

Bernoulli distribution

The simplest non-uniform distribution

p – probability of success (1)

$1-p$ – probability of failure (0)

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Jacob Bernoulli

(1654-1705)

Swiss mathematician (Basel)

- Law of large numbers
- Mathematical constant $e=2.718...$



Bernoulli distribution

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = 0(1 - p) + 1(p) = p$$

$$\text{Var}(X) = E(X^2) - (EX)^2 = [0^2(1 - p) + 1^2(p)] - p^2 = p - p^2 = p(1 - p)$$

Refresher: Binomial Coefficients

$$\binom{n}{k} = C_k^n = \frac{n!}{k!(n-k)!}, \text{ called } n \text{ choose } k$$

$$\binom{10}{3} = C_3^{10} = \frac{10!}{3!7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{3 \cdot 2 \cdot 1 \cdot 7!} = 120$$

Number of ways to choose k objects out of n

without replacement and where the **order does not matter**.

Called binomial coefficients because of the binomial formula

$$(p+q)^n = (p+q) \times (p+q) \dots \times (p+q) = \sum_{x=0}^n C_x^n p^x q^{n-x}$$

Binomial Distribution

- **Binomially-distributed** random variable X equals **sum (number of successes) of n independent Bernoulli trials**
- The probability mass function is:

$$f(x) = C_x^n p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \dots, n \quad (3-7)$$

$q = 1-p$

- Based on the binomial expansion:

$$1 = (p+q)^n = \sum_{x=0}^n C_x^n p^x q^{n-x}$$

Binomial variance and standard deviation

Let X be a binomial random variable
with parameters p and n

Variance:

$$\sigma^2 = V(X) = np(1-p)$$

Standard deviation:

$$\sigma = \sqrt{np(1-p)}$$

Poisson Distribution

- Limit of the binomial distribution when
 - n , the **number of attempts**, is very **large**
 - p , the **probability of success** is very **small**
 - $E(X) = np = \lambda$ is $O(1)$

The annual numbers of deaths from horse kicks in 14 Prussian army corps between 1875 and 1894

| Number of deaths | of Observed frequency | Expected frequency |
|------------------|-----------------------|--------------------|
| 0 | 144 | 139 |
| 1 | 91 | 97 |
| 2 | 32 | 34 |
| 3 | 11 | 8 |
| 4 | 2 | 1 |
| 5 and over | 0 | 0 |
| Total | 280 | 280 |

From von Bortkiewicz 1898



Siméon Denis Poisson
(1781–1840)
French mathematician
and physicist

Let $\lambda = np = E(x)$, so $p = \frac{\lambda}{n}$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$= \frac{n(n-1)\dots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \sim \frac{n^x}{x!} \left(\frac{\lambda}{n}\right)^x = \frac{\lambda^x}{x!};$$

$$\sum_x \frac{\lambda^x}{x!} = e^\lambda.$$

Normalization requires $\sum_x P(X = x) = 1$.

$$\text{Thus } P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Poisson Mean & Variance

If X is a Poisson random variable, then:

- Mean: $\mu = E(X) = \lambda \approx n \cdot p$
- Variance: $\sigma^2 = V(X) = \lambda \approx n \cdot p \cdot (1 - p) \approx n \cdot p$
- Standard deviation: $\sigma = \lambda^{1/2}$

Note: Variance = Mean

Note: Standard deviation/Mean = $\lambda^{-1/2}$
decreases with λ

Matlab exercise: Poisson distribution

- Generate a **sample of size 100,000** for Poisson-distributed random variable X with $\lambda = 2$
- Plot the approximation to the **Probability Mass Function** based on this sample
- Calculate the mean and variance of this sample and compare it to **theoretical calculations**:
 $E[X] = \lambda$ and $V[X] = \lambda$

Matlab exercise: Poisson distribution

- **Stats=100000; lambda=2;**
- **r2=random('Poisson',lambda,Stats,1);**
- **mean(r2)**
- **var(r2)**
- **[a,b]=hist(r2, 0:max(r2));**
- **p_p=a./sum(a);**
- **figure; stem(b,p_p);**
- **figure; semilogy(b,p_p,'ko-')**

