1. Introduction to Bioinformatics

Ricardo Gonzalo[1] ,Mireia Ferrer ,Álex Sánchez[1,2]

1 Unitat d’Estadística i Bioinformàtica (UEB) VHIR

2 Departament de Genètica Microbiologia i EStadística

We are drowning in information
and starved for knowledge

John Naisbitt

Who on efficient work is bent,
Must choose the fittest
instrument.

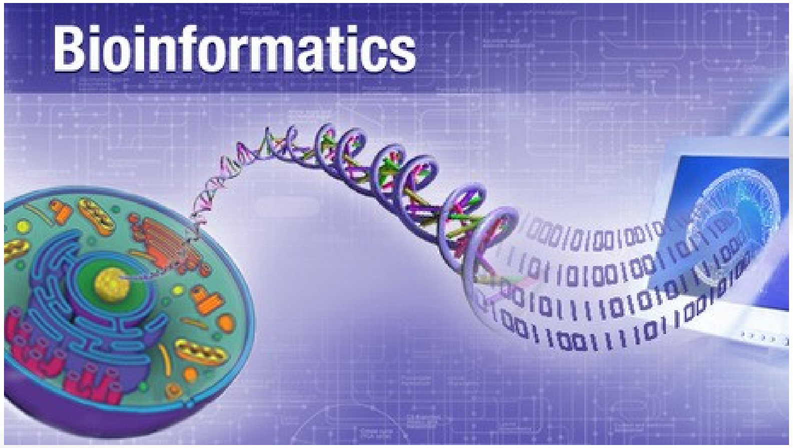
Goehthe (Fausto)

What is Bioinformatics?

A (first) definition

Bioinformatics is the application of computer technology to the management of biological information.

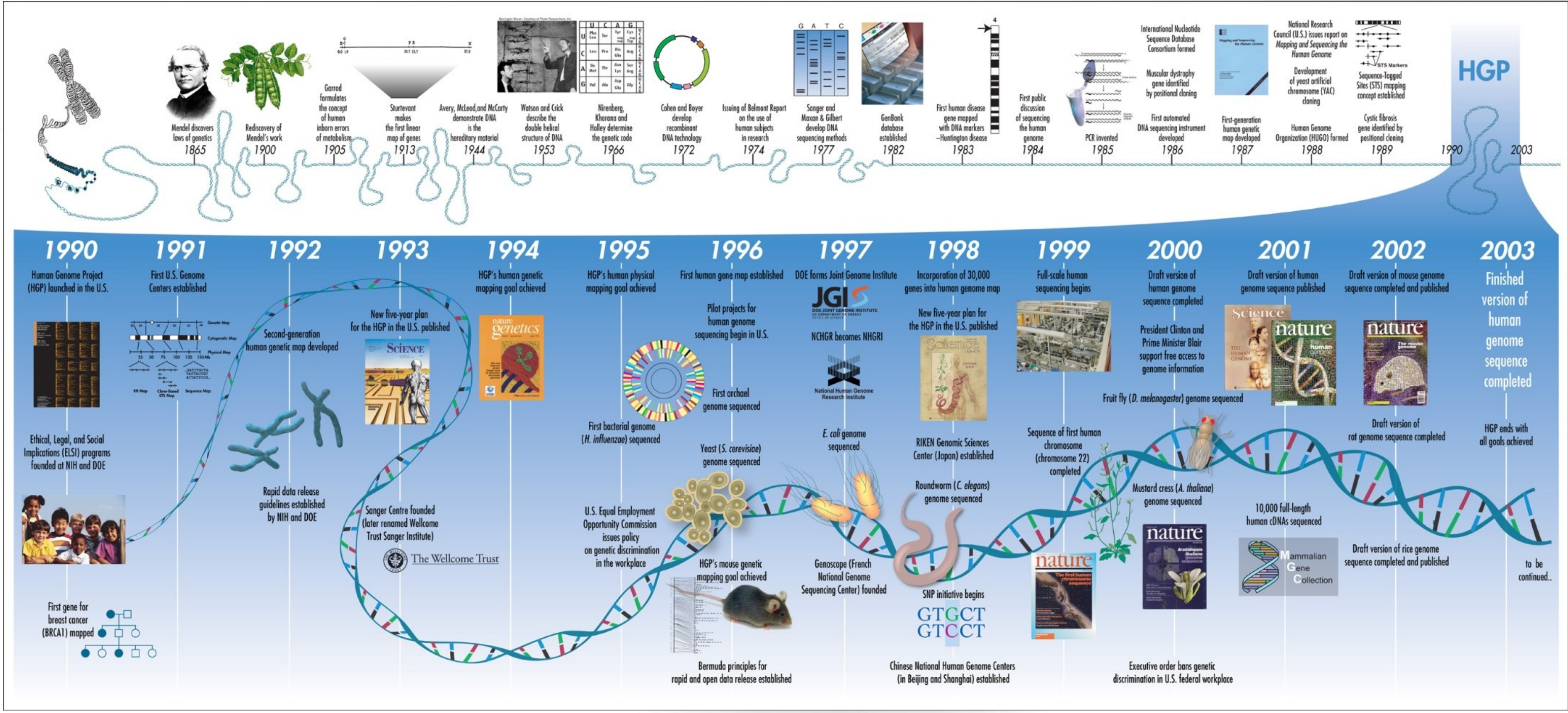
Computers are used to gather, store, analyze and integrate biological information.



A historical approach

- The term appeared in the 70’s
- It became popular/important with the development of the human genome project
- Bioinformatics is entering a big data era that will foster new possibilities.

The Human Genome Project



1865 Mendel discovers laws of genetics

1900 Rediscovery of Mendel's work

1905 Garrod formulates the concept of human inborn errors of metabolism

1913 Sturtevant makes the first linear map of genes

1944 Avery, McLeod,and McCarty demonstrate DNA is the hereditary material

1953 Watson and Crick describe the double helical structure of DNA

1966 Nirenberg, Khorana and Holley determine the genetic code

1972 Cohen and Boyer develop recombinant DNA technology

1974 Issuing of Belmont Report on the use of human subjects in research

1977 Sanger and Maxan & Gilbert develop DNA sequencing methods

1982 GenBank database established

1983 First human disease gene mapped with DNA markers --Huntington disease

1984 First public discussion of sequencing the human genome

1985 PCR invented

1986 International Nucleotide Sequence Database Consortium formed

Muscular dystrophy gene identified by positional cloning

First automated DNA sequencing instrument developed

1987 First-generation human genetic map developed

1988 National Research Council (U.S.) issues report on *Mapping and Sequencing the Human Genome*

Development of yeast artificial chromosome (YAC) cloning

Human Genome Organization (HUGO) formed

1989 Sequence-Tagged Sites (STS) mapping concept established

Cystic fibrosis gene identified by positional cloning

HGP 1990 2003

1990 Human Genome Project (HGP) launched in the U.S.

Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE

First gene for breast cancer (BRCA1) mapped

1991 First U.S. Genome Centers established

1992 Second-generation human genetic map developed

Rapid data release guidelines established by NIH and DOE

1993 New five-year plan for the HGP in the U.S. published

Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)

The Wellcome Trust

1994 HGP's human genetic mapping goal achieved

1995 HGP's human physical mapping goal achieved

First bacterial genome (*H. influenzae*) sequenced

U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace

1996 First human gene map established

Pilot projects for human genome sequencing begin in U.S.

First archael genome sequenced

Yeast (*S. cerevisiae*) genome sequenced

HGP's mouse genetic mapping goal achieved

Bermuda principles for rapid and open data release established

1997 DOE forms Joint Genome Institute

JGI DOE JOINT GENOME INSTITUTE

NCHGR becomes NHGRI

National Human Genome Research Institute

E. coli genome sequenced

Genoscope (French National Genome Sequencing Center) founded

1998 Incorporation of 30,000 genes into human genome map

New five-year plan for the HGP in the U.S. published

RIKEN Genomic Sciences Center (Japan) established

Roundworm (*C. elegans*) genome sequenced

SNP initiative begins

GTGCT
GTCCT

Chinese National Human Genome Centers (in Beijing and Shanghai) established

1999 Full-scale human sequencing begins

Sequence of first human chromosome (chromosome 22) completed

2000 Draft version of human genome sequence completed

President Clinton and Prime Minister Blair support free access to genome information

