Human Genome Project

Prof.B.Vaseeharan
Department of Animal Health and Management, Alagappa University

History of the Human Genome Project

Introduction

- Until the early 1970's, DNA was the most difficult cellular molecule for biochemists to analyze.
- DNA is now the easiest molecule to analyze – we can now isolate a specific region of the genome, produce a virtually unlimited number of copies of it, and determine its nucleotide sequence overnight.

Molecular Biology Of The Cell. Alberts et al. 491-495

- At the height of the Human Genome Project, sequencing factories were generating DNA sequences at a rate of 1000 nucleotides per second.
- Technical breakthroughs that allowed the Human Genome Project to be completed have had an enormous impact on all of biology…..

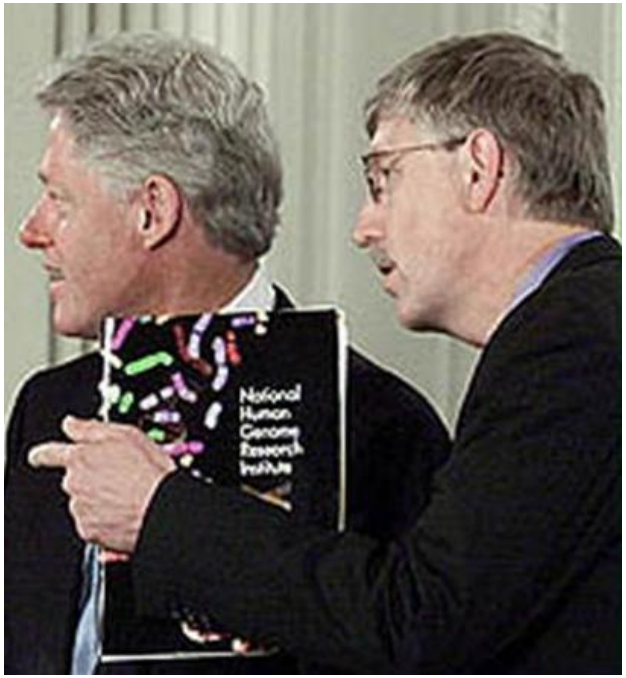
Molecular Biology Of The Cell. Alberts et al. 491-495

The Human Genome Project Began in 1990

- The Mission of the HGP: The quest to understand the human genome and the role it plays in both health and disease.

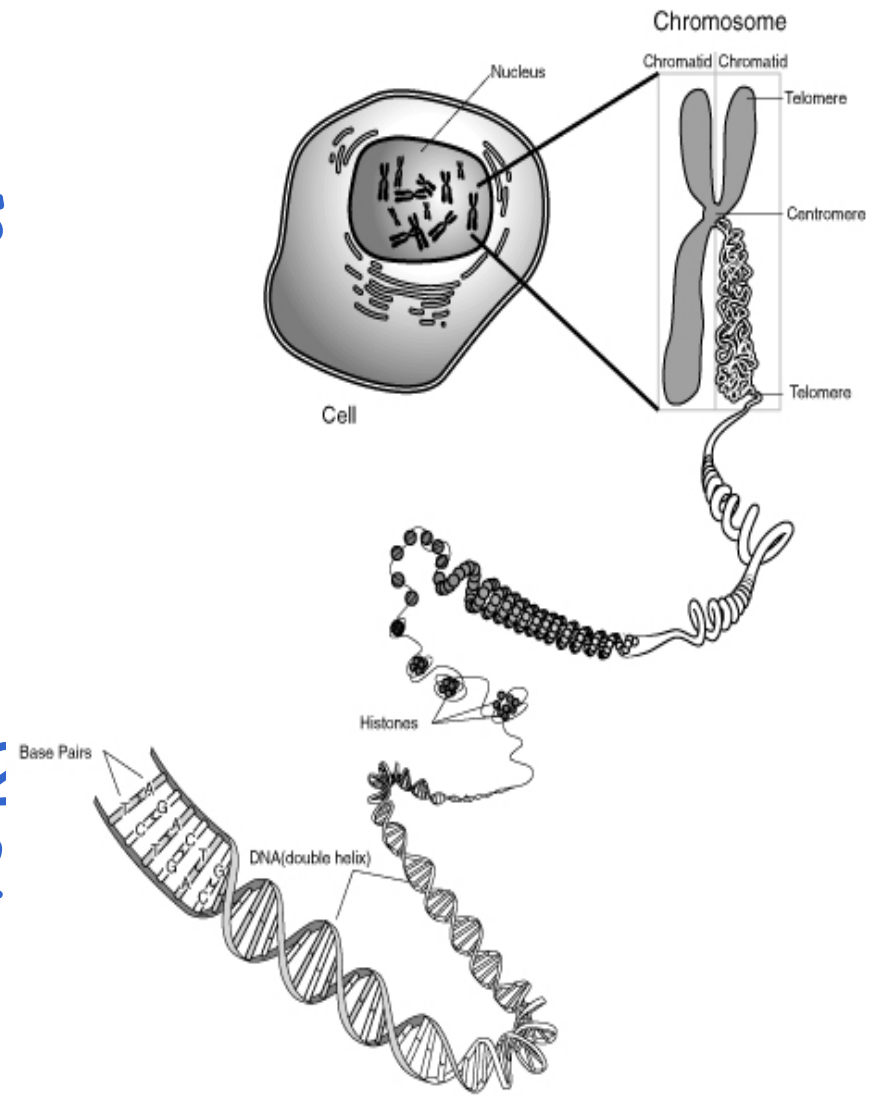
"The true payoff from the HGP will be the ability to better diagnose, treat, and prevent disease."

--- Francis Collins, Director of the HGP and the National Human Genome Research Institute (NHGRI)



The genome is our Genetic Blueprint

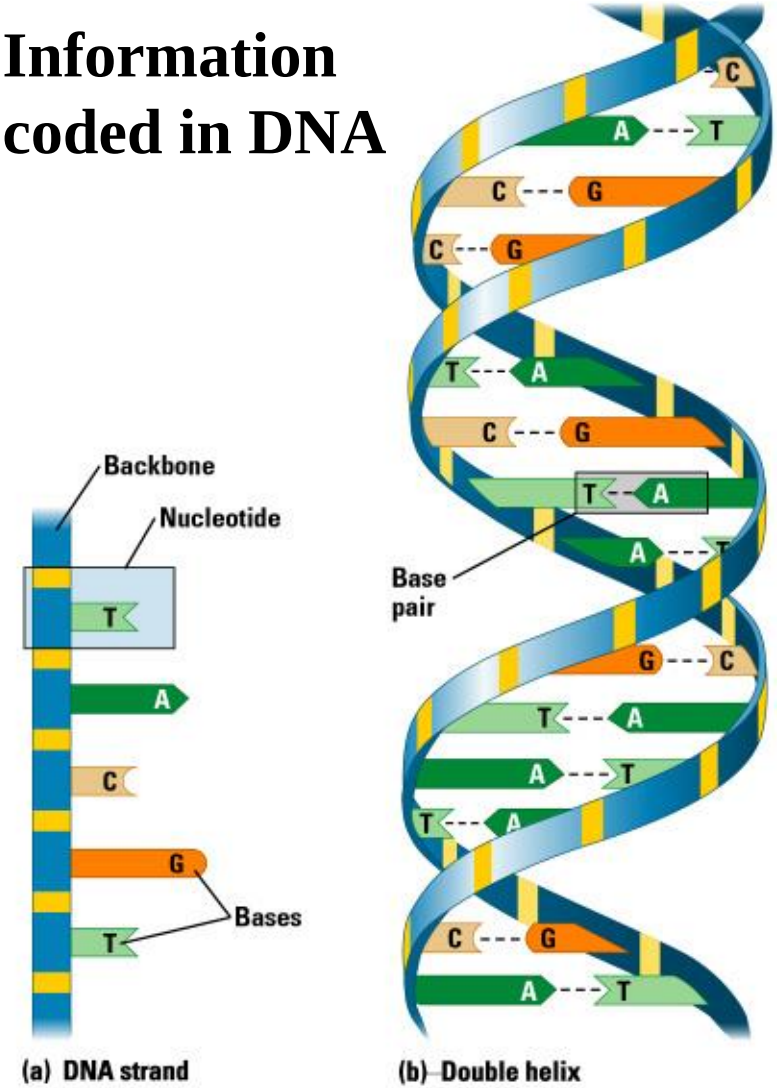
- Nearly every human cell contains 23 pairs of chromosomes
  - 1 - 22 and XY or XX
    - XY = Male
    - XX = Female
- Length of chr 1-22, X, Y together is ~3.2 billion bases (about 2 meters diploid)



The Genome is Who We Are on the inside!

