DNA as
Biological Information

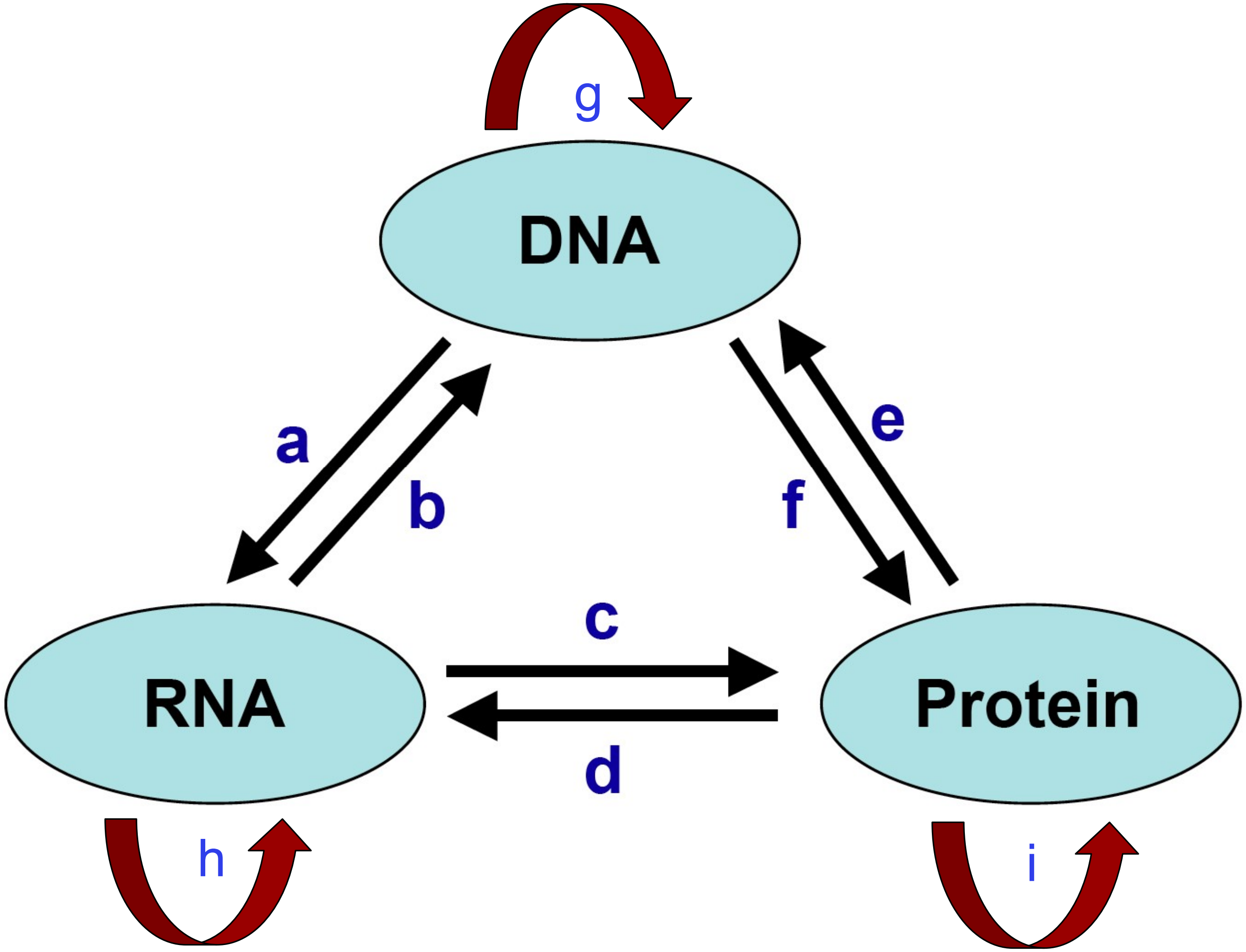
Rasmus Wernersson
Henrik Nielsen
Carolina Barra

Learning objectives

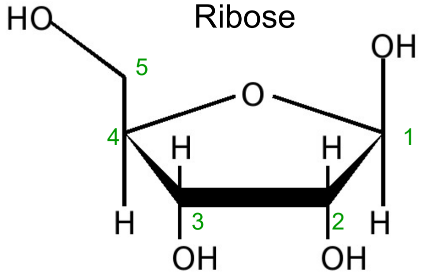
- Interpret DNA as Biological information
- Describe DNA sequencing techniques and DNA data
- Identify file formats used to store DNA data
- Recognize how information is stored in GenBank database

DNA as *information*

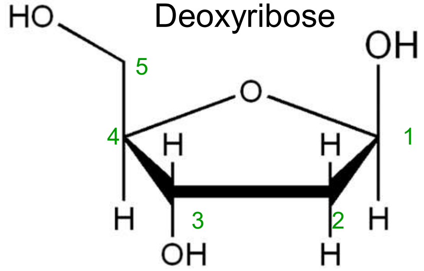
Information flow in biological systems



DNA sequences = summary of information



Ribose

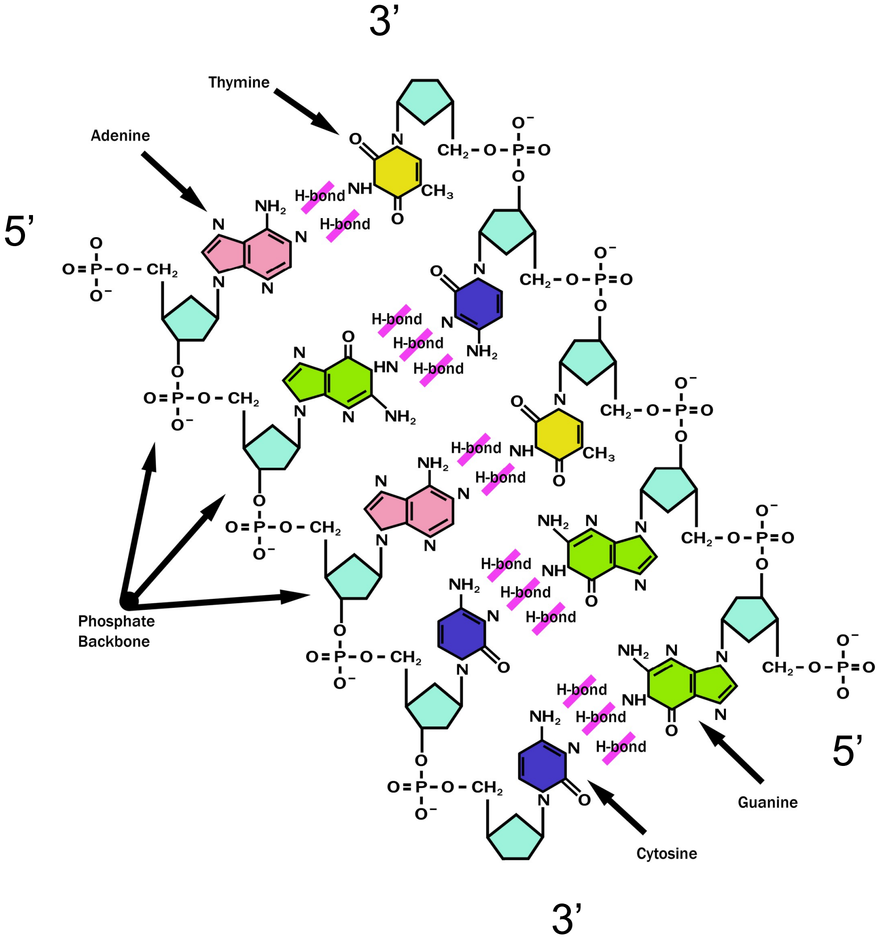


Deoxyribose

5’ AGCC 3’

3’ TCGG 5’

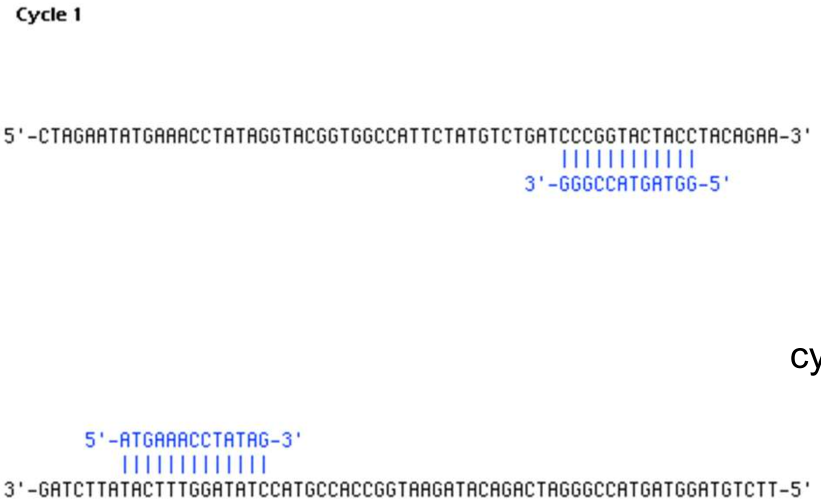
5’ ATGGCCAGGTAA 3’

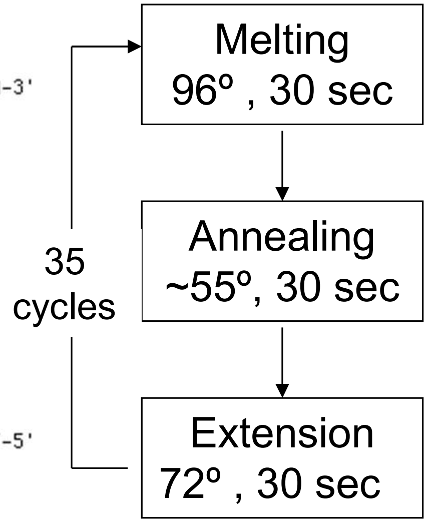


DNA backbone: `http://en.wikipedia.org/wiki/DNA`
(Deoxy)ribose: `http://en.wikipedia.org/`

From molecules to computer files

Reminder: PCR


Cycle 1
5'-CTAGAATATGAAACCTATAGGTACGGTGGCCATTCTATGTCTGATCCCGGTACTACCTACAGAA-3'
3'-GGGCCATGATGG-5'
5'-ATGAAACCTATAG-3'
3'-GATCTTATACTTTGGATATCCATGCCACCGGTAAGATACAGACTAGGGCCATGATGGATGTCTT-5'



Melting
96º , 30 sec

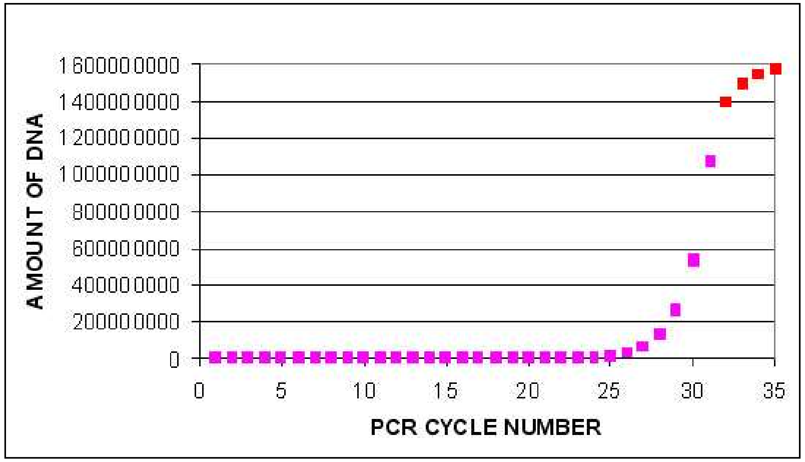
Annealing
~55º, 30 sec

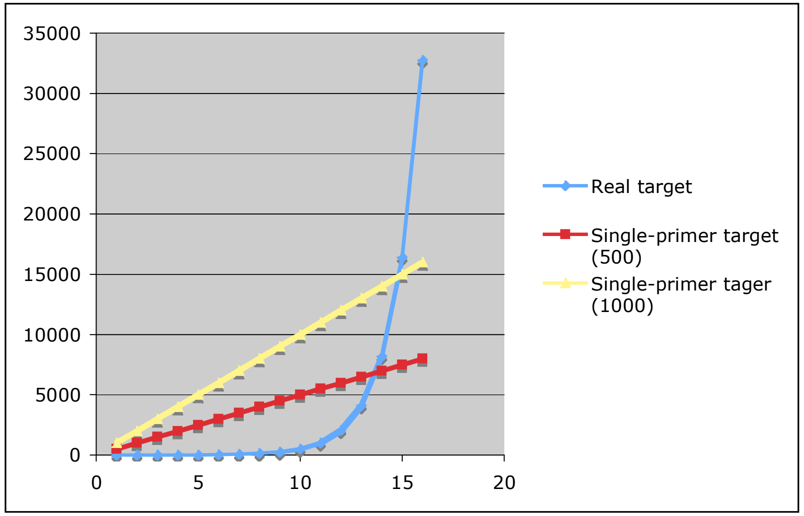
Extension
72º , 30 sec

35
cycles

Animation: http://depts.washington.edu/~genetics/courses/genet371b-aut99/PCR_contents.html

