Lander-Waterman Statistics for Shotgun Sequencing

Math 283: Ewens & Grant 5.1

Math 186: Not in book

Prof. Tesler

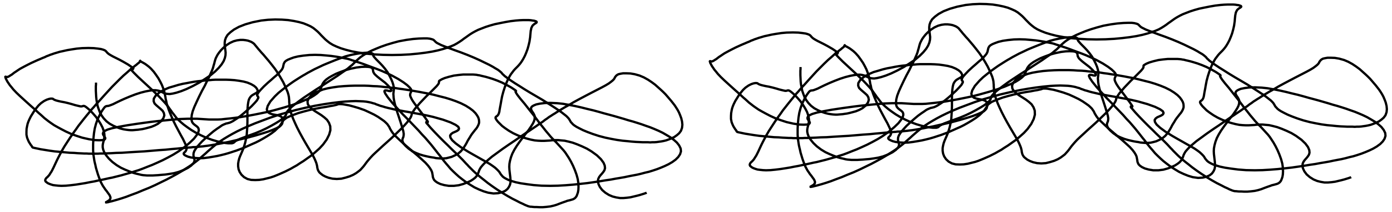
Math 186 & 283
Fall 2019

Genome sequencing

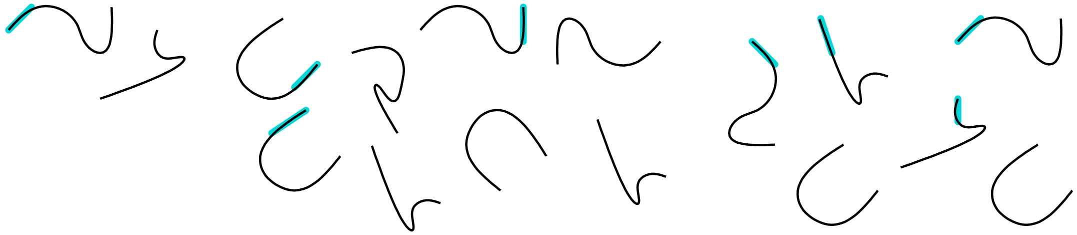
- Frederick Sanger (and others) shared a Nobel Prize in Chemistry in 1980 for developing a method to sequence short regions of DNA.
- Sanger sequencing technology is highly automated and can read approximately 500–1000 consecutive nucleotides from one end of a DNA sequence. If the sequence is larger than that, the rest of it will not be read.
- There is no current technology to simply read the whole genome sequence from one end to the other.
- The human genome is 3 billion nucleotides long. Sequencing it using the Sanger method requires breaking it into little pieces, sequencing the pieces separately, and fitting them back together.

Overview of genome sequencing in the human genome project

Start with many copies of genome



Fragment them and sequence reads



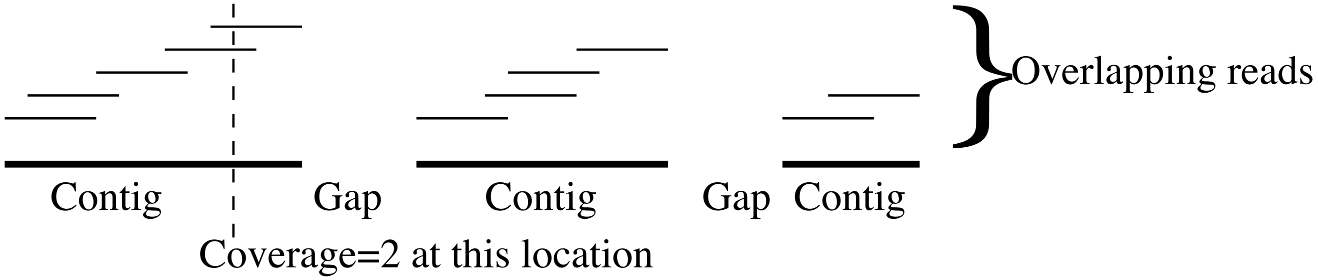
Find overlapping reads

```
             ACGTAGAATCGACCATG...
...AACATAGTTGACGTAGAATC
```

Merge overlapping reads into contig

```
...AACATAGTTGACGTAGAATCGACCATG...
```

Many contigs



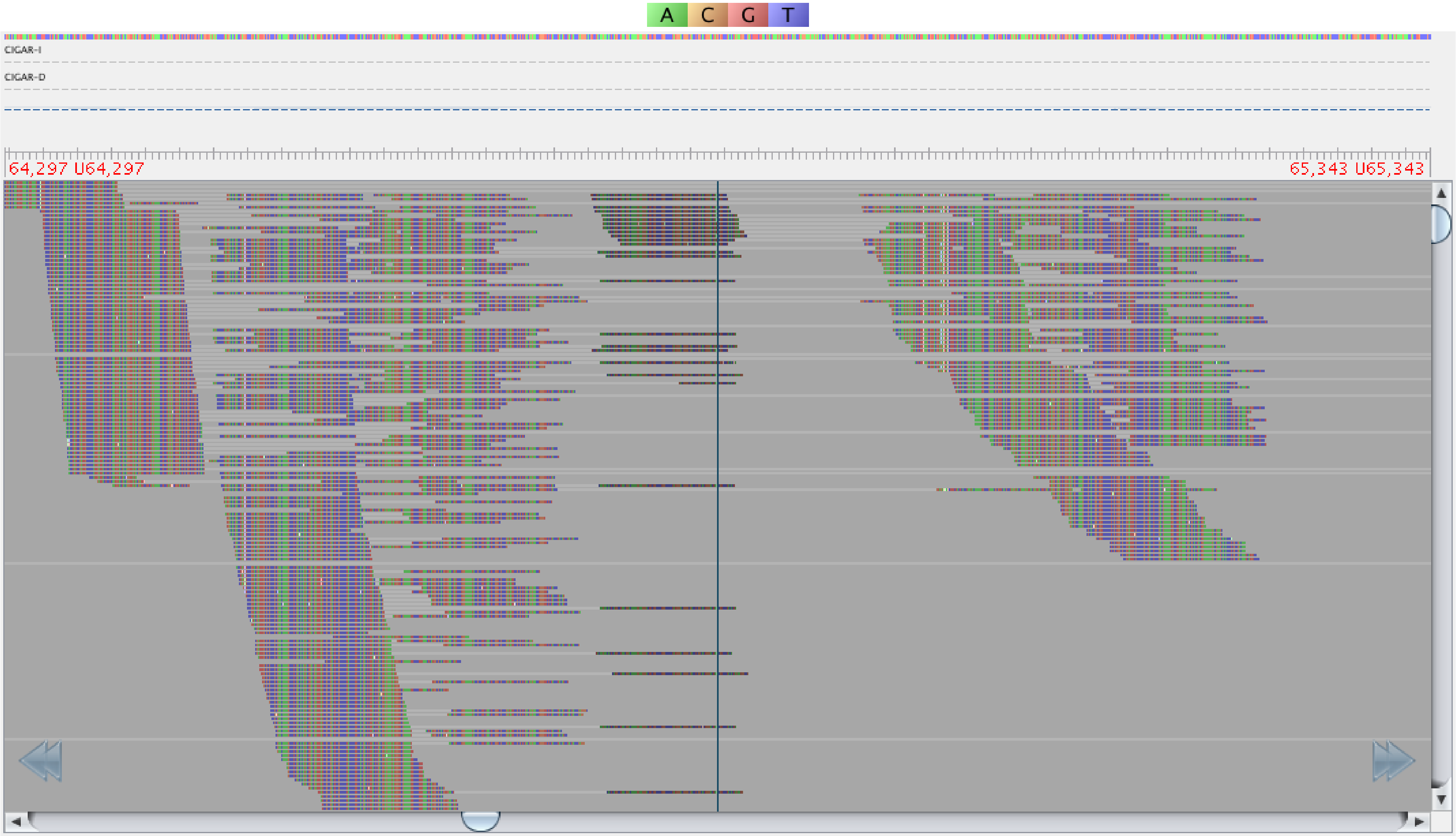
Genome length G

$G \approx 3$ billion

Read length L

$L \approx 500$
(only one end,
only some fragments)

Assembly screenshot



Software: Screenshot of *Tablet* (http://bioinf.scri.ac.uk/tablet/)
Milne et al. (2013) *Briefings in Bioinformatics* 14(2):193–202

Dataset: A single-cell MDA *P. gingivalis* dataset we used in
McLean et al. (2013) *Genome Research* 23(5):867–877

Platform: Illumina GA IIx, 100 bp paired-end reads

Whole genome shotgun sequencing

- Start with a sample with many copies of the same DNA.
- Break it at random into many smaller pieces.
  Randomly select a lot of these pieces to be *sequenced*.
- Approximately the first 500 nucleotides are read from one end of the pieces (the actual number varies from piece to piece).
- These small sequenced regions are called *reads*.
  We do not know their location in the genome or their strand.
- Find overlapping reads using *sequence alignment*, and merge them to form *contigs*.
- Additional information *(scaffolds)* is used to place the contigs into the proper order and direction on chromosomes:
  - *Paired reads* with an approximate known distance apart.
  - Low-resolution *linkage maps*, *optical maps*, . . . , which give approximate positions or spacing of markers over long distances.
- Additonal experiments are needed to sequence the gaps.

BAC-by-BAC sequencing

- Start with many copies of the genome.
- Break them into pieces $\approx$ 100000–300000 nucleotides long and create *Bacterial Artificial Chromosomes* (BACs) from each piece.
- Do shotgun sequencing on each separate BAC; since this is smaller than the whole genome, it's much easier to assemble.
- After assembling BACs, fit overlapping BACs together.

Human genome project

- **Public effort:** Lander et al., *Nature*, Feb. 15, 2001.

  A consortium of government labs and universities.
  Used BAC-by-BAC sequencing.