- Chromosomes consist of DNA
  - molecular strings of A, C, G, & T
  - base pairs, A-T, C-G
- Genes
  - DNA sequences that encode proteins
  - less than 3% of human genome

Information coded in DNA



5000 bases per page

CA[redacted]GCATGTGAGAGCTTCTAATATCTAAATTAATGTTGAATCATTATTCAGAAACAGAGAGCTAACTGTTATCCCATCCTGACTTTATTCTTTATG AGAAAAATACAGTGATTCC
AA[redacted]CAAGTTAGTGCTGCTTGCTTTATAAATGAAGTAATATTTTAAAAGTTGTGCATAAGTTAAAATTCAGAAATAAAACTTCATCCTAAAACTCTGTGTGTTGCTTTAAATAATC
AGAGCATCTGC TACTTAATTTTTTGTGTGTGGGTGCACAATAGATGTTTAATGAGATCCTGTCATCTGTCTGCTTTTTTATTGTAAAACAGGAGGGGTTTTAATACTGGAGGAACAA
CTGATGTACCTCTGAAAAGAGA AGAGATTAGTTATTAATTGAATTGAGGGTTGTCTTGTCTTAGTAGCTTTTATTCTCTAGGTACTATTTGATTATGATTGTGAAAATAGAATTTATCC
CTCATTAAATGTAAAATCAACAGGAGAATAGCAAAAACTTATGAGATAGATGAACGTTGTGTGAGTGGCATGGTTTAATTTGTTTGGAAGAAGCACTTGCCCCAGAAGATACACAAT
GAAATTCATGTTATTGAGTAGAGTAGTAATACAGTGTGTTCCCTTGTGAAGTTCATAACCAAGAATTTTAGTAGTGGATAGGTAGGCTGAATAACTGACTTCCTATC ATTTTCAGGTT
CTGCGTTTGATTTTTTTTACATATTAATTTCTTTGATCCACATTAAGCTCAGTTATGTATTTCCATTTTATAAATGAAAAAAAATAGGCACTTGCAAATGTCAGATCACTTGCCTGTGGT
CATTCGGGTAGAGATTTGTGGAGCTAAGTTGGTCTTAATCAAATGTCAAGCTTTTTTTTTTCTTATAAAATATAGGTTTTAATATGAGTTTTAAAATAAAATTAATTAGAAAAAGGCAA
ATTACTCAATATATATAAGGTATTGCATTTGTAATAGGTAGGTATTTCATTTTCTAGTTATGGTGGGATATTATTCAGACTATAATTCCCAATGAAAAAACTTTAAAAAATGCTAGTGA
TTGCACACTTAAAACACCTTTTAAAAAGCATTGAGAGCTTATAAAATTTTAATGAGTGATAAAACCAAATTTGAAGAGAAAAGAAGAACCCAGAGAGGTAAGGATATAACCTTACC
AGTTGCAATTTGCCGATCTCTACAAATATTAATATTTATTTTGACAGTTTCAGGGTGAATGAGAAAGAAACCAAAACCCAAGACTAGCATATGTTGTCTTCTTAAGGAGCCCTCCCCT
AAAAGATTGAGATGACCAAATCTTATACTCTCAGCATAAGGTGAACCAGACAGACCTAAAGCAGTGGTAGCTTGGATCCACTACTTGGGTTTGTGTGTGGCGTGACTCAGGTAATCT
CAAGAATTGAACATTTTTTTAAGGTGGTCCTACTCATACACTGCCCAGGTATTAGGGAGAAGCAAATCTGAATGCTTTATAAAAATACCCTAAAGCTAAATCTTACAATATTCTCAAG
AACACAGTGAA ACAAGGCAAAATAAGTTAAAATCAACAAAAACAACATGAAACATAATTAGACACACAAAGACTTCAAACATTGGAAAATACCAGAGAAAGATAATAAATAT
TTTACTCTTTAAAAATTTAGTTAAAAGCTTAAACTAATTGTAGAGAAAA AACTATGTTAGTATTATATTGTAGATGAAATAAGCAAAACATTTAAAATACAAATGTGATTACTTAAAT
TAAATATAATAGATAATTTACCACCAGATTAGATACCATTGAAGGAATAATTAATATACTGAAATACAGGTCAGTAGAATTTTTTTCAATTCAGCATGGAGATGTAAAAAATGAAAA
TTAATGCAAAAAATAAGGGCACAAAAAGAAATGAGTAATTTTGATCAGAAATGTATTAAAATTAATAAACTGGAAATTTGACATTTAAAAAAAGCATTGTCATCCAAGTAGATGTG
TCTATTAAATAGTTGTTCTCATATCCAGTAATGTAATTATTATTCCCTCTCATGCAGTTCAGATTCTGGGGTAATCTTTAGACATCAGTTTTGTCTTTTATATTATTTATTCTGTTTACTAC
[redacted]TTGCTAATGATATTTTTAATTTCTGACATTCTGGAGTATTGCTTGTAAAAGGTATTTTTAAAAATACTTTATGGTTATTTTTGTGATTCCTATTCCTCTATGGACACCAAGGCT
[redacted]TTTCTTTGGTTTCTTCTGTTACTTCTATTTTCTTAGTGTTTATATCATTTCATAGATAGGATATTCTTTATTTTTTATTTTTATTTAAATATTTGGTGATTCTTGGTTTTCTCAGCC
[redacted]TCAAGTGTTCTTATTAAGCATTATTATTAAATAAAGATTATTTCCTCTAATCACATGAGAATCTTTATTTCCCCCAAGTAATTGAAAATTGCAATGCCATGCTGCCATGTGG
TACAGCATGGGTTTGGGCTTGCTTTCTTCTTTTTTTTTTAACTTTTATTTTAGGTTTGGGAGTACCTGTGAAAGTTTGTTATATAGGTAAACTCGTGTCACCAGGGTTTGTTGTACAGATCA
TTTTGTCACCTAGGTACCAAGTACTCAACAATTATTTTTCCTGCTCCTCTGTCTCCTGTCACCCTCCACTCTCAAGTAGACTCCGGTGTCTGCTGTTCCATTCTTTGTGTCCATGTGTTCTC

How much data make up the human genome?

- 3 pallets with 40 boxes per pallet x 5000 pages per box x 5000 bases per page = 3,000,000,000 bases!
- To get accurate sequence requires 6-fold coverage.
- Now: Shred 18 pallets and reassemble.

