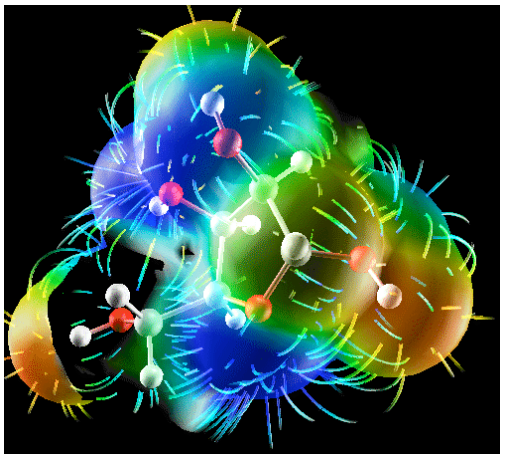
Informatics Challenges

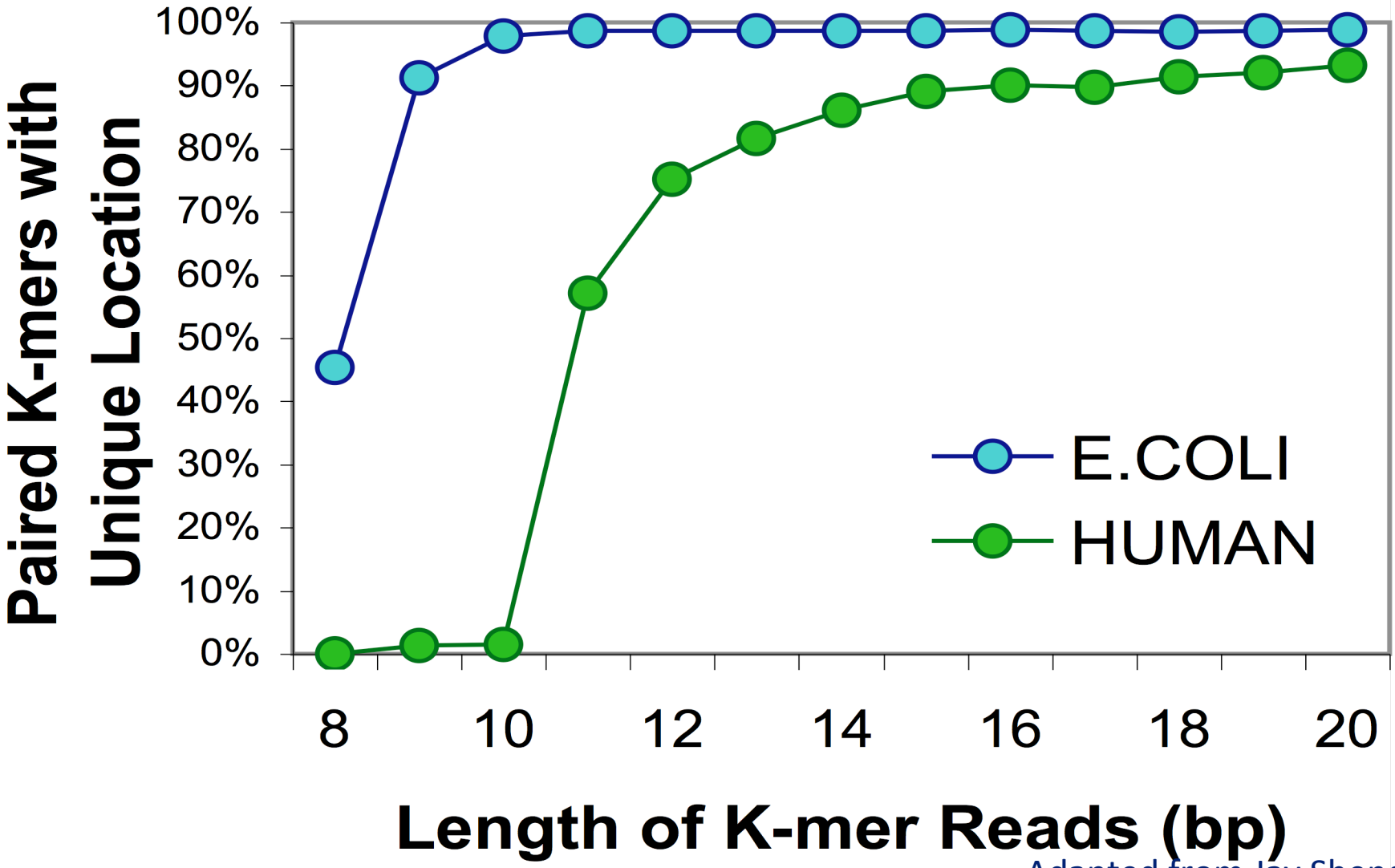
- Data storage
  - 6+ TB for microread raw image files
    - Toss them out: calculate on the fly
- Computation Speed
  - Faster to align long reads
    - Exponential with number of reads if comparing to each other
- Software
  - Getting better
  - Assembly, mapping
  - counting, variation



