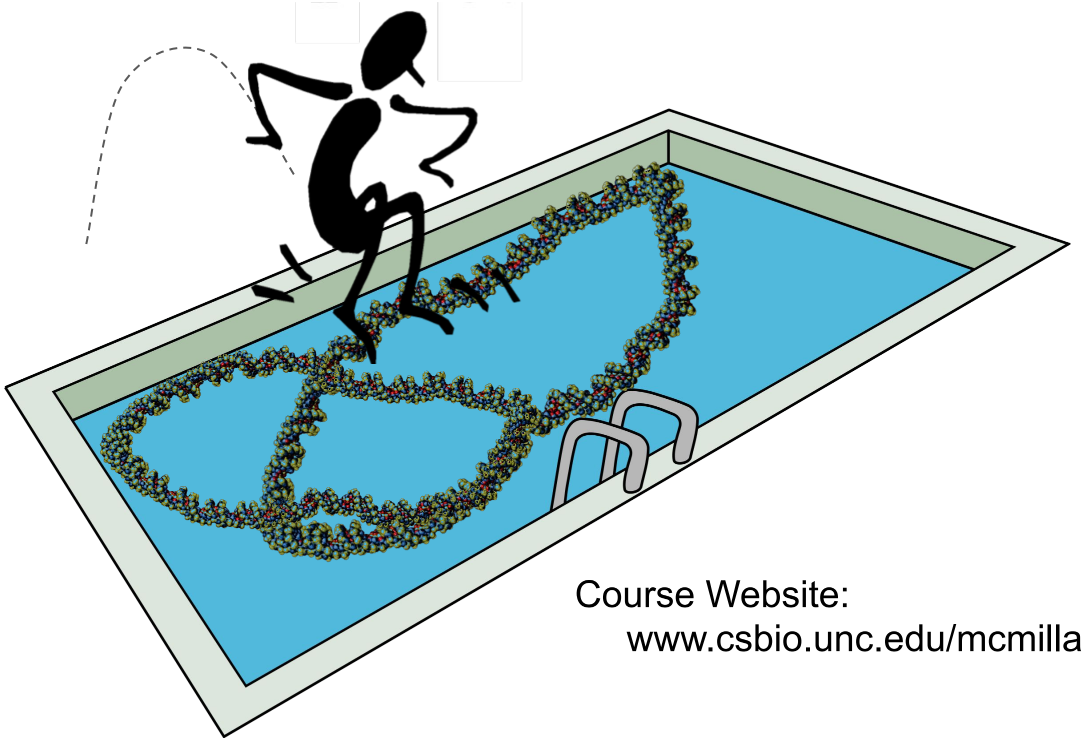
BCB 716 - Sequence Analysis



Course Website:
www.csbio.unc.edu/mcmillan/index.py?run=Courses.BCB716F22

Jumping into Genomes

Logistics

- **Course Website:**
  https://csbio.unc.edu/mcmillan/index.py?run=Courses.BCB716F22

  Or just follow the links from
  https://csbio.unc.edu/mcmillan

  [Courses | BCB 716 | Fall 2022] (and bookmark it)
- **Look there for announcements and zoom links.**
- **Include "BCB716" in subject line of all emails**
- **Course grading:**

○ 3 Problem Sets	-	60 %
○ Final	-	40 %

A Quick Exercise

https://forms.gle/T5NTS34FcXTFwtKV8

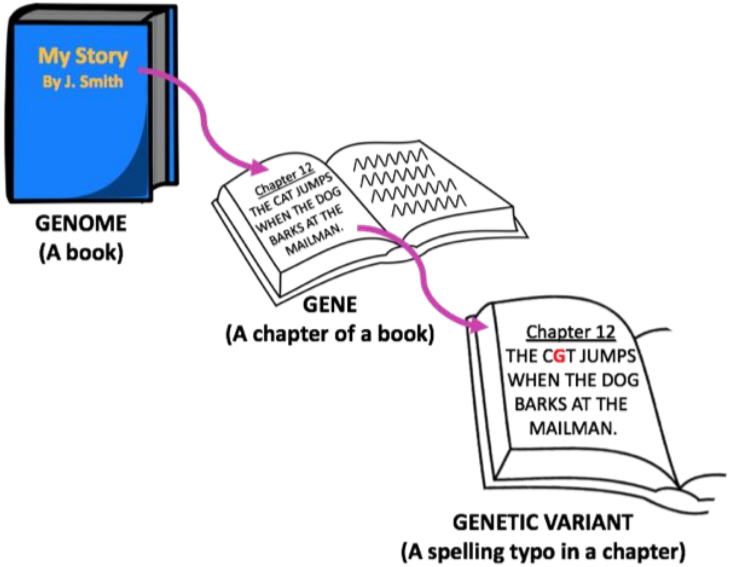
- You will need a UNC longleaf account

  https://its.unc.edu/research-computing/request-a-cluster-account/

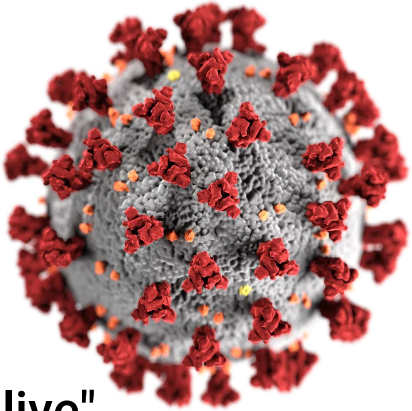
  Request on in my name if you must mcmillan@cs.unc.edu

What is a Genome?

