Bioinformatics (statistics/databases/big data)

Bruno Pot

Applied Maths

A BIOMERIEUX COMPANY

Veyrier-du-Lac
07/04/2016

What are the challenges in our domain?

- The quality of a *bioinformatics* analysis will depend on:

- **Quality**
  - of *sample preparation* (soil, stool, sputum, swaps, blood …)
  - of *wet lab protocols* (DNA preparation, library preparation)

    → as standardized as possible
  - sequencing *equipment* (quality, read length, sequencing depth,…)

- Can *bioinformatics* help you to judge the robustness and reliability of your sequencing preparations?

DNA extraction

Library preparation

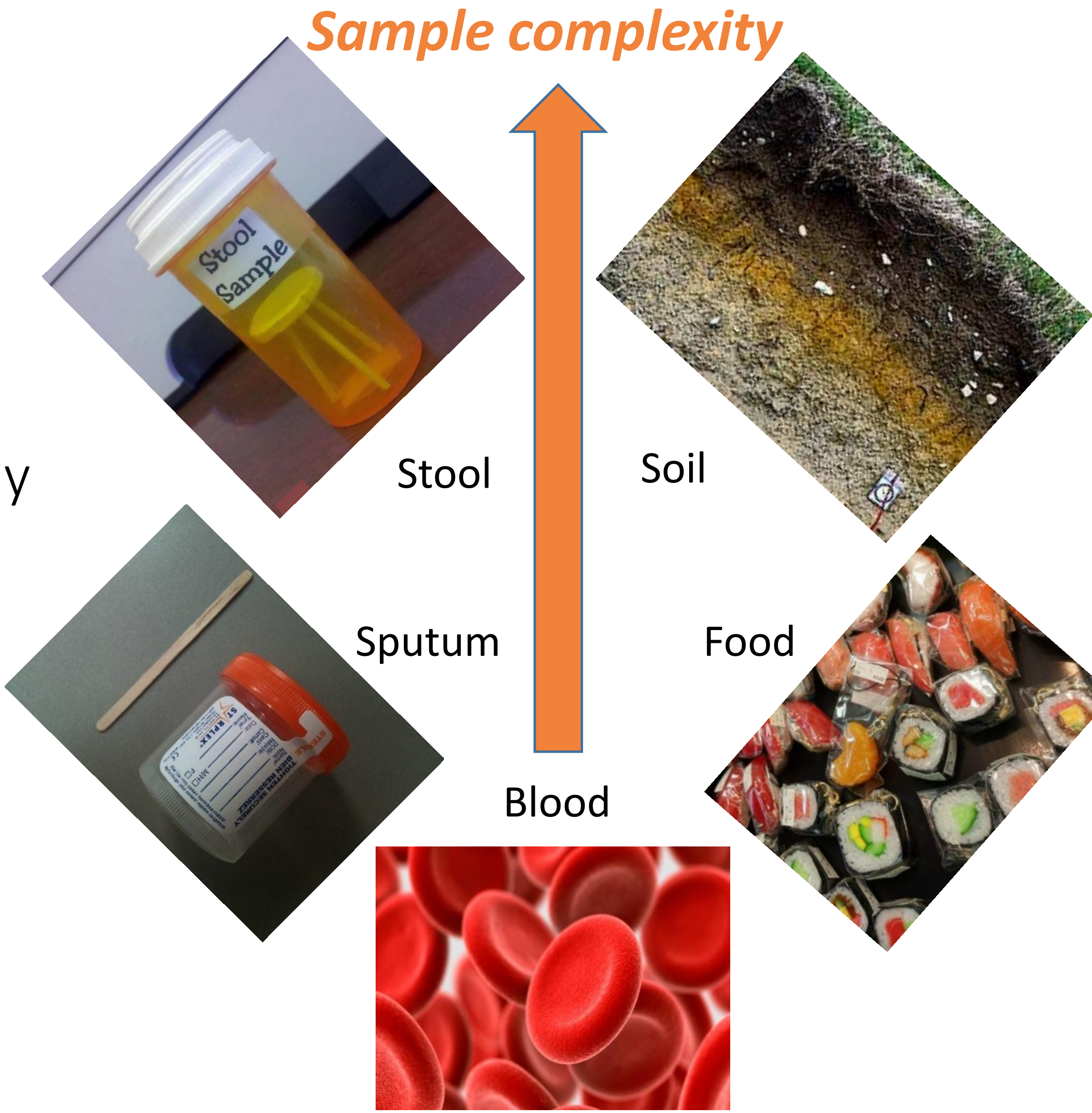
Sequencing depth

What are the challenges in our domain?

- *Bioinformatics* analysis will depend on the sample origin (read: **complexity**)

- This will depend on the availability of sufficently '**Complete**' and '**Reliable**' reference **databases** for proper identification of (all) (relevant) µorganisms and functionalities, and

- require the necessary **validation** processes (inter- and intra run controls; artificial mixes)



High-throughput data processing !

