# Setting up a galaxy instance - starting problems and experiences

**Jochen Bick, Animal Physiology at ETH Zurich**

Mammals

University of Zurich

Embryo

Rock climbing

Animal Physiology

De novo assembly

Ortholog databases

Susanne Ulbrich

Python

R

Programming

Cow

Perl

PhD Student

RNA-Seq

Pig

Bioinformatics

Stefan Bauersachs

Jochen Bick

Germany

ETH Zurich

MySQL

Deer

Linux

Galaxy project

Functional genomics

Ensembl

NGS Analysis

Hiking

NCBI

Endometrium

Bioconductor

Horse

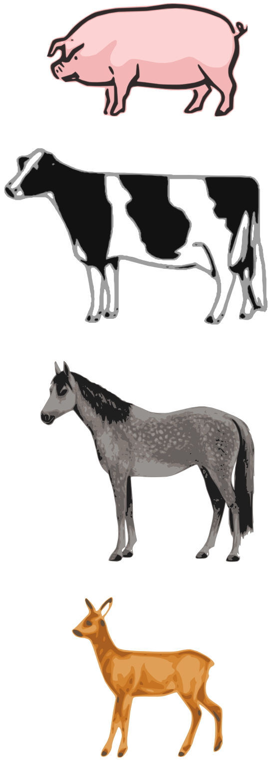
Across-species comparison

Friedrich Schiller University Jena

Animal Physiology

# Background of my research

Functional Genomics

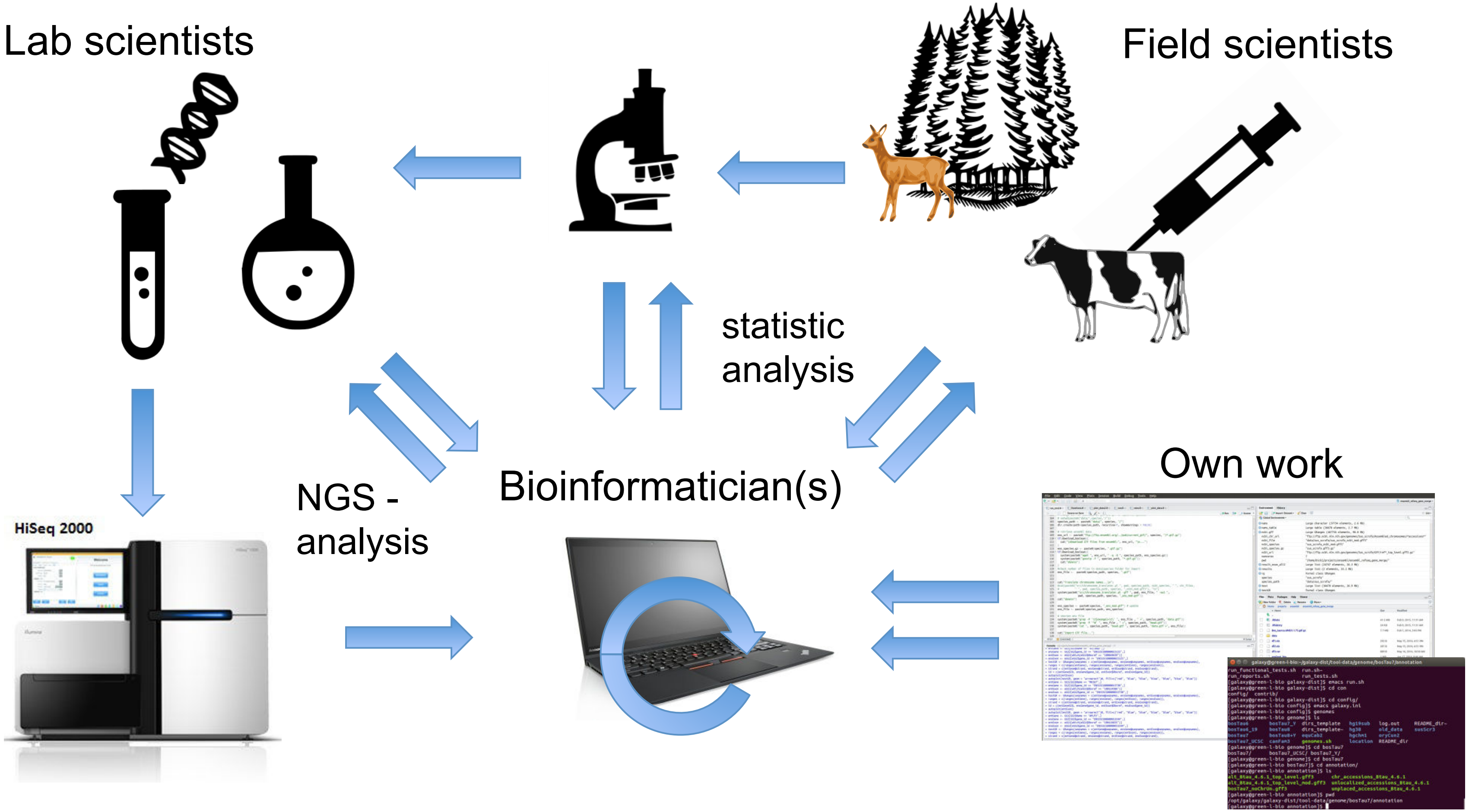


- Research on endometrium and embryo tissues in mammals (e.g. pig, cattle, horse and roe deer)
- Analysis of RNA-Seq data of different stages/ days in cyclic and pregnant mammals
- Comparison of gene expression levels within and across-species
- Development of bioinformatic tools for galaxy server

# Why do we use Galaxy?

- Our group had experiences with an older galaxy version
- We tried CLC Genomics Workbench (ETH has a full license)
  - Less options for modification
  - Black box software and closed source tools (commercial software)
  - Smaller community
  - Less flexible and no easy possibility to edit or add own tools
- Heterogeneous researcher team mostly biologists
- Add own programs/tools and pipelines
- Use own added genomes, files and file formats
- One interface for all different research areas

