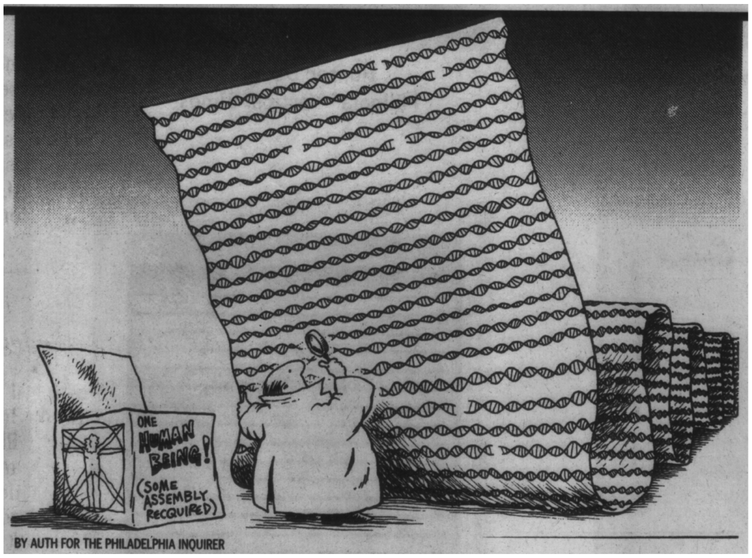
> "I guess I'm just hopelessly fascinated by the realities that you can assemble out of connected fragments."
> --Junot Diaz

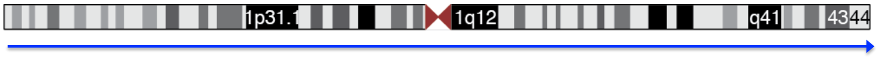


BY AUTH FOR THE PHILADELPHIA INQUIRER

1

Sequence read lengths remain limiting

Chr1: 249 Mb

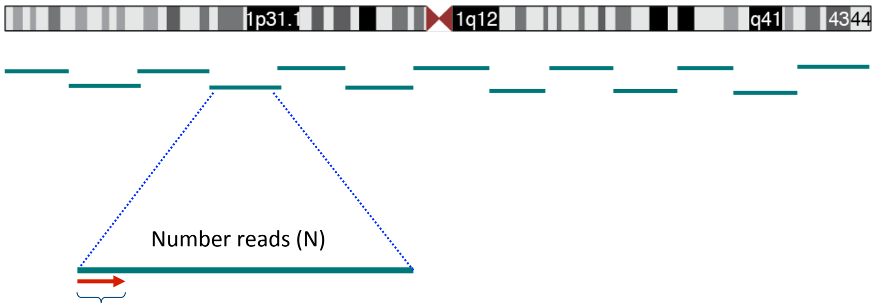


249 Mb sequencing read

Current platforms:

- Sanger: A very small number (1-10,000) reads (700-1000 bp) but lowest error rates
- Illumina: A very large number (2 billion) of short reads (75-200 bp) but error rate 0.1-1%
- PacBio: A moderate number (~500,000) of long reads (~10 kb) but error rate as high as 14% (but have been falling).

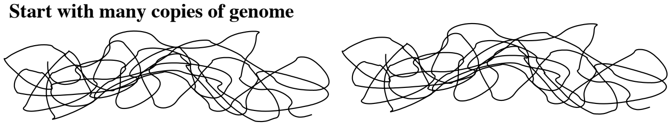
The challenge:



Read length (L)
500 bp

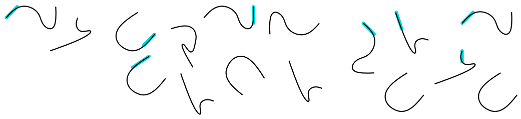
Overview of whole genome shotgun sequencing

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

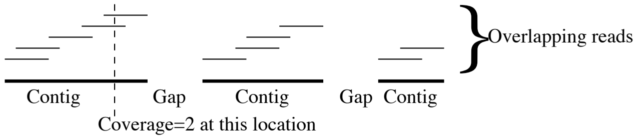
```
                      ACGTAGAATCGACCATG...
       ...AACATAGTTGACGTAGAATC
```

