

# Inter-EURLs Working Group on NGS (NEXT GENERATION SEQUENCING)



*EURL Lm*  
European Union Reference Laboratory for  
*Listeria monocytogenes*  
<http://eur-listeria.anses.fr>



Bioinformatics analysis of NGS data: approaches and opportunities (command-line tools, commercial software, web servers)

**Joakim Skarin**

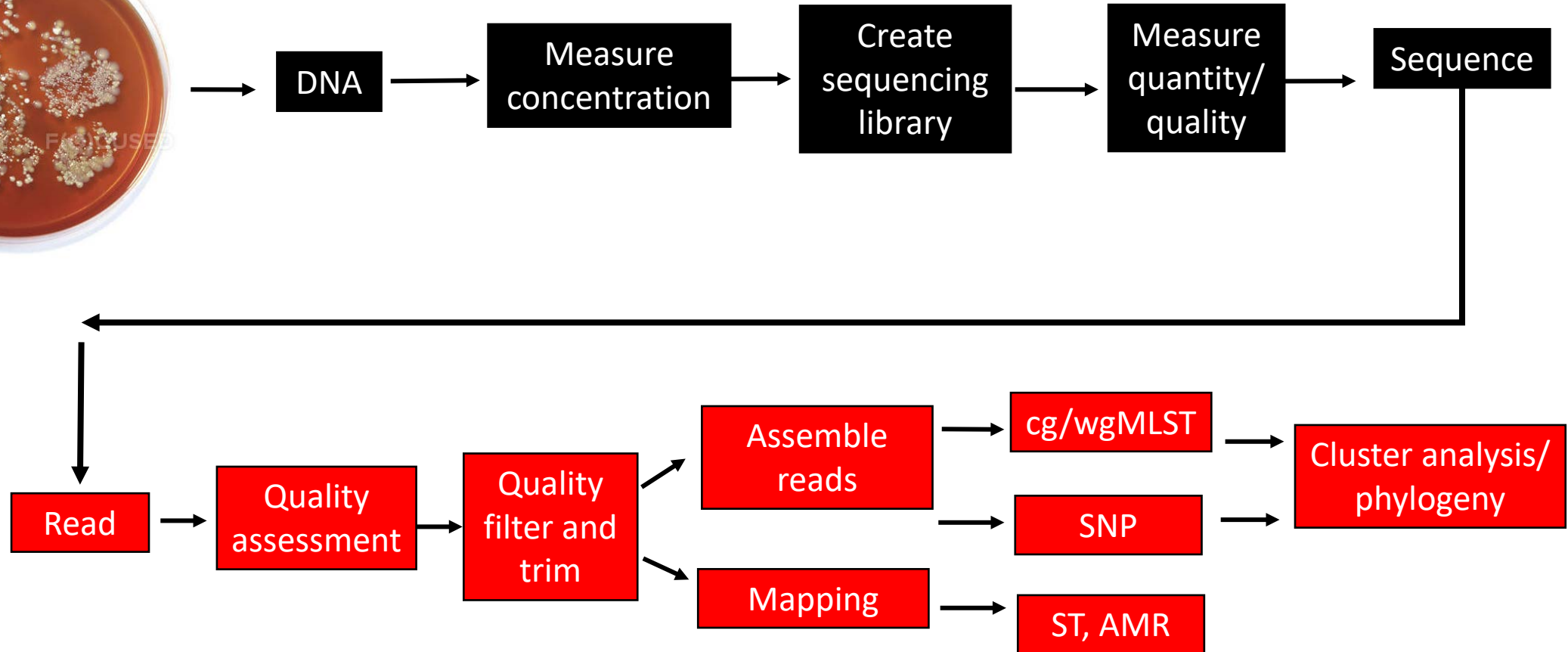
EURL Foodborne Viruses

Swedish Food Agency

- Analyzing millions of sequence reads usually requires several powerful software that performs different tasks (quality trimming, assembly, SNP-calling etc)
- When setting up NGS-analyzing capabilities in your lab, there are different approaches you can choose from
- Hire a bioinformatician and buy a Linux computer?
- Use free webservers?
- Buy expensive commercial software?
- Or do you use all of the above in combination?

...AGCCTAGGGATGCGCGACACGT  
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC  
CAACCTCGGACGGACCTCAGCGAA...

# A general whole genome sequencing workflow



# GUI vs CLI

## **GUI = graphical user interface**

GUI is a form of user interface that allows users to interact with electronic devices through graphical icons

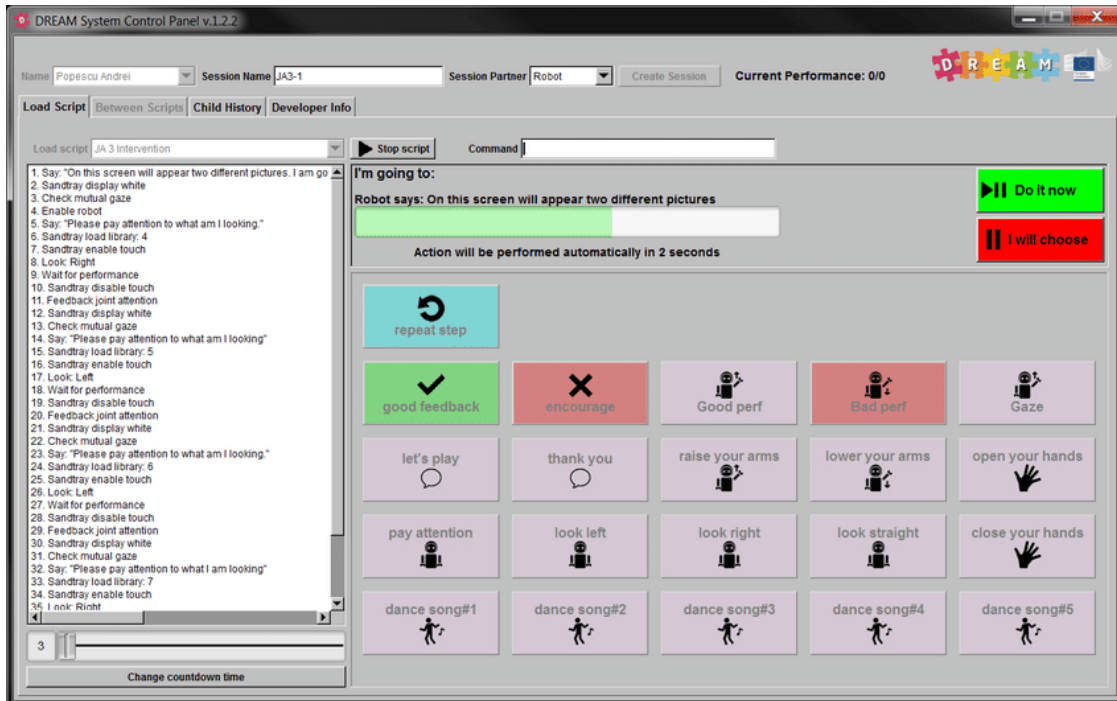
## **CLI = command-line interface**

CLI permits users to put in writing commands in terminal or console window to interact with an operating system

