Sequence Assembly and Next Generation Sequencing

BIOL 7711
Computational Bioscience

Biochemistry and Molecular Genetics
Computational Bioscience Program
Consortium for Comparative Genomics
University of Colorado School of Medicine

David.Pollock@uchsc.edu
www.EvolutionaryGenomics.com

Evolution
Genomics
Biodiversity

Computation on Nucleotides

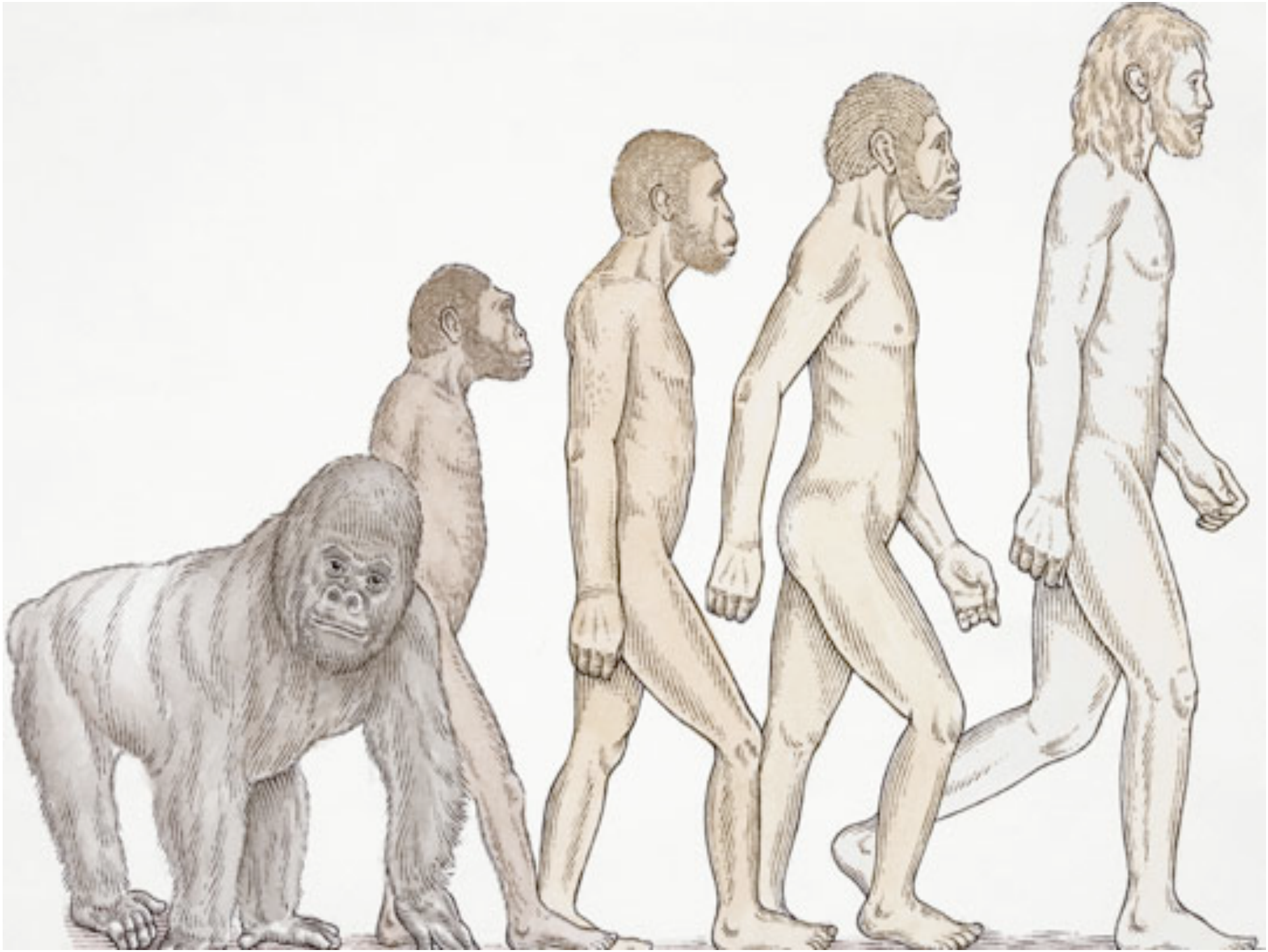
- Alignment--generally more uncertain than amino acids
  - Occasionally provides a more correct answer
- Sequence Assembly
- Sequence Annotation
  - Genes, splice sites
  - Regulatory regions, TFBS
  - Chromatin binding
- Mutation processes
- Route to information much faster, cheaper

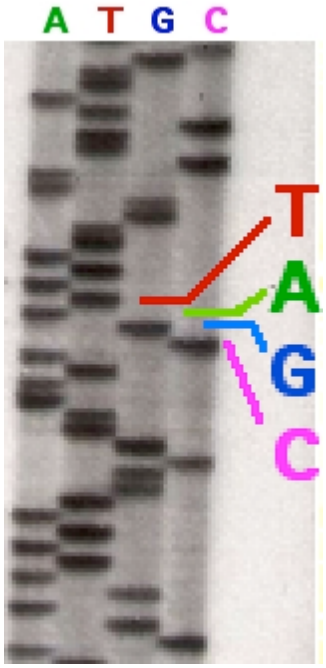
Computation on Nucleotides

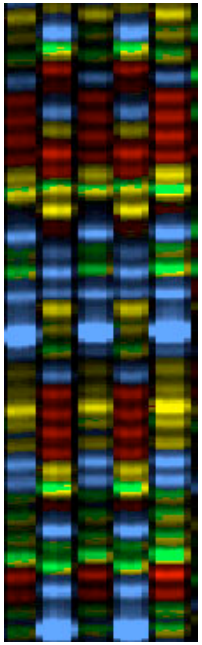
- Alignment--generally more uncertain than amino acids
  - Occasionally provides a more correct answer
- **Sequence Assembly**
- Sequence Annotation
  - Genes, splice sites
  - Regulatory regions, TFBS
  - Chromatin binding
- Mutation processes
- Route to information much faster, cheaper
  - **High-throughput next-generation sequencing**