Reminder: PCR

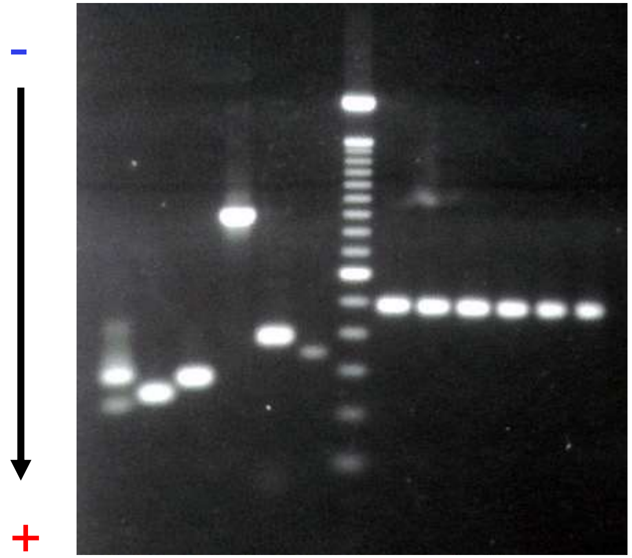


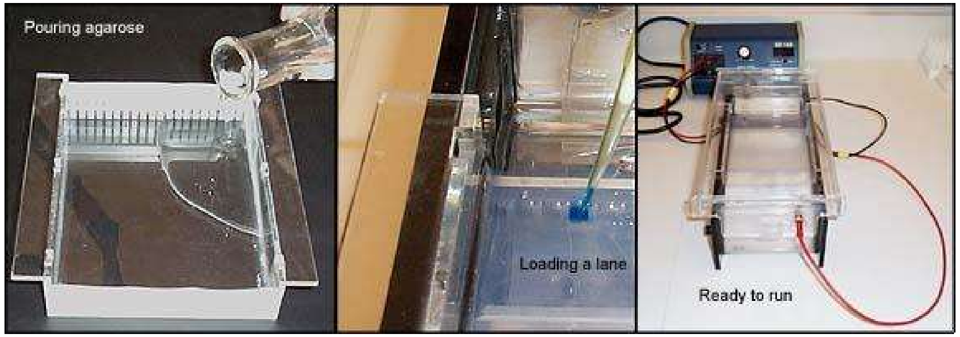


Animation: `http://www.people.virginia.edu/~rjh9u/pcranim.html`
PCR graph: `http://pathmicro.med.sc.edu/pcr/realtime-home.htm`

Gel electrophoresis

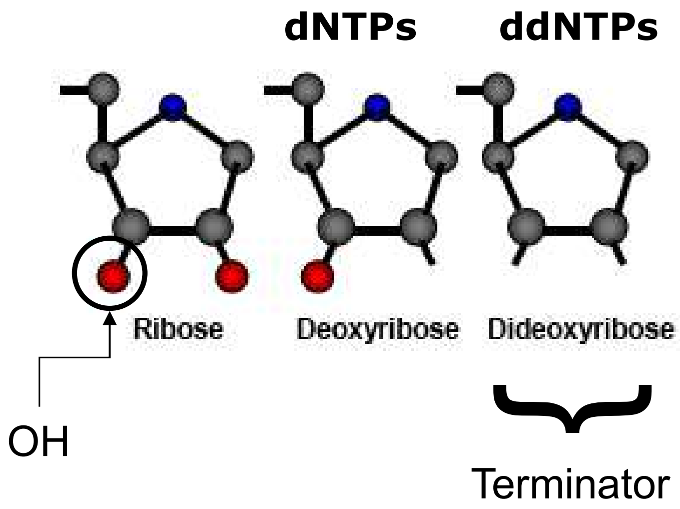
- DNA fragments are separated using gel electrophoresis
  - Typically 1% agarose
  - Colored with EtBr or ZybrGreen (glows in UV light).
  - A DNA ”ladder” is used for identification of known DNA lengths.

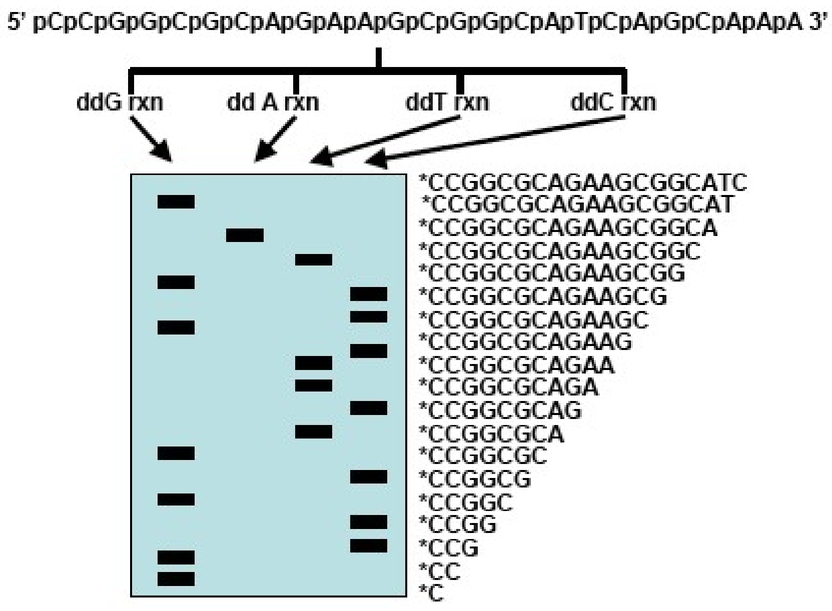




Gel picture: http://www.pharmaceutical-technology.com/projects/roche/images/roche3.jpg
PCR setup: http://arbl.cvmbs.colostate.edu/hbooks/genetics/biotech/gels/agardna.html

The Sanger method of DNA sequencing





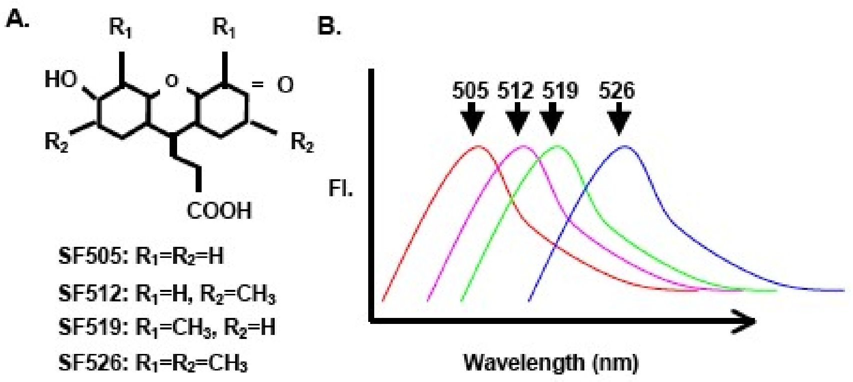
X-ray sequencing gel

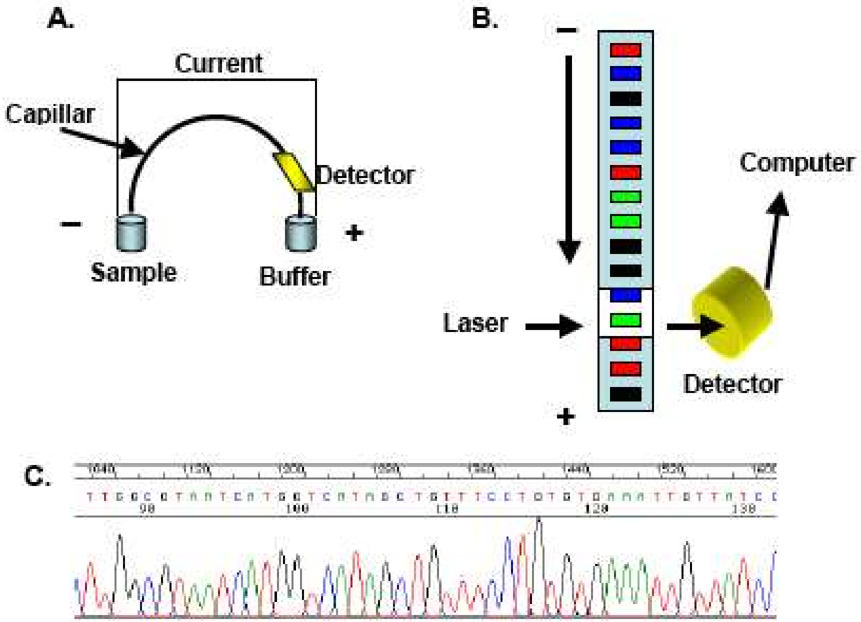
video: https://www.youtube.com/watch?v=FvHRio1yyhQ

Images: http://www.idtdna.com/support/technical/TechnicalBulletinPDF/DNA_Sequencing.pdf

Automated sequencing

- The major break-through of sequencing has happened through *automation.*
- Fluorescent dyes.
- Laser based scanning.
- Capillary electrophoresis
- Computer based base-calling and assembly.

