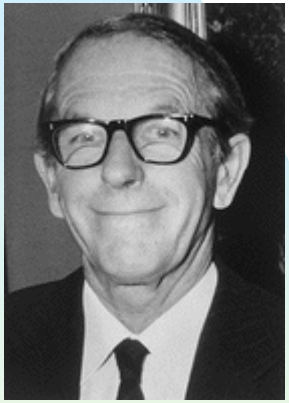
Genome Sequence Analysis

Ping Xu
The Philips Institute,
Virginia Commonwealth University

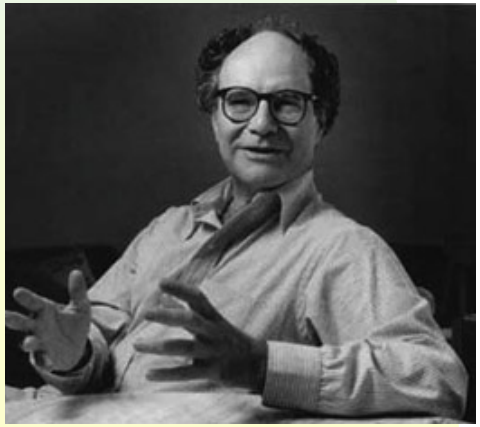
For Bioinformatics and Bioengineering Summer Institute (BBSI)

Pioneers of sequencing



Late 70's - First DNA Sequencing Technologies

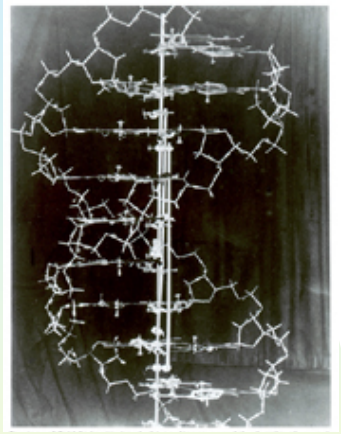
- Fred Sanger (Cambridge)- Enzymatic sequencing
- Walter Gilbert (Harvard)- Chemical sequencing

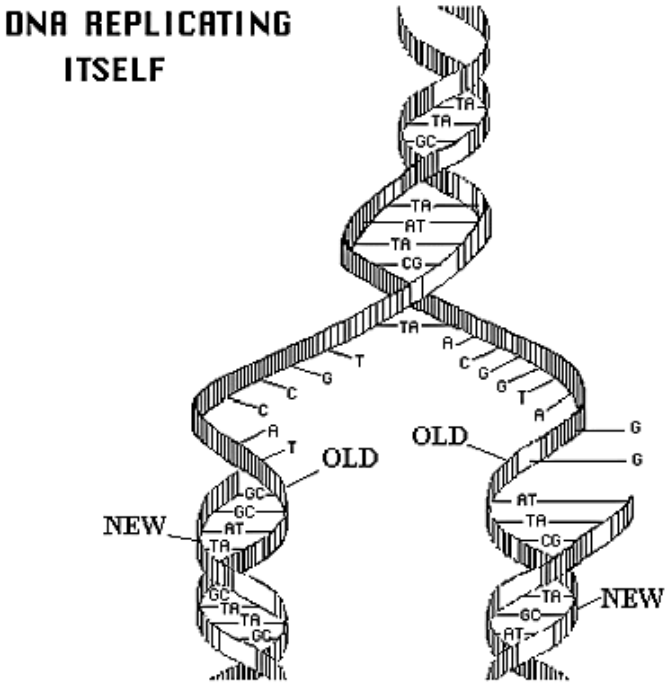


Shared Nobel Prize in Chemistry 1980

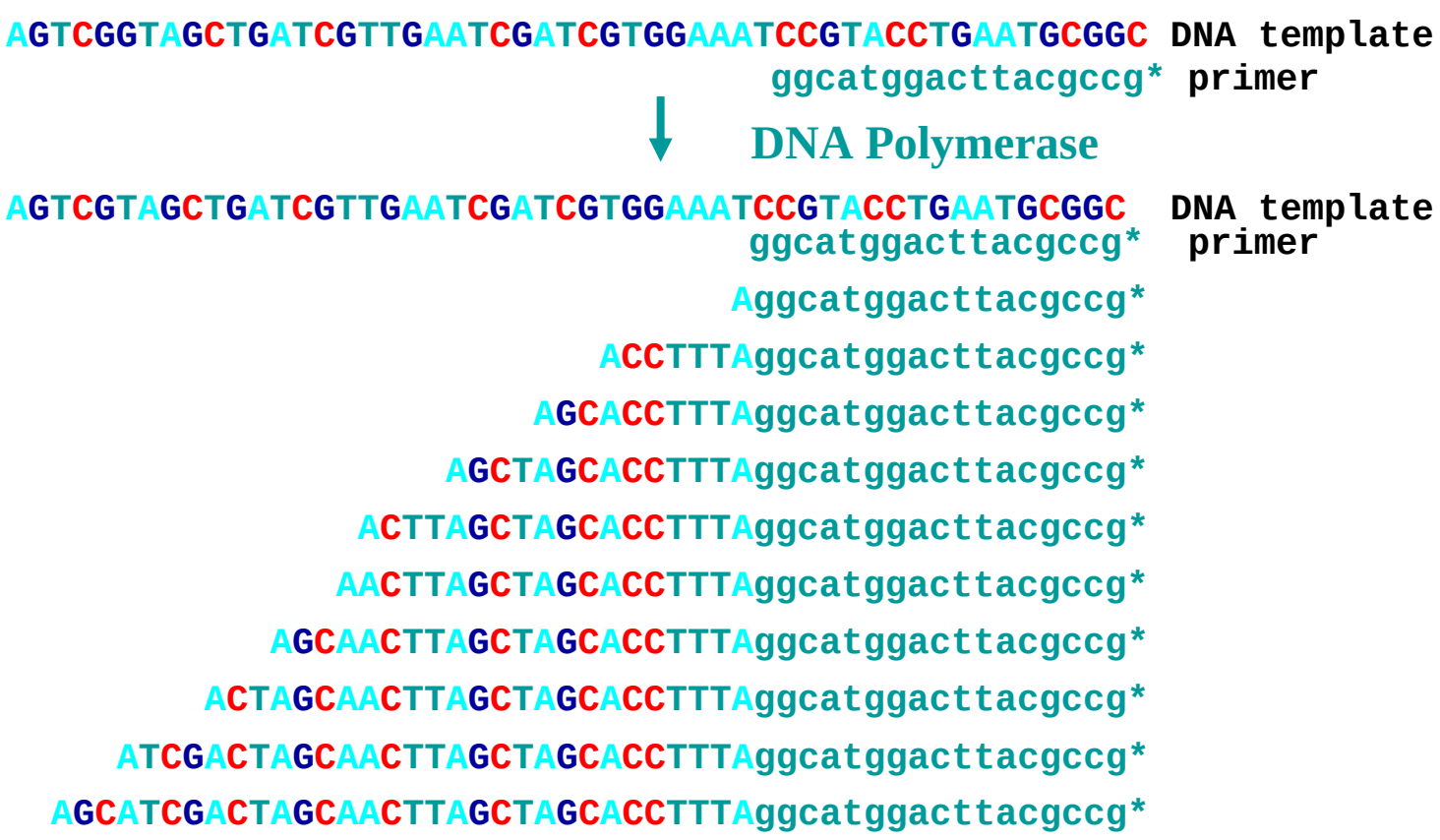
(with Paul Berg)