# Why do we need Galaxy?



Lab scientists

Field scientists

statistic
analysis

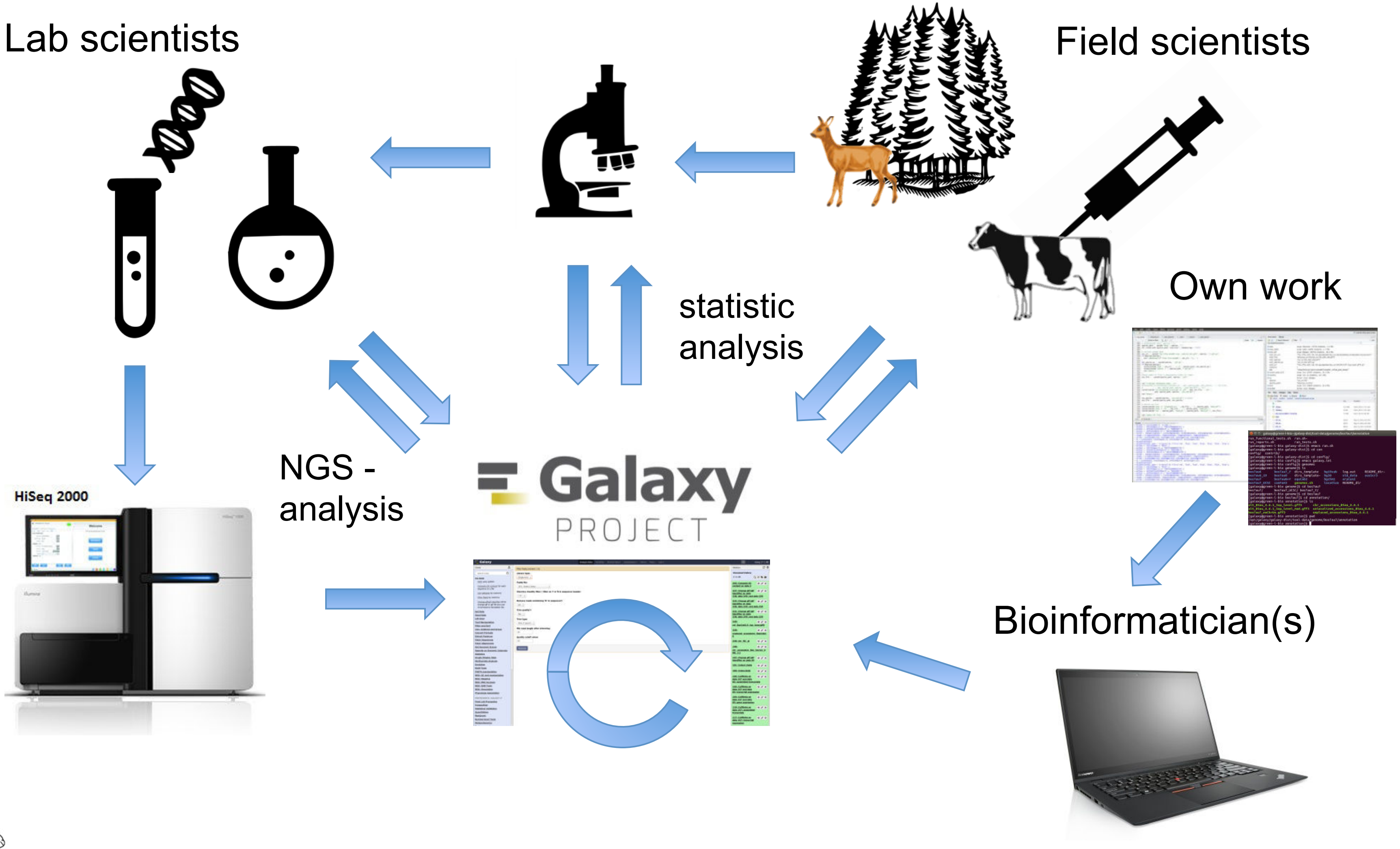
NGS -
analysis

Bioinformatician(s)

Own work

```
galaxy@green-l-bio:~/galaxy-dist/tool-data/genome/bosTau7/annotation
run_functional_tests.sh  run.sh~
run_reports.sh           run_tests.sh
[galaxy@green-l-bio galaxy-dist]$ emacs run.sh
[galaxy@green-l-bio galaxy-dist]$ cd con
config/  contrib/
[galaxy@green-l-bio galaxy-dist]$ cd config/
[galaxy@green-l-bio config]$ emacs galaxy.ini
[galaxy@green-l-bio config]$ genomes
[galaxy@green-l-bio genome]$ ls
bosTau6       bosTau7_Y  dirs_template   hg19sub   log.out    README_dir~
bosTau6_19    bosTau8    dirs_template~  hg38      old_data   susScr3
bosTau7       bosTau8+Y  equCab2         hgchm1    oryCun2
bosTau7_UCSC  canFam3    genomes.sh      location  README_dir
[galaxy@green-l-bio genome]$ cd bosTau7
bosTau7/      bosTau7_UCSC/ bosTau7_Y/
[galaxy@green-l-bio genome]$ cd bosTau7
[galaxy@green-l-bio bosTau7]$ cd annotation/
[galaxy@green-l-bio annotation]$ ls
alt_Btau_4.6.1_top_level.gff3      chr_accessions_Btau_4.6.1
alt_Btau_4.6.1_top_level_mod.gff3  unlocalized_accessions_Btau_4.6.1
bosTau7_noChrUn.gff3               unplaced_accessions_Btau_4.6.1
[galaxy@green-l-bio annotation]$ pwd
/opt/galaxy/galaxy-dist/tool-data/genome/bosTau7/annotation
[galaxy@green-l-bio annotation]$ 
```

