Bioinformatics Quick Start

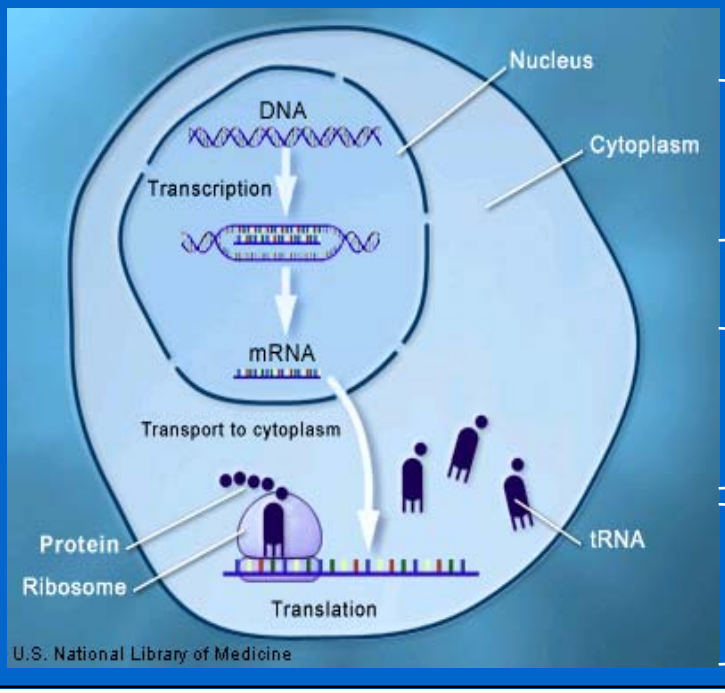
Medha Bhagwat
National Center for Biotechnology Information
National Institutes of Health

NCBI

Outline

1. What is a genome?
2. What is genomics?
3. What is bioinformatics?
4. Applications of genomics/bioinformatics
5. Future implications

NCBI



Information storage

Information carrier

Protein - active machinery

Proteins are Body's Worker Molecules

Hemoglobin carries oxygen to every part of the body.

Ion channel proteins control brain signaling by allowing small molecules into and out of nerve cells.

Enzymes in saliva, the stomach and the small intestine are proteins that help you digest food.

Muscle proteins called actin and myosin enable all muscular movement.

Antibodies are proteins that help defend your body against foreign invaders such as bacteria and viruses.

DNA

Basic Unit (alphabet): Nucleotide (base)
Only 4: A, T, G, and C
Double-stranded

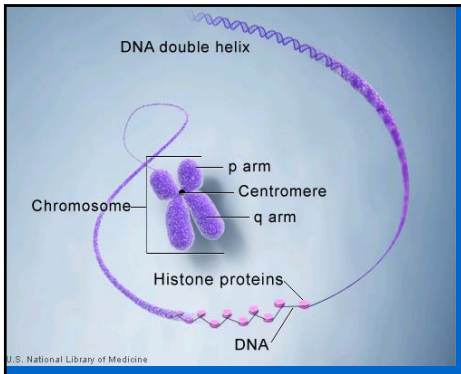
```
..AGCTGCATGCTAGCTGACGTCA....
  | | | | | | | | | | | | | | | | | | | | | |
..TCGACGTACGATCGACTGCAGT....
```

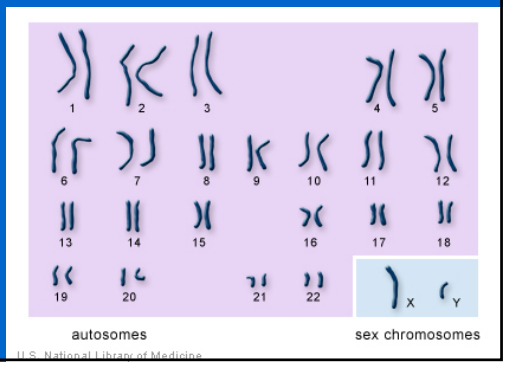
"Words" (genes) to encode proteins, RNA etc.

