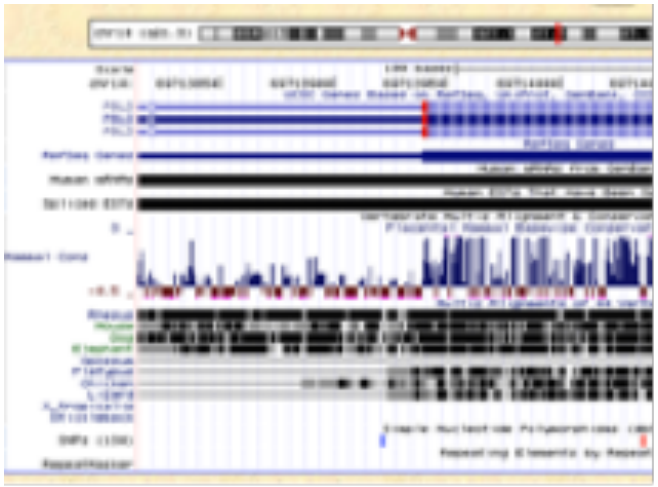
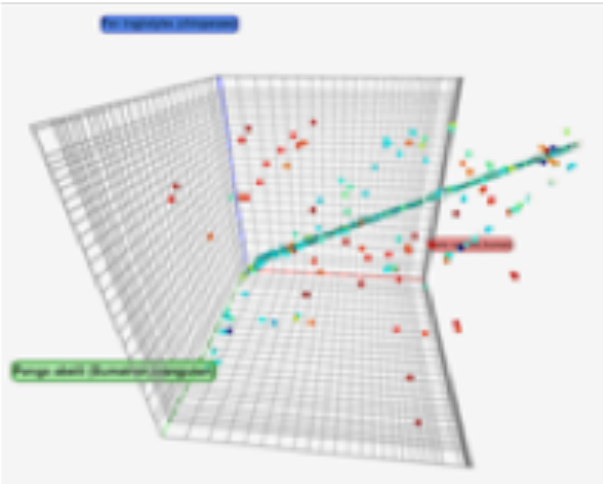
Computational Genomics

History of Genomics





National Human Genome Research Institute

Division of Intramural Research

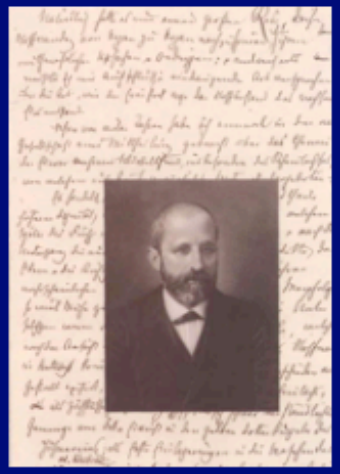
The Genomic Landscape: *circa* 2010

Eric Green, M.D., Ph.D.
Director, NHGRI

Foundational Milestones in Genetics & Genomics

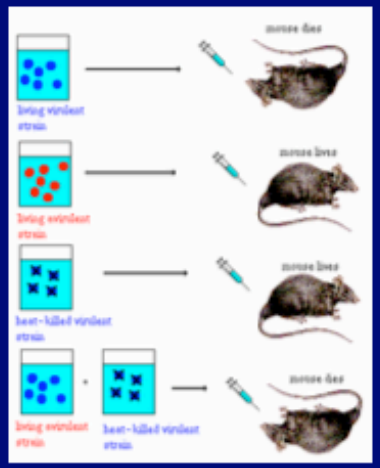
Mendel

1865



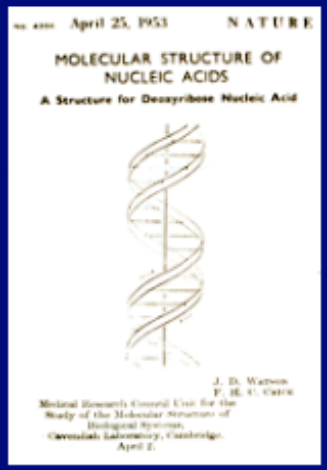
Miescher

1871



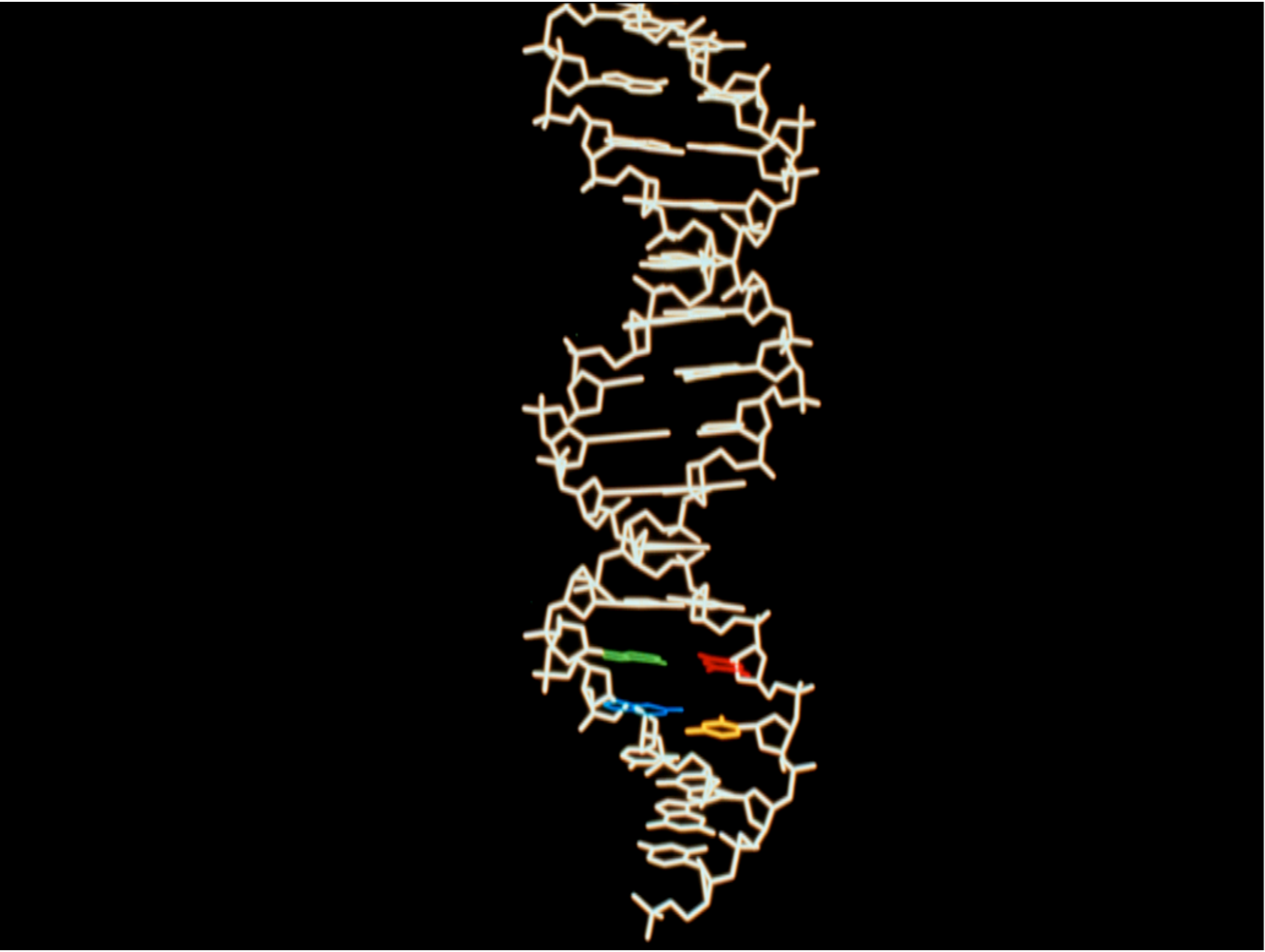
Avery

1944



Watson & Crick

1953



Human Genome Project

1990

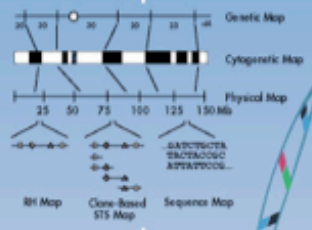
Human Genome Project (HGP) launched in the U.S.

Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE

First gene for breast cancer (BRCA1) mapped

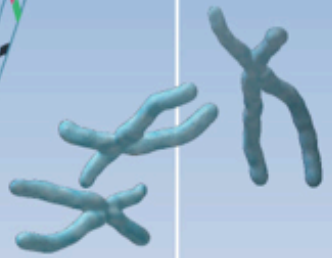
1991

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

Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)

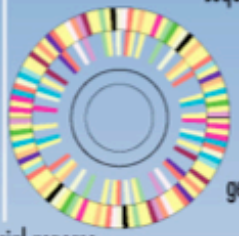
The Wellcome Trust

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

Bermuda principles for rapid and open data release established

Collins et al. (2003)

Human Genome Project

1997

DOE forms Joint Genome Institute

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

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
GTCCT

Chinese National Human Genome Centers (in Beijing and Shanghai) 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*) genome sequenced

Executive order bans genetic discrimination in U.S. federal workplace

2001

Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian Gene Collection

2002

Draft version of mouse genome sequence completed and published

Draft version of rat genome sequence completed

Draft version of rice genome sequence completed and published

2003

Finished version of human genome sequence completed

HGP ends with all goals achieved

to be continued..

Collins et al. (2003)

Outline

I. Fundamentals of Genome Mapping & Sequencing

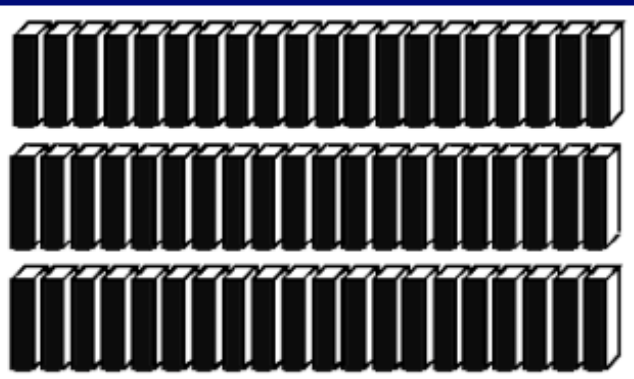
II. Mapping & Sequencing in the Human Genome Project

