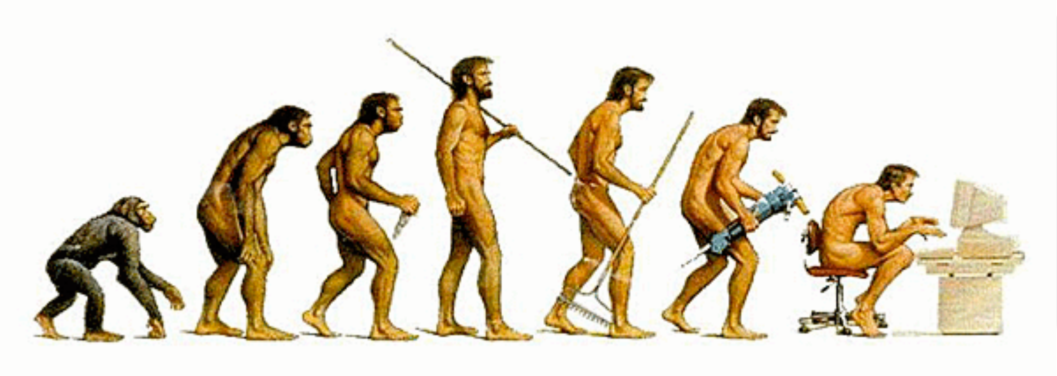
MeetJob 2013

La professione del bioinformatico



Paolo Uva
CRS4 Bioinformatica
Oristano, 23 Aprile 2013

A typical day at work…

Morning

- Discuss with the PI the design of the new experiment
- Received 0.5 TB data from the Wet Lab and transferred to our analysis platform
- Start Quality Control ….
- …something looks strange ➔ go back to the lab!!

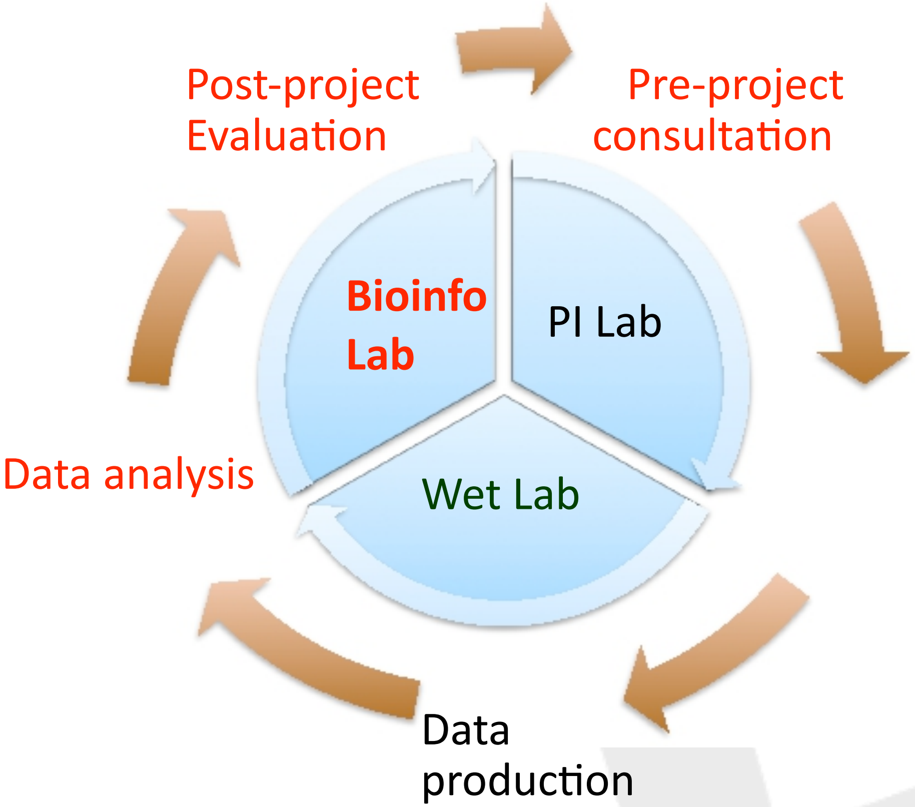
Afternoon

- Mr. X published a new software (Unix based) for the analysis of our data…
- …download, install and run a test using 48 processors on the computer cluster
- Write report
- Before leaving, re-launch a custom software overnight

Agenda

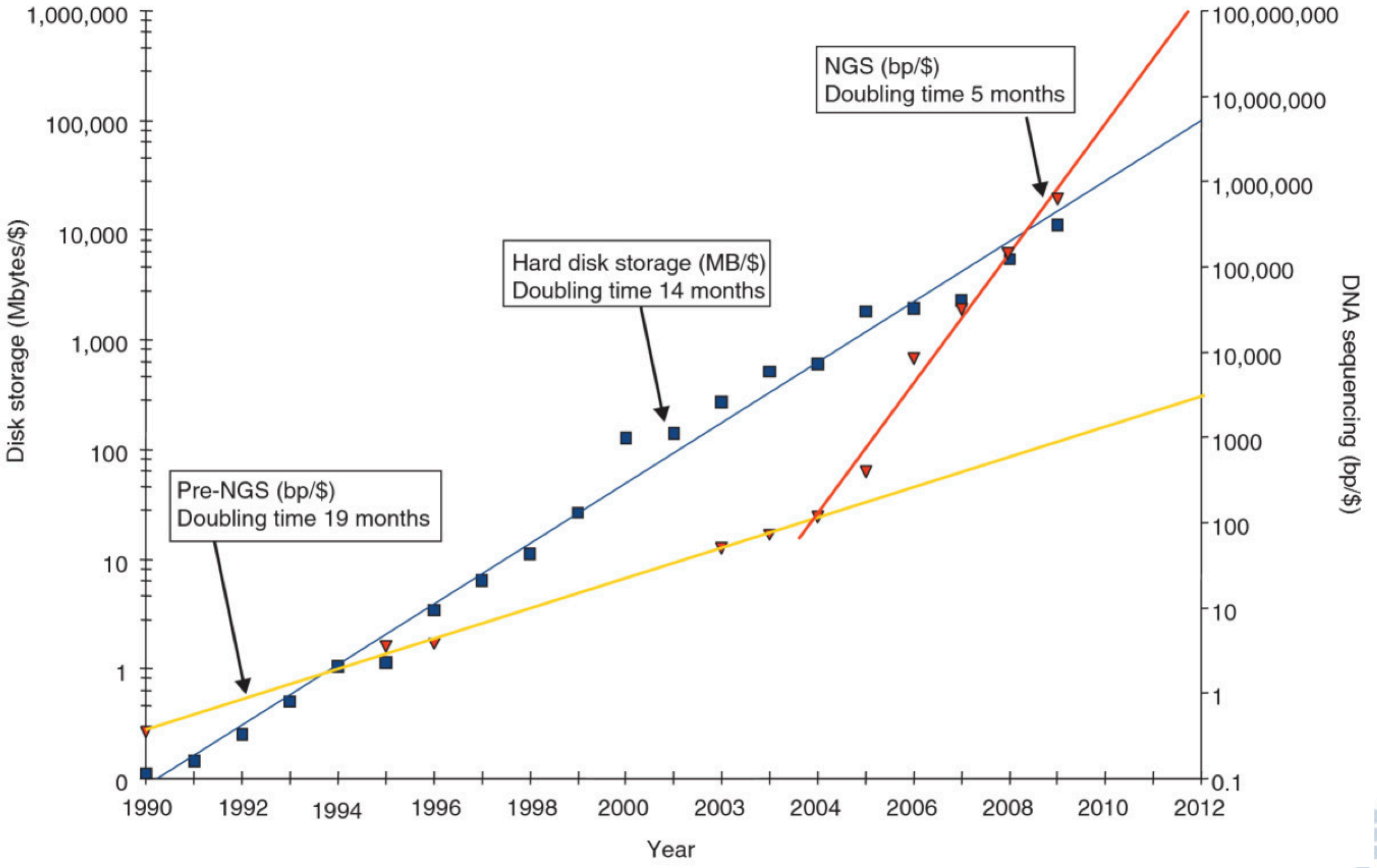
- Bioinformatics today
  - Next Generation Sequencing
- The “ideal” Bioinformatician
- How to become a BI
  - Required skills
- Bioinformatics at CRS4

