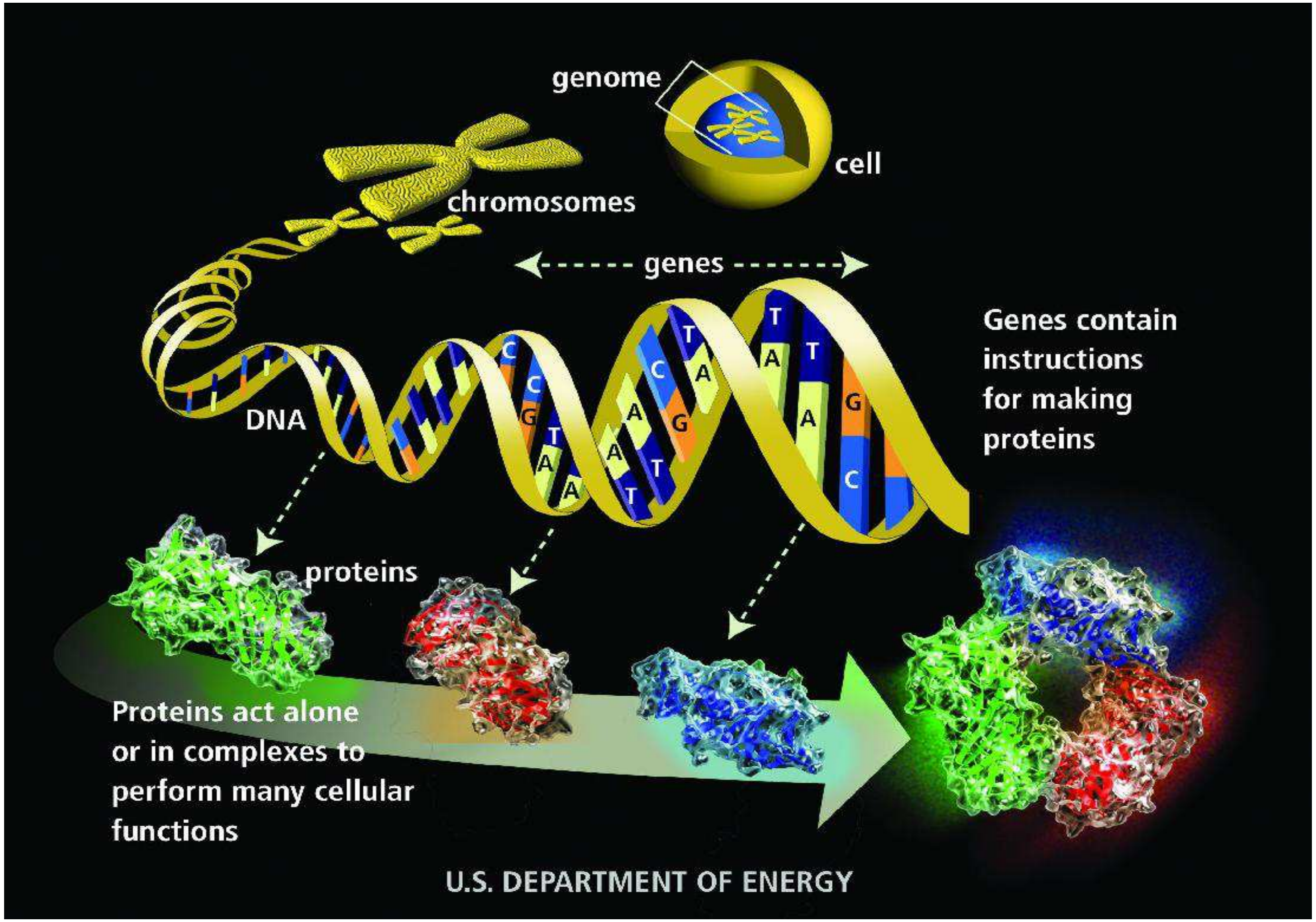
Human Genome Project


genome
cell
chromosomes
genes
DNA
C
C
G
T
A
A
T
C
A
A
G
T
T
A
T
A
T
A
G
C
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

ckground of the human genome project

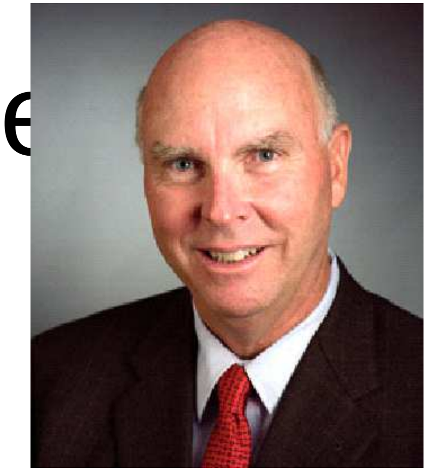
e Human Genome Project was a large-scale international collaboration that an in the 1990s.

e Human Genome Project (HGP) was first proposed by the U.S. National search Council in 1988.

e goals were to create genetic, physical, and sequence maps of the human ome.

parallel, genomes of model organisms were to be studied.

Dr. J. Craig Vente



Earned his Ph.D. in Physiology and Pharmacology from the iversity of California at San Diego and became a researcher he National Institutes of Health. While serving first as a Section

ief and then as a Lab Chief in the National Institute of Neurological Disorders and troke, he developed expressed sequence tags or EST’s, a revolutionary new strategy for ene discovery. In 1992, he and his wife, Dr. Claire Fraser, founded The Institute for Genomic Research known as TIGR, where he served as President and Chief Scientific Officer until 1998. Dr. Venter and his team at TIGR decoded the genome of the acterium Haemophilus influenzae, making it the first free-living organism to have its ull DNA deciphered and to date have sequenced over 30 genomes. He now serves as Chairman of the Board of Trustees of TIGR.

n 1998, he founded Celera Genomics and announced that Celera would decode the uman genome faster and more economically than the publicly funded consortium of cientists. At the White House press conference announcing the sequencing of the human enome, President Bill Clinton called it “the most important, most wondrous map ever roduced by mankind.”

Costs of Human Genomic Sequencing

- *Clone by clone*
- $0.30 per finished base
- $130 million per year for 7 years
- Total $900 million spent by end of 2003
- *Shotgun*
- $0.01 per raw base
- $130 million for 3 years would provide
- 10× coverage/redundancy plus an additional $90 million for informatics

http://www-management.wharton.upenn.edu/pennings/coursedocuments/executive_education_courses/557/genome%20Paper%201.doc

Reducing Costs and Speeding Up Sequencing

Technological developments dramatically decreased DNA sequencing's cost while increasing its speed and efficiency. For example, it took 4 years for the international Human Genome Project to produce the first billion base pairs of sequence and less than 4 months to produce the second billion base pairs. In the month of January 2003, the DOE team sequenced 1.5 billion bases. The cost of sequencing has dropped dramatically since the project began and is still dropping rapidly.

Source DNA

For the publicly funded HGP, human DNA was isolated from blood (female) and sperm (male) collected from a large number of donors.

For the work privately funded by Celera Genomics, DNA resources used for these studies came from anonymous donors of European, African, American (North, Central, South), and Asian ancestry.

Research Challenges

❿Gene number, exact locations, and functions

❿Gene regulation

❿DNA sequence organization

❿Chromosomal structure and organization

❿Noncoding DNA types, amount, distribution, information content, and functions

❿Coordination of gene expression, protein synthesis, and post-translational events

❿Interaction of proteins in complex molecular machines

❿Predicted vs experimentally determined gene function

Introduction

The human genome is the complete set of DNA in *Homo sapiens*.

