High-throughput sequencing and big data: implications for personalized medicine?

Dominick J. Lemas, PhD

Postdoctoral Fellow in Pediatrics - Neonatology

Mentor: Jacob E. (Jed) Friedman, PhD

University of Colorado
Anschutz Medical Campus

What is Big Data?

1) Top Tech phrase of 2013[1]

2) Messy, Noisy, Imprecise

3) Datafication

4) Repurposed

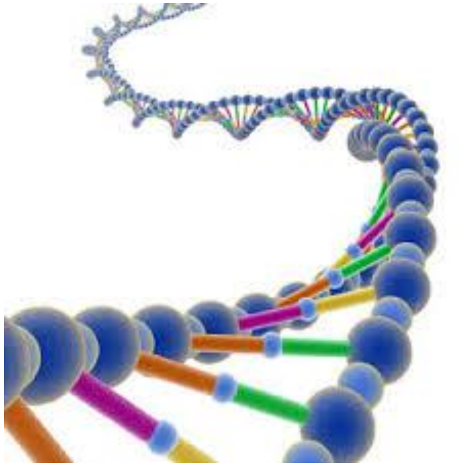
5) N= All

6) Privacy?

BIG DATA
A Revolution That Will Transform How We Live, Work and Think
Viktor Mayer-Schönberger and Kenneth Neil Cukier

[1] http://www.languagemonitor.com/

What is the human genome?



1. Human body contains trillions of cells.
2. Each cell contains a nucleus.
3. Each nucleus contains 2 complete genomes.
4. . . . .

OR if the genome was a book!

- **There are 23 chapters → chromosomes**
- **Each chapter contains several hundred stories→ genes**
- **Each story is composed of paragraphs → exons**
- **Interrupted by advertisements → introns**
- **Each paragraph is made up of words → codons**
- **Each word is written in letters → bases . . . A,C,T,G**

Genome. Matt Ridley. 1999

Goals of Human Genome Project

1) Generate working draft of 90% of the human genome (2001).

2) Obtain complete, high-quality genomic sequence (2003).

3) Make all data publically available.

4) Develop novel sequencing technologies.

5) Map Sequence Variation.

6) Interpret functions of genome.

7) Develop comparative genomic strategies.

8) Ethical, legal and social implications (ELSI).

9) Bioinformatics and Computational Biology

10) Training

Benefits of Sequencing Human Genome

1) Molecular Medicine

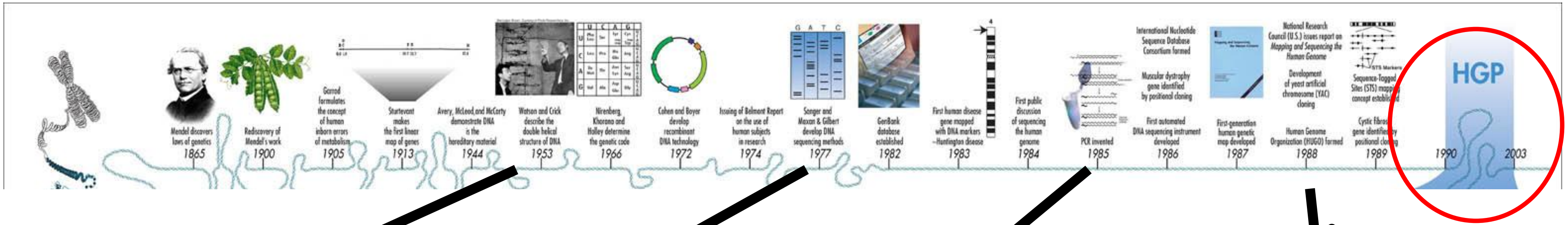
2) Energy and Environmental Applications

3) Bioarcheology, anthropology, evolution, human migration

4) DNA forensics

5) Agriculture, livestock breeding, and bioprocessing

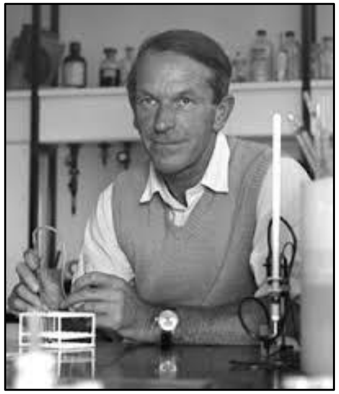
Sequencing Milestones: the early days


Mendel discovers laws of genetics
1865
Rediscovery of Mendel's work
1900
Garrod formulates the concept of human inborn errors of metabolism
1905
Sturtevant makes the first linear map of genes
1913
Avery, McLeod, and McCarty demonstrate DNA is the hereditary material
1944
Watson and Crick describe the double helical structure of DNA
1953
U C A G
Nirenberg, Khorana and Holley determine the genetic code
1966
Cohen and Boyer develop recombinant DNA technology
1972
Issuing of Belmont Report on the use of human subjects in research
1974
G A T C
Sanger and Maxam & Gilbert develop DNA sequencing methods
1977
GenBank database established
1982
4
First human disease gene mapped with DNA markers —Huntington disease
1983
First public discussion of sequencing the human genome
1984
PCR invented
1985
International Nucleotide Sequence Database Consortium formed
Muscular dystrophy gene identified by positional cloning
First automated DNA sequencing instrument developed
1986
First-generation human genetic map developed
1987
National Research Council (U.S.) issues report on Mapping and Sequencing the Human Genome
Development of yeast artificial chromosome (YAC) cloning
Human Genome Organization (HUGO) formed
1988
STS Markers
Sequence-Tagged Sites (STS) mapping concept established
Cystic Fibrosis gene identified by positional cloning
1989
HGP
1990
2003

1953

Watson & Crick

1977

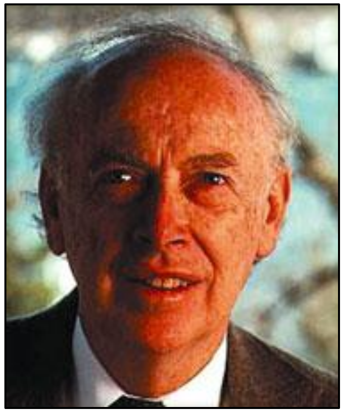


Fredrick Sanger

1985

Charles DeLisi

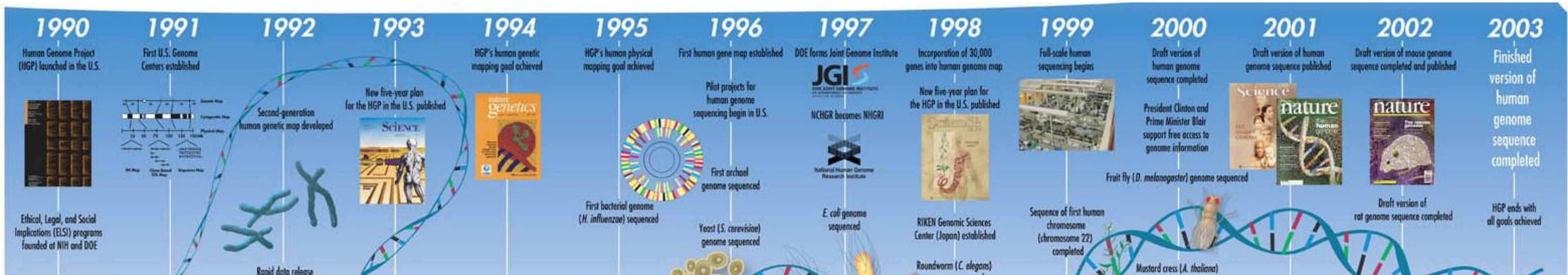
1988

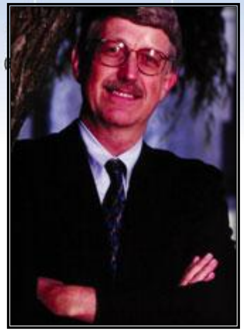


James D. Watson

Collins. 2001.Nature **422**, 835-847

Sequencing Milestones: HGP


1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
1991
First U.S. Genome Centers established
1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the HGP in the U.S. published
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
1995
HGP's human physical mapping goal achieved
First bacterial genome (H. influenzae) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archael genome sequenced
Yeast (S. cerevisiae) genome sequenced
HGP's mouse genetic mapping goal achieved
1997
DOE forms Joint Genome Institute
JGI
NCHGR becomes NHGRI
National Human Genome Research Institute
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
RIKEN Genomic Sciences Center (Japan) established
Roundworm (C. elegans) genome sequenced
SNP initiative begins
GTGCT
n Genome Centers
hai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (D. melanogaster) genome sequenced
Mustard cress (A. thaliana)
2001
Draft version of human genome sequence published
2002
Draft version of mouse genome sequence completed and published
Draft version of rat genome sequence completed
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued...