- **Celera:** Venter et al., *Science*, Feb. 16, 2001.

  Celera is a private company.
  They used whole genome random shotgun sequencing.
  Many people were skeptical that it would work, due to the large coverage required and the large number of repeats that make it impossible to detect overlaps correctly.

- Craig Venter is a UCSD alumnus:
  BS in Biochemistry (1972)
  PhD in Physiology and Pharmacology (1975)

Genome assembly statistics

Parameter names

G = genome length in nucleotides

≈ 3 billion in human

$\approx$ 300000 for a BAC

L = read length in nucleotides (assume 500)

N = number of reads sequenced

NL = number of nucleotides in all sequenced reads

a = NL/G is the *coverage* (average number of times each nucleotide in the whole genome is sequenced)

- $1\times$ (1 times) coverage of the human genome requires $N = aG/L = 1(3 \cdot 10^9)/500 = 6$ million reads.
- $10\times$ coverage requires $N = 60$ million reads.

$1\times$ coverage is not sufficient

- $1\times$ coverage is the minimum for the whole genome to be covered without any gaps, but the reads wouldn't have any overlaps!

```
Genome    ACAGAGTCCAGT
-------------------------------------------
Reads             CAGT
              AGTC
          ACAG
-------------------------------------------
Contigs   ACAG|AGTC|CAGT     3 contigs, not 1
```

- Fully covering the genome with overlaps requires coverage above $1\times$. And since the overlaps are random, it needs to be a lot bigger.

```
Genome    ACAGAGTCCAGT
----------------------
Reads             CAGT
                TCCA
              AGTC
            AGAG
          ACAG
----------------------
Contigs   ACAGAGTCCAGT
```

Genome assembly statistics – Questions

Assume reads are distributed uniformly through the genome.
In terms of the number of reads N or the coverage $a = NL/G$, estimate

- How many contigs are there?
- How big are the contigs?
- How many reads are in each contig?
- How big are the gaps?

After doing shotgun sequencing (with reads drawn at random), additional experiments targeted at the gaps are performed.

Probability some read starts at a position x

- In each chromosome, a read of length L could start anywhere except the last $L-1$ positions.

- In a genome of length G with c chromosomes, there are $G - c \cdot (L-1)$ possible starting positions.

- For human, $c \cdot (L-1) = 23(499) = 11477 \ll G$ so we will approximate that there are G possible starting positions. (That is, we will ignore the end effects.)

- The probability that one of the N reads starts at any specific nucleotide is N/G.

Probability some read hits an interval

Let I be any specific interval of L consecutive nucleotides.
What is the probability that at least one read starts in I?

Binomial distribution

$$p = P(\text{no read starts in } I) = (1 - N/G)^L$$
$$q = P(\geqslant 1 \text{ reads start in } I) = 1 - (1 - N/G)^L$$

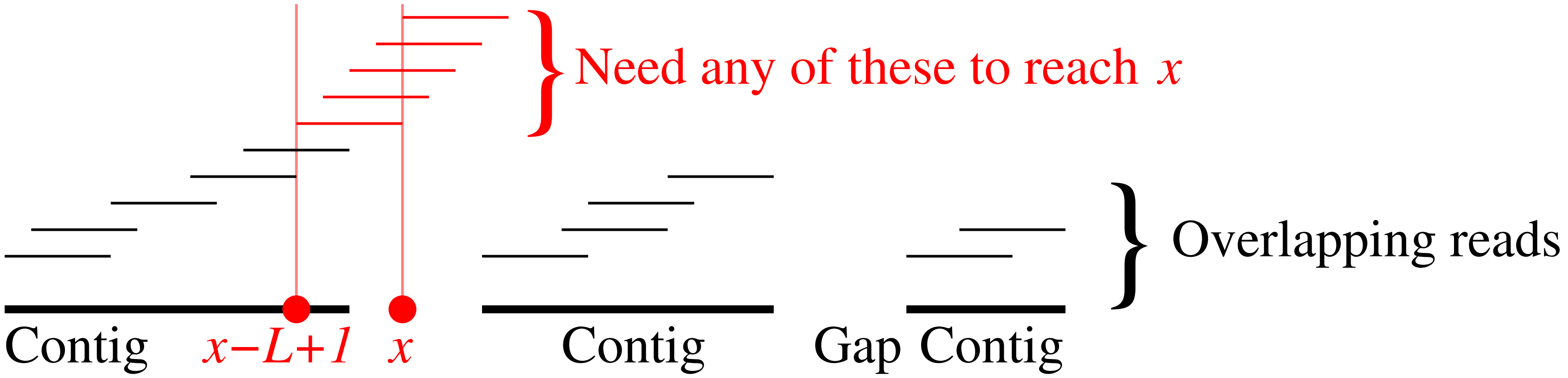
Poisson distribution

The expected # reads starting within I is $(N/G) \cdot L = a$ (the coverage).

$$p = P(\text{no read starts in } I) = e^{-a}\frac{a^0}{0!} = e^{-a}$$
$$q = P(\geqslant 1 \text{ reads start in } I) = 1 - e^{-a}$$

- Poisson is more accurate in the high-coverage cases when it's likely there are multiple read-starts at the same position.

Is position x in a contig or in a gap?



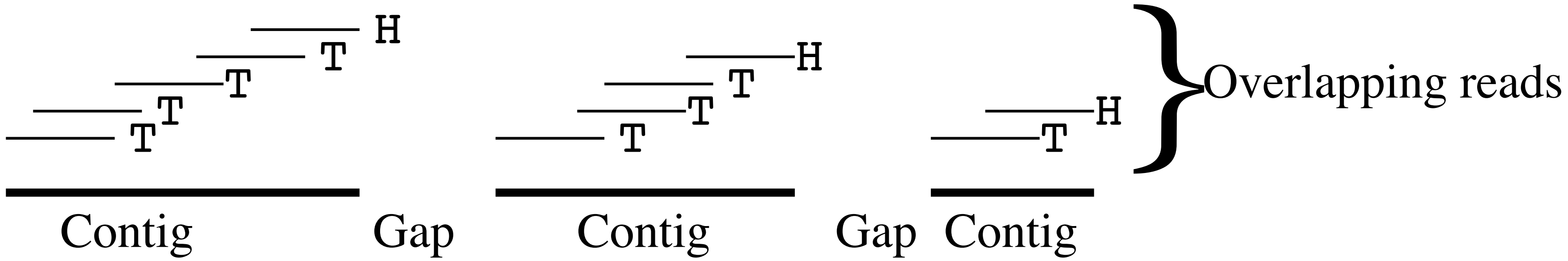
If any read starts in $[x - L + 1, x]$ (in red above), then x is in a contig; otherwise, x is in a gap.

How much of the genome was sequenced?

- Position x is in a gap if no read starts in $[x-L+1,x]$.
  This is an interval of length L.
- We showed this has probability $p=e^{-a}$.
- Estimate:
  Nucleotides in gaps: $pG=e^{-a}G$
  Nucleotides in contigs: $qG=(1-e^{-a})G$

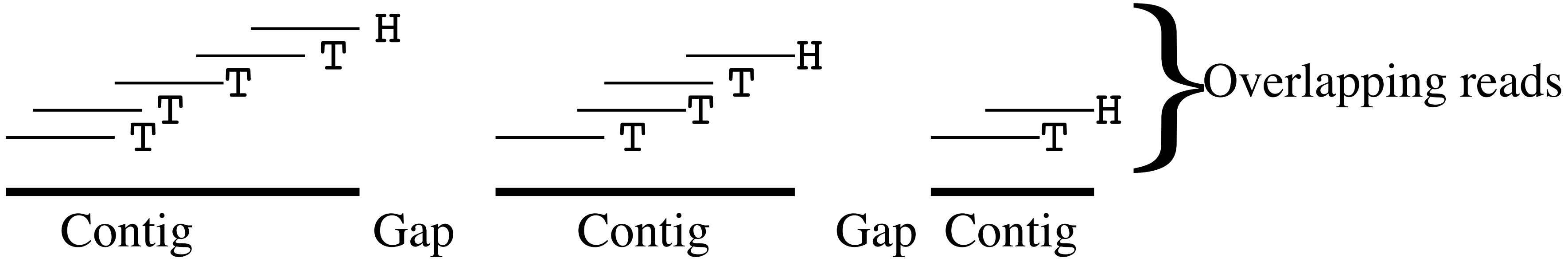
- To have 99% of the genome in contigs and 1% in gaps:
  Fraction in contigs: $q=1-e^{-a}=0.99$
  Fraction in gaps: $p=e^{-a}=0.01$
  Coverage: $a=-\ln(0.01)\approx 4.6$
- This is staggering — for the human genome, if the sequenced reads contain 4.6×3 billion = 13.8 billion nucleotides, you still expect to miss about 30 million (1% of genome size) positions within the genome.

How many contigs are formed?



- Each contig has a unique rightmost read.
- The probability that a read is rightmost is the same as the probability that no other read starts within that read,
$$\exp(-N(L-1)/G) \approx p = \exp(-a) \ .$$
- Label the rightmost reads “heads” (H) and the others “tails” (T). The number of contigs is the number of heads, so it has a binomial distribution with parameters N and p.
- **Expected number of contigs:**
$$Np = Ne^{-a} = Ne^{-NL/G} = (aG/L)e^{-a}$$

How many reads per contig?



- With reads labelled “heads” and “tails,” the number of reads in the first contig is the same as the position of the first heads; i.e., the geometric distribution.
- The expected number of reads per contig is

$$1/p = e^a$$

- We can also deduce this as the number of reads divided by expected number of contigs:

$$N/(Ne^{-a}) = e^a$$

How long are the contigs?