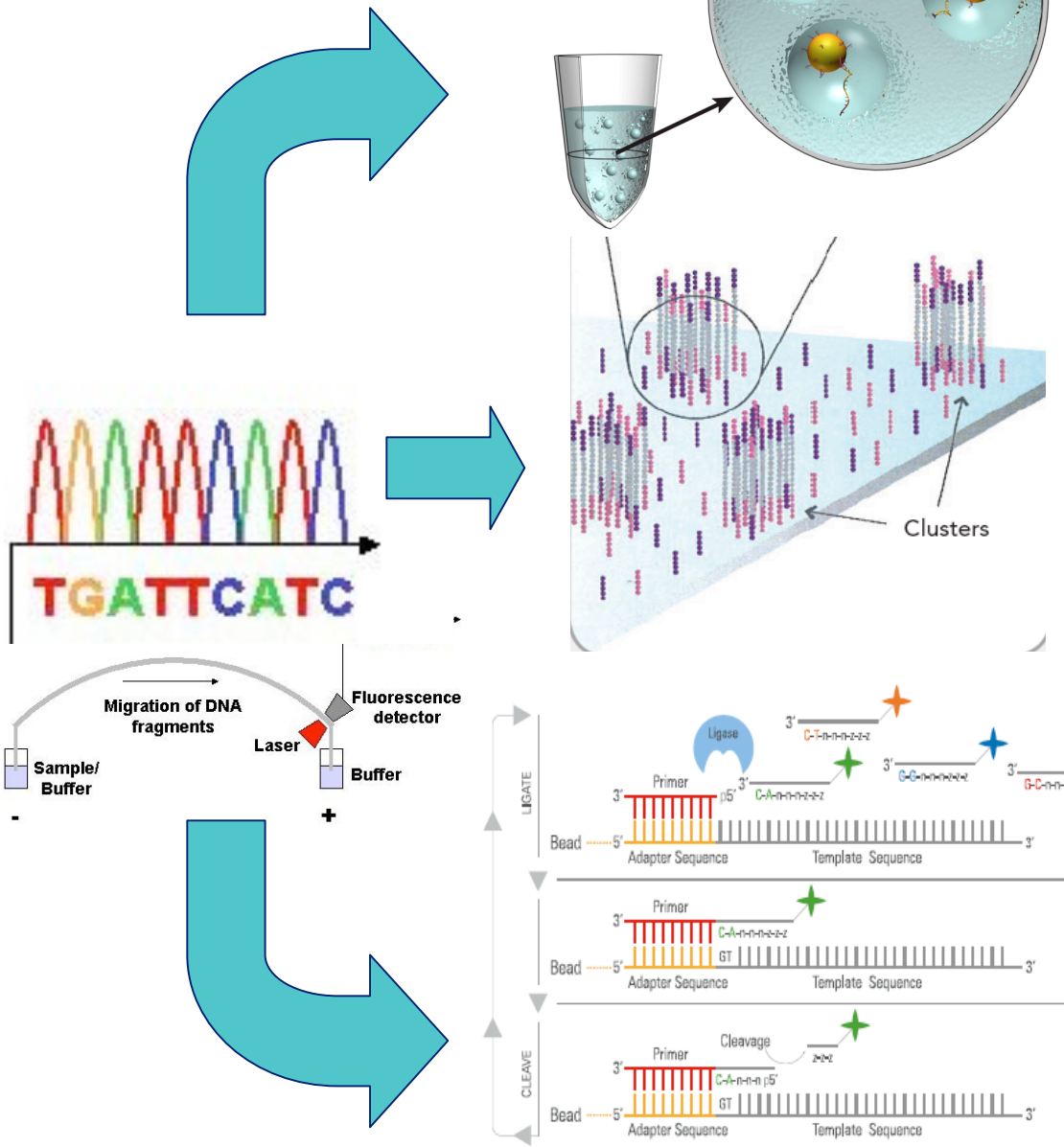
The Evolution of Sequencing

- Sanger sequencing
  - Gels
  - Cycle sequencing
  - Fluorescence
  - Capillary electrophoresis
- Sequencing, the “Next Generation”
  - “Sequencing by synthesis”
    - Pyrosequencing (Roche/454)
    - Cluster sequencing (Illumina/Solexa)
  - Sequencing by ligation (ABI/SOLiD)



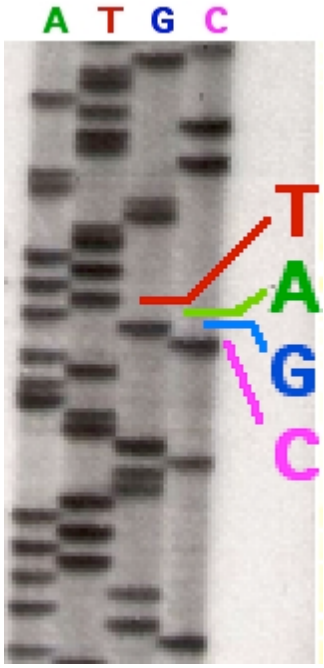

A T G C
T
A
G
C

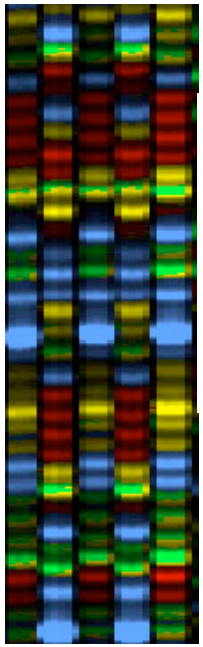


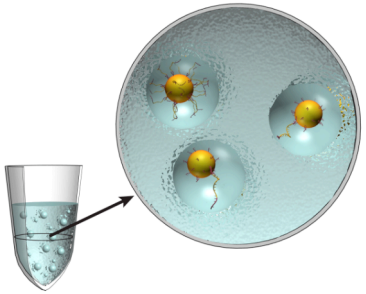

TGATTCATC
Migration of DNA fragments
Fluorescence detector
Laser
Sample/ Buffer
Buffer
-
+
Clusters
LIGATE
CLEAVE
Ligase
Primer
Bead
Adapter Sequence
Template Sequence
Cleavage

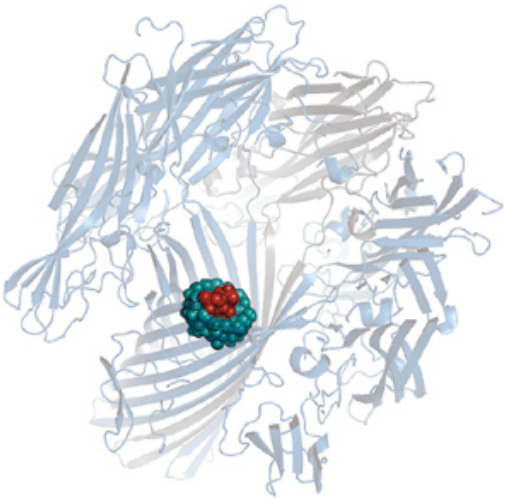
The Evolution of Sequencing

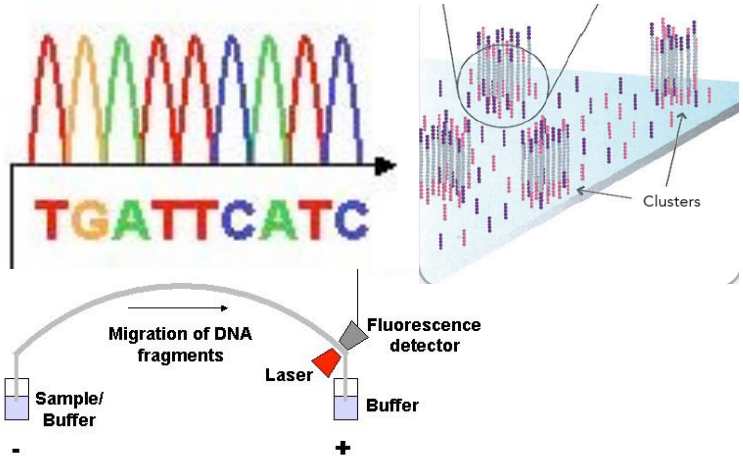
- Sanger sequencing
  - Gels
  - Cycle sequencing
  - Fluorescence
  - Capillary electrophoresis
- Sequencing, the “Next Generation”
  - “Sequencing by synthesis”
    - Pyrosequencing (Roche/454)
    - Cluster sequencing (Illumina/Solexa)
  - Sequencing by ligation (ABI/SOLiD)
- Next “Next Generation”
  - 4th Generation?
  - Single molecule sequencing

