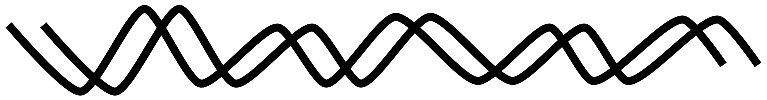
Genome Sequencing

CMSC 423

Carl Kingsford

Genome Sequencing

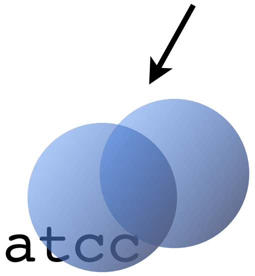


ACCGTCCAATTGG...
TGGCAGGTTAACC...

E.g. human: 3 billion bases
split into 23 chromosomes

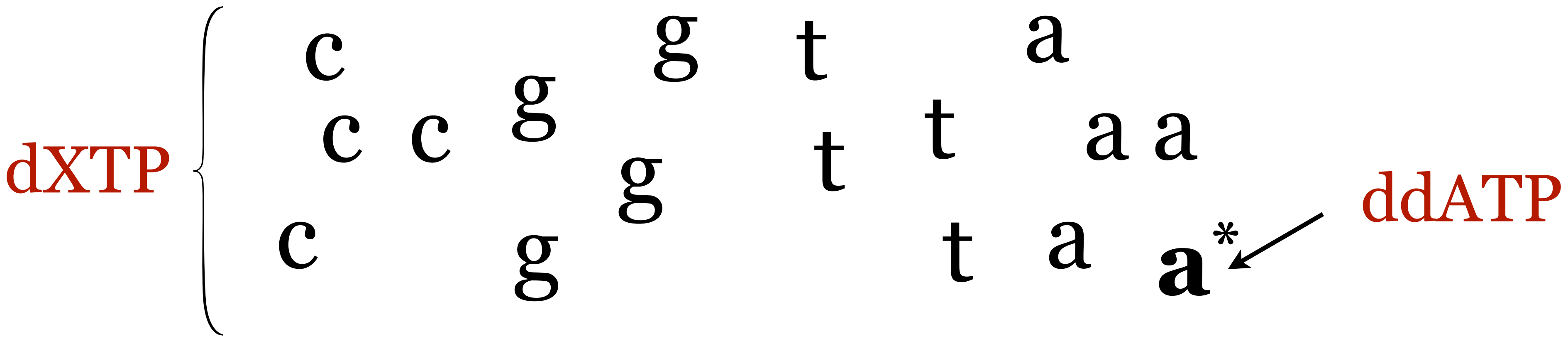
Main tool of traditional sequencing: DNA Synthesis

DNA polymerase: enzyme that will
grow a complementary DNA strand.



```
gacgatcggtttatcc
ctgctagccaaataggctaatactacgga
```

Sanger Sequencing: Finding the As



```
gacgatcgg tttA*
ctgctagcc aaaTaggcTaaTacTacgga
```

```
gacgatcgg tttAtccgA*
ctgctagcc aaaTaggcTaaTacTacgga
```

```
gacgatcgg tttA*
ctgctagcc aaaTaggcTaaTacTacgga
```

```
gacgatcgg tttAtccgAttA*
ctgctagcc aaaTaggcTaaTacTacgga
```

```
gacgatcgg tttAtccgAttAtgA*
ctgctagcc aaaTaggcTaaTacTacgga
```

```
gacgatcgg tttAtccgAttA*
ctgctagcc aaaTaggcTaaTacTacgga
```

```
gacgatcgg tttAtccgAttAtgA*
ctgctagcc aaaTaggcTaaTacTacgga
```

```
gacgatcgg tttAtccgA*
ctgctagcc aaaTaggcTaaTacTacgga
```

