

Gens, Ordinadors i Malalties



Vall d'Hebron
Institut de Recerca

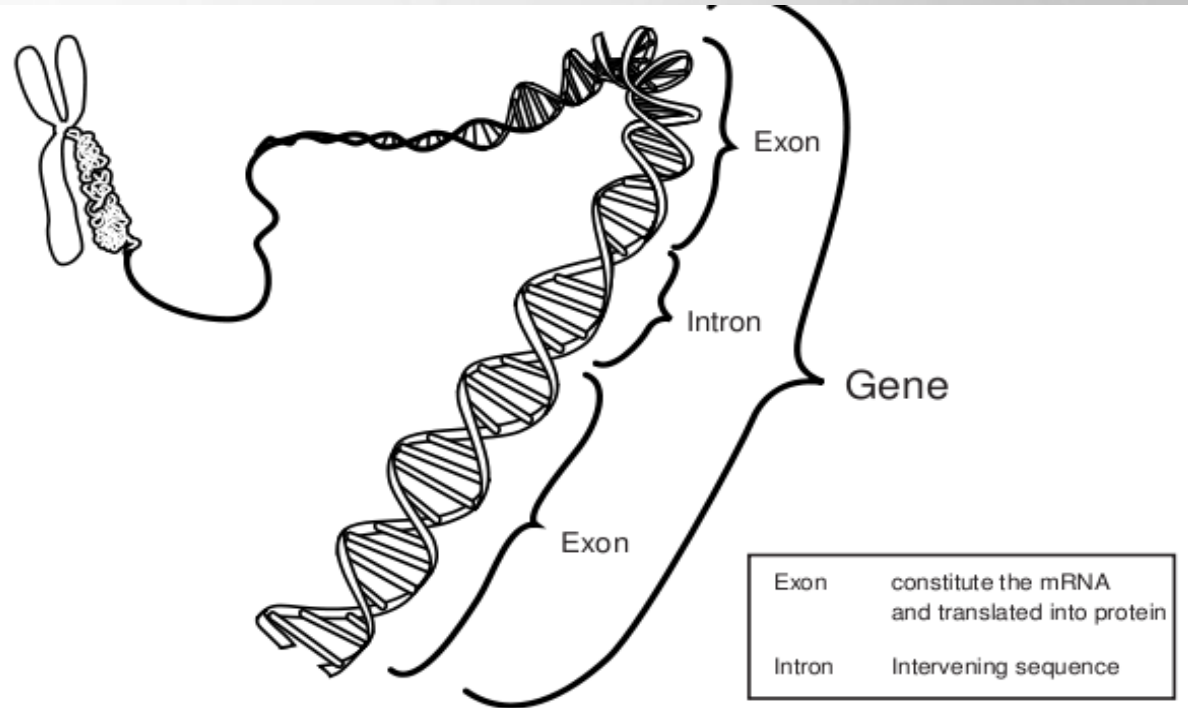
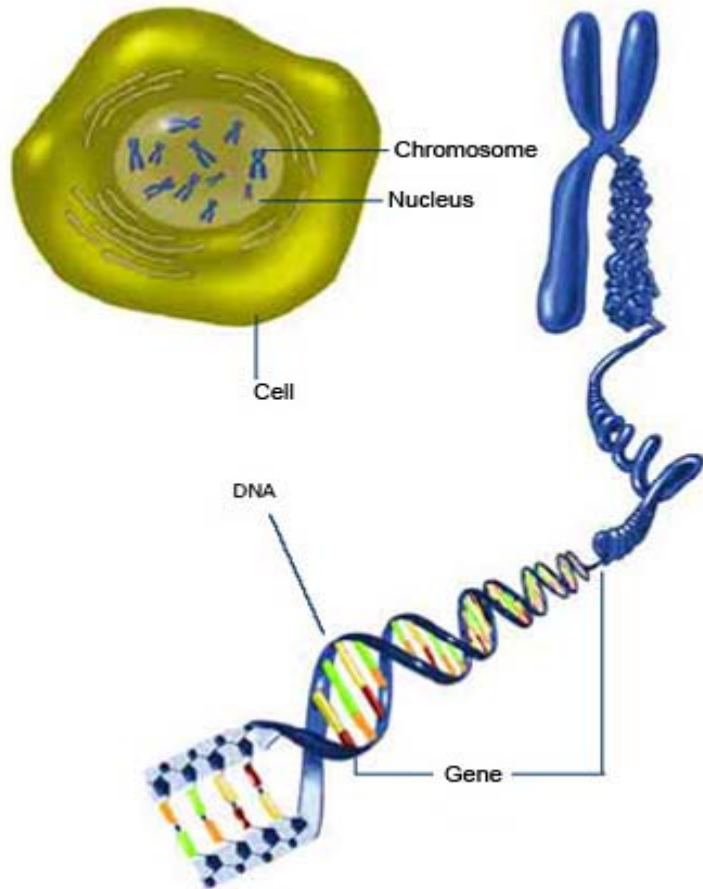
**Biologia i Ordinadors al
segle XXI**

Àlex Sánchez

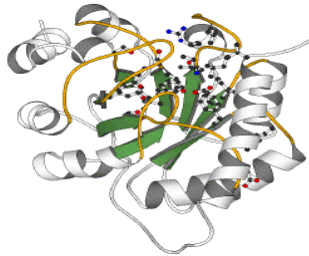
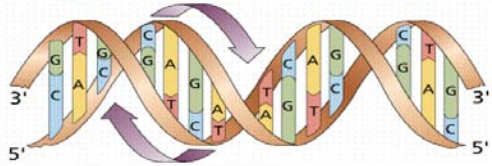
*Unitat d'Estadística i
Bioinformàtica VHIR*

*Departament d'Estadística
Facultat de Biologia. UB*

Genes are made of DNA



The Central Dogma of Molecular Biology



DNA

transcription

mRNA

translation

Protein

CCTGAGCCAAC TATTGATGAA



CCUGAGCCAACU AUUGAUGAA



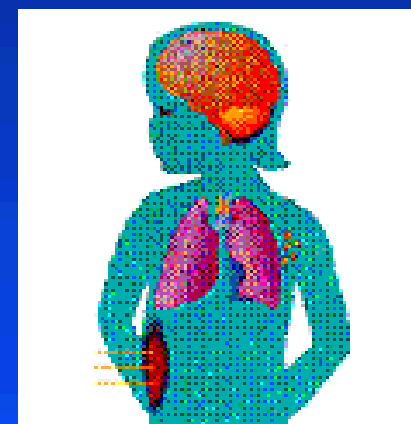
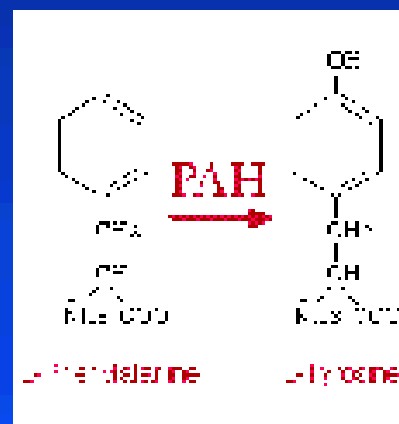
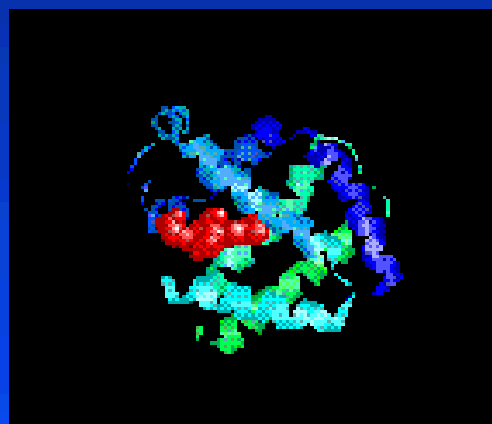
PEPTIDE