Enzymatic sequencing





Enzymatic sequencing

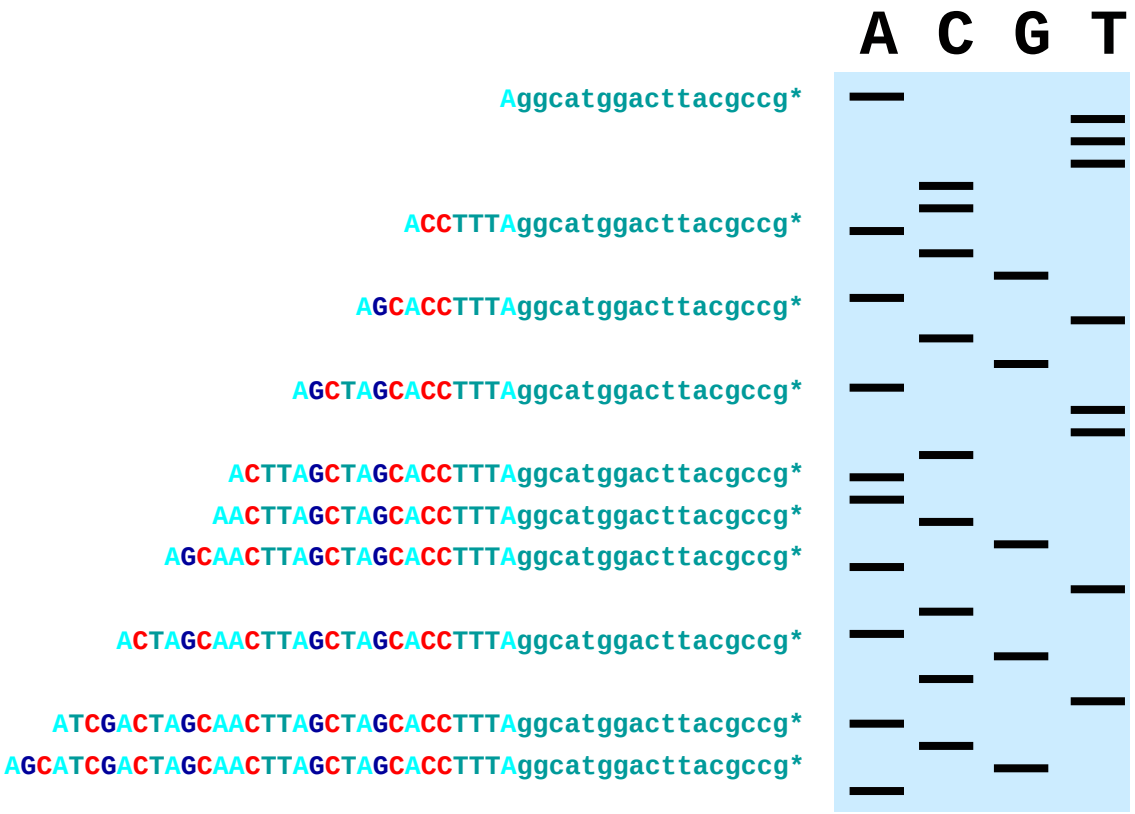


This is the ‘A’ tube reaction.
The ‘G’, ‘C’, and ‘T’ tube reactions must be run in parallel.

Enzymatic sequencing

"**A**" **tube:** all four dNTP's, ddATP and DNA polymerase
"**C**" **tube:** all four dNTP's, ddCTP and DNA polymerase
"**G**" **tube:** all four dNTP's, ddGTP and DNA polymerase
"**T**" **tube:** all four dNTP's, ddTTP and DNA polymerase

Enzymatic sequencing



Acrylamide gel electrophoresis,

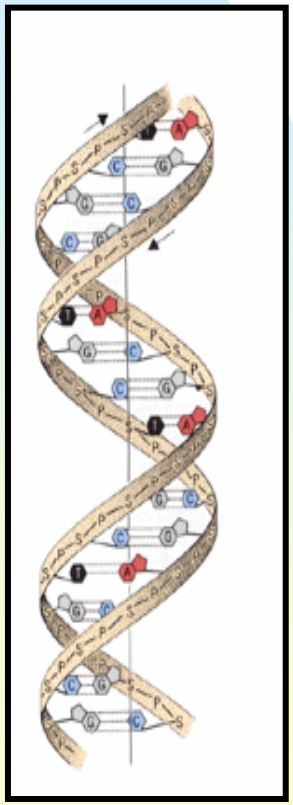
Autoraodiography on X-Ray film

Enzymatic sequencing

It involves:

- The DNA template.
- Labeled (either radioactive or florescent) specific primer or labeled ddNTPs to label synthesized DNA ends.
- dNTPs.
- DNA polymerase to syntheses.
- Base specific chain termination by 2', 3' ddNTPs because they lack a hydroxyl residue at the 3' position of deoxyribose.
- The use of polyacrylamide gel (or capillary) to separate single-stranded DNA chains in differing in length by a single nucleotide

Automated DNA sequencing



Mid 80's - Two major advances:

- Fluorescent sequencing
  - -replace radiolabel with dye
  - - loss of sensitivity
- Polymerase chain reaction (PCR)
  - - amplifies signal (sensitivity)
  - - permits fluorescent labels

PCR Reaction

Fluorescent DNA sequencing

Involves:

- The use of a specific primer for extension by Taq DNA polymerase.
- Fluorescent end labeling DNA by dye primer or dye terminator
- Base-specific chain termination by 2',3' ddNTPs
- The use of polyacrylamide gel or capillary with fluorescent detector to fraction DNA

