Project handling - guidelines

Good practices in bioinformatics

Sarah Bonnin

CRG – May 7th, 2018

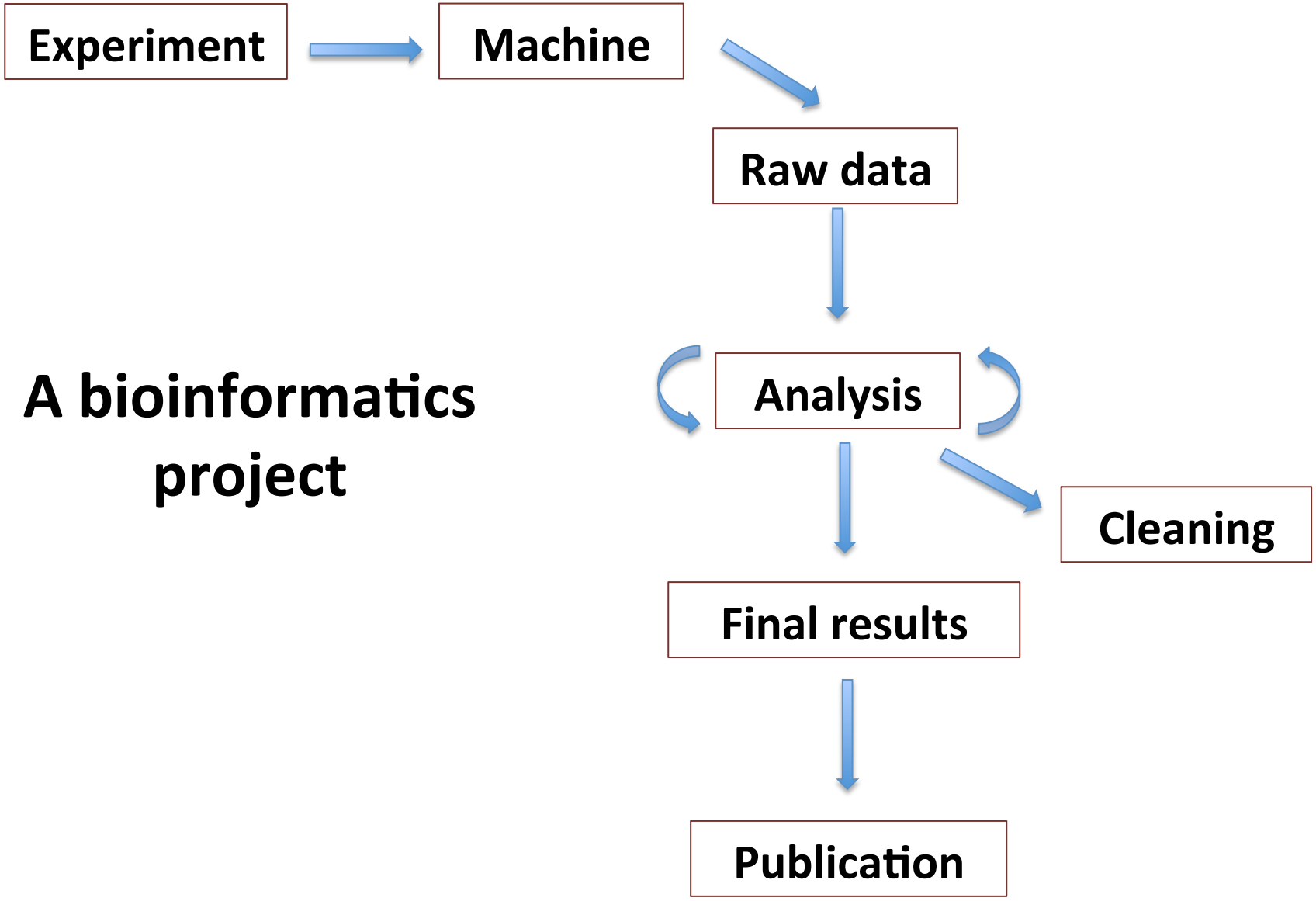
Learning objectives

- How to organize a bioinformatics project.
- Suggestions on how to properly structure the project / data.

Learning outcomes

- Create the base structure of a project
- Know what is a README file and how to write it.
- Which data to store, which data to trash.

A bioinformatics project



Challenges

- Data organization and management.
- Space management.
- Analysis/pipeline efficiency.
- Reproducibility of the analysis:
  - by **someone else**
  - by **yourself in months/years !**

Raw data

- "data that comes out of a machine".
- Often large / heavy.