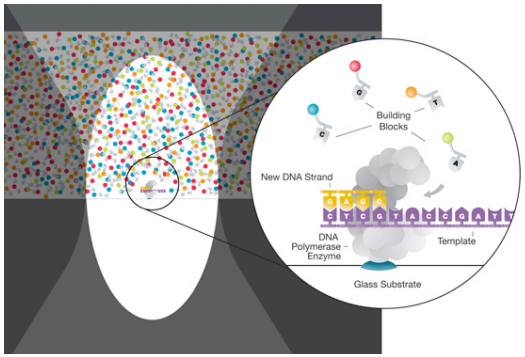

A T G C
T
A
G
C

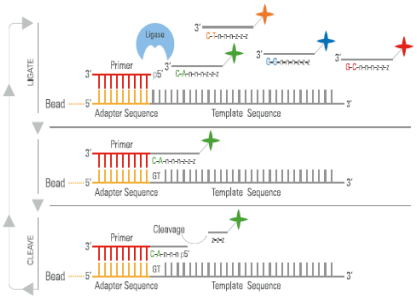


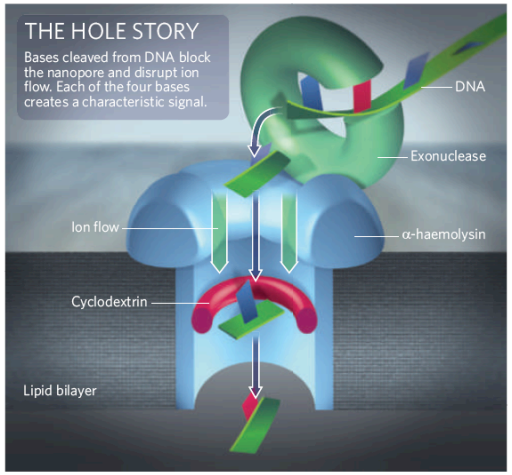




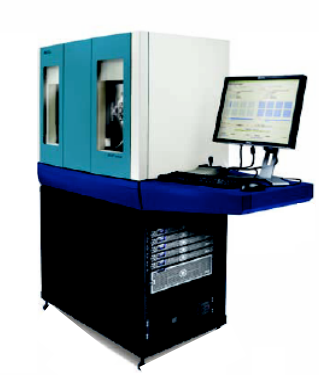

TGATTCATC
Clusters
Migration of DNA fragments
Fluorescence detector
Laser
Sample/ Buffer
Buffer
-
+


Building Blocks
New DNA Strand
DNA Polymerase Enzyme
Template
Glass Substrate


Ligase
Primer
Bead
Adapter Sequence
Template Sequence
LIGATE
CLEAVE
Cleavage


THE HOLE STORY
Bases cleaved from DNA block the nanopore and disrupt ion flow. Each of the four bases creates a characteristic signal.
DNA
Exonuclease
Ion flow
α-haemolysin
Cyclodextrin
Lipid bilayer

Macro versus Micro Reads



Applied Biosystems
SOLiD

Read Length
35 - 75bp **<=>** 250 - 450bp

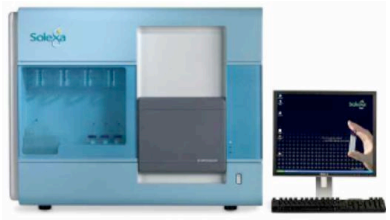
Base Pairs Per Run
3 - 10 Gb **<=>** 0.1 - 0.5 Gb

Base Pairs Per Day
1 - 1.5 Gb **<=>** 0.2 - 1.0 Gb

Number of Sequences
100 M **<=>** 1.2 M

Roche / 454 FLX

Run Time
3 - 7 days **<=>** 0.5 days



Illumina / Solexa
Genetic Analyzer

Reagent Cost per Run
~$4K - $12K **<=>** $6K

Error Rate
Varies, different characteristics

Technology and Informatics

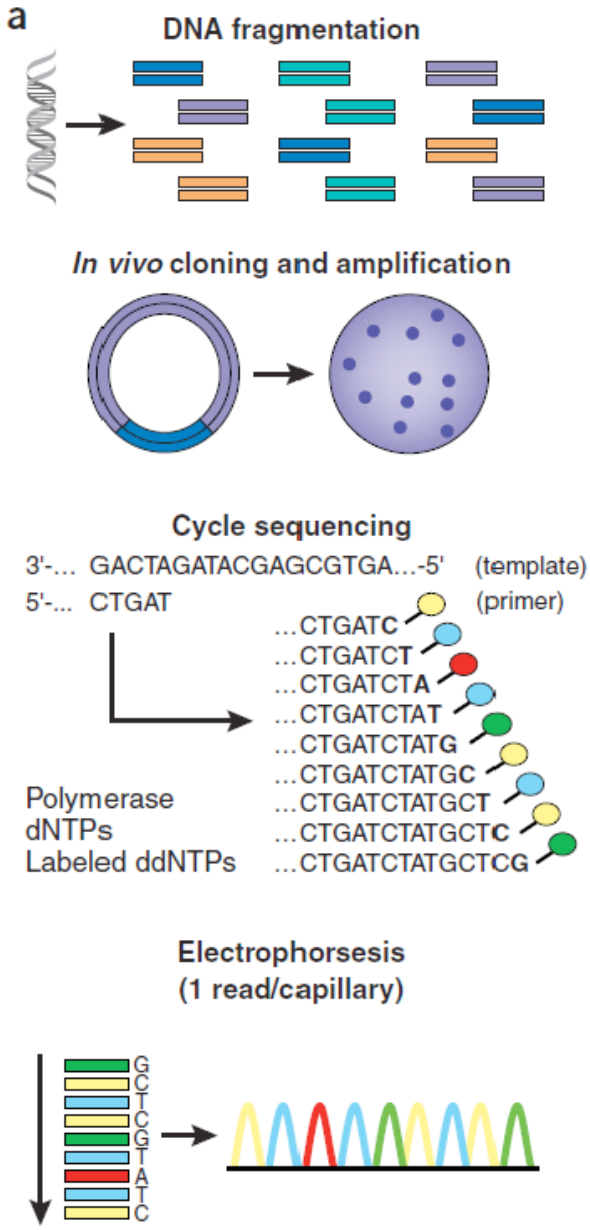
PR Space versus Science Space

- Flow and phasing
- Data quality and Error rate
  - Variation along sequence
  - Quality scores (Equivalence?)
- Length distribution versus average
- Raw versus recovered sequence
  - How much coverage with different methods?
- Tagging (barcodes) and multiplexing
  - Variation in coverage

Next-Gen Basics

- Library creation
  - Shearing, size selection
    - Size distribution
  - Specific primer sequences (adaptors) flank target sequence
    - Allows amplification
      - Opportunity for extra “mutation”
- Tagging (barcodes)
  - Proportion of sequence wasted
- Ligation or amplification (454)
- Paired ends

Sanger Sequencing



- DNA is fragmented
- **Cloned** to a vector
  - Plasmid, BAC
  - Linkage
- Cyclic sequencing
- Separation by electrophoresis
- Read fluorescent tags

