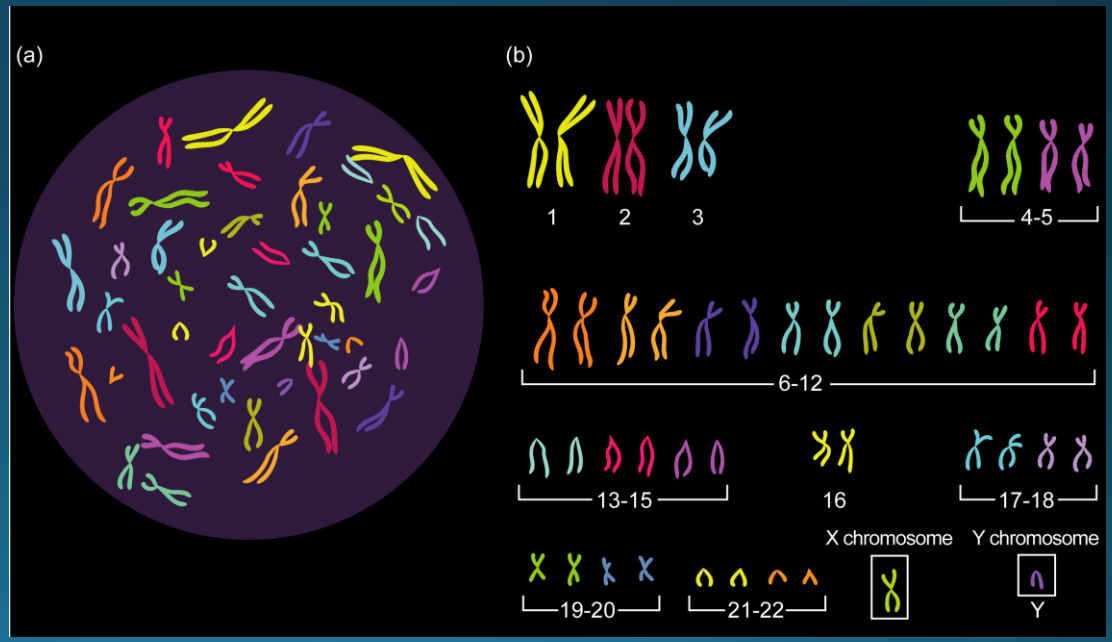
Overview and Opportunities

Genome Assemblers

What is a genome?

- A **genome** is an organism's complete set of DNA, including all of its genes. Each genome contains all of the information needed to build and maintain that organism[1].
- Example:
  - Human Genome
    - 3 billion DNA base pairs
      - 23 pairs of chromosomes
      - Mitochondrial DNA



https://gcps.desire2learn.com/d2l/lor/viewer/viewFile.d2lfile/6605/4821/Cells4_print.html