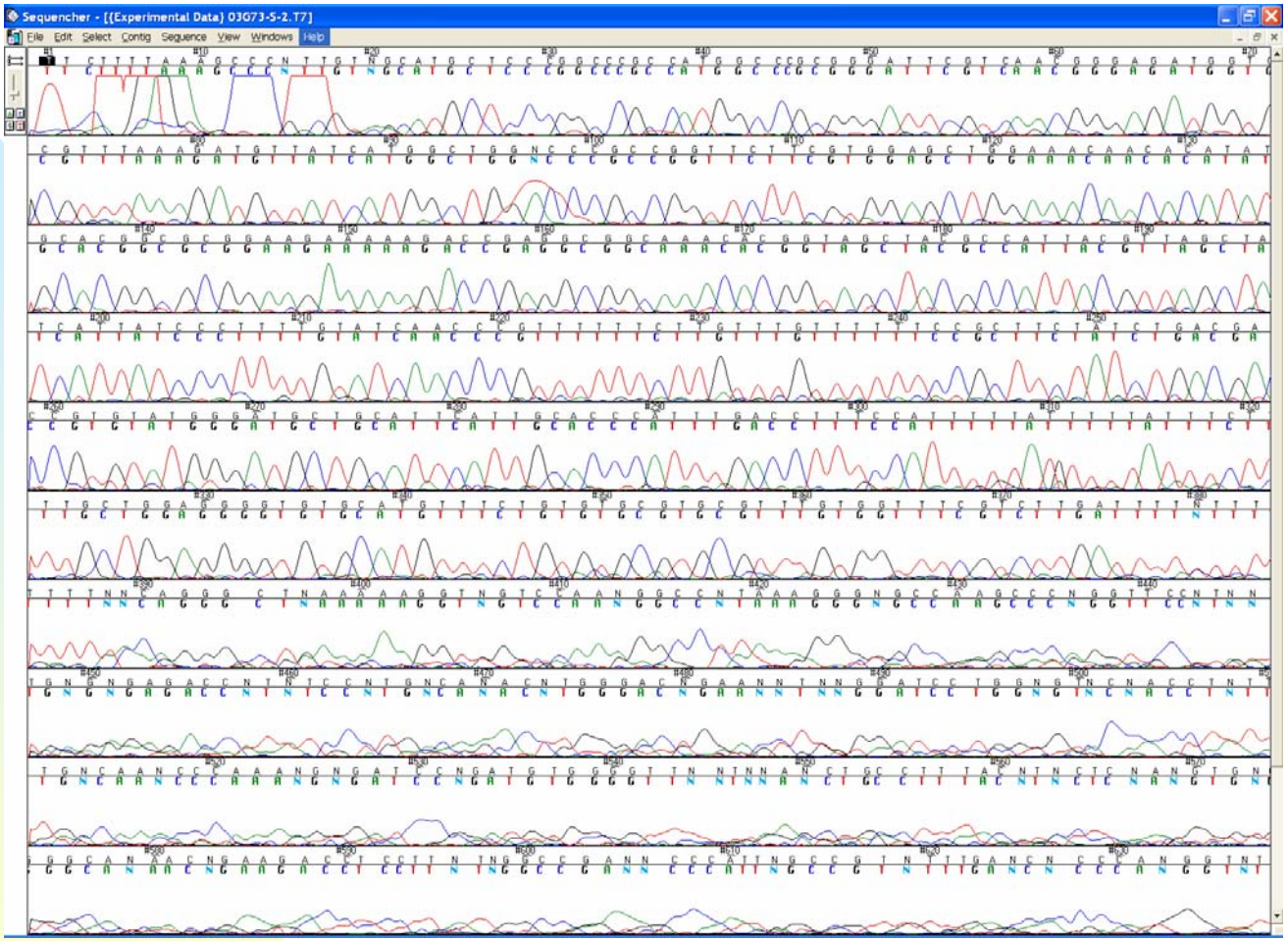
Labeling method:

- Dye primer (four separating reaction/one lane)
- Dye terminator (one reaction/one lane)

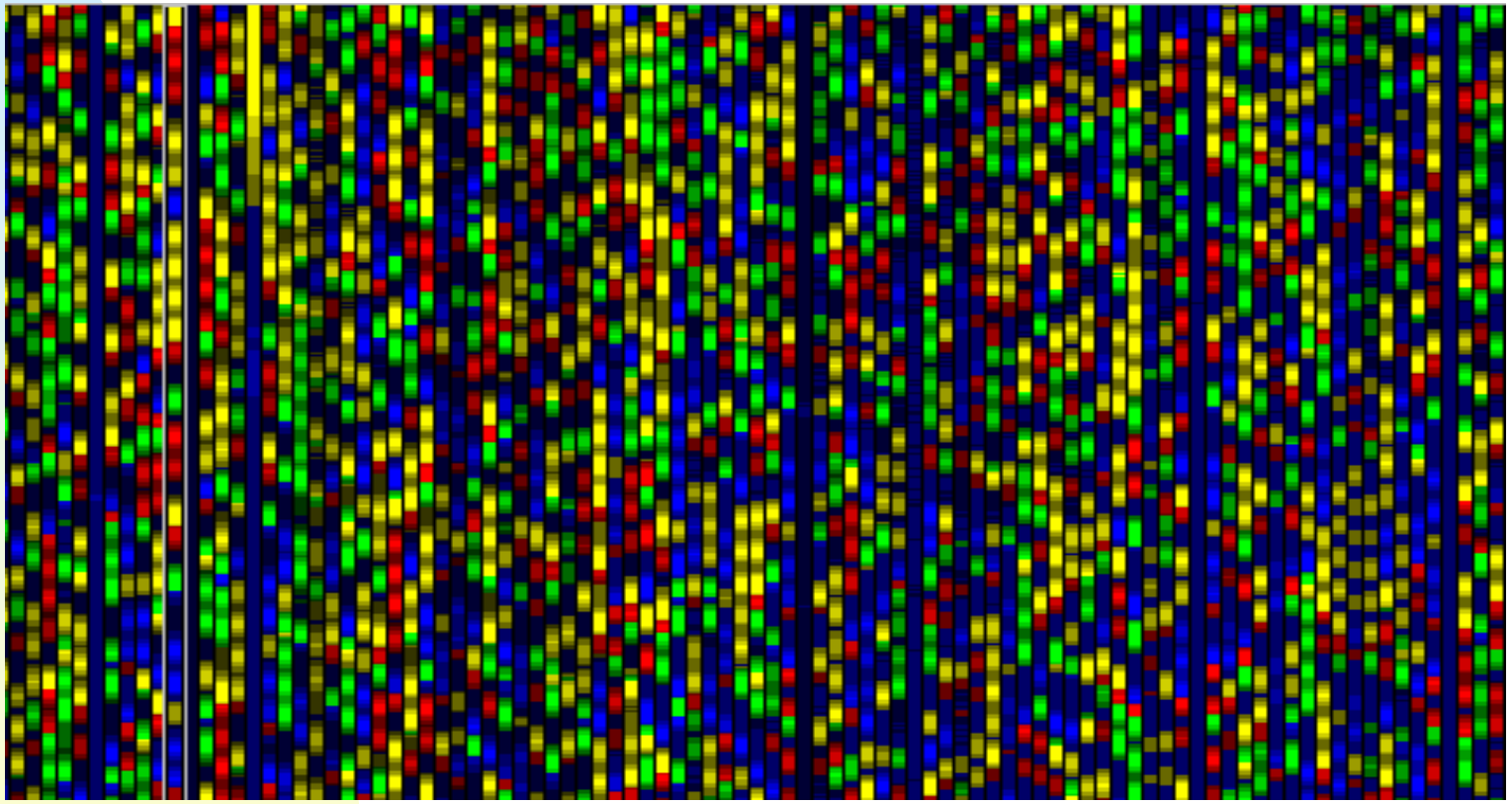
Disadvantage:

- The fluorescent color is less sensitive than radioactive labeling. The method needs more products from synthesis with Taq DNA polymerase for multiple cycle synthesis