III. Comparative Sequencing

IV. New Frontiers in Genomics

Genome Sizes

Human Genome
Mouse Genome



~3,000,000,000 bp

Fruit Fly Genome

~160,000,000 bp

Nematode Genome

~100,000,000 bp

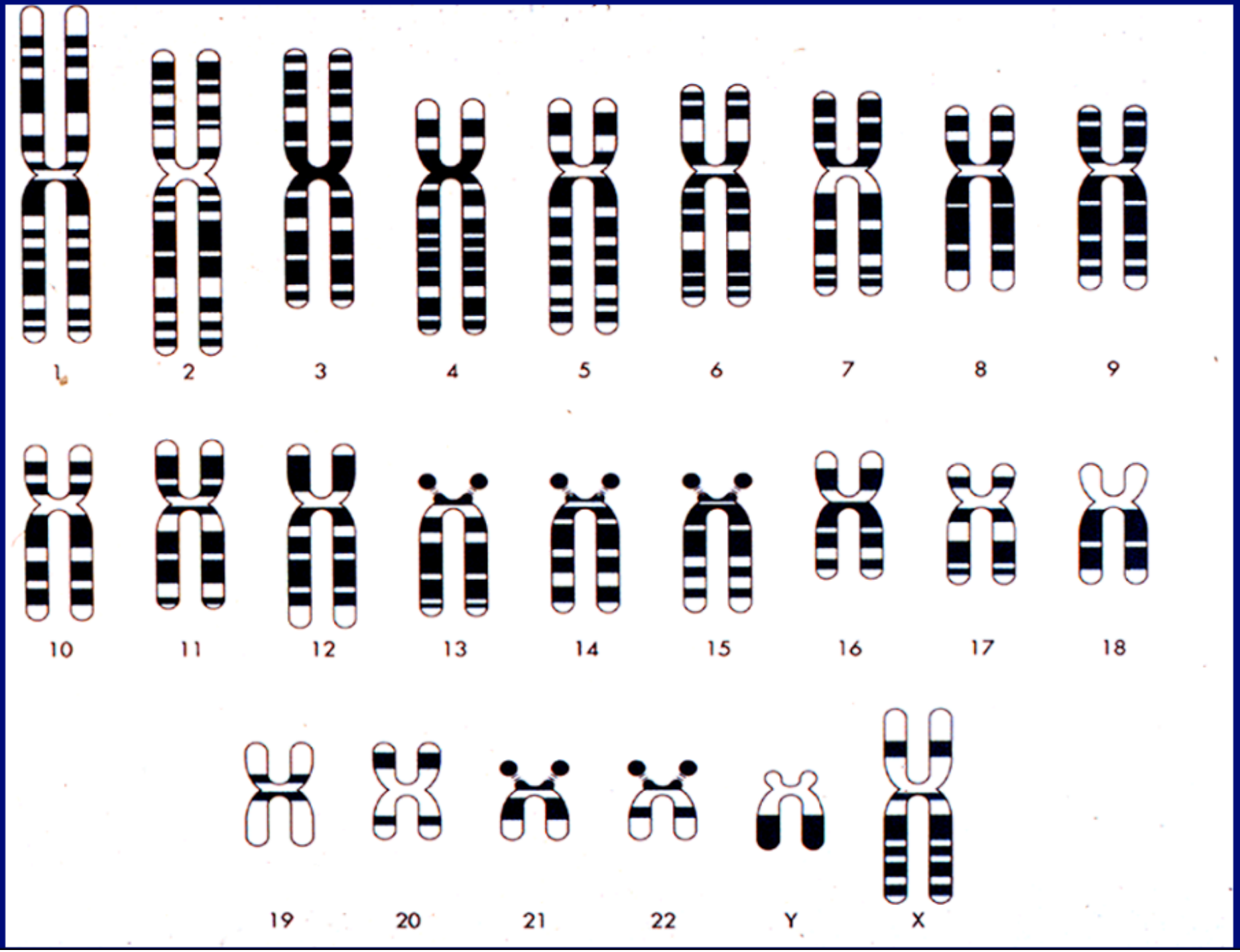
Yeast Genome

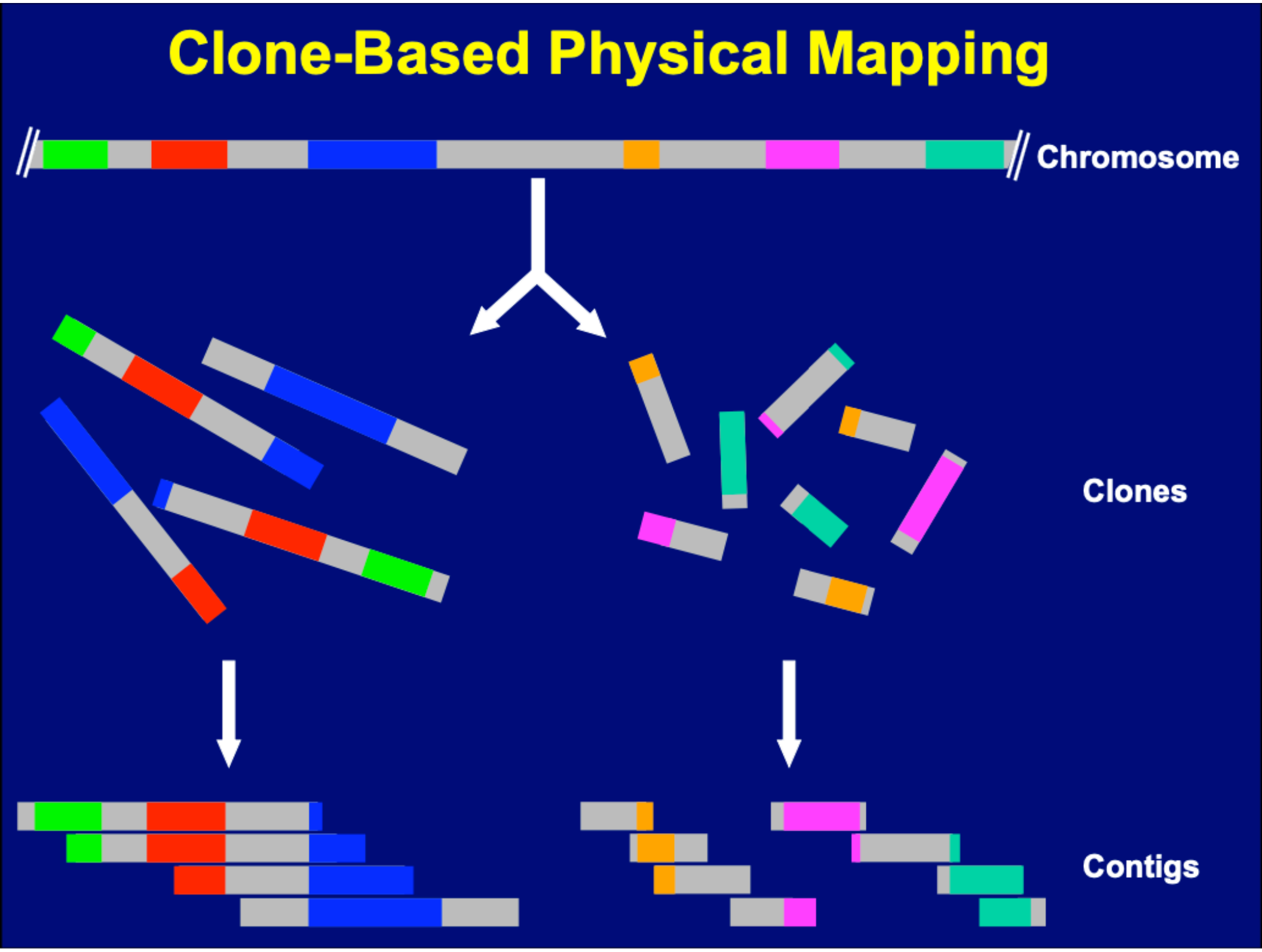
~15,000,000 bp

***E. coli* Genome**

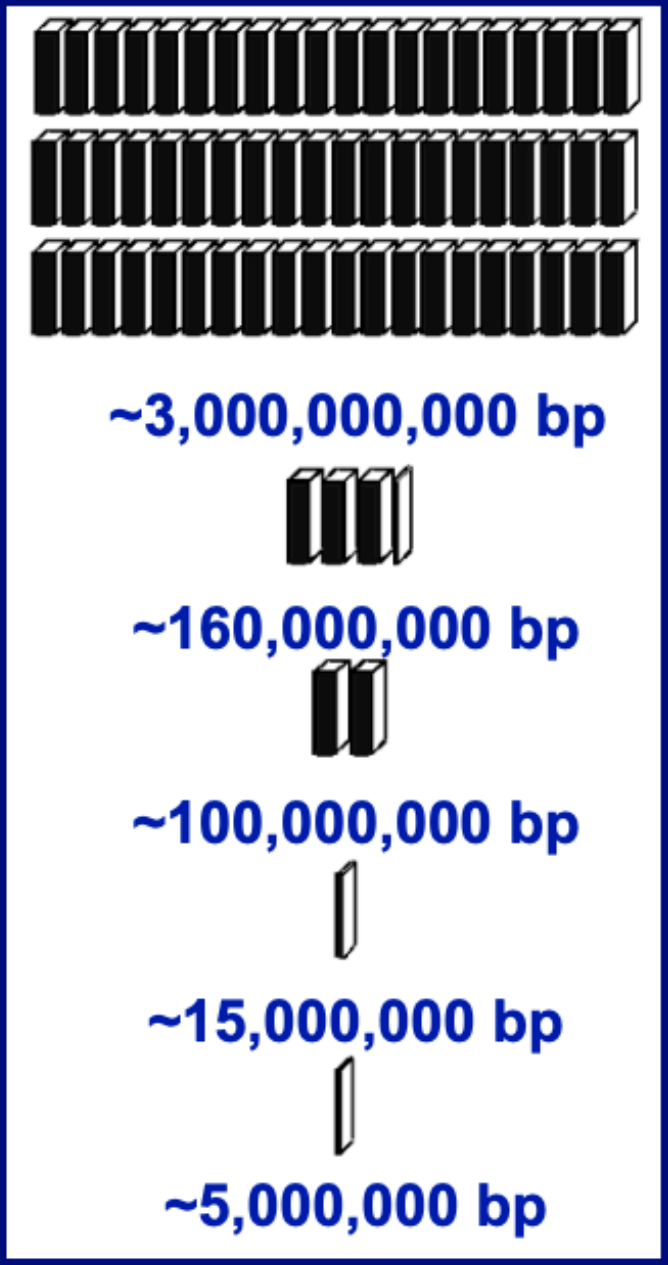
~5,000,000 bp

The Human Cytogenetic Map

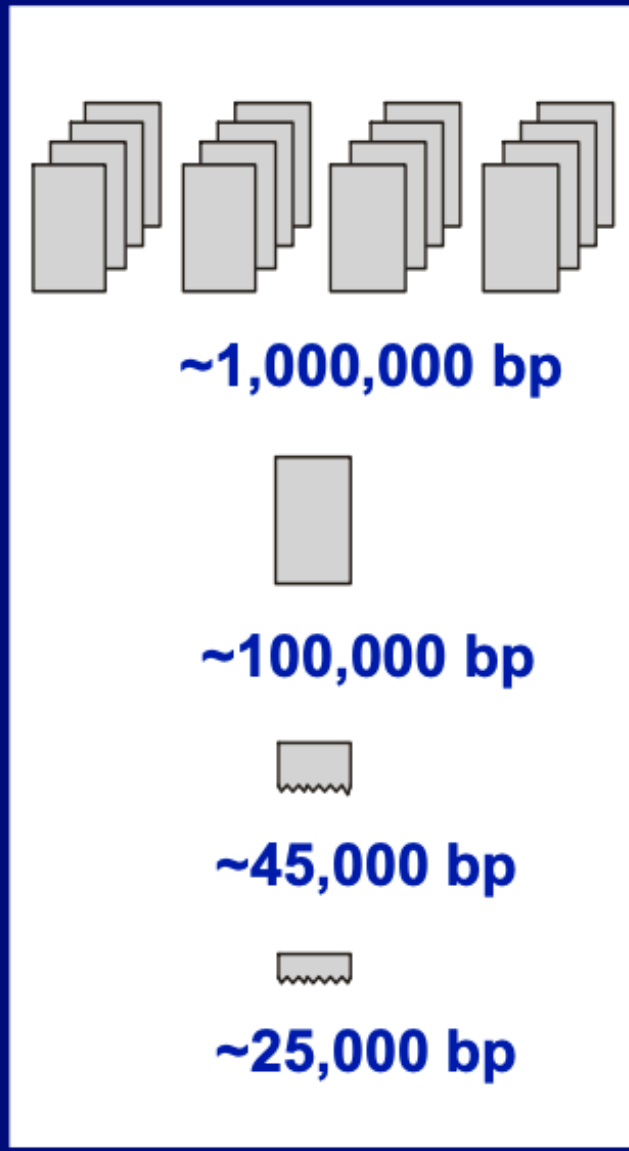


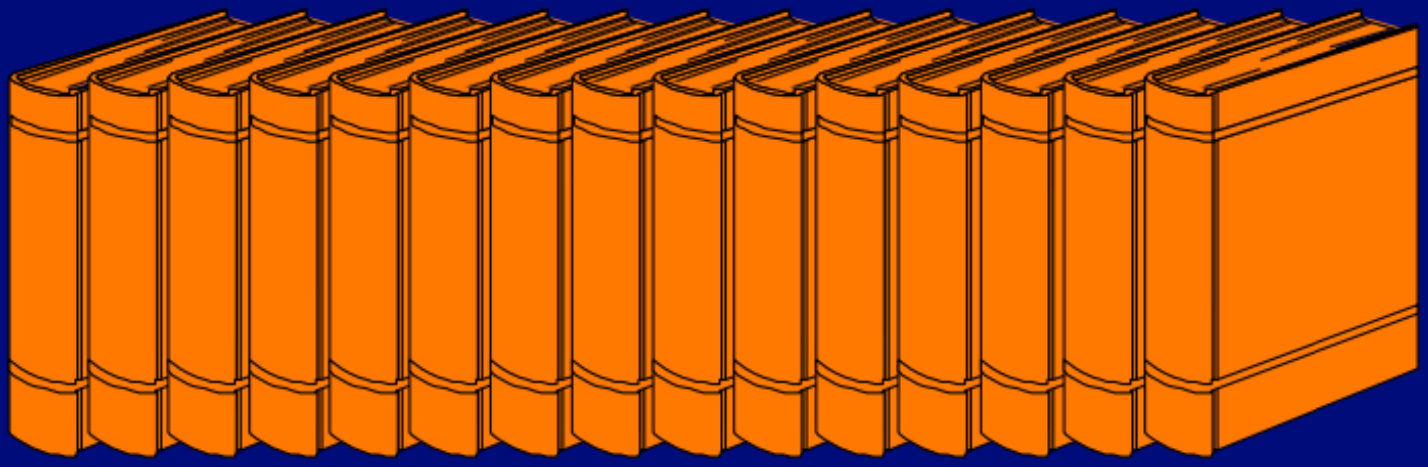

Clone-Based Physical Mapping
Chromosome
Clones
Contigs

Genome Sizes

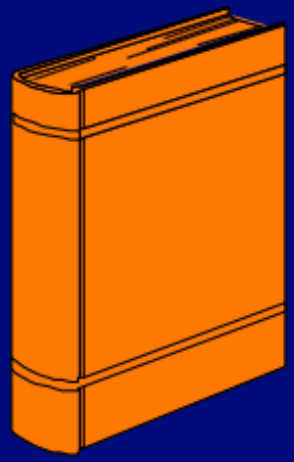


Cloning Capacity

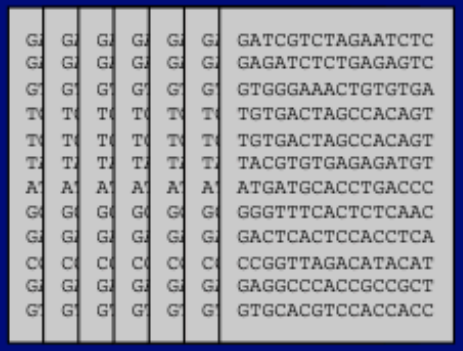




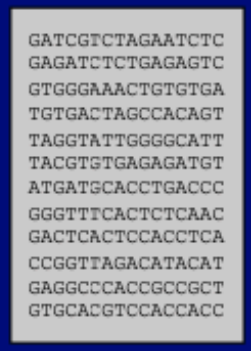
Genome
(~3000 Mb)



