↓

Networks & Function

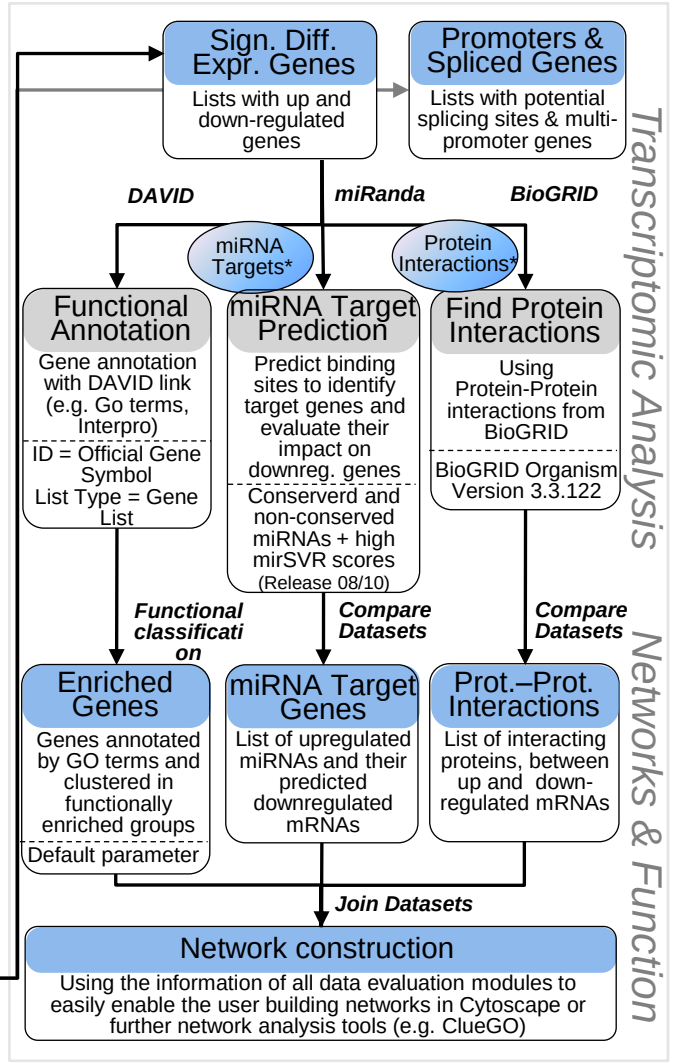
Legend:



Data Processing



Data Evaluation and Annotation



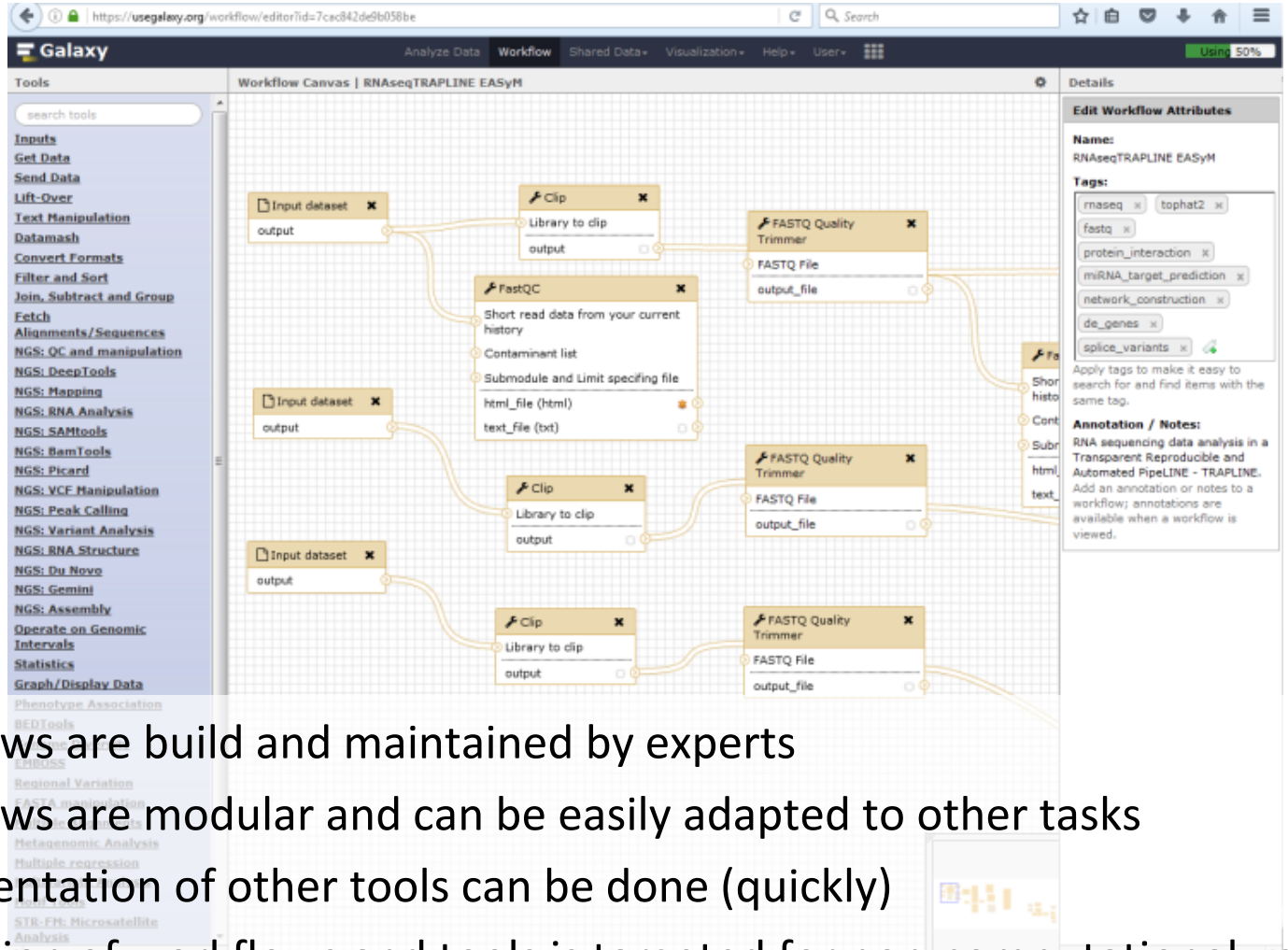
Wolfien *et al.* 2016, *BMC Bioinformatics*

Example Galaxy workflow: TRAPLINE.ga


```
"12": {
    "annotation": "",
    "content_id": "toolshed.g2.bx.psu.edu/repos/devteam/tophat2/tophat2/2.1.0",
    "id": 12,
    "input_connections": {
        "refGenomeSource|ownFile": {
            "id": 3,
            "output_name": "output"
        },
        "singlePaired|input": {
            "id": 8,
            "output_name": "trimmed_reads_paired_collection"
        }
    },
    "inputs": [
        {
            "description": "runtime parameter for tool TopHat",
            "name": "refGenomeSource"
        },
        {
            "description": "runtime parameter for tool TopHat",
            "name": "singlePaired"
        }
    ],
    "label": null,
    "name": "TopHat",
    "outputs": [
        {
            "name": "align_summary",
            "type": "txt"
        },
        {
            "name": "fusions",
            "type": "tabular"
        },
```

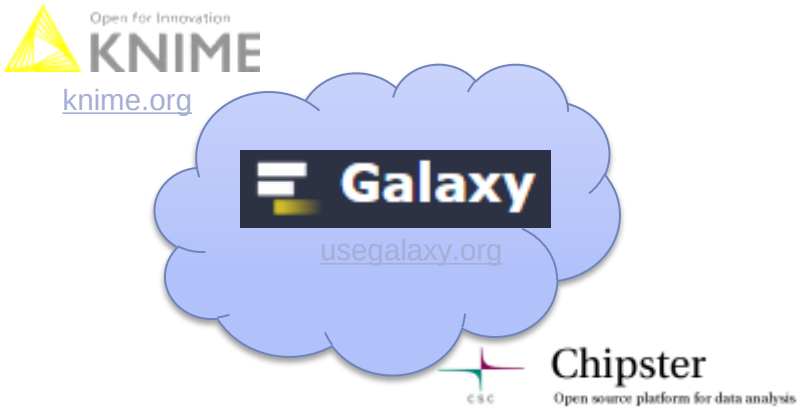

Specific xml file with tools, parameters and meta data!