Complete sequence of nucleotide basepairs.

Its initial sequencing (2003) came 50 years after the publication of the double-stranded helical structure of DNA by Crick and Watson (1953).

In 2001 the sequencing extensive draft versions of the human genome were reported separately by the International Human Genome Sequencing Consortium (IHGSC) and by Celera Genomics.

ght goals of Human Genome Project (1998–2003)

[1] Human DNA sequence

[2] Develop sequencing technology

[3] Identify human genome sequence variation

[4] Functional genomics technology

[5] Comparative genomics

[6] ELSI: ethical, legal, and social issues

[7] Bioinformatics and computational biology

[8] Training and manpower

ain conclusions of human genome project

We now appreciate that human have about the same number of protein-coding genes as fish and plants, and not that many more genes than worms and flies.

Human (Homo sapiens):	31,000 to 38,000
Fugu rubripes (pufferfish):	8,500*
Arabidopsis thaliana (thale cress):	27,600**
Caenorhabditis elegans (worm):	20,300*
Drosophila melanogaster (fly):	13,900*

* 2017 estimate from Ensembl

** 2017 estimate from TAIR

in conclusions of human genome project

98% of the genome does not code for genes

>50% of the genome consists of repetitive DNA derived from transposable elements:

- LINEs (20%)
- SINEs (13%)
- LTR retrotransposons (8%)
- DNA transposons (3%)

re than 1.4 million single nucleotide polymorphisms (SNPs; single base pair nges) were identified.

era initially identified 2.1 million SNPs.

gmental duplication is a frequent occurrence in the human genome.

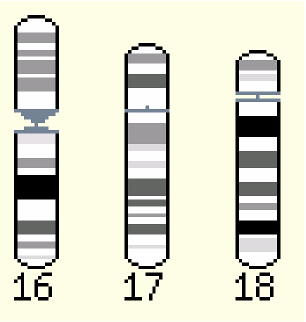
Three gateways to access the human genome

NCBI map viewer
www.ncbi.nlm.nih.gov

Ensembl Project (EBI/Sanger Institute)
www.ensembl.org

UCSC (Golden Path)
www.genome.ucsc.edu

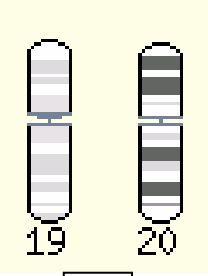
Each of these three sites provides essential resources to study the human genome (and other genomes)



E

Chrom.	Length	# Genes	# Pseudo-genes	Gap size (Mb)	Accession
16	90 Mb	796	778	11.5	NC_000016.9
17	81 Mb	1,266	274	3.4	NC_000017.10
18	78 Mb	337	171	3.4	NC_000018.9

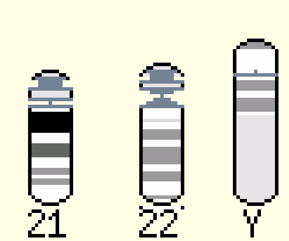
Chromosome 18 has the lowest gene density of any autosome (4.4 genes per megabase) and encodes only 337 genes (about a quarter of the number of the similar-sized chromosome 17). One region of chromosome 18 has only 3 genes across 4.5 megabases. The sparse number of genes may explain why some individuals with trisomy 18 (Edwards syndrome) survive to birth, while all other autosomal trisomies (except trisomy 13 and trisomy 21) are embryonic lethal.



F

Chrom.	Length	# Genes	# Pseudo-genes	Gap size (Mb)	Accession
19	59 Mb	1,461	321	3.3	NC_000019.9
20	63 Mb	727	168	3.5	NC_000020.10

Chromosome 19 has the highest gene density with 26 protein-coding genes per megabase.



G

Chrom.	Length	# Genes	# Pseudo-genes	Gap size (Mb)	Accession
21	48 Mb	796	778	13.0	NC_000021.8
22	51 Mb	545	134	16.4	NC_000022.10
Y	59 Mb	78	n/a	33.7	NC_000024.9

The Y chromosome was the most technically difficult to sequence because of its extraordinarily repetitive nature. It has short pseudoautosomal regions at the ends that recombine with the X chromosome. A large central region, spanning 95% of its length, is termed the male-specific region (MSY). There are 23 megabases of euchromatin including 8 Mb on Yp and 14.5 Mb on Yq. There are three notable heterochromatic regions:

(1) a centromeric region of about 1 Mb,

(2) a block of ~40 Mb on the long arm, and

(3) an island of 400 kilobases comprised of over 3,000 tandem repeats of 125 base pairs.

Of 156 transcription units, about half encode proteins.

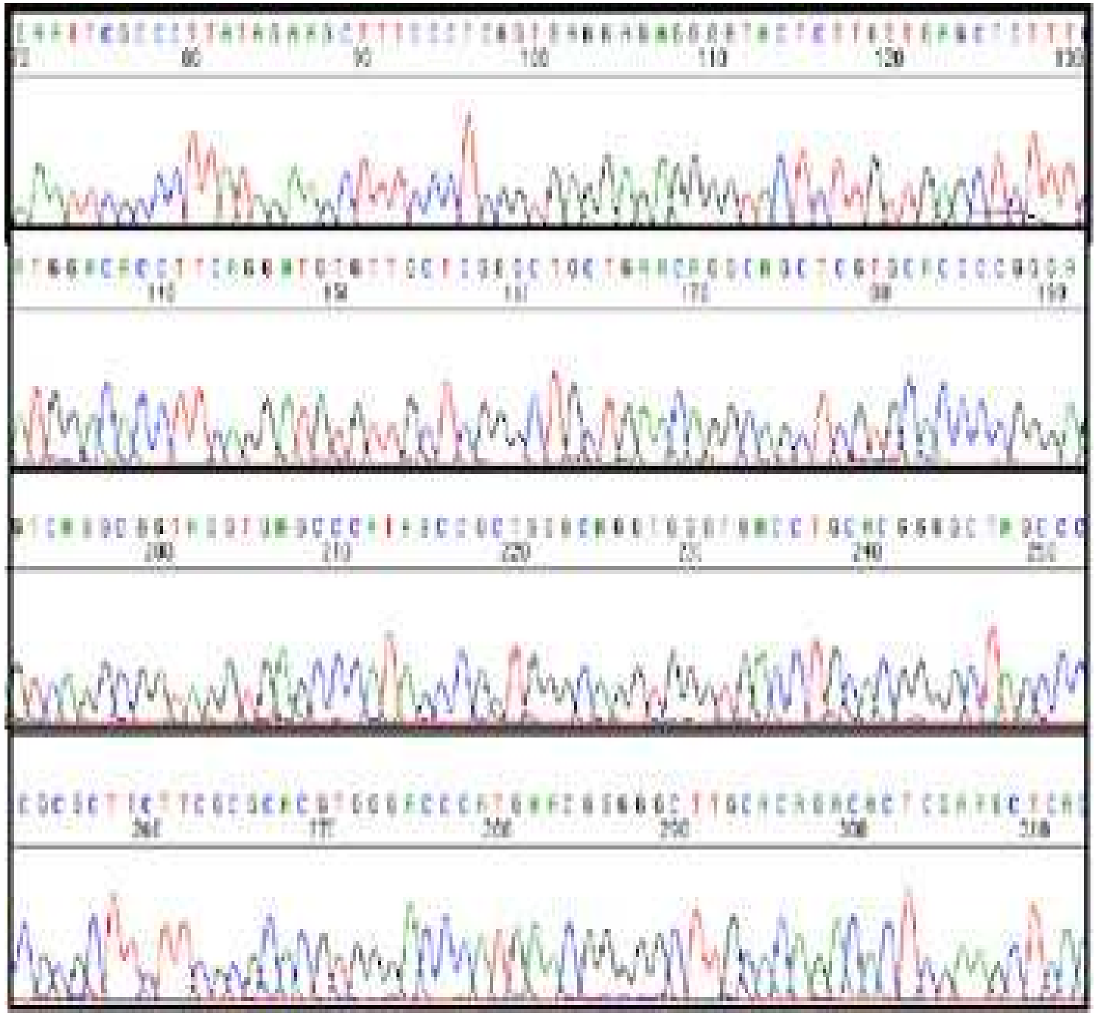
NA Sequencing

DNA sequencing involves:

1) Isolation and purification of DNA from individuals

2) Chemical sequencing to establish the order of nucleotides

3) Arranging the DNA sequences into the proper order



Shotgun Sequencing

Bacteriophage fX174, was the first genome to be sequenced, a viral genome wit nly 5,368 base pairs (bp).

Frederic Sanger, in another revolutionary discovery, invented the method c "shotgun" sequencing, a strategy based on the isolation of random pieces of DNA rom the host genome to be used as primers for the PCR amplification of the entir enome.

The amplified portions of DNA are then assembled by their overlapping regions t orm contiguous transcripts (otherwise known as contigs).

The final step involved the utilization of custom primers to elucidate the gap etween the contigs thus giving the completely sequenced genome.

anger first used "shotgun" sequencing five years later to complete the bacterio hage l sequence that was significantly larger, 48,502 bp. This method allowe equencing projects to proceed at a much faster rate thus expanding the scope o ealistic sequencing venture.

Sanger first used "shotgun" sequencing five years later to complete the bacterio hage l sequence that was significantly larger, 48,502 bp.

Shotgun Sequencing

- A team headed by J. Craig Venter from the Institute for Genomic Research (TIGR) and Nobel laureate Hamilton Smith of Johns Hopkins University, sequenced the 1.8 Mb bacterium with new computational methods developed at TIGR's facility in Gaithersburg, Maryland.
- Previous sequencing projects had been limited by the lack of adequate computational approaches to assemble the large amount of random sequences produced by "shotgun" sequencing.
- In conventional sequencing, the genome is broken down laboriously into ordered, overlapping segments, each containing up to 40 Kb of DNA.
- These segments are "shotgunned" into smaller pieces and then sequenced to reconstruct the genome.
- Venter's team utilized a more comprehensive approach by "shotgunning" the entire 1.8 Mb H. Influenzae genome.

Shotgun Sequencing

- Previously, such an approach would have failed because the software did not exist to assemble such a massive amount of information accurately.
- Software, developed by TIGR, called the TIGR Assembler was up to the task, reassembling the approximately 24,000 DNA fragments into the whole genome.
- After the H. Influenzae genome was "shotgunned" and the clones purified sufficiently the TIGR Assembler software required approximately 30 hours of central processing unit time on a SPARCenter 2000 containing half a gigabyte of RAM .

Shotgun Sequencing

- Venter's H. Influenzae project had failed to win funding from the National Institute of Health indicating the serious doubts surrounding his ambitious proposal.
- It simply was not believed that such an approach could sequence the large 1.8 Mb sequence of the bacterium accurately.
- Venter proved everyone wrong and succeeded in sequencing the genome in 13 months at a cost of 50 cents per base which was half the cost and drastically faster than conventional sequencing.

Anticipated Benefits of Genome Research

Molecular Medicine

- improve diagnosis of disease
- detect genetic predispositions to disease
- create drugs based on molecular information
- use gene therapy and control systems as drugs
- design “custom drugs” (pharmacogenomics) based on individual genetic profiles

Microbial Genomics

- rapidly detect and treat pathogens (disease-causing microbes) in clinical practice
- develop new energy sources (biofuels)
- monitor environments to detect pollutants
- protect citizenry from biological and chemical warfare
- clean up toxic waste safely and efficiently

Anticipated Benefits of Genome Research-cont.

Risk Assessment

- evaluate the health risks faced by individuals who may be exposed to radiation (including low levels in industrial areas) and to cancer-causing chemicals and toxins

Bioarchaeology, Anthropology, Evolution, and Human Migration

- study evolution through germline mutations in lineages
- study migration of different population groups based on maternal inheritance
- study mutations on the Y chromosome to trace lineage and migration of males
- compare breakpoints in the evolution of mutations with ages of populations and historical events

Anticipated Benefits of Genome Research-cont.

DNA Identification (Forensics)

- identify potential suspects whose DNA may match evidence left at crime scenes
- exonerate persons wrongly accused of crimes
- identify crime and catastrophe victims
- establish paternity and other family relationships
- identify endangered and protected species as an aid to wildlife officials (could be used for prosecuting poachers)
- detect bacteria and other organisms that may pollute air, water, soil, and food
- match organ donors with recipients in transplant programs
- determine pedigree for seed or livestock breeds
- authenticate consumables such as caviar and wine

Anticipated Benefits of Genome Research-cont.

Agriculture, Livestock Breeding, and Bioprocessing

- grow disease-, insect-, and drought-resistant crops
- breed healthier, more productive, disease-resistant farm animals
- grow more nutritious produce
- develop biopesticides
- incorporate edible vaccines incorporated into food products
- develop new environmental cleanup uses for plants like tobacco

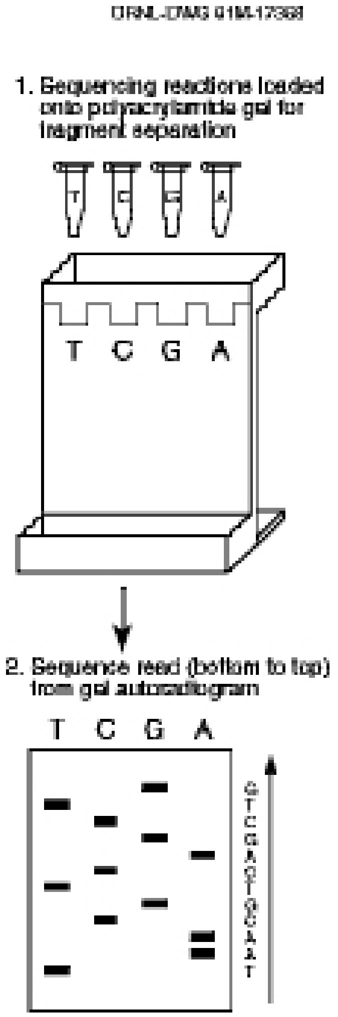
Dideoxy Sequencing

ideoxy sequencing (Sanger method) uses an enzymatic procedure to ynthesize DNA chains of varying lengths, stopping DNA replication at ne of the four bases and then determining the resulting fragment lengths. ach sequencing reaction tube (T, C, G, and A) in the diagram contains

