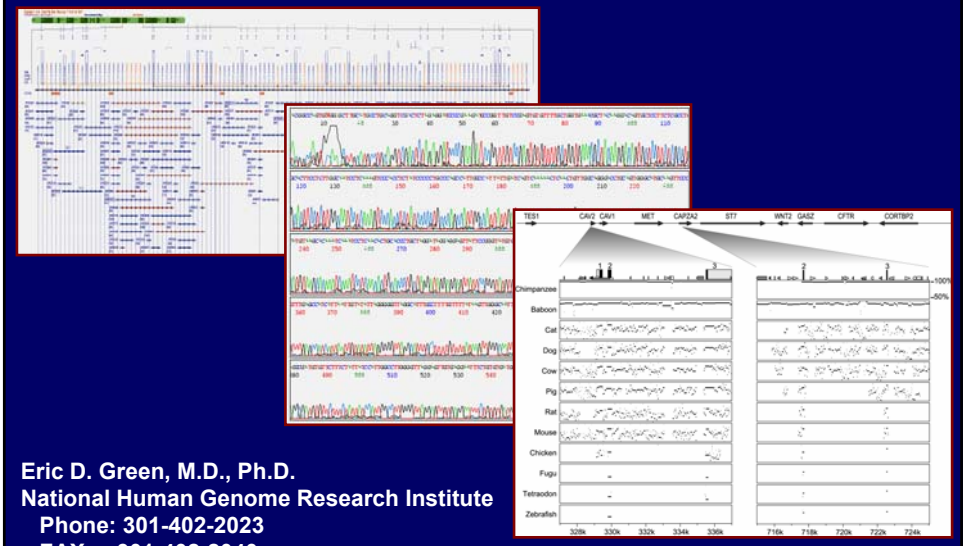
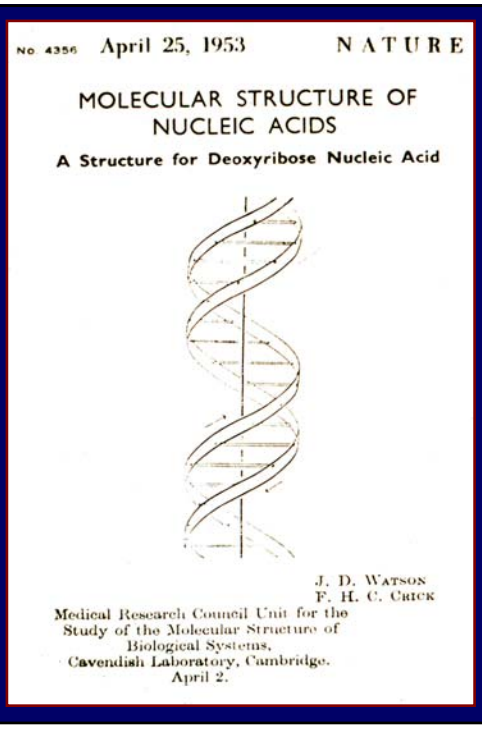
Techniques for Genome Mapping & Sequencing



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



No. 4356 April 25, 1953 NATURE

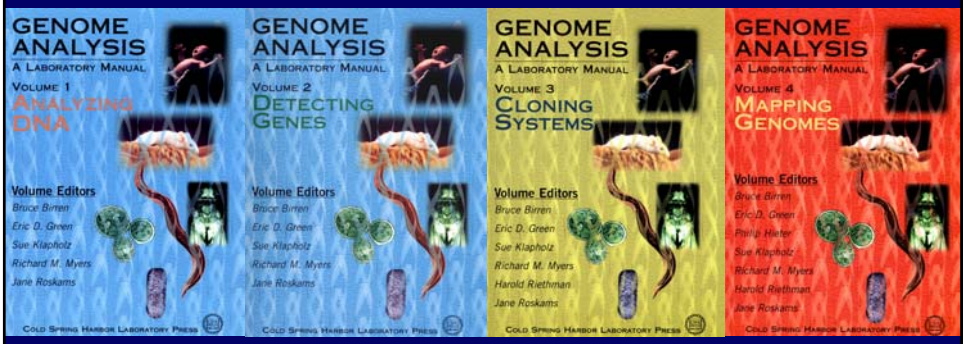
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

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.

Genome Analysis Series: CSHL Press



Outline

I. Fundamentals of Physical Mapping

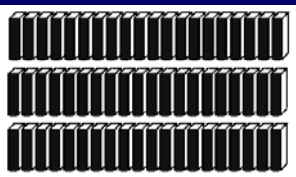
II. Fundamentals of Genome Sequencing

III. Mapping & Sequencing in the Human Genome Project…and Beyond

IV. Future Challenges (i.e., What's Next?)

Genome Sizes

Human Genome
Mouse Genome



~3,000,000,000 bp

Fruit Fly Genome

~160,000,000 bp

Nematode Genome

~100,000,000 bp

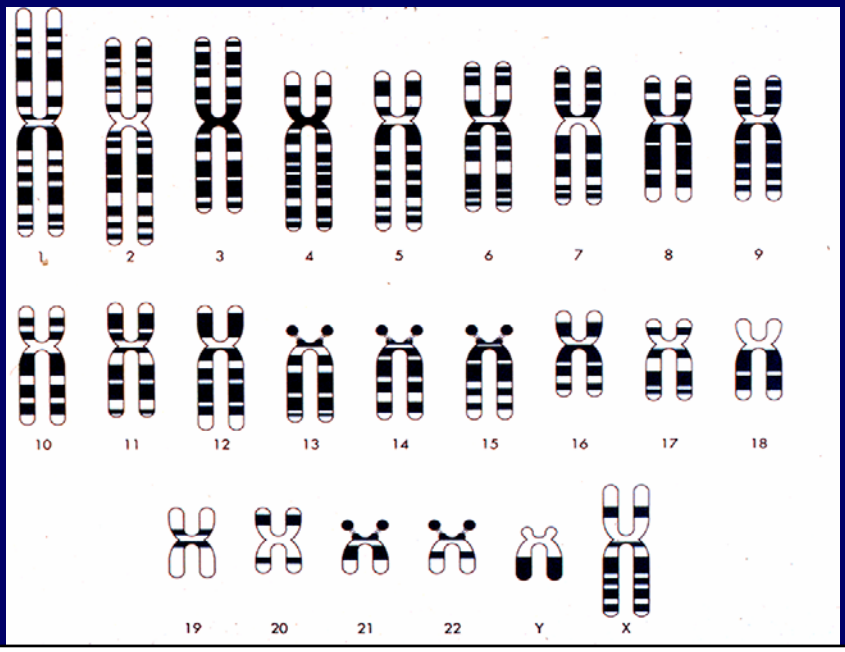
Yeast Genome

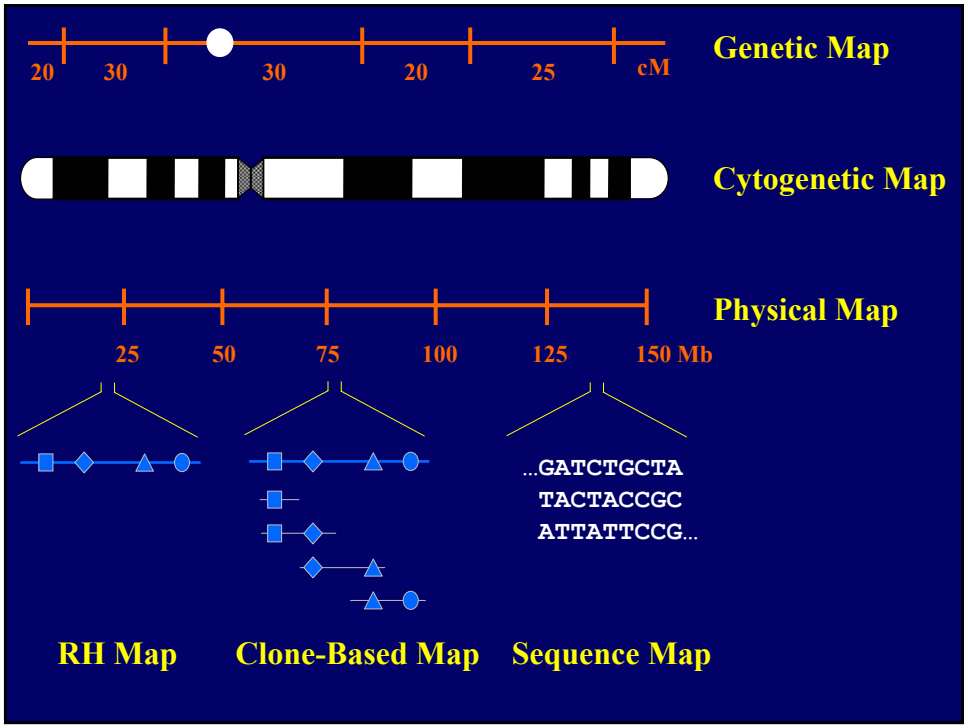
~15,000,000 bp

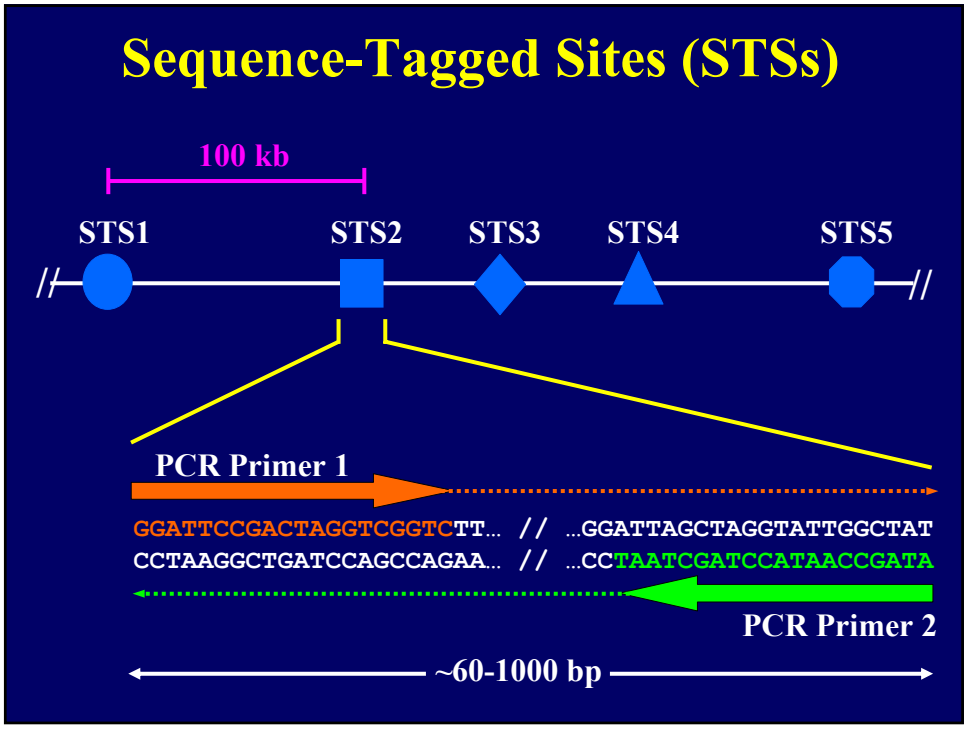
***E. coli* Genome**

~5,000,000 bp

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


Sequence-Tagged Sites (STSs)
100 kb
STS1
STS2
STS3
STS4
STS5
PCR Primer 1
GGATTCCGACTAGGTCGGTCTT... // ...GGATTAGCTAGGTATTGGCTAT
CCTAAGGCTGATCCAGCCAGAA... // ...CCTAATCGATCCATAACCGATA
PCR Primer 2
~60-1000 bp

