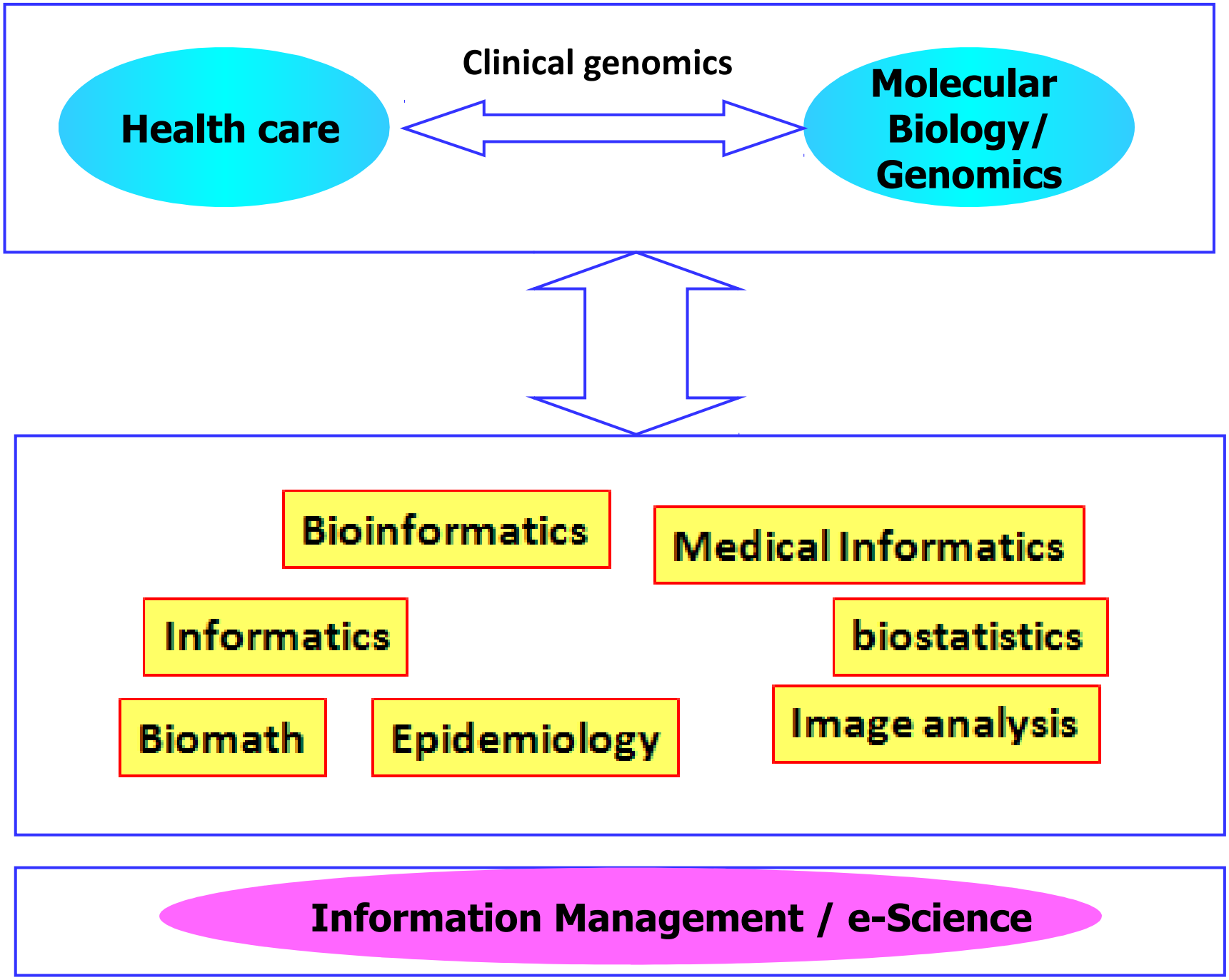
Introduction to bioinformatics

DNA technology course

Barbera van Schaik
Bioinformatics Laboratory
Academic Medical Centre (AMC)
b.d.vanschaik@amc.uva.nl


Clinical genomics
Health care
Molecular Biology/ Genomics
Bioinformatics
Medical Informatics
Informatics
biostatistics
Biomath
Epidemiology
Image analysis
Information Management / e-Science

Lost in translation (biology and informatics)

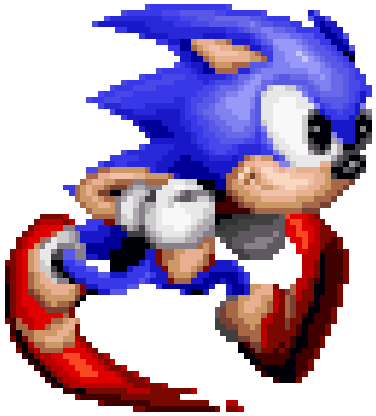
Inheritance

Protocol

NGS

Infrastructure

String



Sequence

Sonic hedgehog

Virus

Definitions of bioinformatics

Adapted definition according to Wikipedia

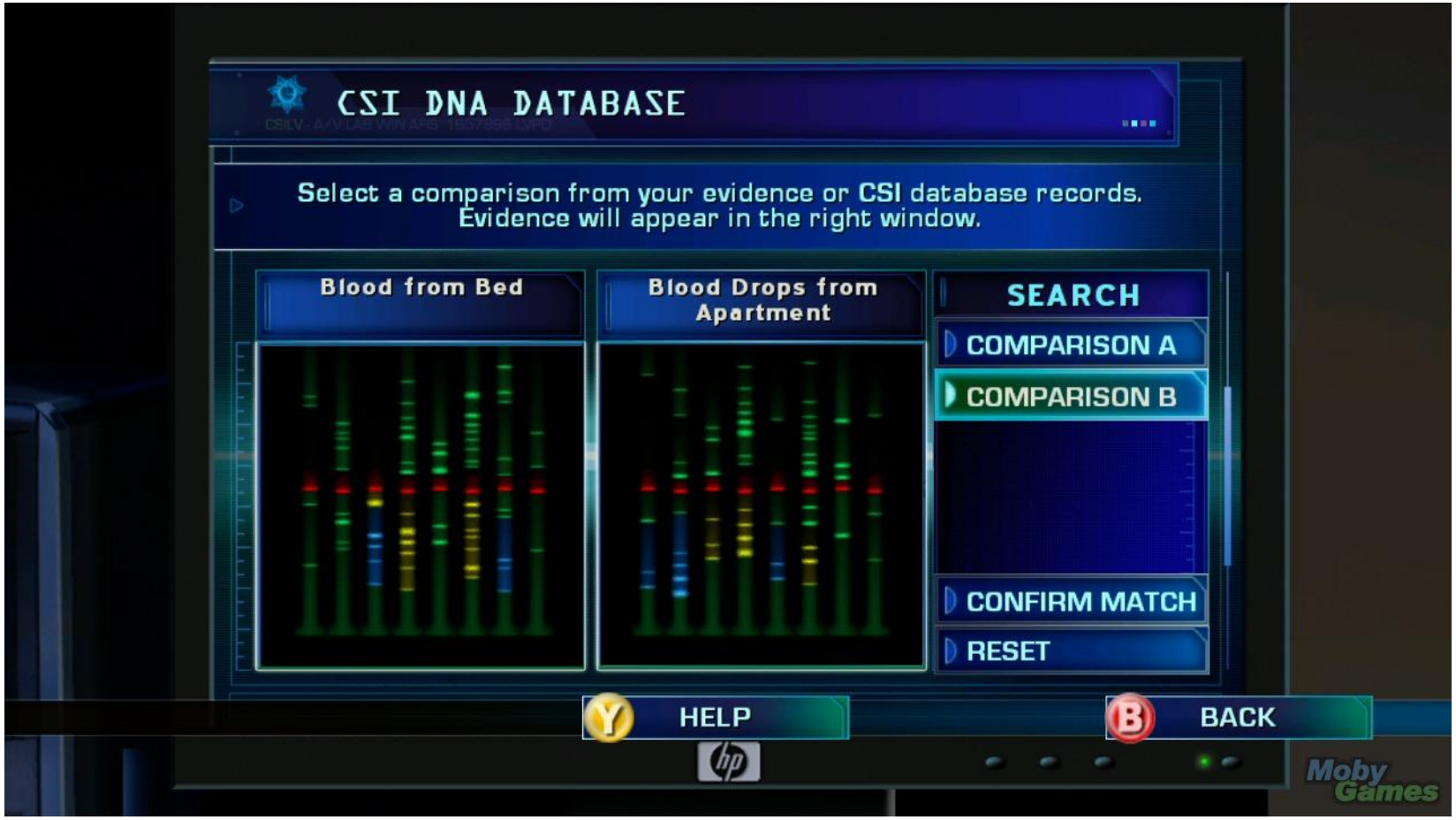
The application of information technology and statistics to the field of molecular biology.

The creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management, analysis and interpretation of biological data.

What is bioinformatics?

Extraction of biological knowledge from complex data

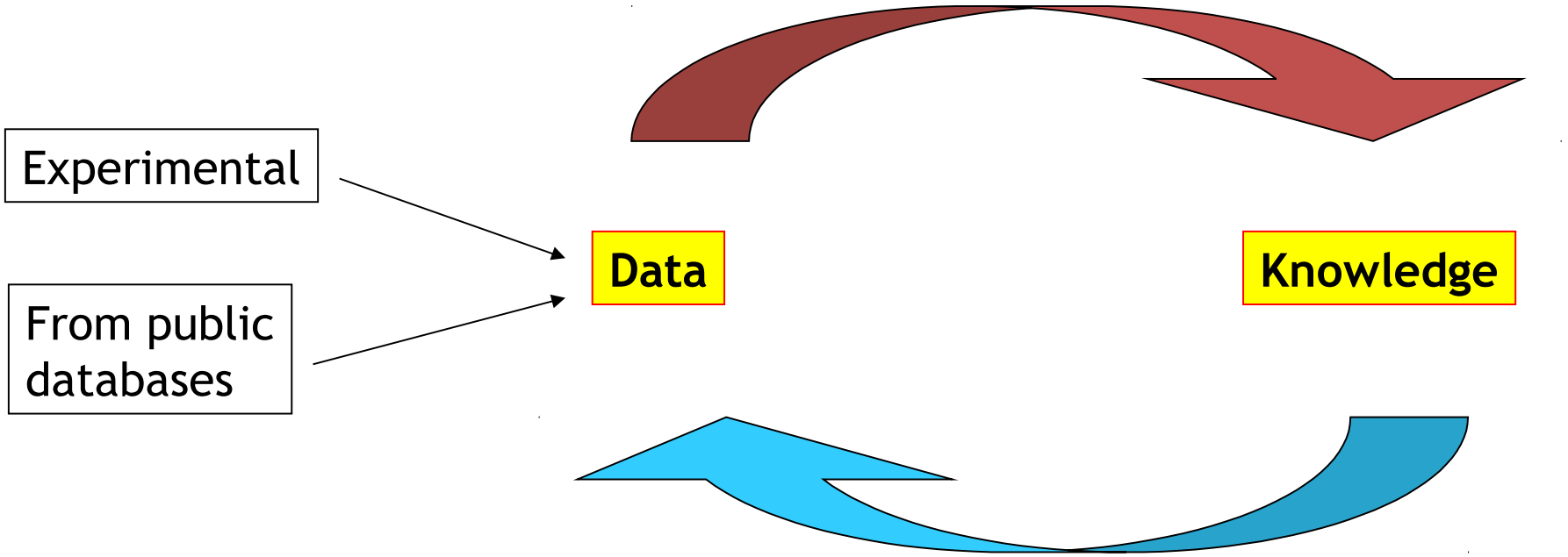
What is bioinformatics?