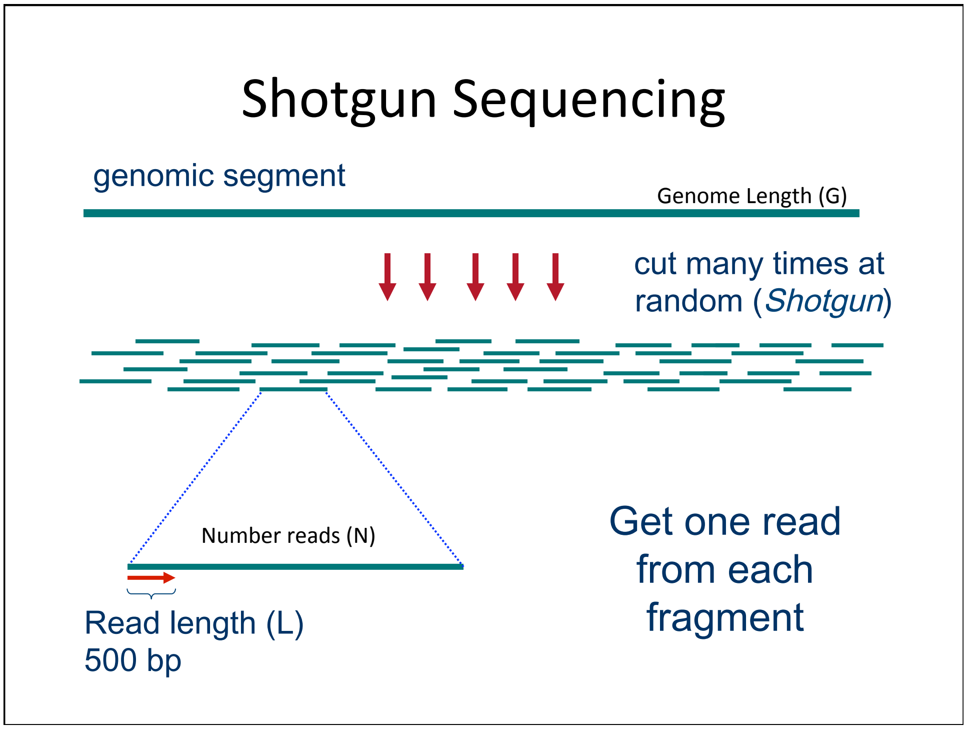
Leverage redundancy
to identify overlaps.

Merge overlapping reads into contig

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

Many contigs





Shotgun Sequencing

genomic segment

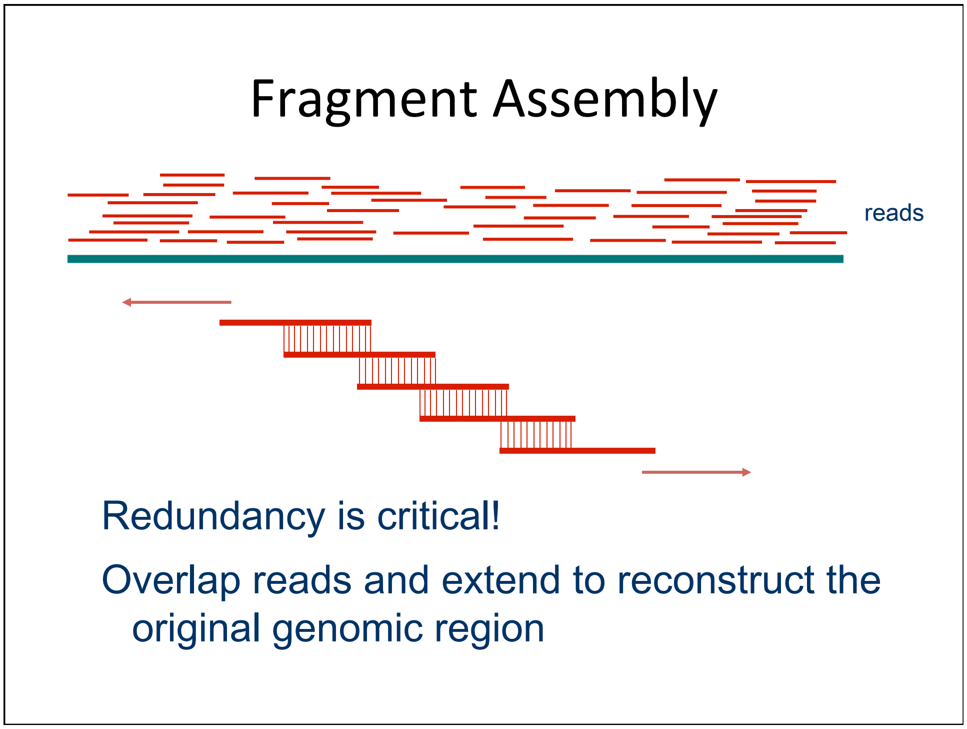
Genome Length (G)

cut many times at random (*Shotgun*)

Number reads (N)

Read length (L)
500 bp

Get one read from each fragment



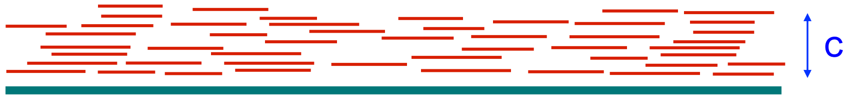
Fragment Assembly

reads

Redundancy is critical!