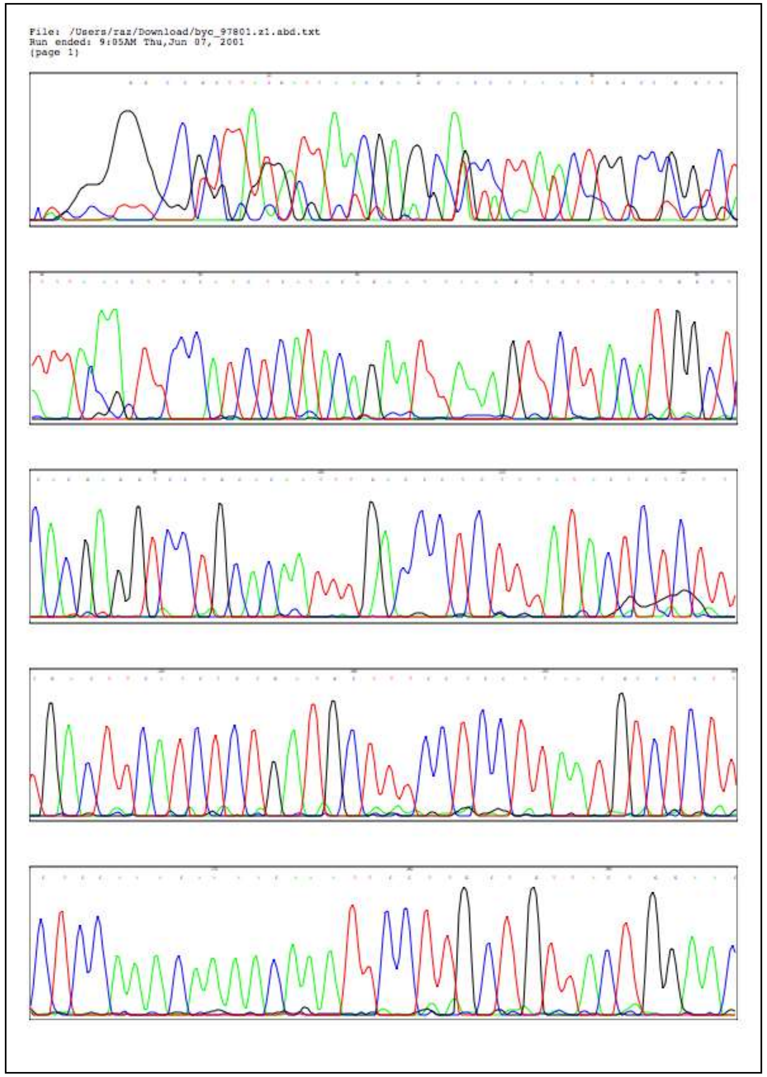




Images: `http://www.idtdna.com/support/technical/TechnicalBulletinPDF/DNA_Sequencing.pdf`

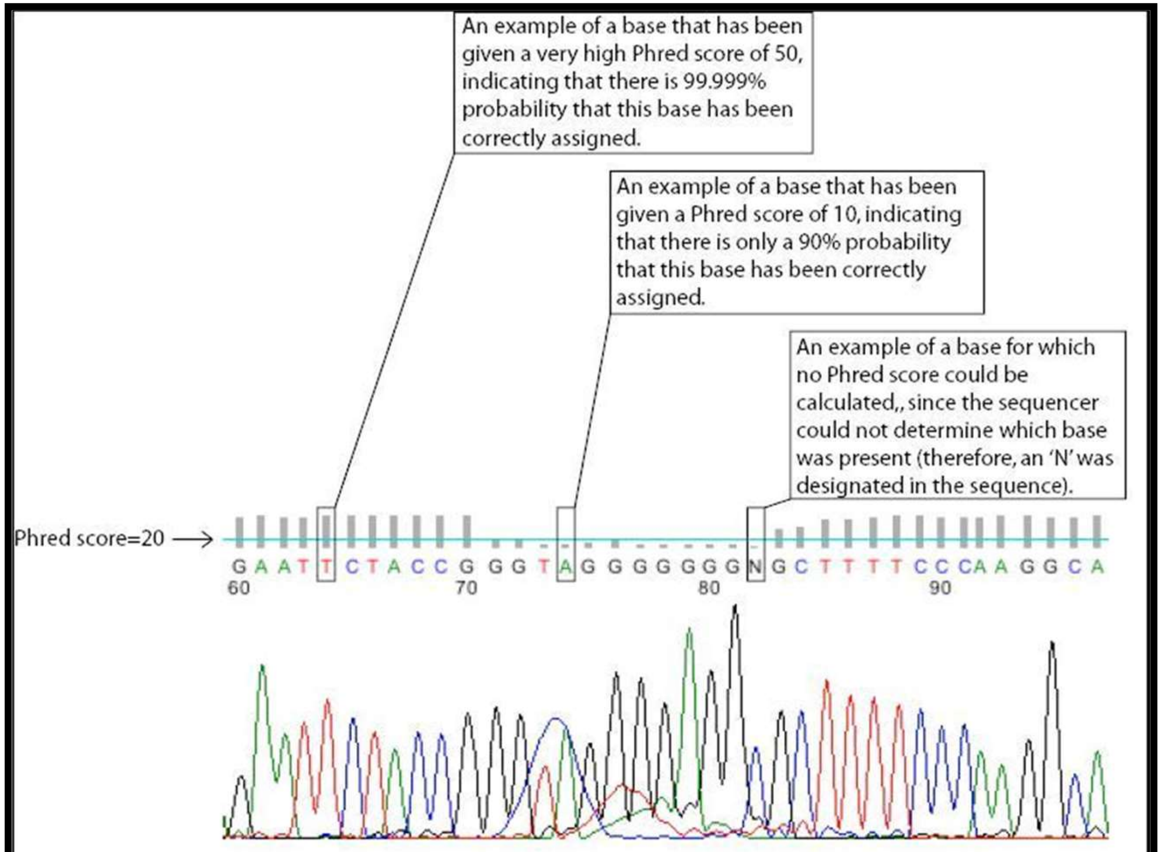
Handout exercise: ”base-calling”

- Handout: Chromatogram
- Groups of 2-3.
- Tasks:
  - Identify “difficult” regions
  - Identify likely errors
  - Try to estimate the best interval to use



File: /Users/raz/Download/byc_97801.z1.abd.txt
Run ended: 9:05AM Thu,Jun 07, 2001
(page 1)

Automatic assignment of quality



An example of a base that has been given a very high Phred score of 50, indicating that there is 99.999% probability that this base has been correctly assigned.

An example of a base that has been given a Phred score of 10, indicating that there is only a 90% probability that this base has been correctly assigned.

An example of a base for which no Phred score could be calculated,, since the sequencer could not determine which base was present (therefore, an 'N' was designated in the sequence).

Phred score=20 ⟶
G A A T T C T A C C G G G T A G G G G G G N G C T T T T C C C A A G G C A
60
70
80
90

Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy.

DNA sequencing - history

1972 Recombinant DNA technology [Paul Berg].

1976 The first sequenced genome, the bacteriophage MS2 [Walter Fiers *et al*.]

1977 DNA sequencing by chemical cleavage [Allan Maxam & Walter Gilbert]
DNA sequencing by enzymatic synthesis [Fred Sanger].

1982 *GenBank* (public database of DNA sequences).

1987 The first automatic sequencer, *Prism 373* [Applied Biosystems].

1990 *Human Genome Project* is launched.

1995 The first genome of a free-living organism, the bacterium *Haemophilus influenzae* (1.8 Mb) [The Institute for Genomic Research (TIGR)].

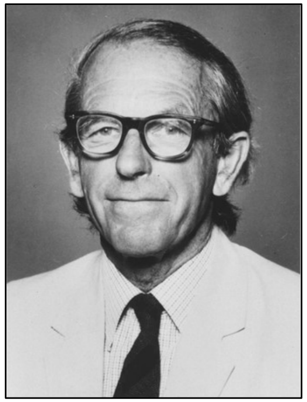
1996 The first genome of a eukaryote, Baker's Yeast, *Saccharomyces cerevisiae* (12.1 Mb) [International consortium].

1998 The first genome of an animal, the round worm *Caenorhabditis elegans* (97Mb) [Sanger Center and collaborators].

2001 The first "drafts" of the human genome (3Gb) [Human Genome Project Consortium (Nature, 15 Feb) + Celera (Science, 16 Feb)].

2000→ Development of several generations of "Next Generation Sequencing" (NGS)

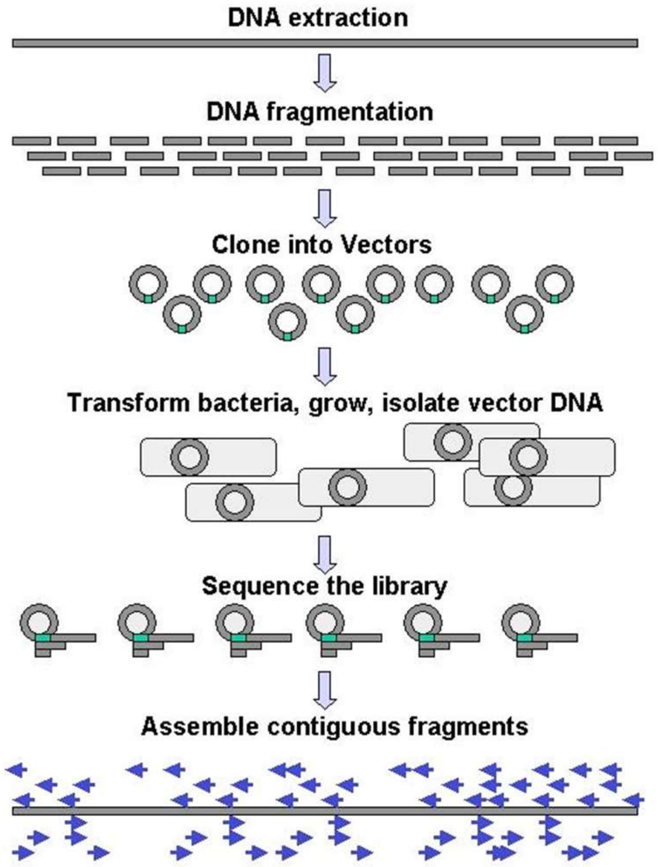
October 2022 *GenBank release 252* contains 20.35 trillion bases and 3.10 billion records (including *Whole Genome Shotgun* (WGS) sequences).



Frederick Sanger

Two Nobel Prizes – one for protein sequencing (1958) and one for DNA sequencing (1980)

“Shotgun sequencing”