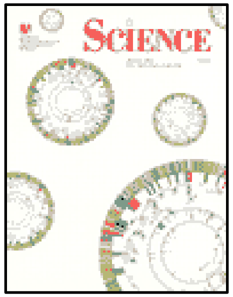


Francis Collins

Bermuda Principle

CELERA

Craig Venter



H. influenzae **Sequenced**

C. elegans **Sequenced**

D. Melanogaster **Sequenced**

Collins. 2001.Nature **422**, 835-847

University of Colorado
Anschutz Medical Campus

The International Human Genome Sequencing Consortium

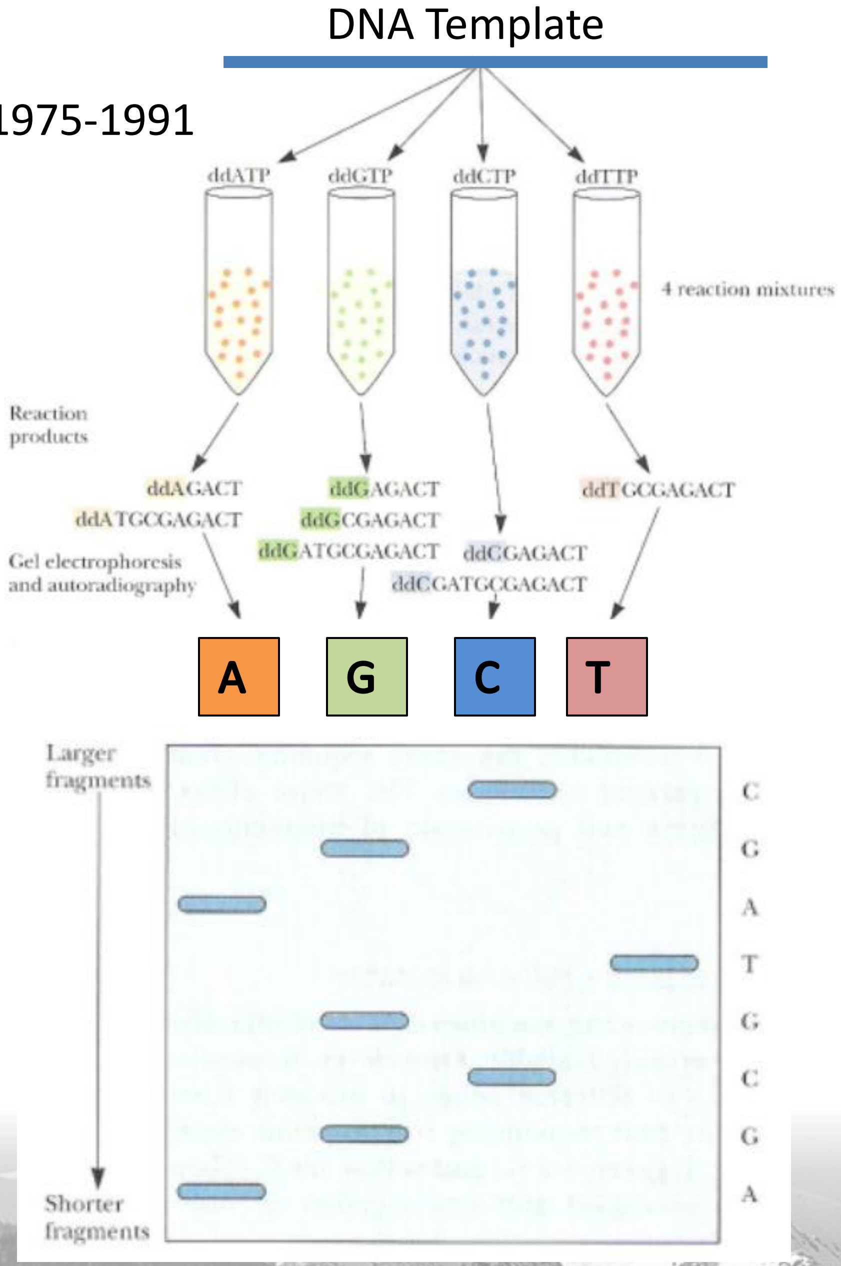
G5- Completed Bulk of Sequencing

- **Whitehead Institute/MIT Center for Genome Research, Cambridge, Mass., U.S.**
- **The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, U.K.**
- **Washington University School of Medicine Genome Sequencing Center, St. Louis, Mo., U.S.**
- **U. S. Department of Energy Joint Genome Institute, Walnut Creek, Calif., U.S.**
- **Baylor College of Medicine Human Genome Sequencing Center, Houston, Tex., U.S.**

- RIKEN Genomic Sciences Center, Yokohama, Japan
- Genoscope and CNRS UMR-8030, Evry, France
- GTC Sequencing Center, Waltham, Mass., U.S.
- Department of Genome Analysis, Jena, Germany
- Beijing Genomics Institute/Human Genome Center, Beijing, China
- Multimegabase Sequencing Center, Seattle, Wash., U.S.
- Stanford Genome Technology Center, Stanford, Calif., U.S.
- Stanford Human Genome Center, Stanford, Calif., U.S.
- University of Washington Genome Center, Seattle, Wash., U.S.
- Department of Molecular Biology, Tokyo, Japan
- University of Texas Southwestern Medical Center at Dallas, Dallas, Texas, U.S.
- University of Oklahoma's Advanced Center for Genome Technology, Norman, Okla., U.S.
- Max Planck Institute for Molecular Genetics, Berlin, Germany
- Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center, Cold Spring Harbor, N.Y., U.S.
- GBF - German Research Centre for Biotechnology, Braunschweig, Germany