Size → Sequence

gacgatcggttt**A***

gacgatcggttt**A***

gacgatcggttt**A**tccg**A***

gacgatcggttt**A**tccg**A***

gacgatcggttt**A**tccg**A**tt**A***

gacgatcggttt**A**tccg**A**tt**A***

gacgatcggttt**A**tccg**A**tt**A**tg**A***

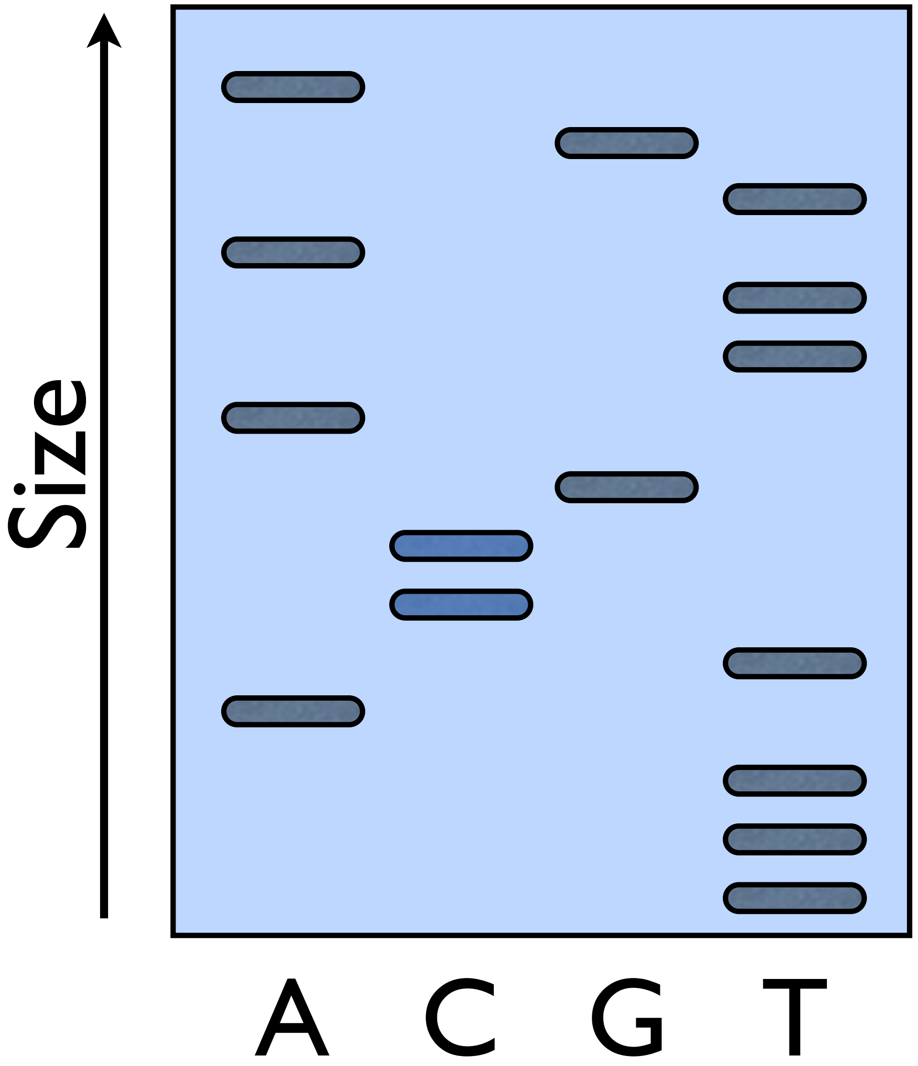
gacgatcggttt**A**tccg**A**tt**A**tg**A***

