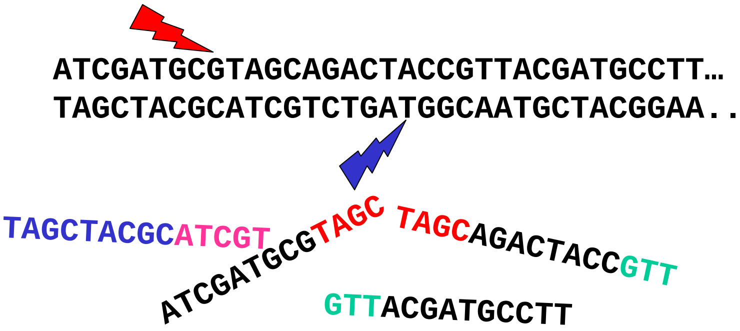
Contig Assembly



ATCGATGCGTAGCAGACTACCGTTACGATGCCTT…
TAGCTACGCATCGTCTGATGGCAATGCTACGGAA. .

TAGCTACGCATCGT

ATCGATGCGTAGC

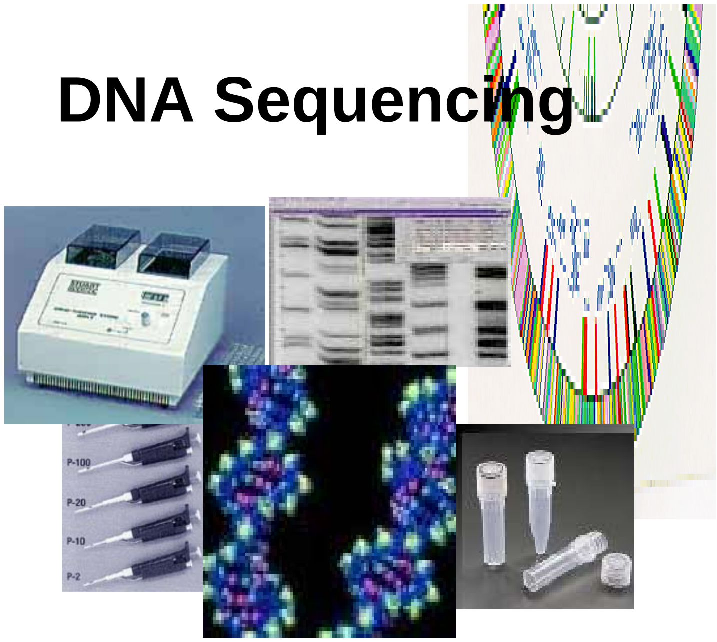
TAGCAGACTACCGTT

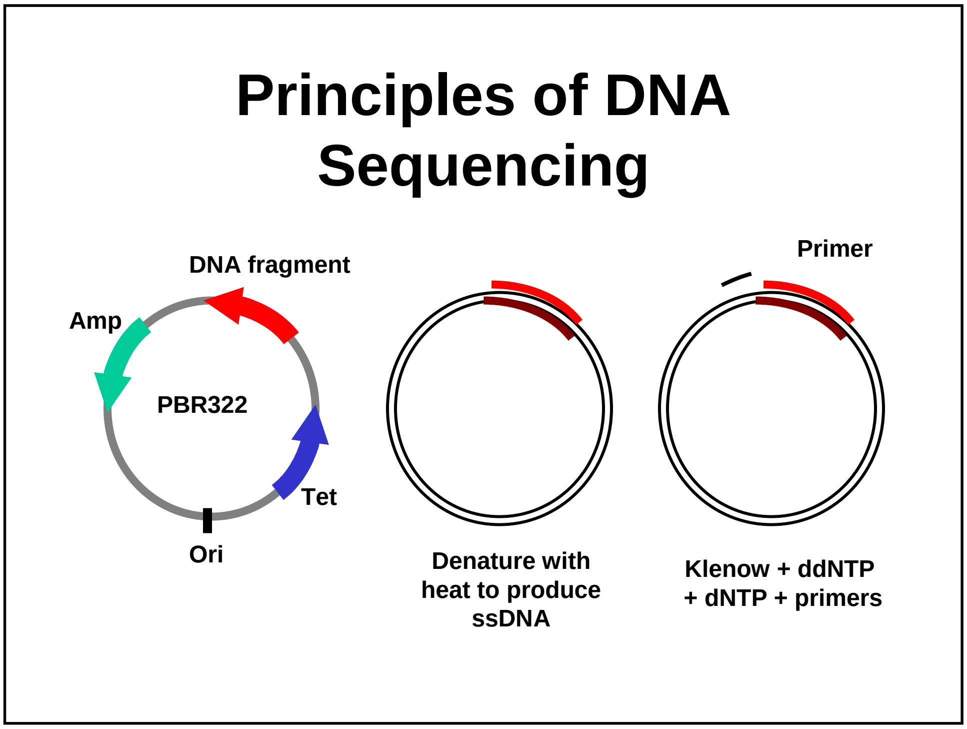
GTTACGATGCCTT

David Wishart, Ath 3-41

david.wishart@ualberta.ca

DNA Sequencing

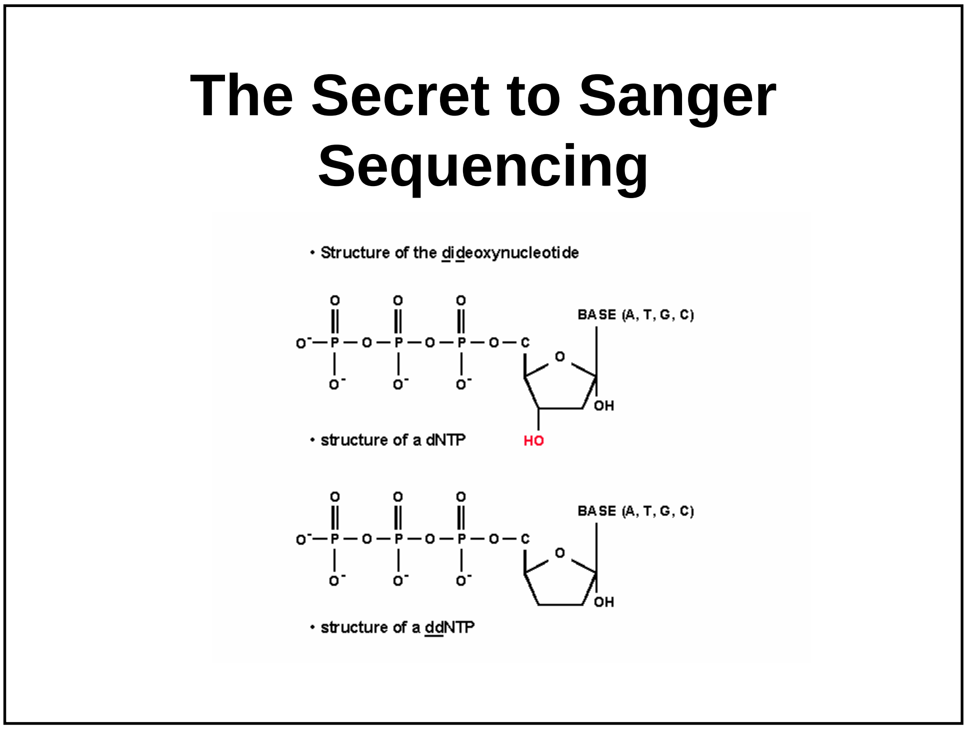




Principles of DNA Sequencing

Denature with heat to produce ssDNA

Klenow + ddNTP + dNTP + primers



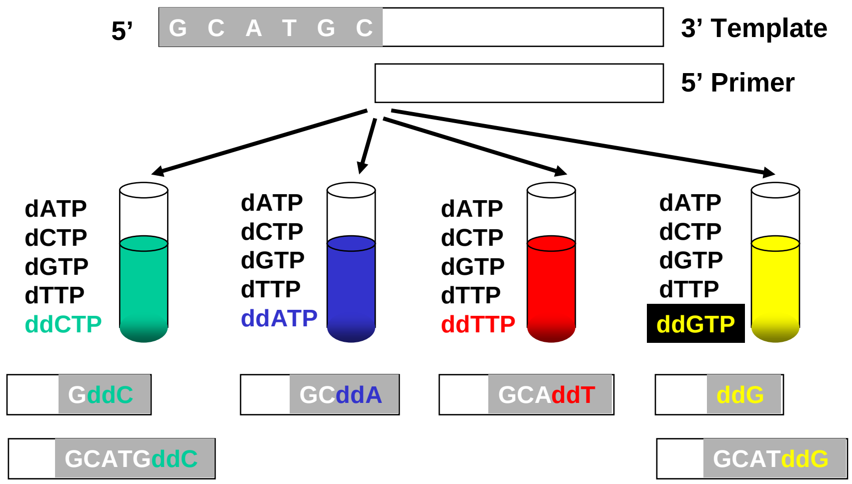
The Secret to Sanger Sequencing

- Structure of the dideoxynucleotide

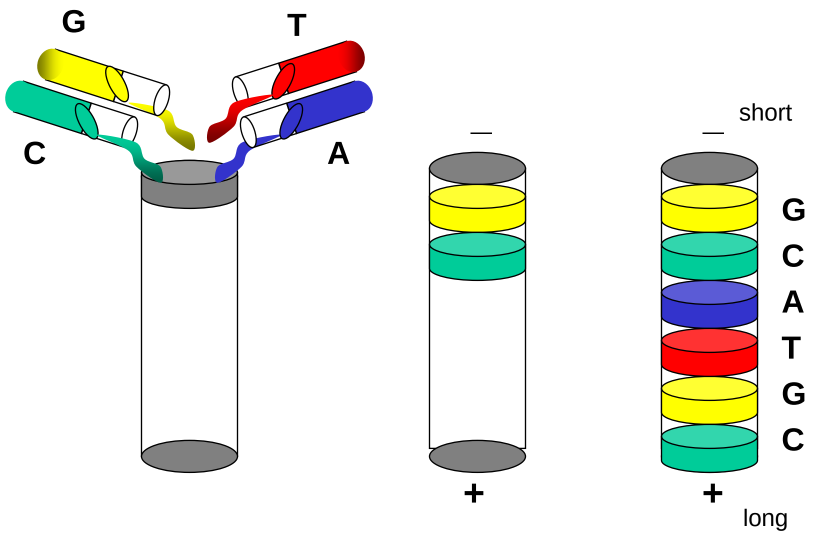
- structure of a dNTP

- structure of a ddNTP

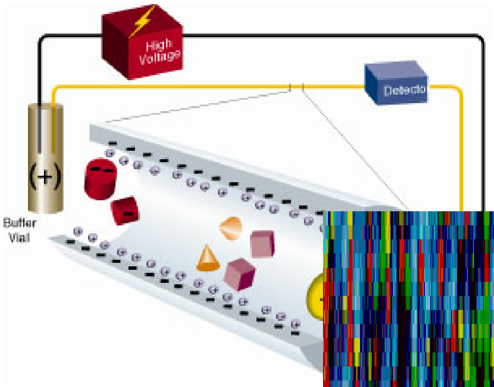
Principles of DNA Sequencing



Principles of DNA Sequencing



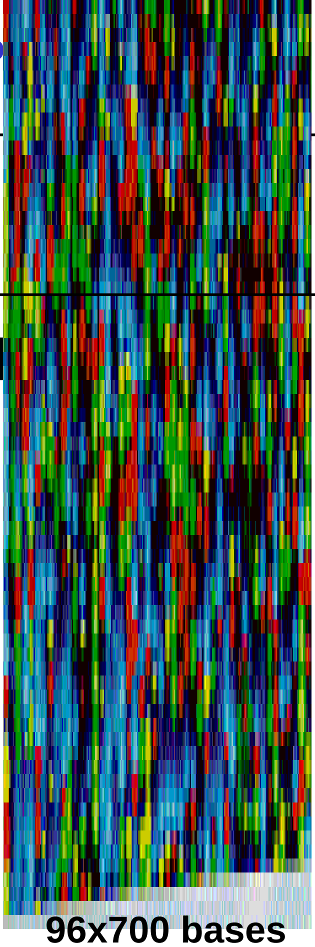
Capillary Electrophoresis



Separation by Electro

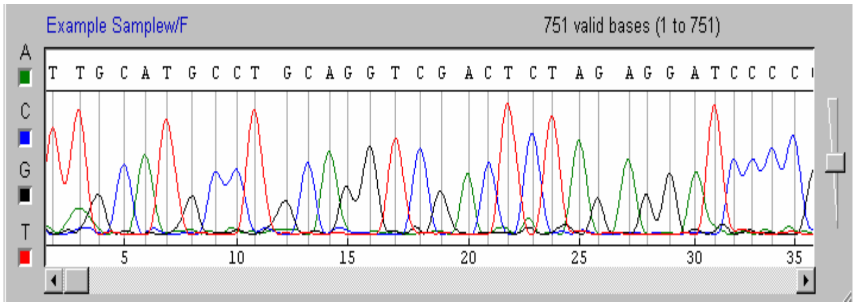
Multiplexed
Fluorescent

ABI 3700



96x700 bases

High Throughput DNA Sequencing



Example Samplew/F

751 valid bases (1 to 751)

T TG C A T G C C T G C A G G T C G A C T C T A G A G G A T C C C C

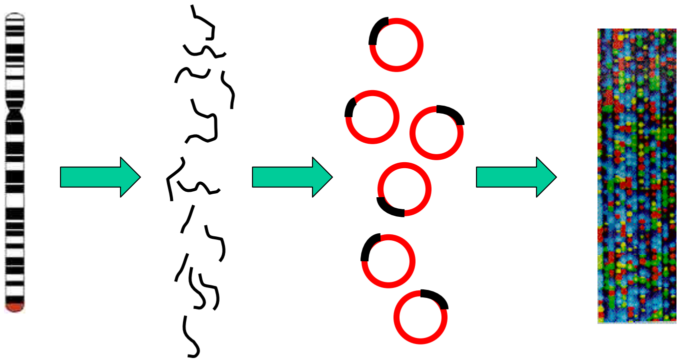
