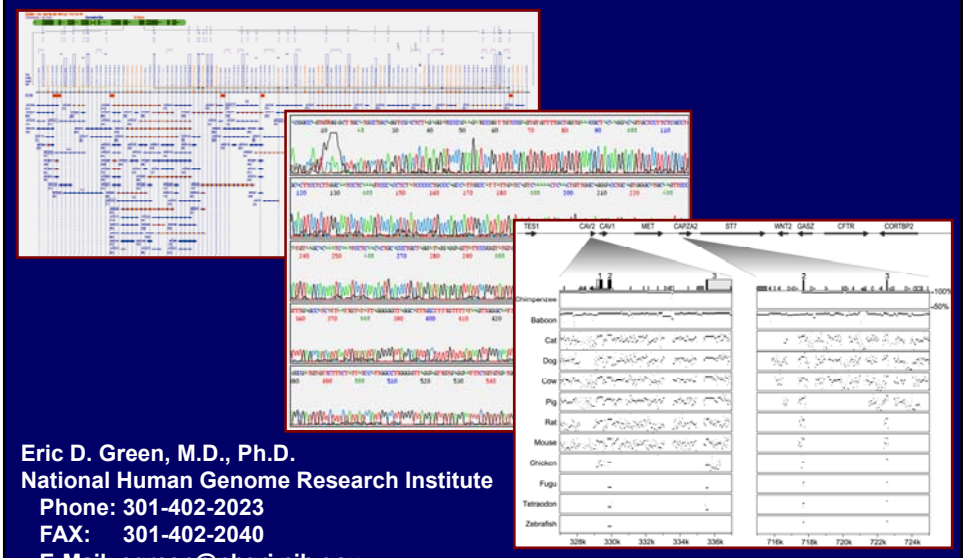
Techniques for Analyzing Genomes I



Eric D. Green, M.D., Ph.D.
National Human Genome Research Institute
Phone: 301-402-2023
FAX: 301-402-2040
E-Mail: egreen@nhgri.nih.gov

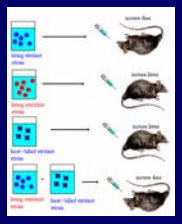
Foundational Milestones in Genetics & Genomics

Mendel

1865

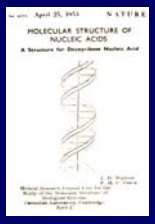
Miescher

1871



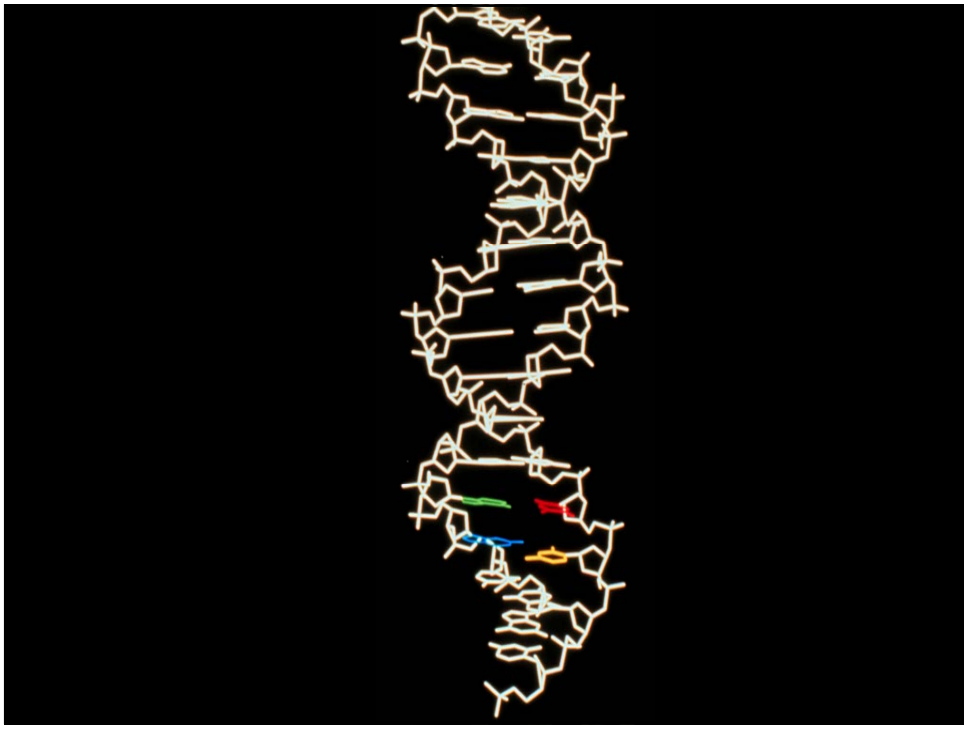
Avery

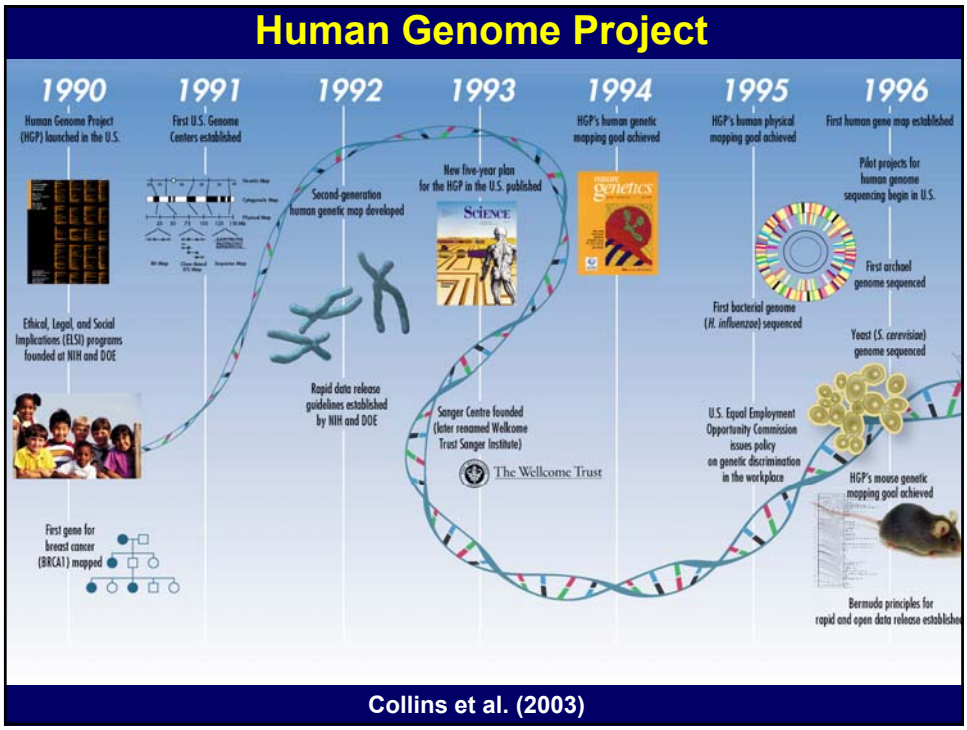
1944



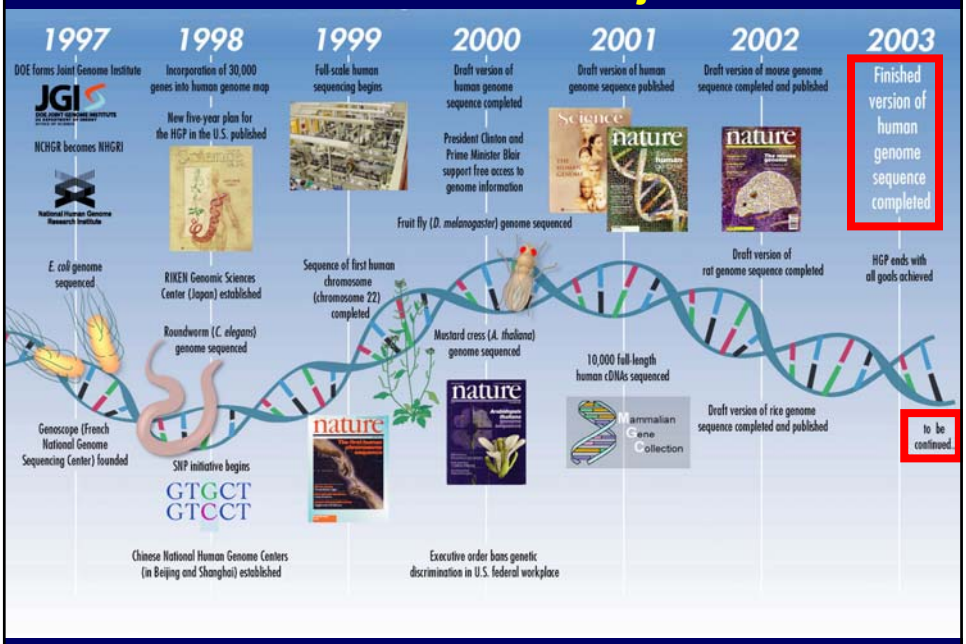
Watson & Crick

1953




Human Genome Project
1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
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
SCIENCE
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
nature genetics
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
Bermuda principles for rapid and open data release establishe
Collins et al. (2003)

Human Genome Project



Collins et al. (2003)