DNA extraction

DNA fragmentation

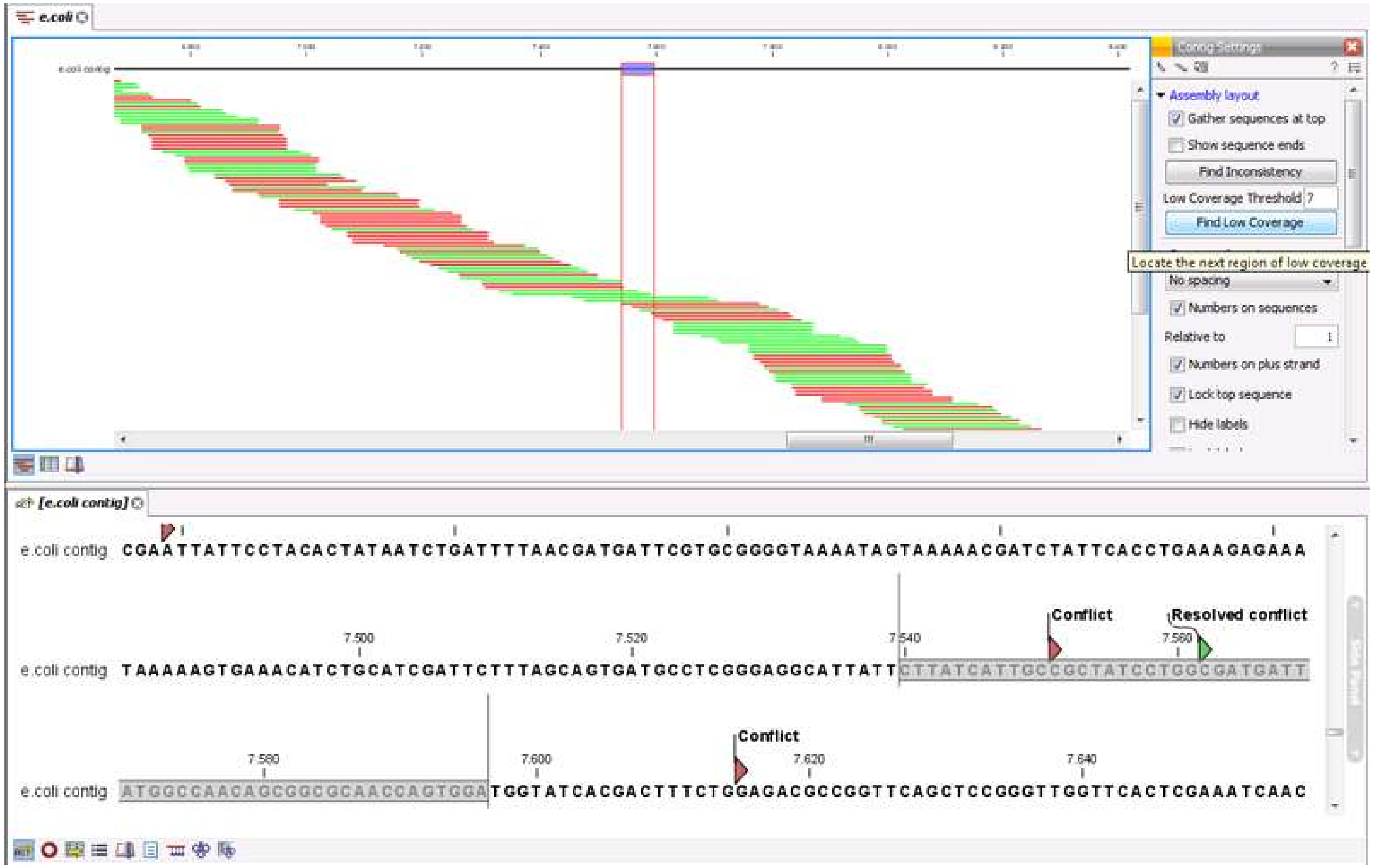
Clone into Vectors

Transform bacteria, grow, isolate vector DNA

Sequence the library

Assemble contiguous fragments

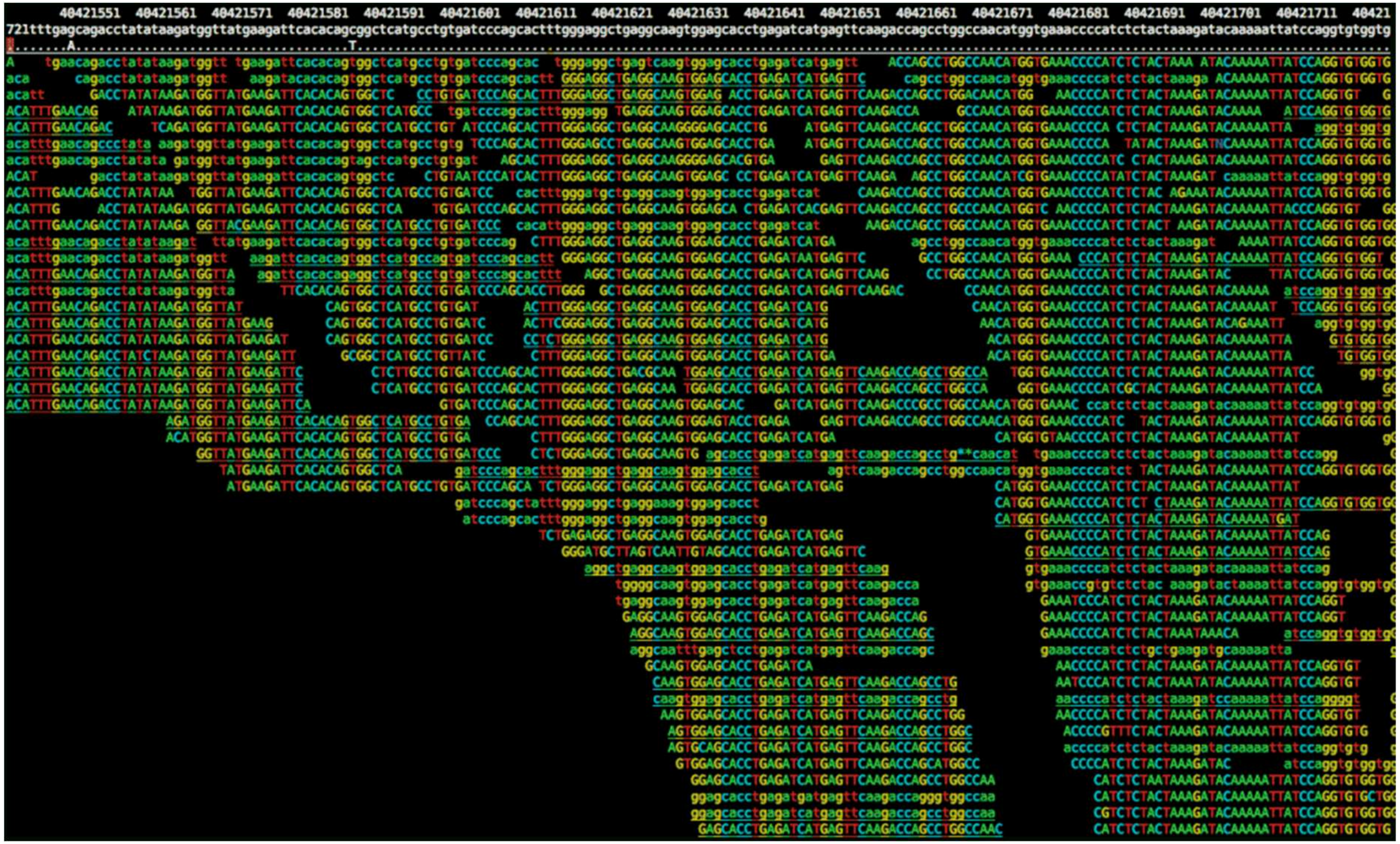
Sequence read mapping



- ☑ Gather sequences at top
- □ Show sequence ends
- Find Inconsistency
- Low Coverage Threshold 7
- Find Low Coverage
- Locate the next region of low coverage
- No spacing
- ☑ Numbers on sequences
- Relative to 1
- ☑ Numbers on plus strand
- ☑ Lock top sequence
- □ Hide labels

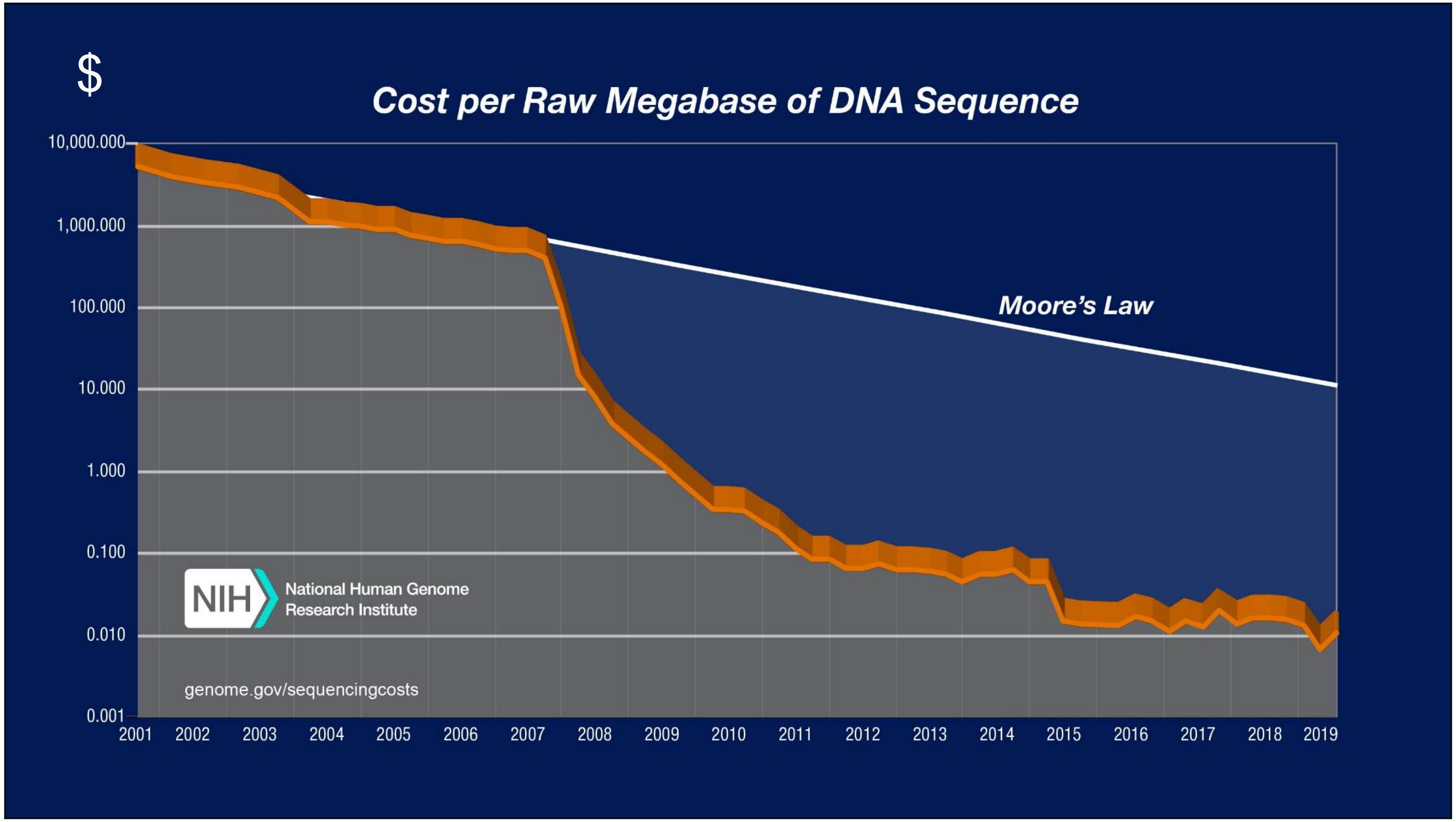
```
e.coli contig  CGAATTATTCCTACACTATAATCTGATTTTAACGATGATTCGTGCGGGGTAAAATAGTAAAAACGATCTATTCACCTGAAAGAGAAA
e.coli contig  TAAAAAGTGAAACATCTGCATCGATTCTTTAGCAGTGATGCCTCGGGAGGCATTATTCTTATCATTGCCGCTATCCTGGCGATGATT
e.coli contig  ATGGCCAACAGCGGCGCAACCAGTGGATGGTATCACGACTTTCTGGAGACGCCGGTTCAGCTCCGGGTTGGTTCACTCGAAATCAAC
```

NGS read mapping

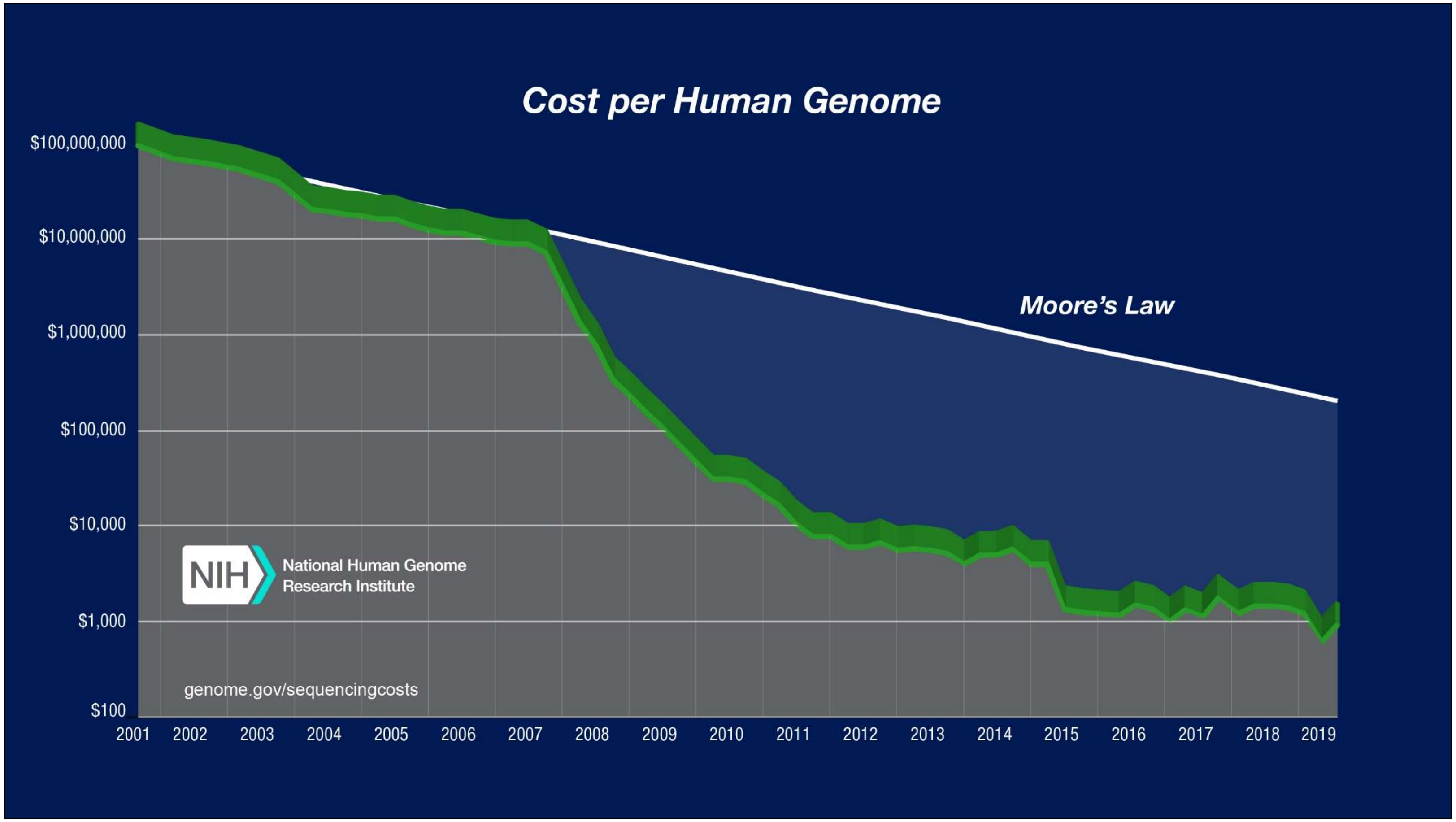


```
      40421551  40421561  40421571  40421581  40421591  40421601  40421611  40421621  40421631  40421641  40421651  40421661  40421671  40421681  40421691  40421701  40421711  40421
721tttgagcagacctatataagatggttatgaagattcacacagcggctcatgcctgtgatcccagcactttgggaggctgaggcaagtggagcacctgagatcatgagttcaagaccagcctggccaacatggtgaaaccccatctctactaaagatacaaaaattatccaggtgtggtg
........A...........................T...............................................................................................................................................
```