Fluorescent Sequencing

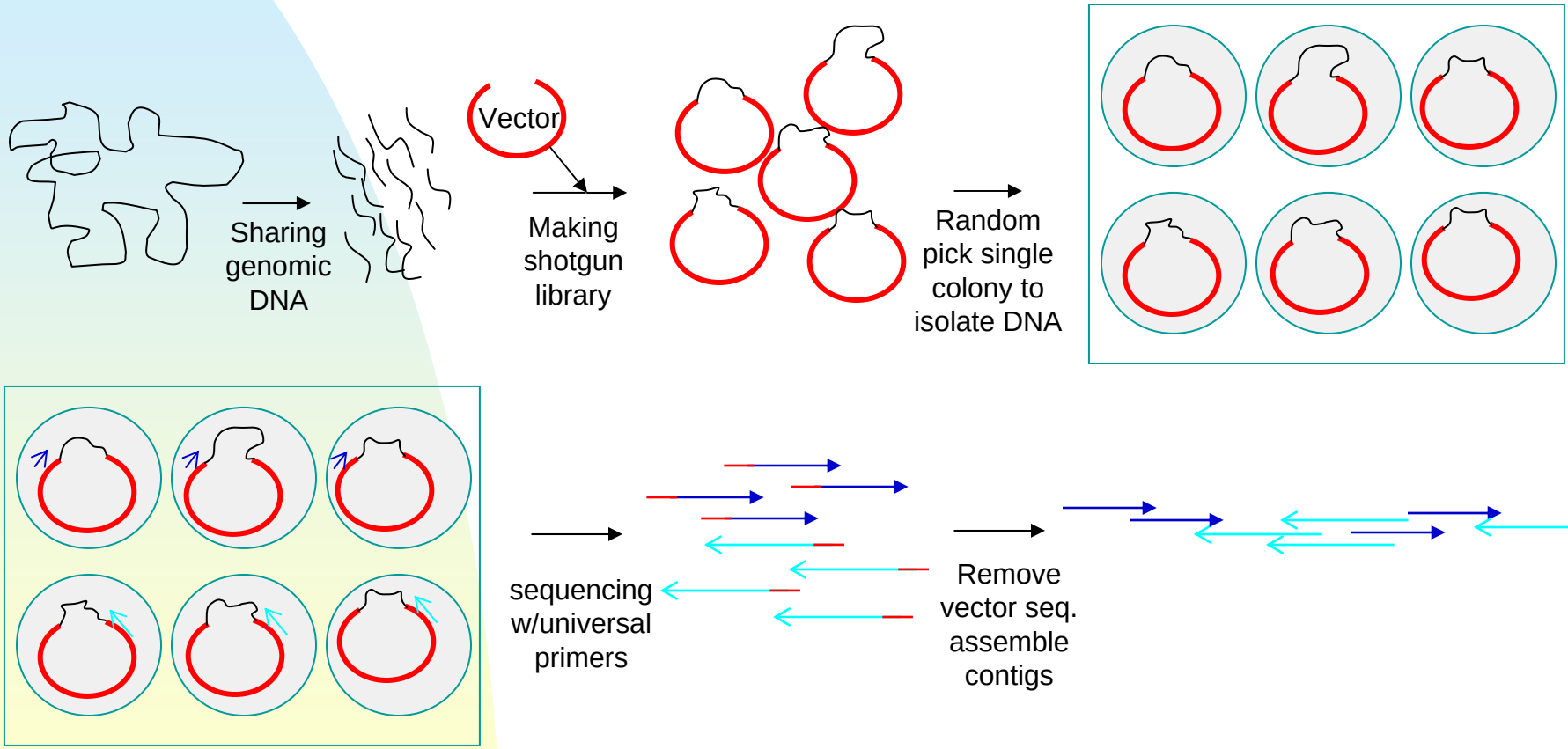

Sequencher - [(Experimental Data) 03G73-S-2.T7]
File Edit Select Contig Sequence View Windows Help

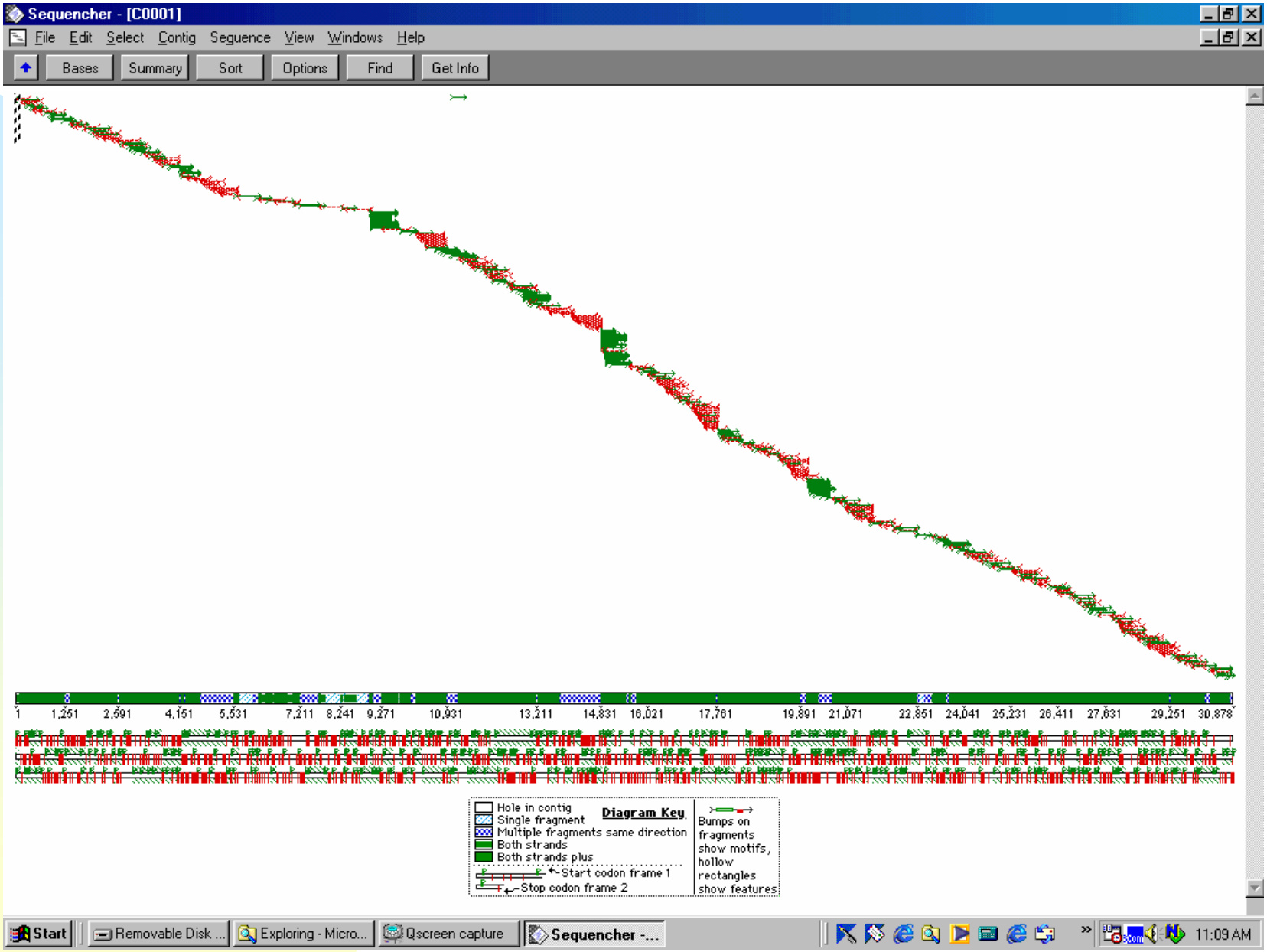
ABI 3700 Sequencer



Raw data from ABI 3700 Prism Sequencer

Shotgun Sequence Strategy




Sequencher - [C0001]
File Edit Select Contig Sequence View Windows Help
Bases Summary Sort Options Find Get Info
1 1,251 2,591 4,151 5,531 7,211 8,241 9,271 10,931 13,211 14,831 16,021 17,761 19,891 21,071 22,851 24,041 25,231 26,411 27,631 29,251 30,878
Diagram Key
Hole in contig
Single fragment
Multiple fragments same direction
Both strands
Both strands plus
Start codon frame 1
Stop codon frame 2
Bumps on fragments show motifs, hollow rectangles show features
Start
Removable Disk ...
Exploring - Micro...
Qscreen capture
Sequencher -...
11:09 AM