Fruit fly (*D. melanogaster*) genome sequenced

Mustard cress (*A. thaliana*) genome sequenced

Executive order bans genetic discrimination in U.S. federal workplace

2001 Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian Gene Collection

2002 Draft version of mouse genome sequence completed and published

Draft version of rat genome sequence completed

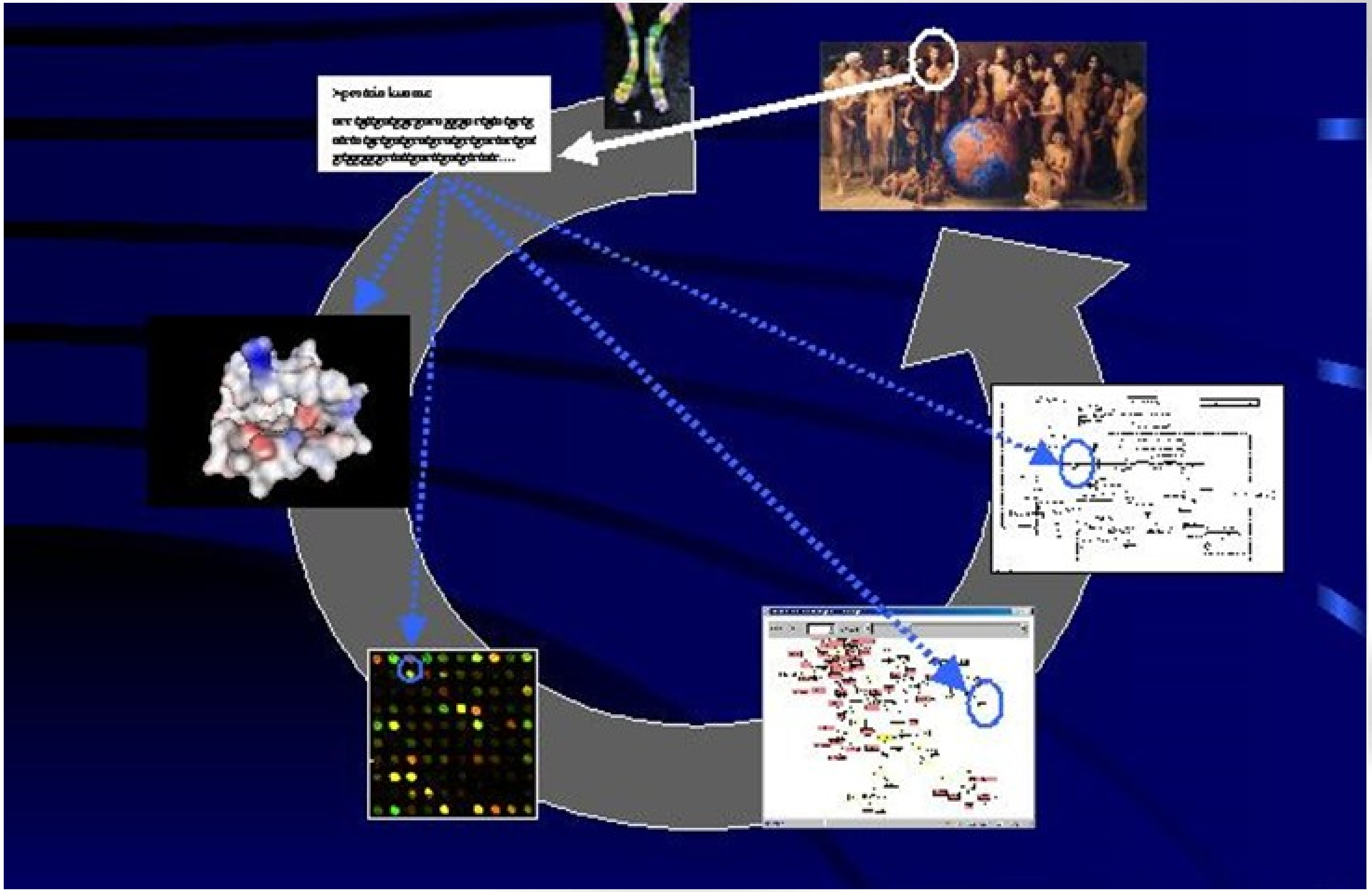
Draft version of rice genome sequence completed and published

2003 Finished version of human genome sequence completed

HGP ends with all goals achieved

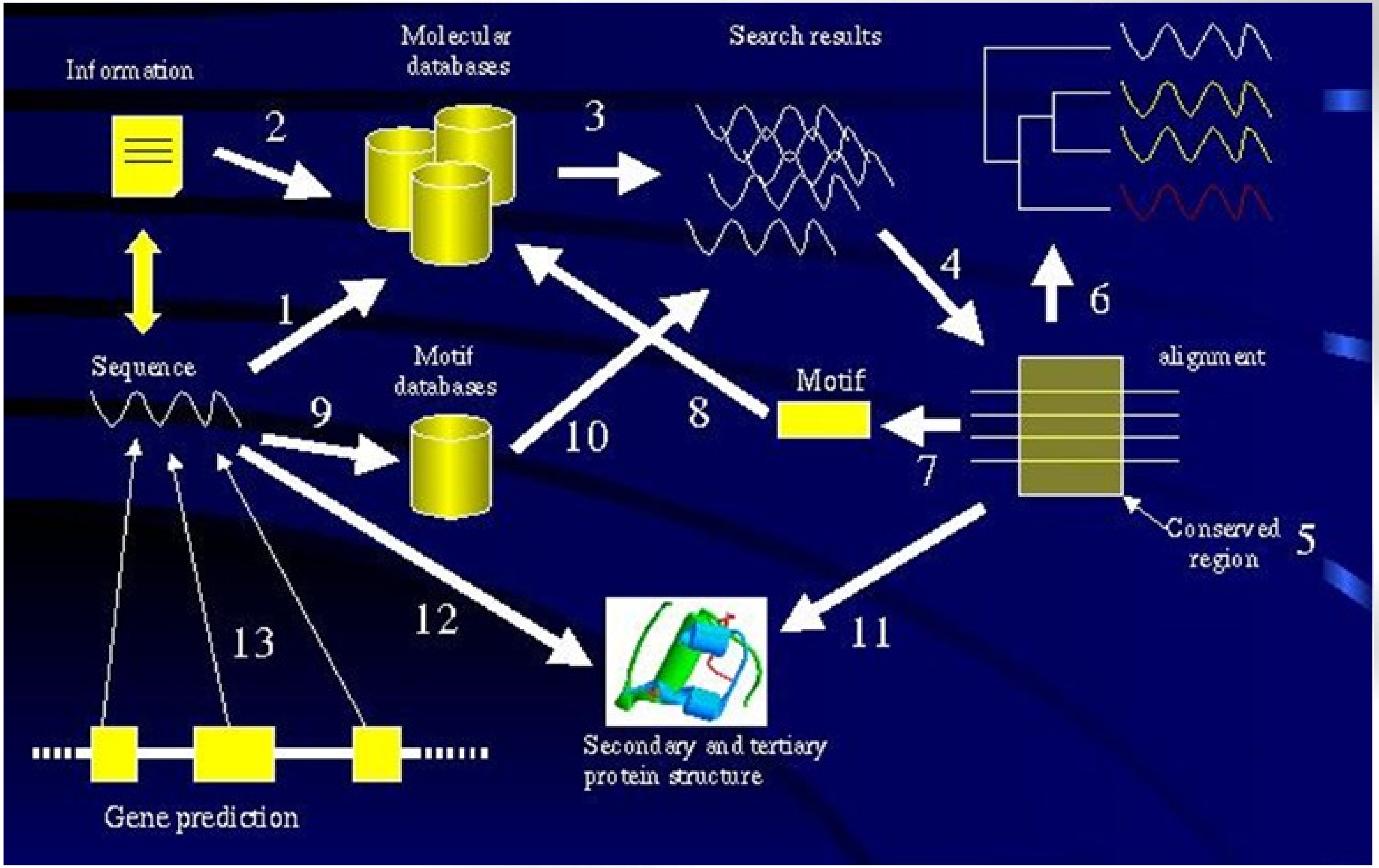
to be continued..

Pre genomics era vision in the lab



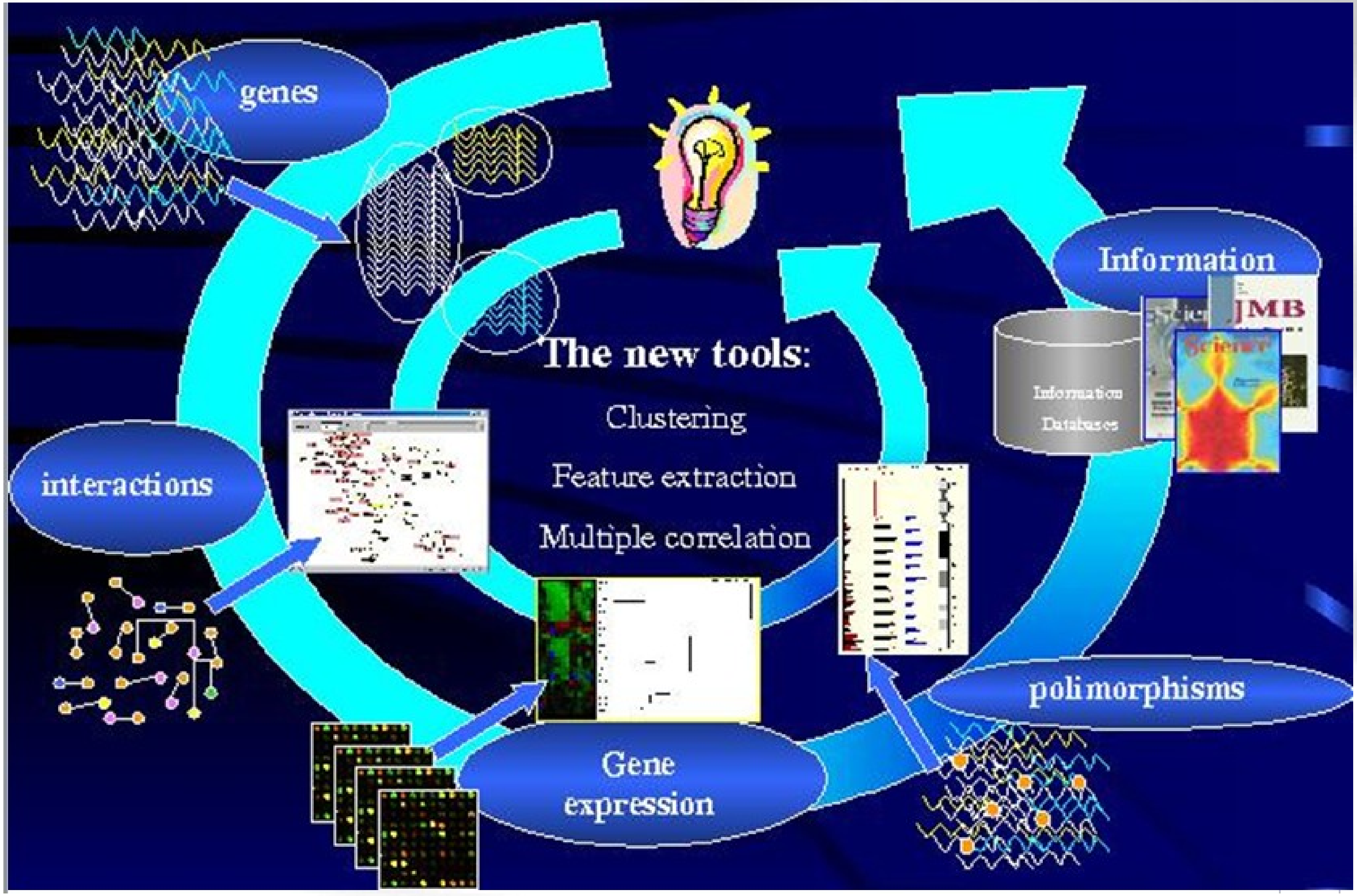
Adapted from a presentation by J. dopazo