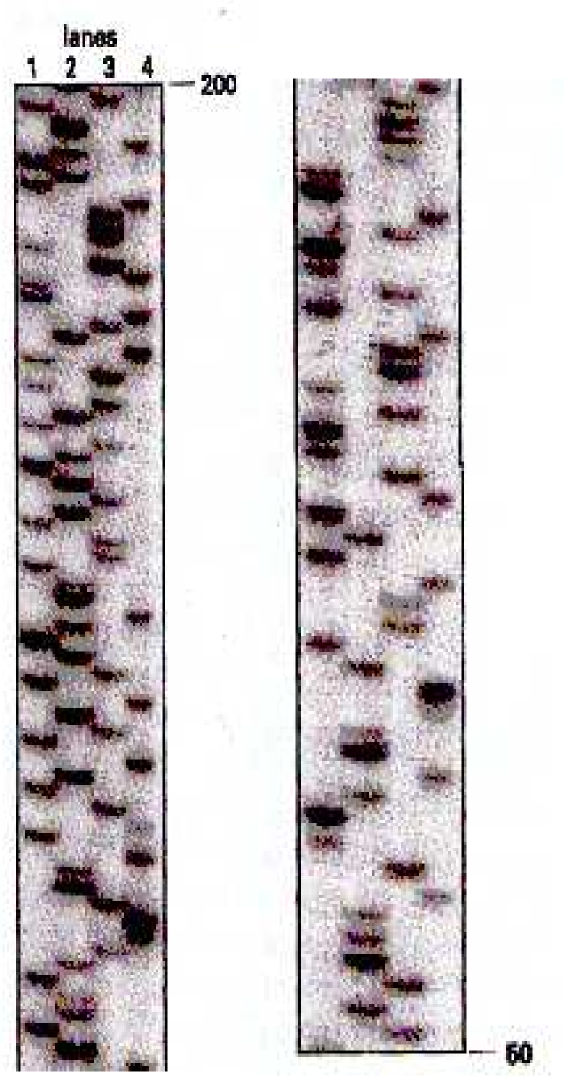
- DNA template, a primer sequence, and a DNA polymerase to initiate synthesis of a new strand of DNA at the point where the primer is hybridized to the template;
- the four deoxynucleotide triphosphates (dATP, dTTP, dCTP, and dGTP) to extend the DNA strand;
- one labeled deoxynucleotide triphosphate (using a radioactive element or dye); and
- one dideoxynucleotide triphosphate, which terminates the growing chain wherever it is incorporated. Tube A has didATP, tube C has didCTP, etc.

Dideoxy Sequencing

r example, in the A reaction tube the ratio of the
ATP to didATP is adjusted so that each tube will
ve a collection of DNA fragments with a didATP
corporated for each adenine position on the
mplate DNA fragments. The fragments of varying
ngth are then separated by electrophoresis and the
sitions of the nucleotides analyzed to determine
quence. The fragments are separated on the basis
size, with the shorter fragments moving faster and
pearing at the bottom of the gel. Sequence is read
om bottom to top.

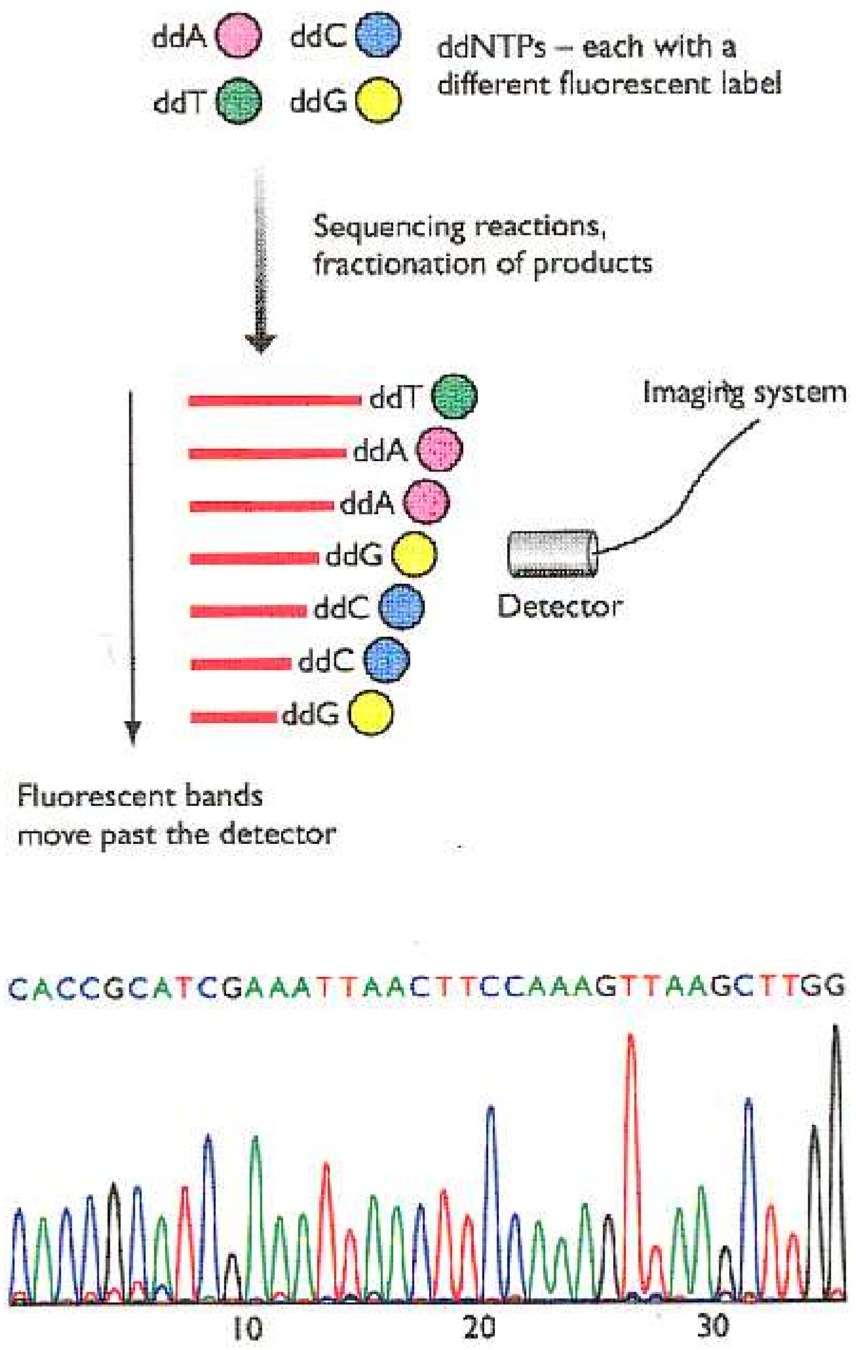

ORNL-DWG 91M-17368
1. Sequencing reactions loaded onto polyacrylamide gel for fragment separation
T
C
G
A
T C G A
2. Sequence read (bottom to top) from gel autoradiogram
T C G A
GTCGACTGCAAT

Dideoxy Sequencing



Polyacrylamide gel with small pores is used to fractionate single-stranded DNA. In the size range 10 to 500 nucleotides, DNA molecules that differ in size by only a single nucleotide can be separated

Lane 1- partial replicas terminating in G
Lane 2- partial replicas terminating in A
Lane 3- partial replicas terminating in T
Lane 4- partial replicas terminating in C


ddA
ddC
ddT
ddG
ddNTPs – each with a different fluorescent label
Sequencing reactions, fractionation of products
ddT
ddA
ddA
ddG
ddC
ddC
ddG
Imaging system
Detector
Fluorescent bands move past the detector
CACCGCATCGAAATTAACTTCCAAAGTTAAGCTTGG
10
20
30

Brown. Genomes 2

Capillary Electrophoresis

- Technique combines the use of gel-filled capillary tubes with a unique laser scanning system to sequence each of the four different types of bases -- adenine, cytosine, guanine, and thymine -- in a sample of DNA.
- Capillary array electrophoresis cuts the time of the Sanger dideoxy method down by replacing the slab with hundreds of tiny gel-filled capillaries, about 100 microns (four thousandths of an inch) in internal diameter, that can be bundled into a single array for automated detection.