- **An organism's complete set of genetic instructions**
- **Genomes are inherited**
- **Genomes evolve**
- **Genomes define and distinguish between organisms**
  - Within a species
  - Between species



An RNA Genome



Viral genomes are some of the smallest

- Sequence is encoded in a single molecular sequence that folds and bonds to itself for stability

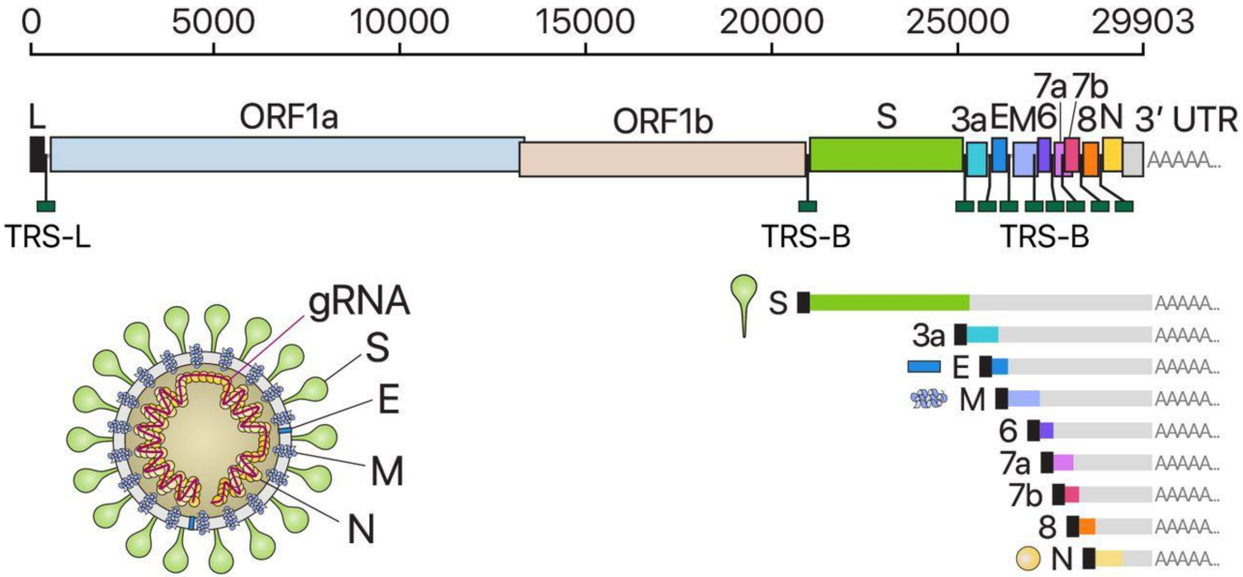
Characteristics of Viral genomes:

- Small, dense, and tricky
- Viral genomes code for functional proteins in order to "live", but rely on a host's machinery to perform essential functions
- Small genomes (3K - 30K bases) with a few "key" genes

SARS-CoV-2, the virus that causes COVID-19

- 29903 bases of the original Wuhan isolate
- 10 (11?) genes, 4 structural, 2 with primary functions

An RNA Genome

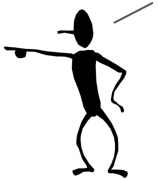


SARS-CoV-2, the virus that causes COVID-19

- 29903 bases of the original Wuhan isolate
- 10 (11?) genes, 4 structural, 2 with primary functions

COVID-19 FASTA file

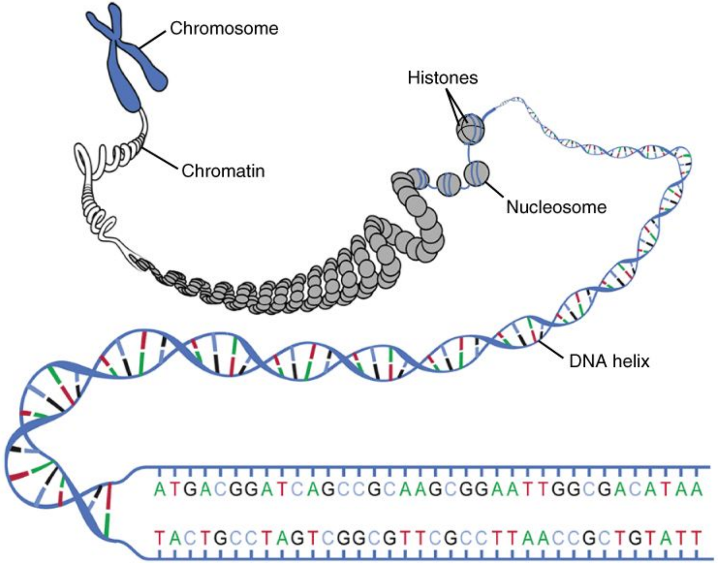
```
>NC_045512.2 |Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
CGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
           ;
     Skip 490 lines
           :
ACAGTGAACAATGCTAGGGAGAGCTGCCTATATGGAAGAGCCCTAATGTGTAAAATTAAT
TTTAGTAGTGCTATCCCCATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAA
```



This file is FASTA formatted. FASTA is the standard format used for genomes and sequences in general.. A FASTA file is composed of one or more sequences, each with a header that starts a '>' followed by lines of nucleotides that when concatenated form the sequence.

DNA Genomes

- Most living organisms have DNA based genomes
- DNA is a long polymer bonded to a second “complementary” sequence
- It therefore contains two sequences each composed of 4 nucleotides (Adenine, Cytosine, Guanine, Thymine)
- Each base binds with another specific base Thymine with Adenine and Cytosine with Guanine
- Nucleotide sequences on either strand can encode functions, sometimes in both directions.