Large Scale Sequencing

- Goal is to determine the nucleic acid sequence of molecules ranging in size from a few hundred bp to $>10^9$ bp
- The methodology requires an extensive computational analysis of raw data to yield the final sequence result

Shotgun Sequencing

- High throughput sequencing method that employs automated sequencing of random DNA fragments
- Automated DNA sequencing yields sequences of 500 to 1000 bp in length
- To determine longer sequences you obtain fragmentary sequences and then join them together by overlapping
- Overlapping is an alignment problem, but different from those we have discussed up to now

Shotgun Sequencing



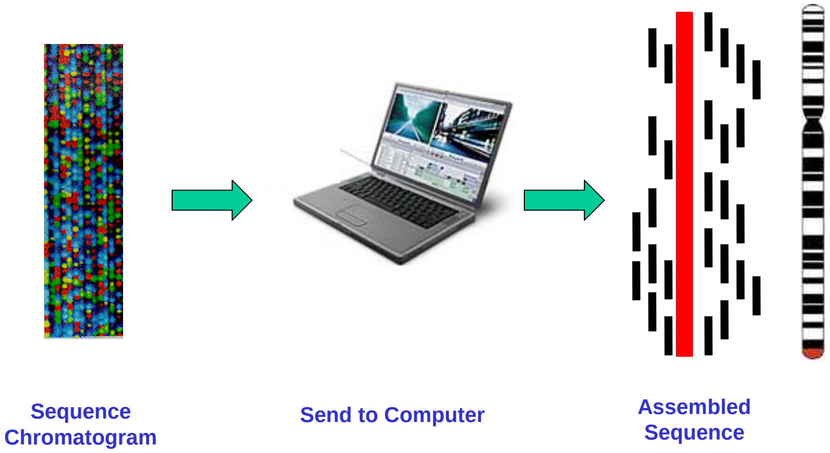
Isolate Chromosome

ShearDNA into Fragments

Clone into Seq. Vectors

Sequence

Shotgun Sequencing



Sequence Chromatogram

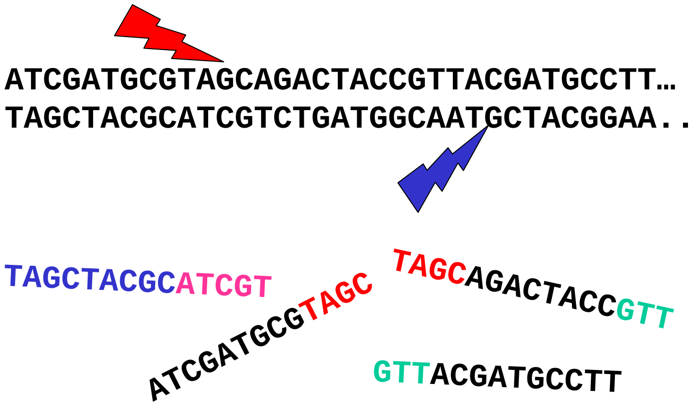
Send to Computer

Assembled Sequence

Analogy

- You have 10 copies of a movie
- The film has been cut into short pieces with about 240 frames per piece (10 seconds of film), at random
- Reconstruct the film

Multi-alignment & Contig Assembly



ATCGATGCGTAGCAGACTACCGTTACGATGCCTT…
TAGCTACGCATCGTCTGATGGCAATGCTACGGAA..