Bioinformatics today



Ability to influence experimental design, early involvement

Sequencing cost decreases

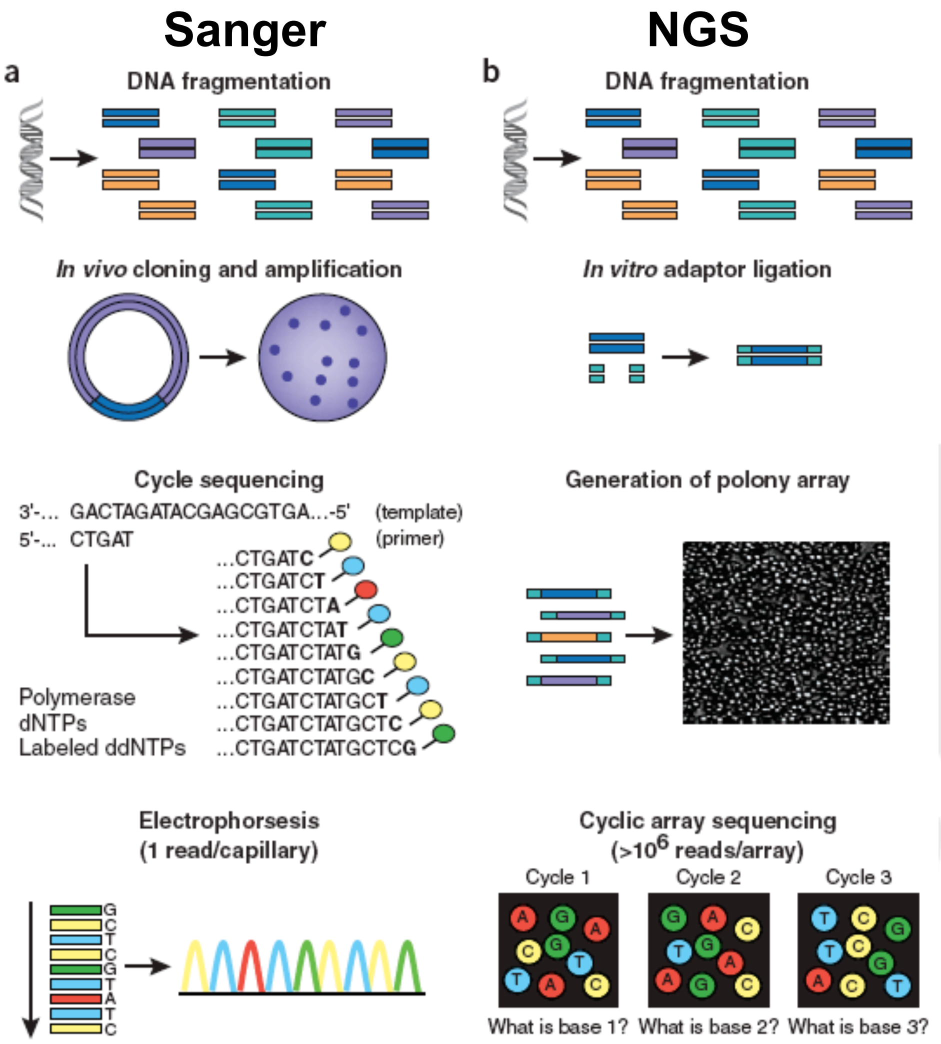


Sequencing (r)evolution

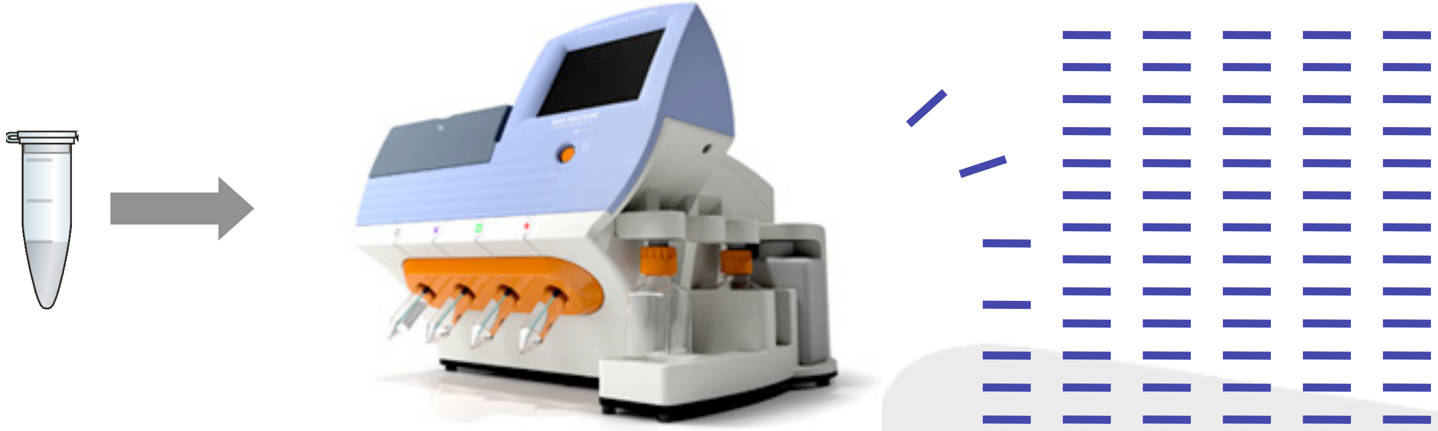
Pre-Sequencing
parallelization
speed

Sequencing
huge amount of nt
sequenced
inexpensive

Post Sequencing
Quality control