The Beginning of the Project

- Most the first 10 years of the project were spent improving the technology to sequence and analyze DNA.
- Scientists all around the world worked to make detailed maps of our chromosomes and sequence model organisms, like worm, fruit fly, and

Human Genome Project, 1993 Revised Goals

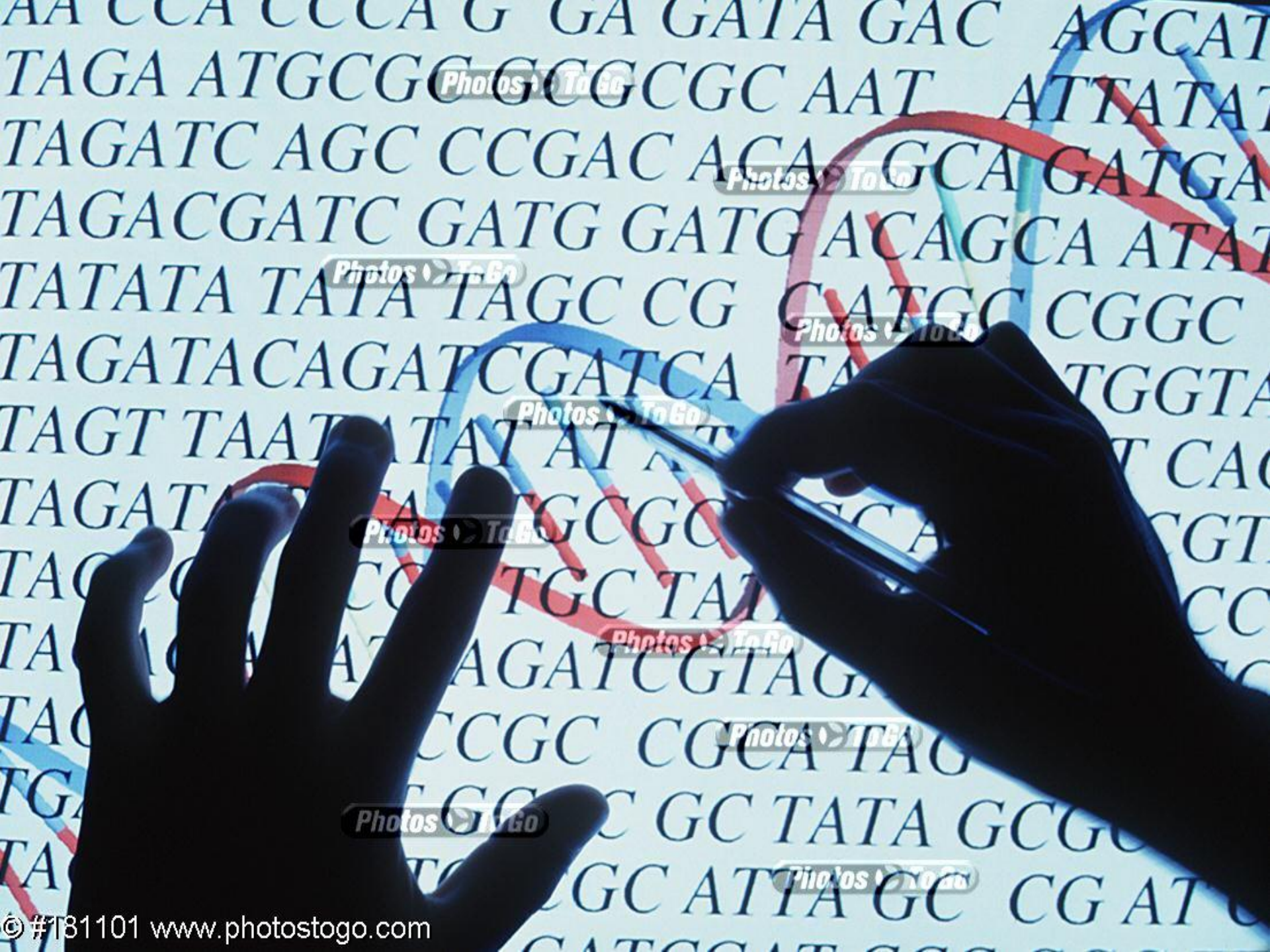
- Revised because of rapid progress
  - Automated DNA sequencing technology
  - Genetic markers could be assayed using PCR
  - Better cloning vectors for large genomes
  - Better computational methods for genome assembly
- Greater focus on genes (<1% of genome)
- Successful international collaboration

Human Genome Project, 1998 New Five Year Plan

- Finish complete human genome sequence by 2003 (50th anniversary of double helix by Watson and Crick)
  - Draft sequence finished in July, 2000
  - 'Complete' sequence to be published in 2001

Human Genome Project, 1998 New Five Year Plan

- Complete sequence of *C elegans* by 1998
  - Published in 1998
- Complete sequence of *D melanogaster* by 2002
  - Published in 2000
- Complete *M musculus* genome sequence by 2005
  - Published in 2002 (draft) – also rice, rat in draft


AA CCA CCCA G GA GATA GAC AGCAT
TAGA ATGCGC GCGCGC AAT ATTATAT
TAGATC AGC CCGAC ACA GCA GATGA
TAGACGATC GATG GATG ACAGCA ATAT
TATATA TATA TAGC CG CATGC CGGC
TAGATACAGATCGATCA TGGTA
TAGT TAAT ATAT AT
TAGAT CAG
TAG TGC TA
TA AGATCGTAGA
TAG CCGC CGCA TAG
TGA GC TATA GCGC
TA GC ATTA GC CG AT
© #T81101 www.photostogo.com

Goals:

- identify all the approximate 30,000 genes in human DNA,
- determine the sequences of the 3 billion chemical base pairs that make up human DNA,
- store this information in databases,
- improve tools for data analysis,
- transfer related technologies to the private sector, and
- address the ethical, legal, and social issues (ELSI) that may arise from the project.

Milestones:

- 1990: Project initiated as joint effort of U.S. Department of Energy and the National Institutes of Health
- June 2000: Completion of a working draft of the entire human genome (covers >90% of the genome to a depth of 3-4x redundant sequence)
- February 2001: Analyses of the working draft are published
- April 2003: HGP sequencing is completed and Project is declared finished two years ahead of schedule

http://doegenomes.org

ww.sanger.ac.uk/HGP/overview.shtml U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

What does the draft human genome sequence tell us?

By the Numbers

- The human genome contains 3 billion chemical nucleotide bases (A, C, T, and G).
- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.
- The total number of genes is estimated at around 30,000--much lower than previous estimates of 80,000 to 140,000.
- Almost all (99.9%) nucleotide bases are exactly the same in all people.
- The functions are unknown for over 50% of discovered genes.

What does the draft human genome sequence tell us?

How It's Arranged

- The human genome's gene-dense "urban centers" are predominantly composed of the DNA building blocks G and C.

- In contrast, the gene-poor "deserts" are rich in the DNA building blocks A and T. GC- and AT-rich regions usually can be seen through a microscope as light and dark bands on chromosomes.

- Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between.

- Stretches of up to 30,000 C and G bases repeating over and over often occur adjacent to gene-rich areas, forming a barrier between the genes and the "junk DNA." These CpG islands are believed to help regulate gene activity.