Central Paradigm of Bioinformatics

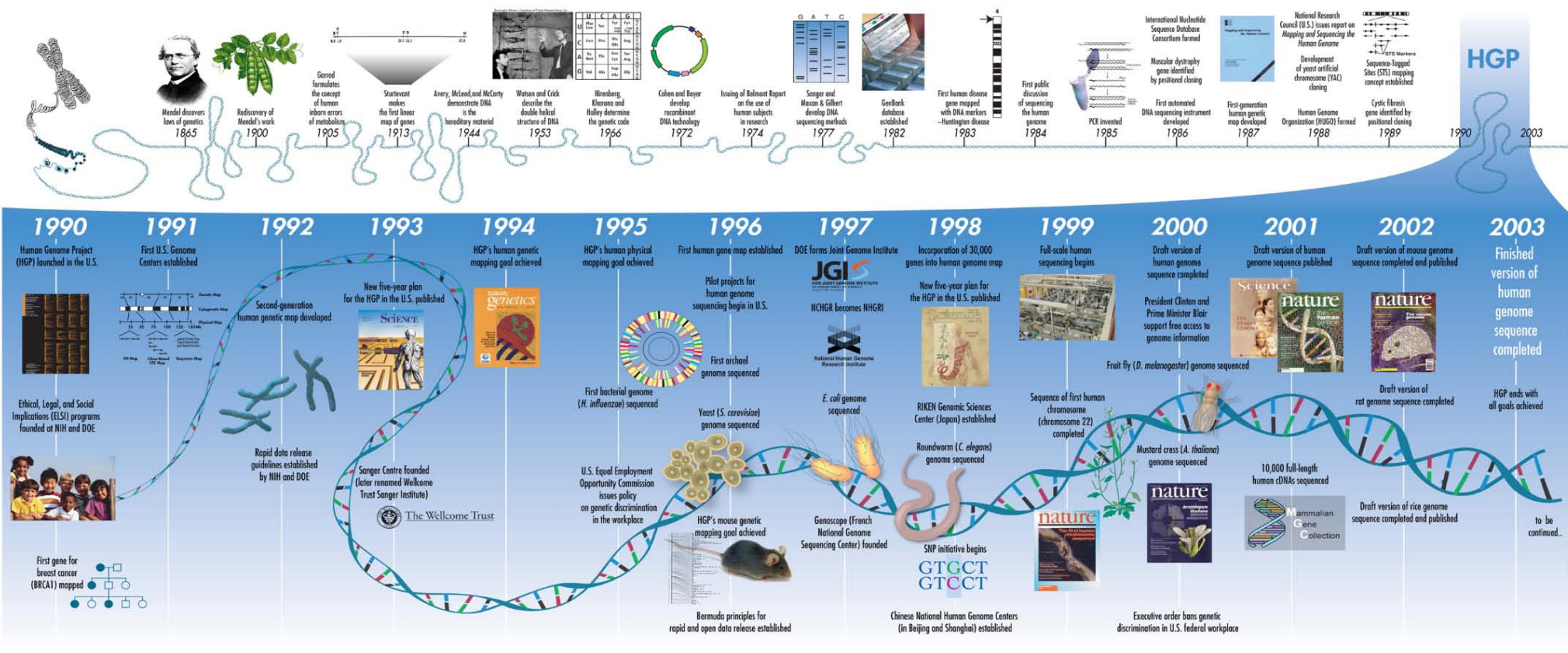
Genetic Information → Molecular Structure → Biochemical Function → Symptoms (Phenotype)

```

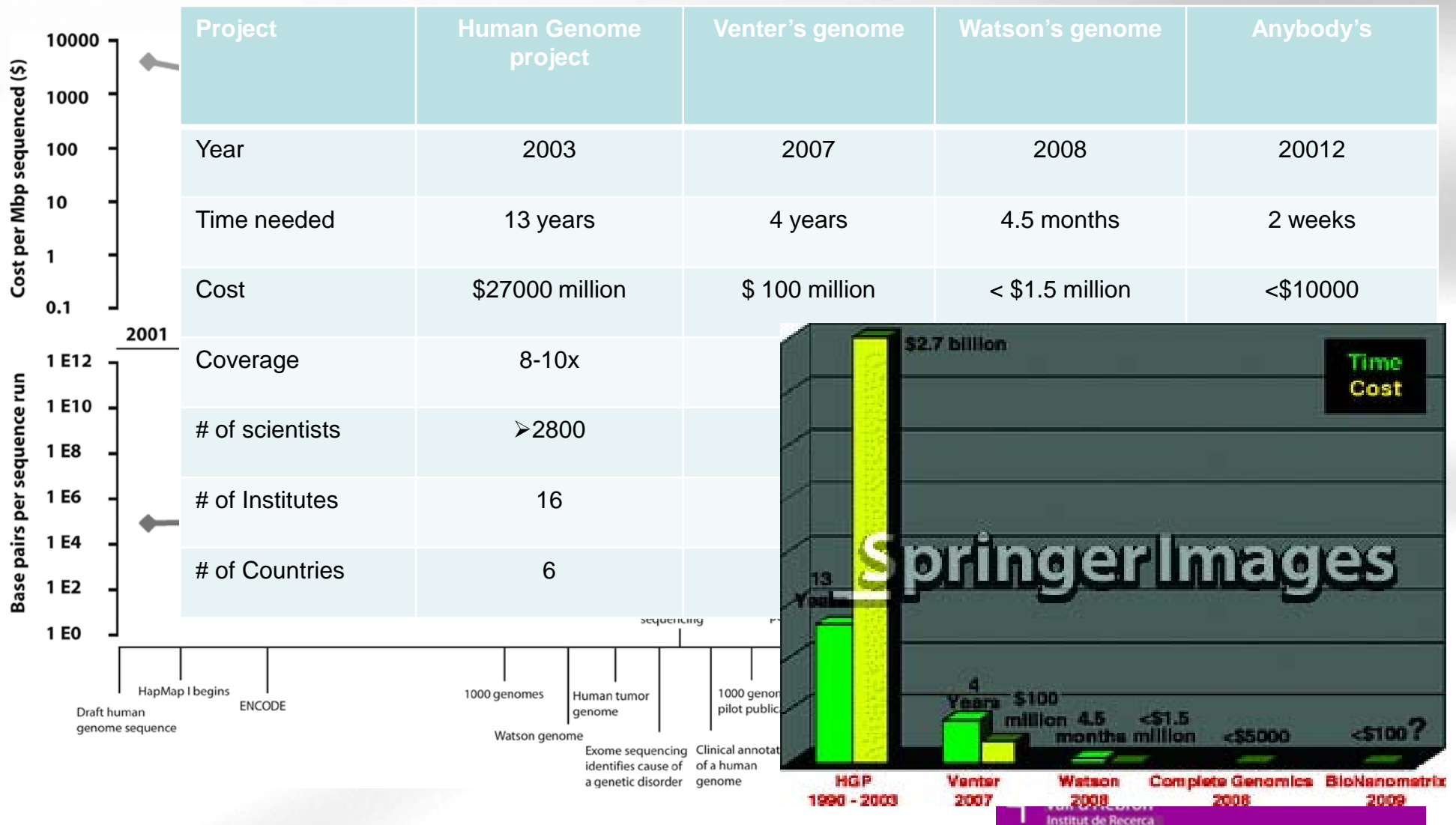
SEAAINGHIVA
VEYQTUSRVVI
VSTATVSRALA
GVTTTVSHVIN
EGVSAVSAILN
GVSEMTEDILN
TAYATIKVEVE
GSGPTVSRRLA
MSIATITRGSN
LSBETVSRILK
FDISRLSHLPK
LSPSRLANLPK
MTVETISRLLC
TLEPHLRLPK
    
```