Outline

I. Fundamentals of Genome Mapping

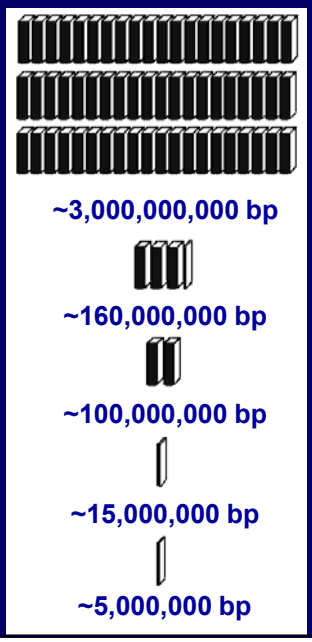
II. Fundamentals of Genome Sequencing

III. Mapping & Sequencing in the Human Genome Project

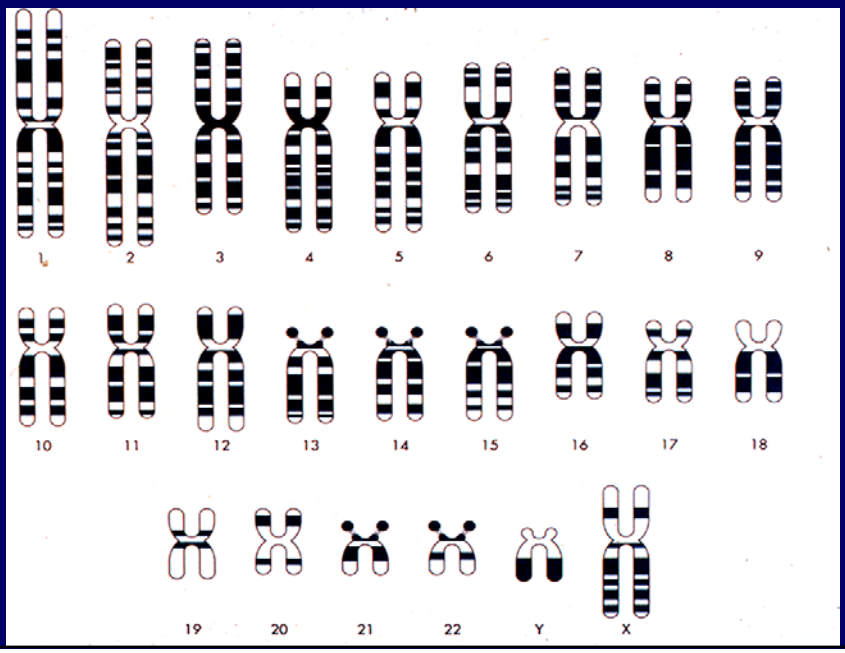
IV. Comparative Sequencing

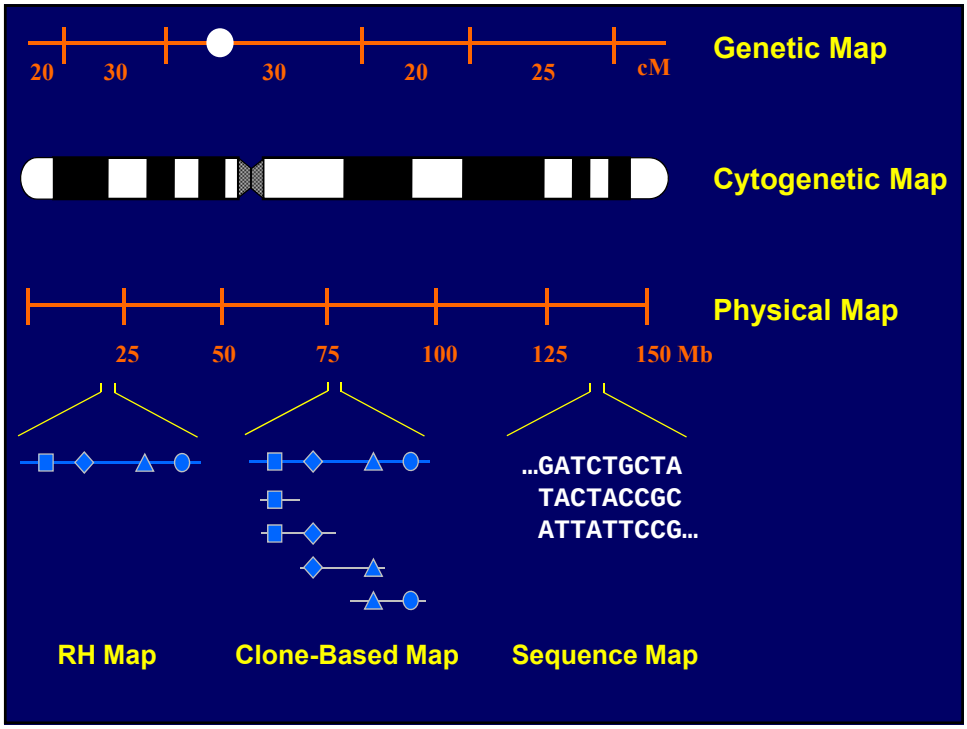
V. New Frontiers in Genome Analysis

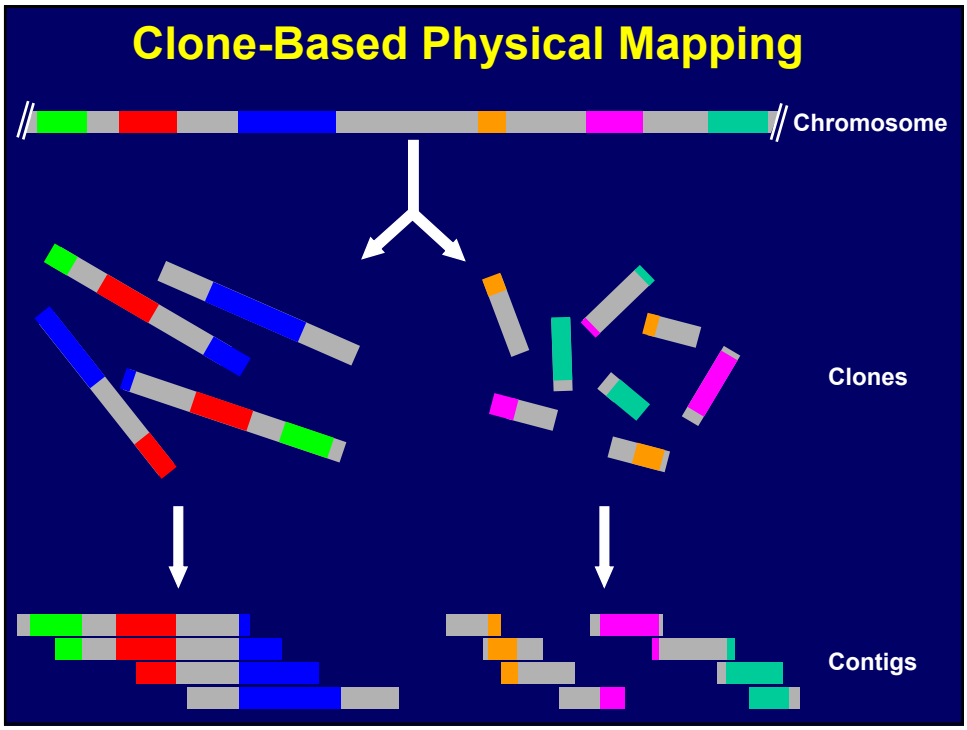
Genome Sizes

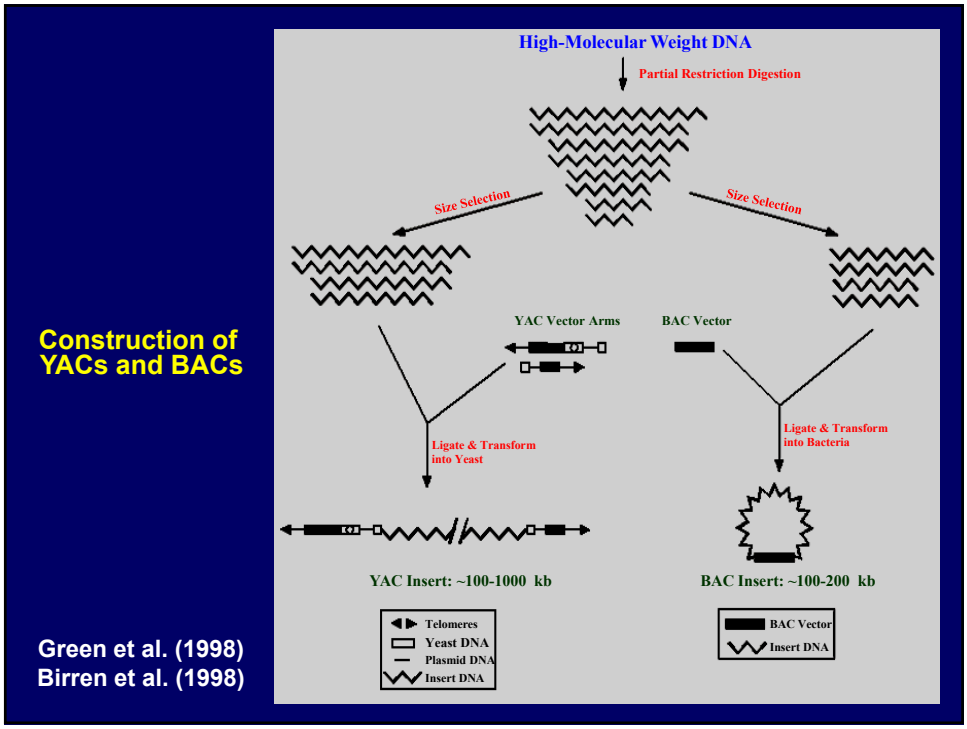


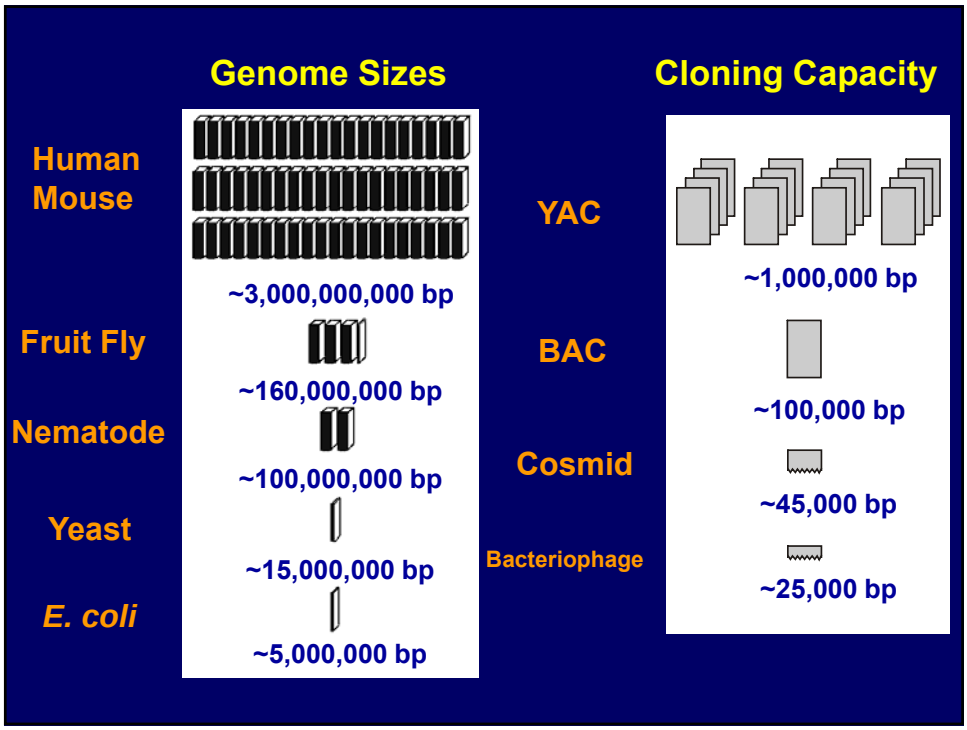
The Human Cytogenetic Map




Genetic Map
20
30
30
20
25
cM
Cytogenetic Map
Physical Map
25
50
75
100
125
150 Mb
...GATCTGCTA
TACTACCGC
ATTATTCCG...
RH Map
Clone-Based Map
Sequence Map


Clone-Based Physical Mapping
Chromosome
Clones
Contigs

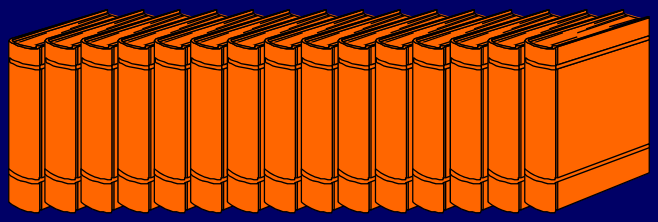

Construction of YACs and BACs
High-Molecular Weight DNA
Partial Restriction Digestion
Size Selection
Size Selection
YAC Vector Arms
BAC Vector
Ligate & Transform into Yeast
Ligate & Transform into Bacteria
YAC Insert: ~100-1000 kb
BAC Insert: ~100-200 kb
Telomeres
Yeast DNA
Plasmid DNA
Insert DNA
BAC Vector
Insert DNA
Green et al. (1998)
Birren et al. (1998)


Genome Sizes
Cloning Capacity
Human
Mouse
~3,000,000,000 bp
Fruit Fly
~160,000,000 bp
Nematode
~100,000,000 bp
Yeast
~15,000,000 bp
E. coli
~5,000,000 bp
YAC
~1,000,000 bp
BAC
~100,000 bp
Cosmid
~45,000 bp
Bacteriophage
~25,000 bp

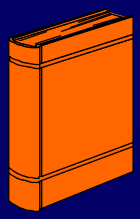
Bacterial Artificial Chromosomes (BACs)

- Bacterial-Based Cloning System
- Based on the *E. coli* F Factor (Fertility Plasmid): Replication Control
- Cloned Inserts: 100-200 kb, Circular DNA
- Low Copy Number

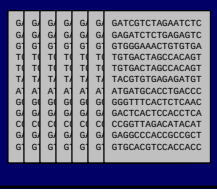
  Low Yields of DNA by Standard Methods
  Reasonably Stable
- See Birren et al. (1998)
- Availability of BAC Libraries from Many Vertebrate Species
  (e.g., www.chori.org/bacpac)



Genome
(~3000 Mb)



Chromosome
(~130 Mb)

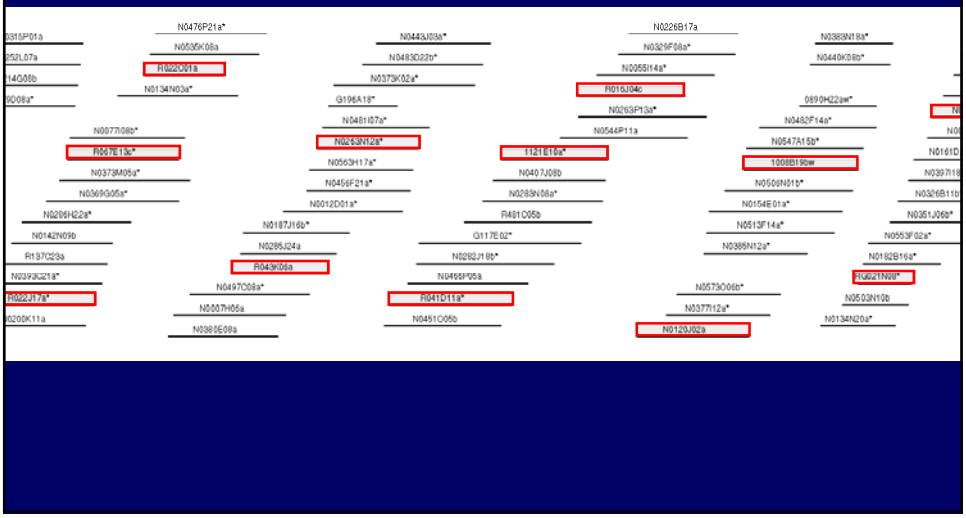


YAC
(~0.5-1.0 Mb)

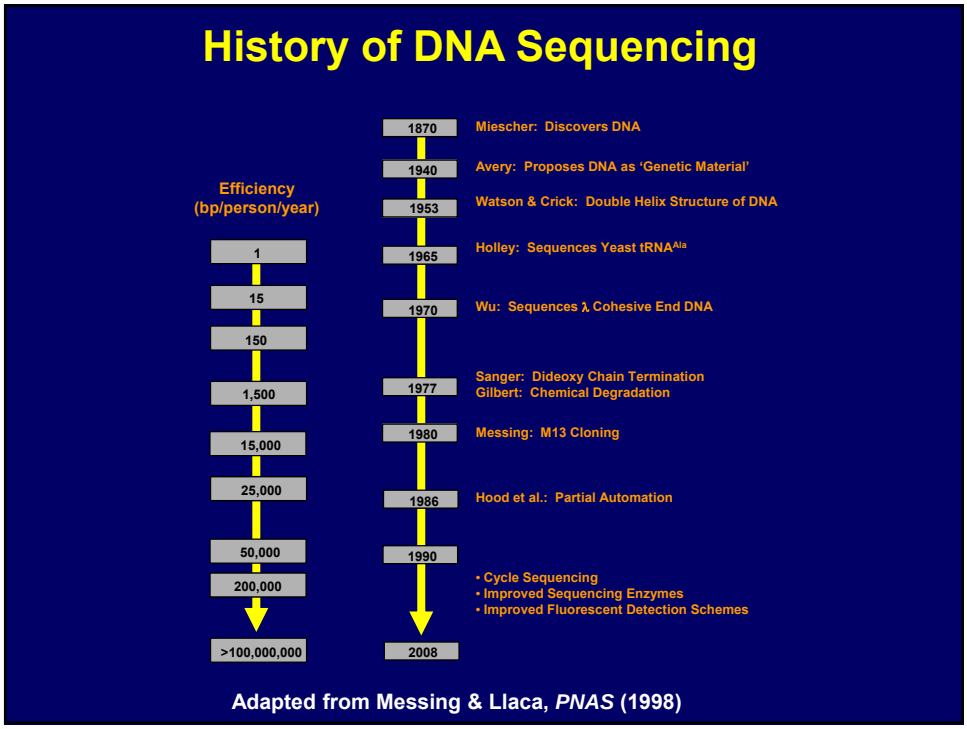
BAC
(~0.1-0.2 Mb)

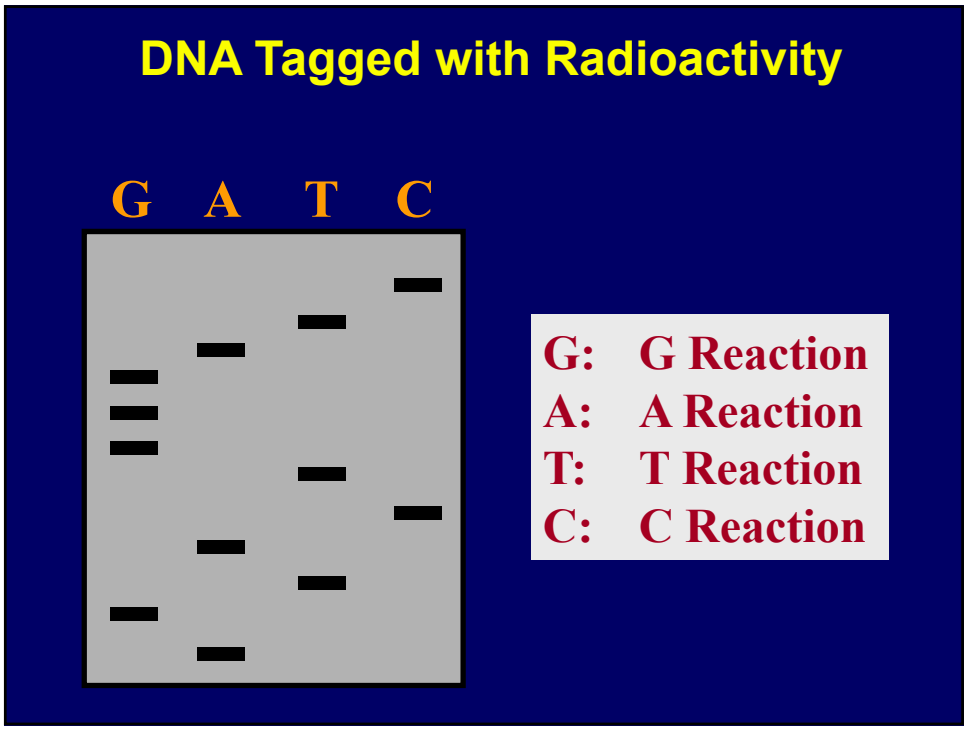
Sequence-Ready BAC Contig Map

Marra et al. (1997)



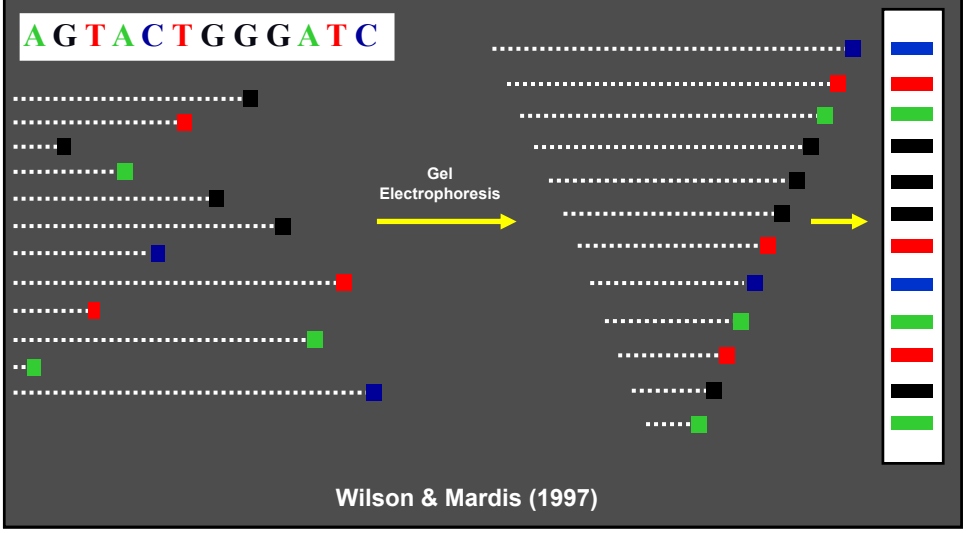
DNA Sequencing


History of DNA Sequencing
Efficiency (bp/person/year)
1
15
150
1,500
15,000
25,000
50,000
200,000
>100,000,000
1870
1940
1953
1965
1970
1977
1980
1986
1990
2008
Miescher: Discovers DNA
Avery: Proposes DNA as 'Genetic Material'
Watson & Crick: Double Helix Structure of DNA
Holley: Sequences Yeast tRNA Ala
Wu: Sequences λ Cohesive End DNA
Sanger: Dideoxy Chain Termination
Gilbert: Chemical Degradation
Messing: M13 Cloning
Hood et al.: Partial Automation
• Cycle Sequencing
• Improved Sequencing Enzymes
• Improved Fluorescent Detection Schemes
Adapted from Messing & Llaca, PNAS (1998)