Command-line  
interface ->>>>

```
John@ubuntu: ~  
John@ubuntu:~$ ls  
John_directory John_file  
John@ubuntu:~$ ls -l  
total 8  
drwxrwxr-x 2 John John 40 Oct 1 11:10 John_directory  
-rw-rw-r-- 1 John John 5120 Oct 1 11:17 John_file  
John@ubuntu:~$ ls -l -h  
total 8.0K  
drwxrwxr-x 2 John John 40 Oct 1 11:10 John_directory  
-rw-rw-r-- 1 John John 5.0K Oct 1 11:17 John_file  
John@ubuntu:~$ ls -lh John_file  
-rw-rw-r-- 1 John John 5.0K Oct 1 11:17 John_file  
John@ubuntu:~$ ls -l --human-readable John_file  
-rw-rw-r-- 1 John John 5.0K Oct 1 11:17 John_file  
John@ubuntu:~$
```

[https://www.linuxjournal.com/sites/default/files/styles/max\\_1300x1300/public/u%5Buid%5D/command-line-syntax-example.png](https://www.linuxjournal.com/sites/default/files/styles/max_1300x1300/public/u%5Buid%5D/command-line-syntax-example.png)



Graphical user  
interface

<<<<<-

<https://www.researchgate.net/profile/Pablo-Gomez-Esteban/publication/316926825/figure/fig4/AS:494261989842947@1494852649800/Graphical-User-Interface-GUI-component-used-by-the-therapist-to-control-the-robot-On.png>

## GUI

## CLI

### Public server

### Commercial software

### Local Linux or Mac-computer

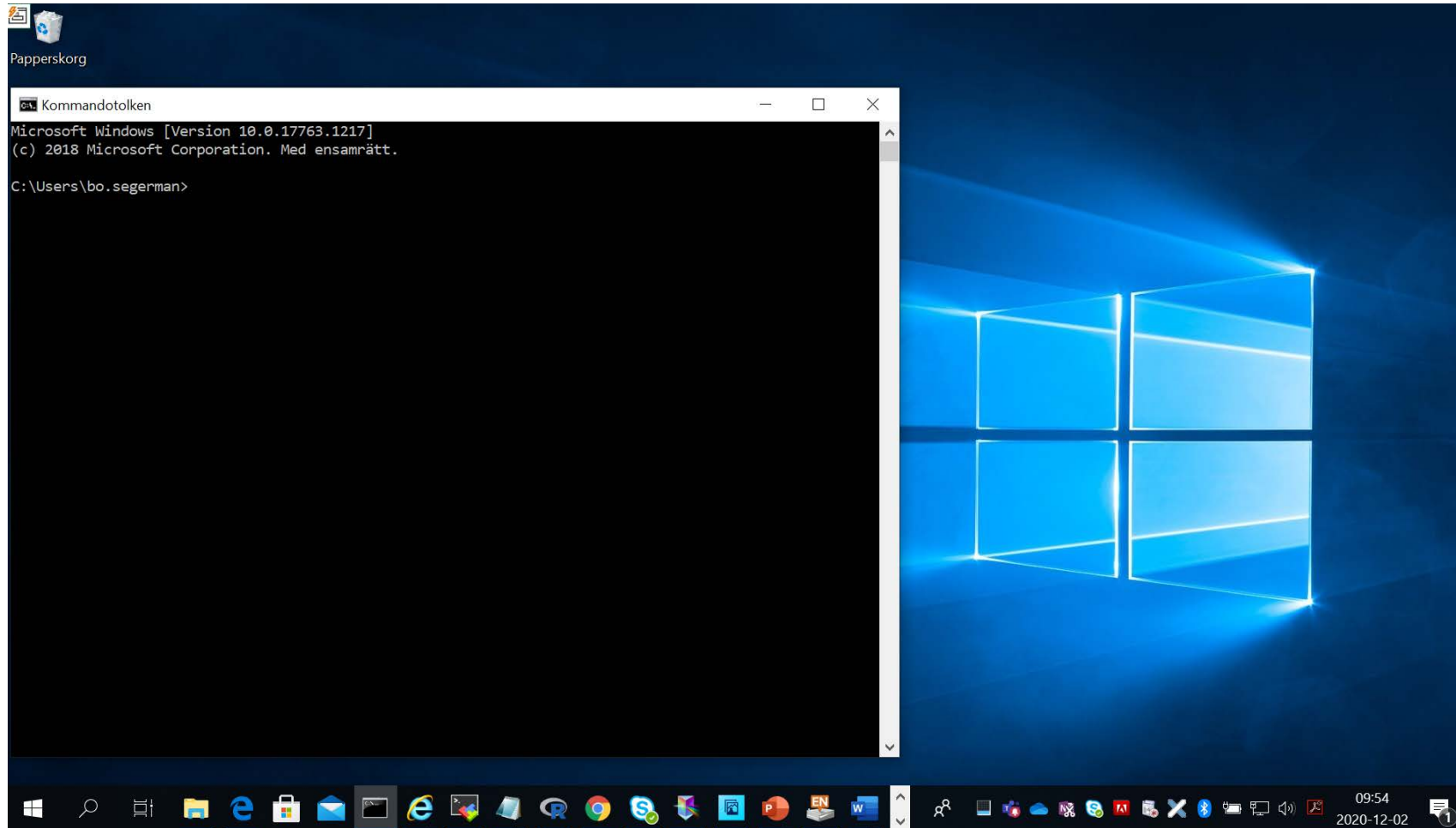
- + No installation
- + Requires little bioinformatics skills
- + Up-to-date nomenclature/DB
- Dependent on service provider
- Downtime of server
- Internet connection
- Can take a long time to get results
- Difficult to analyse large datasets

- + Easy installation
- + Requires little bioinformatics skills
- + Database, easy to use, backup
- Expensive
- Dependent on a company
- Limited to functions in software
- Requires powerful computer

- + No cost for software
- + Free to adjust analysis methods/settings
- + Easy to run once setup
- + Often faster
- + Automation
- Requires bioinformatics skills
- Requires computer knowledge to install and create pipelines
- Requires powerful computer

There are actually some free GUI software as well. Fastqc is one example.

# Windows



GUI operating system

Command Line interface  
available  
(CMD or powershell)

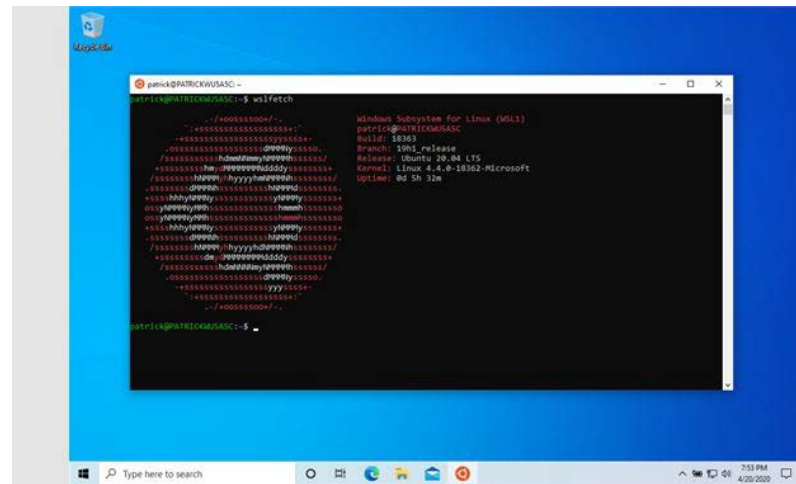
However, few bioinformatic  
programs support these