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

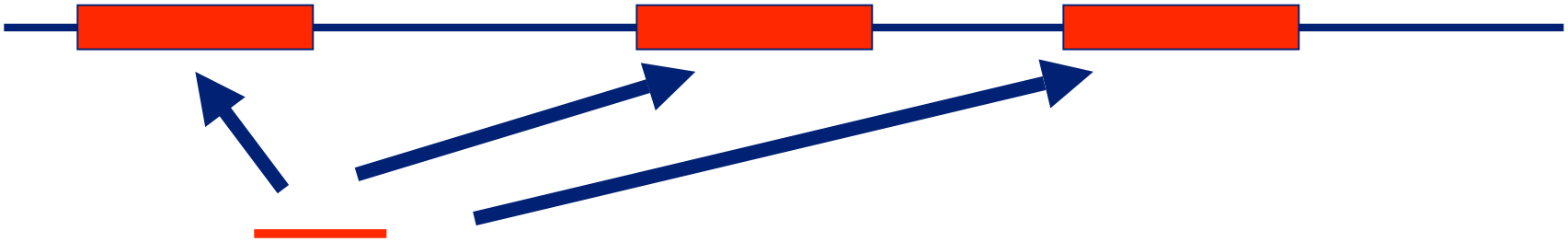
Read Length & Resequencing



Adapted from Jay Shendure

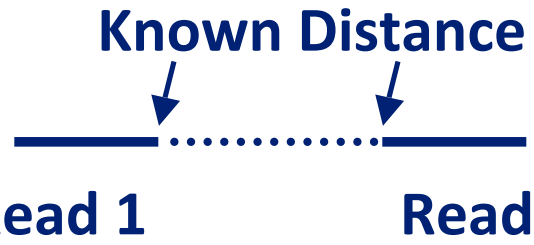
Mapping Unique Reads

Repetitive DNA **Unique DNA**



Single reads map to multiple positions if they hit repetitive DNA

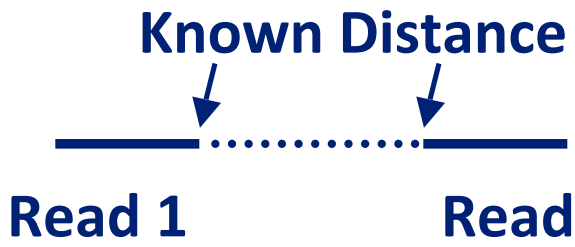
Paired End Reads

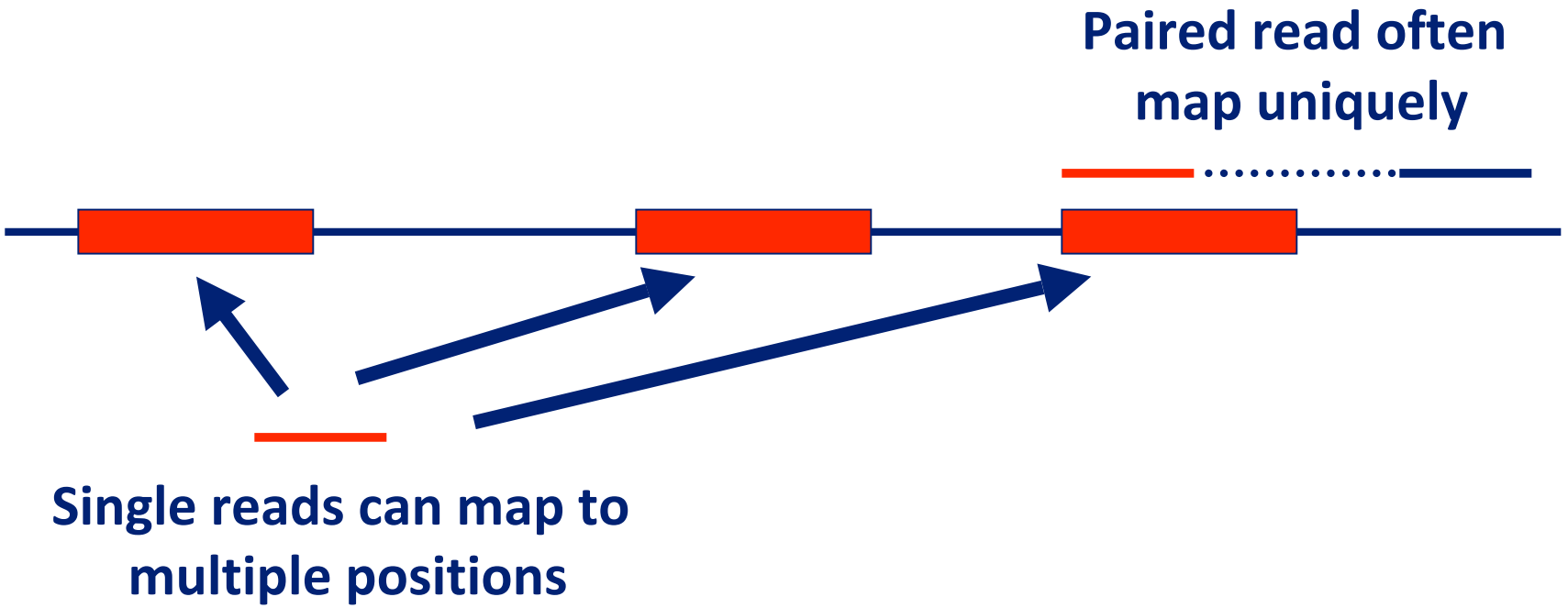


Solexa: paired end is both ends of ~300 bp fragment (shorter than a 454 read, shorter than most TEs)

