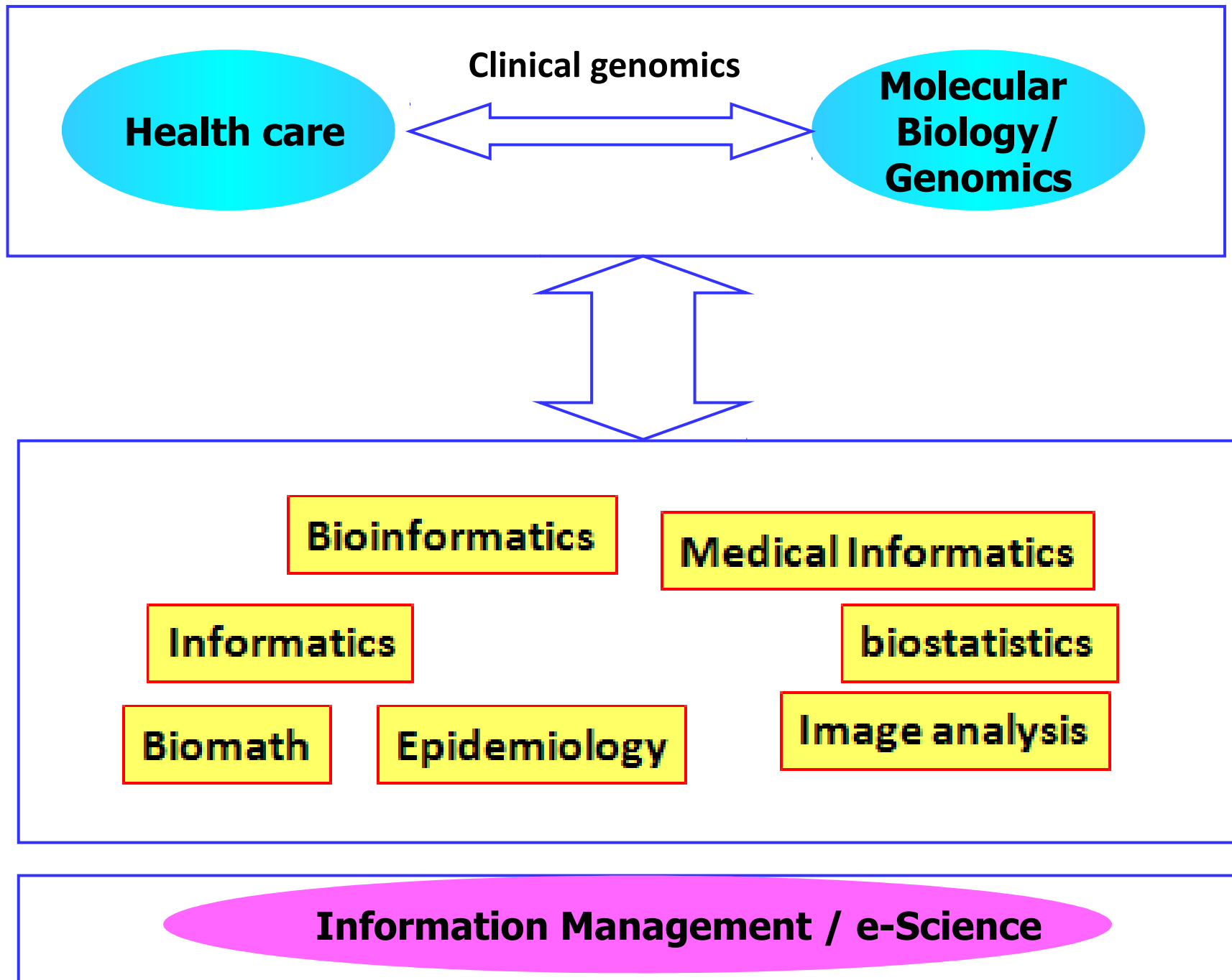


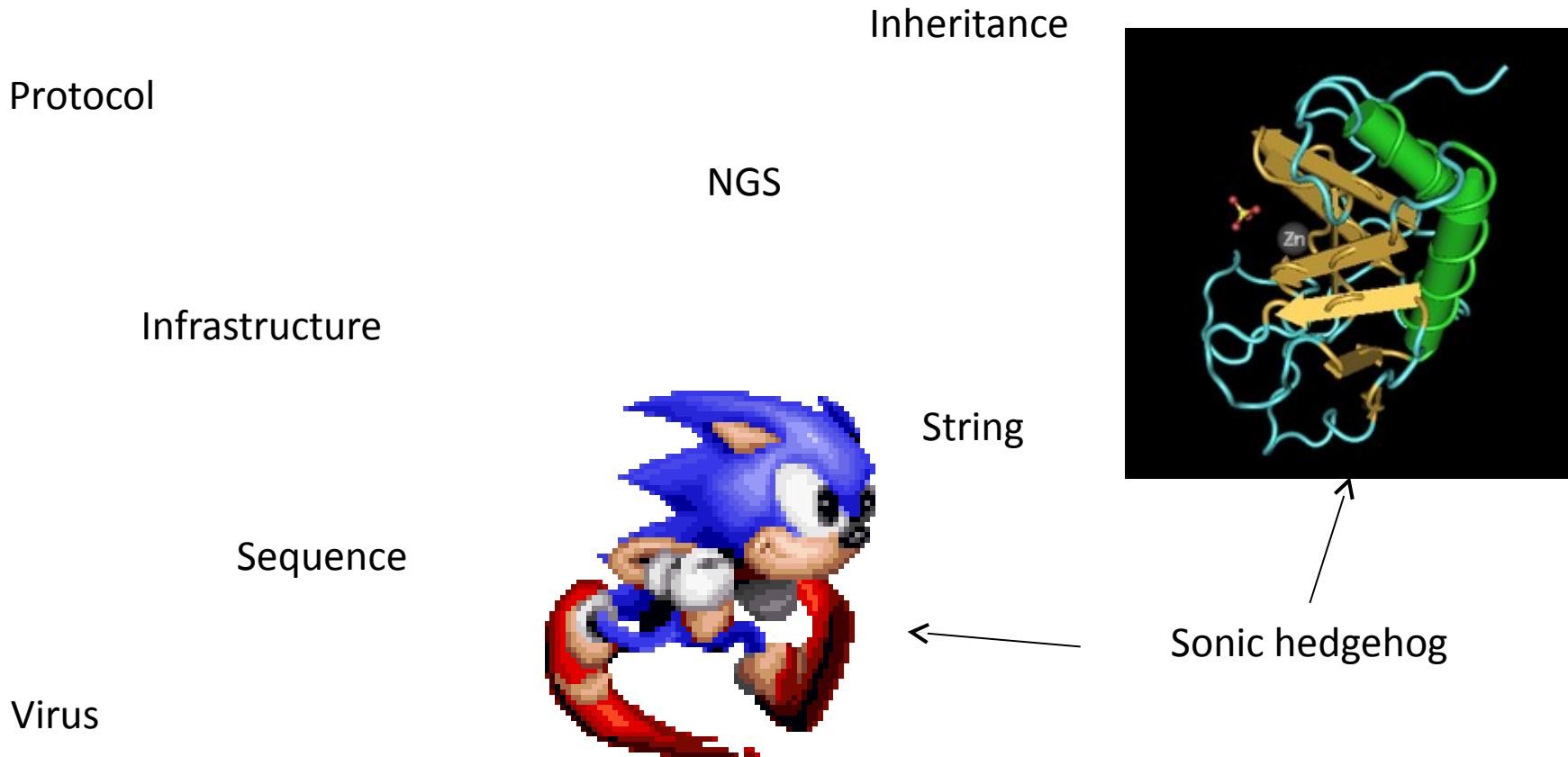
Introduction to bioinformatics

DNA technology course

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



Lost in translation (biology and informatics)



Definitions of bioinformatics

Adapted definition according to Wikipedia

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

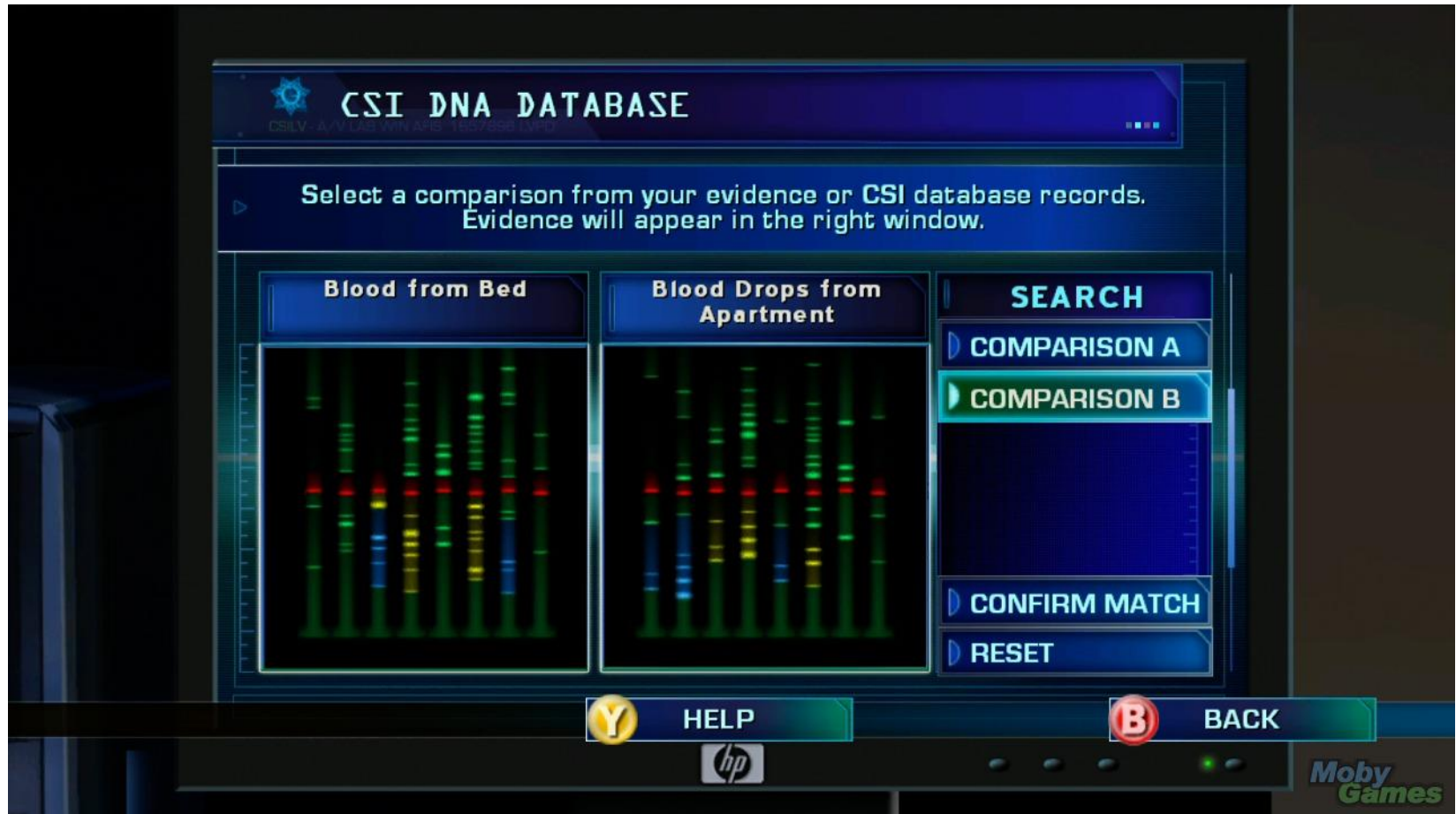
The creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management, analysis and interpretation of biological data.

