What is Bioinformatics?

TWO
Introduction to Bioinformatics

Sami Khuri
Dept of Computer Science
San José State University
June 2016

What is Bioinformatics?

- The Human Genome Project (HGP)
- Mapping
- Model Organisms
- Types of Databases
- Applications of Bioinformatics
- Genome Research

From the Preface

- We believe that to perform a proper analysis it is not sufficient to understand how to use a program and the kind of results (and errors!) it can produce.
- It is of also necessary to have some understanding of the technique used by the program and the science on which it is based.

Preface and Note to the Reader

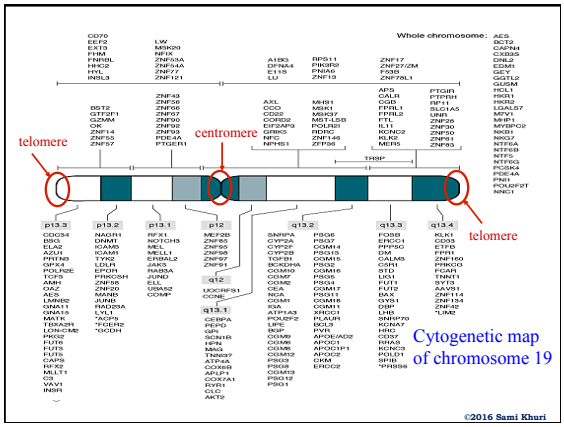
- All research workers in the areas of biomolecular science and biomedicine are now expected to be competent in several areas of sequence analysis and often, additionally, in protein structure analysis and other more advanced bioinformatics techniques.
- The book is designed to be accessible both to students who wish to obtain a working knowledge of the bioinformatics applications, as well as to students who want to know how the applications work and maybe write their own.

The Human Genome Project

- The **HGP** is a multinational effort, begun by the USA in 1988, whose aim is to produce a complete physical map of all human chromosomes, as well as the entire human DNA sequence.
- The ultimate goal of genome research is to find all the **genes** in the **DNA sequence** and to develop tools for using this information in the study of **human biology** and **medicine**.
- The primary goal of the project is to make a series of descriptive diagrams (called **maps**) of each human chromosome at increasingly finer resolutions.

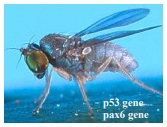
Bioinformatics and the Internet

- The recent enormous increase in biological data has made it necessary to use computer information technology to collect, organize, maintain, access, and analyze the data.
- Computer speed, memory, exchange of information over the Internet has greatly facilitated bioinformatics.
- The bioinformatics tools available over the Internet are accessible, generally well developed, fairly comprehensive, and relatively easy to use.

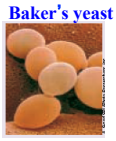


Other Species

As part of the HGP, genomes of other organisms, such as bacteria, yeast, flies and mice are also being studied.



Baker's yeast



DNA repair
Cell division

Chimps are infected with SIV
Very rarely progress to AIDS

Other Sequenced Genomes

Model Organisms

- A **model organism** is an organism that is extensively studied to understand particular biological phenomena.
- **Why have model organisms?** The hope is that discoveries made in model organisms will provide insight into the workings of other organisms.
- **Why is this possible?** This works because evolution reuses fundamental biological principles and conserves metabolic, regulatory, and developmental pathways.

Studying Human Diseases

Organism	Human Diseases
E. coli	DNA repair; colon cancer and other cancers
Yeast	Cell cycle; cancer, Werner syndrome
Drosophila	Cell signaling; cancer
C. elegans	Cell signaling; diabetes
Zebrafish	Developmental pathways; cardiovascular disease
Mouse	Gene expression; Lesch-Nyhan disease, cystic fibrosis, fragile-X syndrome, and many other diseases

Goals of the HGP

- To ***identify*** all the approximately 20,000-25,000 genes in human DNA,
- To ***determine*** the sequences of the 3.2 billion chemical base pairs that make up human DNA,
- To ***store*** this information in databases,
- To ***improve*** tools for data analysis,
- To ***address*** the ethical, legal, and social issues (ELSI) that may arise from the project.