Cost of sequencing



Cost of sequencing the human genome



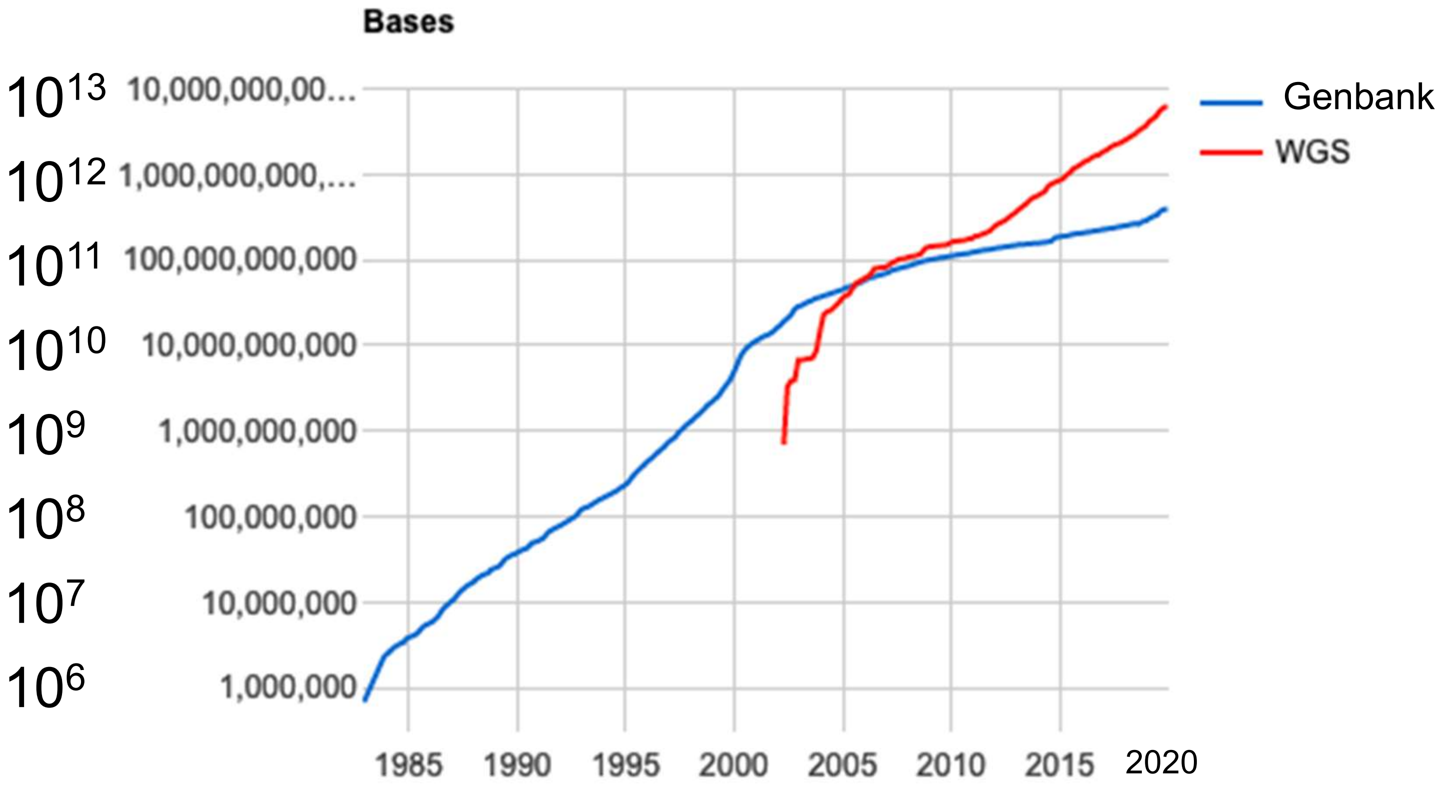
DNA data bases and data formats

Background - Nucleotide databases

- **GenBank**, http://www.ncbi.nlm.nih.gov/Genbank/
- National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), USA
- Established in 1982.
- **EMBL**, http://www.ebi.ac.uk/embl/
- European Bioinformatics Institute (EBI), England
- Established in 1980 by the European Molecular Biology Laboratory, Heidelberg, Germany
- Now part of **ENA**, the European Nucleotide Archive, http://www.ebi.ac.uk/ena/
- **DDBJ**, http://www.ddbj.nig.ac.jp/
- National Institute of Genetics, Japan

- *Together they form*
- International Nucleotide Sequence Database Collaboration, http://www.insdc.org/

Nucleotide database growth



NCBI: Growth in public available sequence databases

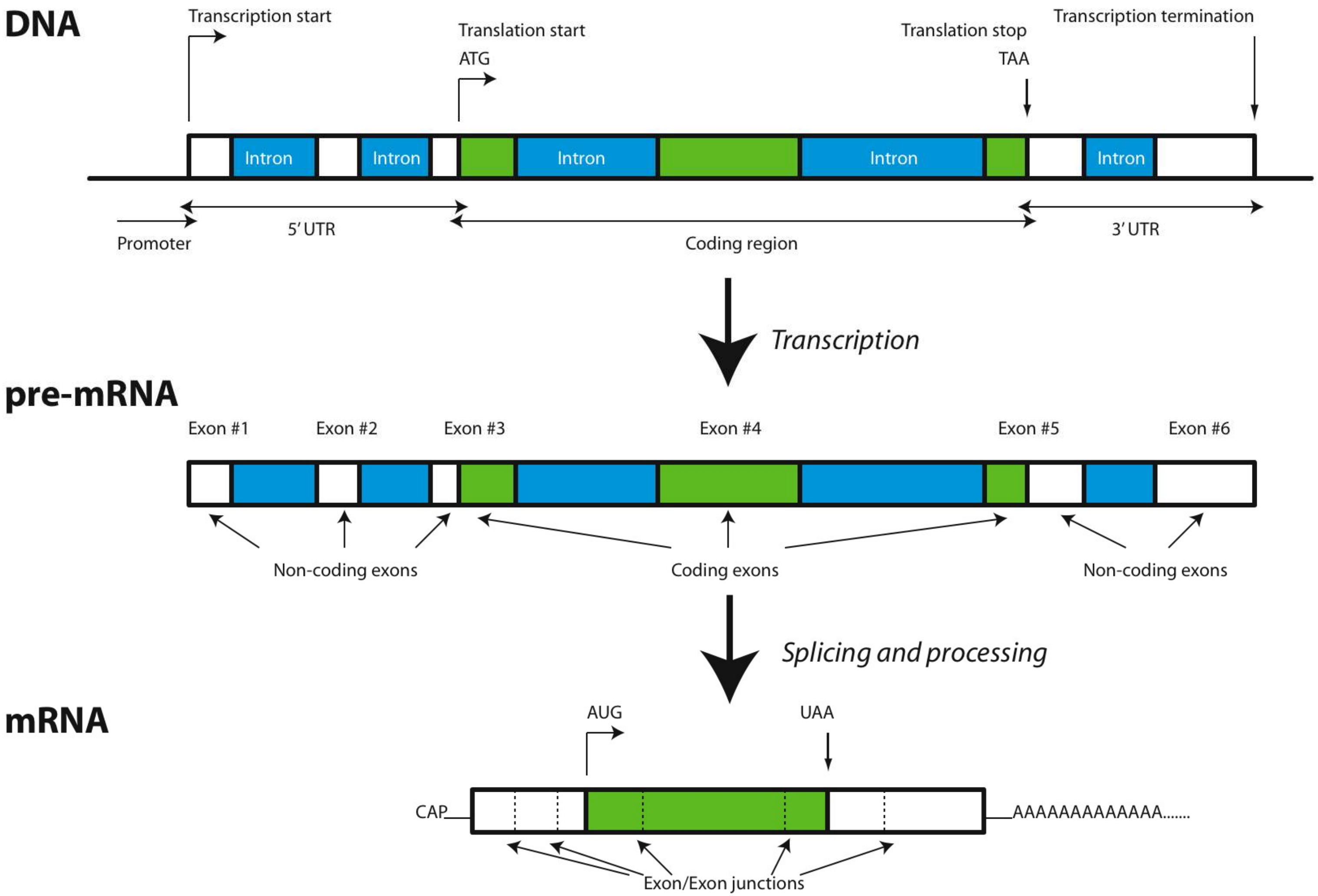
FASTA format

```
>alpha-D
ATGCTGACCGACTCTGACAAGAAGCTGGTCCTGCAGGTGTGGGAGAAGGTGATCCGCCAC
CCAGACTGTGGAGCCGAGGCCCTGGAGAGGTGCGGGCTGAGCTTGGGGAAACCATGGGCA
AGGGGGGCGACTGGGTGGGAGCCCTACAGGGCTGCTGGGGGTTGTTCGGCTGGGGGTCAG
CACTGACCATCCCGCTCCCGCAGCTGTTCACCACCTACCCCCAGACCAAGACCTACTTCC
CCCACTTCGACTTGCACCATGGCTCCGACCAGGTCCGCAACCACGGCAAGAAGGTGTTGG
CCGCCTTGGGCAACGCTGTCAAGAGCCTGGGCAACCTCAGCCAAGCCCTGTCTGACCTCA
GCGACCTGCATGCCTACAACCTGCGTGTCGACCCTGTCAACTTCAAGGCAGGCGGGGGAC
GGGGGTCAGGGGCCGGGGAGTTGGGGGCCAGGGACCTGGTTGGGGATCCGGGGCCATGCC
GGCGGTACTGAGCCCTGTTTTGCCTTGCAGCTGCTGGCGCAGTGCTTCCACGTGGTGCTG
GCCACACACCTGGGCAACGACTACACCCCGGAGGCACATGCTGCCTTCGACAAGTTCCTG
TCGGCTGTGTGCACCGTGCTGGCCGAGAAGTACAGATAA
>alpha-A
ATGGTGCTGTCTGCCAACGACAAGAGCAACGTGAAGGCCGTCTTCGGCAAAATCGGCGGC
CAGGCCGGTGACTTGGGTGGTGAAGCCCTGGAGAGGTATGTGGTCATCCGTCATTACCCC
ATCTCTTGTCTGTCTGTGACTCCATCCCATCTGCCCCCATACTCTCCCCATCCATAACTG
TCCCTGTTCTATGTGGCCCTGGCTCTGTCTCATCTGTCCCCAACTGTCCCTGATTGCCTC
TGTCCCCCAGGTTGTTCATCACCTACCCCCAGACCAAGACCTACTTCCCCCACTTCGACC
TGTCACATGGCTCCGCTCAGATCAAGGGGCACGGCAAGAAGGTGGCGGAGGCACTGGTTG
AGGCTGCCAACCACATCGATGACATCGCTGGTGCCCTCTCCAAGCTGAGCGACCTCCACG
CCCAAAAGCTCCGTGTGGACCCCGTCAACTTCAAAGTGAGCATCTGGGAAGGGGTGACCA
GTCTGGCTCCCCTCCTGCACACACCTCTGGCTACCCCCTCACCTCACCCCCTTGCTCACC
ATCTCCTTTTGCCTTTCAGCTGCTGGGTCACTGCTTCCTGGTGGTCGTGGCCGTCCACTT
CCCCTCTCTCCTGACCCCGGAGGTCCATGCTTCCCTGGACAAGTTCGTGTGTGCCGTGGG
CACCGTCCTTACTGCCAAGTACCGTTAA
```