... one of the results *might* be a tool you can use

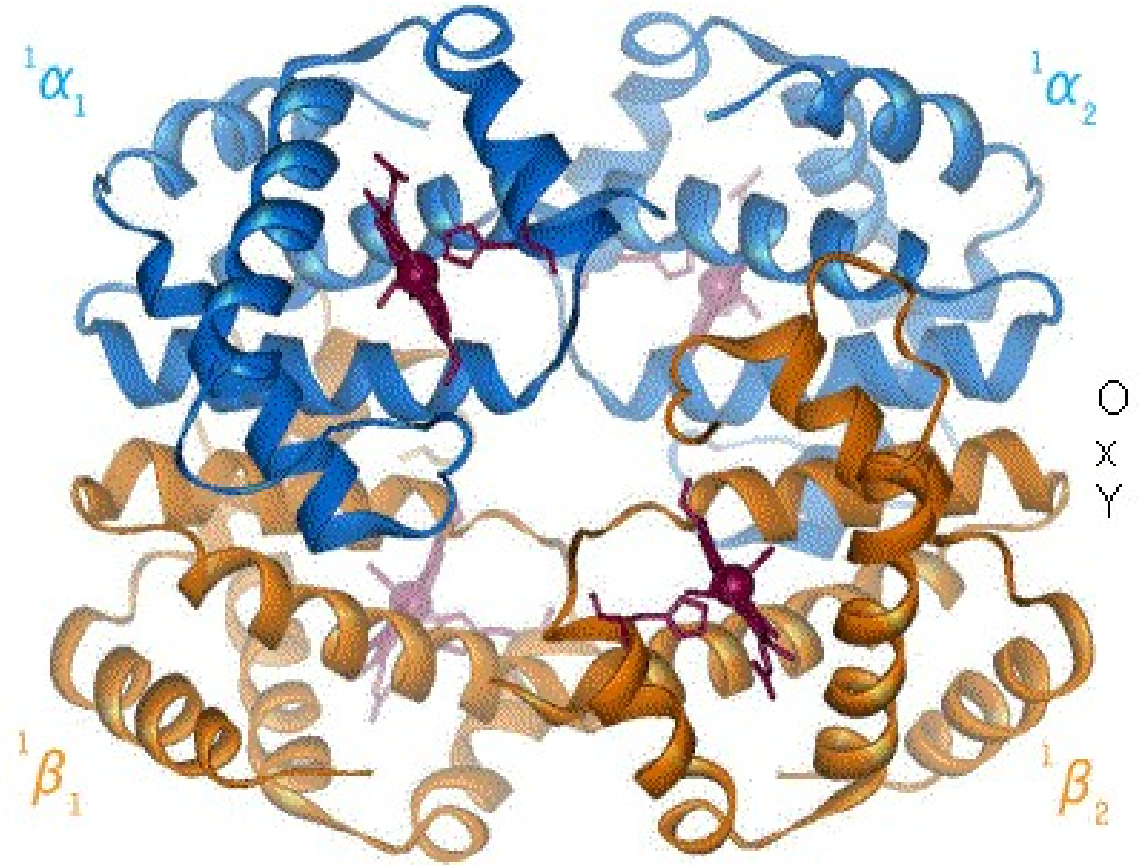
Bioinformatics

*convert data to knowledge
*generate new hypotheses



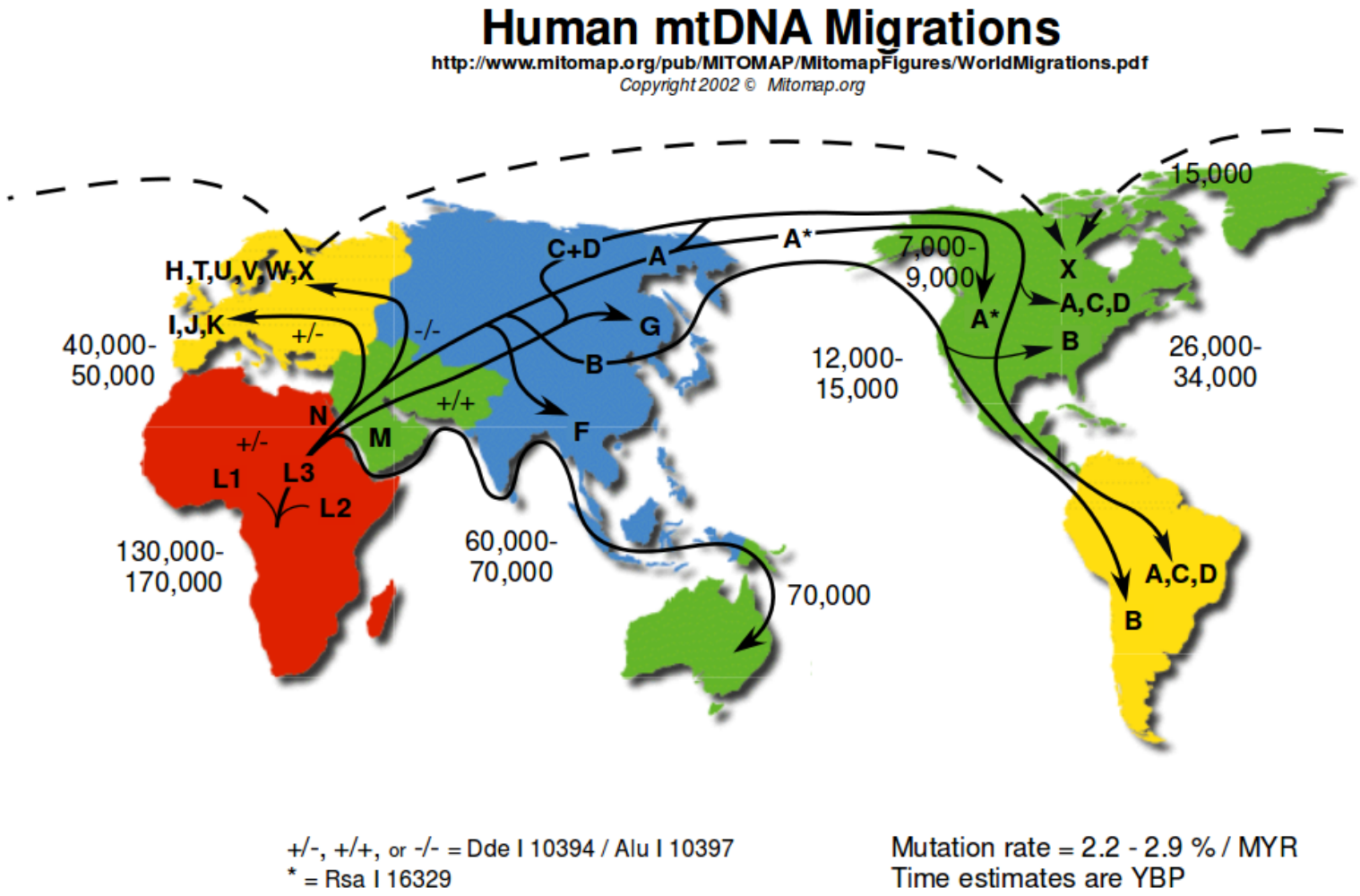
And information management.....

How does molecule A interact with protein B?

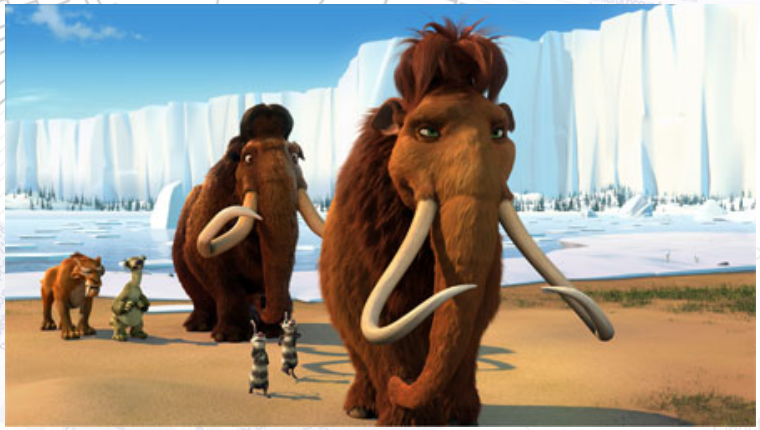


A schematic visual model of oxygen-binding process, showing all four monomersand hemes, and protein chains only as diagramatic coils, to facilitate visualization into the molecule. (http://en.wikipedia.org/wiki/Hemoglobin)

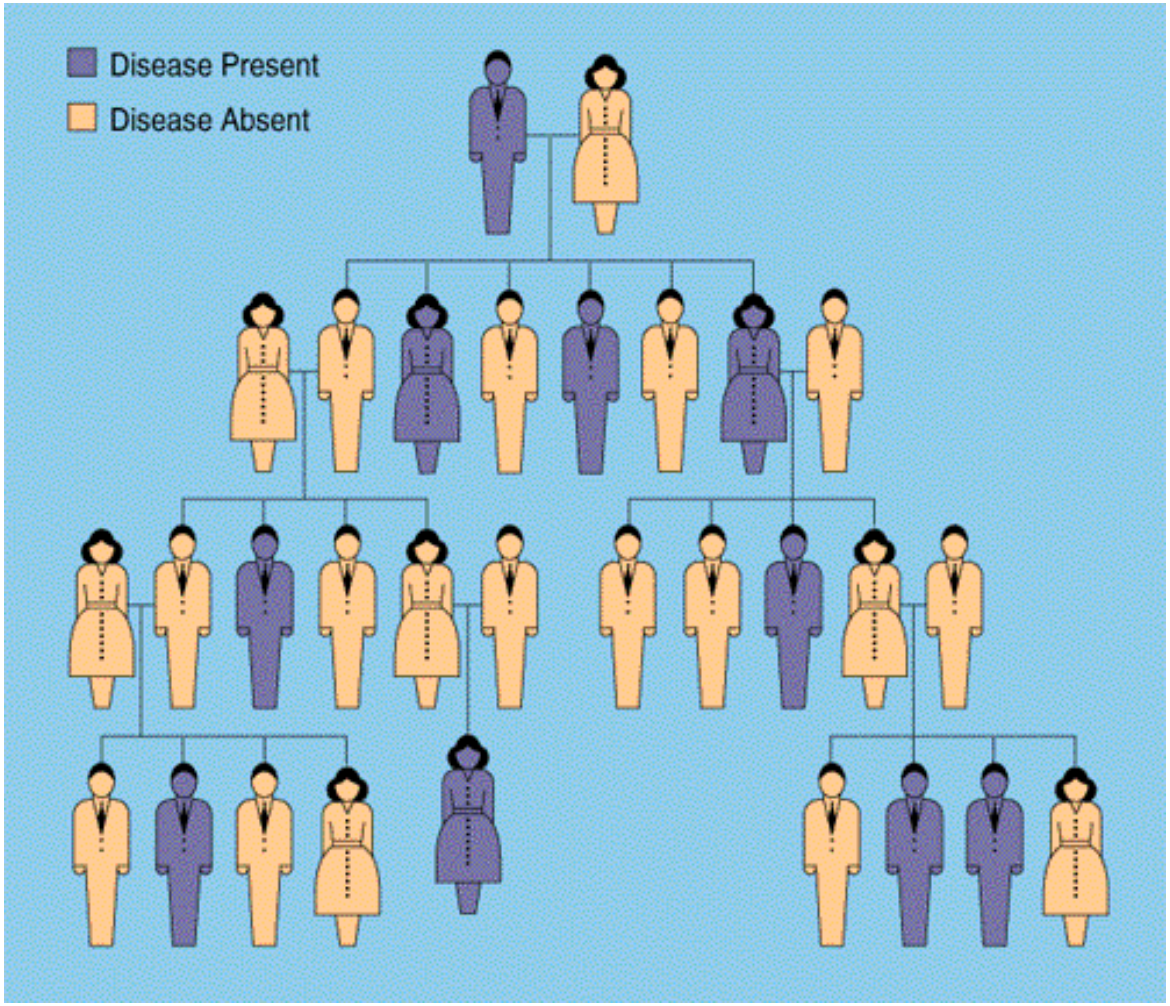
Study migration



Study evolution



Which gene(s) causes disease X?



Genome analysis

Bioimage informatics

Sequence analysis

Databases and ontologies

Data and text mining

Subjects in bioinformatics

Phylogenetics

Systems biology

Structural bioinformatics

Genetics and population analysis

Gene expression

http://www.oxfordjournals.org/our_journals/bioinformatics/for_authors/scope_guidelines.html

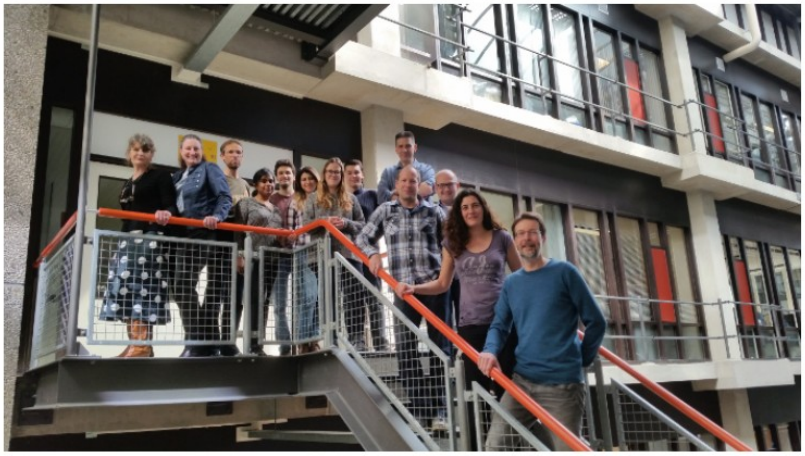
Bioinformatics
Groups and institutes

Bioinformatics Laboratory (AMC)
Medical Bioinformatics

- Part of the KEBB
- You are welcome if you need bioinformatics expertise
- www.bioinformaticslaboratory.nl

Bioinformatics Laboratory

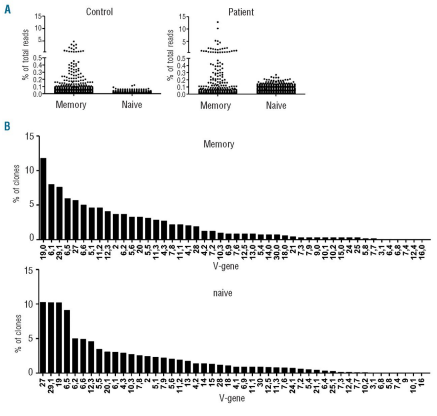
The Bioinformatics Laboratory was initiated in 1997 to advance biomedical research in the Academic Medical Center (AMC, Amsterdam) and is headed by Antoine van Kampen. The laboratory is part of the department of Clinical Epidemiology, Biostatistics and Bioinformatics (KEBB). The group members represent a broad range of (bio)informatics, biostatistics, and e-bioscience expertise and engage in research, education, and data analysis. Van Kampen is appointed part time at the Biosystems Data Analysis group of the Science Faculty to facilitate collaborations between the AMC, the BDA group, and other groups at the Science Faculty.