Micro Reads: SOLiD & Solexa

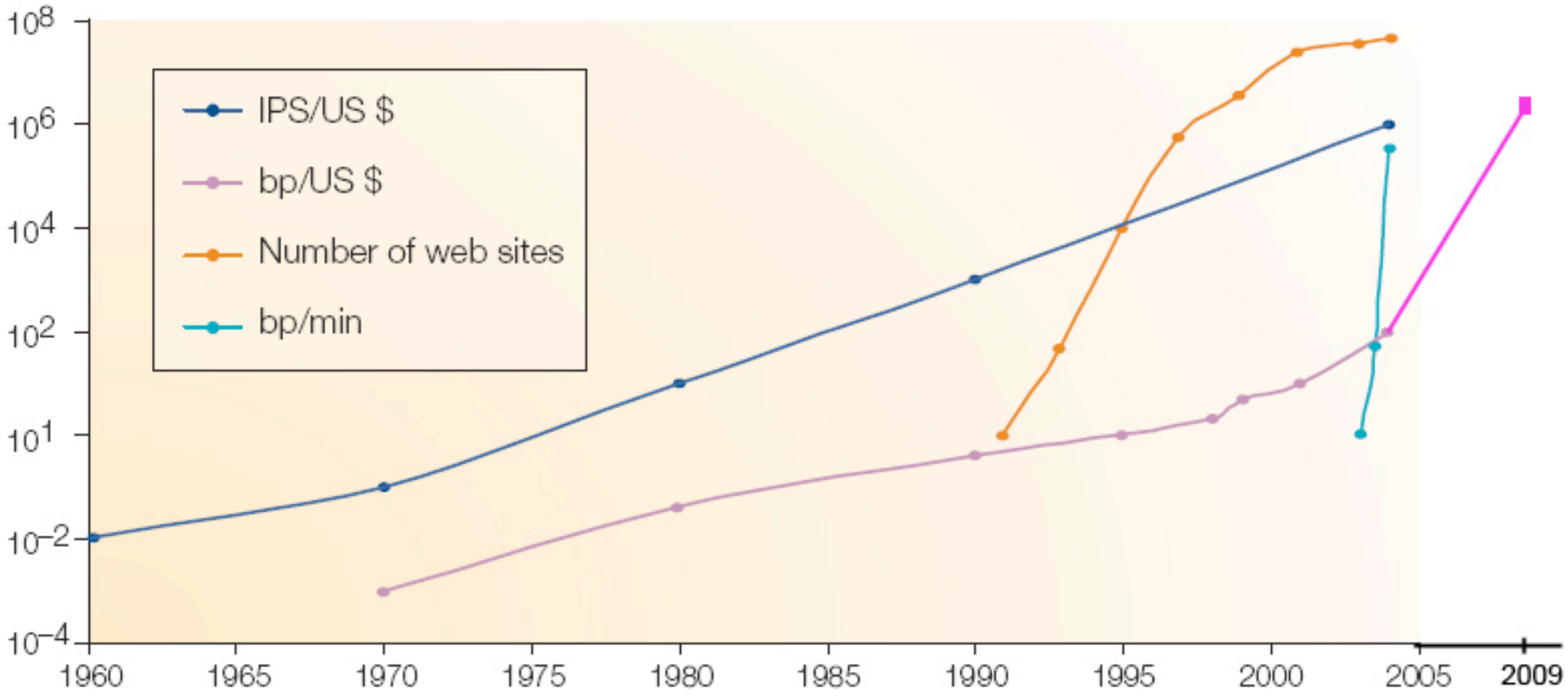
- Reseqeuncing
- SNP detection
- Micro RNA (23 bp)
- Counting (e.g., transcriptome profiling, ChIPSeq)
  - Adjustable dynamic range ($$)
- Hard to place near repetitive elements
- Harder to assemble *de novo*
- *75-100 bp reads intermediate*

Macro Reads: 454, PacBio

Sort of Solexa?

- Many fewer reads
- Much longer
- De novo sequencing
- Amplicons and tagging
- Repetitive regions

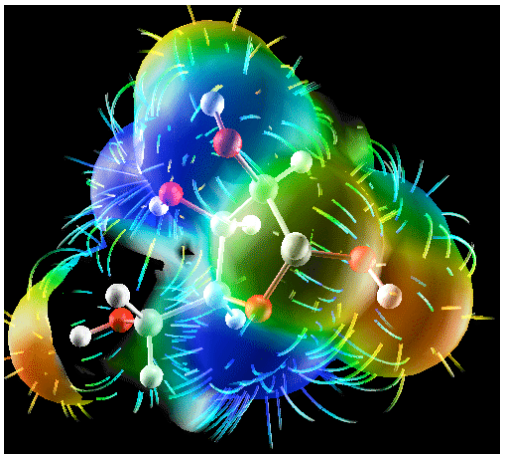
Getting Cheaper all the Time



Adapted from Shendure et al 2004

Informatics Challenges

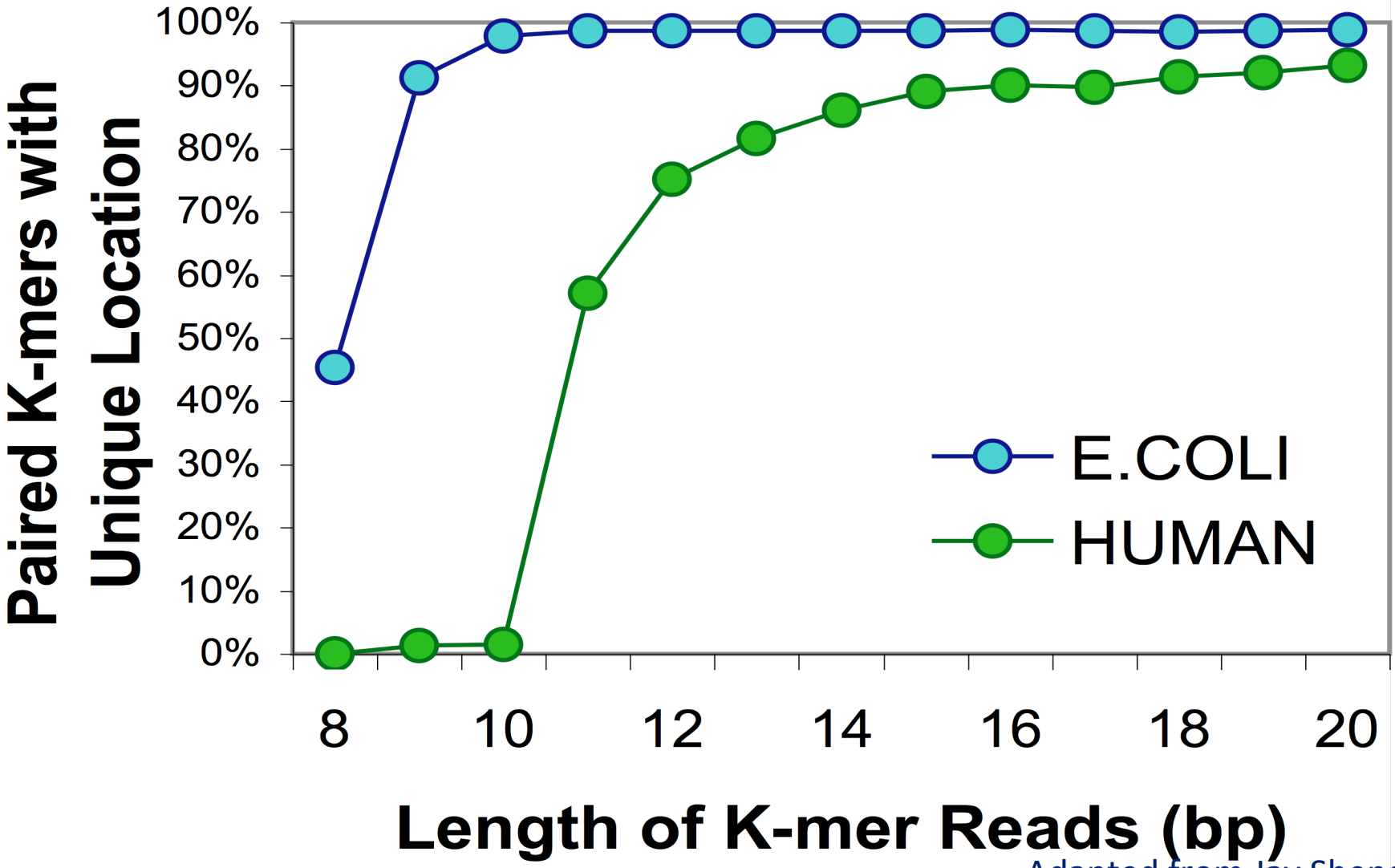
- Data storage
  - 6+ TB for microread raw image files
    - Toss them out: calculate on the fly
- Computation Speed
  - Faster to align long reads
    - Exponential with number of reads if comparing to each other
- Software
  - Getting better
  - Assembly, mapping
  - counting, variation



