Customized workflow development and data integration concepts in Systems Medicine

Markus Wolfien

GMDS Workshop "Methods in Sysems Medicine"

www.sbi.uni-rostock.de

Universität Rostock

Traditio et Innovatio

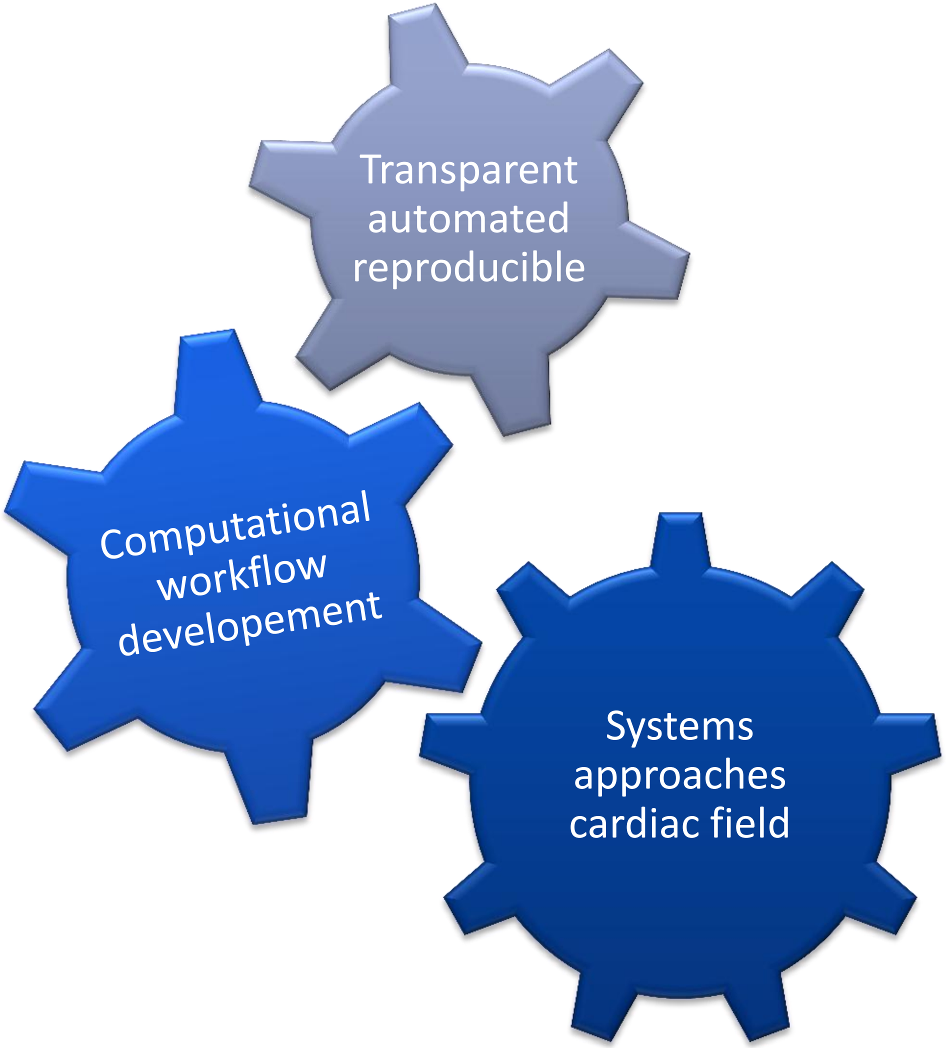
SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

Objectives for my talk

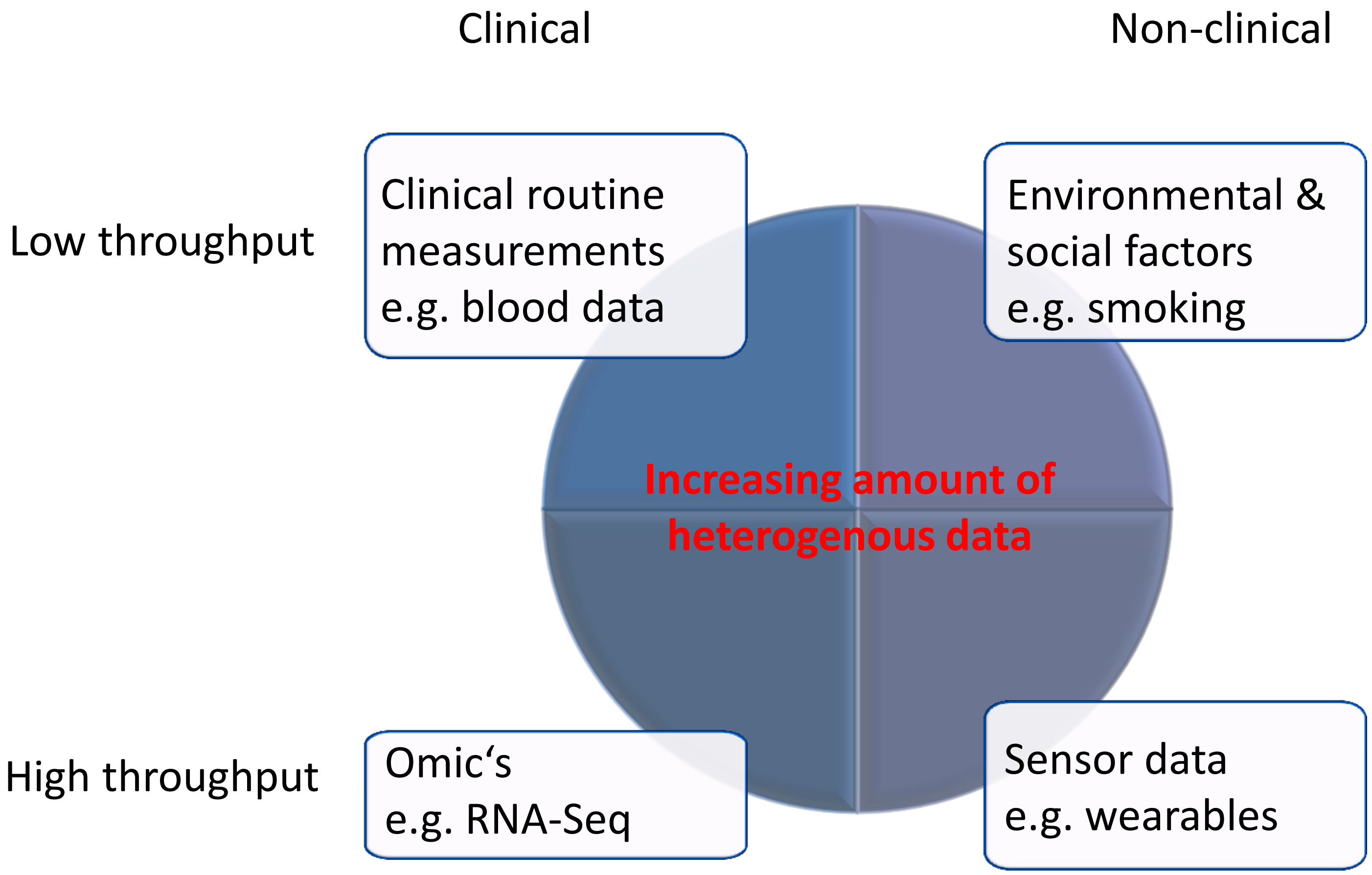
- Why using workflows?

- How are workflows being developed?

- What can be done to analyze and integrate diverse data?


Transparent automated reproducible
Computational workflow developement
Systems approaches cardiac field

Data being generated is steadily increasing



Clinical

Non-clinical

Low throughput

Clinical routine measurements
e.g. blood data

Environmental & social factors
e.g. smoking

Increasing amount of heterogenous data

High throughput

Omic‘s
e.g. RNA-Seq

Sensor data
e.g. wearables

The struggle for the right approaches

What data I want to analyze?
Data input

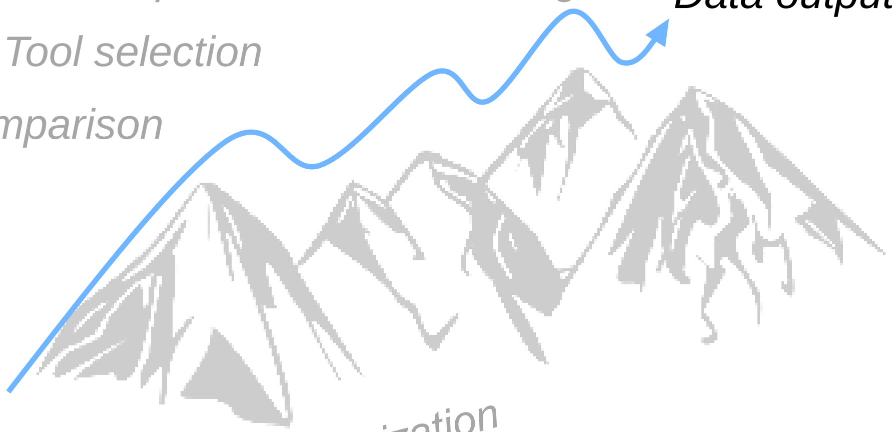
Type of analyses

Tool comparison

Tool selection

Implementation & testing

Proper analysis!
Data output



Parameter optimization

Programming language?
Analysis approach?

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

Workflows!