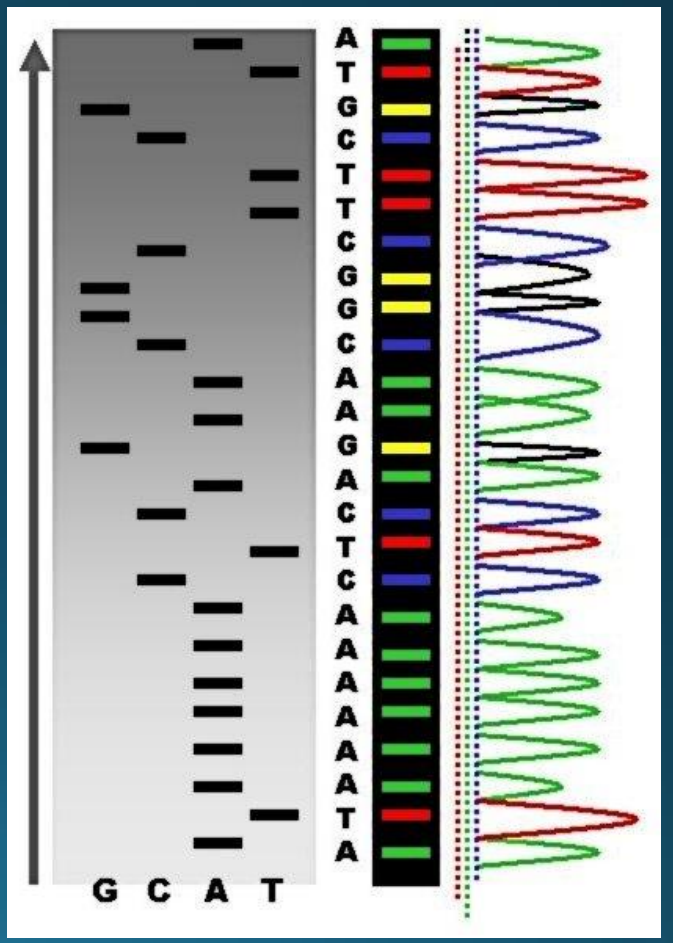
[1]National Institute of Health

Why understanding DNA is important?

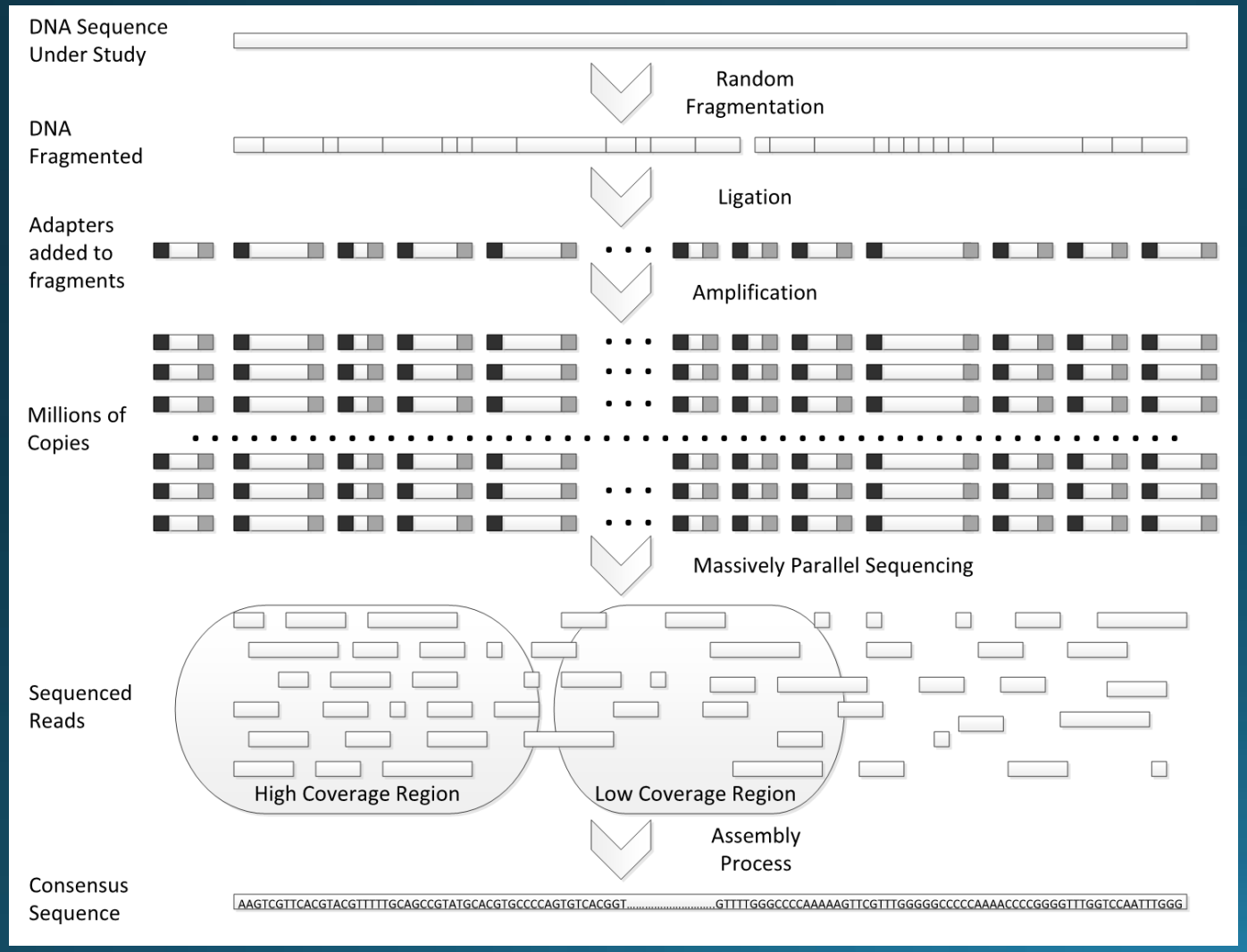
- Disease Diagnosis and Treatment
- Paternity and Legal Impact
- Forensics and DNA
- Agricultural
- Conservation
- Human History