Bioinformatic analysis



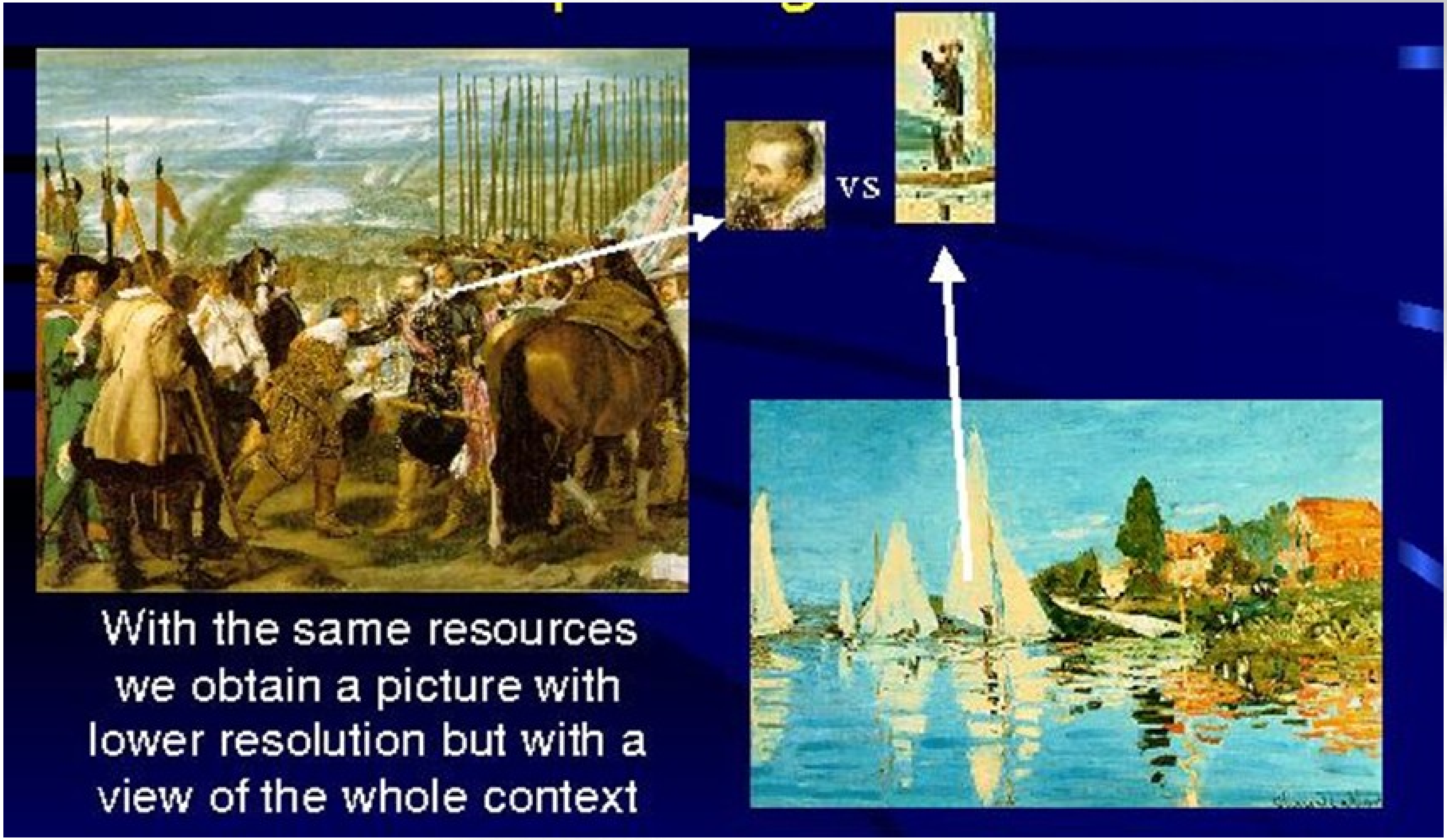
Adapted from a presentation by J. dopazo

Post-genomic vision



Adapted from a presentation by J. dopazo

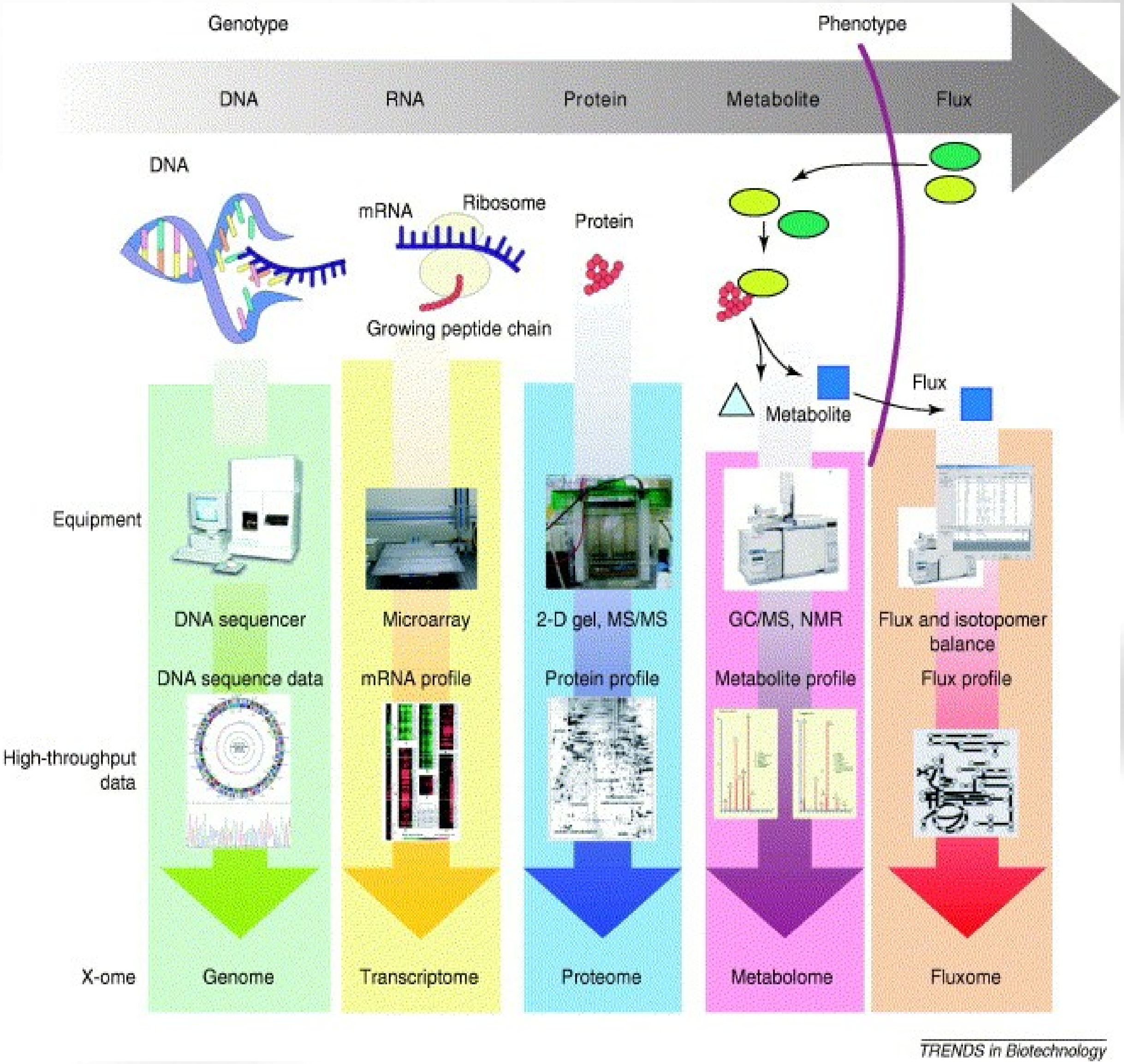
The (first) paradigm shift



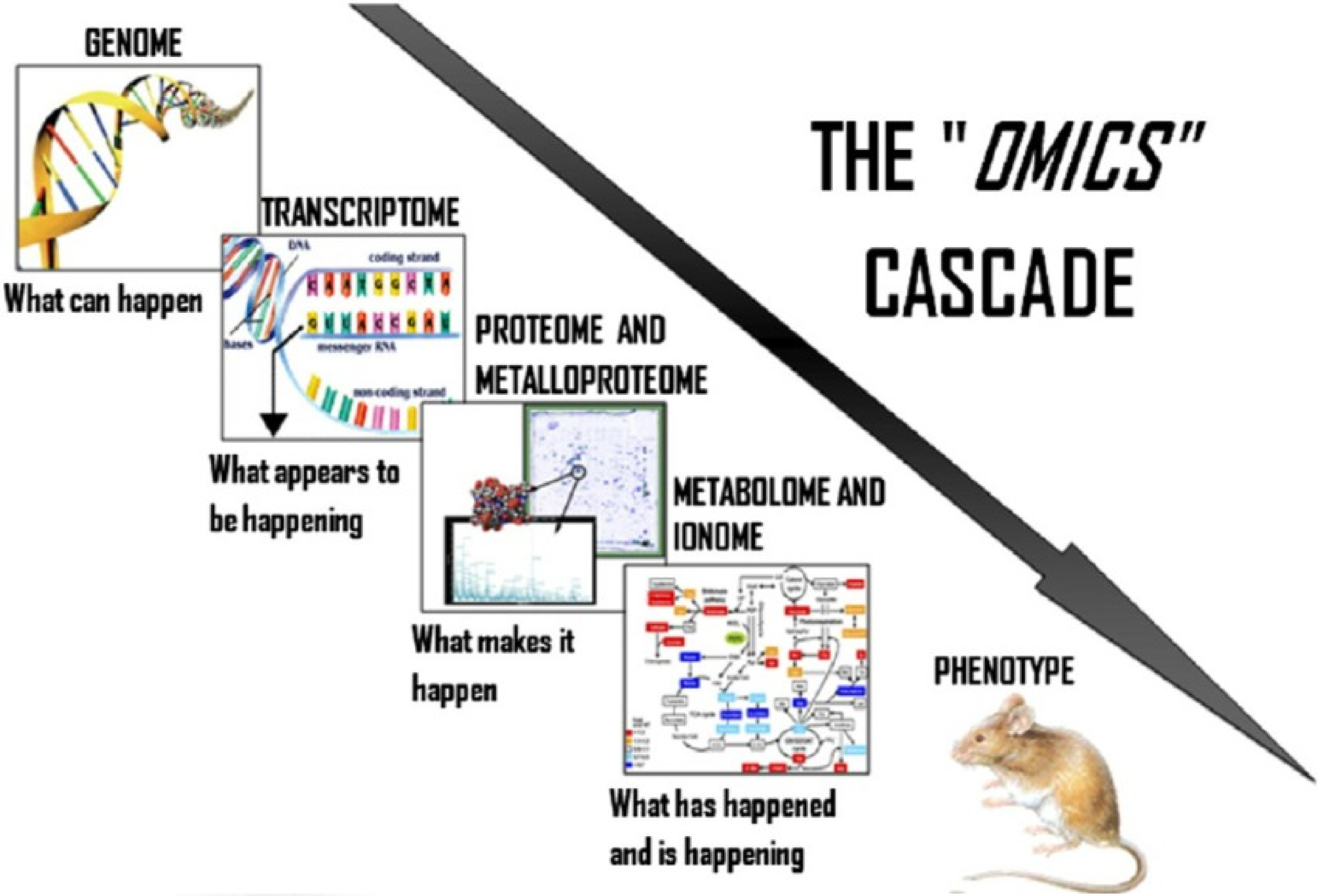
With the same resources we obtain a picture with lower resolution but with a view of the whole context