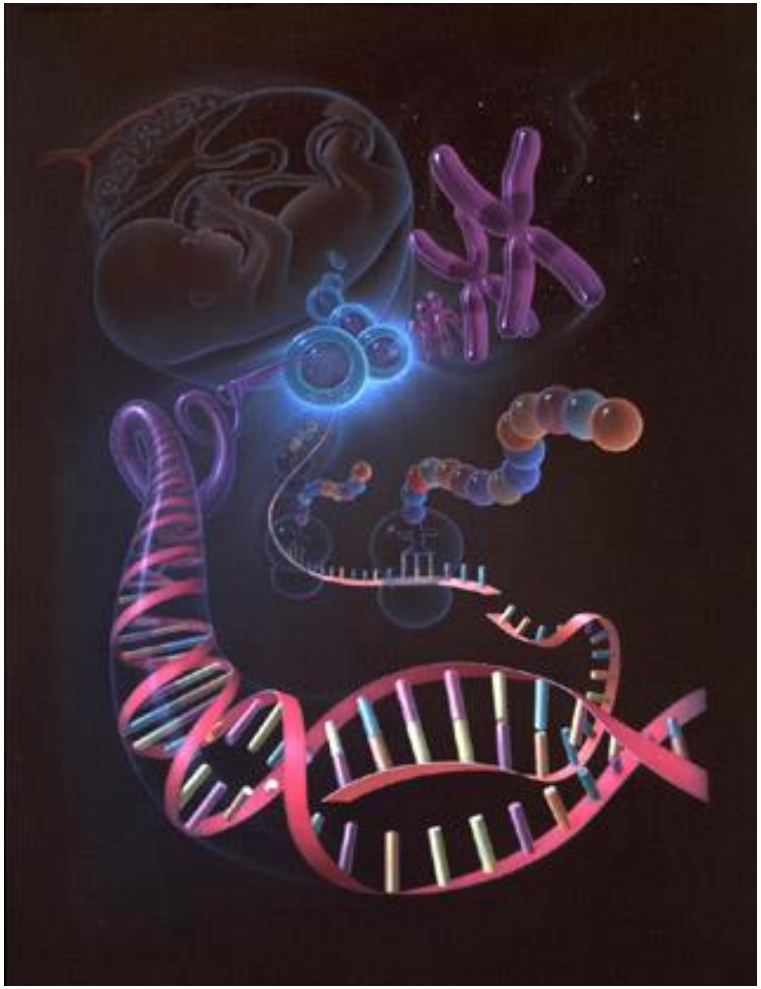
- Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231).

How does the human genome stack up?

Organism	Genome Size (Bases)	Estimated Genes
Human (*Homo sapiens*)	3 billion	30,000
Laboratory mouse (*M. musculus*)	2.6 billion	30,000
Mustard weed (*A. thaliana*)	100 million	25,000
Roundworm (*C. elegans*)	97 million	19,000
Fruit fly (*D. melanogaster*)	137 million	13,000
Yeast (*S. cerevisiae*)	12.1 million	6,000
Bacterium (*E. coli*)	4.6 million	3,200
Human immunodeficiency virus (HIV)	9700	9

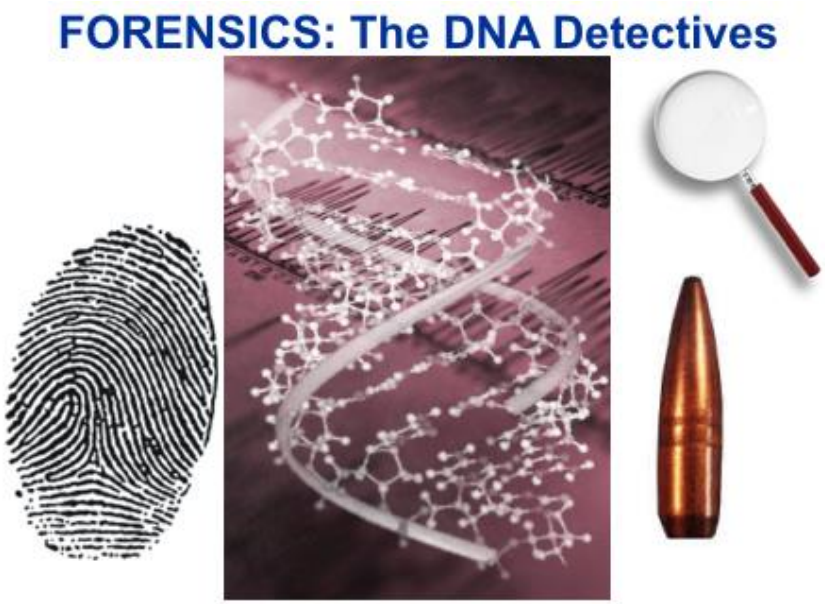
Benefits of Human Genome Project research

- improvements in medicine.
- microbial genome research for fuel and environmental cleanup.
- DNA forensics.
- improved agriculture and livestock.
- better understanding of evolution and human migration.
- more accurate risk assessment.



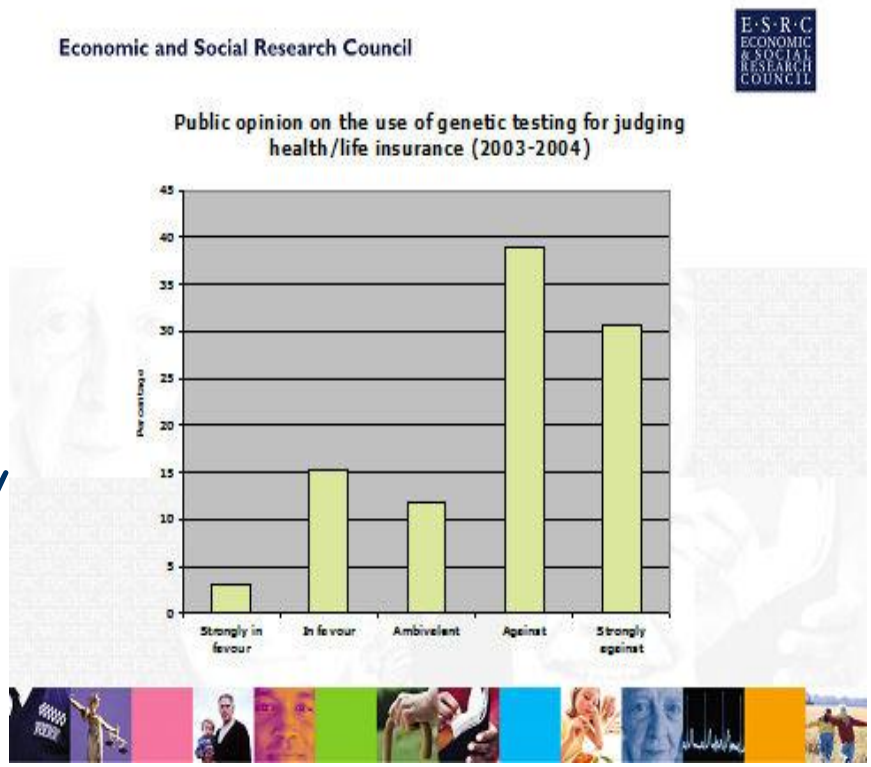
How is each area benefited specifically by the Human Genome Project?

- Improvements in medicine: improved diagnosis of disease.
- Microbial research: new energy sources, bio fuels.
- DNA forensics: identifying potential suspects at a crime scene.
- Agriculture: more nutritious produce.
- Evolution and human migration: study migration of different population groups based on female genetic inheritance.
- Risk assessment: reduce the likelihood of heritable mutations.



Ethical, legal and social implications of the Human Genome Project

- fairness in the use of genetic information.
- privacy and confidentiality.
- psychological impact and stigmatization.
- genetic testing.
- reproductive issues.
- education, standards, and quality control.
- commercialization.
- conceptual and philosophical implications.



What are the implications of the Human Genome Project specifically to each of these areas?

Some questions to consider:

- Fairness and privacy: who should have access to your genetic information?
- Psychological stigmatization: how does knowing your predisposition to disease affect an individual?
- Genetic testing: should screening be done when there is no treatment available?