Provide an infrastructure to set up, execute and monitor tool environments

Medical “Big Data” and the need for new analyses

GenePattern

broadinstitute.org

GeneProf

geneprof.org

Grape

big.crg.cat/services/grape

Chipster

Open source platform for data analysis

chipster.csc.fi

python™

python.org

R

mapman.gabipd.org

Galaxy

usegalaxy.org

Open for Innovation KNIME

knime.org

R

r-project.org/

GeneTalk

gene-talk.de

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

bioconductor.org

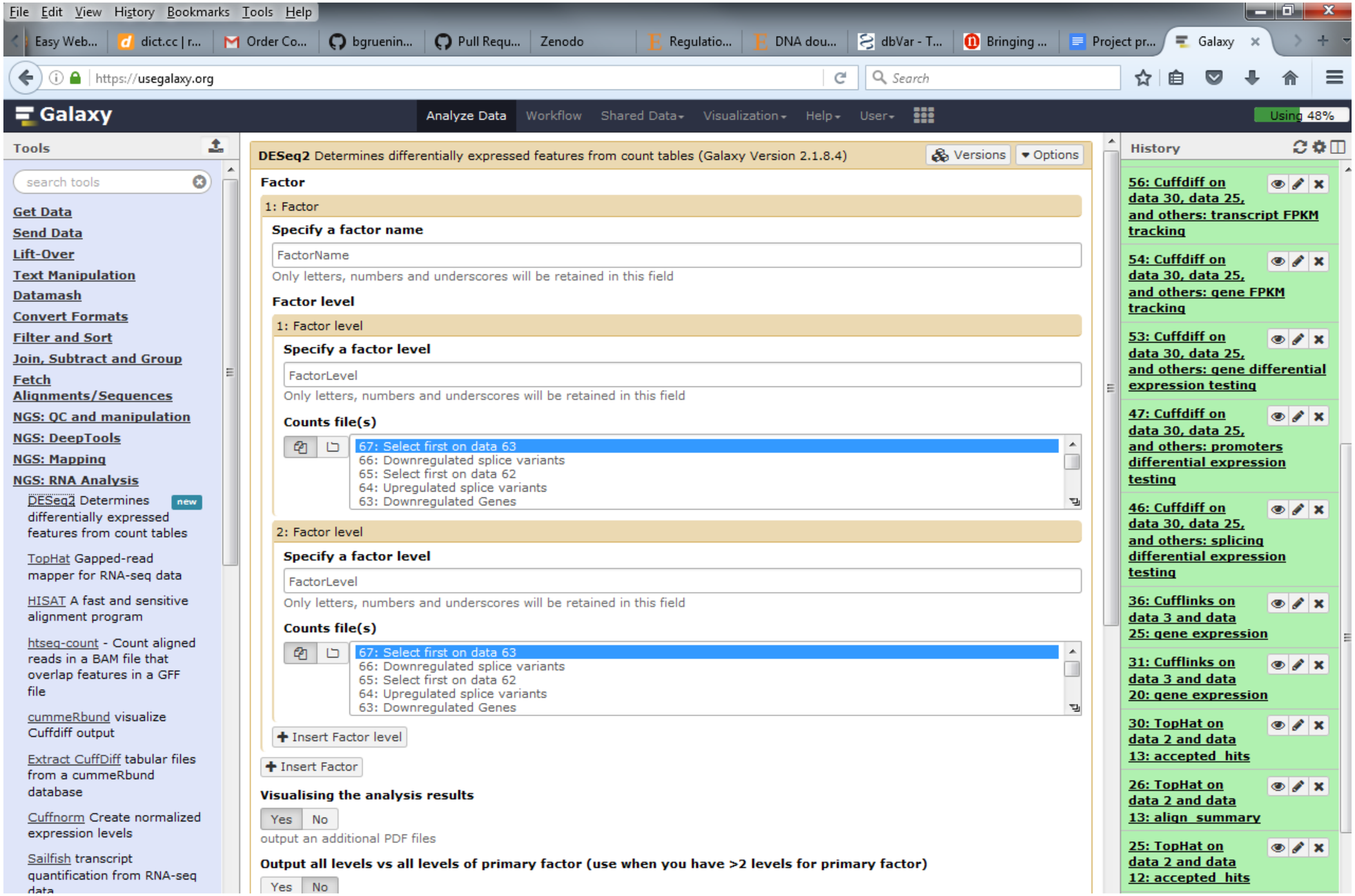
geneXplain

genexplain.com

BaseSpace®

illumina.com

The Galaxy interface and tours: usegalaxy.org


File Edit View History Bookmarks Tools Help
Easy Web... dict.cc | r... Order Co... bgruenin... Pull Requ... Zenodo Regulatio... DNA dou... dbVar - T... Bringing ... Project pr... Galaxy
https://usegalaxy.org
Search
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 48%
Tools
search tools
History
Versions
Options

Get Data
Send Data
Lift-Over
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
NGS: RNA Analysis

DESeq2 Determines differentially expressed features from count tables (new)

TopHat Gapped-read mapper for RNA-seq data

HISAT A fast and sensitive alignment program

htseq-count - Count aligned reads in a BAM file that overlap features in a GFF file

cummeRbund visualize Cuffdiff output

Extract CuffDiff tabular files from a cummeRbund database

Cuffnorm Create normalized expression levels

Sailfish transcript quantification from RNA-seq data

DESeq2 Determines differentially expressed features from count tables (Galaxy Version 2.1.8.4)

Factor

1: Factor

Specify a factor name

FactorName

Only letters, numbers and underscores will be retained in this field

Factor level

1: Factor level

Specify a factor level

FactorLevel

Only letters, numbers and underscores will be retained in this field

Counts file(s)

67: Select first on data 63
66: Downregulated splice variants
65: Select first on data 62
64: Upregulated splice variants
63: Downregulated Genes

2: Factor level

Specify a factor level

FactorLevel

Only letters, numbers and underscores will be retained in this field

Counts file(s)

67: Select first on data 63
66: Downregulated splice variants
65: Select first on data 62
64: Upregulated splice variants
63: Downregulated Genes

+ Insert Factor level

+ Insert Factor

Visualising the analysis results

Yes No

output an additional PDF files

Output all levels vs all levels of primary factor (use when you have >2 levels for primary factor)

Yes No

56: Cuffdiff on data 30, data 25, and others: transcript FPKM tracking

54: Cuffdiff on data 30, data 25, and others: gene FPKM tracking

53: Cuffdiff on data 30, data 25, and others: gene differential expression testing

47: Cuffdiff on data 30, data 25, and others: promoters differential expression testing

46: Cuffdiff on data 30, data 25, and others: splicing differential expression testing

36: Cufflinks on data 3 and data 25: gene expression

31: Cufflinks on data 3 and data 20: gene expression

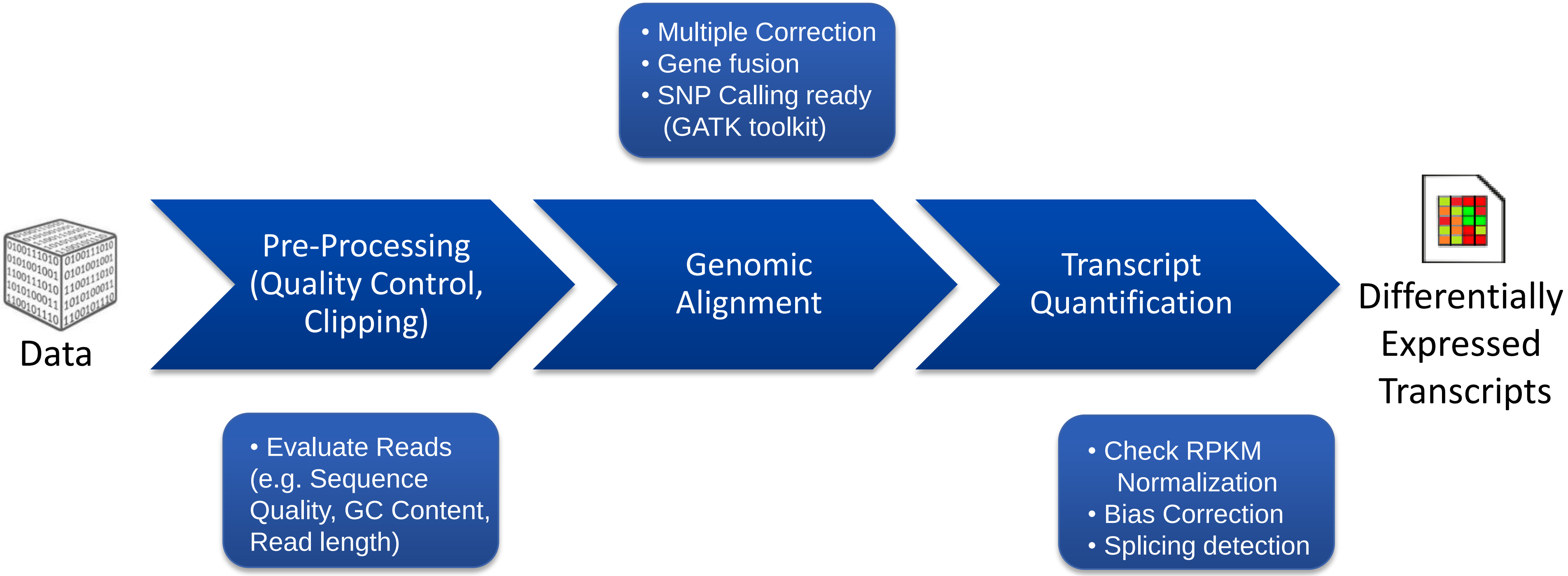
30: TopHat on data 2 and data 13: accepted_hits

26: TopHat on data 2 and data 13: align_summary

25: TopHat on data 2 and data 12: accepted_hits

Basic workflow for data processing

Example: RNA-Seq data analysis



Data

Pre-Processing (Quality Control, Clipping)