What is bioinformatics?

Extraction of biological knowledge from complex data

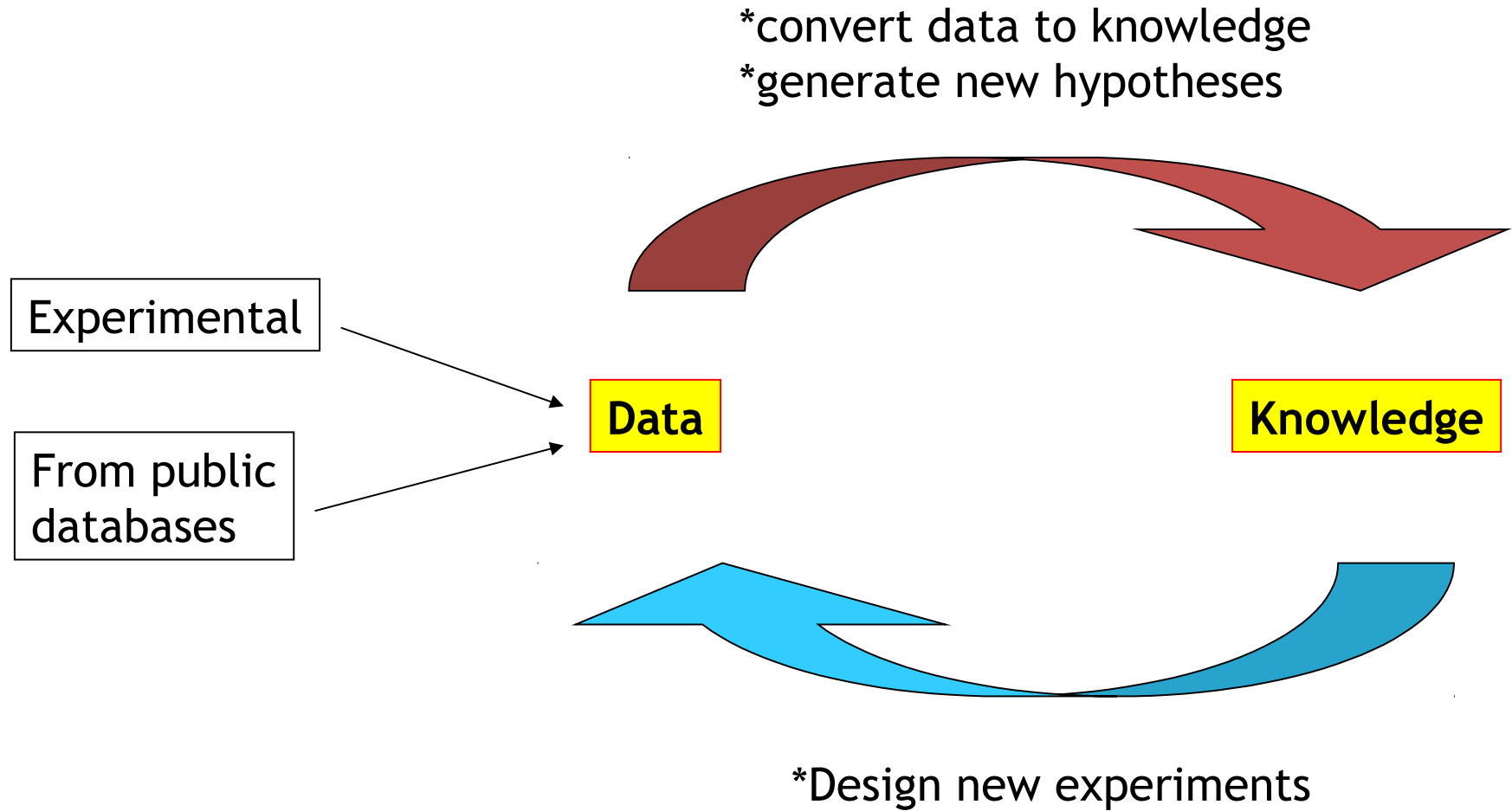


What is bioinformatics?



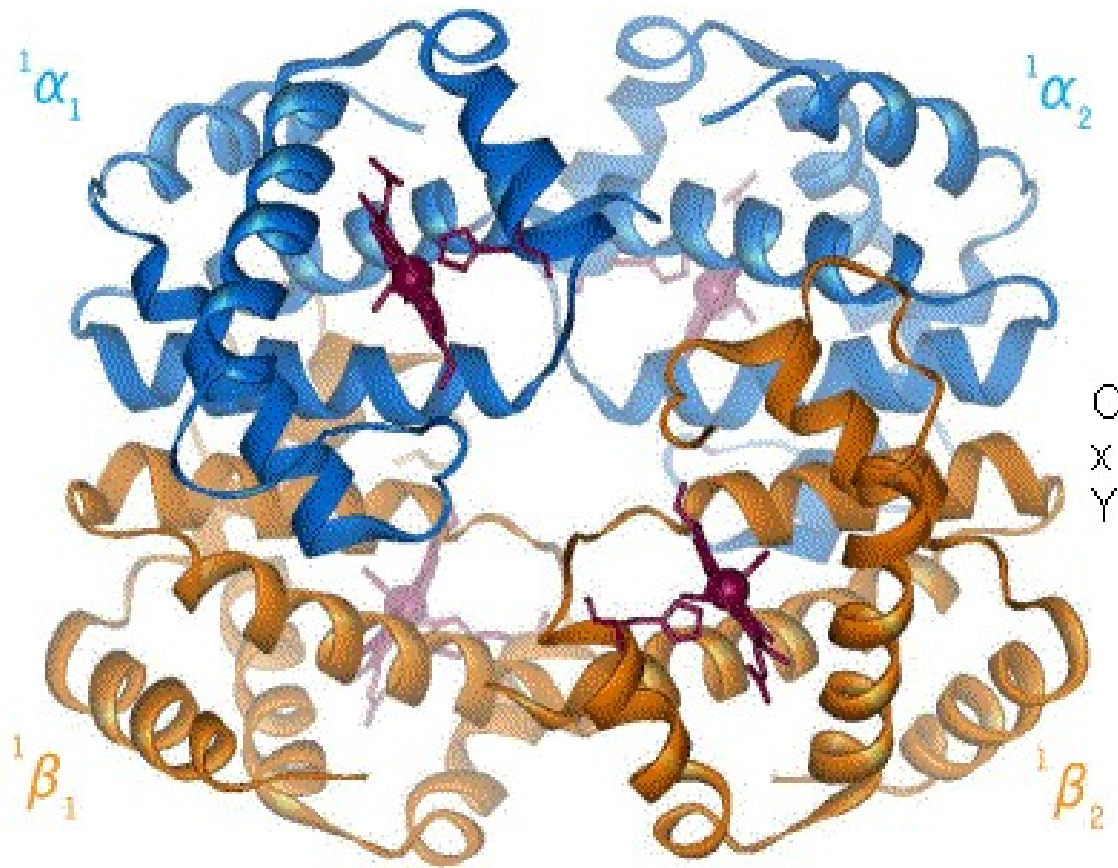
... one of the results *might* be a tool you can use

Bioinformatics



And information management.....

How does molecule A interact with protein B?



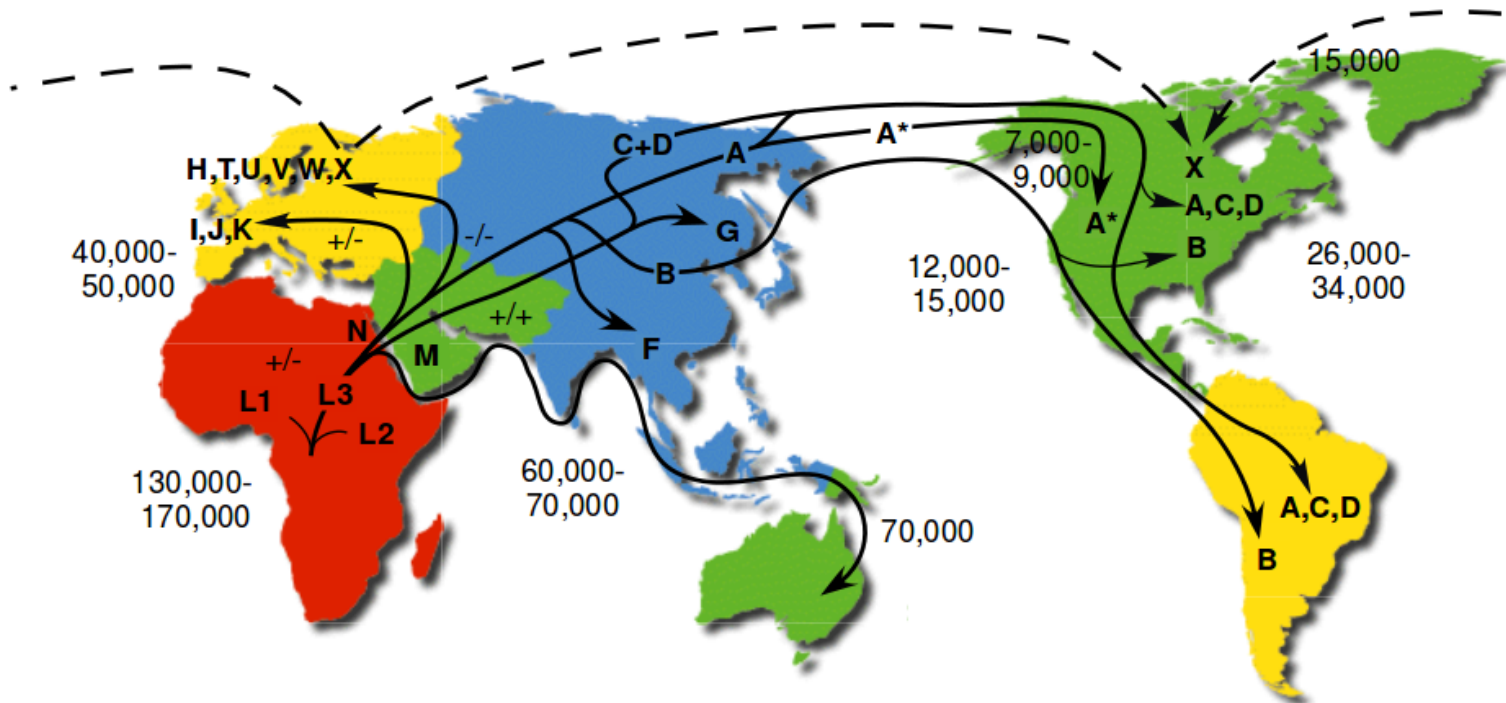
A schematic visual model of oxygen-binding process, showing all four monomers and hemes, and protein chains only as diagramatic coils, to facilitate visualization into the molecule. (<http://en.wikipedia.org/wiki/Hemoglobin>)