→ **Necessary to reproduce the analysis, if ever needed!**

→ **Required for <u>publishing</u>!**

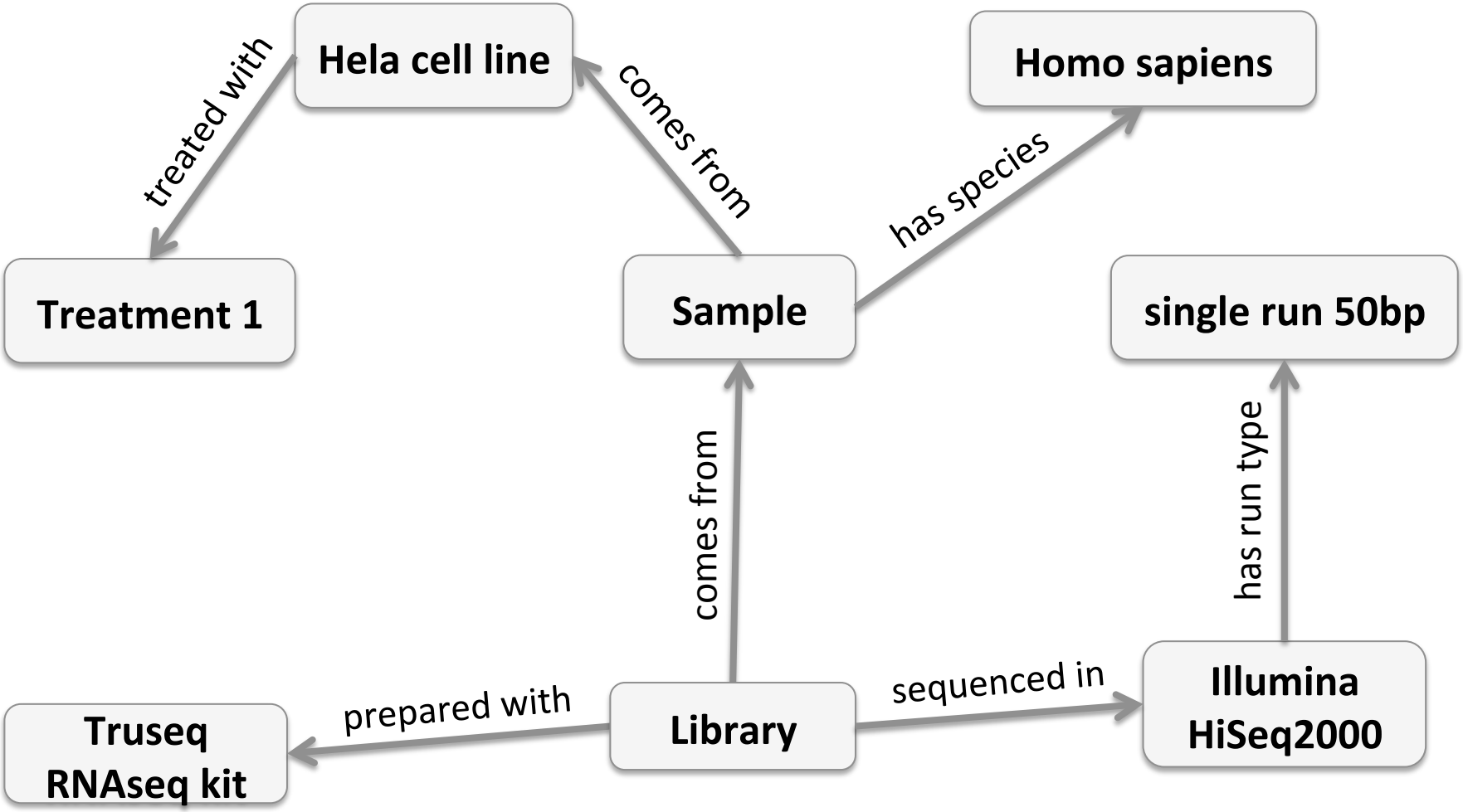
Raw data files examples

- High Throughput Sequencing
  - fastq (Illumina, PacBio)
- Mass spectrometry:
  - mzML, mzXML, netCDF, mzData
- Microarrays:
  - text (Agilent) or CEL (Affymetrix) files

Metadata

- **Samples**
  - origin (tissue, cell line)
  - experimental specificities (treatments, times, age)
- **Experimental design**
  - number of samples and conditions
- **Experimental protocol**
  - Starting material: concentration, volume.
  - Reagents, kits.
  - Machine types/models.

