SevenBridges

Applied Bioinformatics

sevenbridges.com

Applied Bioinformatics

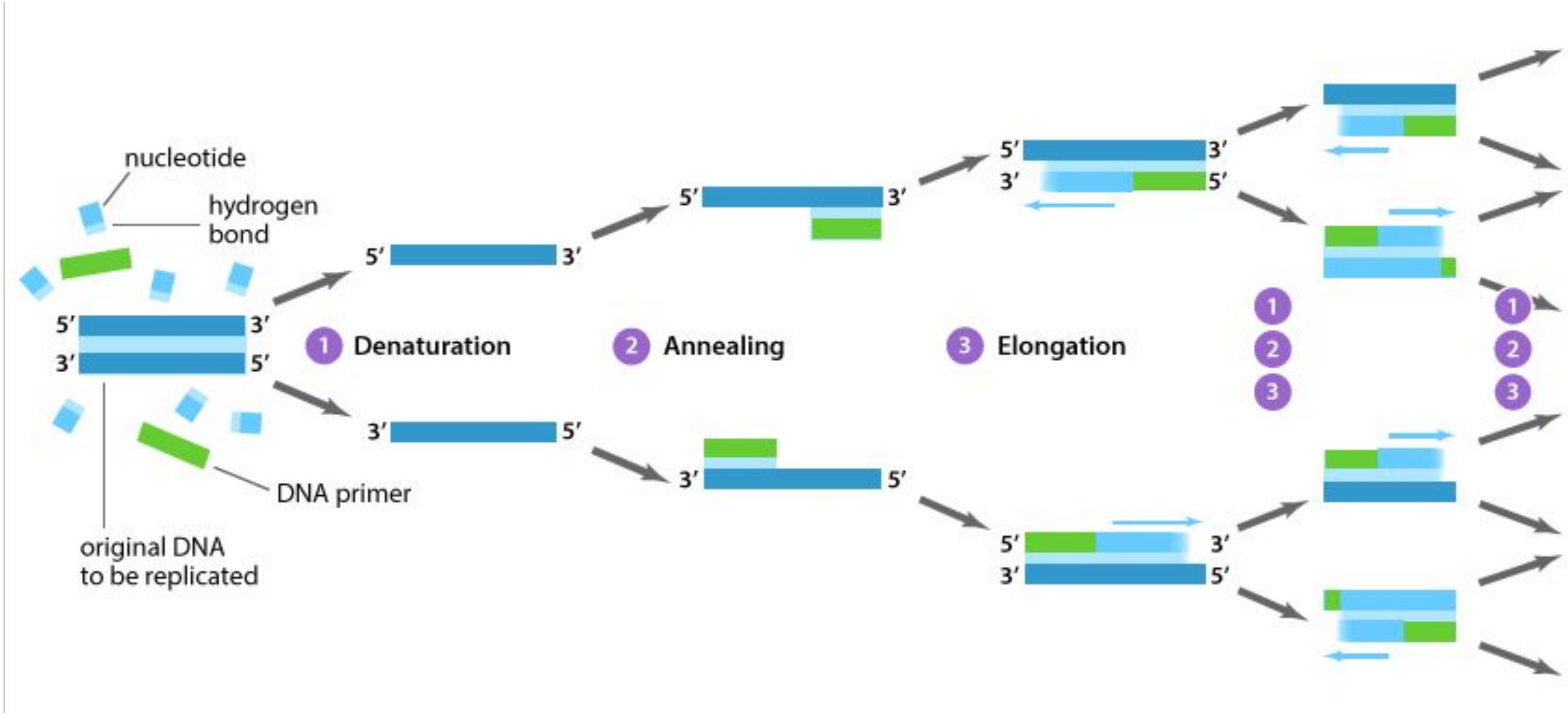
The Reference Genome

Mladen Lazarevic
Milica Kojicic
Milan Kovacevic
milan.kovacevic@sbgenomics.com

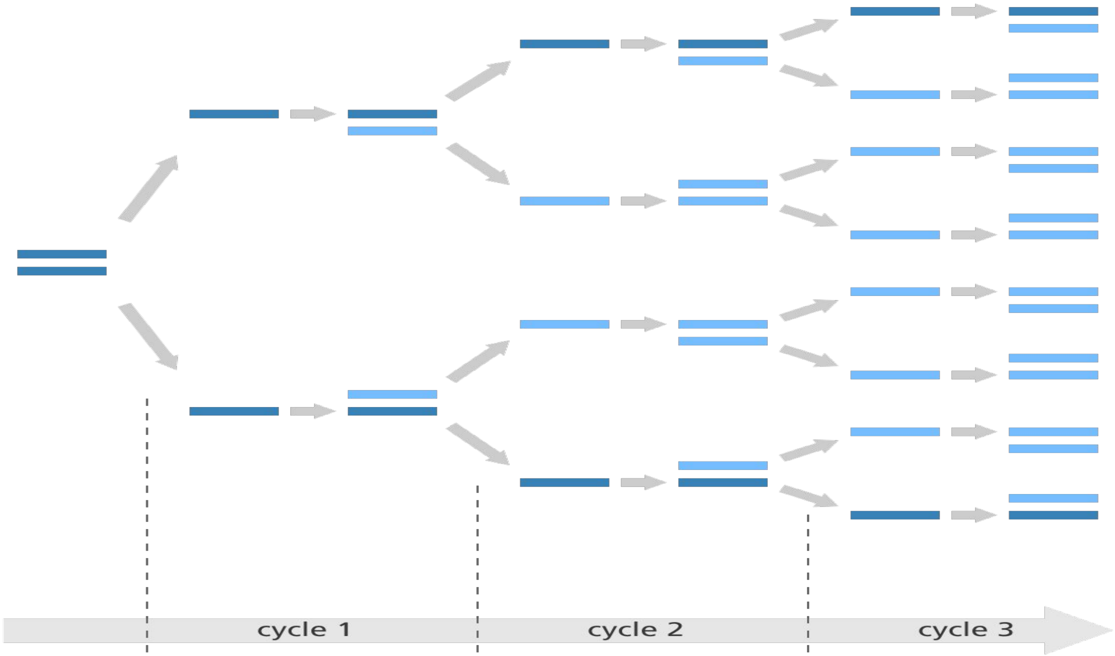
Agenda

- PCR and Sequencing - recapitulation
- Human Genome Project
- File formats related to the reference genome
- Example tasks