Using workflow development



- 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

Medical “Big Data” and the need for new …



+133 different workflow management systems!

Almost no interoperability!

Need for a common line!

https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems

CWL - Open standard designed to express workflows and their tooling goups in YAML structured text files

- Common format for bioinformatics tool execution
  - Inputs & outputs are fully specified
- Community based standards effort, not a specific software package
- Designed for shared-nothing cluster & cloud environments
  - Tool executions are isolated from one another & from parent process
- Designed for containers (e.g. Docker, BioContainer)
- Well defined execution process:
  - 1. Collect & validate inputs
  - 2. Map input file paths to locations inside container
  - 3. Build tool command line
  - 4. Build Docker invocation
  - 5. Execute
  - 6. Collect & validate outputs

COMMON WORKFLOW LANGUAGE

commonwl.org

Example.yaml: samtools [sort]

```
class: CommandLineTool
cwlVersion: draft-3
description: Sort by chromosomal coordinates
```

File type and meta data

```
requirements:
        - class: DockerRequirement
        dockerPull: scidap/samtools:v1.2-216-gdffc67f
```

Runtime environment

```
inputs:
        - id: input
        type: File
        inputBinding:
                position: 1
        - id: output_name
        type: string
        inputBinding:
                position: 2
```

Input parameters

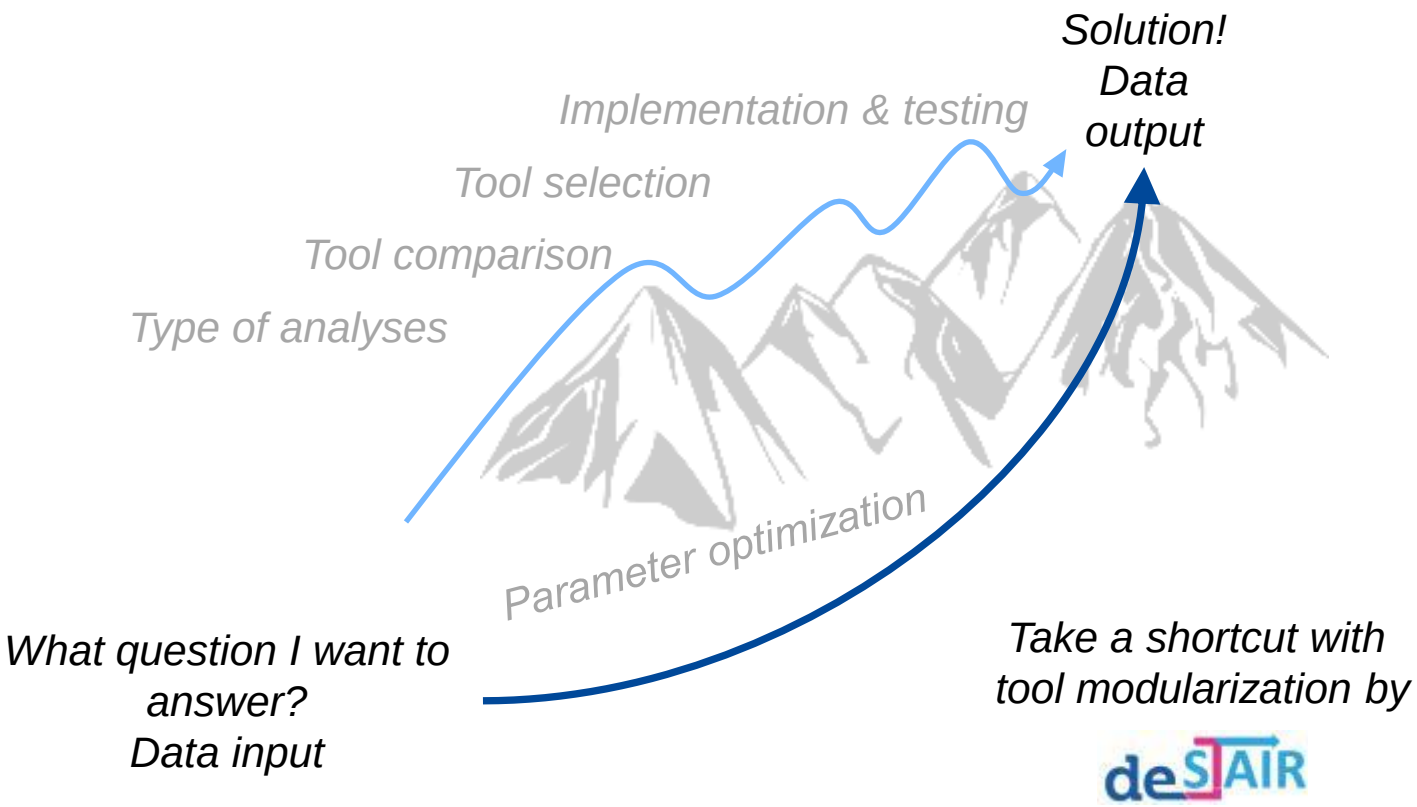
```
outputs:
        - id: output
        type: File
        outputBinding:
                glob: $(inputs.output_name)
```

Output parameters

```
baseCommand: [samtools, sort]
```

Executable

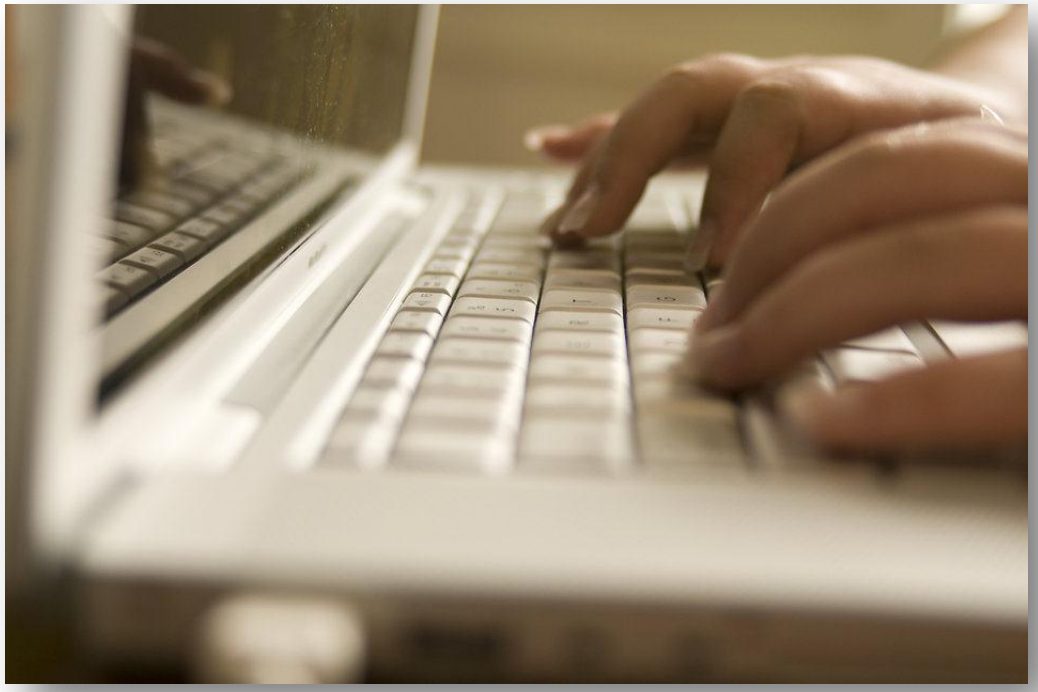
Why using workflows?