Study migration

Human mtDNA Migrations

<http://www.mitomap.org/pub/MITOMAP/MitomapFigures/WorldMigrations.pdf>

Copyright 2002 © Mitomap.org

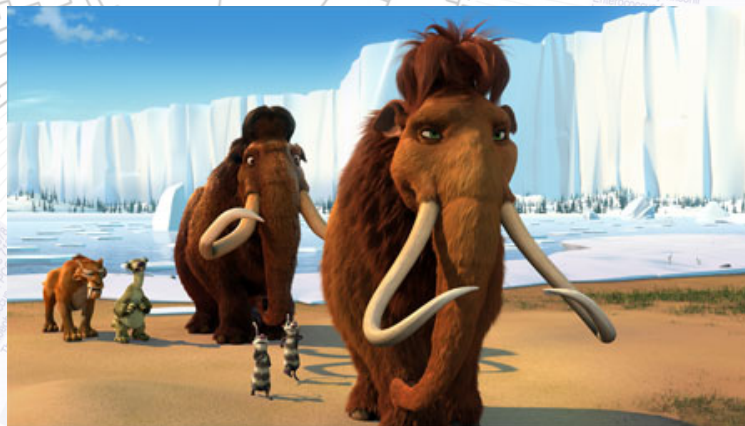


+/-, +/+, or -/- = Dde I 10394 / Alu I 10397

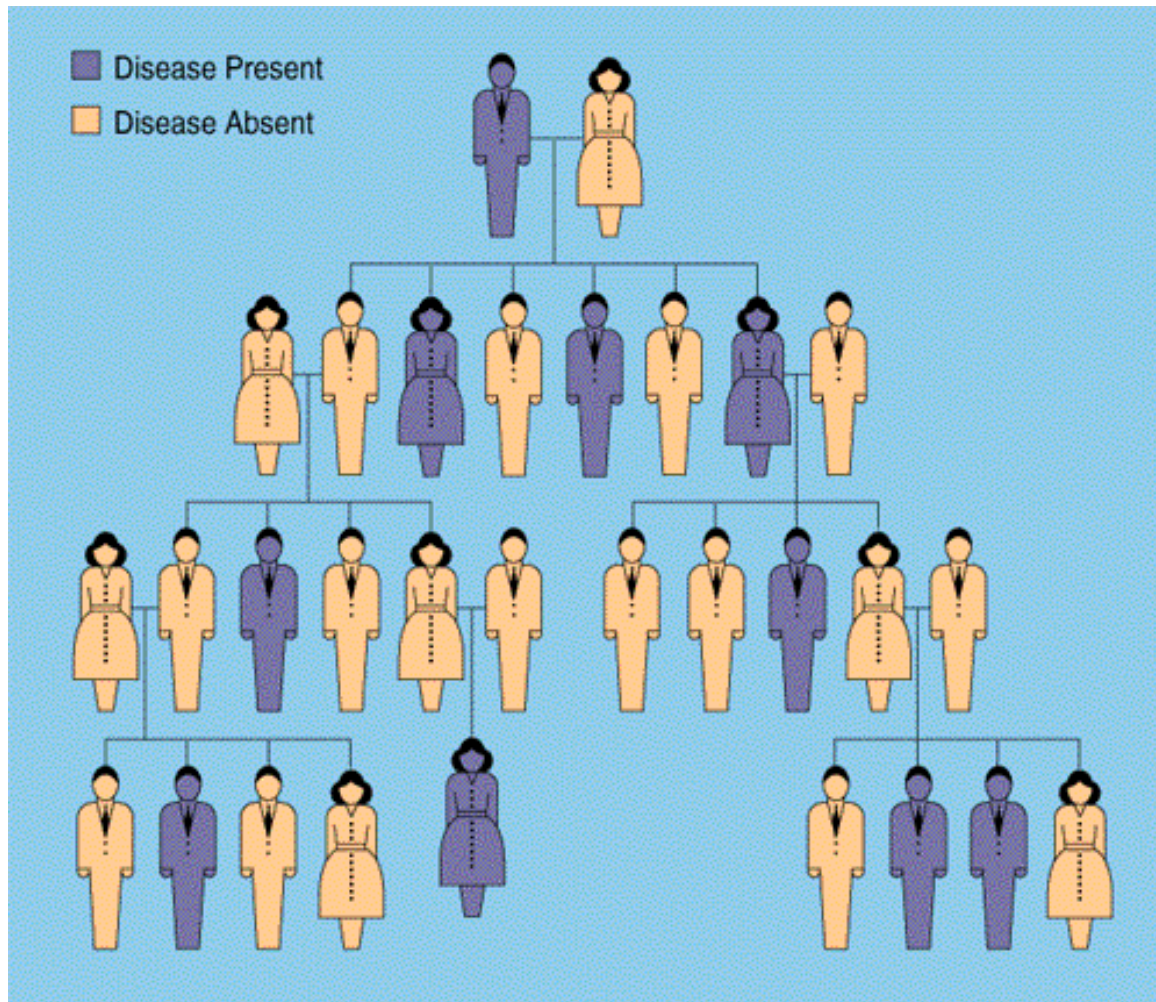
* = Rsa I 16329

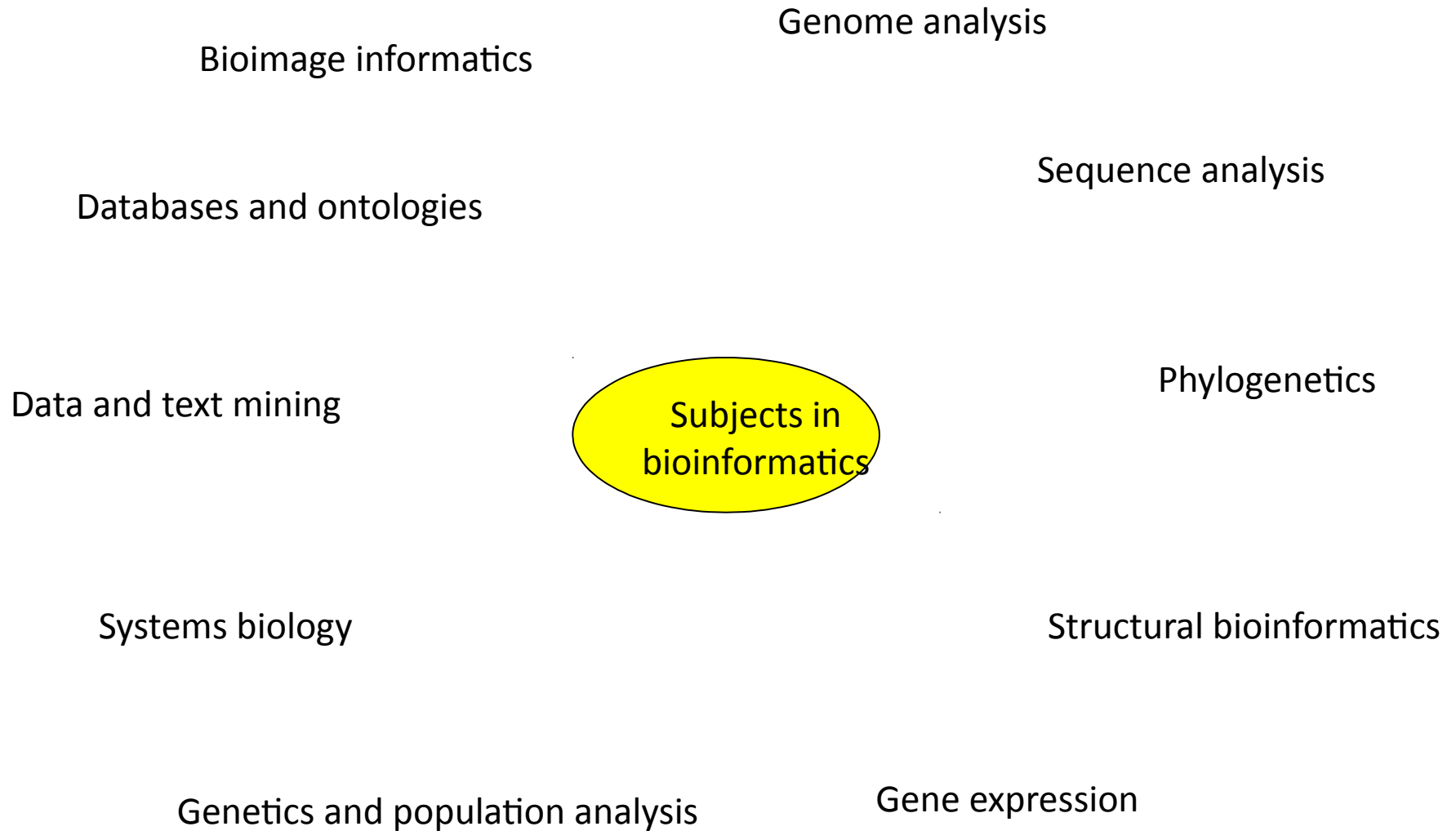
Mutation rate = 2.2 - 2.9 % / MYR

Time estimates are YBP



Which gene(s) causes disease X?