# If you use Windows - use a GUI solution (typically commercial) OR Use a method to access a UNIX compatible system in Windows

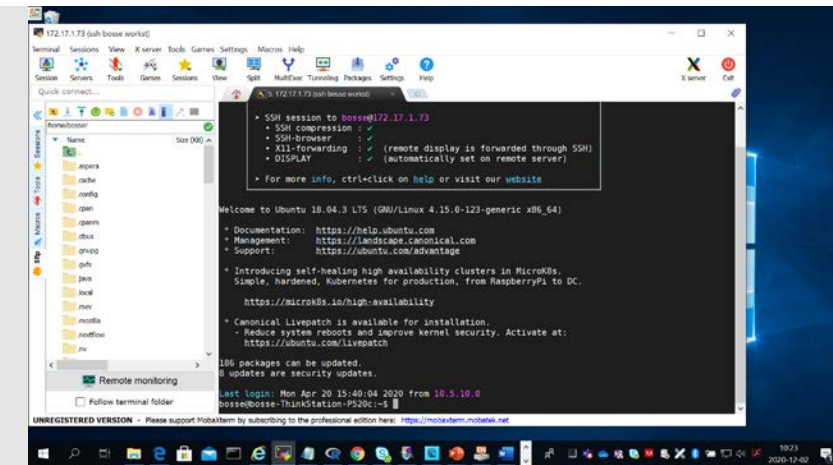
**Dual boot (choose Windows or Linux when starting up)**