Overlap reads and extend to reconstruct the original genomic region

Read Coverage



Length of genomic segment: ***G***

Number of reads: ***n***

Length of each read: ***L***

Coverage ***C = n L / G***

Therefore ..

Getting 1X coverage of the human genome requires:

$N = c*G / L = 1(3x10^9) / 500 =$ 6 million reads

And 10X requires 60 million reads!

Lander-Waterman statistics: questions

- As the coverage increases, more and more areas of the genome are **likely** to be covered. Ideally, you want to see 1 long contig per chromosome.

Genome Coverage

- If you sequence at 10x coverage how much of the genome will be sequenced <u>at least</u> 5 times?

Lander and Waterman (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2(3):231-239.

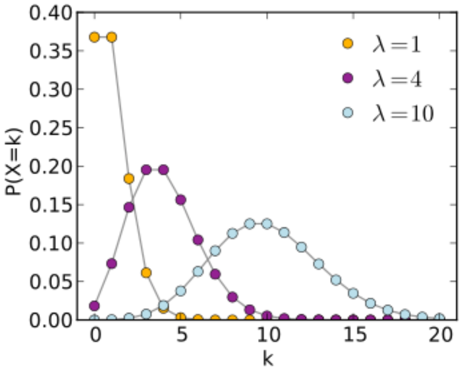
Lander-Waterman model: Poisson distribution

- a discrete frequency distribution that gives the probability of a number of independent events occurring in a fixed time.

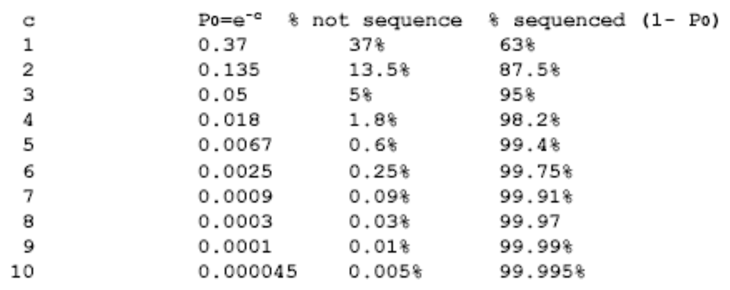
$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

Average coverage = lambda

Probability of getting k reads for a base given the average coverage lambda



Poisson Distribution



c	$Po=e^{-c}$	% not sequence	% sequenced (1- Po)
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%

Example

- Average coverage = 5x
- Probability of a given base being sequenced 10 times is:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

$5^{10}e^{-5}/10!$ = 0.018 or about 2% of bases will have 10x coverage.

- If you sequence at 10x coverage how much of the genome will be sequenced at least 5 times?

$$1 - [f(0,10) + f(1,10) + f(2,10) + f(3,10) + f(4,10)] = 0.97$$

Determining sequence overlap

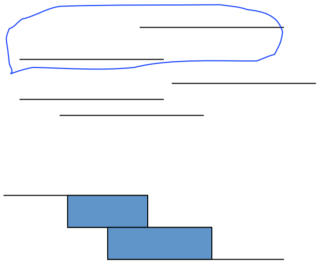
- Given a pair of fragments s_1 and s_2, do they belong together?

Yes, if a prefix of s_2 matches a suffix of s_1

- How would you compute such a match?

Overlap Detection

- Compute the best prefix-suffix alignments between each pair of fragments.
- Keep the "high-scoring" ones as evidence of true overlap.
- What is the problem?



Overlap detection problem

- Consider the number of fragments. The statistics say that we need good coverage (c=8, 10) to get most of the base-pairs.
  - G = 3000Mb, L=500
  - Coverage LN/G = 10
  - N = $10*3*10^9/500 = 6*10^7$
  - Number of comparisons needed = $3.6 * 10^{15}$
    - Not good! (Only a small fraction are true overlaps)

- Repeats at read ends can be assembled in multiple ways.