# Why do we need Galaxy?



Lab scientists

Field scientists

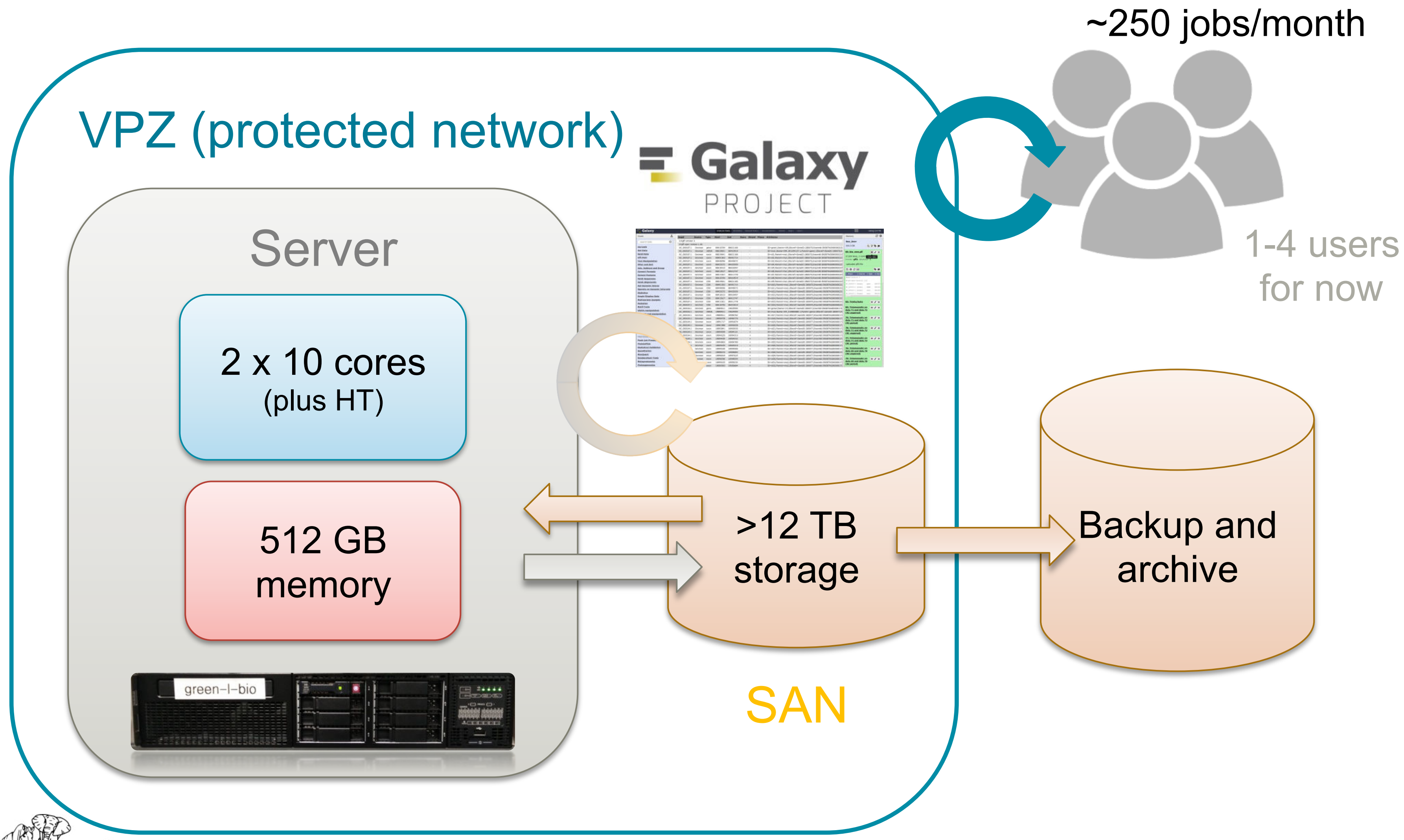
Own work

statistic analysis

NGS - analysis

Bioinformatician(s)

# Specifications of our server



~250 jobs/month

VPZ (protected network)

Galaxy PROJECT

Server

2 x 10 cores
(plus HT)