Bioinformatics

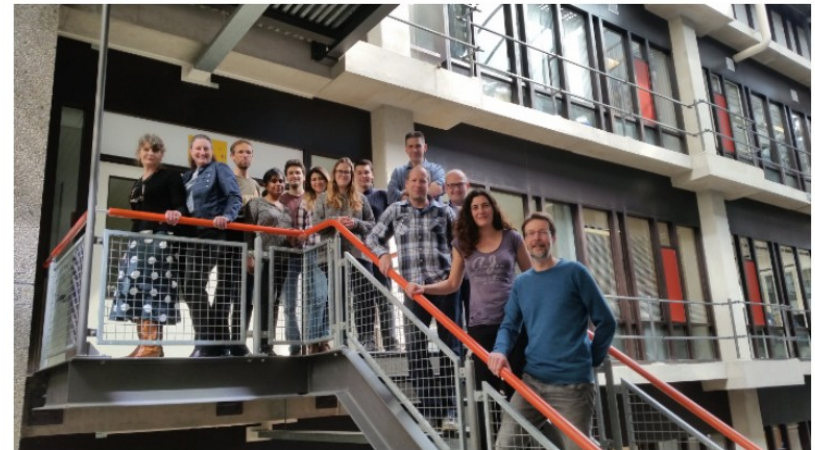
Groups and institutes

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

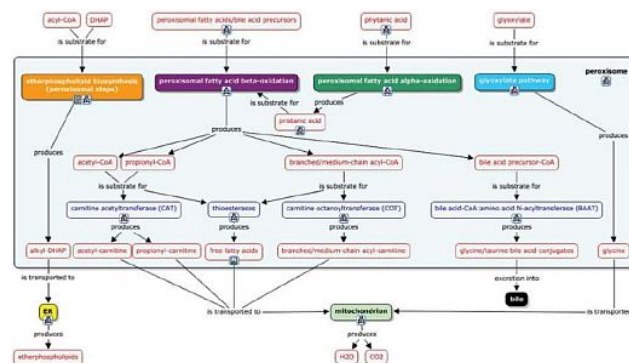
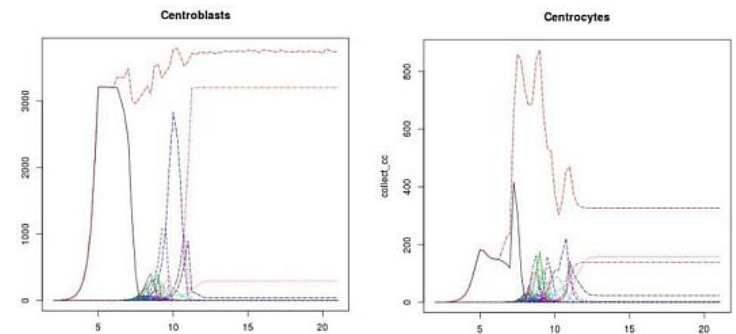
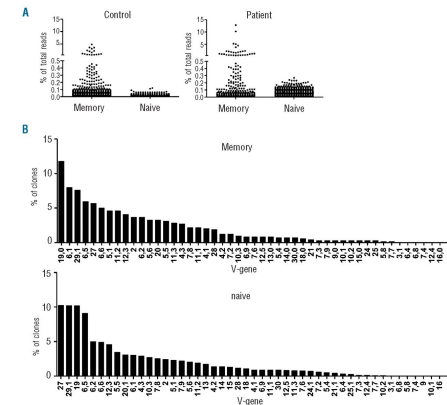
Systems biology (bottom-up)

Mathematical modelling

Information management

Knowledge bases

Science gateways



Dutch Techcentre for Life Sciences (DTL)

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



DTL | 

DUTCH TECHCENTRE FOR LIFE SCIENCES

 [FACILITIES](#) [COURSES](#) [FAIR DATA](#) [ELIXIR](#) [PARTNERS](#) [COMMUNITY](#) [ABOUT](#)

COMMUNITY

DTL is all about the vibrant community of technical experts, (data) engineers, programmers and a wide range of researchers working on biological questions. Check our calendar to find meetings and courses relevant to your line of work.

[MEETINGS](#)

[COURSES](#)



RECENT & UPCOMING

[News](#) [Events](#) [Courses](#) [Jobs](#) [Tweets](#)



ECCB 2016

ECCB 2016: Extended submission deadlines

All deadlines for abstract submission to the European Conference on Computational Biology (ECCB) have been extended. The new deadlines are 5 April 2016 (Proceedings & Highlights) and... [Continue reading →](#)

WE ARE HIRING!



DTL | 

DTL seeks two Software Engineers Linked Data to provide leadership and innovative software development services. Candidates

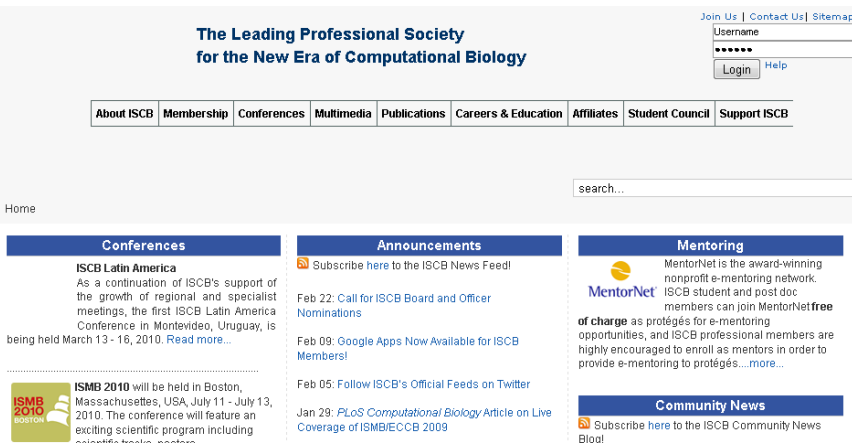
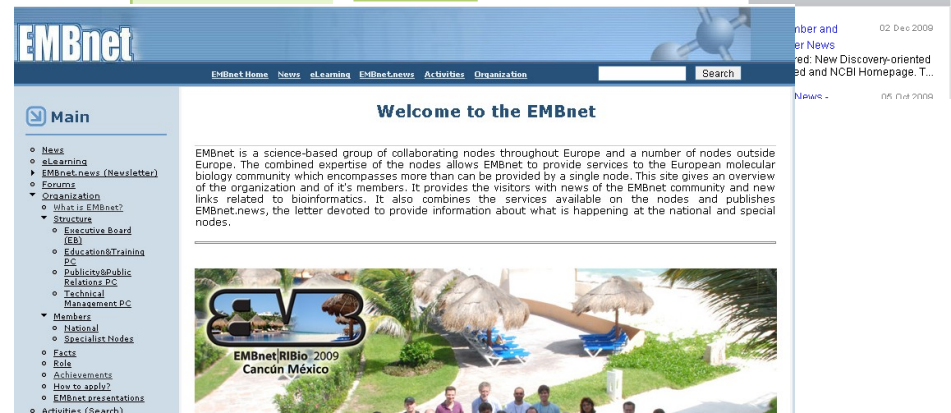
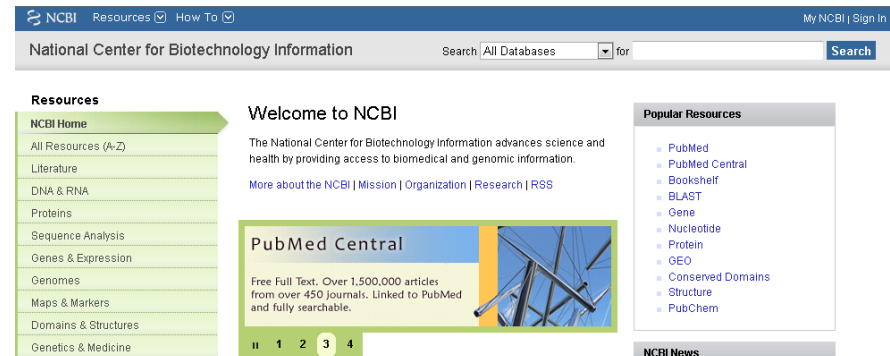
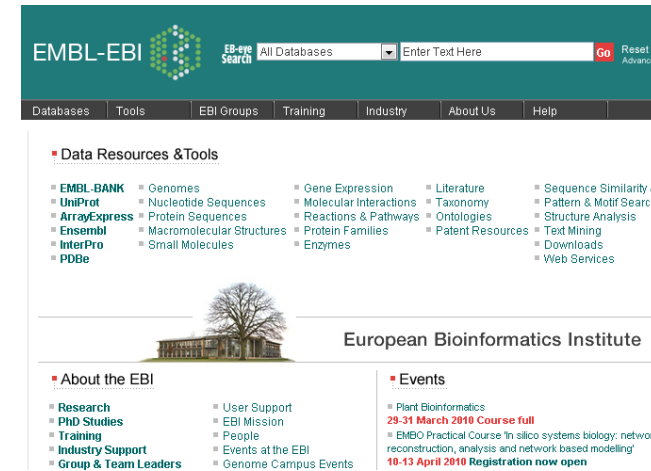
PARTNER VIEWS: SURF



Anwar Osseyran, PAC member for SURF: "SURF is a great supporter of life sciences, and believes in good

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.