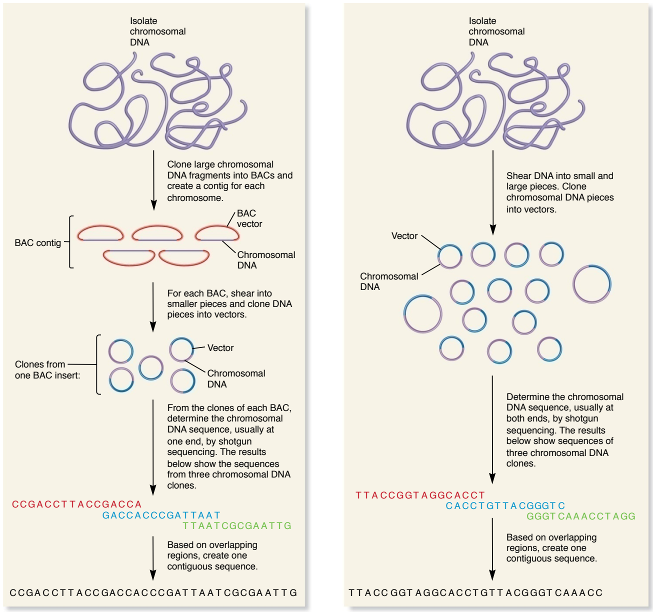
```
TCTTGGTCATGTCAT
     GTCATGTCATACGTC
               ACGTCGTCATGTCAT
                    GTCATGTCATTGGTCCC        correct

                  or

        TCTTGGTCATGTCAT
             GTCATGTCATTGGTCCC

         ACGTCGTCATGTCAT                      incorrect
              GTCATGTCATACGTC
```

Repeats complicate assembly!



Clone large chromosomal DNA fragments into BACs and create a contig for each chromosome.

For each BAC, shear into smaller pieces and clone DNA pieces into vectors.

From the clones of each BAC, determine the chromosomal DNA sequence, usually at one end, by shotgun sequencing. The results below show the sequences from three chromosomal DNA clones.

```
CCGACCTTACCGACCA
           GACCACCCGATTAAT
                     TTAATCGCGAATTG
```

Based on overlapping regions, create one contiguous sequence.

```
CCGACCTTACCGACCACCCGATTAATCGCGAATTG
```

(a) The public genome sequencing strategy (map-based)

Shear DNA into small and large pieces. Clone chromosomal DNA pieces into vectors.

Determine the chromosomal DNA sequence, usually at both ends, by shotgun sequencing. The results below show sequences of three chromosomal DNA clones.

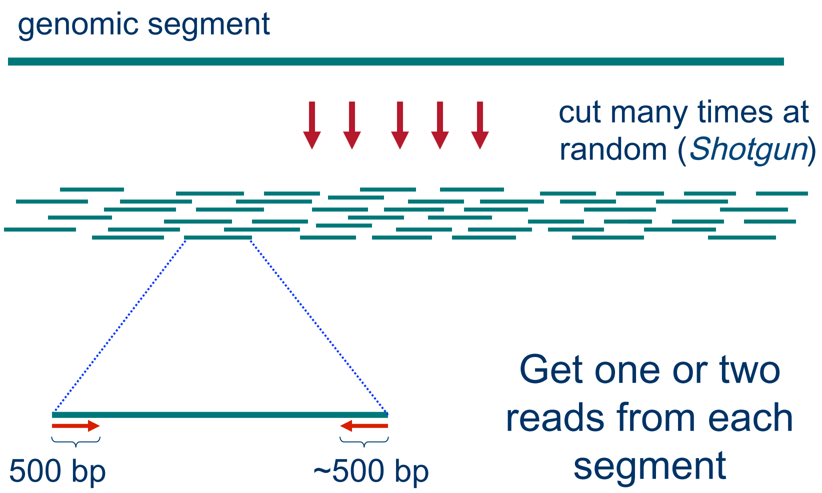
```
TTACCGGTAGGCACCT
           CACCTGTTACGGGTC
                     GGGTCAAACCTAGG
```

Based on overlapping regions, create one contiguous sequence.

```
TTACCGGTAGGCACCTGTTACGGGTCAAACC
```

