BMOL-F-413 Bioinformatique

Introduction

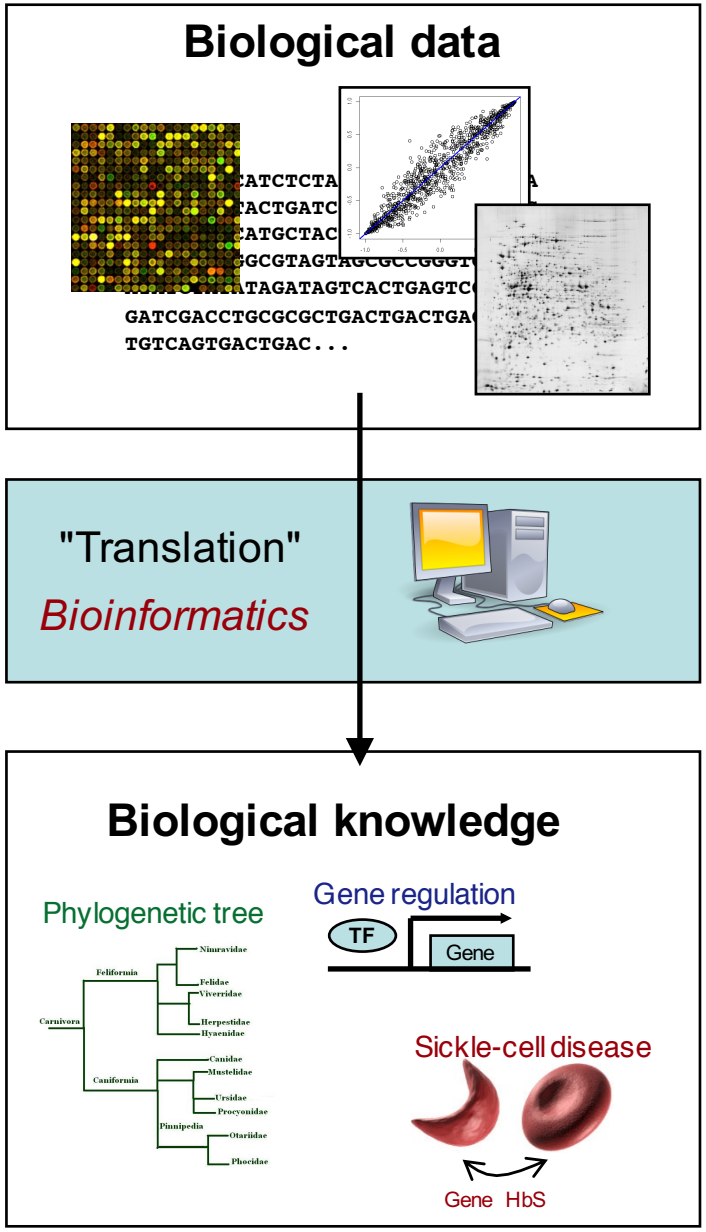
What is Bioinformatics?

- **"Bioinformatics** is an integration of mathematical, statistical and computer methods to analyse biological, biochemical and biophysical data" *(Georgia Inst of Tech., USA)*

- **"Bioinformatics** is the study of biological information as is passes from its storage site in the genome to the various gene products in the cell. …it involves the creating and development of advanced information and computational technologies for problems in molecular biology…" *(Stanford University, USA)*

- **Bioinformatics :** Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data (NIH, USA).

What is Bioinformatics?



Bioinformatics is a translational science

Academic results need translation into products that benefit the people that make up our society. Bioinformatics is such a ***translational science***, where abstract knowledge of mathematicians and informaticians is transferred and applied to data, generated by the hands-on research of biologists and doctors. Bioinformatics makes the world-wide research infrastructure of e-sciences and ICT available to the life sciences.

Bioinformatics includes research into data, derived in different domains (biological, agro, food, medical, behavioural, health-related, pharmacological, and others), that is aimed at increasing knowledge relevant for these domains. This includes the development of tools and approaches related to ***acquiring, storing, organizing, modelling, analyzing, and visualizing*** such data.

(NBIC, The Netherlands)

Source: http://www.nbic.nl/whatis/bioinformatics/

Some landmarks

1869 DNA first isolated

First radio message 1901

1909 The term "gene"

1937 X-ray cristallography of biomolecules

1941 One gene - one enzyme

ENIAC 1948

1951 protein sequencing methods (Sanger)

1952 DNA double helix

First integrated circuit 1958

1958 First protein sequence (insulin)

1961 genetic code

1962 protein structure, X-ray (myoglobin)

1964 First DNA sequence

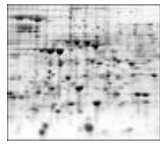
1965 Atlas of protein sequences

1967 Use sequences to build trees

UNIX 1969

Cray1 supercomputer 1971

1970 DP alignment algorithm



Basic 1976

1975 2D electrophoresis

1977 DNA sequencing

Epson Printer 1978

Intel 8086 (8 MHz)

1978 substitution matrices (Dayhoff)

1979 first DNA database (GenBank)

Some landmarks

Commodore 64 1982

Graphical interface & mouse (Mac) 1984

Amiga 1985

Perl 1987

Internet 1989

PC 386 1991

Linux 1991

Java 1995

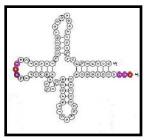
Deep Blue beats Kasparov 1997

Google 1998

imac 2000

Facebook 2004

1981 RNA secondary structure



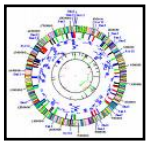
1983 Polymeras Chain Reaction

1987 Multiple sequence alignment (Clustal)

1988 FastA

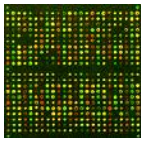
1990 Recombinant DNA

1990 BLAST



1995 First genome sequence: *H influenza*

1995 microarray



1996 yeast genome sequence

1998 C. elegans genome sequence

1998 Hierarchical clustering (microarray)

2000 Arabidopsis genome sequence

2001 miRNA

2004 Human genome sequence

Some landmarks

MacBook 2006

Wii (Nintendo) 2006

First 1 TB hard disk drive 2007

Roadrunner (IBM) 2009
(10^{15} operations / second)

iPad (Apple) 2010

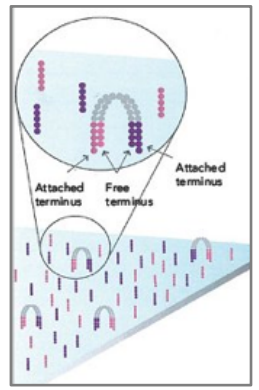
Google Chrome / Chromebook 2011

iCloud 2011

XBox One (Microsoft) 2013
& PS4 (Sony)

AppleWach 2015

2006 Next Generation Sequencing
(Solexa / Illumina)



2007 Human Metabolome Project

2007 Global Ocean Sampling Expedition
(Metagenomics, Publication, C. Venter)

2008 Velvet (short read assembly)

2008 Chip-Seq

2010 Global map of human gene expression

2010 Database of Genomic Variants

2010 1000-genome project
(first published results)

2012 ENCODE (encyclopedia of DNA elements)

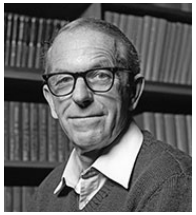
2015 TARA project (Oceans Expedition)

Sequencing

The primary source of data to be analyzed with bioinformatic tools are DNA and protein sequences

History of sequencing

1951: Frederick Sanger: first protein sequence (insulin)



Sanger F, Tuppy H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochem. J. 49: 463
Sanger F, Tuppy H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem. J. 49: 481–490.

1970: Ray Wu: first method for determining DNA sequences involving a location-specific primer extension strategy.

Wu R (1970) Nucleotide sequence analysis of DNA I: Partial sequence of the cohesive ends of bacteriophage and 186 DNA. J Mol Biol 51:501-521

1977: Frederick Sanger: primer-extension strategy

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74: 5463-7.

1977: Walter Gilbert & Allan Maxam: DNA sequencing by chemical degradation

Maxam AM, Gilbert W (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74: 560-4.

1986: Leroy E. Hood: First semi-automated DNA sequencing machine.

1987: Applied Biosystems: first fully automated sequencing machine (Capillary array sequencer, ABI 3700)

1995: Craig Venter

1996: Pal Nyren & Mostafa Ronaghi: pyrosequencing

>2000: Illumina/Solexa: sequencing methods based on reversible dye-terminators technology, and engineered polymerases => massive parallel sequencing

History of sequencing

Next generation sequencing

	Read length	Accuracy	Price	Speed
Roche 454	up to 700bp	99.9%	10$ / million bases	0.7 Gb / 24 hours
Illumina HiSeq	< 100bp	99.7%	0.07$ / million bases	600 Gb / 11 days
SOLID ABI	< 100bp	99.94%	0.13$ / million bases	120 Gb / 7 days
IonTorrent	~200bp	98.30%	1$ / million bases	1 Gb / 2 hours
PacBio	1500bp	87.00%	2$ / million bases	100 Mb / 2 hours

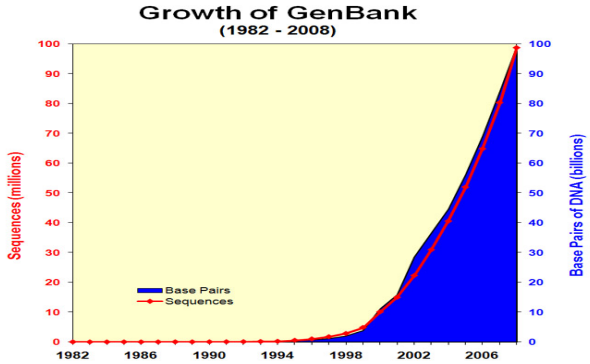
See: Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24:133-41

Source: S. Brohée (BiGRe, ULB)

Sequencing faster and faster

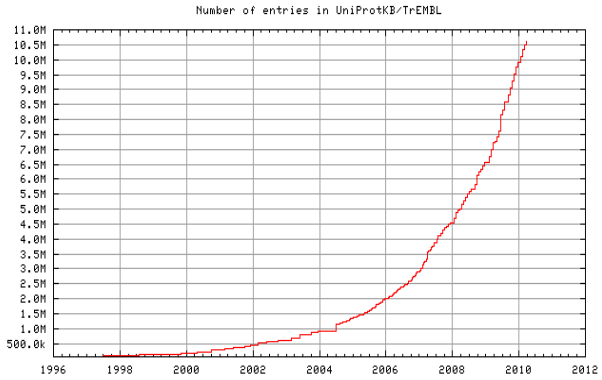
Many genes have been sequenced

Release 211 of Dec 2015 of **GenBank** contains about 189 232 925 gene sequences (203 939 111 071 bases).
http://www.ncbi.nlm.nih.gov/genbank/statistics



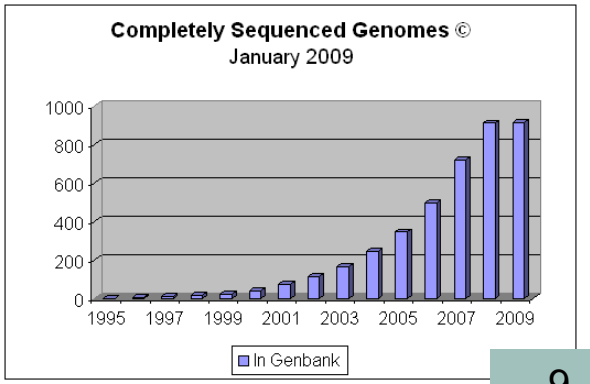
Many proteins have been sequenced

Release 2016_01 of 20-Jan-2016 of **UniProtKB/TrEMBL** contains 59 718 159 sequence (19 944 314 533 aa)
http://www.ebi.ac.uk/swissprot/sptr_stats/full/index.html

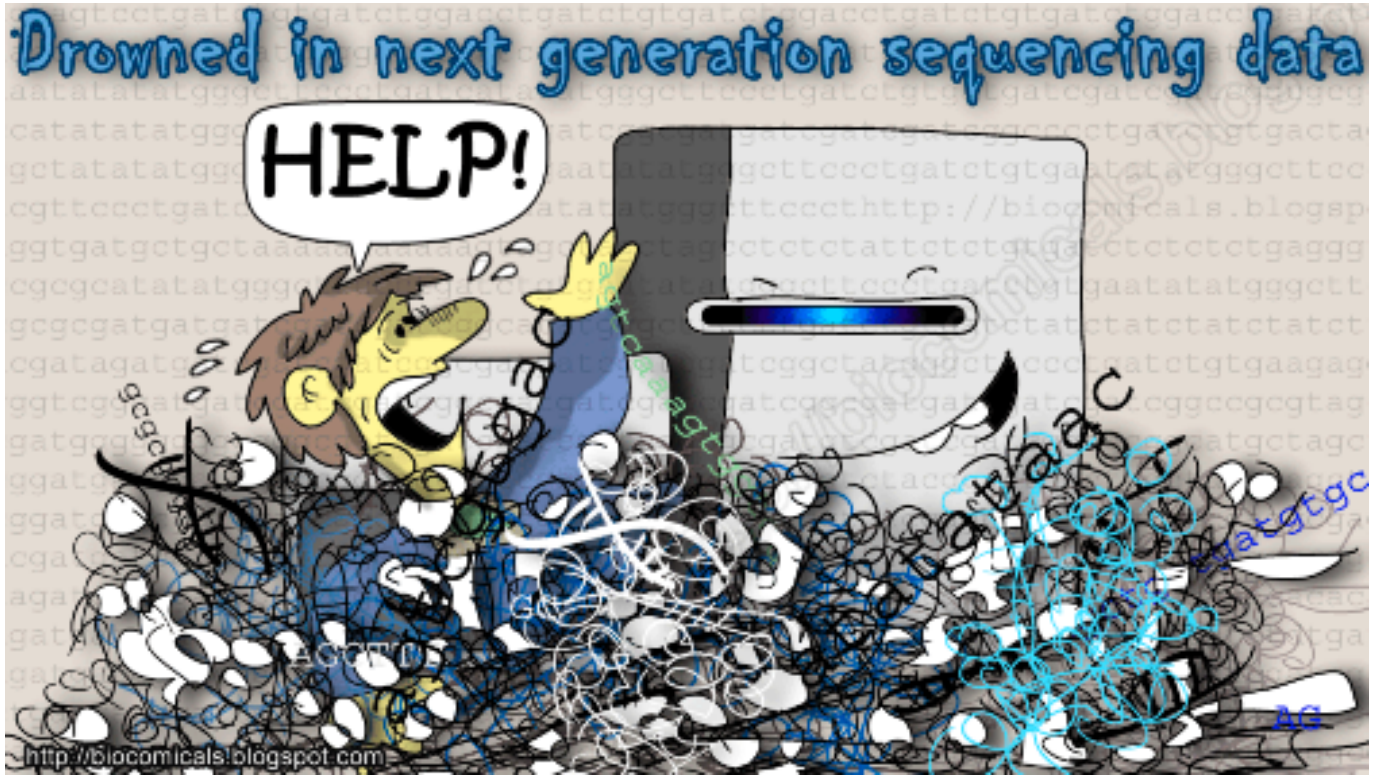


Entire genomes have been sequenced

Release of Jan 2016 of **GOLD** database contains more than 60 000 completely sequenced genomes.
https://gold.jgi.doe.gov/statistics