Some other issues:

- Reproductive issues: use of genetic information in decision making.
- Clinical issues: implementation of standards and quality control measures in testing procedures.

Human Genome Project, 1984-86

- DOE was interested in genetic research regarding health effects from radiation and chemical exposure
- NIH was interested in gene sequencing / mutations and their biomedical implications of genetic variation

Human Genome Project, 1988

- Reports from OTA, NRC, and an Ad Hoc Advisory Committee on Complex Genomes
- All reports supported the development of the Human Genome Project, with parallel projects for other model organisms
- Congress agreed to appropriate funds to support research to determine the structure of complex genomes

Human Genome Project, 1989

- Congress required NIH and DOE to prepare a detailed plan for the appropriations hearings
- NHGRI was created as a new division of NIH with budget estimated at 1% of total for NIH

Human Genome Project, 1990 Five Year Plan

- Construct a high resolution genetic map of the human genome
- Produce physical maps of all chromosomes
- Determine genome sequence of human and other model organisms
- Develop capabilities (technologies) for collecting, storing, distributing and analyzing data

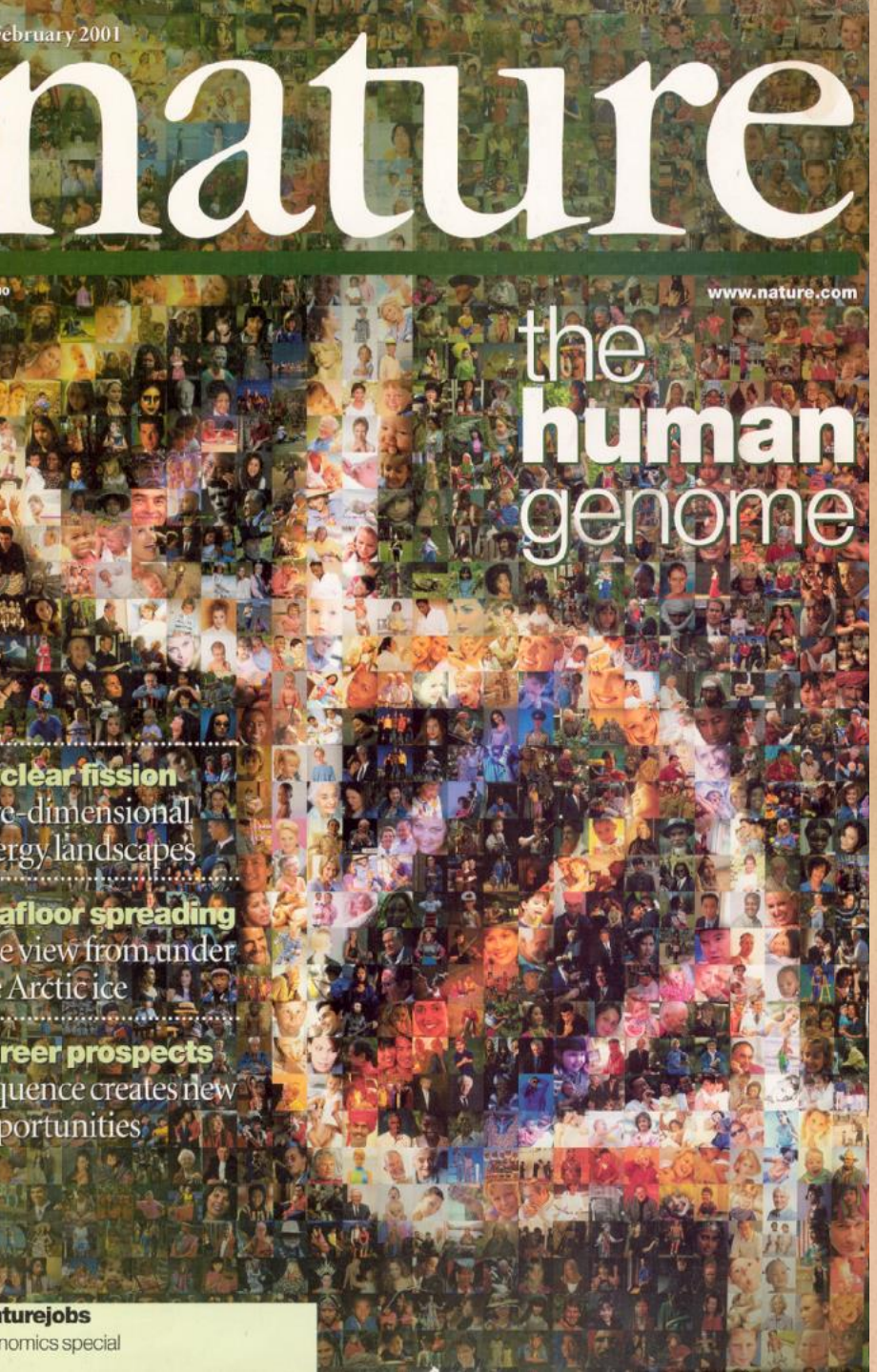
Human Genome Project, 1990 Additional Goals

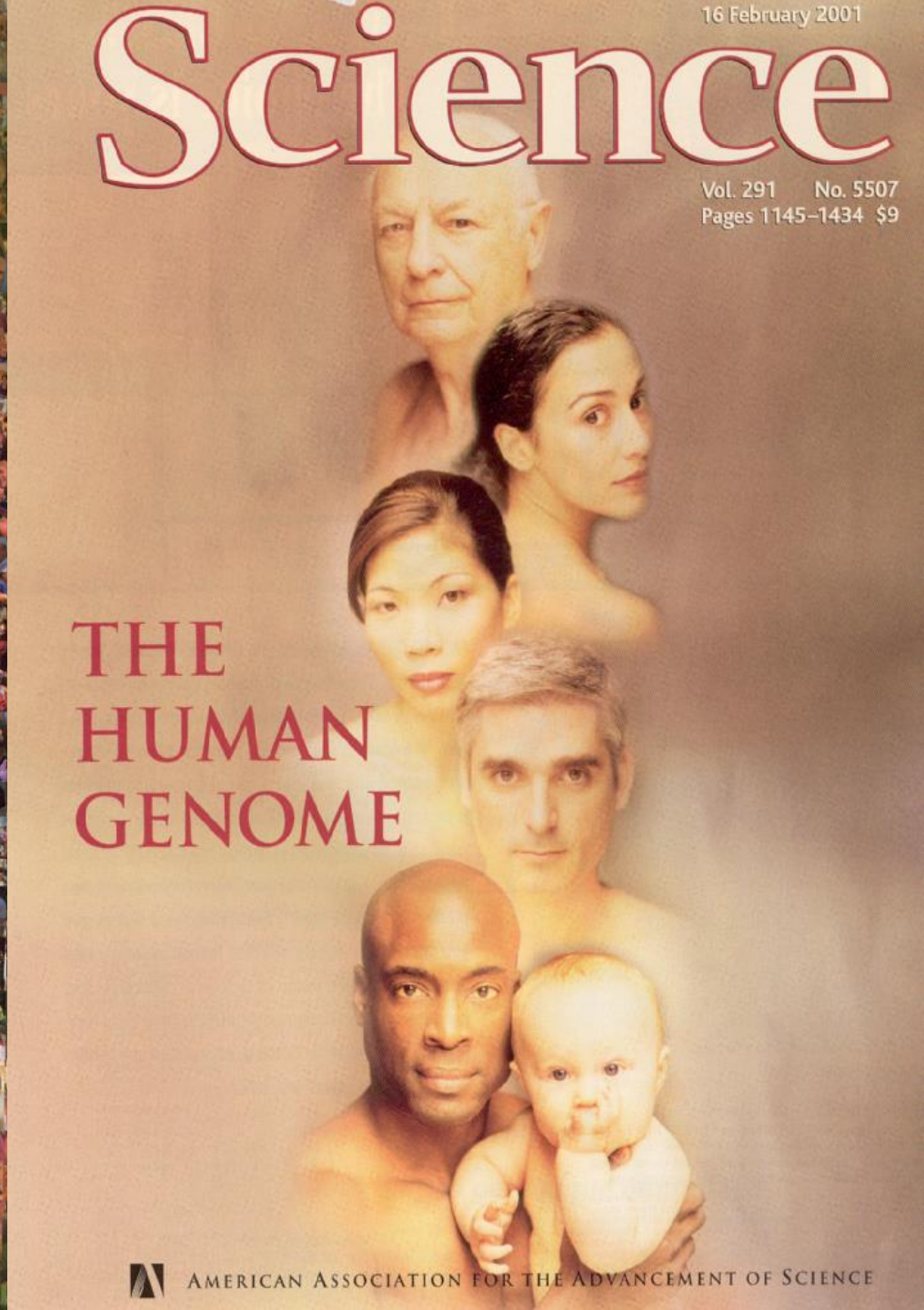
- Ethical, legal, social issues (ELSI)
- Research training
- Technology transfer
- Human Genome Project began with a recommended budget of $200 million per year, adjusted for inflation
  - 15 years, $3 billion