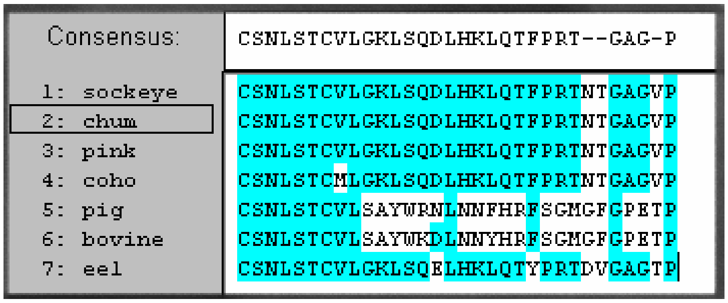
TAGCTACGCATCGT

ATCGATGCGTAGC

TAGCAGACTACCGTT

GTTACGATGCCTT

Multiple Sequence Alignment



```
Consensus:    CSNLSTCVLGKLSQDLHKLQTFPRT--GAG-P

1: sockeye    CSNLSTCVLGKLSQDLHKLQTFPRTNTGAGVP
2: chum       CSNLSTCVLGKLSQDLHKLQTFPRTNTGAGVP
3: pink       CSNLSTCVLGKLSQDLHKLQTFPRTNTGAGVP
4: coho       CSNLSTCMLGKLSQDLHKLQTFPRTNTGAGVP
5: pig        CSNLSTCVLSAYWRNLNNFHRFSGMGFGPETP
6: bovine     CSNLSTCVLSAYWKDLNNYHRFSGMGFGPETP
7: eel        CSNLSTCVLGKLSQELHKLQTYPRTDVGAGTP
```

Multiple alignment of Calcitonins

Multiple Sequence Alignment

- A general method to align and compare more than 2 sequences
- Typically done as a hierarchical clustering/alignment process where you match the two most similar sequences and then use the combined consensus sequence to identify the next closest sequence with which to align

Multiple Alignment Algorithm

- *Take all “n” sequences and perform all possible pairwise (n/2(n-1)) alignments*
- *Identify highest scoring pair, perform an alignment & create a consensus sequence*
- *Select next most similar sequence and align it to the initial consensus, regenerate a second consensus*
- *Repeat step 3 until finished*

Multiple Sequence Alignment

- Developed and refined by many (Doolittle, Barton, Corpet) through the 1980’s
- Used extensively for extracting hidden phylogenetic relationships and identifying sequence families
- Powerful tool for extracting new sequence motifs and signature sequences
- Also applicable to DNA contig assembly

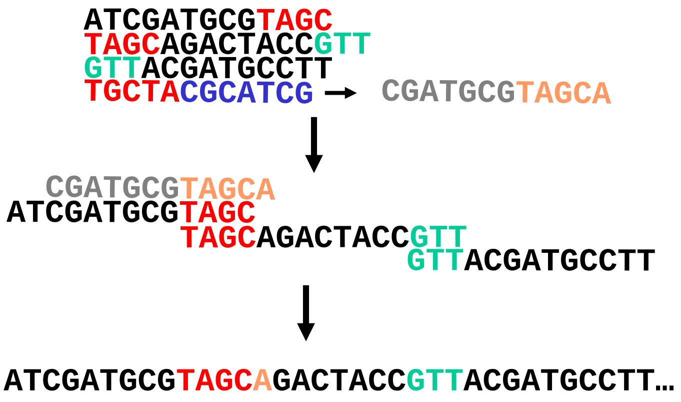
Contig Assembly ≠ Multiple Alignment

1. Only accept a very high sequence identity
2. Accept unlimited number of “end” gaps
3. Very high cost for opening “internal” gaps
4. A short match with high score/residue is preferred over a long match with low score/residue

Contig Assembly Algorithm

- *Read, edit & trim DNA chromatograms*
- *Remove overlaps & ambiguous calls*
- *Read in all sequence files (10-10,000)*
- *Reverse complement all sequences (doubles # of sequences to align)*
- *Remove vector sequences (vector trim)*
- *Remove regions of low complexity*
- *Perform multiple sequence alignment*

Contig Alignment - Process

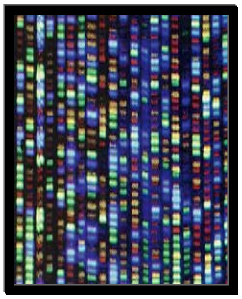


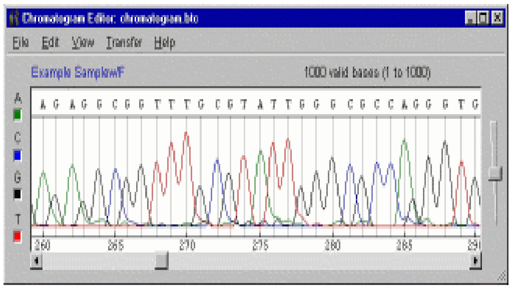
```
    ATCGATGCGTAGC
    TAGCAGACTACCGTT
    GTTACGATGCCTT
    TGCTACGCATCG → CGATGCGTAGCA

  CGATGCGTAGCA
ATCGATGCGTAGC
         TAGCAGACTACCGTT
                     GTTACGATGCCTT

ATCGATGCGTAGCAGACTACCGTTACGATGCCTT…
```

Reading DNA Chromatograms

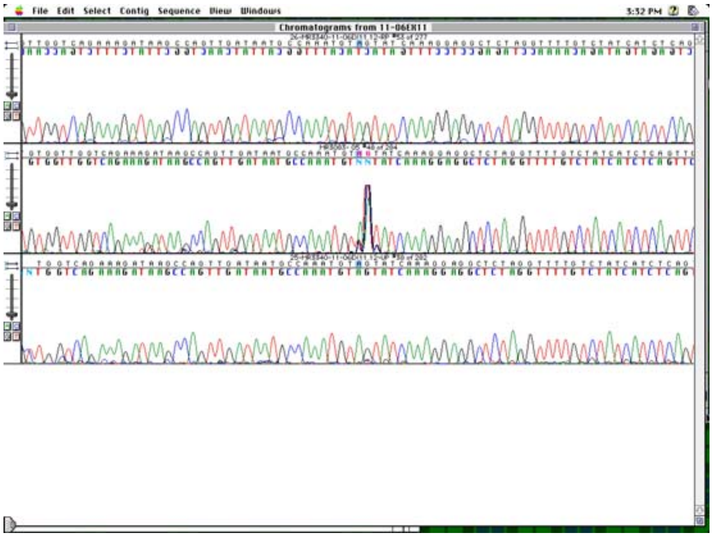



Chromatogram Editor: chromatogram.bio
File Edit View Transfer Help
Example Sample/F
1000 valid bases (1 to 1000)
A C G T
A G A G G C G G T T T G C G T A T T G G G C G C C A G G G T G
260 265 270 275 280 285 29

Gel

ABI Chromatogram

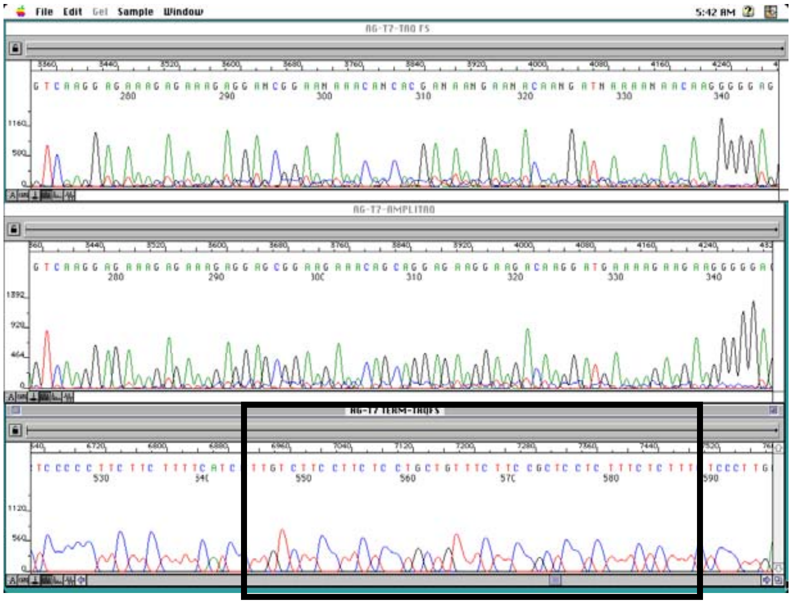
Typical Raw Data


File Edit Select Contig Sequence View Windows
3:32 PM
Chromatograms from 11-06EX11

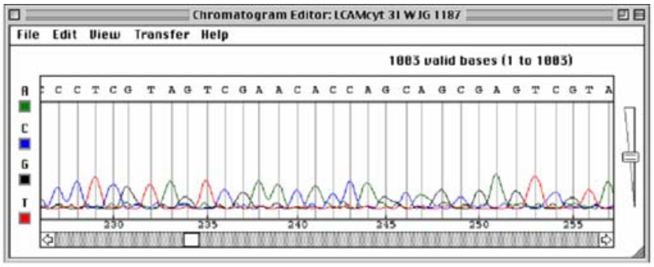
Chromatograms (Problems)

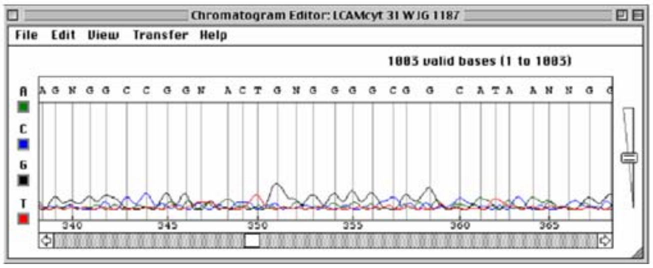
- Degradation of gel resolution (Pile-up or Band Broadening)
- Diminishment or excess of fluorescence intensity (too little or too much DNA tmplte)
- Differential overlap (large peak followed by a small one , ie. “G” dropouts (small G following a big A peak)
- Homopolymeric stretches of A’s and T’s
- Inappropriate spacing (contaminant DNA or poor/noisy primers causing random priming)
- High GC content or GC rich regions
- Secondary structure or inverted repeats of the DNA