Assembling
Aligning



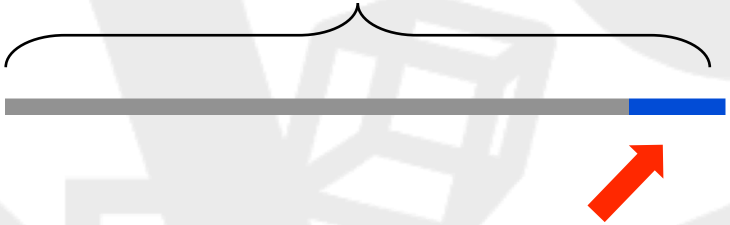
NGS – The basics



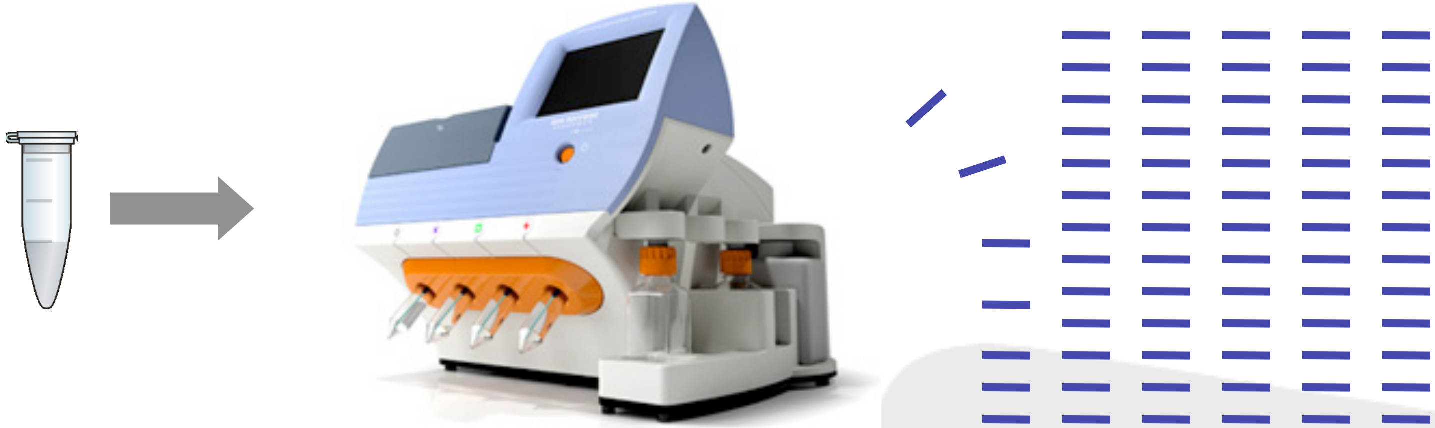
1 sample

Millions of short sequences (reads)

Only the ends of a library of fragments are actually sequenced



NGS – The basics



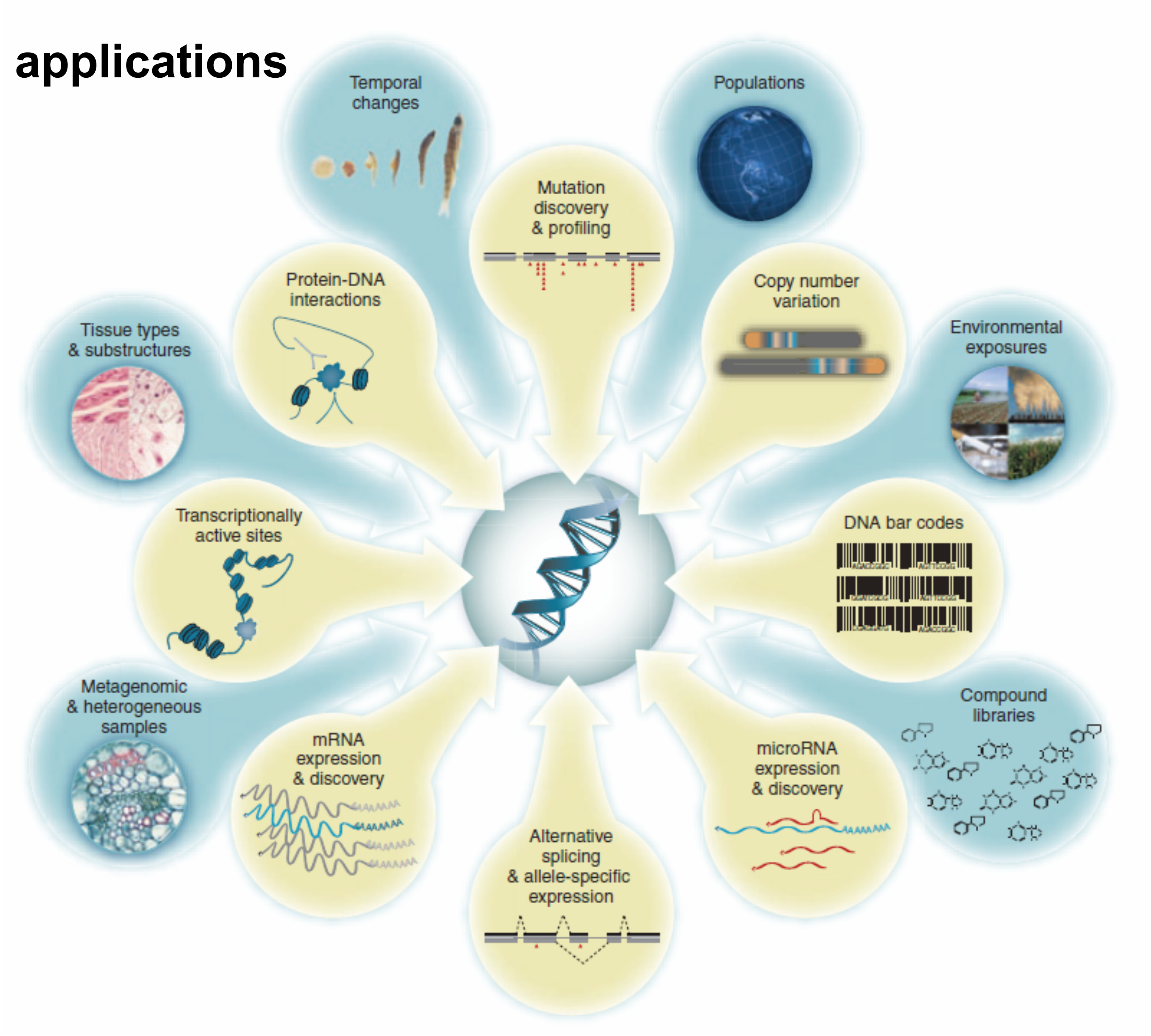
1 sample

Millions of short sequences (reads)