PCR - Polymerase chain reaction



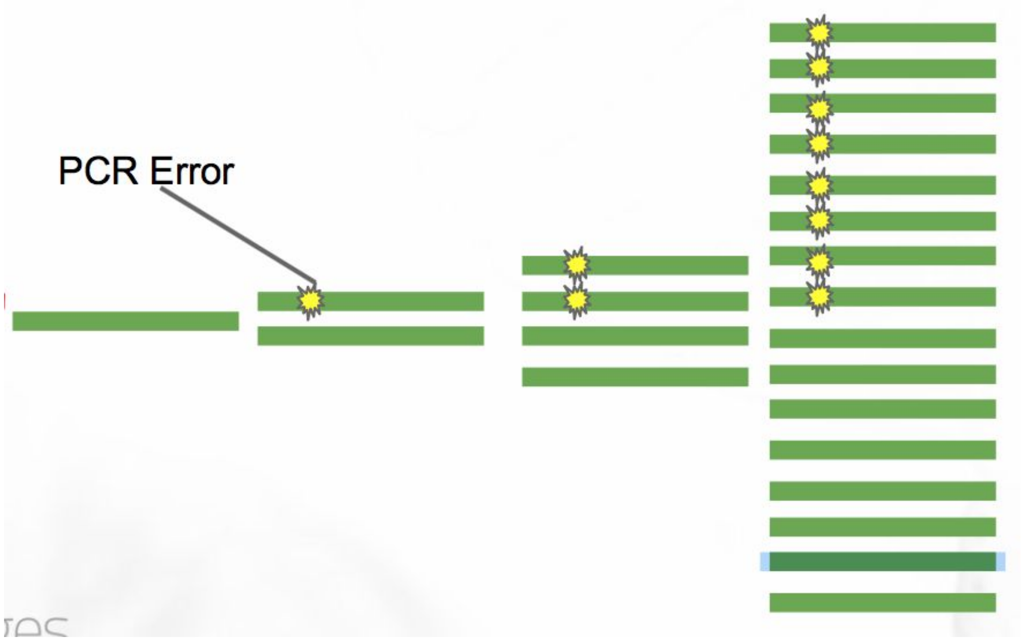
PCR - Polymerase chain reaction



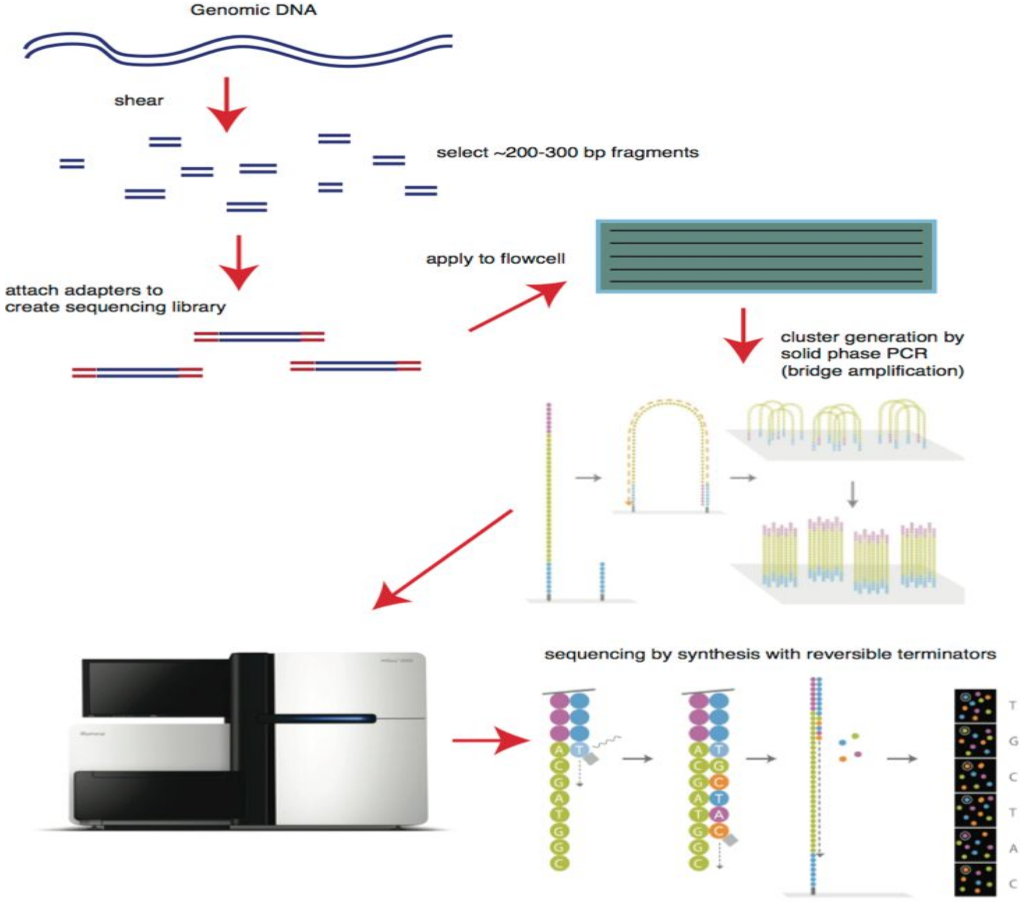
PCR - Polymerase chain reaction

PCR - Error

- 1 in 10k error rate