Band Broadening

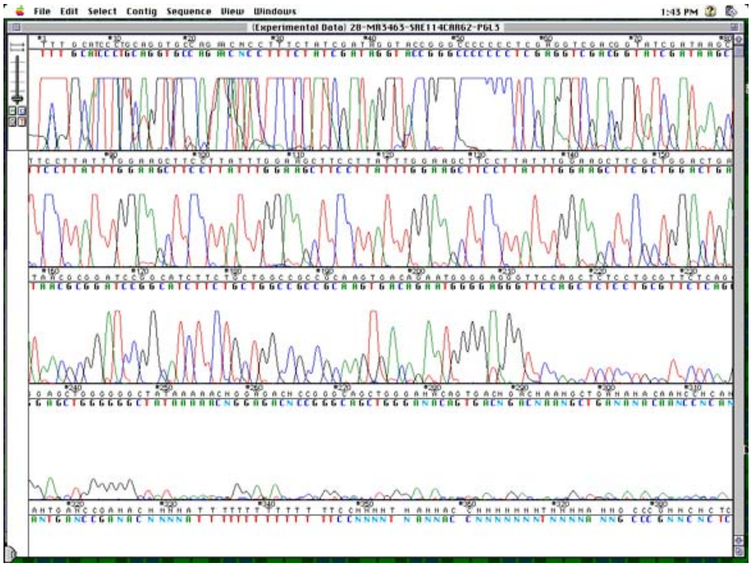

File Edit Gel Sample Window
5:42 AM
AG-T7-TAQ FS
G T C A A G G A G A A A G A G A A A G A G G A N C G G A A N A A A C A N C A C G A N A A N G A A N A C A A N G A T N A A A A N A A C A A G G G G G A G
280
290
300
310
320
330
340
1160
580
0
AG-T7-AMPLITAQ
G T C A A G G A G A A A G A G A A A G A G G A G C G G A A G A A A C A G C A G G A G A A G G A A G A C A A G G A T G A A A A G A A G A A G G G G G A
1392
928
464
AG-T7 TERM-TAQFS
T C C C C C T T C T T C T T T T C A T C T T G T C T T C C T T C T C C T G C T G T T T C T T C C G C T C C T C T T T C T C T T T T C C C T T G
530
540
550
560
570
580
590
1120
560

Diminishing Intensity


Chromatogram Editor: LCAMcyt 31 WJG 1187
File Edit View Transfer Help
1083 valid bases (1 to 1083)
A
C
G
T
C C T C G T A G T C G A A C A C C A G C A G C G A G T C G T A
230
235
240
245
250
255


Chromatogram Editor: LCAMcyt 31 WJG 1187
File Edit View Transfer Help
1083 valid bases (1 to 1083)
A
C
G
T
A G N G G C C G G N A C T G N G G G G C G G C A T A A N N G G
340
345
350
355
360
365

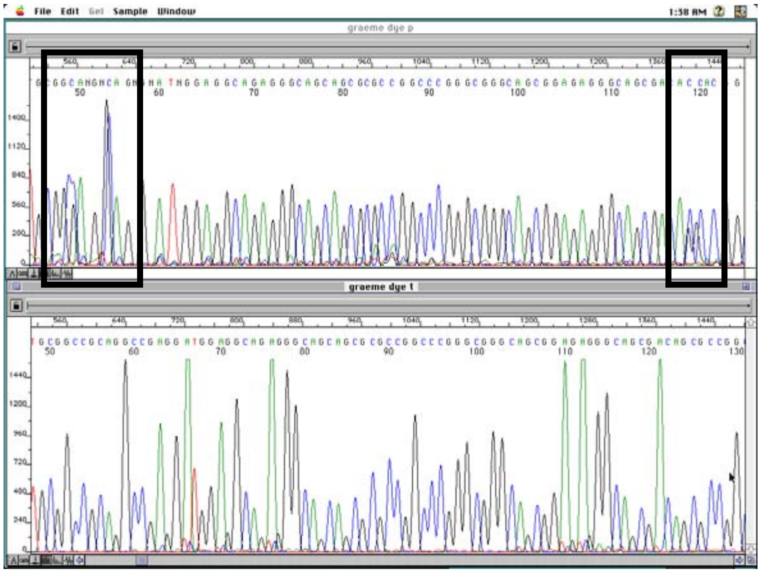
Too Much DNA Template

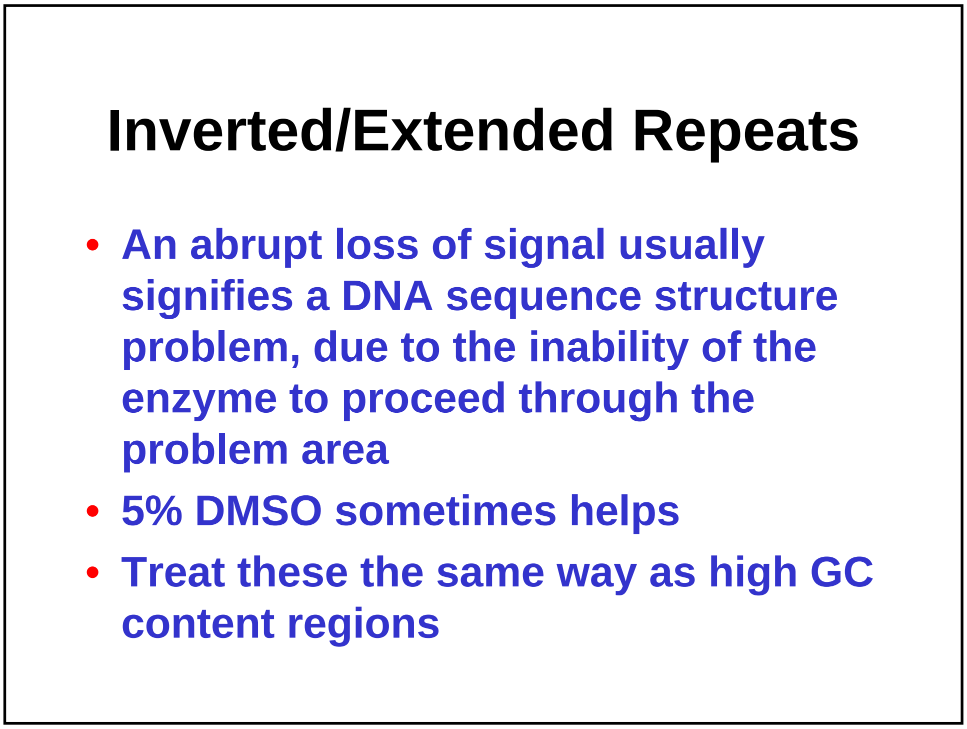

File Edit Select Contig Sequence View Windows
1:43 PM
(Experimental Data) 28-MR3463-SRE114CARG2-PGL3

High G-C Content

- >60% GC content may be difficult to sequence (leads to pile-up)
- Dye terminator performs better than dye primer
- Easiest modification is to add 5% DMSO final concentration to the reaction mix
- Sequence the opposite strand to help resolve ambiguities

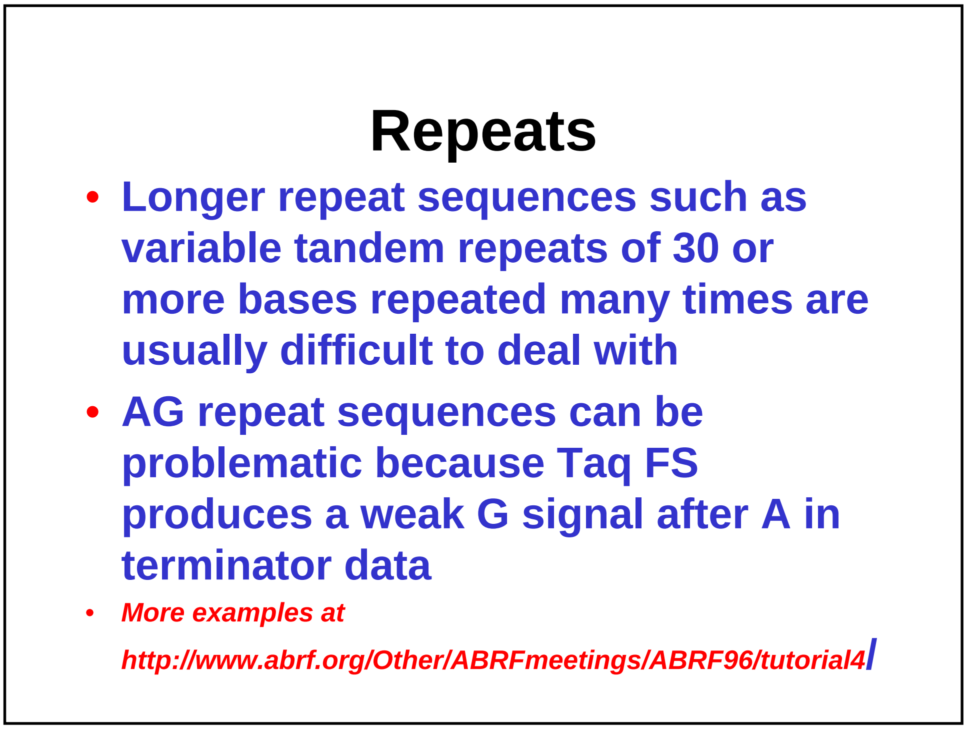
GC Pile Up


File Edit Gel Sample Window
1:38 AM
graeme dye p
GCGGCANGNCA GNGNA TNGGAGGCAGAGGGCAGCAGCGCGCCGGCCCGGGCGGGCAGCGGAGAGGGCAGCGACACCAC G
50 60 70 80 90 100 110 120
1400 1120 840 560 280 0
graeme dye t
TGCGGCCGCAGGCCGAGGATGGAGGCAGAGGGCAGCAGCGCGCCGGCCCGGGCGGGCAGCGGAGAGGGCAGCGACAGCGCCGG
50 60 70 80 90 100 110 120 130
1440 1200 960 720 480 240 0