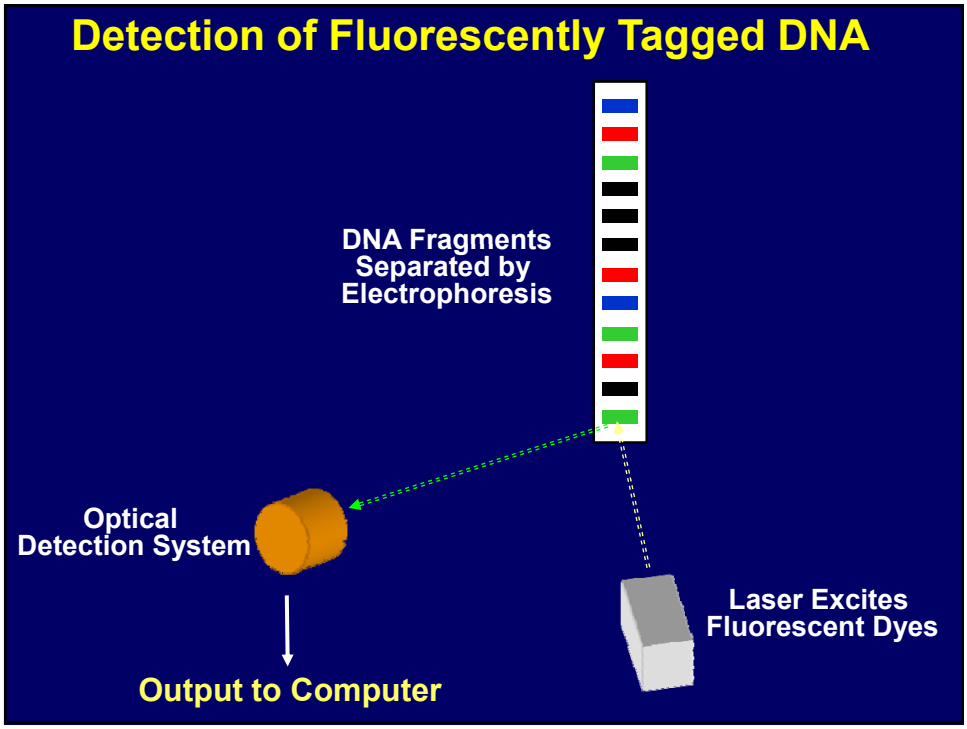

DNA Tagged with Radioactivity
G A T C
G: G Reaction
A: A Reaction
T: T Reaction
C: C Reaction

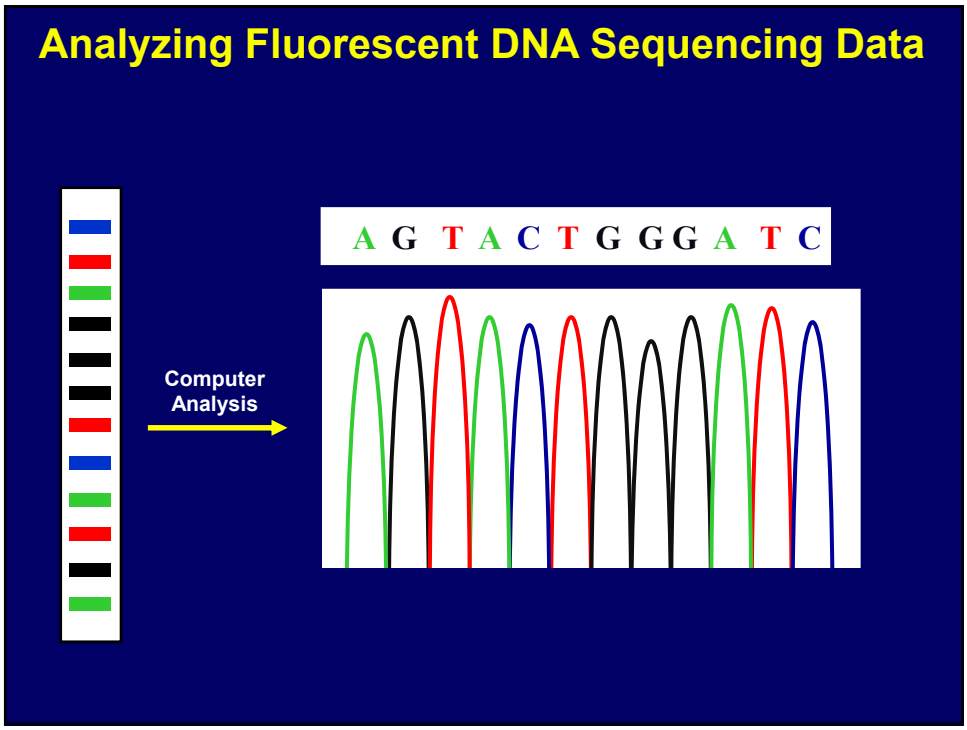
Radioactive Sequencing

Fluorescent DNA Sequencing

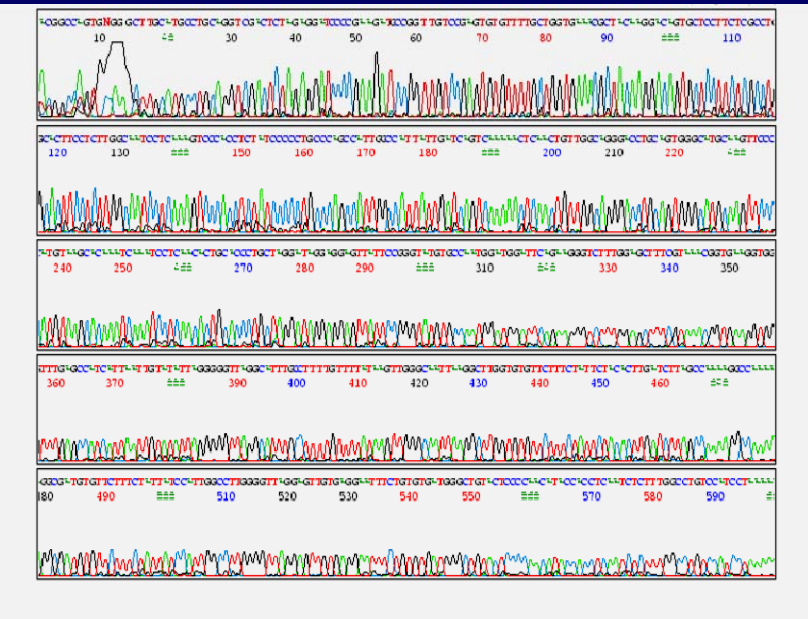


Wilson & Mardis (1997)


Detection of Fluorescently Tagged DNA
DNA Fragments Separated by Electrophoresis
Optical Detection System
Laser Excites Fluorescent Dyes
Output to Computer


Analyzing Fluorescent DNA Sequencing Data
Computer Analysis
A G T A C T G G G A T C

Fluorescent DNA Sequencing Results



Slab Gel-Based DNA Sequencing Instruments

Capillary-Based DNA Sequencing Instruments



Large-Scale cDNA Sequencing

- ESTs: Expressed-Sequence Tags
- SAGE: Serial Analysis of Gene Expression
- Full-Insert (Full-Length) cDNA Sequencing

Mammalian Gene Collection

mgc.nci.nih.gov

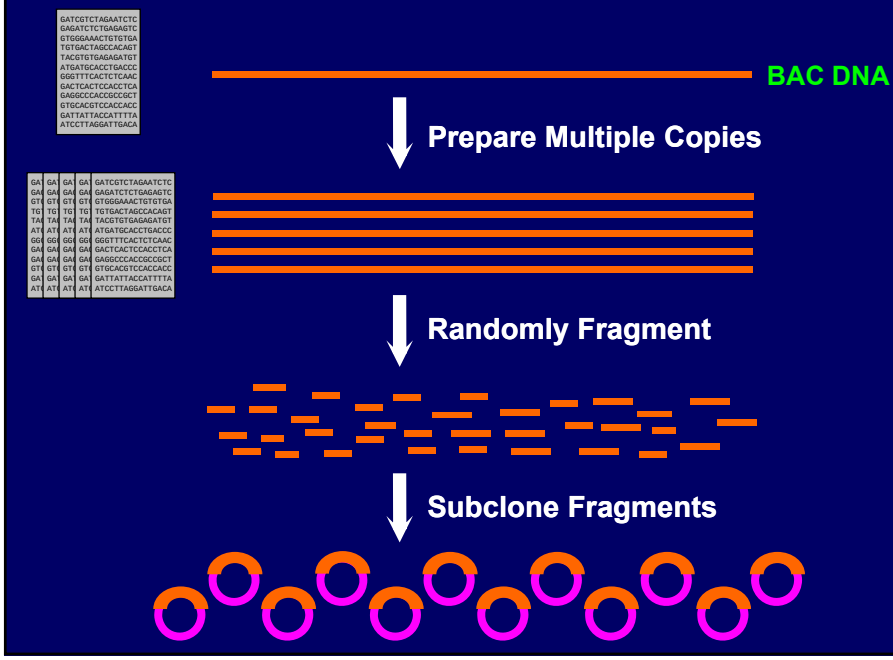
Gerhard et al. (2004)

Large-Scale Genome Sequencing

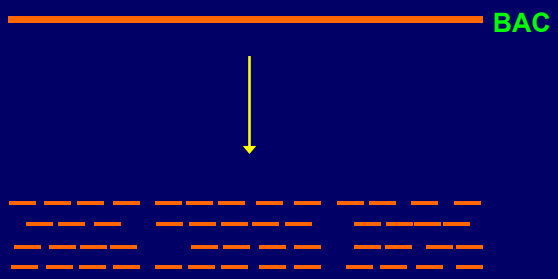
Shotgun Sequencing

Wilson & Mardis (1997)
Green (2001)

Subclone Construction



Shotgun Sequencing Strategy



Poisson Calculations

The sequencing strategy for the shotgun approach follows the Lander and Waterman application of the Poisson distribution

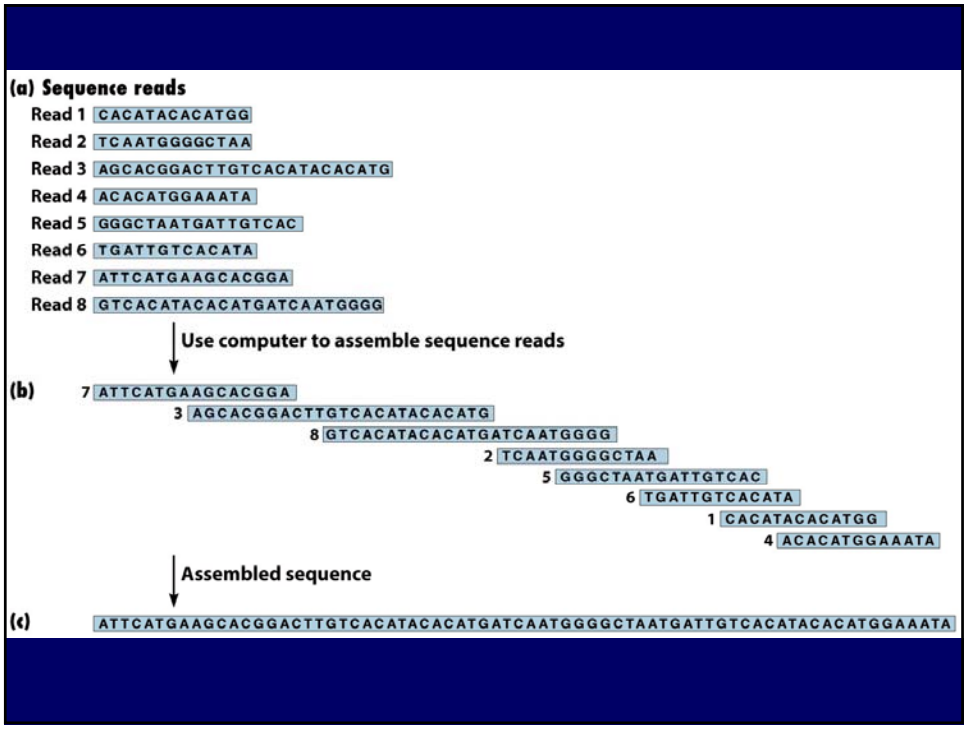
The probability a base is not sequenced is given by:

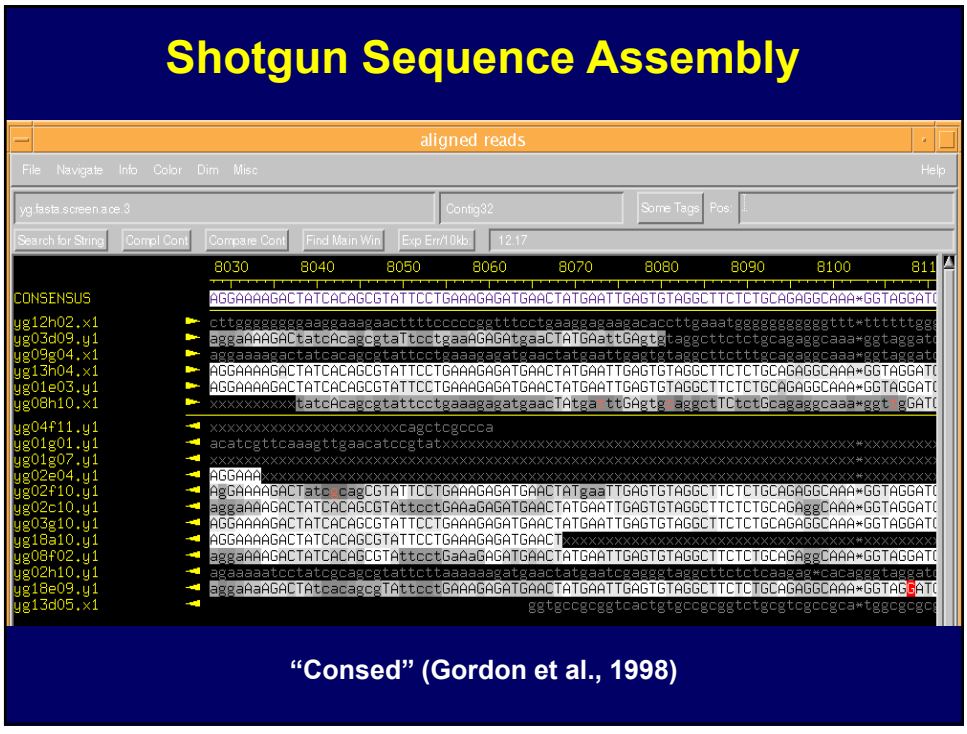
$$P_0 = e^{-c}$$

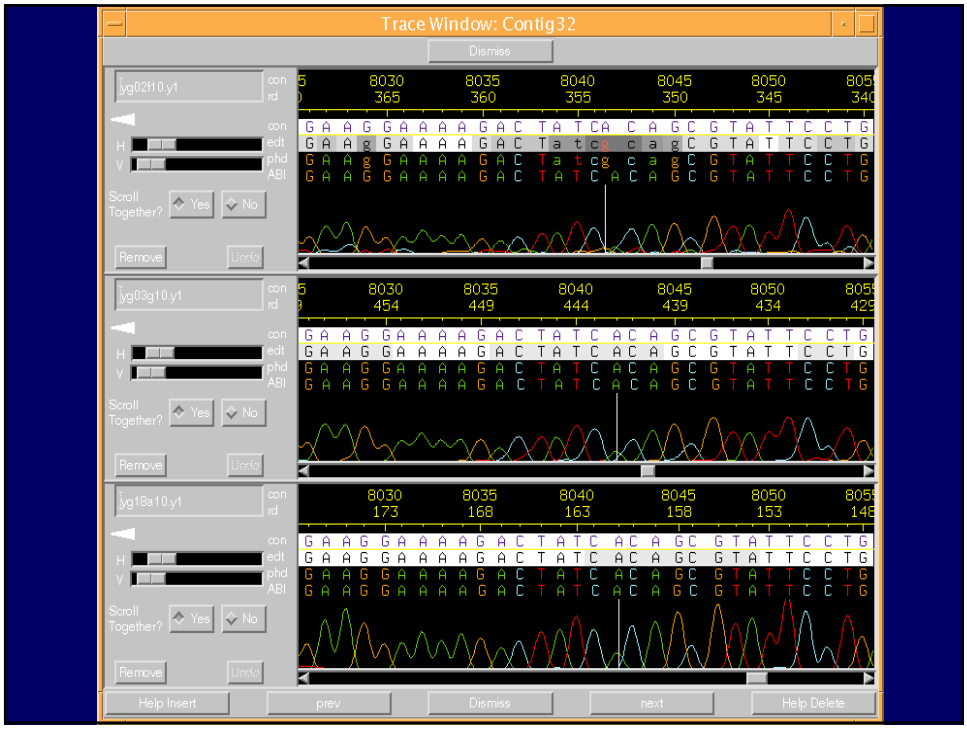
Where:

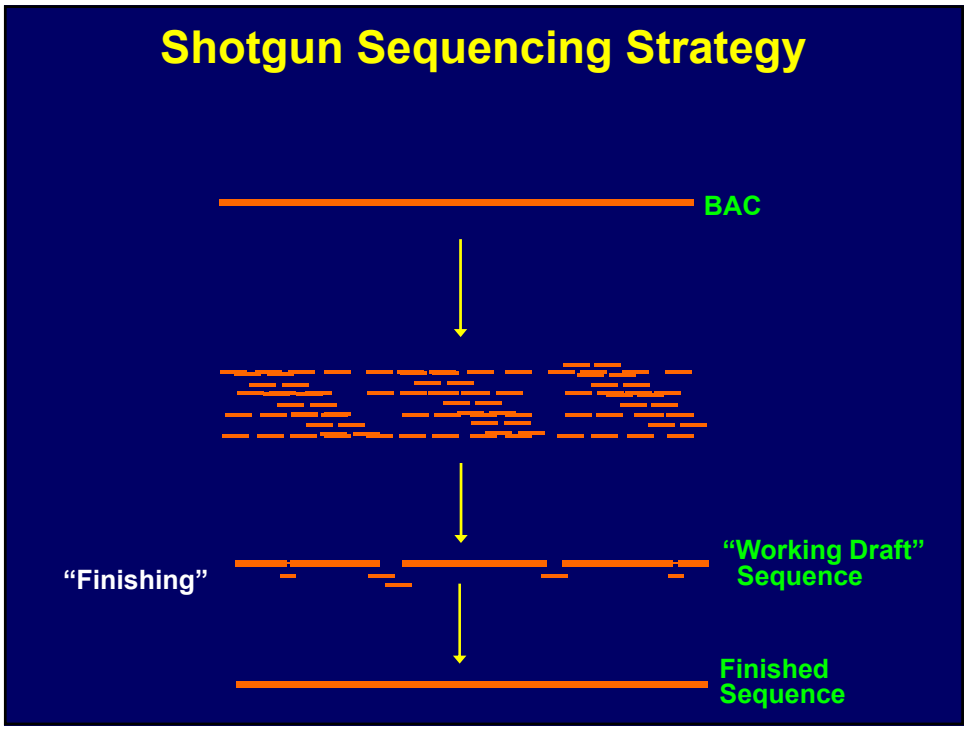
- c = fold sequence coverage (c=LN/G),
- LN = # bases sequenced, i.e. L = average sequencing read length and N = # reads
- G = target sequence length
- e = 2.718 (e=2.718281828459)

Fold Coverage	$P_0=e^{-c}$	% not sequenced	% sequenced
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%


(a) Sequence reads
Read 1 CACATACACATGG
Read 2 TCAATGGGGCTAA
Read 3 AGCACGGACTTGTCACATACACATG
Read 4 ACACATGGAAATA
Read 5 GGGCTAATGATTGTCAC
Read 6 TGATTGTCACATA
Read 7 ATTCATGAAGCACGGA
Read 8 GTCACATACACATGATCAATGGGG
Use computer to assemble sequence reads
(b)
7 ATTCATGAAGCACGGA
3 AGCACGGACTTGTCACATACACATG
8 GTCACATACACATGATCAATGGGG
2 TCAATGGGGCTAA
5 GGGCTAATGATTGTCAC
6 TGATTGTCACATA
1 CACATACACATGG
4 ACACATGGAAATA
Assembled sequence
(c)
ATTCATGAAGCACGGACTTGTCACATACACATGATCAATGGGGCTAATGATTGTCACATACACATGGAAATA


Shotgun Sequence Assembly
aligned reads
File Navigate Info Color Dim Misc Help
Search for String
Compl Cont
Compare Cont
Find Main Win
Exp Err/10kb
8030 8040 8050 8060 8070 8080 8090 8100
CONSENSUS
yg12h02.x1
yg03d09.y1
yg09g04.x1
yg13h04.x1
yg01e03.y1
yg08h10.x1
yg04f11.y1
yg01g01.y1
yg01g07.y1
yg02e04.y1
yg02f10.y1
yg02c10.y1
yg03g10.y1
yg18a10.y1
yg08f02.y1
yg02h10.y1
yg18e09.y1
yg13d05.x1
"Consed" (Gordon et al., 1998)

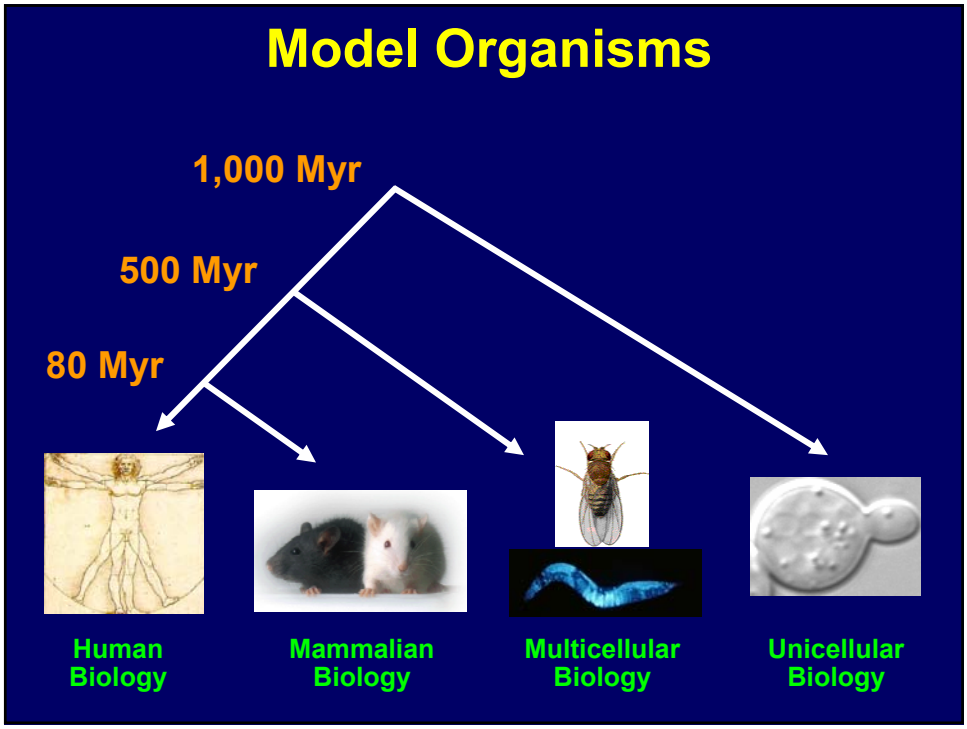

Trace Window: Contig32
Dismiss
Scroll Together?
Yes
No
Remove
Help Insert
prev
Dismiss
next
Help Delete

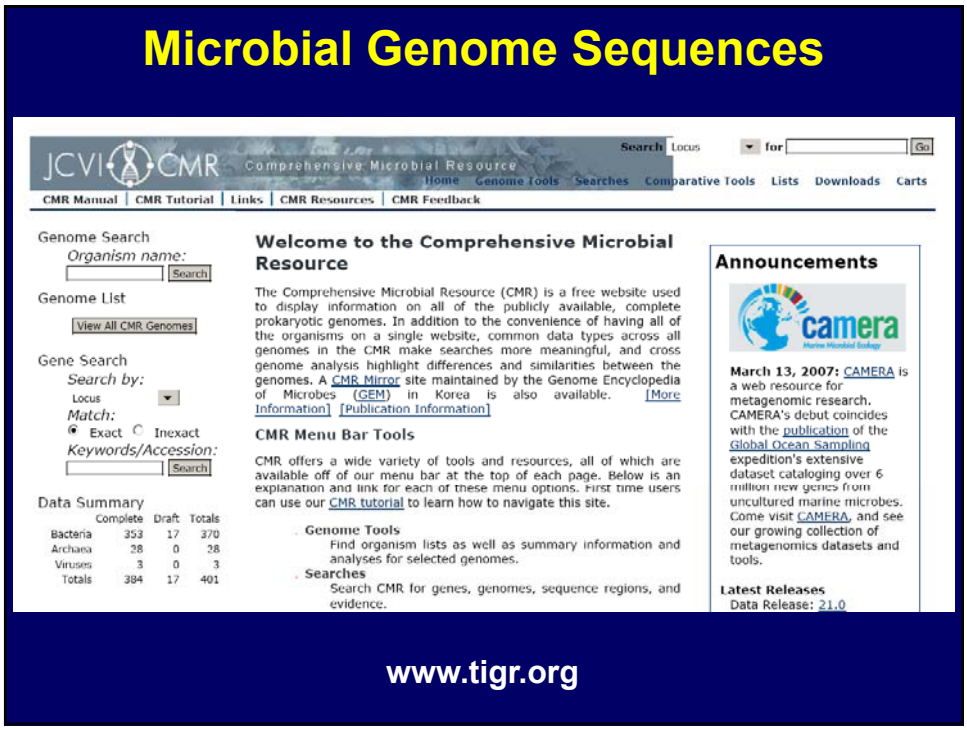

Shotgun Sequencing Strategy
BAC
"Finishing"
"Working Draft" Sequence
Finished Sequence

Sequence Finishing: Resolving Ambiguities

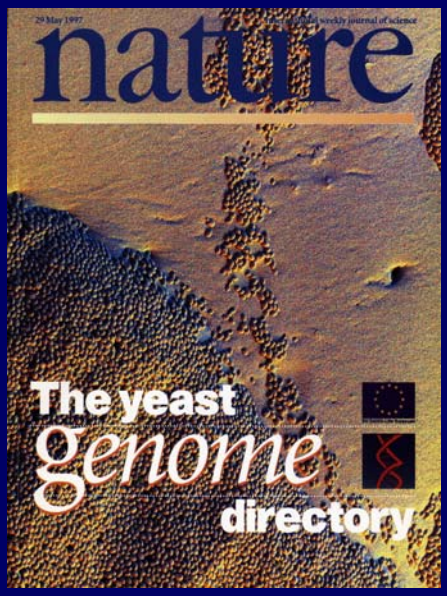
*** Sequence Finishing: Remains Relatively Expensive ***

Historically Significant Genome Sequencing Projects


Model Organisms
1,000 Myr
500 Myr
80 Myr
Human Biology
Mammalian Biology
Multicellular Biology
Unicellular Biology

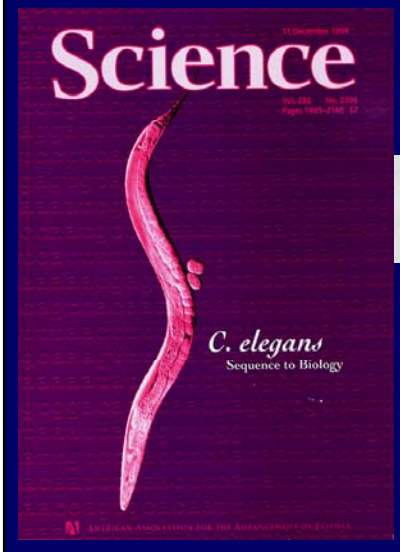

Microbial Genome Sequences
JCVI CMR
Comprehensive Microbial Resource
Search Locus for Go
Home Genome Tools Searches Comparative Tools Lists Downloads Carts
CMR Manual | CMR Tutorial | Links | CMR Resources | CMR Feedback
Genome Search
Organism name:
Search
Genome List
View All CMR Genomes
Gene Search
Search by:
Locus
Match:
Exact Inexact
Keywords/Accession:
Search
Data Summary
Complete Draft Totals
Bacteria 353 17 370
Archaea 28 0 28
Viruses 3 0 3
Totals 384 17 401
Welcome to the Comprehensive Microbial Resource
The Comprehensive Microbial Resource (CMR) is a free website used to display information on all of the publicly available, complete prokaryotic genomes. In addition to the convenience of having all of the organisms on a single website, common data types across all genomes in the CMR make searches more meaningful, and cross genome analysis highlight differences and similarities between the genomes. A CMR Mirror site maintained by the Genome Encyclopedia of Microbes (GEM) in Korea is also available. [More Information] [Publication Information]
CMR Menu Bar Tools
CMR offers a wide variety of tools and resources, all of which are available off of our menu bar at the top of each page. Below is an explanation and link for each of these menu options. First time users can use our CMR tutorial to learn how to navigate this site.
Genome Tools
Find organism lists as well as summary information and analyses for selected genomes.
Searches
Search CMR for genes, genomes, sequence regions, and evidence.
Announcements
camera
March 13, 2007: CAMERA is a web resource for metagenomic research. CAMERA's debut coincides with the publication of the Global Ocean Sampling expedition's extensive dataset cataloging over 6 million new genes from uncultured marine microbes. Come visit CAMERA, and see our growing collection of metagenomics datasets and tools.
Latest Releases
Data Release: 21.0
www.tigr.org

First Eukaryotic Genome Sequence



Goffeau et al. (1997)