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

Genomic Alignment

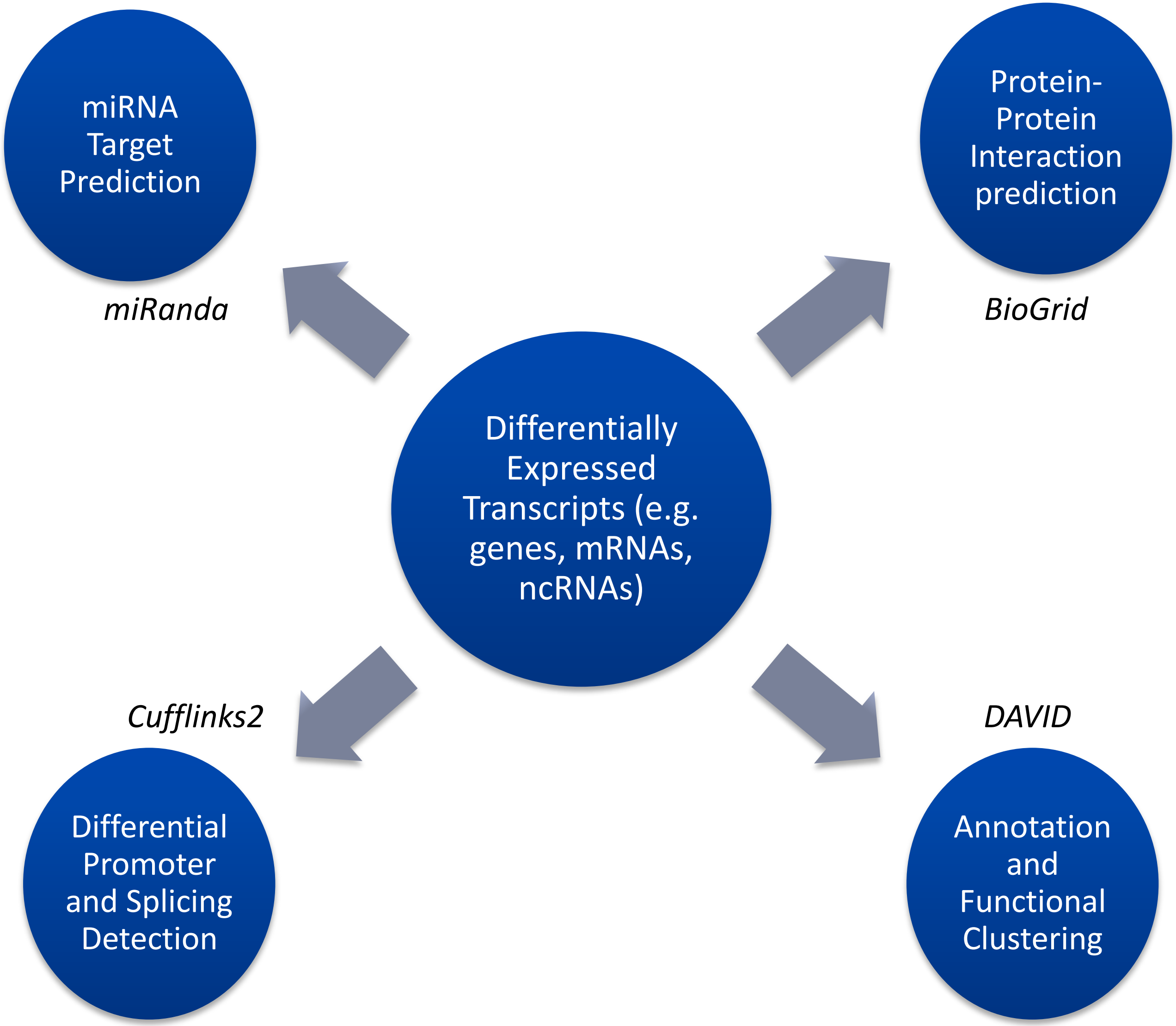
- Multiple Correction
- Gene fusion
- SNP Calling ready (GATK toolkit)

Transcript Quantification

- Check RPKM Normalization
- Bias Correction
- Splicing detection

Differentially Expressed Transcripts

Differential expression is the key of evaluation



miRNA Target Prediction

miRanda

Protein-Protein Interaction prediction

BioGrid

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

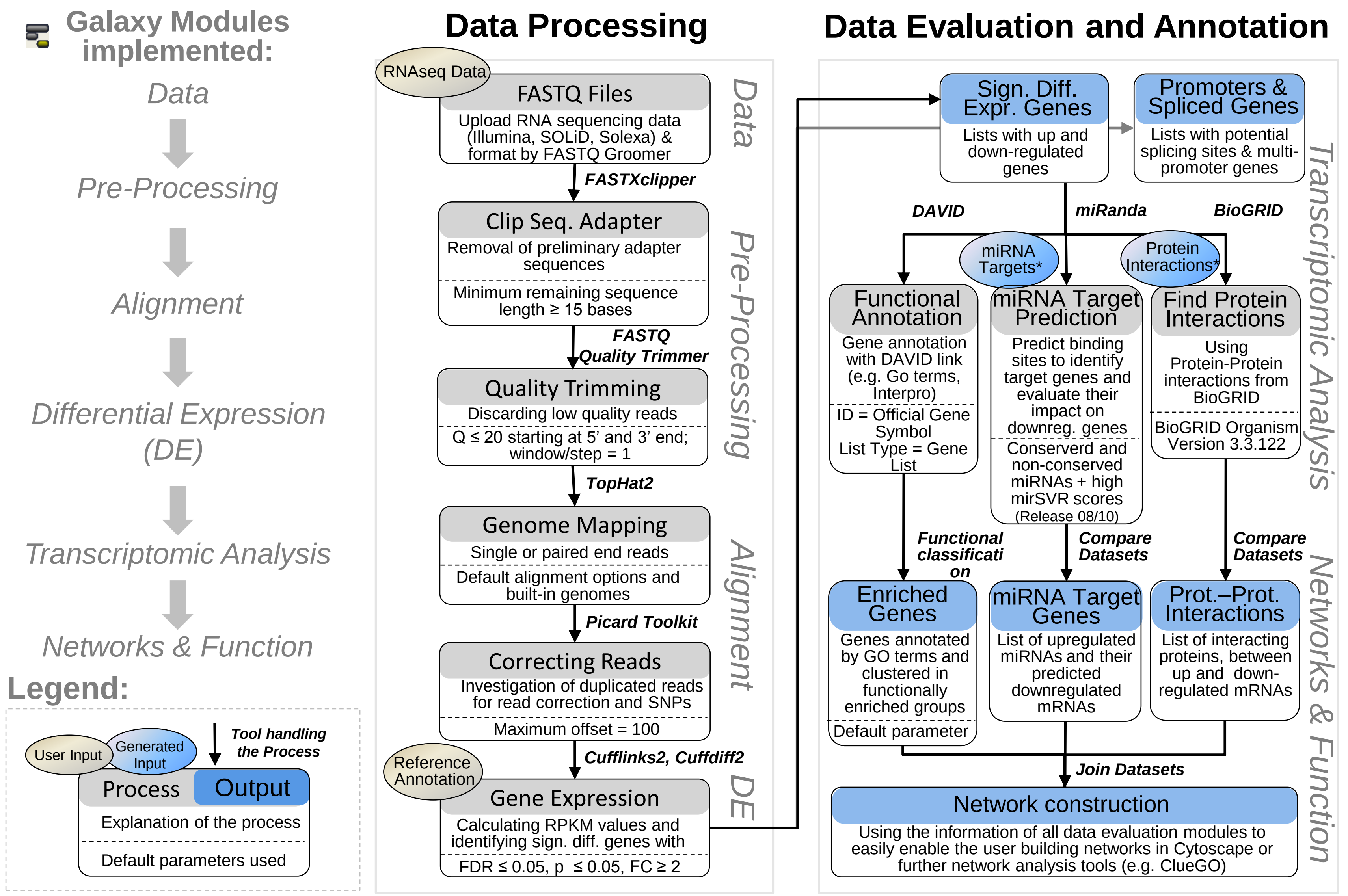
Cufflinks2

DAVID

Differential Promoter and Splicing Detection

Annotation and Functional Clustering

Transparent Reproducible Automated PipeLINE - TRAPLINE



Galaxy Modules implemented:

Data

Pre-Processing

Alignment

Differential Expression (DE)

Transcriptomic Analysis

Networks & Function

Legend:

User Input

Generated Input

Tool handling the Process

Process

Output

Explanation of the process

Default parameters used

Data Processing

Data

RNAseq Data

FASTQ Files

Upload RNA sequencing data (Illumina, SOLiD, Solexa) & format by FASTQ Groomer

FASTXclipper

Pre-Processing

Clip Seq. Adapter

Removal of preliminary adapter sequences

Minimum remaining sequence length ≥ 15 bases

FASTQ Quality Trimmer

Quality Trimming

Discarding low quality reads

Q ≤ 20 starting at 5' and 3' end; window/step = 1

TopHat2

Alignment

Genome Mapping

Single or paired end reads

Default alignment options and built-in genomes

Picard Toolkit

Correcting Reads

Investigation of duplicated reads for read correction and SNPs

Maximum offset = 100

Cufflinks2, Cuffdiff2

DE

Reference Annotation

Gene Expression

Calculating RPKM values and identifying sign. diff. genes with

FDR ≤ 0.05, p ≤ 0.05, FC ≥ 2

Data Evaluation and Annotation

Transcriptomic Analysis

Sign. Diff. Expr. Genes

Lists with up and down-regulated genes

Promoters & Spliced Genes

Lists with potential splicing sites & multi-promoter genes

DAVID

miRanda

BioGRID

miRNA Targets*

Protein Interactions*

Functional Annotation

Gene annotation with DAVID link (e.g. Go terms, Interpro)

ID = Official Gene Symbol
List Type = Gene List

miRNA Target Prediction

Predict binding sites to identify target genes and evaluate their impact on downreg. genes

Conserved and non-conserved miRNAs + high mirSVR scores (Release 08/10)