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

Hands on part 6 14:15 – 15:45

“Clinical use case for RNA-Seq, combining all previous processing steps and linking results to further resources“



Our use case for today

www.pinkribbon-deutschland.de

Most common cancer in women worldwide

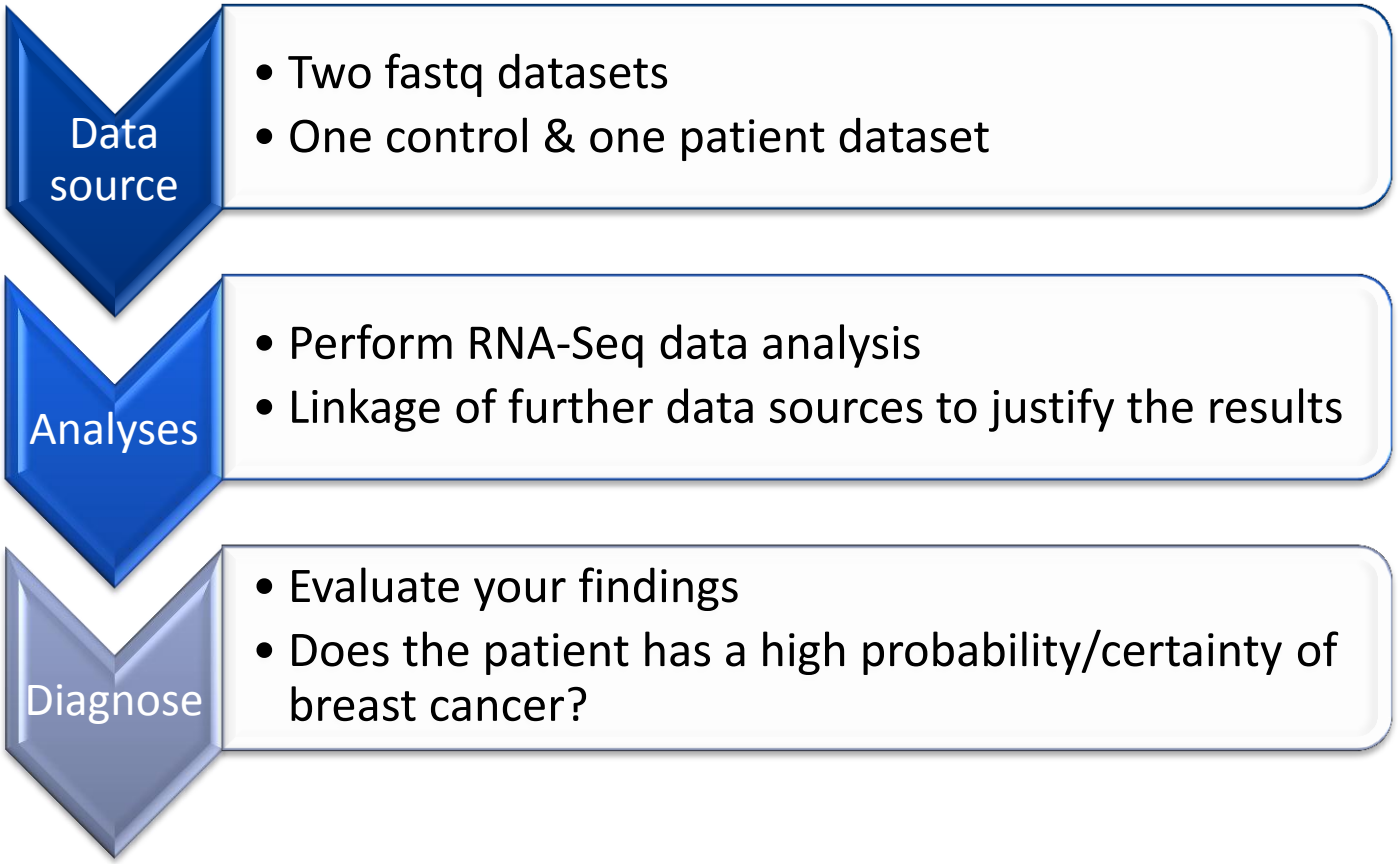
Leading cause of death from cancer in women worldwide

Predictive factors that identify a benefit

Many different variations and subtypes

Many different therapeutically approaches

Our use case for today – breast cancer screening



Get your data set!

https://usegalaxy.org/u/mwolfien/h/rna-seq-workshop-kiel

Basic workflow for data processing

Data

Pre-Processing (Quality Control, Clipping)