Metadata


Hela cell line
Homo sapiens
treated with
comes from
has species
Treatment 1
Sample
single run 50bp
comes from
has run type
prepared with
Library
sequenced in
Truseq
RNAseq kit
Illumina
HiSeq2000

Analysis

- **README** file (text file) : your lab book !
  - Tells the story of your project.
  - Explains HOW and WHY each step is done.

→ Following it should enable one to **reproduce the analysis and the results**.

README files

Mostly used in software deployment:

- Program name and version
- Date of distribution
- Introduction / short description
- Installation requirements / dependencies
- Configuration
- Example code
- Contact / author / license information
- etc.

README adapted to a project: suggestions

- Project title + short description
- Author
- Date
- Analysis steps:
  - linear
  - name and version of software
  - comments!

Analysis

Keep track of:

- Programs and methods:
  - references / versions
  - arguments / options
- Genome and annotation versions
- Intermediate files / temporary data

Analysis

Naming of files: good / explicit:

- sample1_results.txt 
- 201805_WT_rep1_counts_mm10_htseq.txt 

Cleaning up!

- Risk of drowning in data
  - remove what can be removed!

- Risk of bankruptcy
  - more data = more storage = **€€€€**
  - 7€ / Tb / month

Cleaning up
Temporary files

- Created by a program
  - hold information temporarily
- Usually deleted by the program
- Kept if:
  - program abnormally stopped / failed run
  - program defaults to keeping temporary files

Cleaning up
Intermediate / log files

- Various attempts (options) to run a program:
  - remove the non final versions (report which is kept and why in the README)
- "Log" files
  - records of software runs
  - depending on the information they hold, could be removed

Results
What to keep from a project?

- Raw data
  - keep original files (with original file names)
- Relevant documentation
  - README
  - metadata
- Annotation and genome version
  - mm10 or mm9? ENSEMBL or RefSeq?
- Final processed data / results

Example project
as done in the Bioinformatics core facility

RNA-seq project

Assessment of gene expression levels using Next-Generation Sequencing technologies

Goal:

Identify differentially expressed genes between KO (Knock out of the Mstn gene) and WT (Wild Type).

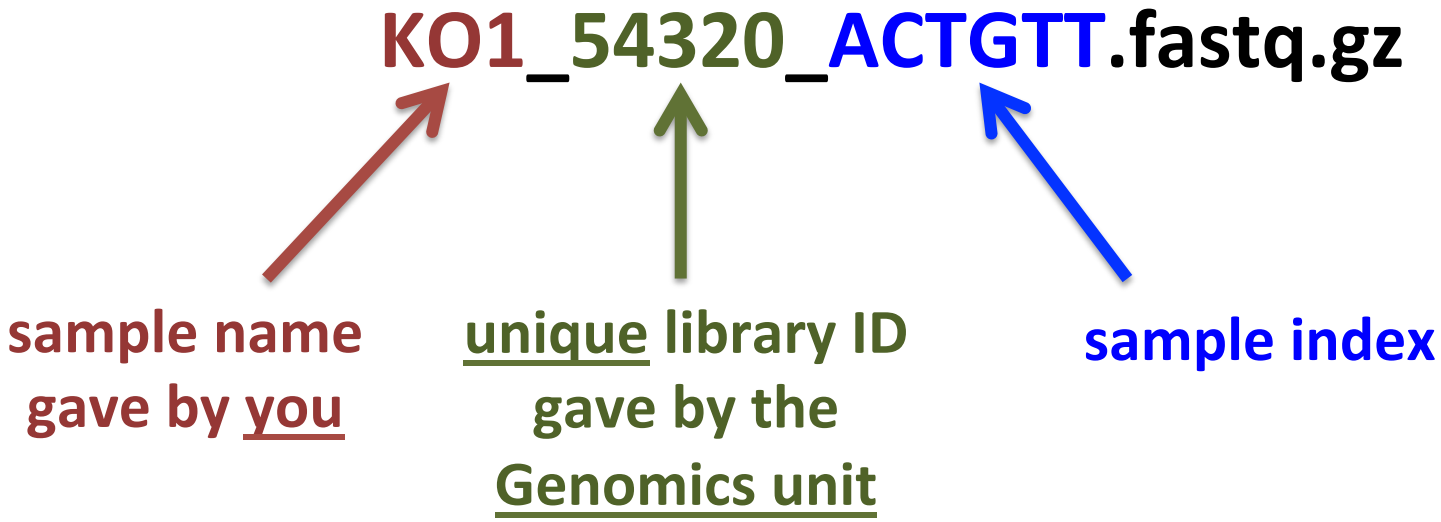
Project organization

|-- data	Raw data
|-- docs	README + relevant documentation
|-- analysis	Analysis and intermediate steps
|-- src	Scripts that are used for the analysis
|-- results	Final results

Raw data: fastq files

KO1_54320_ACTGTT.fastq.gz
KO2_54321_TCTAGT.fastq.gz
KO3_54322_CCAGTA.fastq.gz

WT1_54323_AGTGCA.fastq.gz
WT2_54324_ACTGTT.fastq.gz
WT3_54325_GTTGAG.fastq.gz



Docs: README

Title: RNA-seq project: Knock out of the Myostatin (Mstn) gene.