Chromosome
Histones
Chromatin
Nucleosome
DNA helix
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

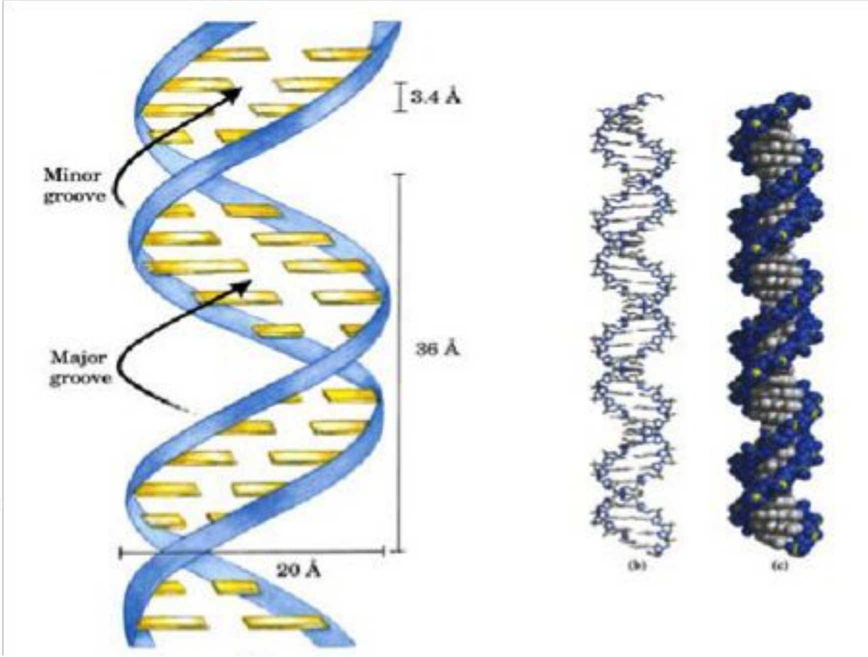
DNA Schematic

- **Many more details are required to give a complete picture of DNA**
  - Complementary strands are antiparallel and, thus, oriented
  - Not a simple twist, but DNA has a major and minor grooves which are recognized by interacting with proteins

- **Rather than keep track of all the details we will often consider DNA as a string of nucleotides**

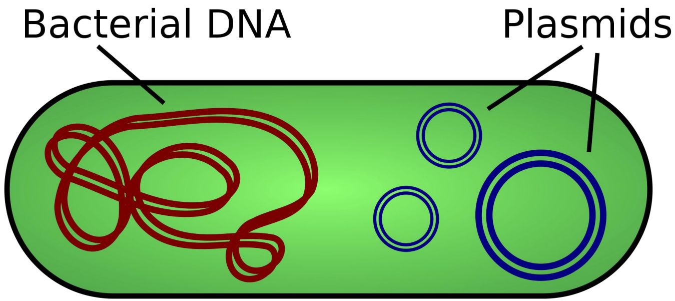
  `5'...ACGGATAGCATGGA...3'`

- **By convention DNA sequences are always ordered in the 5'-to-3' direction. Not coincidentally, this is also the order in which they are synthesized using an important class of molecules call polymerases.**



E. Coli Genome

Escherichia Coli, or E. Coli for short, is a common model organism and an example of a simple prokaryotic bacterial genome composed of 5.3 Mbp.



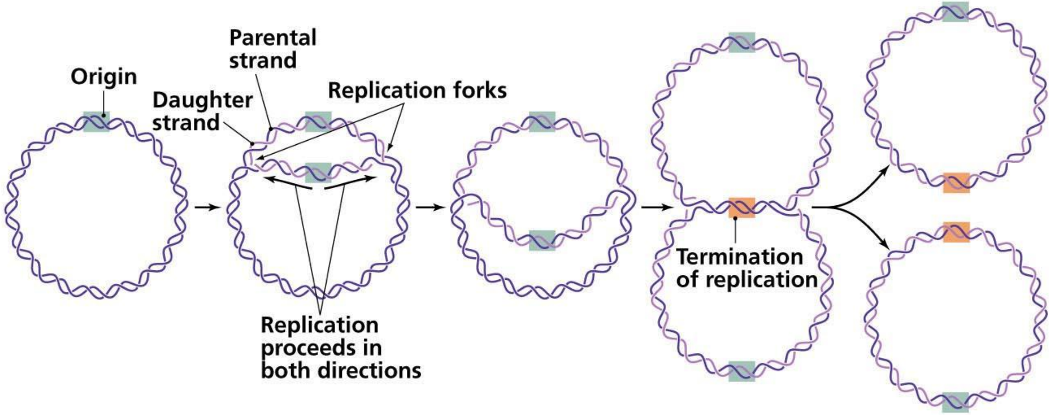
Characteristics of Bacterial DNA

- **A “circular” primary chromosome (a few million bases) with essential genes**
- **Smaller chromosomes or circular plasmids (10-100K bases) with a few additional genes**
- **There can be multiple plasmid sequences with variable numbers of copies**

DNA what's the point

Life ≡ Reproduction ≡ Replicating a Genome

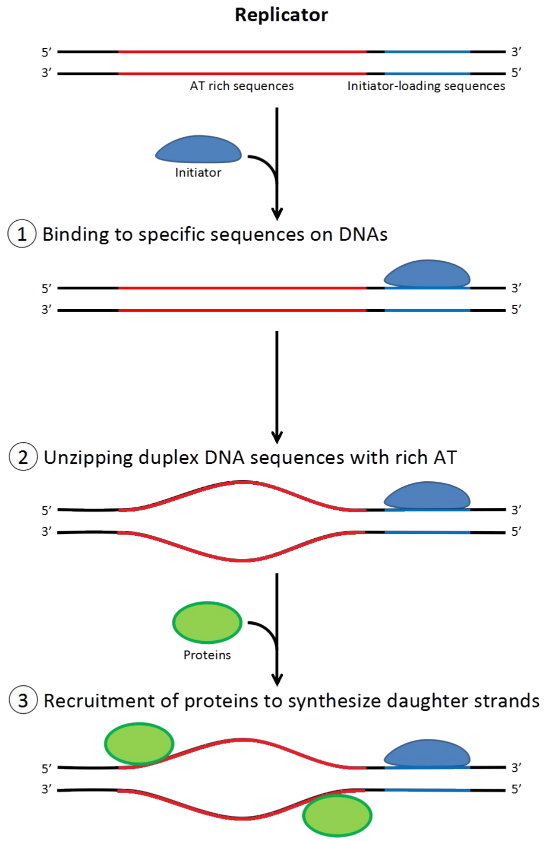
One of the most incredible things about DNA is that it provides instructions for replicating itself.