512 GB
memory

1-4 users
for now

>12 TB
storage

SAN

Backup and
archive

# Quick setup and starting problems

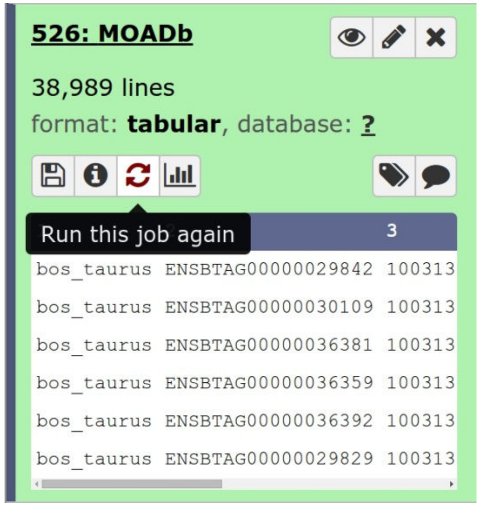
- Cloning, small modification, first start from desktop PC
- Minor obstacles:
  - Learn the basic structure of galaxy
  - Folders/XML/loc-files/config-files
  - Understanding galaxy.ini (universe.ini)
- Bigger problems
  - We first used SQLite (default)
    - Increasing number of users made work impossible (no concurrency handling)
    - Transferring to PostgreSQL is not possible
- How to communicate with users in case of problems?
  - Via email, see errors in Galaxy Report Tool
  - Share history and solve problems based on examples

# Starting problems

- Import histories from different galaxy setups
  - Almost impossible
  - Galaxy setups need to be 100% identical
- Data and tool integration from old galaxy setups is possible