(b) Celera's strategy: Whole-genome shotgun sequencing

Paired end (or mate pair) reads



genomic segment

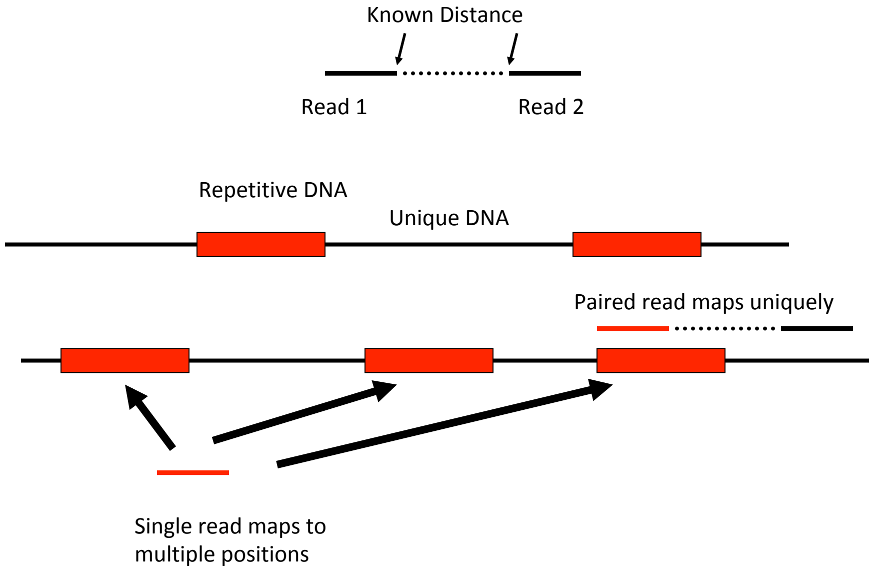
cut many times at random (*Shotgun*)

500 bp

~500 bp

Get one or two reads from each segment

Paired End Reads are Important!



Known Distance

Read 1

Read 2

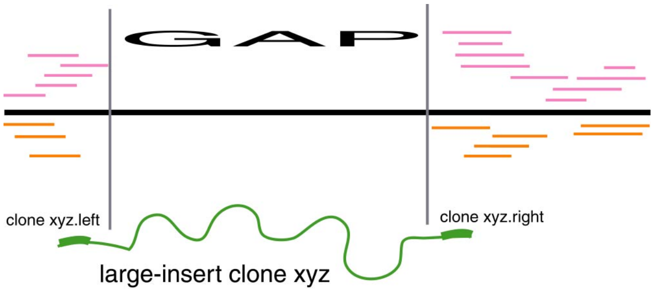
Repetitive DNA

Unique DNA

Paired read maps uniquely

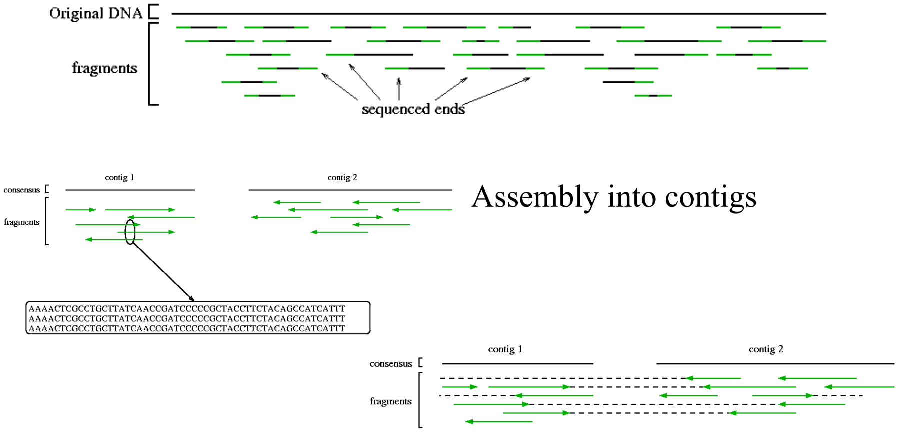
Single read maps to multiple positions

In addition to anchoring repeats, paired end (or mate pair) reads also permit scaffolding



- Mate-pairs allow you to merge contigs into scaffolds

Unifying view of assembly



```
AAAACTCGCCTGCTTATCAACCGATCCCCCGCTACCTTCTACAGCCATCATTT
AAAACTCGCCTGCTTATCAACCGATCCCCCGCTACCTTCTACAGCCATCATTT
AAAACTCGCCTGCTTATCAACCGATCCCCCGCTACCTTCTACAGCCATCATTT
```

Assembly into contigs

Scaffolding using paired end

22