Different DNA/RNA sources for different purposes:

- **RNAs**: transcriptomic analysis (RNA-Seq), non-coding
- **genomic DNA**: genome assembly, protein binding sites detection (ChIP-Seq), epigenetics, metagenomics, SNPs detection

NGS applications


Temporal changes
Populations
Mutation discovery & profiling
Protein-DNA interactions
Copy number variation
Tissue types & substructures
Environmental exposures
Transcriptionally active sites
DNA bar codes
Metagenomic & heterogeneous samples
Compound libraries
mRNA expression & discovery
microRNA expression & discovery
Alternative splicing & allele-specific expression

CRS4 IDEAS BECOME LIFE

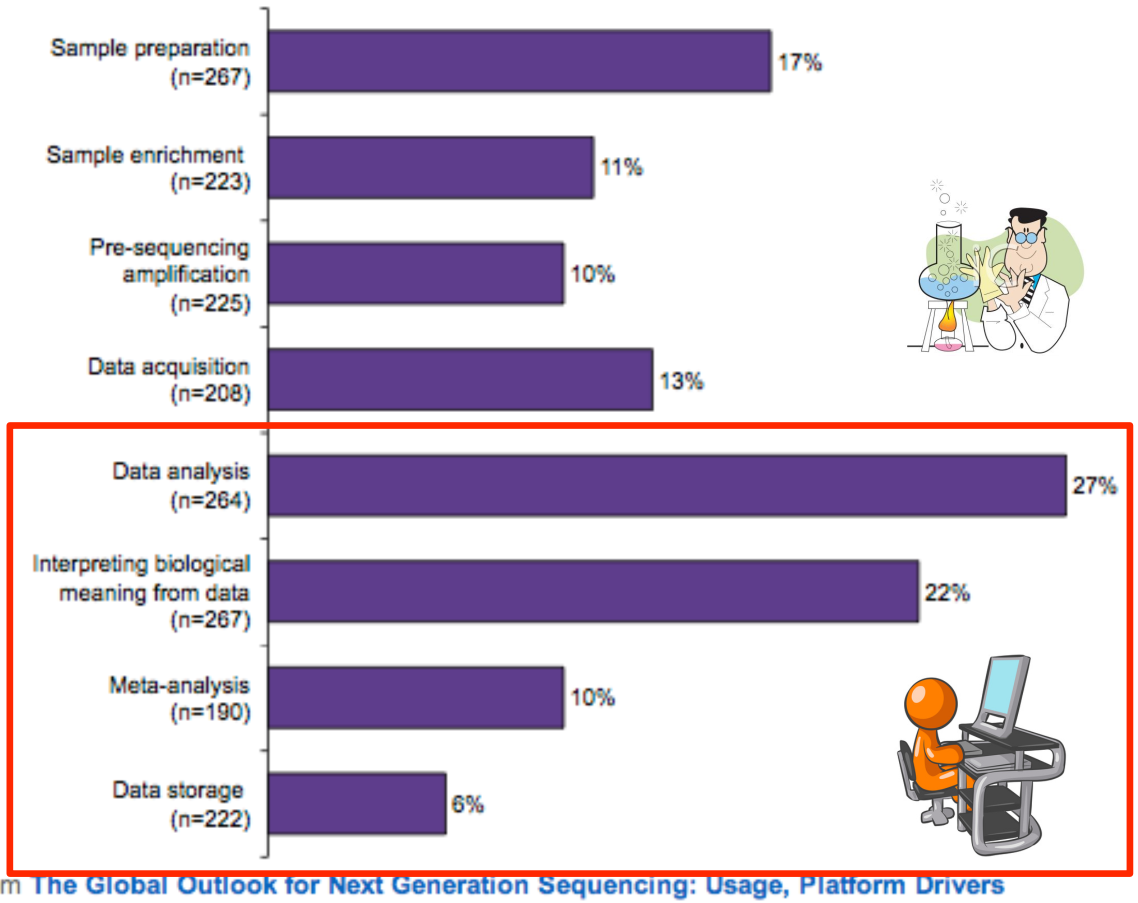
NGS use-cases

Genomics

Application	Data source	Analysis strategy
Variant calling	Genomic DNA from individuals (healthy vs disease)	Alignment of reads to reference genome and detection of variants
De novo sequencing	Genomic DNA, possibly with external data (from closely related species)	Piece together reads to assemble contigs, scaffolds, and (ideally) the whole genome
ChIP-Seq	DNA bound to protein, captured via antibody (Chromatin ImmunoPrecipitation)	Align reads to reference genome, identify peaks and motifs
Metagenomics	Entire RNA or DNA from a microbial/viral community	Alignment of reads to genomes, composition of the community and phylogenetic analysis
Transcriptomics	RNA (mRNA or total RNA)	Alignment of reads to gene, detection of splice junctions and transcript quantification

What takes the most amount of time?