- Download genomes, index files, …
- Generate missing index files
- Synchronize species paths/names/locations in *.loc files
- Which tool needs which file at which location?

# Tools and tool-testing

- Avoid XML errors ➡ tools do not appear in interface
- XML syntax and command syntax can get tricky
- $PATH was not accessible from galaxy instance
  - $PATH set in run.sh
- Tool-testing
  - Re-running can execute old commands
  - Catch tool output instead of STDERR
    - R scripts often have that problem
    - STDERR wrapper by Assaf Gordon helps



**526: MOADb**

38,989 lines

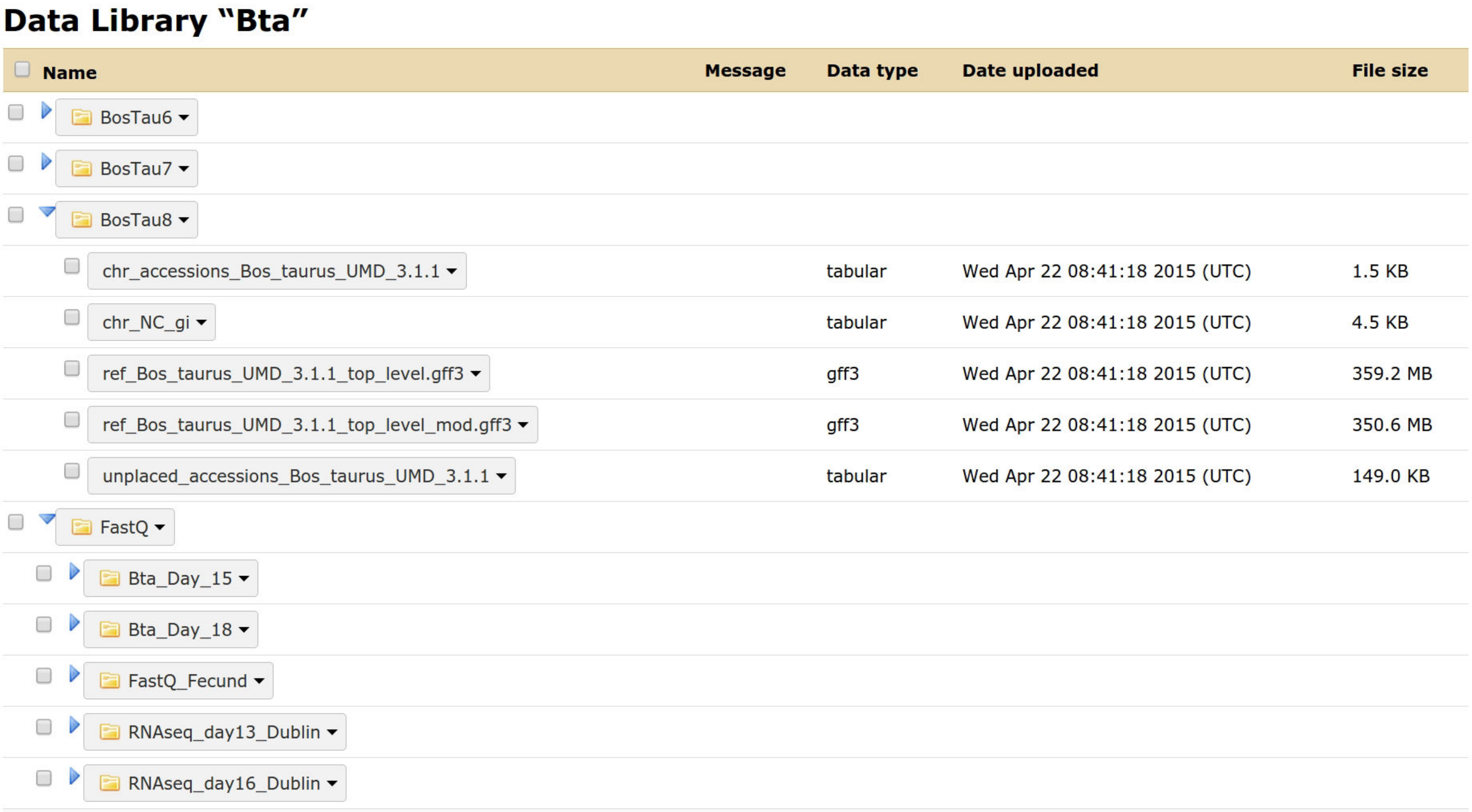
format: **tabular**, database: ?

```
bos_taurus ENSBTAG00000029842 100313
bos_taurus ENSBTAG00000030109 100313
bos_taurus ENSBTAG00000036381 100313
bos_taurus ENSBTAG00000036359 100313
bos_taurus ENSBTAG00000036392 100313
bos_taurus ENSBTAG00000029829 100313
```

