Bioinformatics for biologists

Making sense of the Next Generation Sequencing data

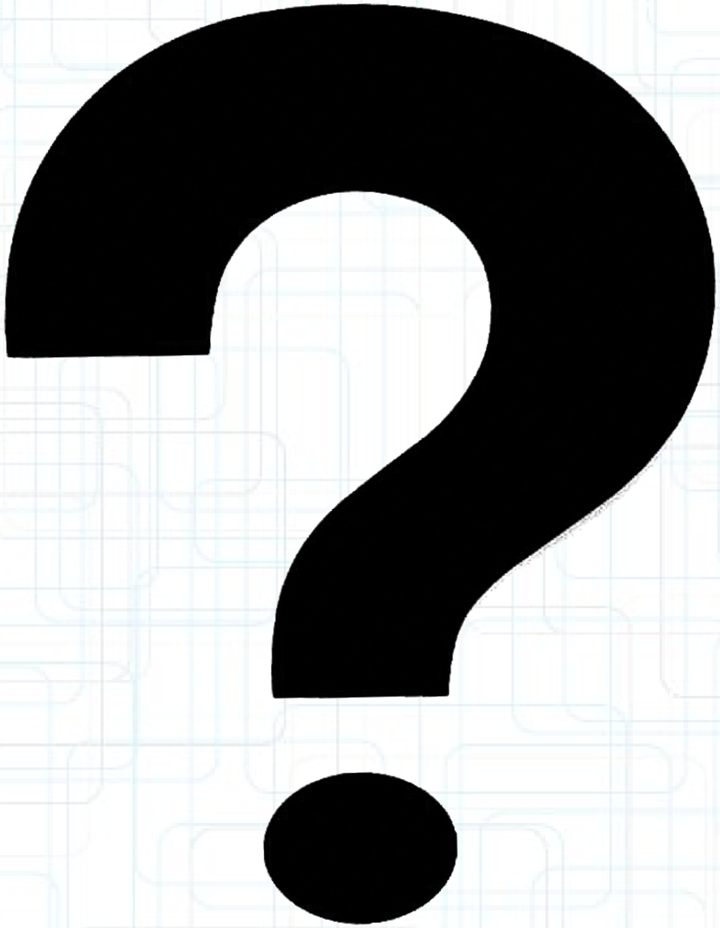
Konstantin Okonechnikov
Max Planck Institute For Infection Biology