Double helical

The double-helical structure of DNA

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=stryer.figgrp.146





Protein

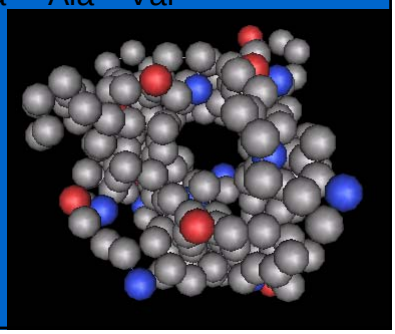
Alphabet: amino acids

There are 20 amino acids
Encoded by codons (triplets of nucleotides)

ATGTGCAGCCTAGCTGCCGTC

Met—Cys—Ser—Leu —Ala — Ala —Val

Water channel protein

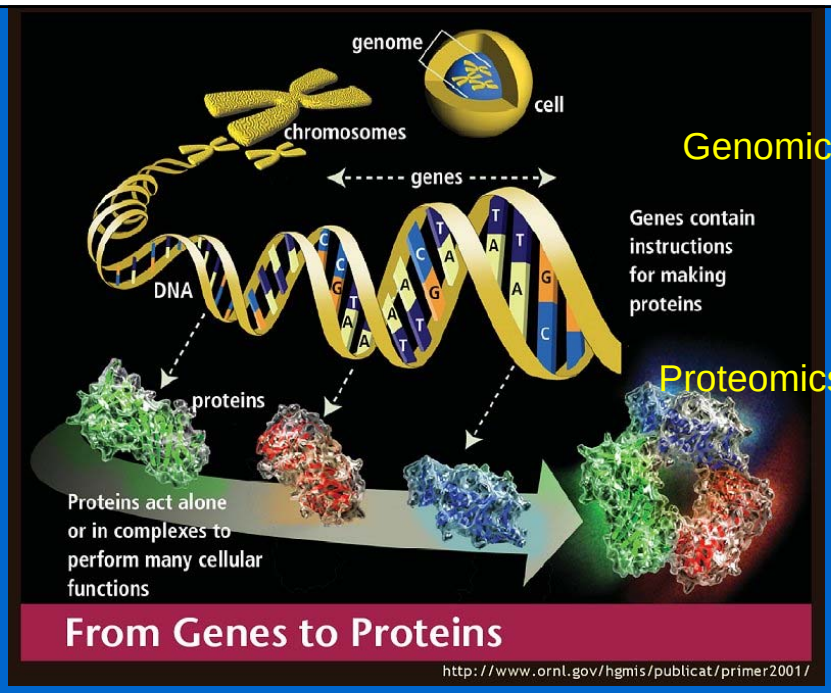


Genome (DNA)

Exact spelling of a word is necessary

CAT	DAT
RAT	GAT
MAT	KAT
FAT	CBT
BAT	CCT
EAT	CAQ
HAT	CAC

Some changes in amino acids lead to diseases and some indicate normal differences among humans.



Genomics

Proteomics

http://genomics.energy.gov/

Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   - How to access the genome data?
   - How to access the analysis tools?
4. Applications of genomics/bioinformatics
   - Analysis of human and other genomes
5. Future implications

NCBI

Additional Information

Genetics Home Reference

http://ghr.nlm.nih.gov/

A Science Primer

http://www.ncbi.nlm.nih.gov/About/primer/

Bookshelf

http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books

Talking Glossary of Genetic Terms

http://www.genome.gov/10002096

Bioinformatics

Variety of definitions

By Luscombe et al Method Inform Med 2001; 40:346-58

Bioinformatics is conceptualizing biology in terms of molecules
(in the sense of Physical chemistry)
and applying "informatics techniques"
(derived from disciplines such as applied
math, computer science and statistics)
to understand and organize the information associated
with these molecules, on a large scale.

Bioinformatics is a management information system for molecular biology and has many practical applications.