Adapted from a presentation by J. dopazo

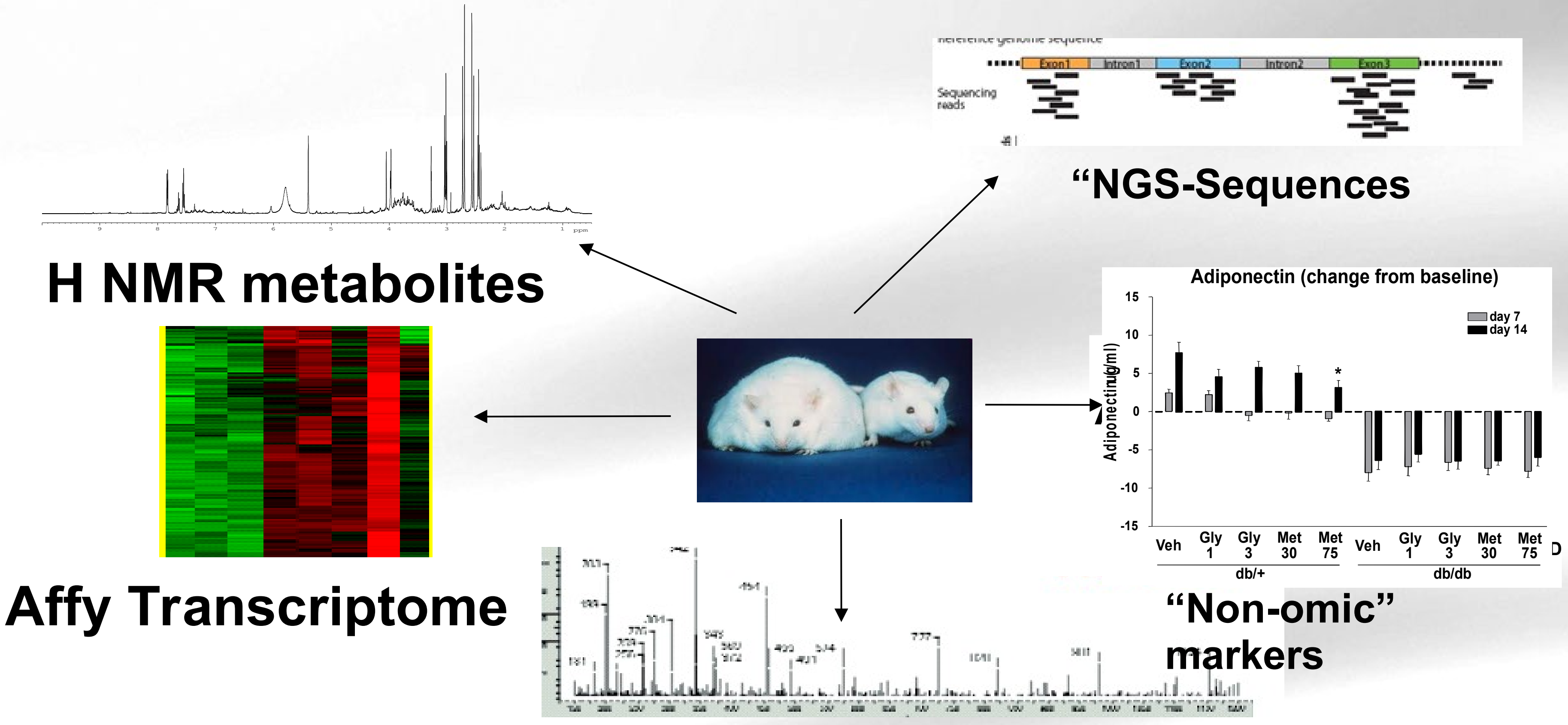
Omics technologies



The “Omics cascade”