Summary / short description: study of the gene expression changes provoked by a KO of Mstn in Mus musculus muscle cells

Date: April 2018

Author: Sarah Bonnin – sarah.bonnin@crg.eu

Experimental design: 6 samples in triplicates (3 Knock Out and 3 Wild Type)

Docs: README

```
# 1. Quality control of the raw data:
# FastQC v0.11.5
cd [path_to_project]/analysis/fastqc

fastqc [path_to_project]/data/*.fastq.gz

# 2. Mapping samples to reference genome
# STAR v2.5.3a
cd [path_to_project]/analysis/star

for fq in [path_to_project]/data/*.fastq.gz
do
      qsub –N star –v $fq [path_to_project]/src/star.sh
done

# 3. Differential expression analysis
# DESeq2 v1.14.1
cd [path_to_project]/analysis/deseq2
…
```

Analysis folder

One folder per step in the analysis:

- |-- deseq2
- |-- fastqc
- |-- star

You can name the folders by their respective order in the analysis pipeline:

- |-- 1_fastqc
- |-- 2_star
- |-- 3_deseq2

Scripts / **src folder**

- Scripts run – reported in the README
  - launched locally
  - launched on the CRG cluster

→ The scripts can also be kept in a **common repository** for the lab/group folder or in your personal folder.


DATA
PROCESSING
CENTER
RAW
DATA
BAR-B-QUED
DATA
©1992,1994 HARLEY SCHWADRON

Results folder

Only **final results** and reports:

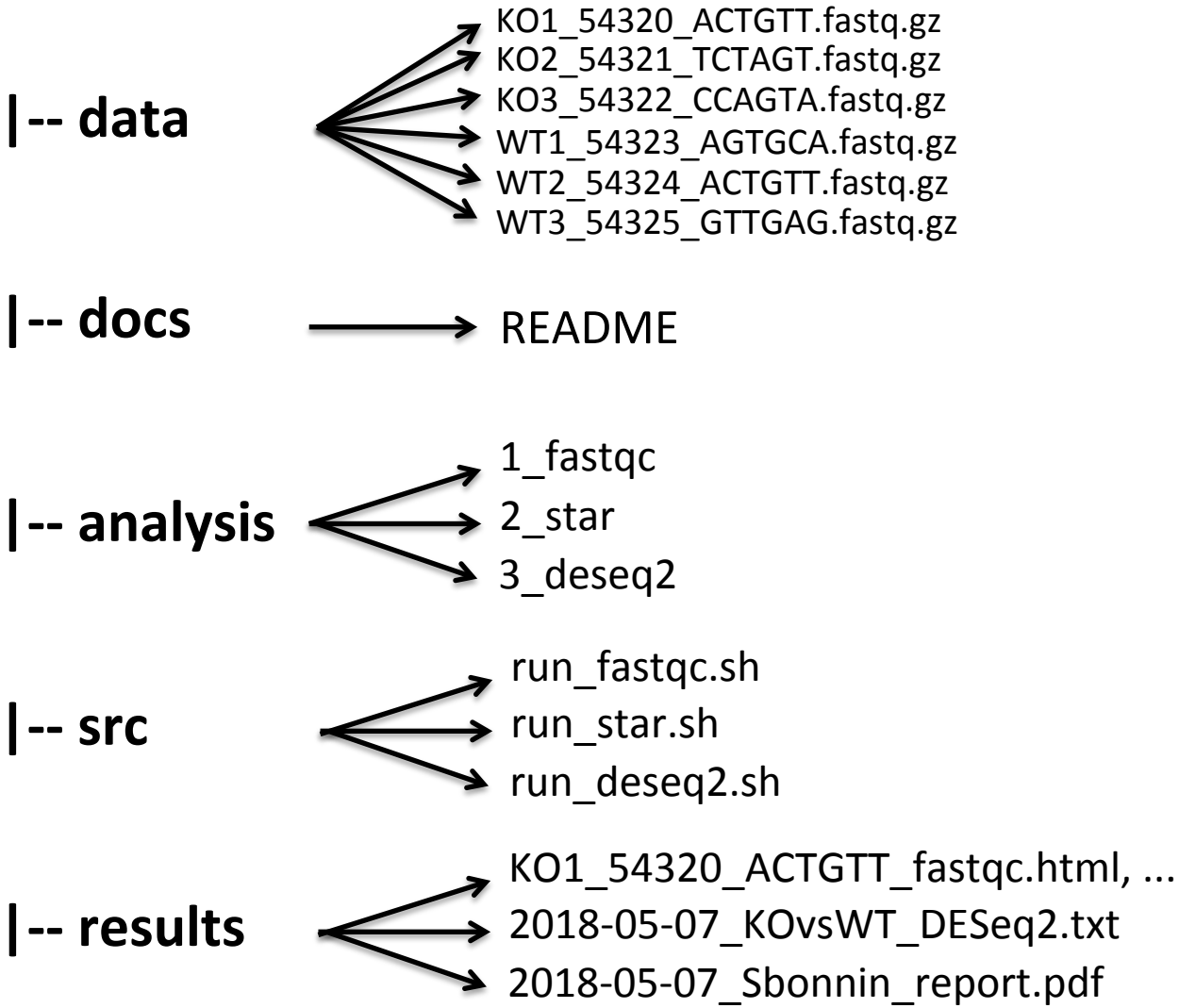
→FastQC reports.