- The expected size of the sequenced region is

$$(1-e^{-a})G$$

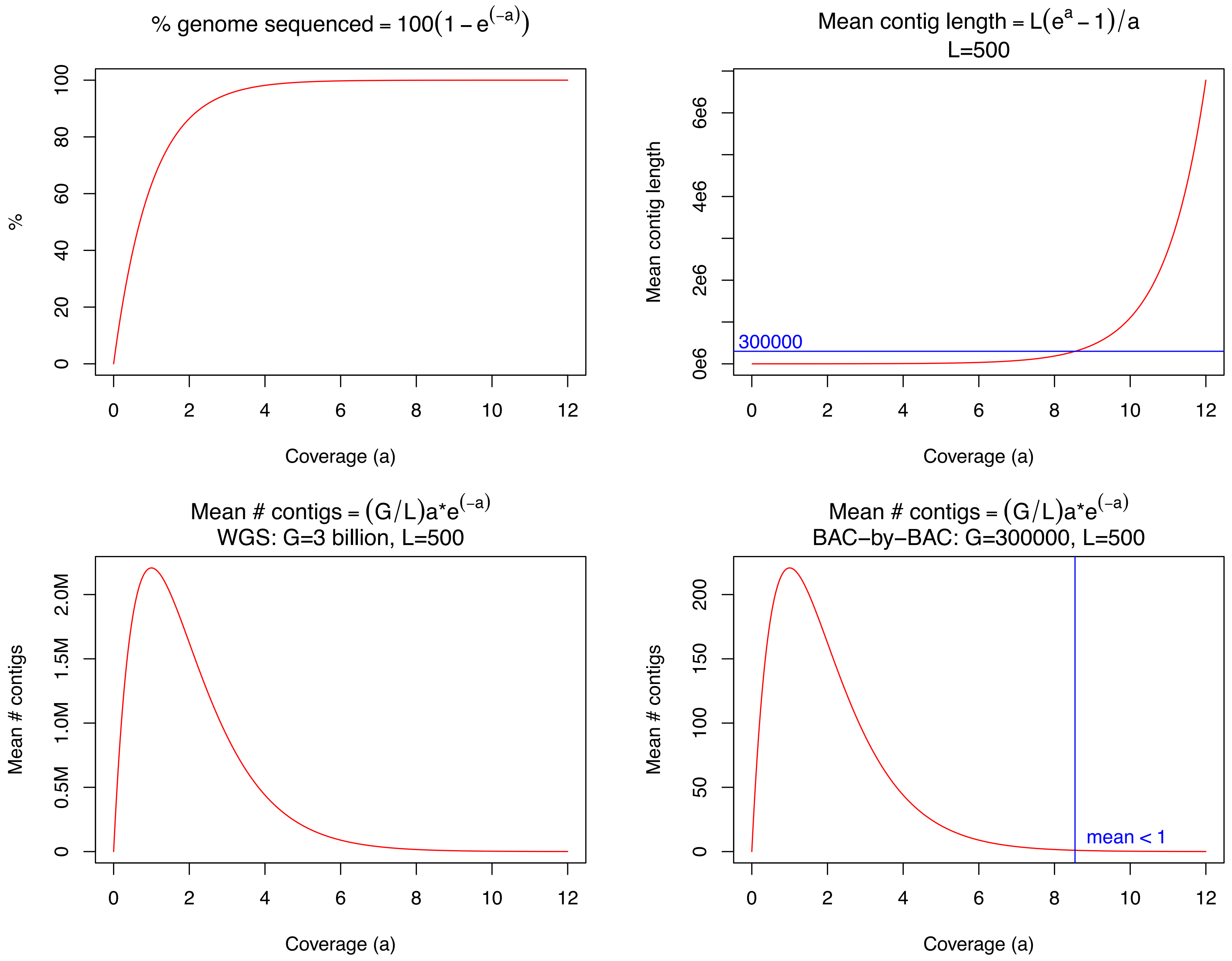
- The expected number of contigs is

$$Ne^{-a} = (aG/L)e^{-a}$$

- The mean contig size is the ratio of those:

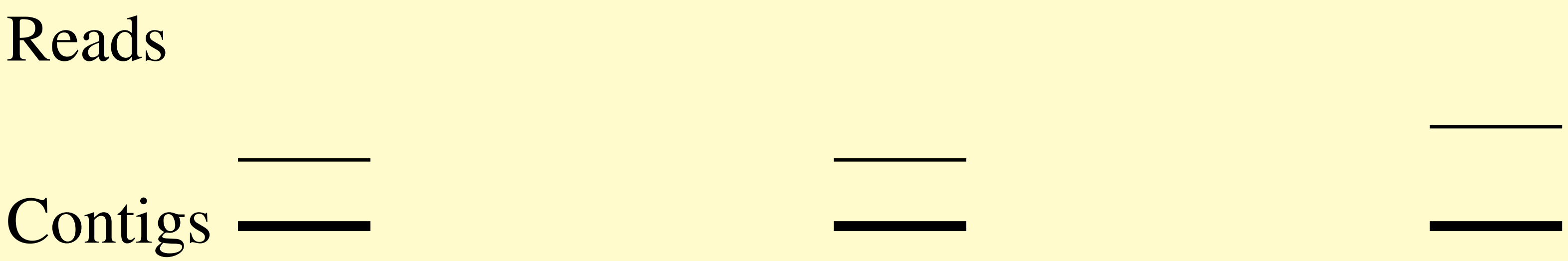
$$\frac{(1-e^{-a})G}{Ne^{-a}} = \frac{(e^{a}-1)G}{N} = \frac{(e^{a}-1)L}{a}$$

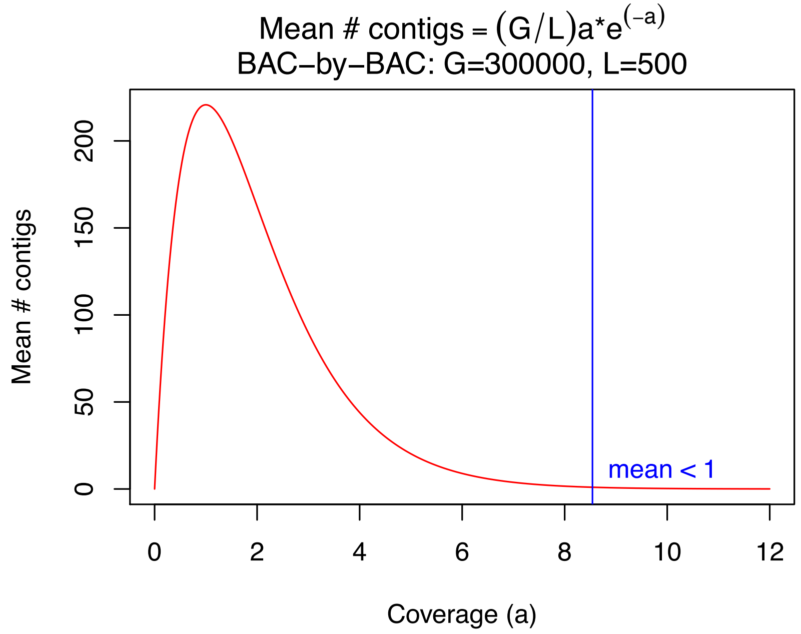
Plots of estimates for human genome sequencing



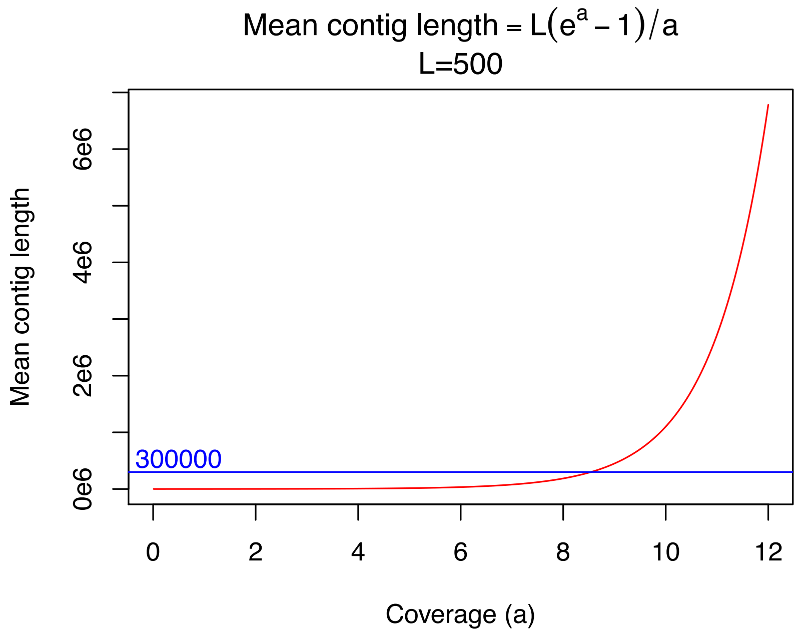
Assembly progression

As we process reads, initially, the reads are scattered through the genome, each forming its own small contig.