(diagnostics, quality control,…)

Mindshifts

Requires

- Extended computer power (Cloud, Cluster, ~~PC~~)
- Extended storage capacity (fast memory!)
- Faster data transfer

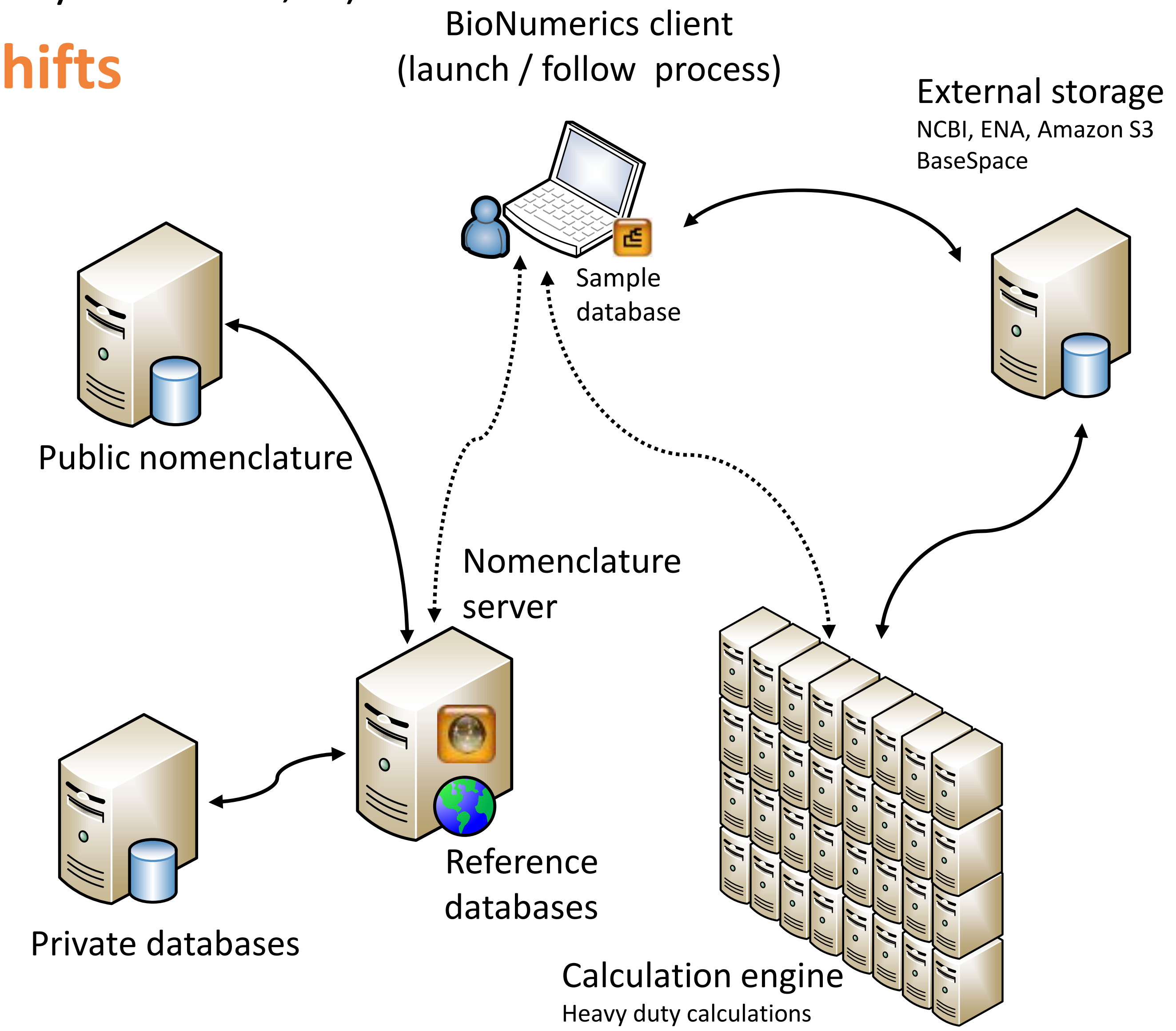
HARDware

- Advanced (fast) algorithms
- Read analysis strategy (assemble-identify/ identify)
- Advanced visualization tools

SOFTware

- Userfriendliness for non-specialized users
  - software installation level
  - software usage level
  - software maintenance level
  - user management and confidentiality management