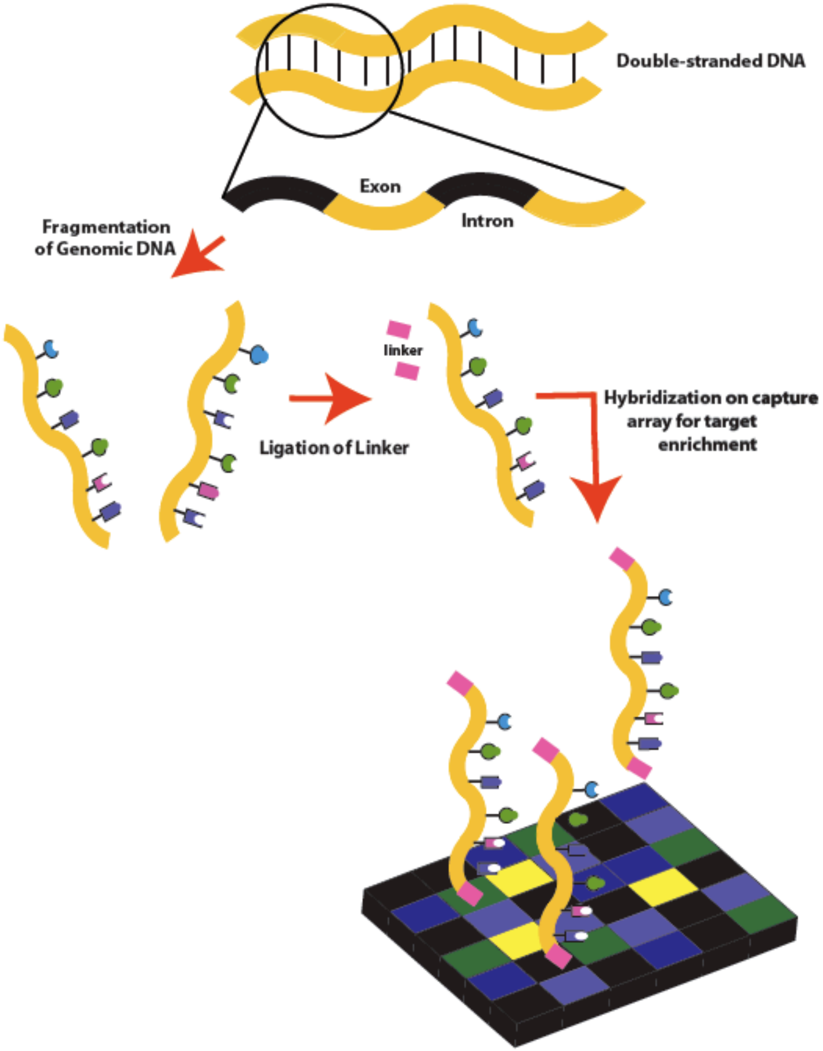
Average Percent of Time Spent on Next Generation Sequencing Workflow Steps



From **The Global Outlook for Next Generation Sequencing: Usage, Platform Drivers & Workflow** (2011) ©BioInformatics LLC

Survey of 267 scientists currently using Next Generation Sequencing in their work (2011)

Example of analysis workflow
Exome sequencing

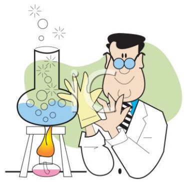


Double-stranded DNA
Exon
Intron
Fragmentation of Genomic DNA
linker
Ligation of Linker
Hybridization on capture array for target enrichment

Sequencing of genomic DNA enriched for ***coding*** regions:

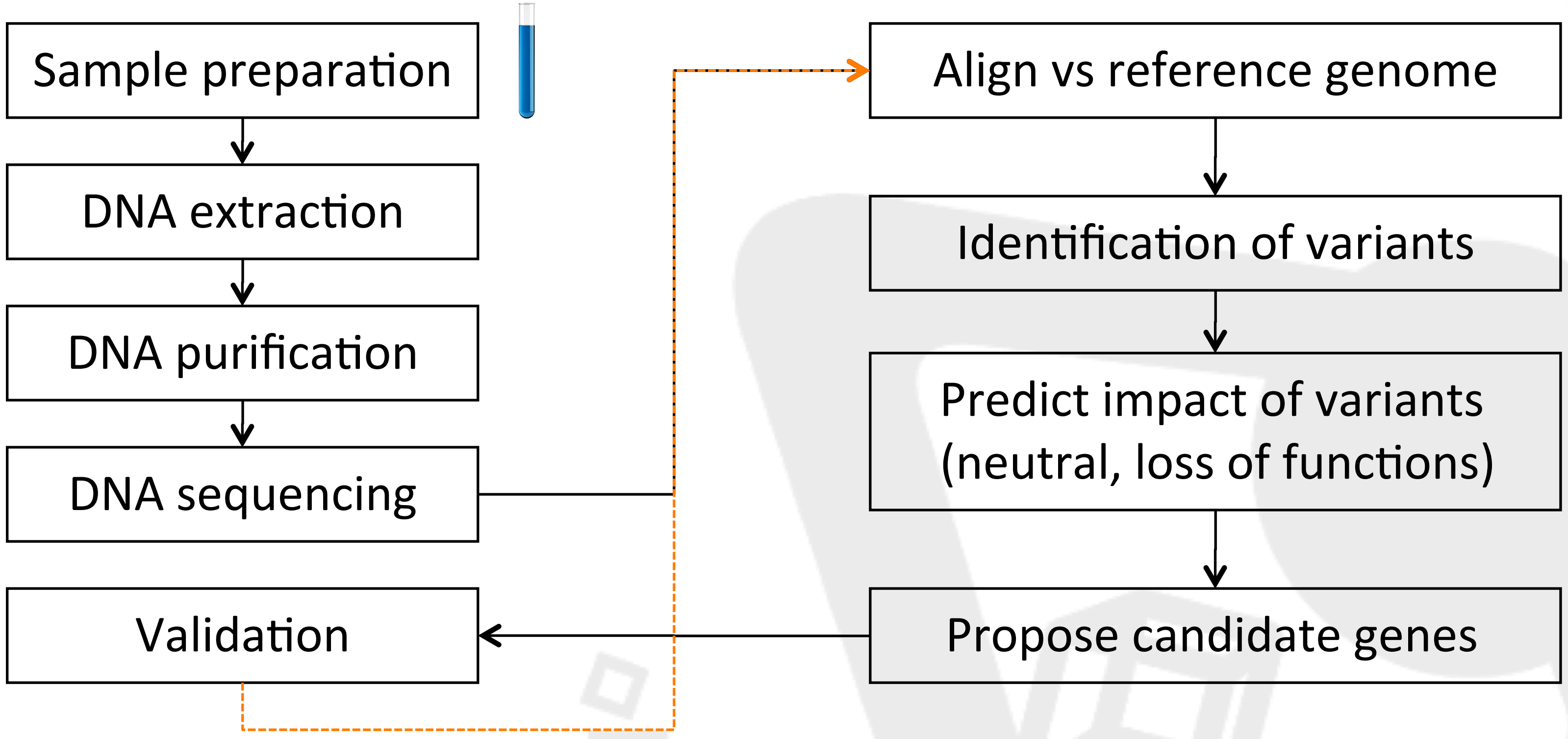
- clinical relevance (e.g. mendelian disorders)
- commercial applications (23andMe)
- technique used by increasing number of hospitals to investigate un-responsive cases