New paradigm shift: ***Integromics***



H NMR metabolites

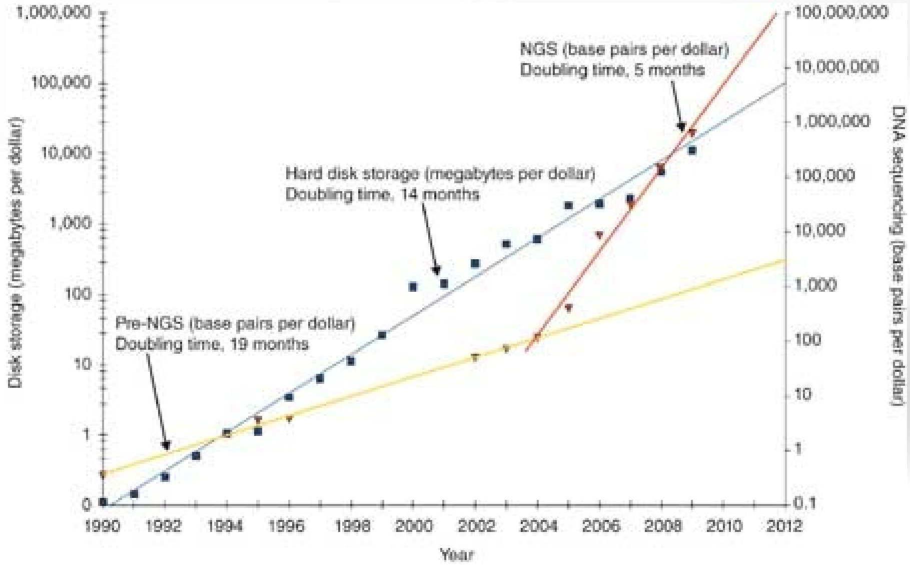
“NGS-Sequences

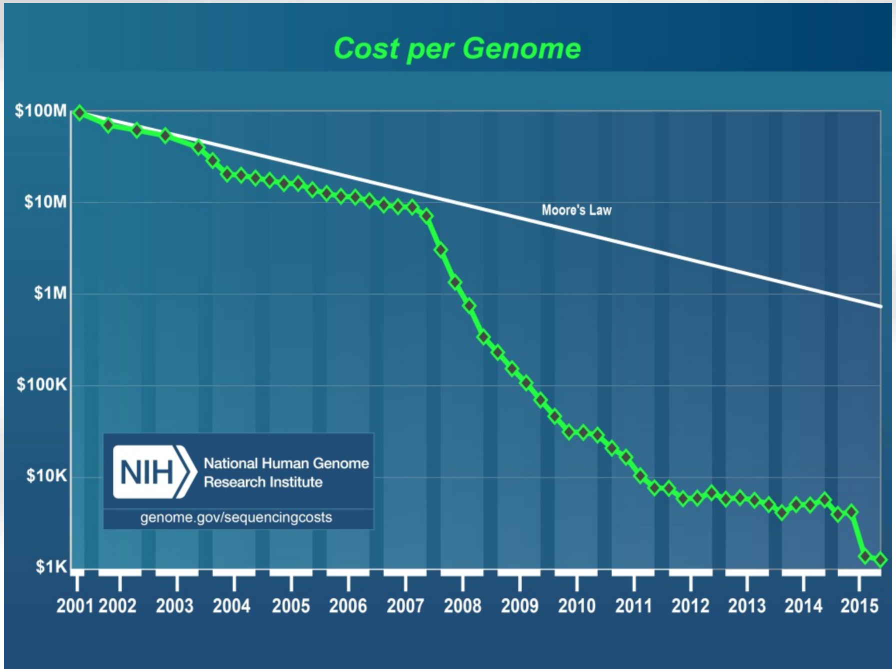
Affy Transcriptome

“Non-omic” markers

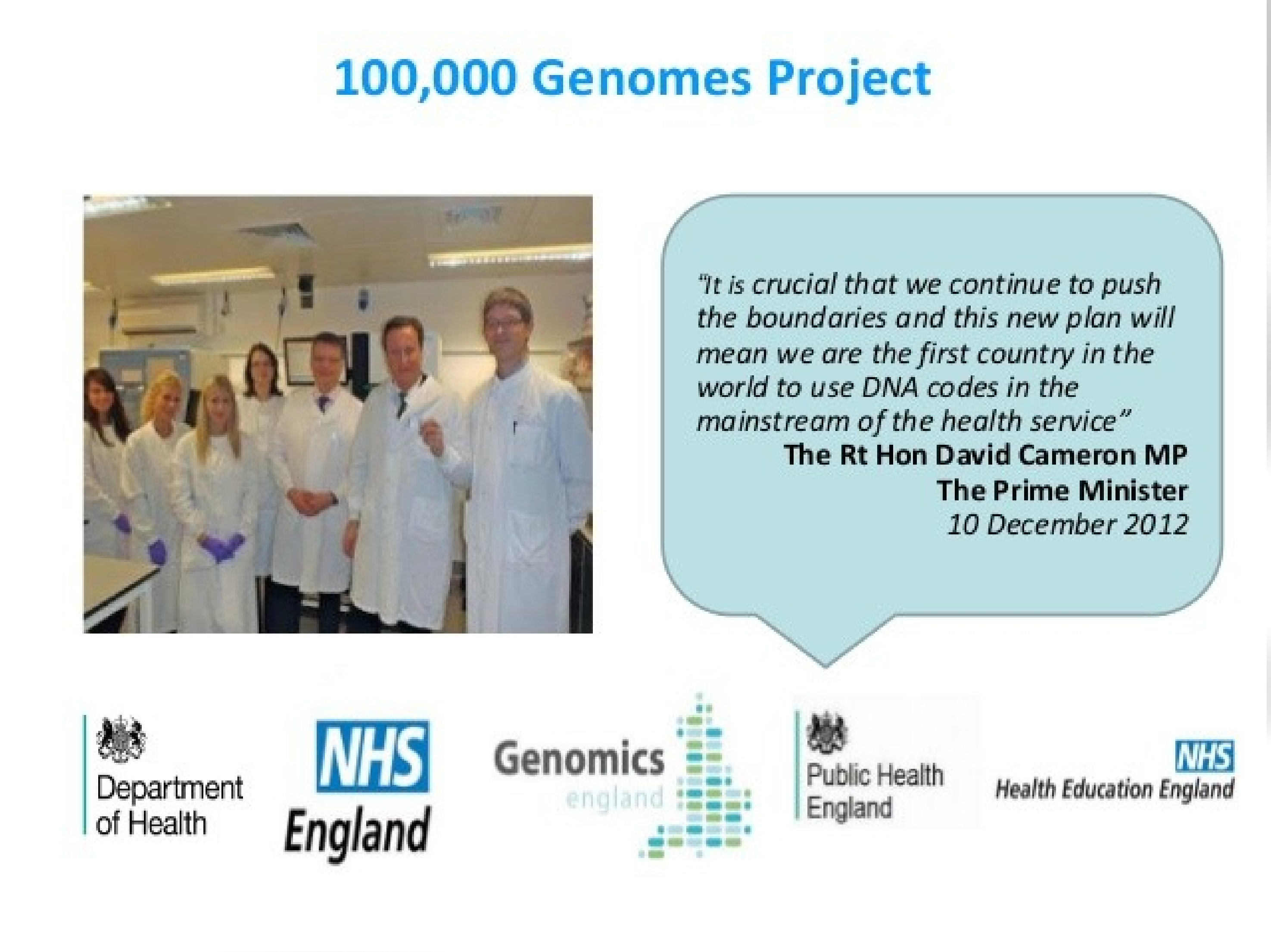
LC-MS proteomicss

“Nextgen” sequencing revolution





New paradigm shift: *Precision Medicine*



100,000 Genomes Project

"It is crucial that we continue to push the boundaries and this new plan will mean we are the first country in the world to use DNA codes in the mainstream of the health service"

The Rt Hon David Cameron MP
The Prime Minister
10 December 2012

Department of Health

NHS England

Genomics england

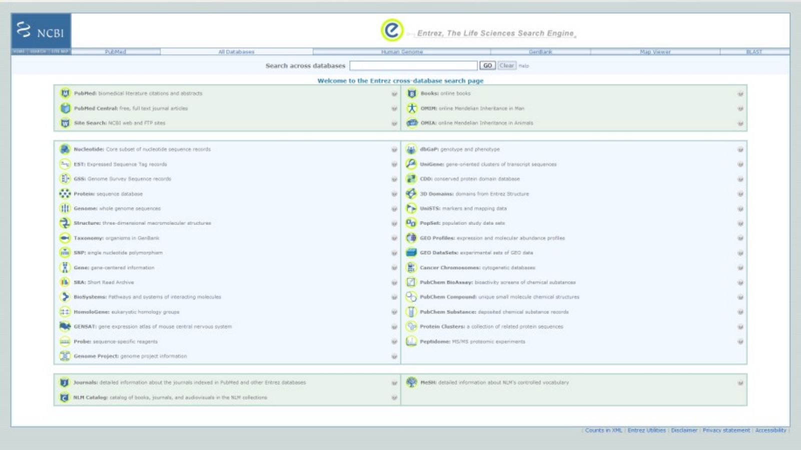
Public Health England

NHS Health Education England

Bioinformatics domains

- Information management
  - Databases, databanks
  - Algorithms and tools for database querying and searching
- Information Modelling
  - Protein structure characterization
- Analysis & Interpretation of results
  - Genome sequencing and analysis
  - Comparative genomics
  - Transcriptomics and gene expression
  - Metagenomics
  - Proteomics, metabolomics, ...
- Biological system modeling

Information management



Entrez, The Life Sciences Search Engine

Search across databases

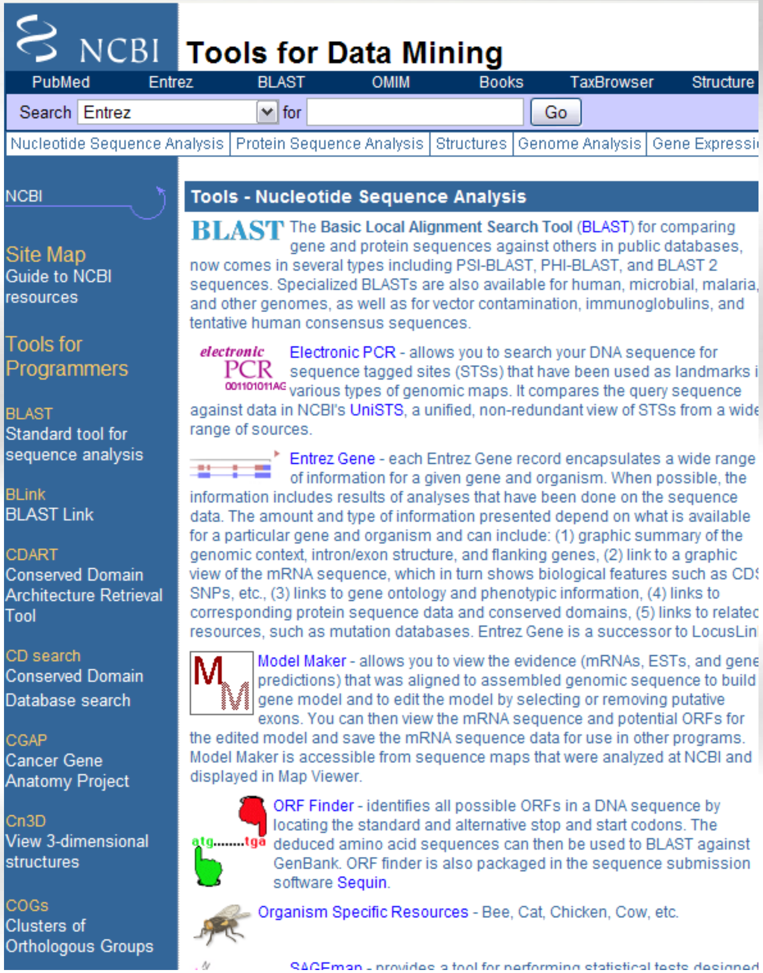
Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts
PubMed Central: free, full text journal articles
Site Search: NCBI web and FTP sites
Books: online books
OMIM: online Mendelian Inheritance in Man
OMIA: online Mendelian Inheritance in Animals

Nucleotide: Core subset of nucleotide sequence records
EST: Expressed Sequence Tag records
GSS: Genome Survey Sequence records
Protein: sequence database
Genome: whole genome sequences
Structure: three-dimensional macromolecular structures
Taxonomy: organisms in GenBank
SNP: single nucleotide polymorphism
Gene: gene-centered information
SRA: Short Read Archive
BioSystems: Pathways and systems of interacting molecules
HomoloGene: eukaryotic homology groups
GENSAT: gene expression atlas of mouse central nervous system
Probe: sequence-specific reagents
Genome Project: genome project information
dbGaP: genotype and phenotype
UniGene: gene-oriented clusters of transcript sequences
CDD: conserved protein domain database
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
PopSet: population study data sets
GEO Profiles: expression and molecular abundance profiles
GEO DataSets: experimental sets of GEO data
Cancer Chromosomes: cytogenetic databases
PubChem BioAssay: bioactivity screens of chemical substances
PubChem Compound: unique small molecule chemical structures
PubChem Substance: deposited chemical substance records
Protein Clusters: a collection of related protein sequences
Peptidome: MS/MS proteomic experiments

Journals: detailed information about the journals indexed in PubMed and other Entrez databases
NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections
MeSH: detailed information about NLM's controlled vocabulary

| Counts in XML | Entrez Utilities | Disclaimer | Privacy statement | Accessibility |



Tools for Data Mining

Site Map
Guide to NCBI resources

Tools for Programmers

BLAST
Standard tool for sequence analysis

BLink
BLAST Link

CDART
Conserved Domain Architecture Retrieval Tool

CD search
Conserved Domain Database search

CGAP
Cancer Gene Anatomy Project

Cn3D
View 3-dimensional structures

COGs
Clusters of Orthologous Groups

Tools - Nucleotide Sequence Analysis

The **Basic Local Alignment Search Tool** (BLAST) for comparing gene and protein sequences against others in public databases, now comes in several types including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, malaria, and other genomes, as well as for vector contamination, immunoglobulins, and tentative human consensus sequences.

Electronic PCR - allows you to search your DNA sequence for sequence tagged sites (STSs) that have been used as landmarks i various types of genomic maps. It compares the query sequence against data in NCBI's UniSTS, a unified, non-redundant view of STSs from a wide range of sources.