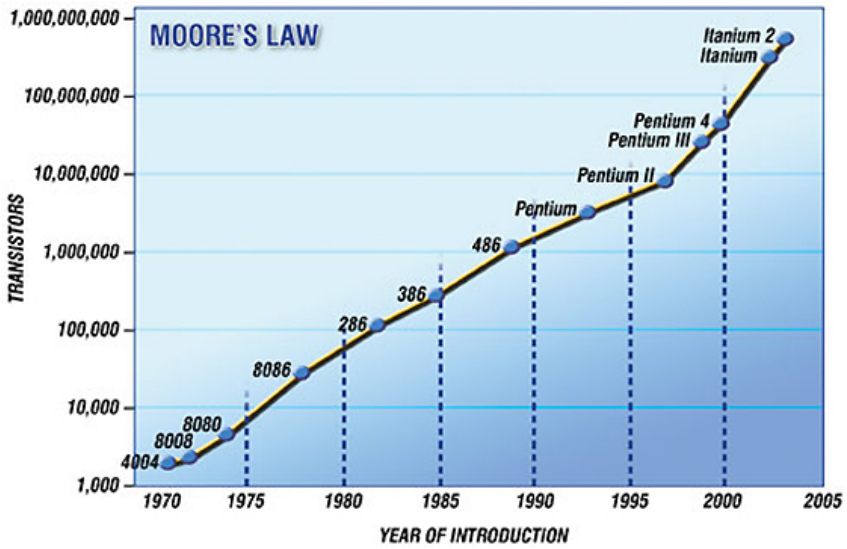
Sequencing faster and faster



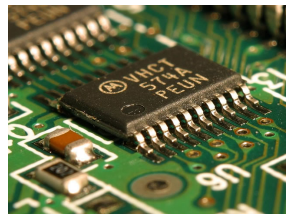
Sequencing faster and faster

The Moore's law

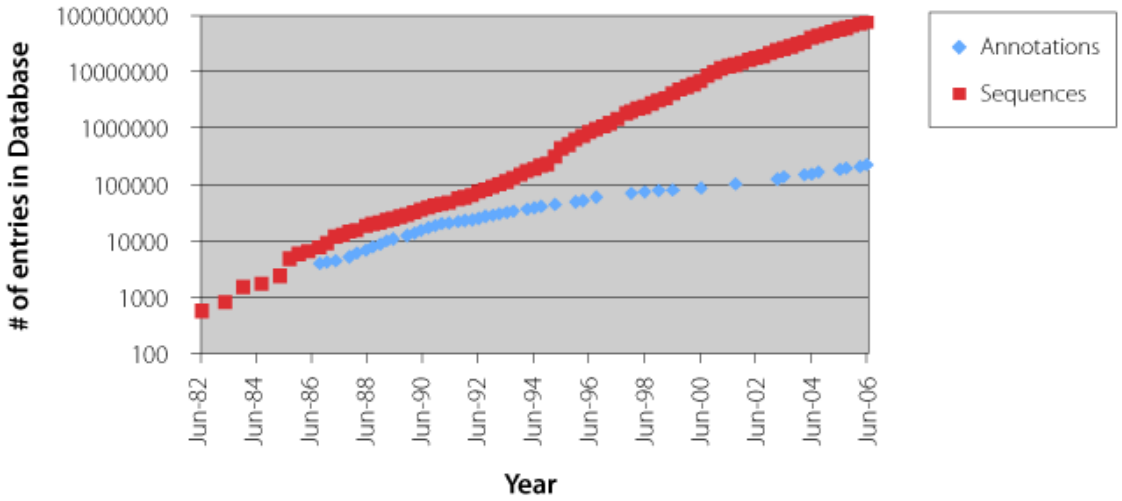
The number of transistors that can be placed on an integrated circuit doubles approximately every two years (from Gordon E. Moore, 1965).



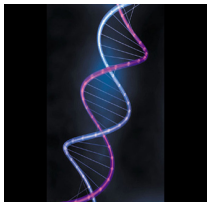
Computer transistors



Growth of sequences and annotations since 1982



DNA sequences



Will cloud computing help genomics handle post-Moore's law data loads?

http://scienceblogs.com/

Sequencing faster and faster

Archon X prize for Genomics

Understanding our genomes may help delay or even prevent disease. For those suffering from genetic illnesses, personal genetic information can determine which medicines will drive their disease into remission without negative side-effects.

The Archon X PRIZE for Genomics challenges scientists and engineers to create better, cheaper and faster ways to sequence genomes. The knowledge gained by compiling and comparing a library of human genomes will create a new era of preventive and personalized medicine and transform medical care from reactive to proactive

http://genomics.xprize.org/

Human Genome

"A scientific milestone of enormous proportions, the sequencing of the human genome will impact all of us in diverse ways-from our views of ourselves as human beings to new paradigms in medicine" (Science, 16 Feb 2001).

Human Genome Project

More and more genomes are available...

- The first completed genomes from viruses, phages and organelles were deposited into the EMBL Database in the early 1980's.

- Since then, major developments in sequencing technology resulted in hundreds of complete genome sequences being added to the database, including Archaea, Bacteria and Eukaryota.

Human Draft Genome

The completion of the human draft genome sequence was announced and published in February 2001 in *Nature* and *Science*. Since the beginning of the Human Genome Project, the international Human Genome Sequencing Consortium has been submitting human draft sequence data to the International Nucleotide Sequence Databases DDBJ/EMBL/GenBank.

Source: http://www.ebi.ac.uk/genomes/

Human Genome Project

The goals of the Human Genome Project are:

- to determine the sequences of the 3 billion chemical base pairs that make up human DNA
- to identify all the approximately 20,000-25,000 genes in human DNA
- to store this information in databases
- to improve tools for data analysis
- to transfer related technologies to the private sector
- to address the ethical, legal, and social issues (ELSI) that may arise from the project.

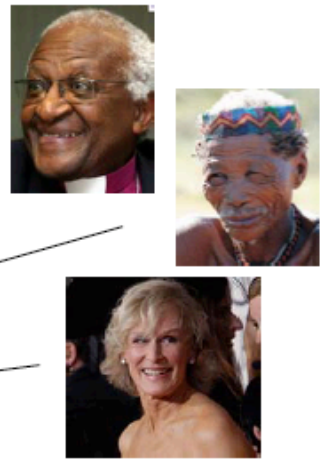
http://www.genome.gov/10001772

http://www.sanger.ac.uk/HGP/

Human Genome Project

First human personal genomes

First versions of human genome	2001-2004
Craig Venter	2007
James Watson	2008
AML (leukemia) patient, normal & cancer tissue	2008
Yoruba, Ibadan, Nigeria (anonymous)	2008
YH (anonymous Han Chinese)	2008
Dan Stoicescu (swiss millionaire)	2008
Stephen Quake (Stanford univ. professor)	2009
Seong-Jin Kim, Korean researcher	2009
James Lupski (Charcot-Marie-Tooth disease)	2010
Desmund Tutu and !Gubi, Southern African	2010
Glenn Close (actress)	2010



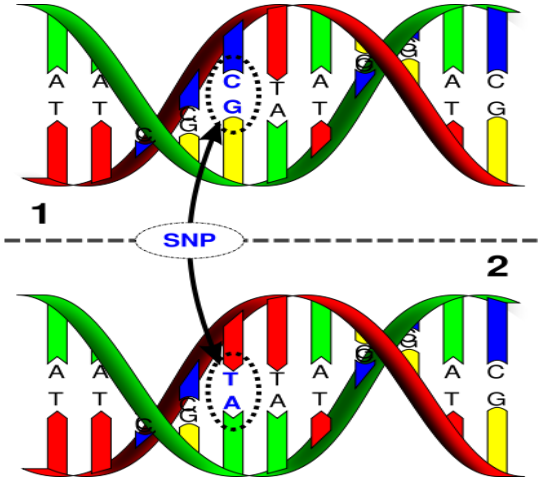
Source: T. Samuelsson

Human Genome Project

Human Genome Project: Revolution postponed?

- The human genome project (HGP) has failed sa far to produce the medical miracles that scientists promised.
- Bill Clinton (2000) predicted that the HGP would "revolutionize the diagnosis, prevention, and treatment of most - if not all - human diseases".
- In 2010, the scientific community finds itself sobered and divided. Goldstein (director of the Center for Human Genome Variation at Duke Univ): "It is fair to say that we are not going to be personalizing the treatment of common disease next year".
- Some steps are done however in the direction of a better understanding and quantification of the human genetic variations. Database of single-nucleotide polymorphisms (SNP) and their (statistical) association to diseases are now being developed (cf. the **HapMap Project**: http://www.genome.gov/10001688).

Source: *Sci. Am.,* October 2010, pp. 42-49



A SNP is a single nucleotide polymorphism. The Human Genome actually has very little variation between individuals. It is estimated that for every 1000 base pairs of DNA, there is only 1base pair of DNA is different. But with 3.3 billion base pairs to consider, there is still a lot of variation to understand!

Ex: Sickle cell anemia (blood disorder) is due to SNP in the beta globin HBB gene.

Source: http://www.snpedia.com

The "1000 genomes" project

2008

1000 Genomes

A Deep Catalog of Human Genetic Variation

Home About Partners Data Contact Internal

INTERNATIONAL CONSORTIUM ANNOUNCES THE 1000 GENOMES PROJECT

Major Sequencing Effort Will Produce Most Detailed Map Of Human Genetic Variation to Support Disease Studies

An international research consortium has been formed to create the most detailed and medically useful picture to date of human genetic variation. The 1000 Genomes Project will involve sequencing the genomes of at least a thousand people from around the world. The project will receive major support from the Wellcome Trust Sanger Institute in Hinxton, England, the Beijing Genomics Institute Shenzhen in China and the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH).

Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

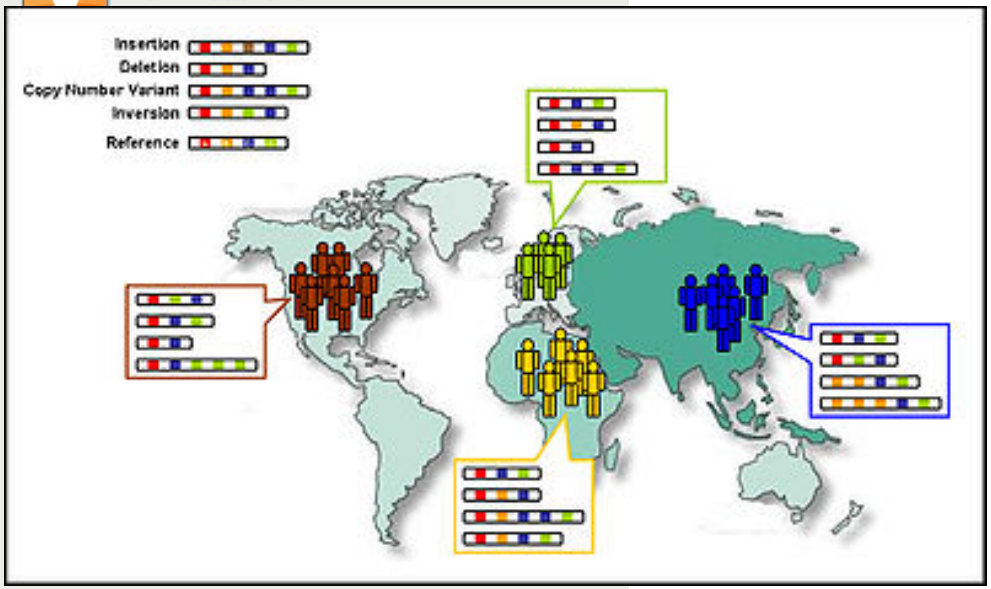
PRESS RELEASE

TUESDAY JAN. 22, 2008

International Consortium Announces the 1000 Genomes Project

LINKS

Download the



Source: http://www.1000genomes.org/

Kaiser (2008) A Plan to Capture Human Diversity in 1000 Genome. *Science* 319: 395

The "1000 genomes" project

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother-father-child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

28 OCTOBER 2010 | VOL 467 | NATURE | 1061

The "1000 genomes" project

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and comp
platforms. We undertook three proje
populations; high-coverage sequencin
individuals from seven populations. W
approximately 15 million single nucle
structural variants, most of which were
majority of common variation, over 95%
data set. On average, each person is fou
genes and 50 to 100 variants previously i
to inform association and functional stu
substitution mutations to be approxima
signatures of natural selection, and ide
due to selection at linked sites. These m

Highlights

- Each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders.
- From the two trios [mother-father-child], we directly estimate the rate of base substitution mutations to be approximately 10^{-8} per base pair per generation.

28 OCTOBER 2010 | VOL 467 | NATURE | 1061

The "1000 genomes" project

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

NATURE | VOL 491 | 1 NOVEMBER 2012

The "1000 genomes" project

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to provide a validated haplotype map of deletions, and more than 14,000 larger profiles of rare and common variants, which is further increased by the actio consequence are key determinants of t across biological pathways, and that eac such as motif-disrupting changes in tr accessible single nucleotide polymorphi low-frequency variants in individuals f

Highlights

- Some of the rarest DNA variants tend to cluster in relatively restricted geographic areas.
- This study lead to the most complete inventory of the millions of variations between people's DNA sequences

NATURE | VOL 491 | 1 NOVEMBER 2012

The "100 000 genomes" project

launched in 2012...

The 100,000 Genomes Project

The project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.

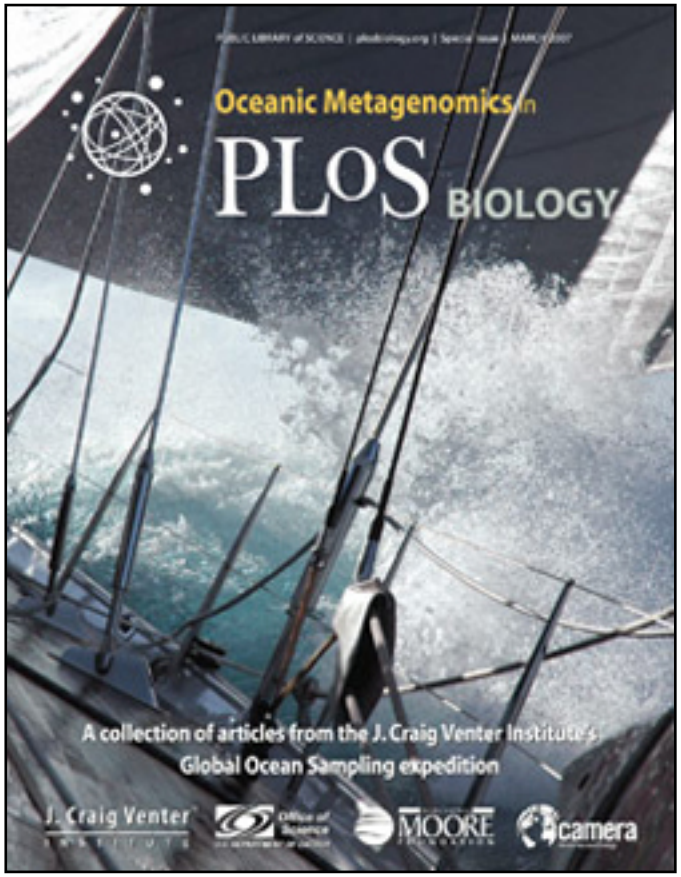
The aim is to create a new genomic medicine service for the NHS – transforming the way people are cared for. Patients may be offered a diagnosis where there wasn't one before. In time, there is the potential of new and more effective treatments.