Genomic Alignment

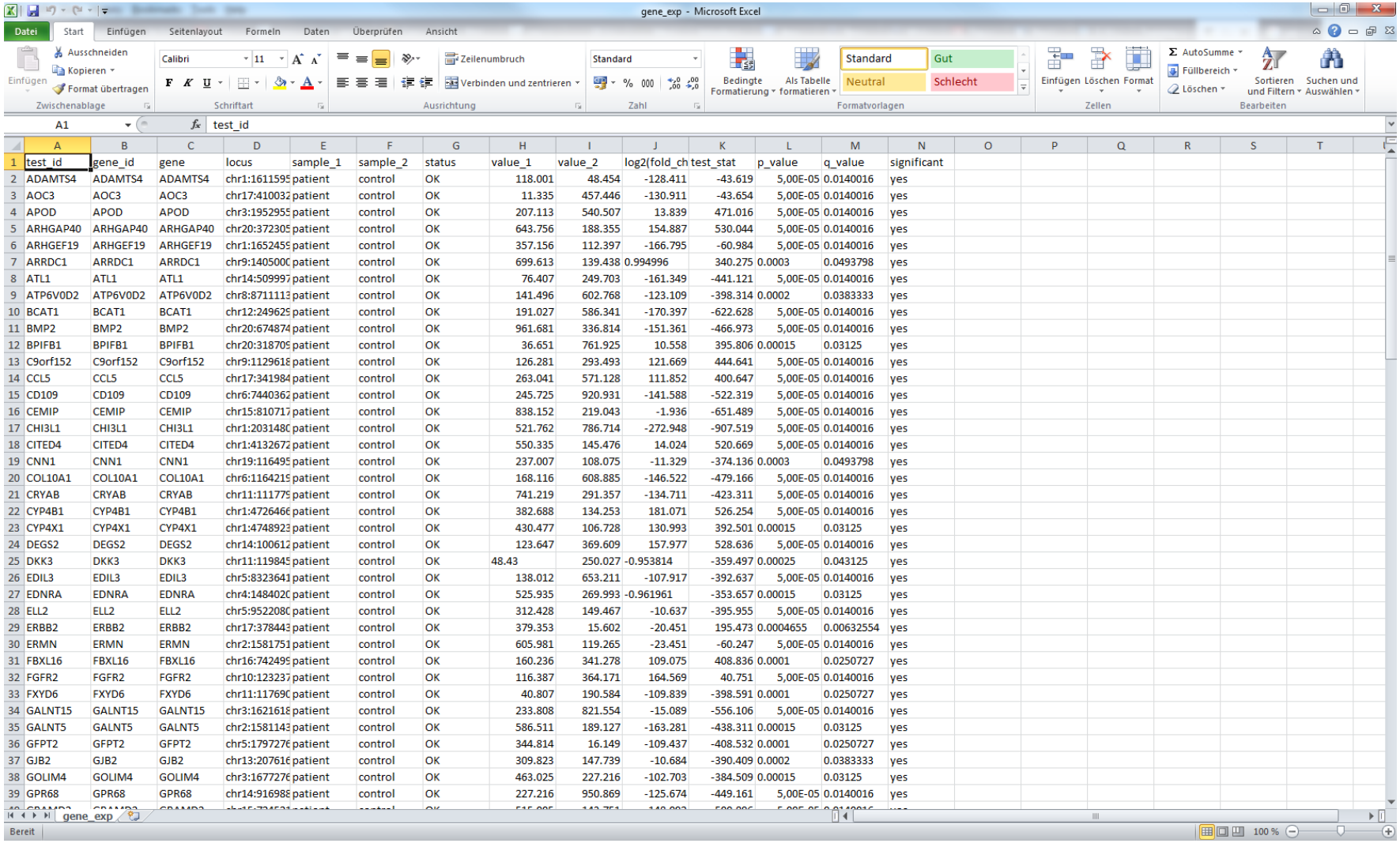
Transcript Quantification

Differentially Expressed Transcripts

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

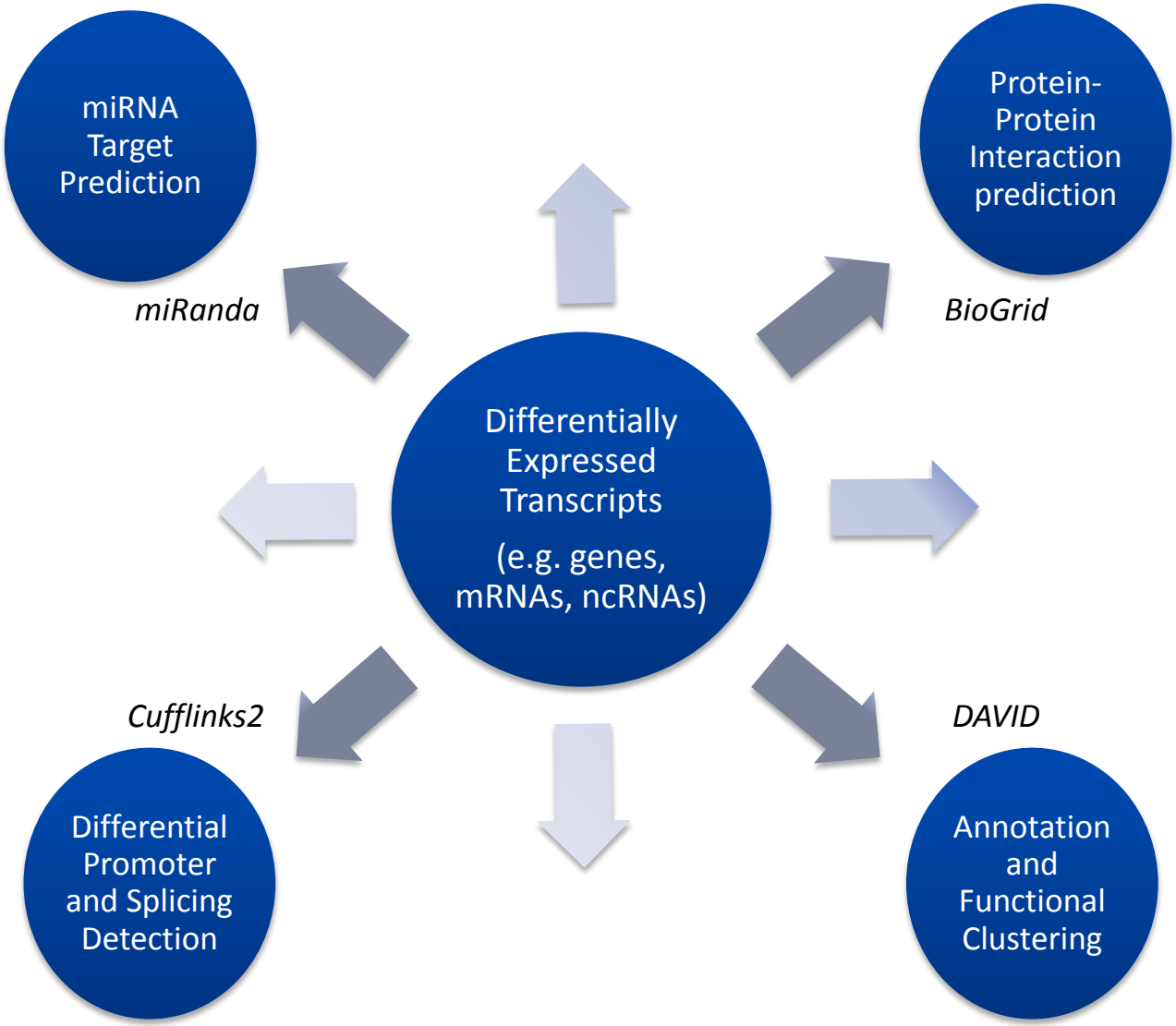
- Check RPKM Normalization
- Bias Correction
- Splicing detection

Explore your data!



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	0.03125	yes
C9orf152	C9orf152	C9orf152	chr9:1129618	patient	control	OK	126.281	293.493	121.669	444.641	5,00E-05	0.0140016	yes
CCL5	CCL5	CCL5	chr17:341984	patient	control	OK	263.041	571.128	111.852	400.647	5,00E-05	0.0140016	yes
CD109	CD109	CD109	chr6:7440362	patient	control	OK	245.725	920.931	-141.588	-522.319	5,00E-05	0.0140016	yes
CEMIP	CEMIP	CEMIP	chr15:810717	patient	control	OK	838.152	219.043	-1.936	-651.489	5,00E-05	0.0140016	yes
CHI3L1	CHI3L1	CHI3L1	chr1:203148C	patient	control	OK	521.762	786.714	-272.948	-907.519	5,00E-05	0.0140016	yes
CITED4	CITED4	CITED4	chr1:4132672	patient	control	OK	550.335	145.476	14.024	520.669	5,00E-05	0.0140016	yes
CNN1	CNN1	CNN1	chr19:116495	patient	control	OK	237.007	108.075	-11.329	-374.136	0.0003	0.0493798	yes
COL10A1	COL10A1	COL10A1	chr6:1164215	patient	control	OK	168.116	608.885	-146.522	-479.166	5,00E-05	0.0140016	yes
CRYAB	CRYAB	CRYAB	chr11:111775	patient	control	OK	741.219	291.357	-134.711	-423.311	5,00E-05	0.0140016	yes
CYP4B1	CYP4B1	CYP4B1	chr1:472646C	patient	control	OK	382.688	134.253	181.071	526.254	5,00E-05	0.0140016	yes
CYP4X1	CYP4X1	CYP4X1	chr1:4748923	patient	control	OK	430.477	106.728	130.993	392.501	0.00015	0.03125	yes
DEGS2	DEGS2	DEGS2	chr14:100612	patient	control	OK	123.647	369.609	157.977	528.636	5,00E-05	0.0140016	yes
DKK3	DKK3	DKK3	chr11:119845	patient	control	OK	48.43	250.027	-0.953814	-359.497	0.00025	0.043125	yes
EDIL3	EDIL3	EDIL3	chr5:8323641	patient	control	OK	138.012	653.211	-107.917	-392.637	5,00E-05	0.0140016	yes
EDNRA	EDNRA	EDNRA	chr4:148402C	patient	control	OK	525.935	269.993	-0.961961	-353.657	0.00015	0.03125	yes
ELL2	ELL2	ELL2	chr5:952208C	patient	control	OK	312.428	149.467	-10.637	-395.955	5,00E-05	0.0140016	yes
ERBB2	ERBB2	ERBB2	chr17:37844	patient	control	OK	379.353	15.602	-20.451	195.473	0.0004655	0.00632554	yes
ERMN	ERMN	ERMN	chr2:1581751	patient	control	OK	605.981	119.265	-23.451	-60.247	5,00E-05	0.0140016	yes
FBXL16	FBXL16	FBXL16	chr16:742495	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/)

DisGeNET

Home About Search Browser Downloads Cytoscape RDF Help

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.

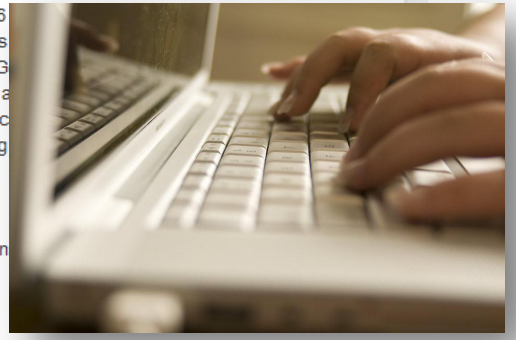
DisGeNET database has been cited by several papers. Some of them can be reviewed here.