Example of analysis workflow
Exome sequencing



Wet-lab

Data analysis



Sample preparation

DNA extraction

DNA purification

DNA sequencing

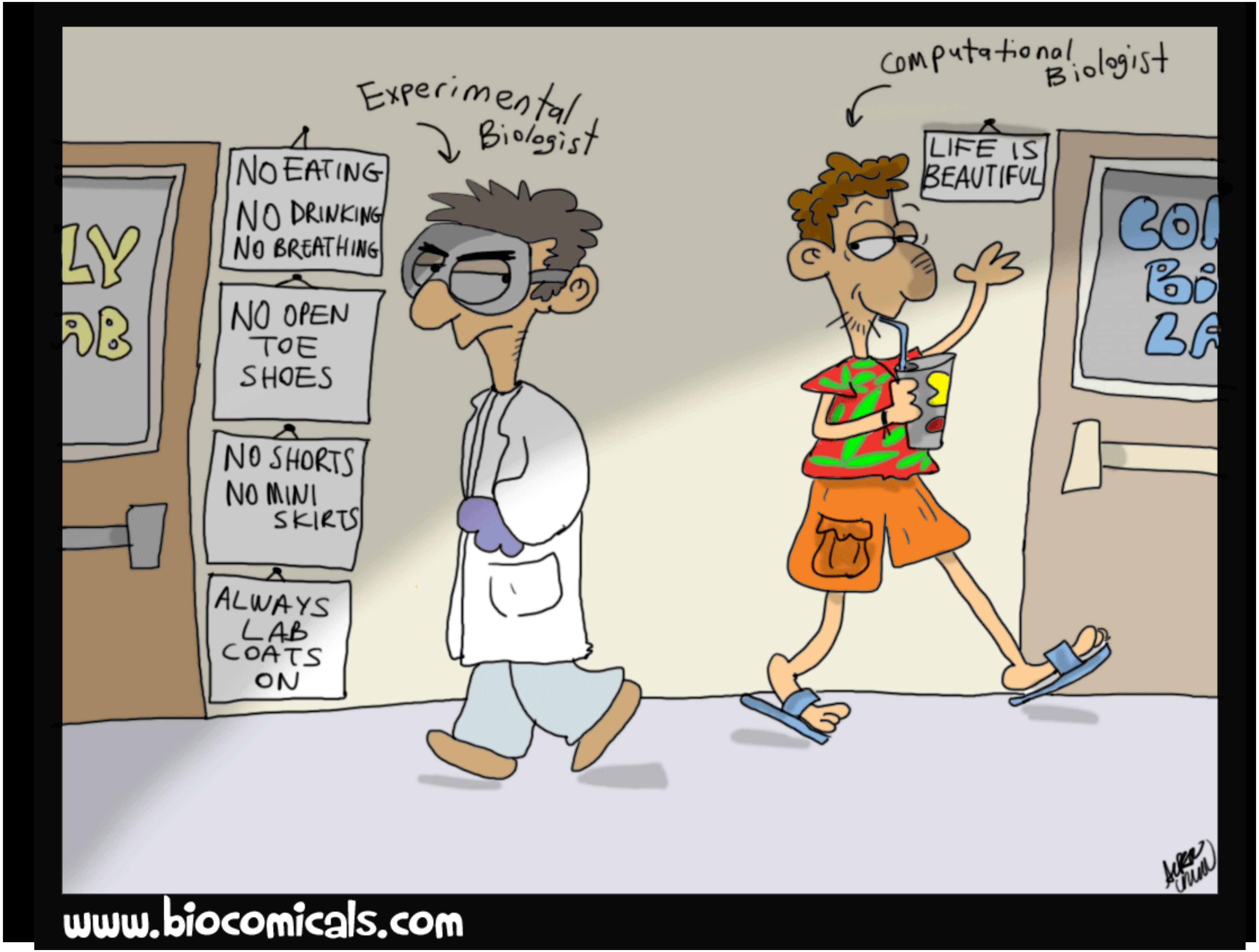
Validation

Align vs reference genome

Identification of variants

Predict impact of variants (neutral, loss of functions)

Propose candidate genes



www.biocomicals.com

4 simple reasons for learning bioinformatics

- Acquire a new perspective about the biological questions involved in my research lines
- Better understand published work and knowledge of additional analysis tools
- Thinking in statistical terms
- Acquire the skills to begin a new line of research

Job offer #1 - Description

xxx Institute is seeking a **Bioinformatician**. The candidate will work on analysis pipelines associated to NGS data (exome, whole-genome, RNA-Seq):

- To run existing analysis pipelines and perform QC and data analyses
- To participate in problem-solving discussions
- To apply bioinformatics solutions for the analysis of complex genomic datasets

Job offer #1 - Qualifications

- BSc in bioinformatics, mathematics, computer sciences, statistics or **molecular biology**
- Demonstrated **computer skills** (bash, Perl, Python, C++, Java) are required
- Previous experience in processing and analyzing NGS data, genome annotation or in developing sequence analysis pipelines is a major asset

Job offer #2 - Description

The xxx Institute is seeking a **Junior Bioinformatician**. The candidate will be working closely with senior programmers, statisticians on projects that can include:

- Optimization and parallelization of algorithms for the analysis of genome-sequencing data (mRNA and DNA microarrays, RNA-seq, whole genome-sequencing)
- Testing, maintenance and extension of existing sequence -analysis pipelines
- Automation of routine programming and data-analysis tasks
- Development of analysis-to-database interfaces to automate and optimize data aggregation.

Job offer #2 - Qualifications

- B.Sc. or equivalent education in computational biology, engineering, mathematics, computer sciences or **molecular biology**
- Strong background with **Unix/Linux** tools
- Experience in **software** and **database programming** (Perl, SQL)
- Statistical background and experience with **R/Matlab**
- Exposure to biological sequence algorithmic and analysis tools is a major asset
- Knowledge of molecular and/or cancer biology beneficial

How to become a bioinformatician

- Knowledge of **programming language**

How to become a bioinformatician

- Knowledge of **programming language**
- Good understanding of **statistics**

How to become a bioinformatician

- Knowledge of **programming language**
- Good understanding of **statistics**
- Knowledge of underlying **biology**

Who needs a "bioinformatician"?