Physical Mapping: General Principles

- Importance of Physical Maps:

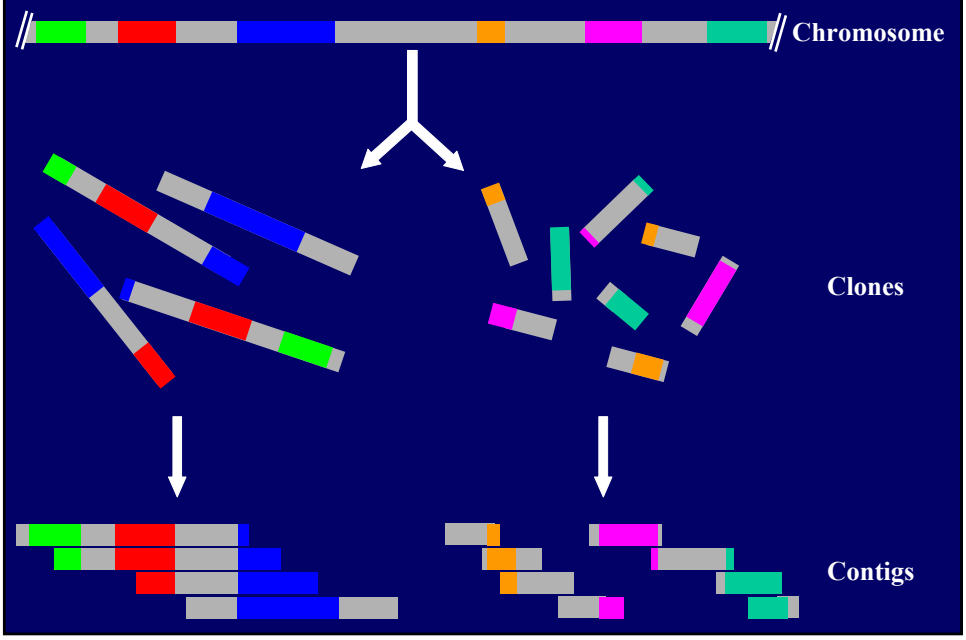
  Localization and Isolation of Genes (e.g., Positional Cloning)
  Study of Genome Organization and Evolution
  Framework for Genome Sequencing

- Physical Mapping Involves Ordering Clones and/or Landmarks

- General Types of Physical Maps:

  Landmark Only (e.g., Radiation Hybrid Maps)
  Clone-Based
  Sequence

Clone-Based Physical Mapping



Clones for Physical Mapping: General Points

- Want Cloned DNA to Accurately Reflect the Source Genome

  Problem of Instability
  Problem of Chimerism

- Development of 'Array Mentality' for Clone Libraries

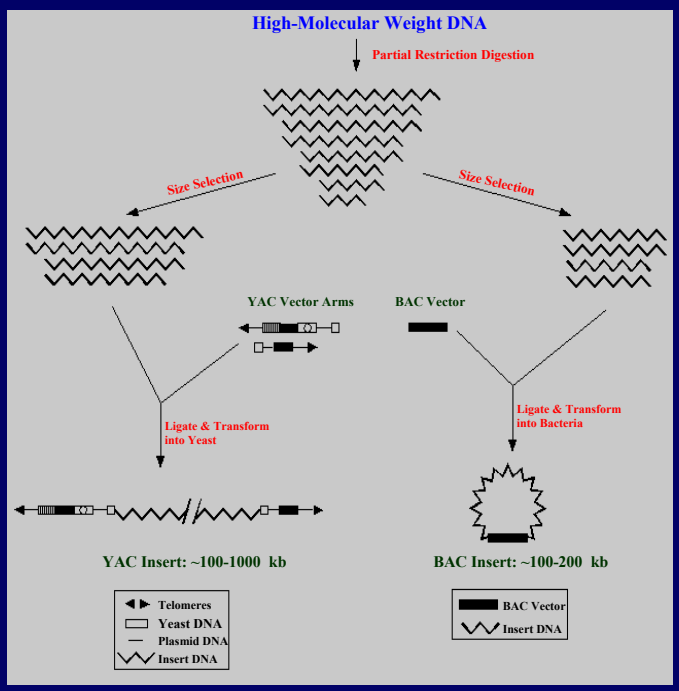
  Clones Arrayed in Individual Wells of Microtiter Plates
  Various Densities (e.g., 96-and 384-Well Plates)

- Advantages of Arrayed Libraries ('Reference Libraries')

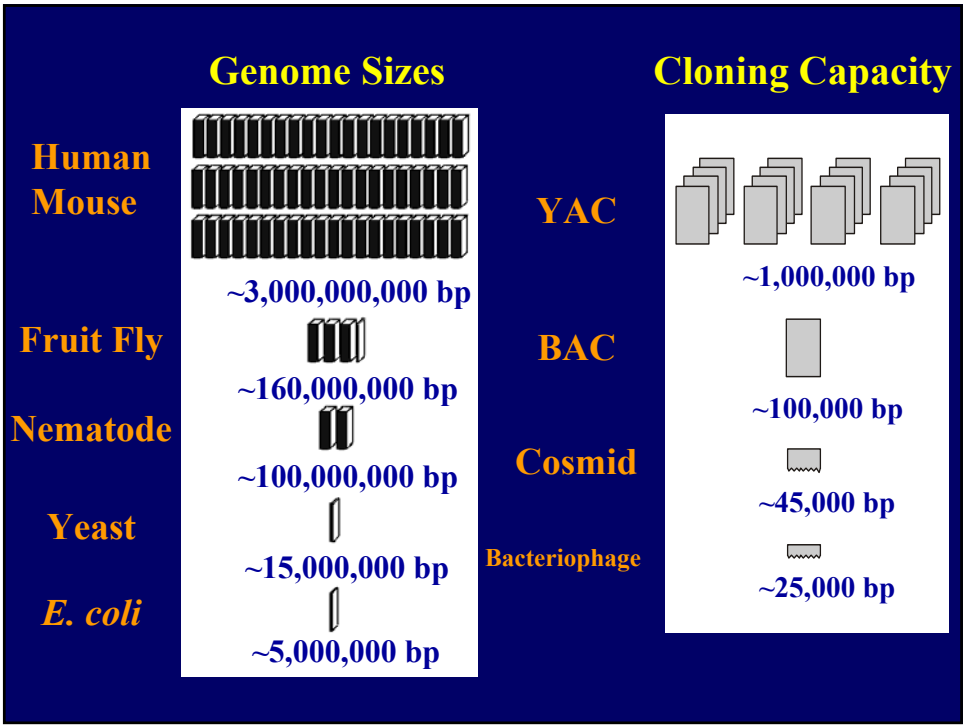
  1. Simplicity of Storing and Transferring Clone Collections
  2. Convenient Format for Retrieving Clones of Interest
  3. Ability to Assimilate Data on Common Clones
  4. Repeated PCR-Based Screening
  5. Repeated Hybridization-Based Screening

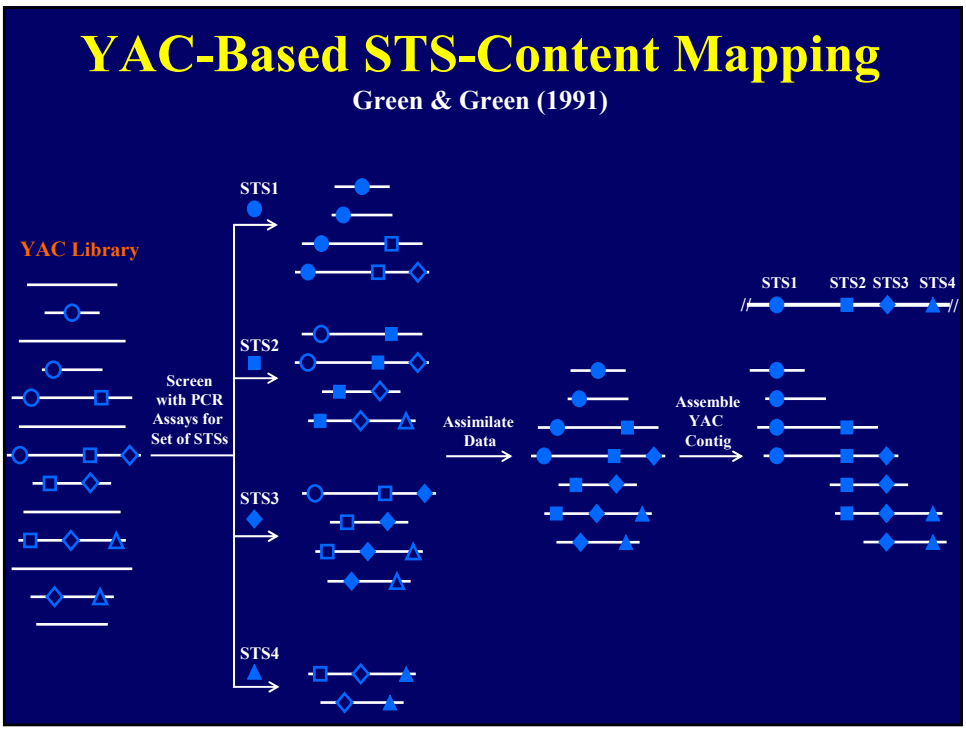
- Trade-Offs with Large vs. Small Inserts

Construction of YACs and BACs



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


YAC-Based STS-Content Mapping
Green & Green (1991)
YAC Library
Screen with PCR Assays for Set of STSs
STS1
STS2
STS3
STS4
Assimilate Data
Assemble YAC Contig
STS1 STS2 STS3 STS4

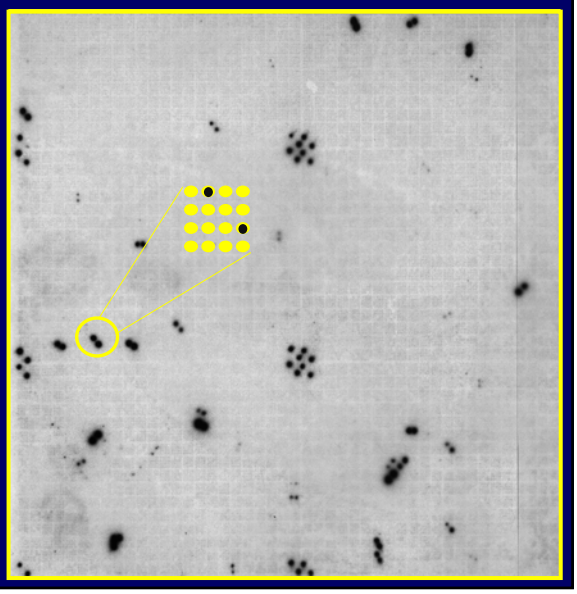
Bacterial Artificial Chromosomes (BACs)