→differential expression analysis table.

→analysis report.

What is needed to send to users/collaborators.

Project organization



Submission of raw data to public repository

→ ***Required for publishing.***

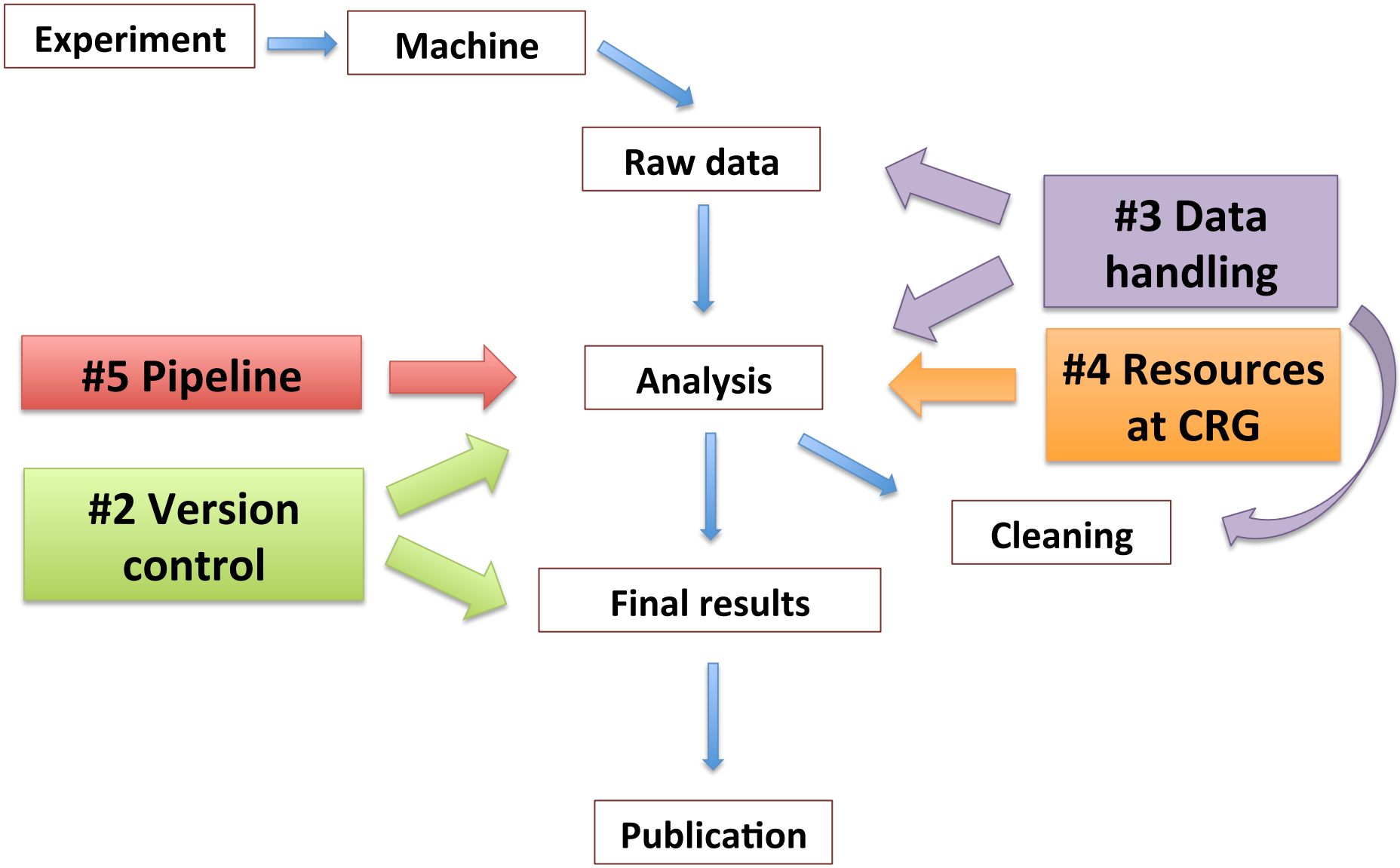
Examples of repositories:

- GEO (Gene Expression Omnibus)
- ArrayExpress
- SRA (Short Read Archive)

Submission of raw data to public repository

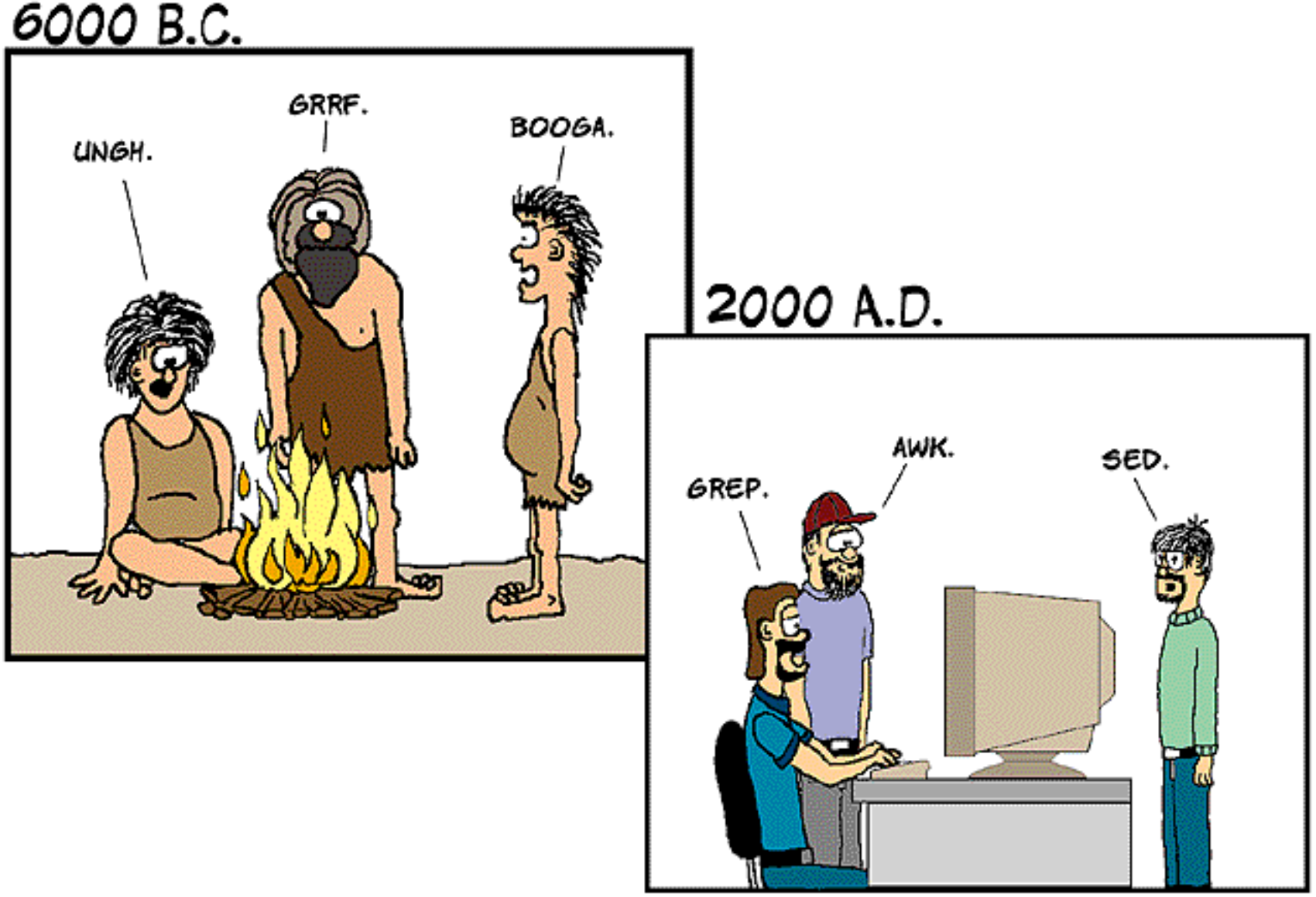
Requirements:

- **Raw data (fastq files)**
- **Processed data (counts per gene)**
- **Metadata file:**
  - experimental protocol
  - sequencing protocol
  - analysis protocol

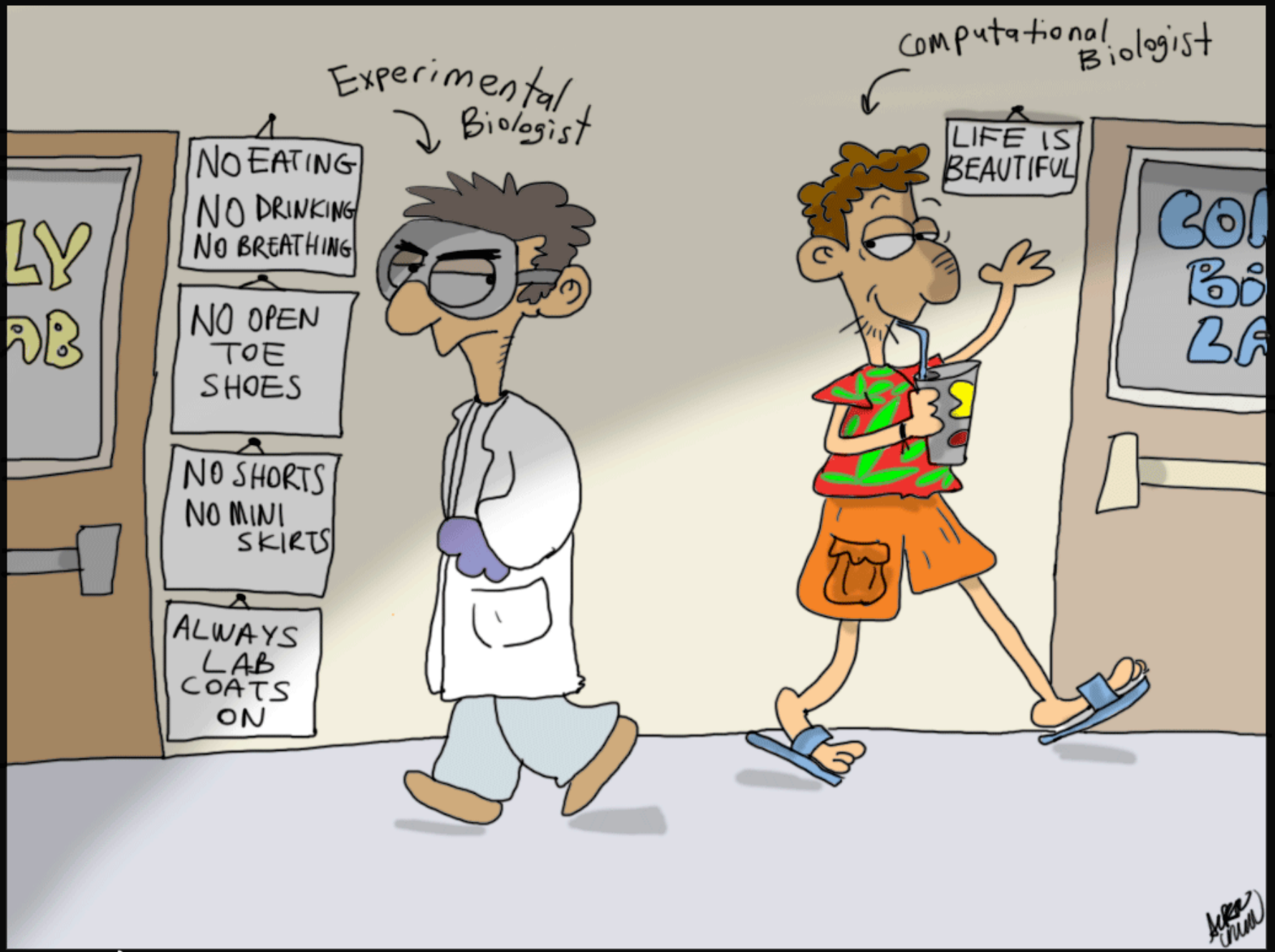

Experiment
Machine
Raw data
#3 Data handling
#5 Pipeline
Analysis
#4 Resources at CRG
#2 Version control
Cleaning
Final results
Publication