Mean # contigs = $(G/L)a^*e^{(-a)}$

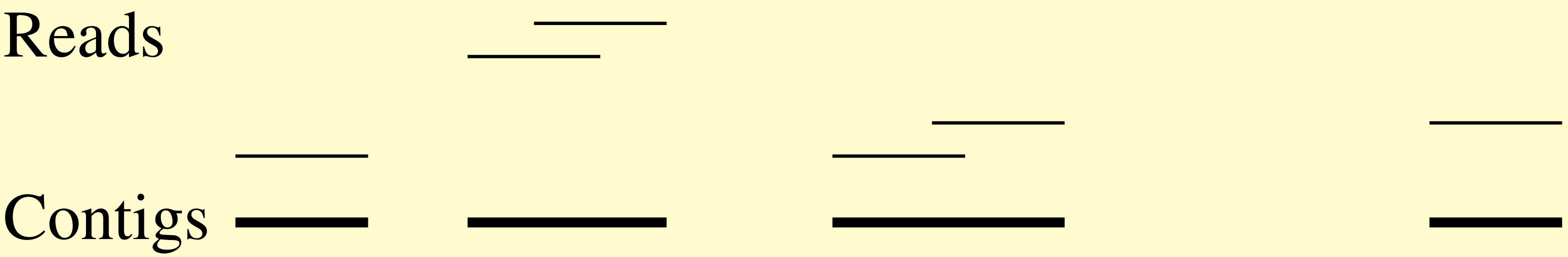
BAC–by–BAC: G=300000, L=500

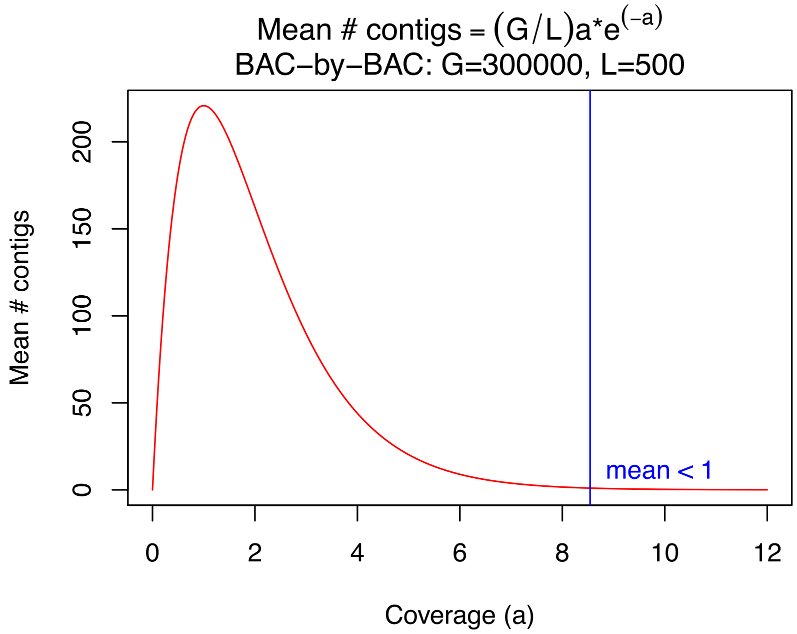

Mean contig length = $L(e^a - 1)/a$

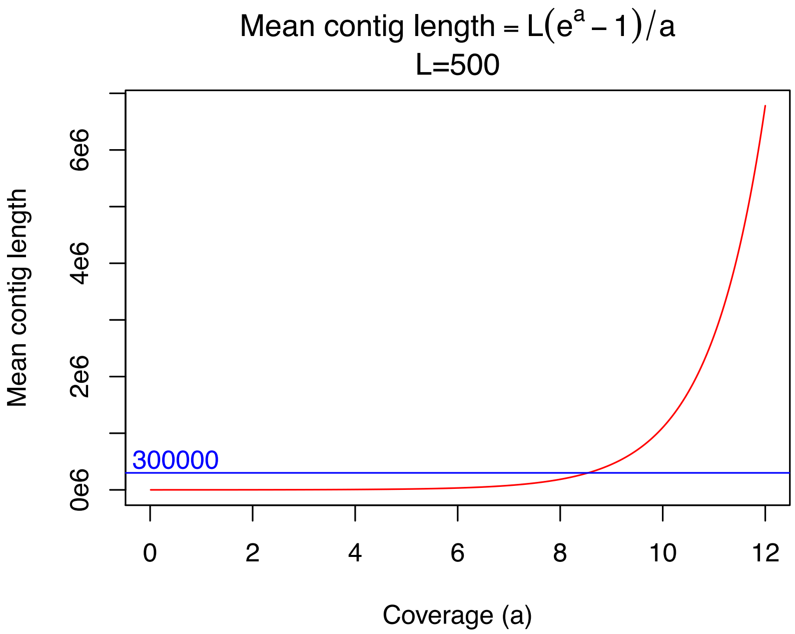
L=500

Assembly progression

As coverage increases, some reads may overlap, so the # of contigs will not increase as much, and the contig size will start to increase.

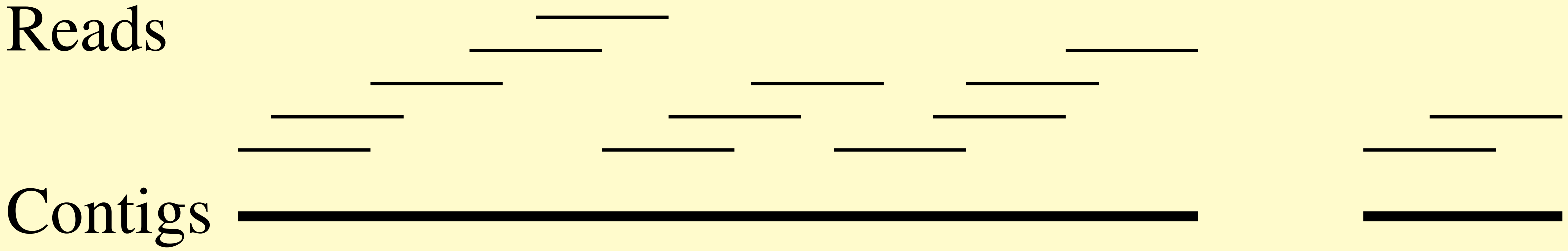


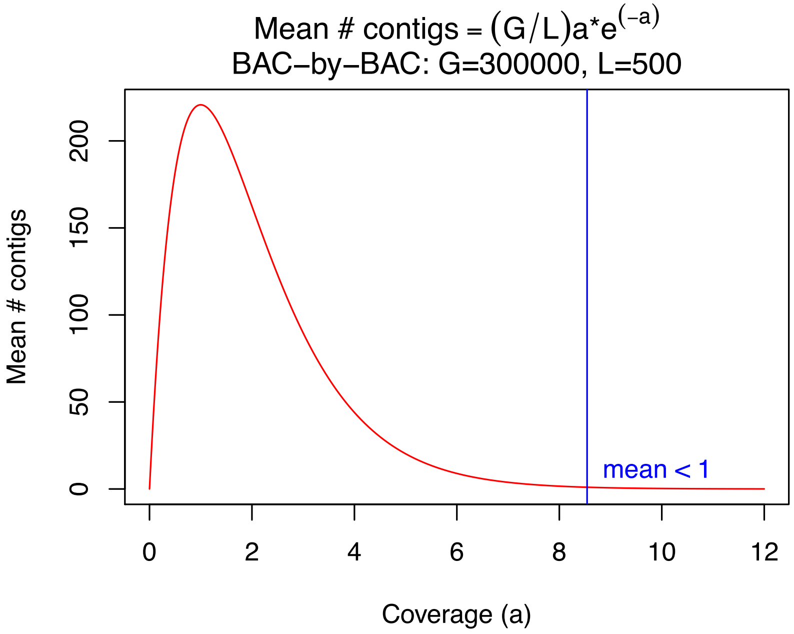

Mean # contigs = $(G/L)a{*}e^{(-a)}$
BAC–by–BAC: G=300000, L=500

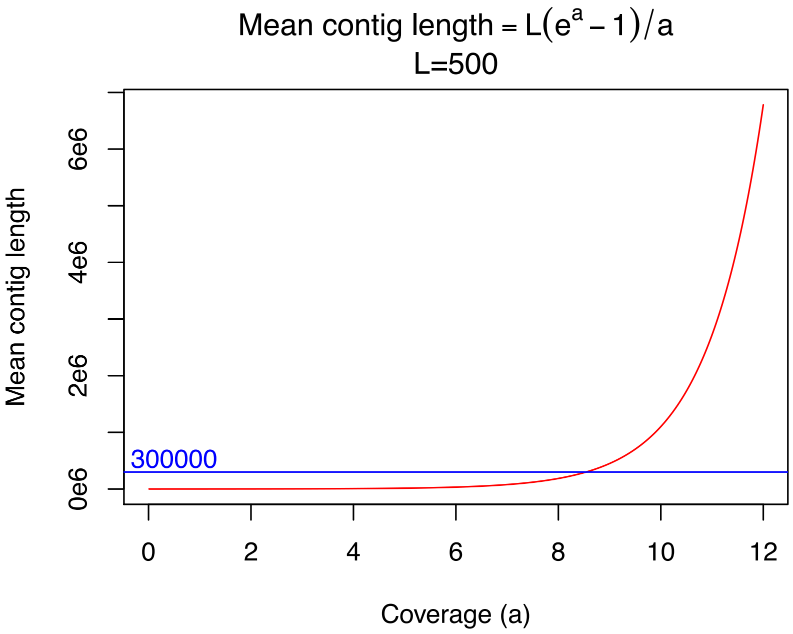

Mean contig length = $L(e^a - 1)/a$
L=500

Assembly progression

Eventually reads are so dense that contigs merge together, so the # contigs decreases while the size increases a lot.