Origin
Parental strand
Daughter strand
Replication forks
Replication proceeds in both directions
Termination of replication

Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.

Biological Insights

- Replication is performed by a DNA polymerase, and the initiation of replication is mediated by a protein called *DnaA*.
- DnaA binds to short (≈ 9 nucleotides long) segments within the replication origin known as a *DnaA* box (≈ 500 bases).
- A DnaA box is a signal telling *DnaA* to “bind here!”
- DnaA can bind to either strand. Thus, both the *DnaA* box and its reverse-complement are equal targets.
- By convention the first base in a bacterial DNA sequence begins with the *DnaA* gene

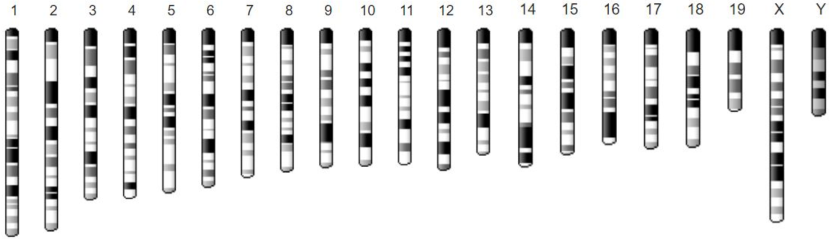

Replicator
5′
3′
3′
5′
AT rich sequences
Initiator-loading sequences
Initiator
1 Binding to specific sequences on DNAs
2 Unzipping duplex DNA sequences with rich AT
Proteins
3 Recruitment of proteins to synthesize daughter strands

E. Coli Genome Sequence

```
>CP003289.1 Escherichia coli O104:H4 str. 2011C-3493, complete genome
CATTATCGACTTTTGTTCGAGTGGAGTCCGCCGTGTCACTTTCGCTTTGGCAGCAGTGTCTTGCCCGATTGCAGGATGAG
TTACCAGCCACAGAATTCAGTATGTGGATACGCCCATTGCAGGCGGAACTGAGCGATAACACGCTGGCCCTGTACGCGCC
AAACCGTTTTGTCCTCGATTGGGTACGGGACAAGTACCTTAATAATATCAATGGACTGCTAACCAGTTTCTGCGGAGCGG
ATGCCCCACAGCTGCGTTTTGAAGTCGGCACCAAACCGGTGACGCAAACGCCACAAGCGGCAGTGACGAGCAACGTCGCG
                      Skipped 65910 lines

>CP003291.1 Escherichia coli O104:H4 str. 2011C-3493 plasmid pAA-EA11, complete sequence
GCCTCGCAAAACATTGCTCTATTCATGCACCACTCTATGTTCTCTATTTTCTTACAGAATAAAGAGGCCAGTTTCCAGAG
CTCAAAAGCAAGAGCAAGATCCATATTCTTCTCTGCAGAAGCTGCTTGTTTCGACAAGGAATGAAAATTCTTTCCAGCCT
TCATCAATACCTCCCAATACAATGTTCTCTTTTCTGAAATTTATAACATGATAAGATATCGGAAGGATAGAATAATGACA
CAATGGAGATATTCAAATAGGGGCCAGAACGCTGCTGCACCAGAAAAACCCGGAATTAAGAGATTATGGAAAAGGACTTA
                      Skipped 924 lines

>CP003290.1 Escherichia coli O104:H4 str. 2011C-3493 plasmid pESBL-EA11, complete sequence
GTTGGGATGACGCCAGACCAACCTCAAATGCGGAACCGTCTTCTGTATGTAACATCAGATTCCCTTTGCCTCGCCAGACA
GAGCCTGCCCTGTTAAGCGGGAAAAGCGAGATGGCATGTGACAACGGGGGATACAGGGGAACCGGGCCGGAGCCTCCGGA
TGCCGCGAGAATGGTATTCACCCAGGCGCGTCGGGGATCGCCAAATGTTGTCGTGGTGCCACATACGCCCCAGCCTTCAA
TGGCTGATTTCAGAATGGCCTGGTTTCGTGCGCATATTTCACGTGTTTTTCCCCACGTGGAGGCCATGATGGTCATTATG
                      Skipped 1103 lines

>CP003292.1 Escherichia coli O104:H4 str. 2011C-3493 plasmid pG-EA11, complete sequence
CTAGCTGAAAAACTTGGAGTTAGCAGAAGCACAATTATTCGGTGGCTCAATTACTTAGAATCAAAAAATGCATTAGTTAG
AATCCCCGTTGCTGGTAAGGTTTGTGCGTATGCCCTCGATCCACATGAAGTCTGGAAGGGATACAACACTACGAAAAACC
ATGCAGCGTTTGTCACTAAAACACTGGTCAACAAAGACGGTGATATTCAGCGCCGAATCATGGCCATGTTTTCAAATTGA
GCTAGCGGCAGGCGGACAATCAGGGGCTACGTGTTAACGTTCTGACCATGATTGTCTATCCTGCATTGCTCTTTTGCCGC
                      Skipped 17 lines
```

Laboratory mouse genome

- A mammalian genome with 21 nuclear chromosomes
- 2.728 Billion base pairs
- Chromosome 1 is the largest with 195 Mbp
- Chromosome 19 is the smallest with 61 Mbp
- Genome was sequenced concurrently with the human Genome and the 1st draft was finished about the same time
- It is a genome of a single common "isogenic" mouse strain C57BL/6J
- It is commonly used as a *reference* genome for all other mouse strains and wild mice.
- A "sample-based" reference



What is a reference genome?