4th Gen PR Space
The 2nd Coming

- 1 Kb sequences, highly accurate
- Fast, cheap
  - $300 genome (10x) in 30 minutes (??)
- Less front-end preparation and labor
- What is required for personal genomics?
- 10,000 vertebrate genomes project

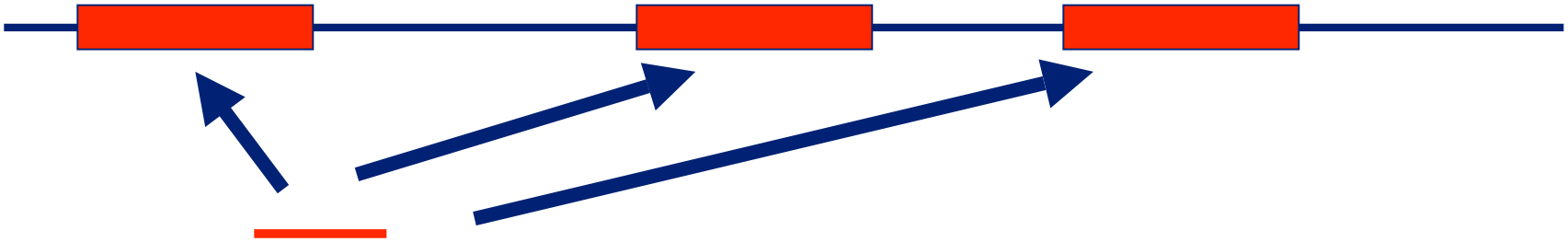
Read Length & Resequencing



Adapted from Jay Shendure

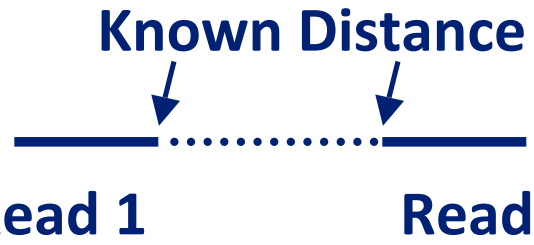
Mapping Unique Reads

Repetitive DNA **Unique DNA**



Single reads map to multiple positions if they hit repetitive DNA

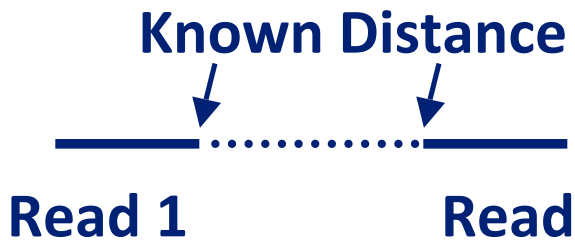
Paired End Reads

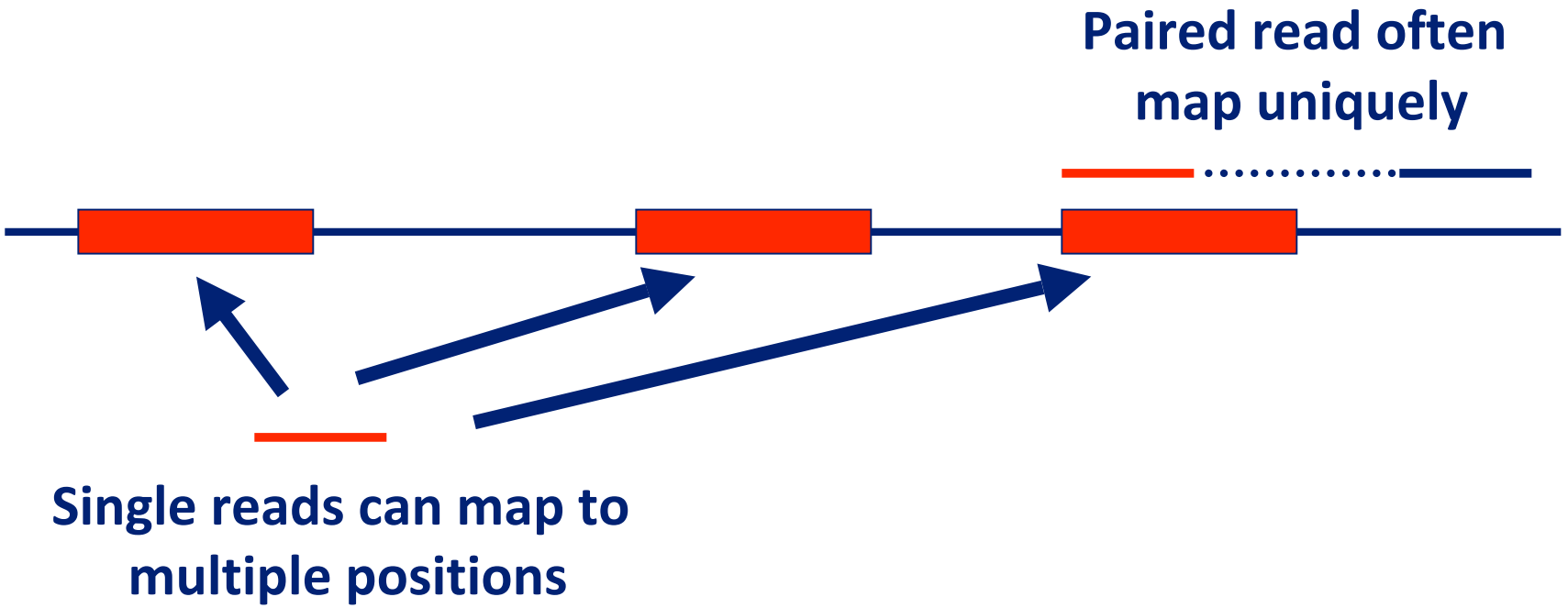


Solexa: paired end is both ends of ~300 bp fragment (shorter than a 454 read, shorter than most TEs)