BI&IT support



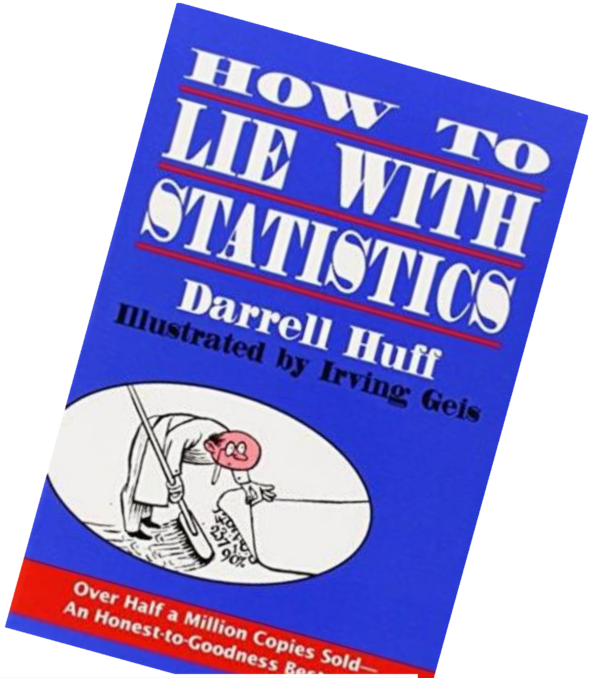
Statistics tell it &ll…

Read statistics (length, numbers, quality,..): QUALITY evaluation

- Filtering (singletons, chimaeras, barcoding…)
- Assembly statistics
- Matching statistics

Analytical statistics RELIABILITY evaluation

- Diversity (indices)
- Distribution / geography (plots)
- Timeframe analyses (plots)
- Cluster algorithms
- PCA and other visualization algorithms
- Correlation versus association versus cause
- (M)anova analyses
- Significance tests
- …


HOW TO LIE WITH STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bo

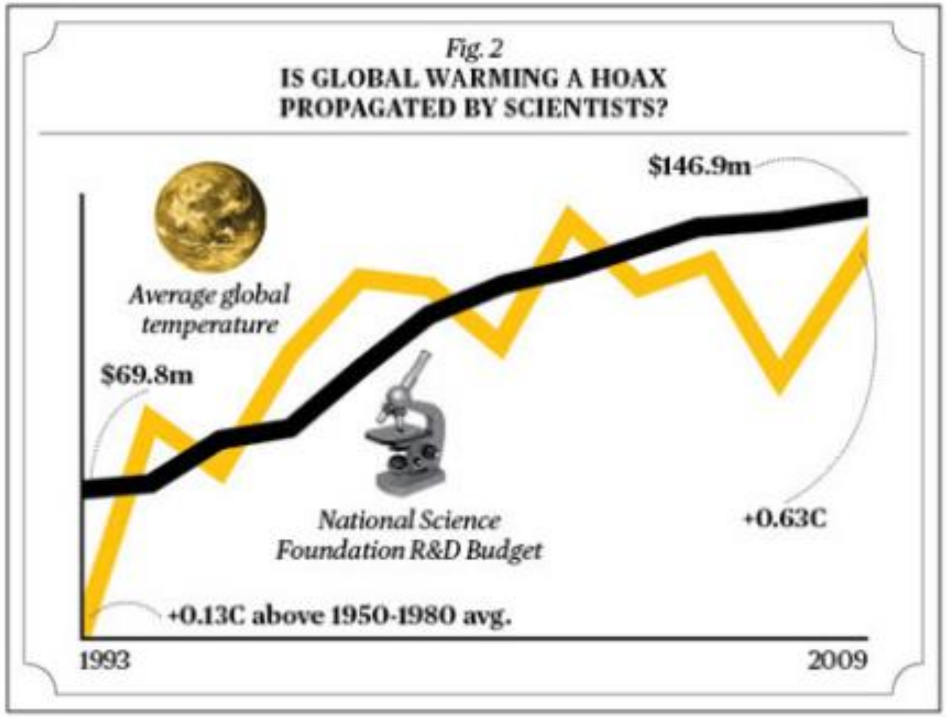


Fig. 2
IS GLOBAL WARMING A HOAX PROPAGATED BY SCIENTISTS?

Conclusions

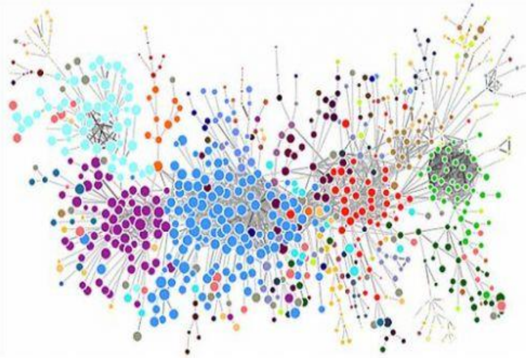
Reliability issues

- Different wet lab protocols will give different outcomes on the same sample (e.g. library preparation)
- Different pipelines will yield different outcomes on the same dataset

  → Artifical *in silico* data are difficult to reconstruct with existing pipelines (due to need for filtering, etc…)
- Different statistical settings will give different interpretations on the same datasets

Complexity issues

- Amplicon-based metagenomics approach verus full shot gun metagenomics
- Causalilty questions
- Interpretation of results in non-perfect conditions (geography, time analysis, different populations, climates, disease types, …)



Practical hurdles

- Amount of data (storage, analysis, visualization,…)
- Complexity of the bioinformatics analysis (several quality dependent steps)
- Manage complexity of the IT infrastructure required
- Lack of adapted databases for ‘new’ subjects