gacgatcggtttatccgattat**G***

gacgatcggtttatcc**G***

gacgatcggtttat**C***

gacgatcggtttatc**C***



Size → Sequence

gacgatcggttt**A***

gacgatcggttt**A***

gacgatcggttt**A**tccg**A***

gacgatcggttt**A**tccg**A***

gacgatcggttt**A**tccg**A**tt**A***

gacgatcggttt**A**tccg**A**tt**A***

gacgatcggttt**A**tccg**A**tt**A**tg**A***

gacgatcggttt**A**tccg**A**tt**A**tg**A***

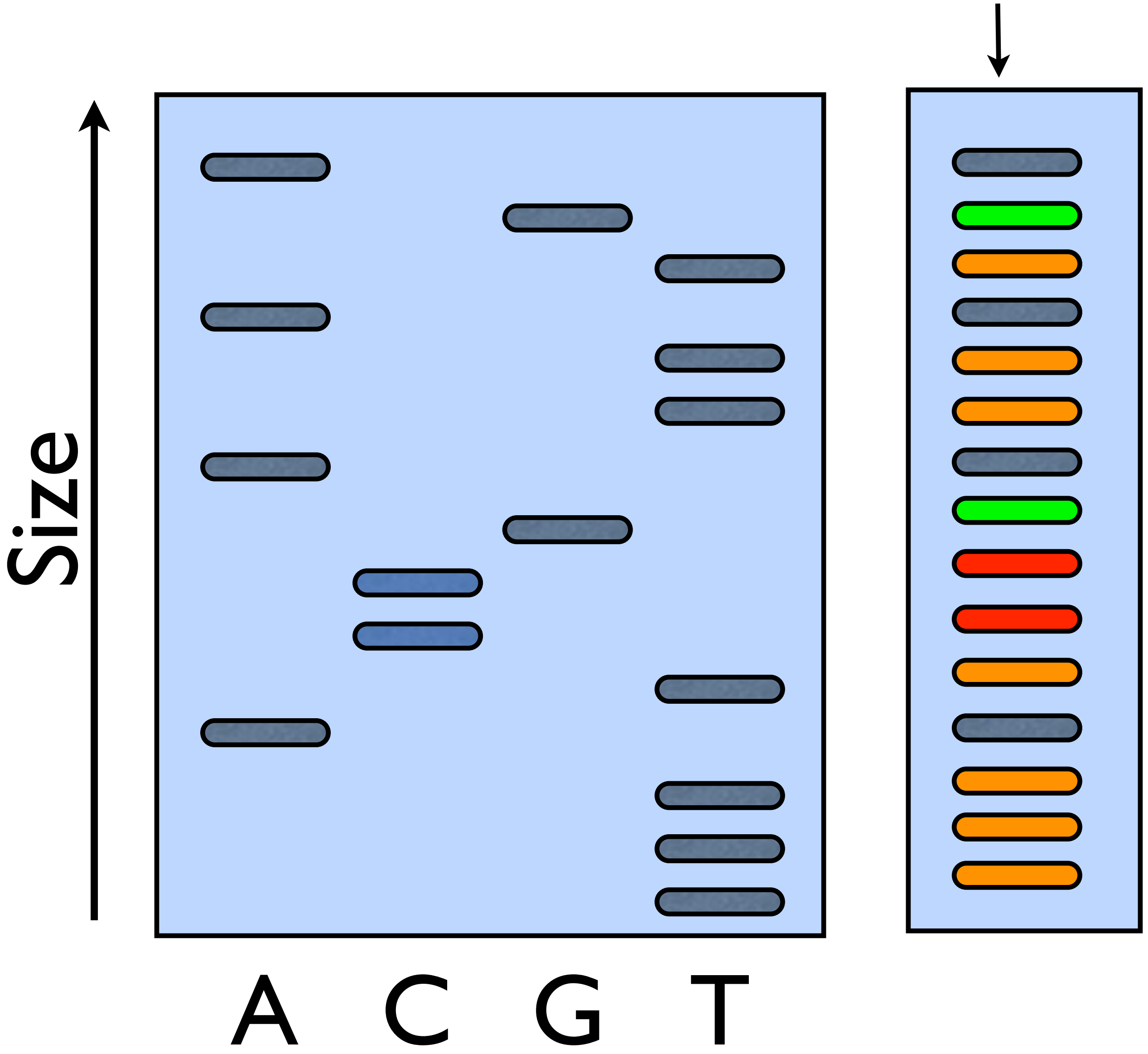
gacgatcggtttatccgattat**G***