Focus

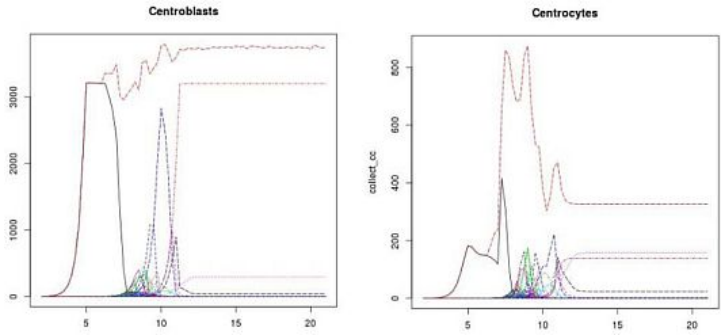
Systems genomics (top-down)

(Statistical) analysis of (multi-)omics data



Systems biology (bottom-up)

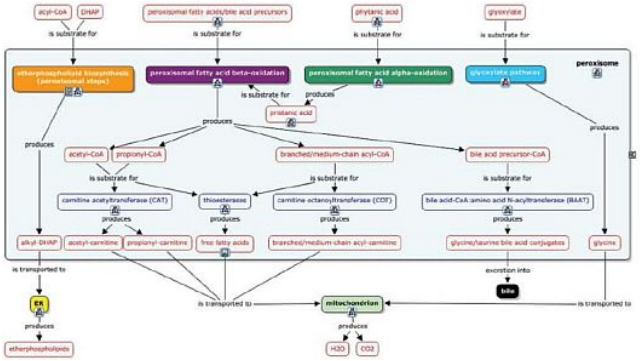
Mathematical modelling



Information management

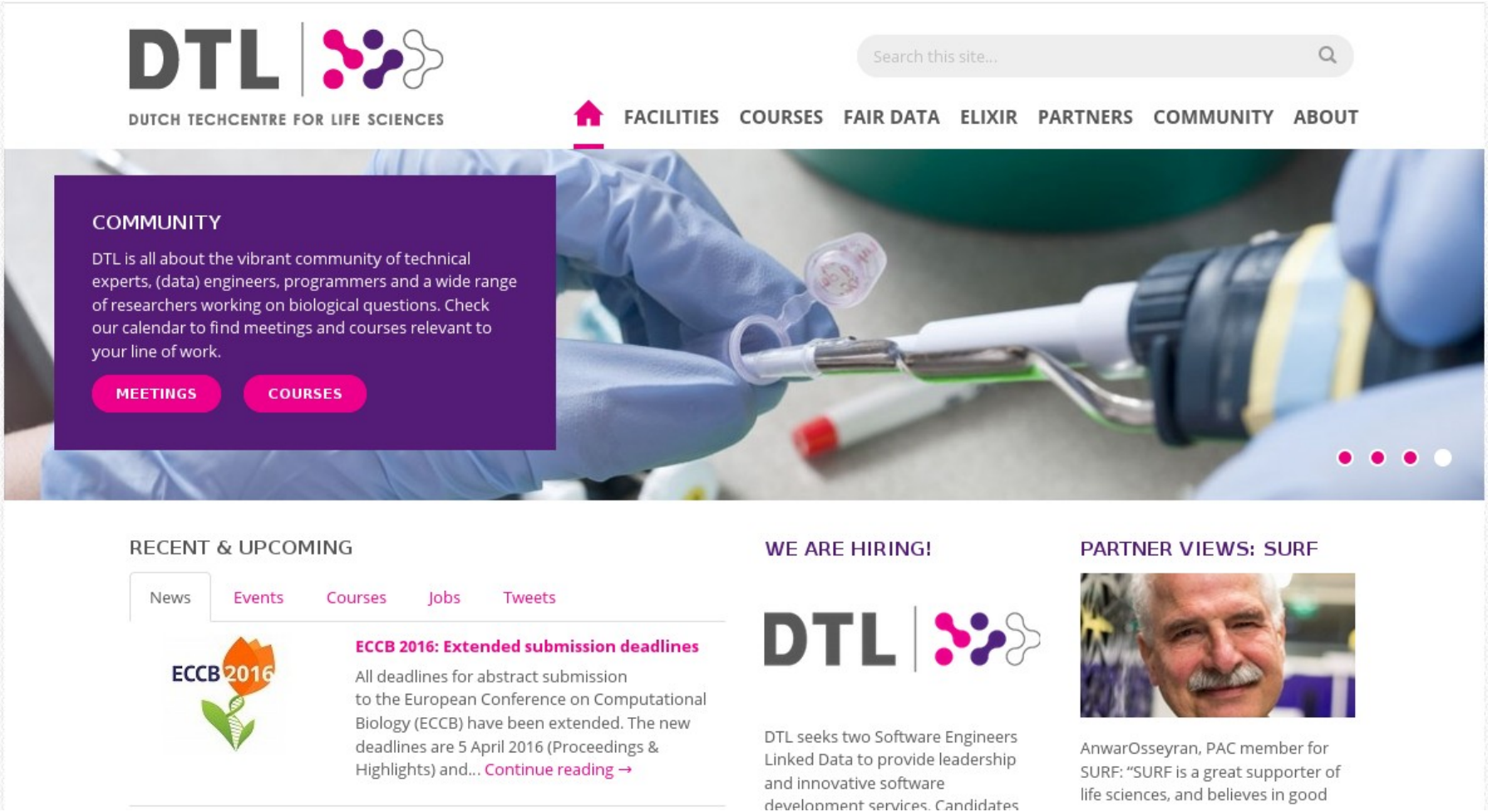
Knowledge bases

Science gateways



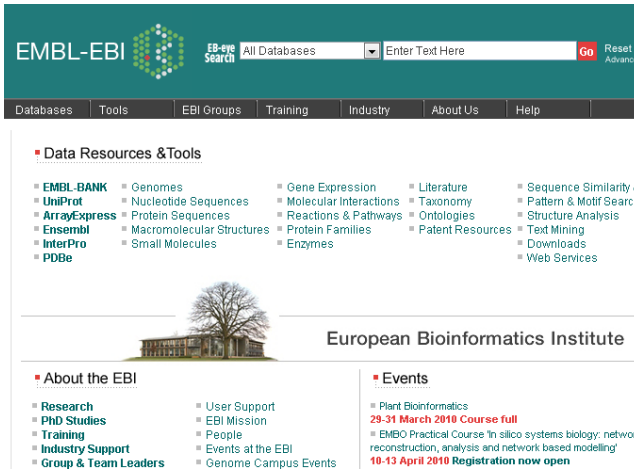
Dutch Techcentre for Life Sciences (DTL)

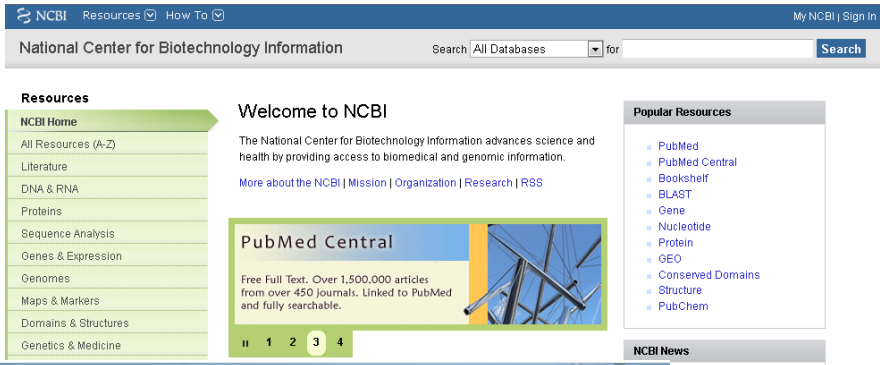
- Support and Education
- http://www.dtls.nl/

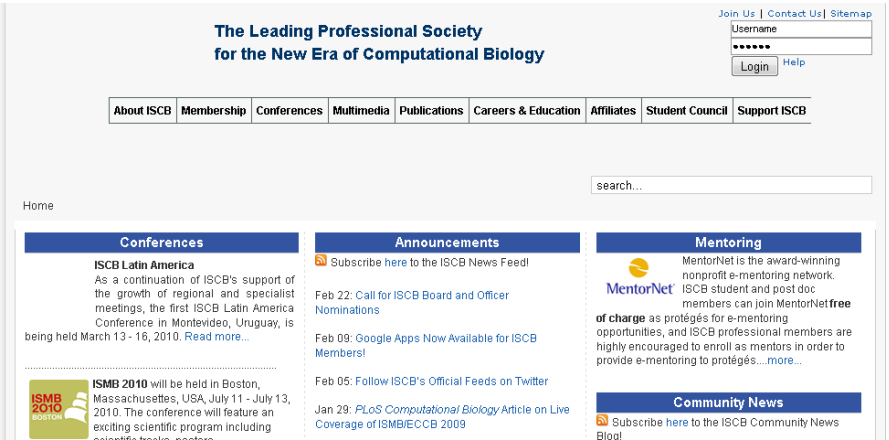


Other bioinformatics organisations

- European Bioinformatics Institute (EBI)
  - http://www.ebi.ac.uk/
- National Center for Biotechnology Information (NCBI)
  - http://www.ncbi.nlm.nih.gov/
- EMBnet
  - http://www.embnet.org/
- International Society for Computational Biology (ISCB)
  - http://www.iscb.org/







Bioinformatics

Extraction of biological knowledge from complex data

@AMC: genomics, metabolomics and proteomics data

What is genomics?

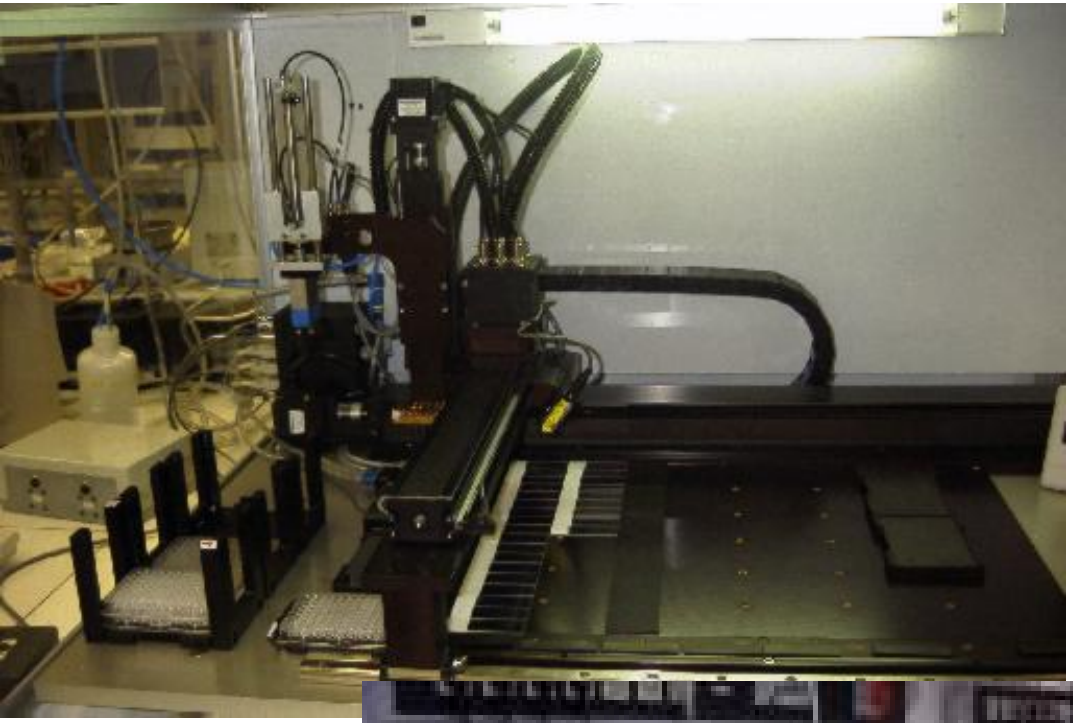
The application of **high-throughput** automated technologies to molecular biology.