Sanger Sequencing

1975-1991



DNA Template

ddATP
ddGTP
ddCTP
ddTTP

4 reaction mixtures

Reaction products

ddAGACT
ddATGCGAGACT
ddGAGACT
ddGCGAGACT
ddGATGCGAGACT
ddCGAGACT
ddCGATGCGAGACT
ddTGCGAGACT

Gel electrophoresis and autoradiography

A
G
C
T

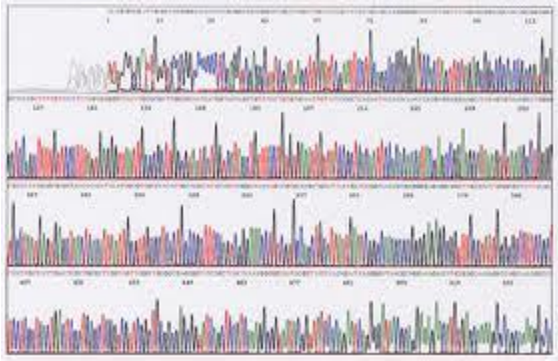
Larger fragments

Shorter fragments

C
G
A
T
G
C
G
A

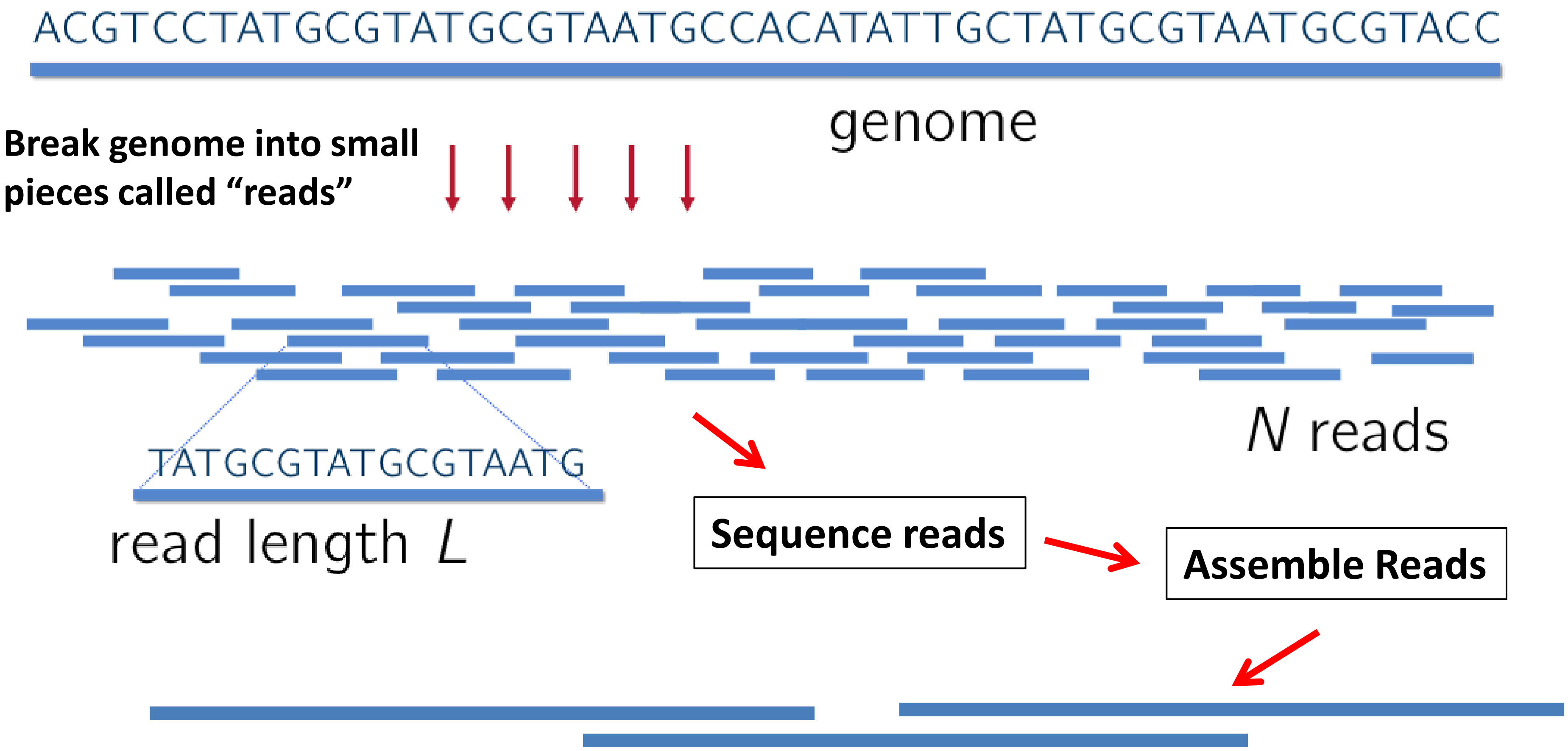
The ABI Prism 3700 /3730

1992-2000's



- $300,000/machine
- Sequence 50-100K bp/hr

Whole-Genome Sequencing



ACGTCCTATGCGTATGCGTAATGCCACATATTGCTATGCGTAATGCGTACC

genome

Break genome into small pieces called "reads"

TATGCGTATGCGTAATG

read length L

N reads

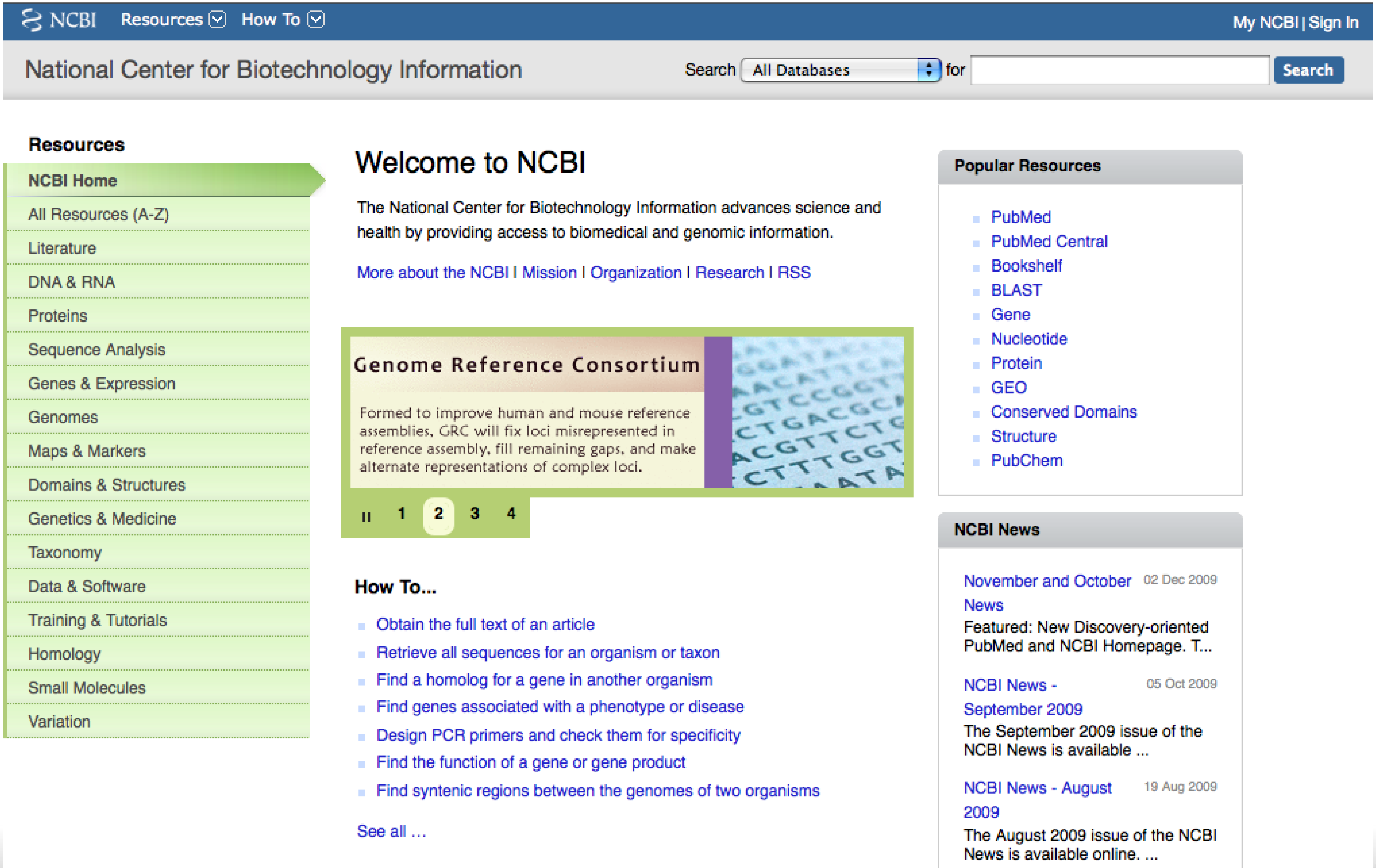
Sequence reads

Assemble Reads

Computational Challenges

- Coverage?
- Imputation?
- Alignment?
- Formatting?
- Analysis?

Where does the data live?



NCBI Resources How To

My NCBI | Sign In

National Center for Biotechnology Information

Search All Databases for Search

Resources

- NCBI Home
- All Resources (A-Z)
- Literature
- DNA & RNA
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structures
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

More about the NCBI | Mission | Organization | Research | RSS

Genome Reference Consortium

Formed to improve human and mouse reference assemblies, GRC will fix loci misrepresented in reference assembly, fill remaining gaps, and make alternate representations of complex loci.

II 1 2 3 4

How To...

- Obtain the full text of an article
- Retrieve all sequences for an organism or taxon
- Find a homolog for a gene in another organism
- Find genes associated with a phenotype or disease
- Design PCR primers and check them for specificity
- Find the function of a gene or gene product
- Find syntenic regions between the genomes of two organisms

See all ...

Popular Resources

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

NCBI News

November and October News 02 Dec 2009
Featured: New Discovery-oriented PubMed and NCBI Homepage. T...

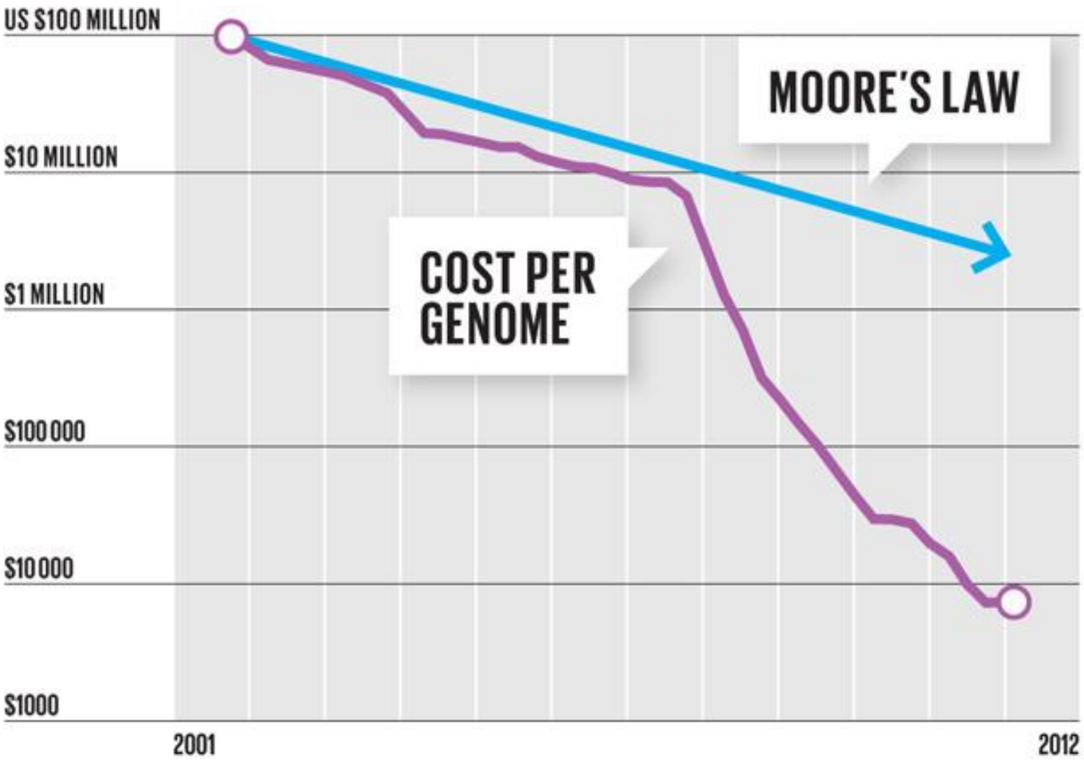
NCBI News - September 2009 05 Oct 2009
The September 2009 issue of the NCBI News is available ...