The human genome project



Decreasing costs & Increasing capabilities of genome sequencing

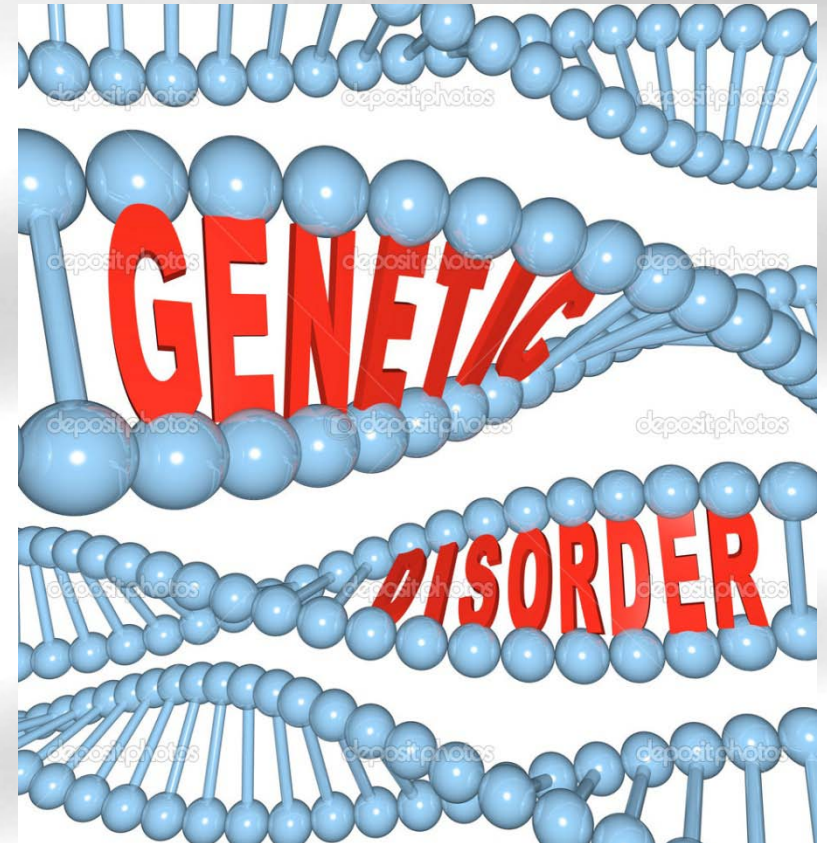


Diseases and genetic diseases

- Traditionally : 3 types of diseases
 - 1. genetically determined
 - 2. environmentally determined
 - 3. 1. + 2.
- Today : distinctions are blurred
 - up to 20% of pediatric in-patients have genetic abnormality
 - about 50% of spontaneous abortuses have chromosomal aberration
 - only mutations that are not lethal are reservoir of genetic diseases

Genetic disease

- Genetic disorders are caused by abnormalities in the genetic material.
- Abnormalities can range from a small mutation in a single gene to the addition or subtraction of an entire chromosome or set of chromosomes.

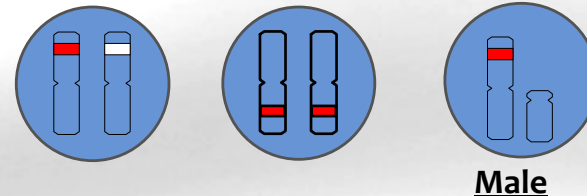


Classification of genetic disorders

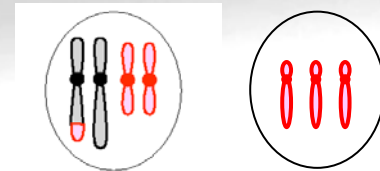
- Multifactorial



- Single gene



- Chromosomal



- Mitochondrial

- Somatic mutations (cancer)

Gene identification of inherited disease

- Every gene has a specific task
- Disease genes carry mutations, which change the protein, which alters the way the task is usually performed.
- The mutation may be
 - within a gene/protein or
 - within a regulatory part of the genome that, e.g., affects the amount of protein being produced.



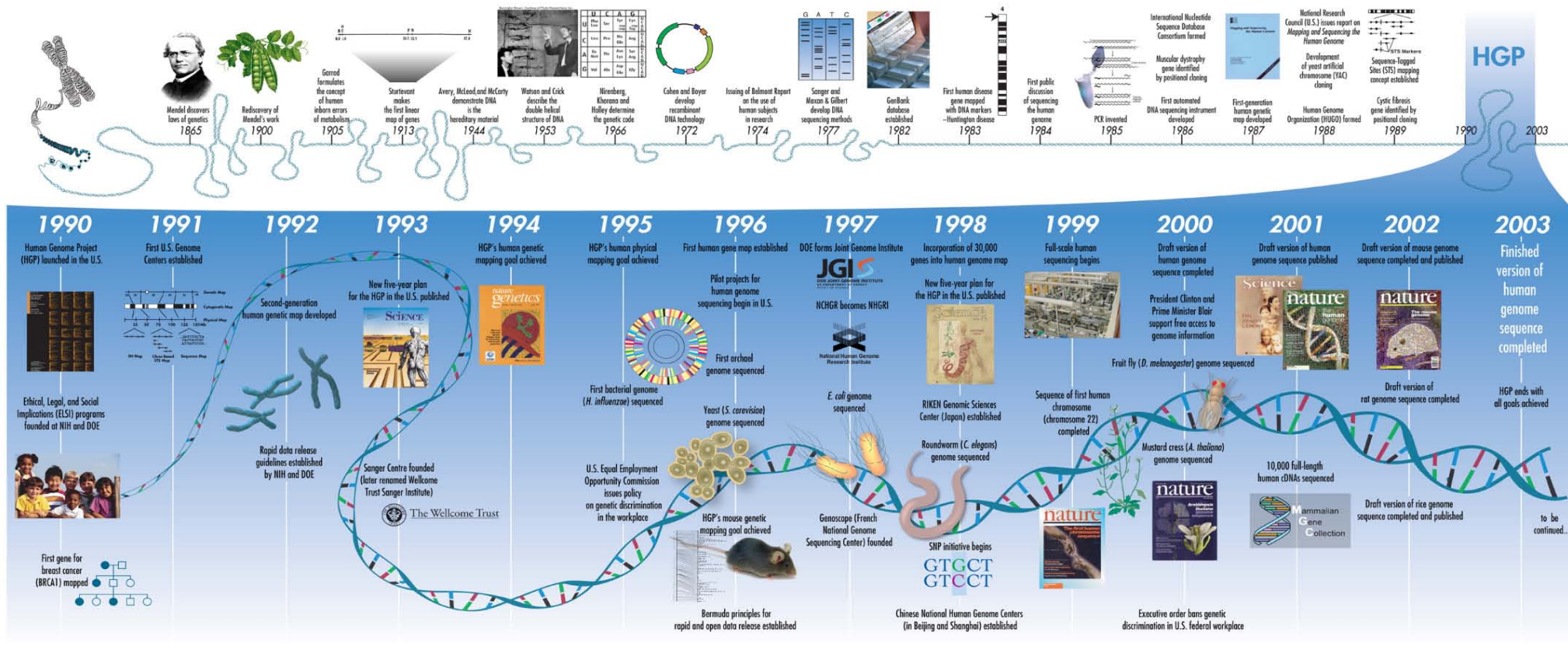
Modern approaches to disease gene identification and analysis