HGP Finished Before Deadline

- In 1991, the USA Congress was told that the HGP could be done by 2005 for $3 billion.
- It ended in 2003 for $2.7 billion, because of efficient computational methods.

What is Bioinformatics? Set of Tools

- The use of computers to collect, analyze, and interpret biological information at the molecular level.
- A set of software tools for molecular sequence analysis

What is Bioinformatics? A Discipline

- The field of science, in which **biology**, **computer science**, and **information technology** merge into a single discipline.

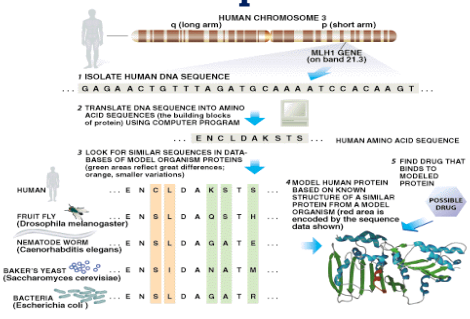
Definition of NCBI (National Center for Biotechnology Information)

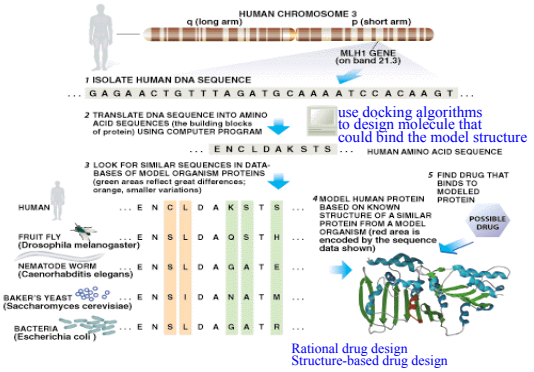
- The ultimate goal of **bioinformatics** is to enable the discovery of new biological insights and to create a global perspective from which unifying principles in biology can be discerned.

Why Study Bioinformatics (I)

- Bioinformatics is intrinsically interesting.
- Bioinformatics offers the prospect of finding better drug targets earlier in the drug development process.
  - By looking for genes in model organisms that are similar to a given human gene, researchers can learn about protein the human gene encodes and search for drugs to block it.

How can Bioinformatics Help?





Scientific American July 2000

Why Study Bioinformatics (II)

- Molecular biology is the new frontier of 21st century science.
  - DNA, RNA, genes, stem cells, etc.. are everywhere in the news.
- Science Magazine celebrated its 125th anniversary by issuing twenty five big questions facing science over the next quarter-century.

www.sciencemag.org/sciext/125th

Science: Top 25 Questions (I)

* What Is the Universe Made Of?
* What is the Biological Basis of Consciousness?
- **What Do Humans Have So Few Genes?**
- **To What Extent Are Genetic Variation and Personal Health Linked?**

* Can the Laws of Physics Be Unified?
* How Much Can Human Life Span Be Extended?
- **What Controls Organ Regeneration?**
- **How Can a Skin Cell Become a Nerve Cell?**
- **How Does a Single Somatic Cell Become a Whole Plant?**

* How Does Earth's Interior Work?
* Are We Alone in the Universe?
* How and Where Did Life on Earth Arise?

Science: Top 25 Questions (II)

- **What Determines Species Diversity?**
- **What Genetic Changes Made Us Uniquely Human?**

* How Are Memories Stored and Retrieved?
- **How Did Cooperative Behavior Evolve?**
- **How Will Big Pictures Emerge from a Sea of Biological Data?**

* How Far Can We Push Chemical Self-Assembly?
* What Are the Limits of Conventional Computing?
- **Can We Selectively Shut Off Immune Responses?**
- Do Deeper Principles Underlie Quantum Uncertainty and Nonlocality?
- **Is an Effective HIV Vaccine Feasible?**

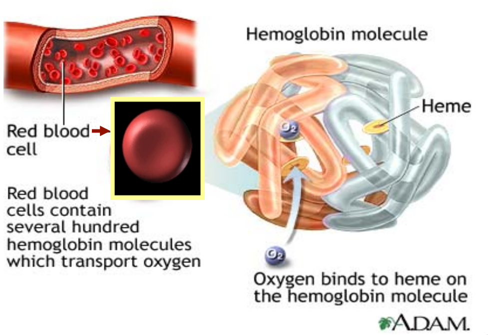
* How Hot Will the Greenhouse World Be?
* What Can Replace Cheap Oil -- and When?