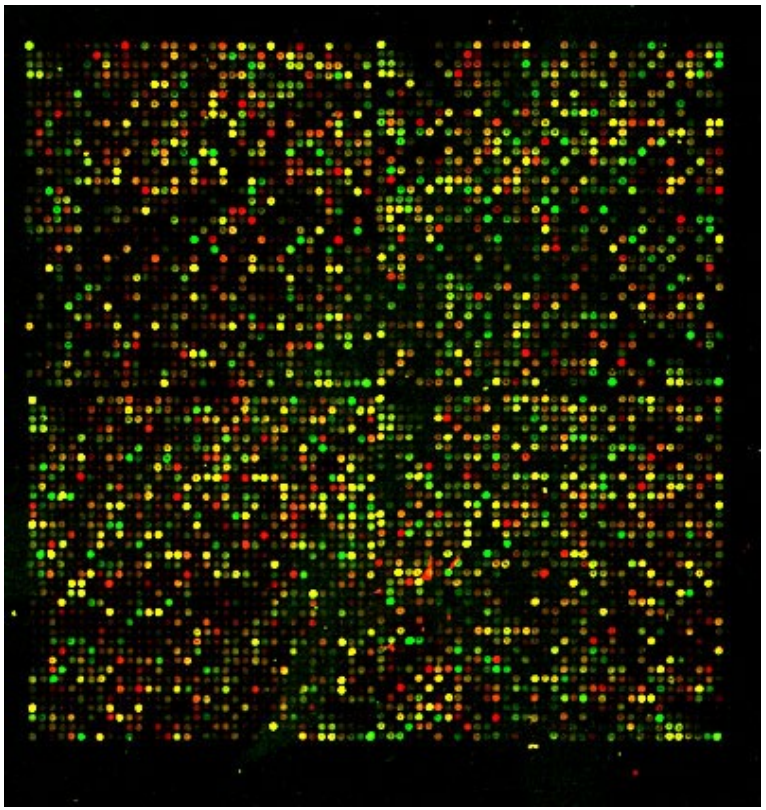
OR

The experimental study of complete genomes.

Sample storage

DNA microarrays





High throughput sequencing

Ion Proton

Run: 2-4 hrs
Data: 10 GB
Reads: 60-80 million
Length: 200bp

Illumina MiSeq

Run: 4-55 hours
Data: 15 GB
Reads: 25 million
Length: 2x300bp

Illumina HiSeq (VUmc)

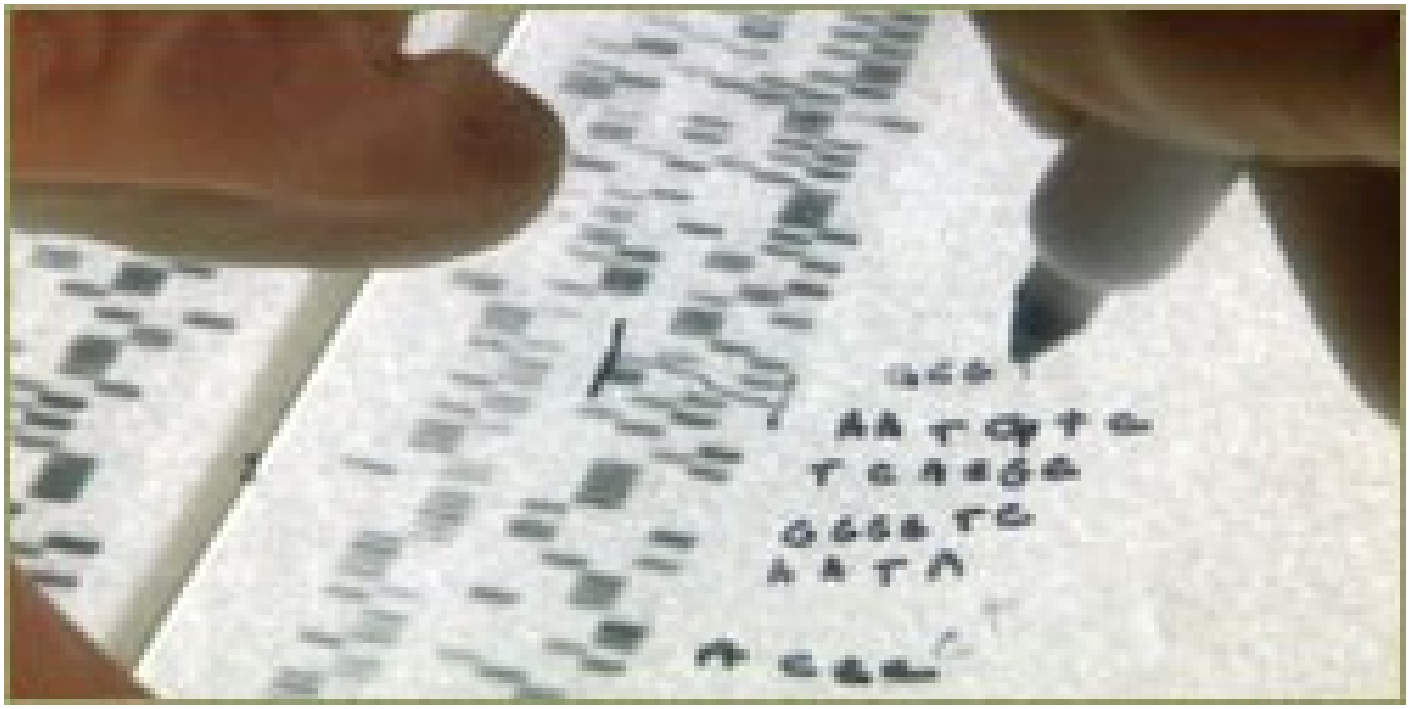
Run: 1-3.5 days
Data: 1500 GB
Reads: 5 billion
Length: 2x150bp

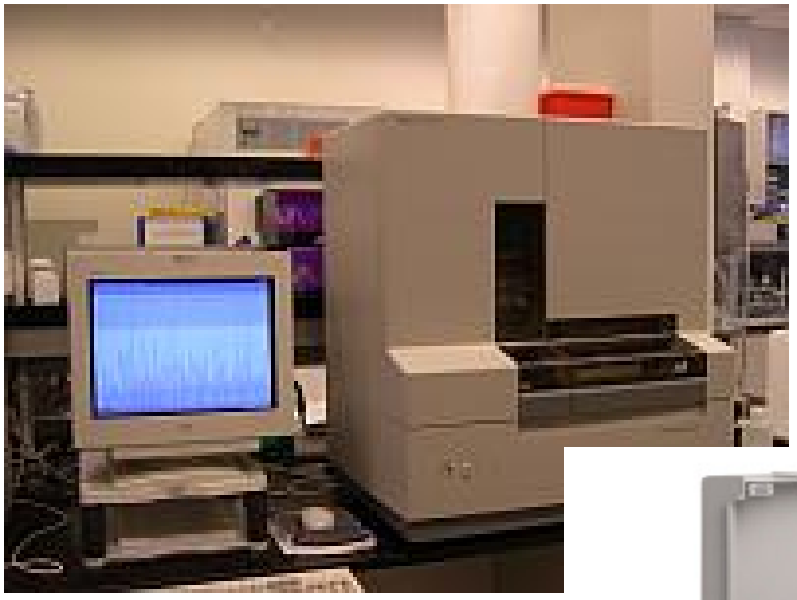
2005-now: Next generation sequencing
Millions to billions of sequences

DNA sequencing

Sanger sequencing

From DNA
to
Autoradiograms

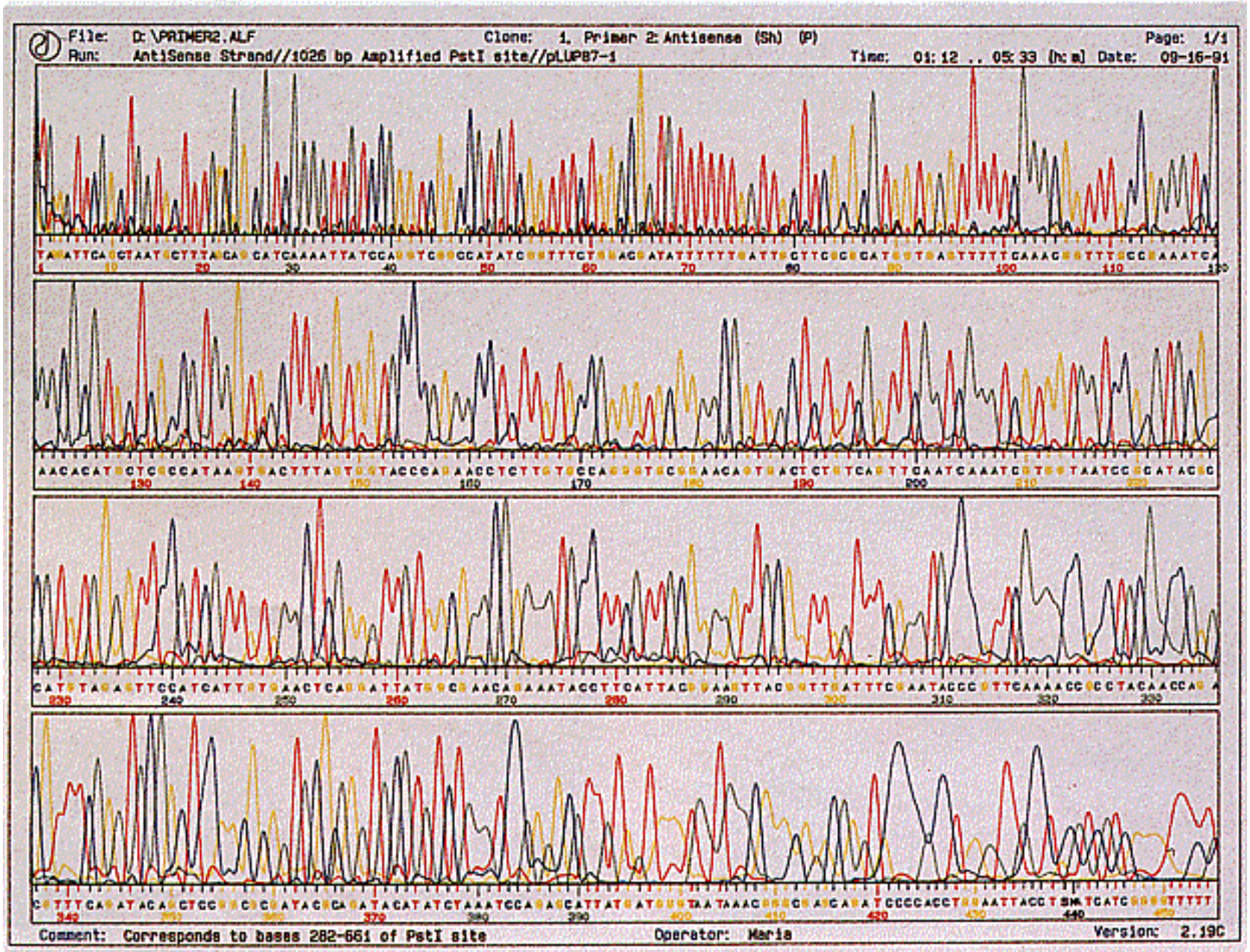




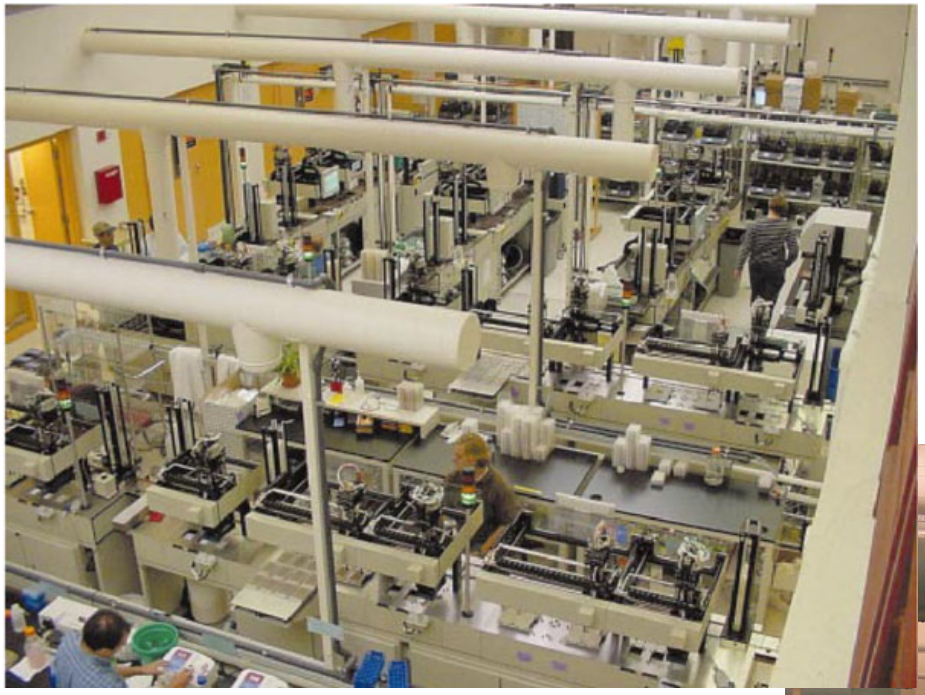
Automated Sequencing

ABI 3100 Automated Capillary DNA Sequencer

Electropherogram



Sequencing factories



Whitehead institute

Custom-designed factory-style conveyor belt robots:

perform all functions from purifying DNA from bacterial cultures through setting up and purifying sequencing reactions.