Genome sequencing - a reminder



Genomic DNA

shear

select ~200-300 bp fragments

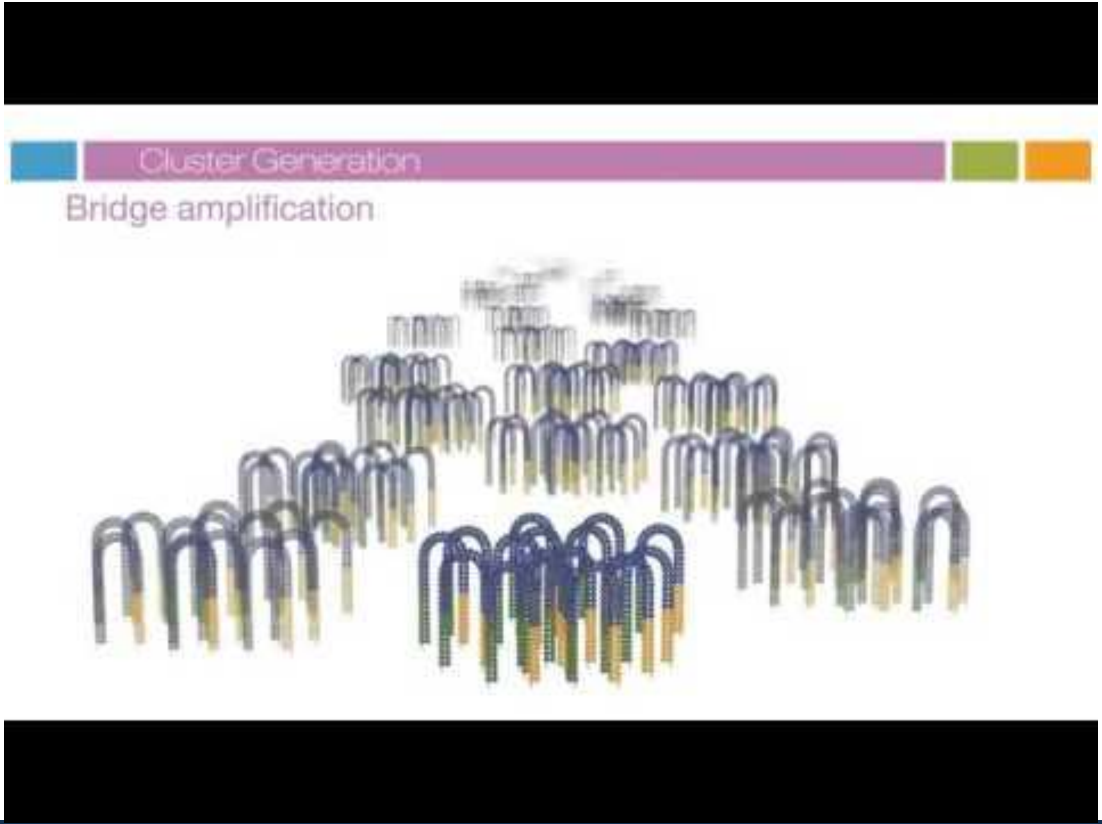
attach adapters to create sequencing library

apply to flowcell

cluster generation by solid phase PCR (bridge amplification)