Find Protein Interactions

Using Protein-Protein interactions from BioGRID

BioGRID Organism Version 3.3.122

Networks & Function

Functional classificati on

Compare Datasets

Compare Datasets

Enriched Genes

Genes annotated by GO terms and clustered in functionally enriched groups

Default parameter

miRNA Target Genes

List of upregulated miRNAs and their predicted downregulated mRNAs

Prot.–Prot. Interactions

List of interacting proteins, between up and down-regulated mRNAs

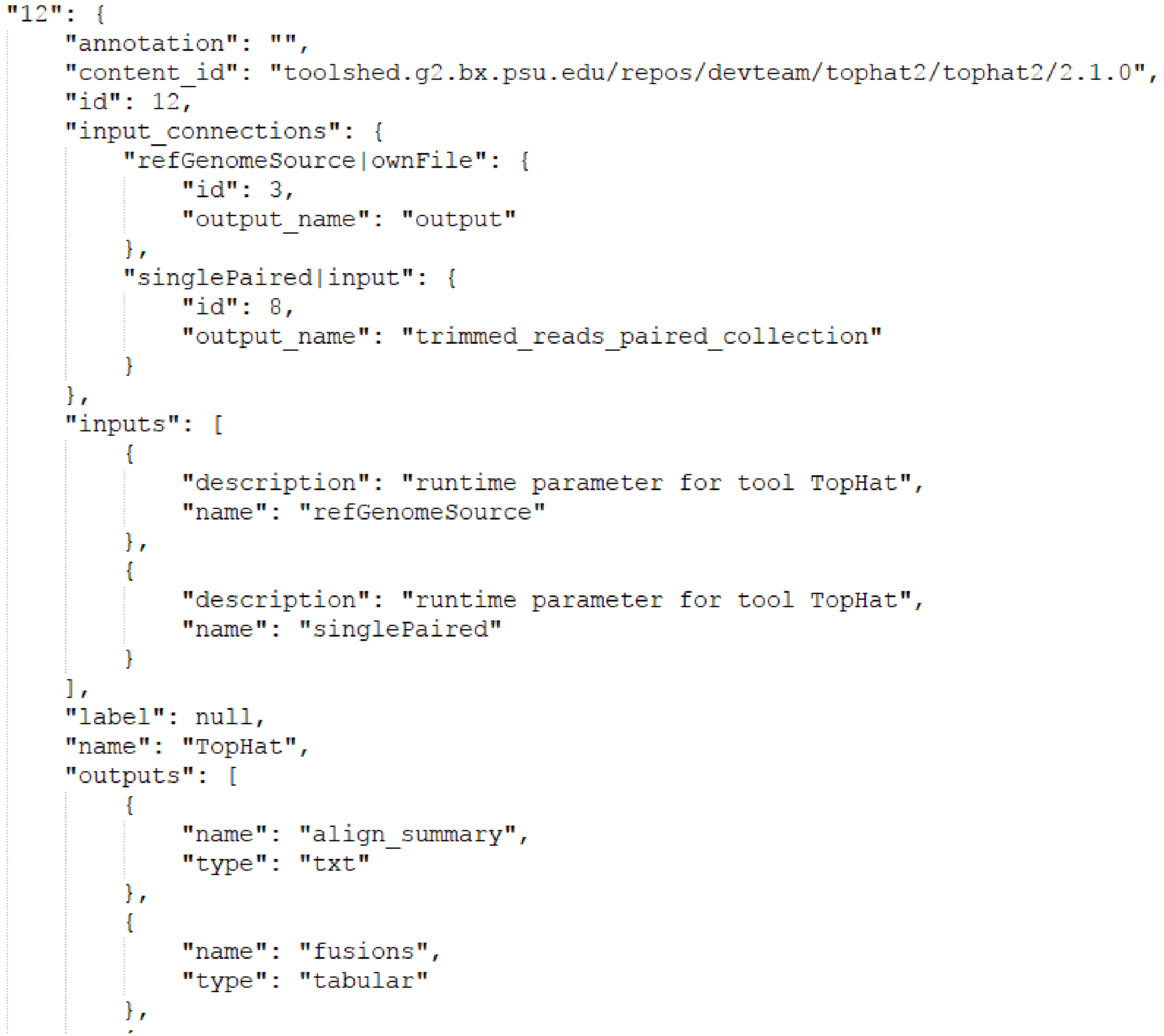
Join Datasets

Network construction

Using the information of all data evaluation modules to easily enable the user building networks in Cytoscape or further network analysis tools (e.g. ClueGO)