Automated sequencing

2001

The HGP consortium publishes its working draft in *Nature* (15 February), and Celera publishes its draft in *Science* (16 February).

Human genome: 3 billion bases

In total 23 billion bases were sequenced (7.5-fold coverage)

This comprises 23 Gbyte of data

Public Human Genome project

-1990-2003

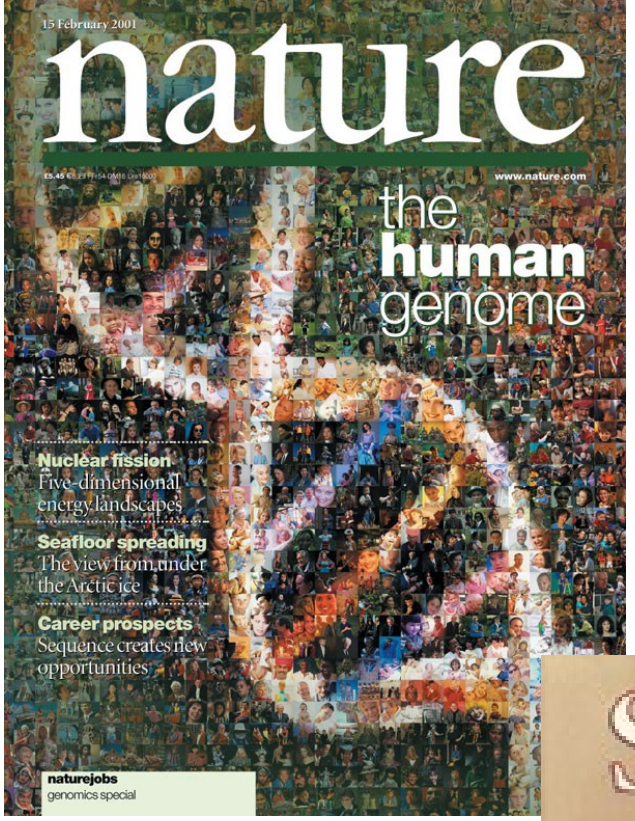
- 3 billion US dollar

Private Celera genome project

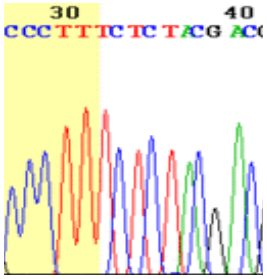
-1998 - 2001

- Craig Venter

- 300 million US dollar 2001



Traditional versus high throughput DNA sequencing



http://en.wikipedia.org/

Sanger, one run:

? hours (human genome took 15 years)

1-384 sequences

300-1000 nt per sequence

1 KB - 384 KB data

= 1,000-384,000 bases

Year: 1963 – now

http://www.454.com/

Roche 454, one run:

7.5 hours

1,000,000 sequences

750 nt per sequence

35 GB data (including images)

1 GB data (excluding images)

= 750,000,000 bases

Year: 2005 – 2015

http://www.illumina.com

Illumina, one run:

1-3.5 days

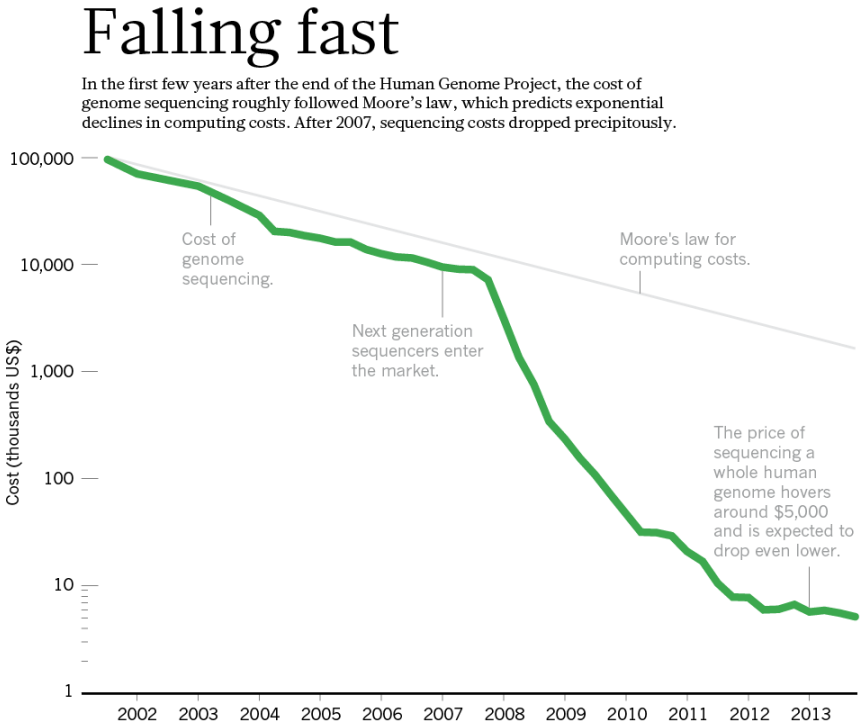
5 billion sequences

2x 150 nt per sequence

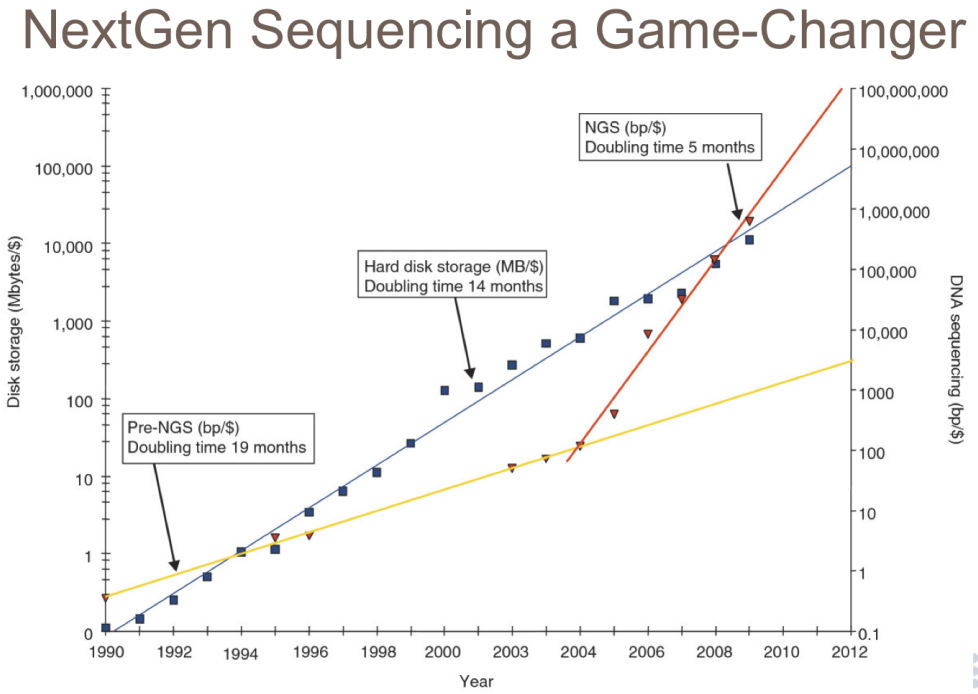
1.5 TB data

Year: 2008 – now

Data size challenges



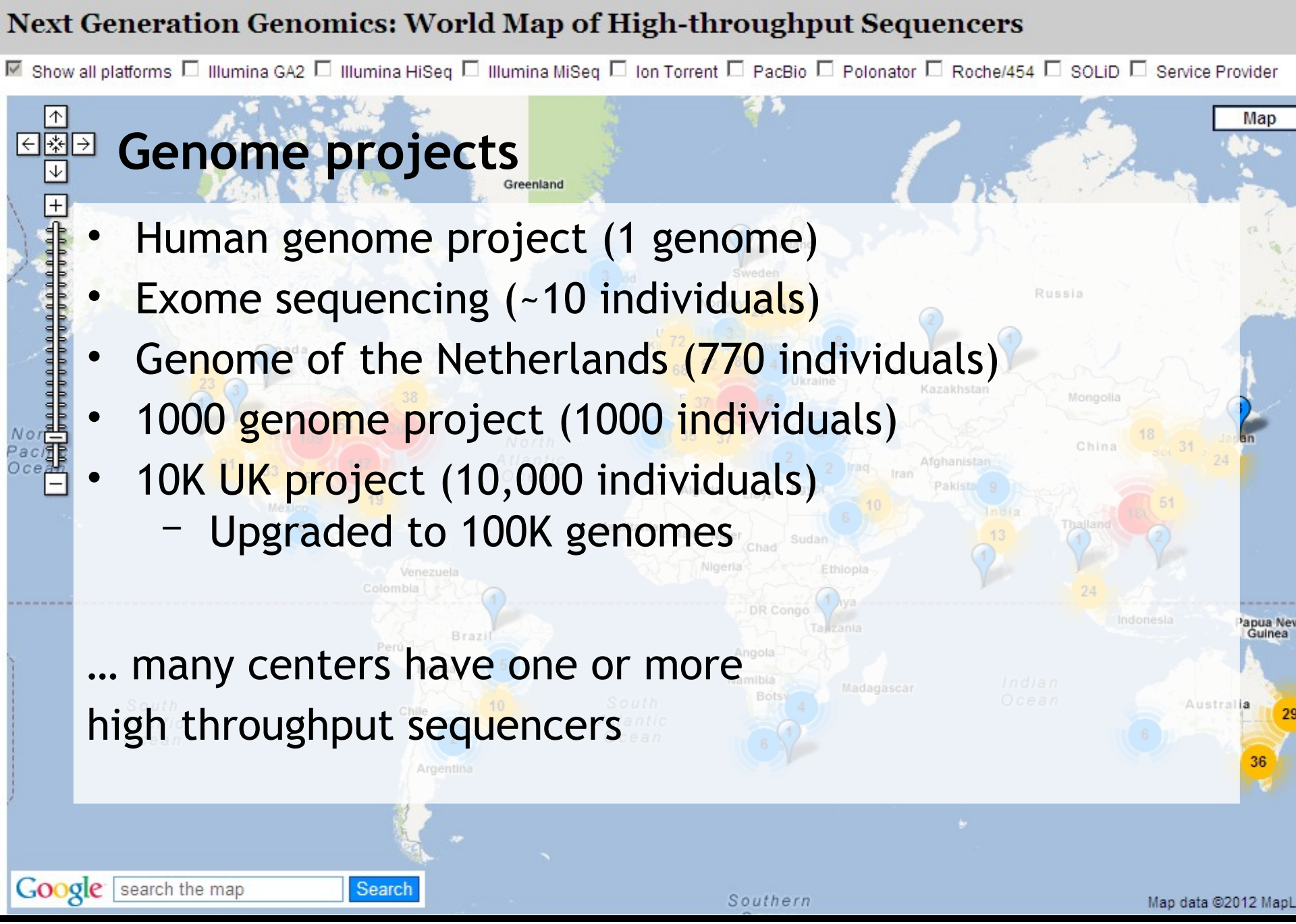
Hayden, Nature, 2014



Stein, Genome Biology, 2010

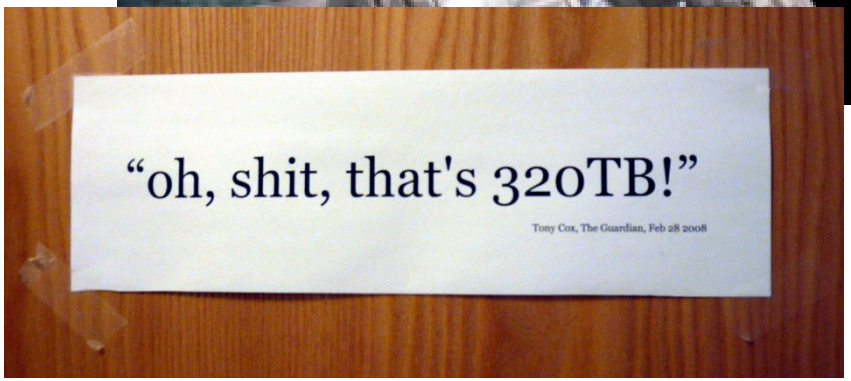
COMPUTING
Sequencing rate is higher than Moore's law