Sequence Assemble Excise

CTGTTGGTCGAGATCCTCA

AAGACCATTT

TAGGACTGTTACAAAAGCTCAAG

GACCATTTTAATCATTCAT

CCTCTCAGGCTGATGTCA

AAAGCTCAAGAAAGAAGGA

CTGATTGTAGGAC

TTGATATTATCAGTGAG

ACATAGTATCGTTCCTGTTG

GGAAAGACCATTTTAATC

TTAATCATTCATCATGAC

ATCAGTGAGGAACTGAT

TCCTGAAGATGTTA

GAGATCCTCACGAATA

TGTTAAAAAGACCTCTC

TCCTCACGAATATGAGC

GAATATGAGCCTCTTCCTGA

GCCACGAAATTCACGAACA

AGGAACTGATTGTAGGACT

TCCTGAAGATGTTA

GAATATGAGCCTCTTCCTGA

GCTTCCATGGATGTTTATTAAGGG

TTTGAAATTTTCGCAT

CTGTTGGTCGAGATCCTCA

AAAGTTTTCGCATAAAT

ATAAATAAAGC

TTCGCATAAATAAAGCTTCC

TCCTCACGAATATGAGC

TCGCATAAATAAAGCTTCCATGCAT

GAGATCCTCACGAATA

CCTCTCAGGCTGATGTCA

TGTTAAAAAGACCTCTC

ACATAGTATCGTTCCTGTTG

GCCACGAAATTCACGAACA

CATAAATAAAGCTTCCA

CCATGCATGCTTGAAAGTTTTC

Contig 1

ACATAGTATCGTTCCTGTTG
CTGTTGGTCGAGATCCTCA
GAGATCCTCACGAATA
TCCTCACGAATATGAGC
GAATATGAGCCTCTTCCTGA
TCCTGAAGATGTTA
TGTTAAAAAGACCTCTC
CCTCTCAGGCTGATGTCA

Contig 2

TTGATATTATCAGTGAG
ATCAGTGAGGAACTGAT
AGGAACTGATTGTAGGACT
CTGATTGTAGGAC
TAGGACTGTTACAAAAGCTCAAG
AAAGCTCAAGAAAGAAGGA
GGAAAGACCATTTTAATC
AAGACCATTT
GACCATTTTAATCATTCAT
TTAATCATTCATCATGAC

Extra

GCCACGAAATTCACGAACA

Contig 1

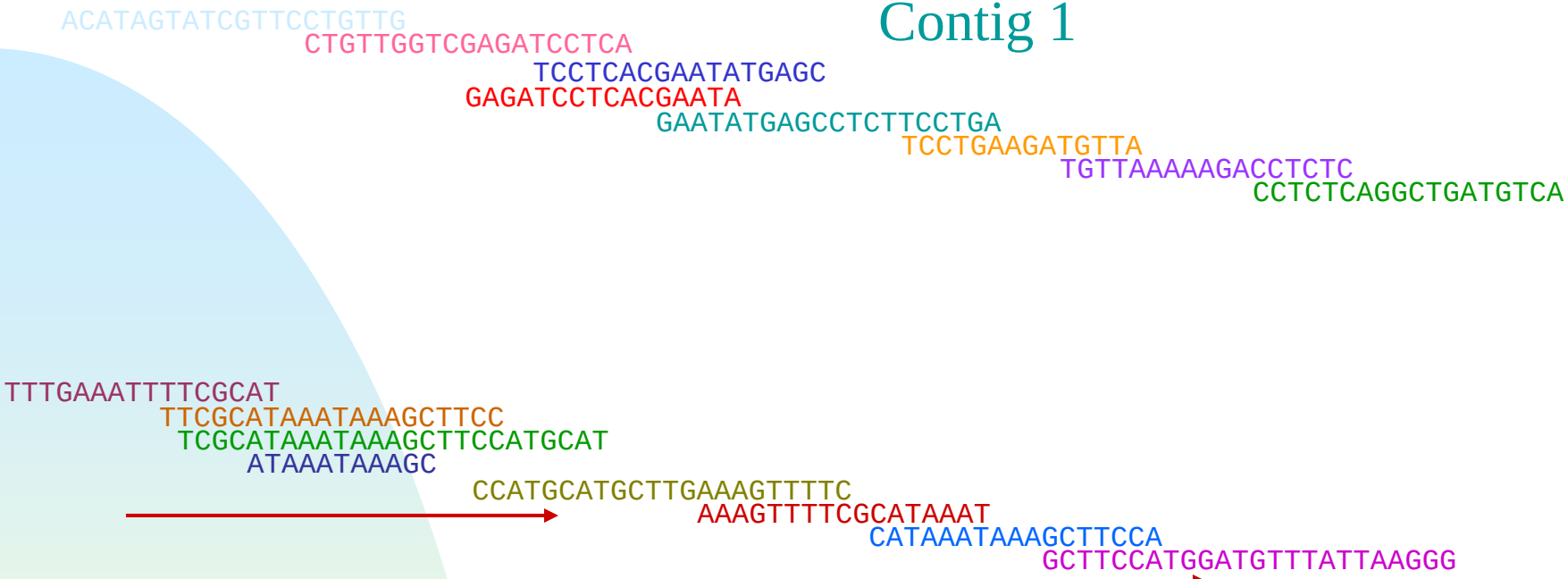
ACATAGTATCGTTCCTGTTG
CTGTTGGTCGAGATCCTCA
GAGATCCTCACGAATA
TCCTCACGAATATGAGC
GAATATGAGCCTCTTCCTGA
TCCTGAAGATGTTA
TGTTAAAAAGACCTCTC
CCTCTCAGGCTGATGTCA

TTTGAAATTTTCGCAT
TTCGCATAAATAAAGCTTCC
TCGCATAAATAAAGCTTCCATGCAT
ATAAATAAAGC
CCATGCATGCTTGAAAGTTTTC
AAAGTTTTCGCATAAAT
CATAAATAAAGCTTCCA
GCTTCCATGGATGTTTATTAAGGG