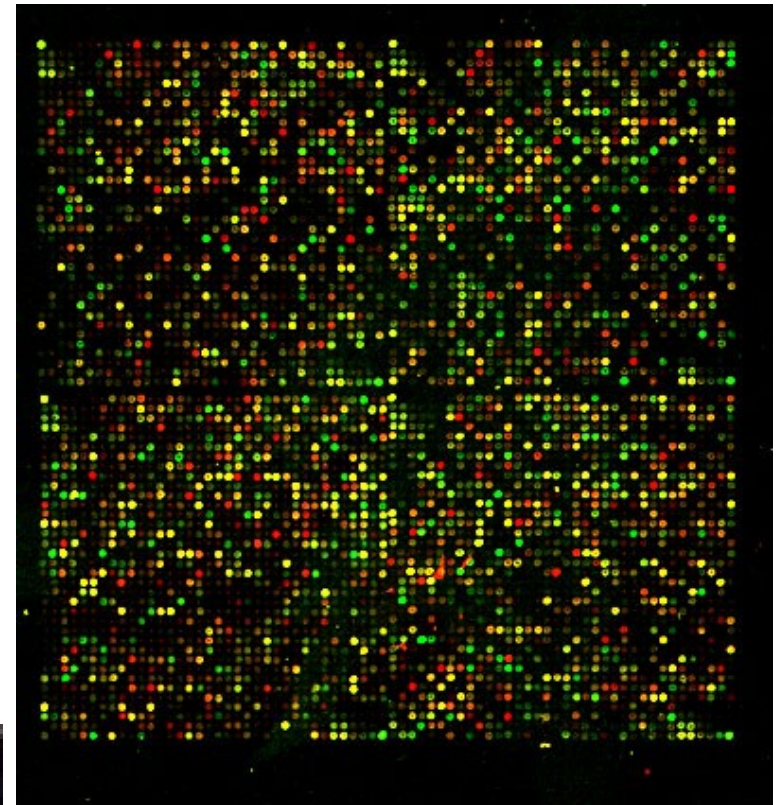
OR

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

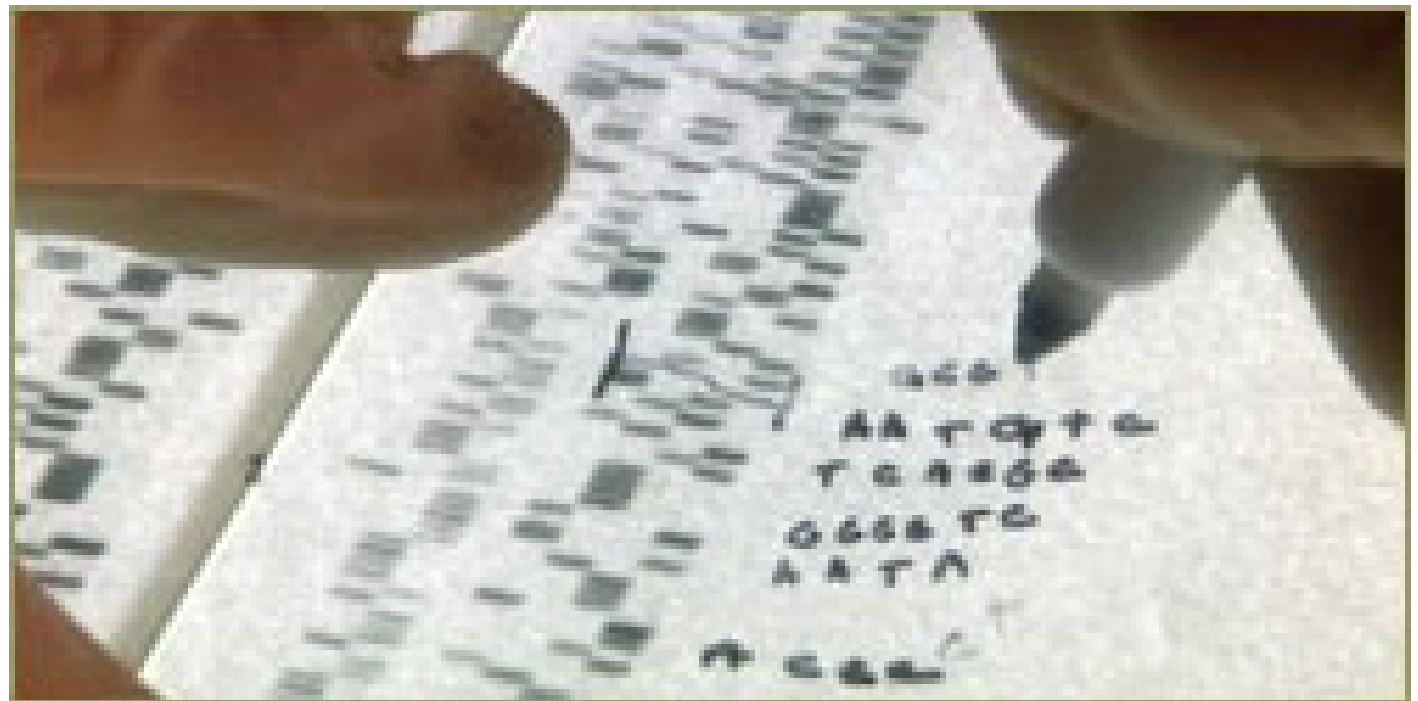
DNA sequencing

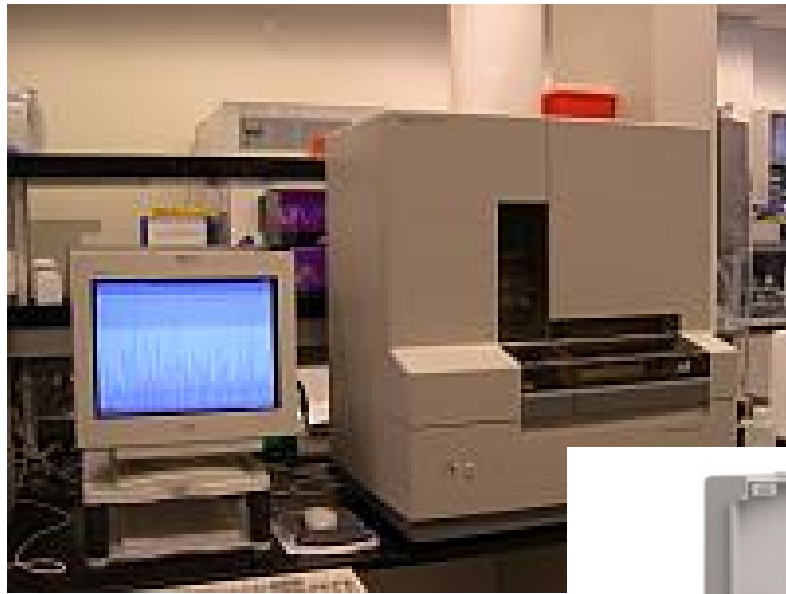




Sanger sequencing

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

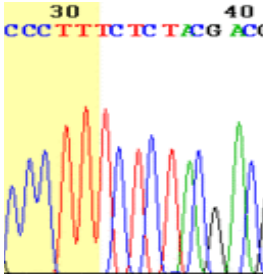
- 1990-2003
- 3 billion US dollar

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

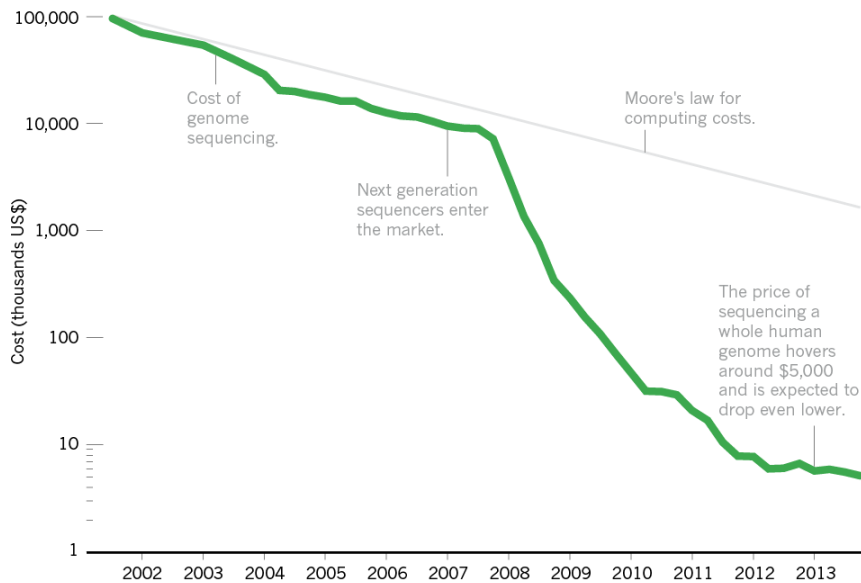
1.5 TB data

Year: 2008 – now

Data size challenges

Falling fast

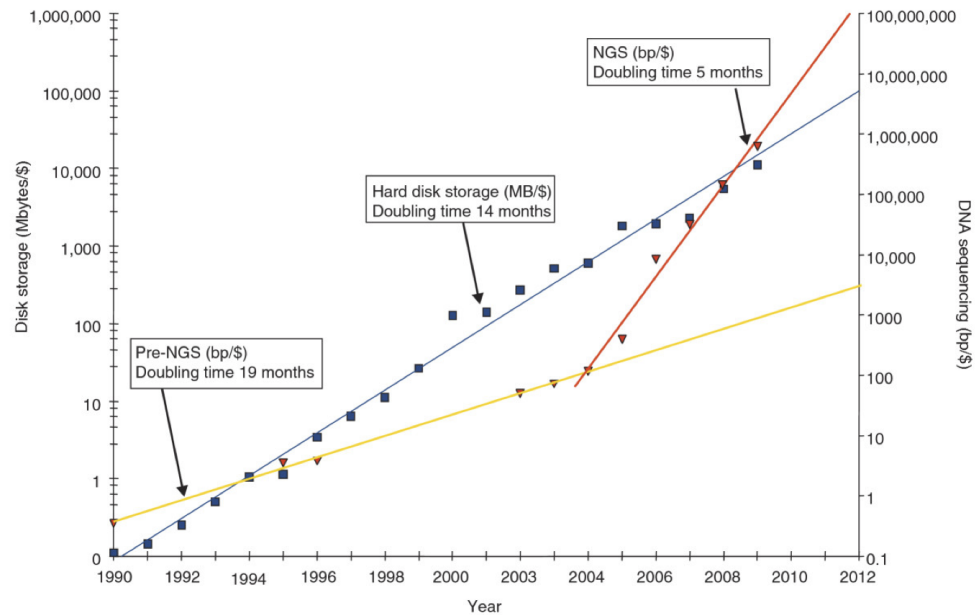
In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore's law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.



Hayden, Nature, 2014

COMPUTING
Sequencing rate is higher
than Moore's law

NextGen Sequencing a Game-Changer



Stein, Genome Biology, 2010

STORAGE
Sequencing costs lower
than data storage

Next Generation Genomics: World Map of High-throughput Sequencers

☒ Show all platforms ☐ Illumina GA2 ☐ Illumina HiSeq ☐ Illumina MiSeq ☐ Ion Torrent ☐ PacBio ☐ Polonator ☐ Roche/454 ☐ SOLiD ☐ Service Provider

Genome projects

- Human genome project (1 genome)
- Exome sequencing (~10 individuals)
- Genome of the Netherlands (770 individuals)
- 1000 genome project (1000 individuals)
- 10K UK project (10,000 individuals)
 - Upgraded to 100K genomes

... many centers have one or more high throughput sequencers



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

“oh, shit, that's 320TB!”

