

Chapter 5.1. Lander-Waterman Statistics for Shotgun Sequencing

Prof. Tesler

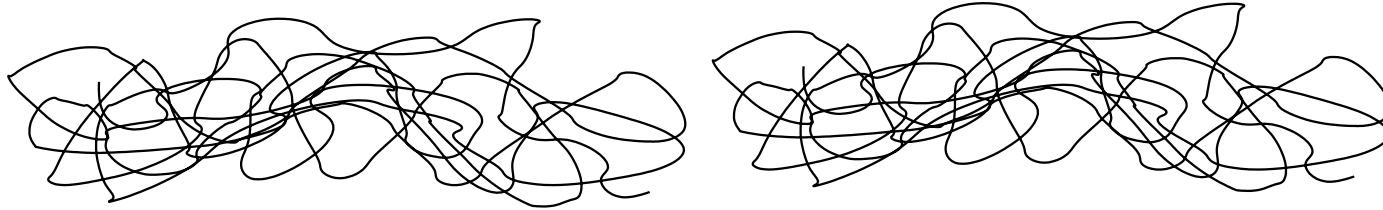
Math 283
April 9, 2008

Whole genome shotgun sequencing

- Frederick Sanger (and others) shared a Nobel Prize in Chemistry in 1980 for developing a method to sequence short regions of DNA.
- Today, 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.

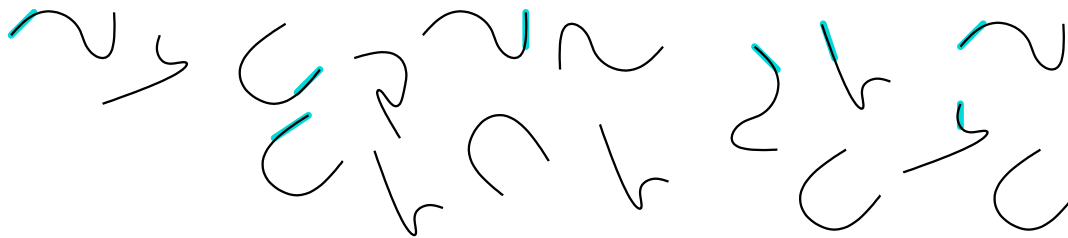
Overview of whole genome shotgun sequencing

Start with many copies of genome



Genome length G
 $G \approx 3$ billion

Fragment them and sequence reads



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

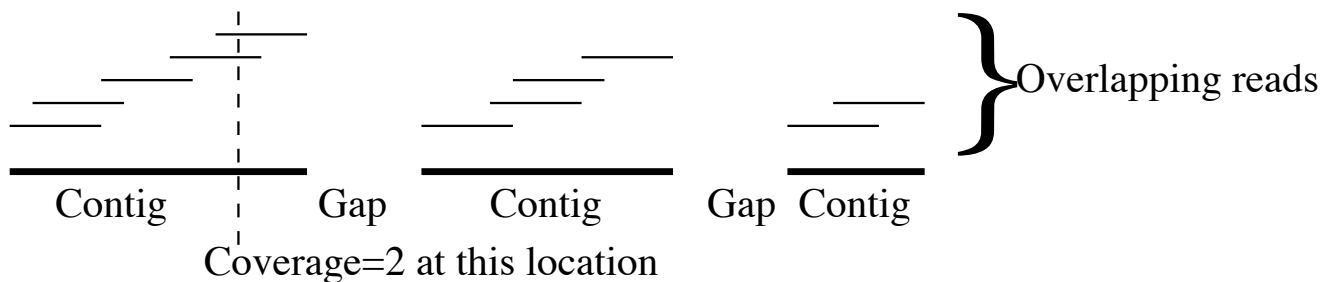
Find overlapping reads

ACGTAGAATCGACCATG...
...AACATAGTTGACGTAGAATC

Merge overlapping reads into contig

...AACATAGTTGACGTAGAATCGACCATG...

Many contigs



Shotgun sequencing

- Start with a sample with many copies of the same DNA.
- Break it at random into many smaller pieces, and randomly select a large number of these pieces to be *sequenced*.
- Approximately the first 500 nucleotides are read from one end of the pieces (the actual number will vary from piece to piece).
- These small sequenced regions are called *reads*.
We do not know their location in the genome or their strand.
- Combine overlapping reads into *contigs*.
- *Sequence alignment* is used to detect the overlaps. There are complications in this step that will be mentioned later.
- 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-res *Linkage maps* or other genetic maps.
- Additional 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.
- Once the BACs have been assembled, fit overlapping BACs together.

Human genome assemblies

- **Celera:** Celera is a company that did 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 (which we have not considered) that make it impossible to detect overlaps correctly.
- **Public effort:** The “public effort” to sequence the human genome used BACs.