NCBI

Bioinformatics

I. Organize data in databases
- researchers can access current data
- submit new data

II. Develop tools and resources to analyze data

III. Interpret data in a biologically useful manner
- global analysis of data to uncover common principles that apply across many systems

NCBI

National Center for Biotechnology Information

http://www.ncbi.nlm.nih.gov

Created as a part of NLM in 1988
- To establish public databases
  - U.S. National DNA Sequence Database
- To perform research in computational biology
- To develop software tools for sequence analysis
- To disseminate biomedical information

NCBI

NCBI Databases

Primary Databases

Genomes (DNA)
↓
mRNA
↓
Protein

Organisms

Expression

Structures

Small compounds

Conserved Domains

Publications
Books

Derived Databases

Gene

Homologene

Unigene

RefSeq

NCBI

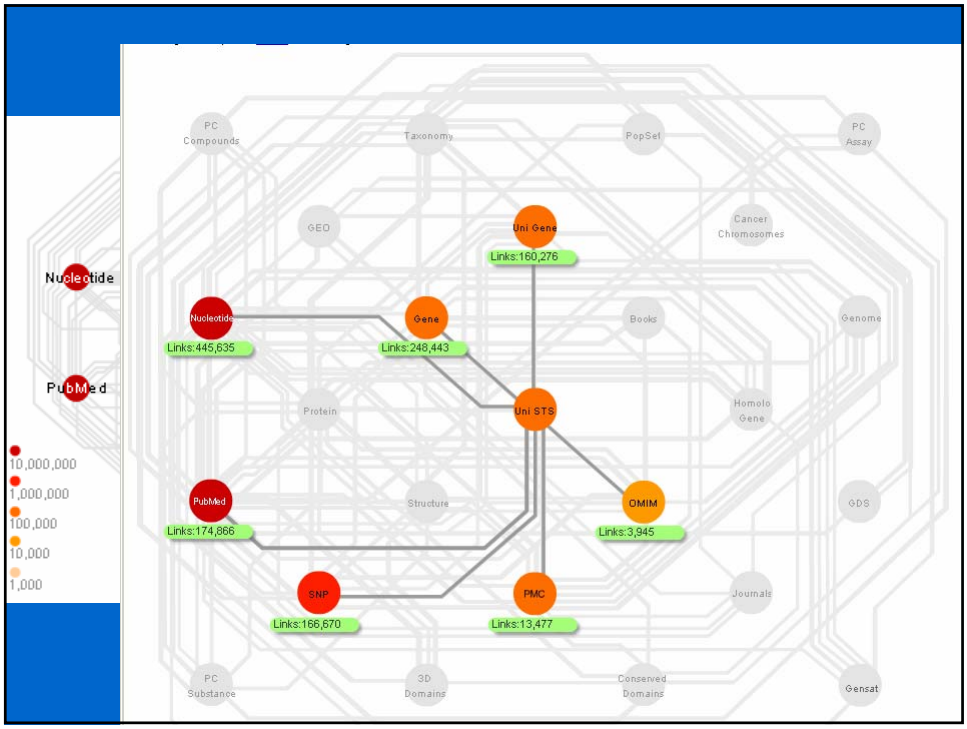
NCBI Databases

Primary	Derived
Archival/repository	Curated
Redundant	Non-redundant
Submitter owner	NCBI owner
Sequenced	Combined/edited
Ex: GenBank	Ex: RefSeq

NCBI

NCBI Databases

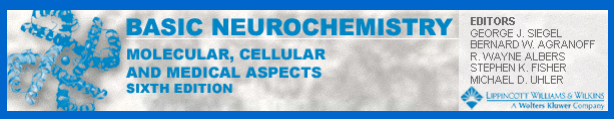
http://www.ncbi.nlm.nih.gov/sites/gquery