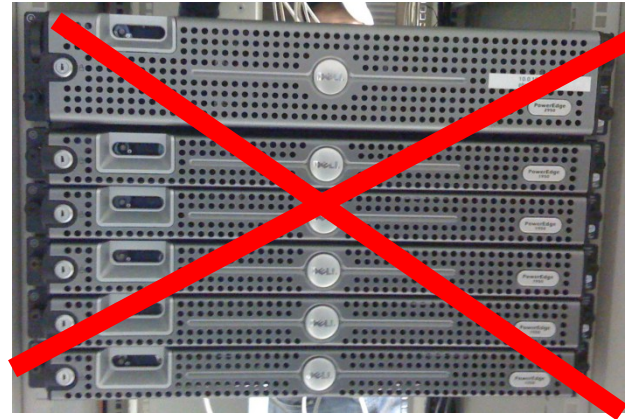
Tony Cox, The Guardian, Feb 28 2008

Sign @ Wellcome-Sanger, Cambridge, UK

Existing hardware



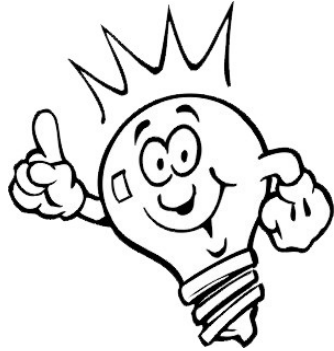
PC



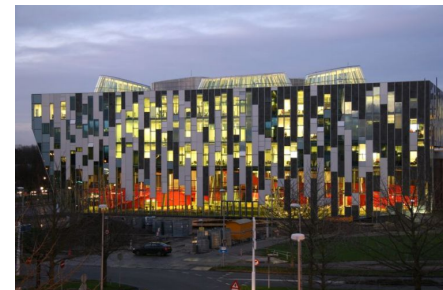
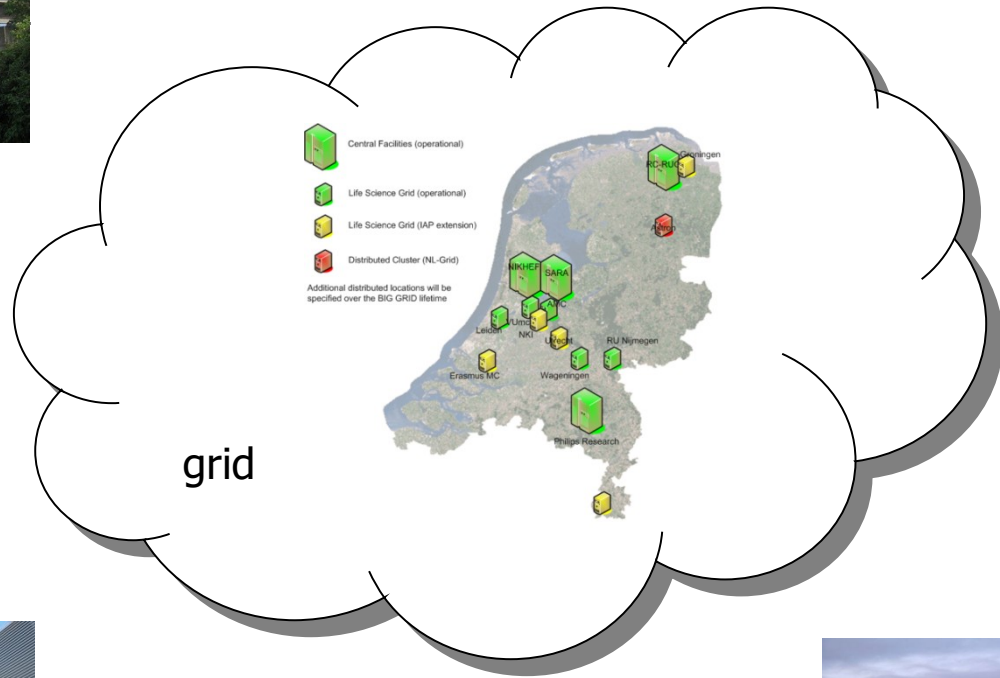
Small cluster



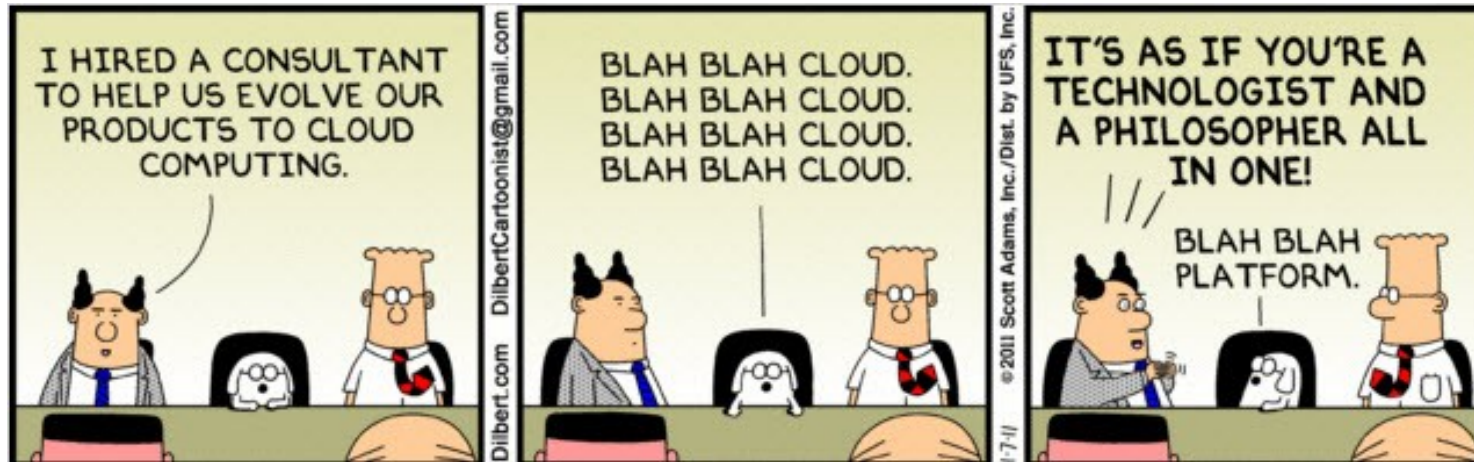
Buy a bigger cluster (centralized model)