The project will also enable new medical research. Combining genomic sequence data with medical records is a ground-breaking resource. Researchers will study how best to use genomics in healthcare and how best to interpret the data to help patients. The causes, diagnosis and treatment of disease will also be investigated. We also aim to kick-start a UK genomics industry. This is currently the largest national sequencing project of its kind in the world.

NHS = National Health Service (England)

https://www.genomicsengland.co.uk/the-100000-genomes-project/

Metagenomics



Craig Venter Institute

http://www.jcvi.org/

2007: large-scale metagenomic project

"More than six million new genes, thousands of new protein families, and incredible degree of microbial diversity discovered from ocean microbes using whole environment shotgun sequencing and new computational tools."

Reference: http://biology.plosjournals.org/

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* 5(3): e77

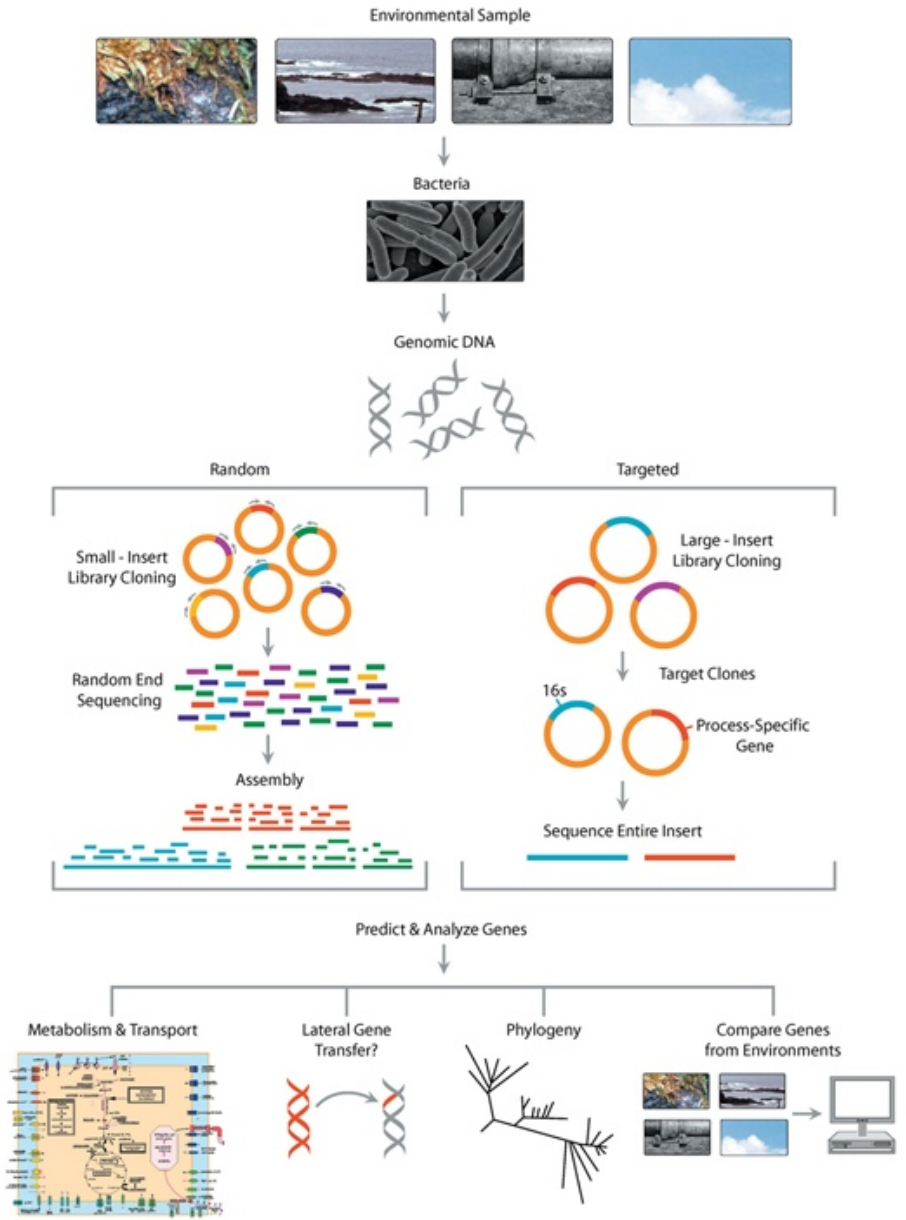
+ other papers in the same issue of *PLoS biol*.

See also:

Daniel (2005) The metagenomics of soil, *Nat Rev Microbio* 3: 470-478.
Gill et al (2006) Metagenomic analysis of the human distal gut microbiome. *Science*. 312:1355-9

- One gram of forest soil contains an estimated 4.10^7 prokaryotic cells, whereas one gram of cultivated soils and grasslands contains an estimated 2.10^9 prokaryotic cells.
- The human body (about 10^{13} cells) hosts an estimated 100 trillion (10^{14}) bacterial cells from at least 500 species of bacteria.

Metagenomics



Source: http://camera.calit2.net/education/what-is-metagenomics

Metagenomics

A metagenomics project begins with the isolation of DNA from a mixed microbial population collected from any given environment (e.g. sewage digestor, dental plaque, gut, your computer keyboard, ocean samples, etc). Environmental DNA is then sheared into fragments that are used in construction of a DNA clone library (either small- or medium-insert libraries (2-15 kb insert size) or large-insert (up to 150 kb insert size)), that may be sequenced in either a random or targeted fashion.

In a **random sequencing approach**, the clones are randomly chosen and sequenced, and the resulting sequences are assembled into larger contiguous pieces ("contigs") by matching up overlapping sequences. The resulting data are contigs of different lengths as well as shorter unassembled fragments. The availability of completely sequenced "reference" genomes may assist in the assembly process for closely related genomes. In the absence of this, contigs may be assigned to various "bins" based on their G+C content, codon usage, sequence coverage, presence of short n-mers (nucleotide frequency), and other parameters, allowing them to be sorted into groups that can be viewed as a "species". Coding sequences (CDSs, or colloquially "genes") are then predicted from these sequence data using various methods. Often in the random sequencing approach, identified genes may not be attributable to a particular microbial species (i.e., there is no taxonomic or phylogenetic affiliation). These nonetheless represent abilities of the general microbial community and may reveal characteristics of their environment.

In a **"targeted" sequencing approach**, clones are first screened for the presence of a desirable gene (e.g., by PCR amplification) or a gene function (by functional assay). Sequencing targeted large-insert clones in their entirety allows the possibility of recovering complete operons, e.g., those encoding metabolic pathways.

Metagenomics

soil

ocean

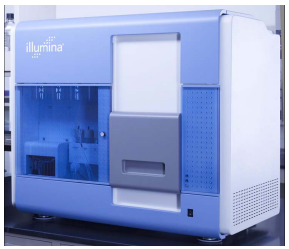
gut

sampling

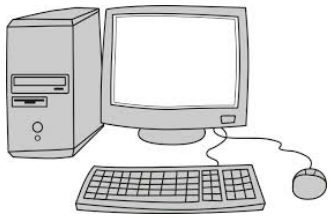
DNA extraction

DNA amplification with "universal" primers

DNA sequencing



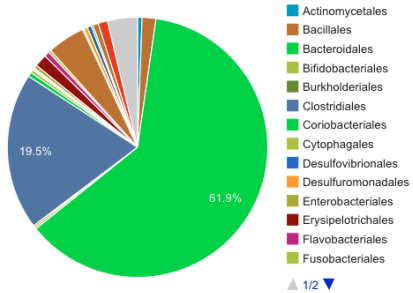
sequence processing (quality filtering, taxonomic assignment)



reference database

microbial composition

Metagenomics workflow



16S sequencing workflow

Source: K. Faust

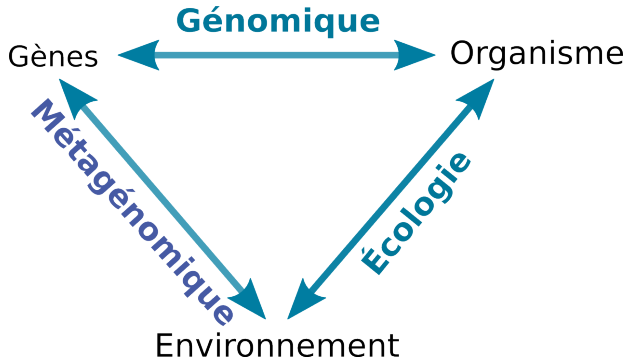
Metagenomics

Metagenomics: Tools & Applications

- Unbiased method to investigate microbial inhabitants of an ecosystem ("who is there")
- Ecosystem: soil, ocean, waste water, human body
- (Bioinformatics) Tools: whole genome assembly, discovery of new variants of known genes, discovery of associations (co-occurence), etc.
- Ecological and medical applications.
  - Ex 1: Biomarkers for diseases (ex: obesity <-> specific species)

    Erickson AR et al (2012) Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One. 7:e49138
  - Ex 2: Effect of antibiotics

    Peterfreund GL et al (2012) Succession in the gut microbiome following antibiotic and antibody therapies for Clostridium difficile. PLoS One. 7:e46966



Source: Karoline Faust (KUL)

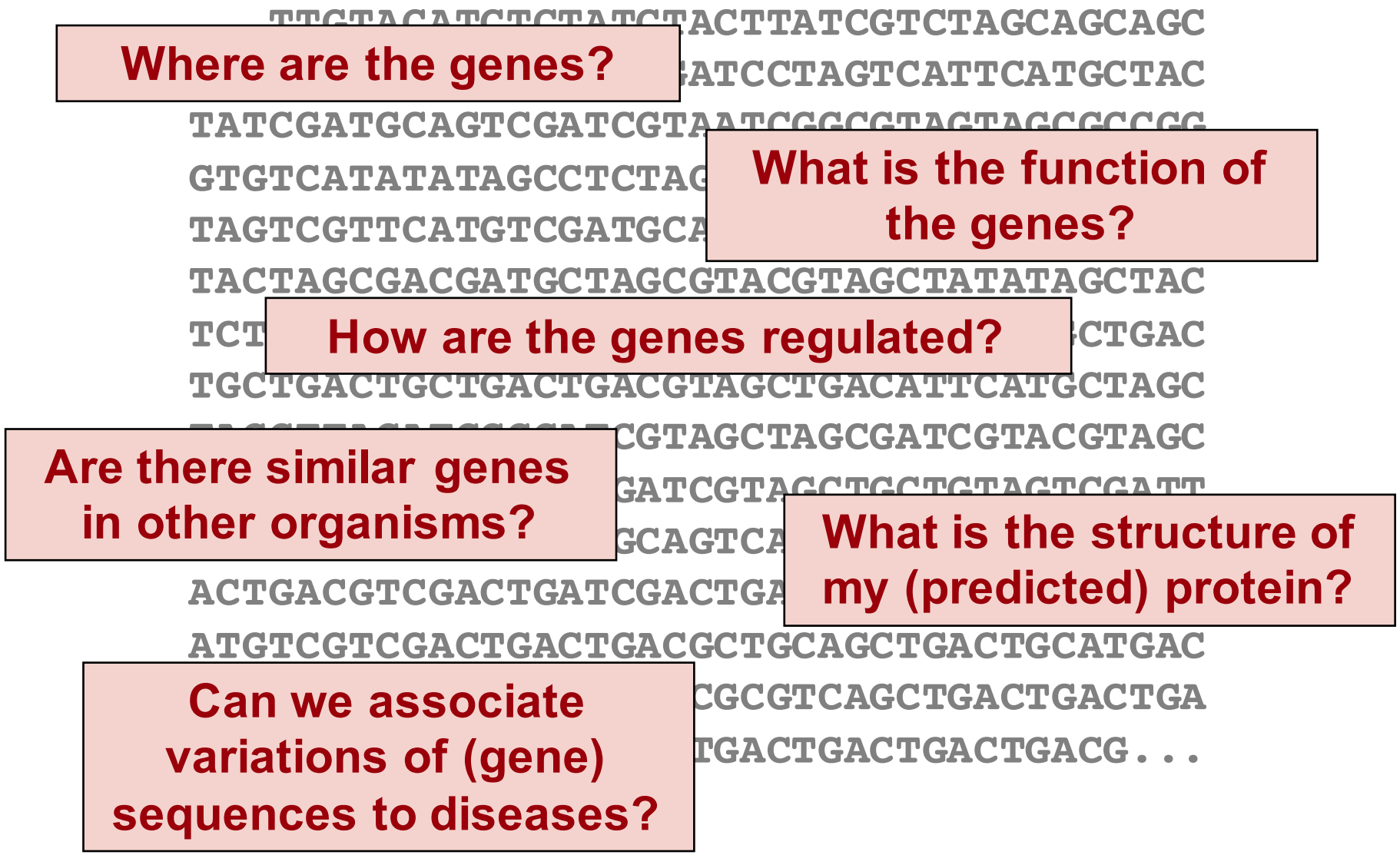
We have the genome.

```
...TTGTACATCTCTATCTACTTATCGTCTAGCAGCAGC
TACTGATCGTAGTCTCGTGATCCTAGTCATTCATGCTAC
TATCGATGCAGTCGATCGTAATCGGCGTAGTAGCGCCGG
GTGTCATATATAGCCTCTAGCGCTAGCAGCTGATCGATC
TAGTCGTTCATGTCGATGCAGCTAGTCTAGTCGTATCTA
TACTAGCGACGATGCTAGCGTACGTAGCTATATAGCTAC
TCTGATATACTGCCGCTAGTACGTACTGCAGCAGCTGAC
TGCTGACTGCTGACTGACGTAGCTGACATTCATGCTAGC
TAGCTTACATCGCGATCGTAGCTAGCGATCGTACGTAGC
GCCTAGCGGTACTTGCGATCGTAGCTGCTGTAGTCGATT
GTGCGATAGTCACTGTGCAGTCAGTCGATCGATCGACTG
ACTGACGTCGACTGATCGACTGACTGACTGACTGACTGC
ATGTCGTCGACTGACTGACGCTGCAGCTGACTGCATGAC
GTCGACTGATGACTGACTGCGCGTCAGCTGACTGACTGA
CTGACTGACTGTCAGTGACTGACTGACTGACTGACG...
```

We have the genome. And so what?