# Tools – Visualization of NGS data

- Very complicated and time consuming
  - Trackster is easy to implement
  - Generate necessary files for each genome
  - ➢ Trackstar handling works intuitive
- ➢ Stay inside Galaxy
  - It is not comfortable to download big files (FastQ, .bam, .sam)
  - Quick access to rethink/modify/fine-tune your analysis
  - It is nice to still have a workflow

# Data Libraries



## Data Library “Bta”

| ☐ Name | Message | Data type | Date uploaded | File size |
|---|---|---|---|---|
| ☐ ▸ BosTau6 ▾ | | | | |
| ☐ ▸ BosTau7 ▾ | | | | |
| ☐ ▾ BosTau8 ▾ | | | | |
| ☐ chr_accessions_Bos_taurus_UMD_3.1.1 ▾ | | tabular | Wed Apr 22 08:41:18 2015 (UTC) | 1.5 KB |
| ☐ chr_NC_gi ▾ | | tabular | Wed Apr 22 08:41:18 2015 (UTC) | 4.5 KB |
| ☐ ref_Bos_taurus_UMD_3.1.1_top_level.gff3 ▾ | | gff3 | Wed Apr 22 08:41:18 2015 (UTC) | 359.2 MB |
| ☐ ref_Bos_taurus_UMD_3.1.1_top_level_mod.gff3 ▾ | | gff3 | Wed Apr 22 08:41:18 2015 (UTC) | 350.6 MB |
| ☐ unplaced_accessions_Bos_taurus_UMD_3.1.1 ▾ | | tabular | Wed Apr 22 08:41:18 2015 (UTC) | 149.0 KB |
| ☐ ▾ FastQ ▾ | | | | |
| ☐ ▸ Bta_Day_15 ▾ | | | | |
| ☐ ▸ Bta_Day_18 ▾ | | | | |
| ☐ ▸ FastQ_Fecund ▾ | | | | |
| ☐ ▸ RNAseq_day13_Dublin ▾ | | | | |
| ☐ ▸ RNAseq_day16_Dublin ▾ | | | | |

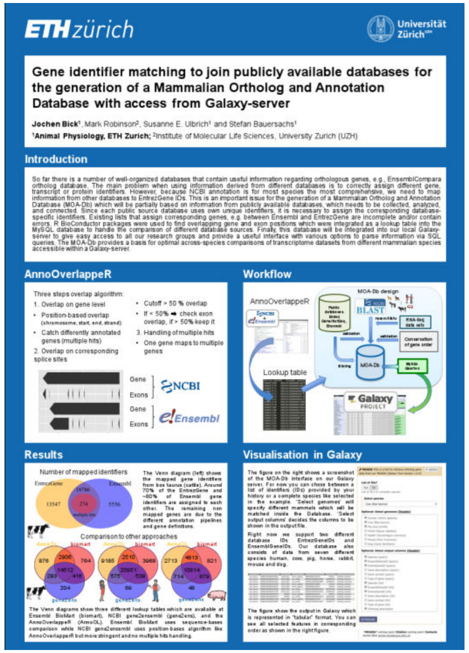
# Acknowledgements

Hans-Rudolf Hotz

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Supervisors:
Stefan Bauersachs
Susanne Ulbrich
Mark Robinson