Inverted/Extended Repeats

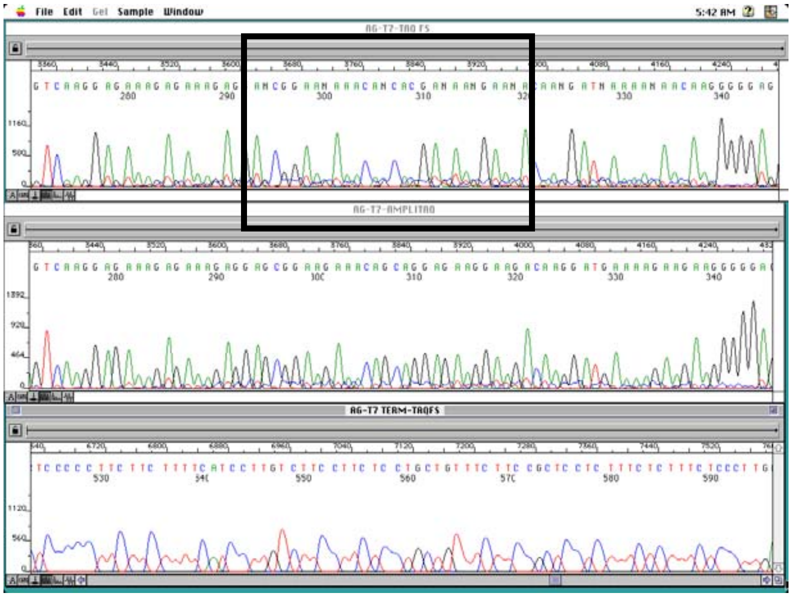
- An abrupt loss of signal usually signifies a DNA sequence structure problem, due to the inability of the enzyme to proceed through the problem area
- 5% DMSO sometimes helps
- Treat these the same way as high GC content regions



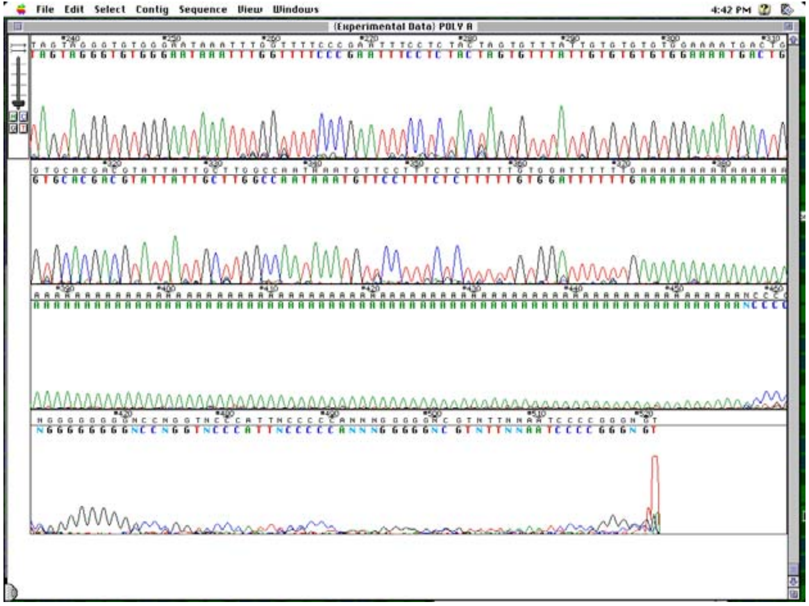
Repeats

- Longer repeat sequences such as variable tandem repeats of 30 or more bases repeated many times are usually difficult to deal with
- AG repeat sequences can be problematic because Taq FS produces a weak G signal after A in terminator data
- *More examples at http://www.abrf.org/Other/ABRFmeetings/ABRF96/tutorial4/*

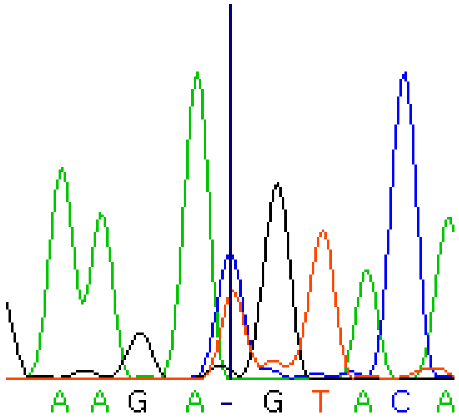
Weak G after A



Homopolymer Stretches


File Edit Select Contig Sequence View Windows
4:42 PM
(Experimental Data) POLY A

Base Calling



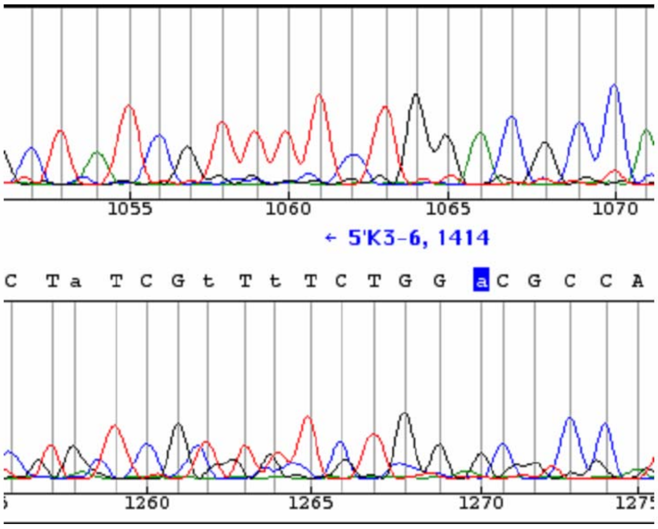
Imperfect Raw Data

- The data from sequencers varies in quality along the length of a single scan
- The base calls can be ambiguous, but there is still some information
- Need a quantitative analysis, not qualitative, to maximize information

Quality Factors

- Simplest approach is human inspection, but not automatable
- Although computationally more difficult, quantitative factors provide a significant improvement in the assembly process
- Particularly important in high-throughput sequencing projects

Human Inspection



Automated Base Calling with Phred

- The Phred software reads DNA sequencing trace files, calls bases, and assigns a quality value to each called base
- The quality value is a log-transformed error probability, specifically

$$Q = -10 \log10(Pe)$$

- where Q and Pe are respectively the quality value and error probability of a particular base call

Phred

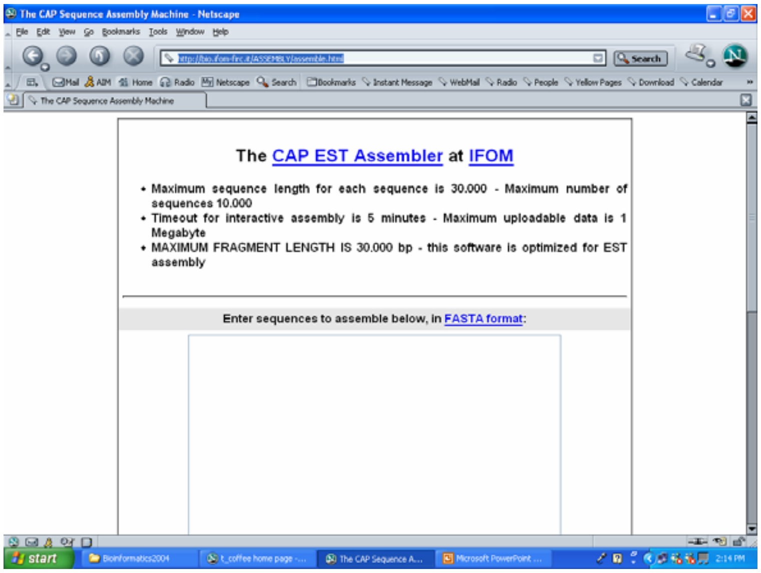
- The Phred quality values have been thoroughly tested for both accuracy and power to discriminate between correct and incorrect base-calls
- Phred can use the quality values to perform sequence trimming

Ewing B, Green P: Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186-194 (1998)

Sequence Assembly Programs

- Phred - base calling program that does detailed statistical analysis (UNIX)
  http://www.phrap.org/
- Phrap - sequence assembly program (UNIX)
  http://www.phrap.org/
- TIGR Assembler - microbial genomes (UNIX)
  http://www.tigr.org/softlab/assembler/
- The Staden Package (UNIX)
  http://www.mrc-lmb.cam.ac.uk/pubseq/
- GeneTool/ChromaTool/Sequencher (PC/Mac)

http://bio.ifom-firc.it/ASSEMBLY/assemble.html


The CAP Sequence Assembly Machine - Netscape
File Edit View Go Bookmarks Tools Window Help
http://bio.ifom-firc.it/ASSEMBLY/assemble.html
Search
Mail AIM Home Radio Netscape Search Bookmarks Instant Message WebMail Radio People Yellow Pages Download Calendar
The CAP Sequence Assembly Machine
start Bioinformatics2004 t_coffee home page -... The CAP Sequence A... Microsoft PowerPoint ...
2:14 PM