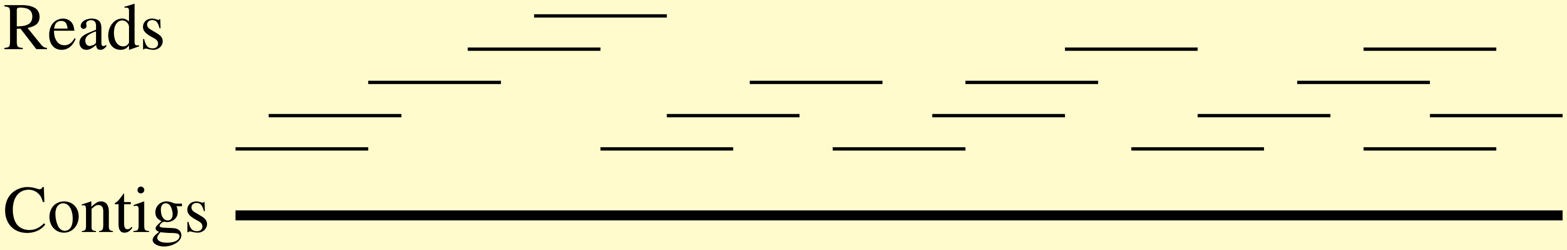


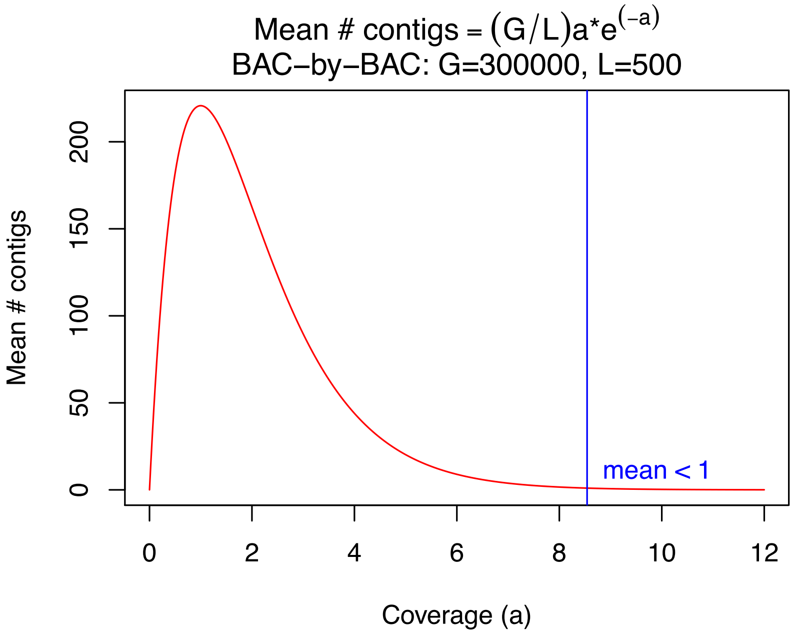

Mean # contigs = $(G/L)a*e^{(-a)}$
BAC–by–BAC: G=300000, L=500

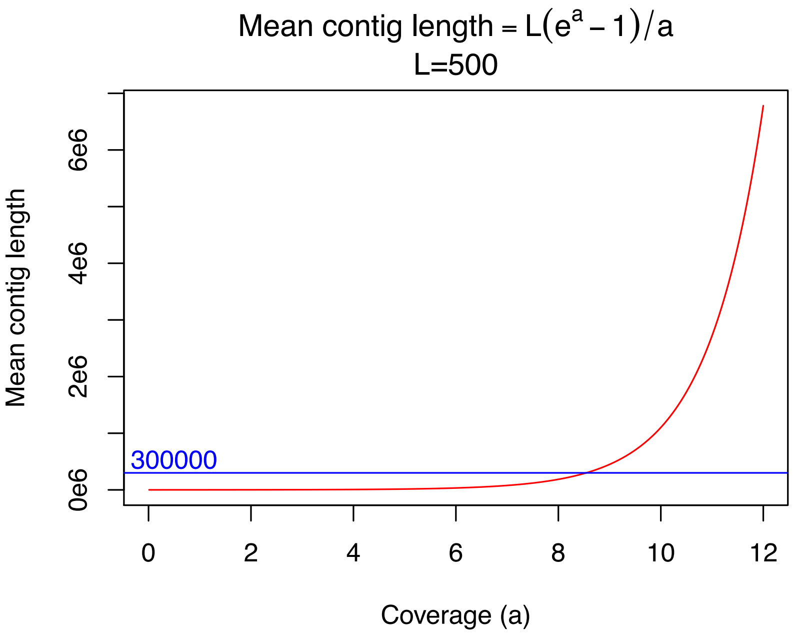

Mean contig length = $L(e^{a} - 1)/a$
L=500

Assembly progression

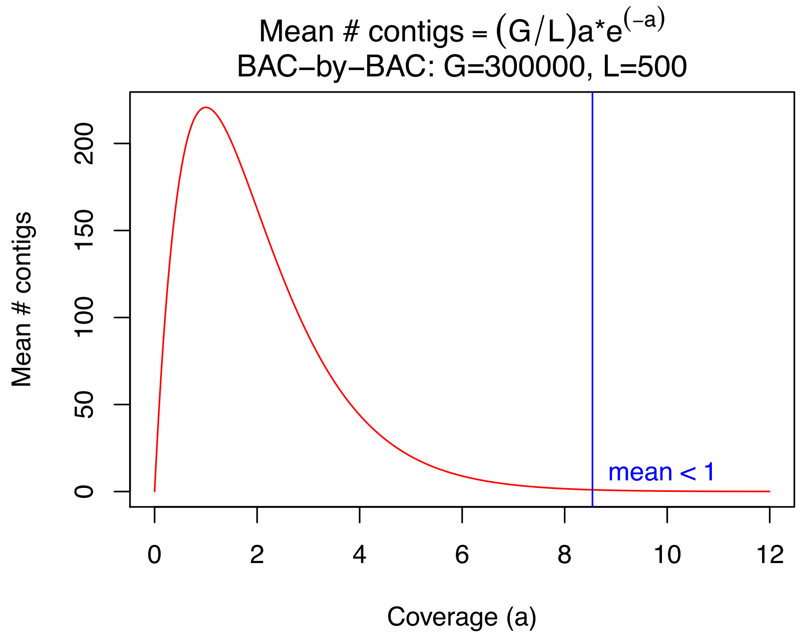
We want a small number of contigs, with the contigs very large. Ideally, one contig per chromosome, covering the whole chromosome.

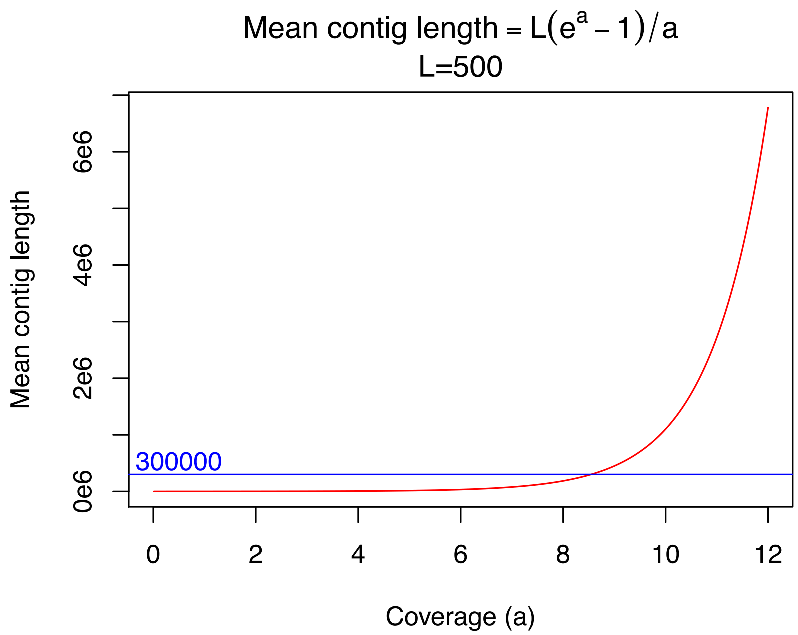



Mean # contigs = $(G/L)a^*e^{(-a)}$
BAC–by–BAC: G=300000, L=500


Mean contig length = $L(e^a - 1)/a$
L=500

Garbage in, garbage out


Mean # contigs $= (G/L)a*e^{(-a)}$
BAC–by–BAC: G=300000, L=500

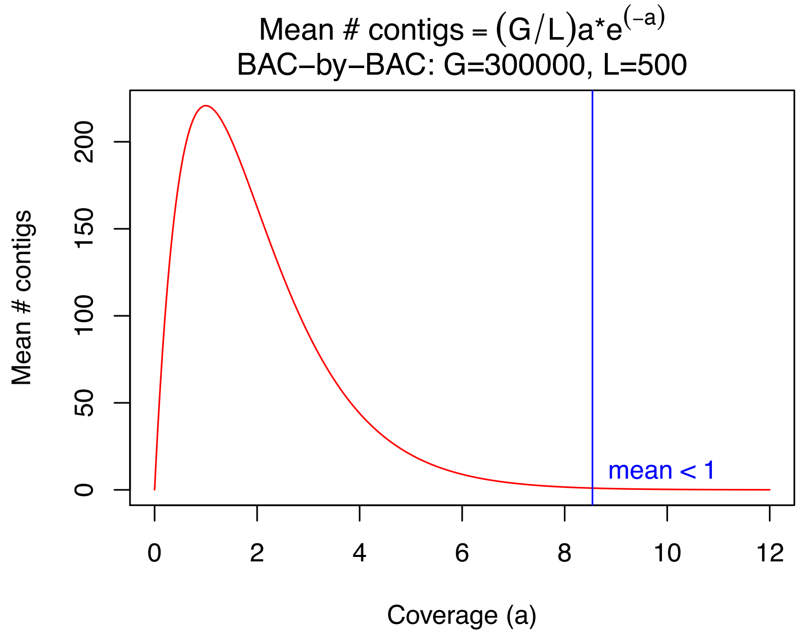

Mean contig length $= L(e^a - 1)/a$
L=500

At high coverage, some values are nonsense:

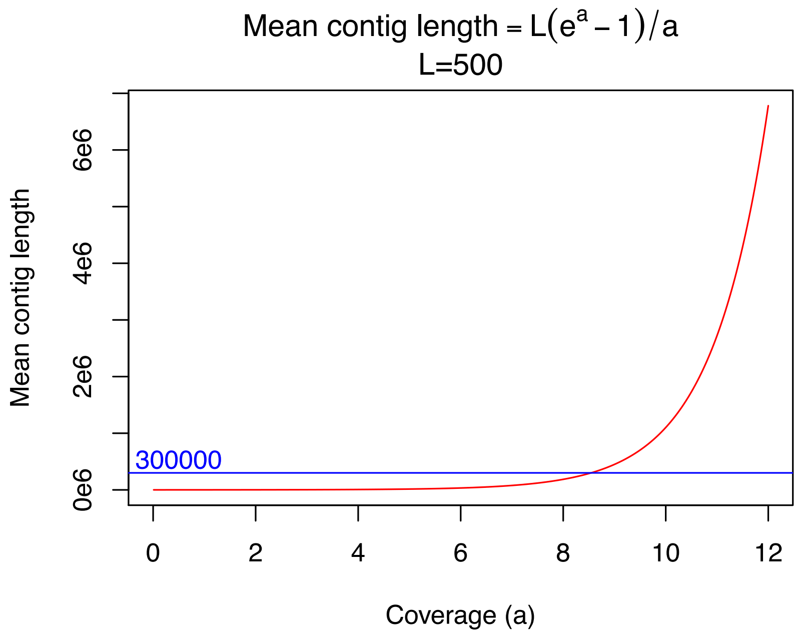
- contig length larger than the BAC
- mean number of contigs below 1.