- The release of genomic sequences, full-length cDNA sequences, of human and model organisms offer invaluable resources for studying genetic diseases.
- It is not possible, however, to access such an impressive quantity of biological information without the use of powerful informatics resources
- ***That is where bioinformatics enters the scene***

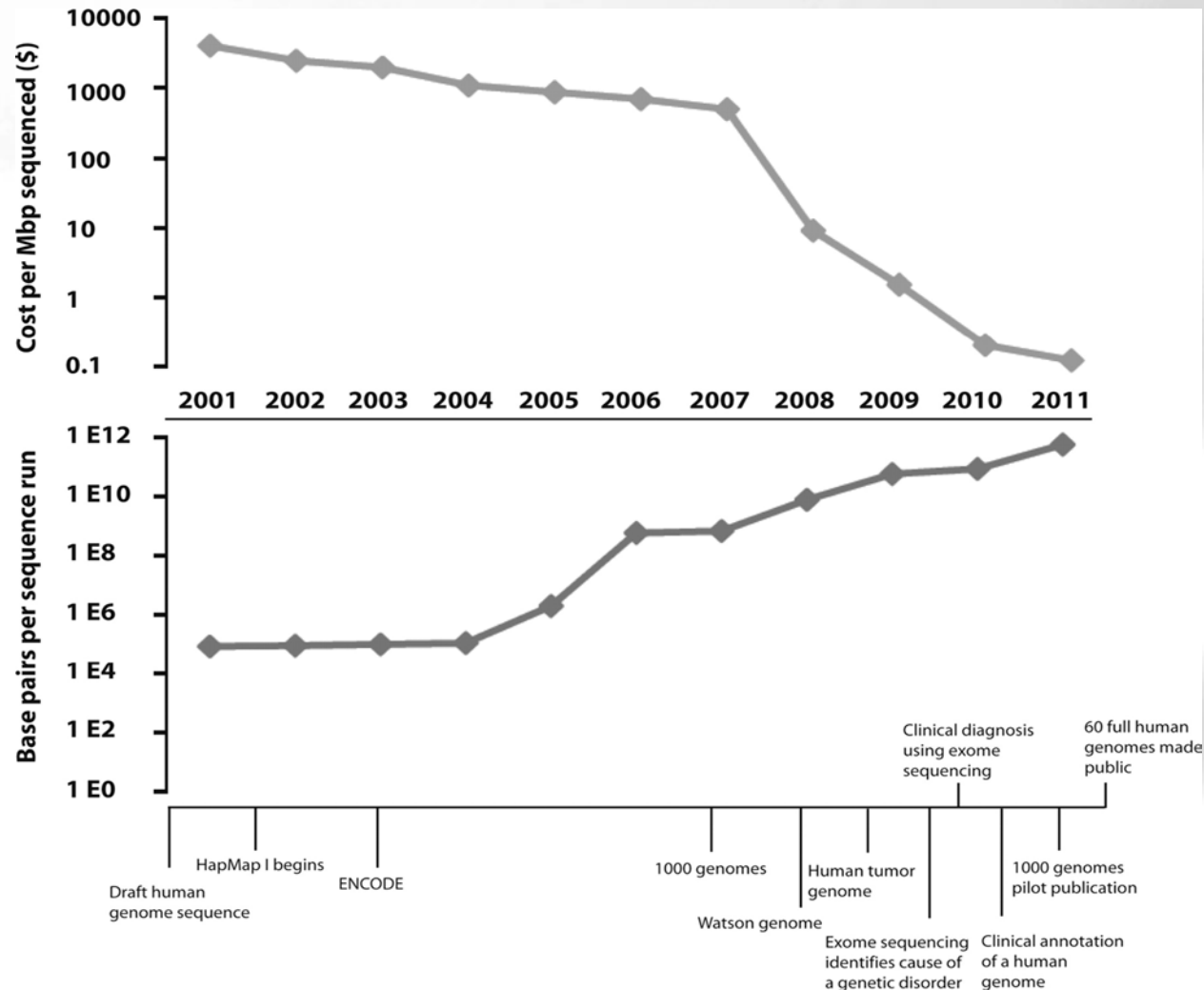
The data explosion(s)

- Bioinformatics has evolved in parallel with
 - The increase in biological data available
 - The ability to generate (and *manage*) them.
- We distinguish (up to now) ...
 - Pre-genomic age
 - Early post-genome age
 - Late post-genome age