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

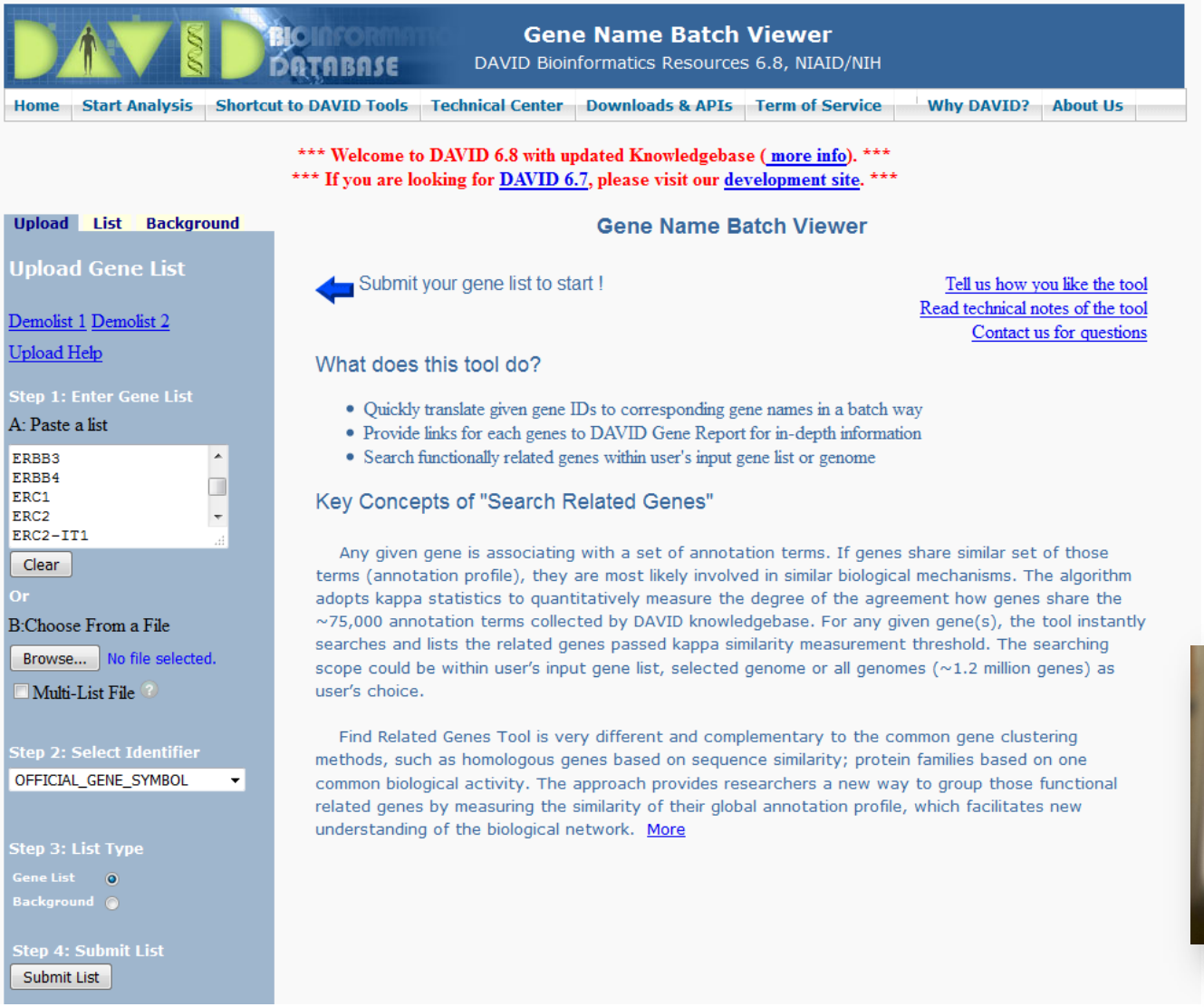
DisGeNET
@DisGeNET

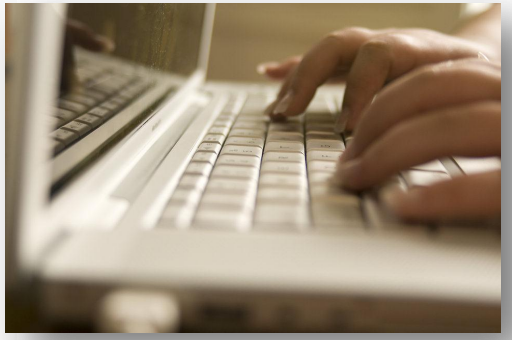
Check out the new publication describing the DisGeNET platfotm in NAR database issue nar.oxfordjournals.org/con



Linking and integrating data

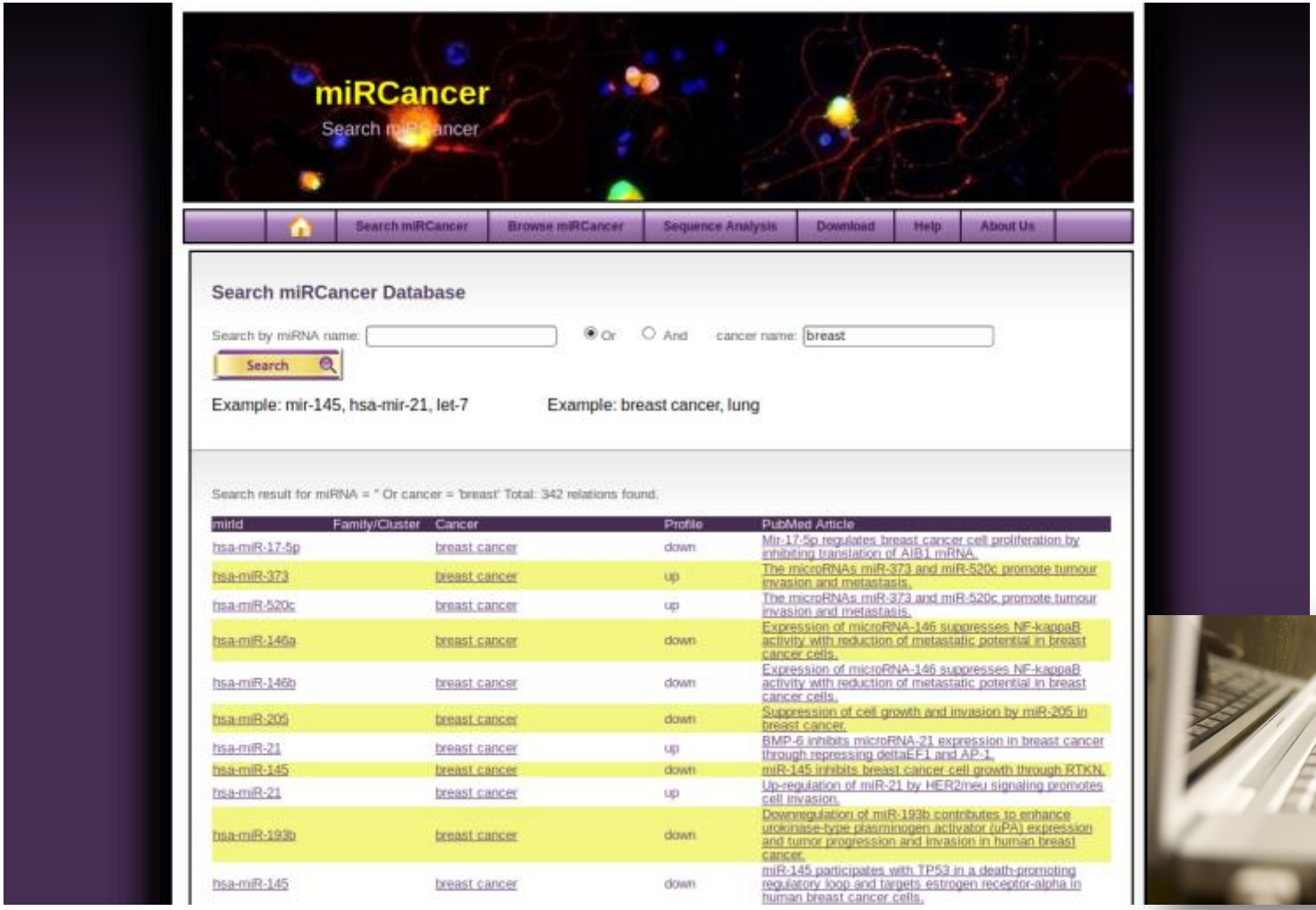
- David (https://david.ncifcrf.gov/list.jsp)

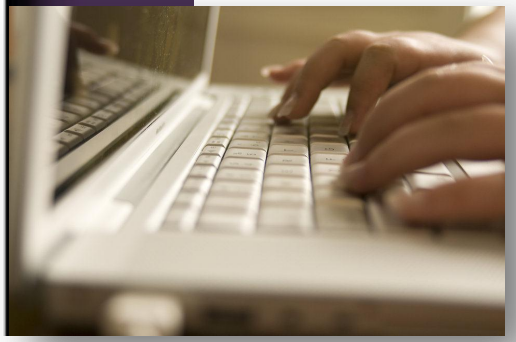




Linking and integrating data

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





Linking and integrating data

miRNA	Regulation	Target
140-5b	up	
148b	Down	
150	Up	
106b	Up	
143	Down	
19b	Up	
21	up	
…	…	…