**Run a "Virtual" Linux as an app in Windows**

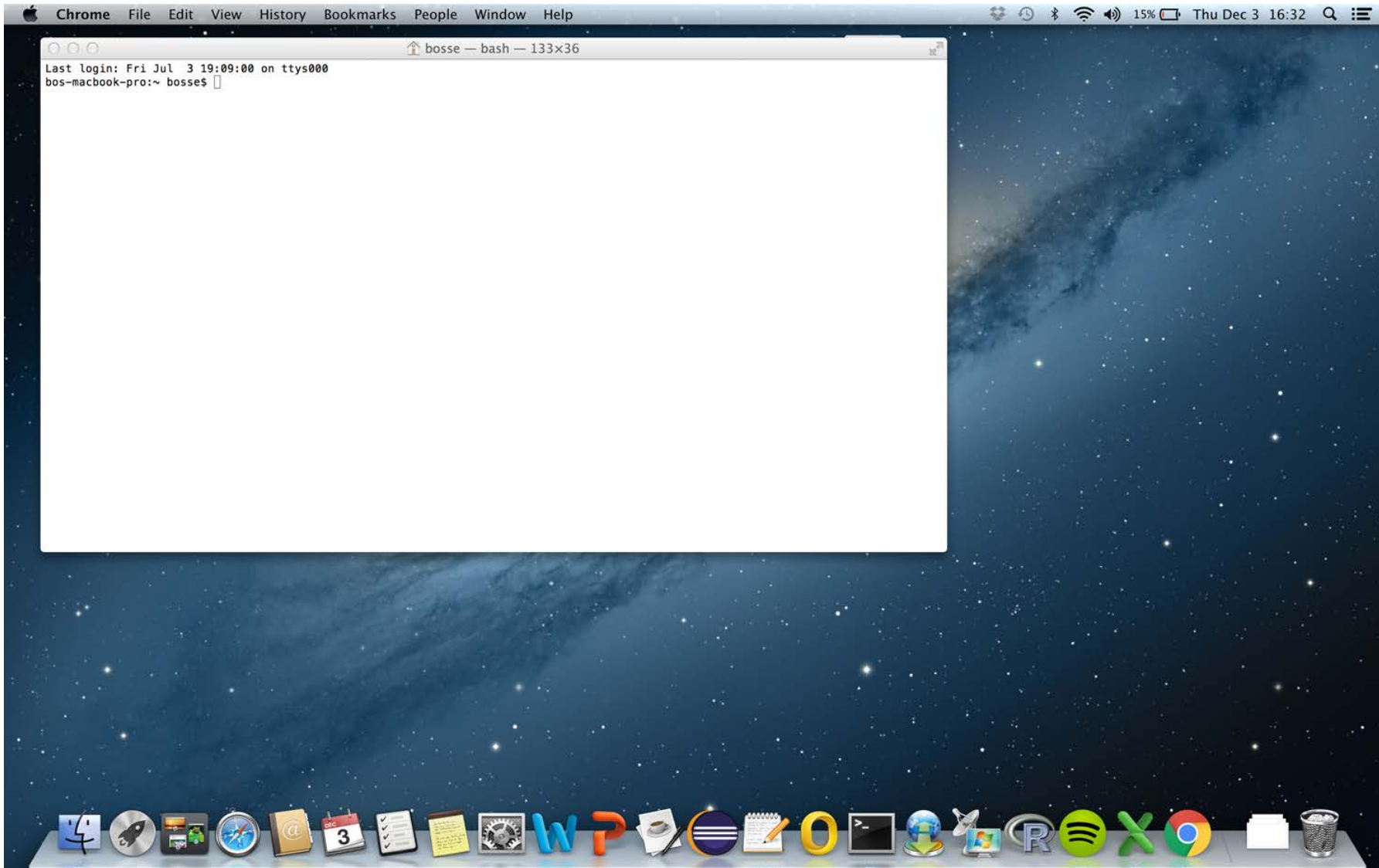


**Connect to a Linux computer (server) and run remotely**





# Mac



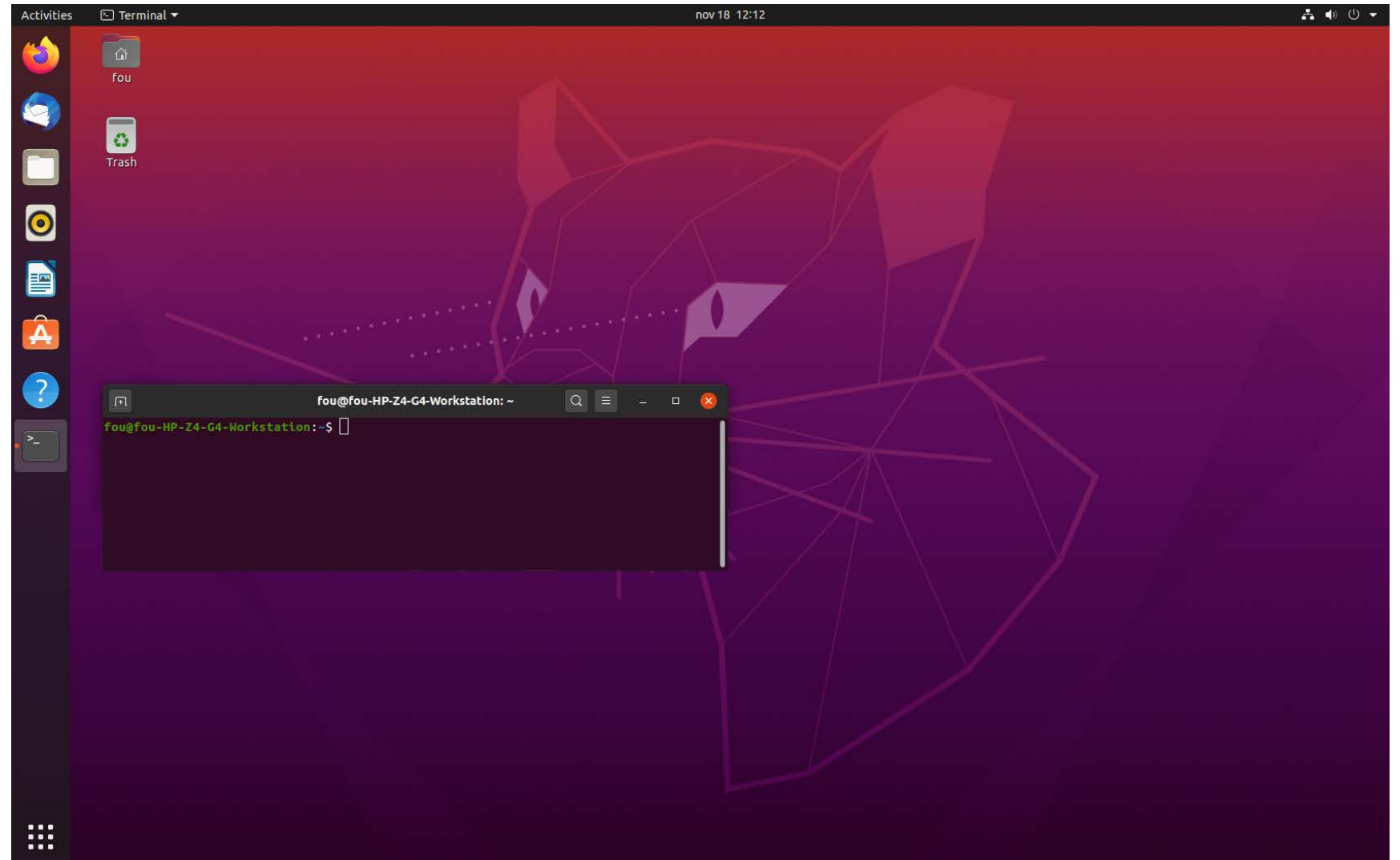
GUI operating system

Modern Mac OS contains a true UNIX compatible terminal

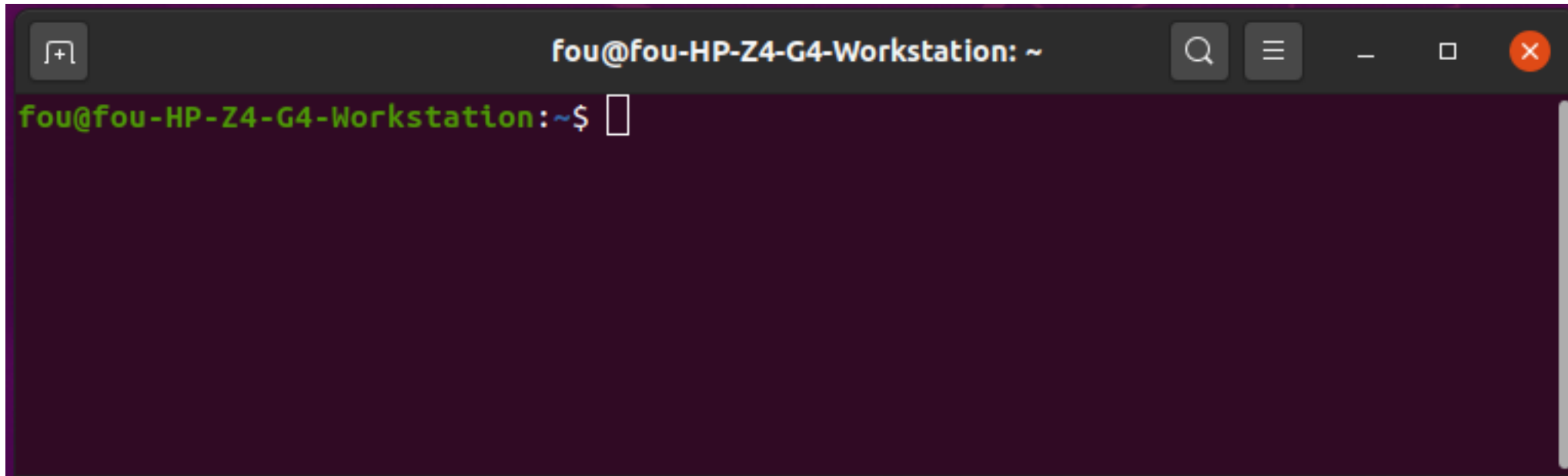
# Linux comes in different distributions

- RedHat
- CentOS
- Fedora
- openSUSE
- Debian
- **Ubuntu**

.....



The terminal (where you put in commands)