DNA Sequencing Technology

- 1976-2002: First Generation Sequencers
  - Capillary Sequencers (Sanger Method)
  - 700 bp **reads**
  - Cost: $15,000,000 per genome
- 2002-Present: Next Generation Sequencers
  - Massively Parallel
  - Cost: $9,000-$10,000



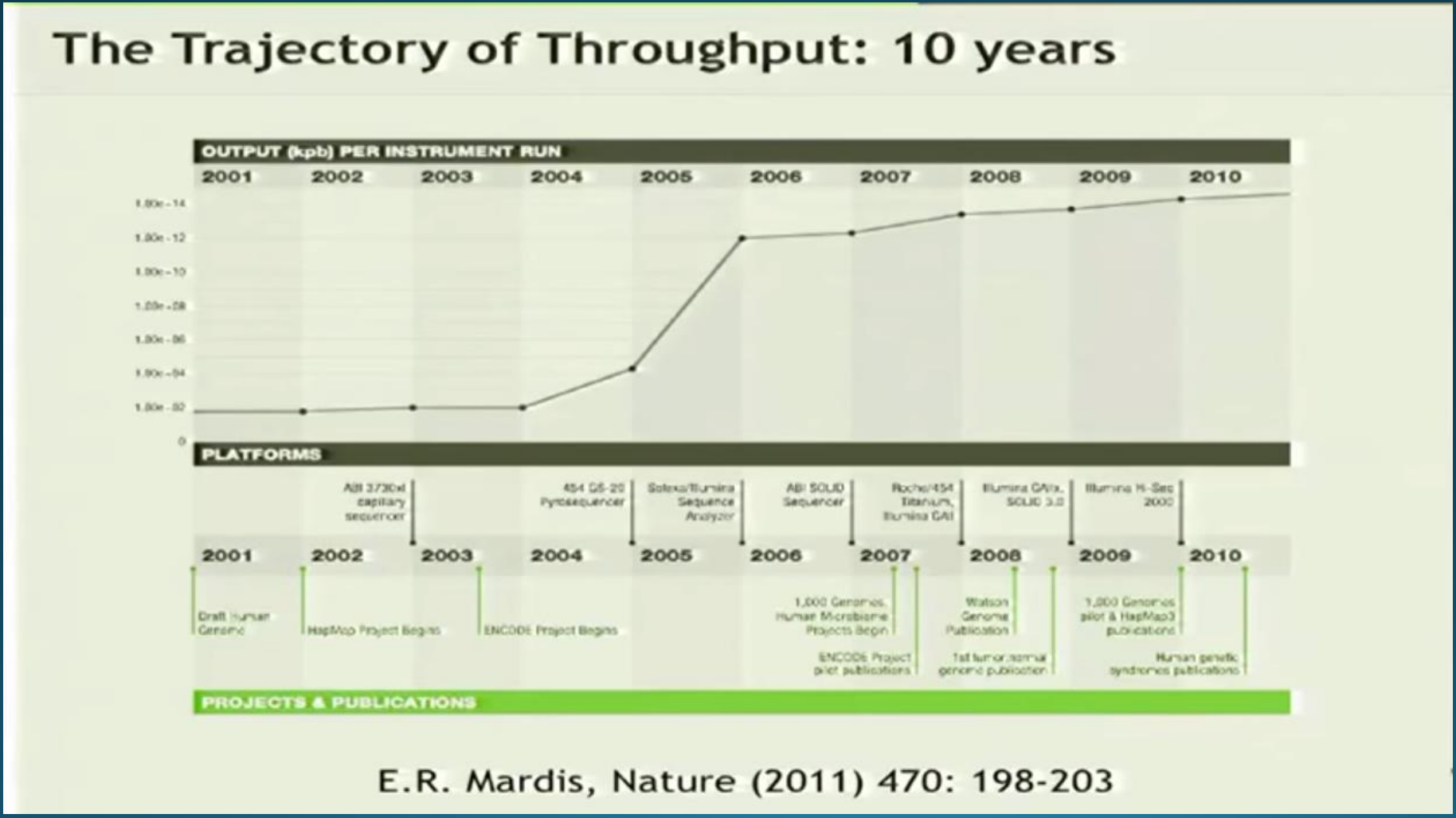
Massively Parallel Sequencing



Sequencer	Read Length (bp)	Accuracy	Reads	Advantage
Illumina	50SE, 50PE	98%	3G	High Coverage
SOLiD	50+35bp	99.94%	1.2-1.4M	Accuracy
Roche 454	700	99.9%	1M	Read Length

(Liu, et al., 2012)

DNA Sequencers Throughput



E.R. Mardis, Nature (2011) 470: 198-203

Assembly Process: Computational Problem

- Brute Force:
  - $O\left(\frac{N!}{(N-M)!(M)!} * R\right)$
- Dynamic Programming
  - $O(N * M * R)$

  where N is the length of the reference genome, M is the size of the read and R the number of reads produced.