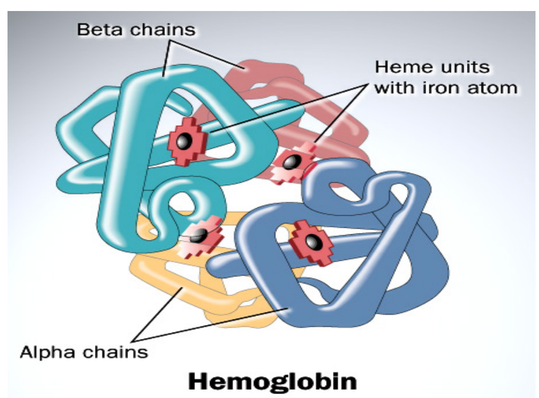

```
HBA_HUMAN     ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS- 49
HBA_HORSE     ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS- 49
HBA_CHICK     --------MVLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYFPHF-DLS- 50
HBB_HUMAN     --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST 50
HBB_BOSMU     ---------MLTAEEKAAVTAFWGKVK--VDEVGGEALGRLLVVYPWTQRFFESFGDLSS 49
HBB_HORSE     --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN 50
HBB_MACGI     --------VHLTAEEKNAITSLWGKVA--IEQTGGEALGRLLIVYPWTSRFFDHFGDLSN 50
MYG_PHYCA     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT 51
GLB5_PETMA    PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT 60
LGB2_LUPLU    --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE 52
                       *:  :   :   :  .              :  .:   * :   *  :   .

HBA_HUMAN     ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL 100
HBA_HORSE     ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL 100
HBA_CHICK     ----HGSAQIKGHGKKVVAALIEAANHIDD-----IAGTLSKLSDLHAHKLRVDPVNFKL 101
HBB_HUMAN     PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL 105
HBB_BOSMU     ADAVMNNPKVKAHGKKVLDSFSNGMKHLDD-----LKGTFAALSELHCDKLHVDPENFKL 104
HBB_HORSE     PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL 105
HBB_MACGI     AKAVMANPKVLAHGAKVLVAFGDAIKNLDN-----LKGTFAKLSELHCDKLHVDPENFKL 105
MYG_PHYCA     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF 106
GLB5_PETMA    ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV 118
LGB2_LUPLU    VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV 109
                    . .:  *.  :   .                    :  *.  *  .  :   : .

HBA_HUMAN     LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------ 141
HBA_HORSE     LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------ 141
HBA_CHICK     LGQCFLVVVAIHHPAALTPEVHASLDKFLCAVGTVLTAKYR------ 142
HBB_HUMAN     LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------ 146
HBB_BOSMU     LGNVLVVVLARHFGKEFTPVLQADFQKVVVGVANALAHRYH------ 145
HBB_HORSE     LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------ 146
HBB_MACGI     LGNIIVICLAEHFGKEFTIDTQVAWQKLVAGVANALAHKYH------ 146
MYG_PHYCA     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 153
GLB5_PETMA    LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY------- 149
LGB2_LUPLU    VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA--- 153
              :   :   :               .         .   :
```


Red Blood Cells





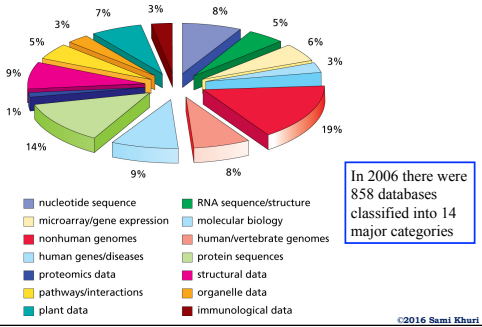
What do Bioinformaticians do?