Genome assembly statistics

Parameter names

G = genome length in nucleotides
 ≈ 3 billion in human
 ≈ 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.

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

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

Poisson distribution

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

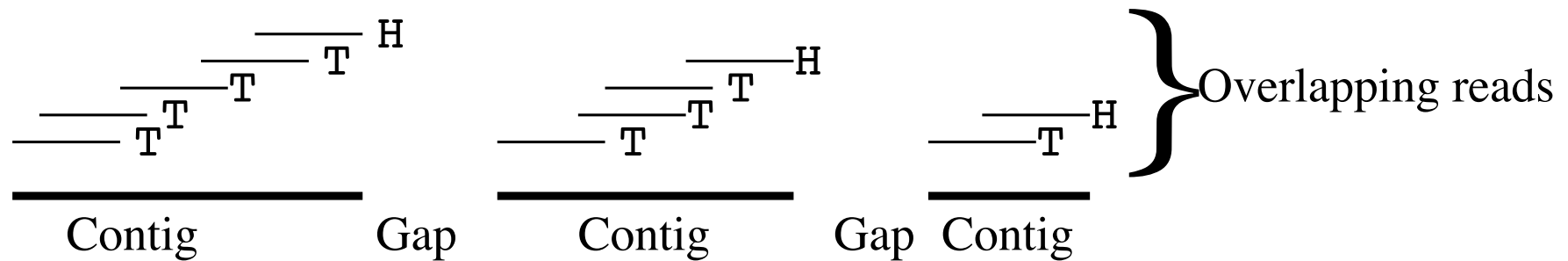
$$p = P(\text{no read starts in } I) = e^{-a} \frac{a^0}{0!} = e^{-a}$$
$$q = P(\geq 1 \text{ reads start in } I) = 1 - e^{-a}$$

- We'll use the Poisson distribution, but it's also possible to do everything with the binomial distribution.
- Poisson is more accurate in the high-coverage cases when it's likely there are multiple read-starts at the same position.

How much of the genome was sequenced?

- The nucleotide at x is in a gap if no read starts within the interval $[x - L + 1, x]$ (which has length L).
- We just computed the probability of this as $p = e^{-a}$. So the amount of the genome in gaps is $\approx pG = e^{-a}G$ nucleotides.
- The probability that the nucleotide at x is in a contig is $q = 1 - e^{-a}$, so the amount of the genome that is sequenced is $\approx qG = (1 - e^{-a})G$ nucleotides.
- To have 99% of the genome in contigs and 1% in gaps requires
$$q = 1 - e^{-a} = .99 \quad p = e^{-a} = .01$$
giving coverage $a = -\ln(.01) \approx 4.6$.
- 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?



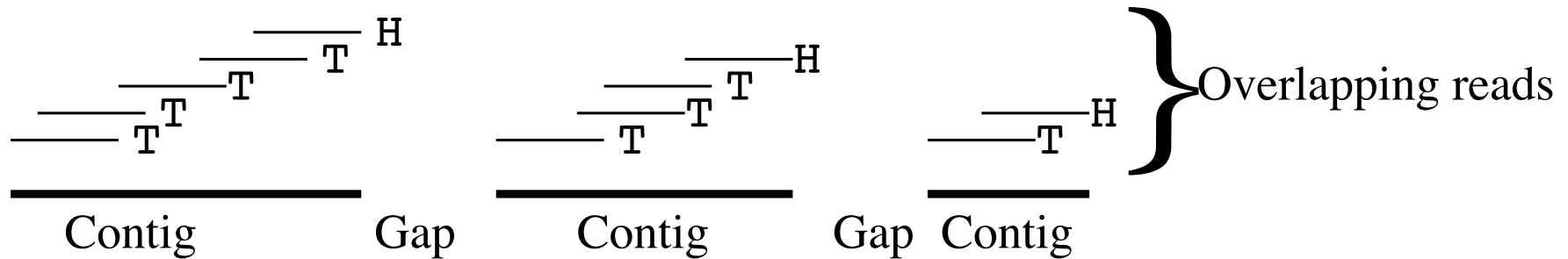
- Each contig has a unique rightmost read.
- The probability that a read is rightmost is the same as the probability that no other read starts within that read,

$$\exp(-N(L-1)/G) \approx p = \exp(-a) .$$

- 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:

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

Human whole genome shotgun sequencing

$G = 3$ billion, $L = 500$

Coverage a	# reads $N =$ aG/L (millions)	# nuc. read aG (billions)	% genome sequenced $(1 - e^{-a})$ $\cdot 100\%$	mean # contigs $(G/L)a \cdot e^{-a}$	mean contig length $(e^a - 1)L/a$	mean # reads/contig e^a
0.5	3M	1.5B	39.35%	1819592.0	648.7	1.6
1.0	6M	3.0B	63.21%	2207276.6	859.1	2.7
1.5	9M	4.5B	77.69%	2008171.4	1160.6	4.5
2.0	12M	6.0B	86.47%	1624023.4	1597.3	7.4
3.0	18M	9.0B	95.02%	896167.2	3180.9	20.1
4.0	24M	12.0B	98.17%	439575.3	6699.8	54.6
5.0	30M	15.0B	99.33%	202138.4	14741.3	148.4
6.0	36M	18.0B	99.75%	89235.1	33535.7	403.4
7.0	42M	21.0B	99.91%	38299.0	78259.5	1096.6
8.0	48M	24.0B	99.97%	16102.2	186247.4	2981.0
9.0	54M	27.0B	99.99%	6664.1	450115.8	8103.1
10.0	60M	30.0B	100.00%	2724.0	1101273.3	22026.5
11.0	66M	33.0B	100.00%	1102.3	2721506.4	59874.1
12.0	72M	36.0B	100.00%	442.4	6781408.0	162754.8
13.0	78M	39.0B	100.00%	176.3	17015861.2	442413.4
14.0	84M	42.0B	100.00%	69.8	42950117.3	1202604.3
15.0	90M	45.0B	100.00%	27.5	108967212.4	3269017.4
16.0	96M	48.0B	100.00%	10.8	277690922.5	8886110.5
17.0	102M	51.0B	100.00%	4.2	710439757.5	24154952.8
18.0	108M	54.0B	100.00%	1.6	1823888003.8	65659969.1
18.5	111M	55.5B	100.00%	1.0	2925810452.7	108254987.8

Human BAC sequencing

$$G = 300,000, L = 500$$

Coverage a	# reads $N =$ aG/L	# nuc. read aG	% genome sequenced $(1 - e^{-a})$ $\cdot 100\%$	mean # contigs $(G/L)a \cdot e^{-a}$	mean contig length $(e^a - 1)L/a$	mean # reads/contig e^a
0.5	300	150,000	39.35%	182.0	648.7	1.6
1.0	600	300,000	63.21%	220.7	859.1	2.7
1.5	900	450,000	77.69%	200.8	1160.6	4.5
2.0	1,200	600,000	86.47%	162.4	1597.3	7.4
2.5	1,500	750,000	91.79%	123.1	2236.5	12.2
3.0	1,800	900,000	95.02%	89.6	3180.9	20.1
3.5	2,100	1,050,000	96.98%	63.4	4587.9	33.1
4.0	2,400	1,200,000	98.17%	44.0	6699.8	54.6
4.5	2,700	1,350,000	98.89%	30.0	9890.8	90.0
5.0	3,000	1,500,000	99.33%	20.2	14741.3	148.4
5.5	3,300	1,650,000	99.59%	13.5	22153.8	244.7
6.0	3,600	1,800,000	99.75%	8.9	33535.7	403.4
6.5	3,900	1,950,000	99.85%	5.9	51087.8	665.1
7.0	4,200	2,100,000	99.91%	3.8	78259.5	1096.6
7.5	4,500	2,250,000	99.94%	2.5	120469.5	1808.0
8.0	4,800	2,400,000	99.97%	1.6	186247.4	2981.0
8.5	5,100	2,550,000	99.98%	1.0	289045.2	4914.8
9.0	5,400	2,700,000	99.99%	0.7	450115.8	8103.1
9.5	5,700	2,850,000	99.99%	0.4	703090.9	13359.7
10.0	6,000	3,000,000	100.00%	0.3	1101273.3	22026.5

Garbage in, garbage out

- Notice at high coverage, some values are nonsense (contig length larger than the BAC), which indicates the approximations we made are not valid. More careful analysis is required to determine exactly when the approximations are valid.
- Even when the values appear to be valid, beware: this is a statistical model, not a physical law. In a physical law ($E = mc^2$, $F = ma$, $PV = nRT$, etc.), you expect the formulas to be obeyed, at least for certain conditions and within certain measurement errors. But the Lander-Waterman statistics are only estimates, and are not too precise.

Low coverage assemblies ($2\times$) — Cat and others

- NHGRI has approved funding for numerous low-coverage genome assemblies (cat, alpaca, armadillo, shrew, peromyscus, tenrec, bat, guinea pig, elephant, ...). Related genomes will be used to make plausible guesses to arrange the contigs.