First Animal Genome Sequence

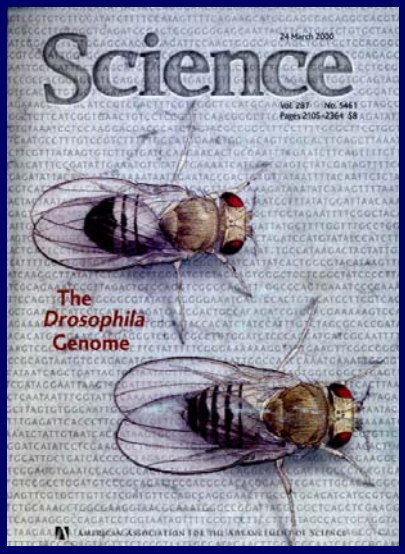


Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

The *C. elegans* Sequencing Consortium*

C. elegans Sequencing Consortium (1998)

Second Animal Genome Sequence



THE *DROSOPHILA* GENOME

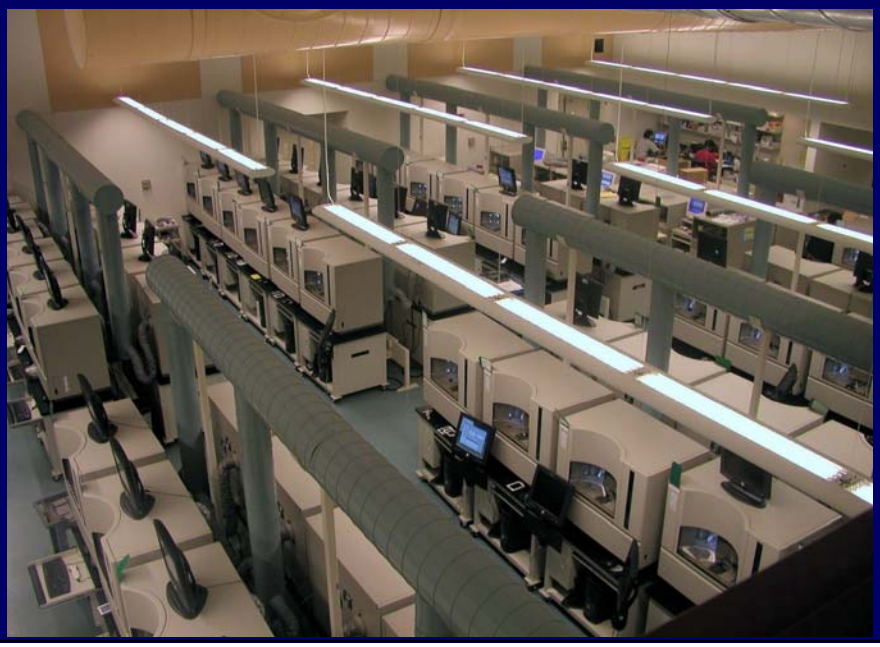
REVIEW

The Genome Sequence of *Drosophila melanogaster*

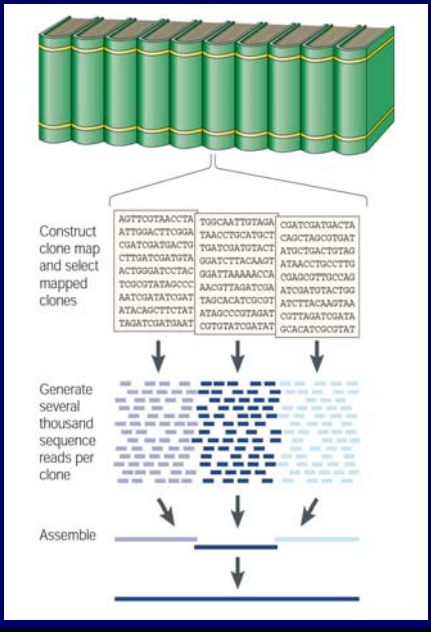
Mark D. Adams,[1*] Susan E. Celniker,[2] Robert A. Holt,[1] Cheryl A. Evans,[1] Jeannine D. Gocayne,[1] Peter G. Amanatides,[1] Steven E. Scherer,[3] Peter W. Li,[1] Roger A. Hoskins,[2] Richard F. Galle,[2] Reed A. George,[2] Suzanna E. Lewis,[4] Stephen Richards,[2] Michael Ashburner,[5] Scott N. Henderson,[1] Granger G. Sutton,[1] Jennifer R. Wortman,[1] Mark D. Yandell,[1] Qing Zhang,[1] Lin X. Chen,[1] Rhonda C. Brandon,[1] Yu-Hui C. Rogers,[1] Robert G. Blazej,[2] Mark Champe,[2] Barret D. Pfeiffer,[2] Kenneth H. Wan,[2] Clare Doyle,[2] Evan G. Baxter,[2] Gregg Helt,[6] Catherine R. Nelson,[6] George L. Gabor Miklos,[7] Josep F. Abril,[8] Anna Agbayani,[2] Hui-Jin An,[1] Cynthia Andrews-Pfannkoch,[1] Danita Baldwin,[1] Richard M. Ballew,[1] Anand Basu,[1] James Baxendale,[1] Leyla Bayraktaroglu,[9] Ellen M. Beasley,[1] Karen Y. Beeson,[1] P. V. Benos,[10] Benjamin P. Berman,[2] Deepali Bhandari,[1] Slava Bolshakov,[11] Dana Borkova,[12] Michael R. Botchan,[13] John Bouck,[2] Peter Brokstein,[6] Phillipe Brottier,[14] Kenneth C. Burtis,[15] Dana A. Busam,[1] Heather Butler,[16] Edouard Cadieu,[17] Angela Center,[1] Ishwar Chandra,[1] J. Michael Cherry,[18] Simon Cawley,[19] Carl Dahlke,[1] Lionel B. Davenport,[1] Peter Davies,[1] Beatriz de Pablos,[20] Arthur Delcher,[1] Zuoming Deng,[1] Anne Deslattes Mays,[1] Ian Dew,[1] Suzanne M. Dietz,[1] Kristina Dodson,[1] Lisa E. Doup,[1] Michael Downes,[21] Shannon Dugan-Rocha,[3] Boris C. Dunkov,[22] Patrick Dunn,[1] Kenneth J. Durbin,[3] Carlos C. Evangelista,[1] Concepcion Ferraz,[23] Steven Ferriera,[1] Wolfgang Fleischmann,[5] Carl Fosler,[1] Andrei E. Gabrielian,[1] Neha S. Garg,[1] William M. Gelbart,[4] Ken Glasser,[1] Anna Glodek,[1] Fangcheng Gong,[1] J. Harley Gorrell,[9] Zhiping Gu,[1] Ping Guan,[1] Michael Harris,[1] Nomi L. Harris,[2] Damon Harvey,[4] Thomas J. Heiman,[1] Judith R. Hernandez,[3] Jarrett Houck,[1] Damon Hostin,[1] Kathryn A. Houston,[2] Timothy J. Howland,[1] Ming-Hui Wei,[1] Chinyere Ibegwam,[1] Mena Jalali,[1] Francis Kalush,[1] Gary H. Karpen,[21] Zhaoxi Ke,[1] James A. Kennison,[24] Karen A. Ketchum,[1] Bruce E. Kimmel,[2] Chinnappa D. Kodira,[1] Cheryl Kraft,[1] Saul Kravitz,[1] David Kulp,[6] Zhongwu Lai,[1] Paul Lasko,[25] Yiding Lei,[1] Alexander A. Levitsky,[1] Jiayin Li,[1] Zhenya Li,[1] Yong Liang,[1] Xiaoying Lin,[26] Xiangjun Liu,[1] Bettina Mattei,[1] Tina C. McIntosh,[1] Michael P. McLeod,[3] Duncan McPherson,[1] Gennady Merkulov,[1] Natalia V. Milshina,[1] Clark Mobarry,[1] Joe Morris,[6] Ali Moshrefi,[2] Stephen M. Mount,[27] Mee Moy,[1] Brian Murphy,[1] Lee Murphy,[28] Donna M. Muzny,[3] David L. Nelson,[3] David R. Nelson,[29] Keith A. Nelson,[1] Katherine Nixon,[2] Deborah R. Nusskern,[1] Joanne M. Pacleb,[2] Michael Palazzolo,[2] Gjange S. Pittman,[1] Sue Pan,[1] John Pollard,[1] Vinita Puri,[1] Martin G. Reese,[6] Knut Reinert,[1] Karin Remington,[1] Robert D. C. Saunders,[30] Frederick Scheeler,[1] Hua Shen,[2] Bixiang Christopher Shue,[1] Inga Silén-Kiamos,[14] Michael Simpson,[1] Marian P. Skupski,[1] Tom Smith,[1] Eugene Spier,[1] Allan C. Spradling,[31] Mark Stapleton,[2] Renee Strong,[2] Eric Sun,[1] Robert Svirskas,[32] Cyndee Tector,[1] Russell Turner,[1] Eli Venter,[1] Aihui H. Wang,[1] Xin Wang,[1] Zhen-Yuan Wang,[1] David A. Wassarman,[33] George M. Weinstock,[3] Jean Weissenbach,[14] Sherita M. Williams,[1] Trevor Woodage,[1] Kim C. Worley,[3] David Wu,[1] Song Yang,[2] Q. Alison Yao,[1] Jane Ye,[1] Ru-Fang Yeh,[10] Jayshree S. Zaveri,[1] Ming Zhan,[1] Guangren Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Xianggun H. Zheng,[1] Fei N. Zhong,[1] Wenyan Zhong,[1] Xiaojun Zhou,[6] Shiaoping Zhu,[1] Xiaohong Zhu,[1] Hamilton O. Smith,[1] Richard A. Gibbs,[3] Eugene W. Myers,[1] Gerald M. Rubin,[34] J. Craig Venter[1]

Adams et al. (2000)

Human Genome Sequencing Centers

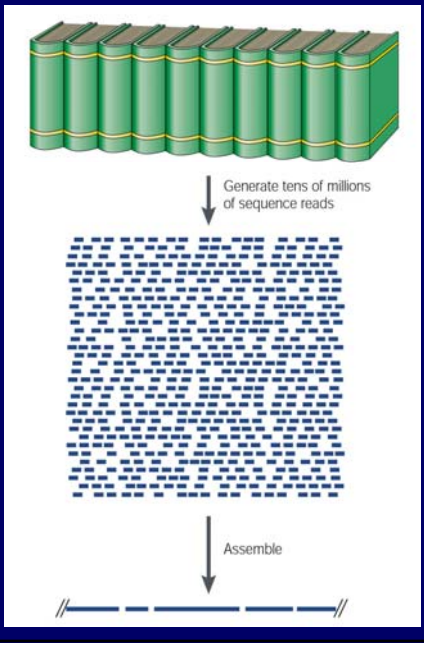


Clone-Based Shotgun Sequencing



Green (2001)

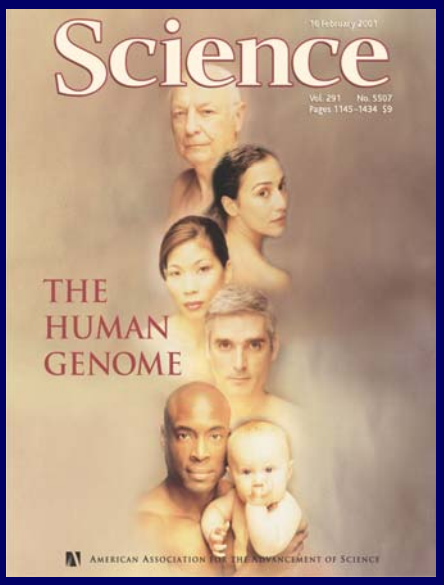
Whole-Genome Shotgun Sequencing



Green (2001)

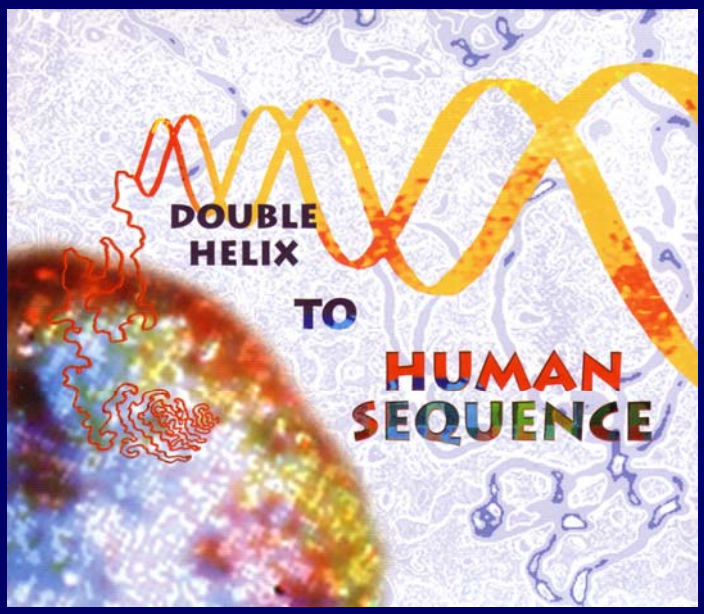
February, 2001 Draft Sequence

International Human Genome Sequencing Consortium (2001)

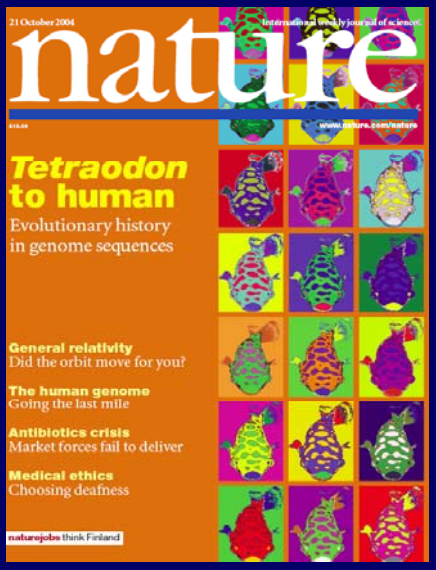


Venter et al. (2001)

April, 2003 Completion



October, 2004 Publication




articles

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium (2004)

CNN's #1 Medical Story of Past 25 Years

CNN.com

PRINT THIS

Powered by Clickability

Click to Print

SAVE THIS | EMAIL THIS | Close

Top 25: Medical stories

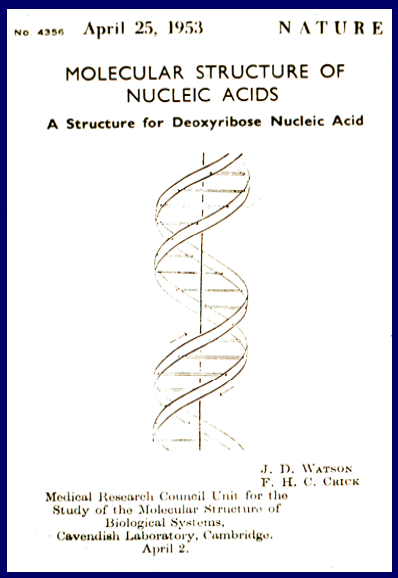
Human genome mapping ranks No. 1 in health news

Tuesday, March 29, 2005 Posted: 4:24 PM EST (2124 GMT)

(CNN) -- Much of the marvel of medicine has to do with discovery. Mapping the human genome, the complete sequence of DNA, gave scientists a blueprint for building a person, making it the No. 1 medical story, according to a distinguished panel CNN gathered to rank the top 25 medical stories of the past quarter-century.

Two men from two separate groups -- Francis Collins of the National Institutes of Health and Craig Venter of Celera Genomics Inc., a pharmaceutical-development company -- worked independently to discover the sequence of the human genome and identify the genes that it contains. This

April, 1953 → April, 2003



No. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

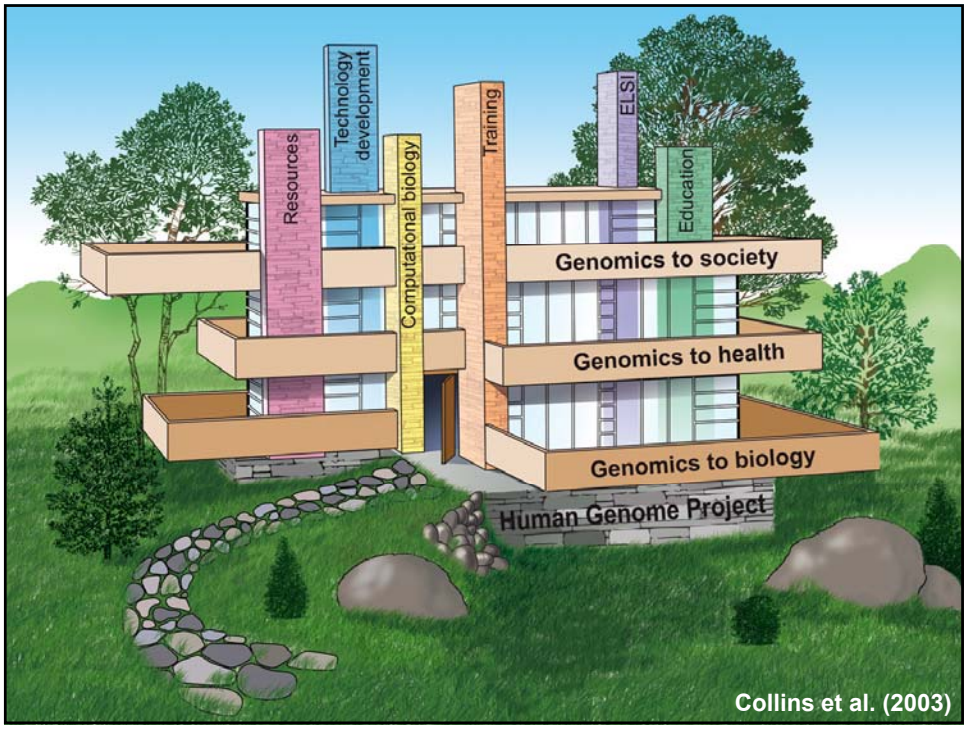
J. D. Watson
F. H. C. Crick

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge. April 2.