Wolfien, *BMC Bioinf.*, 2016

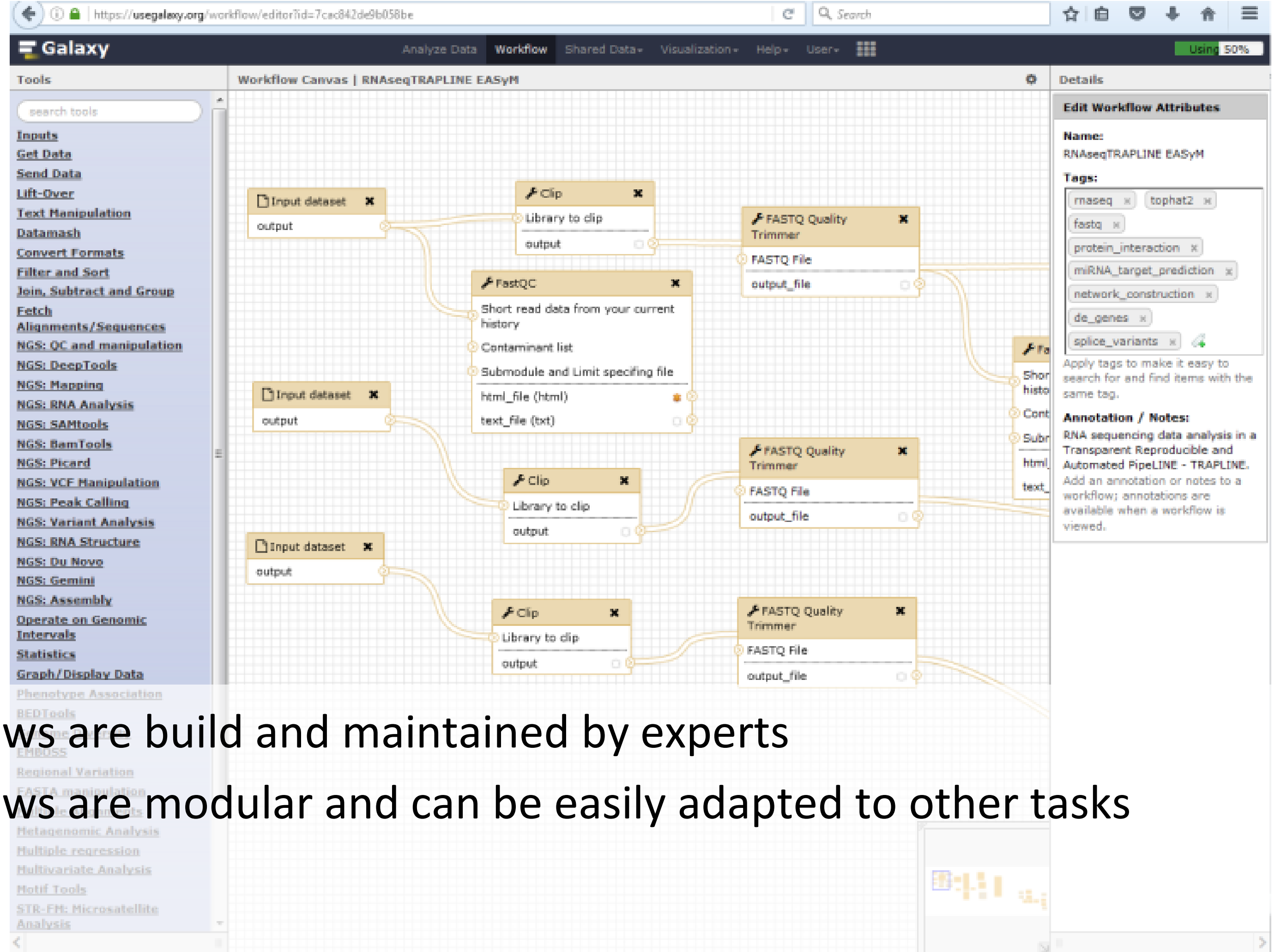
Example Galaxy workflow: TRAPLINE.ga



```
"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



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

Part of German Network for Bioinformatics Infrastructure

deNBI

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

deSTAIR

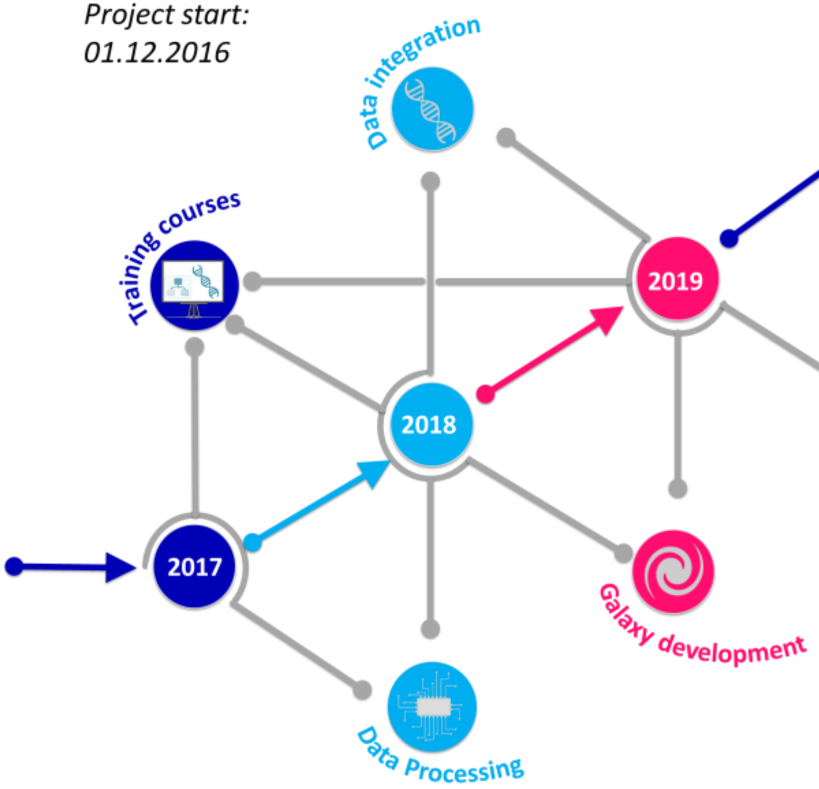
sbi.uni-rostock.de/destair

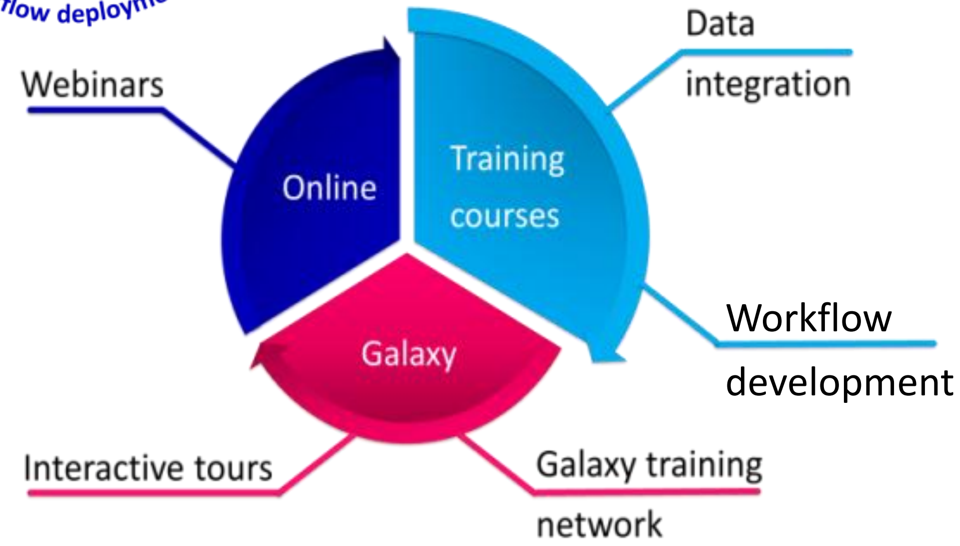
Structured Analysis and Integration of RNA-Seq experiments (de.STAIR)

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





Supporting new data analysis approaches

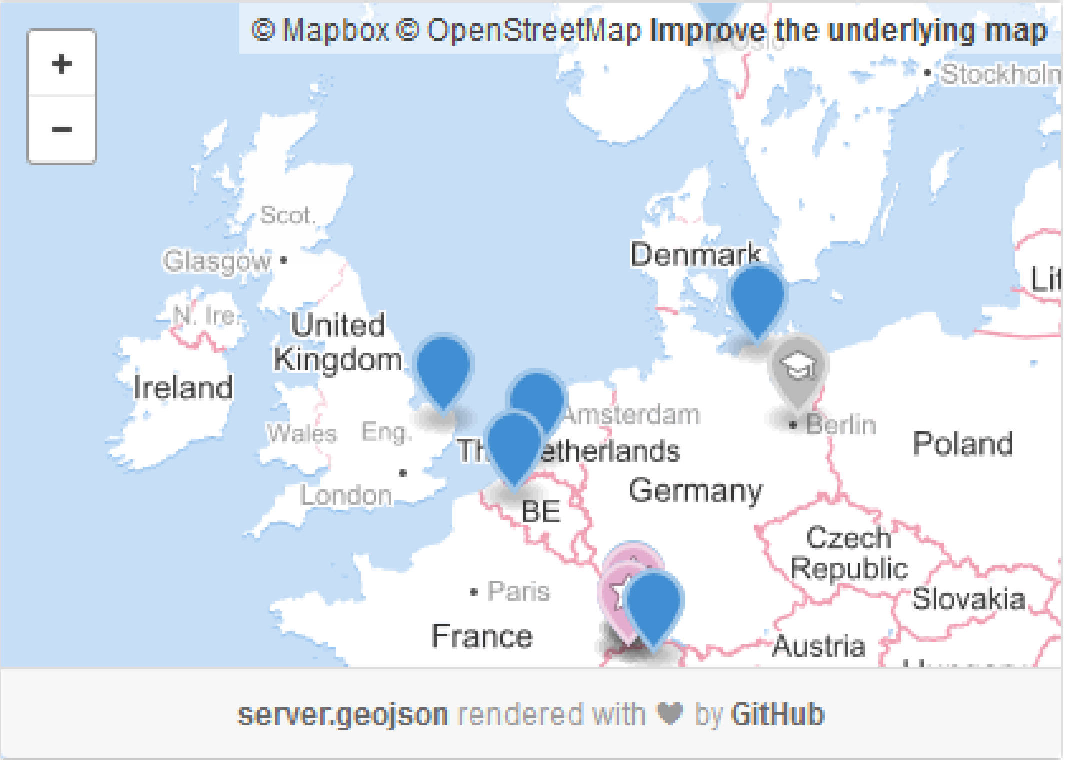
- Key performance of Galaxy
  - Accessibilty
  - Reproducibility
  - Transparency

jupyter

jupyter.org/

Galaxy

usegalaxy.org



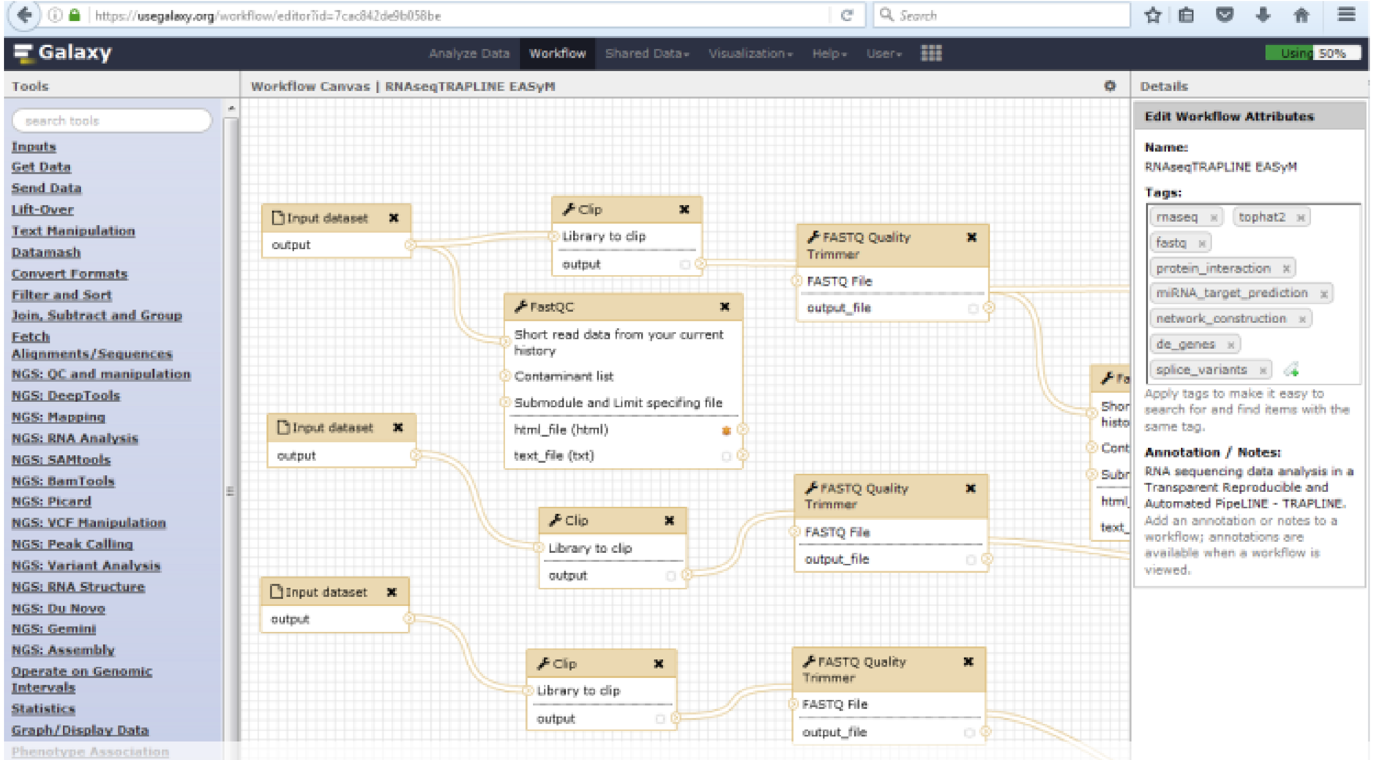
elixir

elixir-europe.org

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

denbi.de

Using workflow development



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

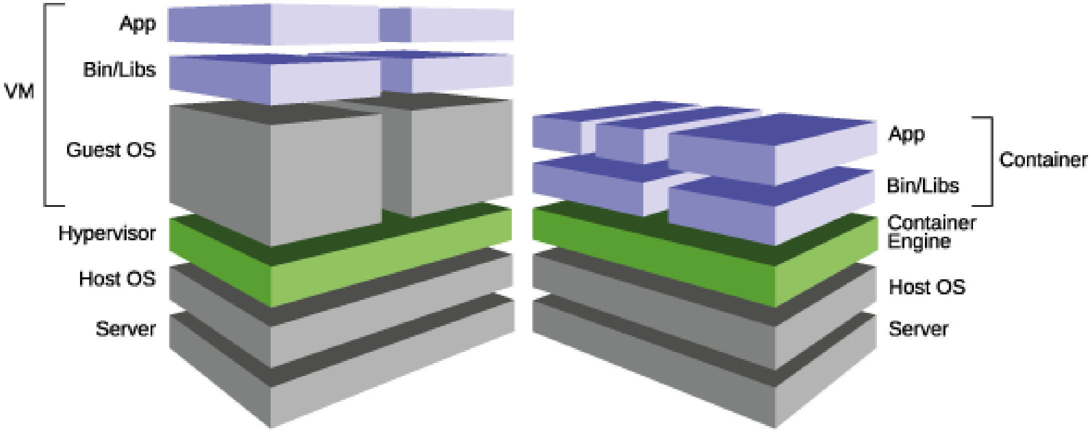
Containerization!
New workflow technologies eliminate
“works on my machine” problems

Containerization!

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

- Build for scale
- Extensible and flexible
- Using optimized system libraries
- E.g. used by ebay, GE, illumina, Spotify

biocontainers.pro



Example Dockerfile: TRAPLINE + Docker