Chromosome
(~130 Mb)

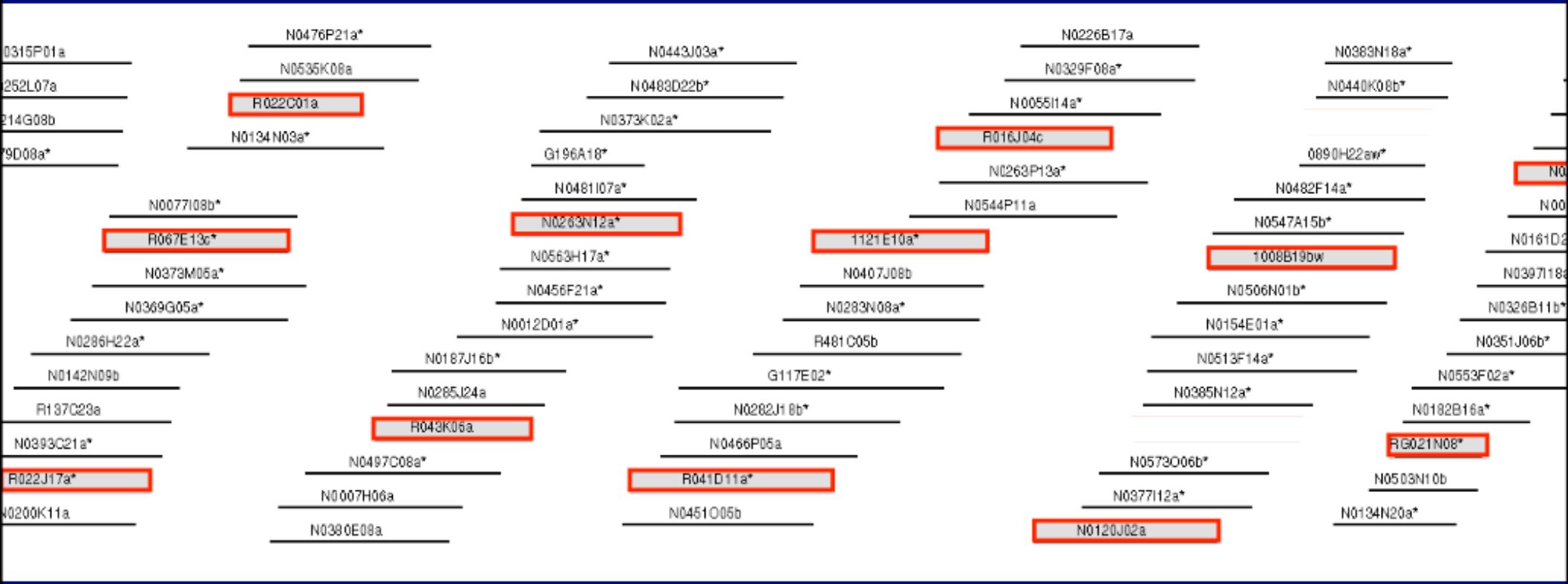


YAC
(~0.5-1.0 Mb)



BAC
(~0.1-0.2 Mb)

Sequence-Ready BAC Contig Map



JARED C. ROACH,*,† CECILIE BOYSEN,† KAI WANG,* AND LEROY HOOD*

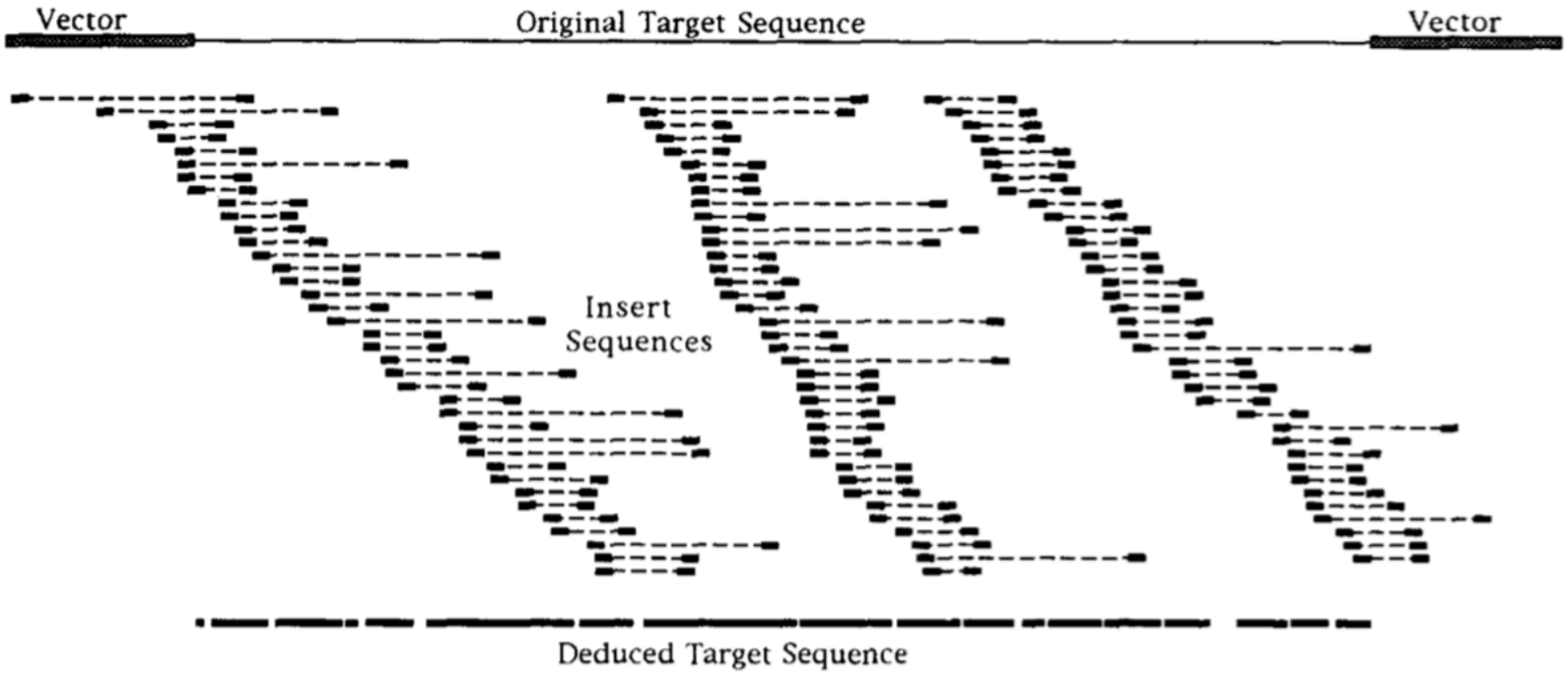
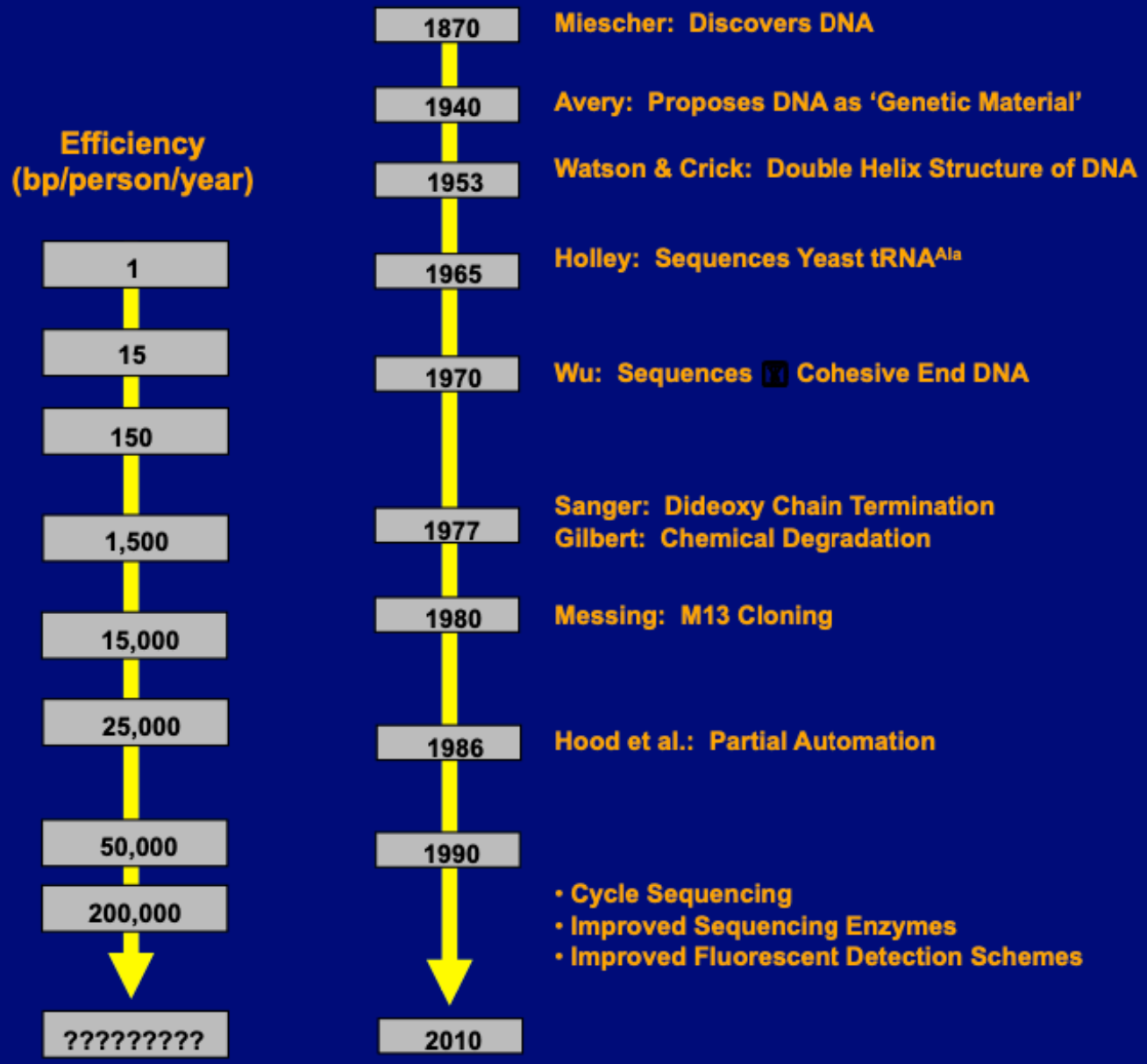


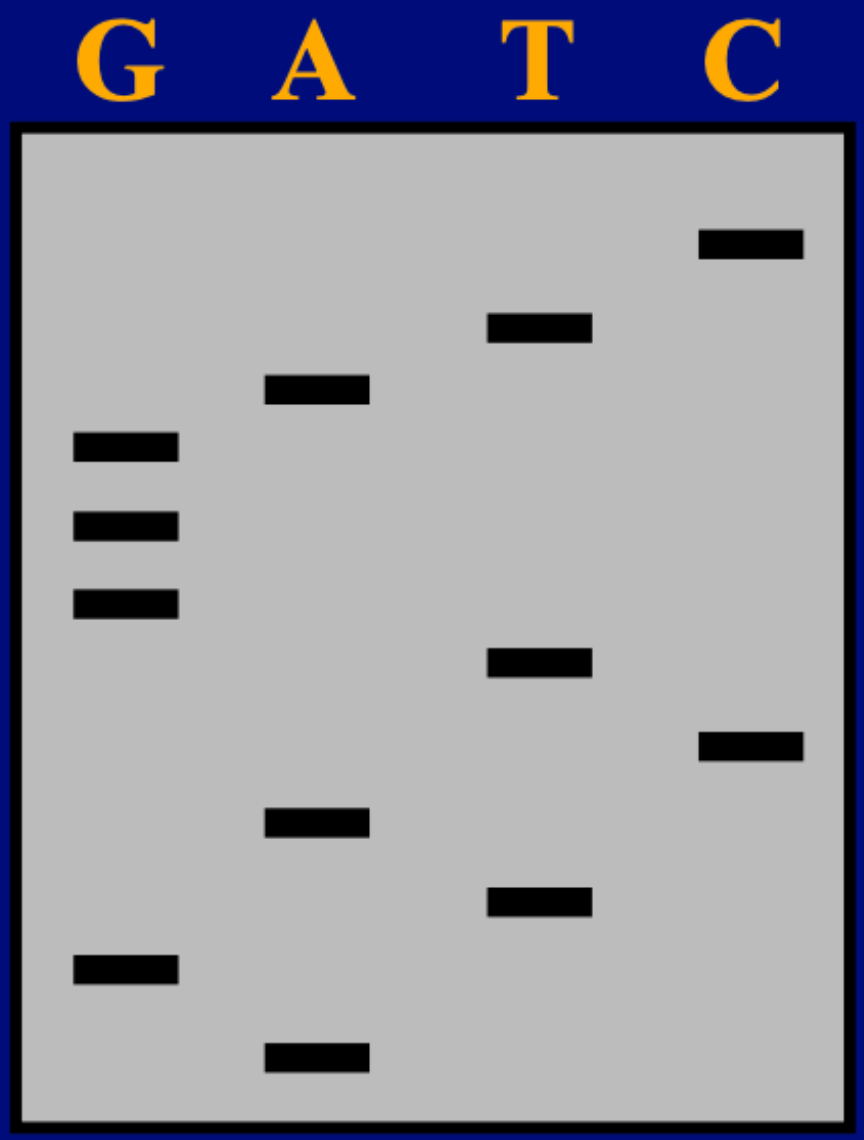
FIG. 1. A model "double-barrel shotgun" assembly. A 2.25 sequence redundancy produces 18 contigs that span 90% of an original target cosmid at 99.9% accuracy. Contig orientation and order are determined as shown. All but one gap are less than 400 bp; the remaining is 751 bp. More statistics are presented in Table 1.