NCBI News - August 2009 19 Aug 2009
The August 2009 issue of the NCBI News is available online. ...

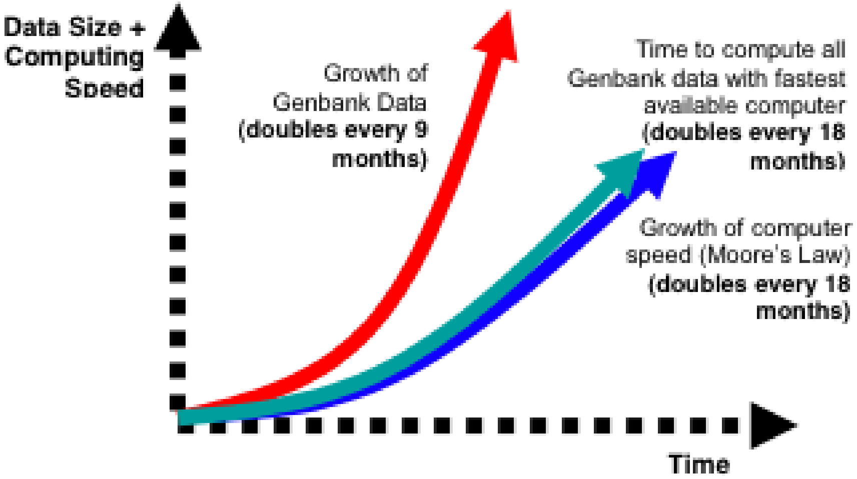
Sequencing has gotten Cheaper and Faster

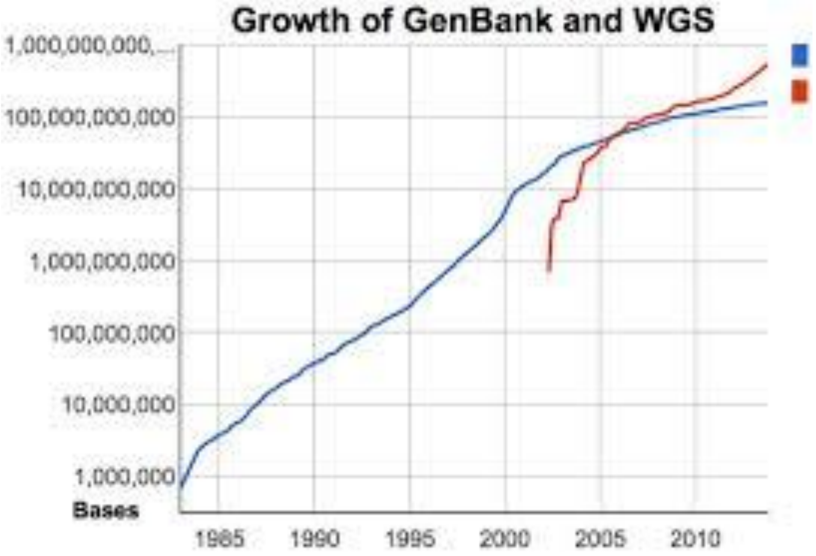
Cost of one human genome

HGP:	$ 3 billion
2004:	$30,000,000
2008:	$100,000
2010:	$10,000
2011:	**$4,000**
2012-13:	$1,000
???:	$300



BIG DATA & Sequence


Data Size + Computing Speed
Growth of Genbank Data (doubles every 9 months)
Time to compute all Genbank data with fastest available computer (doubles every 18 months)
Growth of computer speed (Moore's Law) (doubles every 18 months)
Time



🟥 Whole-Genome Shotgun (WGS)

🟦 GenBank

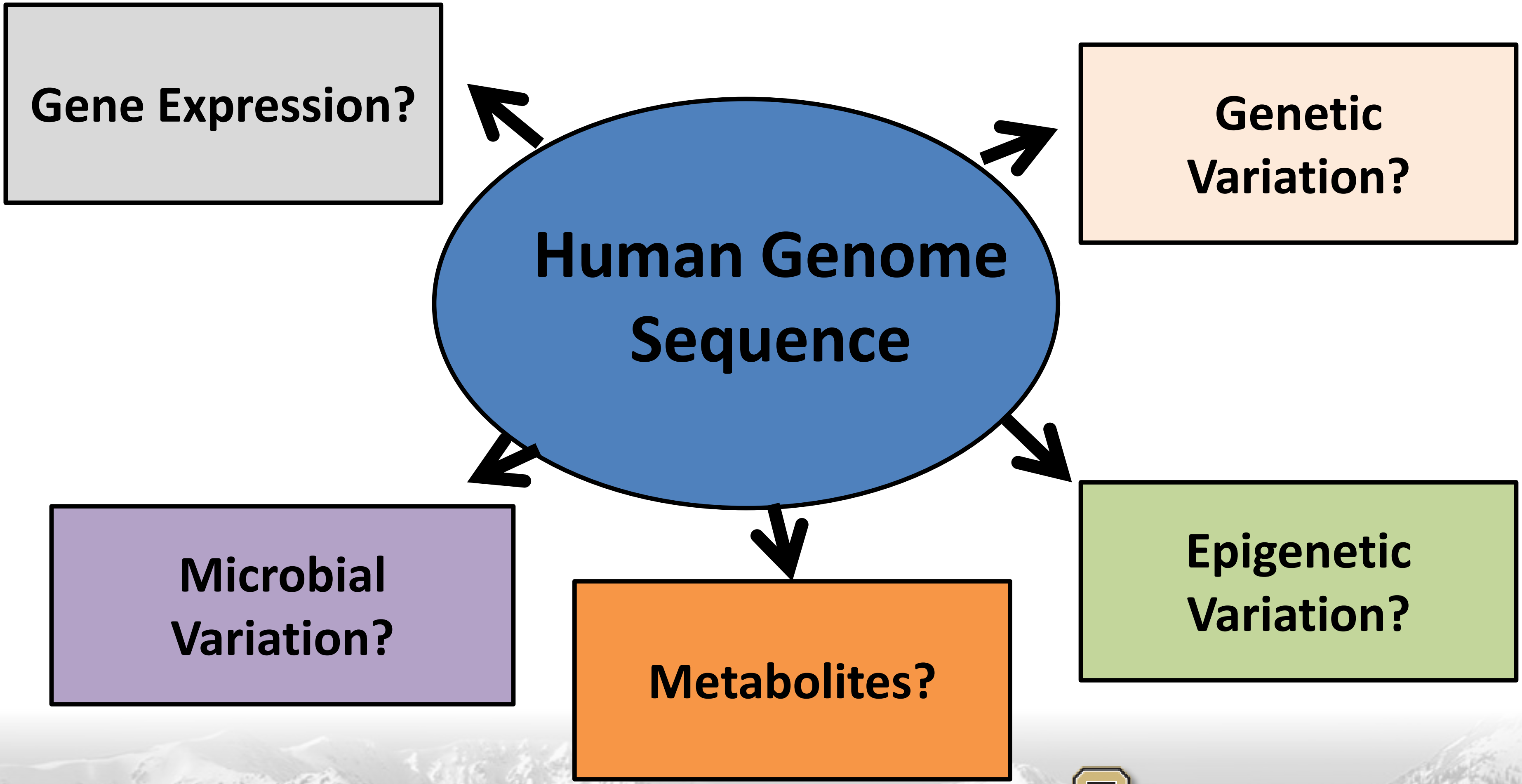
http://plone.hpcf.upr.edu/Members/humberto/class/2006/bioinformatics/bioinformatics
ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

So What Did We Learn?

- <3% of genome encodes ~20,000 genes.
- More than half of genome is repetitive.
- Identification of ~2,850 gene impact rare diseases.
- ~1,100 markers affecting common disease & ~150 targets for cancer.
- “Big Science” can win.
- Cost of sequencing per base has been reduced by magnitude of ~100k.

Lander. 2011. Nature **470** , 187-197

Emerging Applications of Sequence data



Human Genome Sequence

Gene Expression?