http://www.lbl.gov/Science-Articles/Archive/automated-DNA-sequencing.html

Capillary Electrophoresis

- The dideoxy method was improved by the use of fluorescent labels. The primer is synthesized and split into four batches, each of which is labeled with a different fluorescent dye.
- Each dye labeled primer is used in a sequencing reaction with one of the dideoxynucleotides.
- The reaction products are pooled and analysed in a single lane of a sequencing gel. A four-colour fluorescence detector monitors the DNA as it migrates to the bottom of the gel. The fluorescence signature is used to identify the terminal nucleotide.

http://www.lbl.gov/Science-Articles/Archive/automated-DNA-sequencing.html

Capillary Electrophoresis

- A method was then developed based on this where fluorescent dideoxynucleotides were used in a single sequencing reaction. The fragments are then separated in a single lane of a sequencing gel and identified by the fluorescent signature.
- The limitations of gel electrophoresis soon became apparent. Gels take a long time to run and have a limited reproducibility. An automated method was much more desirable, but the automation of gels requires complex robotic handling.
- The use of capillaries allows much higher electrical fields to be used making the separations faster. Flexible capillaries are also easily incorporated into an automated instrument making sequencing cheap, fast and efficient.
- However a single capillary is still a bottleneck in the sequencing process. A gel is easily capable of running up to 96 samples simultaneously. To overcome this a instrument fitted with an array of capillaries was developed. In the first instruments the detector moved across the array, but the time lag means some information can be missed. Now all the capillaries are simultaneously monitored using an array of photodiodes.

http://www.chemsoc.org/exemplarchem/entries/2003/leeds_chromatography/chromatography/cge.htm

Capillary Electrophoresis

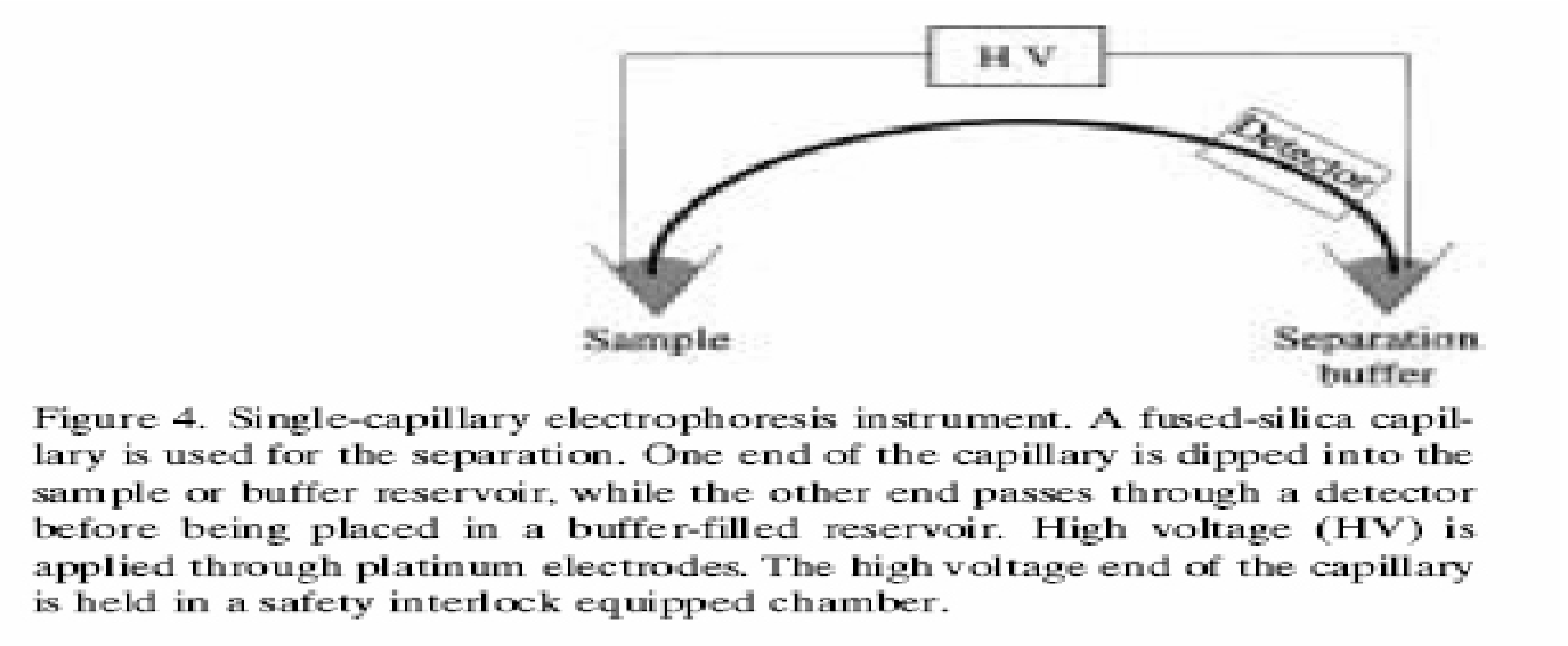


Figure 4. Single-capillary electrophoresis instrument. A fused-silica capillary is used for the separation. One end of the capillary is dipped into the sample or buffer reservoir, while the other end passes through a detector before being placed in a buffer-filled reservoir. High voltage (HV) is applied through platinum electrodes. The high voltage end of the capillary is held in a safety interlock equipped chamber.

http://elchem.kaist.ac.kr/BK21_SMS.web/2001_instanal/FIG/20011029/0003_ANG_2000_04463_DNA_CAE.pdf

Capillary Electrophoresis

- Capillary holds a sieving medium, which allows separation of DNA fragments based on their size
- Sample is injected into the capillary by placing the end of the capillary in the sample solution, and applying electric current.
- After injection is complete, the sample is replaced by running buffer and electric field is reapplied to drive the samples through the capillary.
- Laser-induced fluorescence detector near the end of the capillary records the fluorescent signal in four different channels to resolve the fluorescent signal from the four dyes.