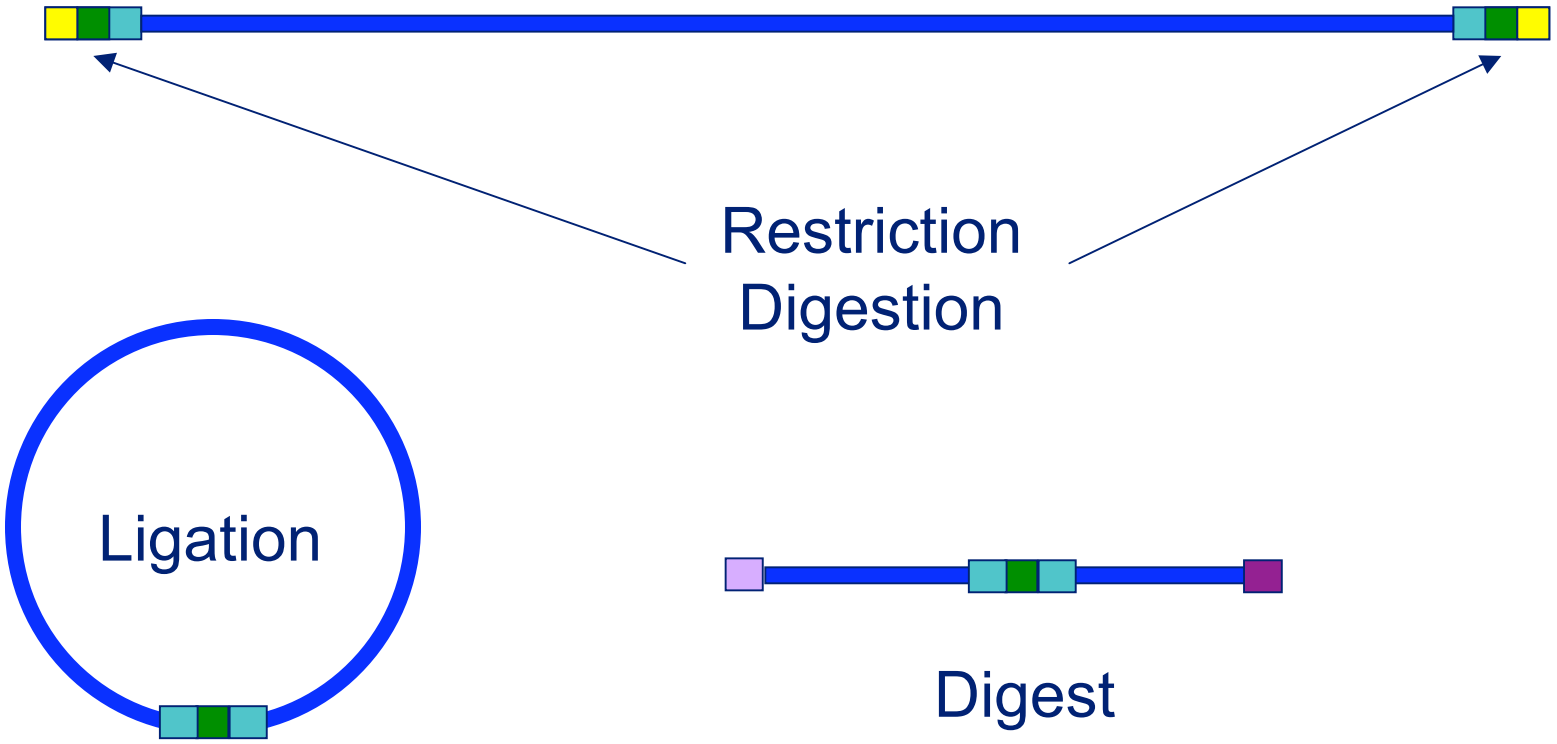
454 paired ends are:
~3Kb
~8Kb
~20Kb

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