Credit: XKCD
comics

WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS

WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE



WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

WHY AREN'T ECONOMISTS RICH

WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS THERE PHLEGM

WHY DO AMERICANS CALL IT SOCCER

WHY IS PSYCHIC WEAK TO BUG

WHY ARE MY EARS RINGING

WHY DO CHILDREN GET CANCER

WHY ARE THERE SO MANY AVENGERS

WHY IS POSEIDON ANGRY WITH ODYSSEUS

WHY ARE THE AVENGERS FIGHTING THE X MEN

WHY IS THERE ICE IN SPACE

WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD

WHY IS SPACE BLACK

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL OUTSIDE MY WINDOW

WHY IS OUTER SPACE SO COLD

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL ON THE DOLLAR BILL

WHY ARE THERE PYRAMIDS ON THE MOON

WHY ARE THERE GHOSTS

WHY DO OWLS ATTACK PEOPLE

WHY IS NASA SHUTTING DOWN

WHY ARE THERE GHOSTS

WHY ARE AK 47s SO EXPENSIVE

WHY ARE THERE MALE AND FEMALE BIKES

WHY ARE THERE GHOSTS

WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE

WHY ARE THERE TINY SPIDERS IN MY HOUSE

WHY ARE THERE GHOSTS

WHY ARE THERE GODS

WHY DO SPIDERS COME INSIDE

WHY ARE THERE GHOSTS

WHY ARE THERE TWO SPOCKS

WHY ARE THERE HUGE SPIDERS IN MY HOUSE

WHY ARE THERE GHOSTS

WHY IS LIFE SO BORING

WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE

WHY ARE THERE GHOSTS

WHY ARE CIGARETTES LEGAL

WHY ARE THERE SPIDERS IN MY ROOM

WHY ARE THERE GHOSTS

WHY ARE THERE DUCKS IN MY POOL

WHY ARE THERE SO MANY SPIDERS IN MY ROOM

WHY ARE THERE GHOSTS

WHY IS JESUS WHITE

WHY DO SPIDER BITES ITCH

WHY ARE THERE GHOSTS

WHY IS THERE LIQUID IN MY EAR

WHY IS DYING SO SCARY

WHY ARE THERE GHOSTS

WHY DO Q TIPS FEEL GOOD

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS

WHY AREN'T THERE DINOSAUR GHOSTS

WHY DO I SAY UH

WHY IS SEA SALT BETTER

WHY ARE THERE TREES IN THE MIDDLE OF FIELDS

WHY IS THERE NOT A POKEMON MMO

WHY IS THERE LAUGHING IN TV SHOWS

WHY ARE THERE DOORS ON THE FREEWAY

WHY ARE THERE SO MANY SVCHOST.EXE RUNNING

WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA

WHY ARE THERE SCARY SOUNDS IN MINECRAFT

WHY IS THERE KICKING IN MY STOMACH

WHY ARE THERE TWO SLASHES AFTER HTTP

WHY ARE THERE CELEBRITIES

WHY DO SNAKES EXIST

WHY DO OYSTERS HAVE PEARLS

WHY ARE DUCKS CALLED DUCKS

WHY DO THEY CALL IT THE CLAP

WHY ARE KYLE AND CARTMAN FRIENDS

WHY IS THERE AN ARROW ON AANG'S HEAD

WHY ARE TEXT MESSAGES BLUE

WHY ARE THERE MUSTACHES ON CLOTHES

WHY ARE THERE MUSTACHES ON CARS

WHY ARE THERE MUSTACHES EVERYWHERE

WHY ARE THERE SO MANY BIRDS IN OHIO

WHY IS THERE SO MUCH RAIN IN OHIO

WHY IS OHIO WEATHER SO WEIRD

WHY ARE THERE MALE AND FEMALE BIKES

WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KUNGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD

WHY DO TREES DIE

WHY IS THERE NO SOUND ON CNN

WHY AREN'T POKEMON REAL

WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY ARE THERE TINY SPIDERS IN MY HOUSE

WHY DO SPIDERS COME INSIDE

WHY ARE THERE HUGE SPIDERS IN MY HOUSE

WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE

WHY ARE THERE SPIDERS IN MY ROOM

WHY ARE THERE SO MANY SPIDERS IN MY ROOM

WHY DO SPIDER BITES ITCH

WHY IS DYING SO SCARY

WHY IS THERE NO GPS IN LAPTOPS

WHY DO KNEES CLICK

WHY AREN'T THERE E GRADES

WHY IS SEX SO IMPORTANT



WHY ARE THERE GHOSTS



WHY IS MT VESUVIUS THERE

WHY DO THEY SAY T MINUS

WHY ARE THERE OBELISKS

WHY ARE WRESTLERS ALWAYS WET

WHY ARE OCEANS BECOMING MORE ACIDIC

WHY IS ARWEN DYING

WHY AREN'T MY QUAIL LAYING EGGS

WHY AREN'T MY QUAIL EGGS HATCHING

WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY IS THERE AN OWL IN MY BACKYARD