Contig 3

Extra

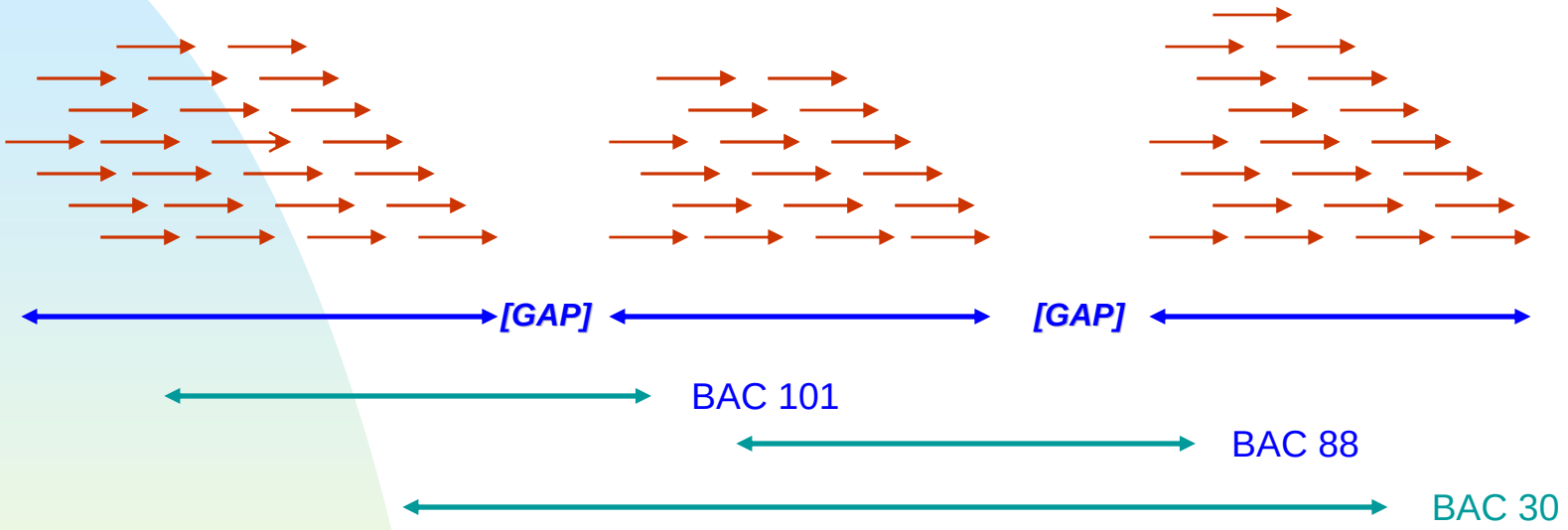
GCCACGAAATTCACGAACA



Extra

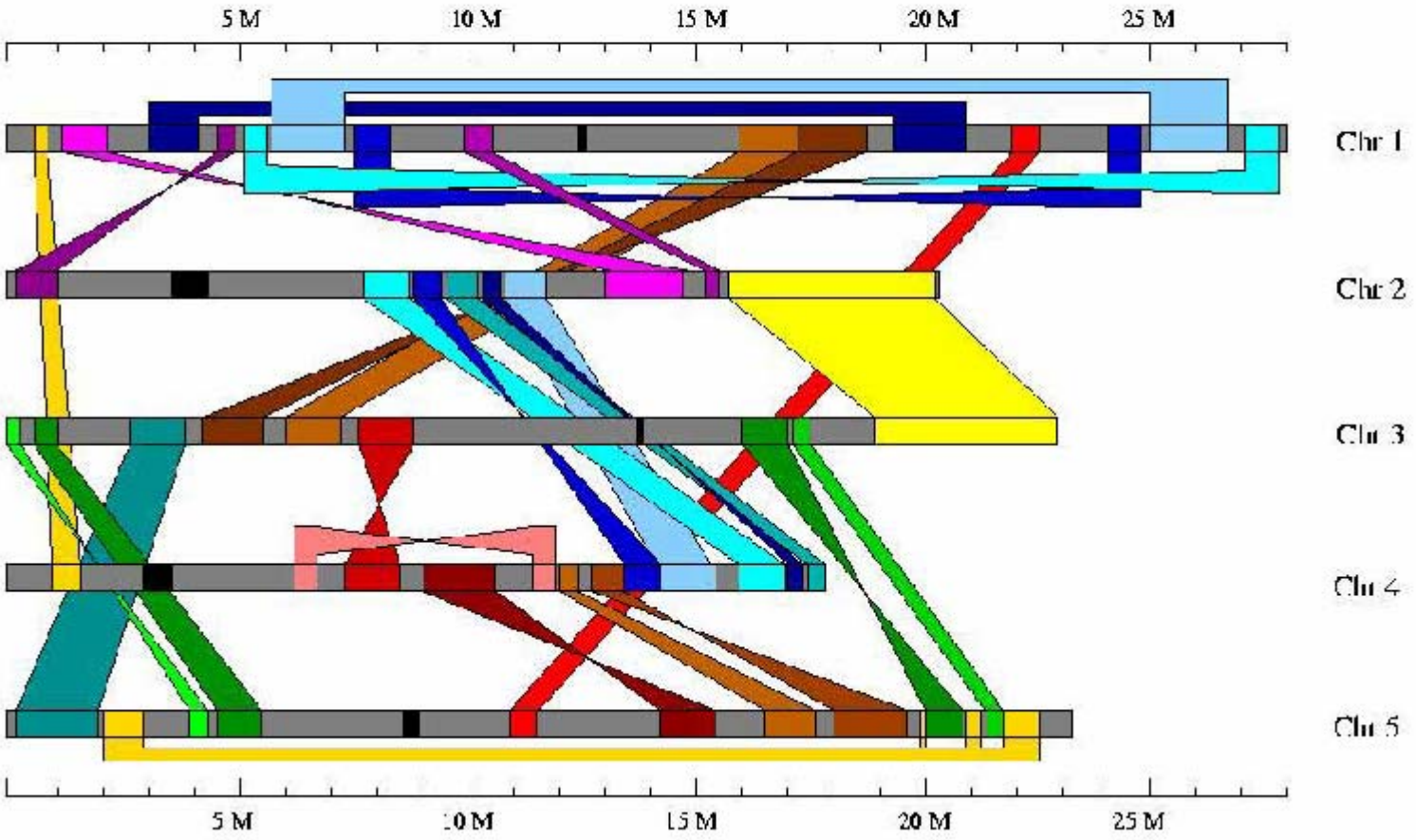
GCCACGAAATTCACGAACA

Gap Closure Strategies



Shotgun contigs aligned by large insert clones (BAC101 and BAC30 selected)

Segmental duplications



From Rob Martienssen
Cold Spring Harbor Laboratory

Hybrid strategy for complete genome sequence

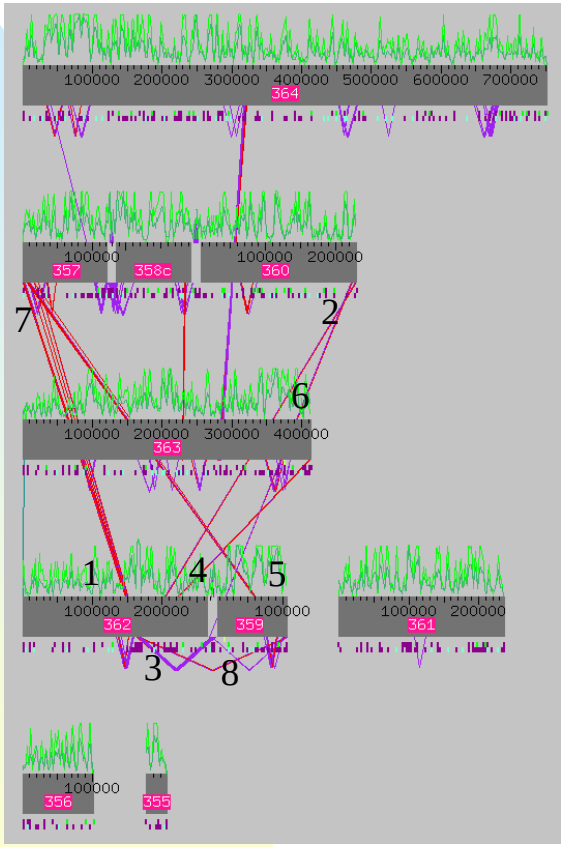
1. Combine shotgun sequencing with genome mapping
2. Shotgun sequencing to 10 X coverage
3. BAC map clones
4. Combine some percentage of sequencing of mapped clones with shotgun sequences
5. Overlay whole genome reads on reads from mapped clones when completed