more interesting research at my poster P03



ETH zürich | Universität Zürich

**Gene identifier matching to join publicly available databases for the generation of a Mammalian Ortholog and Annotation Database with access from Galaxy-server**

**Jochen Bick**[1], Mark Robinson[2], Susanne E. Ulbrich[1] and Stefan Bauersachs[1]

[1]**Animal Physiology, ETH Zurich;** [2]Institute of Molecular Life Sciences, University Zurich (UZH)

**Introduction**

So far there is a number of well-organized databases that contain useful information regarding orthologous genes, e.g., EnsemblCompara ortholog database. The main problem when using information derived from different databases is to correctly assign different gene, transcript or protein identifiers. However, because NCBI annotation is for most species the most comprehensive, we need to map information from other databases to EntrezGene IDs. This is an important issue for the generation of a Mammalian Ortholog and Annotation Database (MOA-Db) which will be partially based on information from publicly available databases, which needs to be collected, analyzed, and connected. Since each public source database uses own unique identifiers, it is necessary to assign the corresponding database-specific identifiers. Existing lists that assign corresponding genes, e.g. between Ensembl and EntrezGene are incomplete and/or contain errors. R BioConductor packages were used to find overlapping gene and exon positions which were integrated as a lookup table into the MySQL database to handle the comparison of different database sources. Finally, this database will be integrated into our local Galaxy-server to give easy access to all our research groups and provide a useful interface with various options to parse information via SQL queries. The MOA-Db provides a basis for optimal across-species comparisons of transcriptome datasets from different mammalian species accessible within a Galaxy-server.

**AnnoOverlappeR**

Three steps overlap algorithm:
1. Overlap on gene level
- Position-based overlap (chromosome, start, end, strand)
- Catch differently annotated genes (multiple hits)
2. Overlap on corresponding splice sites
- Cutoff > 50 % overlap
- If < 50% ➡ check exon overlap, if > 50% keep it
3. Handling of multiple hits
- One gene maps to multiple genes

**Workflow**

**Results**

The Venn diagram (left) shows the mapped gene identifiers from bos taurus (cattle). Around 70% of the EntrezGene and ~80% of Ensembl gene identifiers are assigned to each other. The remaining non mapped genes are due to the different annotation pipelines and gene definitions.

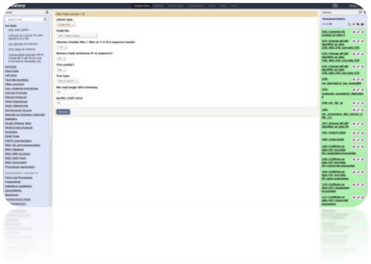
The Venn diagrams show three different lookup tables which are available at Ensembl BioMart (biomart), NCBI gene2ensembl (gene2ens), and the AnnoOverlappeR (AnnoOL). Ensembl BioMart uses sequence-bases comparison while NCBI gene2ensembl uses position-bases algorithm like AnnoOverlappeR but more stringent and no multiple hits handling.

**Visualisation in Galaxy**

The figure on the right shows a screenshot of the MOA-Db interface on our Galaxy server. For now you can chose between a list of identifiers (IDs) provided by your history or a complete species like selected in the example. 'Select genomes' will specify different mammals which will be matched inside the Database. 'Select output columns' decides the columns to be shown in the output file.

Right now we support two different database IDs EntrezGeneIDs and EnsemblGeneIDs. Our database also consists of data from seven different species human, cow, pig, horse, rabbit, mouse and dog.

The figure show the output in Galaxy which is represented in "tabular" format. You can see all selected features in corresponding order as shown in the right figure.



Server admin:
Andreas Burger

Thank you for your attention