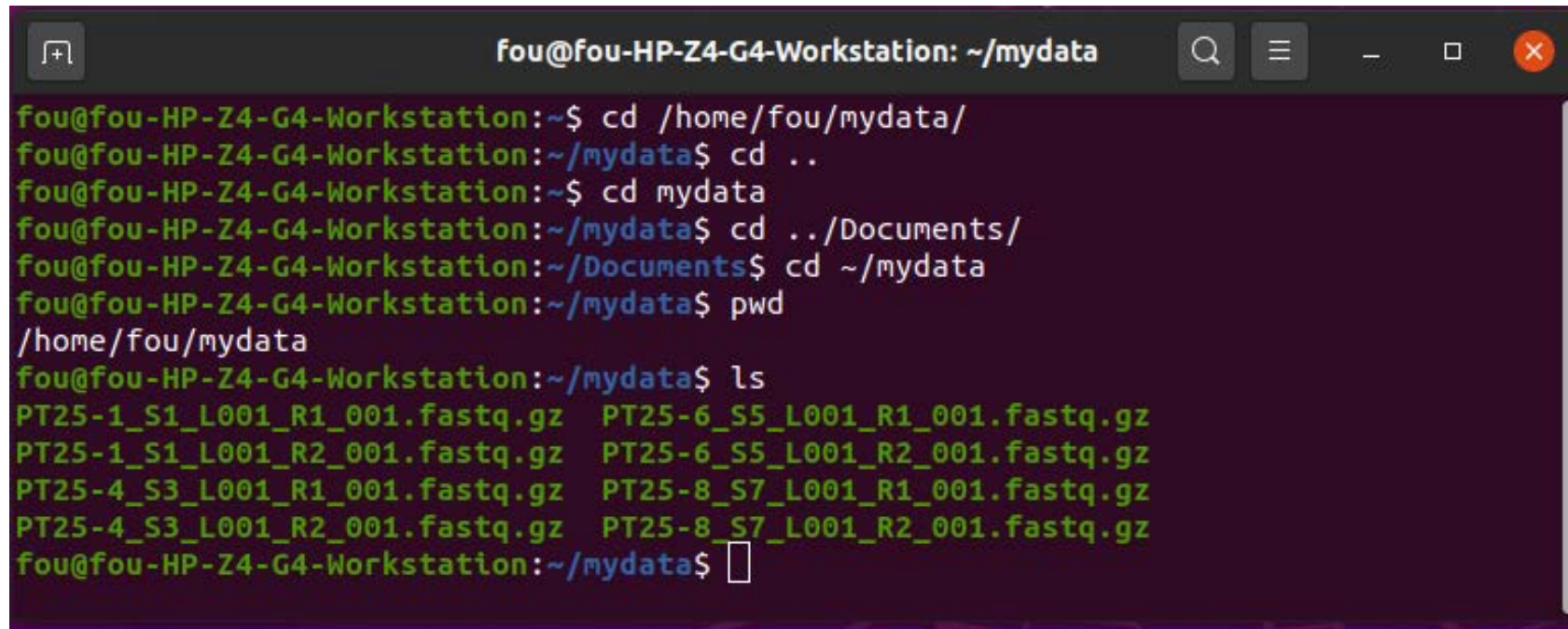
# Navigating the filesystem

pwd  
ls

print the current working directory  
list content of current directory

cd Music  
cd ..  
cd -

goto directory Music  
Up one level  
goto previous directory



```
fou@fou-HP-Z4-G4-Workstation: ~/mydata
fou@fou-HP-Z4-G4-Workstation:~$ cd /home/fou/mydata/
fou@fou-HP-Z4-G4-Workstation:~/mydata$ cd ..
fou@fou-HP-Z4-G4-Workstation:~$ cd mydata
fou@fou-HP-Z4-G4-Workstation:~/mydata$ cd ../Documents/
fou@fou-HP-Z4-G4-Workstation:~/Documents$ cd ~/mydata
fou@fou-HP-Z4-G4-Workstation:~/mydata$ pwd
/home/fou/mydata
fou@fou-HP-Z4-G4-Workstation:~/mydata$ ls
PT25-1_S1_L001_R1_001.fastq.gz  PT25-6_S5_L001_R1_001.fastq.gz
PT25-1_S1_L001_R2_001.fastq.gz  PT25-6_S5_L001_R2_001.fastq.gz
PT25-4_S3_L001_R1_001.fastq.gz  PT25-8_S7_L001_R1_001.fastq.gz
PT25-4_S3_L001_R2_001.fastq.gz  PT25-8_S7_L001_R2_001.fastq.gz
fou@fou-HP-Z4-G4-Workstation:~/mydata$
```

# Why use CLI?

If you use the exact same method all the time, according to a set method available in a commercial software – use that.

However, if you do any kind of development, research and try new methods – use Linux and CLI as well.

Most research articles involving new bioinformatics methods contains a link at the end to a git-repository or available via conda. This can be accessed via the terminal window and you then have access to their software/method on your computer

- Working with files and text with built-in functions in Linux:
  - If you want to extract the first 1000 reads from a file containing millions of reads:  
**head -n 4000 large.fastq > small.fastq**
  - Finding a sequence motif in a reads-file:  
**grep "ATCGGGC" reads.fastq**

# Why use CLI?

## **Pipelines**

Start one script/pipeline that performs a series of operations on your data. Saves time and minimizes human errors and hands-on time. Pipeline can be optimized for your specific needs (sequencing technology, analysis settings, reporting etc)

Example of a pipeline structure:

For all samples of a sequencing run, perform:

”Extract 100,000 reads | quality-trim | assemble | calculate assembly metrics | typing analysis | write report”

Since they are difficult to create – we should all decide on and share pipelines to harmonize our capabilities!  
(Still requires computer know-how to get them going and maintaining them)

# Commercial software: CLC Genomics Workbench

An all-round NGS-  
analysis software.  
Does "everything"

CLC Genomics Workbench 10.0.1

File Edit View Download Toolbox Workspace Help

Show New Save Import Export Graphics Print Launch Undo Redo Cut Copy Paste Delete

Navigation Area

CLC\_Data

- Blast2GO Example Data
  - 100seqs
    - 100seqs\_cloudblast
    - 100seqs\_mapping
    - 100seqs\_annotation\_ips
  - Workflow
  - Enrichment Analysis
    - 1\_annotations
    - 2\_test\_set
    - 3\_reference\_set
  - 100 seqs annotation table
  - Example Data
  - FOLDER
  - Multi BLAST (153974 sequences)
  - PSORTb Results
  - Blast2GO GO Data
  - NOG project
  - Rfam Result
  - Final Gene Distribution

Toolbox

- Translate Longest ORF
- Batch Rename
- BDA
- Retrieve Blast Top-Hit
- Fisher's Exact Test
  - Fisher's Exact Test
  - Reduce to Most Specific Term
- Enriched GO Graph
- Enriched Bar Chart
- Orthologous Groups
- Obtain COG Ortholog Groups
- Merge GOs from COG
- Psortb
- PSORTb
- Merge GOs from PSORTb
- Rfam
  - Rfam
  - Hit Distribution
  - Biotypes Pie
  - Biotypes Distribution
  - Hit Distribution
- Import
  - Import Blast Results
  - Import InterPro Results
  - Import Annotations
- Export
  - Export Annotations
  - Export Table
  - Export Generic

Processes | Toolbox | Favorites

Pr \* 100seqs\_annot... x | \* GO Distributi... x

### GO Distribution by Level (2) - Top 20

Category	GO Term	#Seqs
BP	metabolic process	62
	cellular process	55
	single-organism process	28
	biological regulation	18
	response to stimulus	15
	localization	8
	cellular component organization or biogenesis	6
	signaling	5
	developmental process	4
	reproduction	4
MF	catalytic activity	45
	binding	35
	structural molecule activity	10
	molecular function regulator	8
	transporter activity	5
	antioxidant activity	4
	signal transducer activity	3
	nucleic acid binding transcription factor activity	2
	electron carrier activity	2
	CC	cell
organelle		30
membrane		25
macromolecular complex		15
extracellular region		5
membrane-enclosed lumen		4
cell junction		2
symplast	1	

Pr \* Data Distribu... x

### Data Distribution Pie Chart [100seqs\_annotation\_ips]

Category	Count
B2G Annotated	88
Blasted without Hits	4
With Blast Hits	5
With GO Mapping	3

Workflow x

```

    graph TD
      A[Workflow Input] --> B[Blast2GO Project]
      B --> C[CloudBlast]
      C --> D[Blast2GO Project]
      D --> E[Blast2GO Project]
      E --> F[Mapping]
      F --> G[Blast2GO Project]
      G --> H[Blast2GO Project]
      H --> I[Annotation]
      I --> J[Blast2GO Project]
      J --> K[Blast2GO Project]
      K --> L[InterPro Scan]
      L --> M[Blast2GO Project]
      M --> N[Blast2GO Project]
      N --> O[Merge GOs from InterPro Scan]
      O --> P[Blast2GO Project]
      P --> Q[Statistics]
      Q --> R[Blast2GO Chart]
      R --> S[Blast2GO Chart]
  