- They analyze and interpret data
- Develop and implement algorithms
- Design user interface
- Design database
- Automate genome analysis
- They assist molecular biologists in data analysis and experimental design.

Databases for Storage and Analysis

- Databases store data that need to be analyzed
- By comparing sequences, we discover:
  - How organisms are related to one another
  - How proteins function
  - How populations vary
  - How diseases occur
- The improvement of sequencing methods generated a lot of data that need to be:
  - stored
  - organized
  - curated
  - annotated
  - managed
  - networked
  - accessed
  - assessed

Types of Databases



In 2006 there were 858 databases classified into 14 major categories

Three Major Databases

- GenBank from the NCBI (National Center of Biotechnology Information), National Library of Medicine http://www.ncbi.nlm.nih.gov
- EBI (European Bioinformatics Institute) from the European Molecular Biology Library http://www.ebi.ac.uk
- DDBJ (DNA DataBank of Japan) http://www.ddbj.nig.ac.jp

GenBank Taxonomic Sampling

Homo sapiens	62.1%
Mus musculus	7.7%
Drosophila melanogaster	6.1%
Caenorhabditis elegans	3.3%
Arabidopsis thaliana	2.9%
Oryza sativa	1.3%
Rattus norvegicus	0.8%
Danio rerio	0.6%
Saccharomyces cerevisiae	0.6%

GenBank

GenBank is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information these records contain.

What does NCBI do?

NCBI: established in 1988 as a national resource for molecular biology information.

- it creates public databases,
- it conducts research in computational biology,
- it develops software tools for analyzing genome data, and
- it disseminates biomedical information,

all for the better understanding of molecular processes affecting human health and disease.

Applications of Genome Research

Current and potential applications of Genome Research include:

- Molecular Medicine
- Microbial Genomics
- Risk Assessment
- Bioarcheology, Anthropology, Evolution and Human Migration
- DNA Identification
- Agriculture, Livestock Breeding and Bioprocessing

Molecular Medicine

- Improve the **diagnosis** of disease
- Detect genetic **predispositions** to disease
- Create drugs **based on molecular information**
- Use **gene therapy** and control systems as drugs
- Design **custom drugs** on individual genetic profiles.

Microbal Genomics

- Swift detection and treatment in clinics of disease-causing microbes: pathogens
- Development of new energy sources: biofuels
- Monitoring of the environment to detect chemical warfare
- Protection of citizens from biological and chemical warfare
- Efficient and safe clean up of toxic waste.

DNA Identification I

- Identify potential suspects whose DNA may match evidence left at crime scenes
- Exonerate persons wrongly accused of crimes
- Establish paternity and other family relationships
- Match organ donors with recipients in transplant programs

Louis XVII

Louis XVII: son of Louis XV1 and Marie-Antoinette who died from tuberculosis in 1795 at the age of 12

DNA and Human Trafficking

13 Haitian Children Returned To Their Families Thanks To DNA Analyses: DNA-Prokids Bolivia

Natural disasters frequently turn into human tragedies, such as family separations. The Haiti earthquake of January 12, was followed by emotive worldwide solidarity actions. But this can not outshine extremely serious incidents, like the fact that the human trafficking mafias could take advantage of the catastrophe to get children off the island.

Last January, more than seventy people from Haiti arrived at Santa Cruz de la Sierra (Bolivia), via Lima. Visa problems stopped them on their way to Brazil or Argentina. Bolivian Police suspicions opened a deep investigation and proved that the 25 Haitian children in the group were not accompanied by their relatives. In February, their families in Haiti started to look for them.

The Bolivian Attorney General's Office requested the collaboration of the Laboratory of Forensic Genetics of the Bolivia Forensic Research Institute, which applied the DNA-Prokids action protocol. The genetic research results were unquestionable: eight parents (seven mothers and a father) looking for their 13 children have recovered them, thanks to the DNA identification (two mothers looked for two children each, a mother looked for three children, four mothers looked for a child each, a father looked for two children).