- **A representative sequence that models the gene organization of a species**
- **FASTA files downloadable from the Genome Reference Consortium**
- **It is the basis for calling genomic variants**
- **Sequence differences that distinguish**
  - Individuals
  - Subspecies
  - Used as a "scaffold" for modeling other species
- **Types of variants**
  - Single base changes or Single Nucleotide Polymorphisms (SNPs)
  - Short inserts or deletions (INDELS)
  - Variable length simple repeat patterns (microsatellites)
  - Structural variations (Larger insertions, deletions, Inversions, translocations)

The Genome Reference Consortium

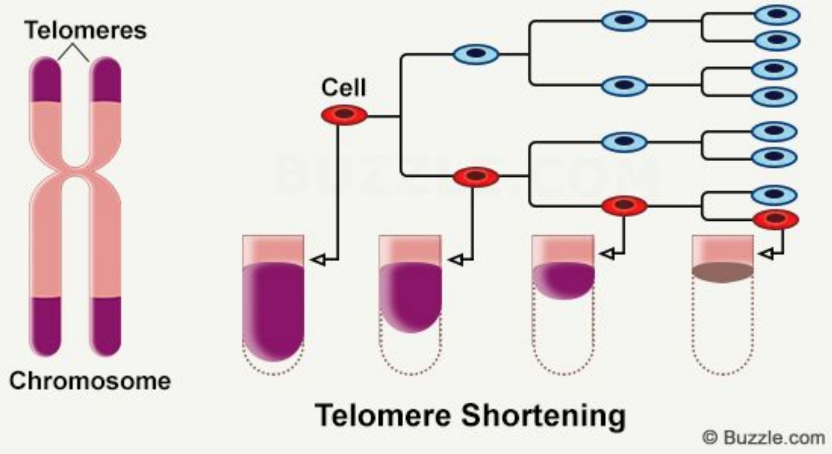
Putting sequences into a chromosome context

The original model for representing the genome assemblies was to use a single, preferred tiling path to produce a single consensus representation of the genome. Subsequent analysis has shown that for most mammalian genomes a single tiling path is insufficient to represent a genome in regions with complex allelic diversity. The GRC is now working to create assemblies that better represent this diversity and provide more robust substrates for genome analysis.