How Did the Draft Sequence Develop?

Draft Sequences, 2001

- International Human Genome Sequencing Consortium ('public project')
  - *Initial Sequencing and Analysis of the Human Genome. Nature 409:860-921, 2001*
- Celera Genomics – Venter JC et al. ('private project')
  - *The Sequence of the Human Genome. Science 291:1304-1351, 2001.*


ebruary 2001
nature
www.nature.com
the human genome
clear fission
e-dimensional
ergy landscapes
afloor spreading
e view from under
Arctic ice
reer prospects
quence creates new
portunities
turejobs
nomics special


16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

International Human Genome Sequencing Consortium

- Collaboration would be open to centers from any nation
  - 20 centers from 6 countries
  - US, UK, China, France, Germany, Japan

- Rapid and unrestricted data release
  - Assembled sequences >2 kb were deposited within 24 hours of assembly

Table 3 Total human sequence deposited in the HTGS division of GenBank

Sequencing centre	Total human sequence (kb)	Finished human sequence (kb)
Whitehead Institute, Center for Genome Research*	1,196,888	46,560
The Sanger Centre*	970,789	284,353
Washington University Genome Sequencing Center*	765,898	175,279
US DOE Joint Genome Institute	377,998	78,486
Baylor College of Medicine Human Genome Sequencing Center	345,125	53,418
RIKEN Genomic Sciences Center	203,166	16,971
Genoscope	85,995	48,808
GTC Sequencing Center	71,357	7,014
Department of Genome Analysis, Institute of Molecular Biotechnology	49,865	17,788
Beijing Genomics Institute/Human Genome Center	42,865	6,297
Multimegabase Sequencing Center; Institute for Systems Biology	31,241	9,676
Stanford Genome Technology Center	29,728	3,530
The Stanford Human Genome Center and Department of Genetics	28,162	9,121
University of Washington Genome Center	24,115	14,692
Keio University	17,364	13,058
University of Texas Southwestern Medical Center at Dallas	11,670	7,028
University of Oklahoma Advanced Center for Genome Technology	10,071	9,155
Max Planck Institute for Molecular Genetics	7,650	2,940
GBF – German Research Centre for Biotechnology	4,639	2,338
Cold Spring Harbor Laboratory Lita Annenberg Hazen Genome Center	4,338	2,104
Other	59,574	35,911
Total	4,338,224	842,027

Total human sequence deposited in GenBank by members of the International Human Genome Sequencing Consortium, as of 8 October 2000.The amount of total sequence (finished plus draft plus predraft) is shown in the second column and the amount of finished sequence is shown in the third column. Total sequence differs from totals in Tables 1 and 2 because of inclusion of padding characters and of some clones not used in assembly. HTGS, high throughput genome sequence.

*These three centres produced an additional 2.4 Gb of raw plasmid paired-end reads (see Table 4), consisting of 0.99 Gb from Whitehead Institute, 0.66 Gb from The Sanger Centre and 0.75 Gb from Washington University.

Challenges

- Data were generated in labs all over the world
- Organism is diploid, extremely large genome
- Large proportion of the human genome consists of repetitive and duplicated sequences
- Cloning bias (under-representation of some region of the genome)

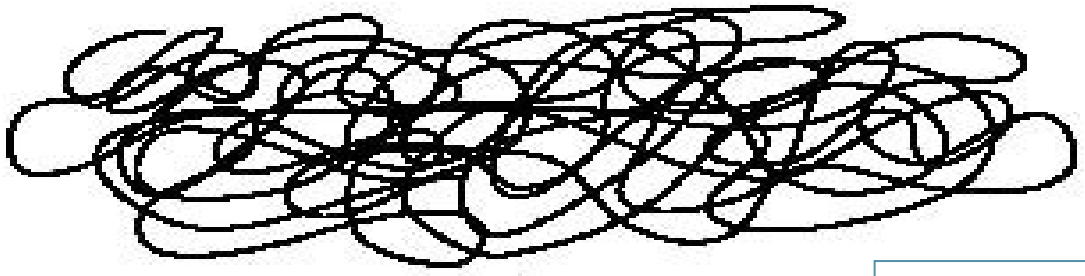
Approaches to Sequence

- Shotgun Phase
  - Hierarchical shotgun sequencing used to produce draft sequence of 90% of the genome

- Finishing Phase
  - Fill in gaps and resolve ambiguities
  - Fragments must be sequenced ~ 10 times to reach accuracy of >99%
    - In 1981, sequencing 12,000 bp took ~1 yr
    - In 2001, sequencing 12,000 bp takes < 1 min