The CAP EST Assembler at IFOM

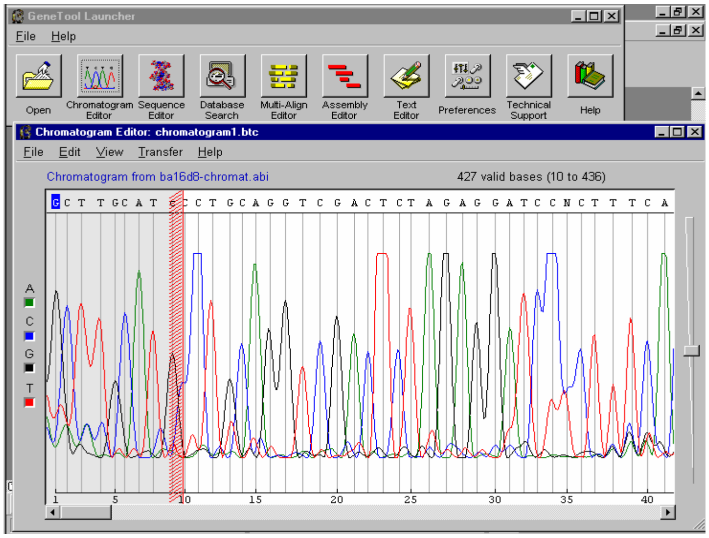
- Maximum sequence length for each sequence is 30.000 - Maximum number of sequences 10.000
- Timeout for interactive assembly is 5 minutes - Maximum uploadable data is 1 Megabyte
- MAXIMUM FRAGMENT LENGTH IS 30.000 bp - this software is optimized for EST assembly

Enter sequences to assemble below, in FASTA format:

Contig Assembly Algorithm

- *Read, edit & trim DNA chromatograms*
- *Remove overlaps & ambiguous calls*
- *Read in all sequence files (10-10,000)*
- *Reverse complement all sequences (doubles # of sequences to align)*
- *Remove vector sequences (vector trim)*
- *Remove regions of low complexity*
- *Perform multiple sequence alignment*

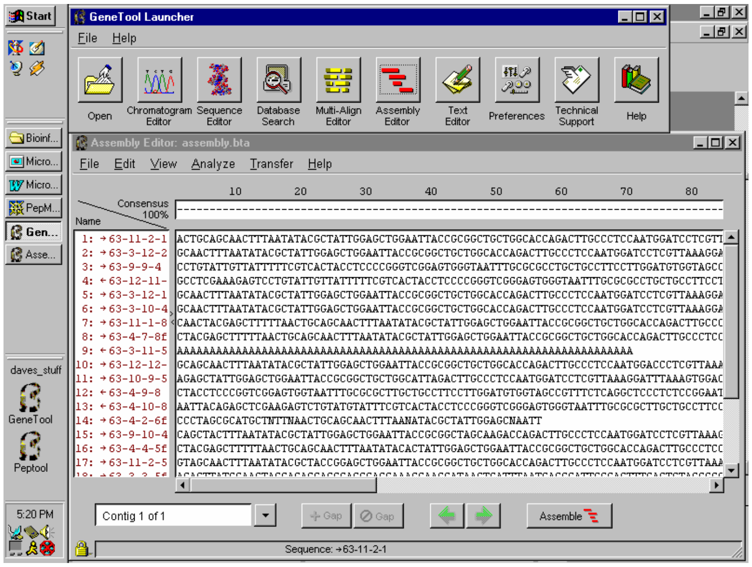
Chromatogram Editing



Chromatogram from ba16d8-chromat.abi

427 valid bases (10 to 436)

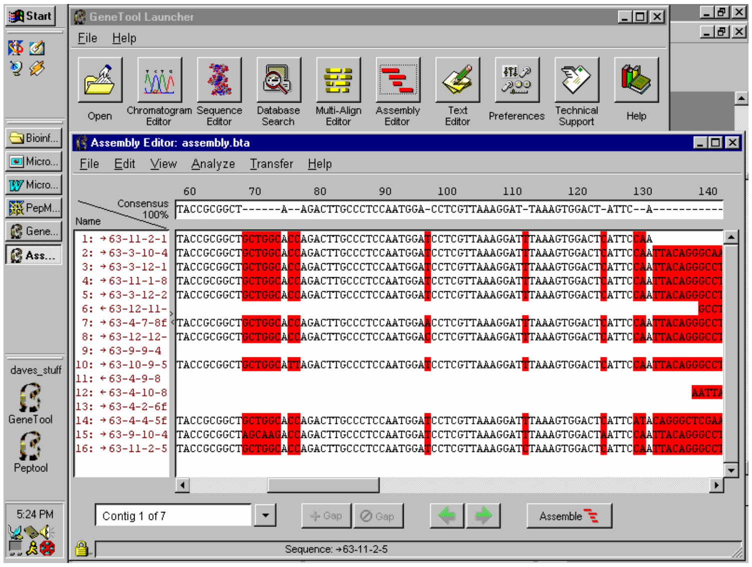
Sequence Loading



```
                 10        20        30        40        50        60        70        80
Consensus 100%   ------------------------------------------------------------------------------------
Name
 1: →63-11-2-1   ACTGCAGCAACTTTAATATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTT
 2: →63-3-12-2   GCAACTTTAATATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGA
 3: →63-9-9-4    CCTGTATTGTTATTTTTCGTCACTACCTCCCCGGGTCGGAGTGGGTAATTTGCGCGCCTGCTGCCTTCCTTGGATGTGGTAGCC
 4: ←63-12-11-   GCCTCGAAAGAGTCCTGTATTGTTATTTTTCGTCACTACCTCCCCGGGTCGGAGTGGGTAATTTGCGCGCCTGCTGCCTTCCT
 5: →63-3-12-1   GCAACTTTAATATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGA
 6: →63-3-10-4   GCAACTTTAATATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGA
 7: →63-11-1-8   CAACTACGAGCTTTTTAACTGCAGCAACTTTAATATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCC
 8: →63-4-7-8f   CTACGAGCTTTTTAACTGCAGCAACTTTAATATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCCTCC
 9: ←63-3-11-5   AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
10: →63-12-12-   GCAGCAACTTTAATATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGACCCTCGTTAAA
11: →63-10-9-5   AGAGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCATTAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATTTAAAGTGGAC
12: ←63-4-9-8    CTACCTCCCGGTCGGAGTGGTAATTTGCGCGCTTGCTGCCTTCCTTGGATGTGGTAGCCGTTTCTCAGGCTCCCTCTCCGGAAT
13: ←63-4-10-8   AATTACAGAGCTCGAAGAGTCTGTATGTATTTCGTCACTACCTCCCGGGTCGGAGTGGGTAATTTGCGCGCTTGCTGCCTTCC
14: →63-4-2-6f   CCCTAGCGCATGCTNTTNAACTGCAGCAACTTTAANATACGCTATTGGAGCNAATT
15: →63-9-10-4   CAGCTACTTTAATATACGCTATTGGAGCTGGAATTACCGCGGCTAGCAAGACCAGACTTGCCCTCCAATGGATCCTCGTTAAAG
16: →63-4-4-5f   CTACGAGCTTTTTAACTGCAGCAACTTTAATATACACTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCCTCC
17: →63-11-2-5   GTAGCAACTTTAATATACGCTACCGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAA
```

Sequence: →63-11-2-1

Sequence Alignment



```
                 60        70        80        90        100       110       120       130       140
Consensus 100%   TACCGCGGCT------A--AGACTTGCCCTCCAATGGA-CCTCGTTAAAGGAT-TAAAGTGGACT-ATTC--A----------
Name
 1: →63-11-2-1   TACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATTTAAAGTGGACTCATTCCAA
 2: →63-3-10-4   TACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATTTAAAGTGGACTCATTCCAATTACAGGGCAA
 3: →63-3-12-1   TACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATTTAAAGTGGACTCATTCCAATTACAGGGCCT
 4: →63-11-1-8   TACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATTTAAAGTGGACTCATTCCAATTACAGGGCCT
 5: →63-3-12-2   TACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATTTAAAGTGGACTCATTCCAATTACAGGGCCT
 6: ←63-12-11-                                                                                  GCCT
 7: →63-4-7-8f   TACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGAACCTCGTTAAAGGATTTAAAGTGGACTCATTCCAATTACAGGGCCT
 8: →63-12-12-   TACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGACCCTCGTTAAAGGATTTAAAGTGGACTCATTCCAATTACAGGGCCT
 9: →63-9-9-4
10: →63-10-9-5   TACCGCGGCTGCTGGCATTAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATTTAAAGTGGACTCATTCCAATTACAGGGCCT
11: ←63-4-9-8
12: ←63-4-10-8                                                                                  AATTA
13: →63-4-2-6f
14: →63-4-4-5f   TACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATTTAAAGTGGACTCATTCATACAGGGCTCGAA
15: →63-9-10-4   TACCGCGGCTAGCAAGACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATTTAAAGTGGACTAATTCCAATTACAGGGCCT
16: →63-11-2-5   TACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAAGGATCTAAAGTGGACTCATTCCAATTACAGGGCCT
```

Sequence: →63-11-2-5

Assembly Parameters

- User-selected parameters
  1. minimum length of overlap
  2. percent identity within overlap
- Non-adjustable parameters
  1. sequence “quality” factors

Phrap

- Phrap is a program for assembling shotgun DNA sequence data
- Uses a combination of user-supplied and internally computed data quality information to improve assembly accuracy in the presence of repeats
- Constructs the contig sequence as a mosaic of the highest quality read segments rather than a consensus
- Handles large datasets