DNA Sequencing

History of DNA Sequencing

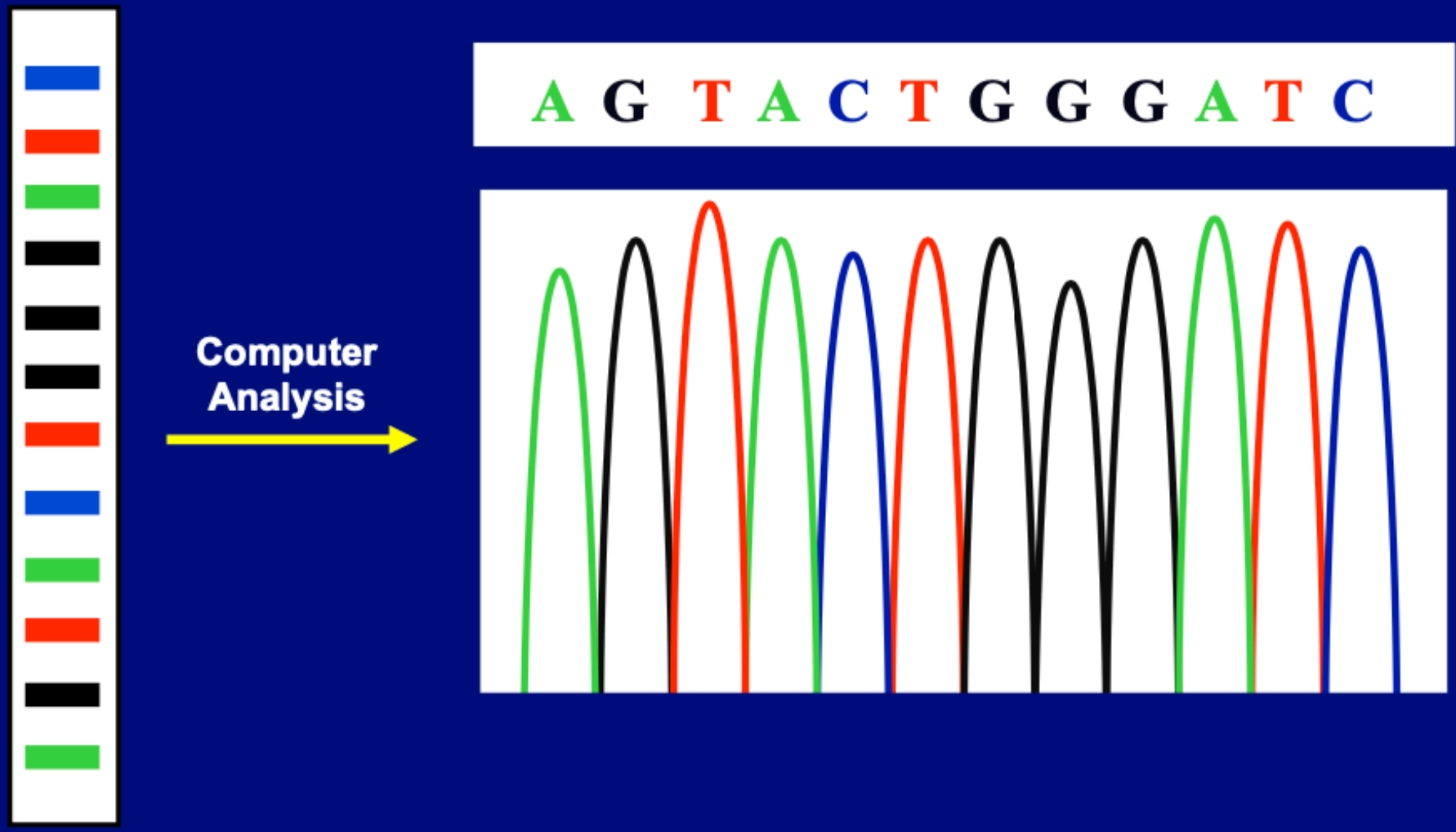


DNA Tagged with Radioactivity



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

Analyzing Fluorescent DNA Sequencing Data



Fluorescent DNA Sequencing Results

Analysis of Gene Expression

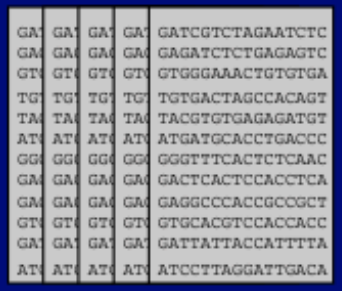
- **ESTs:** Expressed-Sequence Tags
- **SAGE:** Serial Analysis of Gene Expression
- **Full-Insert (Full-Length) cDNA Sequencing**

mgc.nci.nih.gov

Genome Sequencing

Shotgun Sequencing

Subclone Construction



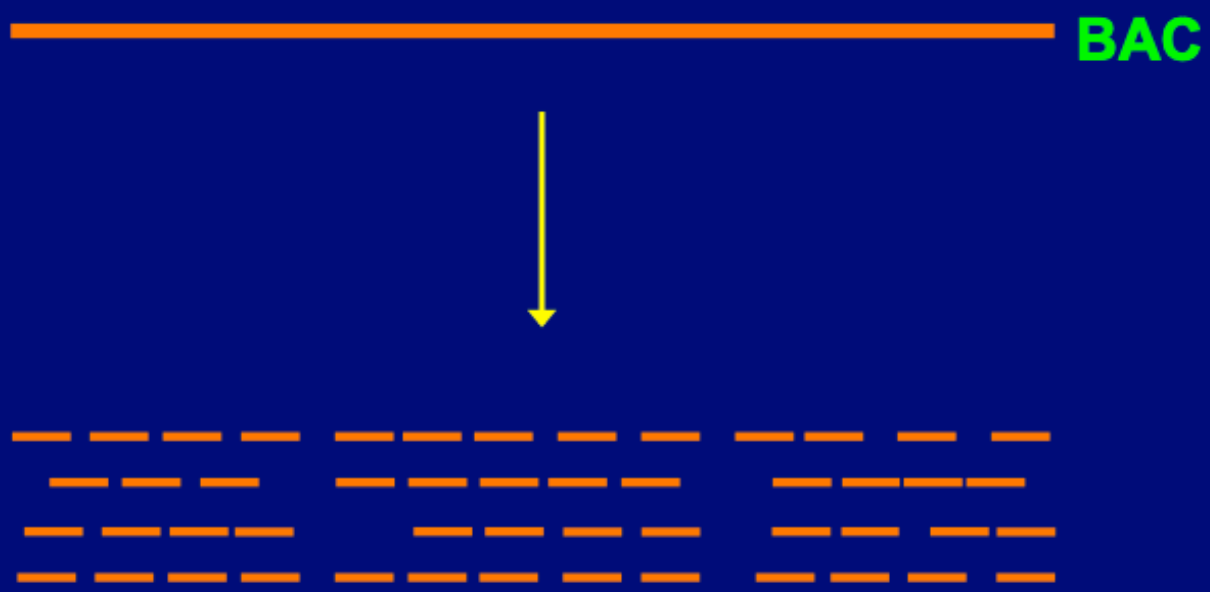
BAC DNA

Prepare Multiple Copies

Randomly Fragment

Subclone Fragments

Shotgun Sequencing Strategy



Poisson Calculations

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

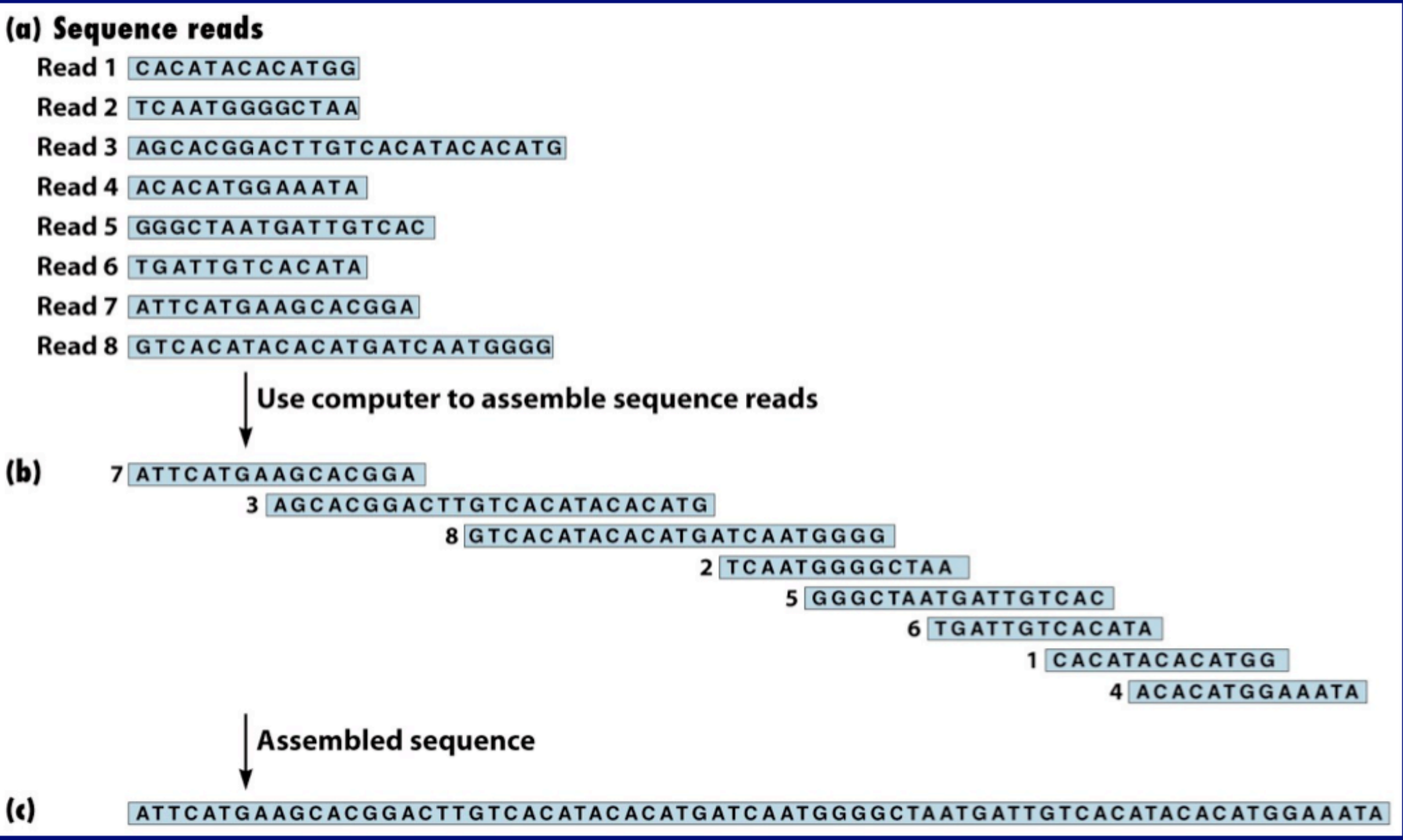
The probability a base is not sequenced is given by:

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

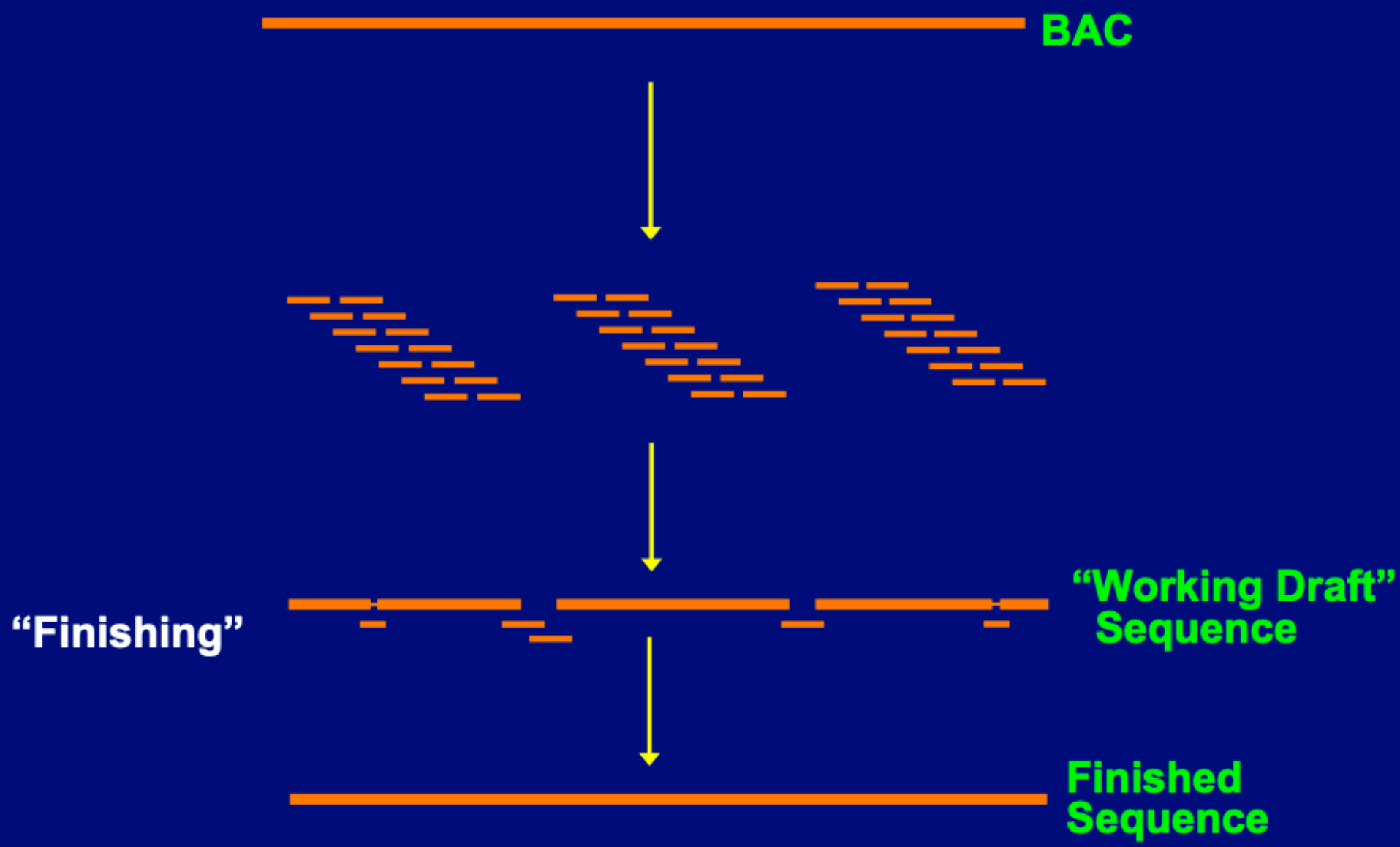
Where:

- 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%

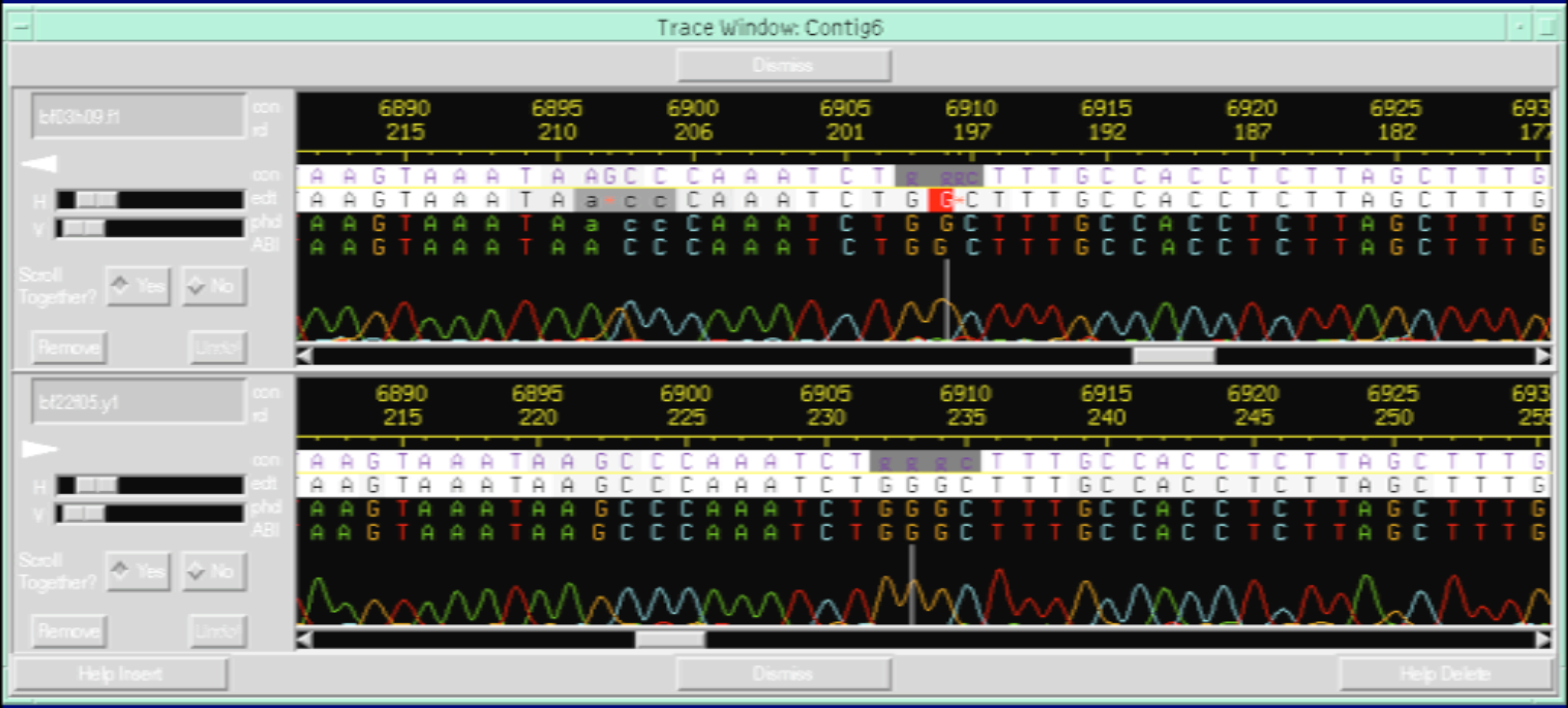

(a) Sequence reads
Read 1 CACATACACATGG
Read 2 TCAATGGGGCTAA
Read 3 AGCACGGACTTGTCACATACACATG
Read 4 ACACATGGAAATA
Read 5 GGGCTAATGATTGTCAC
Read 6 TGATTGTCACATA
Read 7 ATTCATGAAGCACGGA
Read 8 GTCACATACACATGATCAATGGGG
Use computer to assemble sequence reads
(b)
7 ATTCATGAAGCACGGA
3 AGCACGGACTTGTCACATACACATG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

Shotgun Sequencing Strategy




Trace Window: Contig32
Dismiss
yg02f10.y1
con
rd
H
V
con
edt
phd
ABI
Scroll Together?
Yes
No
Remove
Undo
8030
8035
8040
8045
8050
365
360
355
350
345
yg03g10.y1
454
449
444
439
434
yg18a10.y1
173
168
163
158
153
Help Insert
prev
Dismiss
next
Help Delete

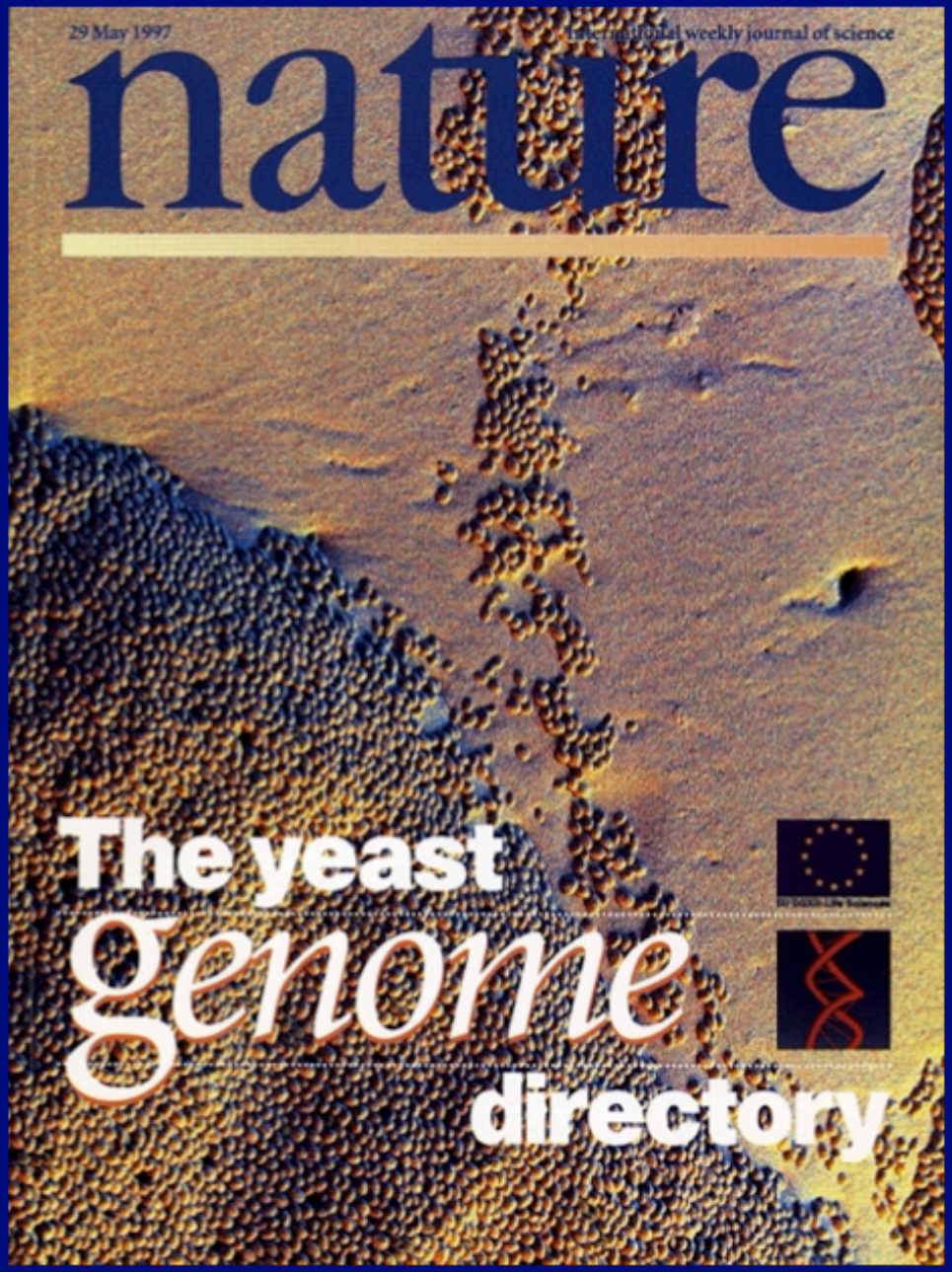
Sequence Finishing: Resolving Ambiguities