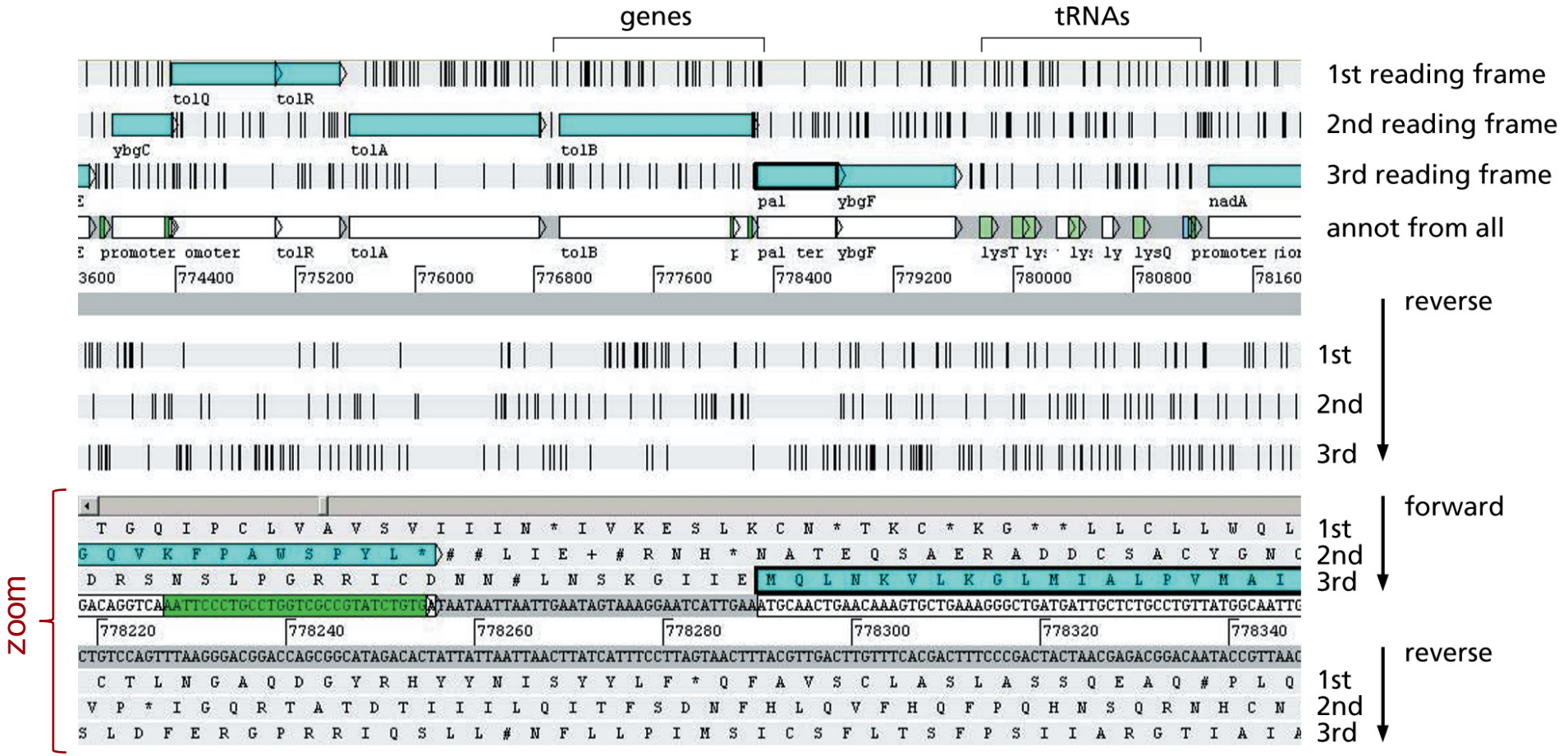
...TTGTACATCTCTATCTACTTATCGTCTAGCAGCAGC
TACTGATCGTAGTCTCGTGATCCTAGTCATTCATGCTAC
TATCGA
GTGTC
TAGTC
TACTA
TCTGA
TGCTG
TAGCT
GCCTA
GTGCG
ACTGA
ATGTC
GTCGA
CTGAC

We have the genome. And so what?



We have the genome. And so what?

Genome annotation

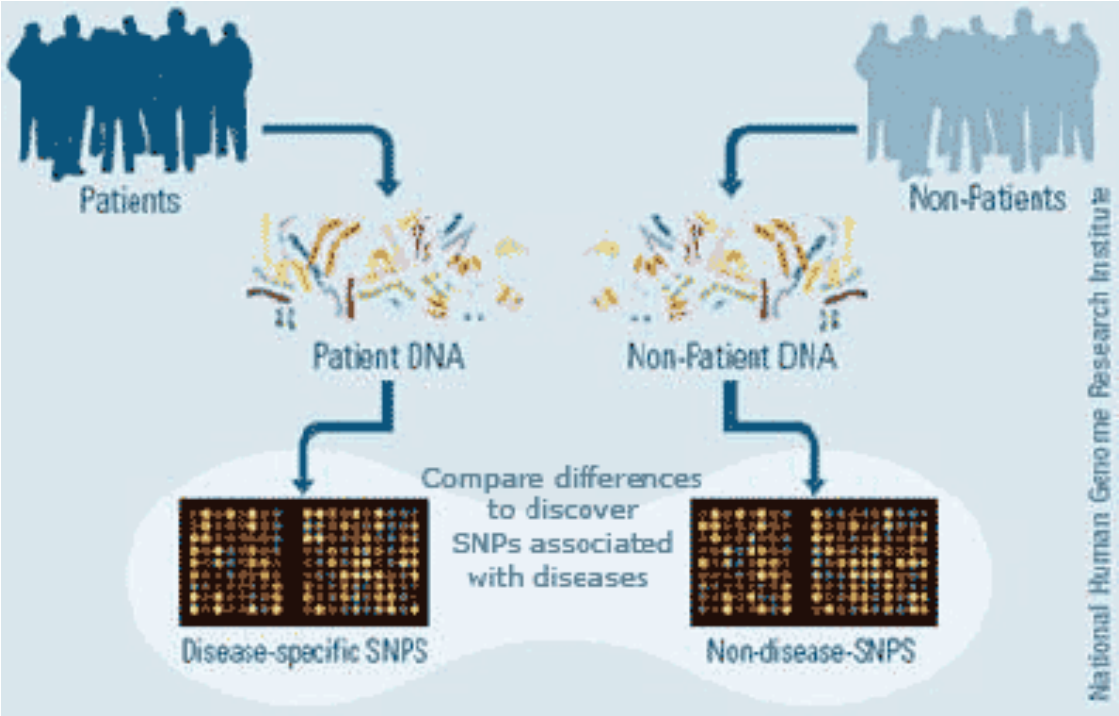


Source: Zvelebil & Baum

We have the genome. And so what?

GWAS

Genome-wide association studies



The less than 1% of DNA that differs among individuals contributes to all aspects that make each of us unique. More than simply physical features, abilities, and personality; health and disease risk differences are also part of this small, but powerful difference between individuals.

Source: http://cpmc.coriell.org/Sections/Medical/GeneInteraction_mp.aspx?PgId=93

Towards personalized medicine...

Towards a personalized medicine...

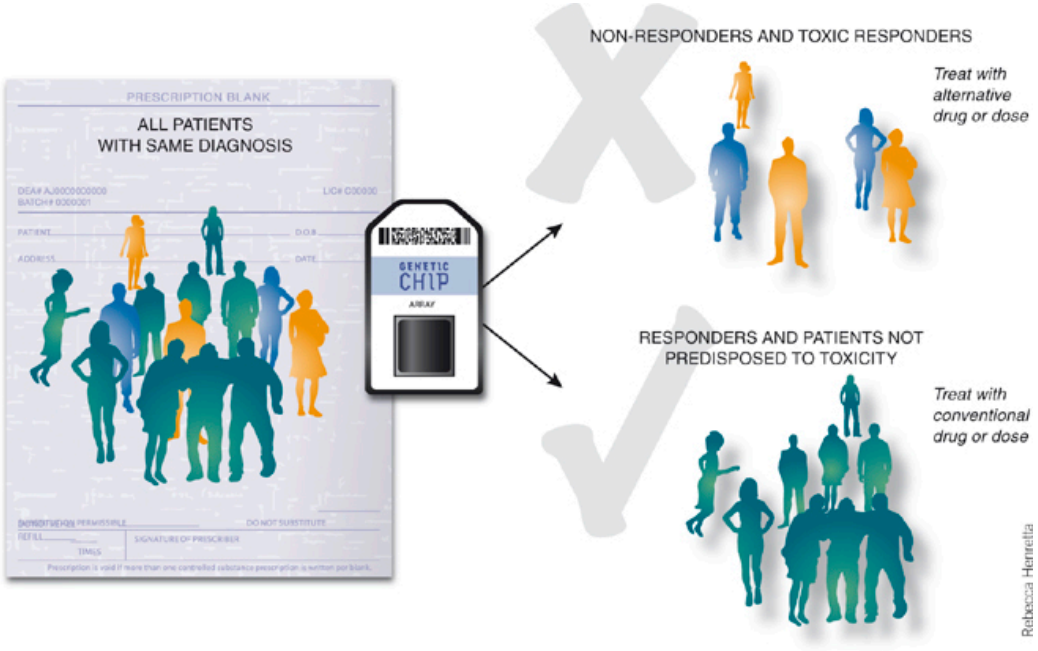


Figure from Piquette-Miller & Grant (2007) The Art and Science of Personalized Medicine, *Clin Pharmacol Ther.* 81:311-5.

More than 1,000 DNA variants associated with diseases and traits have been identified.

"Direct-to-consumer" companies are harnessing these discoveries by offering DNA tests that provide insights into personal genetic traits and disease risks.

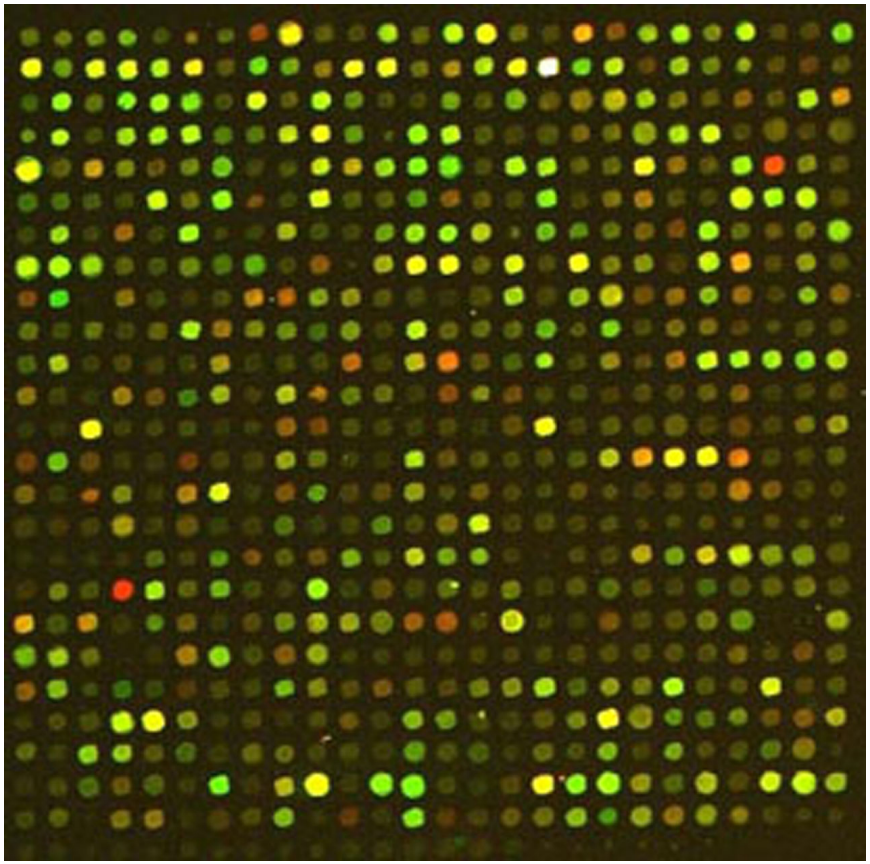
Genetic testing can improve lifestyle choices and increase preventive screening.

However, understanding of the genetic contribution to human disease is far from complete.

Ng, Murray, Levy, Venter (2009) An agenda for personalized medicine, *Nature* 724-726

High throughput technologies

DNA sequences are not the only type of genome-scale data...



Microarray (gene expression)

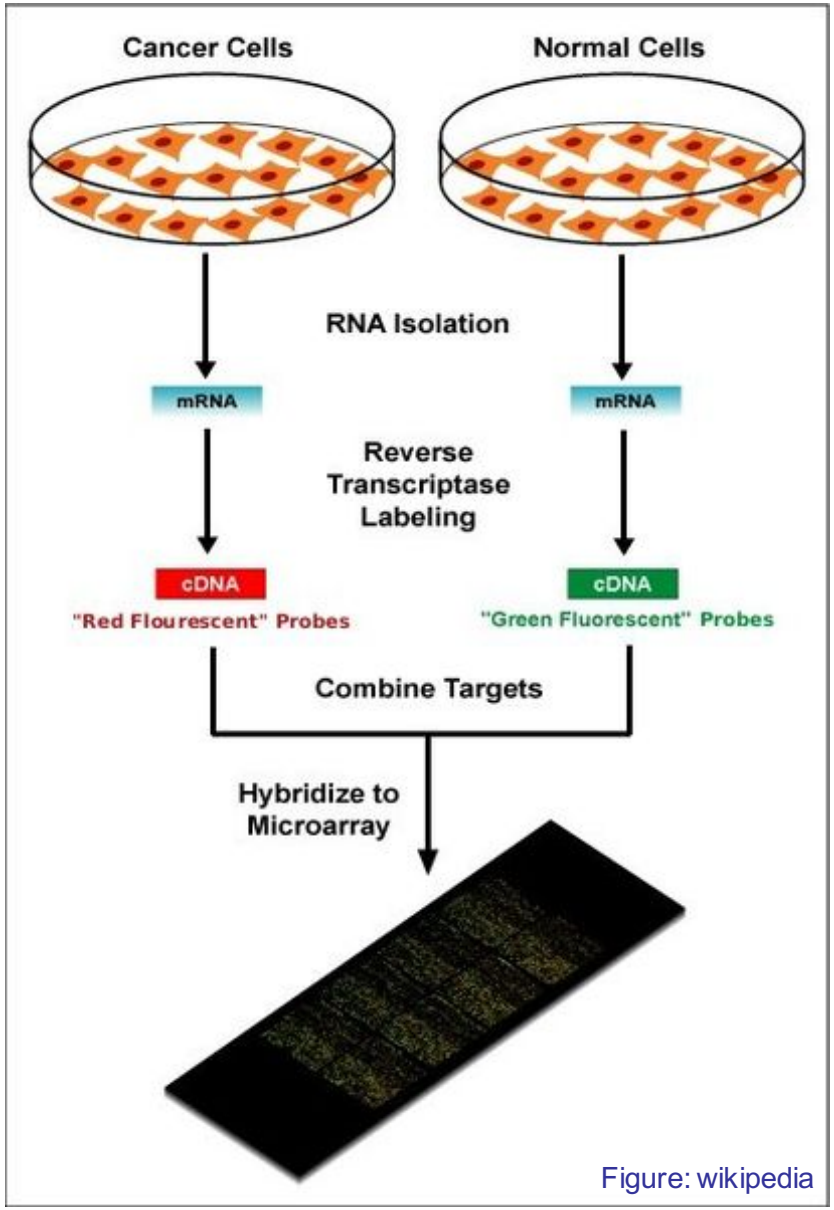
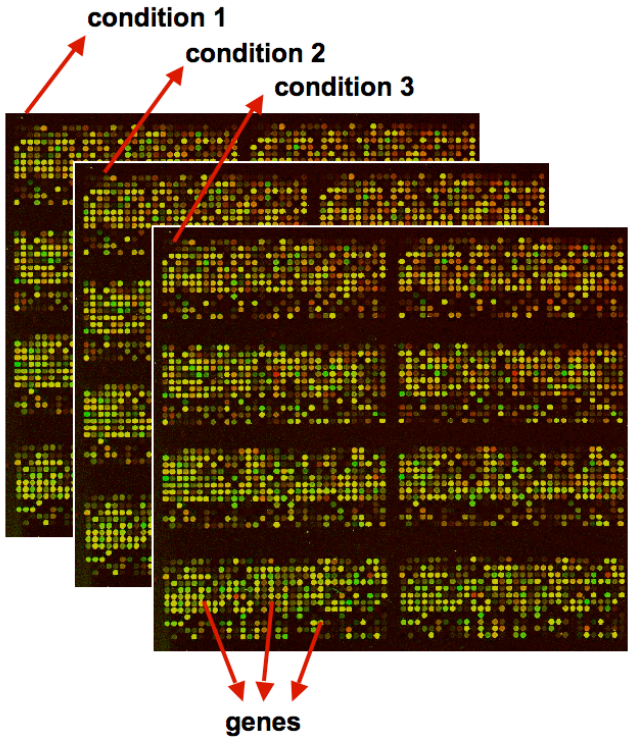