All of the original goals of the Human Genome Project have been accomplished!

What's Next?

Collins et al. (2003)



Mapping the Human Genome — ~1990 to ~2000

Sequencing the Human Genome — ~1998 to ~2003

The Human Genome Project

Interpreting the Human Genome Sequence — ~2003 to ???

Beyond The Human Genome Project

~3,000 bp (0.0001%) of Human Genome Sequence

TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAG
GAACCCGACTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGA
CGCGAAGGAGGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCCACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTC
AGAATCGGGAAAGGGAGGTGCGGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGA
GAAAGCCGCTAGAGCAAATTTGGGGCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAGGCGGGGGAAAAGAGCAAAAGGAAGGGGTGG
TGTGCGGAGTAGGGGTGGGTGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAG
GTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGT
TGTCTCCAAACTTTTTTTCAGGTGAGAAGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCCACGAAAGAGGAGGGCGTGTGTATGGGTTGGGTT
TGGGGTAAAGGAATAAGCAGTTTTTAAAAAGATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTA
AGAAGAGATGGAAGAATGAACTGAAGCTGATTGAATAGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTTGTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGG
ACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATT
TTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTATTTGTCACAGGTAAATGATTCTTCAAAAATTG
AAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGGAATAGATACCGAAGTTATATCCAA
CTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAATAGCATCTAAGTTCGG
AATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGT
TATAAATAGCTCATATATGTGTGTAGGGGGGAAGGAATTTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTTGTGT
TGATGTTATCCACCTTTTGTTACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCAT
AAATTATAGACTTTTTTTAGCAGAGAACTTTGAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTCATGGATA
GTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTGATCTGTCAAAGTGGGTTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTT
AATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTATTACAAAGGCTGCTAAGTGTGTTAAGAGCATACTCATTTCTGTTCTCC
AAAATATTTCATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAAGTTAAGTTCCTACTATTTATAGGAACTGACAATCACCTAAAATACCAATGA
TTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATTAAGTTAGGCAGTATTGCTTAGTTTTCAAA
GTGGTAGGCTTTGGAGTCAGATTATTTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTTAACATCTCTGTG
TGTGACTTGACCTTTAAAATTTGGAGACTGTCATAGGGGTTAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATTATT
ATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAAT
CTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATG
TTTCTATTGTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT

The Human Genome… by the Numbers

~5% of Human Genome Sequence is Constrained Across Mammals (and Presumed Functional)

- 5% of 3B Bases = ~150M Bases
- Do NOT Yet Know the Position of these ~150M Functional Bases
- Lower Bound for the Amount that is Functional

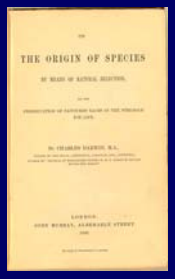
~1.5% Encodes for Protein (Genes)

- Corresponds to ~18-22K Genes
- Many More than ~22K Different Proteins
- Good Inventory at Present

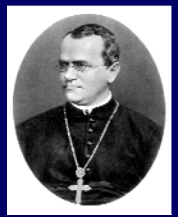
~3.5% Functional But Non-Coding

- Gene Regulatory Elements
- Chromosomal Functional Elements
- Undiscovered Functional Elements (NOT Yet in Textbooks!)
- Poor Inventory at Present

Foundational Milestones in Genetics & Genomics

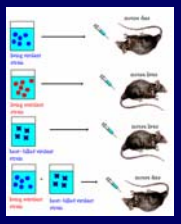


Darwin *1859*

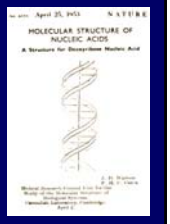


Mendel *1865*

Miescher *1871*

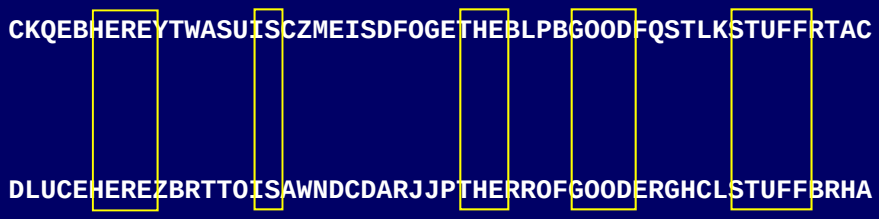


Avery *1944*



Watson & Crick *1953*

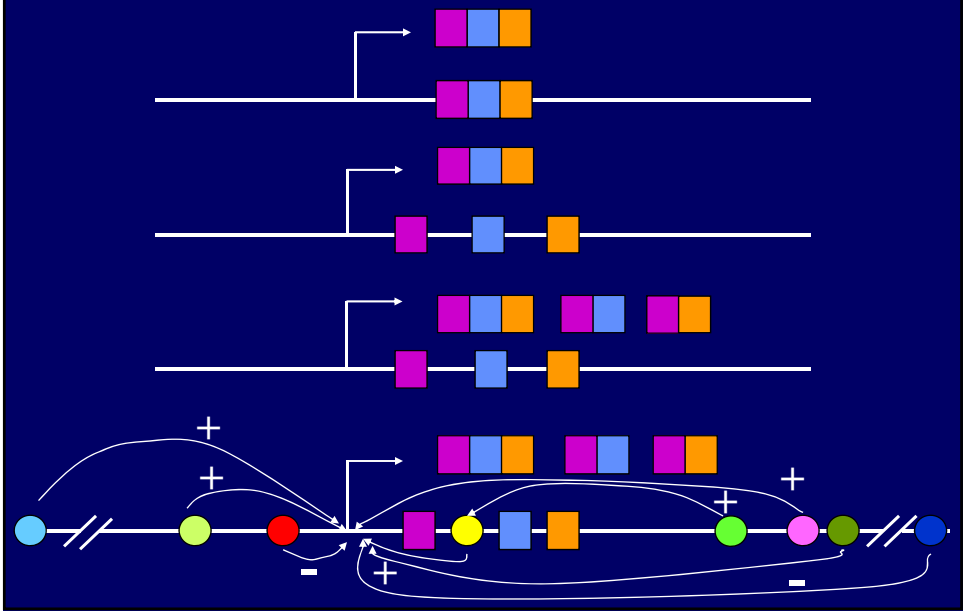
Comparing Genomes is Like Cryptography



Functional Elements: Coding vs. Non-Coding

- Coding Sequences (i.e., Genes)
  - Relatively EASY to Identify
  - Mostly Know What to Look For
  - Complementary Data Sets Available (ESTs, cDNAs)
  - Ever-Improving Computational Gene Predictions
- Non-Coding Functional Sequences
  - HARD to Identify
  - Very Little Known About What to Look For
  - Virtually No Complementary Data Sets Available
  - Poor Computational Predictions

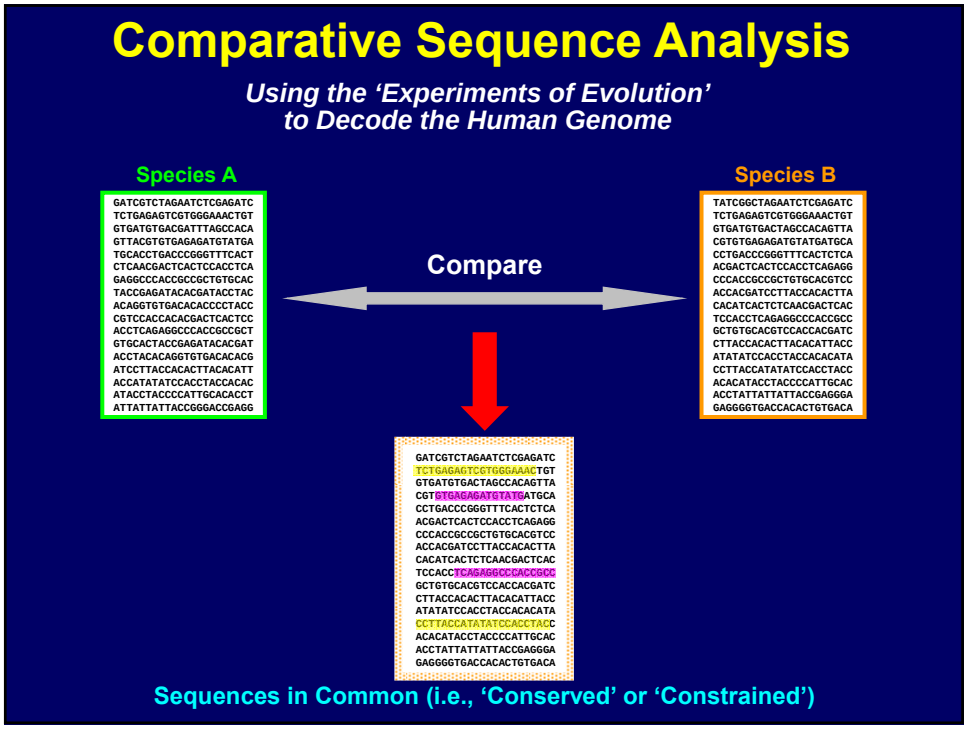
The Language of the Genome

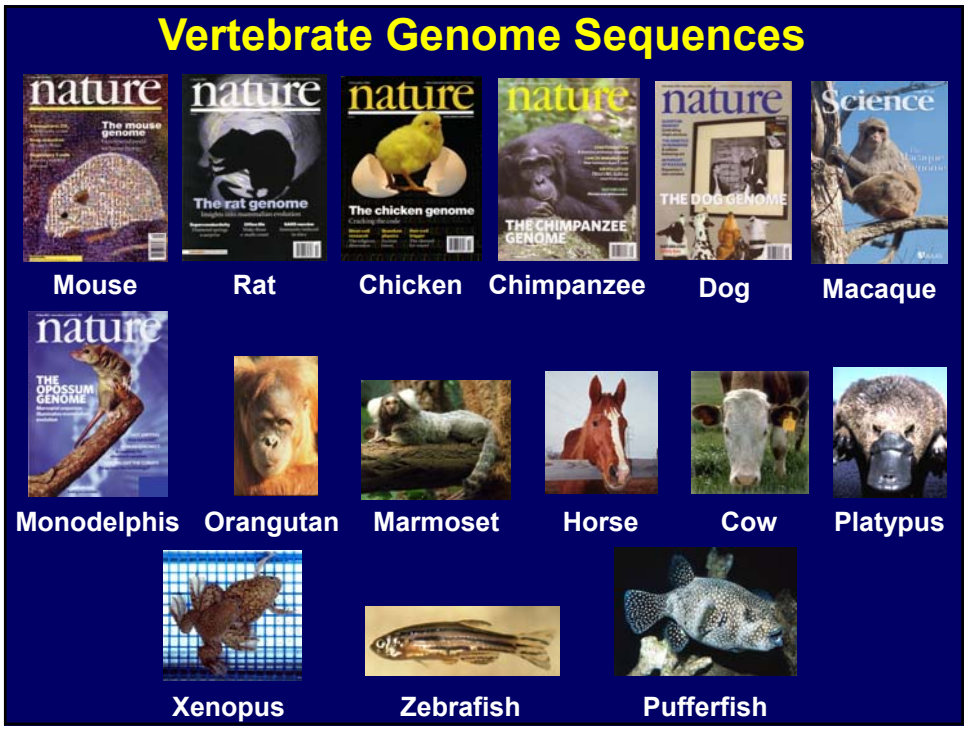


Functional Elements: Coding vs. Non-Coding

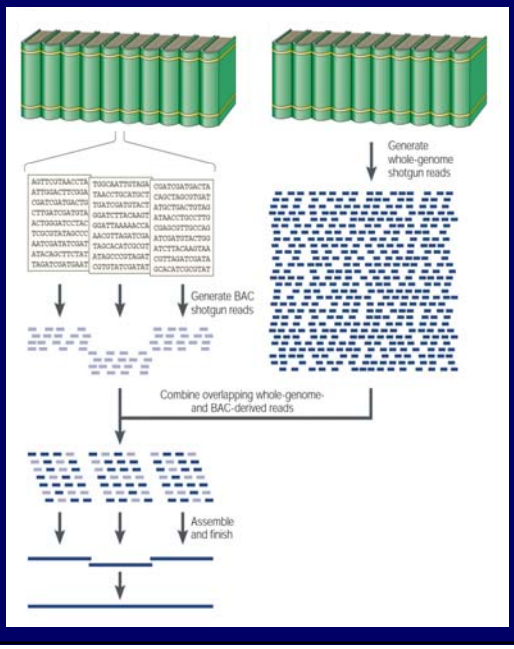
- **Coding Sequences (i.e., Genes)**
  - Relatively EASY to Identify
  - Mostly Know What to Look For
  - Complementary Data Sets Available (ESTs, cDNAs)
  - Ever-Improving Computational Gene Predictions
- **Non-Coding Functional Sequences**
  - HARD to Identify
  - Very Little Known About What to Look For
  - Virtually No Complementary Data Sets Available
  - Poor Computational Predictions

Major role for comparative sequence analysis will be the identification of functionally important, non-coding sequences


Comparative Sequence Analysis
Using the 'Experiments of Evolution' to Decode the Human Genome
Species A
GATCGTCTAGAATCTCGAGATC
TCTGAGAGTCGTGGGAAACTGT
GTGATGTGACGATTTAGCCACA
GTTACGTGTGAGAGATGTATGA
TGCACCTGACCCGGGTTTCACT
CTCAACGACTCACTCCACCTCA
GAGGCCCACCGCCGCTGTGCAC
TACCGAGATACACGATACCTAC
ACAGGTGTGACACACCCCTACC
CGTCCACCACACGACTCACTCC
ACCTCAGAGGCCCACCGCCGCT
GTGCACTACCGAGATACACGAT
ACCTACACAGGTGTGACACACG
ATCCTTACCACACTTACACATT
ACCATATATCCACCTACCACAC
ATACCTACCCCATTGCACACCT
ATTATTATTACCGGGACCGAGG
Compare
Species B
TATCGGCTAGAATCTCGAGATC
TCTGAGAGTCGTGGGAAACTGT
GTGATGTGACTAGCCACAGTTA
CGTGTGAGAGATGTATGATGCA
CCTGACCCGGGTTTCACTCTCA
ACGACTCACTCCACCTCAGAGG
CCCACCGCCGCTGTGCACGTCC
ACCACGATCCTTACCACACTTA
CACATCACTCTCAACGACTCAC
TCCACCTCAGAGGCCCACCGCC
GCTGTGCACGTCCACCACGATC
CTTACCACACTTACACATTACC
ATATATCCACCTACCACACATA
CCTTACCATATATCCACCTACC
ACACATACCTACCCCATTGCAC
ACCTATTATTATTACCGAGGGA
GAGGGGTGACCACACTGTGACA
GATCGTCTAGAATCTCGAGATC
TCTGAGAGTCGTGGGAAACTGT
GTGATGTGACTAGCCACAGTTA
CGTGTGAGAGATGTATGATGCA
CCTGACCCGGGTTTCACTCTCA
ACGACTCACTCCACCTCAGAGG
CCCACCGCCGCTGTGCACGTCC
ACCACGATCCTTACCACACTTA
CACATCACTCTCAACGACTCAC
TCCACCTCAGAGGCCCACCGCC
GCTGTGCACGTCCACCACGATC
CTTACCACACTTACACATTACC
ATATATCCACCTACCACACATA
CCTTACCATATATCCACCTACC
ACACATACCTACCCCATTGCAC
ACCTATTATTATTACCGAGGGA
GAGGGGTGACCACACTGTGACA
Sequences in Common (i.e., 'Conserved' or 'Constrained')