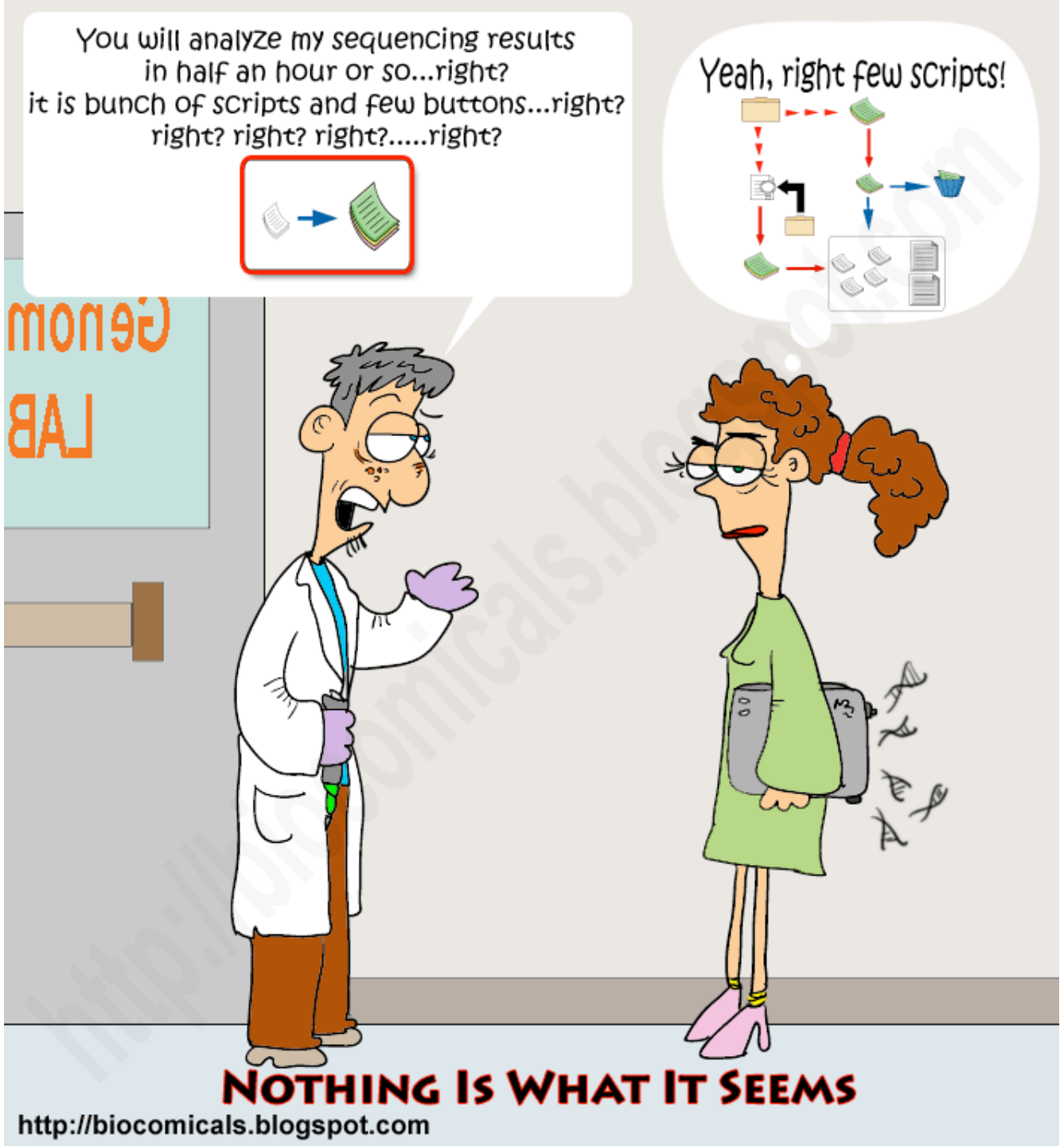
EVOLUTION OF LANGUAGE THROUGH THE AGES.

(THIS 'TOON IS A REPEAT)



HTTP://WWW.USERFRIENDLY.ORG/




You will analyze my sequencing results
in half an hour or so...right?
it is bunch of scripts and few buttons...right?
right? right? right?.....right?
Yeah, right few scripts!
LAB
Genom
NOTHING IS WHAT IT SEEMS
http://biocomicals.blogspot.com