The approximations are clearly invalid there. But there aren't guidelines on exactly where they are valid.

Garbage in, garbage out



Mean # contigs = $(G/L)a*e^{(-a)}$
BAC–by–BAC: G=300000, L=500

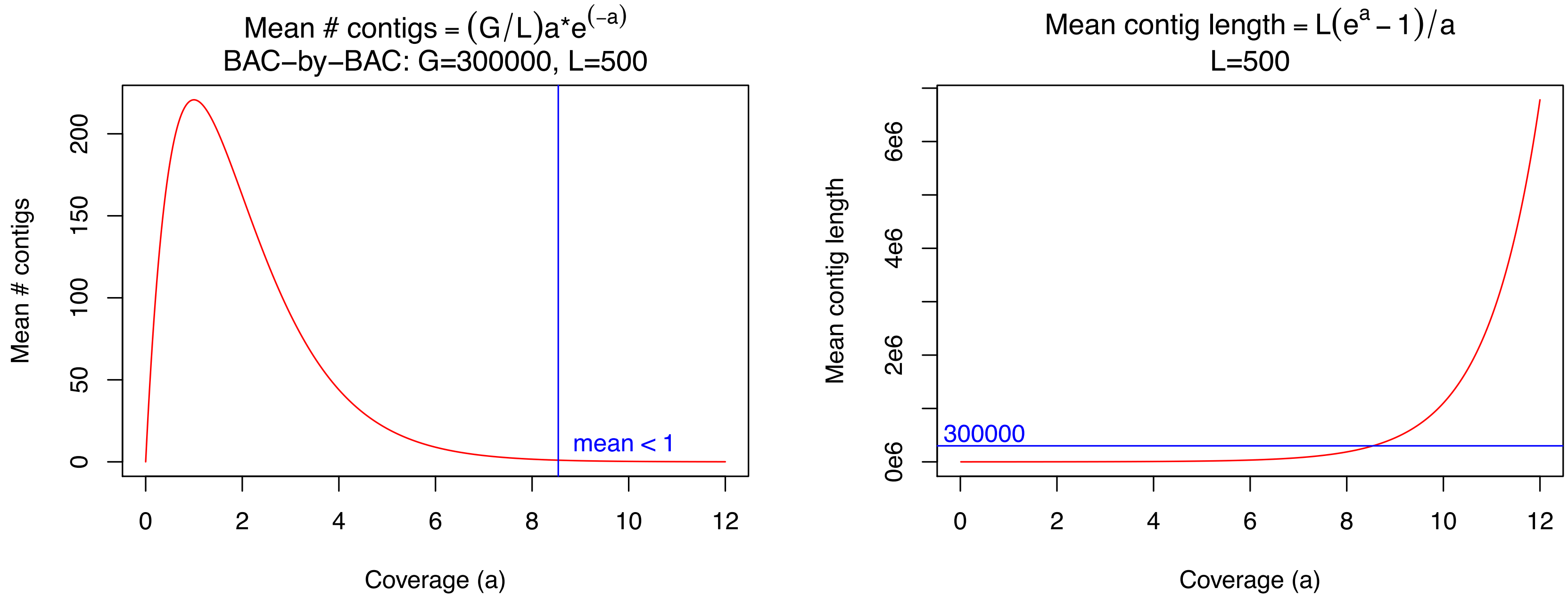


Mean contig length = $L(e^a - 1)/a$
L=500

Even when the values appear to be valid, beware: this is a statistical model about average values, not a physical law.

- In a physical law ($E = mc^2$, $F = ma$, $PV = nRT$, etc.), you expect the formulas to be obeyed, within measurement errors.
- But the Lander-Waterman statistics are only rough estimates; actual values in an assembly dataset will likely deviate.

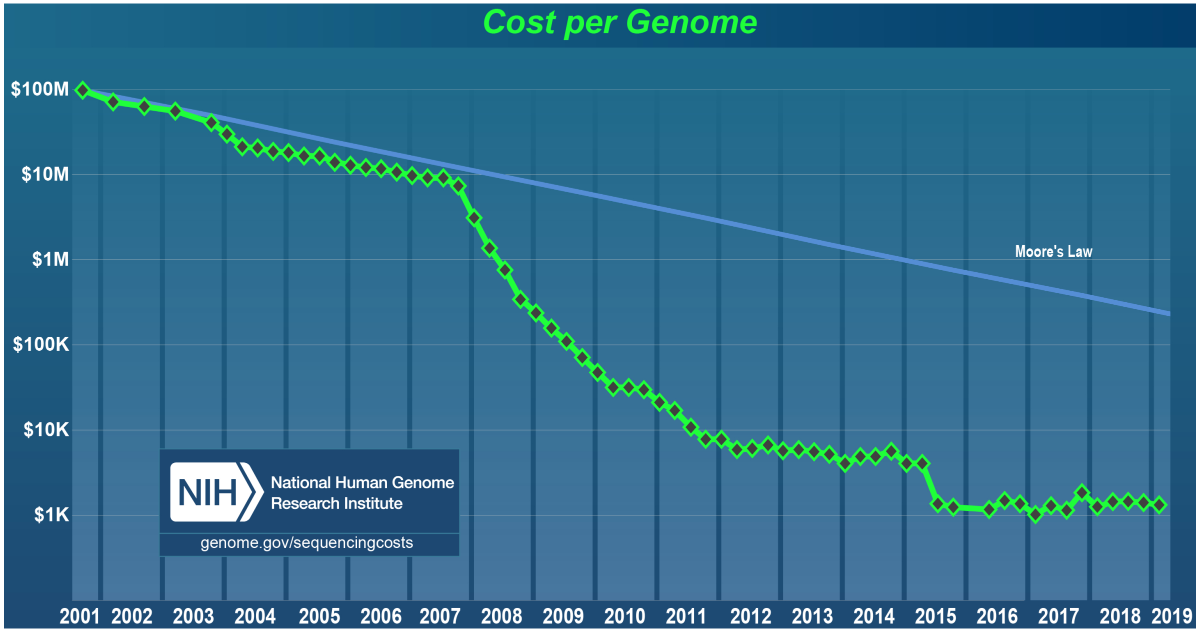
Garbage in, garbage out



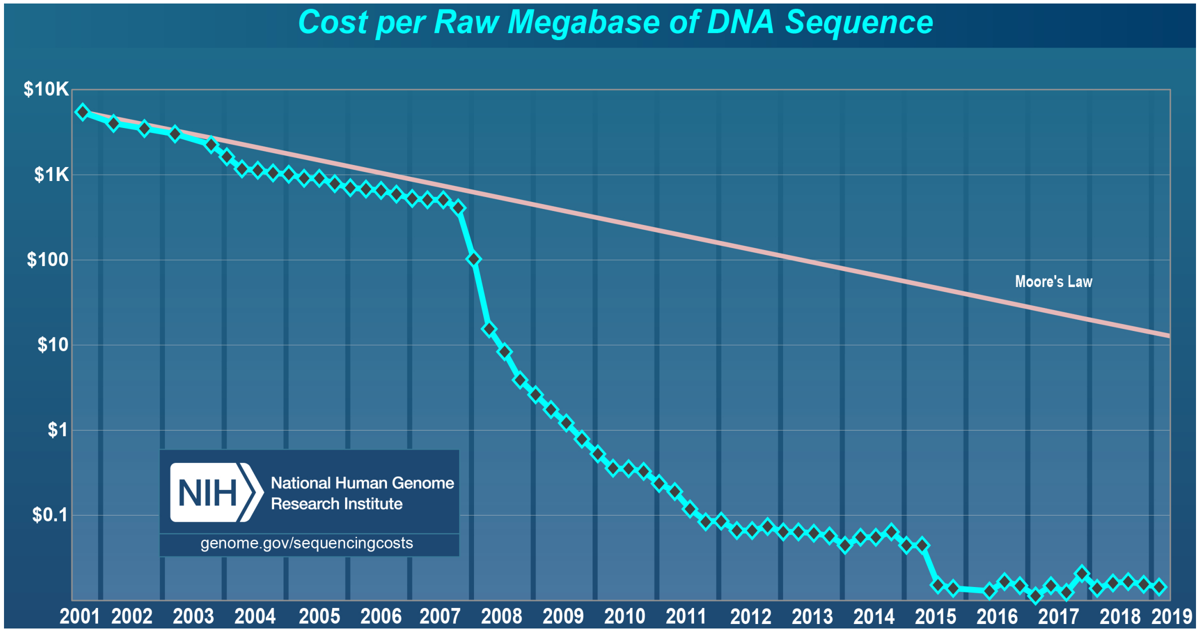
- We need higher coverage than the model suggests due to issues this model doesn't consider, such as repeats, read errors, and contamination.