The human genome project



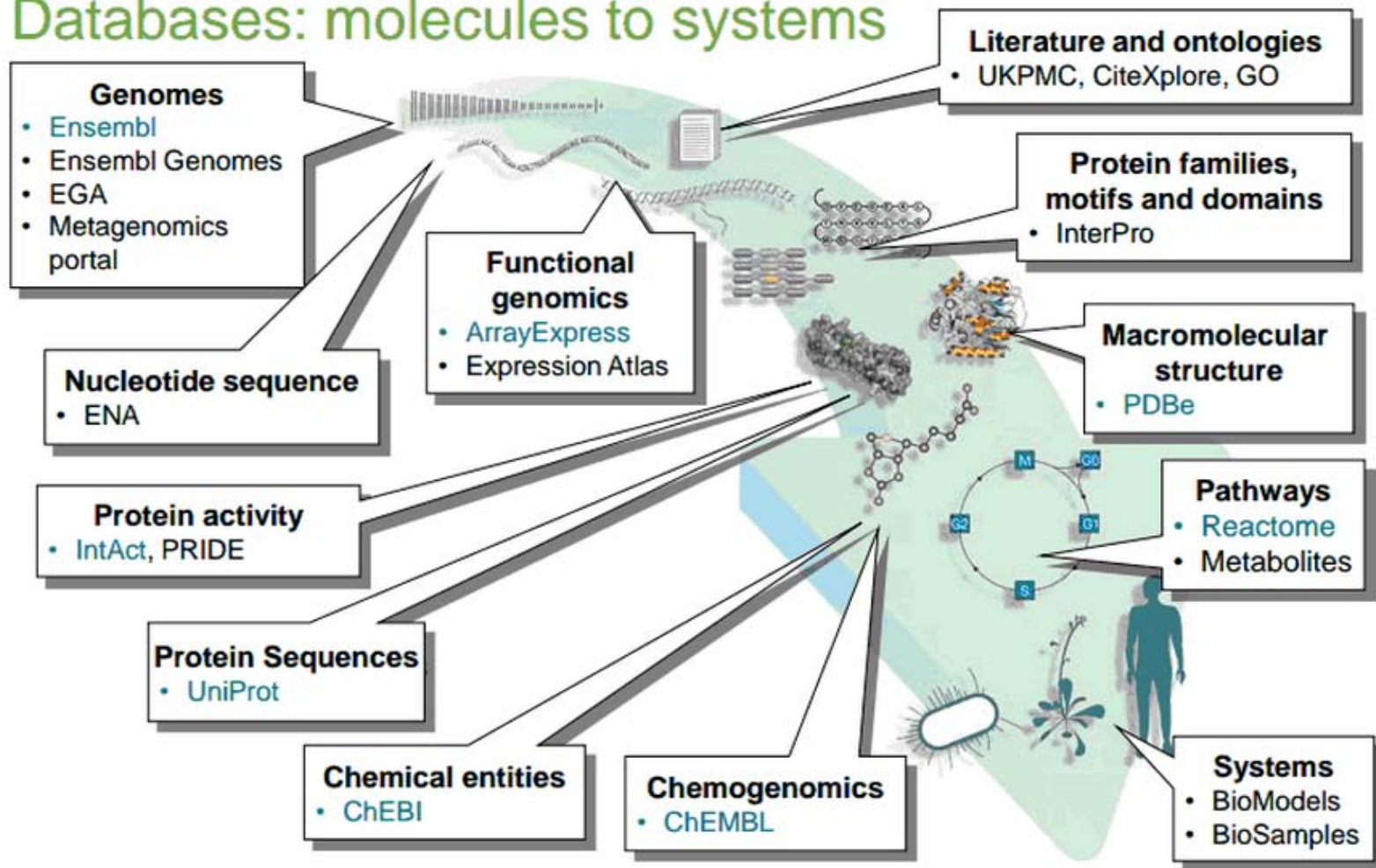
NGS: (much) More per (much) less



<http://www.genome.gov/sequencingcosts>

Today: a wealth of diverse information

Databases: molecules to systems

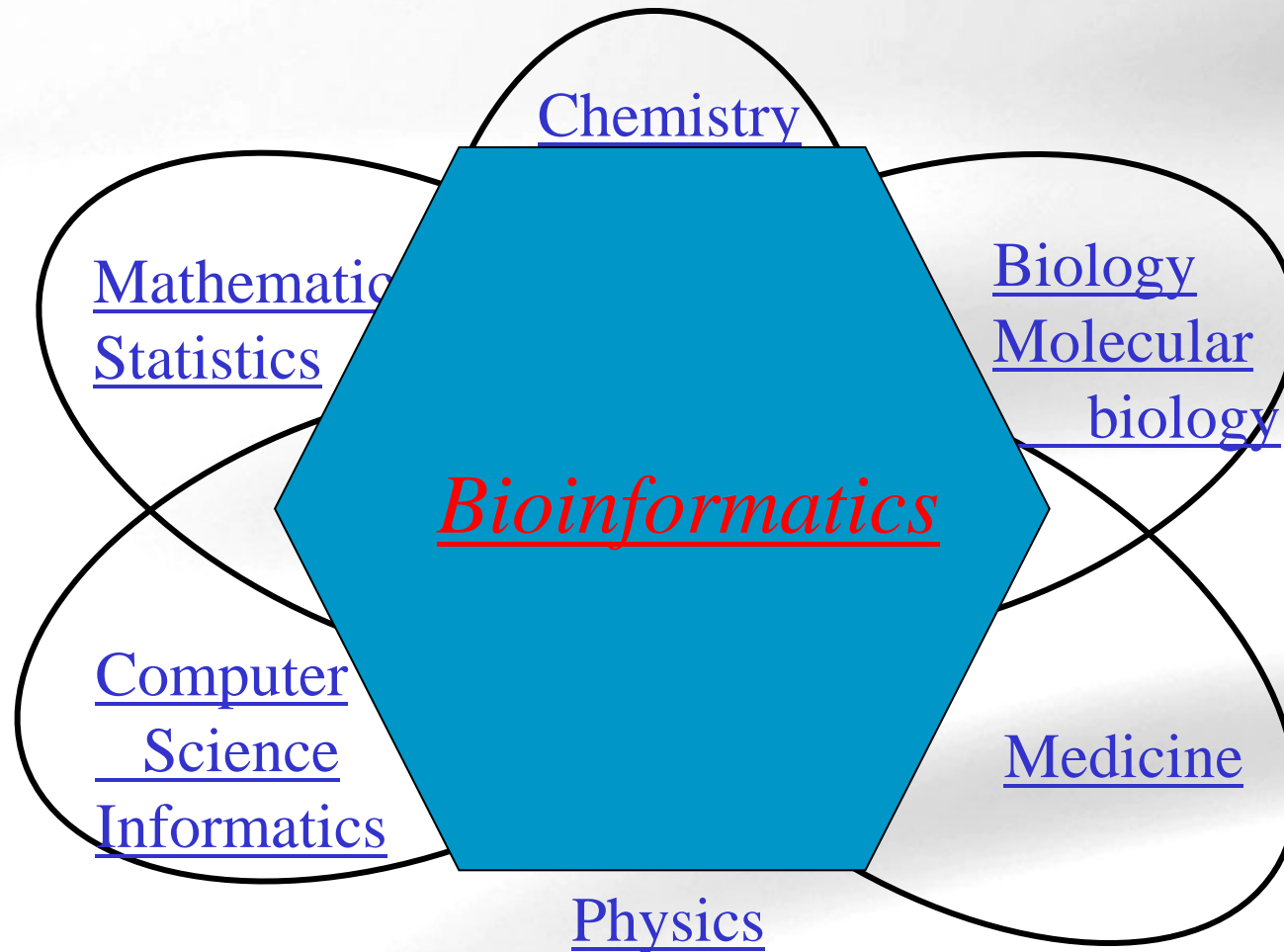


EMBL-EBI



- Born on the lookout for
 - the development of new technologies and
 - its application in the generation of huge amounts of data.
- *It is the interdisciplinary scientific field that develops methods for storing, retrieving, organizing and analyzing biological data.*

BIOINFORMATICS AND INTERDISCIPLINARITY



- Information organization
 - Databanks and databases
 - Algorithms and exploitation tools
- Analysis and interpretation of experimental results
 - Genome analysis and sequencing
 - Comparative genomics
 - Gene expression and transcription
 - Proteomics, protein-protein interaction
- Systems biology modeling