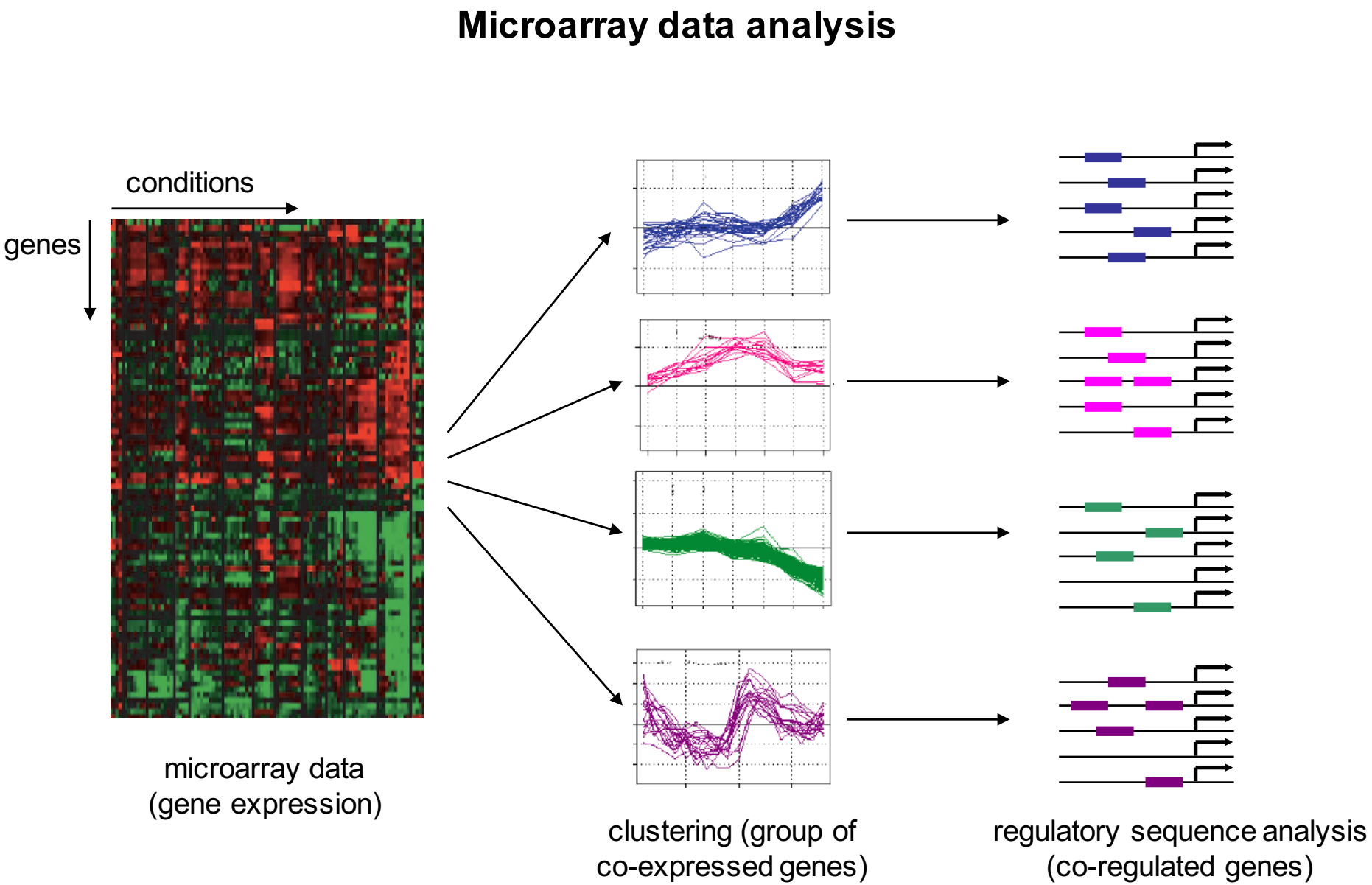
Figure: wikipedia



DNA microarrays can be used to measure the expression levels of large numbers of genes simultaneously (compared to a reference level of expression). When such experiment is repeated over several conditions (environment, mutants, phase of the cycle cycle, etc), it is possible to determine clusters of co-expressed (possibly co-regulated) genes.

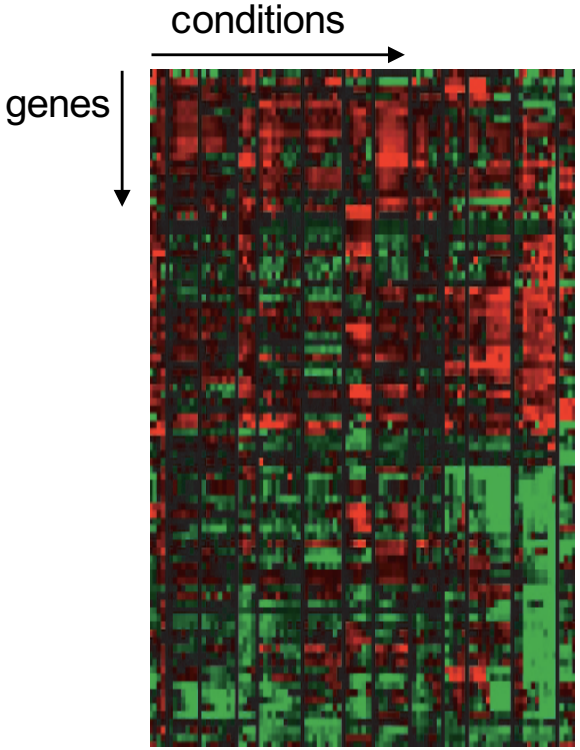
Schena M, Shalon D, Davis RW, Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467-470

Microarray (gene expression)

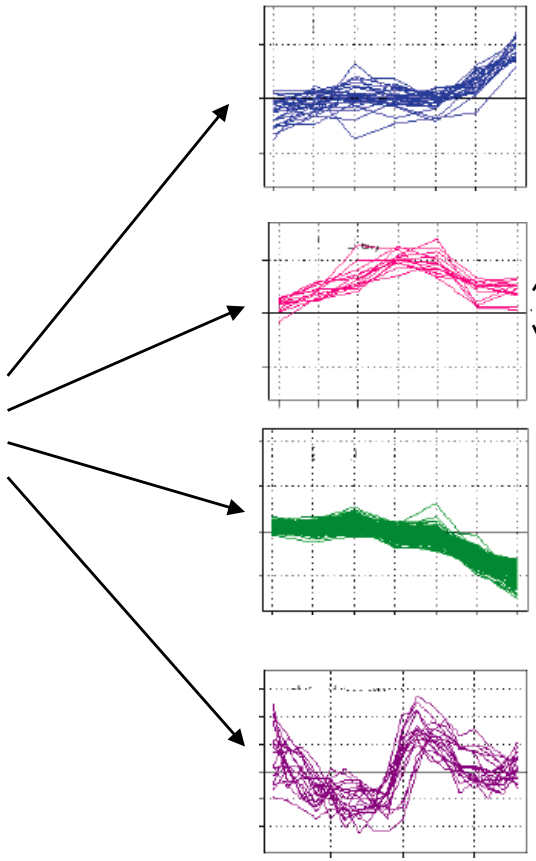


Microarray (gene expression)

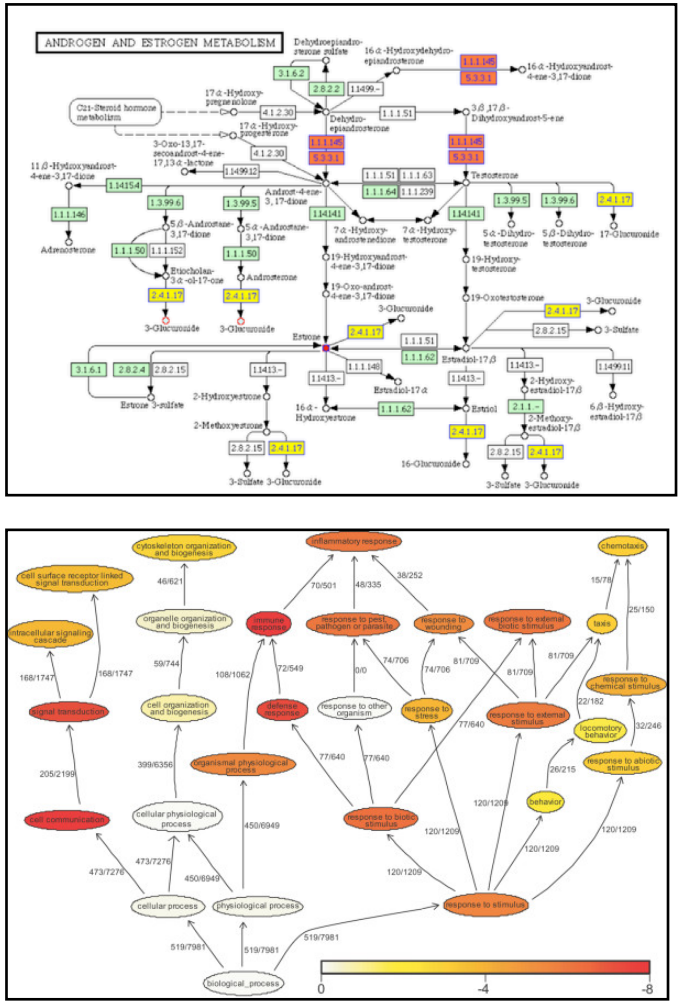
Microarray data analysis



microarray data
(gene expression)



clustering (group of
co-expressed genes)



mapping to metabolic or other
functional databases.

Microarray (gene expression)

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

SCIENCE • VOL. 278 • 24 OCTOBER 1997 • www.sciencemag.org

Molecular Biology of the Cell
Vol. 9, 3273–3297, December 1998

Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization

Paul T. Spellman,*† Gavin Sherlock,*†‡ Michael Q. Zhang,‡ Vishwanath R. Iyer,§ Kirk Anders,* Michael B. Eisen,* Patrick O. Brown,§‖ David Botstein,*¶ and Bruce Futcher‡

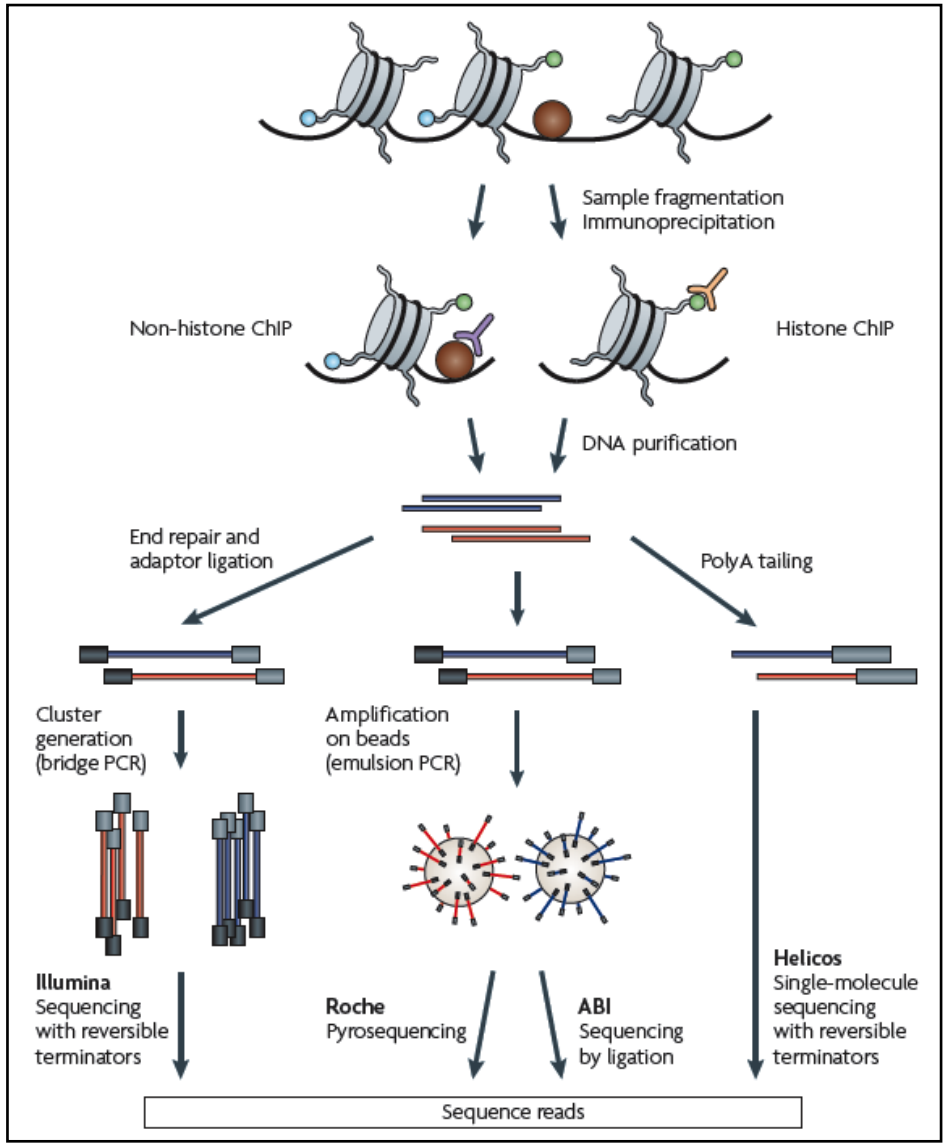
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2]*† D. K. Slonim,[1]† P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5]*

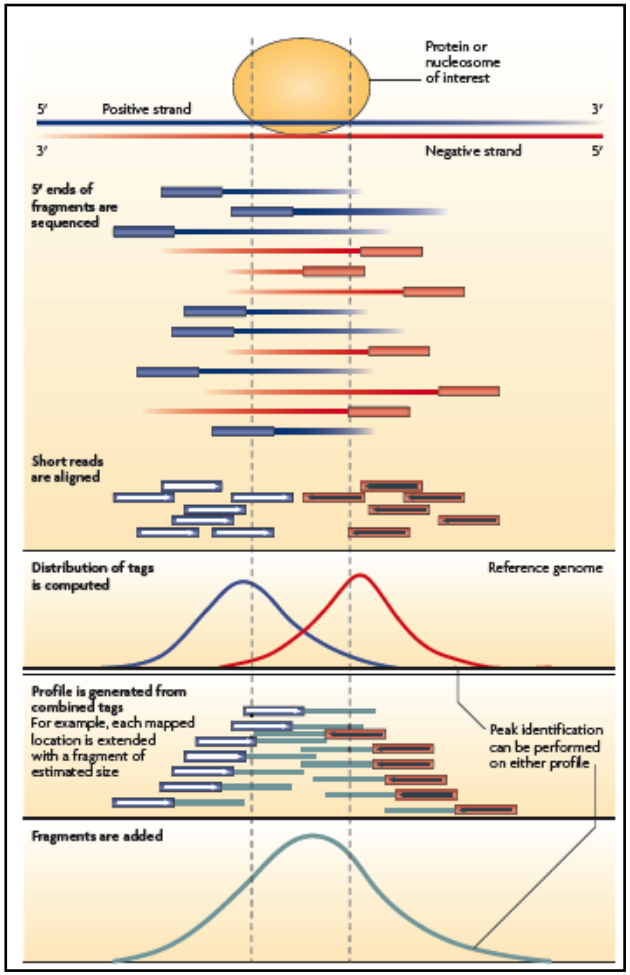
Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

www.sciencemag.org SCIENCE VOL 286 15 OCTOBER 1999

ChIP-seq (DNA binding protein)



Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 10:669-80.