Hierarchical shotgun sequencing

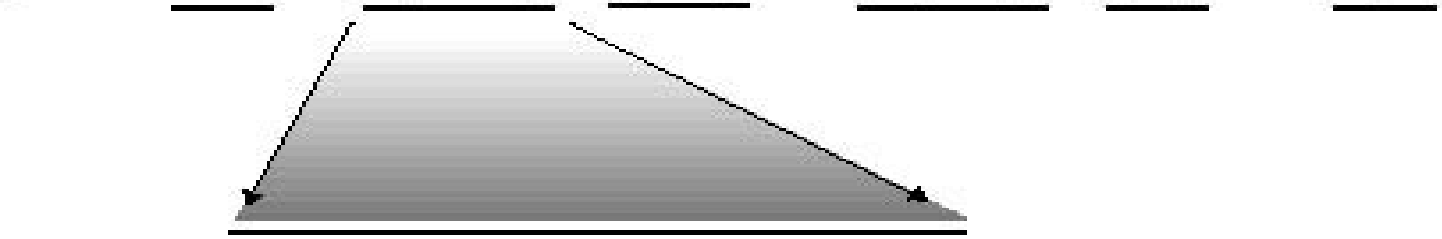
Genomic DNA



Partial Digest

BAC library

Organized mapped large clone contigs



BAC to be sequenced

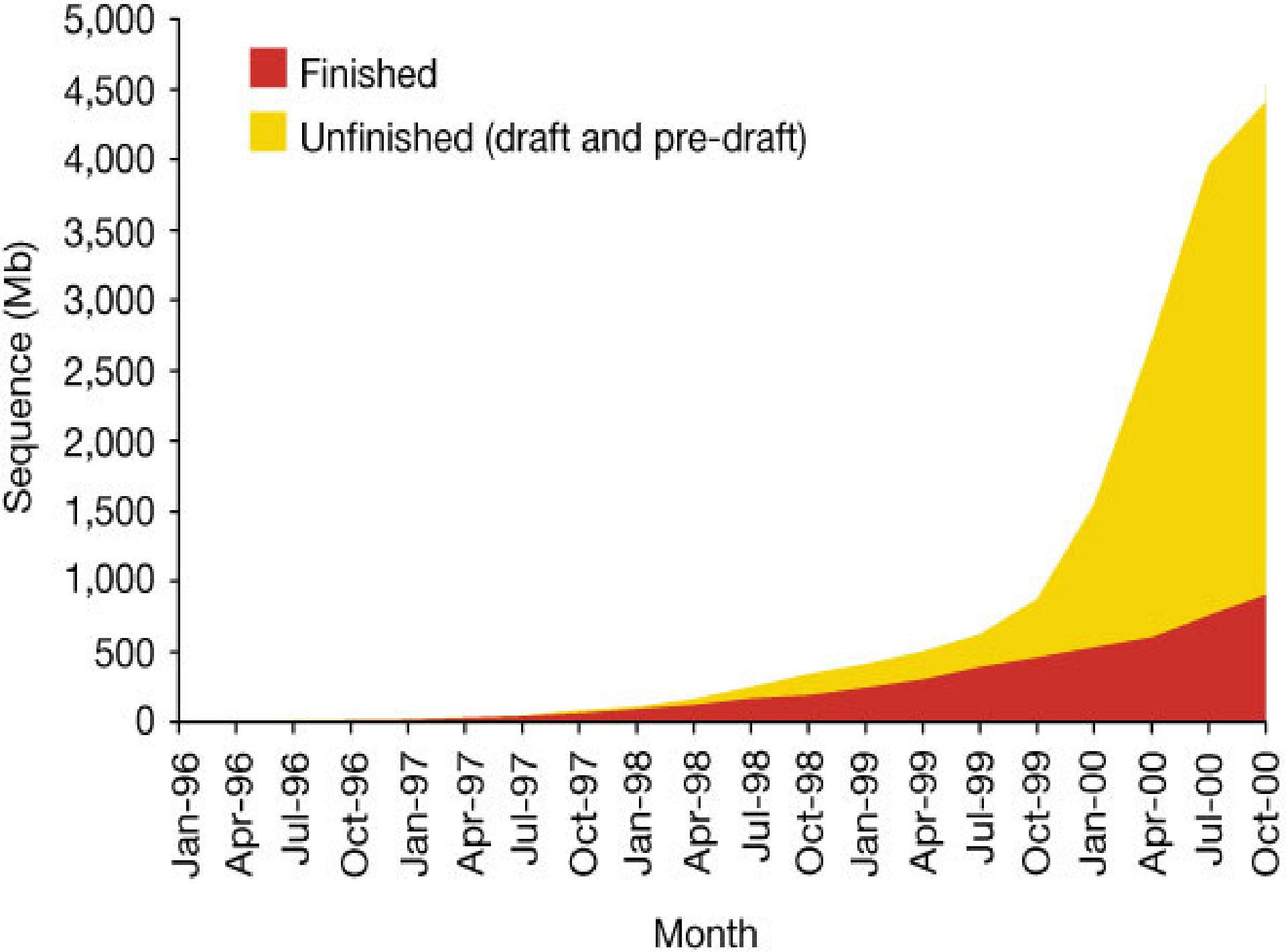
Shotgun clones

Shotgun sequence

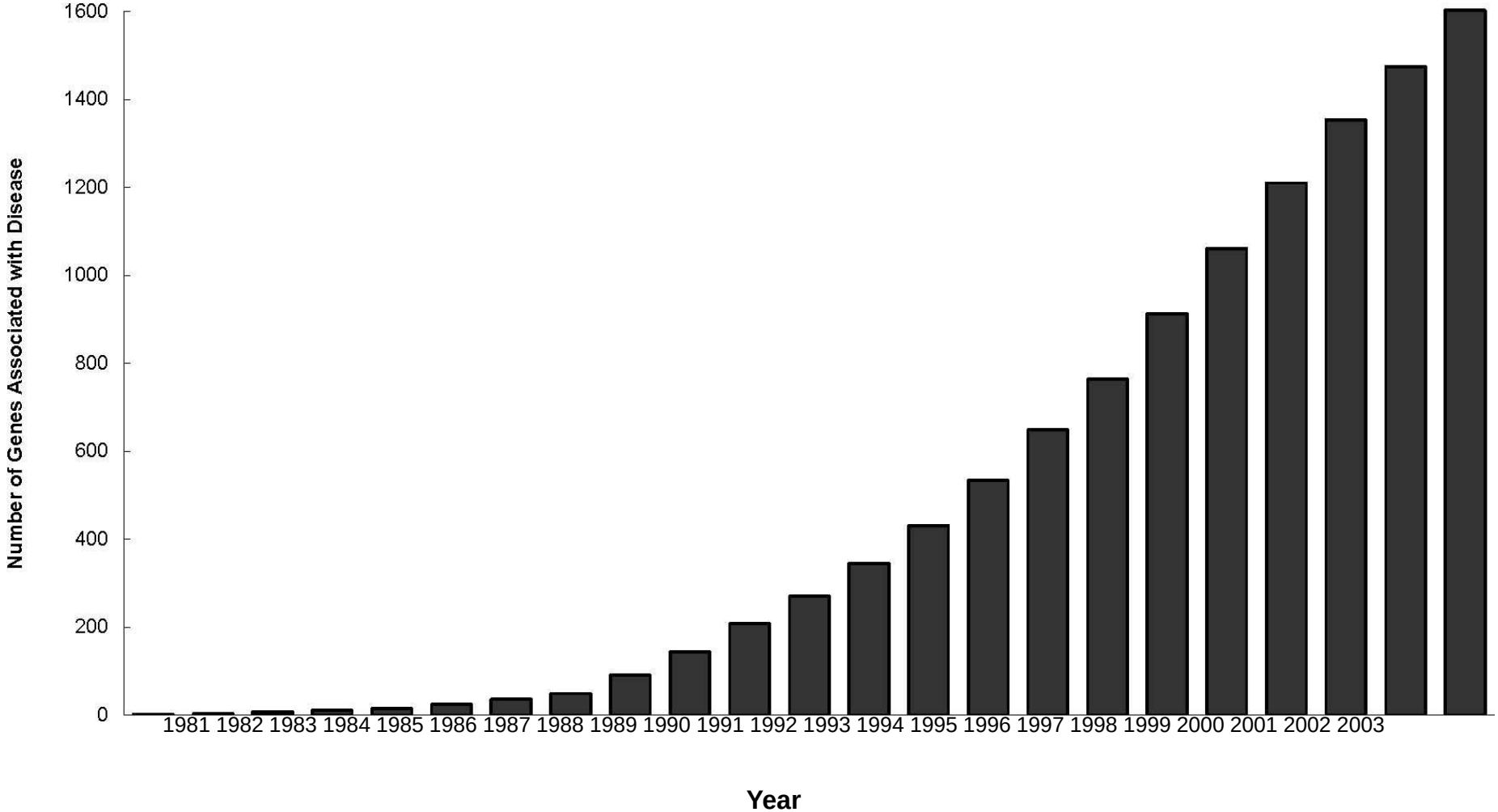
```
...ACCGTAAATGGGCTGATCATGCTTAAA
                   TGATCATGCTTAAACCCTGTGCATCCTACTG...
```