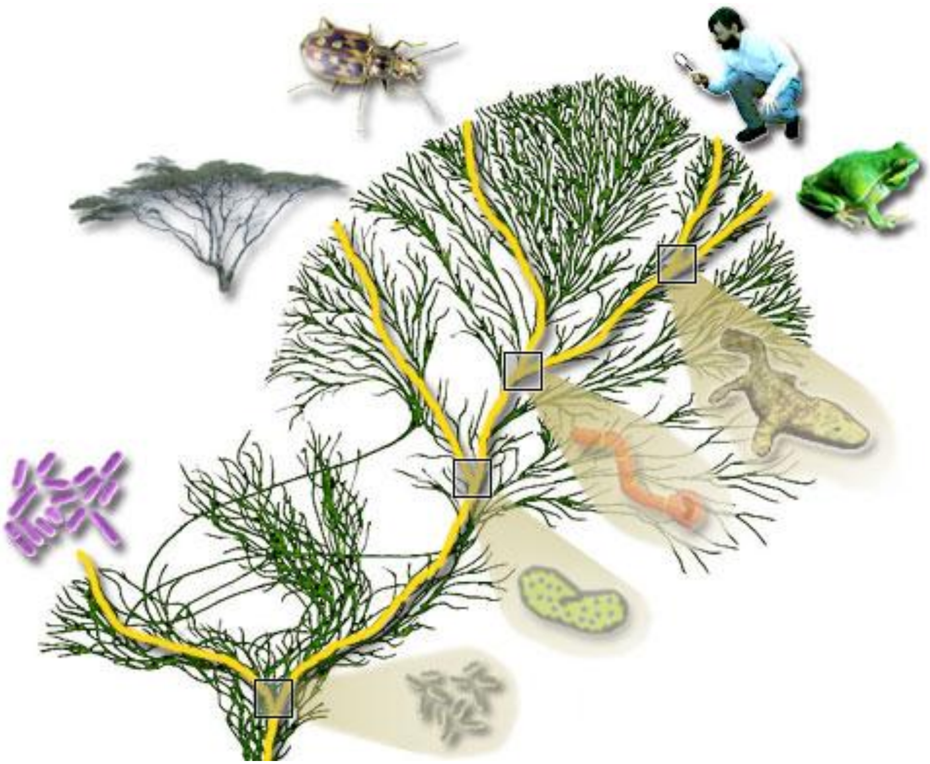
Genetic Variation?

Epigenetic Variation?

Metabolites?

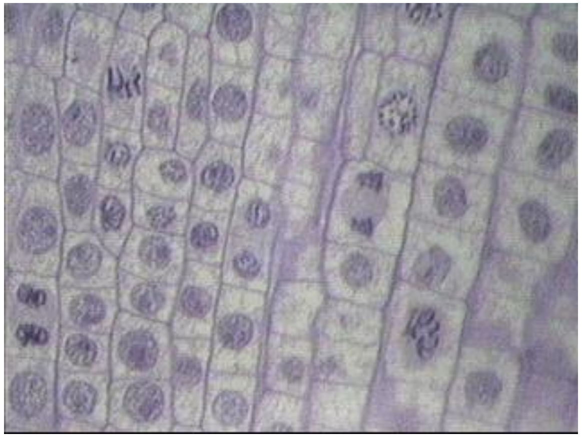
Microbial Variation?