ChIP-sequencing (ChIP-Seq), is a method used to detect protein-DNA interactions. It combines chromatin immunoprecipitation (ChIP) with massive DNA sequencing to identify the binding sites of DNA-associated proteins.

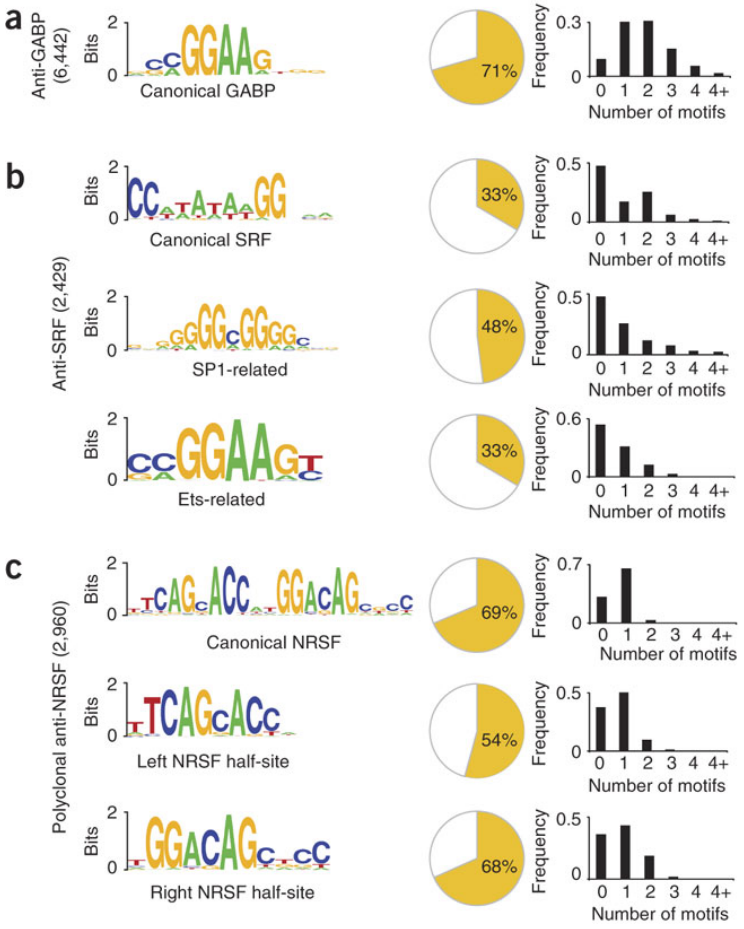
ChIP-seq (DNA binding protein)

Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data

Anton Valouev[1,4], David S Johnson[2,4], Andreas Sundquist[3], Catherine Medina[2], Elizabeth Anton[2], Serafim Batzoglou[3], Richard M Myers[2] & Arend Sidow[1,2]

Molecular interactions between protein complexes and DNA mediate essential gene-regulatory functions. Uncovering such interactions by chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-Seq) has recently become the focus of intense interest. We here introduce quantitative enrichment of sequence tags (QuEST), a powerful statistical framework based on the kernel density estimation approach, which uses ChIP-Seq data to determine positions where protein complexes contact DNA. Using QuEST, we discovered several thousand binding sites for the human transcription factors SRF, GABP and NRSF at an average resolution of about 20 base pairs. MEME motif-discovery tool–based analyses of the QuEST-identified sequences revealed DNA binding by cofactors of SRF, providing evidence that cofactor binding specificity can be obtained from ChIP-Seq data. By combining QuEST analyses with Gene Ontology (GO) annotations and expression data, we illustrate how general functions of transcription factors can be inferred.

NATURE METHODS | VOL.5 NO.9 | SEPTEMBER 2008 | 829



Yeast two-hybrid (protein interactions)

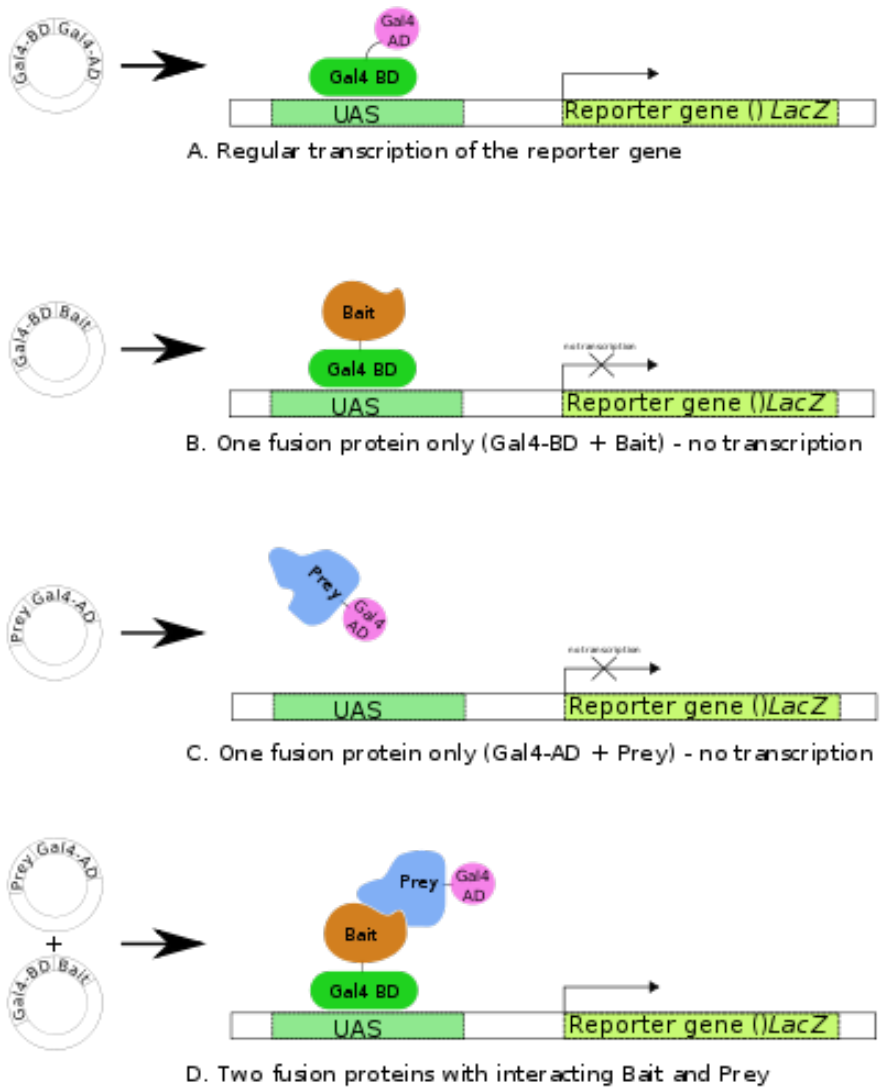


Figure from wikipedia

Two-hybrid screening (also known as yeast two-hybrid system or Y2H) is a technique used to detect physical protein-protein interactions.

Fields S, Song O (1989) A novel genetic system to detect protein–protein interactions. Nature 340:245–246.

Chien CT, Bartel PL, Sternglanz R, Fields S (1991) The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci USA*. 88:9578-82.

Uetz P, Giot L, Cagney G, Mansfield TA et al (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*. 403:623-7.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci.* 98:4569-74

Yeast two-hybrid at genomic scale

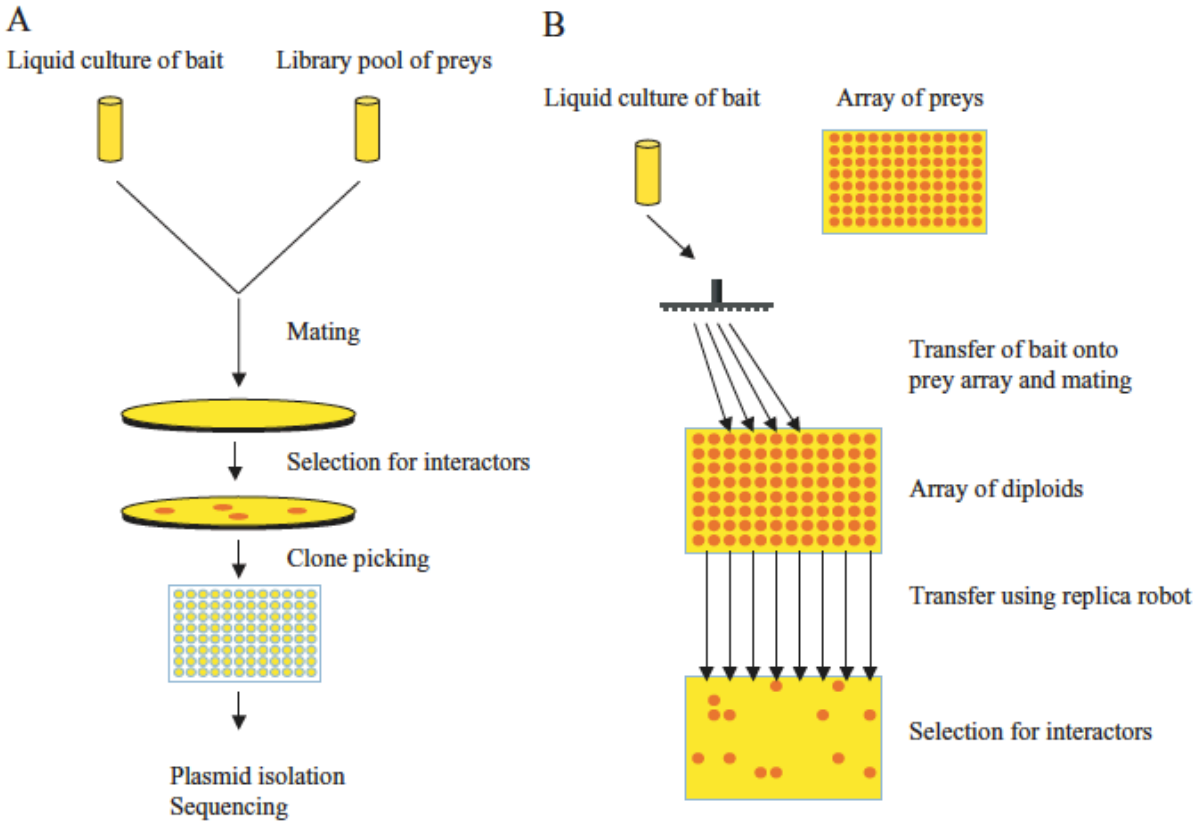


Figure 2.2. High-throughput approaches utilizing the yeast two-hybrid system. (**A**) The library screening approach. A yeast strain expressing a bait under investigation is mixed with a collection of yeast strains each expressing a random prey from a library. Incubation in rich medium allows the two strains to mate and diploids expressing bait and prey are selected. The diploids are then transferred to selective medium to isolate those clones containing interacting baits and preys (selection for interactors). Yeast clones that display growth on selective medium are picked up, transferred into multiwell plates, and processed for plasmid isolation and insert sequencing to identify the interacting prey. (**B**) The matrix or array approach. An array of preys is prepared by spotting yeast clones each expressing a known prey onto plates. The colonies on the array are then picked up by a robot and mated with a yeast strain expressing the bait under investigation. An exact replica of the array is transferred to a fresh plate to select for diploids expressing bait and prey and then to selective medium to select for interacting baits and preys. The identity of the prey in colonies that grow under selection is determined by its position within the array.

Auerbach & Stagljar, Proteomics and Protein–Protein Interactions: Biology, Chemistry, Bioinformatics, and Drug Design, ed. Waksman. Springer, 2005.

Yeast two-hybrid at genomic scale

A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*

Peter Uetz*†, Loic Giot*‡, Gerard Cagney†, Traci A. Mansfield‡, Richard S. Judson‡, James R. Knight‡, Daniel Lockshon†, Vaibhav Narayan‡, Maithreyan Srinivasan‡, Pascale Pochart‡, Alia Qureshi-Emili†§, Ying Li‡, Brian Godwin‡, Diana Conover†§, Theodore Kalbfleisch‡, Govindan Vijayadamodar‡, Meijia Yang‡, Mark Johnston†‖, Stanley Fields†§ & Jonathan M. Rothberg‡

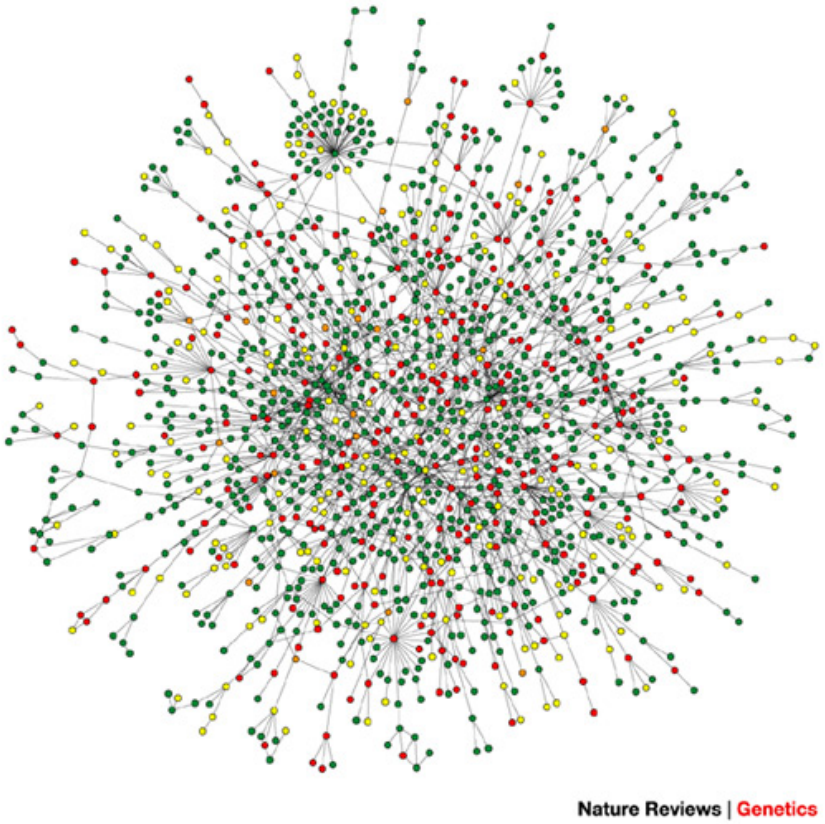
‡ CuraGen Corporation, 555 Long Wharf Drive, 11th Floor, New Haven, Connecticut 06511, USA
† Departments of Genetics and Medicine and § Howard Hughes Medical Institute, University of Washington, Box 357360, Seattle, Washington 98195-7360, USA
** These authors contributed equally to this work*

Two large-scale yeast two-hybrid screens were undertaken to identify protein–protein interactions between full-length open reading frames predicted from the *Saccharomyces cerevisiae* genome sequence. In one approach, we constructed a protein array of about 6,000 yeast transformants, with each transformant expressing one of the open reading frames as a fusion to an activation domain. This array was screened by a simple and automated procedure for 192 yeast proteins, with positive responses identified by their positions in the array. In a second approach, we pooled cells expressing one of about 6,000 activation domain fusions to generate a library. We used a high-throughput screening procedure to screen nearly all of the 6,000 predicted yeast proteins, expressed as Gal4 DNA-binding domain fusion proteins, against the library, and characterized positives by sequence analysis. These approaches resulted in the detection of 957 putative interactions involving 1,004 *S. cerevisiae* proteins. These data reveal interactions that place functionally unclassified proteins in a biological context, interactions between proteins involved in the same biological function, and interactions that link biological functions together into larger cellular processes. The results of these screens are shown here.