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

Historically Significant Genome Sequencing Projects

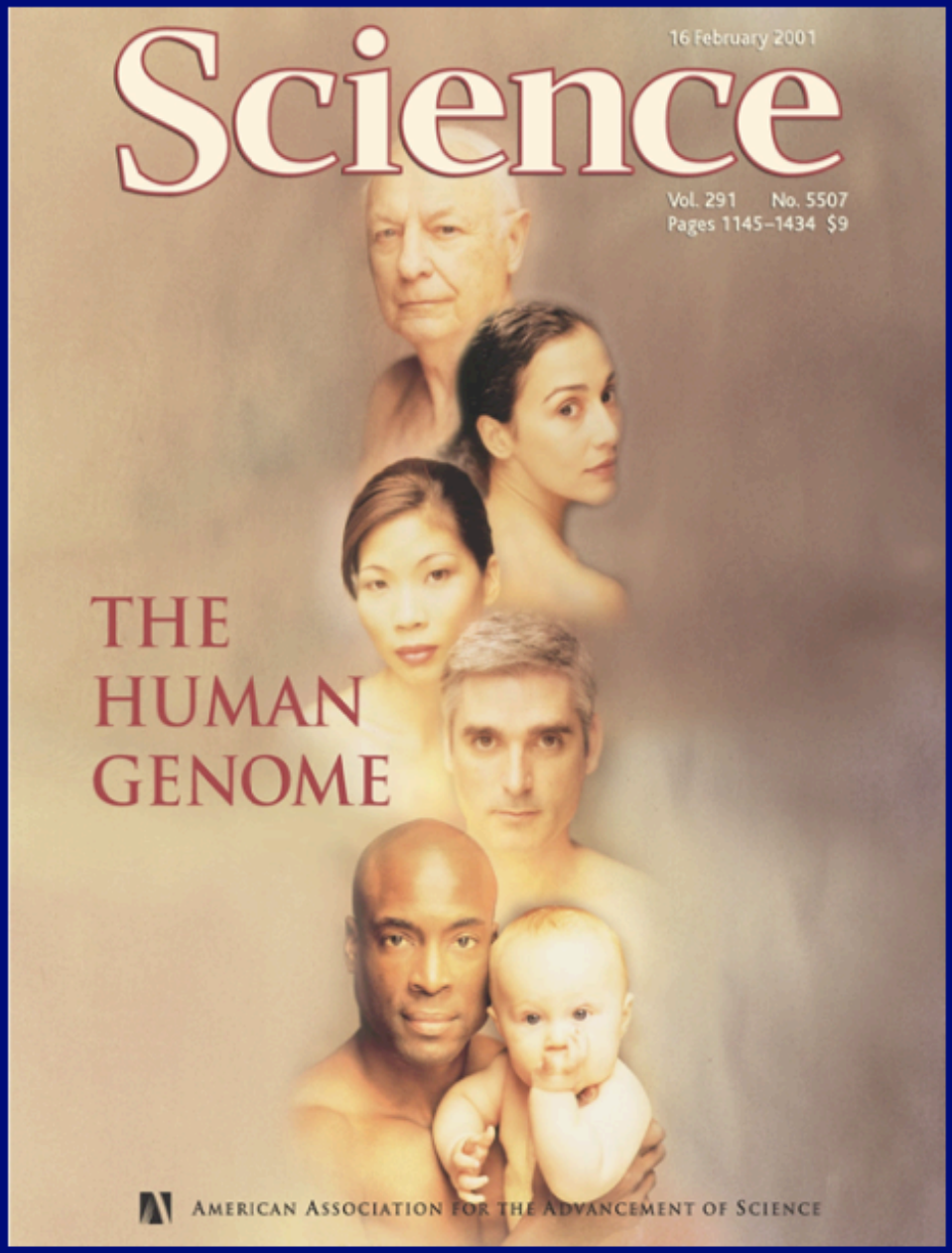
First Eukaryotic Genome Sequence



Goffeau et al. (1997)

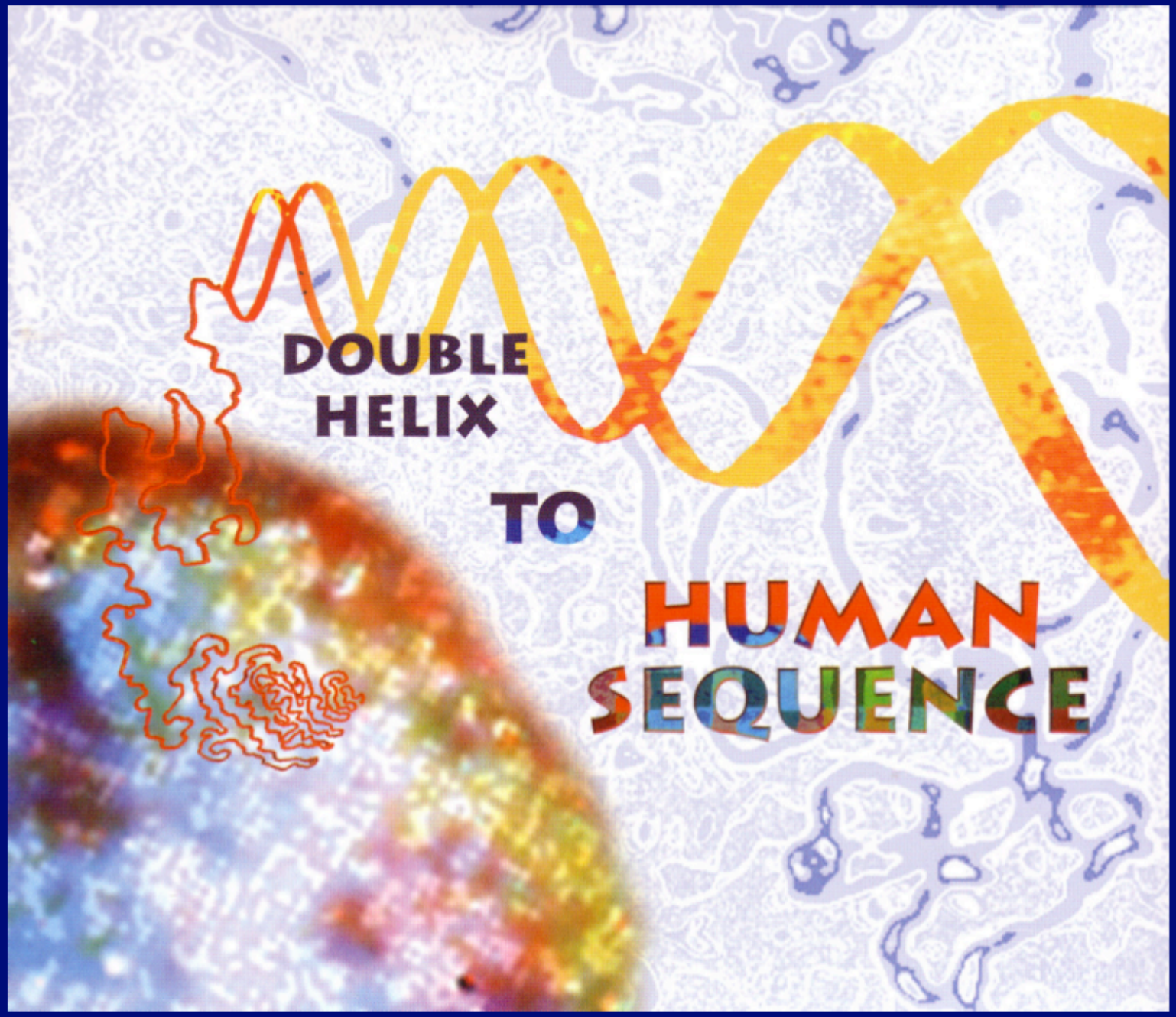
February, 2001 Draft Sequence

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*

* A list of authors and their affiliations appears in the Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human enome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

The Human Genome Project (HGP) was launched in 1990 with the goal of obtaining a highly accurate sequence of the vast majority of the euchromatic portion of the human genome. The initial work followed a two-pronged approach: (1) the mapping of the human and mouse genomes[1–9] to allow the study of inherited disease and provide a crucial scaffold for genome assembly; and (2) the sequencing of organisms with smaller, simpler genomes[10–18] to serve as a testbed for method development and assist in interpreting the human genome. With success along both paths, the sequencing of the human genome itself eventually became feasible. The International Human Genome Sequencing Consortium (IHGSC), an open collaboration involving twenty centres in six countries, was formed to carry out this component of the HGP.

In February 2001, the IHGSC[19] and Celera Genomics[20] each reported draft sequences providing a first overall view of the human genome. These sequences allowed systematic study of the human genome itself, including identification of genes, combinatorial architecture of proteins, regional differences in genome composition, distribution and history of transposable elements, distribution of polymorphism and relationship between genetic recombination and physical distance. Moreover, systematic knowledge of the human genome has enabled new tools and approaches that have markedly accelerated biomedical research.

Both draft sequences, however, had important shortcomings. The IHGSC sequence, for example, omitted ~10% of the euchromatic genome; it was interrupted by ~150,000 gaps; and the order and orientation of many segments within local regions had not been established. The IHGSC thus turned to the challenge of completing the sequence of the euchromatic genome. Operationally, a finished sequence was defined as having an error rate of, at most, one event per 10^4 bases, and the goal for completion was coverage in finished sequence of at least 95% of the euchromatic genome, with the only gaps being those refractory to all available techniques[21] (see http://www.genome.gov/10000923). The goal was challenging because the human genome is replete with such features as dispersed repeats and large segmental duplications, which greatly complicate the determination of genome structure and sequence. In fact, near-complete sequences have been obtained so far only for three multicellular organisms: the nematode[13], mustard weed[14] and the fruitfly[15]. These genomes are all roughly 30-fold smaller than the human genome and have much simpler structure.

We describe here the results of a multiyear effort by the IHGSC towards the goal of a complete human sequence. The number of gaps has been reduced 400-fold to only 341, most of which are associated with segmental duplications and will require new methods for resolution. The assembled near-complete genome sequence has an error rate of only ~1 event per 100,000 bases; it contains 2.85 billion nucleotides and covers ~99% of the euchromatic genome. This paper describes the current genome sequence and the process used to produce it; examines the accuracy and completeness of the sequence; and illustrates biological analyses made possible by the sequence. We do not attempt here a comprehensive analysis of the contents of the human genome. An initial analysis was previously reported[19] and a series of papers is being written describing the individual chromosomes[22–36], including annotation of genes and other features.

Current genome sequence

Finishing process

The process of converting the initial draft sequence into a near-complete sequence is referred to as 'finishing'. It is a complex iterative process that proceeds simultaneously at multiple scales, ranging from single nucleotides to the integrity of whole chromosomes. The fundamental challenge is that genomic regions that are not well represented or readily resolved through random shotgun sequencing tend to be highly enriched in problematic sequences. Resolving such regions required the development of special approaches, which evolved substantially over time and varied among centres.

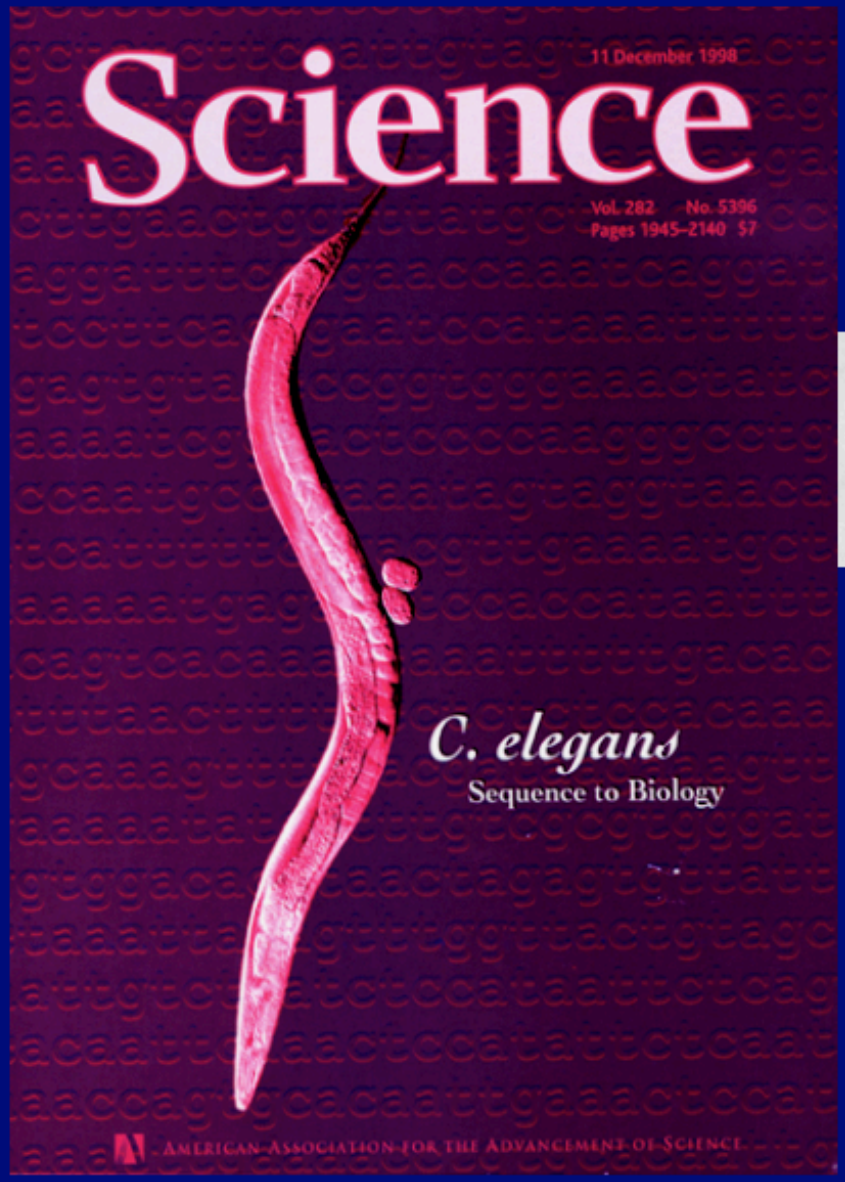
Broadly, the finishing process involved two distinct components: (1) producing finished maps, consisting of continuous and accurate paths of overlapping large-insert clones spanning the euchromatic region of each chromosome arm; and (2) producing finished clones, consisting of continuous and accurate nucleotide sequence across each large-insert clone. In practice, these two components were tightly intertwined in that progress in each often depended on results from the other. The components are described in Boxes 1 and 2. Further information about the finishing process and finishing standards can be found in the Supplementary Information (Note 1) and at http://www.genome.gov/10000923.