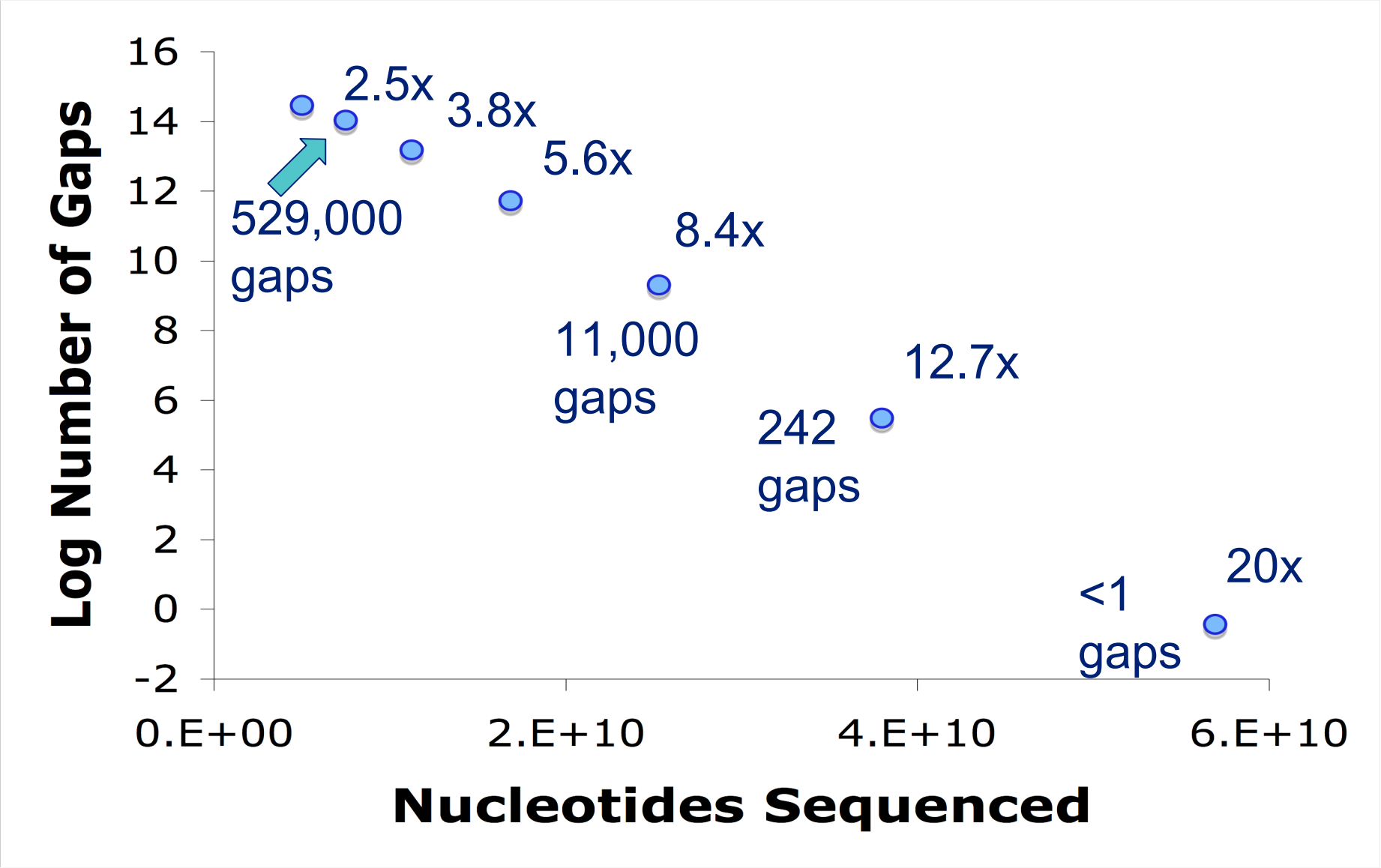
Shotgun Sequencing

- 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