Mapping Genetic Variation



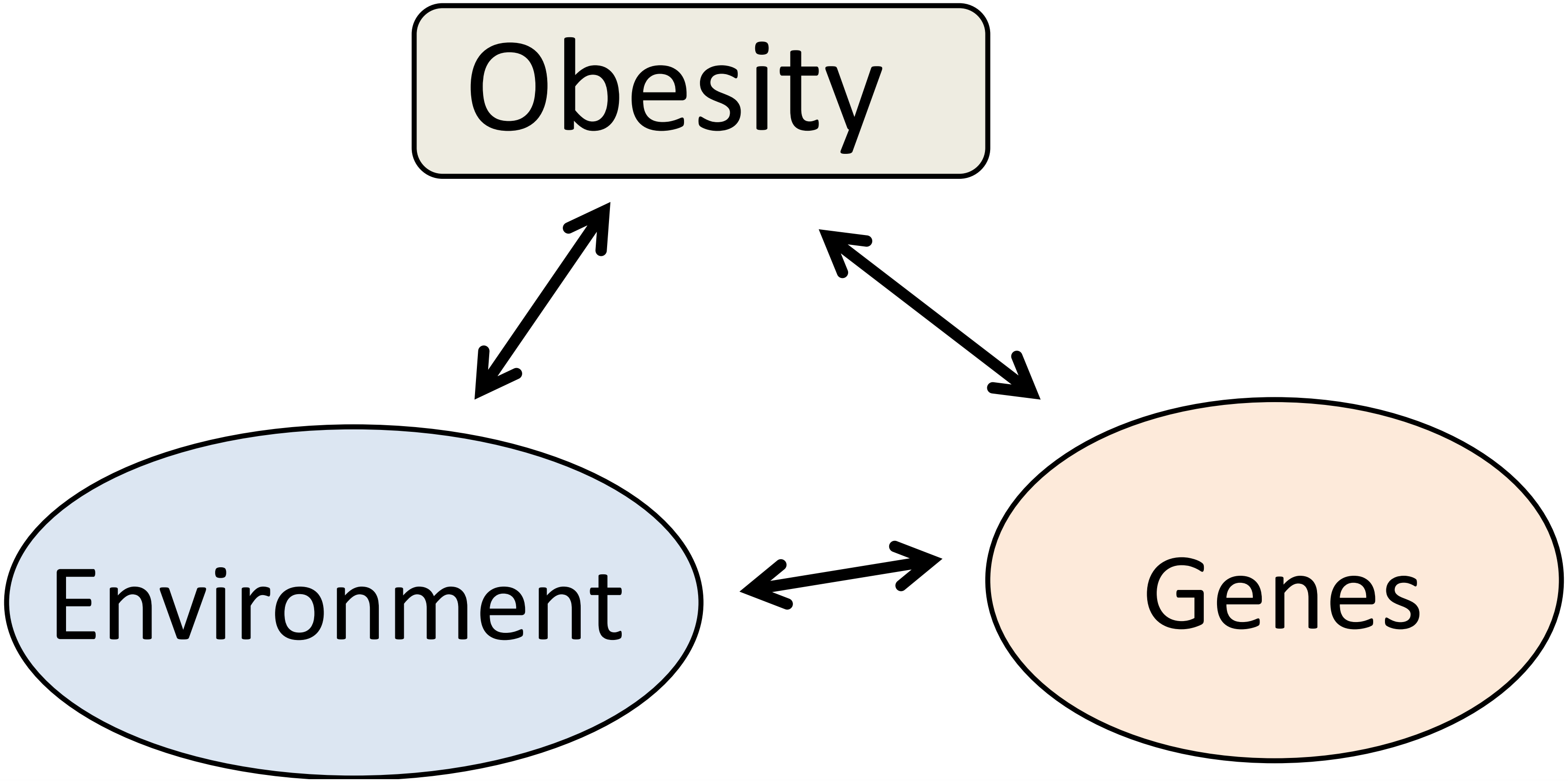
~8.7 million species

~7 billion people

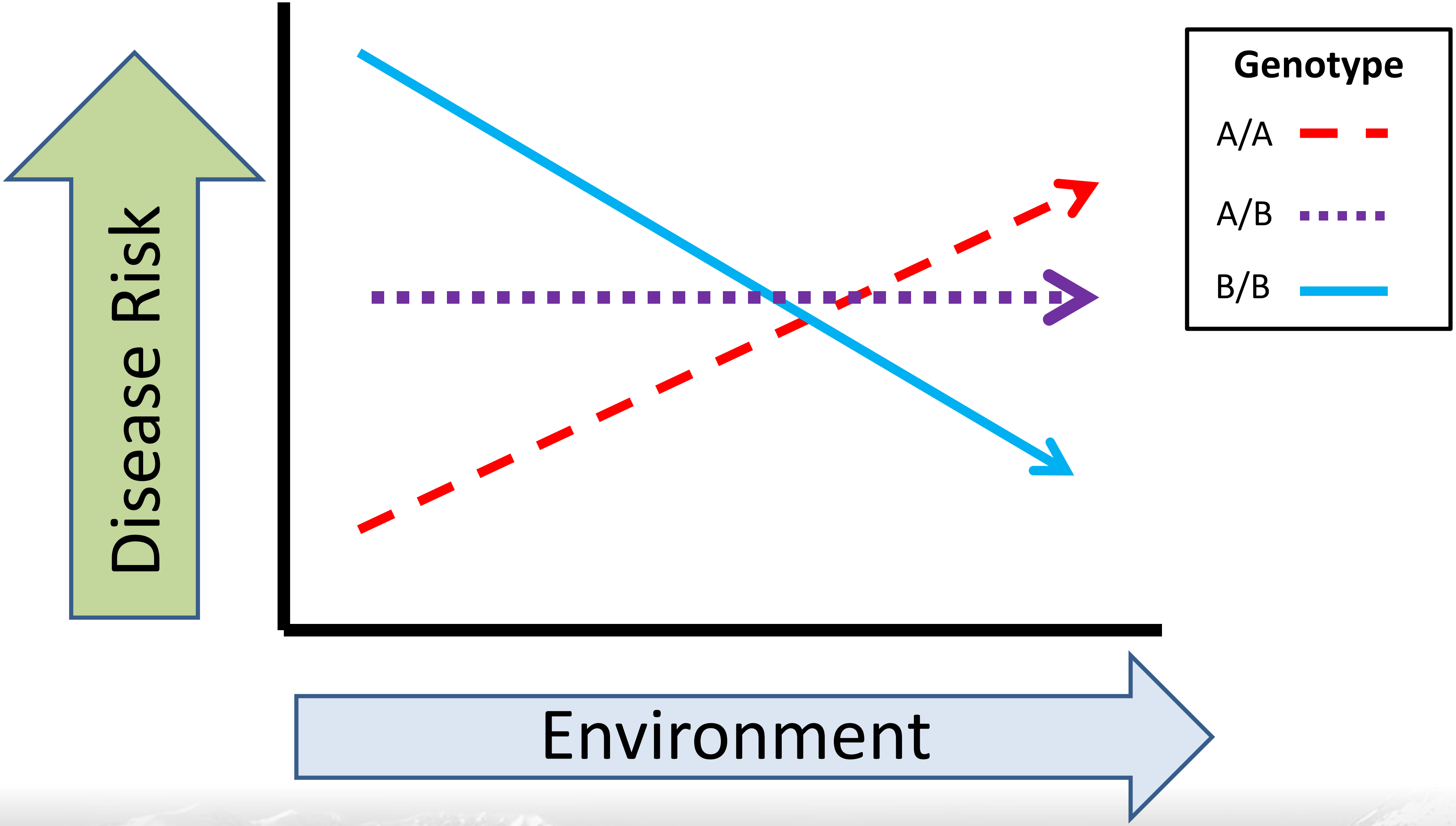


~37 trillion cells/human

Obesity is Complex



Gene-Environment Interactions



Disease Risk

Environment

Genotype

A/A

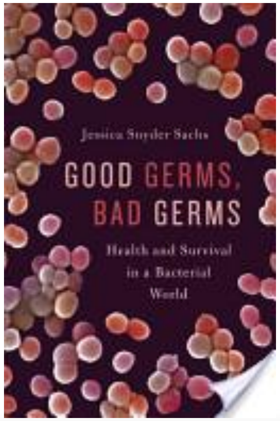
A/B

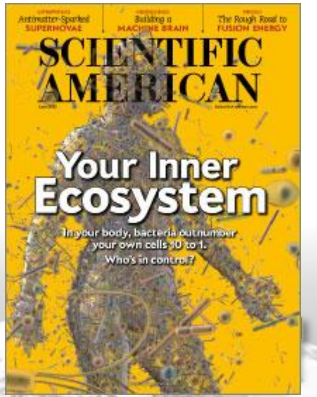
B/B

Variation in the Human Microbiome

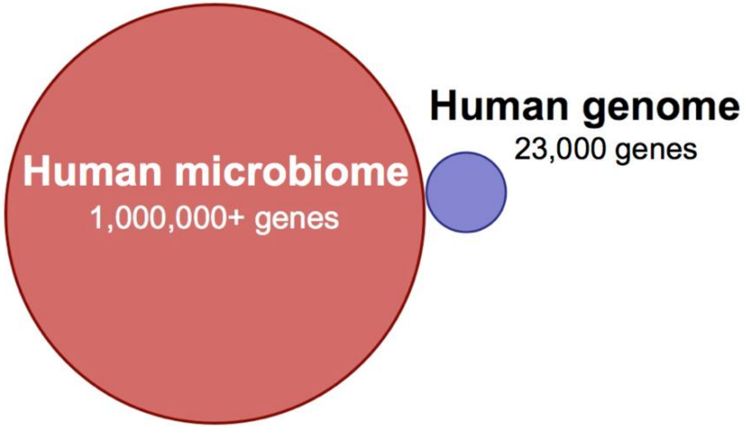
Microbe: tiny living organism, such as bacterium, fungus, protozoan, or virus.

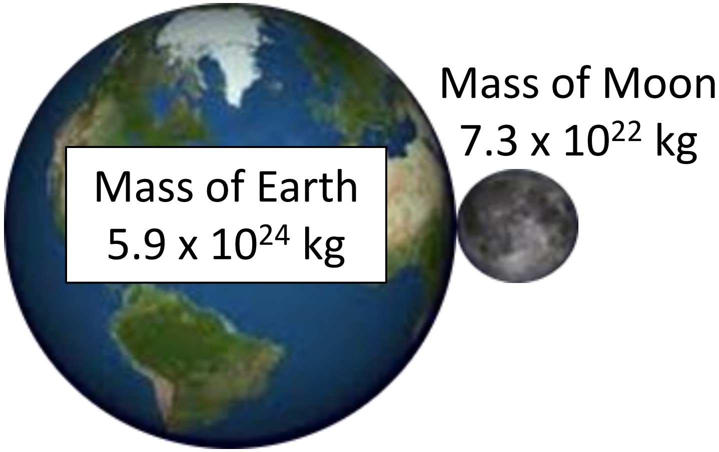
Microbiome: collectively all the microbes in the human body; a community of microbes.




SCIENTIFIC AMERICAN
Your Inner Ecosystem
In your body, bacteria outnumber your own cells 10 to 1. Who's in control?

Variation in Human Microbiome?





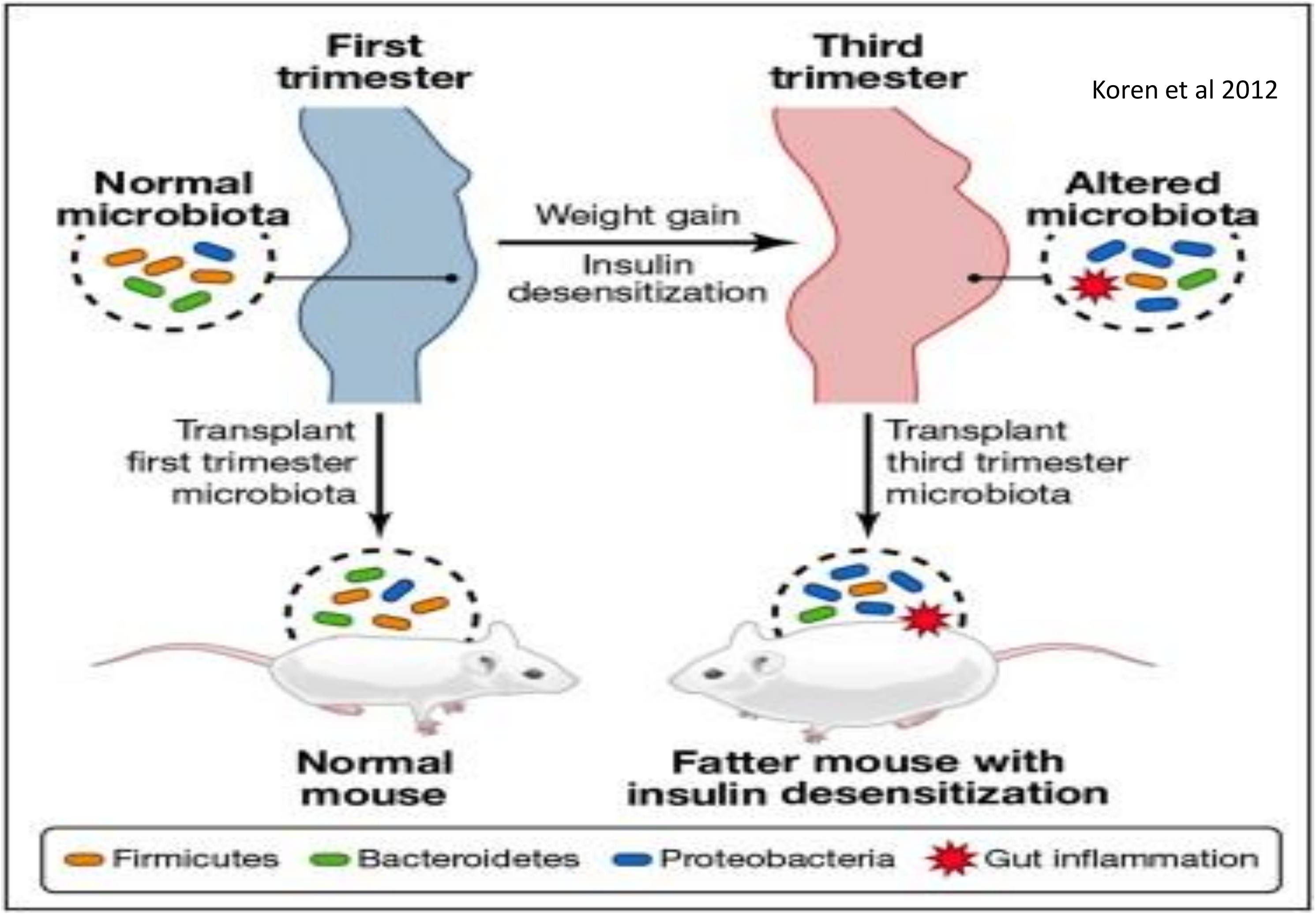
Relative to humans:

- 9 in 10 cells are microbial!
- ~1000 different species.
- ~150x more genes.
- ~3lbs of microbes in the human gut.
- ~60% of stool by dry mass is microbial.