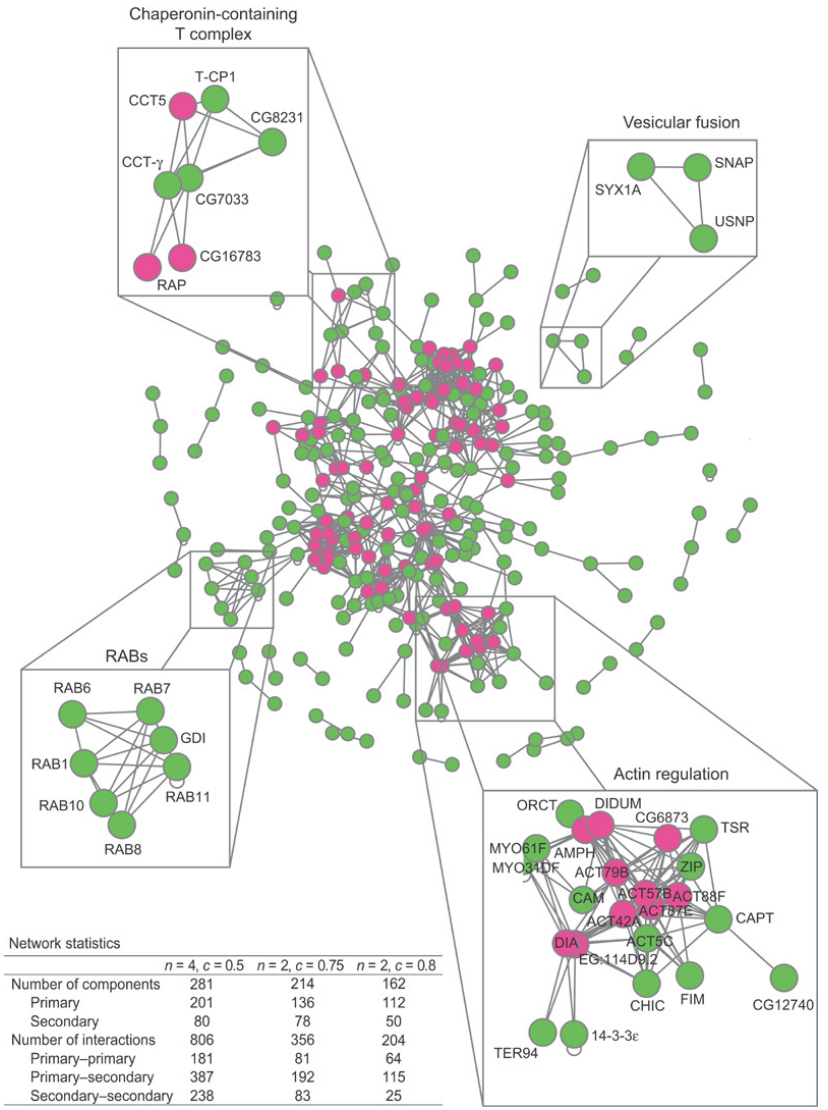
Yeast two-hybrid at genomic scale

Protein-protein interaction networks

Yeast protein interaction network



Barabasi & Oltvai (2004) Nature Rev Genet 5:101-113



Network statistics

	$n = 4, c = 0.5$	$n = 2, c = 0.75$	$n = 2, c = 0.8$
Number of components	281	214	162
Primary	201	136	112
Secondary	80	78	50
Number of interactions	806	356	204
Primary–primary	181	81	64
Primary–secondary	387	192	115
Secondary–secondary	238	83	25

Stuart et al (2007) *Nature* 445:95-101

Large-scale analyses

A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*

Efficient genome-wide mutagenesis of zebrafish genes by retroviral insertions

Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes[D]

Genome-wide association study of 14,000 cases of seven common diseases and

A comprehensive analysis of protein–protein interactions in *Saccharo*

Large-scale analysis of the yeast genome by transposon tagging and gene disruption

Large-scale mapping of human protein–protein interactions by mass spectrometry

A comprehensive two-hybrid analysis to explore the yeast protein interactome

High throughput technologies

- **Genome projects** stimulated drastic improvement of sequencing technology.
- **Post-genomic era**:
  - Genome sequence was not sufficient to predict gene function
  - This stimulated the development of new experimental methods
    - transcriptomics (microarrays)
    - proteomics (2-hybrid, mass spectrometry, ...)

Warning: the "omics" trends

transcriptome, proteome, metabolome, interactome, ORFeome, reactome, diseasome (diseases), lipidomes (lipides), kinome (kinases), ...

- The few real high throughput methods raised a fashion of "omics", which introduced more confusion than progress.
- Some of the "omics" are not associated to any new/high throughput approach, this is just a new name on a previous method, or on an abstract concept.

Source: Jacques van Helden

High throughput technologies

- **The availability of massive amounts of data enables to address questions that could not even be imagined a few years ago**
  - genome-scale measurement of transcriptional regulation
  - genome-scale measurement of protein-protein interaction
  - comparative genomics
- **Most of the downstream analyses require a good understanding of statistics**

Warning: the global trends

- the capability to analyze large amounts of data presents a risk to remain at a superficial level, or to be fooled by forgetting to check the pertinence of the results (with some in-depth examples).
- good news: this does not prevent the authors from publishing in highly quoted journals.

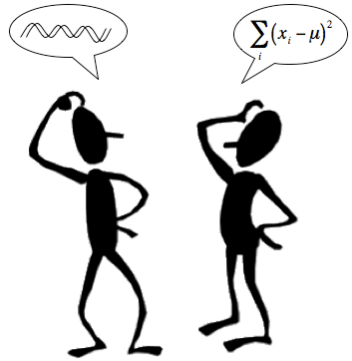
Source: Jacques van Helden

Multidisciplinarity

Bioinformatics is intrinsically a multidisciplinary domain

- **Problem 1: Scientists can not be experts in all of these domains**
  - Biologists (generally) hate statistics and computers
  - Computer scientists (generally) ignore statistics and biology
  - Statisticians and mathematicians (generally)
    - speak a strange language for any other human being
    - spend their time writing formula everywhere

  Solution: multidisciplinary teams and/or multi-lab projects



- **Problem 2: Complexity of the biological domain**
  - Each time you try to formulate a rule, there is a possible counter-example
  - Even the definition of a single word requires a book rather than a sentence (exercise: find a consensual definition of "*gene", of "species"*)

  Solution: find compromise and define standards (ex. Gene Ontology)

Source: Jacques van Helden

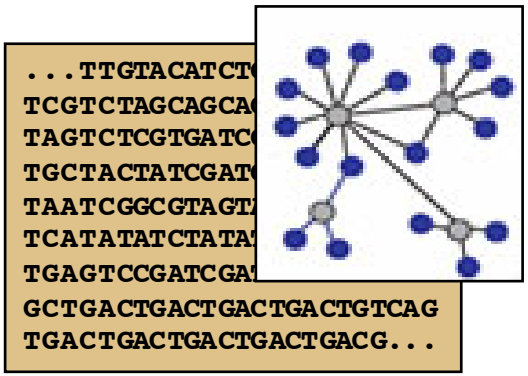
Scope of Bioinformatics

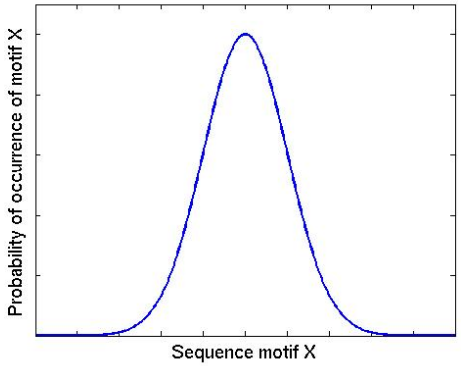
- **Storage and retrieval of biological data**
  databases
- **Sequence analysis**
  sequence alignments, database searches, motif detection
- **Genomics**
  genome annotation, comparative genomics
- **Phylogeny**
  evolution
- **Functional genomics**
  transcriptome, proteome, interactome
- **Analysis of biochemical networks**
  metabolic networks, regulatory networks
- **Systems biology**
  modelling and simulation of dynamical systems
- **Molecular structures**
  visualization and analysis, classification, prediction
- **...**

Source: Jacques van Helden

Scope of Bioinformatics

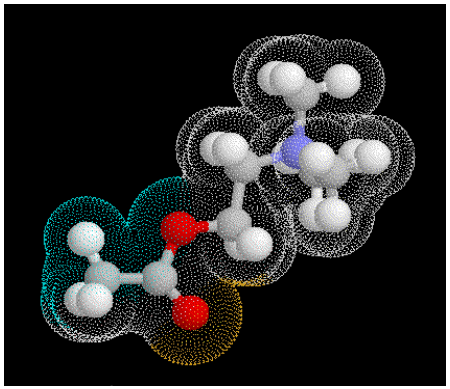
Sequences & networks

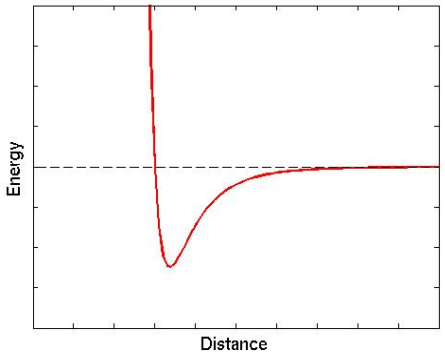




- Where are the genes in my genomic sequence?
- What is the probability to observe a given DNA motif in a DNA sequence?

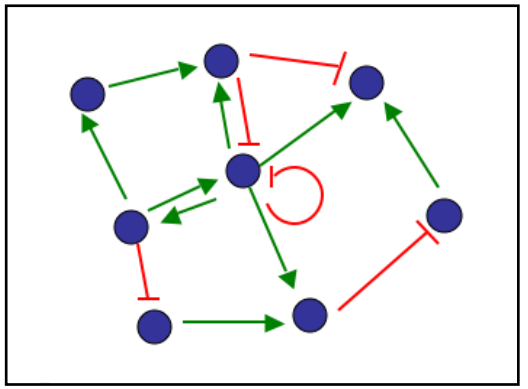
Structural biology

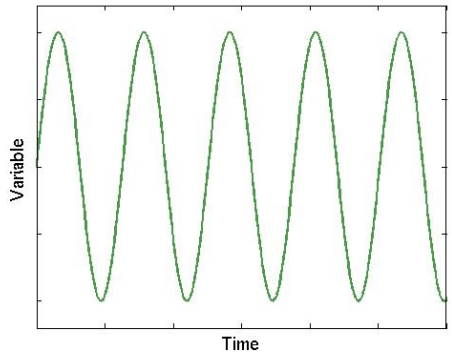




- Will a protein bind a given ligand?
- What will be the optimal protein folding?

Dynamics modeling

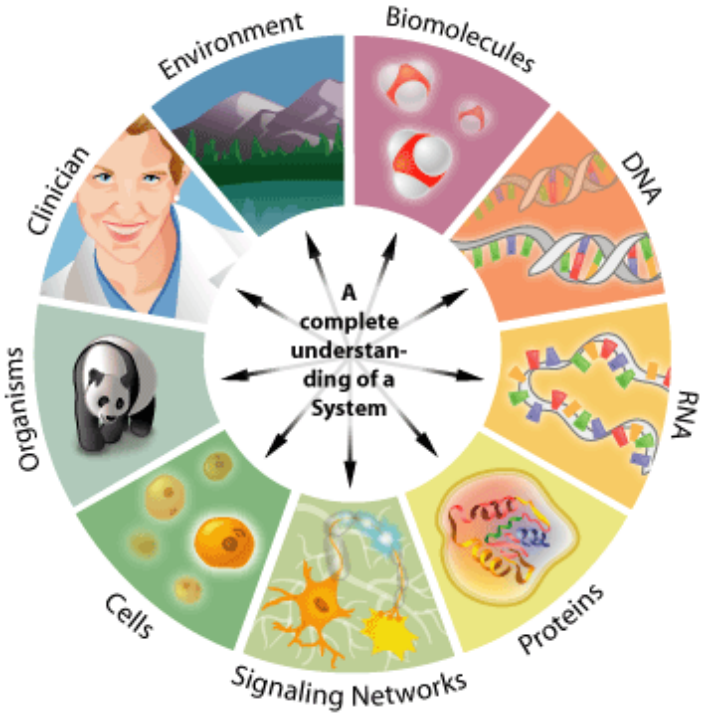




- What kind of dynamical behavior can be expected from a given regulatory network?
- How can I perturb the system in order to have a certain respo

Applications of Bioinformatics

- **Research in biology**
  - Molecular organization of the cell/organism
  - Development
  - Mechanisms of evolution
- **Medicine**
  - Diagnostic of cancers
  - Detecting genes involved in cancer
- **Pharmaceutical research**
  - Mechanisms of drug action
  - Drug target identification
- **Biotechnology**
  - Crop/food production and improvement
  - Drug development and gene therapy
  - Bioengineering
  - Biocleaning, biodegradable plastics, biofuels



Source: Jacques van Helden

Challenges for the bioinformaticians

Progresses and accumulation of data in biology/biophysics stimulate the developement of new methods in bioinformatics:

- **Structure analysis** (since the 50s)
  - Structure prediction
  - Structure comparison (1)
- **Sequencing** (since the 70s)
  - Sequence alignment
  - Sequence search in databases
- **Genomes** (since the 90s)
  - Genome annotation
  - Comparative genomics
  - Functional classifications ("ontologies")
- **Transcriptome** (since 1997)
  - Multivariate analysis (2)
- **Proteome** (~ 2000)
  - Graph analysis

(1) Before, the prediction of **protein structure** relied mainly on energy computation. Now, with the accumulation of structural data, structural alignments can be performed.