Sequencing costs are dropping

Source: NIH/NHGRI website, http://www.genome.gov/sequencingcosts/



Cost per Genome



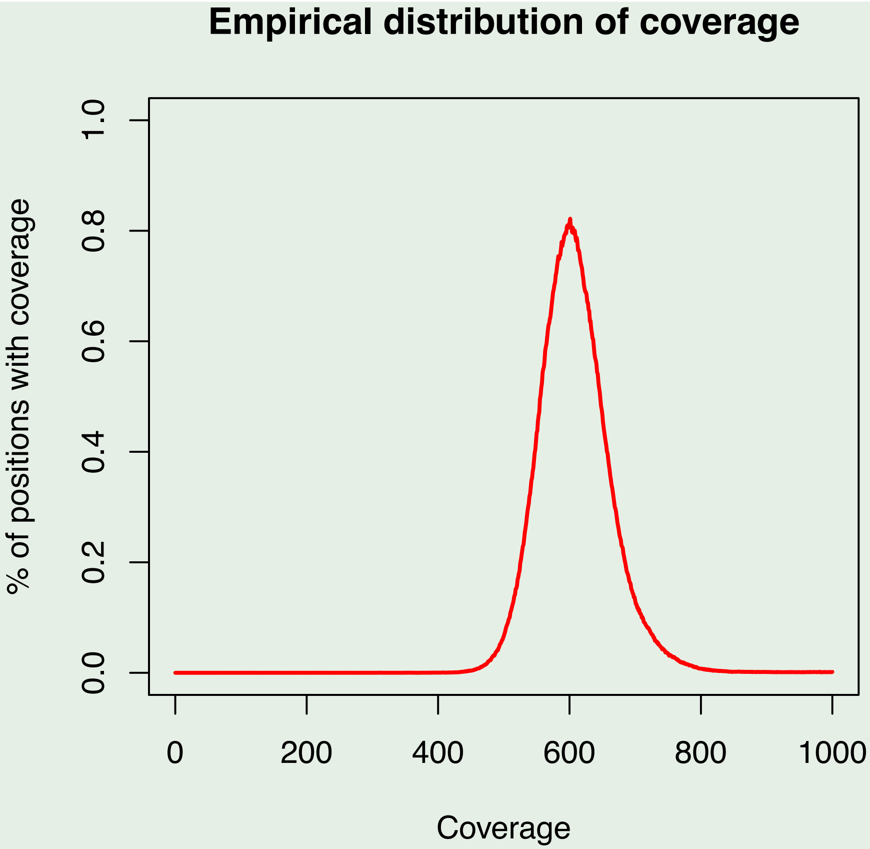
Cost per Raw Megabase of DNA Sequence

Bacterial sequencing

	E. coli	Human
Genome length	4.7 Mbp (million base pairs)	3.4 Gbp (billion base pairs)
Ploidy	haploid	diploid
Coverage of one lane (Illumina 100 base mate-paired reads)	$600\times$	$0.8\times$

The *E. coli* coverage plot resembles the normal distribution ("bell curve").

Chitsaz et al (2011), *Nat. Biotech.*



Complications — Paired reads

- **Double-barreled reads in Human Genome Project:**
  Select fragments with similar lengths.
  Read 500 nucleotides apiece from both ends of a fragment, and also estimate the fragment length.
  This gives correlated reads, which may allow positioning, orienting, or combining their contigs.
  It effectively increases read length.

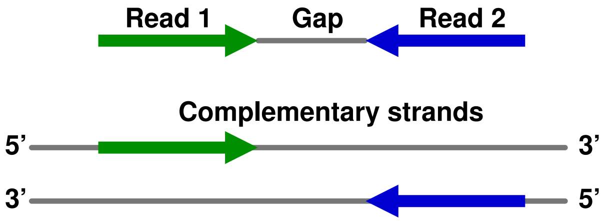
- **Illumina machines:**
  35–250 nt from both ends of each fragment.
  Fragments may be a few hundred bases long (“paired-end protocol”) or thousands long (“mate-pair protocol”).

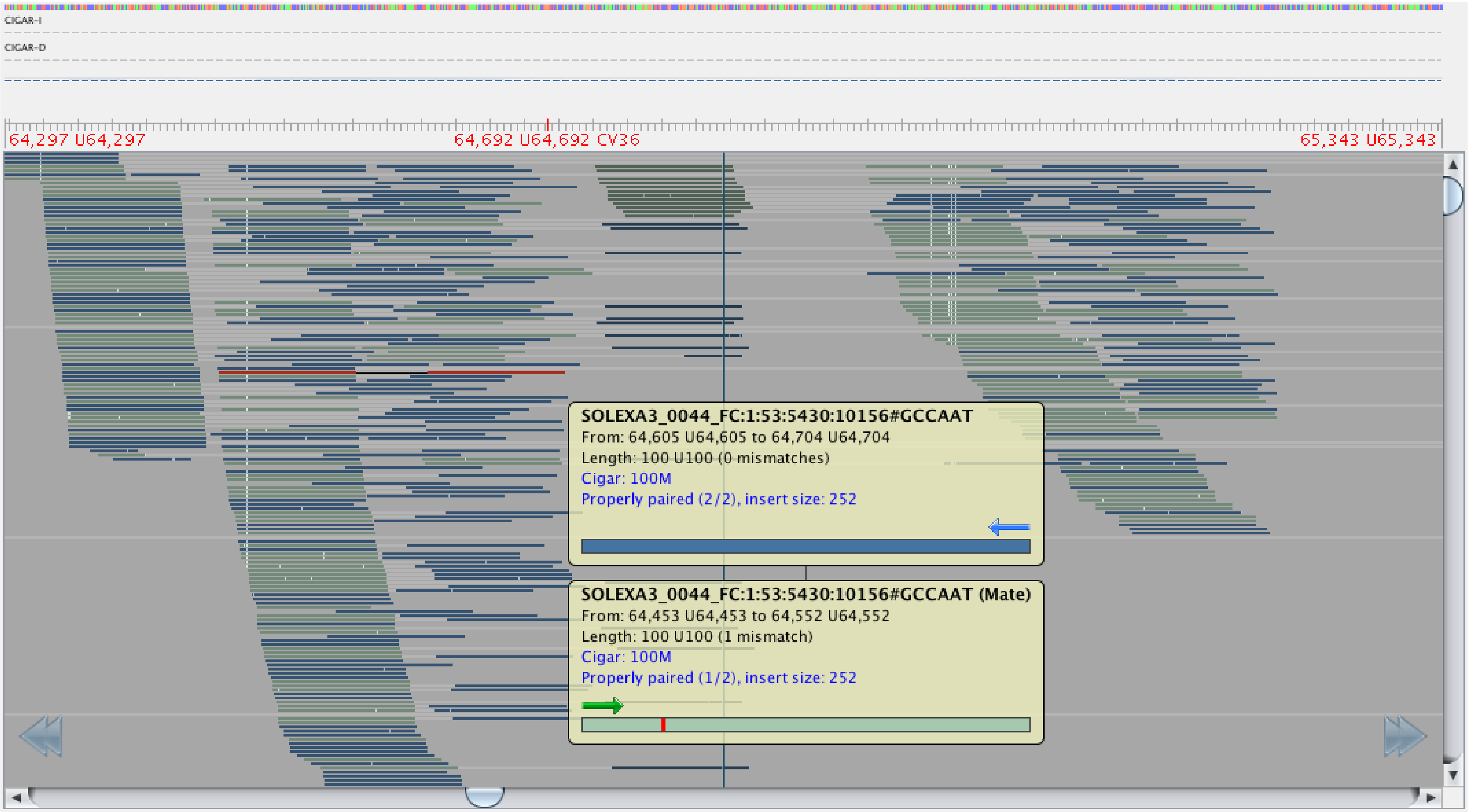
Complications — DNA is double-stranded

- **Double-stranded reads:**
  Both strands contribute to the reads. When a fragment is sequenced, it could be the first 500 nucleotides from either strand. In order to fit together the reads that could come from either strand, the assembly software has to consider every read and its reverse complement.