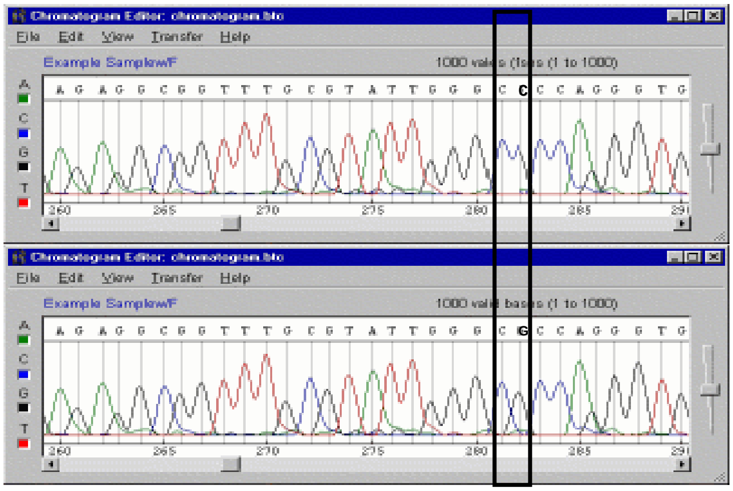
Problems for Assembly

- Repeat regions
  - Capture sequences from non-contiguous regions
- Polymorphisms
  - Cause failure to join correct regions
- Large data volume
  - Requires large numbers of pair-wise comparisons

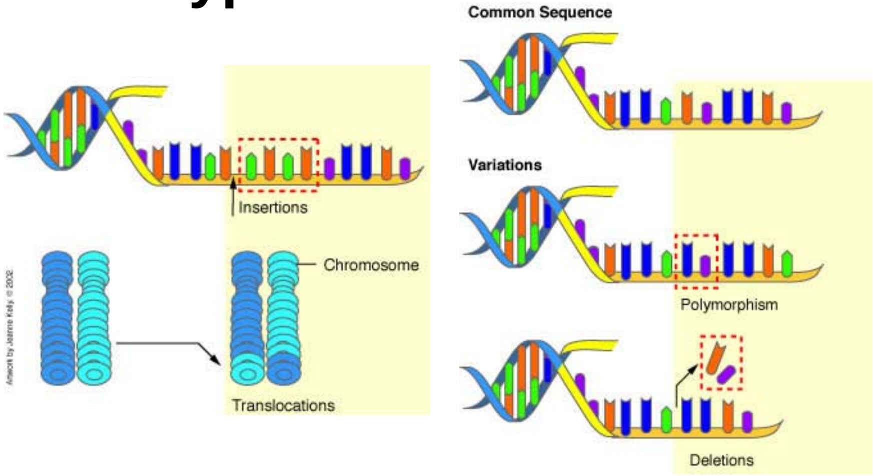
Mutation Detection

Normal

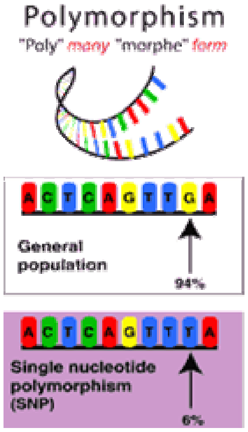
Diseased

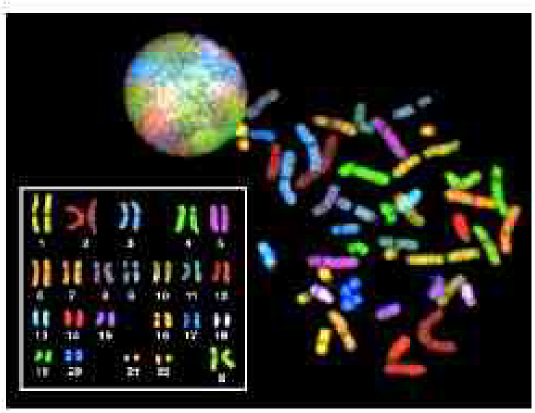

Chromatogram Editor: chromatogram.bio
File Edit View Transfer Help
Example SamplewF
1000 values (bases (1 to 1000)
A C G T
A G A G G C G G T T T G C G T A T T G G G C C C C A G G G T G
260 265 270 275 280 285 29
Chromatogram Editor: chromatogram.bio
File Edit View Transfer Help
Example SamplewF
1000 valid bases (1 to 1000)
A C G T
A G A G G C G G T T T G C G T A T T G G G C G C C A G G G T G
260 265 270 275 280 285 29

Types of Mutations



SNPs & Polymorphisms

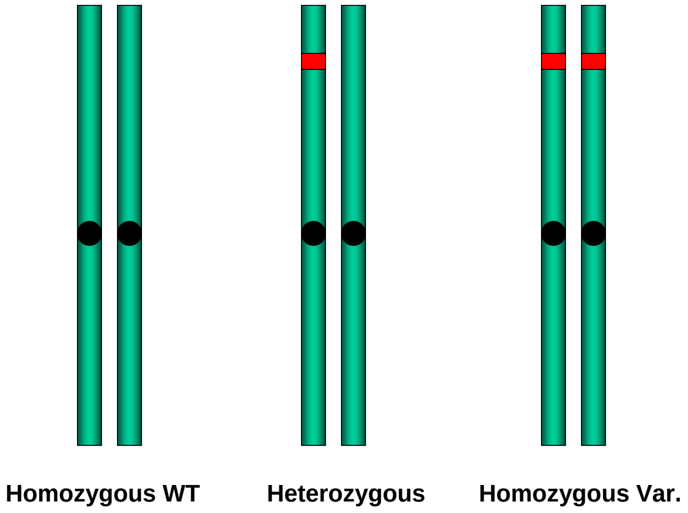




SNPs (Single Nucleotide Polymorphisms)

- Single nucleotide polymorphisms or SNPs are DNA sequence variations that occur when a single nucleotide (A,T,C or G) in the genome sequence is altered
- For a variation to be considered a SNP, it must occur in at least 1% of the population
- If the frequency is less than 1% (although this is somewhat arbitrary) then this variation is called a mutation
- SNPs are classified in three different ways…

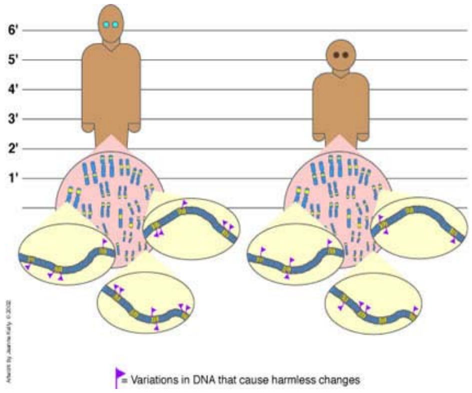
Zygosity and SNPs

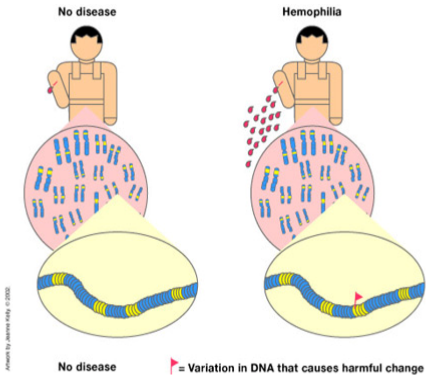


SNPs

- SNPs account for about 90% of all human genetic variation and are believed to occur every 100 to 300 bases along the 3-billion-base human genome
- Approximately 5 million of the ~10 million human SNPs have been catalogued
- SNPs may occur in exons, introns (non coding regions between exons) and intergenic regions (regions between genes)
- SNPs may lead to coding or amino acid sequence changes (non-synonymous) or they may leave the sequence unchanged (synonymous)

Synonymous vs. Non-Synonymous SNPs


6'
5'
4'
3'
2'
1'
Artwork by Jeanne Kelly © 2002
= Variations in DNA that cause harmless changes


No disease
Hemophilia
Artwork by Jeanne Kelly © 2002
No disease
= Variation in DNA that causes harmful change

Hardy Weinberg Equilibrium

Hardy Weinberg Equilibrium

- True SNPs should follow Hardy Weinberg Equilibrium in that
- The choice of a mate is not influenced by his/her genotype at the locus/gene (random mating or panmixia)
- The locus/gene/SNP does not affect the chance of mating at all, either by altering fertility or decreasing survival to reproductive age

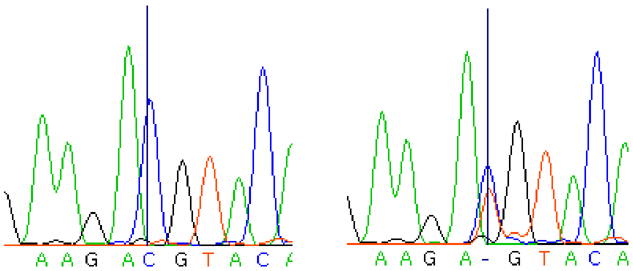
Deviations from HWE

- Marital assortment: "like marrying like"
- Inbreeding
- Population stratification: multiple subgroups are present within the population, each of which mates only within its own group (homogamy)
- Decreased viability of a particular genotype (hemophilia)

Measuring SNPs

- Classical sequencing (homozygotes)
- Chromatogram analysis (heterozygotes)
- Denaturing HPLC
- Rolling Circle Amplification
- Antibody-based detection
- Enzyme- or cleavage-based detection
- Mass spectrometry
- SNP chips or microarrays