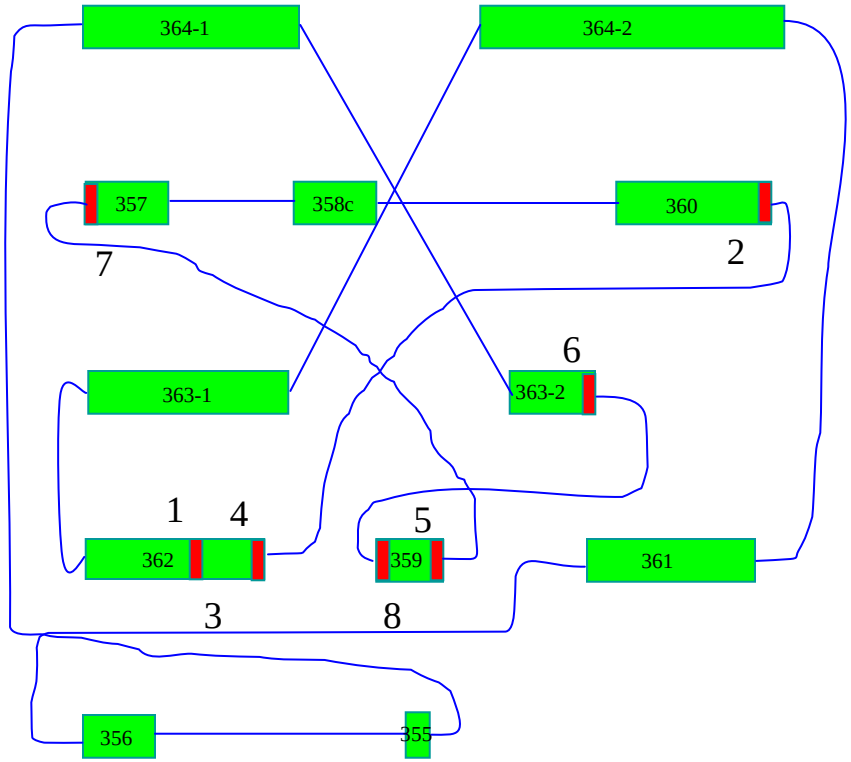
Finishing

Finishing is the process of assembling and refining raw sequence data into a highly accurate final genomic sequence. There are five finishing goals:

1. Filling gaps.
2. Address regions of low sequence quality that may contain sequence errors.
3. Add coverage to single-clone regions
4. Examine high quality discrepancies.
5. Confirm the sequence by comparing the restriction map *in silicon* to a real restriction fingerprint.

Close gaps


100000 200000 300000 400000 500000 600000 700000
364
100000
357
358c
100000 200000
360
7
2
6
100000 200000 300000 400000
363
1
4
5
100000 200000
362
100000
359
100000 200000
361
3
8
100000
356
355


364-1
364-2
357
358c
360
7
2
6
363-1
363-2
1
4
5
362
359
361
3
8
356
355

First Contact...

Haemophilus influenzae:

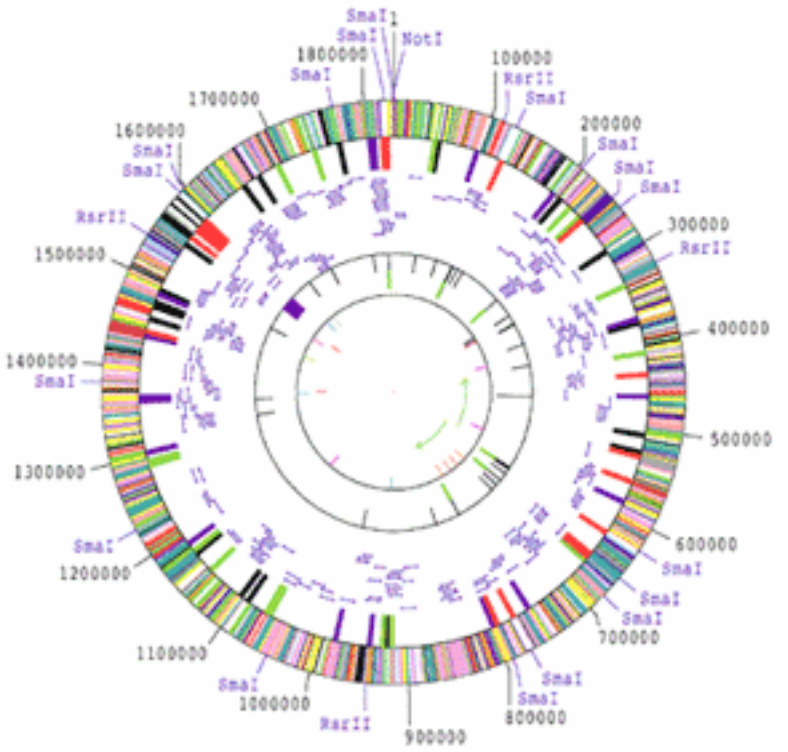
- circ. chromosome

- 1.8 million bp

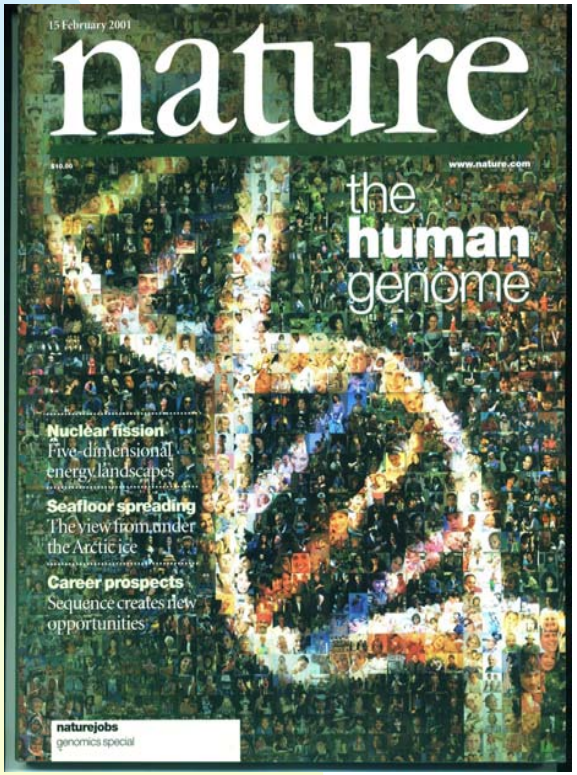
- < 2000 genes

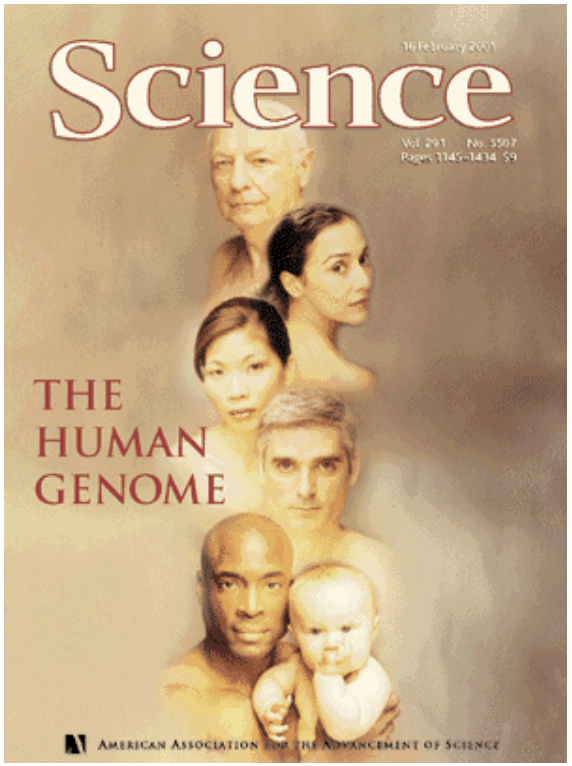
First complete genome of a living organism.