```

Add Element... | Run... | Installation



# Commercial software: Ridom SeqSphere+

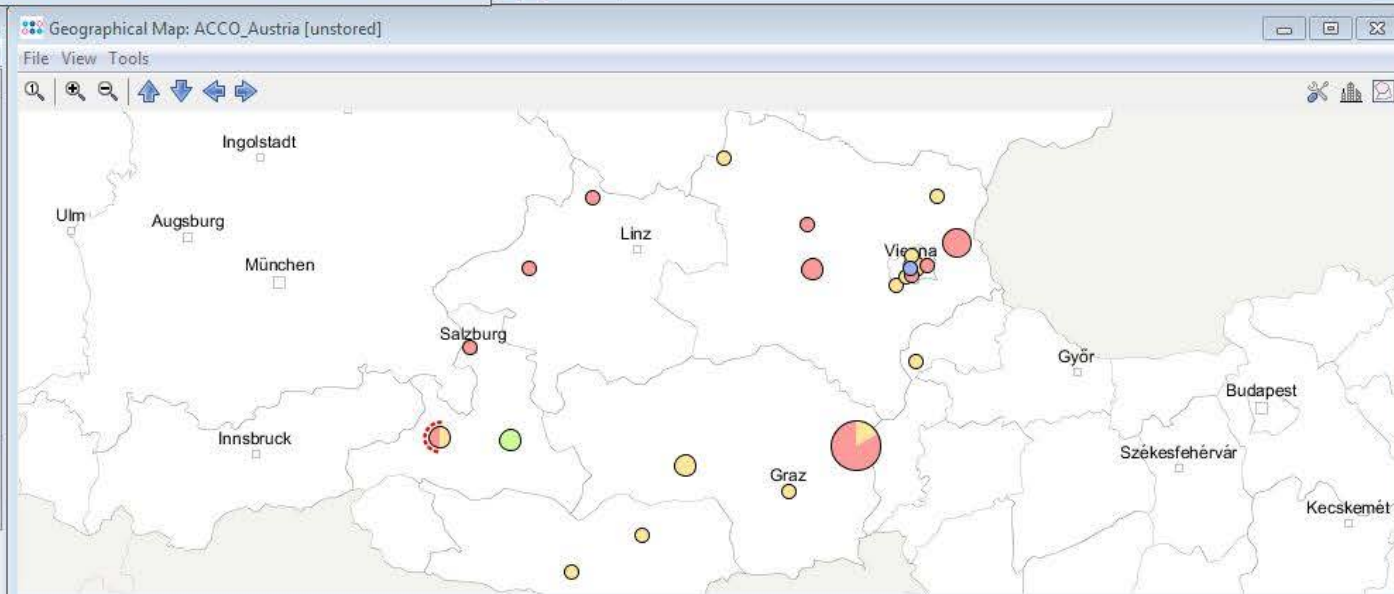
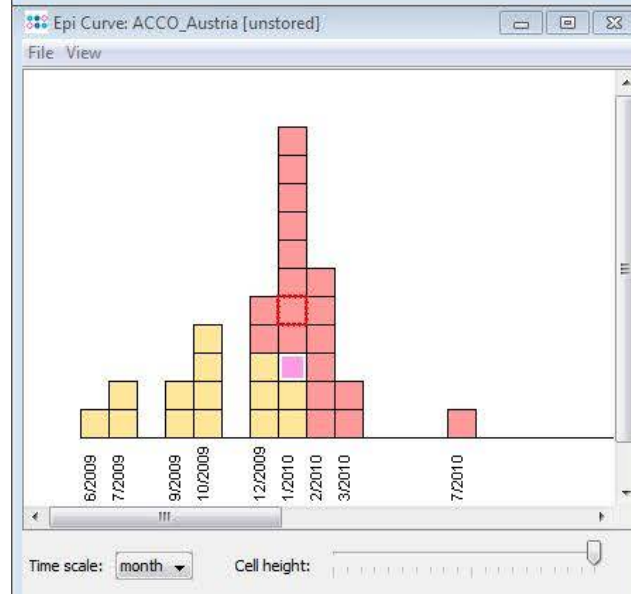
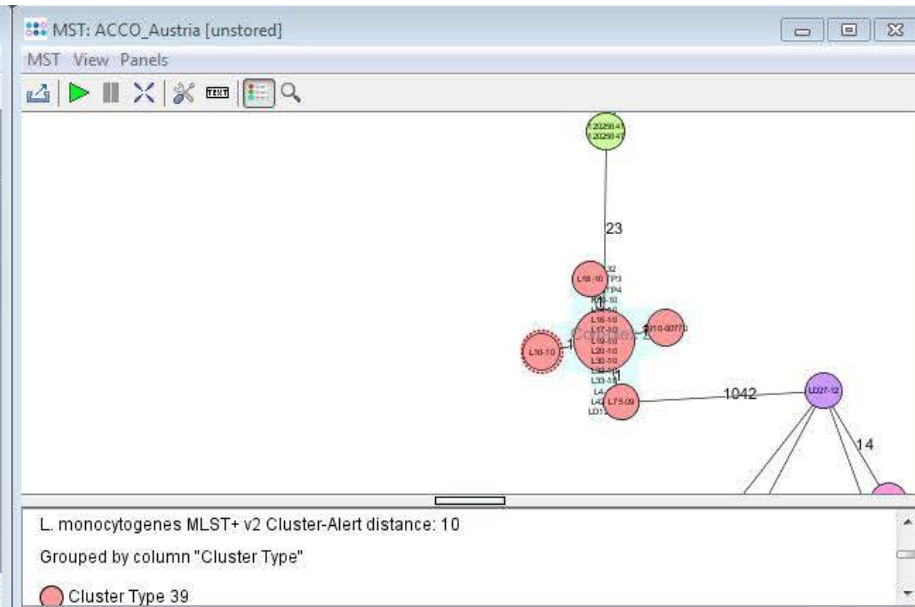
Focus on typing of  
bacterial genomes  
cgMLST and wgMLST

Assembly, mapping  
and quality control of  
data

Comparison Table: ACCO\_Austria [unstored]

#Missing v	Sample ID Sample	Perc. Good L. monocyt	ST L. m	Epi Info Source	Collection Date Sample	Country of Isolation Source	City of Isolation Source	ZIP of Isolation Source	Lat/L Source
4	12025641	99.8	398	Outgroup of ACCO II	Mar 14, 2012	Austria	? (unknown)	? (unknown)	47.3
4	12025647	99.8	398	Outgroup of ACCO II	Mar 14, 2012	Austria	? (unknown)	? (unknown)	47.3
4	16132	99.8	398	ACCO II	Mar 3, 2010	Austria	Hartberg	8230	47.3
4	2010-00770	99.8	398	ACCO II	Feb 2, 2010	Austria	Hartberg	8230	47.3
4	3230TP3	99.8	398	ACCO II	Jan 22, 2010	Austria	Hartberg	8230	47.3
3	3230TP5	99.8	403	ACCO I	Jan 22, 2010	Austria	Hartberg	8230	47.3
4	4548TP4	99.8	398	ACCO II	Jan 26, 2010	Austria	Hartberg	8230	47.3
0	EGD-e	100.0	35	NCBI RefSeq	1924	United Kingdom	Cambridge	? (unknown)	?
4	K70-10	99.8	398	ACCO II	Jul 5, 2010	Austria	Hartberg	8230	47.3
4	L10-10	99.8	398	ACCO II	Jan 12, 2010	Austria	Zell am See	5700	47.3
4	L14-10	99.8	398	ACCO II	Jan 25, 2010	Austria	Rohrbach	4150	48.5
4	L16-10	99.8	398	ACCO II	Jan 29, 2010	Austria	Salzburg Land	5020	47.8
4	L17-10	99.8	398	ACCO II	Jan 30, 2010	Austria	Krems Land	3500	48.4
4	L18-10	99.8	398	ACCO II	Feb 2, 2010	Austria	Gänserndorf	2230	48.3
4	L19-10	99.8	398	ACCO II	Feb 2, 2010	Austria	Gänserndorf	2230	48.3