gacgatcggtttatcc**G***

gacgatcggtttat**C***

gacgatcggtttatc**C***

Single lane: ddXTP
that fluoresce
different colors



Size → Sequence

gacgatcggttt**A***

gacgatcggttt**A***

gacgatcggttt**A**tccg**A***

gacgatcggttt**A**tccg**A***

gacgatcggttt**A**tccg**A**tt**A***

gacgatcggttt**A**tccg**A**tt**A***

gacgatcggttt**A**tccg**A**tt**A**tg**A***

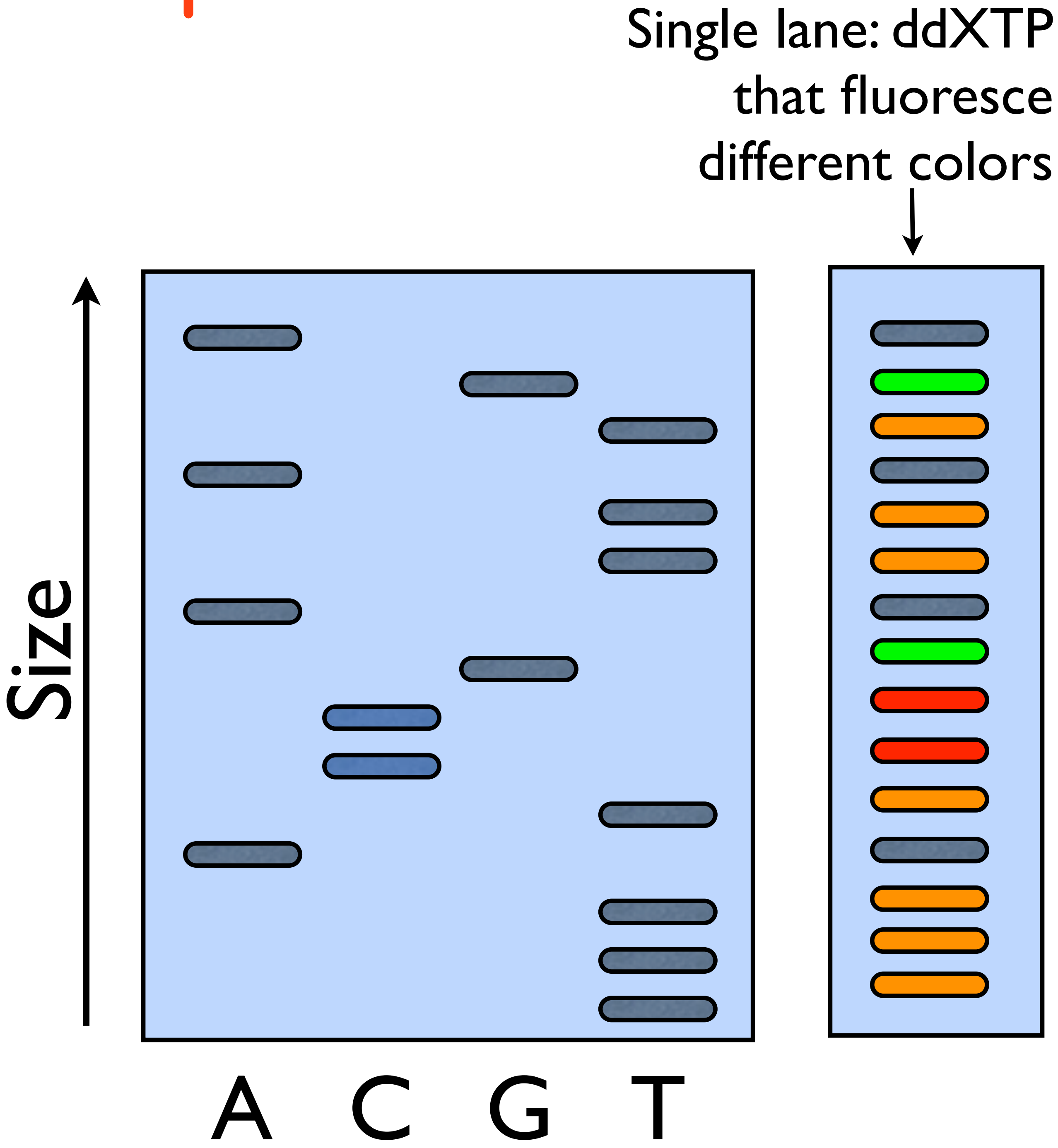
gacgatcggttt**A**tccg**A**tt**A**tg**A***