sequencing by synthesis with reversible terminators

Genome sequencing - recapitulation



Genome sequencing - recapitulation

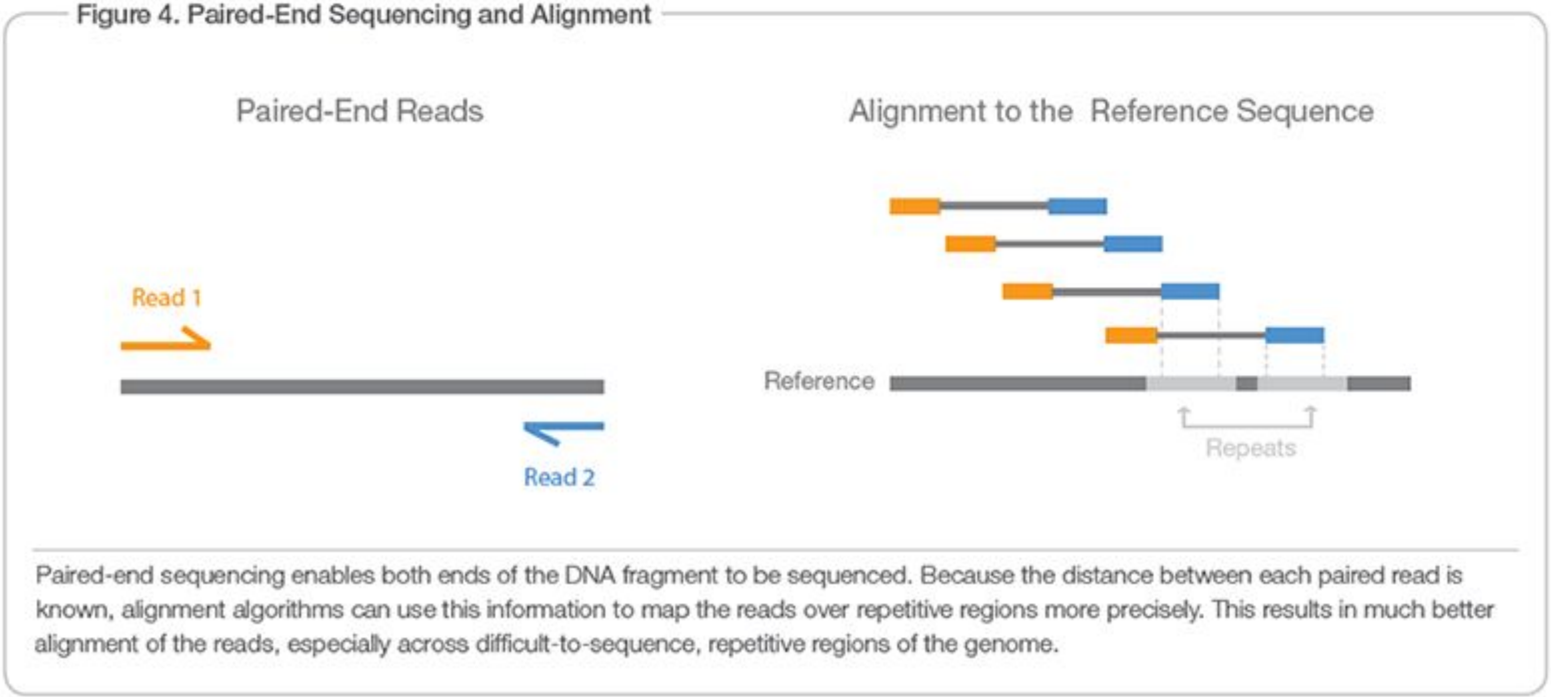


Figure 4. Paired-End Sequencing and Alignment

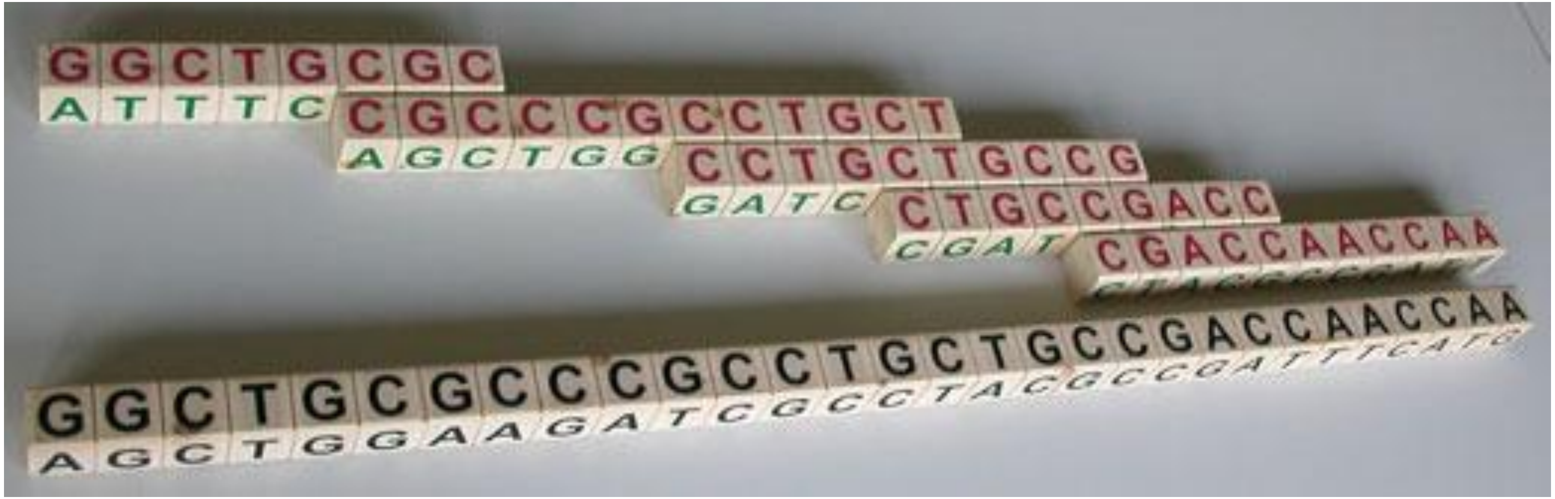
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

DNA Sequencing - Reminder

- We got a FASTQ file with the “reads” - little pieces of the genome

What to do with the sequencing reads?