- Bacterial-Based Cloning System Developed by Shizuya et al. (1992)
- 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
- Relatively Non-Chimeric
- Numerous Libraries Available (see www.chori.org/bacpac)
- See Birren et al. (1998)

Screening BAC Libraries by Hybridization

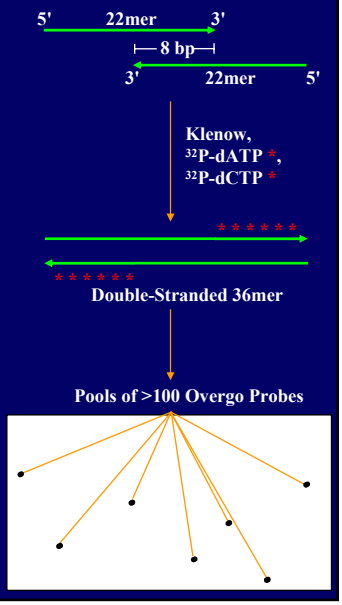
- 6 Fields, 16 x 384 BACs
- ~18,000 Unique Clones
- 4 x 4 Array
- Clones in Duplicate



‘Overgo’ Hybridization Probes

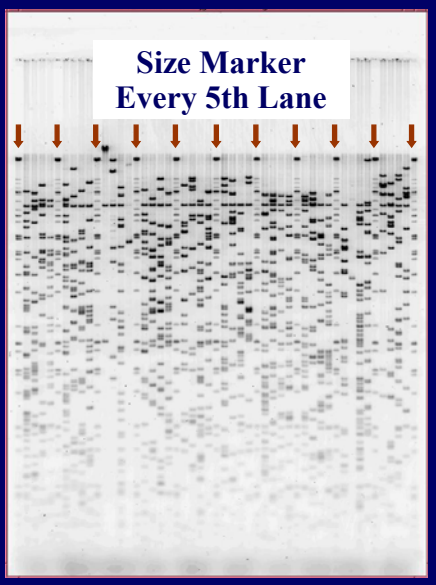
Vollrath (1999)

- Pair of ~22mer Oligonucleotide Primers with 8-bp Overlap
- Primer Extension with Klenow and Both ^{32}P-dATP and ^{32}P-dCTP
- Low Background Allows Pooling of Multiple Overgo Probes



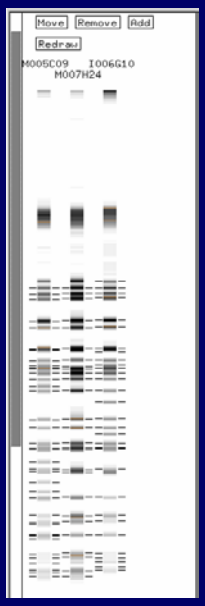
Restriction Enzyme Digest-Based Fingerprint Analysis

- BAC DNA Purification in 96-Well Format
- Single-Enzyme Digestion
- Agarose Gel

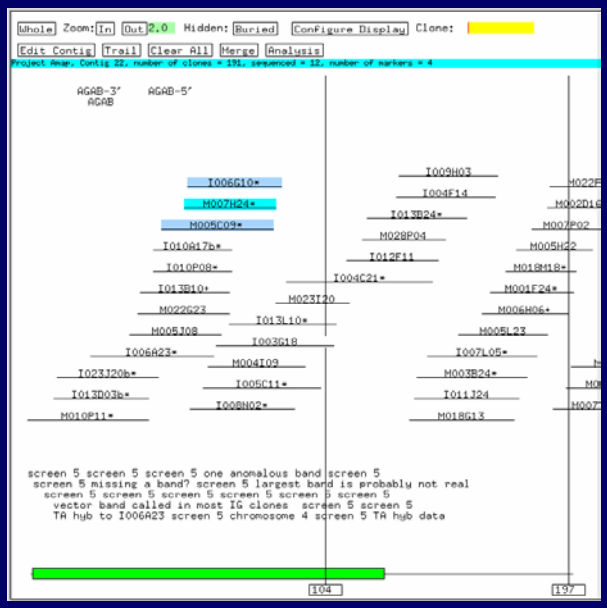


Marra et al. (1997)

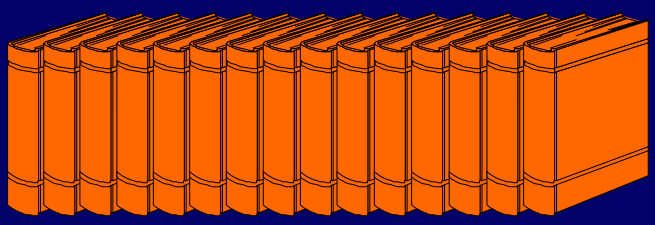
IMAGE



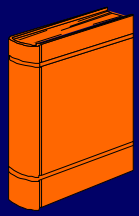
FPC



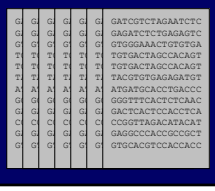
Soderlund et al. (2000) and www.sanger.ac.uk/Software



Genome
(~3000 Mb)



Chromosome
(~130 Mb)

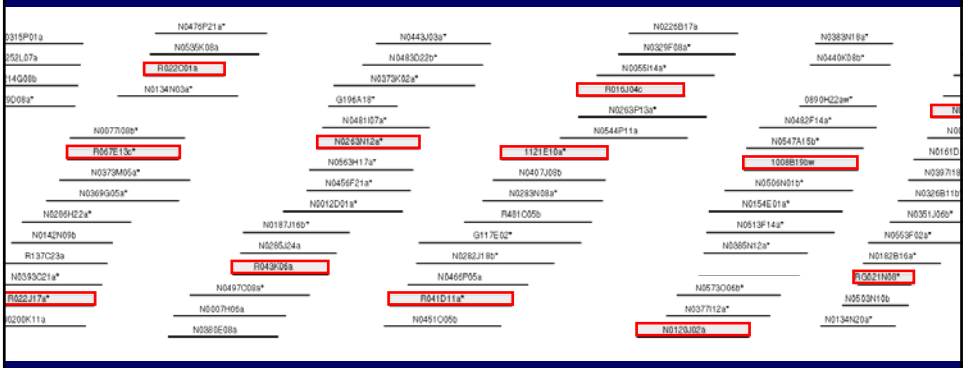


YAC
(~0.5-1.0 Mb)

BAC
(~0.1-0.2 Mb)

Sequence-Ready Contig Map

Marra et al. (1997) and Gregory et al. (1997)



BAC-Based Physical Maps of Human Genome

A physical map of the human genome

The International Human Genome Mapping Consortium*

Nature 409:934-941 (2001)

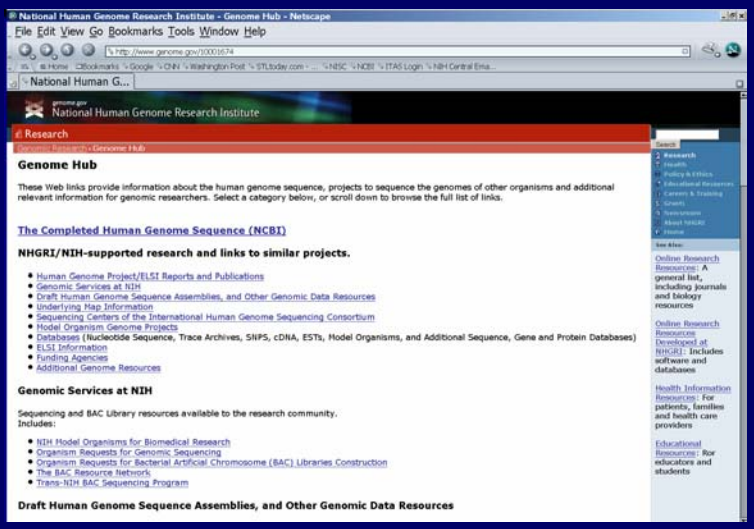
Chromosomes 1, 6, 9, 10, 13, 20, and X
Nature 409:942-943 (2001)

Y Chromosome
Nature 409:943-945 (2001)

Chromosome 12
Nature 409:945-946 (2001)

Chromosome 14
Nature 409:947-948 (2001)

The Genome Hub



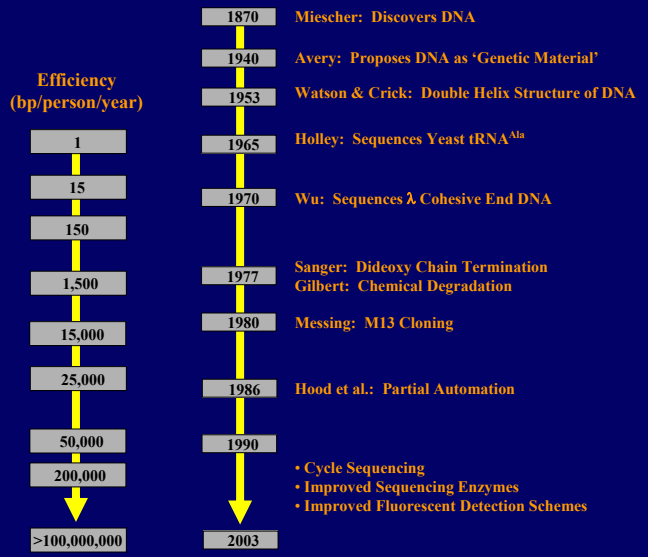
www.genome.gov/10001674

Physical Mapping: Future Prospects

- **Strategies for Physical Mapping are Radically Changing in the Sequence-Based Era**
- **Will Now See a Closer Interplay of Mapping and Sequencing in the Exploration of New Genomes**
- **Construction of New BAC Libraries will Allow Physical Mapping Studies of More Species' Genomes**
- **Sequence-Driven Approaches will Increasingly be Used for Building Comparative Physical Maps**

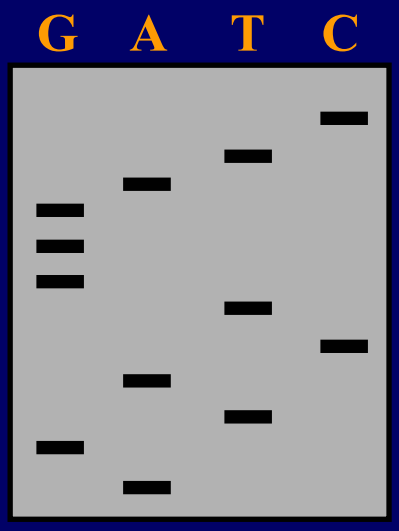
DNA Sequencing

History of DNA Sequencing



Adapted from Messing & Llaca, *PNAS* (1998)