So we got the DNA sequence – now what?

```
CGATCAGGTTACATTTACTGCCCATGCCTGTCTCAGAGGAATTCTGACACGAAAAGGTGGGCACAAATTC
TTAAGCACACTCTGATGGTACAACGTGAGCTGGCACTACAAGCTGTGTTCCTCATCCCGTTTACAAAATT
TTGAGACTGTGTTTGGGCAAGGGGGAGAGAGACAGTGCAGAAGCTCTGAAGCCACTGAATTTCTCTAAAT
GTGTTTAGAGAAGCTTTTCAAACATGCTACATTTGTGGGTCTCAAGTAACCCGAACATTTAAATCCACAG
CTGAGGTGGTAGGTCAGCAACTGGTGTATCCCTAACTCCAAGTTTGTACCAAAGACTTTAACAGAGGCTG
AGTGAAGGATGTGACAGTCACCAGCCACCATATGCCCTGCCAAATGTCCCAGTGTTTACAGAGGGATAGC
AACACACACTGGGGTGGGAAGAAGGAAAGAAGACCAGGCCTGACAAGCATCACAGGATGGATTTTGGGAA
GACTATGGACCTGAAAAGGAGATTCTTCCCCACTCAGGTCTCTCCAGGATGCTGGGGAGATGCTGTTTCC
TGTGGTAGATCCCCAGCATGAACCAGGAGGGCATGTCCGTGGCTCCTGCTCTGAGGCTCACAGTGTCTTT
GGGTGGAGAGGGGATGGATGCACTGGGGTCTGGAGGACATGAGGGACTGGGGGCGCTCGTGGGATCTACT
CTGACACCTGCAGAGACAGGGAGGACCCTGGCCTGGCCAGAAGGGAATGGTGGATCCCAACAGGAAGCTT
GAGGATATGCAGGTTTGTGAGGCCGAGGCTGTGGCACCCGTGGGACATGCCGATGGCTGCTGTTGACCAT
GGGGCAGCTCAGCCAAGTGCTGCCCCCAGCCCCCAGCCCAGCGTGGGGCTGGTGCAGTGCGGCACATCAG
GGCAGGGCAGCCGCCCCATTGGGGCCCCCTCGGGGCTGGGCCTCCCAGGGCAGTCGGGGCCCCCTGAGGC
AGTGGCCCCCCACCCCTTGGTGCCGATAAGATAACGCTGGGGCGGAGGTGCCGACCACTATAAGAGGATG
TCCTGGTGGGCCCTGCTACCACTGAGCCCTGACCGCCACCCCCAGCCGCCACCATGCTGACCGACTCTGA
CAAGAAGCTGGTCCTGCAGGTGTGGGAGAAGGTGATCCGCCACCCAGACTGTGGAGCCGAGGCCCTGGAG
AGGTGCGGGCTGAGCTTGGGGAAACCATGGGCAAGGGGGGCGACTGGGTGGGAGCCCTACAGGGCTGCTG
GGGGTTGTTCGGCTGGGGGTCAGCACTGACCATCCCGCTCCCGCAGCTGTTCACCACCTACCCCCAGACC
AAGACCTACTTCCCCCACTTCGACTTGCACCATGGCTCCGACCAGGTCCGCAACCACGGCAAGAAGGTGT
TGGCCGCCTTGGGCAACGCTGTCAAGAGCCTGGGCAACCTCAGCCAAGCCCTGTCTGACCTCAGCGACCT
GCATGCCTACAACCTGCGTGTCGACCCTGTCAACTTCAAGGCAGGCGGGGGACGGGGGTCAGGGGCCGGG
GAGTTGGGGGCCAGGGACCTGGTTGGGGATCCGGGGCCATGCCGGCGGTACTGAGCCCTGTTTTGCCTTG
CAGCTGCTGGCGCAGTGCTTCCACGTGGTGCTGGCCACACACCTGGGCAACGACTACACCCCGGAGGCAC
ATGCTGCCTTCGACAAGTTCCTGTCGGCTGTGTGCACCGTGCTGGCCGAGAAGTACAGATAAGCCATCGC
```

Reminder: Eukaryotic gene structure



DNA

Transcription start

Translation start

ATG

Translation stop

TAA

Transcription termination

Intron

Intron

Intron

Intron

Intron

Promoter

5' UTR

Coding region

3' UTR

Transcription

pre-mRNA

Exon #1

Exon #2

Exon #3

Exon #4

Exon #5

Exon #6

Non-coding exons

Coding exons

Non-coding exons

Splicing and processing

mRNA

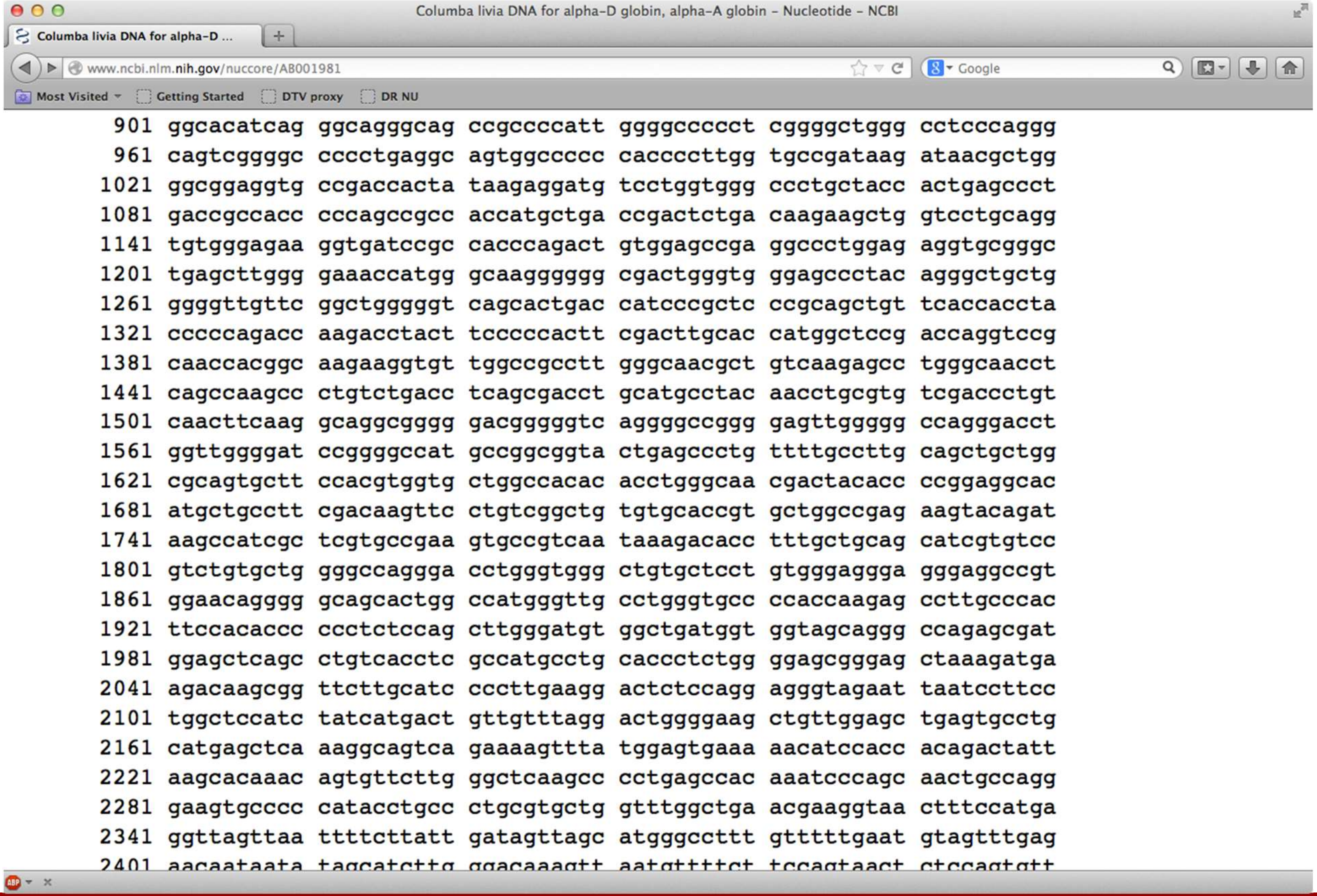
AUG

UAA

CAP

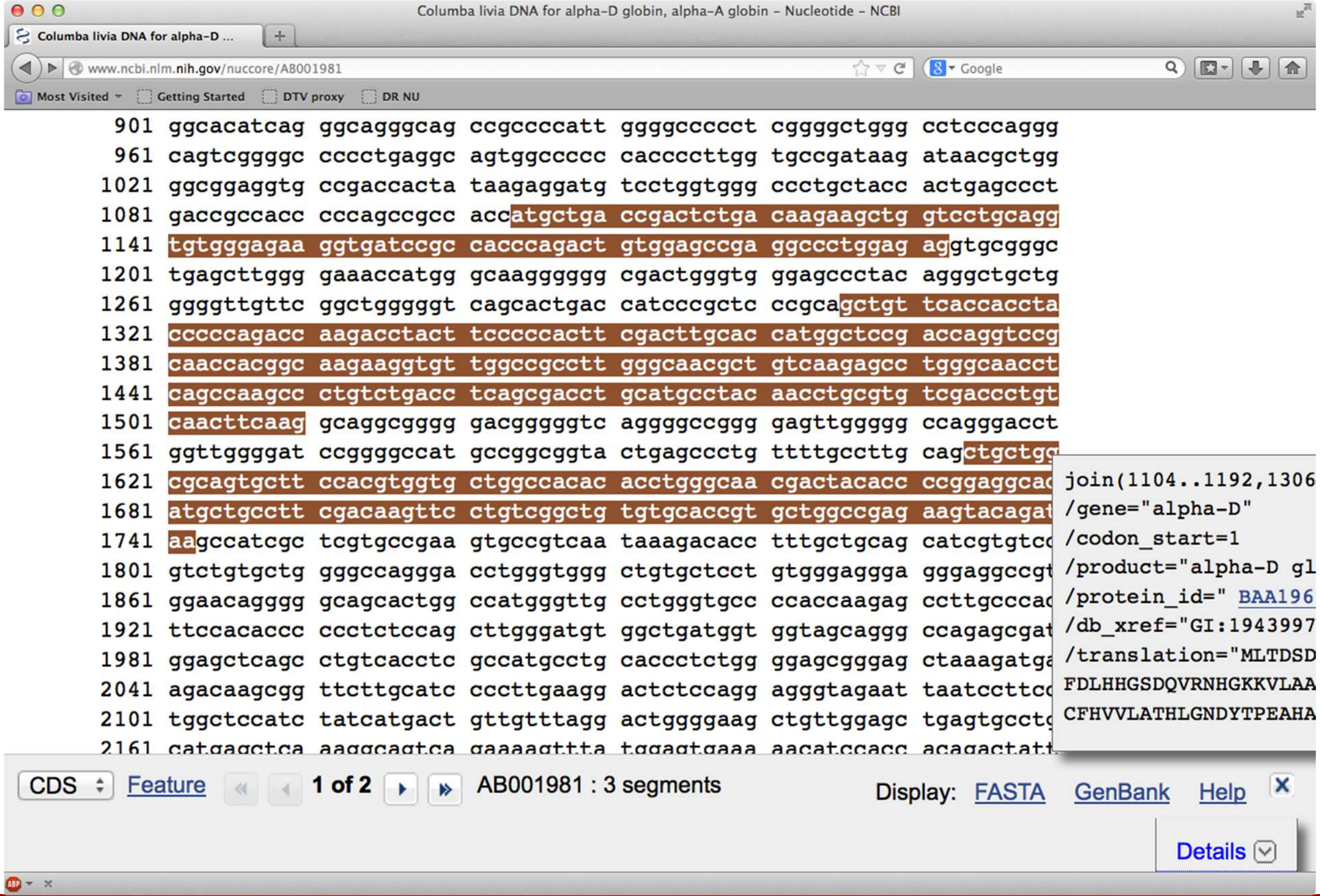
AAAAAAAAAAAA.......