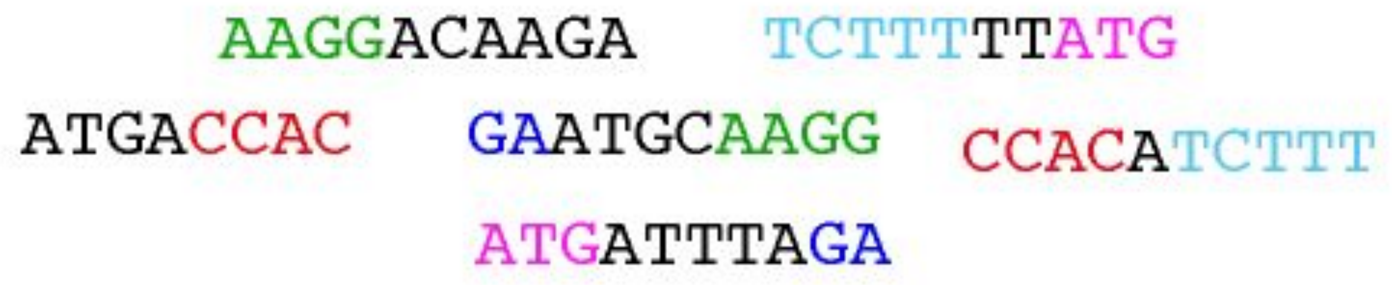
- How do we reconstruct the genome that went into the “shredder”?
- We could try “assembly” - connecting the reads into longer sequences



Genome Assembly

- Greedy algorithm (suboptimal solution):
  1. Calculate pairwise alignments of all fragments.
  2. Choose two fragments with the largest overlap.
  3. Merge chosen fragments.
  4. Repeat step 2 and 3 until only one fragment is left.
- Even the more practical solutions have problems:
  - High computational cost
  - High memory consumption (100s of GB or RAM)

Genome Assembly



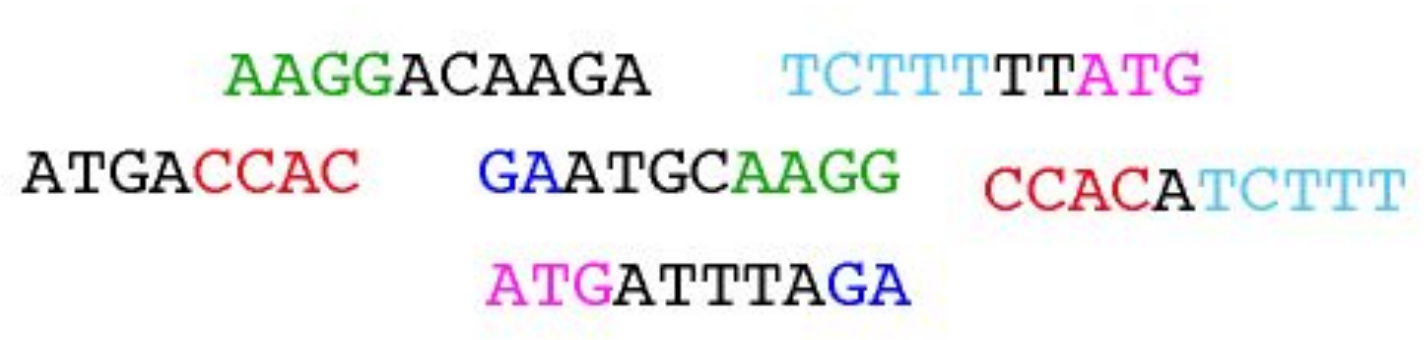
What to do with reads? (an alternative)

- We do have an alternative approach to recover that genome (the genome that went into the "shredder")
- This way is faster and more practical than assembly
- It is based around a *Reference genome, Alignment,* and *Variant Calling*
- But first we will need to define these terms and procedures

The reference genome

- A reference genome is representative example of a species' genome
- It is often built from genomes of multiple individuals
- Every individual differs from this genome in some places
- We can think of *genomic variation* as differences from the reference (genome)
- Reference genomes provide a coordinate system for communicating genomic data
- And, they make analyses easier!

What do with the sequencing reads?
(an alternative)