Entrez Gene - each Entrez Gene record encapsulates a wide range of information for a given gene and organism. When possible, the information includes results of analyses that have been done on the sequence data. The amount and type of information presented depend on what is available for a particular gene and organism and can include: (1) graphic summary of the genomic context, intron/exon structure, and flanking genes, (2) link to a graphic view of the mRNA sequence, which in turn shows biological features such as CDS SNPs, etc., (3) links to gene ontology and phenotypic information, (4) links to corresponding protein sequence data and conserved domains, (5) links to relatec resources, such as mutation databases. Entrez Gene is a successor to LocusLin

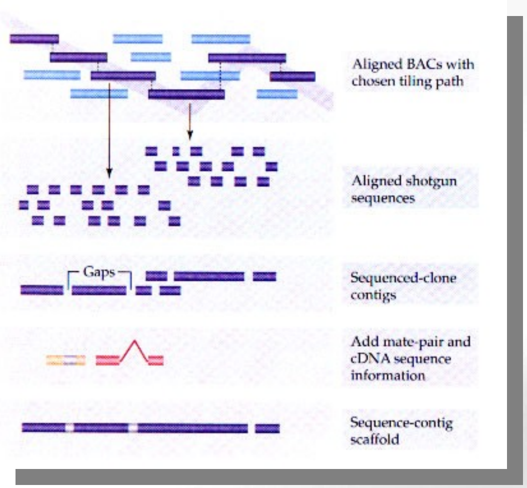
Model Maker - allows you to view the evidence (mRNAs, ESTs, and gene predictions) that was aligned to assembled genomic sequence to build gene model and to edit the model by selecting or removing putative exons. You can then view the mRNA sequence and potential ORFs for the edited model and save the mRNA sequence data for use in other programs. Model Maker is accessible from sequence maps that were analyzed at NCBI and displayed in Map Viewer.

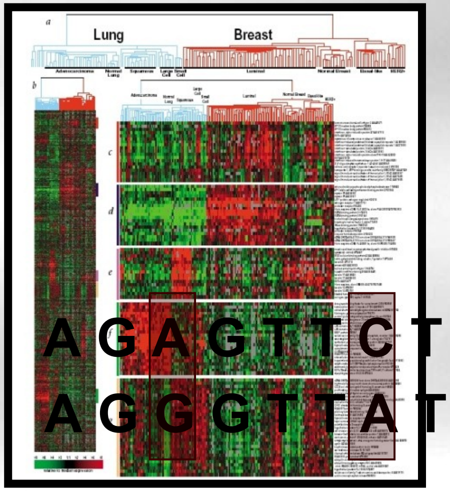
ORF Finder - identifies all possible ORFs in a DNA sequence by locating the standard and alternative stop and start codons. The deduced amino acid sequences can then be used to BLAST against GenBank. ORF finder is also packaged in the sequence submission software Sequin.

Organism Specific Resources - Bee, Cat, Chicken, Cow, etc.

SAGEmap - provides a tool for performing statistical tests designed

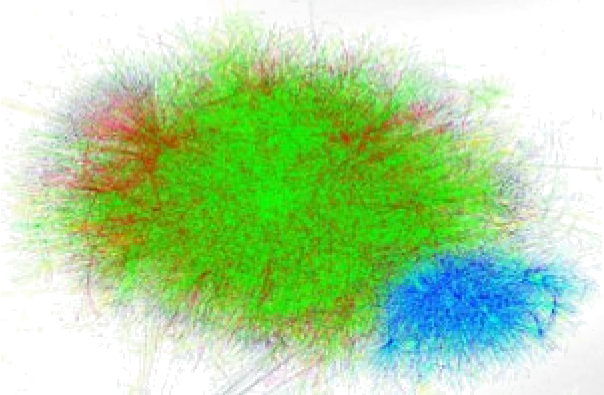
Analysis and interpretation


Aligned BACs with chosen tiling path
Aligned shotgun sequences
Gaps
Sequenced-clone contigs
Add mate-pair and cDNA sequence information
Sequence-contig scaffold