In total, we generated a shotgun sequence from 59,208 large-insert clones (total length ~5.84 gigabases (Gb)) and finished the sequence from 45,742 of these clones (total length ~3.67 Gb). The clones consisted primarily of bacterial artificial chromosomes

©2004 Nature Publishing Group

International Human Genome Sequencing Consortium (2004)

First Animal Genome Sequence

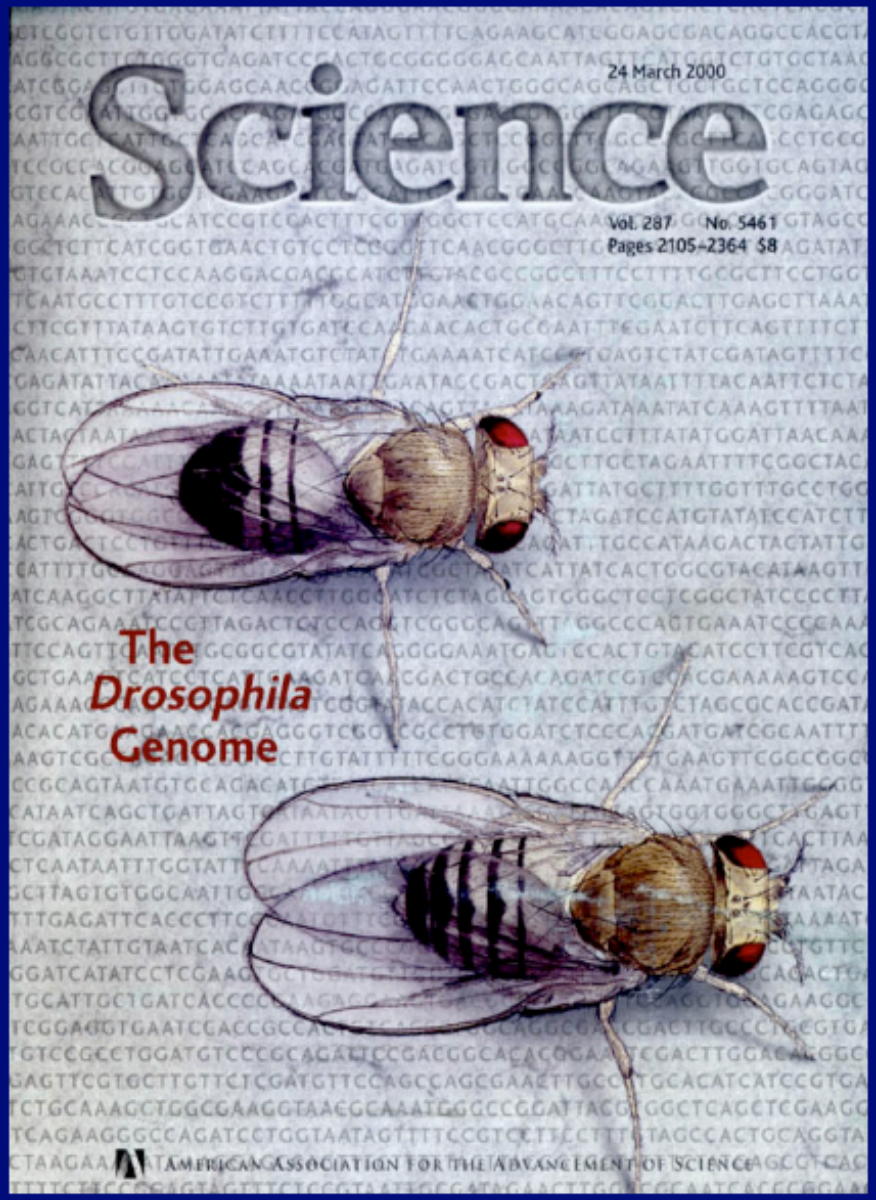


Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

The *C. elegans* Sequencing Consortium*

***C. elegans* Sequencing Consortium (1998)**

Second Animal Genome Sequence



THE *DROSOPHILA* GENOME

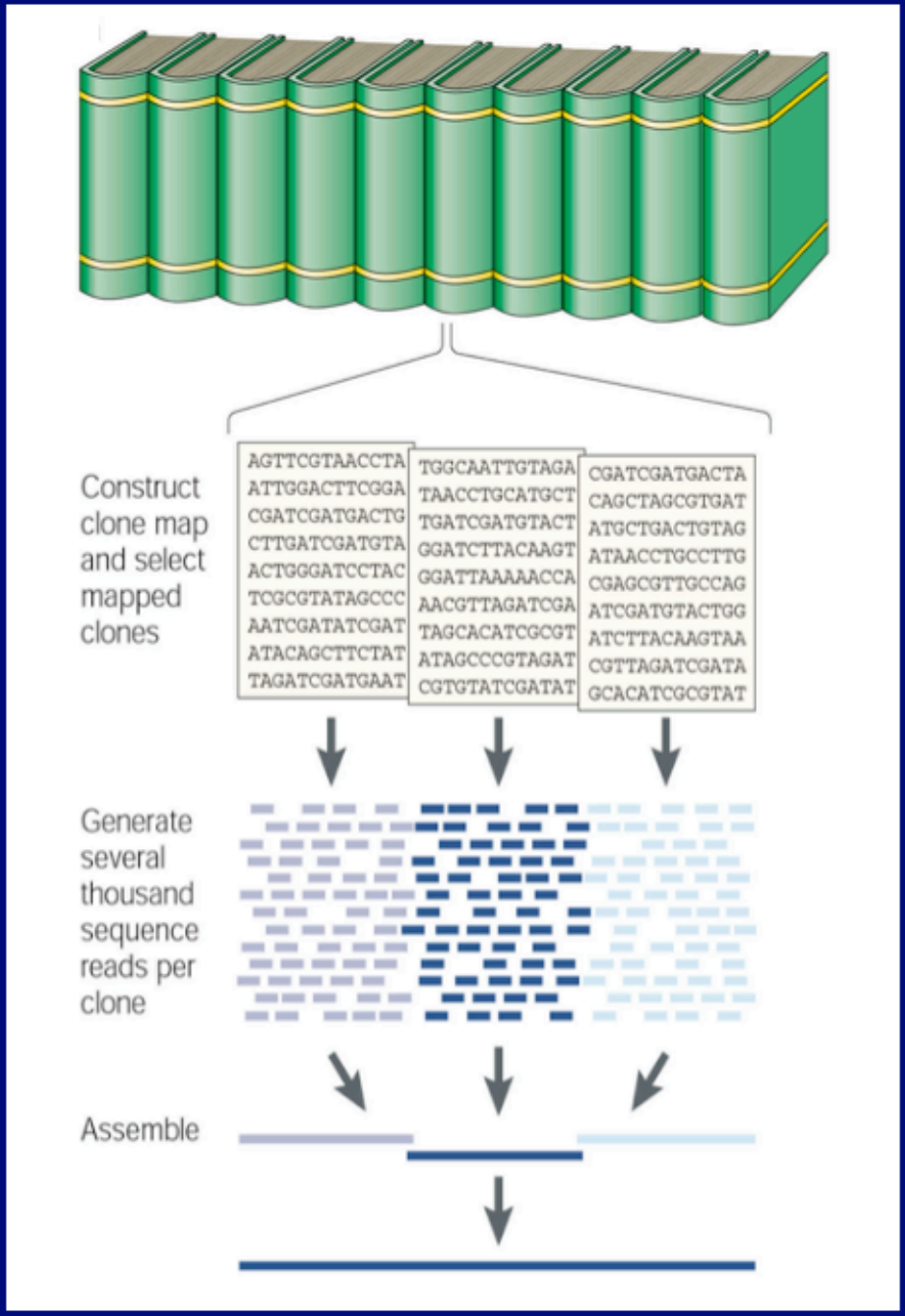
REVIEW

The Genome Sequence of *Drosophila melanogaster*

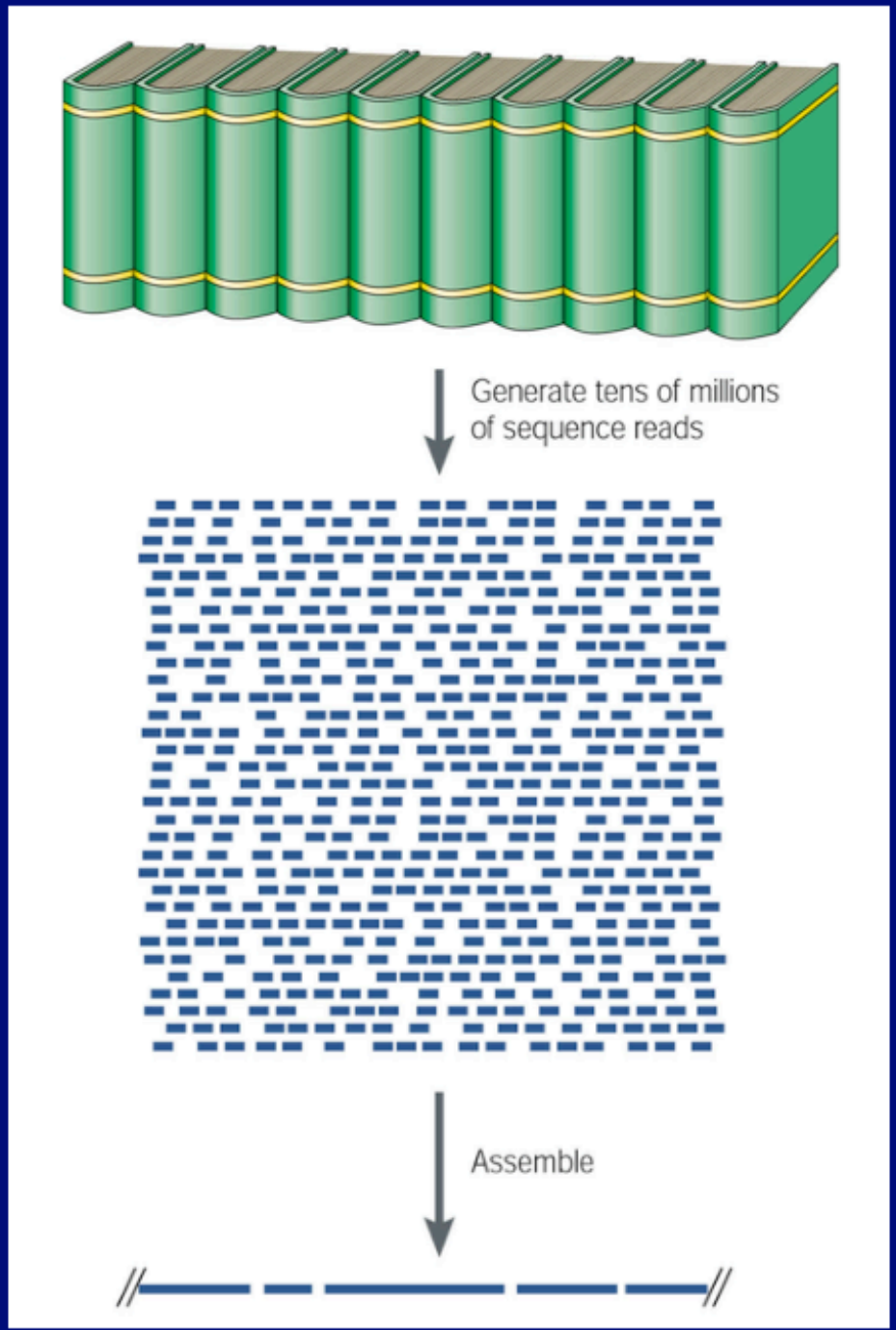
Mark D. Adams,[1*] Susan E. Celniker,[2] Robert A. Holt,[1] Cheryl A. Evans,[1] Jeannine D. Gocayne,[1] Peter G. Amanatides,[1] Steven E. Scherer,[3] Peter W. Li,[1] Roger A. Hoskins,[2] Richard F. Galle,[2] Reed A. George,[2] Suzanna E. Lewis,[4] Stephen Richards,[2] Michael Ashburner,[5] Scott N. Henderson,[1] 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,[4] 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,[3] Peter Brokstein,[4] 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,[1] 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,[9] Ken Glasser,[1] Anna Glodek,[1] Fangcheng Gong,[1] J. Harley Gorrell,[3] 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,[4] Knut Reinert,[1] Karin Remington,[1] Robert D. C. Saunders,[30] Frederick Scheeler,[1] Hua Shen,[3] Bixiang Christopher Shue,[1] Inga Sidén-Kiamos,[11] 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] 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,[19] 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,[3] 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]