Exon/Exon junctions



Columba livia DNA for alpha-D globin, alpha-A globin - Nucleotide - NCBI
Columba livia DNA for alpha-D ...
www.ncbi.nlm.nih.gov/nuccore/AB001981
Google
Most Visited
Getting Started
DTV proxy
DR NU

```
 901 ggcacatcag ggcagggcag ccgccccatt ggggcccccct cggggctggg cctcccaggg
 961 cagtcggggc cccctgaggc agtggccccc caccccttgg tgccgataag ataacgctgg
1021 ggcggaggtg ccgaccacta taagaggatg tcctggtggg ccctgctacc actgagccct
1081 gaccgccacc cccagccgcc accatgctga ccgactctga caagaagctg gtcctgcagg
1141 tgtgggagaa ggtgatccgc cacccagact gtggagccga ggccctggag aggtgcgggc
1201 tgagcttggg gaaaccatgg gcaagggggg cgactgggtg ggagccctac agggctgctg
1261 ggggttgttc ggctgggggt cagcactgac catcccgctc ccgcagctgt tcaccaccta
1321 cccccagacc aagacctact tcccccactt cgacttgcac catggctccg accaggtccg
1381 caaccacggc aagaaggtgt tggccgcctt gggcaacgct gtcaagagcc tgggcaacct
1441 cagccaagcc ctgtctgacc tcagcgacct gcatgcctac aacctgcgtg tcgaccctgt
1501 caacttcaag gcaggcgggg gacgggggtc aggggccggg gagttggggg ccagggacct
1561 ggttggggat ccggggccat gccggcggta ctgagccctg ttttgccttg cagctgctgg
1621 cgcagtgctt ccacgtggtg ctggccacac acctgggcaa cgactacacc ccggaggcac
1681 atgctgcctt cgacaagttc ctgtcggctg tgtgcaccgt gctggccgag aagtacagat
1741 aagccatcgc tcgtgccgaa gtgccgtcaa taaagacacc tttgctgcag catcgtgtcc
1801 gtctgtgctg gggccaggga cctgggtggg ctgtgctcct gtgggaggga gggaggccgt
1861 ggaacagggg gcagcactgg ccatgggttg cctgggtgcc ccaccaagag ccttgcccac
1921 ttccacaccc ccctctccag cttgggatgt ggctgatggt ggtagcaggg ccagagcgat
1981 ggagctcagc ctgtcacctc gccatgcctg caccctctgg ggagcgggag ctaaagatga
2041 agacaagcgg ttcttgcatc cccttgaagg actctccagg agggtagaat taatccttcc
2101 tggctccatc tatcatgact gttgtttagg actggggaag ctgttggagc tgagtgcctg
2161 catgagctca aaggcagtca gaaaagttta tggagtgaaa aacatccacc acagactatt
2221 aagcacaaac agtgttcttg ggctcaagcc cctgagccac aaatcccagc aactgccagg
2281 gaagtgcccc catacctgcc ctgcgtgctg gtttggctga acgaaggtaa ctttccatga
2341 ggttagttaa ttttcttatt gatagttagc atgggccttt gtttttgaat gtagtttgag
2401 aacaataata tagcatcttg ggacaaagtt aatgttttct tccagtaact ctccagtgtt
```

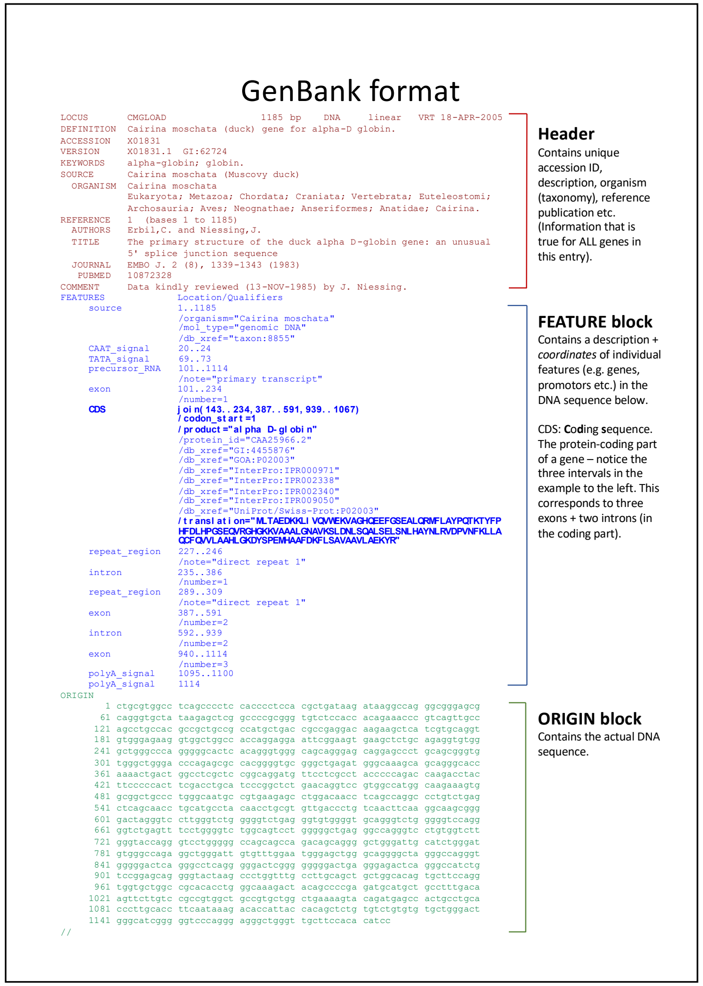

Columba livia DNA for alpha-D globin, alpha-A globin – Nucleotide – NCBI
Columba livia DNA for alpha-D ...
www.ncbi.nlm.nih.gov/nuccore/AB001981
Google
Most Visited
Getting Started
DTV proxy
DR NU

```
 901 ggcacatcag ggcagggcag ccgccccatt ggggccccct cggggctggg cctcccaggg
 961 cagtcggggc cccctgaggc agtggccccc cacccccttgg tgccgataag ataacgctgg
1021 ggcggaggtg ccgaccacta taagaggatg tcctggtggg ccctgctacc actgagccct
1081 gaccgccacc cccagccgcc accatgctga ccgactctga caagaagctg gtcctgcagg
1141 tgtgggagaa ggtgatccgc cacccagact gtggagccga ggccctggag aggtgcgggc
1201 tgagcttggg gaaaccatgg gcaagggggg cgactgggtg ggagccctac agggctgctg
1261 ggggttgttc ggctgggggt cagcactgac catcccgctc ccgcagctgt tcaccaccta
1321 cccccagacc aagacctact tcccccactt cgacttgcac catggctccg accaggtccg
1381 caaccacggc aagaaggtgt tggccgcctt gggcaacgct gtcaagagcc tgggcaacct
1441 cagccaagcc ctgtctgacc tcagcgacct gcatgcctac aacctgcgtg tcgaccctgt
1501 caacttcaag gcaggcgggg gacgggggtc aggggccggg gagttggggg ccagggacct
1561 ggttggggat ccggggccat gccggcggta ctgagccctg ttttgccttg cagctgctgg
1621 cgcagtgctt ccacgtggtg ctggccacac acctgggcaa cgactacacc ccggaggcac
1681 atgctgcctt cgacaagttc ctgtcggctg tgtgcaccgt gctggccgag aagtacagat
1741 aagccatcgc tcgtgccgaa gtgccgtcaa taaagacacc tttgctgcag catcgtgtcc
1801 gtctgtgctg gggccaggga cctgggtggg ctgtgctcct gtgggaggga gggaggccgt
1861 ggaacagggg gcagcactgg ccatgggttg cctgggtgcc ccaccaagag ccttgcccac
1921 ttccacaccc ccctctccag cttgggatgt ggctgatggt ggtagcaggg ccagagcgat
1981 ggagctcagc ctgtcacctc gccatgcctg caccctctgg ggagcgggag ctaaagatga
2041 agacaagcgg ttcttgcatc cccttgaagg actctccagg agggtagaat taatccttcc
2101 tggctccatc tatcatgact gttgtttagg actggggaag ctgttggagc tgagtgcctc
2161 catgagctca aaggcagtca gaaaagttta tggagtgaaa aacatccacc acagactatt
```

```
join(1104..1192,1306
/gene="alpha-D"
/codon_start=1
/product="alpha-D gl
/protein_id=" BAA196
/db_xref="GI:1943997
/translation="MLTDSD
FDLHHGSDQVRNHGKKVLAA
CFHVVLATHLGNDYTPEAHA
```