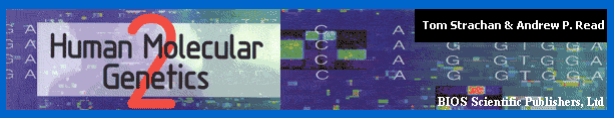


NCBI Databases

Bookshelf

http://www.ncbi.nlm.nih.gov/sites/entrez?db=books

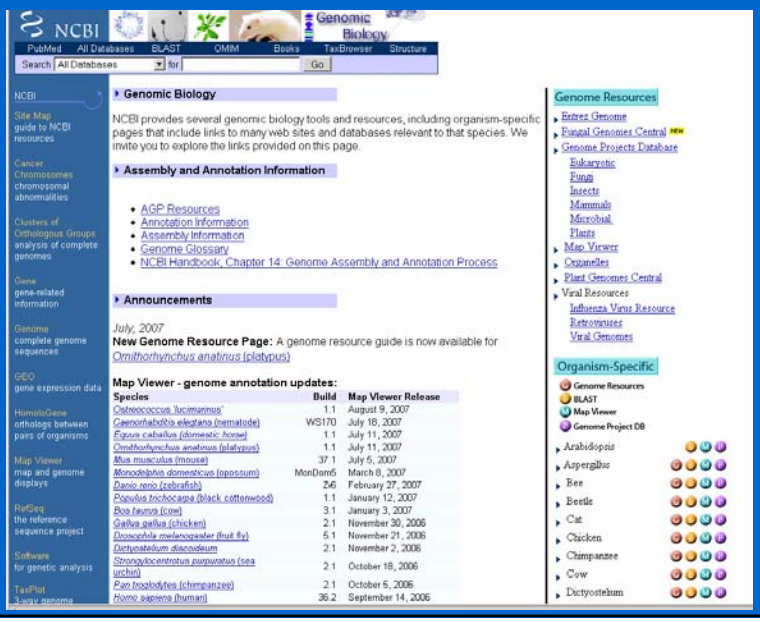




NCBI

Genome Sequence Data and Analysis Tools at NCBI

http://www.ncbi.nlm.nih.gov/Genomes/



Sizes of Different Genomes

Aloe vera	16.0 billion
Rabbit	3.5 billion
Human	3.2 billion
Laboratory mouse	2.6 billion
Fruit fly	137 million
Yeast	12.1 million
Bacterium (*E. coli*)	4.6 million
Human immunodeficiency virus	9700

NCBI

Tools for Data Mining

ez BLAST OMIM Books TaxBrowser Structure

for Go

alysis | Protein Sequence Analysis | Structures | Genome Analysis | Gene Expression

Tools - Nucleotide Sequence Analysis

BLAST 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 00110101AG Electronic PCR - allows you to search your DNA sequence for sequence tagged sites (STSs) that have been used as landmarks in 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 related resources, such as mutation databases. Entrez Gene is a successor to LocusLink.

Model Maker - allows you to view the evidence (mRNAs, ESTs, and gene predictions) that was aligned to assembled genomic sequence to build a 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.

http://www.ncbi.nlm.nih.gov/Tools/

SAGEmap - provides a tool for performing statistical tests designed specifically for differential-type analyses of SAGE (Serial Analysis of Gene Expression) data. The data include SAGE libraries generated by individual labs as well as those generated by the Cancer Genome Anatomy Project (CGAP), which have been submitted to Gene Expression Omnibus (GEO). Gene expression profiles that compare the expression in different SAGE libraries are also available on the Entrez GEO Profiles pages. It is possible to enter a query sequence in the SAGEmap resource to determine what SAGE tags are in the sequence, then map to associated SAGEtag records and view the expression of those tags in different CGAP SAGE libraries.

Spidey - aligns one or more mRNA sequences to a single genomic sequence. Spidey will try to determine the exon/intron structure, returning one or more models of the genomic structure, including the genomic/mRNA alignments for each exon.

Splign - is a utility for computing cDNA-to-Genomic alignments based on a variation of the Needleman-Wunsch algorithm combined with Blast for compartment detection and greater performance.

VecScreen - a tool for identifying segments of a nucleic acid sequence that may be of vector, linker, or adapter origin prior to sequence analysis or submission. VecScreen was developed to combat the problem of vector contamination in public sequence databases.