```
FROM bgruening/galaxy-stable
```

Source Container

```
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" \
```

Tools to be added to the new Container

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

EXPOSE :80
EXPOSE :21
EXPOSE :8080

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

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 tool compilation (for an easy use – one command line or single kitematic.com click!) – Slurm cluster

```
docker run -p 8080:80 mwolfien/trapline
```

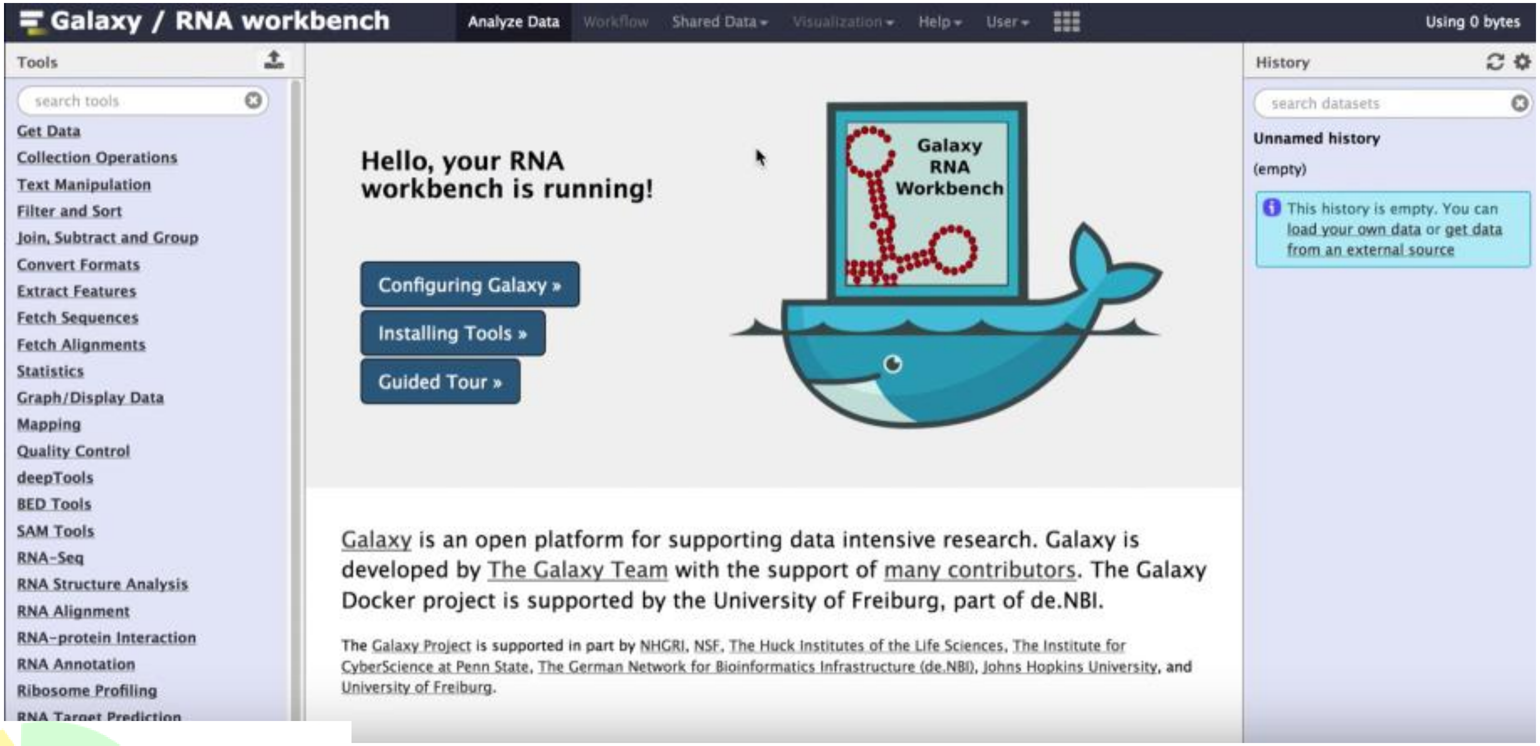
- Docker swarm for single docker container & tools for higher flexibility

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



Hello, your RNA workbench is running!

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors. The Galaxy Docker project is supported by the University of Freiburg, part of de.NBI.

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, The German Network for Bioinformatics Infrastructure (de.NBI), Johns Hopkins University, and University of Freiburg.

Unnamed history

(empty)

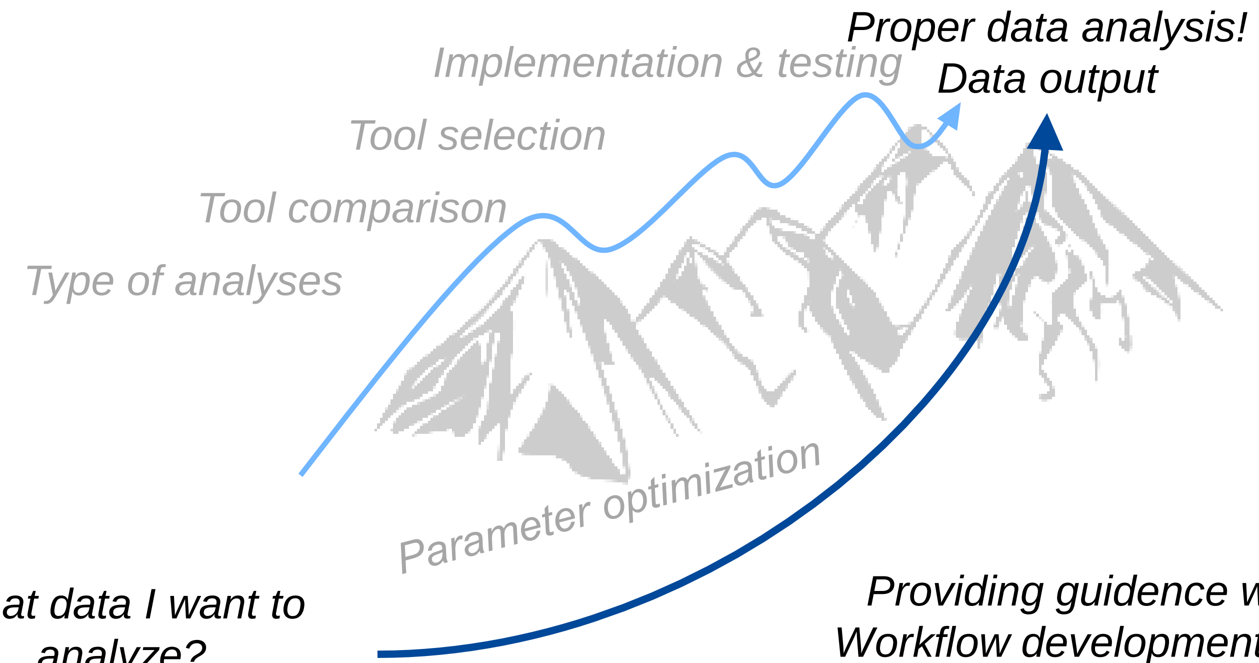
This history is empty. You can load your own data or get data from an external source

Galaxy Training Network

galaxyproject.github.io/training-material

Gruening *et al., NRA, 2017*

Why using workflows?



Proper data analysis!
Data output

Implementation & testing

Tool selection

Tool comparison

Type of analyses

Parameter optimization

What data I want to analyze?
Data input

Providing guidence with Workflow development and tool modularization by e.g.

deSTAIR

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

Medical “Big Data” and the need for new …

Open for Innovation KNIME
knime.org

Galaxy
usegalaxy.org

Chipster
Open source platform for data analysis
chipster.csc.fi

geneXplain
genexplain.com

BaseSpace®
illumina.com

+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

Workflows can be used for complex data integration and interpretation

Our implementation strategy

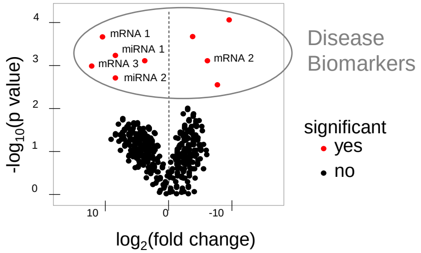
Unexplained symptoms

ATCATG101000111
TAGTG101010101111
TTCATAA00011011
AAGCA1010011

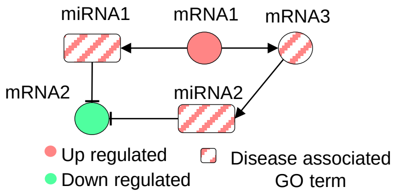
Bioinformatics
Automated and scalable
data processing, e.g. TRAPLINE [2]

docker Galaxy rkt

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