STORAGE
Sequencing costs lower than data storage



http://omicsmaps.com/

image credit: Digital Vision, PhotoDisc, Matt Ray/EHP



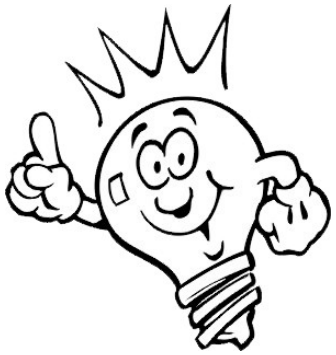
Sign @ Wellcome-Sanger, Cambridge, UK

Existing hardware

PC

Small cluster

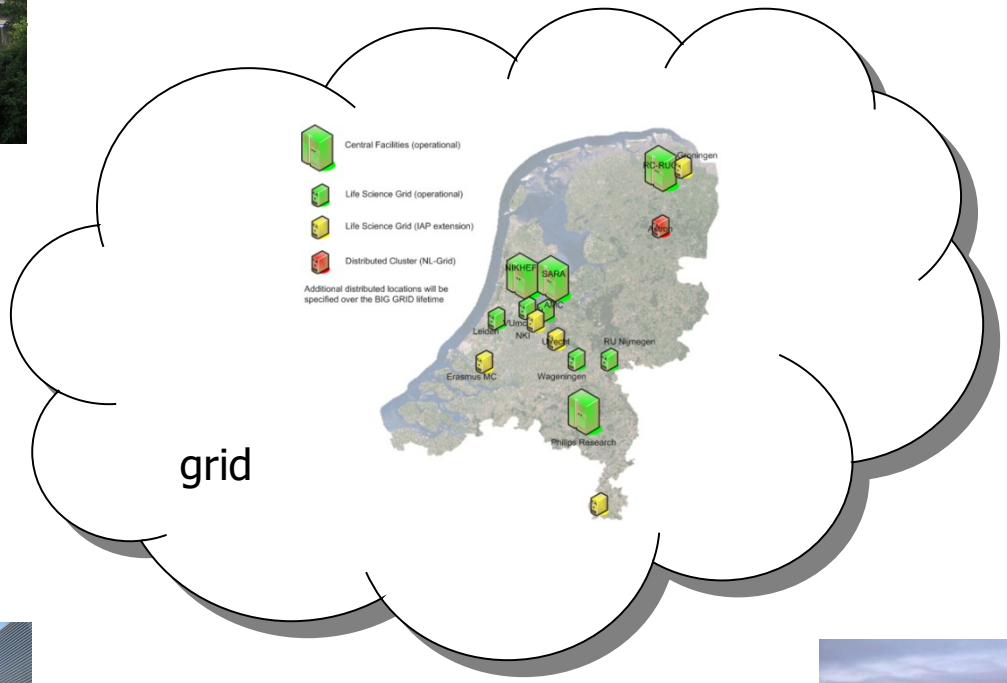
Buy a bigger cluster (centralized model)



Dutch life science grid

BiG Grid
the dutch e-science grid

SURF SARA



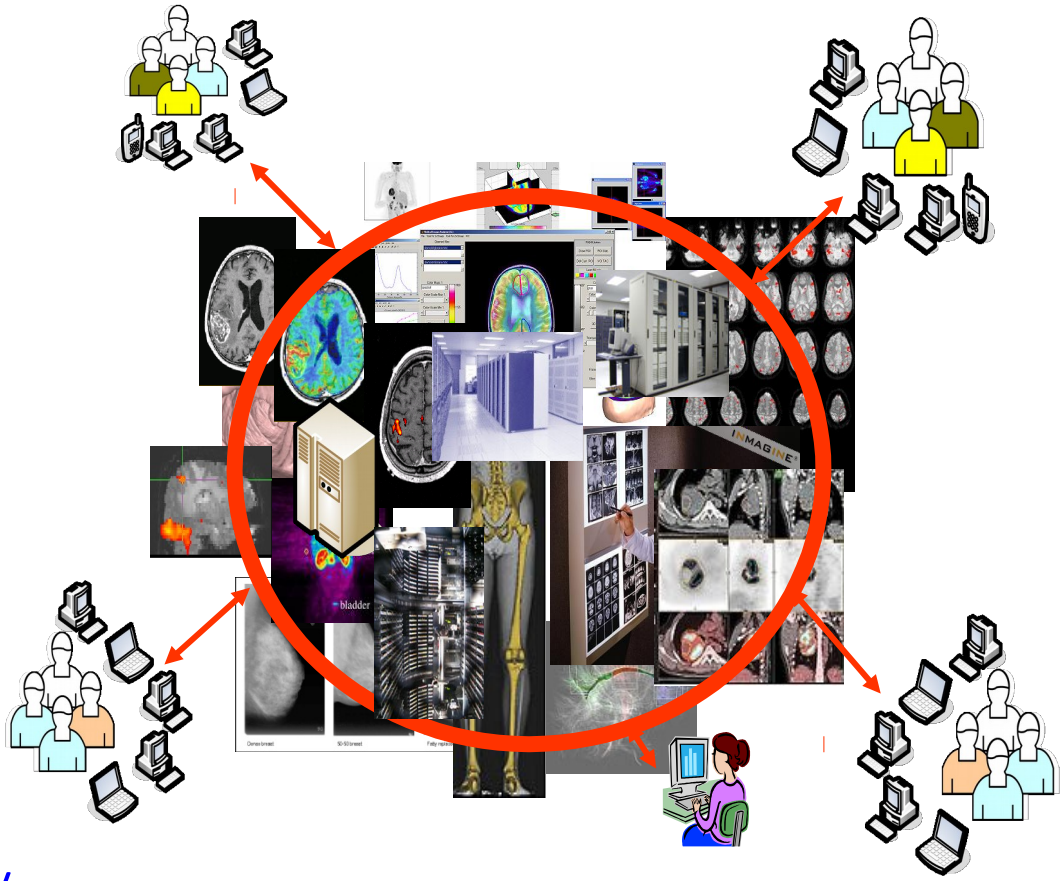
https://www.surfsara.nl/

Cloud computing

Not only hardware

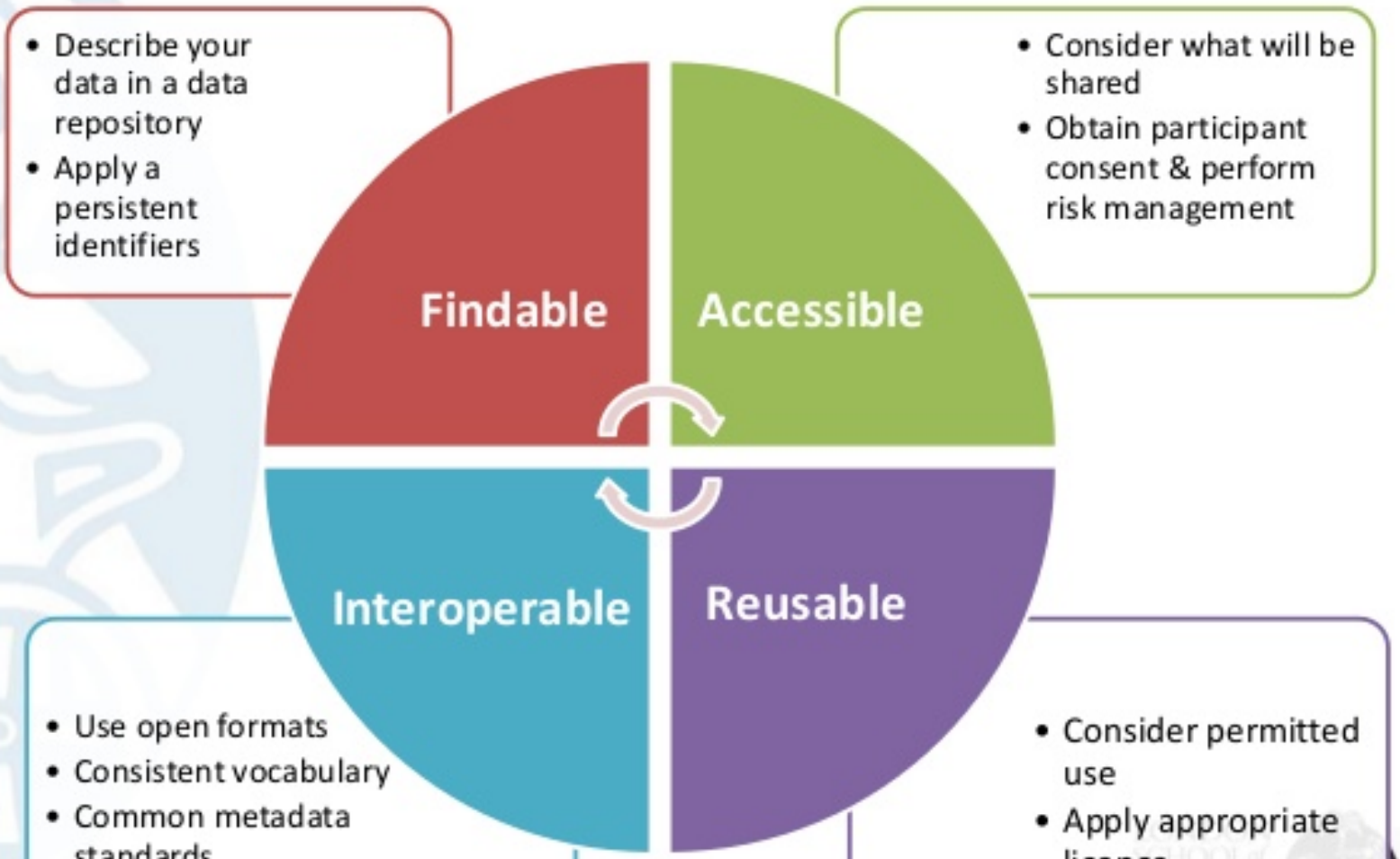
- Software
- Peopleware
- Sharing

e-Science



More info:
http://www.egi.eu/
http://www.ebioscience.amc.nl/

FAIR data

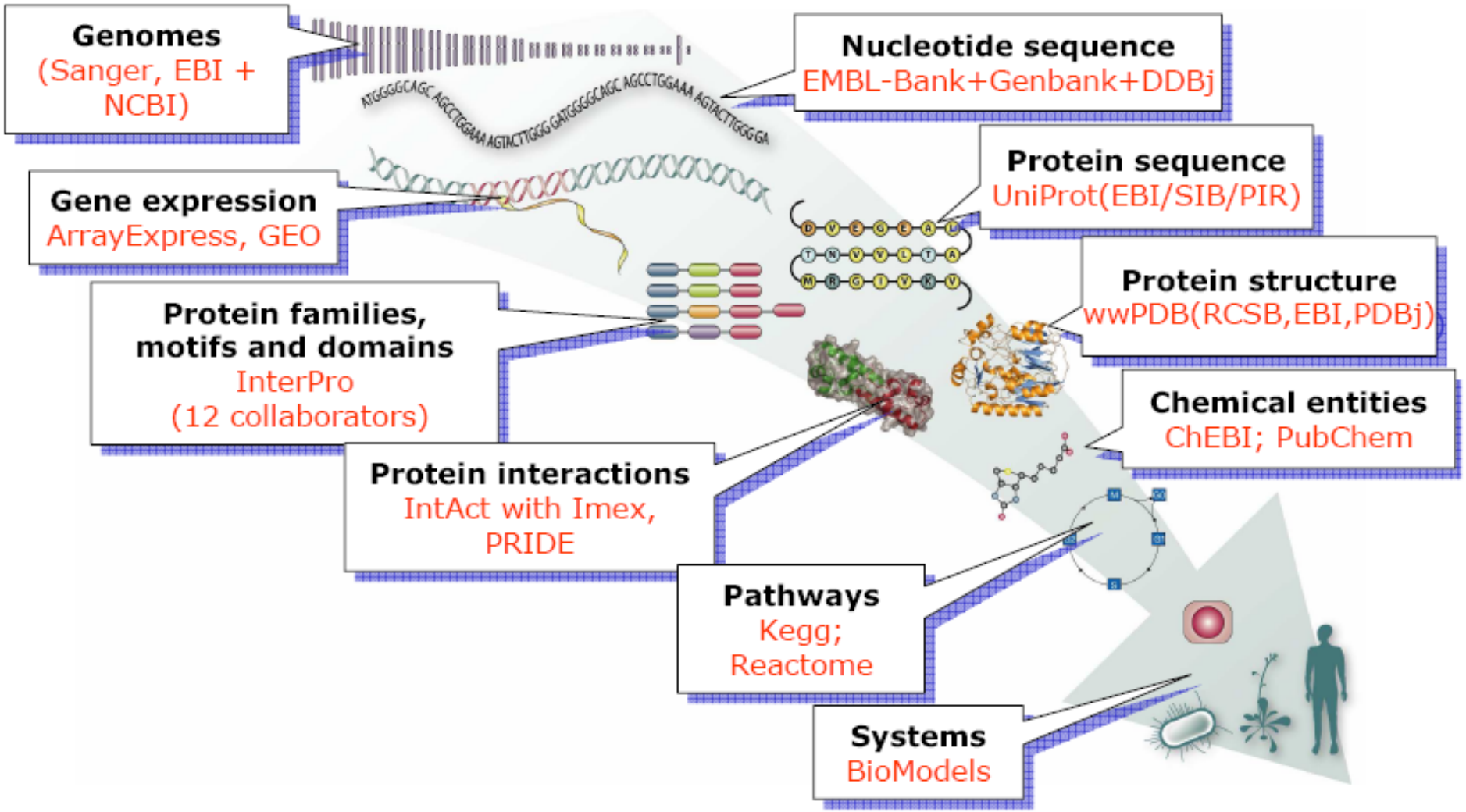


Wilkinson et al. (2016) Scientific Data

Tips and tricks… where to start?