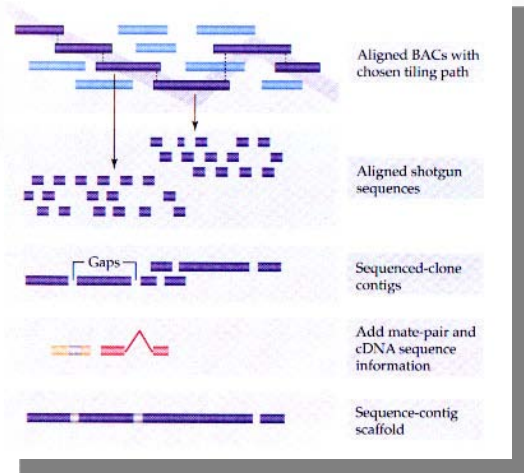
INFORMATION ORGANIZATION

The screenshot shows the NCBI Entrez homepage. At the top, there's a navigation bar with links to PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, and Structure. Below this is a search bar with the text "Search across databases" and a "Go" button. The main content area is titled "Welcome to the Entrez cross database search page" and displays a grid of database categories and their descriptions. The categories include: Published biomedical literature citations and abstracts, Published Central full-text journal articles, Site Search (NCBI web and FTP sites), Nucleotide Core subset of nucleotide sequence records, EST Expressed Sequence Tag records, GDB Genome Survey Sequence records, Protein sequence database, Genomes whole genome sequences, Structures three-dimensional macromolecular structures, Taxonomy organisms in GenBank, SNP single nucleotide polymorphism, Gene gene-oriented information, SRA Short Read Archive, Biophysical Pathways and systems of interacting molecules, Homologous subfamily homolog groups, GENES gene expression data of mouse central nervous system, Probe sequence-specific reagents, Genome Projects genome project information, BioRx protein and phenotype, OMIM online Mendelian Inheritance in Man, OMIM online Mendelian Inheritance in Animals, dbGAP genotype and phenotype, UniGene gene-oriented clusters of transcribed sequences, CDD conserved protein domain database, 3D Rfam domains from Entrez Structure, dbSTS markers and mapping data, PopSet population study data sets, GEO Profiles expression and molecular abundance profiles, GEO DataSets experimental sets of GEO data, Cancer Chromosome 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, PeptideAtlas MS/MS proteomic experiments, Journals detailed information about the journals indexed in PubMed and other Entrez databases, and NLM Catalog listing of books, journals, and audio/video in the NLM collections.

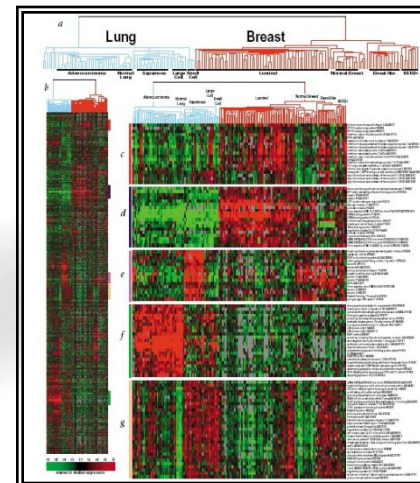
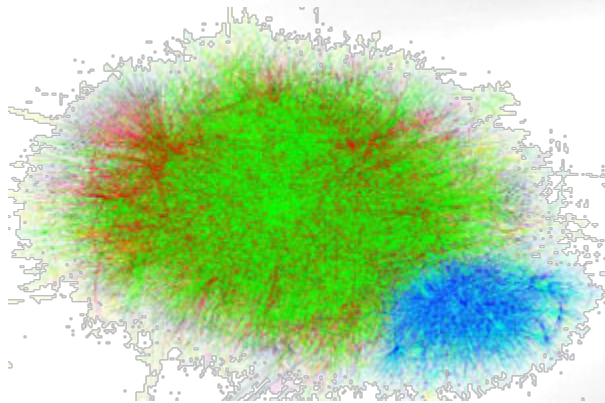
The screenshot shows the NCBI Tools for Data Mining page. The top navigation bar includes links to PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, and Structure. Below this is a search bar with the text "Search Entrez for" and a "Go" button. The main content area is titled "Tools - Nucleotide Sequence Analysis" and lists several tools: BLAST (The Basic Local Alignment Search Tool), Electronic PCR (allows you to search your DNA sequence for sequence tagged sites), Entrez Gene (each Entrez Gene record encapsulates a wide range of information), Model Maker (allows you to view the evidence), ORF Finder (identifies all possible ORFs in a DNA sequence), and Organism Specific Resources (Bee, Cat, Chicken, Cow, etc.). The page also includes a "Site Map" link, a "Tools for Programmers" link, a "BLAST Standard tool for sequence analysis" link, a "BLINK BLAST Link" link, a "CDART Conserved Domain Architecture Retrieval Tool" link, a "CD search Conserved Domain Database search" link, a "CGAP Cancer Gene Anatomy Project" link, a "Cn3D View 3-dimensional structures" link, and a "COGs Clusters of Orthologous Groups" link.



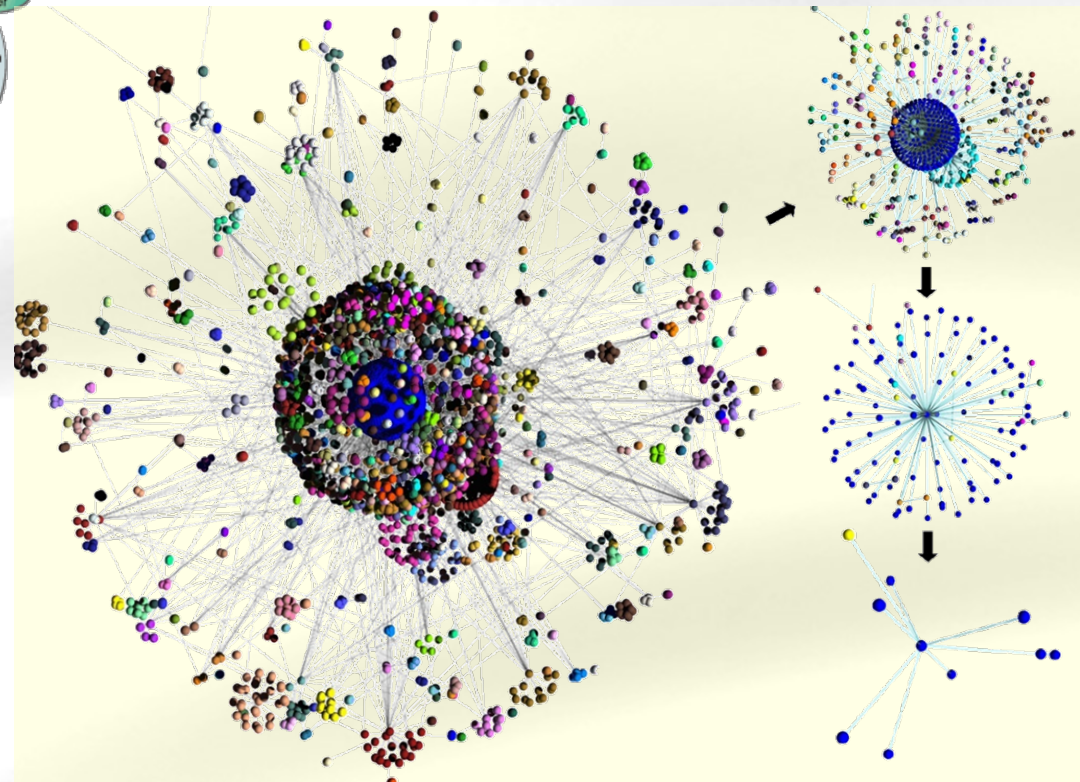
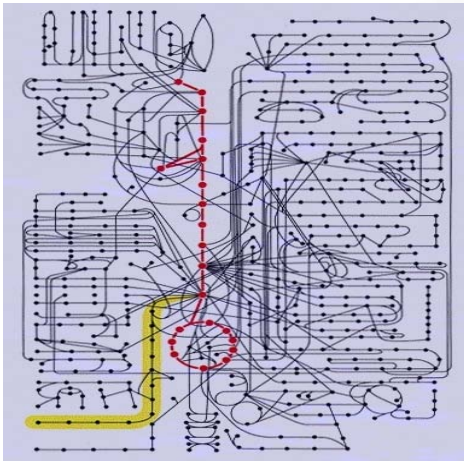
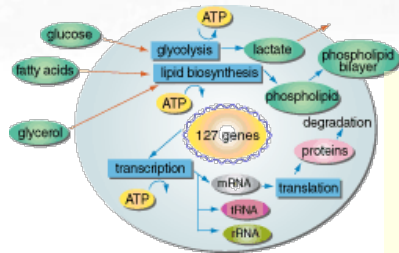
ANALYSIS AND INTERPRETATION