Assembly

```
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...
```


Finished
Unfinished (draft and pre-draft)
Sequence (Mb)
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Jan-96
Apr-96
Jul-96
Oct-96
Jan-97
Apr-97
Jul-97
Oct-97
Jan-98
Apr-98
Jul-98
Oct-98
Jan-99
Apr-99
Jul-99
Oct-99
Jan-00
Apr-00
Jul-00
Oct-00
Month

Cumulative Pace of Gene Discovery 1981-2003[1]



Minimum estimated values are represented

Major Findings of the Draft Sequence


ONCE YOU UNFOLD ONE OF THESE THINGS, IT'S NEVER THE SAME.
Joe Heller
GREEN BAY PRESS-GAZETTE
Joe@hellerton.com
MAP OF THE HUMAN GENOME
Ethical Questions
Legal Tangles
Privacy concerns
Moral Issues
Insurance Applications
Medical Dilemmas

Number of Genes

- Number of genes only ~ 35,000
  - <2% of genome encode genes
  - Fruit fly has 13,000 genes
  - Mustard weed has 26,000

- Proteome is complex
  - 1 gene codes up to ~ 1000 proteins
    - Alternative splicing
  - Variation in gene regulation
  - Post transcription modification

- Hundreds of genes appear to have come from bacteria

Number of Genes

- Estimated from:
  - Comparisons with other genomes
  - Comparisons with identified genes (protein motifs, pseudogenes)
  - Extrapolations from chr 21 and 22
  - Presence of CpG islands
  - Presence of initiator, promoter or enhancer / silencer sequences
  - Evidence of alternative splicing
  - Known expressed sequence tags

Categorization of genes

23.2%	Expression, replication, maintenance
21.1%	Signal transduction
17.5%	Biochemical functions of the cell
38.2%	Other

Mutation Rate and SNPs

- Mutation rate is twice as high in male compared to female meiosis
  - Sperm is the main source of mutation

- More than 1.4 million single nucleotide polymorphisms (SNPs)
  - Occur ~ 2000 base pairs apart, but density varies
  - Used to create haplotypes to differential between maternal / paternal chromosomes
  - May relate to disease susceptibility
  - Used for genome-wide association studies

Genetic Variation

- Comparisons between humans indicate that we are all 99.9% genetically identical
  - No basis for genetic discrimination
  - Translates to ~3 million difference
    - Some have no effect
    - Some cause differences in appearance, behavior, etc.
    - Some effect vulnerability to disease

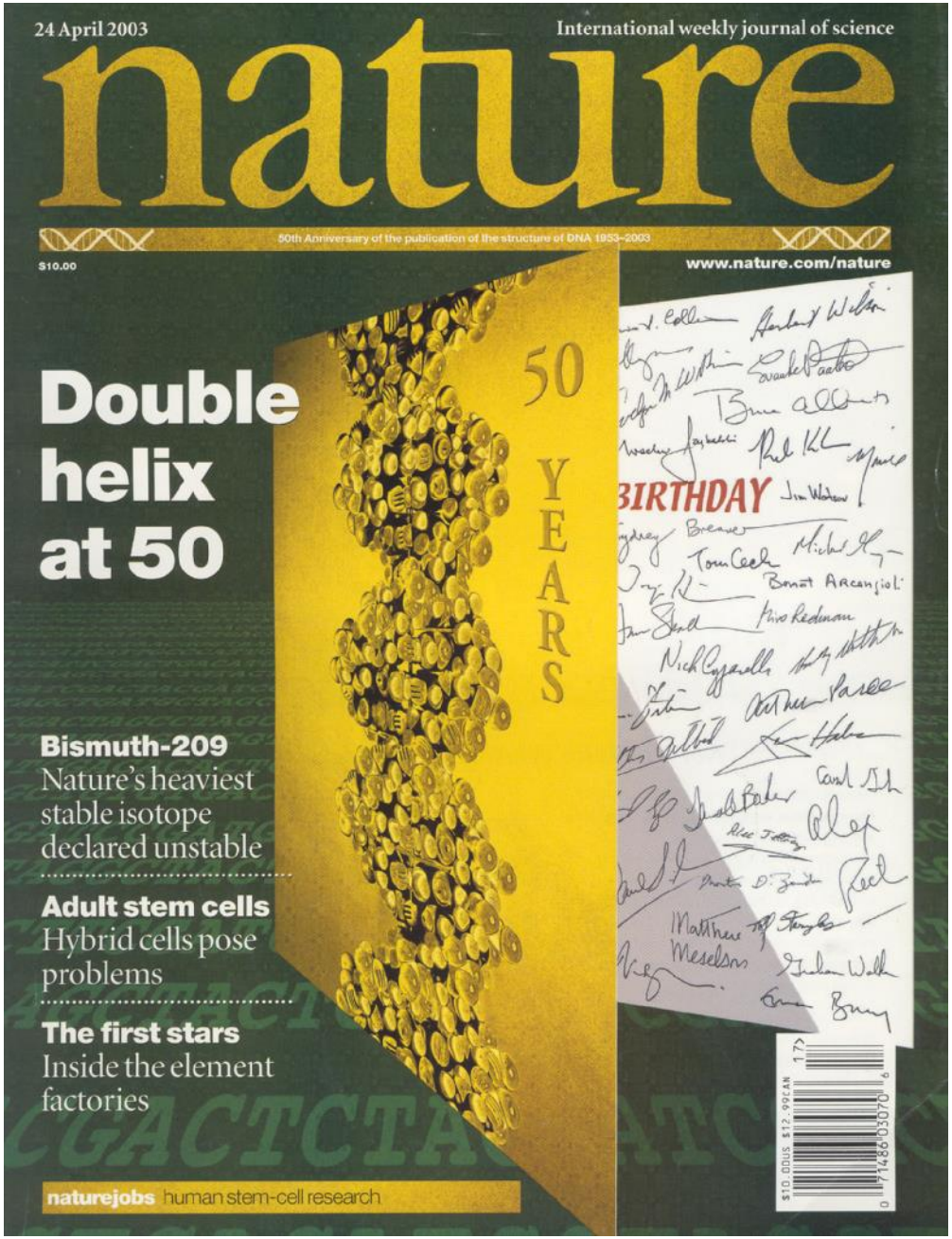
- Marked variation across the human genome
  - Gene rich (chr 19) vs. gene poor (chr 13) regions
  - Chr 21 & 22 (smallest) were sequenced 1st
  - Chr 21 ~ 225 genes; Chr 22 ~ 550 genes
  - Variation in distribution of repeat sequences

Repeat Content

- Account for > 50% of genome
  - Transposon-derived repeats (LINES, SINES)
  - Simple sequence repeats (SSRs – satellite DNA)
  - Segmental duplication
  - Blocks of tandem repeats
- Reshaped the genome
  - Key route to providing an enhanced functional repertoire
- Pseudogenes – inactive genes

In the US Senate, February 27, 2003

- Concurrent Resolutions Designating
  - April, 2003 as "Human Genome Month"
    - IHGSC placed complete human genome sequence in public databases
    - Celera database was offered for purchase
    - NHGRI unveils new plan for the future of genomics
  - April 25 as "DNA Day"
    - Marks the 50th anniversary of the description of the double helix by Watson and Crick


24 April 2003
International weekly journal of science
nature
50th Anniversary of the publication of the structure of DNA 1953–2003
$10.00
www.nature.com/nature
Double helix at 50
50 YEARS
Bismuth-209
Nature's heaviest stable isotope declared unstable
Adult stem cells
Hybrid cells pose problems
The first stars
Inside the element factories
naturejobs human stem-cell research
BIRTHDAY