Viral Genotyping Tool - a web-based program that identifies the genotype (or subtype) of recombinant or non-recombinant viral nucleotide sequences. It works by using BLAST to compare a query sequence to a set of reference sequences for known genotypes. Predefined reference genotypes exist for three major viral pathogens: human immunodeficiency virus 1 (HIV-1), hepatitis C virus (HCV) and hepatitis B virus (HBV), as well as for poliovirus. User-defined reference sequences can be used at the same time. The query sequence is broken into segments for comparison to the reference so that the mosaic organization of recombinant sequences is revealed. The results are displayed graphically using color-coded genotypes. Therefore, the genotype(s) of any portion of the query can quickly be determined.

Genome Sequence Data and Analysis Tools at NCBI

Tools - Protein Sequence Analysis and Proteomics

BLAST 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.

BLink - ("BLAST Link") displays the results of BLAST searches that have been done for every protein sequence in the Entrez Proteins data domain.

CD Search - search the Conserved Domain Database with Reverse Position Specific BLAST.

CDART - when given a protein query sequence, CDART displays the functional domains that make up the protein and lists proteins with similar domain architectures.

Open Mass Spectrometry Search Algorithm (OMSSA) - The OMSSA search service allows proteomics researchers to submit the mass spectra of peptides and proteins for identification. OMSSA then compares these mass spectra to theoretical ions generated from data libraries of known protein sequences and ranks the results using a score derived from classical hypothesis testing.

TaxPlot - a tool for 3-way comparisons of genomes on the basis of the protein sequences they encode. To use TaxPlot, one selects a reference genome to which two other genomes are compared. Pre-computed BLAST results are then used to plot a point for each predicted protein in the reference genome, based on the best alignment with proteins in each of the two genomes being compared.

http://www.ncbi.nlm.nih.gov/Tools/

Tools - Structures

Cn3D - Cn3D is a helper application for your web browser that allows you to view 3-dimensional structures from NCBI's Entrez retrieval service. Cn3D runs on Windows, Macintosh, and Unix.

VAST Search - VAST Search is NCBI's structure-structure similarity search service. It compares 3D coordinates of a newly determined protein structure to those in the MMDB/PDB database.

CD Search - search the Conserved Domain Database with Reverse Position Specific BLAST.

Tools - Genome Analysis

Entrez Genomes - whole genomes of over 1000 organisms. The genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life - bacteria, archaea, and eukaryota - are represented, as well as many viruses, phages, viroids, plasmids, and organelles.. Entrez Genomes provides graphical overviews of complete genomes/chromosomes and the ability to explore regions of interest in progressively greater detail.

COGs - Clusters of Orthologous Groups - a natural system of gene families from complete genomes. Clusters of Orthologous Groups (COGs) were delineated by comparing protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

Map Viewer - shows integrated views of chromosome maps for many organisms, including human and numerous other vertebrates, invertebrates, fungi, protozoa, and plants. Map Viewer is used to view assembled genomes (either draft or complete) and is a valuable tool for the identification and localization of genes and other biological features. Multiple map displays are aligned based on shared marker and gene names when available, and sequence map displays are based on a common sequence coordinate system. Sequence data for chromosome regions of interest can be downloaded, biological annotations can be viewed in graphical format and/or downloaded in tabular format, and gene models can be manipulated in the associated ModelMaker tool.

SKY/M-FISH & CGH Database - The NCI and NCBI SKY/M-FISH and CGH Database is a repository of publicly submitted data from Spectral Karyotyping (SKY), Multiplex Fluorescence In Situ Hybridization (M-FISH), and Comparative Genomic Hybridization (CGH), which are complementary fluorescent molecular cytogenetic techniques. SKY/M-FISH permits the simultaneous visualization of each human or mouse chromosome in a different color, facilitating the identification of chromosomal aberrations; CGH can be used to generate a map of DNA copy number changes in tumor genomes. Collaborative project with the National Cancer Institute. (data submission instructions...)