Explore miRNA cooperativity to justify diagnosis

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

TriplexRNA

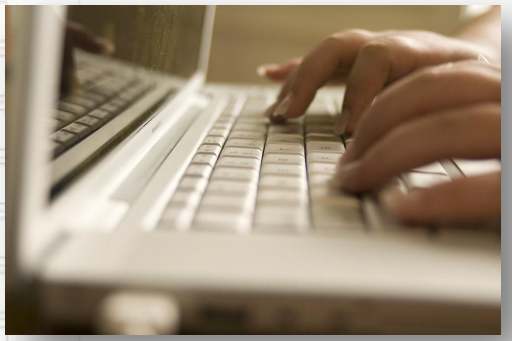
A database of cooperating microRNAs and their mutual targets

Search targets of synergistic microRNA regulation

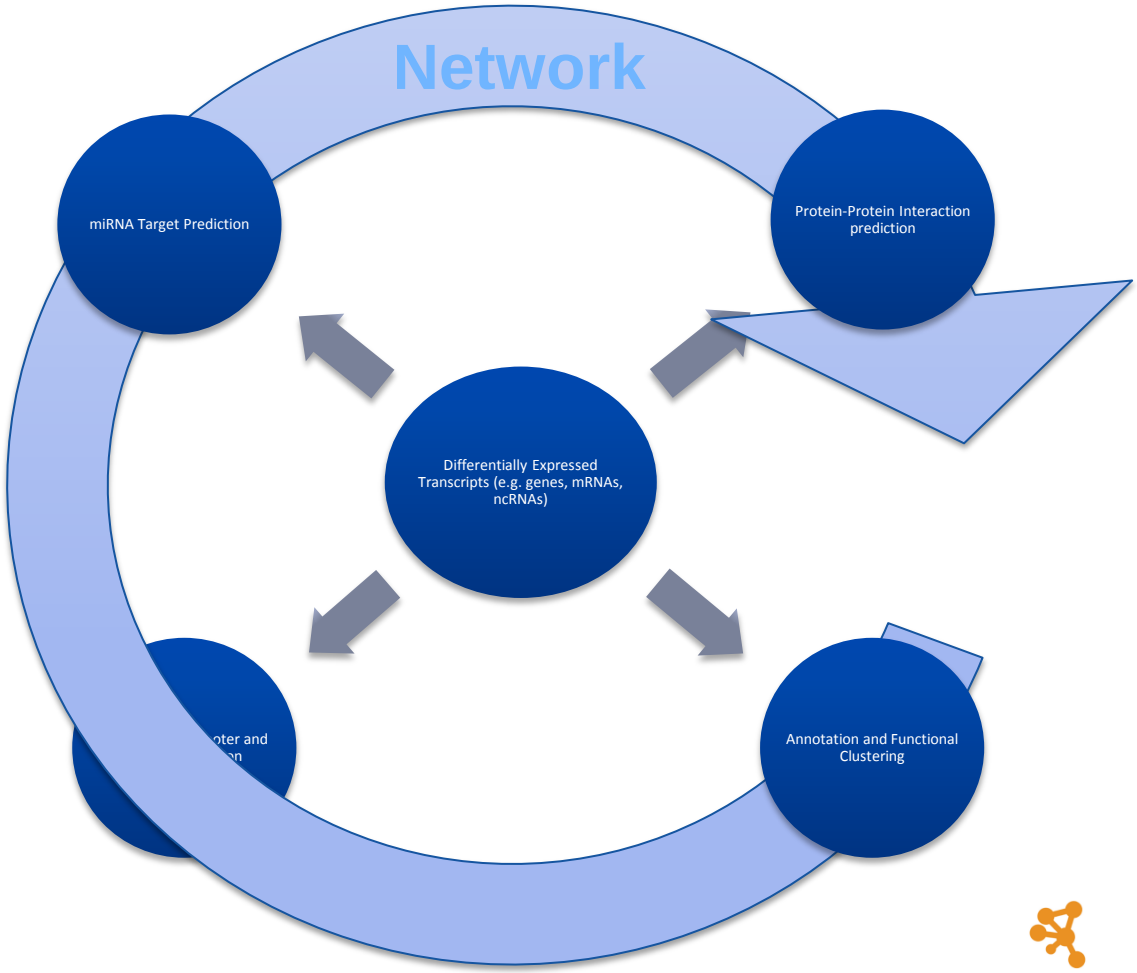
Search in: Human | for: miRNA ID | hsa-miR-140-5p

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



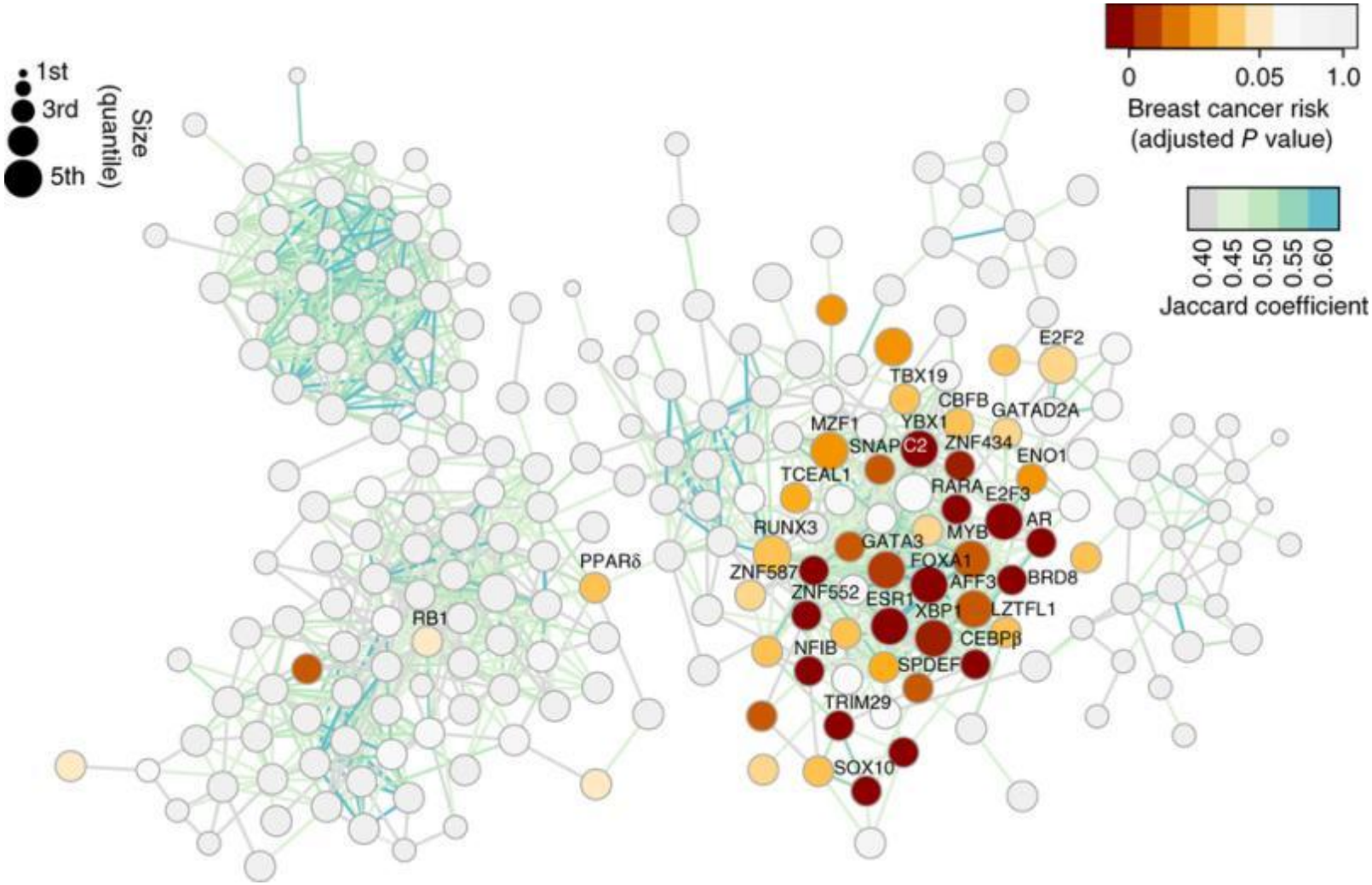
There is nothing more practical than a network