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



New insights

101000111ATCATG
101010101111TTCATAA
00011011AAGCA
1010011TAGTG

Med. Informatics
Data management

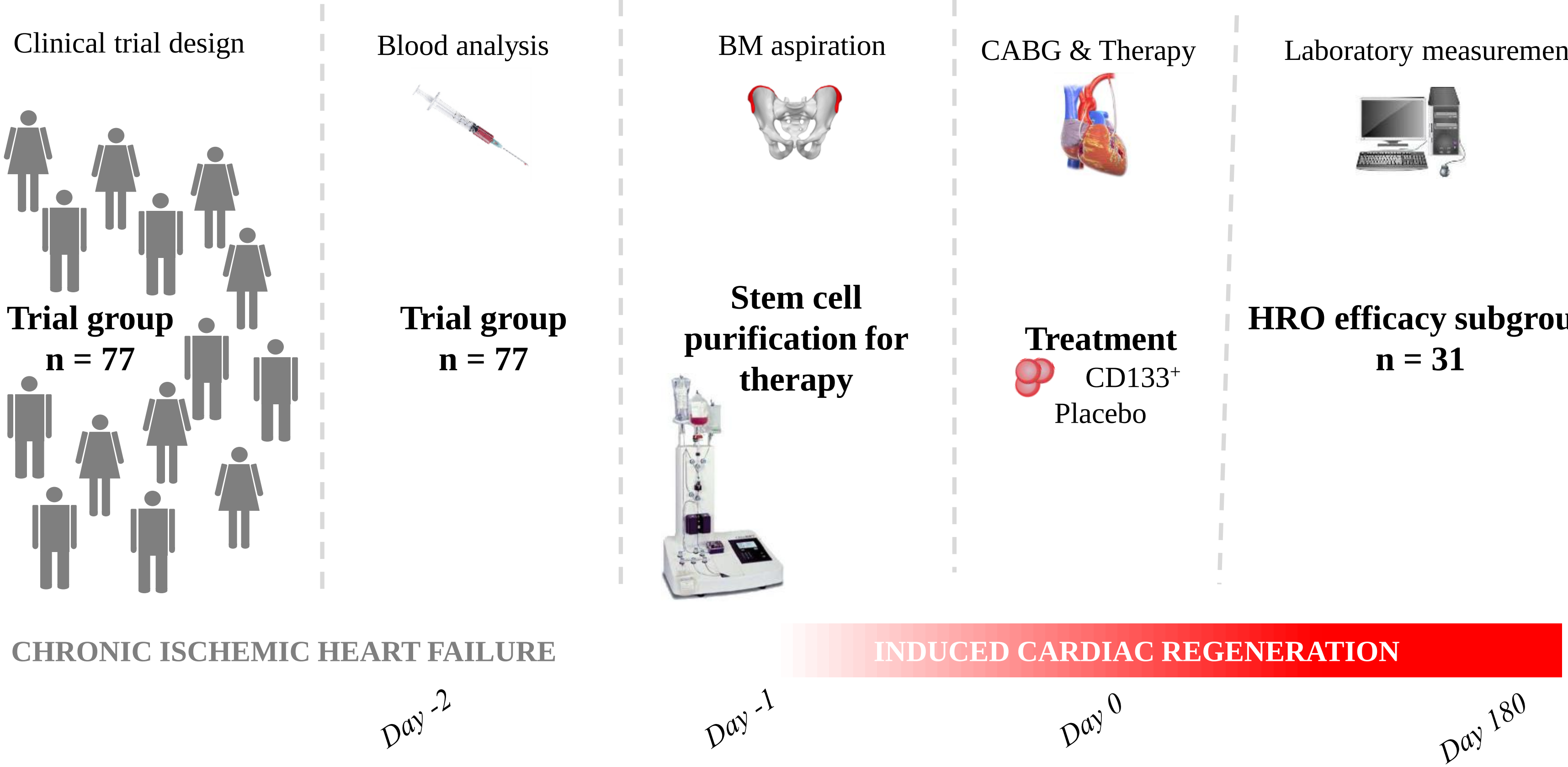
Machine learning

HUB
FAIRDOM

Clinic
Improved diagnosis,
prognosis and therapy

„Medizin 4.0 - Zur Zukunft der Medizin in der digitalisierten Welt“. ISBN: 978-3-9809206-5-0
Poster prize. https://doi.org/10.6084/m9.figshare.4029069.v1

Phase III clinical trial - setting



Clinical trial design

Trial group
n = 77

Blood analysis

Trial group
n = 77

BM aspiration

Stem cell purification for therapy

CABG & Therapy

Treatment
CD133^{+}
Placebo

Laboratory measurements

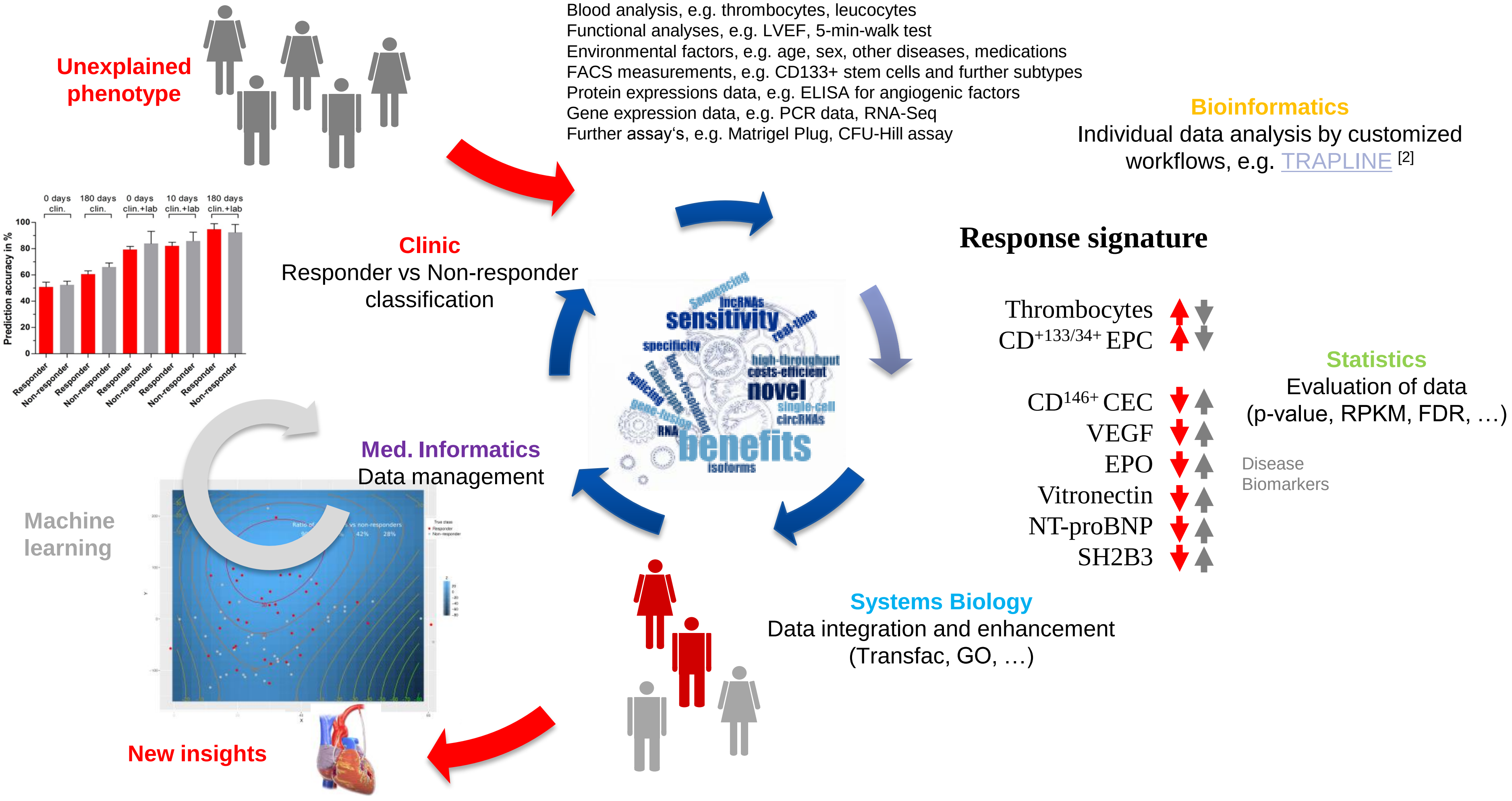
HRO efficacy subgroup
n = 31

CHRONIC ISCHEMIC HEART FAILURE

INDUCED CARDIAC REGENERATION

Steinhoff, Nesteruk, Wolfien, et al., *EBioMedicine*, 2017

Customized workflow strategy for clinical trial



Blood analysis, e.g. thrombocytes, leucocytes
Functional analyses, e.g. LVEF, 5-min-walk test
Environmental factors, e.g. age, sex, other diseases, medications
FACS measurements, e.g. CD133+ stem cells and further subtypes
Protein expressions data, e.g. ELISA for angiogenic factors
Gene expression data, e.g. PCR data, RNA-Seq
Further assay's, e.g. Matrigel Plug, CFU-Hill assay

Bioinformatics
Individual data analysis by customized workflows, e.g. TRAPLINE [2]

Clinic
Responder vs Non-responder classification

Response signature

Thrombocytes ↑ ↓
CD$^{+133/34+}$ EPC ↑ ↓

CD^{146+} CEC ↓ ↑
VEGF ↓ ↑
EPO ↓ ↑
Vitronectin ↓ ↑
NT-proBNP ↓ ↑
SH2B3 ↓ ↑

Statistics
Evaluation of data (p-value, RPKM, FDR, …)

Disease Biomarkers

Med. Informatics
Data management

Machine learning

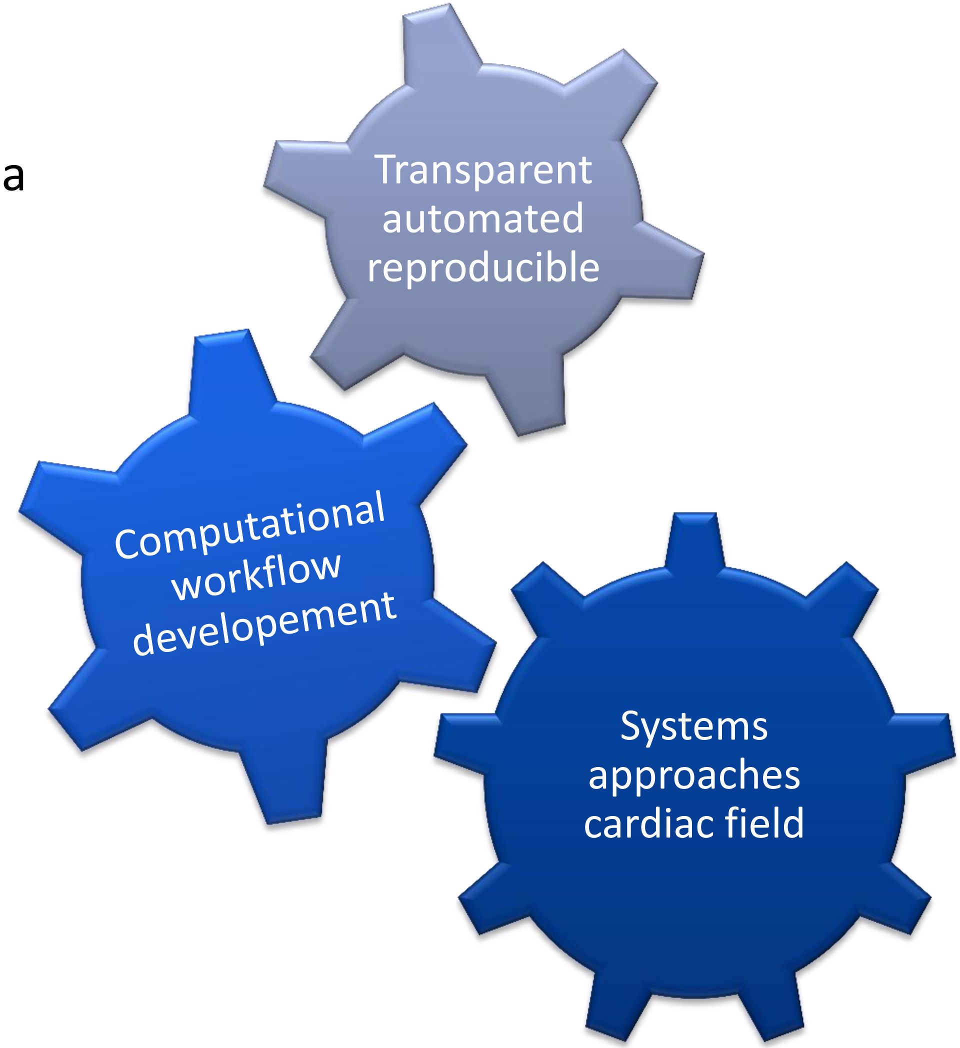
Systems Biology
Data integration and enhancement (Transfac, GO, …)

New insights

Steinhoff, Nesteruk, Wolfien, et al., EBioMedicine, 2017

Summarizing my talk

- Workflows are a great ressource for data analyses
- Containers are used to simplify data analyses
- Integrative workflows support clinical investigations



Acknowledgements

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

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

Gustav Steinhoff (University of Rostock)
Robert David (University of Rostock)

Supported by:

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