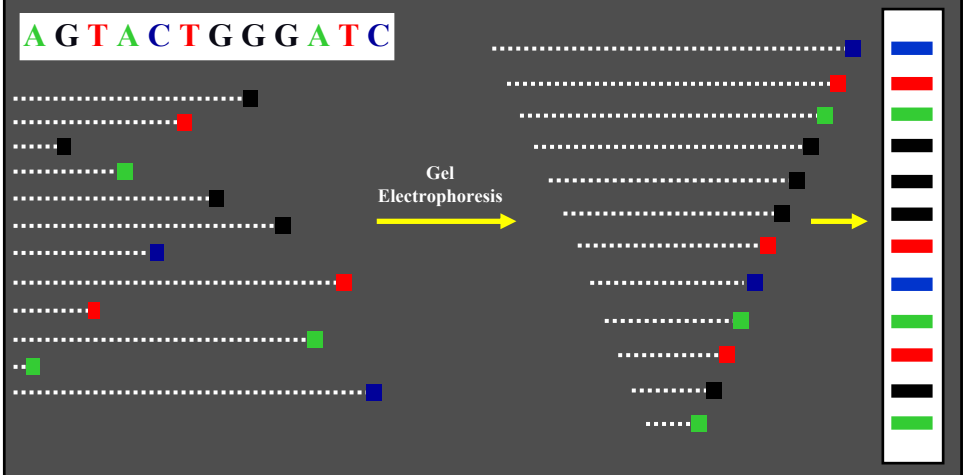
DNA Tagged with Radioactivity



G: G Reaction
A: A Reaction
T: T Reaction
C: C Reaction

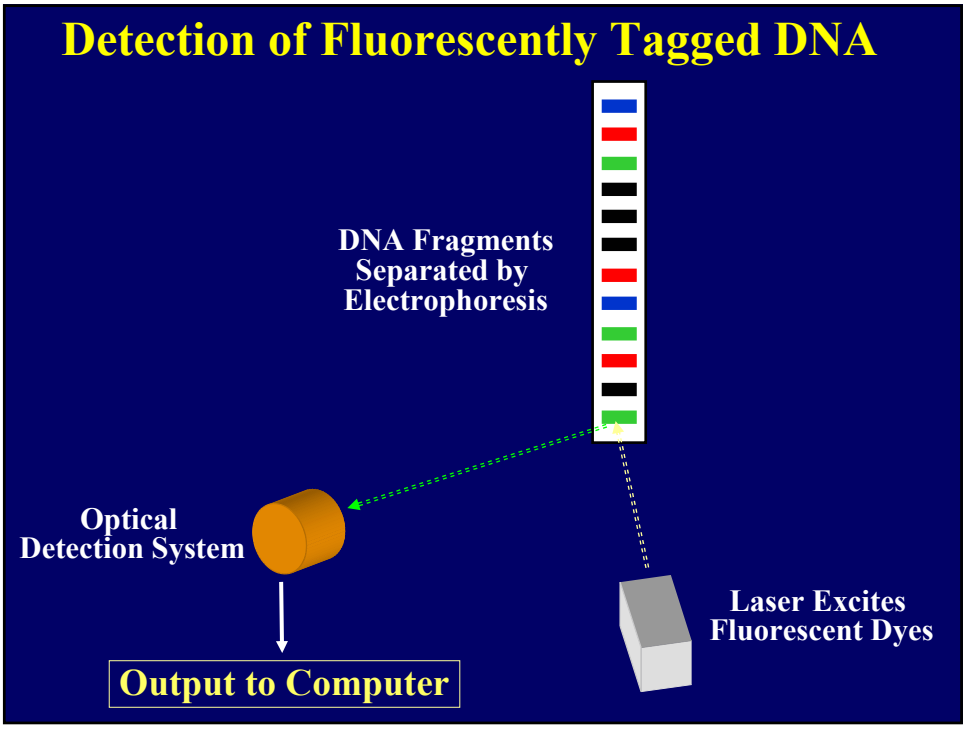
Radioactive Sequencing

Fluorescent DNA Sequencing

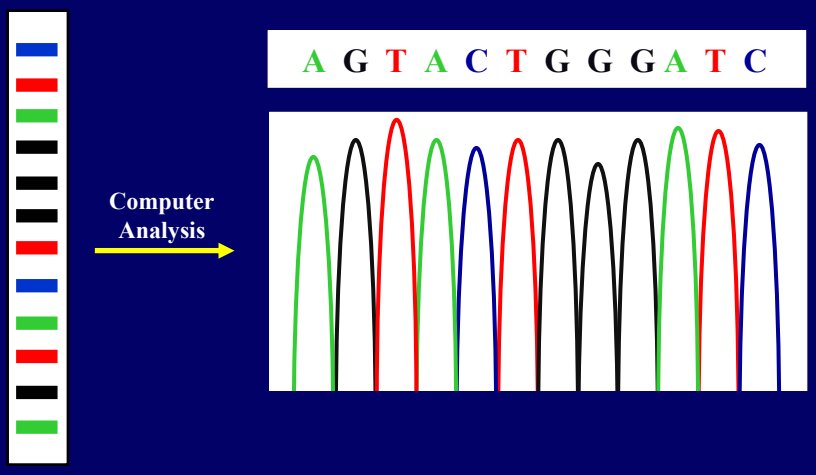


Wilson & Mardis (1997)

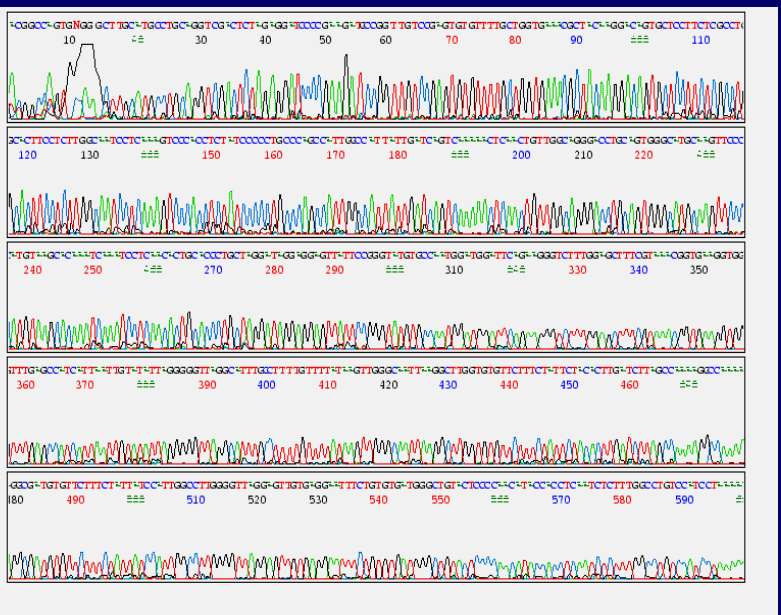
Detection of Fluorescently Tagged DNA



Analyzing Fluorescent DNA Sequencing Data



Fluorescent DNA Sequencing Results



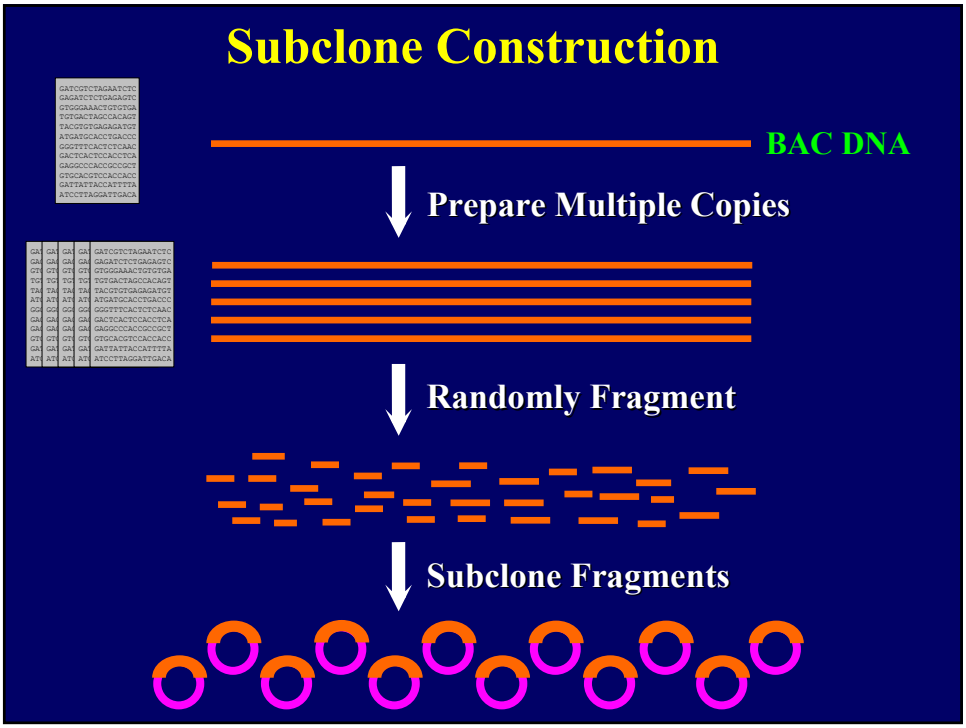
Applied Biosystems 377

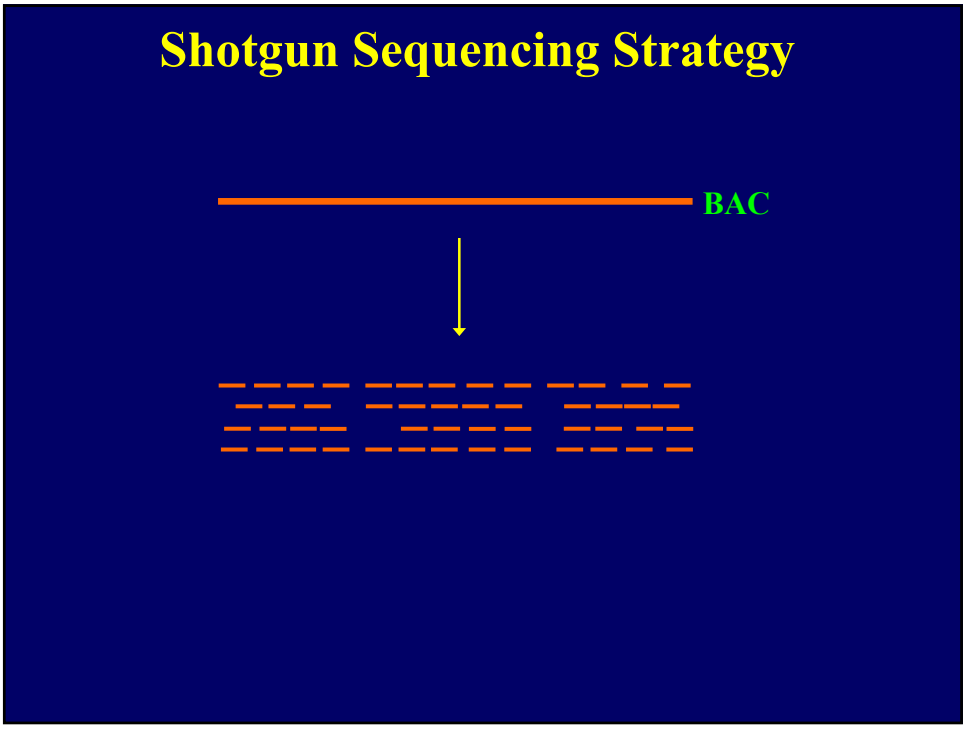
Capillary-Based DNA Sequencing Instruments