gacgatcggtttatccgattat**G***

gacgatcggtttatcc**G***

gacgatcggtttat**C***

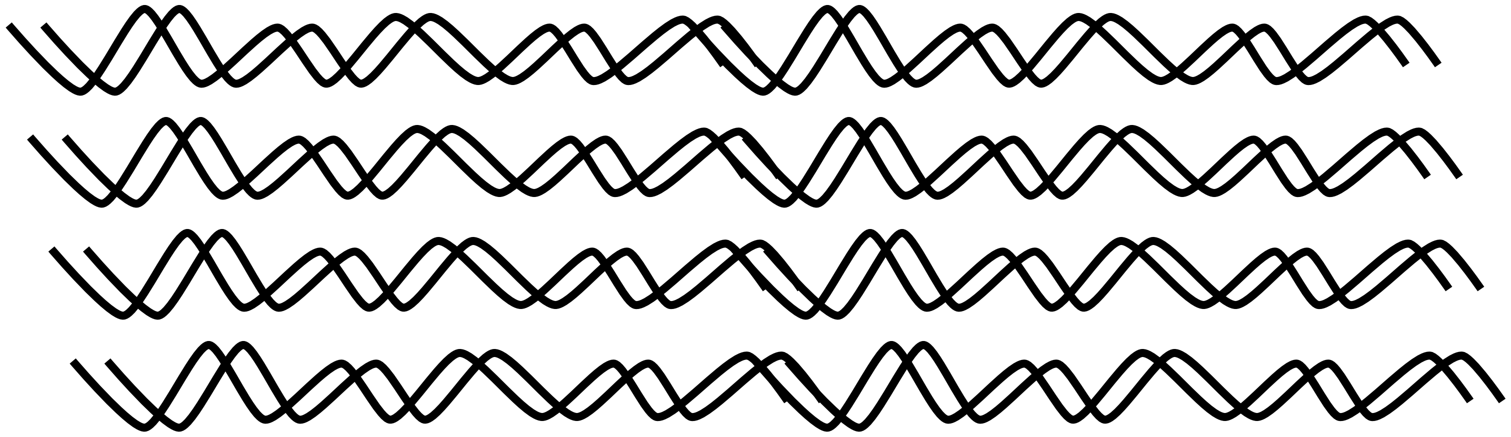
gacgatcggtttatc**C***



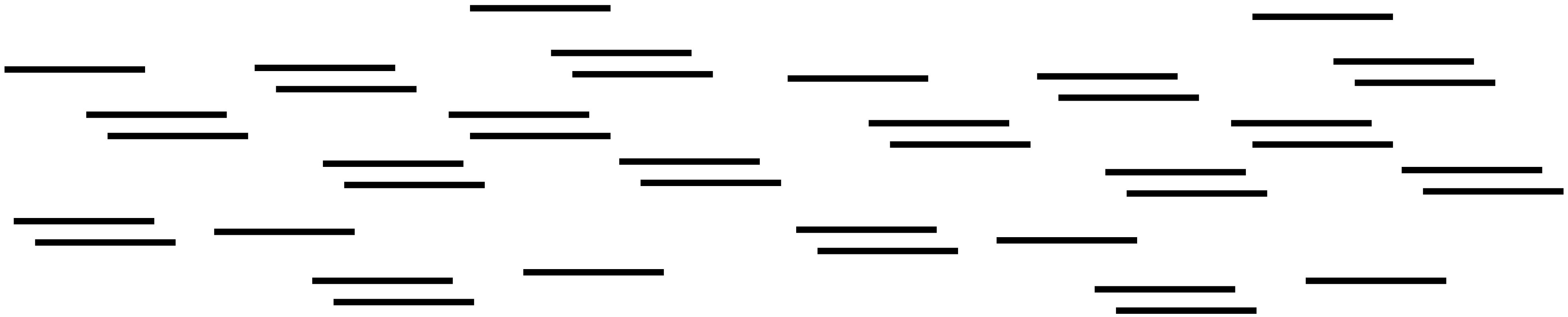
Main problem: larger fragments take a long time to be sorted correctly (or don't sort correctly ever) → 800-1000 letter maximum

Shotgun Sequencing

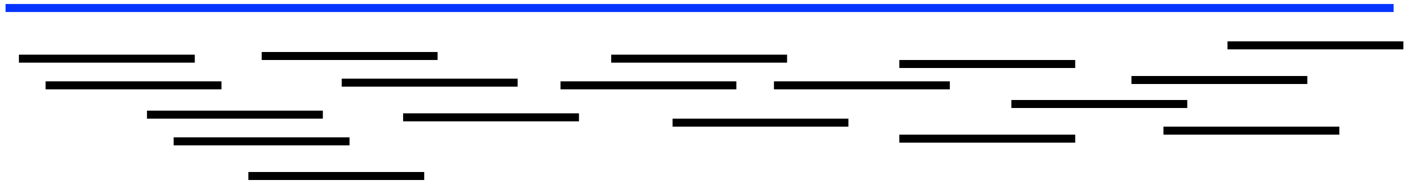
Many copies
of the DNA



Shear it, randomly breaking them into many small pieces,
read ends of each:

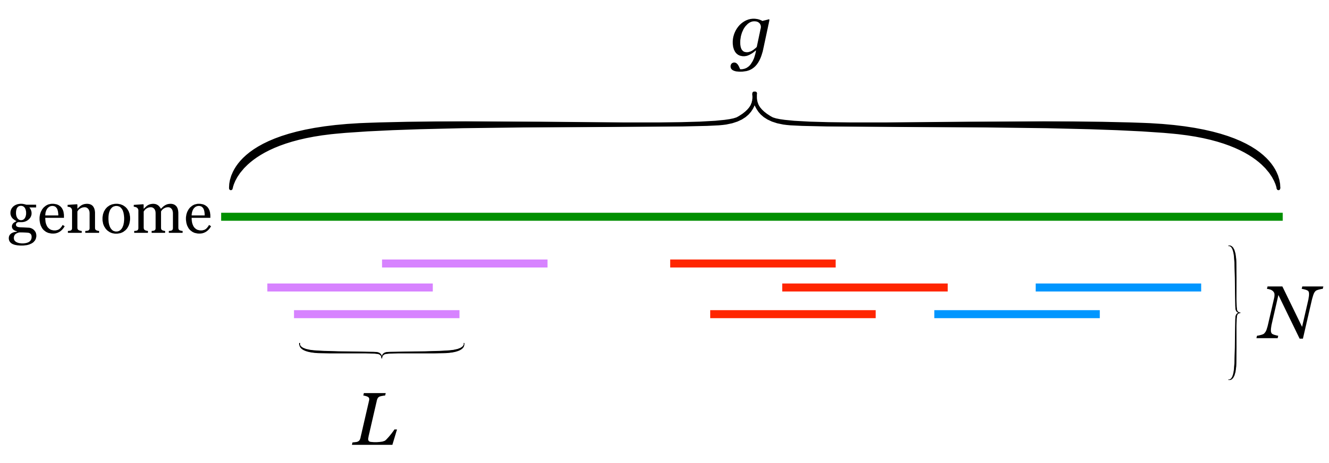


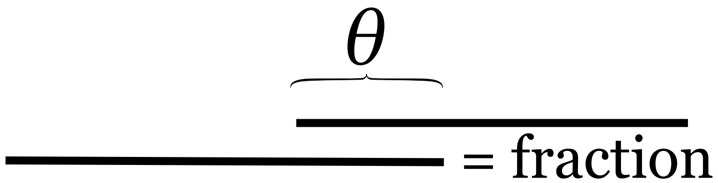
Assemble into original genome:



Lander-Waterman Statistics

How many reads to we need to be sure we cover the whole genome?





= fraction of L required to detect an overlap

An ***island*** is a contiguous group of reads that are connected by overlaps of length ≥ θL. (Various colors above)

Want: Expression for expected # of islands given N, g, L, θ.

Expected # of Islands

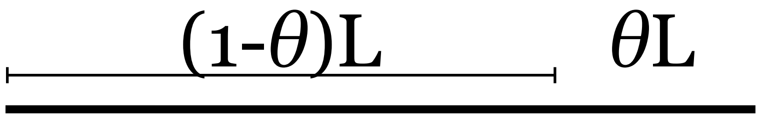
λ := N/g = probability a read starts at a given position (assuming random sampling)

Pr(k reads start in an interval of length x)

x trials, want k "successes," small probability λ of success
Expected # of successes = λx
Poisson approximation to binomial distribution:

$$\Pr(k \text{ reads in length } x) = e^{-\lambda x}\frac{(\lambda x)^k}{k!}$$

Expected # of islands = N × Pr(read is at rightmost end of island)



= N × Pr(0 reads start in (1-θ)L)