Vertebrate Genome Sequences
nature
The mouse genome
Mouse
nature
The rat genome
Rat
nature
The chicken genome
Chicken
nature
THE CHIMPANZEE GENOME
Chimpanzee
nature
THE DOG GENOME
Dog
Science
Macaque
nature
THE OPOSSUM GENOME
Monodelphis
Orangutan
Marmoset
Horse
Cow
Platypus
Xenopus
Zebrafish
Pufferfish

Hybrid Shotgun Sequencing



Green (2001)

Diverse Landscape of Genome Sequencing

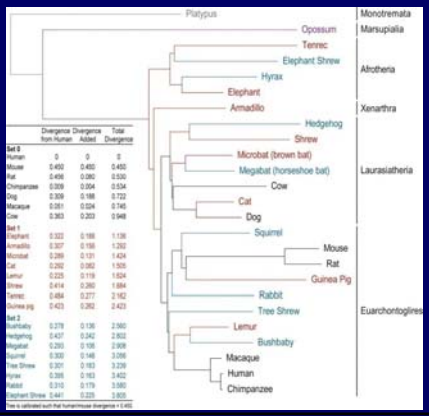
Human
Mouse
Rat
Pufferfish
Zebrafish
Chicken
Chimpanzee
Dog
Cow
Xenopus
Monodelphis
Macaque
Platypus
Marmoset
etc.…

Low-Redundancy, Whole-Genome Shotgun Sequencing

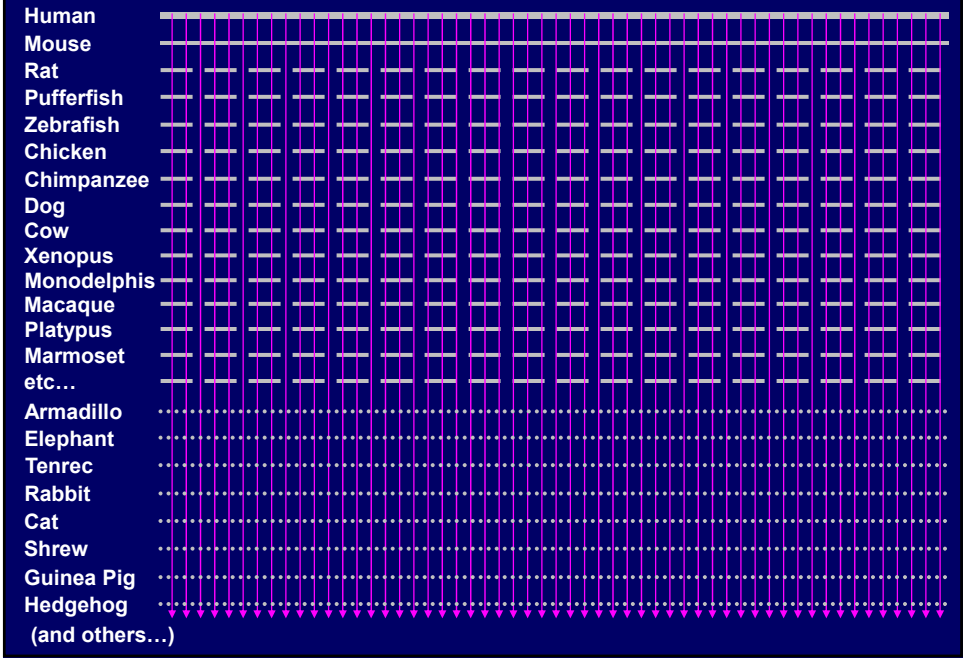
An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing

Elliott H. Margulies*[†], Jade Vinson[†‡], NISC Comparative Sequencing Program*[¶§], Webb Miller[‖], David B. Jaffe[¶], Kerstin Lindblad-Toh[¶], Jean Chang[¶], Eric D. Green*[§], Eric S. Lander[¶], James C. Mullikin*[†**], and Michele Clamp[†**]

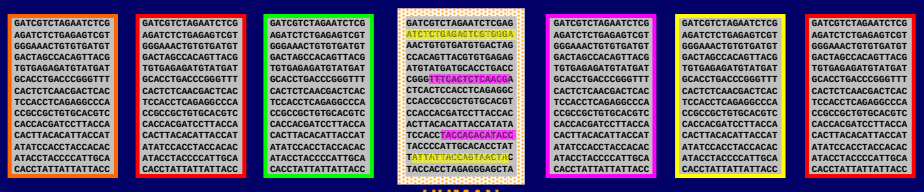
Margulies et al. (2005)



Diverse Landscape of Genome Sequencing



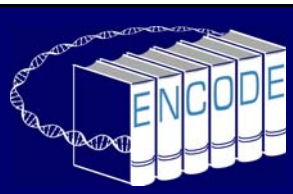
Multi-Species Sequence Comparisons



Multi-Species Conserved Sequences (MCSs)

Margulies et al. (2003)
Thomas et al. (2003)

ENCODE Project



- ENCODE: ENCyclopedia Of DNA Elements
- Goal: Compile a *Comprehensive Encyclopedia* of All Functional Elements in the Human Genome
- Initial Pilot Project: 1% of Human Genome
- Apply Multiple, Diverse Approaches to Study and Analyze that 1% in a Consortium Fashion

SPECIAL SECTION

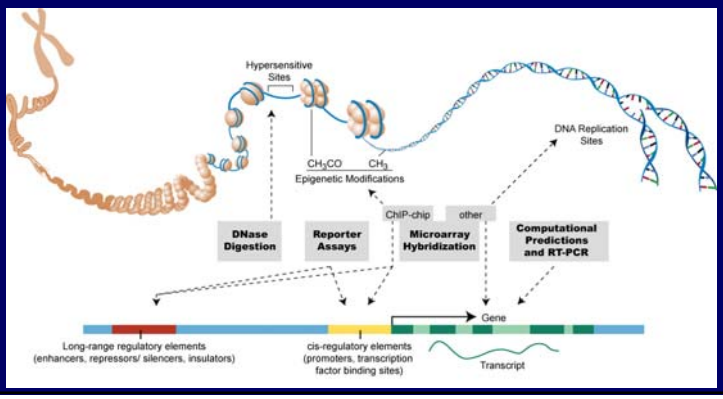
GENES IN ACTION

VIEWPOINT

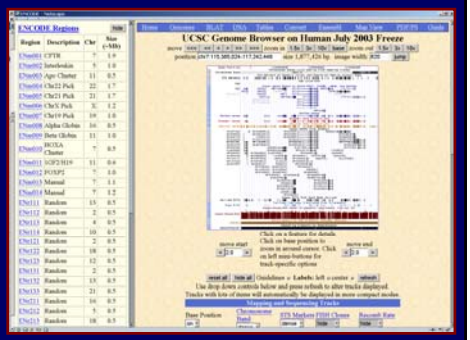
The ENCODE (ENCyclopedia Of DNA Elements) Project

The ENCODE Project Consortium*†

ENCODE Project Consortium (2004)

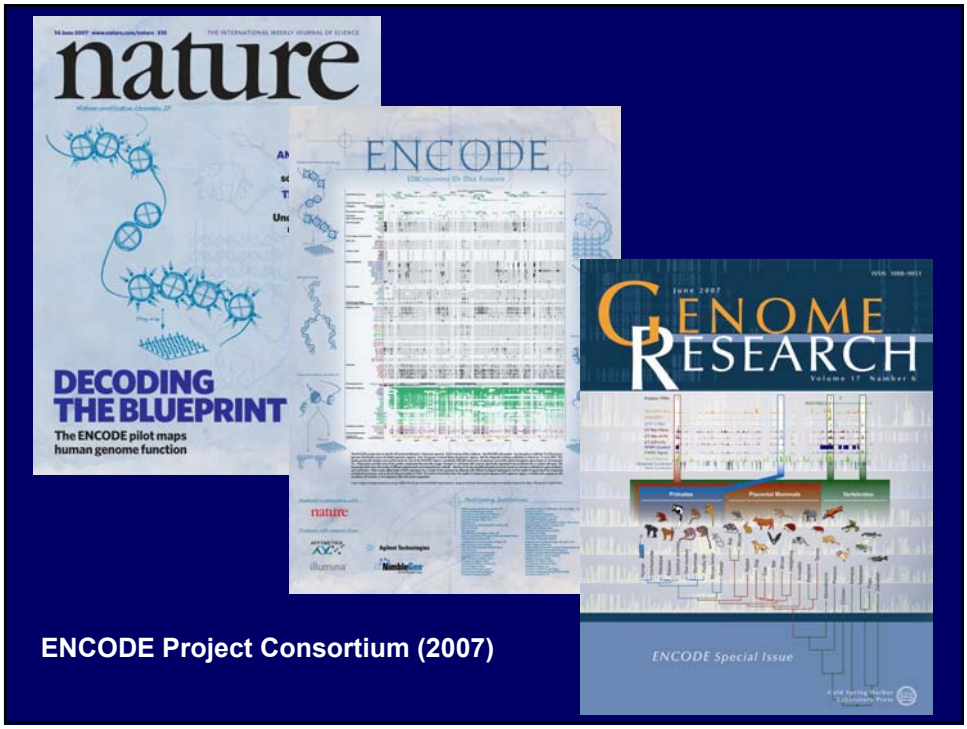


ENCODE Project: Web Sites



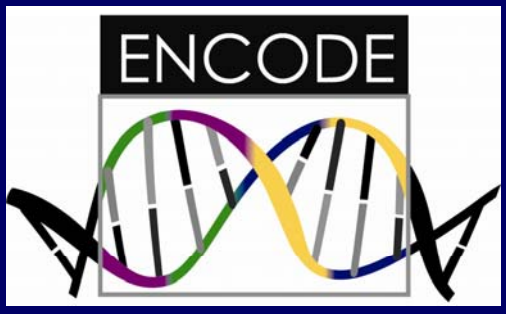
genome.gov/ENCODE

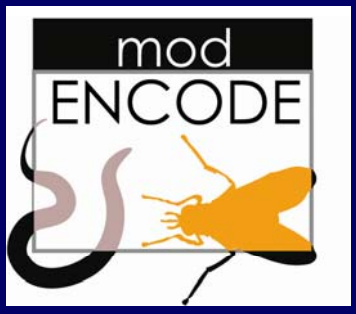
genome.ucsc.edu/ENCODE

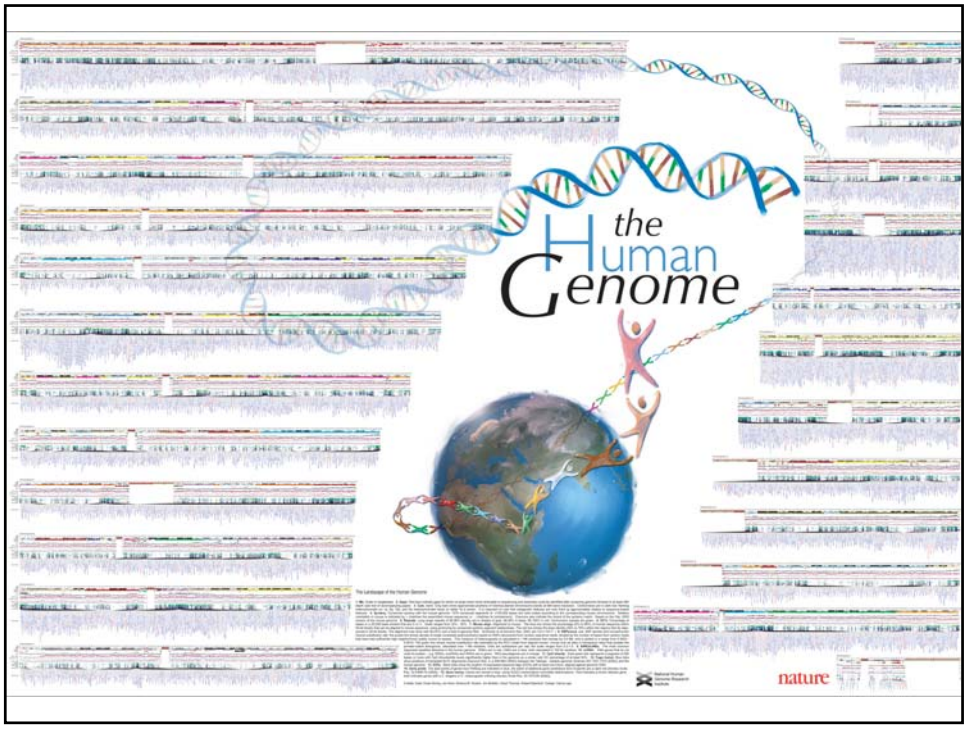


ENCODE Project Consortium (2007)