AB001981 : 3 segments

GenBank format



GenBank format

```
LOCUS       CMGLOAD                 1185 bp    DNA     linear   VRT 18-APR-2005
DEFINITION  Cairina moschata (duck) gene for alpha-D globin.
ACCESSION   X01831
VERSION     X01831.1  GI:62724
KEYWORDS    alpha-globin; globin.
SOURCE      Cairina moschata (Muscovy duck)
  ORGANISM  Cairina moschata
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.
REFERENCE   1  (bases 1 to 1185)
  AUTHORS   Erbil,C. and Niessing,J.
  TITLE     The primary structure of the duck alpha D-globin gene: an unusual
            5' splice junction sequence
  JOURNAL   EMBO J. 2 (8), 1339-1343 (1983)
   PUBMED   10872328
COMMENT     Data kindly reviewed (13-NOV-1985) by J. Niessing.
FEATURES             Location/Qualifiers
     source          1..1185
                     /organism="Cairina moschata"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:8855"
     CAAT_signal     20..24
     TATA_signal     69..73
     precursor_RNA   101..1114
                     /note="primary transcript"
     exon            101..234
                     /number=1
     CDS             join(143..234, 387..591, 939..1067)
                     /codon_start=1
                     /product="alpha D-globin"
                     /protein_id="CAA25966.2"
                     /db_xref="GI:4455876"
                     /db_xref="GOA:P02003"
                     /db_xref="InterPro:IPR000971"
                     /db_xref="InterPro:IPR002338"
                     /db_xref="InterPro:IPR002340"
                     /db_xref="InterPro:IPR009050"
                     /db_xref="UniProt/Swiss-Prot:P02003"
                     /translation="MLTAEDKKLIVQVWEKVAGHQEEFGSEALQRMFLAYPQTKTYFP
                     HFDLHPGSEQVRGHGKKVAAALGNAVKSLDNLSQALSELSNLHAYNLRVDPVNFKLLA
                     QCFQVVLAAHLGKDYSPEMHAAFDKFLSAVAAVLAEKYR"
     repeat_region   227..246
                     /note="direct repeat 1"
     intron          235..386
                     /number=1
     repeat_region   289..309
                     /note="direct repeat 1"
     exon            387..591
                     /number=2
     intron          592..939
                     /number=2
     exon            940..1114
                     /number=3
     polyA_signal    1095..1100
     polyA_signal    1114
ORIGIN
        1 ctgcgtggcc tcagcccctc cacccctcca cgctgataag ataaggccag ggcgggagcg
       61 cagggtgcta taagagctcg gccccgcggg tgtctccacc acagaaaccc gtcagttgcc
      121 agcctgccac gccgctgccg ccatgctgac cgccgaggac aagaagctca tcgtgcaggt
      181 gtgggagaag gtggctggcc accaggagga attcggaagt gaagctctgc agaggtgtgg
      241 gctgggccca gggggcactc acagggtggg cagcagggag caggagccct gcagcgggtg
      301 tgggctggga cccagagcgc cacggggtgc gggctgagat gggcaaagca gcagggcacc
      361 aaaactgact ggcctcgctc cggcaggatg ttcctcgcct acccccagac caagacctac
      421 ttcccccact tcgacctgca tcccggctct gaacaggtcc gtggccatgg caagaaagtg
      481 gcggctgccc tgggcaatgc cgtgaagagc ctggacaacc tcagccaggc cctgtctgag
      541 ctcagcaacc tgcatgccta caacctgcgt gttgaccctg tcaacttcaa ggcaagcggg
      601 gactagggtc cttgggtctg gggtctgag ggtgtggggt gcagggtctg ggggtccagg
      661 ggtctgagtt tcctggggtc tggcagtcct gggggctgag ggccagggtc ctgtggtctt
      721 gggtaccagg gtcctggggg ccagcagcca gacagcaggg gctgggattg catctgggat
      781 gtgggccaga ggctgggatt gtgtttggaa tgggagctgg gcaggggcta gggccagggt
      841 gggggactca gggcctcagg gggactcggg gggggactga gggagactca gggccatctg
      901 tccggagcag gggtactaag ccctggtttg ccttgcagct gctggcacag tgcttccagg
      961 tggtgctggc cgcacacctg ggcaaagact acagccccga gatgcatgct gcctttgaca
     1021 agttcttgtc cgccgtggct gccgtgctgg ctgaaaagta cagatgagcc actgcctgca
     1081 cccttgcacc ttcaataaag acaccattac cacagctctg tgtctgtgtg tgctgggact
     1141 gggcatcggg ggtcccaggg agggctgggt tgcttccaca catcc
//
```

Header
Contains unique accession ID, description, organism (taxonomy), reference publication etc. (Information that is true for ALL genes in this entry).

FEATURE block
Contains a description + *coordinates* of individual features (e.g. genes, promotors etc.) in the DNA sequence below.

CDS: Coding sequence. The protein-coding part of a gene – notice the three intervals in the example to the left. This corresponds to three exons + two introns (in the coding part).

ORIGIN block
Contains the actual DNA sequence.

- Originates from the GenBank database.
- Contains both a DNA sequence and annotations of features (*e.g.* location of genes).

GenBank format - HEADER

```
LOCUS       CMGLOAD                 1185 bp    DNA     linear   VRT 18-APR-2005
DEFINITION  Cairina moschata (duck) gene for alpha-D globin.
ACCESSION   X01831
VERSION     X01831.1  GI:62724
KEYWORDS    alpha-globin; globin.
SOURCE      Cairina moschata (Muscovy duck)
  ORGANISM  Cairina moschata
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.
REFERENCE   1  (bases 1 to 1185)
  AUTHORS   Erbil,C. and Niessing,J.
  TITLE     The primary structure of the duck alpha D-globin gene: an unusual
            5' splice junction sequence
  JOURNAL   EMBO J. 2 (8), 1339-1343 (1983)
   PUBMED   10872328
COMMENT     Data kindly reviewed (13-NOV-1985) by J. Niessing.
```

GenBank format - ORIGIN section

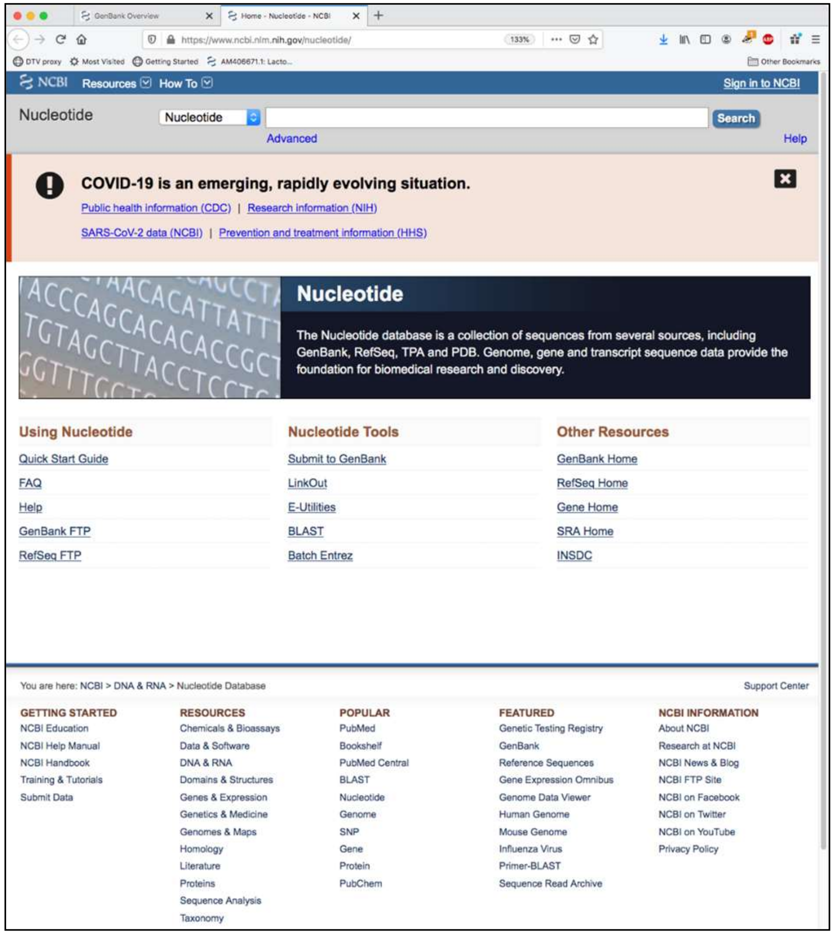
```
ORIGIN
        1 ctgcgtggcc tcagcccctc caccccctcca cgctgataag ataaggccag ggcgggagcg
       61 cagggtgcta taagagctcg gccccgcggg tgtctccacc acagaaaccc gtcagttgcc
      121 agcctgccac gccgctgccg ccatgctgac cgccgaggac aagaagctca tcgtgcaggt
      181 gtgggagaag gtggctggcc accaggagga attcggaagt gaagctctgc agaggtgtgg
      241 gctgggccca gggggcactc acagggtggg cagcagggag caggagccct gcagcgggtg
      301 tgggctggga cccagagcgc cacggggtgc gggctgagat gggcaaagca gcagggcacc
      361 aaaactgact ggcctcgctc cggcaggatg ttcctcgcct acccccagac caagacctac
      421 ttcccccact tcgacctgca tcccggctct gaacaggtcc gtggccatgg caagaaagtg
      481 gcggctgccc tgggcaatgc cgtgaagagc ctggacaacc tcagccaggc cctgtctgag
      541 ctcagcaacc tgcatgccta caacctgcgt gttgaccctg tcaacttcaa ggcaagcggg
      601 gactagggtc cttgggtctg ggggtctgag ggtgtggggt gcagggtctg ggggtccagg
      661 ggtctgagtt tcctggggtc tggcagtcct gggggctgag ggccagggtc ctgtggtctt
      721 gggtaccagg gtcctggggg ccagcagcca gacagcaggg gctgggattg catctgggat
      781 gtgggccaga ggctgggatt gtgtttggaa tgggagctgg gcagggggcta gggccagggt
      841 gggggactca gggcctcagg gggactcggg gggggactga gggagactca gggccatctg
      901 tccggagcag gggtactaag ccctggtttg ccttgcagct gctggcacag tgcttccagg
      961 tggtgctggc cgcacacctg ggcaaagact acagccccga gatgcatgct gcctttgaca
     1021 agttcttgtc cgcccgtggct gccgtgctgg ctgaaaagta cagatgagcc actgcctgca
     1081 cccttgcacc ttcaataaag acaccattac cacagctctg tgtctgtgtg tgctgggact
     1141 gggcatcggg ggtcccaggg agggctgggt tgcttccaca catcc
//
```

GenBank format - FEATURE section

```
FEATURES             Location/Qualifiers
     source          1..1185
                     /organism="Cairina moschata"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:8855"
     CAAT_signal     20..24
     TATA_signal     69..73
     precursor_RNA   101..1114
                     /note="primary transcript"
     exon            101..234
                     /number=1
     CDS             join(143..234,387..591,939..1067)
                     /codon_start=1
                     /product="alpha D-globin"
                     /protein_id="CAA25966.2"
                     /db_xref="GI:4455876"
                     /db_xref="GOA:P02003"
                     /db_xref="InterPro:IPR000971"
                     /db_xref="InterPro:IPR002338"
                     /db_xref="InterPro:IPR002340"
                     /db_xref="InterPro:IPR009050"
                     /db_xref="UniProt/Swiss-Prot:P02003"
                     /translation="MLTAEDKKLIVQVWEKVAGHQEEFGSEALQRMFLAYPQTKTYFP
                     HFDLHPGSEQVRGHGKKVAAALGNAVKSLDNLSQALSELSNLHAYNLRVDPVNFKLLA
                     QCFQVVLAAHLGKDYSPEMHAAFDKFLSAVAAVLAEKYR"
     repeat_region   227..246
                     /note="direct repeat 1"
     intron          235..386
                     /number=1
     repeat_region   289..309
                     /note="direct repeat 1"
     exon            387..591
                     /number=2
     intron          592..939
                     /number=2
     exon            940..1114
                     /number=3
     polyA_signal    1095..1100
     polyA_signal    1114
```

Exercise: GenBank

- The exercise guide is linked from the course programme.
- Read the guide carefully - it contains a lot of information about GenBank.
- Remember your handouts:
  - –GenBank & FASTA format
  - –Eukaryotic gene structure



Nucleotide

Nucleotide

COVID-19 is an emerging, rapidly evolving situation.

Public health information (CDC) | Research information (NIH)

SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)

Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

Using Nucleotide	Nucleotide Tools	Other Resources
Quick Start Guide	Submit to GenBank	GenBank Home
FAQ	LinkOut	RefSeq Home
Help	E-Utilities	Gene Home
GenBank FTP	BLAST	SRA Home
RefSeq FTP	Batch Entrez	INSDC

You are here: NCBI > DNA & RNA > Nucleotide Database

Support Center

GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials
Submit Data

RESOURCES
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy

POPULAR
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

FEATURED
Genetic Testing Registry
GenBank
Reference Sequences
Gene Expression Omnibus
Genome Data Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive

NCBI INFORMATION
About NCBI
Research at NCBI
NCBI News & Blog
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube
Privacy Policy

Learning objectives

- Interpret DNA as Biological information **(4 letters is a representation of DNA)**
- Describe DNA sequencing techniques and DNA data **(not free of errors)**
- Identify file formats used to store DNA data **(annotation is important)**
- Recognize how information is stored in GenBank database