Dutch life science grid



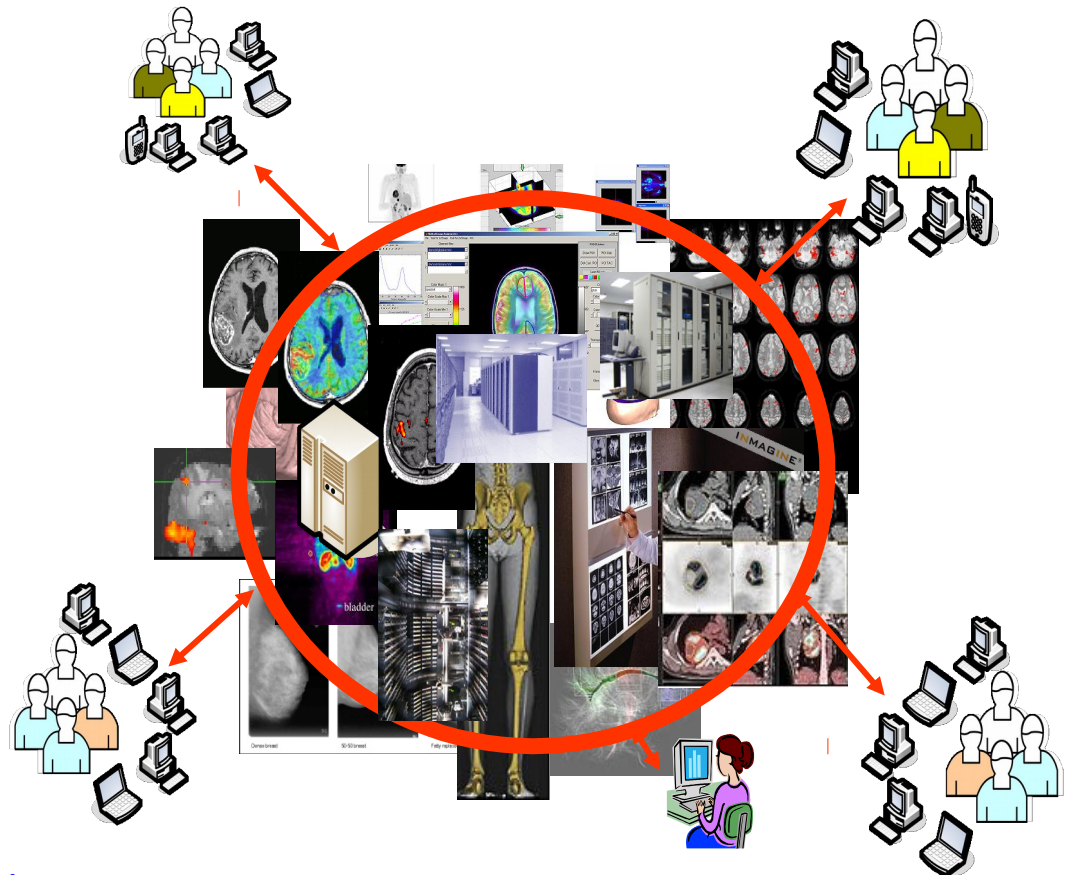
Cloud computing



Not only hardware

- Software
- Peopleware
- Sharing

e-Science

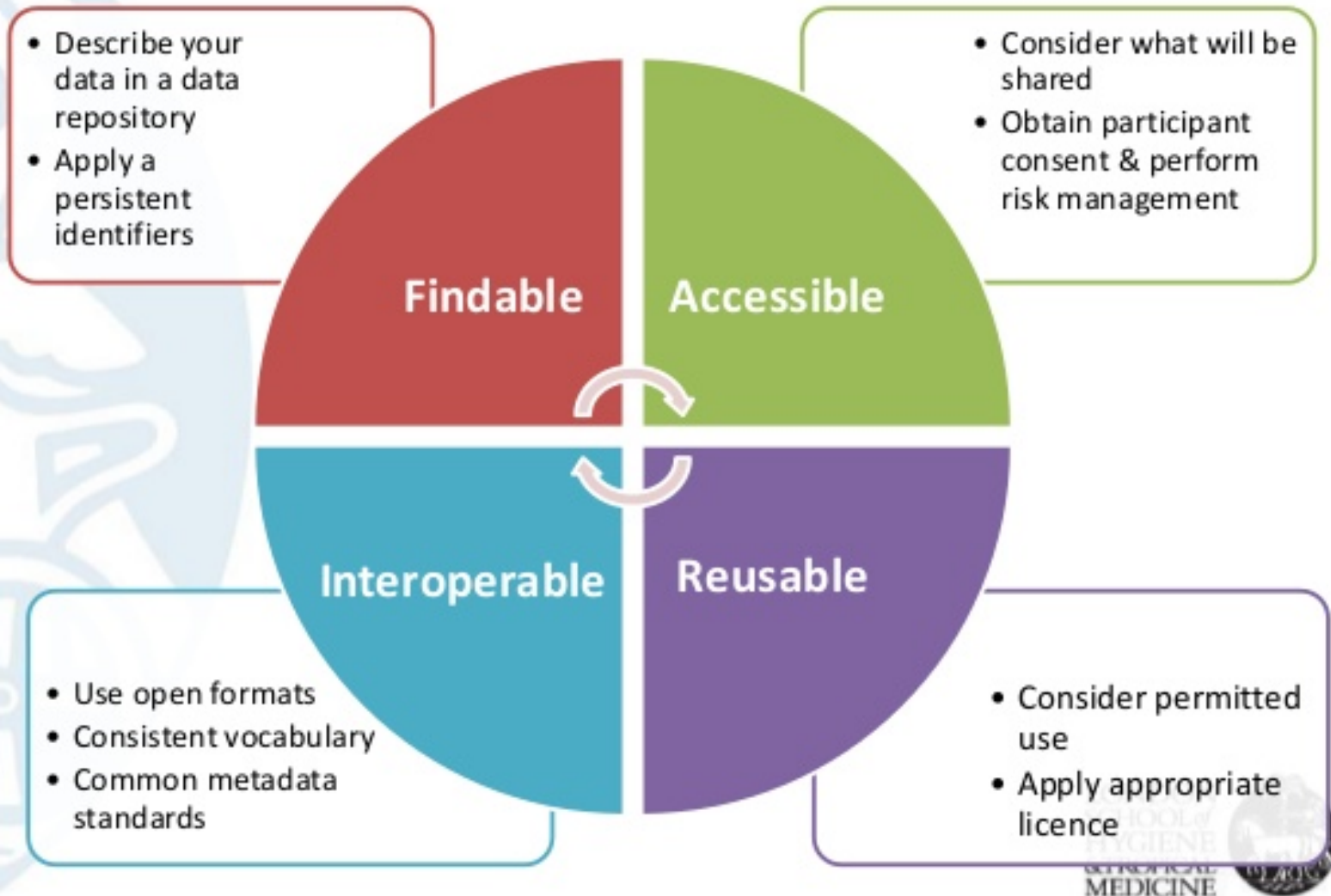


More info:

<http://www.egi.eu/>

<http://www.ebioscience.amc.nl/>

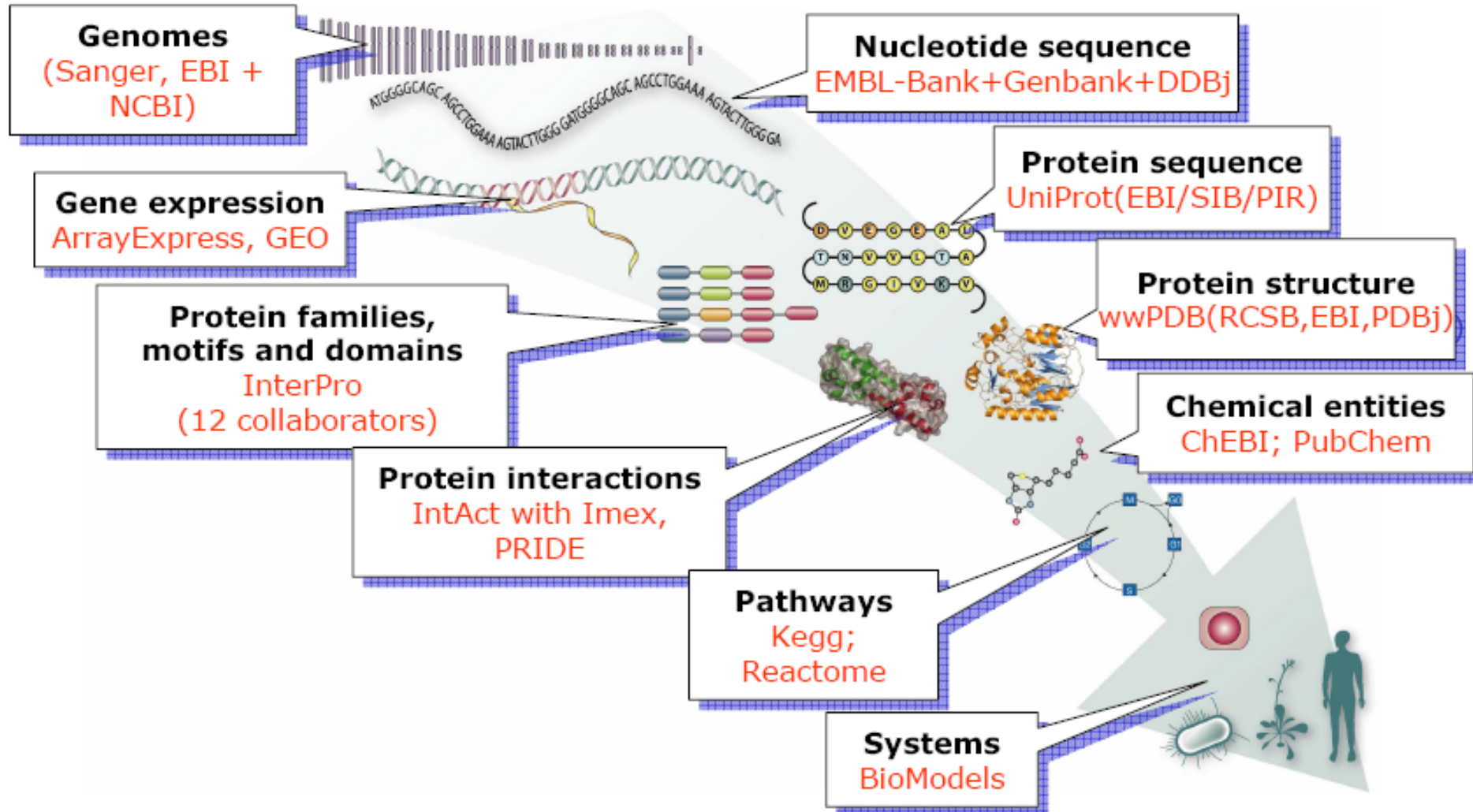
FAIR data



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.

The screenshot displays the Nucleic Acids Research journal website. At the top, the 'OXFORD JOURNALS' logo is on the left, and 'CONTACT US', 'MY BASKET', and 'MY ACCOUNT' links are on the right. A large maroon banner features the journal title 'Nucleic Acids Research' in white. Below this, a navigation bar includes 'ABOUT THIS JOURNAL', 'CONTACT THIS JOURNAL', 'SUBSCRIPTIONS', 'CURRENT ISSUE', 'ARCHIVE', and 'SEARCH'. The main content area shows the breadcrumb path: 'Oxford Journals > Life Sciences > Nucleic Acids Research > Volume 37, Database issue'. A horizontal bar contains four icons with labels: 'Isolation', 'Expression Analysis', 'Localization', and 'Functional Analysis'. Below this, a red text prompt says 'Receive this page by email each issue: [Sign up for eTOCs]'. The 'Contents' section for 'Volume 37, Database issue, January 2009' includes links for '[Index by Author]', 'Articles', and 'Front-Matter/Back-Matter'. On the left, a thumbnail of the journal cover is shown, featuring a molecular structure diagram and the text 'New Open Access'.

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)

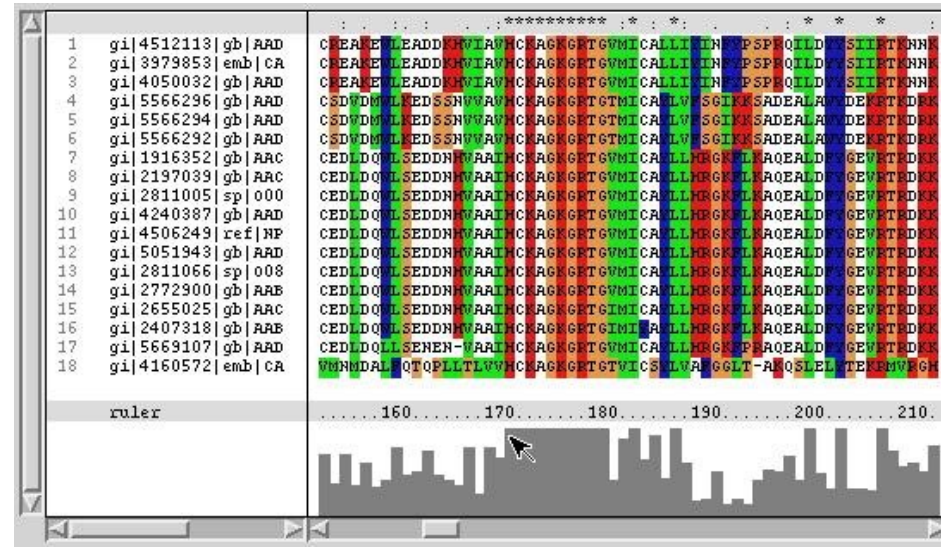
SNPs and mutations (variations)

Sequence analysis

Pairwise alignment: in-exact matching of 2 sequences

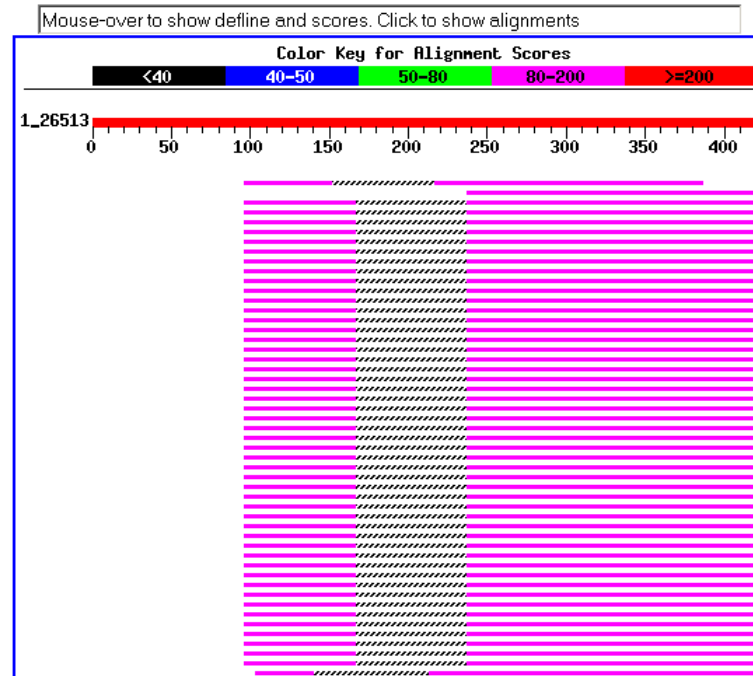
```
CTCCTGAGGCCAAATCTGTCA GTCCATCCTGGCTGAGTCCTCGCAGTCCCGGCAGATCTTGAAGAAAAGA
||||| ||| ||| |||| | |||| | | ||| | | ||| ||| ||| |||
CTCCTGAGGAAAACCTGCCAGTCAGTTGTGGCTAATCCTCAGCAGCACTCAGC---TCTGGAGGAAGAGA
```