- Next Generation Sequencing facilities
- Pharmaceutical companies
- -omics companies
- Research institutes
- Clinics/Hospitals

Readings

- What it takes to be a bioinformatician
  http://www.nygenome.org/blog/what-it-takes-be-bioinformatician
- How not to be a bioinformatician
  http://www.ncbi.nlm.nih.gov/pubmed/22640778

Training

- Bioinformatics Master @ UniCa
- Bioinformatics Training Network
  http://www.biotnet.org/
- EMBL-EBI (UK)
  http://www.ebi.ac.uk/training/
- Cambridge (UK)
  http://www.bio.cam.ac.uk/training/bioinformatics.html

CRS4 Scientific sectors

Biomedicine

- Advanced Genomics
- Bioinformatics
- Bioengineering
- Databases, Support and Services

Data Fusion

- Healthcare Flows
- Distributed Computing
- Visual Computing

Energy and Environment

- Clean Combustion technologies
- Geophysical Imaging
- Environmental Sciences
- Renewable Energy

Information Society

- Digital Media Applications
- Location and Sensor Based Services
- Natural Interaction Technologies

BIOMEDICINE

DATA FUSION

ENERGY and ENVIRONMENT

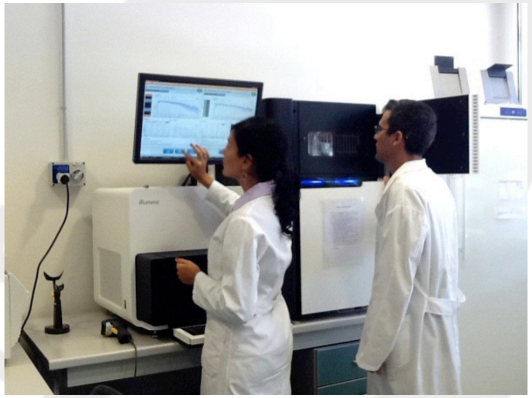
INFORMATION SOCIETY

CRS4 Sequencing Platform

The sequencing platform offers a wide range of advanced genomic characterization techniques:

- High Throughput (12 TBytes of raw sequencing data every ten days)
- Last generation Illumina technology
- Microarray Affymetrix plaftorm

Sequencing applied to complex diseases especially relevant to Sardinia population



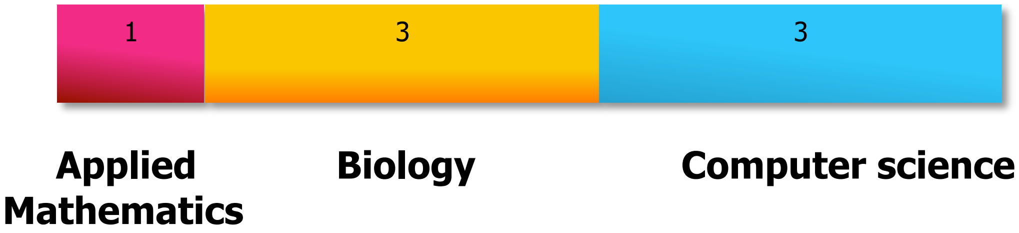
CRS4 Computing Center

CRS4 offers support to the scientific community through the High Performance Computing centre and its applications

- 44 TeraFlops of computational power
- 5 Petabytes of disk space
- 1 GBps connection

Bioinformatics Laboratory

Interdisciplinary research laboratory focused on computational biology



Operational since 2006

Our partners

- Hospitals / IRCCS
- Research centers (e.g. Porto Conte Ricerche, Alghero)
- CNR
- ISS

Staff with experience on both sides of the gap!

- Relevant biology background plus years of bioinformatics exposure

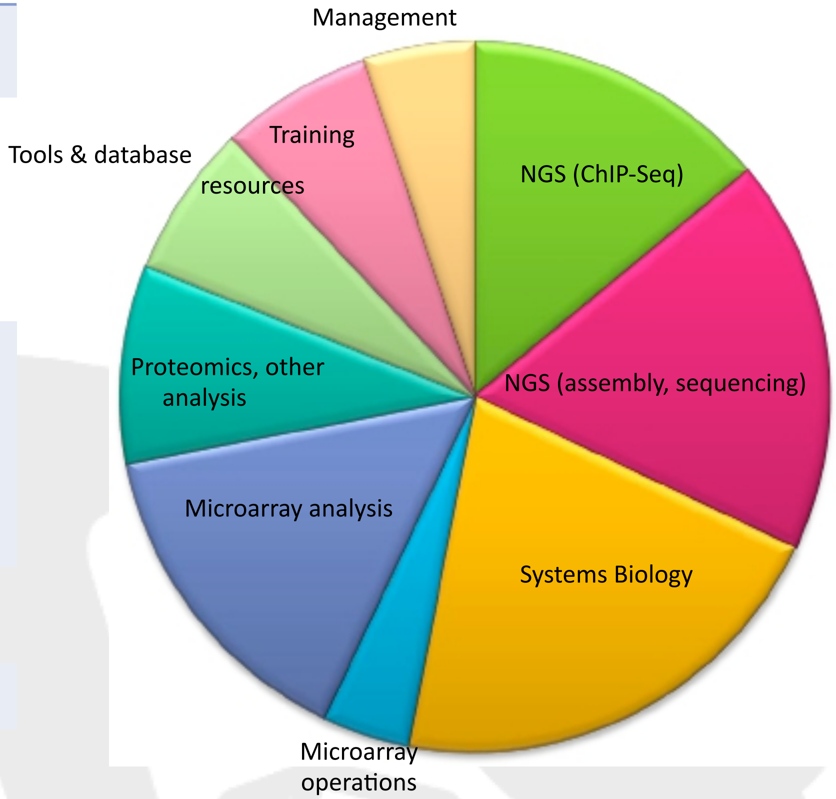
Key strengths include

- Wide expertise in complementary fields
- Multi-disciplinary competence
- Strong international connections

CRS4 Bioinformatics

Our main expertise

Experimental Design	Support in experimental design
Next gen sequencing	Algorithms Data management, processing, QC Analysis pipeline development Analysis projects – ChIP-seq, variation
Microarrays	Operational support, data QC Analysis pipeline, statistical analysis, data integration Analysis projects (expression, SNP/CNV) Illumina, Affymetrix , Agilent, custom arrays, …
Systems biology	Research projects
Other analysis	Motif enrichment, functional mapping
Analysis tools & data resources	Galaxy, Ensembl Open-source databases, tools
Training courses	NGS, microarrays, motif analysis, functional/pathway analysis