$$= Ne^{-\lambda(1-\theta)L}\frac{\lambda^0}{0!} \quad \text{(from above)}$$

$$= Ne^{-\lambda(1-\theta)L}$$

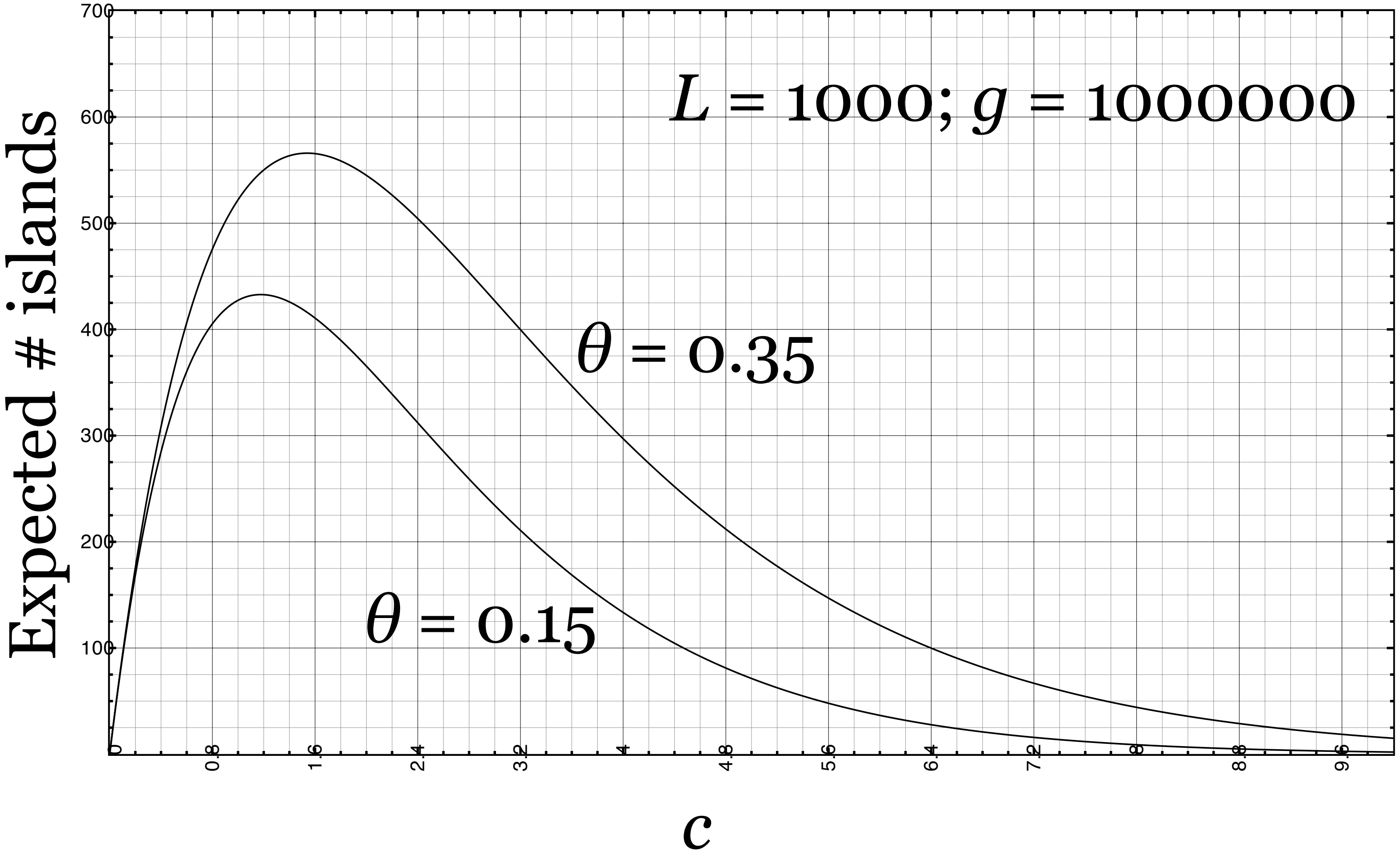
$$= Ne^{-(1-\theta)LN/g}$$

← LN/g is called the **coverage** c.

Expected # of Islands, 2

Rewrite to depend more directly on the things we can control: c and θ

$$\begin{aligned}\text{Expected \# of islands} &= Ne^{-(1-\theta)LN/g} \\ &= Ne^{-(1-\theta)c} \\ &= \frac{L/g}{L/g}Ne^{-(1-\theta)c} \\ &= \frac{g}{L}ce^{-(1-\theta)c}\end{aligned}$$



Summary

- “Sanger” sequencing widely used up through 2006 or 2007, including for the human genome project.

- Won Sanger his second Nobel prize.

- Lander-Waterman statistics estimate the number of islands you will get for a given coverage.

    - Used as a way to guess how much sequencing you need to do for a given technology and genome size.

    - Often hard in practice to guess the genome size g before you’ve sequenced it.