- Assuming a Human Genome on an Illumina Sequencers these numbers may be:
  - N = 3,000,000,000
  - M = 50
  - R = 3,000,000,000

Current Approaches: Indexing

- Reference Genome Indexing
  - Bowtie
    - Burrows-Wheeler Transform
  - BLAT
    - K-mers Hashing
  - SeqMap
    - Pigeonhole Principle based Hashing
  - MAQ
    - Reference Hashing
  - SSAHA
    - Reference Hashing

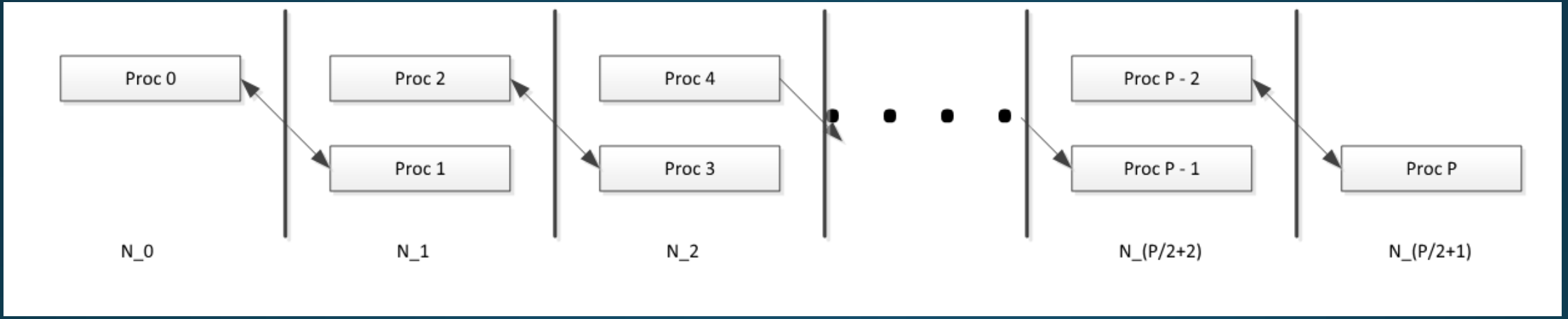
Problem with current approaches

- Single Nucleotide Polymorphism
  - Reference genome is only that: a reference
    - E.g. Around 10,000,000 nucleotides change between humans
- Quality Values
  - On most indexing based algorithms quality values are only used on deciding between 2 possible solutions.
- Greedy outcomes