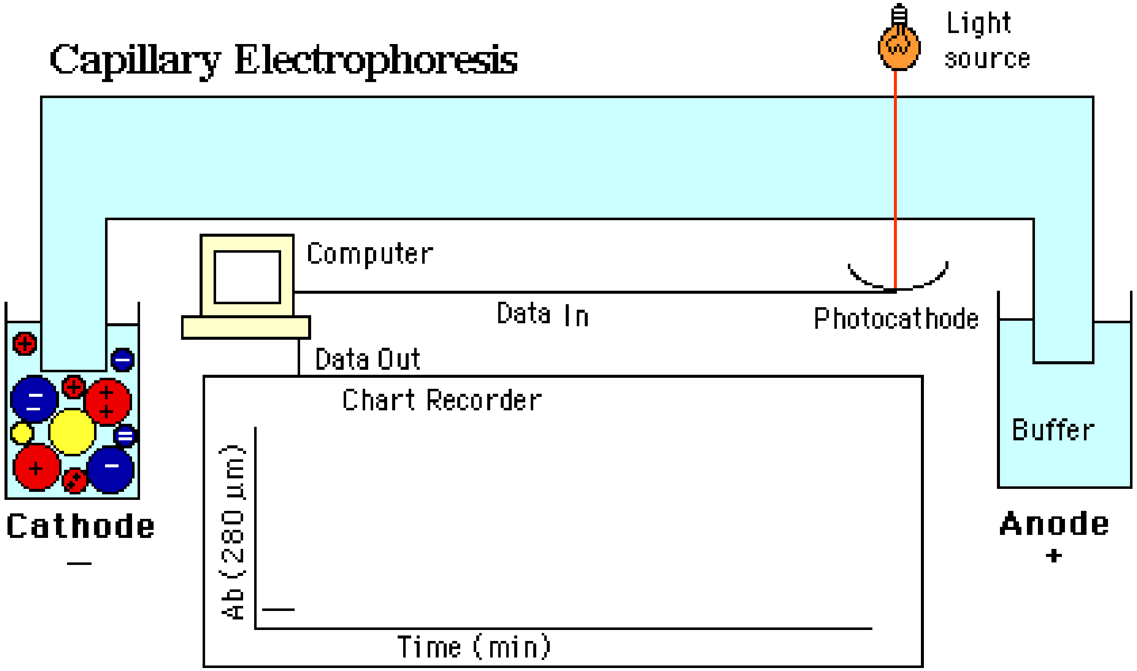
http://elchem.kaist.ac.kr/BK21_SMS.web/2001_instanal/FIG/20011029/0003_ANG_2000_04463_DNA_CAE.pdf

Capillary Electrophoresis

- Small inner diameter of capillary tube reduces Joule heating to negligible levels, allowing the use of extremely high voltage, for rapid sequencing
- Provides a 2-fold improvement over slab gel sequencing methods

http://elchem.kaist.ac.kr/BK21_SMS.web/2001_instanal/FIG/20011029/0003_ANG_2000_04463_DNA_CAE.pdf