From Haiti to Bolivia

DNA in Murder, Suicide Cases and History

- What do these people have in common?
  - Tycho Brahe
  - Salvador Allende
  - Albert DeSalvo
  - Maria Ridulf
  - Luigi Tenco
- They all had their bodies exhumed for DNA testing.

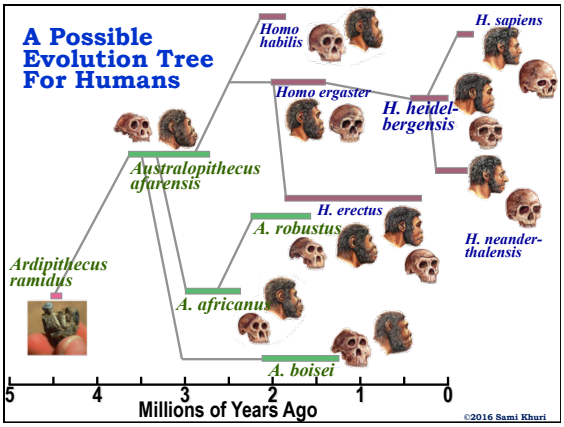
Danish Astronomer: Tycho Brahe (1546 – 1601)

He catalogued more than 1,000 new stars and his stellar and planetary observations helped lay the foundations of early modern astronomy. He was long thought to have died of a bladder infection, which legend suggests was contracted 11 days previously - when he had been too polite to leave the royal banquet table to go to the toilet. Others have suggested he was poisoned. The finger of suspicion had fallen on his assistant, Johannes Kepler, who later became a renowned astronomer himself. In November 2012, Brahe's body was exhumed and scientists concluded that he was probably not poisoned.

Quagga: Zebra or Horse?

Died in Amsterdam zoo in 1883.

A Possible Evolution Tree For Humans



DNA Identification II

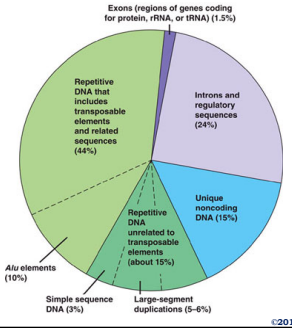
- Identify endangered and protected species as an aid to wildlife officials and also to prosecute poachers
- Detect bacteria and other organisms that may pollute air, water, soil, and food
- Determine pedigree for seed or livestock breeds
- Authenticate consumables such as wine and caviar

Agriculture, Livestock Breeding and Bioprocessing

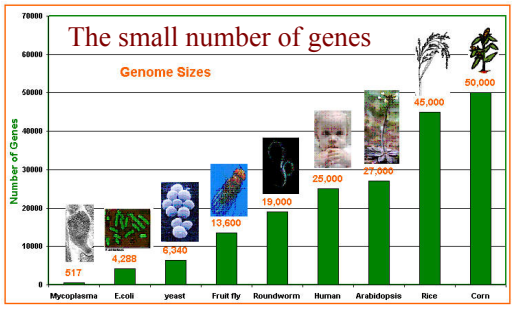
- Grow disease-resistant, insect-resistant, and drought-resistant crops
- Breed healthier, more productive, disease-resistant farm animals
- Grow more nutritious produce
- Develop biopesticides
- Incorporate edible vaccines into food products

What have we learned from the HGP?

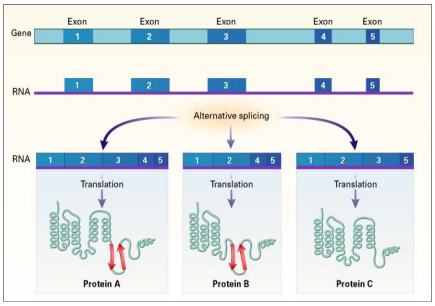
A small portion of the genome codes for proteins, tRNAs and rRNAs



What have we learned from the HGP?