Mouse (GRCm38_68) genome

```
>1 dna:chromosome chromosome:GRCm38:1:1:195471971:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
--
>10 dna:chromosome chromosome:GRCm38:10:1:130694993:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
--
>11 dna:chromosome chromosome:GRCm38:11:1:122082543:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
--
>12 dna:chromosome chromosome:GRCm38:12:1:120129022:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

    ...

>9 dna:chromosome chromosome:GRCm38:9:1:124595110:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
--
>MT dna:chromosome chromosome:GRCm38:MT:1:16299:1 REF
GTTAATGTAGCTTAATAACAAAGCAAAGCACTGAAAATGCTTAGATGGATAATTGTATCC
CATAAACACAAAGGTTTGGTCCTGGCCTTATAATTAATTAGAGGTAAAATTACACATGCA
--
>X dna:chromosome chromosome:GRCm38:X:1:171031299:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
--
>Y dna:chromosome chromosome:GRCm38:Y:1:91744698:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

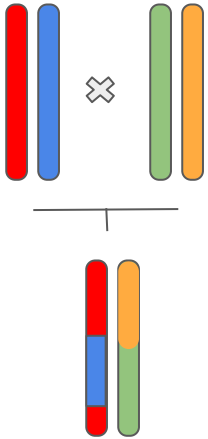
- Most chromosomes/contigs start with 3M 'N's
- These are the "telomeres", a repetitive sequence of the "TTAGGG" hexamer whose length is unknown
- Changes after each mitotic subdivision
- Some chromosomes have gaps with Ns where the assembly is still incomplete

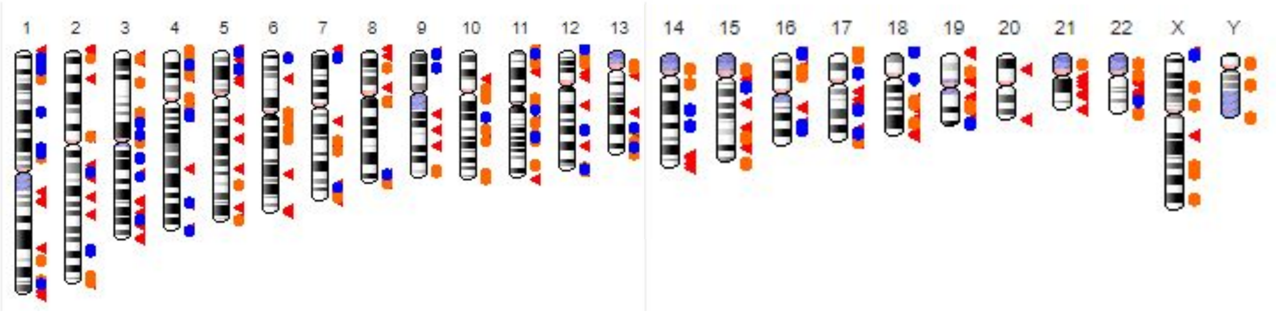


© Buzzle.com

A Human Genome

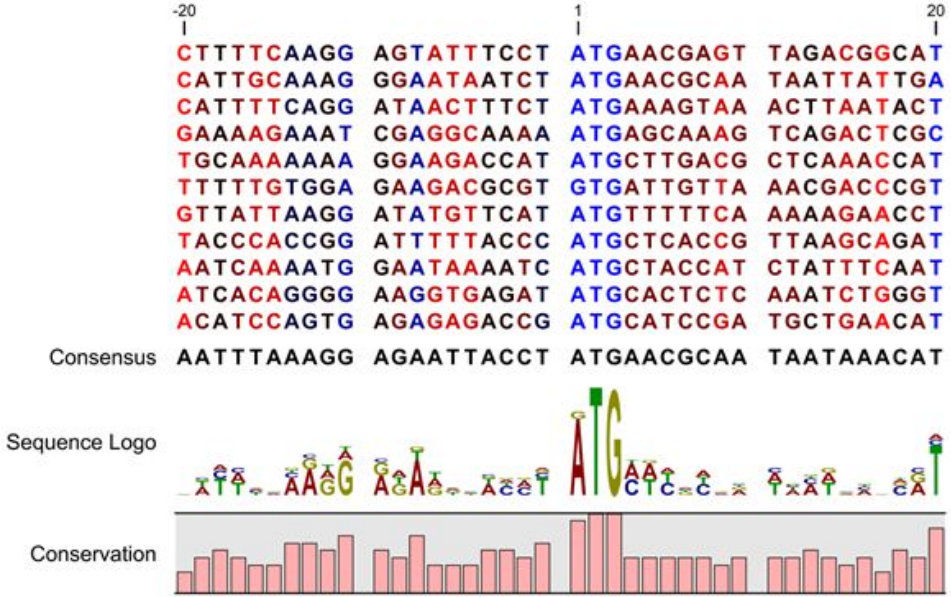
- 24 chromosomes with 3.27 Bbp
- Chr 1 (248 Mbp) is the longest Chr 21 (46 Mbp) is the smallest
- Composed of DNA from 13 volunteers, and 26 haploid genomes
- Humans are "outbred" with two "diploid" chromosome pairs
- C57BL/6J is diploid too, but both chromosomes are identical
- It is "consensus-based" rather than "sample-based"
- Thus it is a genome that no one actually has it.





A Consensus Genome

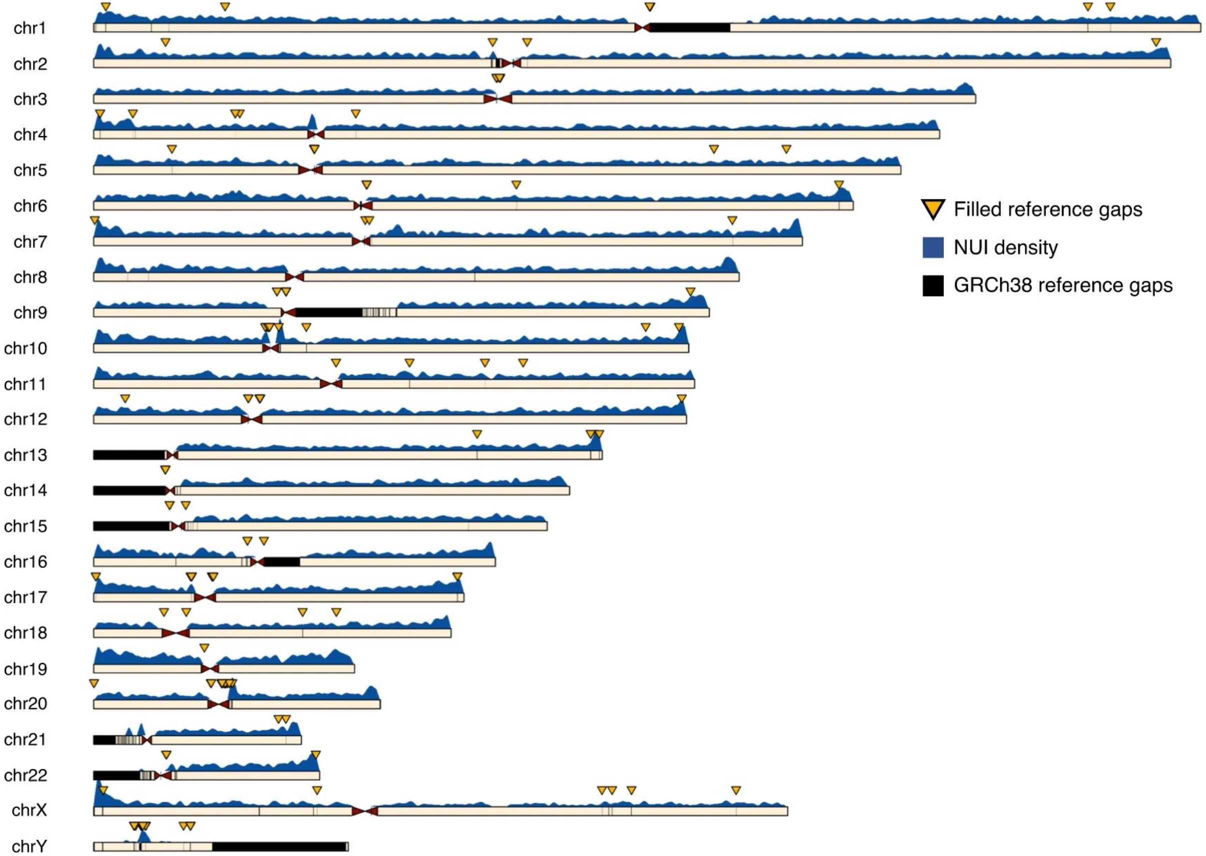
- The reference base is the most common at each position
- It is population dependent (did the 13 contributors represent all of humanity)
- What if there are structural differences between contributors?
- Copy number differences?
- Technical issues resulted in most of the assembly (80%) comes from only 8 people. One male accounts for 66%.