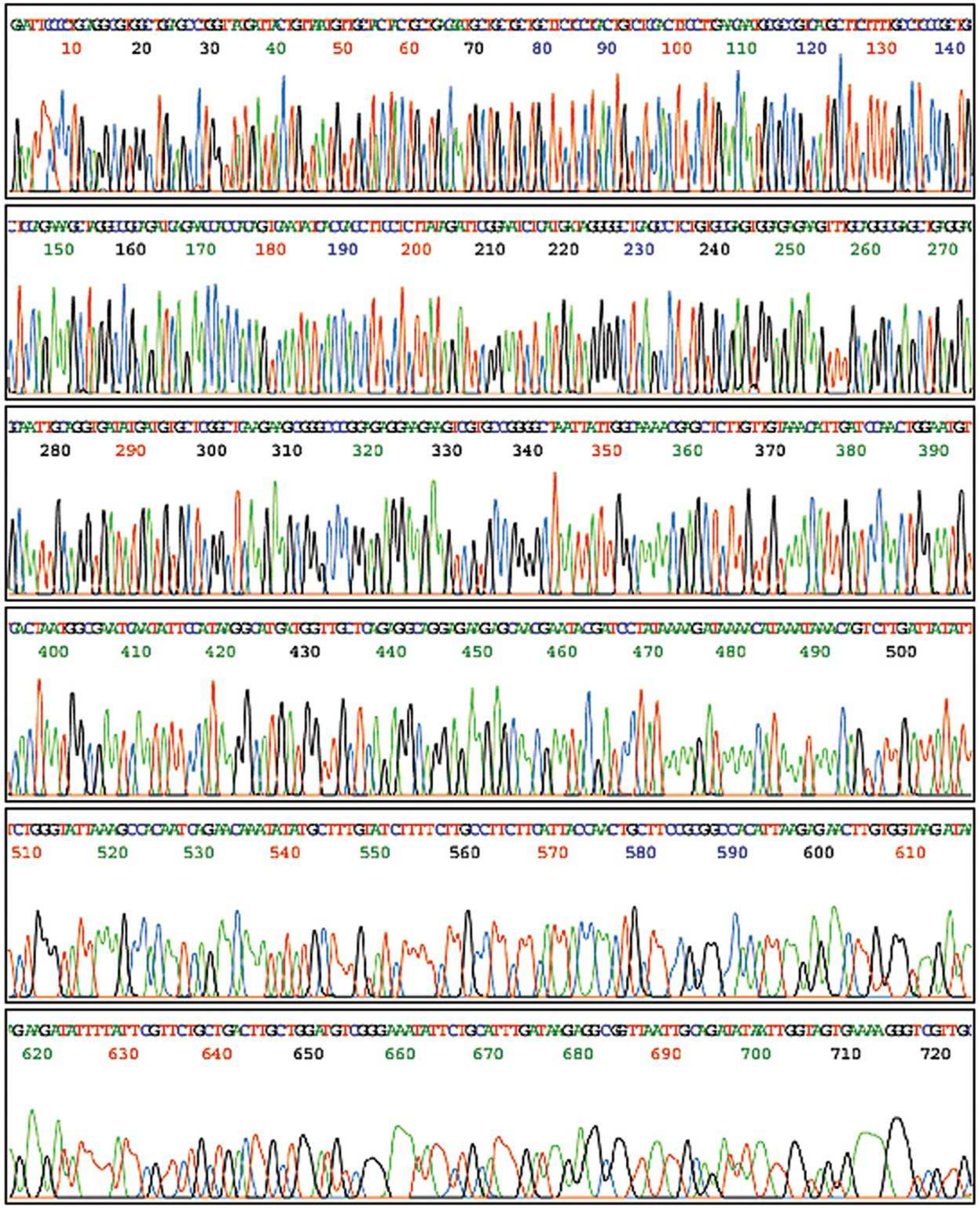
Capillary Electrophoresis



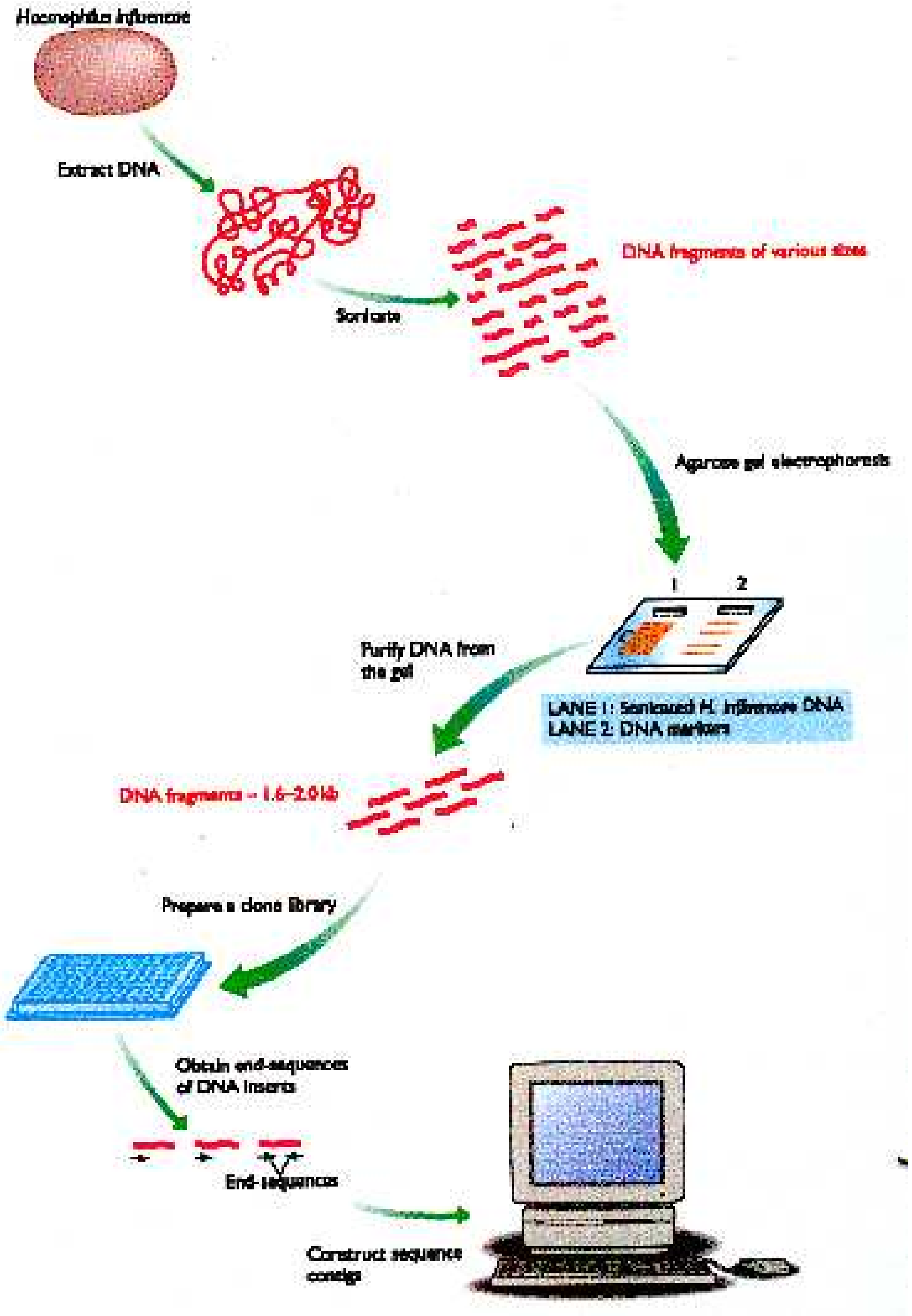
http://ntri.tamuk.edu/ce/ce.html

Automated Sequencing

http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v10n3/images/megabaces.jpg



http://www.ornl.gov/sci/techresources/Human_Genome/education/images.shtml


Haemophilus influenzae
Extract DNA
Sonicate
DNA fragments of various sizes
Agarose gel electrophoresis
1
2
Purify DNA from the gel
LANE 1: Sonicated H. influenzae DNA
LANE 2: DNA markers
DNA fragments – 1.6–2.0 kb
Prepare a clone library
Obtain end-sequences of DNA inserts
End-sequences
Construct sequence contigs

Brown. Genomes 2

mRNA Sequencing

- An alternative strategy for sequencing was presented during the early 90’s by Craig Venter.
- The functional portion of the human DNA supposedly accounts for less than 10% (perhaps less than 5%) of the entire human genome.
- The HGP strategy, i.e. sequencing everything, could be considered as lavishness with resources since the main part of the information lacks relevance.
- The scope of the strategy by Venter and co-workers is to focus the investigations to messenger ribonucleic acid (mRNA) instead of DNA. The point of using mRNA is that it does not include any non-coding DNA. The mRNA molecule can be isolated and used as a template to synthesize a complementary DNA (cDNA) strand, which can then be used to locate the corresponding genes on a chromosome map.
- By using Venter´s method an incomplete copy of the gene, called expressed sequence tag (EST), is acquired. ESTs are useful for localizing and orienting the mapping and sequence data reported from many different laboratories and serve as landmarks on the developing physical map of the human genome.

http://wcn.ntech.se/history.html

Whole Genome Shotgun Sequencing

- Venter and co-workers success in 1995 of sequencing the H. influenzae bacterium introduced a method called "whole genome shotgun sequencing". The shotgun method involves randomly sequencing tiny cloned sections of the genome, with no foreknowledge of where on a chromosome the section originally came from.
- The partial sequences obtained are then reassembled to a complete sequence by use of computers. The advantage with this method is that it eliminates the need for time-consuming mapping.

- By competing (and cooperating) the governmentally financed human genome project (HGP) and the private biotechnology company Celera has completed a reference DNA sequence of the human genome. Both parties made their information simultaneously available in February 2001, by publishing it in on the Internet and in the scientific journals Nature and Science.

http://wcn.ntech.se/history.html

Whole Genome Shotgun Sequencing

- After the sequence is shotgunned the 10 million fragments of the genomic jigsaw puzzle need to be recompiled into the readable base pairs in the proper order.
- This method will be completed using a 10x redundancy to eliminate errors and reduce the possibility of having misses any targeted regions.
- The Celera Assembler is one of the core competencies and makes this Herculean task possible.
- The first pass through the data the shotgunned fragments are compared against each other and equivalent sequences greater than 40 base pairs long identified.
- These 40 base pairs matches are statistically impossible to occur by chance. These matches are then determined to be true or repeat induced. True matches are overlapping sections and are the desired fragments; repeat-induced fragments occur in multiple locations of the genome and do not belong together.

http://www-management.wharton.upenn.edu/pennings/coursedocuments/executive_education_courses/557/genome%20Paper%201.doc

Whole Genome Shotgun Sequencing

- The assembler then searches for overlapping fragments that have a common sequence and are not contested elsewhere in the dataset.
- The uncontested data is assembled into unitigs containing approximately 30 fragments.
- These assembled unitigs are 99 % accurate and repeats are filtered out using the Discriminator algorithm.
- Unitigs passing this filter are identified and renamed U-untigs that are ready for ordering.
- The scaffolding stage starts and the order found by looking at the mate pairs and organizing these into contigs. By constantly looking at these contigs and looking at the orientation the scaffold become complete except for some sequencing gaps.
- This strategy is repeated until the gaps are filled using the Discriminator algorithm and a method using sequence “rocks” and “pebbles”.