Cytoscape

Vanted / CellDesigner

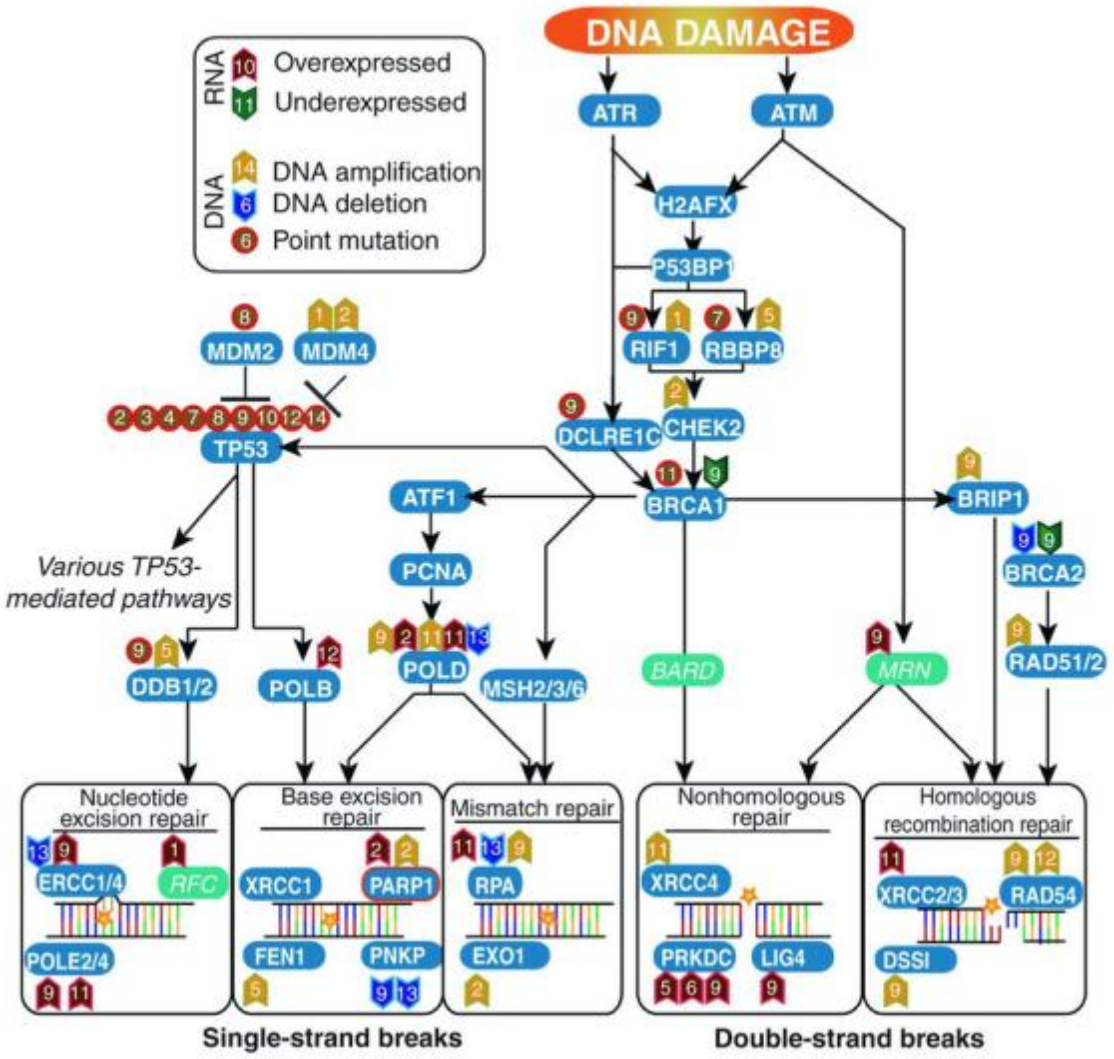
Network comparisons also reveal differences



Castro, *Nat. Gen*, 2016

Cytoscape

What else is done in this field with NGS?



- Craig, D. W. et al. Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. Mol. Cancer Ther. 12 , 104–116 (2013). One of the first papers investigating integration of whole-transcriptome sequencing and genome sequencing for targeted therapy selection in advanced metastatic triple-negative breast cancer

Does the patient has a high risk of cancer

Patient

Our implementation strategy

Unexplained symptoms

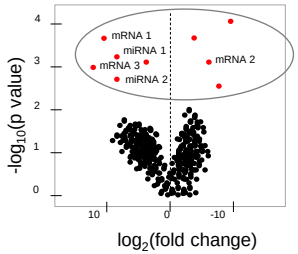
ATCATG101000111
TAGTG101010101111
TTCATAA00011011
AAGCA1010011

Bioinformatics
Automated and scalable data processing, e.g. TRAPLINE [2]

docker Galaxy rkt

Clinic
Improved diagnosis, prognosis and therapy

Statistics
Evaluation of omics data (RPKM, FDR, Bayes, …)



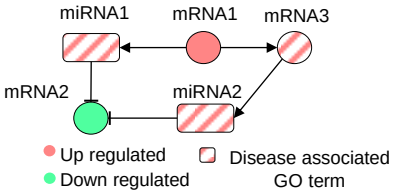
101000111ATCATG
101010101111TTCATAA
00011011AAGCA
1010011TAGTG

Med. Informatics
Data management

Machine learning

HUB FAIRDOM

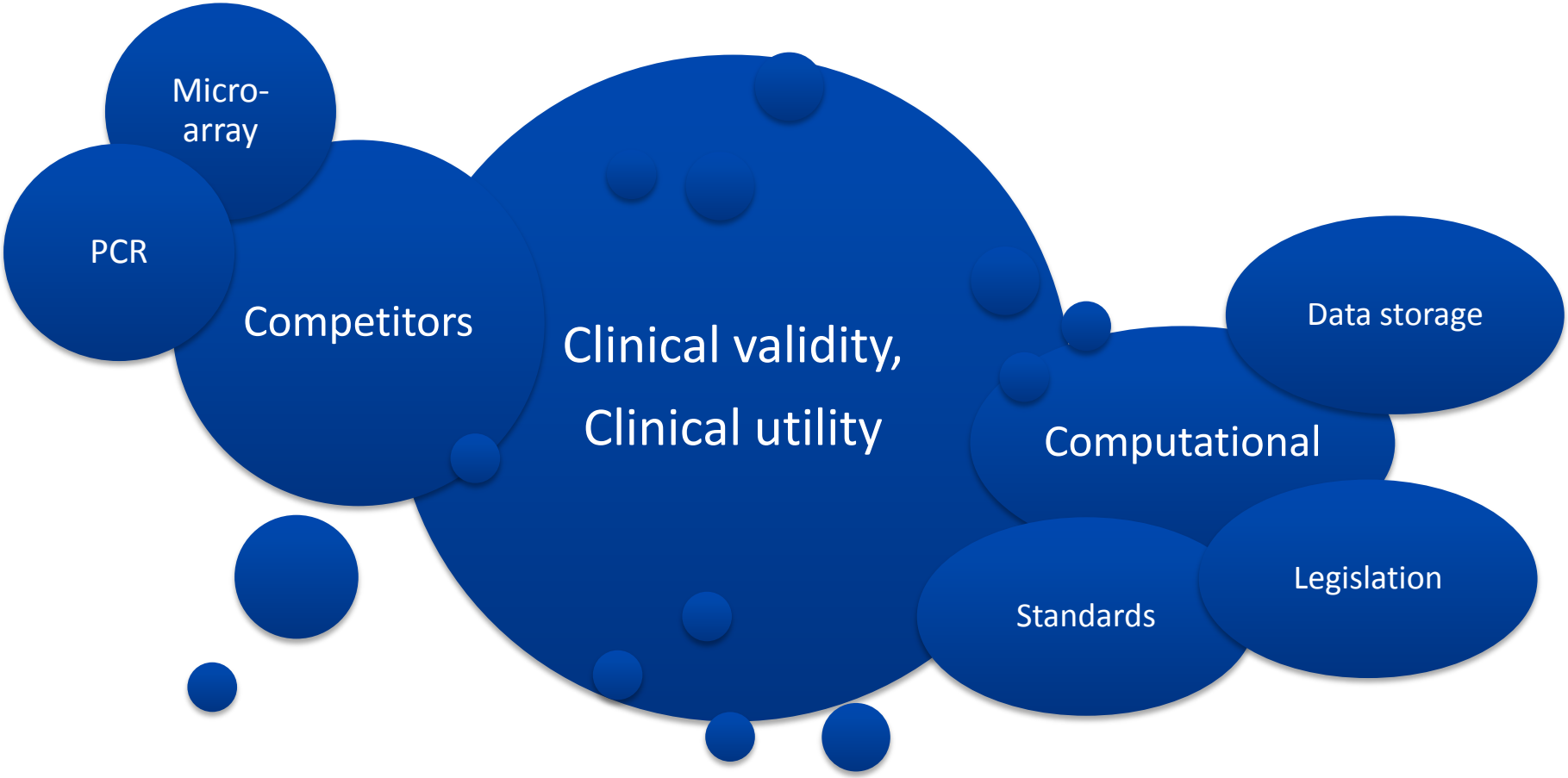
Systems Biology
Data integration (DisGeNet, miRCancer, TriplexRNA, …)