AAGGACAAGA

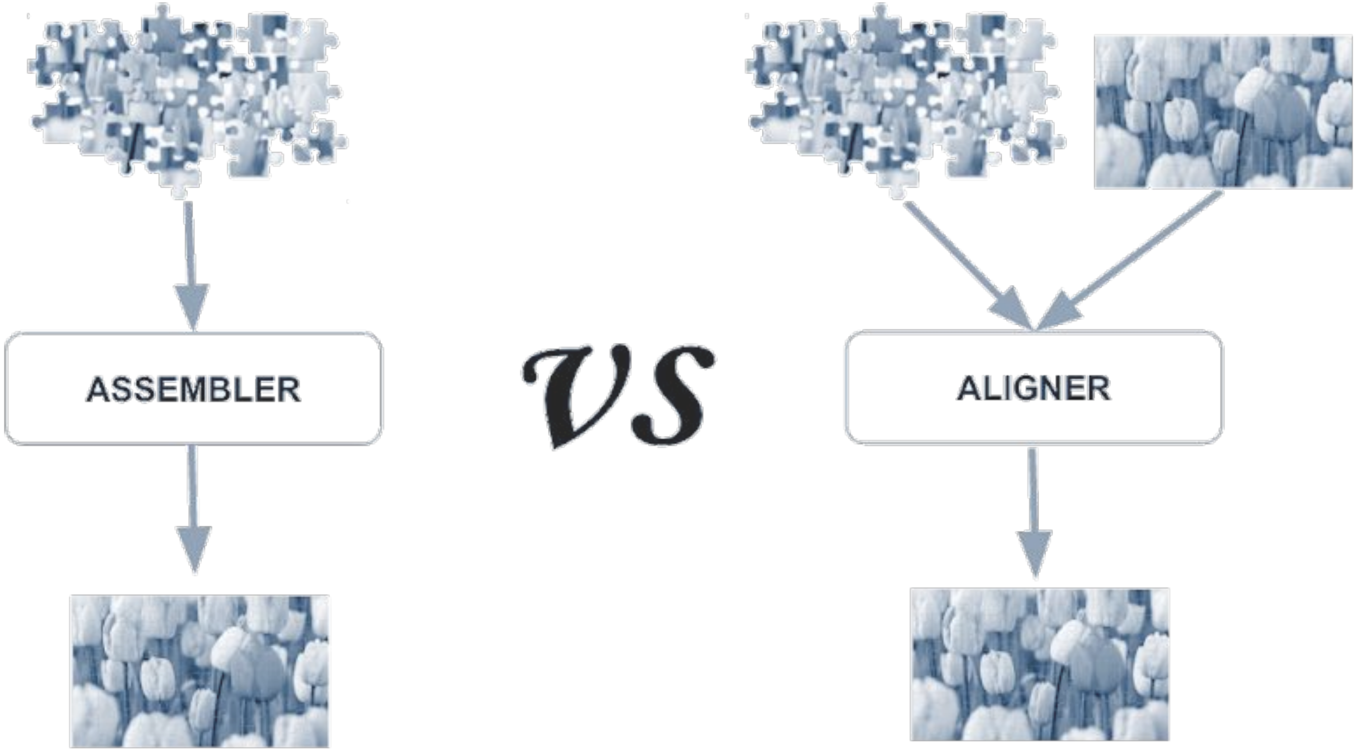
TCTTTTTTATG

ATGACCAC

GAATGCAAGG

CCACATCTTT

ATGATTTAGA

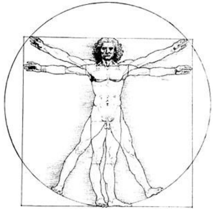


ASSEMBLER

VS

ALIGNER

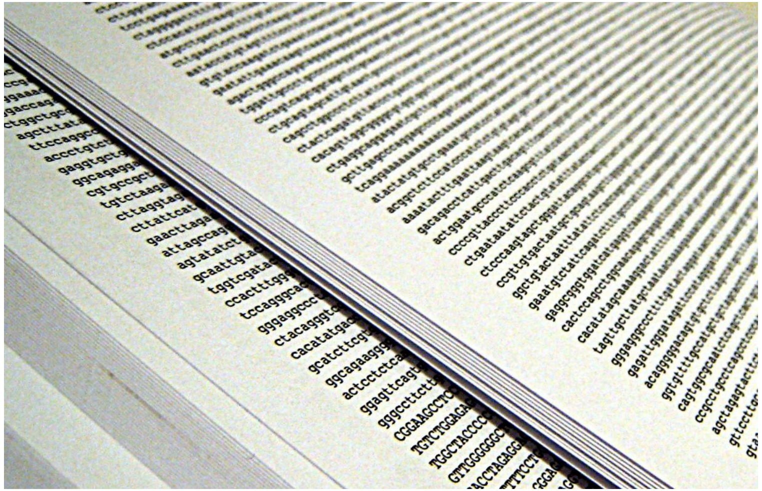
Human Genome Project

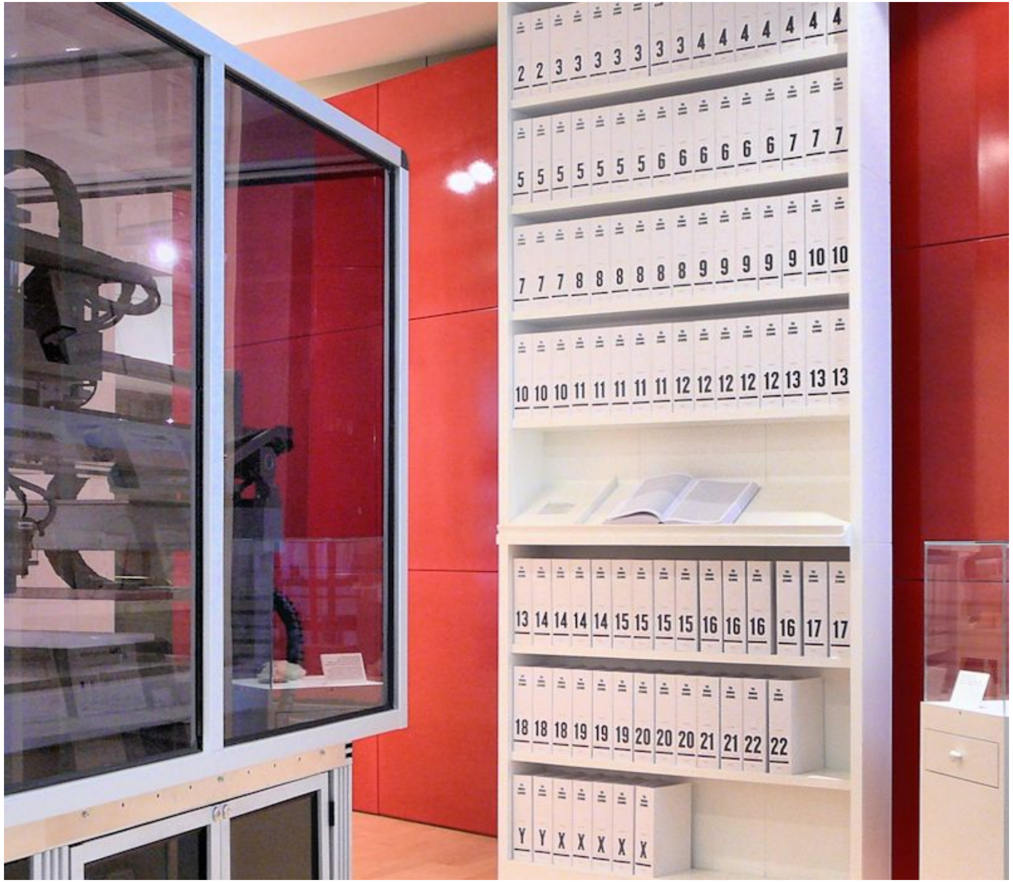


- International scientific initiative
  to create a reference human genome
- Active in years 1990 to 2003, at a cost over $3B
- 70% of the reference came from a single male donor
- In parallel, Celera Corporation launched a privately funded project, with the same goal, but had the intention to patent the sequence
- This caused a race, leading up the release of the first draft
  of the public version of the human genome on July 7, 2000,
  by the UCSC Genome Bioinformatics Group

Human Genome Project

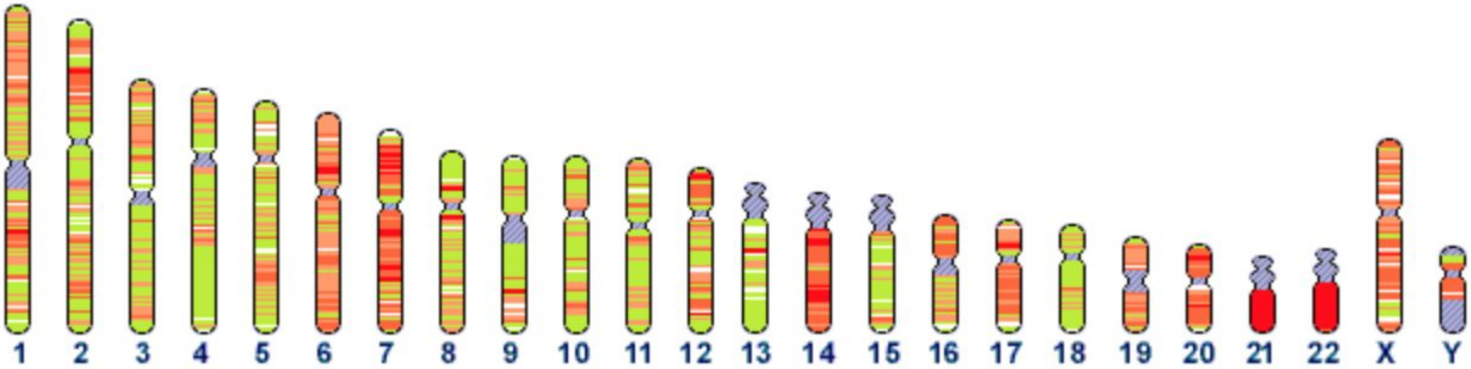
- 130 Books
- Font size: 4
- 43k letters per page





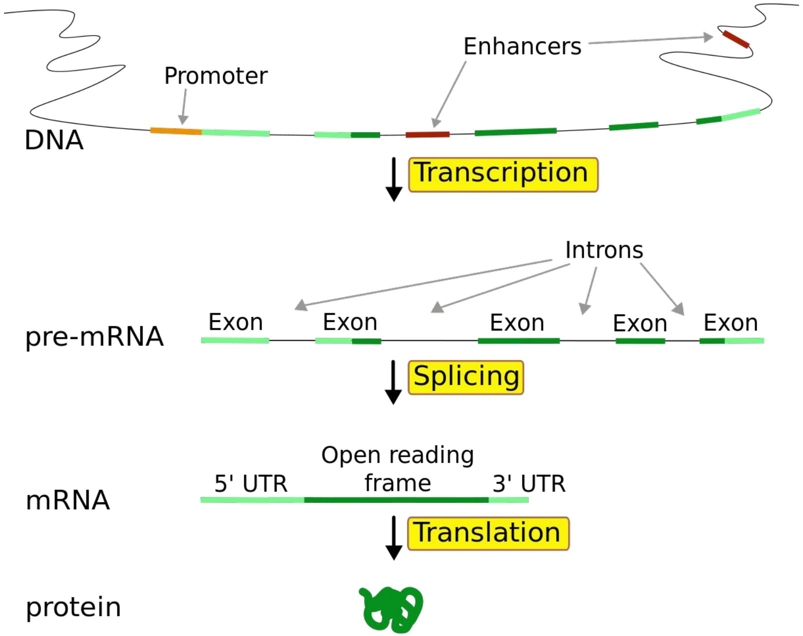
Human genome stats

- The human genome has 23 pairs of chromosomes:
  - Males have 22 autosomes, one X and one Y chromosome
  - Females have 22 autosomes and two X chromosomes
- There is also a separate Mitochondrial DNA contig
- Total length is ~3 Gigabases (3B basepairs)
- About 20 000 genes covering about 30 Megabases
  - 2% is the coding region



Human exome stats

- Exome - part of the human genome coding for proteins/mRNA
- Covers only 2% of genome - much cheaper to sequence



Human genome HG38

- Current version of the chromosomes is *HG38* (though 37 is still widely used)
- Released December 24th 2013
- Additional sequences are added to the genome:
  - Unplaced sequences (some genomes contain them, somewhere)
  - Unlocalized sequences (chromosome known, but coordinates are not)
  - Alternate sequences (some genomes contain them, instead of some parts)
  - Human Herpesvirus 4 type 1
- Patches are commonly added
- http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/