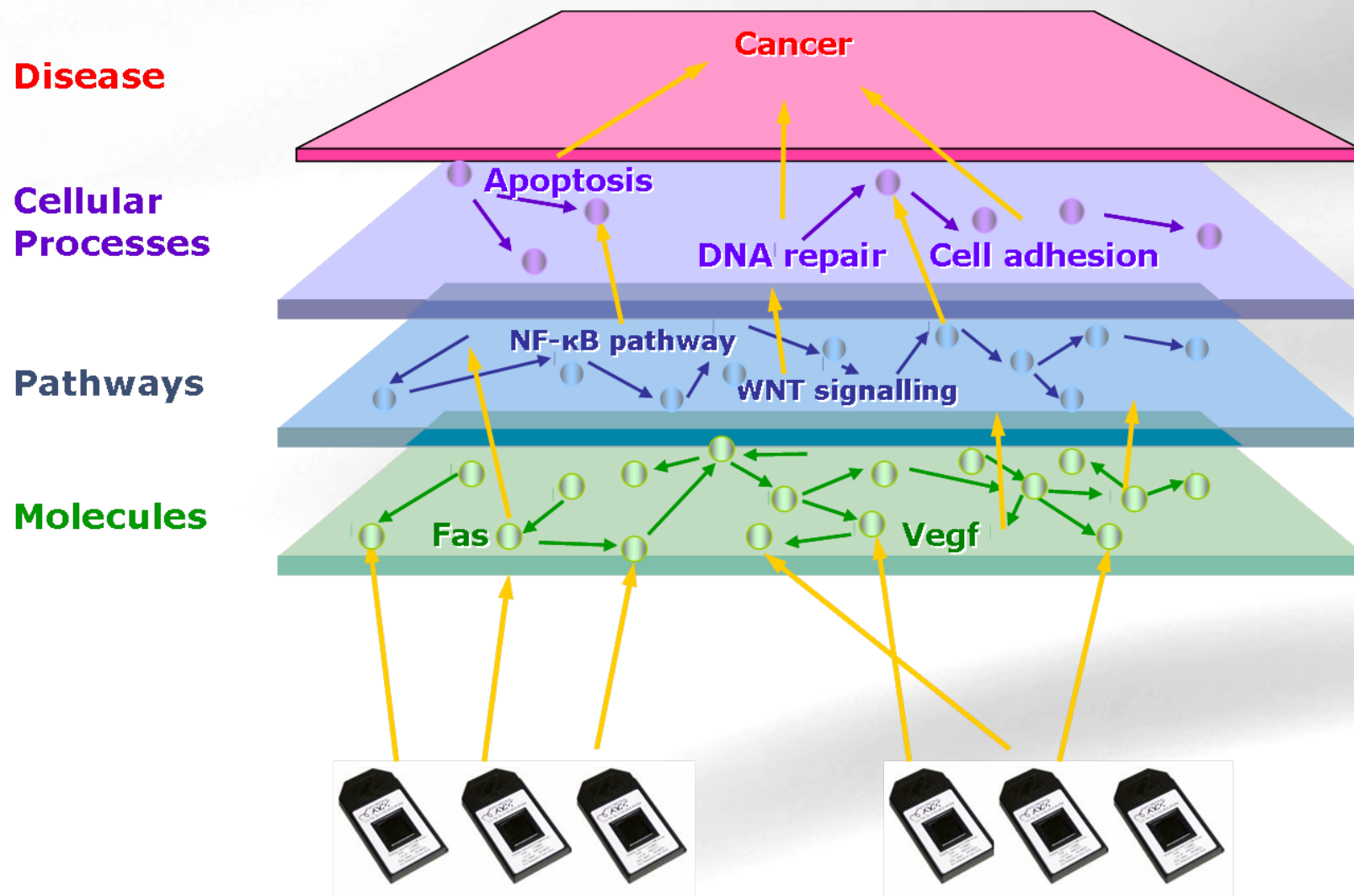


AGAGTTCTGCTCG
AGGGTTATGCGCG



SYSTEMS BIOLOGY MODELING



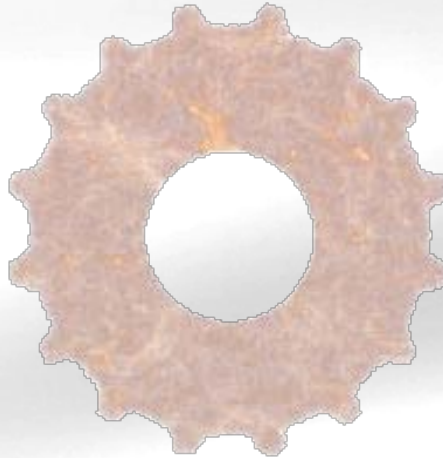


IN SUMMARY...

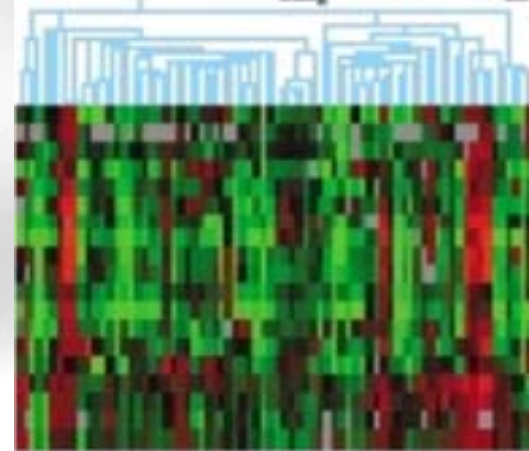


ATGTGCAATGCTT
CGTTACGGCTCAA
TATGCCGCAGTAA
GCTGCAGTATCCG
CCGCAGTAACTGG
GCCGCAG.....

Data



Bioinformatics tools
and resources



Knowledge

How does one do
bioinformatics?

Leon (bioinformatics user)

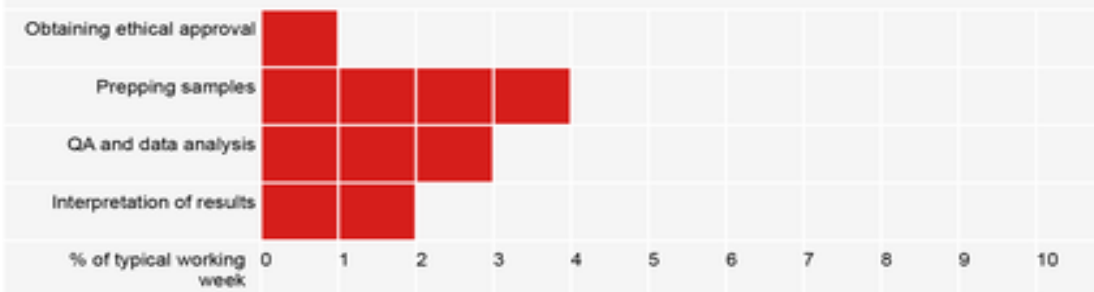
Leon is on his second postdoctoral fellowship, working on quorum sensing in bacteria. "I'm using a combination of transcriptomics, proteomics and metabolomics to understand these pathogenic changes better" he explains. "I end up with big spreadsheets of protein or gene IDs and I'm trying to piece together 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.