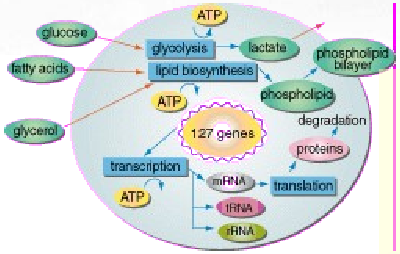


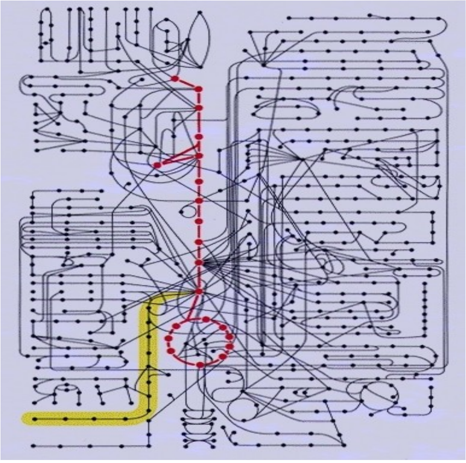
A G A G T T C T G C T C G

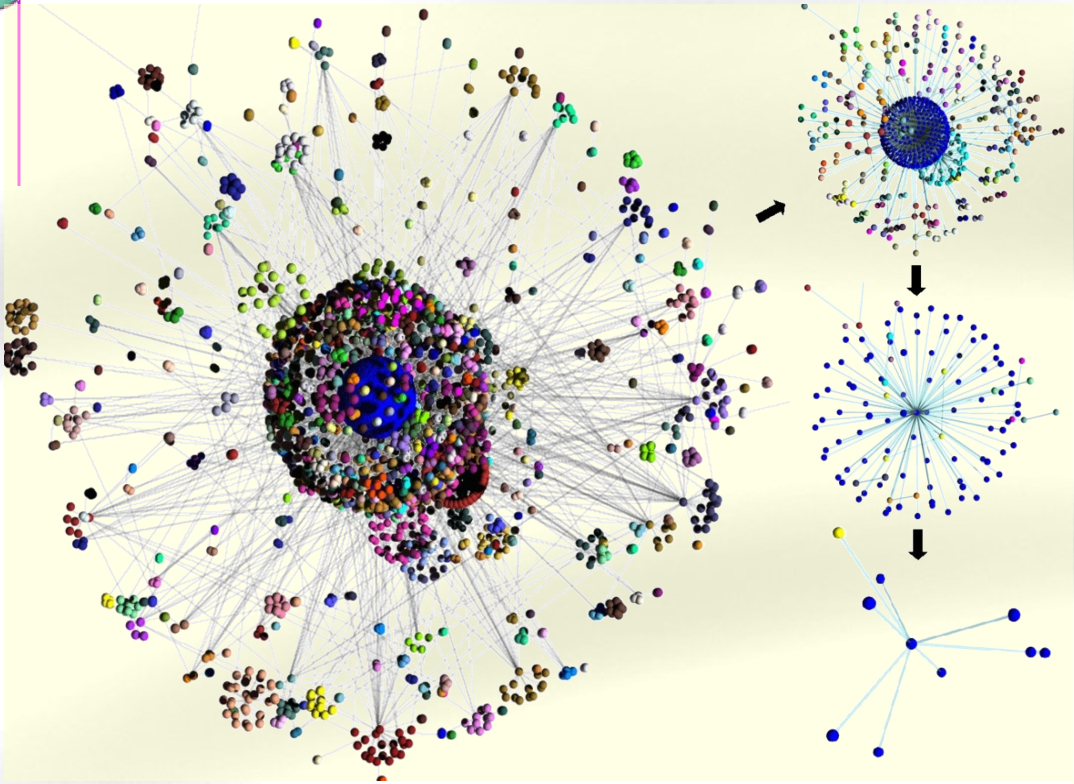
A G G G T T A T G C G C G



Biological system modeling







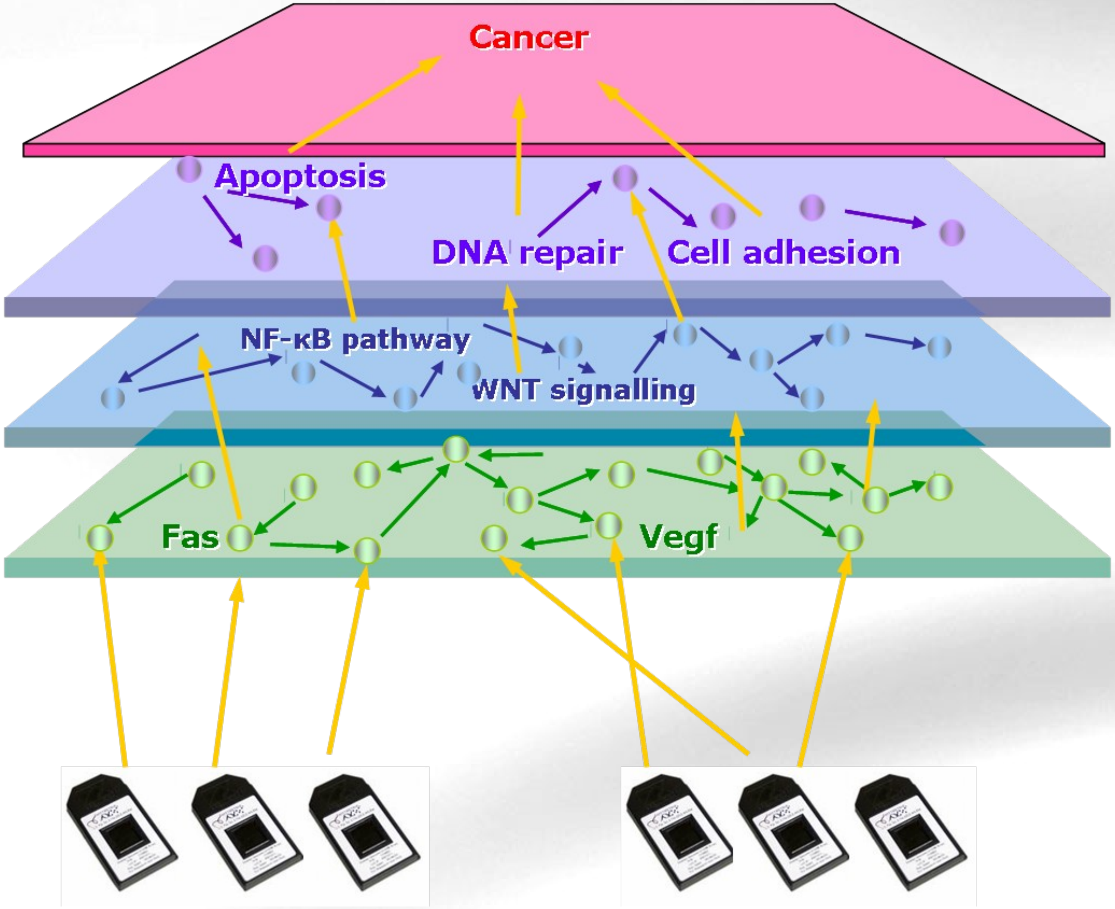
Integrative bioinformatics

Disease

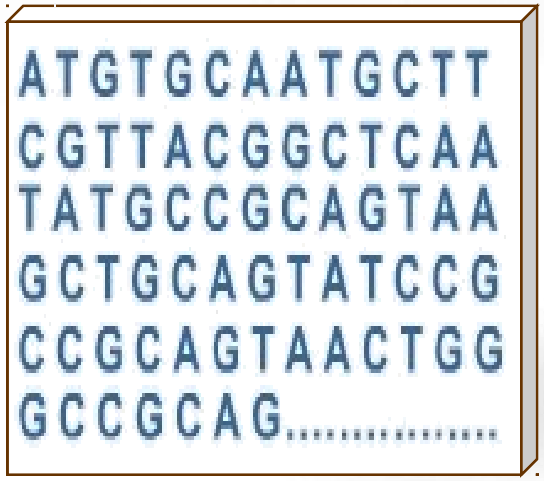
Cellular Processes

Pathways

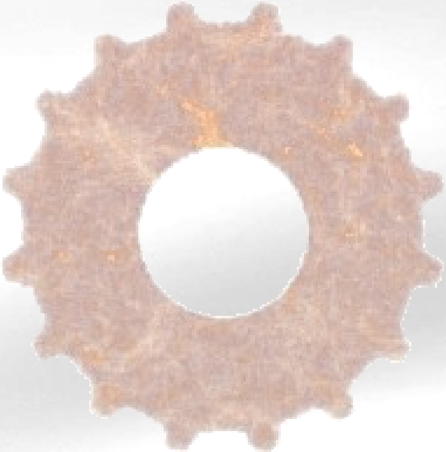
Molecules

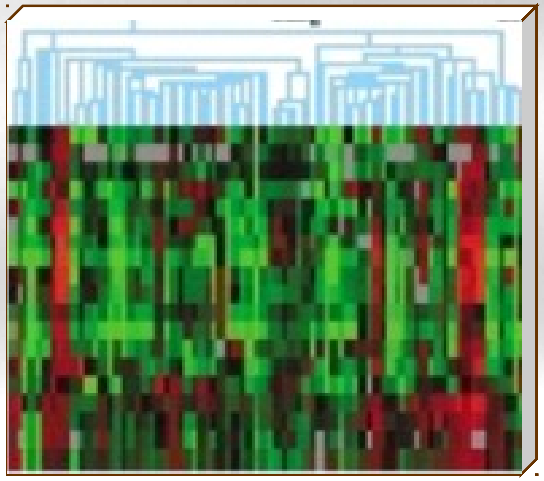


In summary...



ATGTGCAATGCTT
CGTTACGGCTCAA
TATGCCGCAGTAA
GCTGCAGTATCCG
CCGCAGTAACTGG
GCCGCAG...............





Data

Bioinformatics methods and resources

Knowledge

How does one do bioinformatics?

http://biomedicalcomputationreview.org/content/landscape-bioinformatics-education

“Doing bioinformatics”

- Bioinformatics analyses
  - Database searching/querying
  - Sequence analysis, Omics data analysis
  - Systems biology
- Can be done differently
  - From console-based systems
    - Using scripts (perl/python/R) for automating processes
    - Doing data analysis with R
  - Or working with graphical/web interfaces to do (almost) the same things
- Each user has a different preferred approach

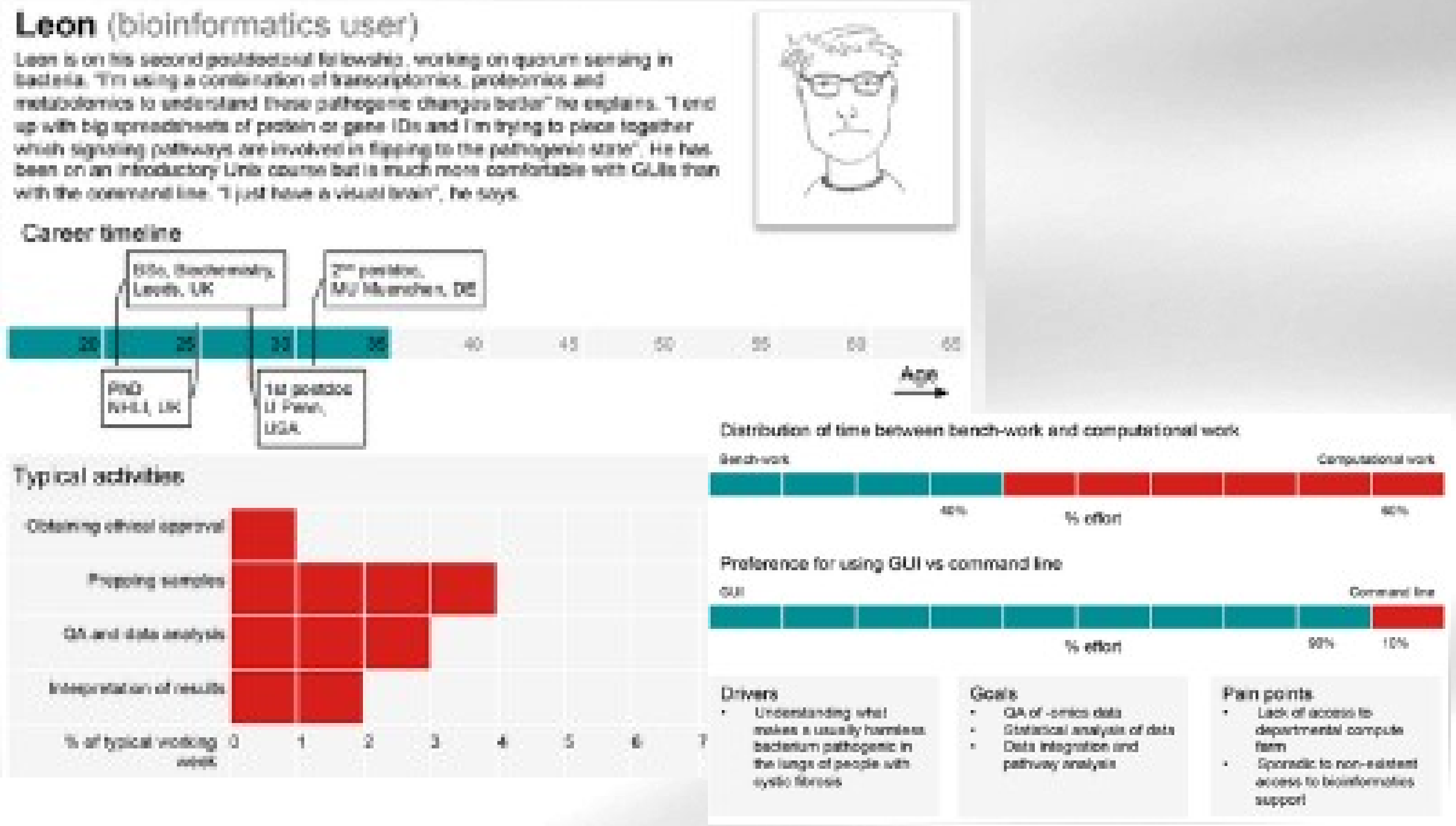
What does a bioinformatician know?

Must have “good background” in

- Some biological discipline
  - Molecular biology, biochemistry, evolution …
- Computer science
  - Operating systems: Linux
  - Programming languages: Python, R, Perl
  - Databases SQL
  - Web development: HTML, PHP, …
- Some “quantitative” science
  - Mathematics, Physics, Statistics

Ideally 1, ½, ⅓ from the previous three!!

A typical “bioinformatics user”



Leon (bioinformatics user)

Leon is on his second postdoctoral fellowship, working on quorum sensing in bacteria. "I'm using a combination of transcriptomics, proteomics and metabolomics to understand these pathogenic changes better" he explains. "I end up with big spreadsheets of protein or gene IDs and I'm trying to piece together which signaling pathways are involved in flipping to the pathogenic state". He has been on an introductory Unix course but is much more comfortable with GUIs than with the command line. "I just have a visual brain", he says.

Career timeline

Typical activities

Obtaining ethical approval
Prepping samples
QA and data analysis
Interpretation of results
% of typical working week
0
1
2
3
4
5
6
7

Distribution of time between bench-work and computational work

Preference for using GUI vs command line

Drivers

- Understanding what makes a usually harmless bacterium pathogenic in the lungs of people with cystic fibrosis

Goals

- QA of -omics data
- Statistical analysis of data
- Data integration and pathway analysis