EST Sequencing
&
SAGE

Shotgun Sequencing

Wilson & Mardis (1997) and Green (2001)


Subclone Construction
BAC DNA
Prepare Multiple Copies
Randomly Fragment
Subclone Fragments


Shotgun Sequencing Strategy
BAC

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:

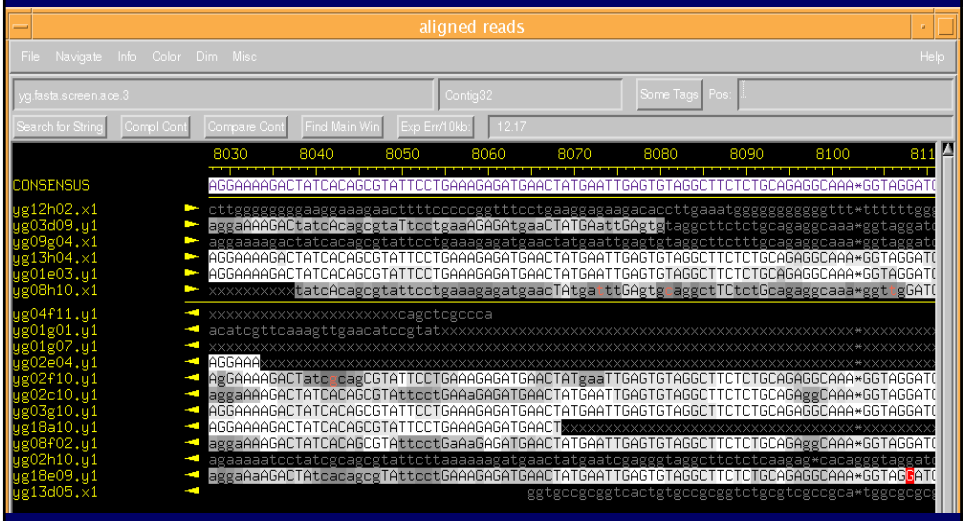
$$P_0 = e^{-c}$$

Where:

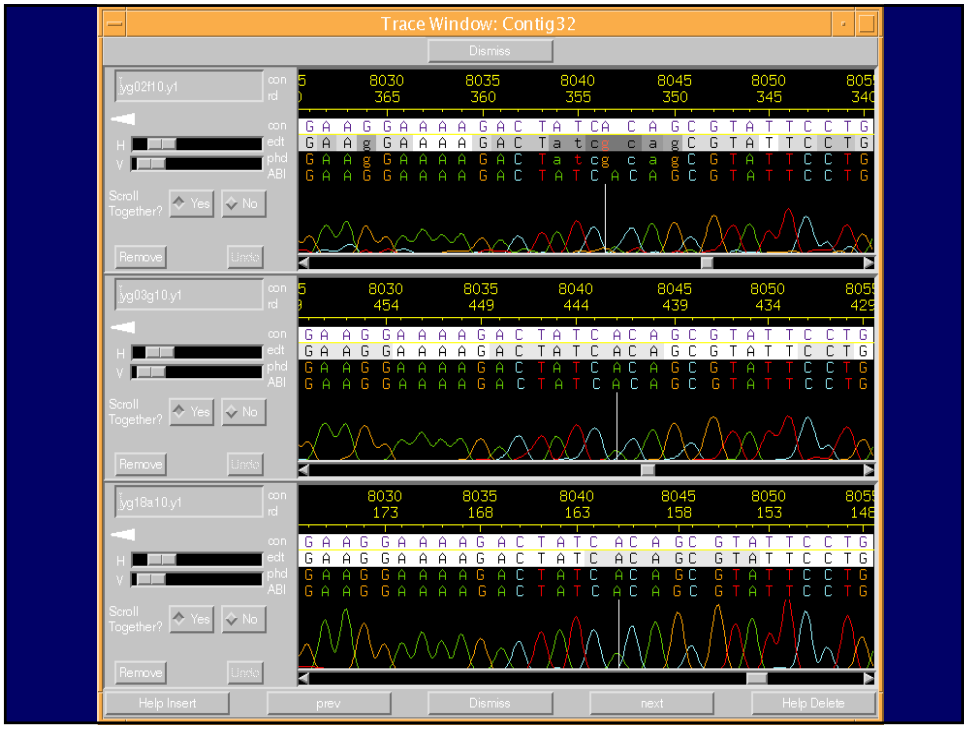
- c = fold sequence coverage (c=LN/G),
- LN = # bases sequenced, i.e. L = average sequencing read length and N = # reads
- G = target sequence length
- e = 2.718 (e=2.718281828459)

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%

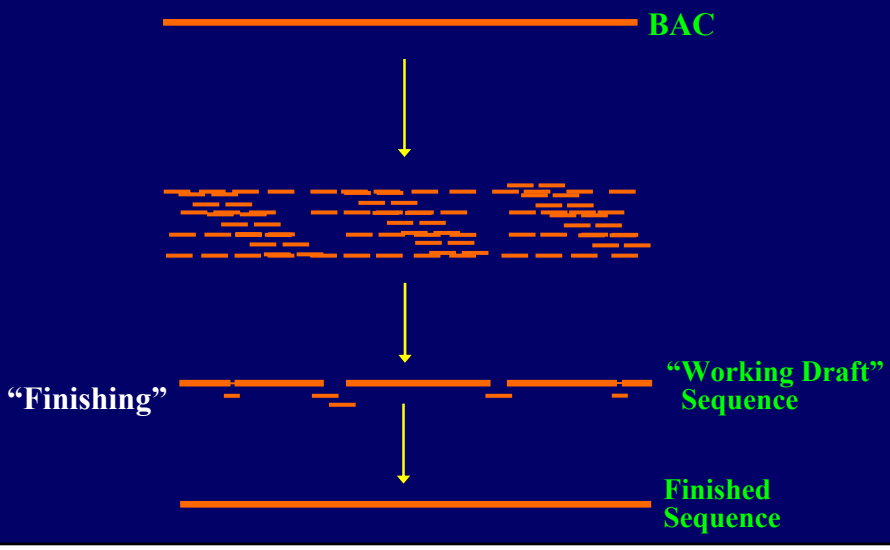
Shotgun Sequence Assembly



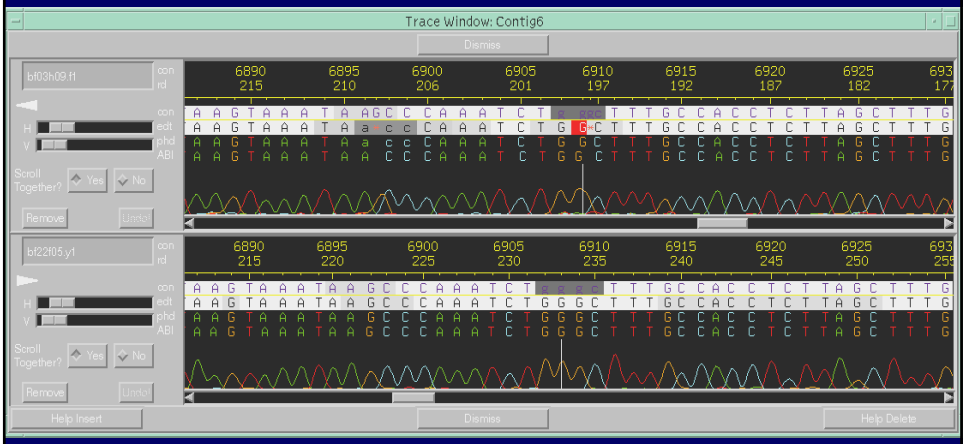
"Consed" (Gordon et al., 1998)