WHY IS THERE AN OWL OUTSIDE MY WINDOW

WHY IS THERE AN OWL ON THE DOLLAR BILL

WHY DO OWLS ATTACK PEOPLE

WHY ARE AK 47s SO EXPENSIVE

WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE

WHY ARE THERE GODS

WHY ARE THERE TWO SPOCKS

WHY IS LIFE SO BORING

WHY ARE CIGARETTES LEGAL

WHY ARE THERE DUCKS IN MY POOL

WHY IS JESUS WHITE

WHY IS THERE LIQUID IN MY EAR

WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT

WHY ARE ULTRASOUND MACHINES EXPENSIVE

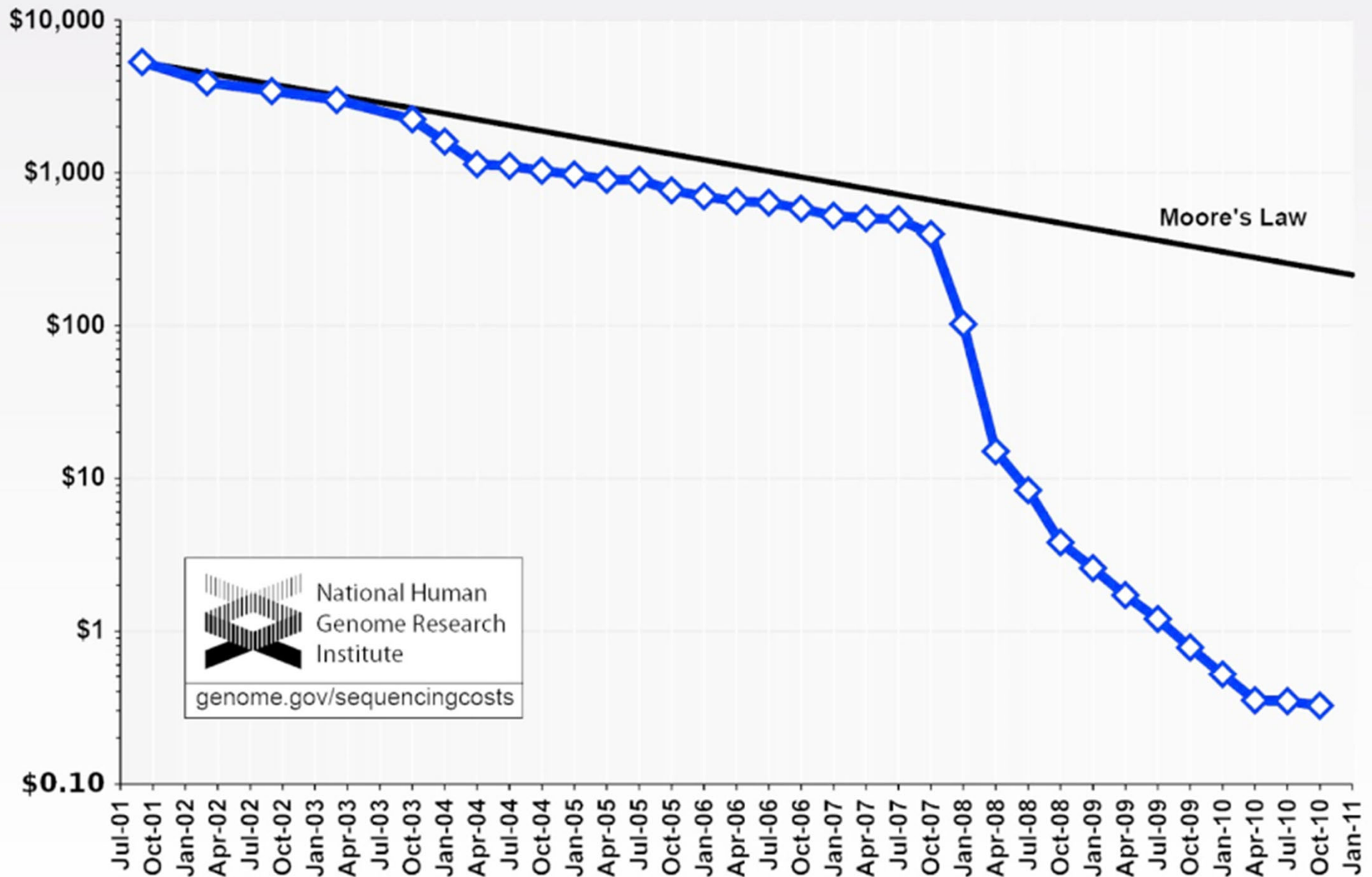
WHY IS STEALING WRONG

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Poisson Distribution in Genome Assembly

Cost per Megabase of DNA Sequence



 National Human
Genome Research
Institute
genome.gov/sequencingcosts

Poisson Example: Genome Assembly

- **Goal:** figure out the sequence of DNA nucleotides (ACTG) **along the entire genome**
- **Problem:** Sequencers generate random **short reads**

TABLE 9.1 Next-generation sequencing technologies compared to Sanger sequencing. Adapted from the companies' websites, [⊕ http://en.wikipedia.org/wiki/DNA_sequencer](http://en.wikipedia.org/wiki/DNA_sequencer), and literature cited for each technology.