TIGR - 1995



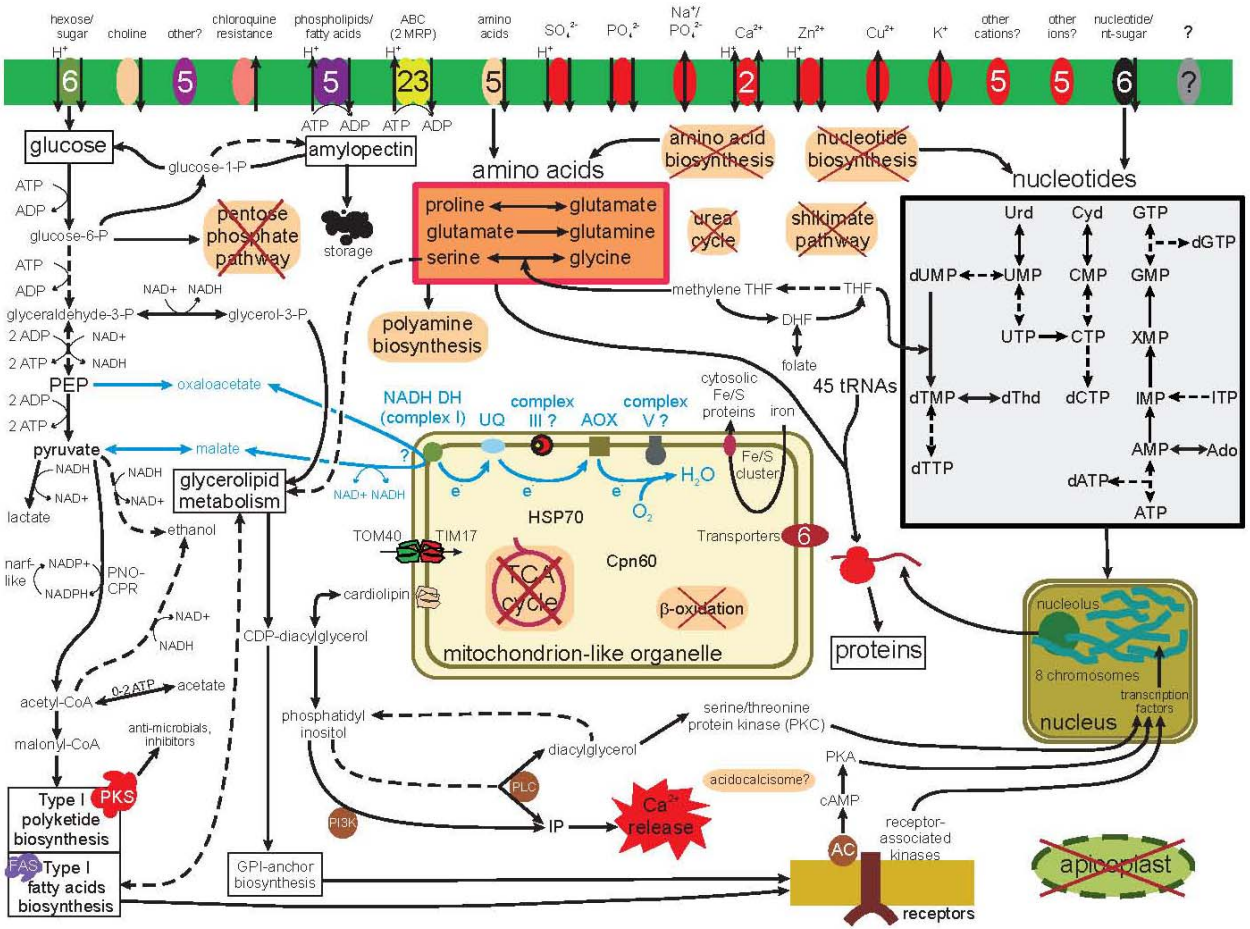
The First Human Genome Draft...





Public Effort, Nature

Private Effort, Science



Xu et al., Nature. 2004 Oct

Importance of Streptococci in human health

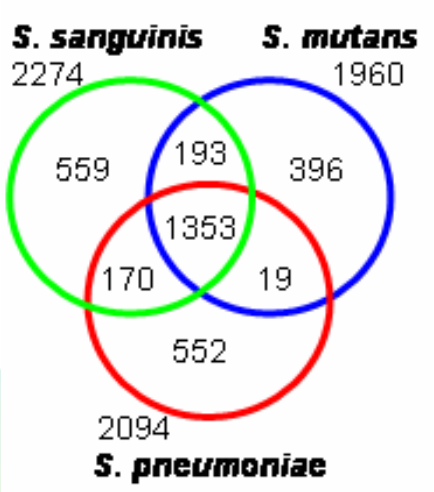
S. agalactiae causes sepsis, pneumonia, and meningitis.

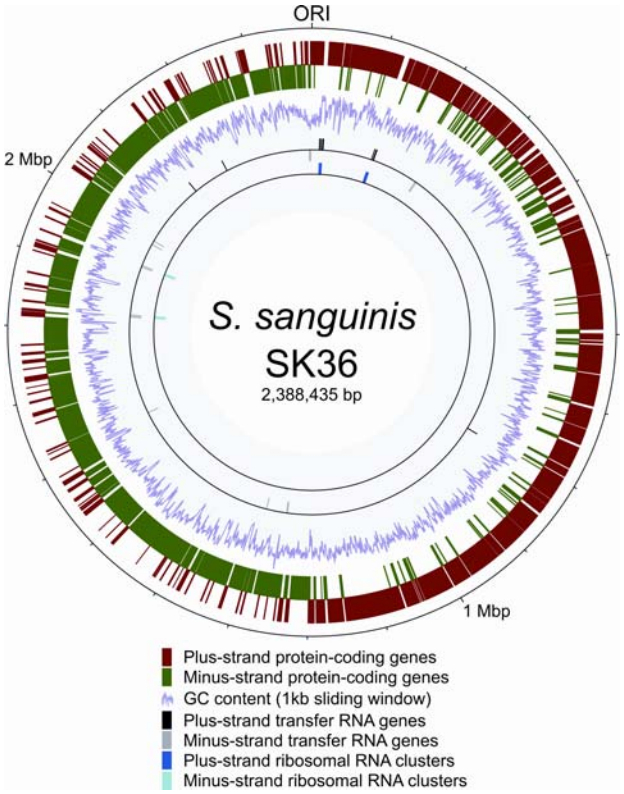
S. mutans causes dental caries (tooth decay) worldwide.

S. pneumoniae causes pneumonia, meningitis, and otitis.

S. pyogenes causes scarlet fever, impetigo, septicemia, etc.

S. sanguis is a leading cause of infective endocarditis.





Xu et al., J. Bacteriol. 2007 Feb

Genome Annotation

1. Repeat identification (RepeatMasker)
2. ORF finding (GenScan, Grail or Glimmer)
3. Homology Searches (BLAST or FASTA)
4. Characterization of proteins(GCG, EMBOSS)