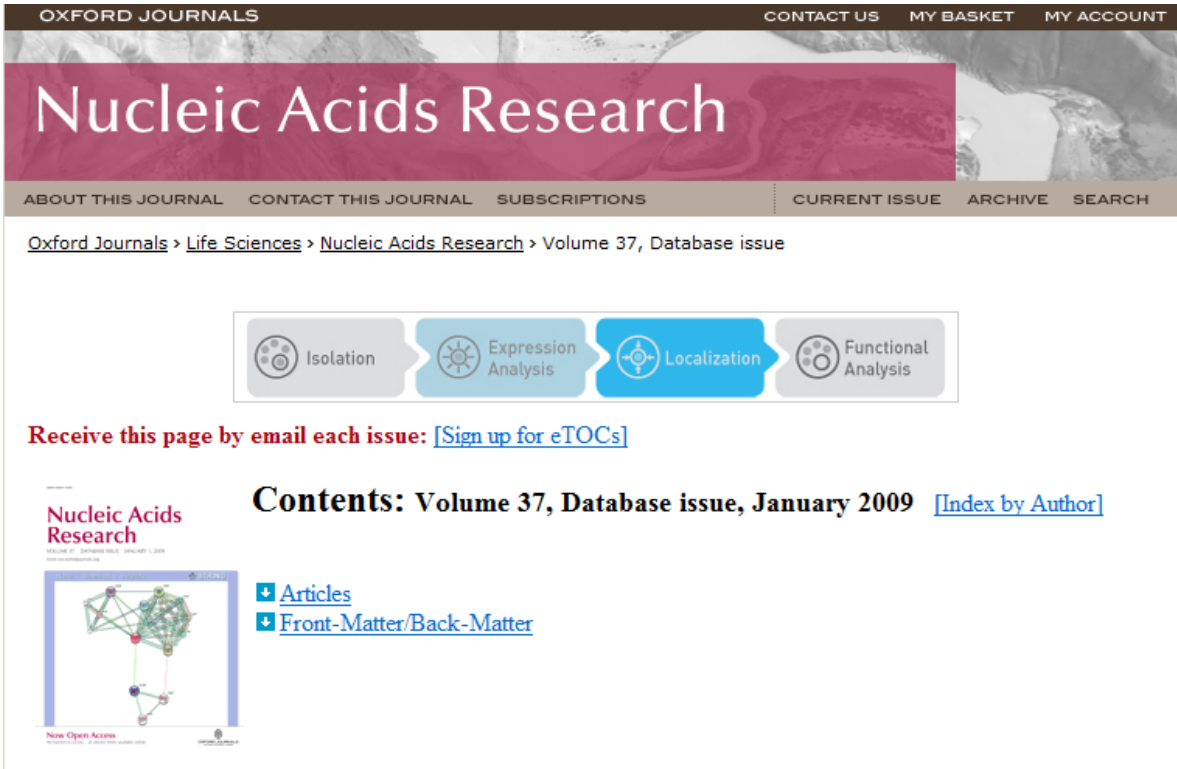
Biological databases
Publicly available tools

Publicly available biological databases



Molecular Biology Database Collection 2015

- Published by Nucleic Acids Research (scientific journal)
- Currently ? databases.
- Latest issue new databases
- Coverage is far from exhaustive.



Database access

- Web-interface
  - Keyword search
  - Browsing / cross-linking
  - BLAST
- From other applications
  - Application programming interface (API)
  - Web services
  - In-house / third party software
- Download
  - Full database
  - Specific records
  - Different formats (text, xml, rdf, etc)

Sequence analysis

Function prediction (similarity, sequence search)

Localisation (genefinding)

Grouping (genes, protein families)

Conservation (motifs, functional blocks)

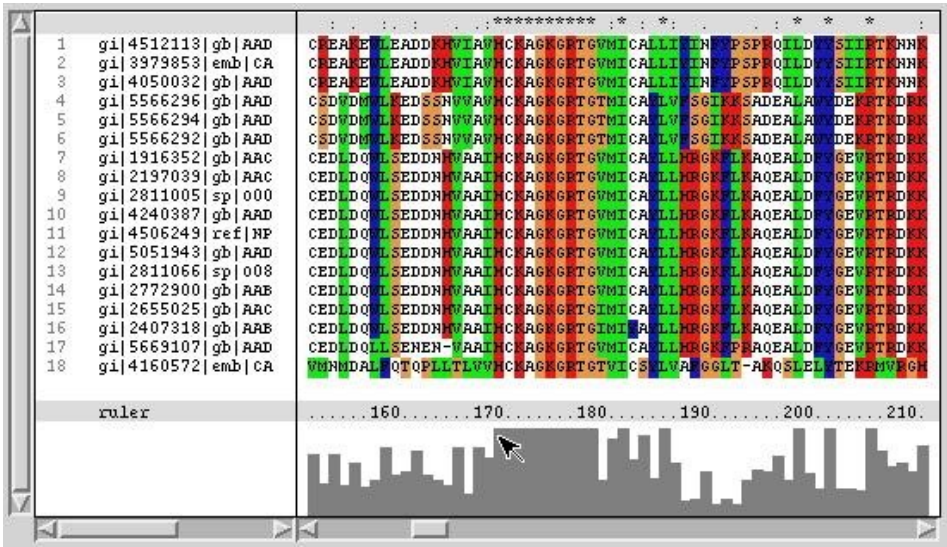
SNPs and mutations (variations)

Sequence analysis

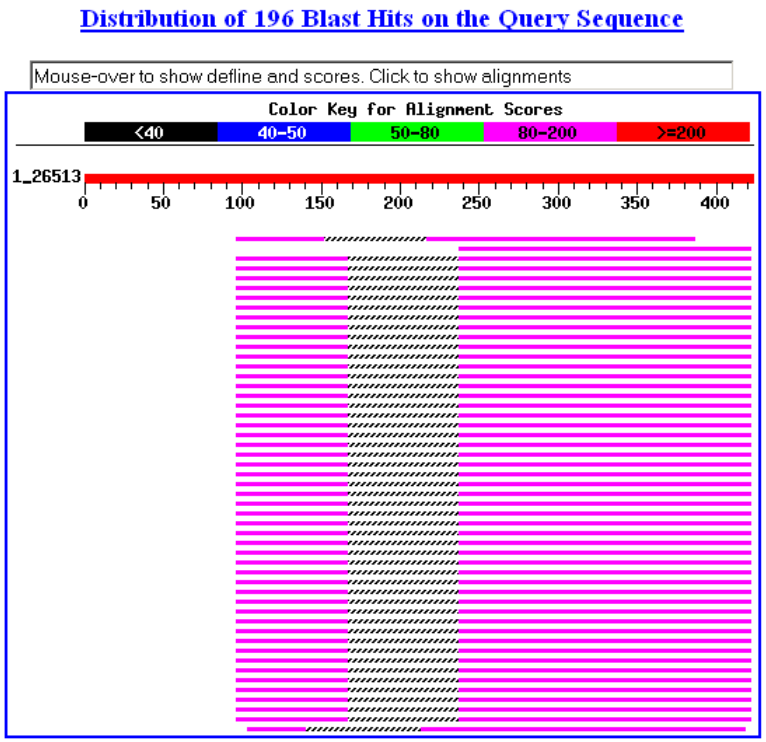
Pairwise alignment: in-exact matching of 2 sequences

```
CTCCTGAGGCAAATCTGTCAGTCCATCCTGGCTGAGTCCTCGCAGTCCCCGGCAGATCTTGAAGAAAAGA
|||||||||  |||  |||  |||||   |   |||||  |   |    ||||   | |  ||    |||  ||  |||  |||
CTCCTGAGGAAAACCTGCCAGTCAGTTGTGGCTAATCCTCAGCAGCACTCAGC---TCTGGAGGAAGAGA
```

Multiple sequence alignment: in-exact matching of >2 sequences

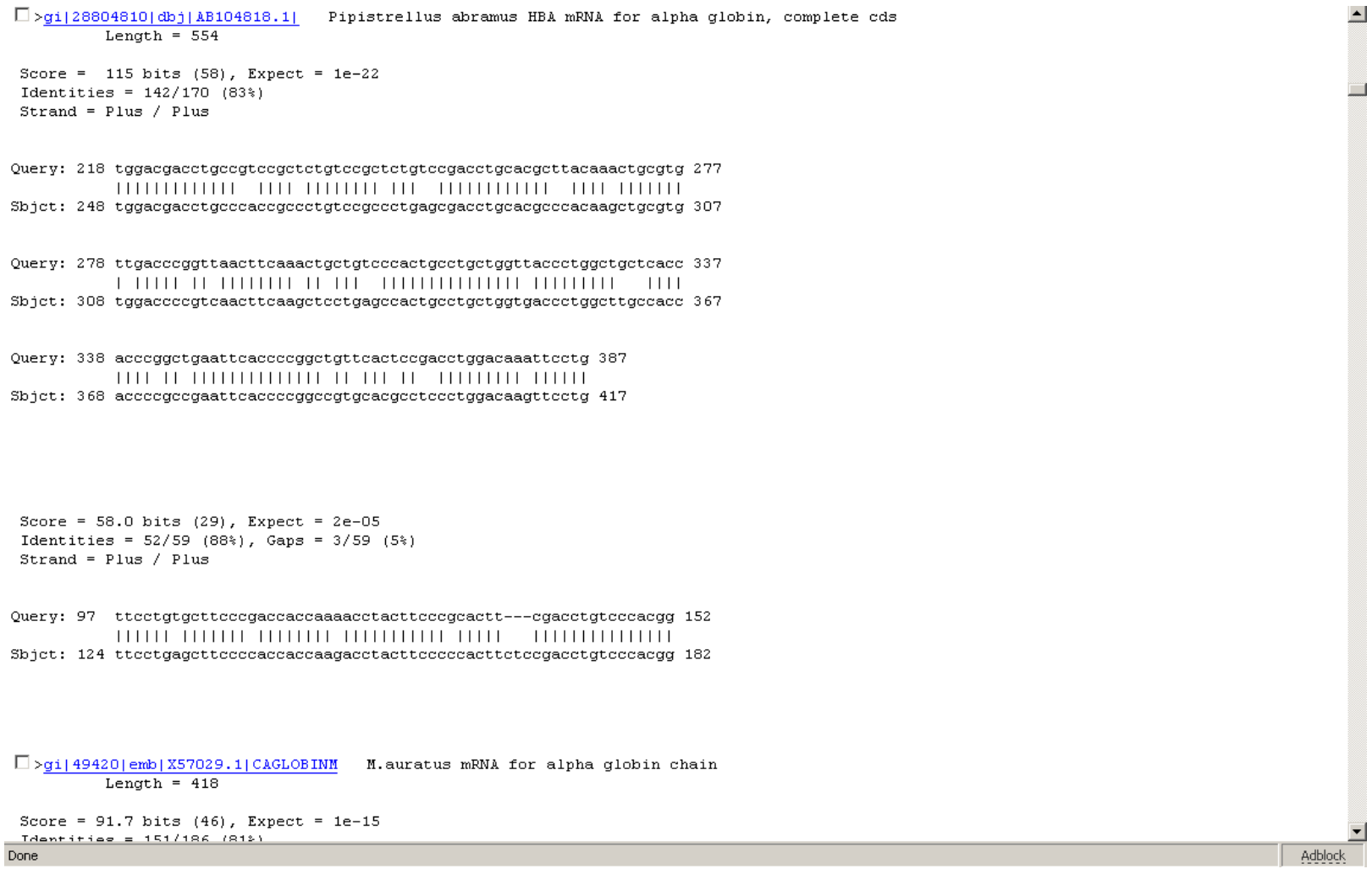


Blast output



Sequences producing significant alignments:	Score (bits)	E Value	
gi|28804810|dbj|AB104818.1| Pipistrellus abramus HBA mRNA f...	115	1e-22	
gi|49420|emb|X57029.1|CAGLOBINM M.auratus mRNA for alpha gl...	92	1e-15	
gi|40225899|gb|BC032122.2| Homo sapiens hemoglobin, alpha 2...	84	3e-13	G U E
gi|30047787|gb|BC050661.1| Homo sapiens hemoglobin, alpha 2...	84	3e-13	G U