Primary functions of the microbiome:

- Stimulate the development of our immune system.
- Resist colonization by pathogens.
- Extract energy from food.

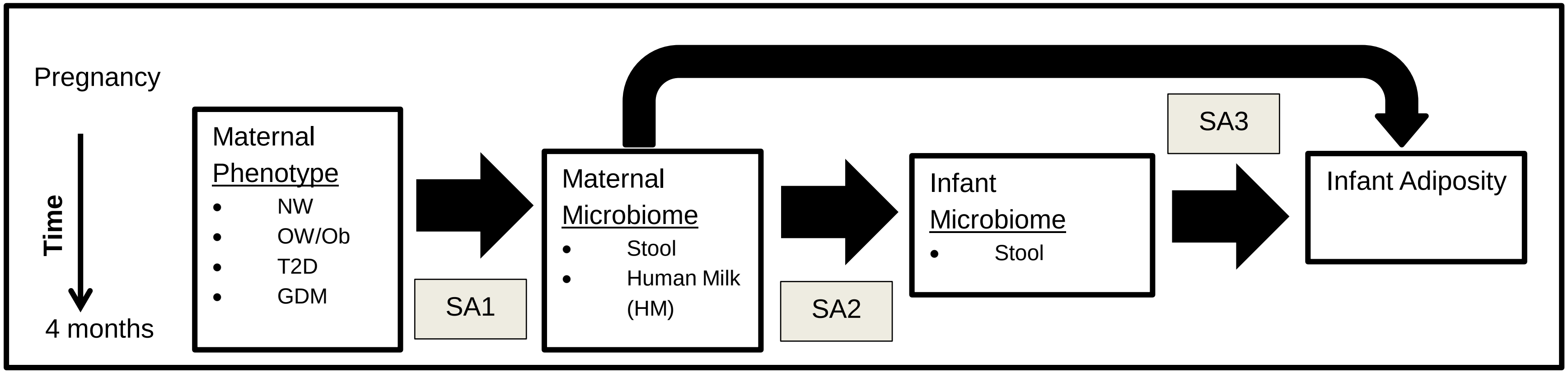
Does the microbiome impact maternal metabolism during pregnancy ?



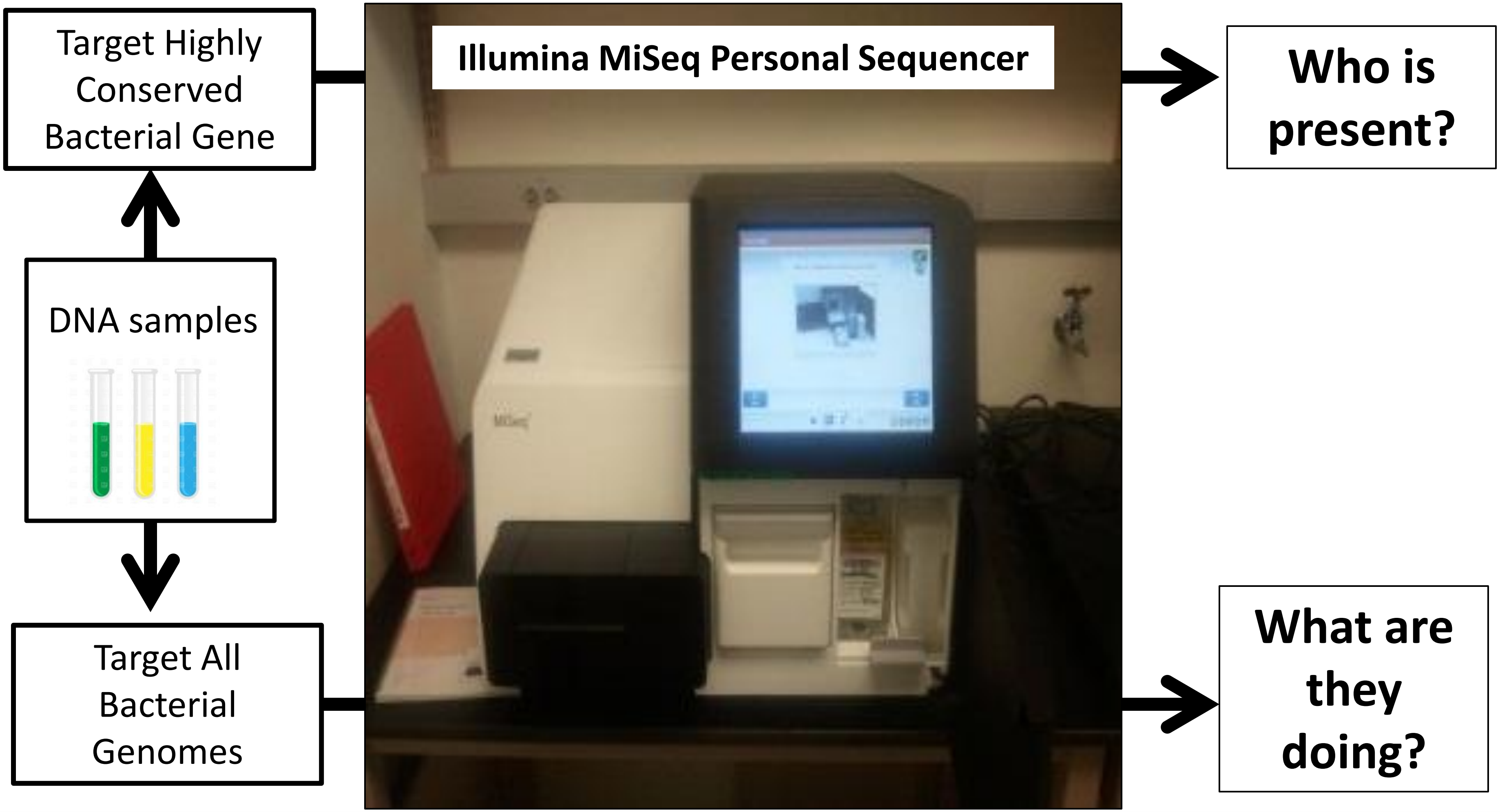
Koren et al 2012

We hypothesize the maternal microbiome in mothers will directly affect the development of the infants microbiome and adiposity during the first year of life.

Overview of Specific Aims



How Do You Measure the Microbiome?



Target Highly Conserved Bacterial Gene

DNA samples

Target All Bacterial Genomes

Illumina MiSeq Personal Sequencer

Who is present?

What are they doing?

Moving toward a Personalized Medicine?