Genome Sequence Data and Analysis Tools at NCBI

Tools - Gene Expression

GEO Gene Expression Omnibus - The Gene Expression Omnibus (GEO) provides several tools to assist with the visualization and exploration of GEO data. Datasets may be viewed as hierarchical cluster heat maps, providing insight into the relationships between samples and co-regulated genes. Individual gene expression profiles showing significant differences between experimental subsets may be located using average subset rank value comparisons. Related gene expression profiles may be identified on the basis of sequence similarity, profile similarity, or homology. Indicators of dataset normalization quality are provided as distribution graphs, and by flagging outliers. Links to other NCBI sequence, mapping and publication database resources are provided where possible.

SAGEmap - provides a tool for performing statistical tests designed specifically for differential-type analyses of SAGE (Serial Analysis of Gene Expression) data. The data include SAGE libraries generated by individual labs as well as those generated by the Cancer Genome Anatomy Project (CGAP), which have been submitted to Gene Expression Omnibus (GEO). Gene expression profiles that compare the expression in different SAGE libraries are also available on the Entrez GEO Profiles pages. It is possible to enter a query sequence in the SAGEmap resource to determine what SAGE tags are in the sequence, then map to associated SAGEtag records and view the expression of those tags in different CGAP SAGE libraries.

The Cancer Genome Anatomy Project (CGAP) - aims to decipher the molecular anatomy of cancer cells. CGAP develops profiles of cancer cells by comparing gene expression in normal, precancerous, and malignant cells from a wide variety of tissues.

UniGene DDD - Digital Differential Display - an online tool to compare computed gene expression profiles between selected cDNA libraries. Using a statistical test, genes whose expression levels differ significantly from one tissue to the next are identified and shown to the user. Additional information about UniGene is above, including a list of organisms represented.

http://www.ncbi.nlm.nih.gov/Tools/

Tools for Programmers

Entrez Programming Utilities - E-Utilities are a set of programs that provide a stable interface into the Entrez retrieval system. The eUtils use a fixed URL syntax that translates a standard set of input parameters into values necessary for various NCBI software components to search for and retrieve data from 23 Entrez databases.

IEB Information Engineering Branch - IEB is responsible for developing NCBI's resources and databases. Access is provided to documentation, access to NCBI software tools and libraries, and announcements.

Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   - How to access the genome data?
   - How to access the analysis tools?
4. Applications of genomics/bioinformatics
   - Analysis of human and other genomes
5. Future implications

NCBI

Applications of Genomics and Proteomics

1. Understand basic biology
2. Diagnosis and treatment of diseases
3. Rationale for drug design
4. Protect plant life
5. Understand bacterial resistance
6. Solve environmental problems
7. Develop new energy sources
8. Improve industrial processes
9. Study evolutionary changes among organisms

Human Genome Sequenced

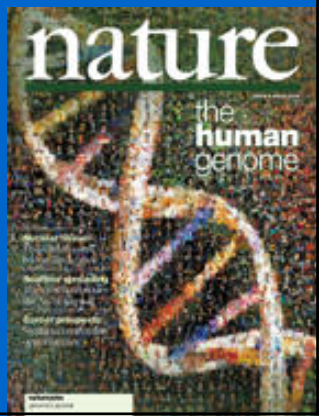
Celera

President Clinton and geneticists J. Craig Venter (left) and Francis Collins (right) celebrate.

The Human Genome Project

www.sciencenews.org

June 23, 2000



The Human Genome

23 pairs of chromosomes
3.2 billion base pairs
Estimated number of genes about 30,000
Only 2% of the human genome “codes”
Average gene size 4000 base pairs
Largest gene dystrophin 2.4 million base pairs
More than 50% in repeat elements or
so called “junk DNA”

NCBI

Analysis of the Human Genome

The DNA sequence of any two people is 99.5-99.9 percent identical.

Sites in the DNA sequence where individuals differ at a single DNA base are called single nucleotide polymorphisms (SNPs).

The SNPs may greatly affect an individual's disease risk.

NCBI

Sickle Cell Anemia

- Sickled red blood cells
- Mutation in the HBB gene that codes for hemoglobin
- one nucleotide change in the 7th codon GAG to GTG
- changing glutamic acid to valine
- interaction between valine and the complementary regions on adjacent molecules results in the formation of polymers that aggregate and distort the shape of the red blood cells

3-D structure of hemoglobin

3-D structure of mutant hemoglobin

Human Genome Sequenced

What are the genetic changes that make us human?