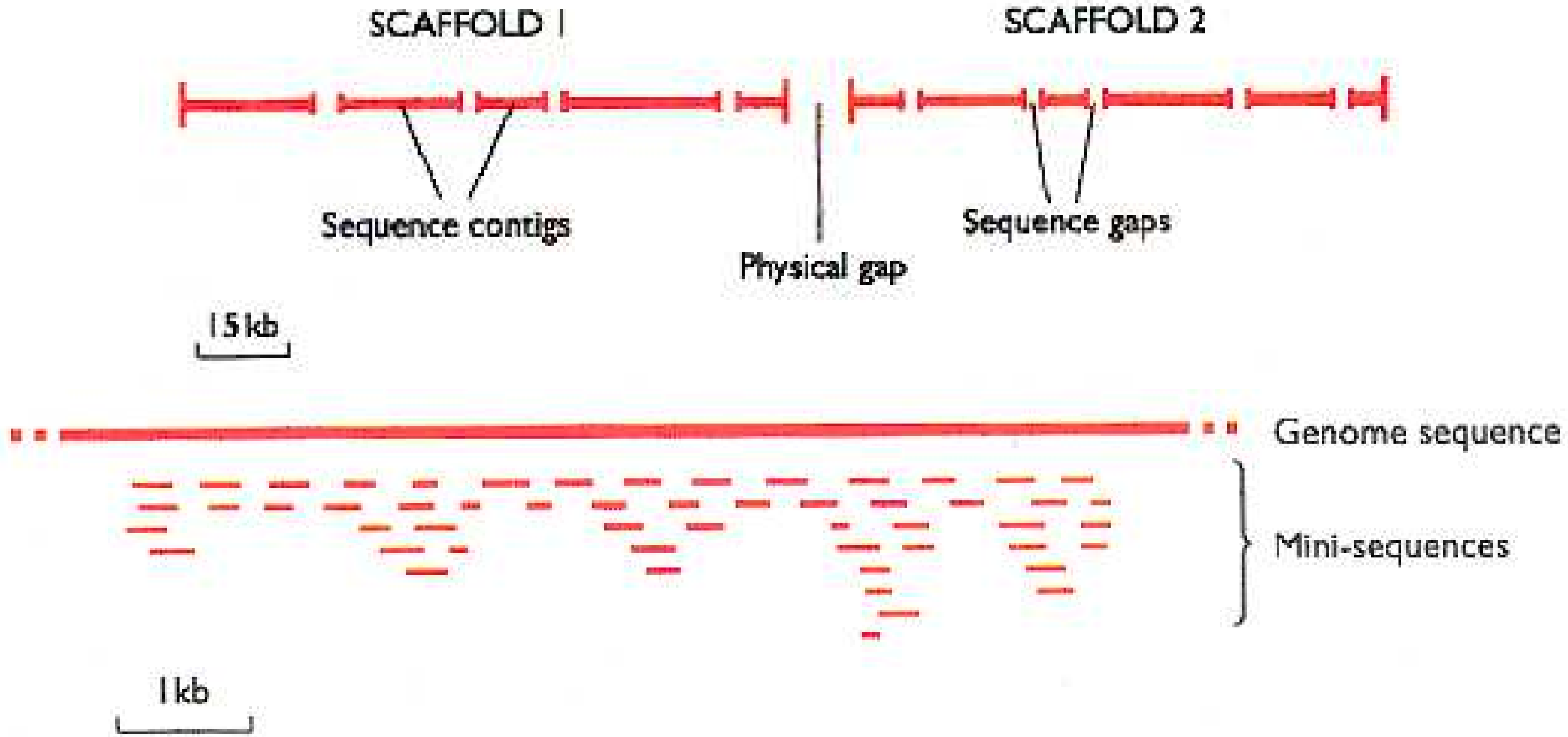
http://www-management.wharton.upenn.edu/pennings/coursedocuments/executive_education_courses/557/genome%20Paper%201.doc

Whole Genome Shotgun Sequencing

- As HGP has been making public the incremental sequence the shotgun approach utilized this data to help eliminate errors and speed the scaffolding process.

http://www-management.wharton.upenn.edu/pennings/coursedocuments/executive_education_courses/557/genome%20Paper%201.doc

Sequence Gaps

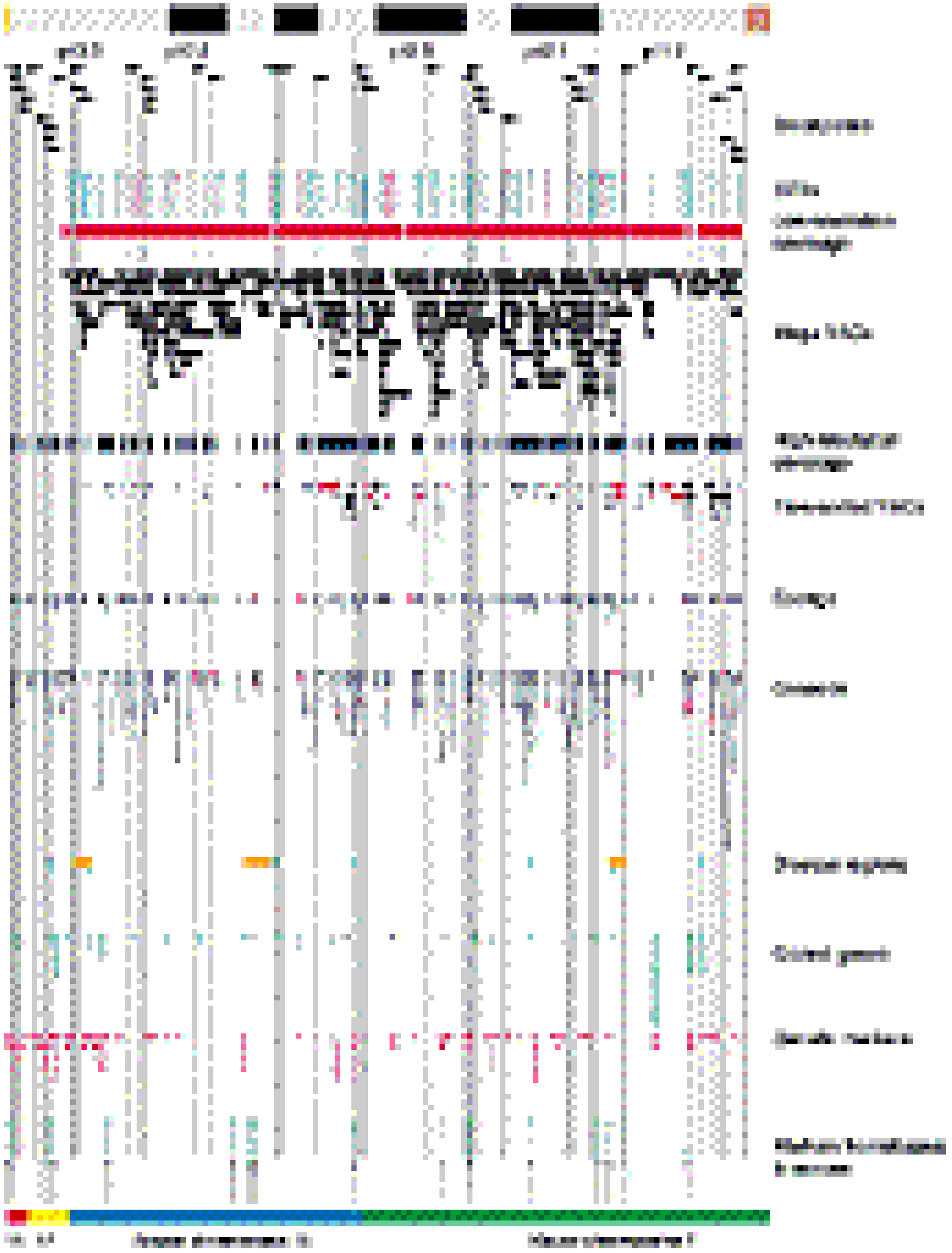


Brown. Genomes 2

Advances

- The following advances in robotics and automation reduced the labor by 80% while combining the microbiological advances:
  - Development of Perkin-Elmer (ABI PRISM 3700) gene sequence.
  - 1000 sample per day
  - 15 minutes instead of 8 hours for first automated sequencers
  - A parallel system of 300 sequencers ($300,000 each)
  - Use of supercomputers to assemble fragments
- Development of process support instrumentation to process 100 K template preps and 200 K sequence reactions per day.
- 24 hour per day unattended operation of sequencers

http://www-management.wharton.upenn.edu/pennings/coursedocuments/executive_education_courses/557/genome%20Paper%201.doc



Map of Chromosome 16

http://www.ornl.gov/sci/techresources/Human_Genome/education/images.shtml

Advances

- In addition to the above advances the field of computational biology (bioinformatics) became increasingly important as the software and processors required to assemble a puzzle of this size still needed to be developed.
- The solution came in advances in processor speeds that have doubled every 18 months and the development of better overlap detection algorithms. It is expected that using 100 networked workstations the entire genome could be assembled in 30-60 computational days.

http://www-management.wharton.upenn.edu/pennings/coursedocuments/executive_education_courses/557/genome%20Paper%201.doc