International HapMap Project

american gut

ENCODE

23andMe

1000 Genomes

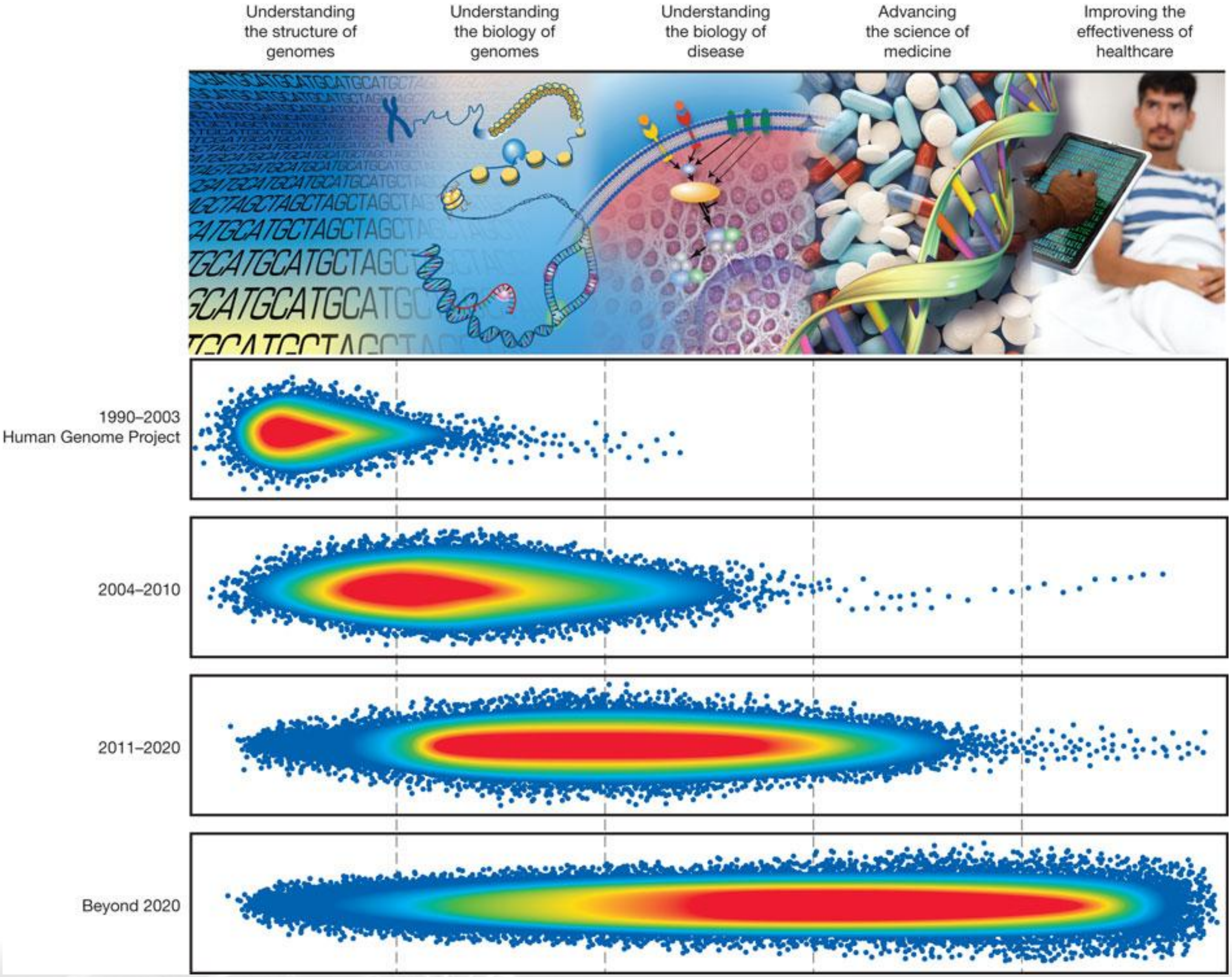
A Deep Catalog of Human Genetic Variation

X PRIZE FOUNDATION

ancestryDNA™

But has sequencing the human genome improved health?

Five Domains of Genomic Research



Green. 2011. *Nature* **470**, 204-213

Bioinformatics and Computational Biology to the Rescue?

Data Analysis- existing tools are becoming inadequate to analyze data.

Data Integration- Need to harmonize disparate data types.

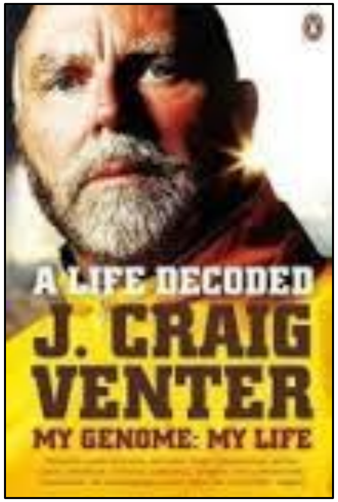
Visualization- Need to accommodate multi-dimensional data.

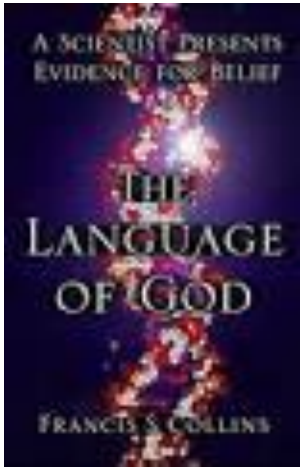
Computational Tools and Infrastructure- storage, capacity, privacy.

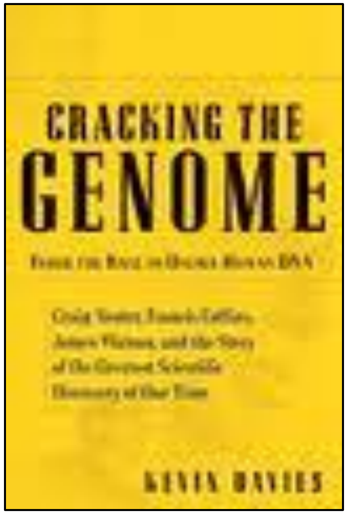
Training- Need biologist, computer science, informatics, data science . . . . . .

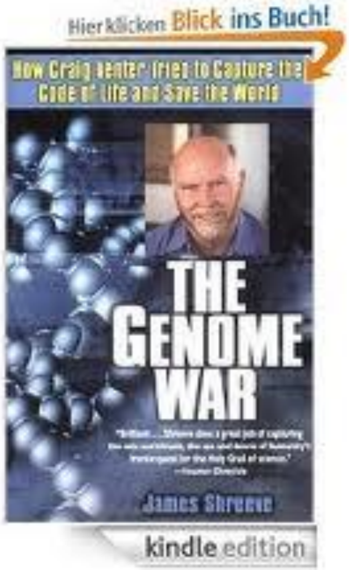
Green. 2011. *Nature* **470**, 204-213

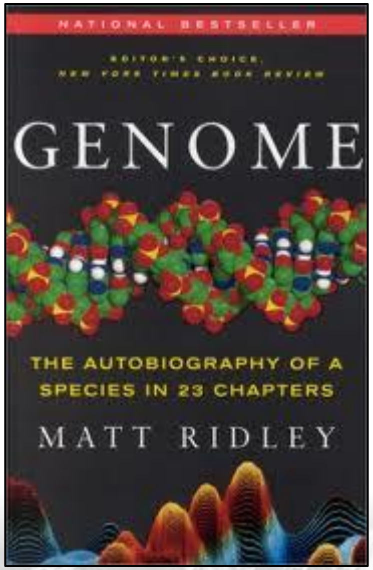
Recommended Reading

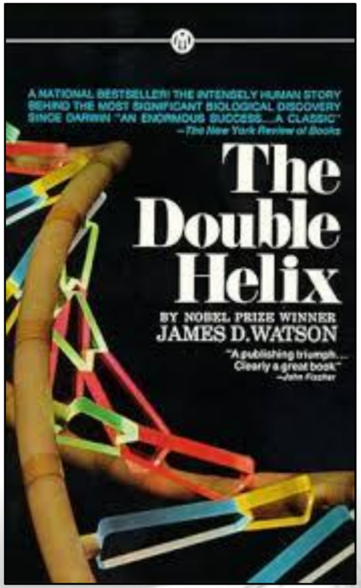


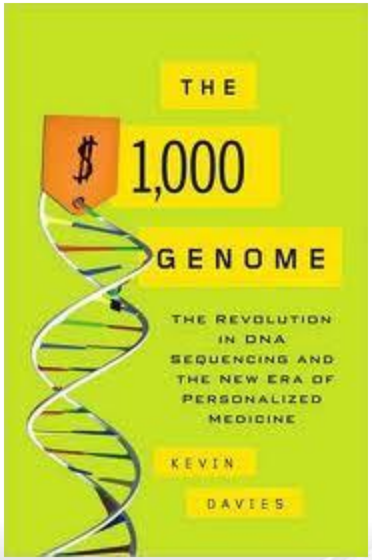



CRACKING THE GENOME
Inside the Race to Unlock Human DNA
Craig Venter, Francis Collins, James Watson, and the Story of the Greatest Scientific Discovery of Our Time
KEVIN DAVIES


Hier klicken Blick ins Buch!
How Craig Venter Tried to Capture the Code of Life and Save the World
THE GENOME WAR
James Shreeve
kindle edition


NATIONAL BESTSELLER
EDITOR'S CHOICE
NEW YORK TIMES BOOK REVIEW
GENOME
THE AUTOBIOGRAPHY OF A SPECIES IN 23 CHAPTERS
MATT RIDLEY


A NATIONAL BESTSELLER! THE INTENSELY HUMAN STORY BEHIND THE MOST SIGNIFICANT BIOLOGICAL DISCOVERY SINCE DARWIN "AN ENORMOUS SUCCESS...A CLASSIC" —The New York Review of Books
The Double Helix
BY NOBEL PRIZE WINNER JAMES D. WATSON
"A publishing triumph... Clearly a great book" —John Fischer


THE $1,000 GENOME
THE REVOLUTION IN DNA SEQUENCING AND THE NEW ERA OF PERSONALIZED MEDICINE
KEVIN DAVIES

Special Thanks to:

University of Colorado Denver

My mentors

- Jacob E. Friedman, PhD
- Daniel N. Frank, PhD
- Stephanie A. Santorico, PhD
- Linda A. Barbour, MD, MSPH
- Dana Dabelea, MD, PhD

Funding Support:

- ADA/Glaxo-SmithKline
- T32-HD07186

Questions?