denbi.de

elixir
elixir-europe.org

RTC
cardiac-stemcell-therapy.com

Bundesministerium
für Bildung
und Forschung
bmbf.de

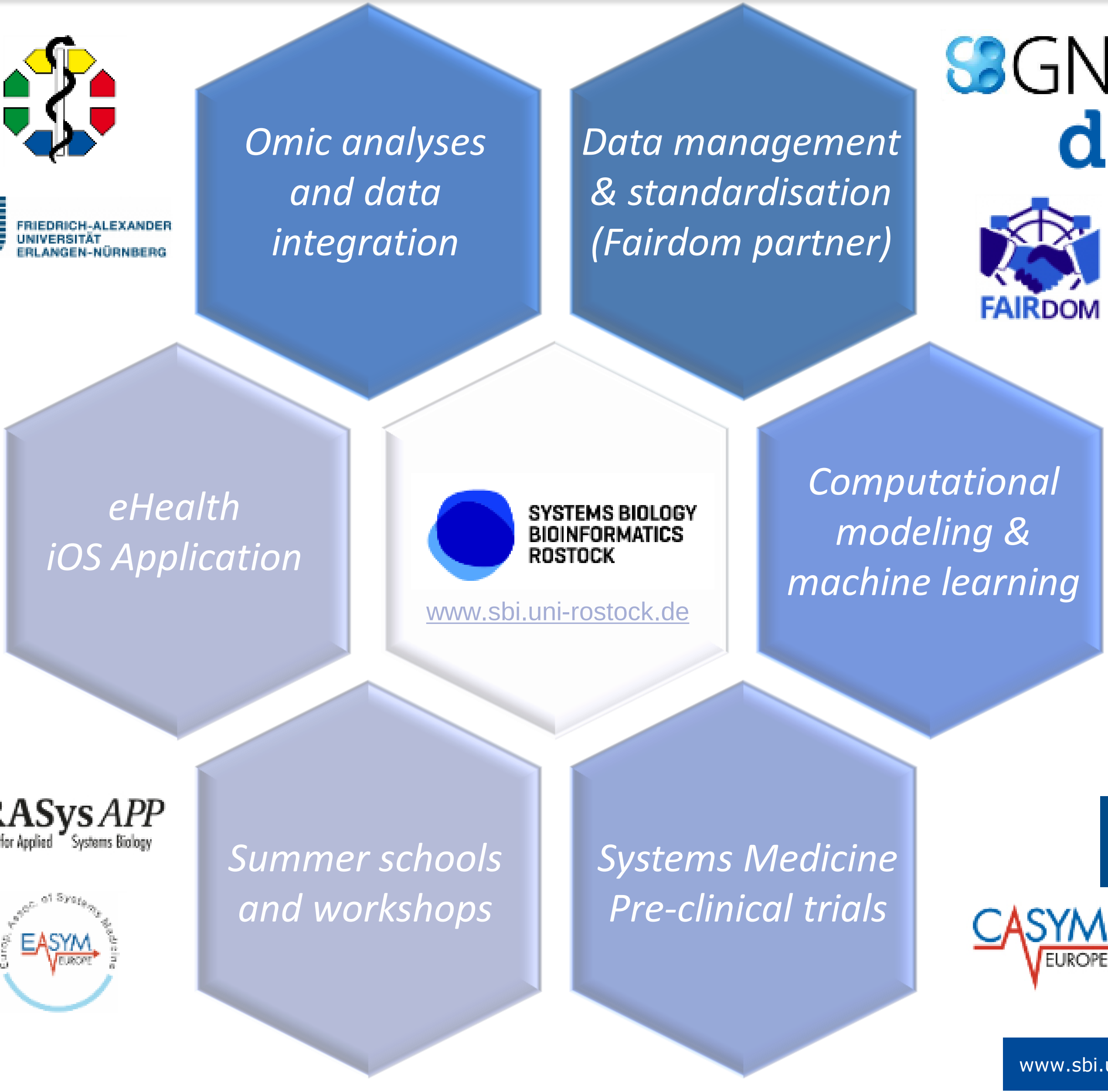
Work where others would like to spend their holidays

Making sense out of data – providing meaning to models

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

PtJ Projektträger Jülich Forschungszentrum Jülich

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG



Omic analyses and data integration
Data management & standardisation (Fairdom partner)
eHealth iOS Application
SYSTEMS BIOLOGY BIOINFORMATICS ROSTOCK
www.sbi.uni-rostock.de
Computational modeling & machine learning
Summer schools and workshops
Systems Medicine Pre-clinical trials

SBGN

de.NBI

FAIRDOM

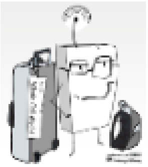
SEMS simulation experiment management system

HICE Aerosols and Health
Helmholtz Virtual Institute of Complex Molecular Systems in Environmental Health

BROAD INSTITUTE

CellProfiler cell image analysis software

ERASysAPP ERA-Net for Applied Systems Biology



EASYM EUROPE Europ. Assoc. of Systems Medicine

Systems MedicineInet

CASYM EUROPE

Universitätsmedizin Rostock

References

- Steinhoff G, Nesteruk J, Wolfien M, Kundt G, The PERFECT Trial Investigators Group. Cardiac Function Improvement and Bone Marrow Response Outcome Analysis of the Randomized Perfect Phase III Clinical Trial of Intramyocardial CD133 + Application After Myocardial Infarction. *EBioMedicine*. 2017. doi.org/10.1016/j.ebiom.2017.07.022
- Lott SC, Wolfien M, Riege K, Bagnacani A, Wolkenhauer O, Hoffmann S, Hess WR. Customized workflow development and data modularization concepts for RNA-Sequencing and metatranscriptome experiments. *Journal of Biotechnology. 2017.* doi.org/10.1016/j.jbiotec.2017.06.1203
- Gruening BA, Fallmann J, Yusuf D, Will S, Erxleben A, Eggenhofer F, Houwaart T, Batut B, Videm P, Bagnacani A, Wolfien M, Lott SC, Hoogstrate Y, Hess WR, Wolkenhauer O, Hoffmann S, Akalin A, Ohler U, Stadler PF, Backofen R. The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. *Nucleic Acids Research*. 2017. doi.org/10.1093/nar/gkx409
- Wolfien M, Rimmbach C, Schmitz U, Jung JJ, Krebs S, Steinhoff G, David R, Wolkenhauer O. TRAPLINE: A standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC Bioinformatics. 2016. doi: 10.1186/s12859-015-0873-9*