Shotgun Sequencing Strategy

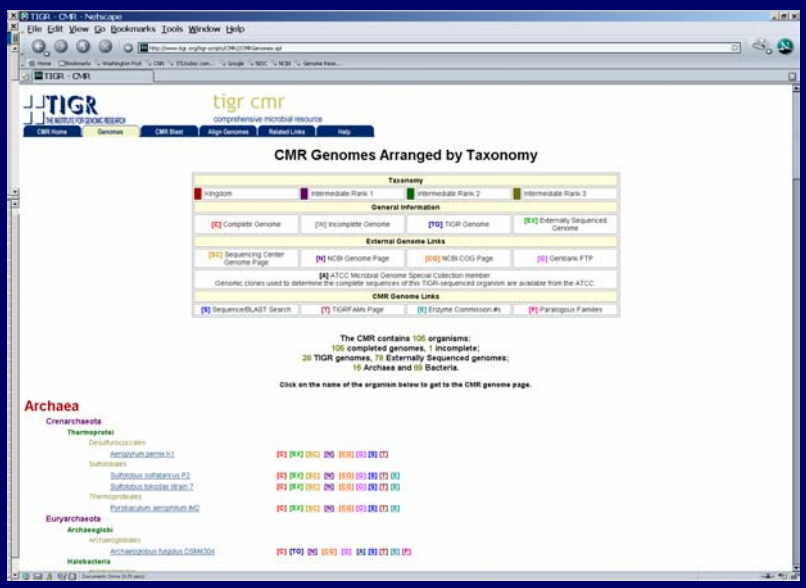


Sequence Finishing: Resolving Ambiguities



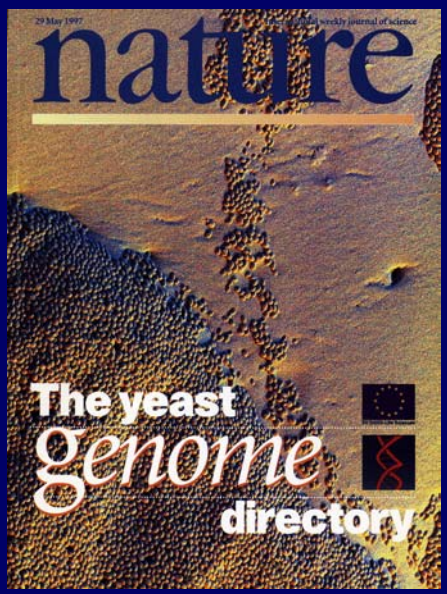
Large-Scale DNA Sequencing Projects

Bacterial Genome Sequences



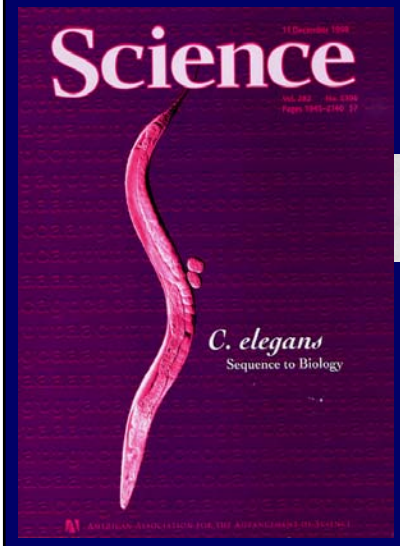
www.tigr.org

First Eukaryotic Genome Sequence



Nature 387:1-105, 1997

First Animal Genome Sequence

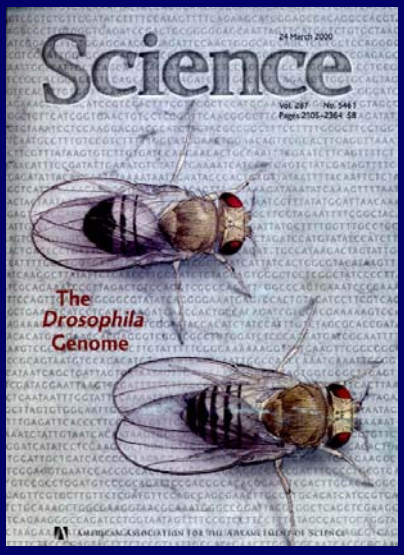


Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

The *C. elegans* Sequencing Consortium*

Science 282:1012-2018, 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,[2] 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,[2] 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,[6] 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,[2] 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,[3] Bruce E. Kimmel,[2] Chinnappa D. Kodira,[1] Cheryl Kraft,[1] Saul Kravitz,[1] David Kulp,[2] 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,[4] 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,[4] Knut Reinert,[1] Karin Remington,[1] Robert D. C. Saunders,[30] Frederick Scheeler,[1] Hua Shen,[3] Bixiang Christopher Shue,[1] Inga Silén-Kiamos,[31] Michael Simpson,[1] Marian P. Skupski,[1] Tom Smith,[1] Eugene Spier,[1] Allan C. Spradling,[31] Mark Stapleton,[2] Renee Strong,[1] Eric Sun,[1] Robert Svirskas,[32] Cyndee Tector,[1] Russell Turner,[1] Eli Venter,[1] Alhul 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] Xiangqun H. Zheng,[1] Fei N. Zhong,[1] Wenyan Zhong,[1] Xiaojun Zhou,[2] 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]

Science 287:2185-2195, 2000

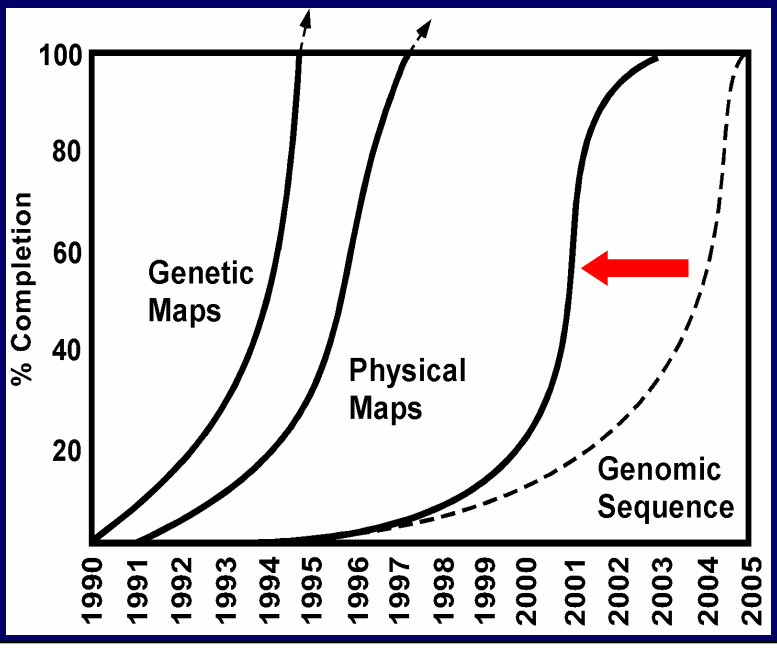
Human Genome Project: 5 Year Goals

New Goals for the U.S. Human Genome Project: 1998–2003

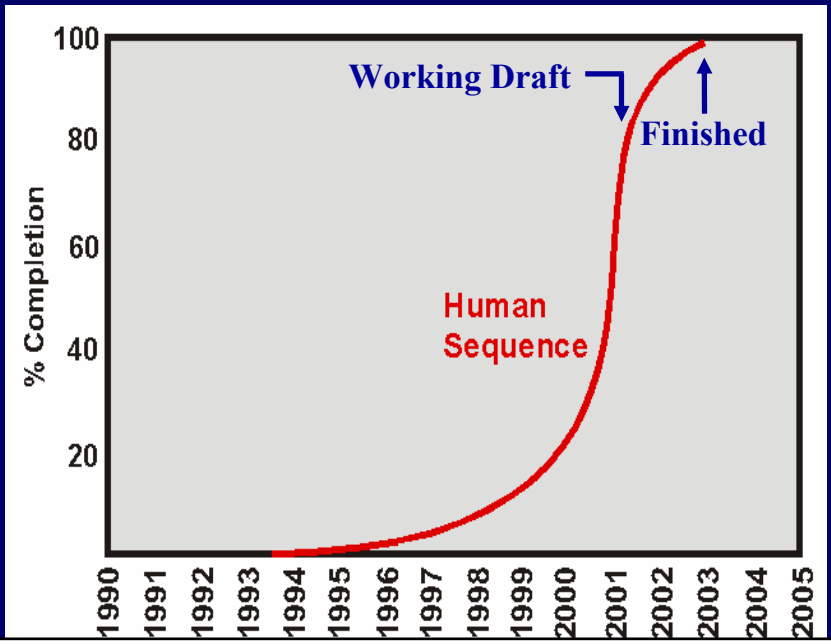
Francis S. Collins,* Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, and the members of the DOE and NIH planning groups

Science 282:682-689, 1998

Revised Timetable for Human Sequencing



Timetable for Human Genome Sequencing



Human Genome Sequencing Centers

Genome Sequencing Center
Washington University
School of Medicine
St. Louis, MO USA