Ingredients an Genome Assembly Algorithm

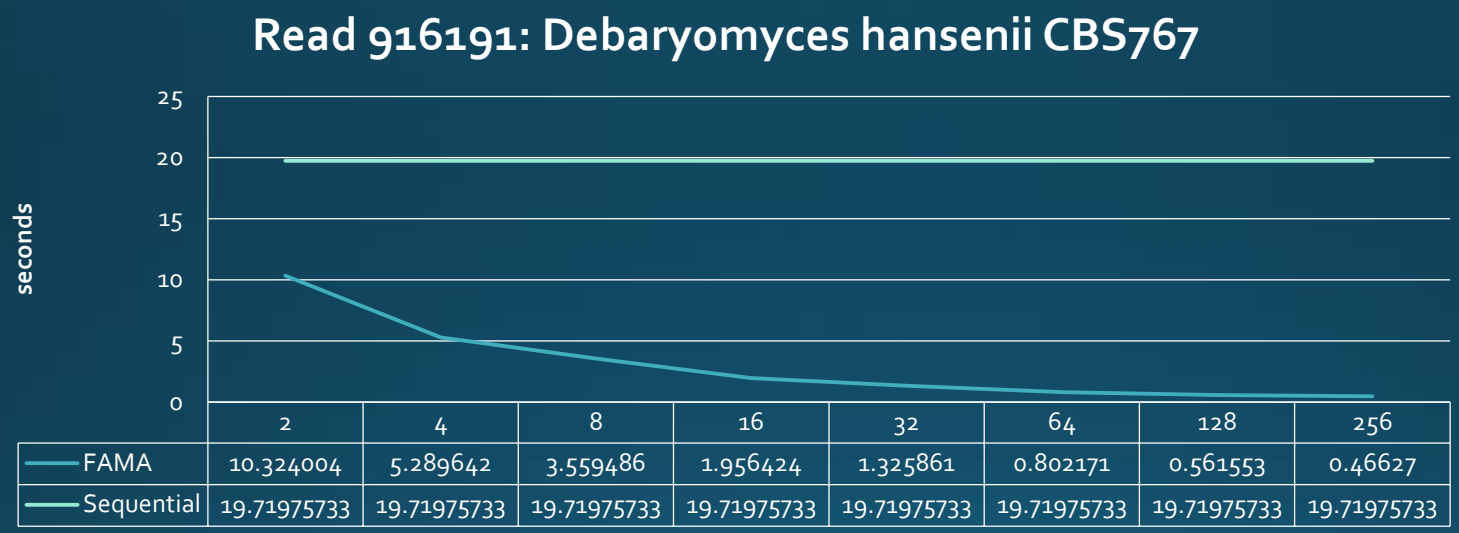
- MUST be FAST (when compared with current solutions)
- MUST provide optimal results
- MUST incorporate quality values
- MUST accept thresholds
- MUST take into account SNPs.

FAMA: Fast Accurate Mapping Assembly



- All computing nodes have a segment of the reference genome.
- Reads are sent to each node.
- A master node gather relevant results only.

FAMA: Preliminary Results



	2	4	8	16	32	64	128	256
FAMA	10.324004	5.289642	3.559486	1.956424	1.325861	0.802171	0.561553	0.46627
Sequential	19.71975733	19.71975733	19.71975733	19.71975733	19.71975733	19.71975733	19.71975733	19.71975733

Next Steps

- Benchmarks against Indexing Methods & Publish
- Apply Methodology to Multi-Sequence Alignment