Polymorphism in Connexin26 (CX26) – Common Cause of Deafness -- ID by Sequencing



Homozyogous for C

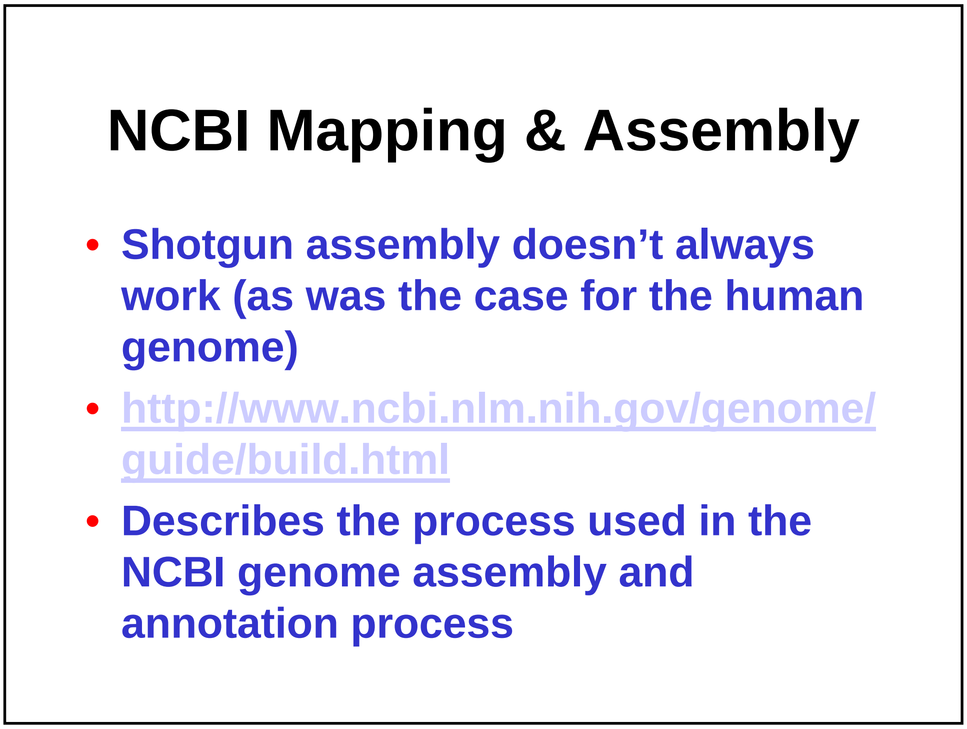
Heterozygous for T/C

The Finished Product

```
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TACAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATTAC
AGATTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
```

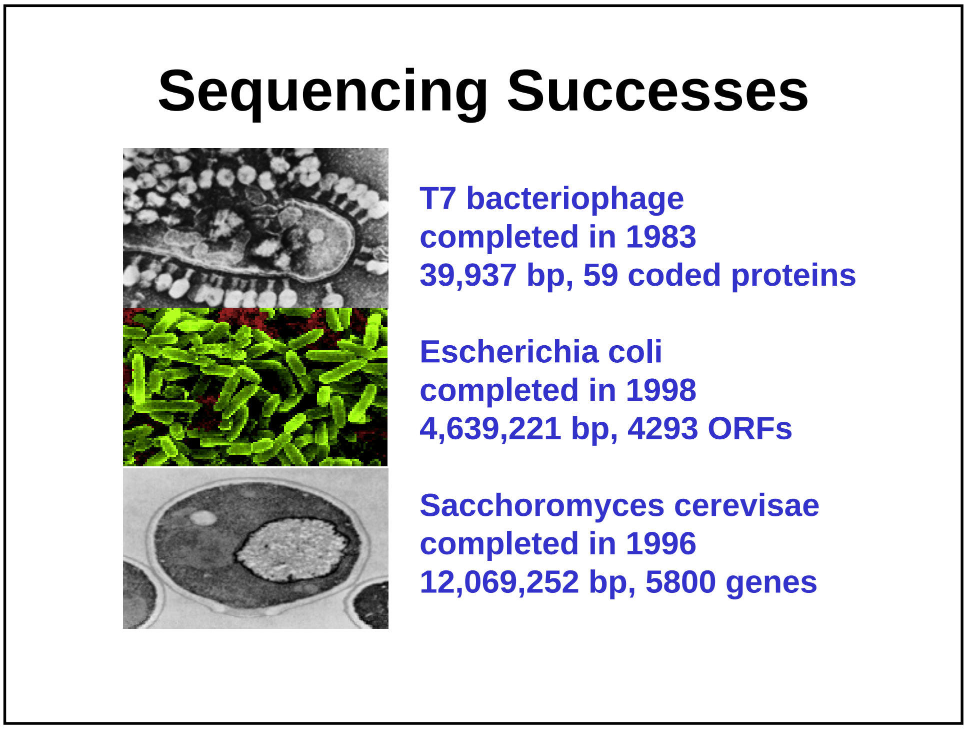
Shotgun Sequencing Summary

- Very efficient process for small-scale (~10 kb) sequencing (preferred method)
- First applied to whole genome sequencing in 1995 (*H. influenzae*)
- Now standard for all prokaryotic genome sequencing projects
- Successfully applied to *D. melanogaster*
- Moderately successful for *H. sapiens*



NCBI Mapping & Assembly

- Shotgun assembly doesn't always work (as was the case for the human genome)
- http://www.ncbi.nlm.nih.gov/genome/guide/build.html
- Describes the process used in the NCBI genome assembly and annotation process



Sequencing Successes

T7 bacteriophage
completed in 1983
39,937 bp, 59 coded proteins

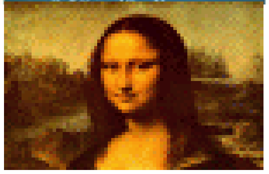
Escherichia coli
completed in 1998
4,639,221 bp, 4293 ORFs

Sacchoromyces cerevisae
completed in 1996
12,069,252 bp, 5800 genes

Sequencing Successes

Caenorhabditis elegans
completed in 1998
95,078,296 bp, 19,099 genes

Drosophila melanogaster
completed in 2000
116,117,226 bp, 13,601 genes

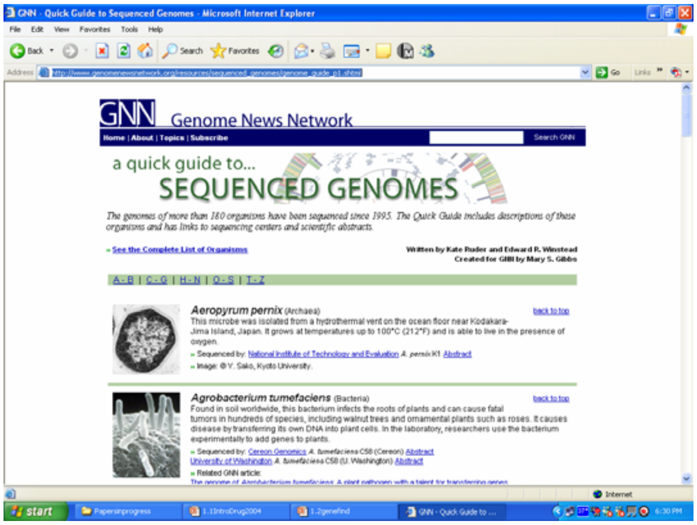


Homo sapiens
Final draft completed in 2003
3,201,762,515 bp, 31,780 genes

Genomes to Date

- 8 vertebrates (human, mouse, rat, fugu, zebrafish)
- 2 plants (arabadopsis, rice)
- 2 insects (fruit fly, mosquito)
- 2 nematodes (C. elegans, C. briggsae)
- 1 sea squirt
- 4 parasites (plasmodium, guillardia)
- 4 fungi (S. cerevisae, S. pombe)
- 200 bacteria and archaebacteria
- 1900+ viruses

Sequenced Genomes



GNN - Quick Guide to Sequenced Genomes - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_p1.shtml
Go
Links
GNN Genome News Network
Home | About | Topics | Subscribe
Search GNN

a quick guide to...

SEQUENCED GENOMES

The genomes of more than 180 organisms have been sequenced since 1995. The Quick Guide includes descriptions of these organisms and has links to sequencing centers and scientific abstracts.

» See the Complete List of Organisms

Written by Kate Ruder and Edward R. Winstead
Created for GNN by Mary S. Gibbs

A - B | C - G | H - N | O - S | T - Z

Aeropyrum pernix (Archaea) back to top

This microbe was isolated from a hydrothermal vent on the ocean floor near Kodakara-Jima Island, Japan. It grows at temperatures up to 100°C (212°F) and is able to live in the presence of oxygen.

» Sequenced by: National Institute of Technology and Evaluation *A. pernix* K1 Abstract

» Image: © Y. Sako, Kyoto University.

Agrobacterium tumefaciens (Bacteria) back to top

Found in soil worldwide, this bacterium infects the roots of plants and can cause fatal tumors in hundreds of species, including walnut trees and ornamental plants such as roses. It causes disease by transferring its own DNA into plant cells. In the laboratory, researchers use the bacterium experimentally to add genes to plants.

» Sequenced by: Cereon Genomics *A. tumefaciens* C58 (Cereon) Abstract
University of Washington *A. tumefaciens* C58 (U. Washington) Abstract

» Related GNN article:
The genome of *Agrobacterium tumefaciens*. A plant pathogen with a talent for transferring genes

http://www.genomenewsnetwork.org/