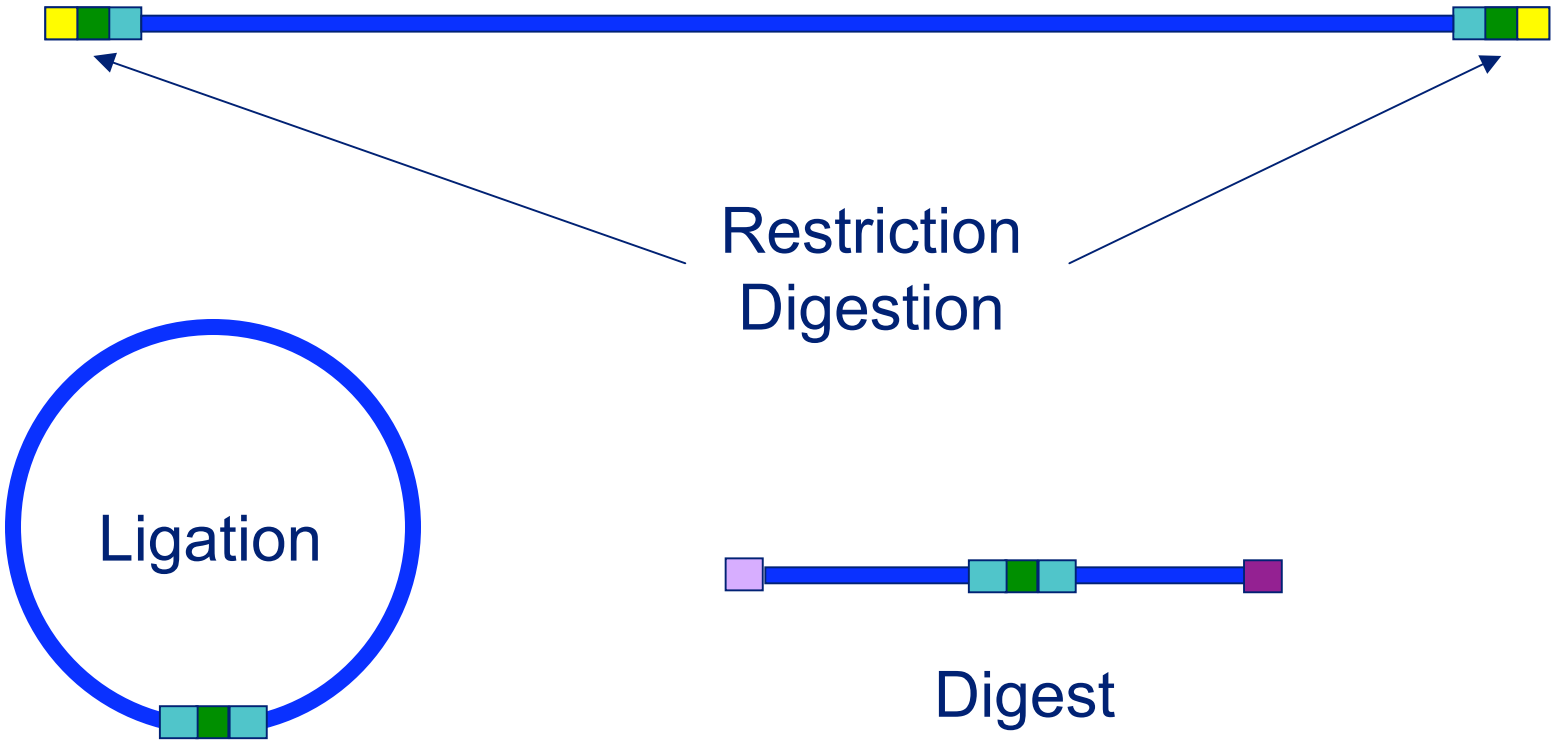
454 paired ends are:
~3Kb
~8Kb
~20Kb

Paired End Reads





454 Paired-End Library Construction



Other Order Information

- FISH mapping
- Recombination map
- BAC paired ends
- Verification by PCR
  - Quite expensive; usually long-term follow-up, only samples

Contig Assembly

- Significant overlap at ends of fragments
  - **IF** overlap fragment is unique in genome, then perfect assembly of contigs (with gaps in between)
  - So, want long enough to be likely unique
  - Want to identify repeat sequences
  - “Shortest Common Superstring” Problem
    - But, tend to delete duplicate regions

Oligo Frequency Model

- $$P(oligo) = \left(\prod_{nuc} freq_{nuc} \right)^{L}$$
- Expected occurrences in genome?
  - Genome length $N=3x10^9$
  - Nucleotide frequencies equal
- What length expected to occur <1 time?
- For that length, what is probability of 2, 3, 5, 10?
  - Use Poisson