New insights

Steinhoff, *ADDR*, 2017

Challanges for a broader usage of NGS in the clinic



Future perspectives of NGS in the clinic

- *"With its unprecedented ability to simultaneously detect global gene transcript levels and diverse RNA species, RNA-seq has the potential to revolutionize clinical testing for a wide range of diseases."* Byron *et al*., *Nat. Rev. Genet*., 2016

  - "*Once the discovery phase is complete, many diagnostic tests will become targeted assays, sensitive enough to detect small numbers of rare transcripts*."

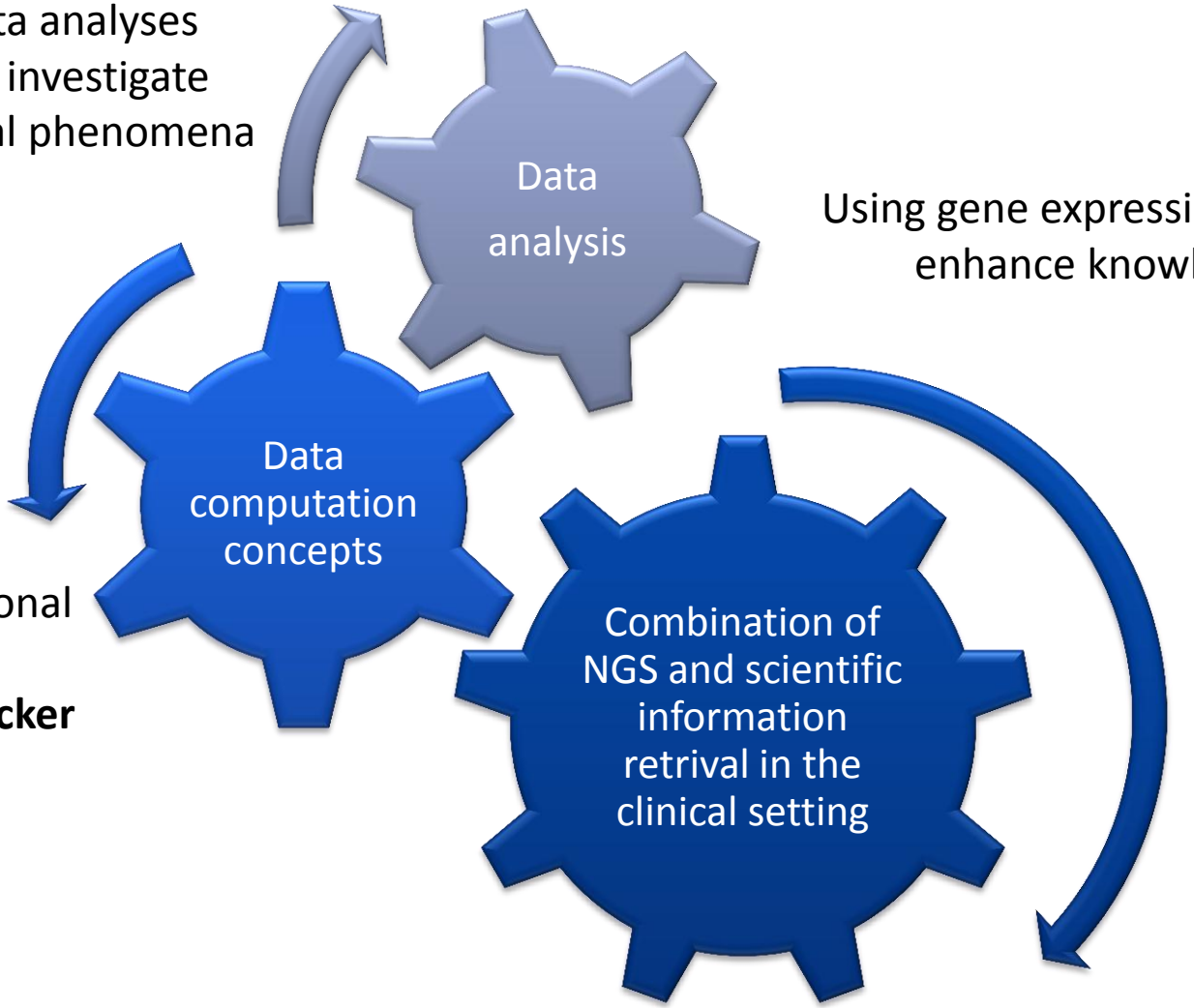
    Andersson *et al., Nat. Genet*., 2015

- "*Feed in latest scientific findings and analyze the same dataset over and over again* […] ". Comment on crowdsourced research in Medicine (*Nature*)

  - "*Value of incorporating RNA sequencing (RNA-seq) with DNA sequencing to evaluate the expression of mutant alleles, to detect both known and novel gene fusions, and to detect splice variants.*" Robinson *et al., Cell*, 2015

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**

Part 7 16:00 – 16:30

“Open discussions about RNA-Seq projects from the participants“

- What data do you have?
- What do you want to analyze?
- Do you feel better informed?

Do you want more NGS? - Further training courses

www.denbi.de/index.php/training-courses

Home Mission Organization Network Services Training Cloud Events News Jobs Help

Training Courses 2018

Training Courses 2018

Training Archive sorted according to date

Training Archive sorted according to de.NBI units

Online training & Media library

de.NBI Youtube channel

2018	Topic	Location
05-08 Mar	de.NBI Winter School on Computational Metabolomics	Lutherstadt Wittenberg
06 Mar	Introduction to genome-wide association studies (GWAS)	Kiel
07 Mar	Introduction to RNA-Seq data analysis with Galaxy	Kiel
07-13 Mar	Applied Metaproteomics Workshop 2018	Magdeburg
08 Mar	Getting Started with the de.NBI-Cloud	Bielefeld
09 Mar	Analysis of Mass Spectrometry and Sequence Data with KNIME - KNIME Spring Summit	Berlin
11 Mar	Computational Proteomics Workshop - DGMS2018	Saarbrücken
11-18 Mar	SegAn Developer Retreat	Mallorca
27-29 Mar	3rd de.NBI Training Course on Metagenome Analysis	Bielefeld
02-06 Apr	OpenMS Developer Meeting 2018	Troia
04-05 Apr	Introduction to Python Programming	Heidelberg
11-12 Apr	Introduction to BRENDA & ProteinsPlus 2018	Braunschweig
23-25 Apr	4th de.NBI Genomics training course	Gießen
24-26 Apr	Tools for Systems biology modeling and data exchange: COPASI, CellNetAnalyzer, SABIO-RK, SEEK	Magdeburg

Do you want more NGS? - Further training courses

Welcome to Galaxy Training!

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

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy	13
Assembly	3
ChIP-Seq data analysis	2
Epigenetics	2
Metagenomics	2
Proteomics	8
Sequence analysis	6
Transcriptomics	5
Variant Analysis	6

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	8
Development in Galaxy	14
Train the trainers	6

Evaluations

Evaluation at :

https://de.surveymonkey.com/r/denbi-course?sc=rbc&id=000111

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:

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
denbi.de

elixir
elixir-europe.org

Bundesministerium
für Bildung
und Forschung
bmbf.de

We hope you enjoyed the training!

Universität Rostock
Traditio et Innovatio

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

Galaxy Training Network