Летняя школа биоинформатики
Москва, 2013

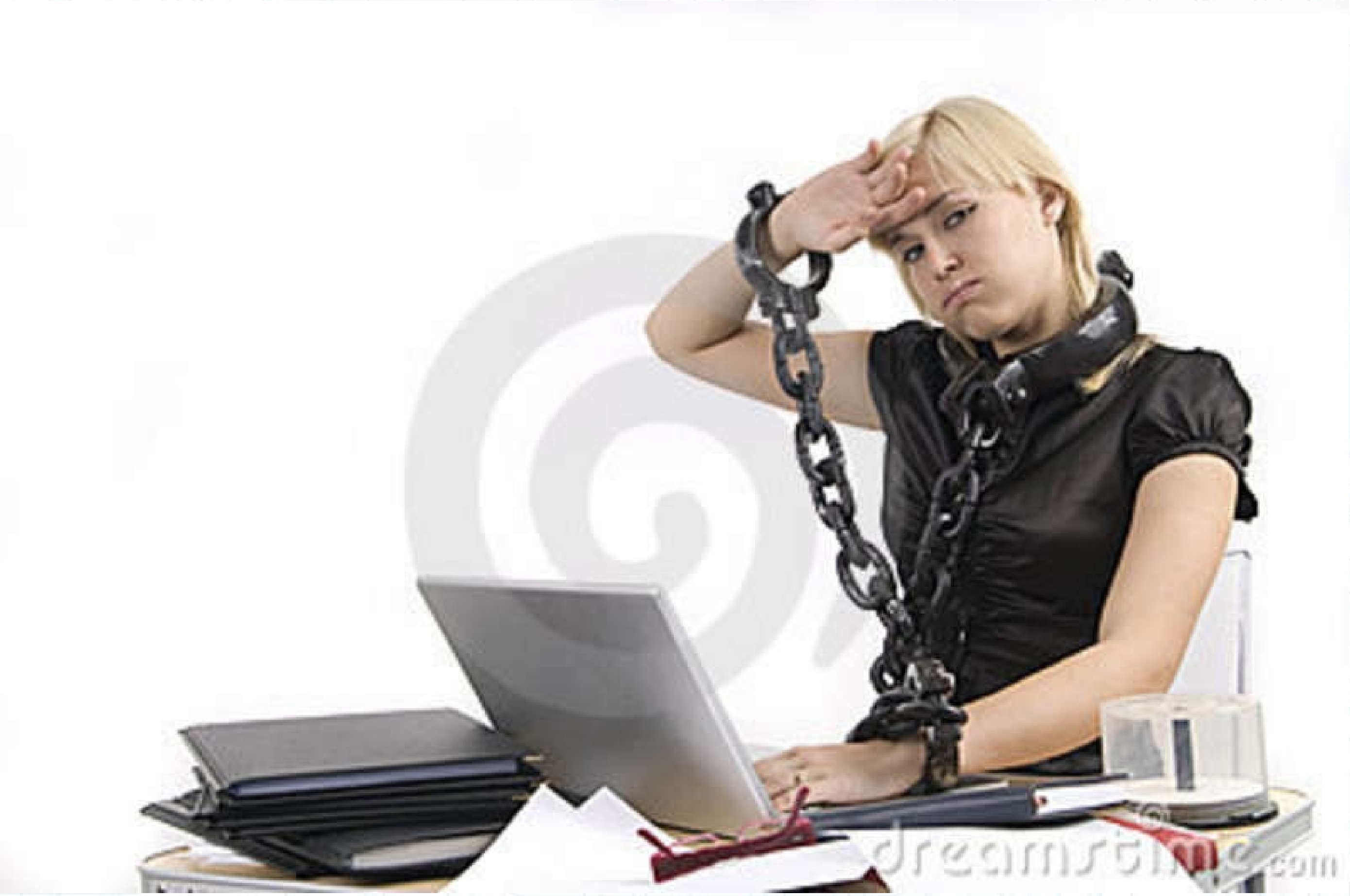
Objectives

- Learn about skills required for in-lab bioinformatics
- Learn about resources and tools in the area of NGS data analysis
- Start acquiring the bioinformatics skills by example of RNA-seq experiment analysis

Who are bioinformaticians?



Seen by the boss



Seen by biologists

Seen by other bioinformatician

Who are bioinformaticians?

- Bioinformatics: a union of scientific **disciplines** and a set of **skills**

Scientific and other disciplines

- What to learn:
  - Algorithms and programming
  - Statistics and data analysis
  - Biology
  - English
- How to learn:
  - Go to university
  - Read books and papers
  - MOE: Coursera, edX, etc

Skills: operating systems

- Windows
  - Mostly commercial software (CLCBio etc)
- Unix-based
  - Most popular bioinformatics tools are available only here
  - A lot of useful commands available by default

Skills: programming languages

- Be problem oriented
- Best general choice: R, Python
    - Available everywhere
    - Easy to learn: link to resources
    - A lot of libraries available (e.g. Bioconductor, Biopython)

Skills: understanding the data

- Data formats: flat-files, XML, etc.. (some links required)
- Data acquisition
  - Databases
  - Raw data
- Data manipulation:
  - Get subsets
  - Clean-up
  - Conversion

Skills: searching for answers

- Somebody already knows the answer
- Places to look:
    - Google
    - Biostar.org (Bioinformatics in general)
    - SeqAnsewrs (NGS)
    - <your favorite forum here>

Tools: data sources

- Big databases
  - NCBI (sequences, genes, proteins, ontologies etc...)
  - Ensembl (mostly genomes and annotations)
  -
- Learn APIs to access from code:
  - REST
  - Http requests...

Tools: algorithms

- Area specific
- To use best tools read papers:
  - Bioinformatics, Nature methods, etc
  - http://seqanswers.com/wiki/Software/list
- Example: Whole Genome Seq
  - Bwa, Samtools
- Example 2: RNA-seq tuxedo pipeline
- Example 3: < your example here >

Tools: visualization

- Genome browsers
  - UCSC
- Sequence and alignment viewer
  - IGV
  - Tablet
  - Unipro UGENE
- Other tools

Tools: workflow management

- Goal: better maintenance, visualization, reproducibility
- Big frameworks:
    - Galaxy
    - Taverna
    - Unipro UGENE
    - Knime

Demonstration: Unipro UGENE

- Website: http://ugene.unipro.ru/
- Demonstration and data:
- Pluses: looks sexy, free, available for many platforms
- Minues: bugs are possible

Demonstration: Galaxy

- Website:

  http://usegalaxy.org/
- Pluses: available from web, established community
- Minuses: difficult to debug, can not upload big data – local installation and setup is required

Demonstration: simple RNA-seq analysis

- Turorial from J. Goecks from Galaxy communty:
- Gene expression studies

https://main.g2.bx.psu.edu/u/jeremy/p/galaxy-rna-seq-analysis-exercise

Спасибо за внимание!