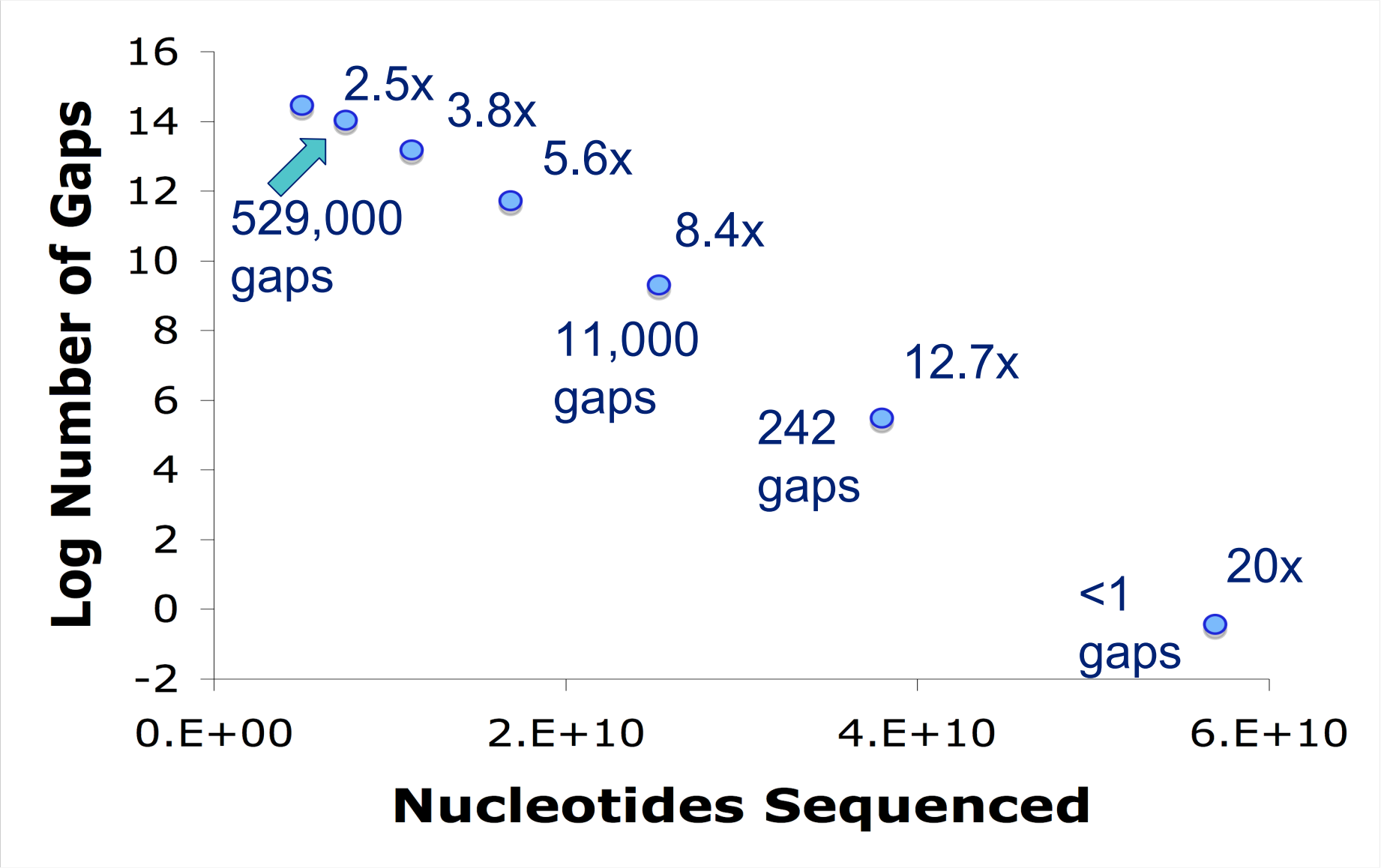
Shotgun Sequencing

- Random fragments
- Coverage (C), or redundancy, is average number of times a nucleotide should be sequenced
  - C=NL/G
    - Number reads sequenced
    - Length of read (average)
    - Genome size
- How many nucleotides covered at least once?
  - Poisson approximation: $1 - e^{-c}$

More Shotgun Rough Expectations

- Average contig length: $(L/c)e^{c}$
- Number of gaps: Ne^{-c}
- Average gap length: L/c

A Quick Visual



But It's Not That Simple

- Calculations assume you know where the reads go
- Sequencing errors
  - Quality scores, low error in the first place
- Sampling bias
  - Cloning bias is particularly bad
    - Some sequences are poison
- Repetitive sequence
  - TEs, mini-satellites, microsatellites, low complexity, tandem repeats
  - Gene paralogs (really want to get these right!)
- The more free unplaced ends, the more likely to have spurious overlap (orientation, revcomp)

More Concerns

- Over-collapsing
  - Leaves extra unplaceable fragments
  - More reads with no place to go
  - Shortest common superstring => biased
- BAC ends, paired end info
  - Drastically reduce the possibilities of where a contig can go
  - Supercontigs
- Polymorphisms

Screening Reads

- Prefer accurate reads: 98%+
  - Clip reads if predicted accuracy drops
- Remove vector and other contaminants
  - Alignment & identification
  - Especially other recent genome projects
- Screen for known repeats

Evaluate Overlap

- Compare each fragment with each other
  - This is why Illumina seqs so hard! N^2
- Example
  - >= 40 bp overlap, <= 6 mismatches
  - Either a true overlap, or a repeat
  - Figure this out asap!
  - Fragments with excessive numbers of overlaps are probably repeats

Unambiguous Contigs

- Combine fragments with only one possible assembly into longer sequences
  - Perfect matches
  - Match no other: no conflicting overlaps
- Drosophila
  - 3.158M reads => 54K unitigs
  - Still might be wrong
- Can extend unitigs up to one read length into repeat regions

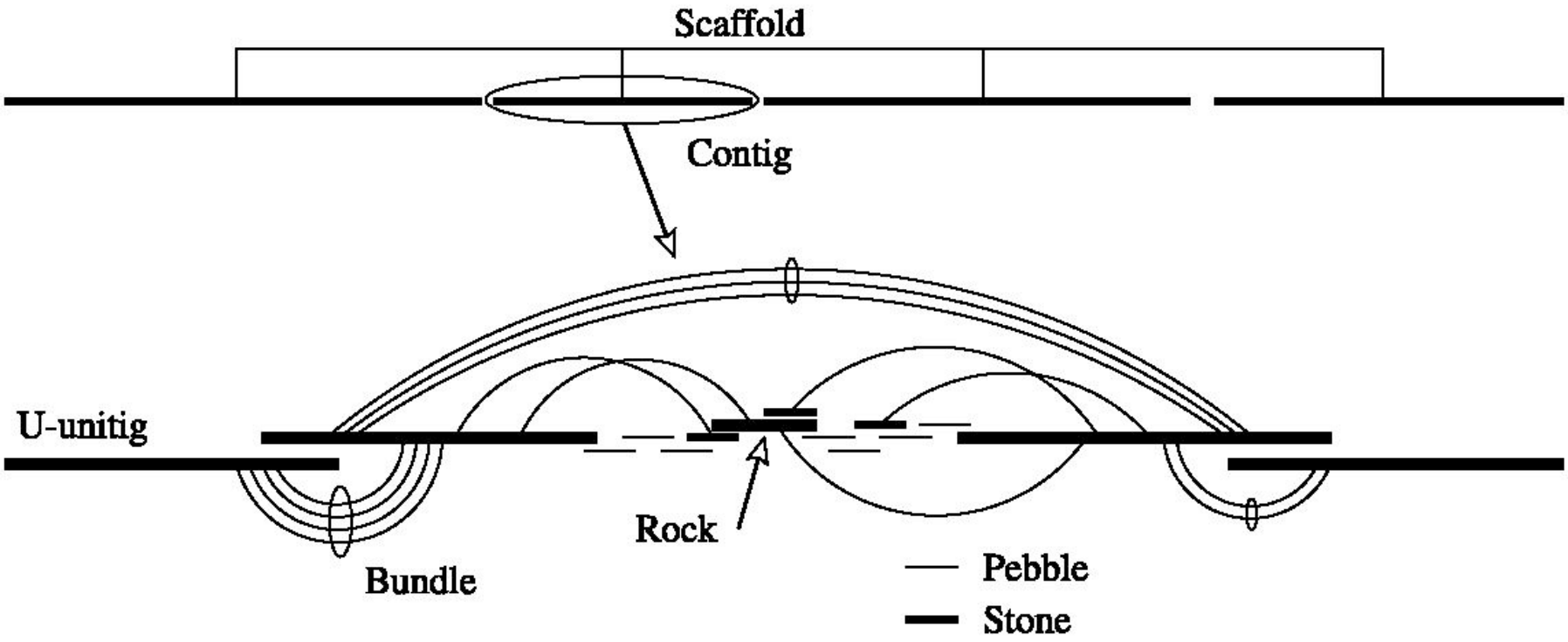
Scaffolds

- Set of ordered, oriented contigs
- Gaps of approximately known size
  - BAC ends in two different contigs
  - BAC library of tight known size range
  - Same concept for other paired end reads
- “Bundle” if more than one placement
  - The more mate pairs, the more reliable
- Map scaffolds with FISH, recombination

Place Repeats

- Placement evidence from mate pairs
  - Multiple = rocks
  - Single = stones
  - None = pebbles
    - Basically just guessing. Statistical