Whitehead Institute/MIT
Genome Sequencing Center

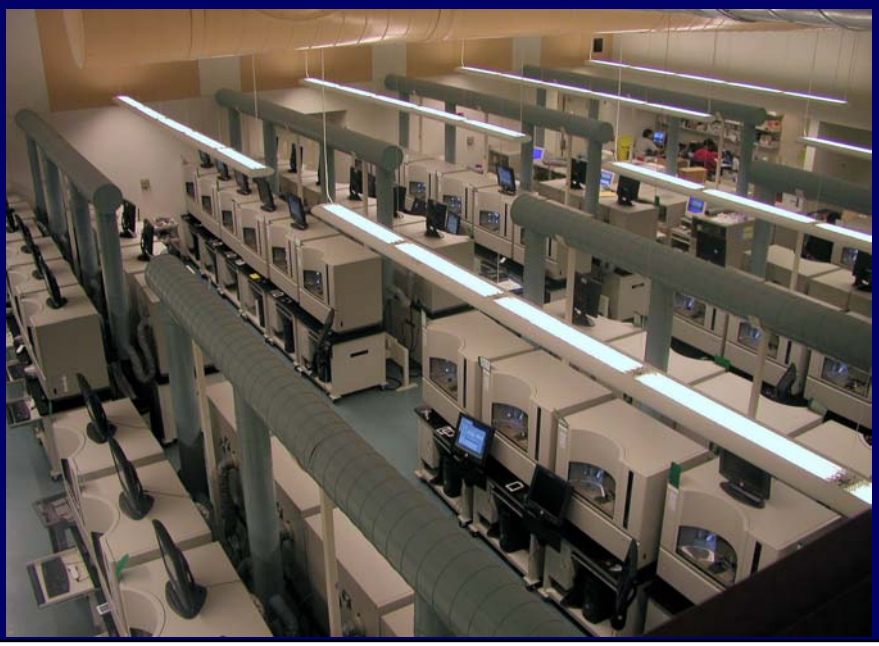
Baylor College of Medicine

HGSC
HUMAN GENOME SEQUENCING CENTER

JGI
JOINT GENOME INSTITUTE

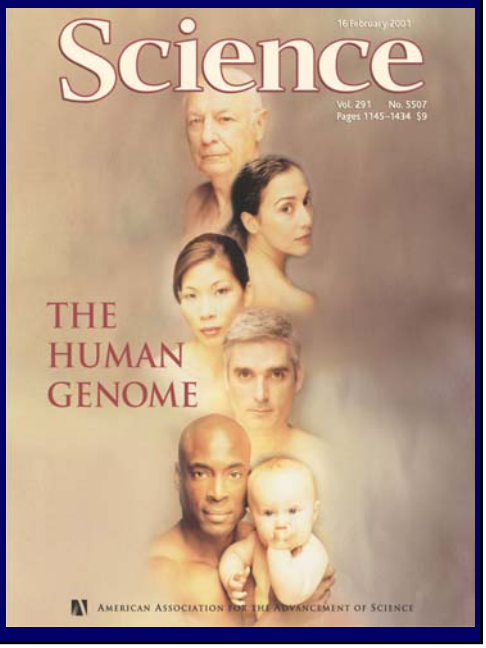
The Sanger Centre

Human Genome Sequencing Centers

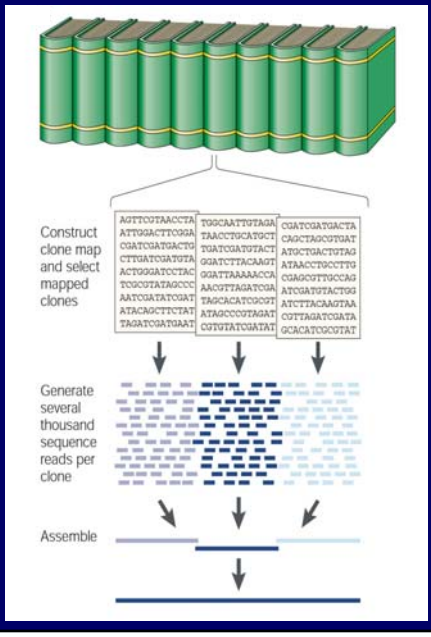


June, 2000 Announcement

February, 2001 Publications

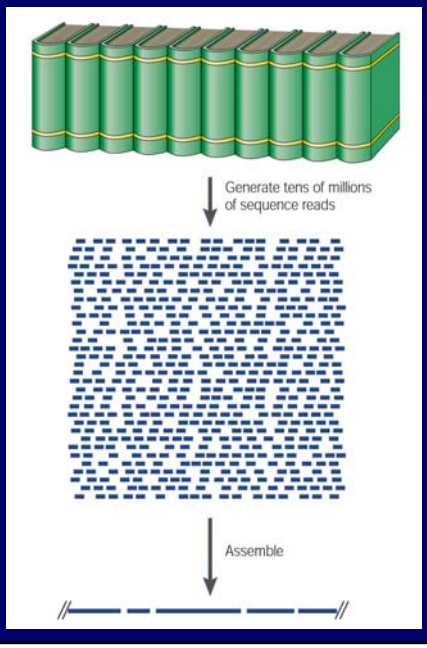


BAC-by-BAC Shotgun Sequencing



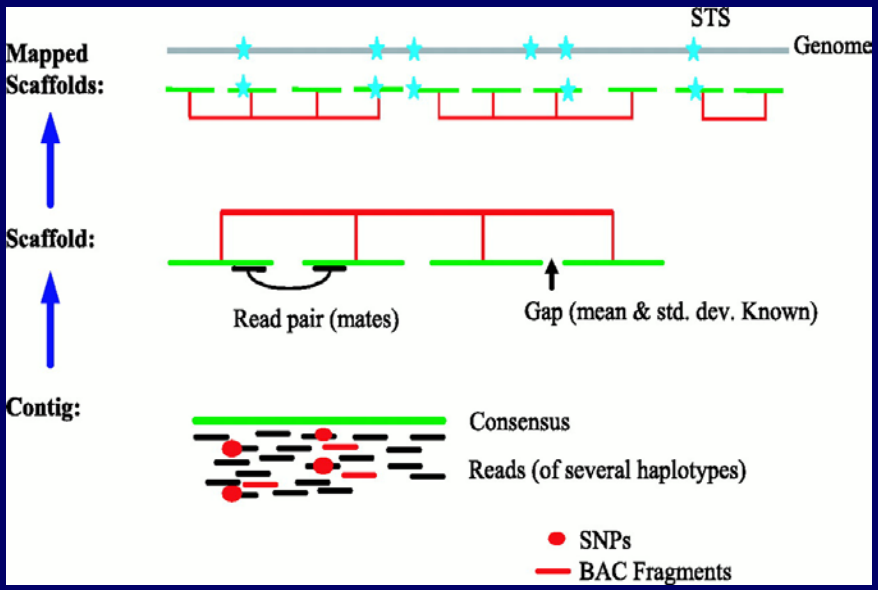
Green (2001)

Whole-Genome Shotgun Sequencing



Green (2001)

Whole-Genome Shotgun Sequence Assembly



Venter et al., 2001

April, 2003 Completion

International Human Genome Sequencing Consortium

- 6 Countries
- 20 Sequencing Centers
- 1000's of Individuals
- ~1,000 bases per second, 24 hours per day, 7 days per week

108TH CONGRESS
1ST SESSION

S. CON. RES. 10

Designating April 2003 as "Human Genome Month" and April 25 as "DNA Day".

IN THE SENATE OF THE UNITED STATES

FEBRUARY 27, 2003

Mr. GREGG (for himself, Mr. KENNEDY, Ms. SNOWE, and Mr. DASCHLE) submitted the following concurrent resolution; which was considered and agreed to

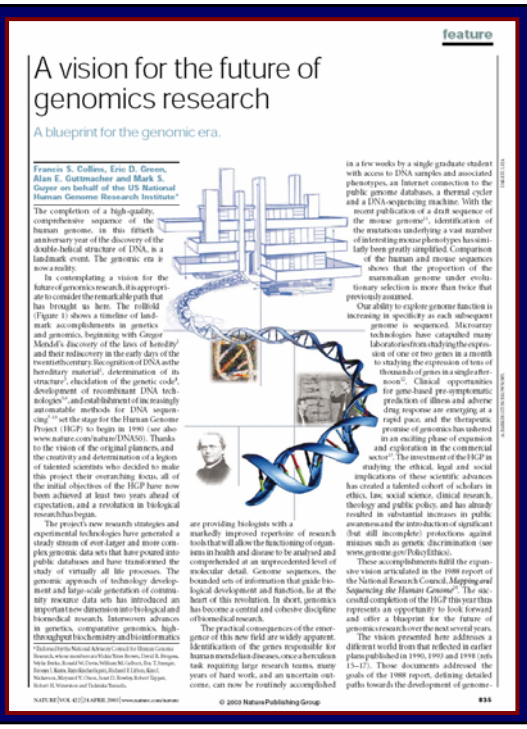
---

CONCURRENT RESOLUTION

Designating April 2003 as "Human Genome Month" and April 25 as "DNA Day".

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

What's Next?



feature

A vision for the future of genomics research

A blueprint for the genomic era.

Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer on behalf of the US National Human Genome Research Institute*

The completion of a high-quality, comprehensive sequence of the human genome, in this fiftieth anniversary year of the discovery of the double-helical structure of DNA, is a landmark event. The genomic era is now a reality.

In contemplating a vision for the future of genomics research, it is appropriate to consider the remarkable path that has brought us here. The rollfold (Figure 1) shows a timeline of landmark accomplishments in genetics and genomics, beginning with Gregor Mendel's discovery of the laws of heredity[1] and their rediscovery in the early days of the twentieth century. Recognition of DNA as the hereditary material[2], determination of its structure[3], elucidation of the genetic code[4], development of recombinant DNA technologies[5,6], and establishment of increasingly automatable methods for DNA sequencing[7–10] set the stage for the Human Genome Project (HGP) to begin in 1990 (see also www.nature.com/nature/DNA50). Thanks to the vision of the original planners, and the creativity and determination of a legion of talented scientists who decided to make this project their overarching focus, all of the initial objectives of the HGP have now been achieved at least two years ahead of expectation, and a revolution in biological research has begun.

The project's new research strategies and experimental technologies have generated a steady stream of ever larger and more complex genomic data sets that have poured into public databases and have transformed the study of virtually all life processes. The genomic approach of technology development and large-scale generation of community resource data sets has introduced an important new dimension into biological and biomedical research. Interwoven advances in genetics, comparative genomics, high-throughput biochemistry and bioinformatics are providing biologists with a markedly improved repertoire of research tools that will allow the functioning of organisms in health and disease to be analysed and comprehended at an unprecedented level of molecular detail. Genome sequences, the bounded sets of information that guide biological development and function, lie at the heart of this revolution. In short, genomics has become a central and cohesive discipline of biomedical research.

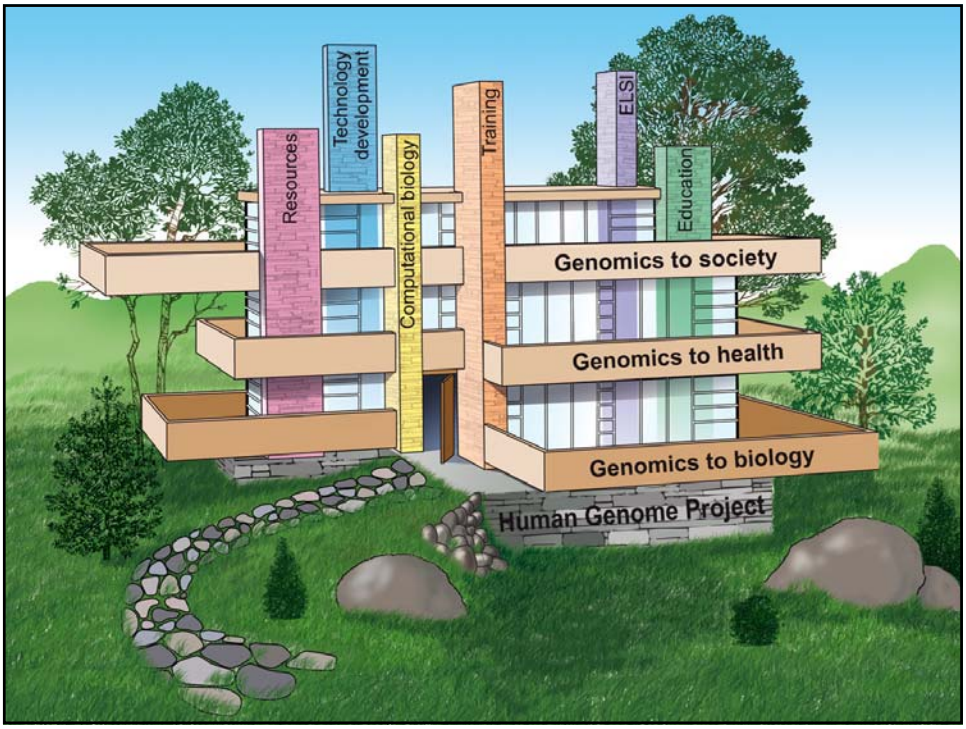
The practical consequences of the emergence of this new field are widely apparent. Identification of the genes responsible for human mendelian diseases, once a herculean task requiring large research teams, many years of hard work, and an uncertain outcome, can now be routinely accomplished in a few weeks by a single graduate student with access to DNA samples and associated phenotypes, an Internet connection to the public genome databases, a thermal cycler and a DNA-sequencing machine. With the recent publication of a draft sequence of the mouse genome[11], identification of the mutations underlying a vast number of interesting mouse phenotypes has similarly been greatly simplified. Comparison of the human and mouse sequences shows that the proportion of the mammalian genome under evolutionary selection is more than twice that previously assumed.

Our ability to explore genome function is increasing in specificity as each subsequent genome is sequenced. Microarray technologies have catapulted many laboratories from studying the expression of one or two genes in a month to studying the expression of tens of thousands of genes in a single afternoon[12]. Clinical opportunities for gene-based pre-symptomatic prediction of illness and adverse drug response are emerging at a rapid pace, and the therapeutic promise of genomics has ushered in an exciting phase of expansion and exploration in the commercial sector[13]. The investment of the HGP in studying the ethical, legal and social implications of these scientific advances has created a talented cohort of scholars in ethics, law, social science, clinical research, theology and public policy, and has already resulted in substantial increases in public awareness and the introduction of significant (but still incomplete) protections against misuses such as genetic discrimination (see www.genome.gov/Policy/Ethics).

These accomplishments fulfil the expansive vision articulated in the 1988 report of the National Research Council, *Mapping and Sequencing the Human Genome*[14]. The successful completion of the HGP this year thus represents an opportunity to look forward and offer a blueprint for the future of genomics research over the next several years.

The vision presented here addresses a different world from that reflected in earlier plans published in 1990, 1993 and 1998 (refs 15–17). Those documents addressed the goals of the 1988 report, defining detailed paths towards the development of genome-

*Full text Of the National Advisory Council for Human Genome Research, whose members are Vivian Bonn-Brown, David R. Burgess, Sylvia Burke, Ronald W. Davis, William M. Gelbart, Eric T. Juengst, Jerome I. Kurtz, Jane Kirschstein, Richard E. Litton, Karl L. Nickerson, Maynard V. Olson, Joan D. Ramsey, Robert Tjian, Robert H. Waterston and Yoshiaki Yasuda.