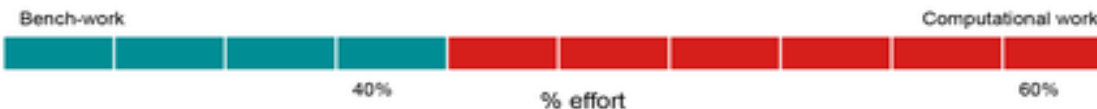
Career timeline



Typical activities



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

•A typical “bioinformatics user.”

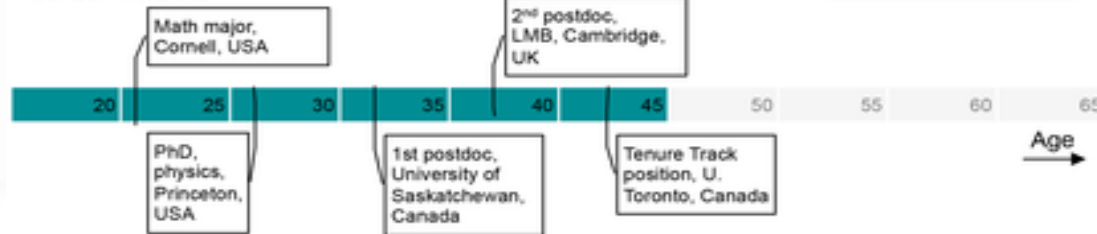
Welch L, Lewitter F, Schwartz R, Brooksbank C, et al. (2014) Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. PLoS Comput Biol 10(3): e1003496. doi:10.1371/journal.pcbi.1003496 <http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003496>

Martha (bioinformatics scientist)

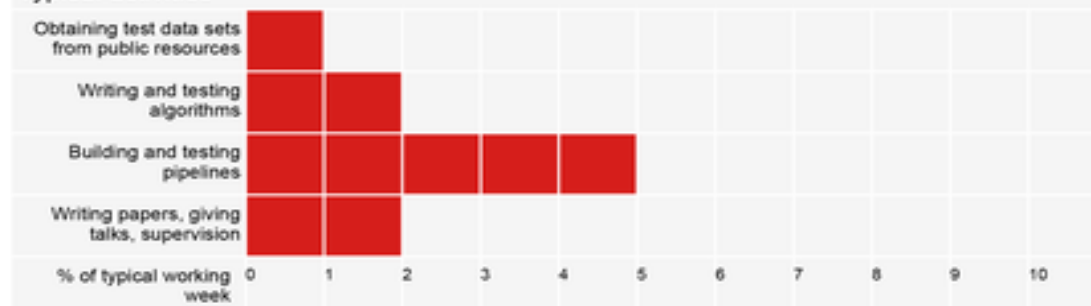
Martha is a senior bioinformatician in an international structural genomics consortium. Her biggest project is on predicting the functions of proteins whose structures have just been solved; she's building a structure-to-function prediction pipeline for the project. This is funded partly by the NIH and 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



Typical activities



Distribution of time between bench work and computational work



Preference using for GUI vs command line



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

•A typical “bioinformatics scientist.”

Welch L, Lewitter F, Schwartz R, Brooksbank C, 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://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003496>



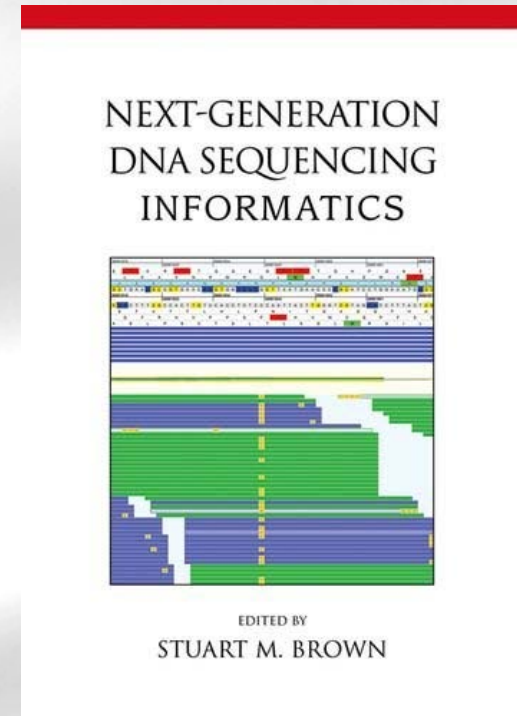
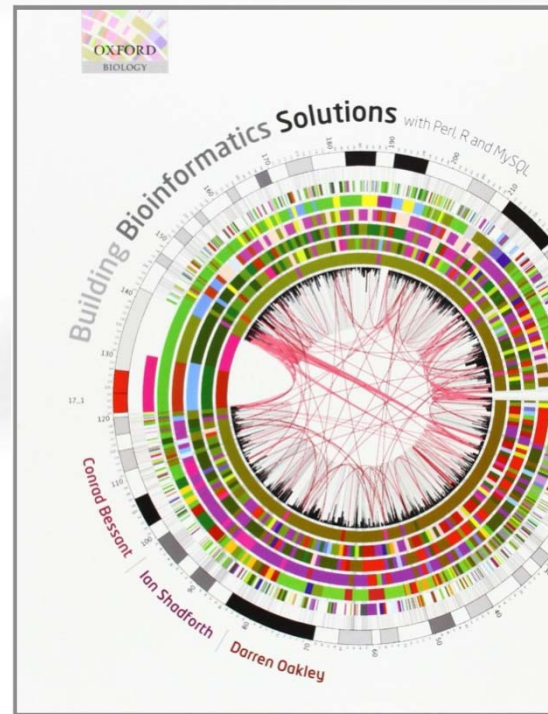
Computer systems for Bioinformatics

- Bioinformatics has been developed in environment which required ...
 - Creating and accessing databases in the web
 - Creating and executing programs in the web
 - Installing and Managing web servers.
 - Ability of file text parsing and batch processing.
 - The need of developing and sharing tools between diverse systems and users.
- Windows was not an option of choice
 - No console (or almost). Difficult to share with non-windows systems.
Difficult to scale-up applications.
- Unix/Linux best suited for this purpose

Bioinformatics “computer skills”

- Some consensus on that these include:
 - SQL and knowledge of databases.
 - Perl or Python.
 - basic Linux.
 - basic bash shell scripting.
 - Some experience with Java or other “traditional languages”.
 - R + Bioconductor.

Some references



TO KNOW MORE

- There are a lot of free courses and training activities:
 - EBI's [Online training and courses](#)
 - NCBI's [tutorials](#)
 - Local training courses
 - [Introducción a la Bioinformática](#) (Alex Sánchez, UEB/UB)
 - [Invitació a la Bioinformàtica](#) (Plataforma BioinfoUAB)
- A great variety of rereference books:
 - [List of books on bioinformatics](#)
- Scientific Societies and Publications:
 - [Bioinformatics](#), [Briefings in Bioinformatics](#)
 - [International Society for Computational Biology](#)