1701 of 1712 columns used for distance calculation | 1 Sample selected | 40 Samples





## Commercial software: Bionumerics

Specializes in typing applications. Will however be discontinued. A totally new software is planned

The screenshot displays the Bionumerics software interface for a sequence named 'inRA001'. The interface is divided into several sections:

- Sequence editor:** Shows a DNA sequence in a grid format. The sequence is: AATGGAAAAA AACGTCACCTG TTACACACCGC CCAAGACATA CTGGAAAAGA CACACAACGG GAAACTCTGC 210  
GATCTAGATG GAGTGAAGCC TCTAATTTTA AGAGATTGTA GTGTAGCTGG ATGGCTCCTC GGGAACCCAA 280  
TGTGTGACGA ATTCTCAAT GTGCCGGAAT GGTC+tacat agtggagaag atcaatccag ccaatgaect 350  
ctgttaccoc ggaatttca acgactatga agaactgaaa cacctattga gcagaataaa coattttgag 420  
aaaattcaga tcatccocaa aagttottgg teagatcatg aagcctcctc aggggtgagc teagcatgtc 490  
cataccaggg aaggtcctcc ttttttagaa atgtggtatg gcttatcaca aaggacaatg catatccac 560
- Sequence viewer:** A graphical representation of the sequence with a scale from 40 to 170. A blue bar labeled 'SoxS' is positioned above the sequence, indicating a specific region.
- Annotation:** A table showing a single feature: a CDS (Coding DNA Sequence) from position 102 to 314, with a length of 212. The feature list table is as follows:

Feature...	Start	End	Length
1 CDS	102	314	212
- Annotation details:** A panel on the right provides detailed information for the selected feature:

```
102..314
/codon_start=1
/translation_table=11
/gene="SoxS"
/product="regulatory protein SoxS"
/protein_id="ABV19788.1"
/db_xref="GI:157080080"
/translation="LPCKQLDRAG"
```
- Bottom navigation:** A series of tabs for different analysis tools: Annotation, Header, Custom Fields, Sequence search, Frame analysis, and Restriction analysis.

# Commercial software: Geneious Prime

Not specific for  
typing. One of the  
more affordable  
solutions for working  
with sequence data in  
general

The screenshot displays the Geneious Prime interface. The top toolbar includes navigation (Back, Forward), file management (Add, Export), and analysis tools (BLAST, Workflows, Align/Assemble, Tree, Primers, Cloning, Help). A search bar is located in the top right corner.

The left sidebar shows a project tree with folders for Local, Reference Features, Shared Databases, Operations, and NCBI. The 'Local' folder is expanded, showing 'Sample Documents' and 'GeneiousDB'.

The main window displays a table of primers with the following data:

Name	Description	Sequence Length	Tm
M13-F (-20)	M13 forward sequencing primer, 20bp upstream	17	54.7
M13-F (-40)	M13 forward sequencing primer, 40bp upstream	17	52.6
M13-F (-46)	M13 forward sequencing primer, 46bp upstream	22	65
M13-R (-26)	M13 reverse sequencing primer, 26bp upstream	17	49.1
M13-R (-46)	M13 reverse sequencing primer, 46bp upstream	24	60.4
SP6 promoter	SP6 promoter sequencing primer, 24-mer	24	54.5
T3 promoter (17bp)	T3 promoter sequencing primer, 17-mer	17	44.2

Below the table, the 'Sequence View' tab is active, showing the sequence alignment of the selected primers. The alignment is displayed as a grid with sequence positions 1, 10, 20, and 24 marked. Green bars above the sequences indicate 'Binding Regions'.

1. M13-F (-20) GTAAAACGACGGCCAGT  
2. M13-F (-40) GTTTTCCTCAGTCACGAC  
3. M13-F (-46) GCCAGGGTTCCTCAGTCACGA  
4. M13-R (-26) CAGGAAACAGCTATGAC  
5. M13-R (-46) GAGCGGATAACAATTCACACAGG

At the bottom of the window, a status bar indicates '1,370 / 14,380 MB Memory' and provides instructions: 'Alt click on a sequence position or annotation, or select a region to zoom in. Alt-shift click to zoom out.'

Public server:  
**Galaxy**

User-friendly service  
where many (most)  
bioinformatics  
operations and  
programs can be used  
and put into  
workflows.

Start with raw reads  
and design own  
workflow

You can share your  
workflow with other  
users

The screenshot shows the Galaxy web interface at <https://usegalaxy.org>. The main content area displays the configuration for the **Cutadapt** tool, which is used to remove adapter sequences from FASTQ/FASTA files. The tool version is 3.7+galaxy0.

**Single-end or Paired-end reads?**  
Single-end

**FASTQ/A file**  
No fastq.gz, fastq or fasta dataset available.  
Should be of datatype "fastq.gz" or "fasta"

**Read 1 Options**

- 3' (End) Adapters**  
+ Insert 3' (End) Adapters
- 5' (Front) Adapters**  
+ Insert 5' (Front) Adapters
- 5' or 3' (Anywhere) Adapters**  
+ Insert 5' or 3' (Anywhere) Adapters

**Cut bases from reads before adapter trimmina**

The left sidebar contains navigation options: Tools (with a search bar and Upload Data button), Get Data, Send Data, Collection Operations, GENERAL TEXT TOOLS (Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash), GENOMIC FILE MANIPULATION, and FASTA/FASTQ (Trim Galore! Quality and adapter trimmer of reads).

# Public server: CGE

- Species identification
- VirulenceFinder
- ResFinder
- MLST
- Etc

# Center for Genomic Epidemiology

Home

Organization

Project

Services

Contact

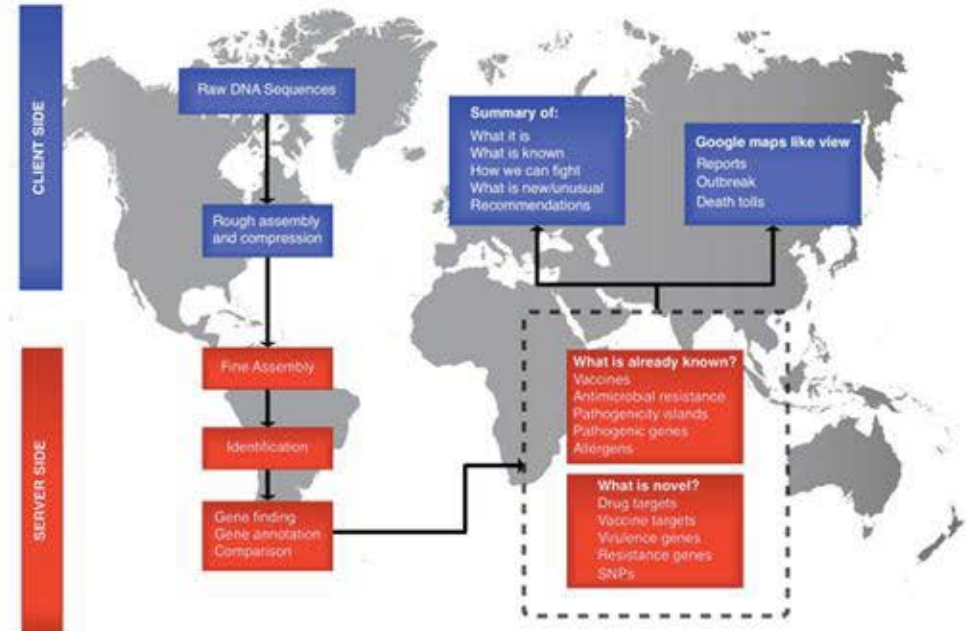
## Services

### Phenotyping:

- Identification of acquired antibiotic resistance genes. [ResFinder](#)
- Prediction of a bacteria's pathogenicity towards human hosts. [PathogenFinder](#)
- Identification of acquired virulence genes. [VirulenceFinder](#)