Expanding ENCODE Portfolio

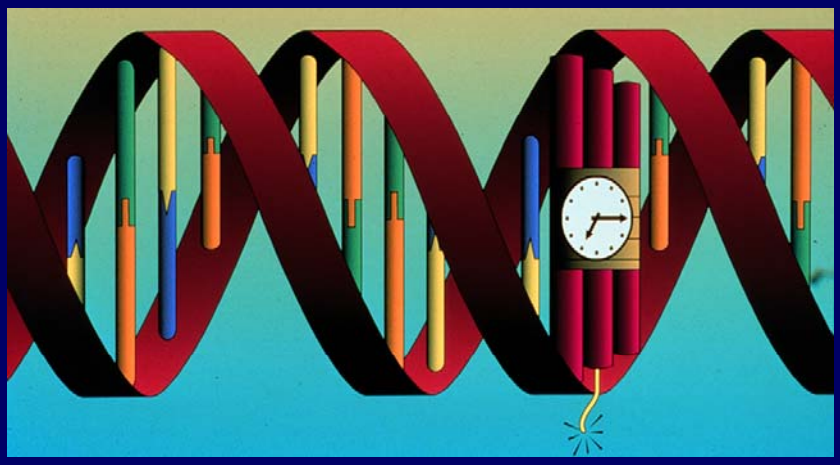




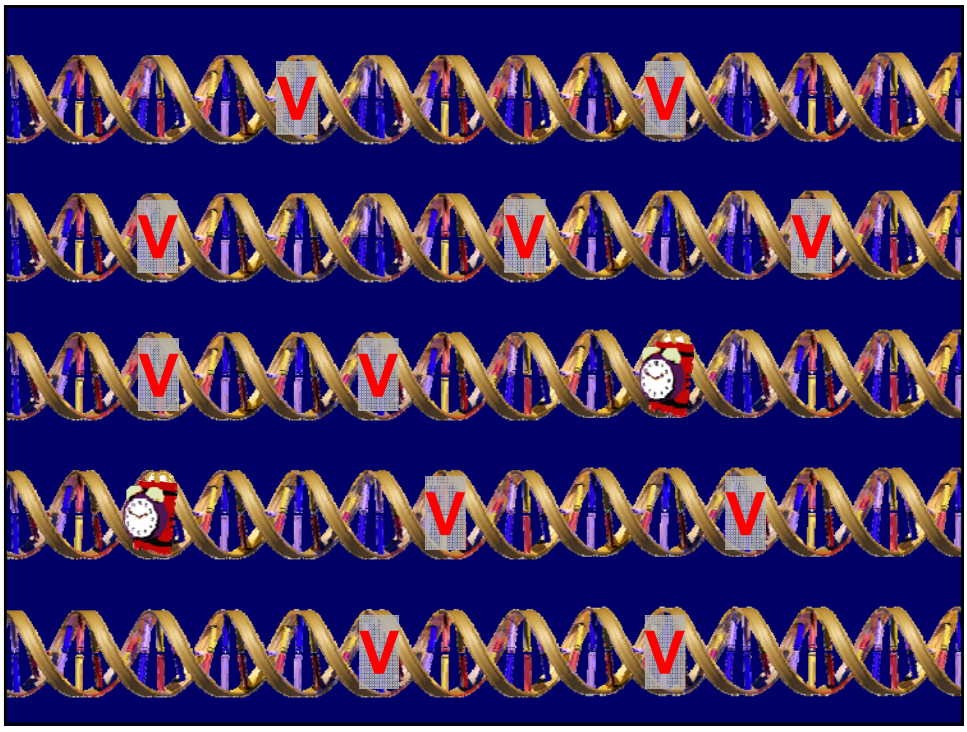

the
Human
Genome
nature


Inter-Species Sequence Comparisons
Intra-Species Sequence Comparisons

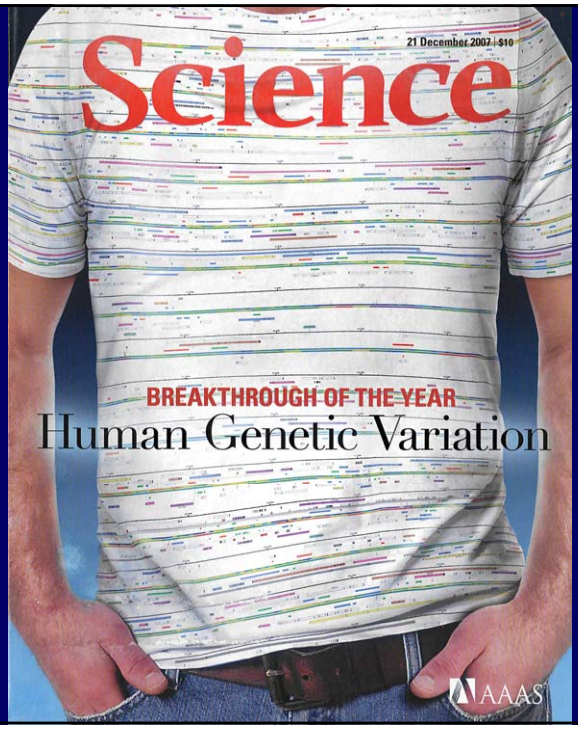
All humans are ~99.9% identical at the DNA sequence level, and yet…

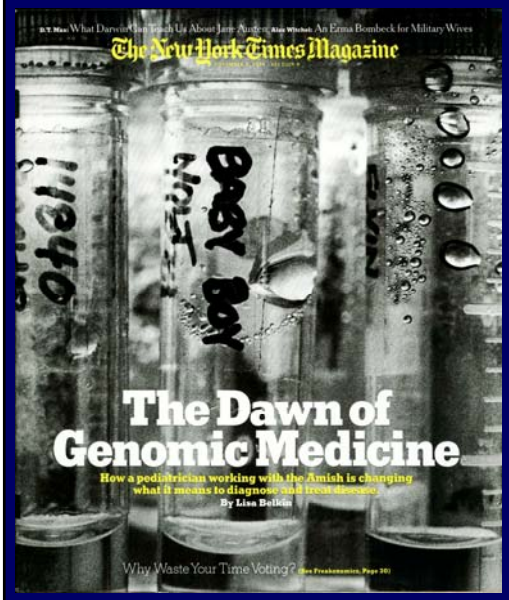


all of us carry a significant number of 'glitches' in our genomes.



International HapMap Consortium (2005, 2007)





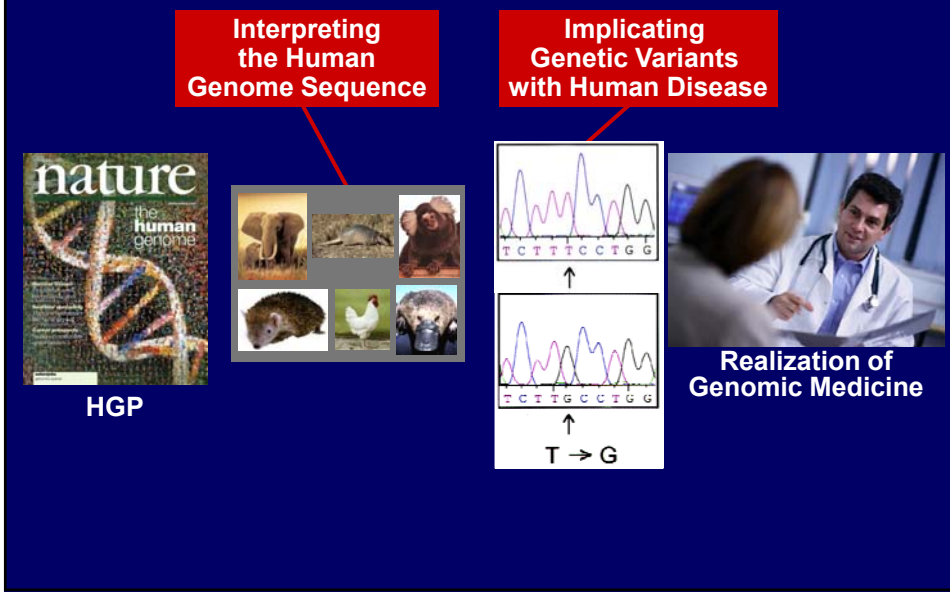
The Pathway to Genomic Medicine

HGP

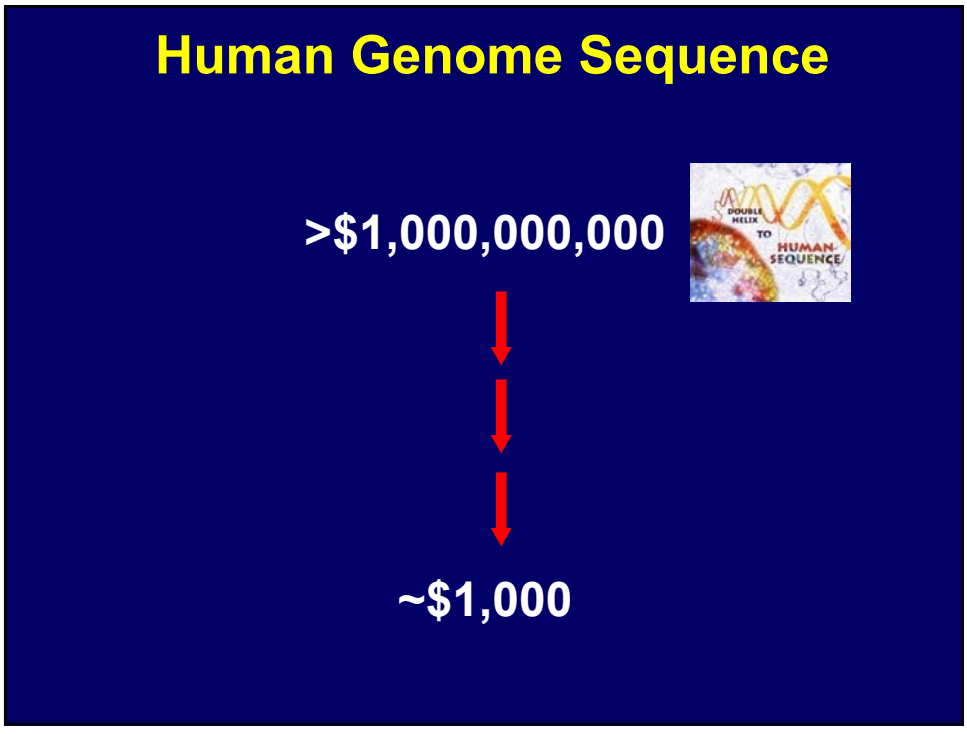
→→ ? →→

Realization of Genomic Medicine

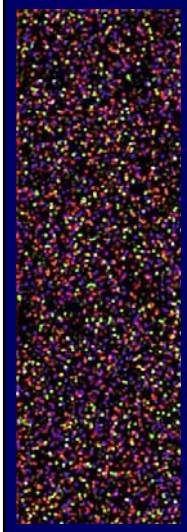
The Pathway to Genomic Medicine

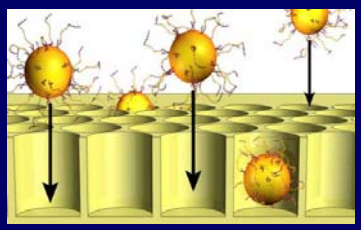


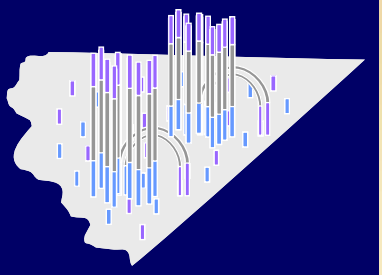
Human Genome Sequence

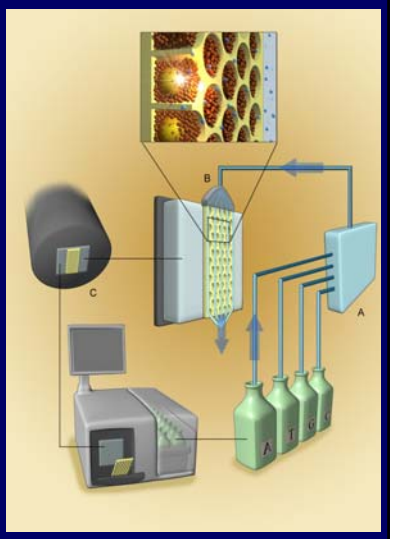


En Route to the $1000 Genome





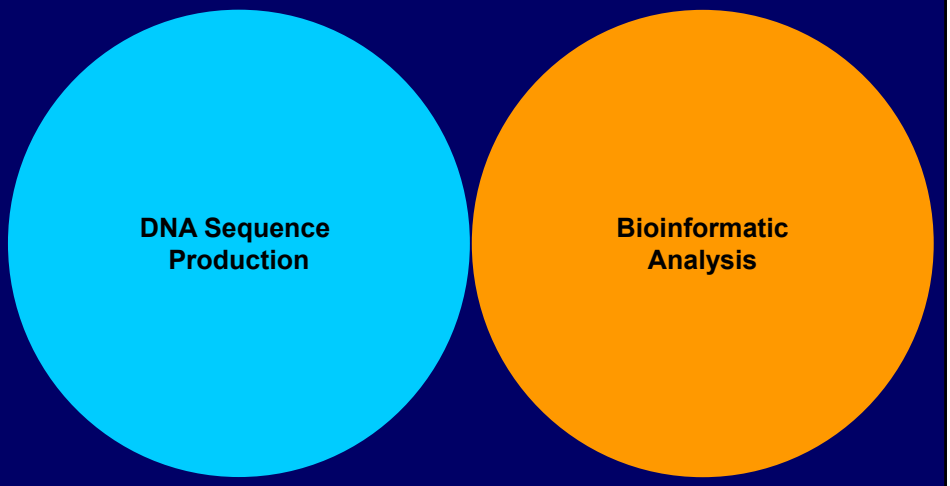




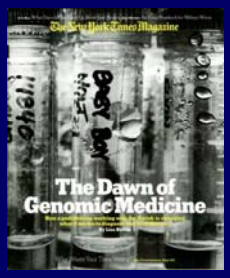
Church (2006)

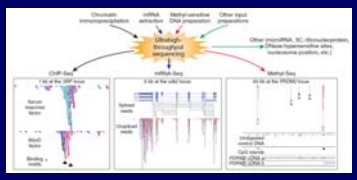
Realities of New DNA Sequencing Technologies…

Changing Infrastructure Requirements



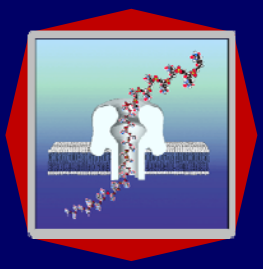
Expanding Universe of Sequence-Based Explorations



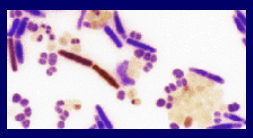


Wold and Myers (2008)

Enard and Paabo (2004)



Collins and Barker (2007)



Turnbaugh et al. (2007)

Green et al. (2006)

The Human Genome Sequence to Genomic Medicine…

…from base pairs to bedside.

Bibliography

Adams MD et al. (2000). The genome sequence of *Drosophila melanogaster*. Science 287:2185-2195.

Birren B et al. (1998). Bacterial artificial chromosomes. In *Genome Analysis: A Laboratory Manual, Vol. 3 Cloning systems* (B Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 241-295.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012-2018.

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69-87.

Church GM (2006). Genomes for all. Sci Am 294:46-54.

Collins FS et al. (2003). A vision for the future of genomics research: a blueprint for the genomic era. Nature 422:835-847.

Collins FS and Barker AD (2007). Mapping the cancer genome: pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. Sci Am 296:50-57.

Enard W and Paabo S (2004). Comparative primate genomics. Annu Rev Genomics Hum Genet 5:351-378.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636-640.

ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799-816.

Gerhard DS et al. (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res 14:2121-2127.

Goffeau A et al. (1997). The Yeast Genome Directory. Nature 387S:1-105.

Gordon D et al. (1998). Consed: a graphical tool for sequence finishing. Genome Res 8:195-202.

Green ED (2001). Strategies for the systematic sequencing of complex genomes. Nature Rev Genet 2:573-583.

Green ED et al. (1998). Yeast artificial chromosomes. In *Genome Analysis: A Laboratory Manual, Vol. 3 Cloning systems* (B Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 297-565.

Green RE et al. (2006). Analysis of one million base pairs of Neanderthal DNA. Nature 444:330-336.

Hillier LW et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695-716.

International HapMap Consortium (2005). A haplotype map of the human genome. Nature 437:1299-1320.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409:860-921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature 431:931-945.

Lindblad-Toh K et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803-819.

Margulies EH et al. (2003). Identification and characterization of multi-species conserved sequences. Genome Res 13:2507-2518.

Margulies EH et al. (2005). An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci 102:4795-4800.

Marra MA et al. (1997). High throughput fingerprint analysis of large-insert clones. Genome Res 7:1072-1084.

Messing J and Llaca V (1998). Importance of anchor genomes for any plant genome project. Proc Natl Acad Sci 95:2017-2020.

Mikkelsen TS et al. (2007). Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. Nature 447:167-177.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

Rat Genome Sequencing Project Consortium (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493-521.

Rhesus Macaque Genome Sequencing and Analysis Consortium (2007). Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222-234.

Thomas JW et al. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788-793.

Turnbaugh PJ et al. (2007). The human microbiome project. Nature 449:804-810.

Venter JC et al. (2001). The sequence of the human genome. Science 291:1304-1351.

Wilson RK and Mardis ER (1997). Fluorescence-based DNA sequencing. In *Genome Analysis: A Laboratory Manual, Vol. 1 Analyzing DNA* (B Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 301-395.

Wilson RK and Mardis ER (1997). Shotgun sequencing. In *Genome Analysis: A Laboratory Manual, Vol. 1 Analyzing DNA* (B Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 397-454.

Wold B and Myers RM (2008). Sequence census methods for functional genomics. Nat Methods 5:19-21.