Cat genome

- Pontius et al., *Initial Sequence and Comparative Analysis of the Cat Genome*, *Genome Res.* 17: 1675–1689 (2007).
- 8 million reads, covering 1.642 Gb assembled sequence ($\approx 60\%$ of 2.7 Gb genome).
- $\approx 820,000$ contigs (N50 length* = 2378 bp) due to low coverage.
- Contigs assembled into $\approx 222,000$ scaffolds (N50 length = 117 kb).
- Human and dog assemblies were used to assist the cat assembly.
- Cat scaffolds initially placed against dog (when alignments were unique).
- Final placement of contigs was based on an RH (Radiation Hybrid) map with 1680 markers.

* *N50 length* L is the max L s.t. 50% of all nucleotides are in contigs of length $\geq L$.

High coverage (“Next generation sequencing”)

- Several companies have new approaches to generating large numbers of reads more cheaply and quickly than Sanger sequencing. However, the lengths are still shorter than in Sanger sequencing and the error rates are still higher.
- Exact comparisons are still hard to obtain. The estimates below do not include the price of actually purchasing the machine, or the cost & time of all the preparatory work before using the machine.

Estimates per run (websites & asking sales people & users)

Platform	# bases	Read length	Time	Cost
Roche/454 GS-FLX	100 Mb	200–300 bp Mean 250 bp	7.5 hours	\$14K
Illumina/Solexa	1 Gb	30–50 bp		\$3K
ABI Solid System				
Fragment library	1–1.5 Gb	≤ 35 bp	1 week	\$2K
Mate-pair library	1.5–2 Gb	2×25 bp	1 week	\$3K–\$4K
and the machine costs \$600K				

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. (See later slides.)

- **Minimum overlap length:**

There should be a minimum overlap length requirement. For example, assuming all four nucleotides occur with equal probability $1/4$, the last character of any read will be overlapped by 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 $Ne^{(1-\theta)a} = \frac{aG}{L}e^{-(1-\theta)a}$.

Complications — DNA is double-stranded

- 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 from the two strands, it is necessary to double the number of reads: for each read obtained by the sequencing machine, the sequence assembly software creates an additional read by taking the reverse complement.
- **Double-barreled reads:**
Read 500 nucleotides apiece from both ends of a fragment, and also estimate the fragment length.
This gives correlated reads, which may allow positioning and orienting two contigs.

Complications — Read errors and repeats

- **Read errors:**

There are usually errors in the reads. 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.

- **Repeats:**

Additionally, there are many repeats in the genome. Reads from different parts of the genome will have identical sequences and thus appear to overlap, even though they should not be fitted together into a contig!

Complications

- In some sequencing experiments, the DNA samples come from different individuals. Thus, there will be slight differences in corresponding parts of the genome (substitutions, deletions, insertions, variable number tandem repeats, etc.) that will make it difficult to detect overlaps, and once detected, difficult to resolve what “consensus” letters should be used when the reads are merged into a contig.

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 $0, 1, 2, \dots$ then

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

Proof.

$$\begin{aligned} \sum_{x=1}^{\infty} P(X \geq x) &= \\ & P(X \geq 1) = P_X(1) + P_X(2) + P_X(3) + \dots \\ & + P(X \geq 2) = + P_X(2) + P_X(3) + \dots \\ & + P(X \geq 3) = + P_X(3) + \dots \\ & + P(X \geq 4) = + \dots \\ & = \frac{1P_X(1) + 2P_X(2) + 3P_X(3) + \dots}{=} \\ & = E(X) \end{aligned}$$

□

Variable read lengths — # reads covering a point

- The number of reads covering point x is

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

- Assume the # reads starting at each nucleotide is either 0 or 1.
- The probability a read starts at any specific point is

$$\frac{\# \text{ reads}}{\text{genome length}} = \frac{N}{G}$$

(ignoring boundary effects)

- The probability the read length is $> k$ is

$$P(L > k) = \sum_{\ell=k+1}^{\infty} P_L(\ell)$$

Variable read lengths — # reads covering a point

- The expected # reads covering x is

$$\begin{aligned} \sum_{k \geq 0} E(\# \text{ reads starting at } x - k \text{ with length } > k) \\ &= \sum_{k \geq 0} 1 \cdot \frac{N}{G} \cdot P(L > k) \\ &= \frac{N}{G} \sum_{k \geq 0} P(L > k) = \frac{N}{G} E(L) \end{aligned}$$

where we used the tail recursion formula in the last step.

- It turns out the probability position X is in some contig becomes $1 - e^{-NE(L)/G}$ instead of $1 - e^{-a}$.
- The answers to the other questions are messier but doable.

References

For more details on the Lander-Waterman genome assembly statistics:

- 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.
- The original papers on this topic by Lander and Waterman, and follow-ups with other coauthors, are listed (and some are available for download) at <http://www-hto.usc.edu/people/msw/publications.html>

The NIH timeline of organisms to be sequenced is at <http://www.genome.gov/10002154>