### Typing:

- Multi Locus Sequence Typing (MLST) from an assembled genome or from a set of reads [MLST](#)
- PlasmidFinder identifies plasmids in total or partial sequenced isolates of bacteria. [PlasmidFinder](#)
- Multi Locus Sequence Typing (MLST) from an assembled plasmid or from a set of reads [pMLST](#)
- Prediction of bacterial species using a fast K-mer algorithm. [KmerFinder](#)



## Welcome to the Center for Genomic Epidemiology

The cost of sequencing a bacterial genome is \$50 and is expected to decrease further in the near future and the equipment needed cost less than \$150 000. Thus, within a few years all clinical microbiological laboratories will have a sequencer in use on a daily basis. The price of genome sequencing is already so low that whole genome sequencing will also find worldwide application in human and veterinary practices as well as many other places where bacteria are handled. In Denmark alone this equals more than 1 million isolates annually in 15-20 laboratories and globally

## News

**Course on the use of the CGE tools in November 2014**  
September 2014  
The course is for clinical microbiologists to learn how to use the CGE tools. The course will be taught in English and take place at the Technical University of Denmark [Course flyer \(pdf\)](#)

**Benchmarking of Methods for Genomic Taxonomy**  
April 2014  
How to optimally determine taxonomy from whole genome sequences. [Link to article...](#)

**CGE tools applied for bacteriophage characterization**  
March 2014  
Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences. [Link to article...](#)

**Evaluation of Whole Genome Sequencing for Outbreak Detection of Salmonella enterica**  
March 2014  
We evaluated WGS for outbreak detection of Salmonella enterica including different approaches for analyzing and comparing with a traditional typing, PFGE. [Link to article...](#)

**Low-bandwidth and non-compute intensive remote identification of**

Public server:

Via EURLs

- Aries (E. coli)
- Starflow (Listeria)

Organism-specific  
WGS analysis servers



Istituto Superiore di Sanita'

ARIES - Advanced Research Infrastructure for Experimentation  
in Genomics - Galaxy Instance at ISS

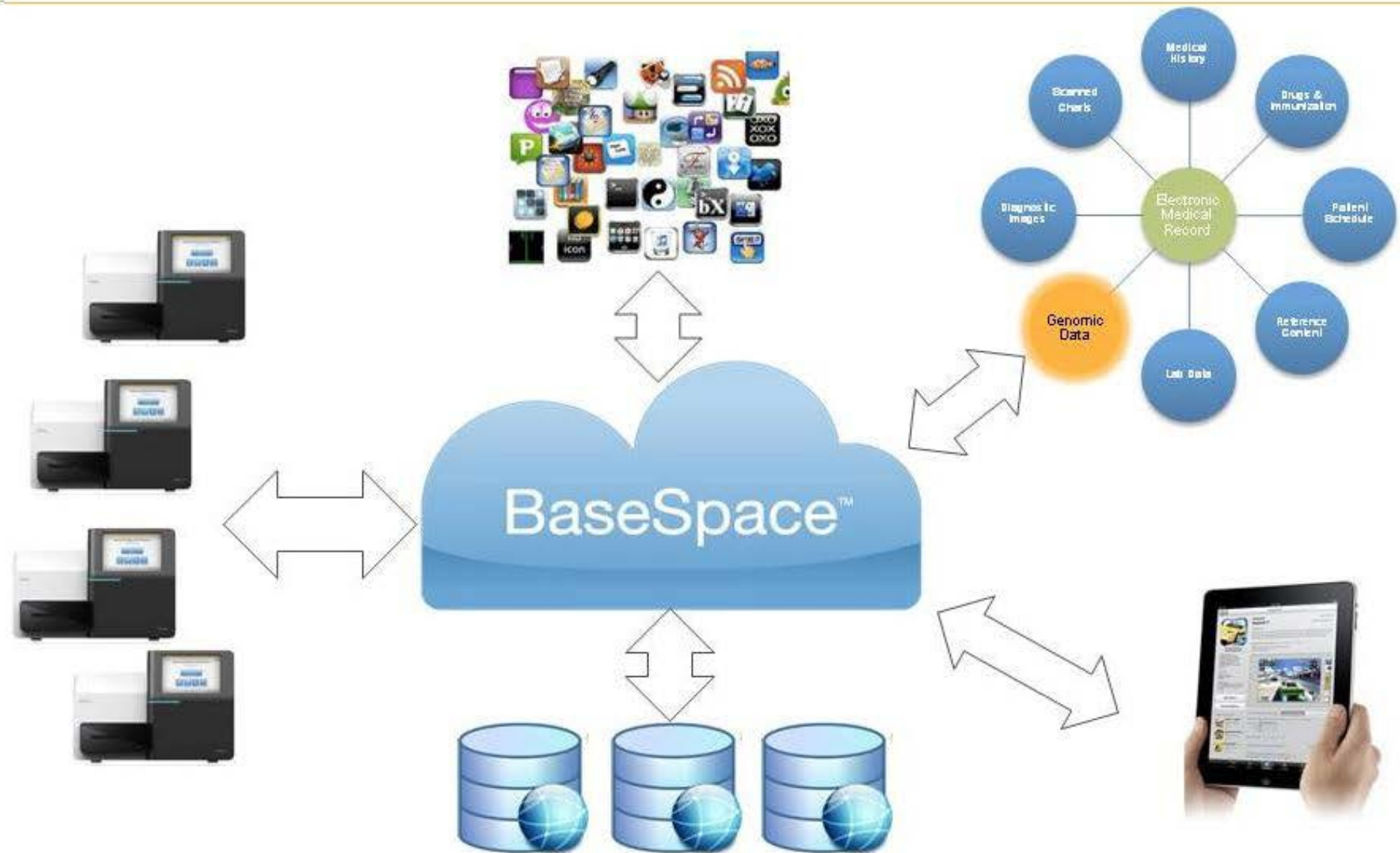


Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology



Public server:  
**Illumina BaseSpace**  
Cloud-based NGS  
analysis service.

Free to use basic  
functions but requires  
(?) paid subscription  
to access everything



## Example – what we use at the Swedish Food Agency



### Windows computer

- Commercial software for cgMLST, used for outbreak investigations and surveillance
- Linux app that controls a pipeline for: QC, trim, assembly, serotyping, AMR, SNP-analysis etc.
- Online servers used to quickly determine a sequence type or AMR or species

### Linux computer

- CLI-workflows written in Python for metagenomics, assembly, QC, contamination checks and SNP-typing etc
- Powerful enough to handle large metagenomics datasets

### Windows laptops

- Commercial software for general NGS analysis

Questions?