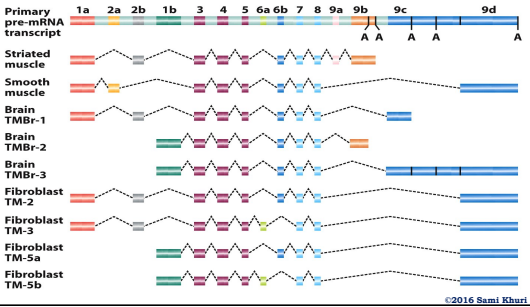


Alternative Splicing

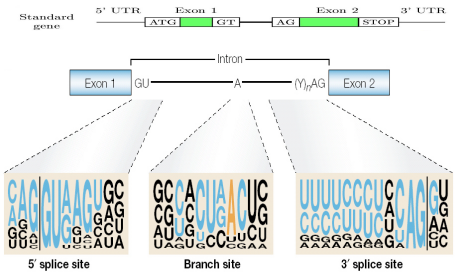


Genomic Medicine by Guttmacher et al., NEJM, 2002

The Alpha-Tropomyosin Gene



Anatomy of an Intron



Building upon the Foundations of HGP

- As we build upon the foundation laid by the **Human Genome Project**, our ability to explore uncharted frontiers will hinge upon melding biological know-how with expertise in computer science, physics, math, clinical research, bioethics, and many other disciplines.
- A firm understanding of the powerful potential of **genomics**, **proteomics**, and **bioinformatics** will be essential to success in this amazing new world.

Discovering Genomics, Campbell, 2007 – Preface by Francis Collins

Genomics is a Way of Seeing Life

- **Genome**: the complete (haploid) DNA content of an organism.
- **Genomics**: the field of genome studies.
- **Genomics**
  - is not just a collection of methods
  - has become an enhanced way of seeing life.
- **Genomics** includes the study of interaction of molecules inside the cell:
  DNA Protein Lipids Carbohydrates
- **Genomics** requires us to analyze, hypothesize, think, and formulate models.

Pathway to Genomic Medicine

Human Genome Project	ENCODE Project	HapMap Project	Genomic Medicine
Sequencing of the human DNA	Interpreting the human genome sequence	Implicating genetic variants with human disease	Personalized medicine Cure for diseases

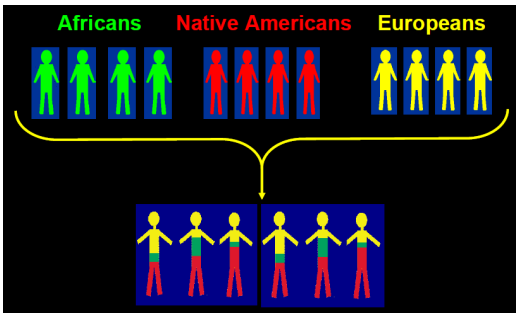
Personalized Medicine

Personalized medicine is the use of diagnostic and screening methods to better manage the individual patient's disease or predisposition toward a disease.

Personalized medicine will enable risk assessment, diagnosis, prevention, and therapy specifically tailored to the unique characteristics of the individual, thus enhancing the quality of life and public health.

Personalized Medicine is Genotype-Specific Treatment.

Origins of African Americans



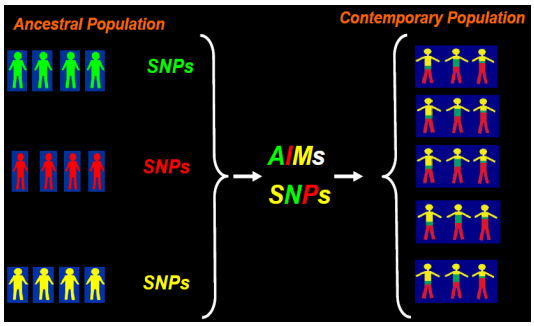
Source: Esteban González Burchard

Ancestry Informative Marker

- An **Ancestry-Informative Marker** (AIM) is a set of polymorphisms for a locus which exhibits substantially different frequencies between populations from different geographical regions.
- By using a number of **AIM**s one can estimate the geographical origins of the ancestors of an individual and ascertain what proportion of ancestry is derived from each geographical region.