Blast output - alignments



>gi|28804810|dbj|AB104818.1| Pipistrellus abramus HBA mRNA for alpha globin, complete cds
Length = 554

Score = 115 bits (58), Expect = 1e-22
Identities = 142/170 (83%)
Strand = Plus / Plus

Query: 218 tggacgacctgccgtccgctctgtccgctctgtccgacctgcacgcttacaaactgcgtg 277
Sbjct: 248 tggacgacctgcccaccgccctgtccgccctgagcgacctgcacgcccacaagctgcgtg 307

Query: 278 ttgacccggttaacttcaaactgctgtcccactgcctgctggttaccctggctgctcacc 337
Sbjct: 308 tggaccccgtcaacttcaagctcctgagccactgcctgctggtgaccctggcttgccacc 367

Query: 338 acccggctgaattcaccccggctgttcactccgacctggacaaattcctg 387
Sbjct: 368 accccgccgaattcaccccggccgtgcacgcctccctggacaagttcctg 417

Score = 58.0 bits (29), Expect = 2e-05
Identities = 52/59 (88%), Gaps = 3/59 (5%)
Strand = Plus / Plus

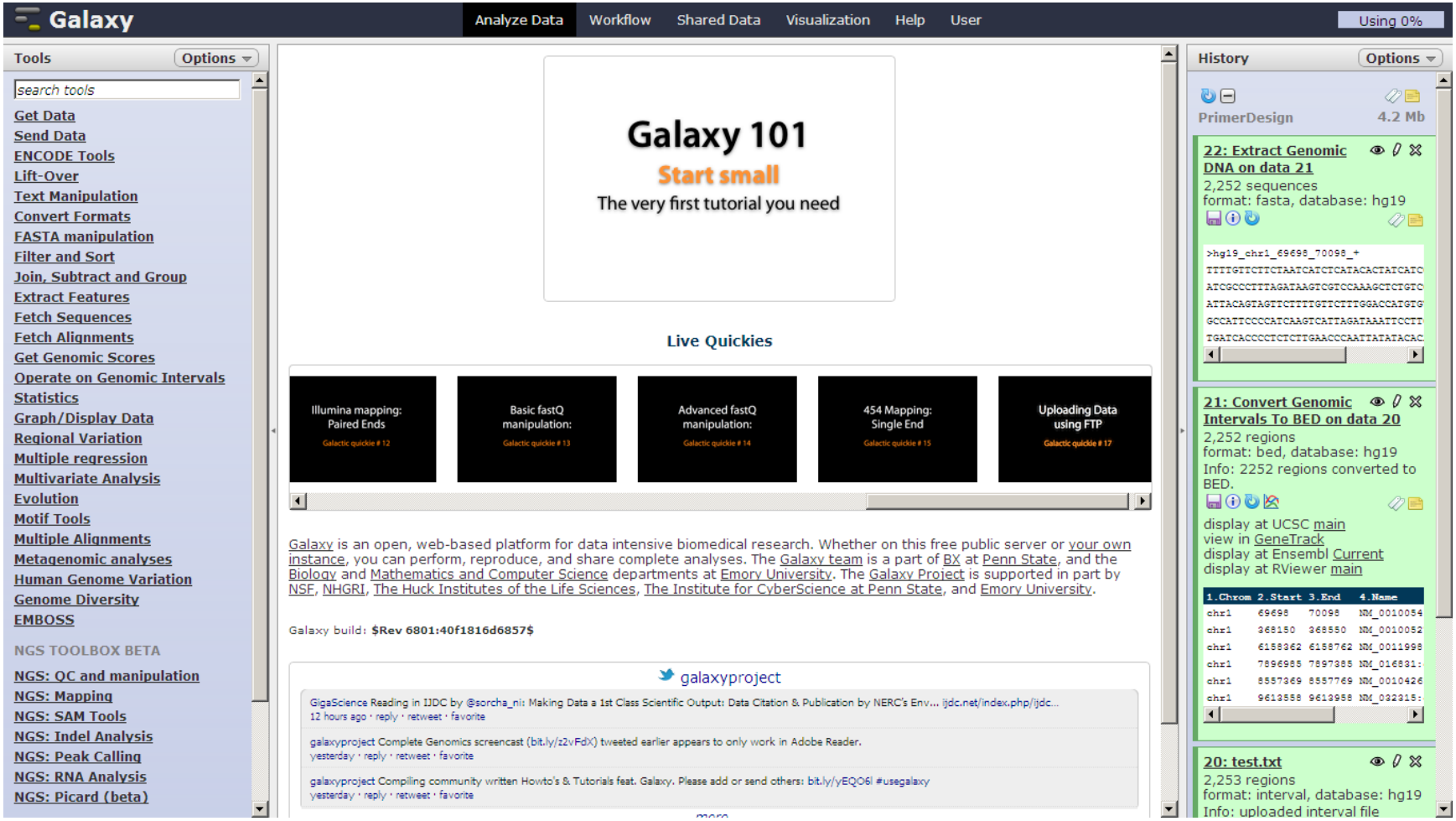
Query: 97 ttcctgtgcttcccgaccaccaaaacctacttcccgcactt---cgacctgtcccacgg 152
Sbjct: 124 ttcctgagcttccccaccaccaagacctacttcccccacttctccgacctgtcccacgg 182

>gi|49420|emb|X57029.1|CAGLOBINM M.auratus mRNA for alpha globin chain
Length = 418

Score = 91.7 bits (46), Expect = 1e-15
Identities = 151/186 (81%)

Done Adblock

Galaxy



Tip: Has instruction videos for a few use cases
e.g. for analyzing next gen sequence data

https://galaxyproject.org/

Nucleic acid research - annual web server issue

bioinformatics.ca
links directory

https://links.bioinformatics.ca/

Bioinformatics Links Directory

Search Bioinformatics Links Directory | Search Directory

Change Text

Username:

Password:

Log in

Create new
Request ne

Main Pag
Citations
Acknowle
News
Suggest a
NAR Colla
RSS Feed
Support

Bioinformatics Links Directory

The Bioinformatics Links Directory features curated links to molecular resources, tools and databases. The links listed in this directory are selected on the basis of recommendations from bioinformatics experts in the field. We also rely on input from our community of bioinformatics users for suggestions. Starting in 2003, we have also started listing all links contained in the NAR Webserver issue.

Computer Related (82)

This category contains links to resources relating to programming languages often used in bioinformatics. Other tools of the trade, such as web development and database resources, are also included here.

DNA (517)

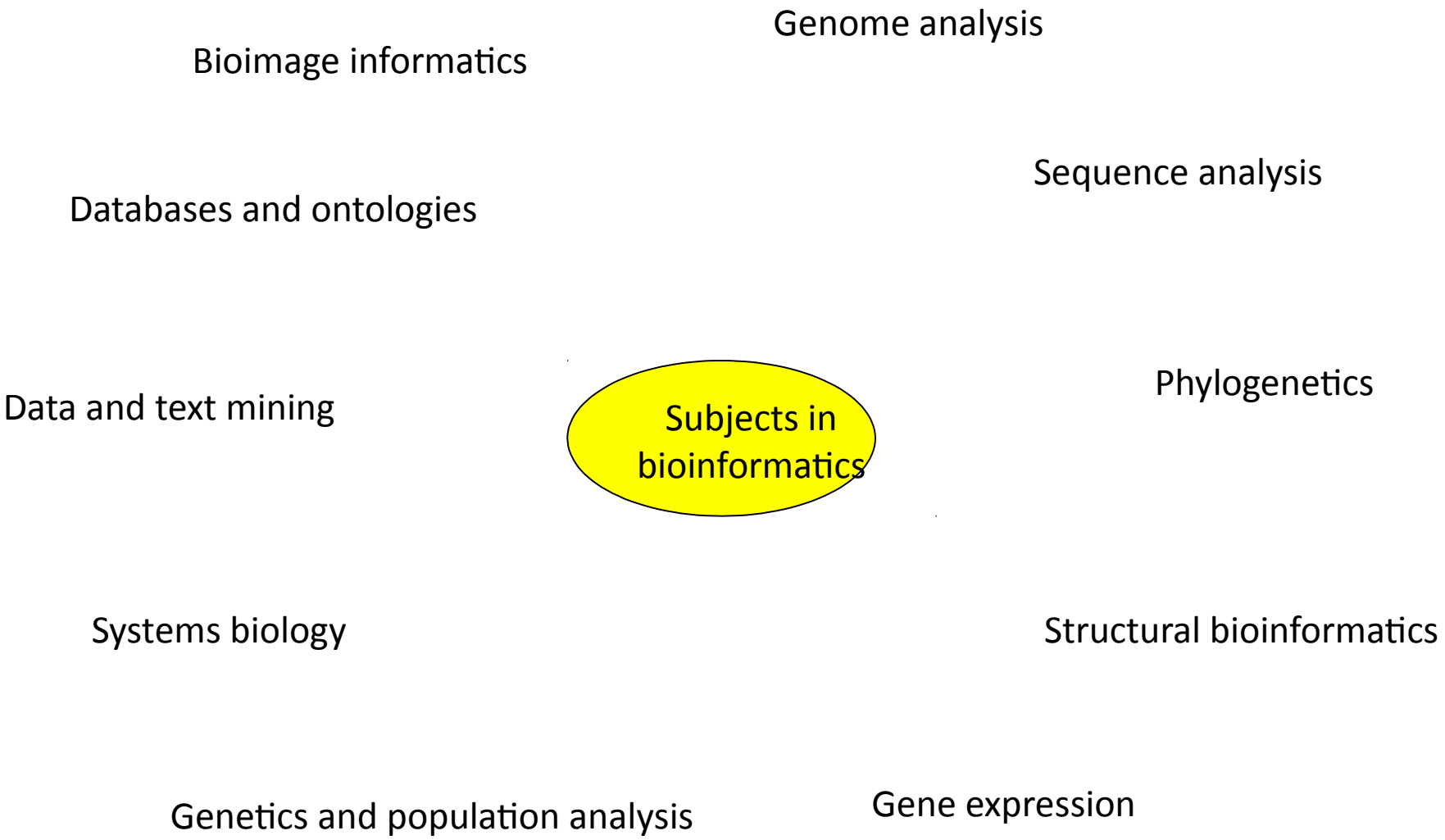
This category contains links to useful resources for DNA sequence analyses such as tools for comparative sequence analysis and sequence assembly. Links to programs for sequence manipulation, primer design, and sequence retrieval and submission are also listed here.

Education (67)

Links to information about the techniques, materials, people, places, and events of the greater bioinformatics community. Included are current

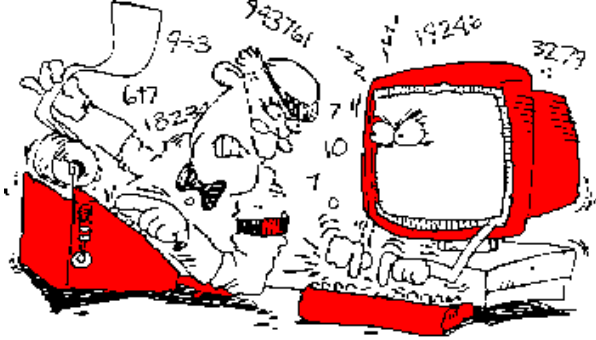
Expression (382)

Links to tools for predicting the expression, alternative splicing, and regulation of a gene sequence are found here. This section also



http://www.oxfordjournals.org/our_journals/bioinformatics/for_authors/scope_guidelines.html

Topics computer exercises



Query public databases

Sequence format conversion

Automatic RNA → protein translation

Primer design

Gene finding

Tomorrow 13:30-15:00

https://bioinformatics.amc.nl/