Scaffold Visual



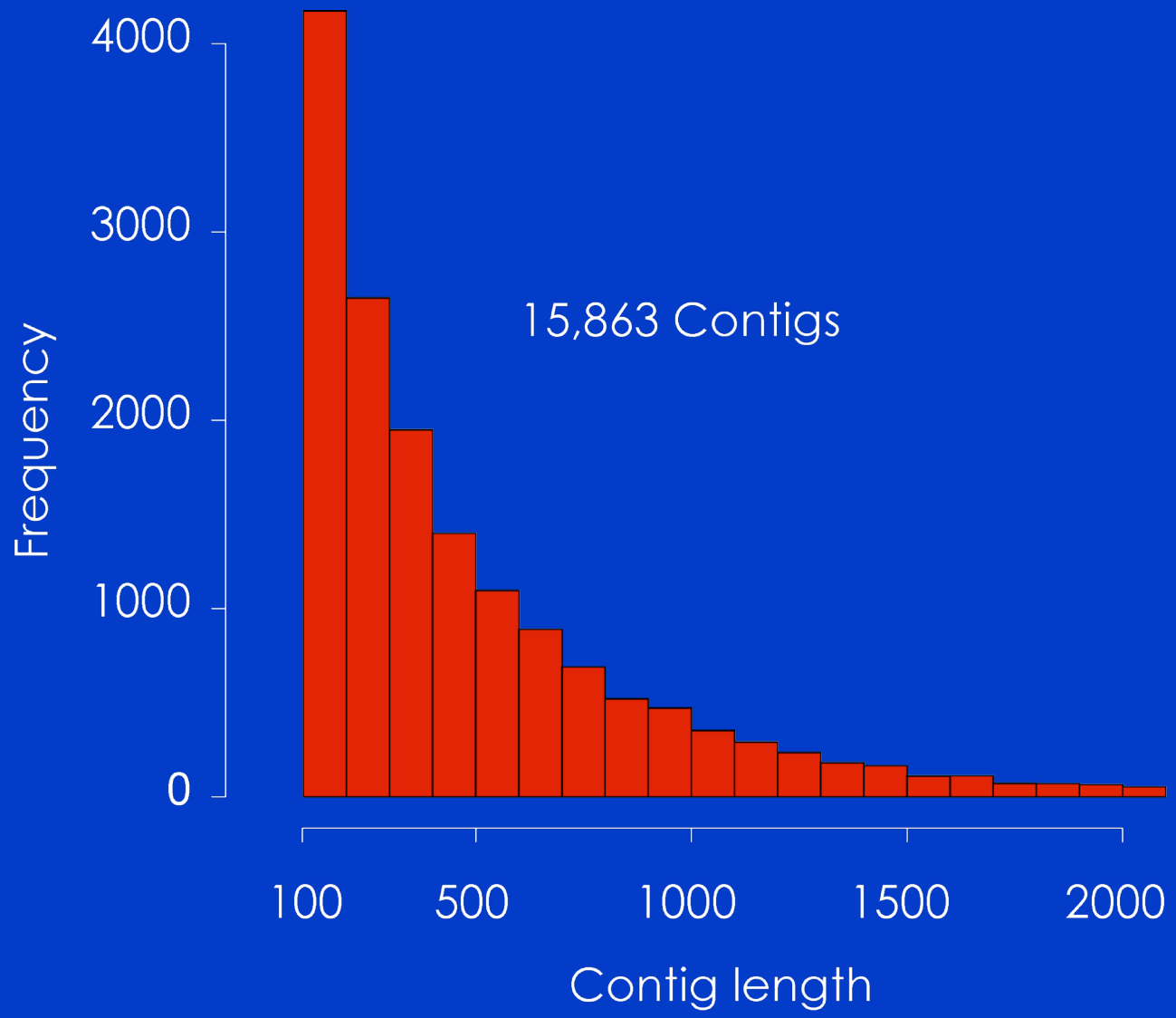
Finishing and Validating

- Manual review and adjustment
- Quality control
  - PCR
- Overlaps, mis-assembly, etc.
- Gaps can reflect unresolvable repeats or low coverage in a region
- More intense sequencing in a region
  - Compare to other sequencing efforts (Celera)

NextGen Assembly

- More faster cheaper shorter error-prone
- Bacterial example: Mycobacterium spp.
  - ~4 Mb genome
- Solexa/Illumina, de novo assembly with VELVET assembler
- 50x coverage = 2Gb, 36 bp reads

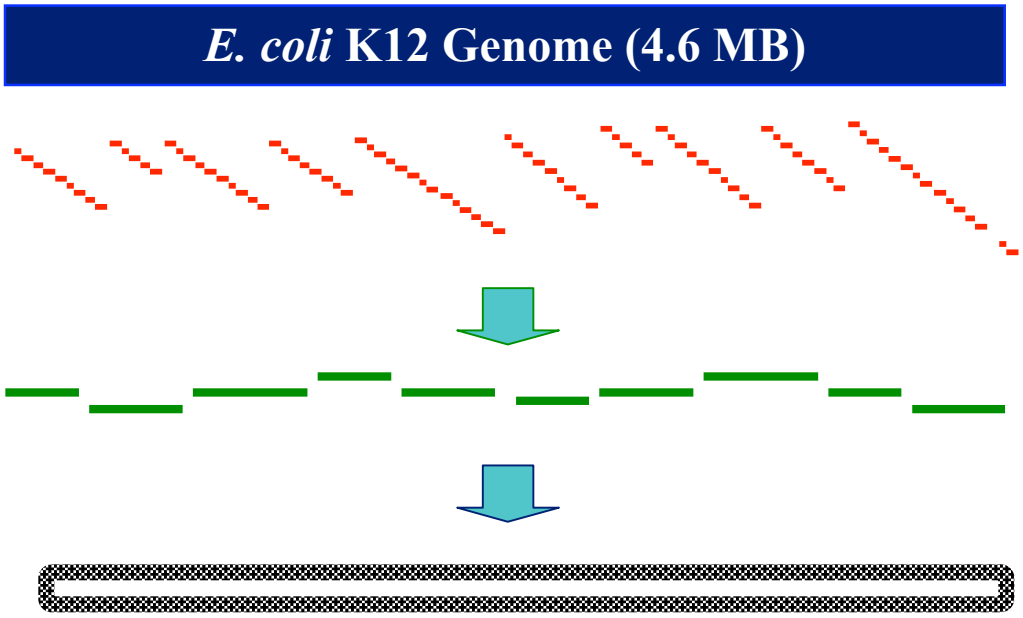
Assembly Result



454 E. coli Assembly

old 250 bp reads

454 Read Type	Genome Coverage	Number of Contigs/ Scaffolds
Shotgun	15×	98
+		
3 Kb Jump	18×	7
+		
20 Kb Jump	20×	1

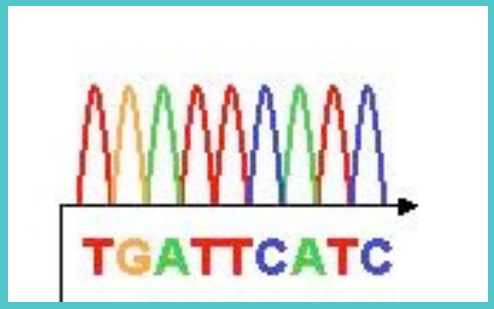


So, this is 20x total coverage

Consensus Accuracy: ~ 99.999%

Open Questions

- What is the most efficient way to combine various sequencing methods?
  - Solexa paired ends versus 454 single reads
    - SOLiD for its accuracy?
  - 454 paired end 3Kb, 8Kb, 20Kb mix
  - Sample repeat regions ahead of time
- Do you have to have BACs for a eukaryote genome?
- Any tricks to finish off gaps efficiently?
- Priors: low heterozygosity, few repeats


TGATTCATC