```
CTTTTCAAGG AGTATTTCCT ATGAACGAGT TAGACGGCAT
CATTGCAAAG GGAATAATCT ATGAACGCAA TAATTATTGA
CATTTTCAGG ATAACTTTCT ATGAAAGTAA ACTTAATACT
GAAAAGAAAT CGAGGCAAAA ATGAGCAAAG TCAGACTCGC
TGCAAAAAAA GGAAGACCAT ATGCTTGACG CTCAAACCAT
TTTTTGTGGA GAAGACGCGT GTGATTGTTA AACGACCCGT
GTTATTAAGG ATATGTTCAT ATGTTTTTCA AAAAGAACCT
TACCCACCGG ATTTTTACCC ATGCTCACCG TTAAGCAGAT
AATCAAAATG GAATAAAATC ATGCTACCAT CTATTTCAAT
ATCACAGGGG AAGGTGAGAT ATGCACTCTC AAATCTGGGT
ACATCCAGTG AGAGAGACCG ATGCATCCGA TGCTGAACAT
Consensus AATTTAAAGG AGAATTACCT ATGAACGCAA TAATAAACAT
```

Human Genome Diversity

- Efforts are underway to create a better Human reference genome
- Combining > 300 sequenced and assembled and truly diverse samples uncovers significant non-reference unique insertions (NUIs)
- Sequences someone has that aren't in the humane reference
- What are these?
  - Genes
  - Structural variants



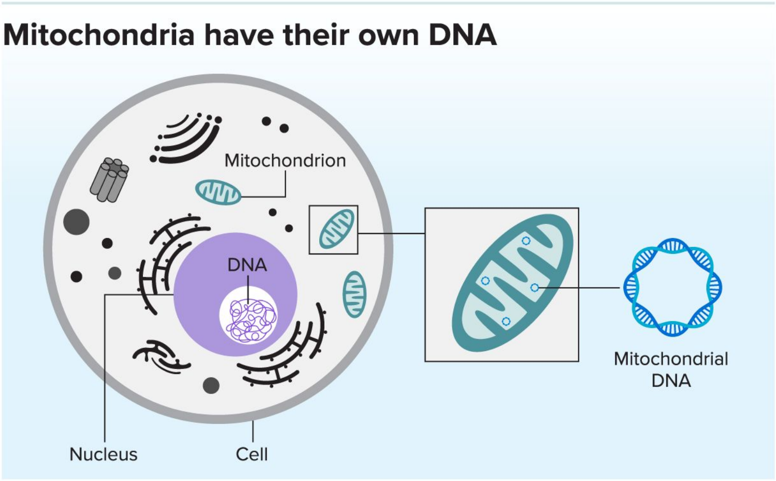
▽ Filled reference gaps

▨ NUI density

■ GRCh38 reference gaps

Mitochondrial Genomes

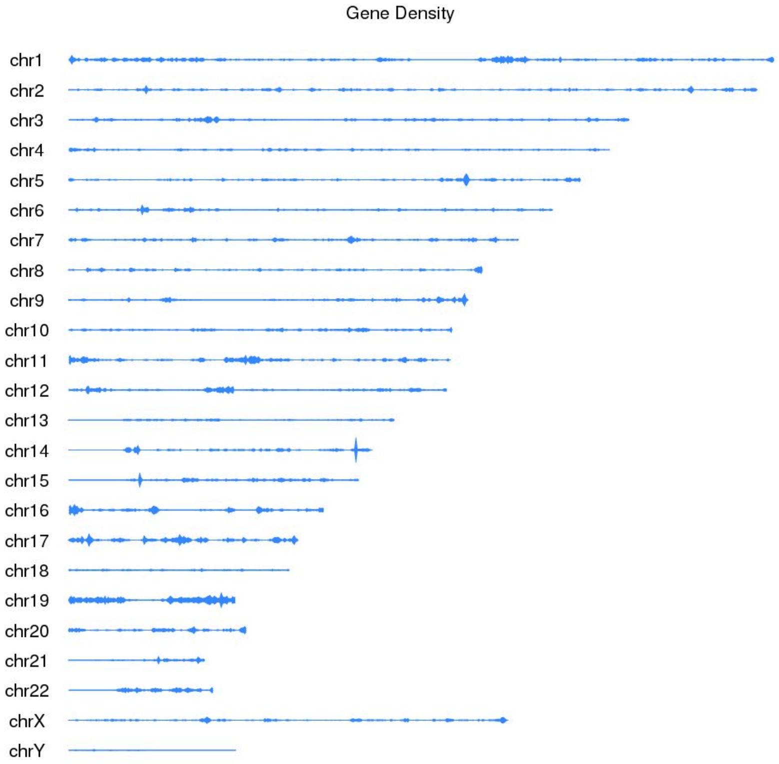
- **Mitochondria organelles have their own DNA sequence**
- **Circular like bacterial genomes**
- **Inherited strictly from maternal genome**
- **Basis for the "Eve" model of out-of-africa**
- **Small, but with many copies**
    - 16,295 in mouse with 37 genes
    - 16,569 in human with 37 genes



KNOWABLE MAGAZINE

Genes in Genomes

- A typical gene is roughly 27,000 bases
- The largest known human gene (dystrophin) has 2.4 million bases
- Estimated number of human genes is roughly 20,000-25,000
- The genome is nearly identical for every human (99.9%)
- Human DNA is 98% identical to chimpanzee DNA.
- The functions are unknown for more than 50% of discovered genes.
- Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between.

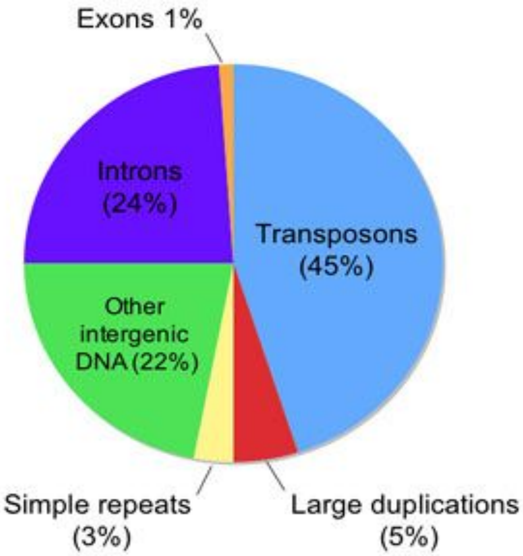


Other Stuff in Genomes

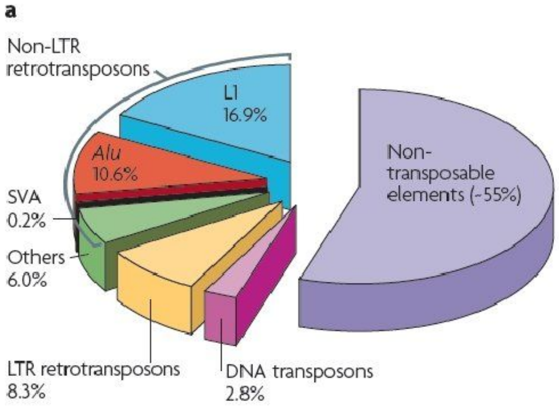
There are huge parts of the genome that are hard to sequence and assemble.