FASTA file format

- A simple format for storing reference files
- Two “types” of lines:
  - Description lines, start with ‘>’, contain sequence name and optional description
  - Sequence lines follow a description line and contain the actual nucleotide sequence
- Sequences are represented by a single description line, followed by one or more sequence lines (usually 70 bases or less in a line)

```
>chr 1
ACTGCCCTCCGTACCTACTGCCCTCCGTACCTACTGCCCTCCGTACCTACTGCCCTCCGTACCT
ACTGCCCTCCGTACCTACTGCCCTCCGTACCTACTGCCCTCCGTACCTACTGCCCTCCGTACCT
ACTGCCCTCCGTACCTACTGCCCTCCGTACCTACTGCCCTCCGTACCTACTGCCCTCCGTACCT
ACTGCCCTCCGTACCTACTGCCCTCCGTACCTACTGCCCTCCG
```

FASTA File IUPAC Codes (and contig names)

- A, C, T, G, U - Nucleotides (Adenine, Cytosine, Guanine, Thymine, Uracil)
- Ambiguous bases:
  - R - A or G
  - Y - C, T, or U
  - N - Any base
  - https://en.wikipedia.org/wiki/FASTA_format

- Often two versions of the Human Reference Genome are found
  - Chromosomes labeled 1, 2, 3... (Human Genome Consortium style)
  - Chromosomes labeled chr1, chr2, chr3... (UCSC style)

FASTA index (fai)

- Some tools build or require indices of the fasta file
  - Needs to be in the same folder, and have the same name as the fasta + .fai extension
- Fai file structure:
  1. The name of the sequence
  2. The length of the sequence
  3. The byte offset of the first base in the file
  4. The number of bases in each fasta line
  5. The number of bytes in each fasta line

Example:

```
1     249250621    52              60    61
2     243199373    253404903       60    61
3     198022430    500657651       60    61
```

Pysam - Python fasta interface

- Pysam - a Python toolkit for working with genomic files
- pysam.Fastafile:
  - Create a fasta file parser: `fasta = pysam.Fastafile(path_to_file)`
  - Get the sequence names in the file: `fasta.references`
  - Get the lengths of the sequences: `fasta.lengths`
  - Retrieve a (part of) sequence: `fasta.fetch(sequence_name, [start], [stop])`
- Fasta coordinates in pysam are zero-based
  - Not true for all file types in that pysam supports
- Other fasta interfaces for Python exist
  - pyfasta - works only on fasta files
  - biopython - a much larger toolkit that complement sequences, etc.
  - We are describing pysam, as it covers all the file types covered in the course

Exercise: FASTA file format (15 minutes)

- An example FASTA file is found under */sbgenomics/project-files/example_human_reference.fasta*
- View the contents of the file
  - You can use “!head filename.ext” in the Notebook to invoke linux head
- Create a pysam Fastafile parser
- Get and print sequence names
  - How many sequences are there?
- Fetch the entire sequence
  - How long is it?
  - Print the first 100 bases

Exercise: FASTA file format (20 minutes)

- Full hg38 FASTA file is located under in *"/sbgenomics/project-files/Homo_sapiens_assembly38.fasta"*
- Create a pysam Fastafile parser
- Get sequence names for all contigs
  - How many contigs are there?
  - Read the names of the contigs
  - How long is chromosome 5?
- Fetch section chromosome 17:43044295-43125370
  - What is the “GC content” on this region? (percent of G and C bases)
  - What is the most common 3-mer?
- Fetch base at chromosome 1:248755121
  - What is the base?
- Fetch region at chromosome 1:50000-50100
  - What is the base composition?

FASTQ file format

- Most common format for storing sequencing reads
- It's spread across four lines. The four lines are:
  - “@” followed by a read name
  - Nucleotide sequence
  - “+” , possibly followed by some info, but ignored by most of the tools
  - Quality sequence

- Example:

```
@ERR294379.100739024 HS24_09441:8:2203:17450:94030#42/1
AGGGAGTCCACAGCACAGTCCAGACTCCCACCAGTTCTGACGAAATGATGAGAGCTCAGA
+
BDDEEF?FGFFFHGFFHHGHGGHCH@GHHHGFAHEGFEHGEFGHCCGGGFEGFGFFDFFHB
```

Genes in HG38 (UCSC Genome Browser)

- A FASTA file contains raw sequences of nucleotide from a genome
- Some sections of these sequences have biological meanings attached
- Examples:
  - Chromosome 17, 43045681 and ends at 43124096 bp (41196312- 41277500 bp in hg19 and v37) is the BRCA1 Gene, implicated in breast cancer
  - Chromosome X, 73.82 Mb - 73.85 Mb is the XIST lncRNA, implicated in X inactivation
  - Chromosome 1, 121.1 Mb - 124.3 Mb is the Chromosome 1 centromere
- USCS Genome browser is a place where different annotation tracks can be explored in depth
- USCS Genome browser is located at:

https://genome-euro.ucsc.edu/cgi-bin/hgGateway

Annotations file formats

- Annotations are stored in few (similar) file types
  - **BED**
  - GTF
  - GFF
  - …
- **BED file format** (tab-separated):

CHROM START END NAME score ticks blocks

- Chrom, start and end are required
- Start is 0-based
- End is open interval (not included)