Complications — Paired reads and double-stranded DNA

Screenshot of Tablet on *P. gingivalis* paired end reads





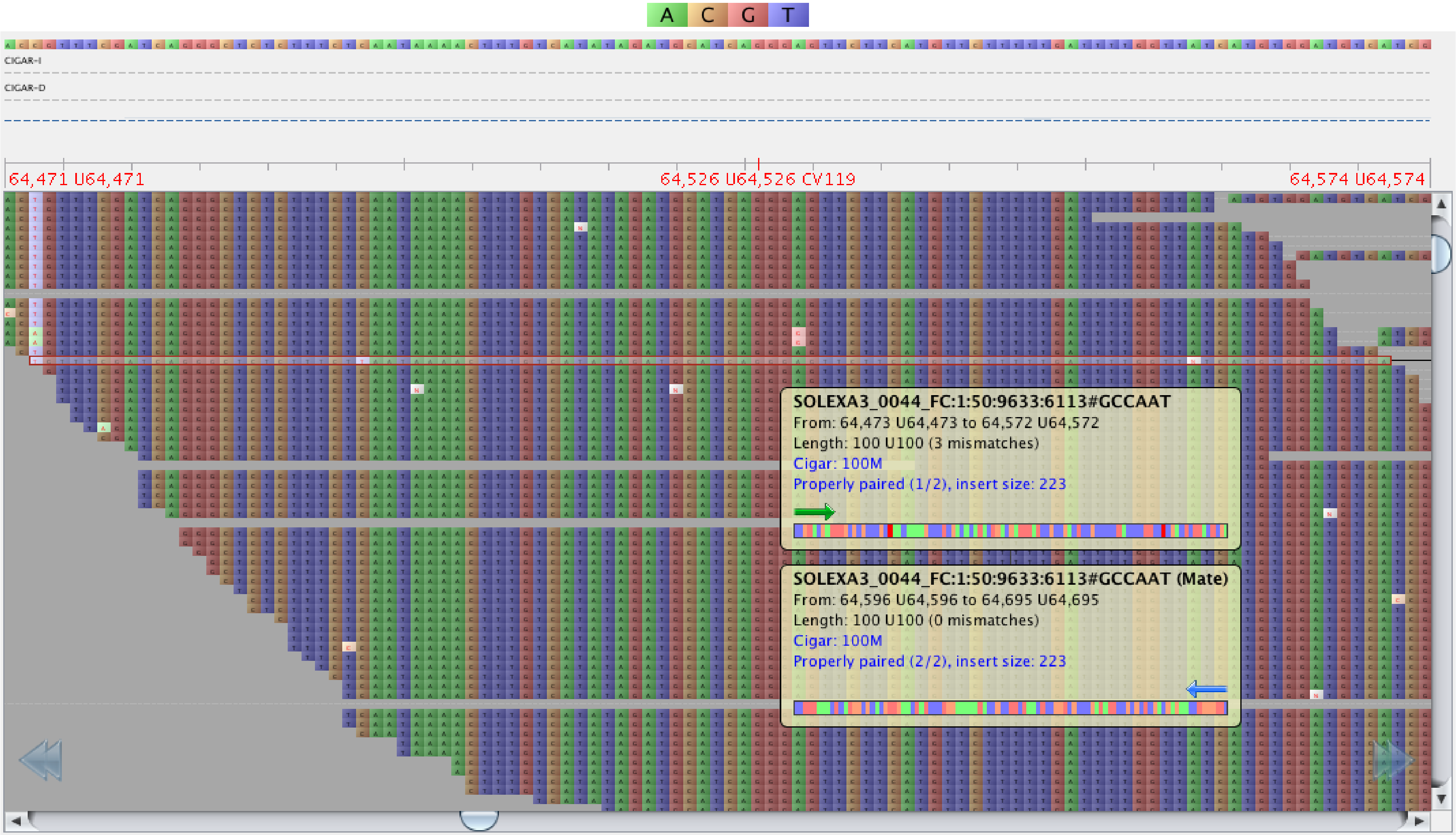
SOLEXA3_0044_FC:1:53:5430:10156#GCCAAT
From: 64,605 U64,605 to 64,704 U64,704
Length: 100 U100 (0 mismatches)
Cigar: 100M
Properly paired (2/2), insert size: 252

SOLEXA3_0044_FC:1:53:5430:10156#GCCAAT (Mate)
From: 64,453 U64,453 to 64,552 U64,552
Length: 100 U100 (1 mismatch)
Cigar: 100M
Properly paired (1/2), insert size: 252

Complications — Read errors

Screenshot of Tablet on *P. gingivalis* mate pairs

- Bases that don’t match the reference genome are highlighted in light colors.
- We sequenced a new strain. Some differences from the reference are due to true differences in the new strain (light colored third column), and others are due to read errors (isolated light spots).



SOLEXA3_0044_FC:1:50:9633:6113#GCCAAT
From: 64,473 U64,473 to 64,572 U64,572
Length: 100 U100 (3 mismatches)
Cigar: 100M
Properly paired (1/2), insert size: 223

SOLEXA3_0044_FC:1:50:9633:6113#GCCAAT (Mate)
From: 64,596 U64,596 to 64,695 U64,695
Length: 100 U100 (0 mismatches)
Cigar: 100M
Properly paired (2/2), insert size: 223

Complications — Repeats

```
Reads            IJKL      IJOP
               GHIJ      GHIJ
             EFGH      EFGH
           CDEF      MNEF
---------------------------------
Genome   ABCDEFGHIJKLMNEFGHIJOPQR
```

- Reads from different parts of the genome have identical sequences and appear to overlap. E.g., the 1$^{\text{st}}$ EFGH and 2$^{\text{nd}}$ GHIJ.
- These reads could also be assembled into

  CDEFGHIJOP and MNEFGHIJKL

  or other incorrect results.
- Recall that you're given the reads as strings, without knowing their positions. There's insufficient information to resolve it.
- When the repeat length exceeds the read length ($6 > 4$ here), more info is needed to match up the regions left/right of each repeat copy.

Complications — Read errors and repeats

- **Read errors:**
  All platforms have errors in the reads. In Sanger sequencing, about 1% of the nucleotides will be misread, so reads from overlapping regions of the genome may have differences. Thus, it's necessary to allow for approximate matches instead of just exact matches. In PacBio sequencing, it's up to 15% error rate!

- **Diploid samples:**
  In a diploid sample, the two sets of chromosomes have slight differences. They each occur in about half of the reads covering a certain location, and appear similar to a repeat of multiplicity two.

- **Repeats:**
  Over generations, repeats that started identical may acquire mutations. If reads have a small number of differences, assemblers have to decide whether these are due to read errors, repeats, diploid variations, etc.

Complications — Mixed samples

Complications arising from mixed samples

- Some experimental samples have cells with different genomes.
- Differences in corresponding parts of the genome (substitutions, deletions, insertions, variable number tandem repeats, rearrangements, etc.) make it difficult to detect overlaps, and once detected, difficult to resolve the “consensus” sequence.
- At the same time, homologous regions in the different genomes may have overlaps, leading the assembler to put them together.

Some experiments with mixed samples

- Celera’s Human Genome assembly: hybrid of 5 people’s DNA.
- Metagenomics: samples from the environment, animal guts, etc., consist of many different bacteria.
- Cancer tumor sequencing.
- HIV has $\approx 10\%$ mutation rate per generation, leading to a population of different HIV genomes in an HIV sample.

Complications — Read length, overlap length

- **Read length:**
  The read length, L, varies, so it should be treated as a random variable instead of a constant. (Coming up next.)

- **Minimum overlap length:**
  There should be a minimum overlap length requirement.
  E.g., if all four nucleotides occur with equal probability 1/4, the last character of any read will overlap the first character of one quarter of all the reads, which is not useful.
  - Typically, the minimum overlap is $\Omega = 100$ nucleotides, which is a fraction $\theta = \Omega/L = 100/500 = .2$ of the read length.
  - The size of the sequenced region doesn't change, but some contigs by our original definition may be split into multiple contigs overlapping by $< \Omega$.
  - Expected # contigs becomes $Ne^{(1-\theta)a} = \frac{aG}{L}e^{-(1-\theta)a}$.

Improved model accounting for variable read lengths

- Before we assumed the read length L was constant, say $L = 500$.
- Now treat it as a random variable with pdf $P_L(\ell)$.
- The pdf could be estimated from experimental data; it doesn't have to be one of the standard distributions.

Tail recursion formula in probability

If X is a random variable with range $0, 1, 2, \ldots$ then

$$E(X) = \sum_{x=1}^{\infty} P(X \geqslant x) = \sum_{x=0}^{\infty} P(X > x)$$

Proof.

$\sum_{x=1}^{\infty} P(X \geqslant x) =$

$$\begin{array}{rcr}
P(X \geqslant 1) & = & P_X(1) + P_X(2) + P_X(3) + \cdots \\
+\, P(X \geqslant 2) & & +\, P_X(2) + P_X(3) + \cdots \\
+\, P(X \geqslant 3) & & +\, P_X(3) + \cdots \\
+\, P(X \geqslant 4) & & +\, \cdots \\
\cline{3-3}
& = & 1P_X(1) + 2P_X(2) + 3P_X(3) + \cdots \\
& = & E(X) \qquad\qquad\qquad\qquad \square
\end{array}$$

Variable read lengths — # reads covering a point

- The number of reads covering point x is

$$\sum_{k\geqslant 0} \text{# reads starting at } x-k \text{ with length } > k$$

- Expected value of the term inside the summation:

$$E(\underbrace{\text{# reads starting at } x-k}_{\text{Expected value}=N/G} \cdot \underbrace{\text{with length } > k}_{P(L>k)})$$

- Expected value of summation:

$$\sum_{k\geqslant 0} \frac{N}{G} \cdot P(L>k) = \frac{N}{G} \cdot \sum_{k\geqslant 0} P(L>k) = \frac{N}{G} \cdot \underbrace{E(L)}_{\text{using tail recursion formula}}$$

- It turns out the fraction of positions in contigs (rather than gaps) becomes $1-e^{-NE(L)/G}$ instead of $1-e^{-a}$.

- The answers to the other questions are messier but doable.

Further reading

Textbooks

- W.J. Ewens and G.R. Grant, *Statistical Methods in Bioinformatics: An Introduction*, 2nd edition, Springer-Verlag, New York, 2005, Chapter 5.1.
- R.C. Deonier, Simon Tavaré, M.S. Waterman, *Computational Genome Analysis: An Introduction*, Springer, New York, 2005, Chapters 4.5, 8.

Original papers

- E.S. Lander and M.S. Waterman (1988), "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, 2:231-239.
- R. Arratia, E.S. Lander, S. Tavare, and M.S. Waterman (1991), "Genomic mapping by anchoring random clones: a mathematical analysis," *Genomics*, 11:806-827.
- E. Port, F. Sun, D. Martin, and M.S. Waterman (1995), "Genomic mapping by end-characterized random clones: a mathematical analysis," *Genomics*, 26:84-100.
- These are available at https://dornsife.usc.edu/labs/msw/research-papers/