They have a history of being mobile, and could segregate in populations

They have genes



the transposable element content of the human genome.



Is Bigger Better? More Advanced?

- The genome size of a species is relatively constant
- Large variations can occur across species lines
- Not strictly correlated with organism complexity
- Genome lengths can vary as much as 100 fold between similar species
- Length and variability are more of an indications of a phylum's susceptibility to mutation

670 Bbp

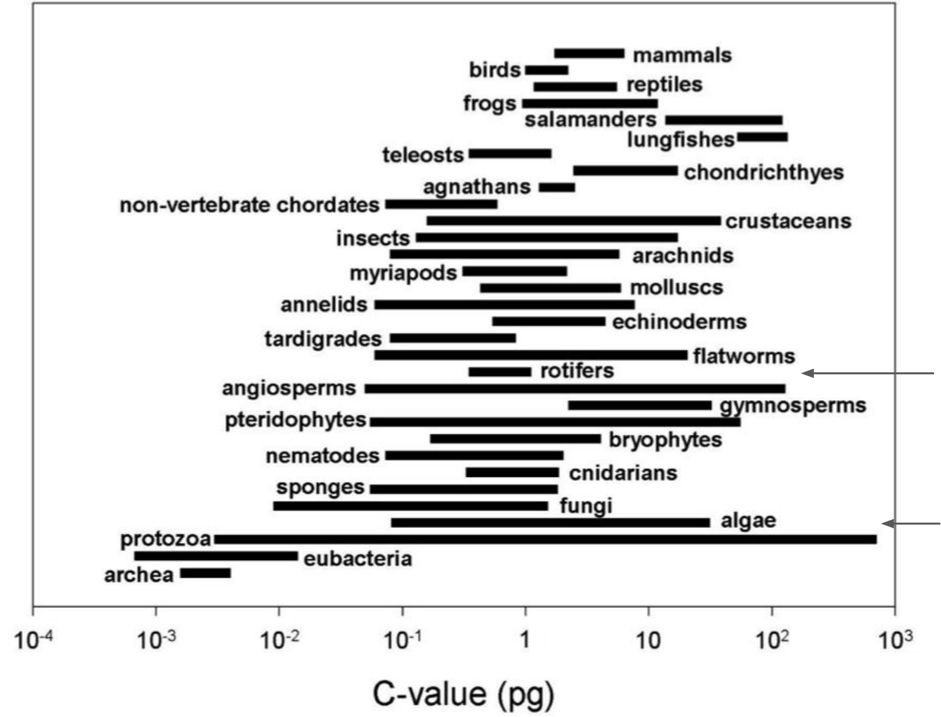
118 Bbp

13 Bbp

168 x

Amoeba (Amoeba dubia) ~ 670 Bbp
Salamander (120.60pg, Necturus lewisi, Gulf coast waterdog) ~118 Bbase pairs
Frog (13.40pg, Ceratophrys ornata (8n), Ornate horned frog) ~13 Bbase pairs
Marbled Lungfish (130pg) ~ 130 Bbp
Orchids (angiosperms) have the the largest variation within a species
(strains that can interbreed and generate fertile progeny) with a range that varies at least 168-fold.

Genome Variation



Orchids

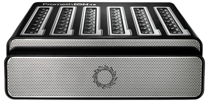
Amoeba

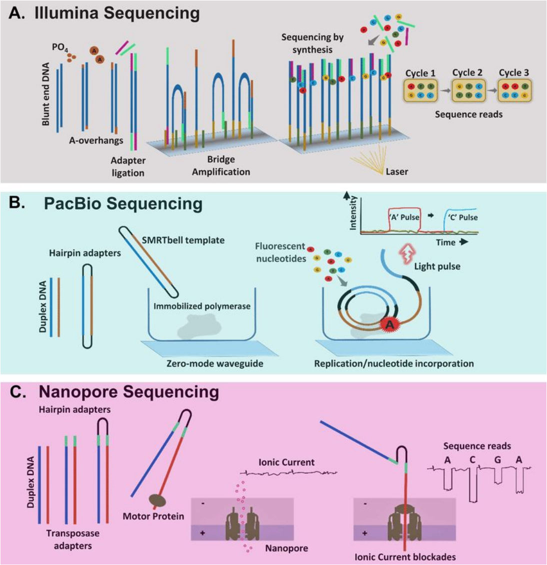
Length and variability are more of an indications of a phylum’s susceptibility to mutations than its complexity

(C-value = the Amount of DNA in an unreplicated gametic nucleus. It is measures in pico Grams, and 1pg = 978M base pairs.)

Next Time

The technology of sequencing




A. Illumina Sequencing
Blunt end DNA
PO_4
A-overhangs
Adapter ligation
Bridge Amplification
Sequencing by synthesis
Laser
Cycle 1
Cycle 2
Cycle 3
Sequence reads
B. PacBio Sequencing
Duplex DNA
Hairpin adapters
SMRTbell template
Immobilized polymerase
Zero-mode waveguide
Fluorescent nucleotides
Intensity
'A' Pulse
'C' Pulse
Time
Light pulse
A
Replication/nucleotide incorporation
C. Nanopore Sequencing
Hairpin adapters
Duplex DNA
Transposase adapters
Motor Protein
Ionic Current
Nanopore
Sequence reads
A C G A
Ionic Current blockades