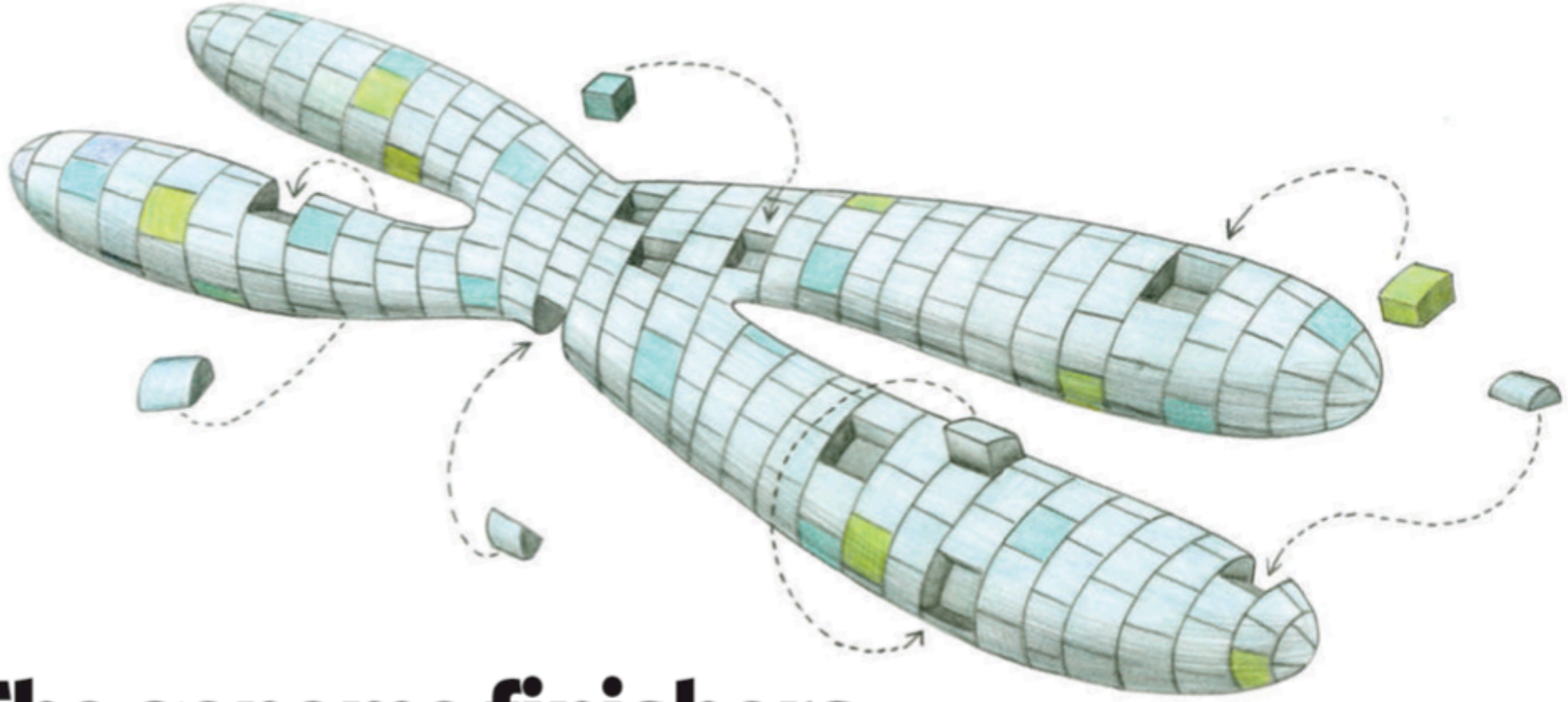
Adams et al. (2000)

Clone-Based Shotgun Sequencing



Whole-Genome Shotgun Sequencing



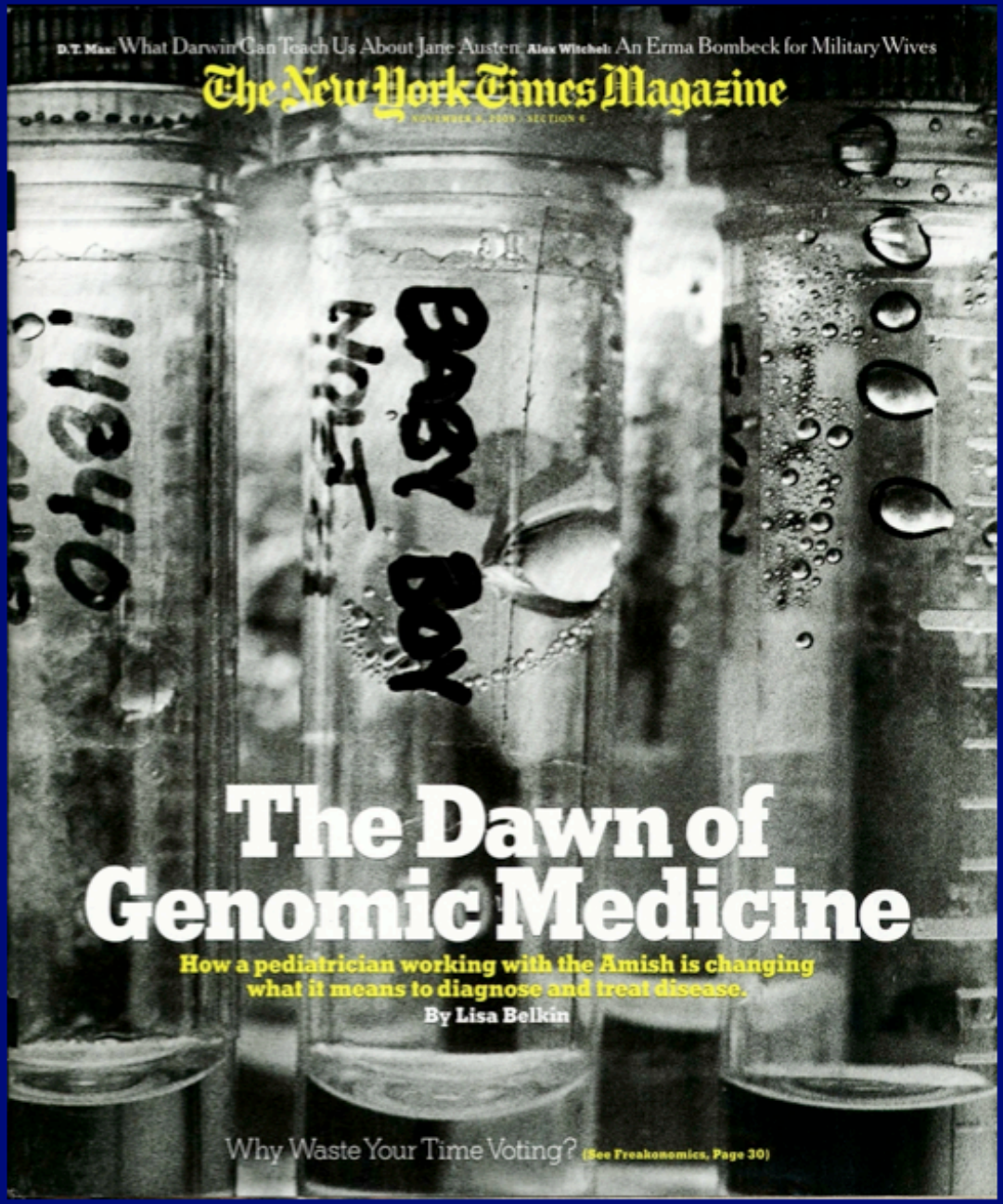


The genome finishers

Dedicated scientists are working hard to close the gaps, fix the errors and finally complete the human genome sequence. **Elie Dolgin** looks at how close they are.

***Nature* (2009)**

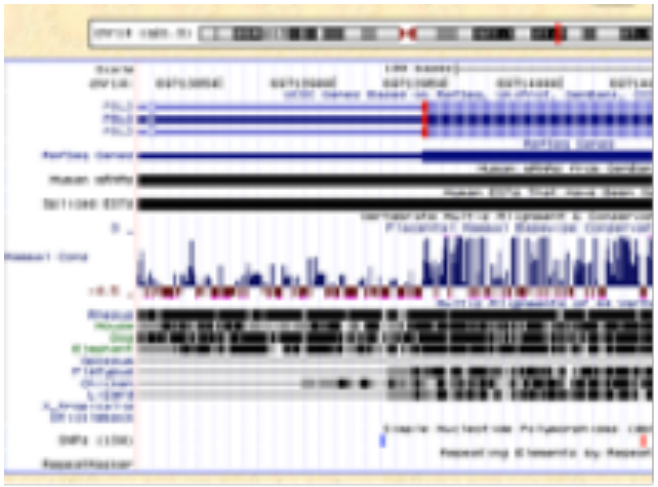
Human Genome Project

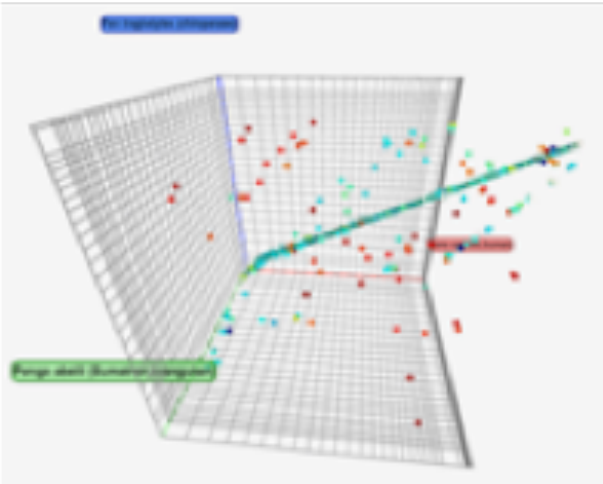

D.T. Max: What Darwin Can Teach Us About Jane Austen Alex Witchel: An Erma Bombeck for Military Wives
The New York Times Magazine
The Dawn of Genomic Medicine
How a pediatrician working with the Amish is changing what it means to diagnose and treat disease.
By Lisa Belkin
Why Waste Your Time Voting? (See Freakonomics, Page 30)


Science
24 October 2003
Vol. 302 No. 5645
Pages 517–728 $10
Genomic Medicine
American Association for the Advancement of Science

Computational Genomics

The Carbon and Clarke Formula





[27] Selection of Specific Clones from Colony Banks by Suppression or Complementation Tests

By LOUISE CLARKE and JOHN CARBON

METHODS IN ENZYMOLOGY, VOL. 68

Copyright © 1979 by Academic Press, Inc.
All rights of reproduction in any form reserved.
ISBN 0-12-181968-X

We have determined the transformant colony bank size needed to obtain a plasmid collection representing 90–99% of the *E. coli* or yeast genome as follows.[2] Given a preparation of cell DNA fragmented to a size such that each fragment represents a fraction (f) of the total genome, the probability (p) that a given unique DNA sequence is present in a collection of N transformant colonies is given by the expression

$$P = 1 - (1-f)^N$$

or

$$N = \ln (1\text{-}P)/\ln (1\text{-}f)$$

A sample calculation for *E. coli* (genome size, 2.7×10^9 daltons) for $P = 0.99$ is

$$N = \frac{\ln (1\text{-}0.99)}{\ln [1 - (8.5 \times 10^6/2.7 \times 10^9)]} = 1437$$

Thus, using a preparation of DNA randomly sheared to an average size of 8.5×10^6 daltons for the construction of annealed hybrid circular DNA, a colony bank of only about 1400 transformants for *E. coli* or 5400 transformants for yeast is adequate to give a probability of 99% that any *E. coli* or yeast gene will be on a hybrid plasmid in one of the clones.

$$N = \frac{ln(1-P)}{ln(1-f)}$$

where,

N is the necessary number of recombinants[16]

P is the desired probability that any fragment in the genome will occur at least once in the library created

f is the fractional proportion of the genome in a single recombinant

f can be further shown to be:

$$f = \frac{i}{g}$$

where,

i is the insert size

g is the genome size

Thus, increasing the insert size (by choice of vector) would allow for fewer clones needed to represent a genome. The proportion of the insert size versus the genome size represents the proportion of the respective genome in a single clone.[14] Here is the equation with all parts considered:

$$N = \frac{ln(1-P)}{ln(1-\frac{i}{g})}$$

Vector selection example [edit]

The above formula can be used to determine the 99% confidence level that all sequences in a genome are represented by using a vector with an insert size of twenty thousand basepairs (such as the phage lambda vector). The genome size of the organism is three billion basepairs in this example.

$$N = \frac{ln(1-0.99)}{ln[1-\frac{2.0\times10^4 basepairs}{3.0\times10^9 basepairs}]}$$

$$N = \frac{-4.61}{-6.7\times10^{-6}}$$

$$N = 688,060 \text{ clones}$$