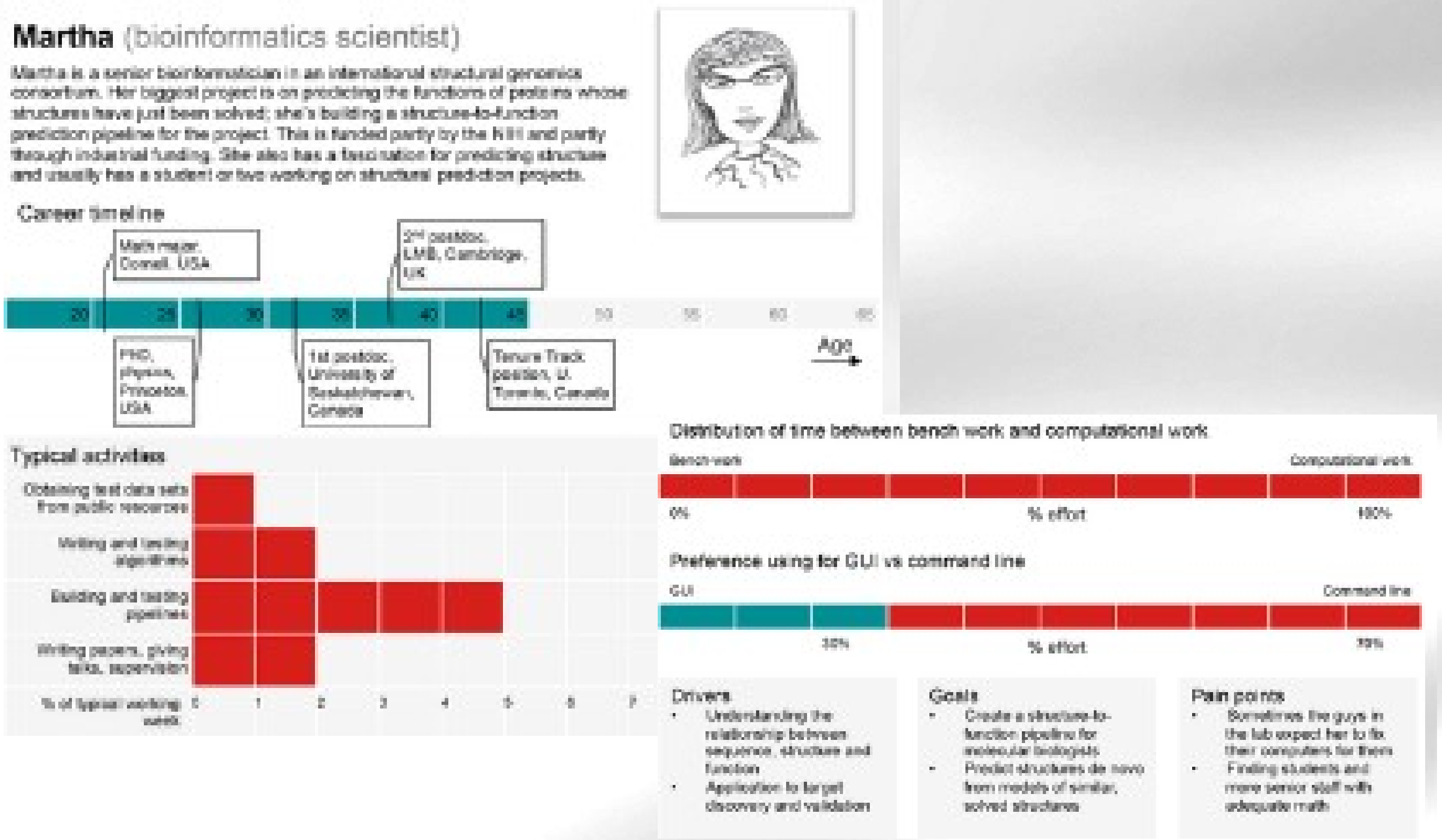
Pain points

- Lack of access to departmental compute farm
- Sporadic to non-existent access to bioinformatics support

Welch L, Lewitter F, Schwartz R, Brooksbank C, Radivojac P, et al. (2014) Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. PLoS Comput Biol 10(3): e1003496. doi:10.1371/journal.pcbi.1003496 http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1003496

Vall d'Hebron
Institut de Recerca
VHIR

A typical “bioinformatics scientist”



Martha (bioinformatics scientist)

Martha is a senior bioinformatician in an international structural genomics consortium. Her biggest project is on predicting the functions of proteins whose structures have just been solved; she's building a structure-to-function prediction pipeline for the project. This is funded partly by the NIH and partly through industrial funding. She also has a fascination for predicting structure and usually has a student or two working on structural prediction projects.

Career timeline
Math major, Cornell, USA
PhD, physics, Princeton, USA
1st postdoc, University of Saskatchewan, Canada
2nd postdoc, LMB, Cambridge, UK
Tenure Track position, U. Toronto, Canada
20
25
30
35
40
45
50
55
60
65
Age
Typical activities
Obtaining test data sets from public resources
Writing and testing algorithms
Building and testing pipelines
Writing papers, giving talks, supervision
% of typical working week
0
1
2
3
4
5
6
7
Distribution of time between bench work and computational work
Bench work
Computational work
0%
% effort
100%
Preference using for GUI vs command line
GUI
Command line
30%
% effort
70%

Drivers
- Understanding the relationship between sequence, structure and function
- Application to target discovery and validation

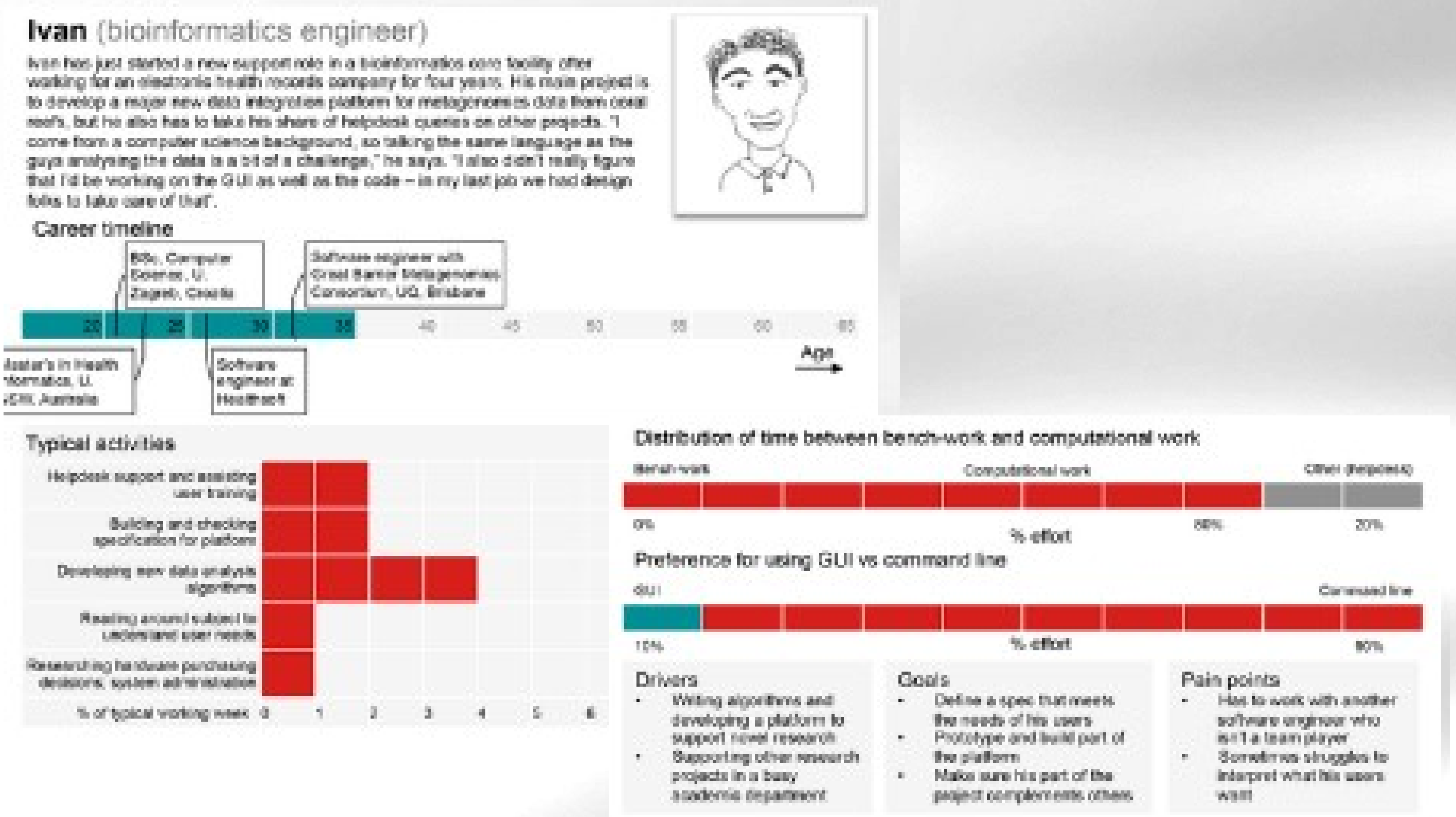
Goals
- Create a structure-to-function pipeline for molecular biologists
- Predict structures de novo from models of similar, solved structures

Pain points
- Sometimes the guys in the lab expect her to fix their computers for them
- Finding students and more senior staff with adequate math

Welch L, Lewitter F, Schwartz R, Brooksbank C, Radivojac P, et al. (2014) Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. PLoS Comput Biol 10(3): e1003496. doi:10.1371/journal.pcbi.1003496 http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1003496

PLOS | COMPUTATIONAL BIOLOGY

A typical “bioinformatics engineer”



Ivan (bioinformatics engineer)

Ivan has just started a new support role in a bioinformatics core facility after working for an electronic health records company for four years. His main project is to develop a major new data integration platform for metagenomics data from coral reefs, but he also has to take his share of helpdesk queries on other projects. "I come from a computer science background, so talking the same language as the guys analysing the data is a bit of a challenge," he says. "I also didn't really figure that I'd be working on the GUI as well as the code – in my last job we had design folks to take care of that".

Career timeline

BSc, Computer Science, U. Zagreb, Croatia
Software engineer with Great Barrier Metagenomics Consortium, UQ, Brisbane
20
25
30
35
40
45
50
55
60
65
laster's in Health nformatics, U. JSW, Australia
Software engineer at Healthsoft
Age

Typical activities

Helpdesk support and assisting user training
Building and checking specification for platform
Developing new data analysis algorithms
Reading around subject to understand user needs
Researching hardware purchasing decisions, system administration
% of typical working week
0
1
2
3
4
5
6

Distribution of time between bench-work and computational work

Preference for using GUI vs command line

Drivers

- Writing algorithms and developing a platform to support novel research
- Supporting other research projects in a busy academic department

Goals

- Define a spec that meets the needs of his users
- Prototype and build part of the platform
- Make sure his part of the project complements others

Pain points

- Has to work with another software engineer who isn't a team player
- Sometimes struggles to interpret what his users want

Welch L, Lewitter F, Schwartz R, Brooksbank C, Radivojac P, et al. (2014) Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. PLoS Comput Biol 10(3): e1003496. doi:10.1371/journal.pcbi.1003496 http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1003496

In summary, bioinformatics …

- Was born with
  - Development of new technologies
  - Its application for generating –increasingly huge- of big masses of biological data.
- *Has become now an interdisciplinar science encompassing all aspects of the Acquisition, Processing, Distribution, Analysis, Integration and Interpretation of biological information.*