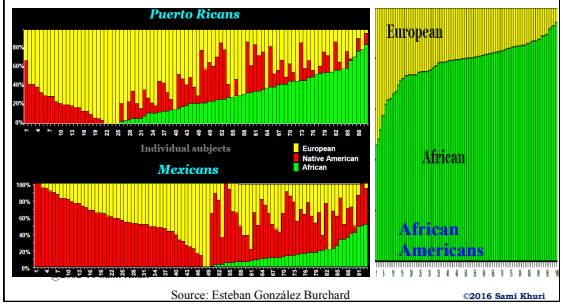
en.wikipedia.org/wiki/

SNPs and AIMs



Source: Esteban González Burchard

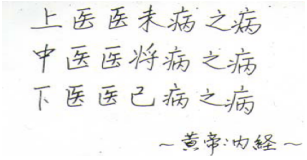
Origins of Latinos and African Americans



Source: Esteban González Burchard

Self-Identified Race: Genetic Ancestry

The Superior Doctor



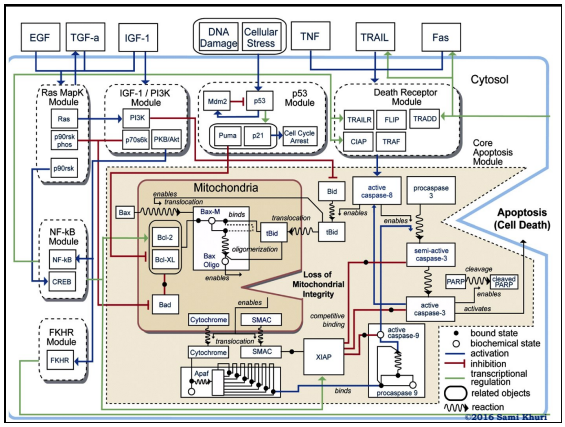
Superior doctors prevent the disease
Mediocre doctors treat the disease before evident
Inferior doctors treat the full blown disease

-Huang Dee: Nai - Ching
(2600 B.C. 1st Chinese Medical Text)

Preventive Medicine

- Prevent disease from occurring
- Identify the cause of the disease
- Treat the cause of the disease rather than the symptoms
- Genomics identifies the cause of disease
- "All medicine may become pediatrics" Paul Wise
- Effects of environment, accidents, aging, penetrance …
- Health care costs can be greatly reduced if
  - invests in preventive medicine
  - one targets the cause of disease rather than symptoms

Wellderly: Healthy Aging

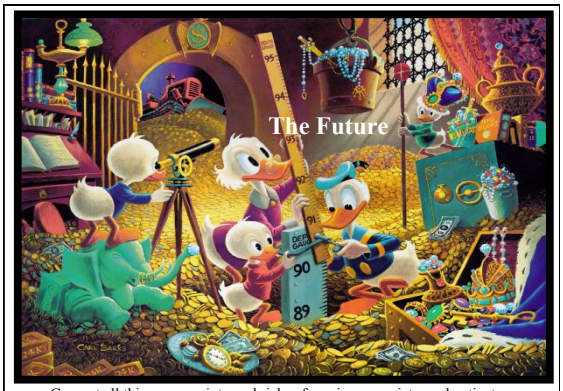


Anatomy Lesson of Dr. Nicolaes Tulp

1632 oil painting by Rembrandt Harmenszoon van Rijn

If Rembrandt was Around Today

Source: Carlos Cordon-Cardo, Columbia University



Convert all this progress into real riches for science, society, and patients

Concluding Remarks

- Biology is becoming an information science
- Progression: **in vivo** to **in vitro** to **in silico**
- Are natural languages adequate in predicting quantitative behavior of biological systems?
  - Need to produce biological knowledge and operations in ways that natural languages do not allow
- "Biology easily has 500 years of exciting problems to work on". Donald Knuth
- Today's biologists need to think quantitatively and from a multidisciplinary perspective.