Resources
Technology development
Computational biology
Training
ELSI
Education
Genomics to society
Genomics to health
Genomics to biology
Human Genome Project


PHASE TWO • INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

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

Beyond the Human Genome Sequence…

Comparing the Human Genetic Blueprint to that of Other Species

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
  - Know Very Little about What to Look For
  - Virtually No Complementary Data Sets Available
  - Poor Computational Predictions

Major role for comparative sequencing is the identification of functionally important non-coding sequences

Whole-Genome Vertebrate Sequencing Efforts

Human

Pufferfish

Sequence of a 'Compact' Vertebrate Genome

RESEARCH ARTICLE

Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*

Samuel Aparicio,[2,3*] Jarrod Chapman,[3] Elia Stupka,[1*] Nik Putnam,[3] Jer-ming Chia,[1] Paramvir Dehal,[3] Alan Christoffels,[1] Sam Rash,[3] Shawn Hoon,[1] Arian Smit,[4] Maarten D. Sollewijn Gelpke,[3] Jared Roach,[4] Tania Oh,[1] Isaac Y. Ho,[3] Marie Wong,[1] Chris Detter,[3] Frans Verhoef,[1] Paul Predki,[3] Alice Tay,[1] Susan Lucas,[3] Paul Richardson,[3] Sarah F. Smith,[5] Melody S. Clark,[5] Yvonne J. K. Edwards,[5] Norman Doggett,[6] Andrey Zharkikh,[7] Sean V. Tavtigian,[7] Dmitry Pruss,[7] Mary Barnstead,[8] Cheryl Evans,[8] Holly Baden,[8] Justin Powell,[9] Gustavo Glusman,[4] Lee Rowen,[4] Leroy Hood,[4] Y. H. Tan,[1] Greg Elgar,[5*] Trevor Hawkins,[3*†] Byrappa Venkatesh,[1*] Daniel Rokhsar,[3*] Sydney Brenner[1,10*]

The compact genome of *Fugu rubripes* has been sequenced to over 95% coverage, and more than 80% of the assembly is in multigene-sized scaffolds. In this 365-megabase vertebrate genome, repetitive DNA accounts for less than one-sixth of the sequence, and gene loci occupy about one-third of the genome. As with the human genome, gene loci are not evenly distributed, but are clustered into sparse and dense regions. Some "giant" genes were observed that had average coding sequence sizes but were spread over genomic lengths significantly larger than those of their human orthologs. Although three-quarters of predicted human proteins have a strong match to *Fugu*, approximately a quarter of the human proteins had highly diverged from or had no pufferfish homologs, highlighting the extent of protein evolution in the 450 million years since teleosts and mammals diverged. Conserved linkages between *Fugu* and human genes indicate the preservation of chromosomal segments from the common vertebrate ancestor, but with considerable scrambling of gene order.

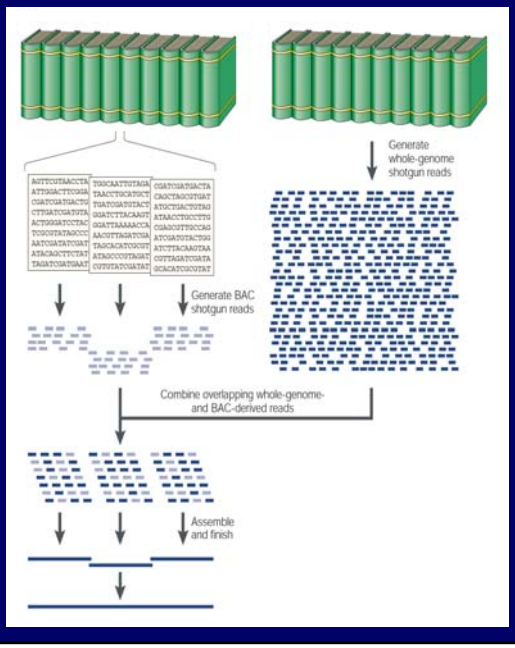
Introduction

Whole-Genome Vertebrate Sequencing Efforts

Human **Mouse**

Pufferfish

Hybrid Shotgun Sequencing



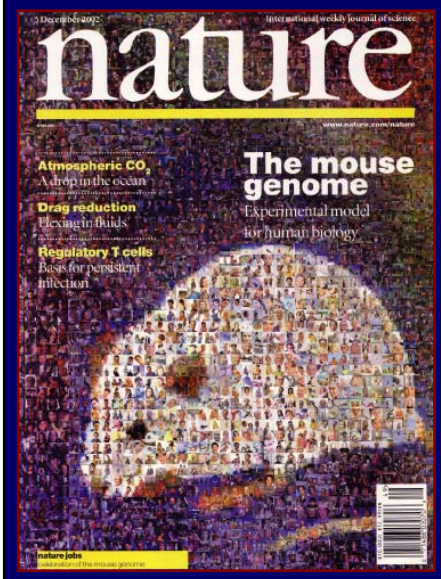
Green (2001)

Hybrid Shotgun Sequencing

What is the optimal mixture???

Redundant Coverage:	**0**	**~5-fold**	**~10-fold**
	Whole-Genome Shotgun Sequence Reads		
	BAC-by-BAC Shotgun Sequence Reads		
Redundant Coverage:	**~10-fold**	**~5-fold**	**0**

Human-Mouse Sequence Comparisons



- ~40% in Alignments
- ~5% Under Selection
  - ~1.5% Protein Coding
  - ~3.5% Non-Coding

Multi-Species Comparative Sequence Analysis

Comparative analyses of multi-species sequences from targeted genomic regions

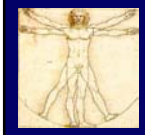
The systematic comparison of genomic sequences from different organisms represents a central focus of contemporary genome analysis. Comparative analyses of vertebrate sequences can identify coding[1–4] and conserved non-coding[5,6] regions, including regulatory elements[7–9], and provide insight into the forces that have rendered modern-day genomes[1]. As a complement to whole-genome sequencing efforts[2,3,10], we are sequencing and comparing targeted genomic regions in multiple, evolutionarily diverse vertebrates. Here we report the generation and analysis of over 12 megabases (Mb) of sequence from 12 species, all derived from the genomic region orthologous to a segment of about 1.8 Mb on human chromosome 7 containing ten genes, including the gene mutated in cystic fibrosis. These sequences show conservation reflecting both functional constraints and the neutral mutational events that shaped this genomic region. In particular, we identify substantial numbers of conserved non-coding segments beyond those previously identified experimentally, most of which are not detectable by pair-wise sequence comparisons alone. Analysis of transposable element insertions highlights the variation in genome dynamics among these species and confirms the placement of rodents as a sister group to the primates.

The NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program aims to sequence and to analyse targeted genomic regions in multiple vertebrates. Our initial target is a genomic segment of about 1.8 Mb on human chromosome 7q31.3

- Targeted Genomic Regions
- BAC-Based Sequencing in Multiple Vertebrates
- Identify Highly Conserved Non-Coding Sequences
- Conserved Sequences Correlate with Functional Elements

Thomas et al. (2003)

Whole-Genome Vertebrate Sequencing Efforts


Human


Mouse


Rat


Chimpanzee


Dog


Cow

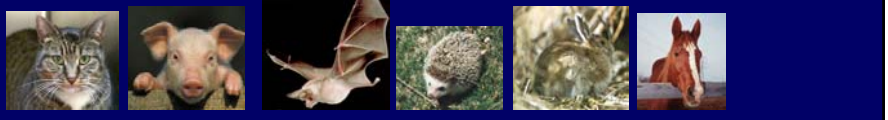

Chicken

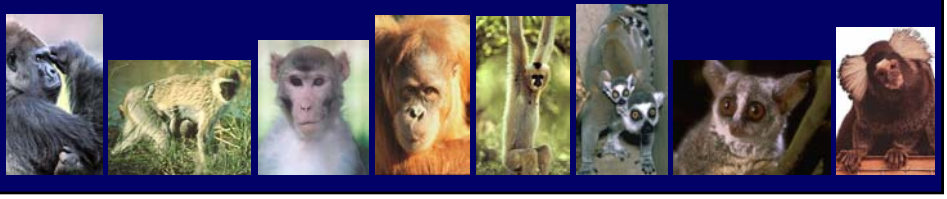

Xenopus


Zebrafish


Pufferfish

Future Genomes to Sequence???





Bibliography

Adams, M.D. et al. (2000). The genome sequence of Drosophila melanogaster. Science 287, 2185-2195.

Aparicio, S. et al. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297, 1301-1310.

Bentley, D.R. et al. (2001). The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20, and X. Nature 409, 942-943.

Birren, B. et al. (1998). Bacterial artificial chromosomes. In Genome analysis: a laboratory manual. Vol. 3 Cloning systems (ed. B. Birren et al.), pp. 241-295. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Bruls et al. (2001). A physical map of human chromosome 14. Nature 409, 947-948.
Burke, D.T. et al. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236, 806-812.

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

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

Collins, F.S. et al. (1998). New goals for the U.S. Human Genome Project: 1998-2003. Science 282, 682-689.

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 Research 8, 195-202.

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

Green, E.D. and Green, P. (1991). Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences. PCR Methods Applic. 1, 77-90.

Green, E.D. et al. (1998). Yeast artificial chromosomes. In Genome analysis: A laboratory manual. Vol. 3 Cloning systems (ed. B. Birren et al.), pp. 297-565. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Gregory, S.G. et al. (1997). Genome mapping by fluorescent fingerprinting. Genome Research 7, 1162-1168.

International Human Genome Mapping Consortium (2001). A physical map of the human genome. Nature 409, 934-941.

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

Marra, M.A. et al. (1997). High throughput fingerprint analysis of large-insert clones. Genome Research 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.

Montgomery, K.T. et al. (2001). A high-resolution map of human chromosome 12. Nature 409, 945-946.

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

Olson, M. et al. (1989). A common language for physical mapping of the human genome. Science 245, 1434-1440.

Riethman, H. et al. (1997). Preparation, manipulation, and mapping of HMW DNA. In Genome analysis: a laboratory manual. Vol. 1 Analyzing DNA (ed. B. Birren et al.), pp. 83-248. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Shizuya, H. et al. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc. Natl. Acad. Sci. 89, 8794-8797.

Soderlund, C. et al. (2000). Contigs built with fingerprints, markers, and FPC V4.7. Genome Res. 10, 1772-1787.

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

Tilford, C.A. et al. (2001). A physical map of the human Y chromosome. Nature 409, 943-945.

Venter, J.C. et al. (2001). The sequence of the human genome. Science 291, 1304-1351.

Vollrath, D. (1999). DNA markers for physical mapping. In Genome analysis: a laboratory manual. Vol. 4 Mapping genomes (ed. B. Birren et al.), pp. 187-215. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Wilson, R.K. and Mardis, E.R. (1997). Fluorescence-based DNA sequencing. In Genome analysis: a laboratory manual. Vol. 1 Analyzing DNA (ed. B. Birren et al.), pp. 301-395. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Wilson, R.K. and Mardis, E.R. (1997). Shotgun sequencing. In Genome analysis: a laboratory manual. Vol. 1 Analyzing DNA (ed. B. Birren et al.), pp. 397-454. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.