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



Blast output

Distribution of 196 Blast Hits on the Query Sequence



Sequences producing significant alignments:

Score E
(bits) Value

gi 28804810 dbj AB104818.1	Pipistrellus abramus HBA mRNA f...	115	1e-22
gi 49420 emb X57029.1 CAGLOB.INM	M.auratus mRNA for alpha gl...	92	1e-15
gi 40225899 gb BC032122.2	Homo sapiens hemoglobin, alpha 2...	84	3e-13
gi 30047787 gb BC050661.1	Homo sapiens hemoglobin, alpha 2...	84	3e-13

G U E
G U
G U E

Done

Adblock

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 tggacgacctgcccaccgccctgtccgccctgagcgcacctgcacgcccacaagctgcgtg 307

Query: 278 ttgacccgggttaacttcaaactgctgtcccactgcctgctgggttacctggctgctcacc 337
|
Sbjct: 308 ttgaccccgctcaacttcaagctcctgagccactgcctgctgggtgacctggcttgccacc 367

Query: 338 acccggctgaattcaccccggtgttccactccgacctggacaaaattcctg 387
|||
Sbjct: 368 acccgcggaattcaccccgccgtgcacgctccctggacaagttcctg 417

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

Query: 97 ttccctgtgcttccccaccacacaaaacctacttccccgactt---cgacctgtcccacgg 152
|||||
Sbjct: 124 ttccctgagcttccccaccacacaaagacctacttccccacttctccgacctgtcccacgg 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

The screenshot displays the Galaxy web interface. At the top, a dark navigation bar contains the 'Galaxy' logo and links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 0%' indicator is on the right. Below the navigation bar, the main content area is divided into three sections. The left sidebar, titled 'Tools', lists various tool categories such as 'Get Data', 'Text Manipulation', 'FASTA manipulation', 'Fetch Sequences', 'Get Genomic Scores', 'Statistics', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Motif Tools', 'Multiple Alignments', 'Metagenomic analyses', 'Human Genome Variation', 'Genome Diversity', 'EMBOSS', 'NGS TOOLBOX BETA', 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: SAM Tools', 'NGS: Indel Analysis', 'NGS: Peak Calling', 'NGS: RNA Analysis', and 'NGS: Picard (beta)'. The central area features a large box titled 'Galaxy 101' with the subtitle 'Start small' and the text 'The very first tutorial you need'. Below this is a 'Live Quickies' section with five cards: 'Illumina mapping: Paired Ends', 'Basic fastQ manipulation:', 'Advanced fastQ manipulation:', '454 Mapping: Single End', and 'Uploading Data using FTP'. A paragraph of text describes Galaxy as an open, web-based platform for data intensive biomedical research. Below this is a 'Galaxy build' string and a section for social media updates from the 'galaxyproject' Twitter account. The right sidebar, titled 'History', shows a list of analyses. The top analysis is '22: Extract Genomic DNA on data 21', which has 2,252 sequences in FASTA format from the hg19 database. Below it is '21: Convert Genomic Intervals To BED on data 20', which has 2,252 regions in BED format from the hg19 database. The bottom analysis is '20: test.txt', which has 2,253 regions in interval format from the hg19 database. A table of genomic intervals is visible for the '21: Convert Genomic Intervals To BED on data 20' analysis.

Galaxy 101
Start small
The very first tutorial you need

Live Quickies

- Illumina mapping: Paired Ends (Galactic quickie # 12)
- Basic fastQ manipulation: (Galactic quickie # 13)
- Advanced fastQ manipulation: (Galactic quickie # 14)
- 454 Mapping: Single End (Galactic quickie # 15)
- Uploading Data using FTP (Galactic quickie # 17)

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy build: \$Rev 6801:40f1816d6857\$

galaxyproject

GigaScience Reading in JDC by @sorchan_ni: Making Data a 1st Class Scientific Output: Data Citation & Publication by NERC's Env... jdc.net/index.php/jdc... 12 hours ago · reply · retweet · favorite

galaxyproject Complete Genomics screencast (bit.ly/z2vFdX) tweeted earlier appears to only work in Adobe Reader. yesterday · reply · retweet · favorite

galaxyproject Compiling community written Howto's & Tutorials feat. Galaxy. Please add or send others: bit.ly/yEQ06l #usegalaxy yesterday · reply · retweet · favorite

History

PrimerDesign 4.2 Mb

22: Extract Genomic DNA on data 21 2,252 sequences format: fasta, database: hg19

```
>hg19_chr1_69698_70098_+
TTTTGTTCTTCTAATCATCTCATACATATCATC
ATCGCCCTTTAGATAAGCTGTCGAAAGCTCTGTC
ATTACAGTAGTCTTTTGTCTCTTGGACCATGTG
GCCATTCCCCATCAAGTCATTAGATAAATTCCTT
TGATCACCCCTCTCTTGAACCCCAATTATACAC
```

21: Convert Genomic Intervals To BED on data 20 2,252 regions format: bed, database: hg19
Info: 2252 regions converted to BED.
display at UCSC main
view in GeneTrack
display at Ensembl Current
display at RViewer main

1.Chrom	2.Start	3.End	4.Name
chr1	69698	70098	NM_0010054
chr1	368150	368550	NM_0010052
chr1	6158262	6158762	NM_0011998
chr1	7896985	7897385	NM_016831:
chr1	8557369	8557769	NM_0010426
chr1	9613558	9613958	NM_032815:

20: test.txt 2,253 regions format: interval, database: hg19
Info: uploaded interval file

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



<https://links.bioinformatics.ca/>

Bioinformatics Links Directory



Search Directory

Change Text

Username:

Password:

Log in

Create new
Request new

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

Main Page

Citations

Acknowledgements

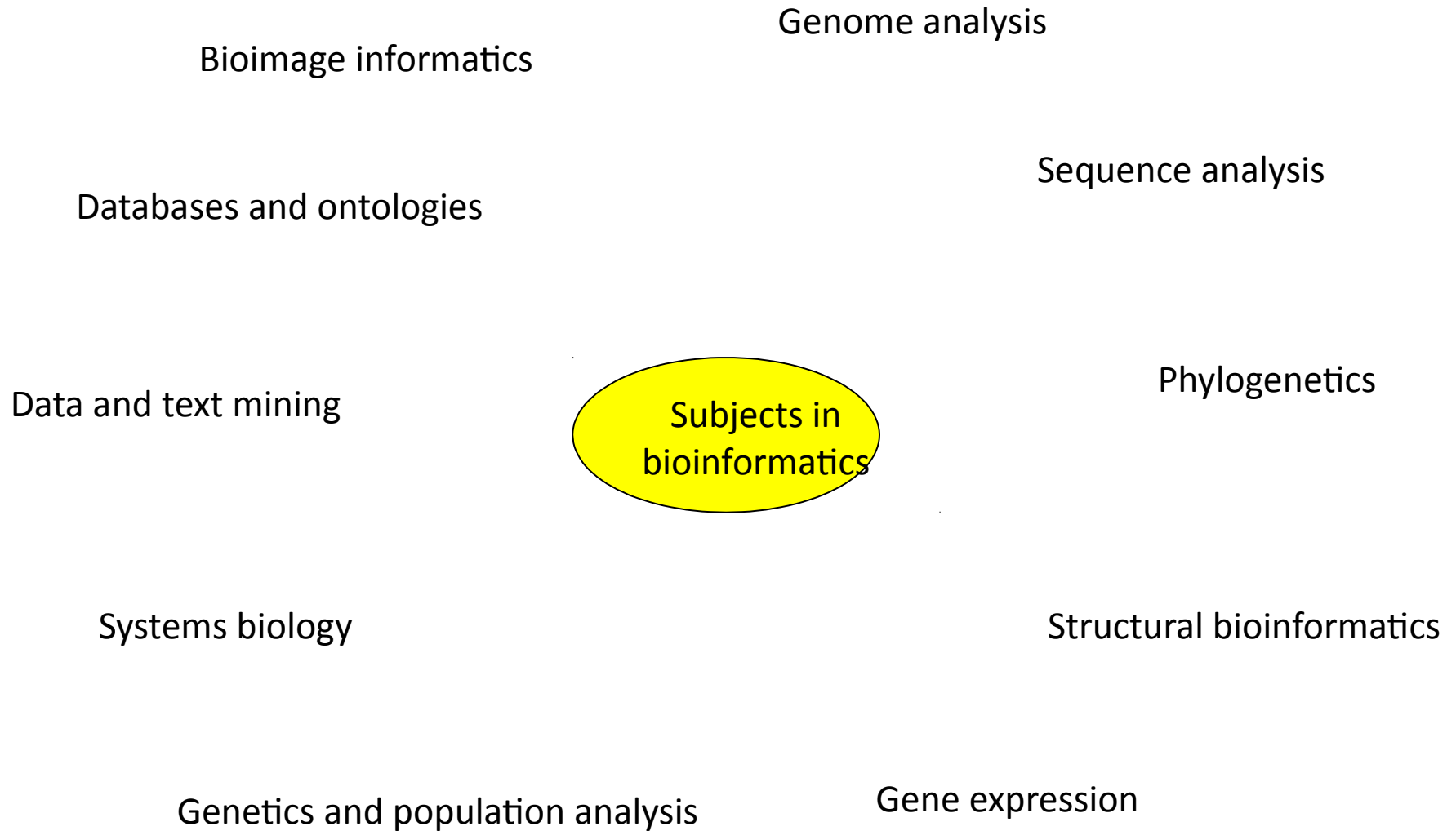
News

Suggest a link

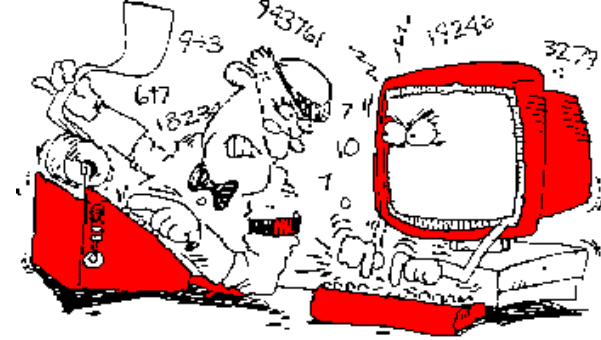
NAR Collaboration

RSS Feeds

Support



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/>