Only 1.2% difference between human and chimpanzee:
- Every 100th nucleotide different
- Could affect thousands of genes
- Many are probably the consequence of 6 million years of genetic drift, with little effect on body or behavior
- Other small changes--perhaps in regulatory, noncoding sequences--may have dramatic consequences

What are the genetic changes that make us human?

Study clinical mutations in the genes that impair key traits and trace the genes' evolution.
For example, FOXP2, the first ever gene associated with human speech.
FOXP2 knock-out mice are squeak less.

A complete understanding of uniquely human traits will, however, include more than DNA such as nurture and nature.

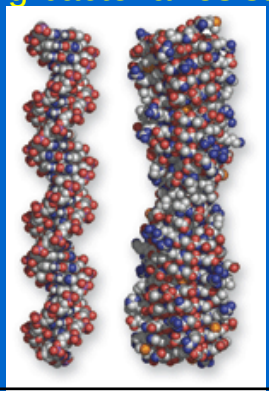
Understand Bacterial Resistance

Fluoroquinolone antibiotics kill Tuberculosis bacteria by binding to DNA-DNA gyrase complex

Tuberculosis bacterium encodes a novel protein mfpA resembling DNA

mfpA competes with DNA for binding to Fluoroquinolone antibiotics thus making bacteria resistant to the antibiotic

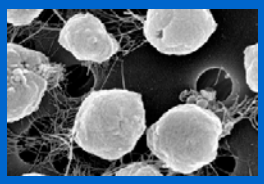
DNA



mfpA protein

Science (2005) 308, 1393

Methanocaldococcus jannaschii



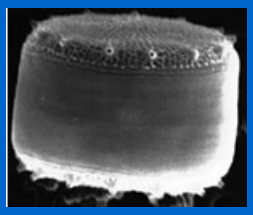
Methane-producing thermophilic archeon

Produces methane, an important energy source

encodes enzymes that withstand high temperatures and pressures possibly useful for industrial processes

Photo: © UC Berkeley Electron Microscope Lab (GNN)

Thalassiosira pseudonana

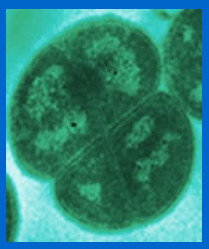


Ocean diatom, a major participant in biological pumping of carbon to ocean depths

has potential for mitigating global climate change

Photo: courtesy of DOE-Genomes to Life

Deinococcus radiodurans



Survives extremely high levels of radiation

has high potential for radioactive waste cleanup

Photo: DOE Joint Genome Institute

What's Next?????

1. HapMap: Genetic variation mapping project
   - Discovery of genes related to diseases
   - Gene Testing
   - Gene Therapy
2. Pharmacogenomics: Pharmacology and genomics
   - Custom effective drugs based on genetic profile
   - Reduce adverse reactions
3. ENCODE: Encyclopedia of functional elements
   - Study expression of genes

NCBI

What's Next?????

Nature, Volume 449 Number 7164
18 October 2007

What's Next?????

Catechol-O-Methyltransfease (COMT) gene variants predict response to buproprion therapy for tobacco dependence.

COMT genotyping could be applied to identify likely responders to burpoprion treatment for smoking cessation.

Biol Psychiatry 2007, 61:111-118

Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   - How to access the genome data?
   - How to access the analysis tools?
4. Applications of genomics/bioinformatics
   - Analysis of human and other genomes
5. Future implications

NCBI

More Information

Volume 291, Issue 5507, Pages 1145-1434

Volume 409, Issue 6822, Pages 745-964

Nature Supplement on "Human Genome"
http://www.nature.com/nature/supplements/collections/humangenome/index.html

Human Genome Project Information
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

NHGRI Fact sheets
http://www.genome.gov/10000202

Questions about NCBI Resources?

E-mail
info@ncbi.nlm.nih.gov

NCBI