(2) The development of **microarray** raised new challenges for the statisticians - they have to deal with data consisting with a large number of genes, a limited number of conditions and no (or very few) replicates.

Adapted from Jacques van Helden

Bioinformatics in practice

- In most cases, "web" biology will be required afterwards to validate the predictions.

- Bioinformatics can
  - reduce the universe of possibilities to a small set of testable predictions.
  - assign a degree of confidence to each prediction.

- The biologist will often have to choose the appropriate degree of confidence, depending on the trade-off between:
  - cost for validating predictions
  - benefit expected from the right predictions

Source: Jacques van Helden

Bioinformatics in practice

Bioinformatics is the field of science in which **biology**, **computer science**, **mathematics** and **information technology** merge into a single discipline.

The ultimate goal of the field is to enable the discovery of new **biological insights** as well as to create a global perspective from which unifying principles in biology can be discerned. There are three important subdisciplines within bioinformatics:

- the development of new **algorithms** and **statistics** with which to assess relationships among members of large data sets
- the **analysis** and **interpretation** of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures
- the **development** and **implementation of tools** that enable efficient access and management of different types of information.

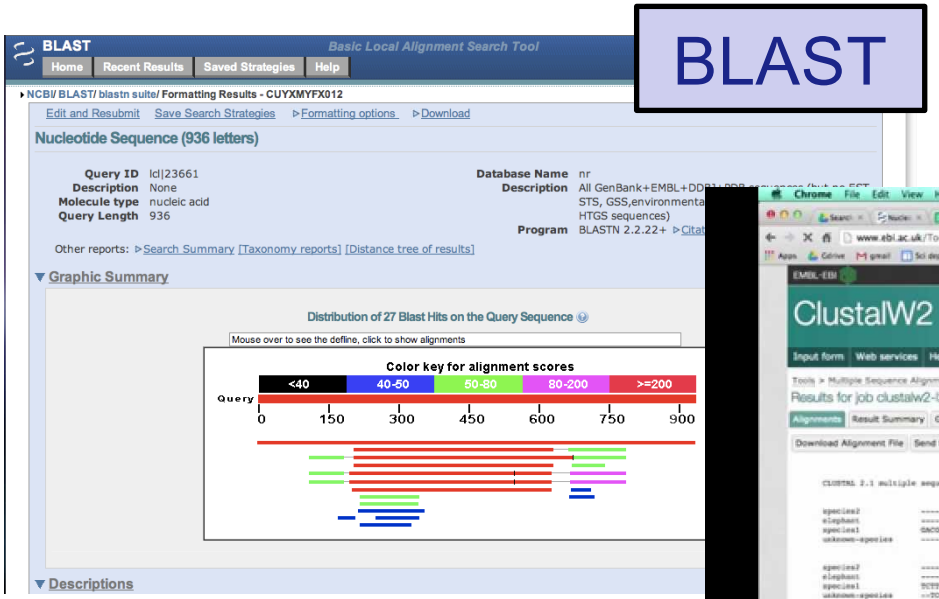
Need to have **biological knowledge** to know what questions to ask.

Source: Jacques van Helden

Why do biologists need bioinformatics?

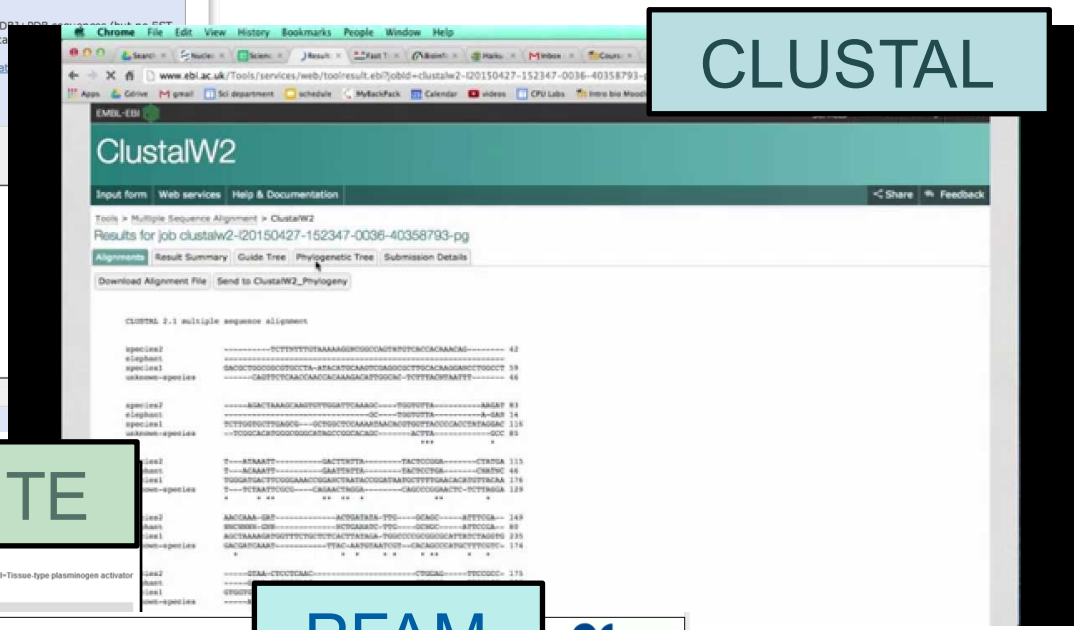
- I have sequenced a gene. What is its function? Has this gene homologs in other organisms? How is this gene regulated?
- I have sequenced a genome. Where are the coding/uncoding sequences? Where are the genes?
- I have sequenced a protein. Can I have an idea of its structure? of its function?
- I am interested in a given organism. Has this organism a given gene? a given protein?
- I have performed a microarray experiment. How can I interpret my results? Which genes are coregulated? In which conditions are my genes up-regulated? down-regulated? How can I classify my genes?
- I have identified a group of genes co-expressed or interacting with each other. Are they co-regulated? Do they participate to the same biological processes/functions?
- I have the sequence of a gene (or a set of genes) of several organisms. Can I infer phylogenetic relationships?

Bioinformatic tools for biologists?

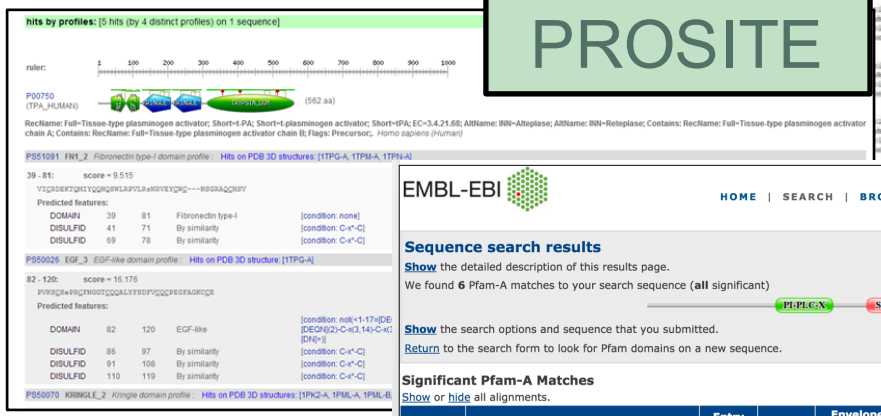


BLAST

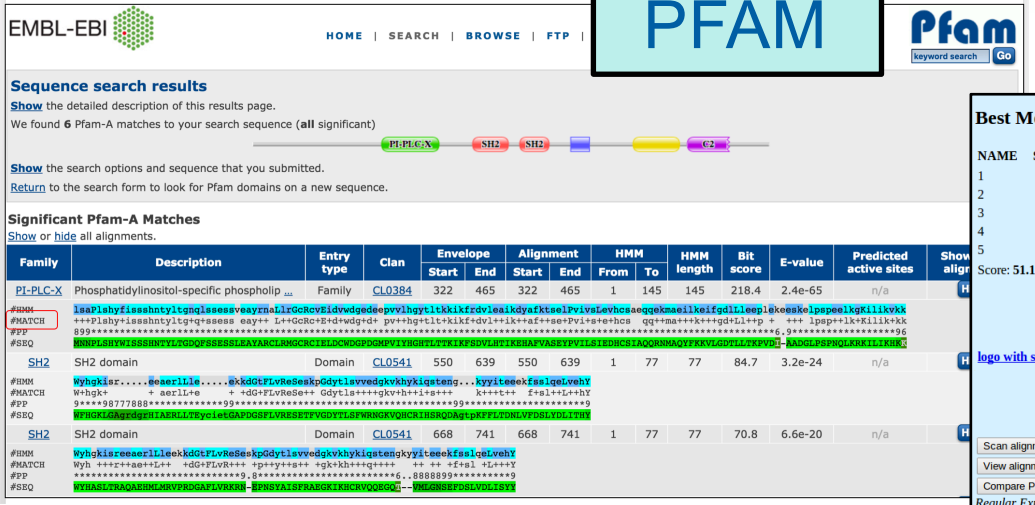
Many softwares, often with web interfaces, will help the biologists to answer these questions...



CLUSTAL

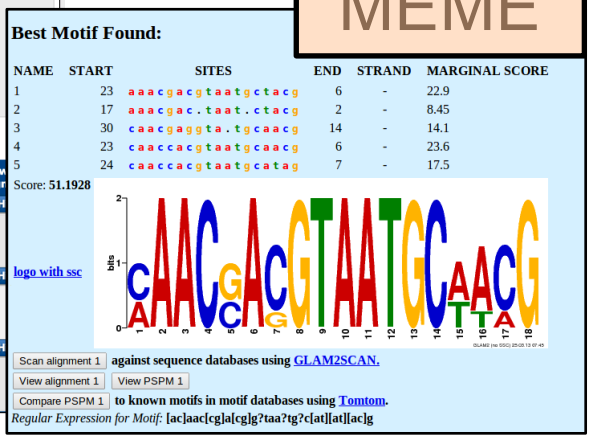


PROSITE

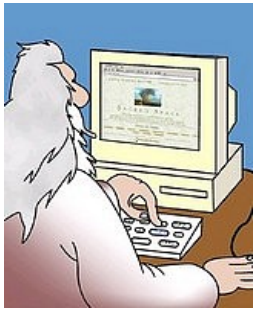


PFAM

MEME

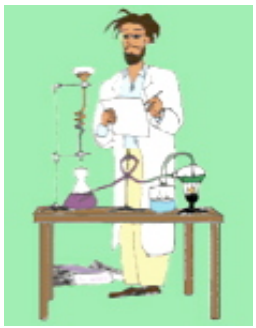


Who has the last word?



While **computer-based analysis** has the benefit of being easily carried out (large memory, fast computation) in an objective way, it cannot guarantee to produce biologically relevant results.

Manual checking (i.e. interpreting the results using the biology knowledge) remains essential!



Ultimately, only **experiments** will **validate** (or not) the bioinformatic predictions. Bioinformatic predictions can be used to reduce the number of possibilities.

Overview of the course

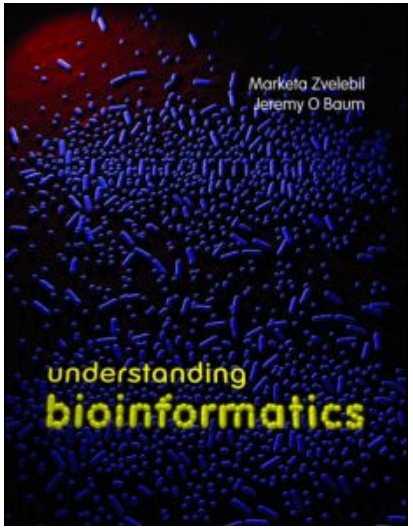
- **Databases**
  - Examples of databases
  - Structure of a database
- **Sequence analysis**
  - Pair-wise sequences alignment
  - Matching sequences against databases
  - Multiple sequences alignment
  - Finding motifs in sequences
  - Gene prediction and genome annotation
- **Large-scale data and network analysis**
  - Gene expression analysis (microarray)
  - Regulatory networks
  - Protein-protein interactions

Didier Gonze
12h theory
+ 12h practicals
+ personal work

Vincent Detours
24h theory &
illustrations/demos

References

- **Zvelebil and Baum** (2007) Understanding Bioinformatics, Garland Science.
- **Samuelsson T** (2012) Genomics and Bioinformatics: An Introduction to Programming Tools for Life Scientists, Cambridge Univ. Press.
- **Pevzner P, Shamir R** (2011) Bioinformatics for Biologists, Cambridge Univ. Press.
- **Mount** (2001) Bioinformatics: *Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, New York.
- **Durbin R, Eddy S, Krogh A, Mitchison G** (1998) *Biological sequence analysis*. Cambridge Univ. Press.
- **Xiong J** (2007) Essential Bioinformatics, Cambridge
- **Westhead, Parish, and Twyman** (2002) *Bioinformatics*. BIOS Scientific Publishers, Oxford.



+ Courses of Jacques van Helden

Further reading

Some review papers

- **Kaminski** (2000) Bioinformatics. A user's perspective, *Am. J. Respir. Cell Mol. Biol.* 23:705-711.
- **Chicurel** (2002) Bioinformatics: bringing it all together, *Nature* 419:751-757.
- **Kanehisa, Bork** (2003) Bioinformatics in the post-sequence era, *Nat. Genet.* 33(Suppl.):305-310.
- **Hanash** (2003) Disease proteomics, *Nature* 422:226-232.
- **Aggarwal, Lee** (2003) Functional genomics and proteomics as a foundation for systems biology. *Briefings in Functional Genomics and Proteomics* 2:175-184.
- **Oltvai, Barabasi** (2002) Systems biology. Life's complexity pyramid, *Science* 298:763-764.
- **Yaspo** (2001) Taking a functional genomics approach in molecular medicine, *Trends Mol. Med.* 7:494-501.
- **Luscombe, Greenbaum, Gerstein** (2001) What is bioinformatics? An introduction and overview, *Yearbook of Medial Informatics*, Intl. Medical Informatics Association , pp 83-100.
- **Ouzounis, Valencia** (2003) Early bioinformatics: the birth of a discipline - a personal perspective. *Bioinformatics* 19, 2176-2190.

Further reading

Some historial papers

- **Zuckerkandl and Pauling** (1965) Molecules as documents of evolutionary history. *J. Theor. Biol.* 8:357–366.
- **Fitch and Margoliash** (1967) Construction of phylogenetic trees. *Science*, 155:279–284.
- **Needleman and Wunsch** (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443–453.
- **Chou and Fasman** (1974) Prediction of protein conformation. *Biochemistry*, 13:222–244.
- **Dayhoff** (1978) Atlas of Protein Sequence and Structure, Vol. 4, Suppl. 3. National Biomedical Research Foundation, Washington, D.C., U.S.A.
- **Doolittle** (1981) Similar amino acid sequences: chance or common ancestry? *Science*, 214: 149–159.
- **Felsenstein** (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.
- **Feng and Doolittle** (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25:351–360
- **Altschul et al.** (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
- **Eisen et al.** (1998) Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc Natl Acad Sci USA* 95:14863-8.

The complete references and additionnal papers are given in the website

Adapted from Ouzounis and Valencia (2003) Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics* 19: 2176-2190.