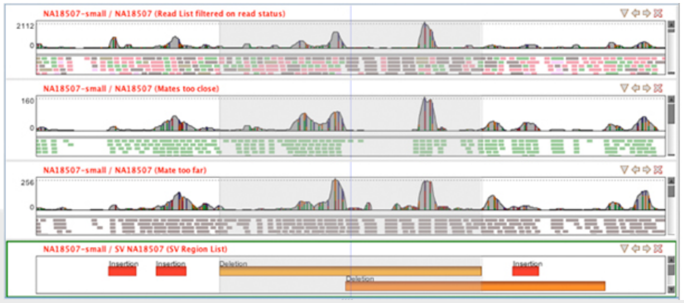


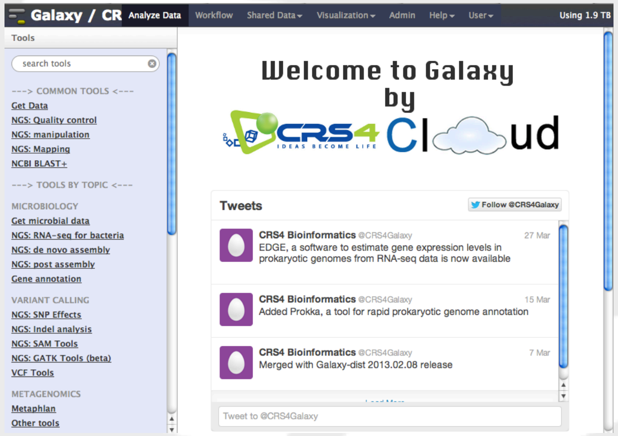
CRS4 Bioinformatics

Case studies

- **Microbiology**: development of computational pipeline for the assembly and annotation of bacterial genome from Next Generation Sequencing (NGS) data (partner: Porto Conte, IZS)

- **High-throughput analysis**: set-up the infrastructure for the analysis and interpretation of NGS datasets


NA18507-small / NA18507 (Read List filtered on read status)
2112
0
NA18507-small / NA18507 (Mates too close)
160
0
NA18507-small / NA18507 (Mate too far)
256
0
NA18507-small / SV NA18507 (SV Region List)
Insertion
Insertion
Deletion
Deletion
Insertion



Welcome to Galaxy by CRS4 Cloud

Tweets

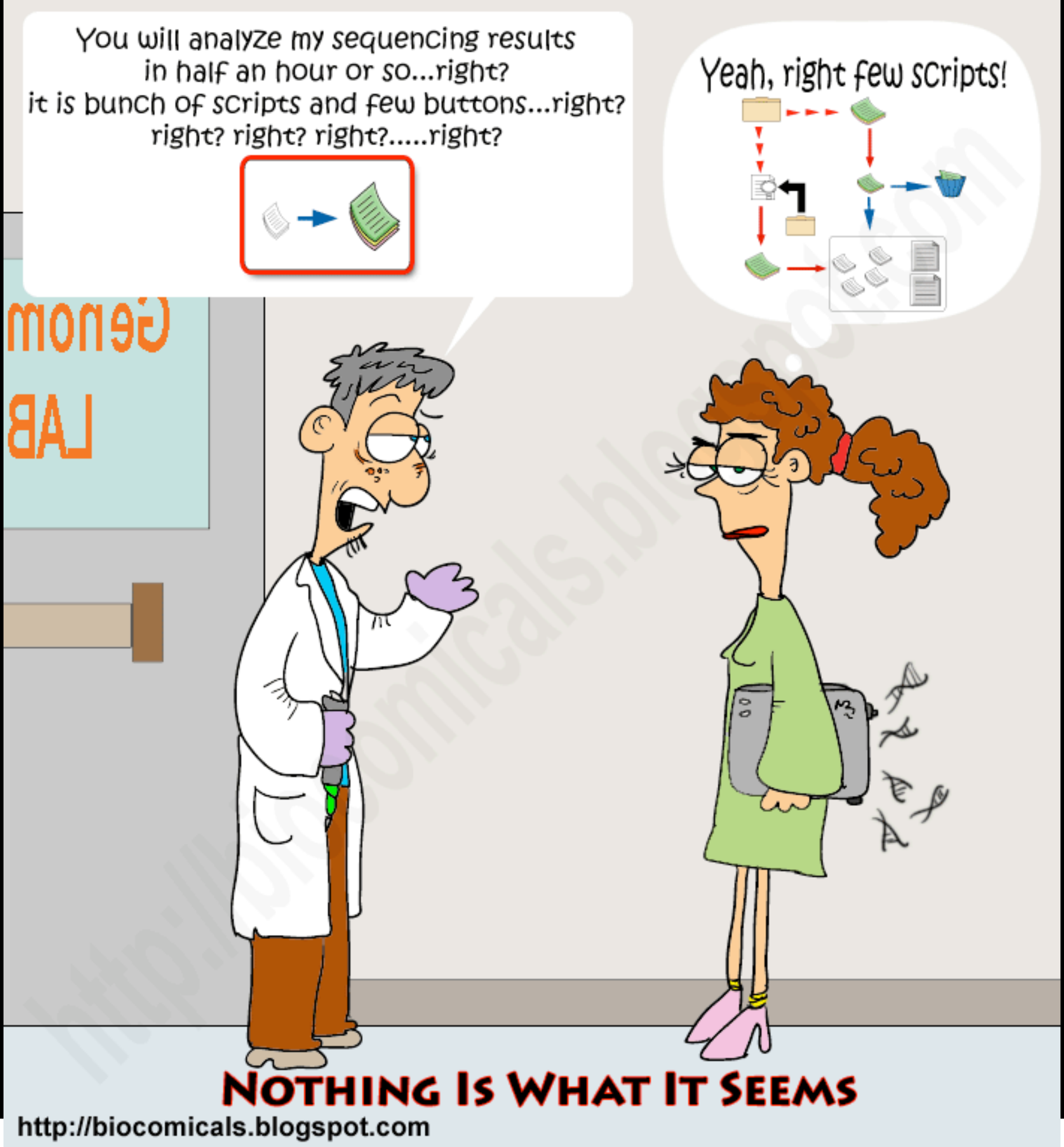
CRS4 Bioinformatics @CRS4Galaxy 27 Mar
EDGE, a software to estimate gene expression levels in prokaryotic genomes from RNA-seq data is now available

CRS4 Bioinformatics @CRS4Galaxy 15 Mar
Added Prokka, a tool for rapid prokaryotic genome annotation

CRS4 Bioinformatics @CRS4Galaxy 7 Mar
Merged with Galaxy-dist 2013.02.08 release

CENTRO DI RICERCA, SVILUPPO E STUDI SUPERIORI IN SARDEGNA
CENTER FOR ADVANCED STUDIES, RESEARCH AND DEVELOPMENT IN SARDINIA

Thank you for your attention


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!
Genom
LAB
NOTHING IS WHAT IT SEEMS
http://biocomicals.blogspot.com