| Technology | Read length (bp) | Reads per run | Time per run | Cost per megabase (US\$) | Accuracy (%) |
|---------------------|------------------|---------------|--------------|--------------------------|--------------|
| Roche 454 | 700 | 1 million | 1 day | 10 | 99.90 |
| Illumina | 50–250 | <3 billion | 1–10 days | ~0.10 | 98 |
| SOLiD | 50 | ~1.4 billion | 7–14 days | 0.13 | 99.90 |
| Ion Torrent | 200 | <5 million | 2 hours | 1 | 98 |
| Pacific Biosciences | 2900 | <75,000 | <2 hours | 2 | 99 |
| Sanger | 400–900 | N/A | <3 hours | 2400 | 99.90 |

- **Solution:** **assemble genome** from short reads using computers. **Whole Genome Shotgun Assembly.**



MinION, a palm-sized gene sequencer made by UK-based Oxford Nanopore Technologies

Short Reads assemble into Contigs

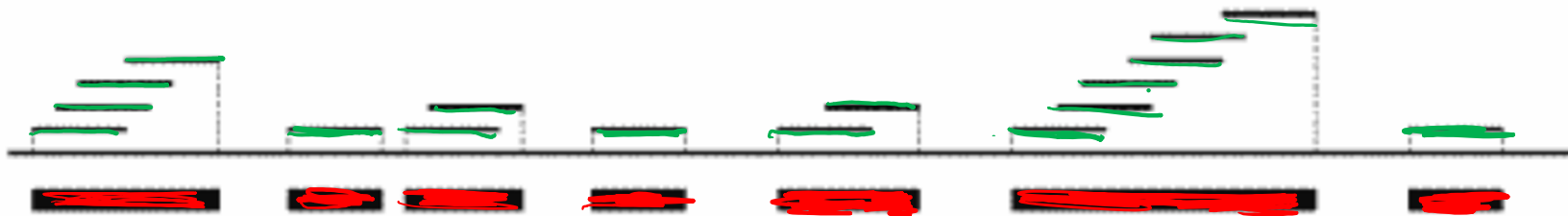
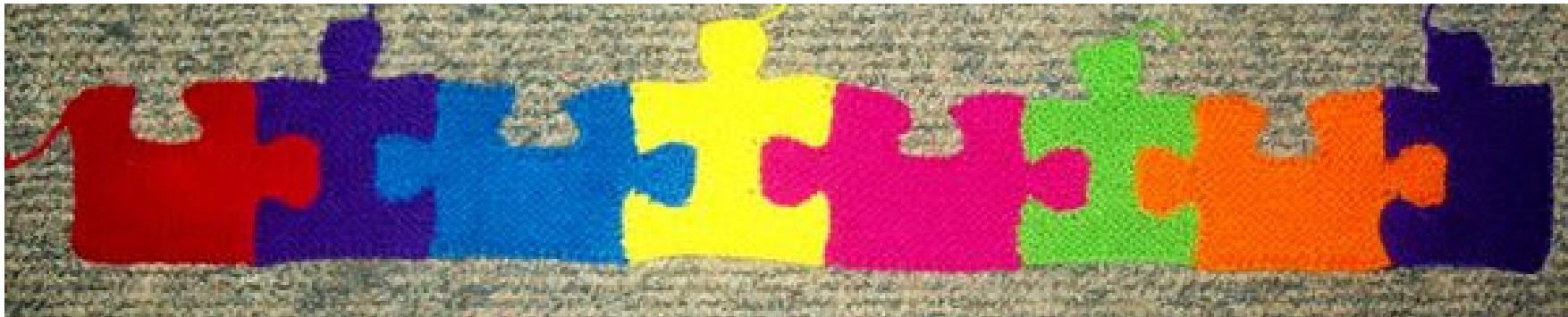


Figure 5.1.



Promise of Genomics



Drew Sheneman, New Jersey -- The Newark Star Ledger, [E-mail Drew](#).

I think I found the corner piece!

How many short reads do we need?

Input

Output

Low coverage:



A few pieces to assemble



many contigs, many gaps



High coverage:



many pieces to assemble



a few contigs, a few gaps



Genome Assembly

Whole-genome “shotgun” sequencing starts by copying and fragmenting the DNA

(“Shotgun” refers to the random fragmentation of the whole genome; like it was fired from a shotgun)

Input: GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTTT
35bp

Copy GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTTT
by GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTTT
PCR: GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTTT
GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTTT

Fragment: GCGTCTA TATCTCGG CTCTAGGCCCTC ATTTTTT
GGC GTCTATAT CTCGGCTCTAGGCCCTCA TTTTTT
GGCGTC TATATCT CGGCTCTAGGCCCT CATTTTTTT
GGCGTCTAT ATCTCGGCTCTAG GCCCTCA TTTTTT

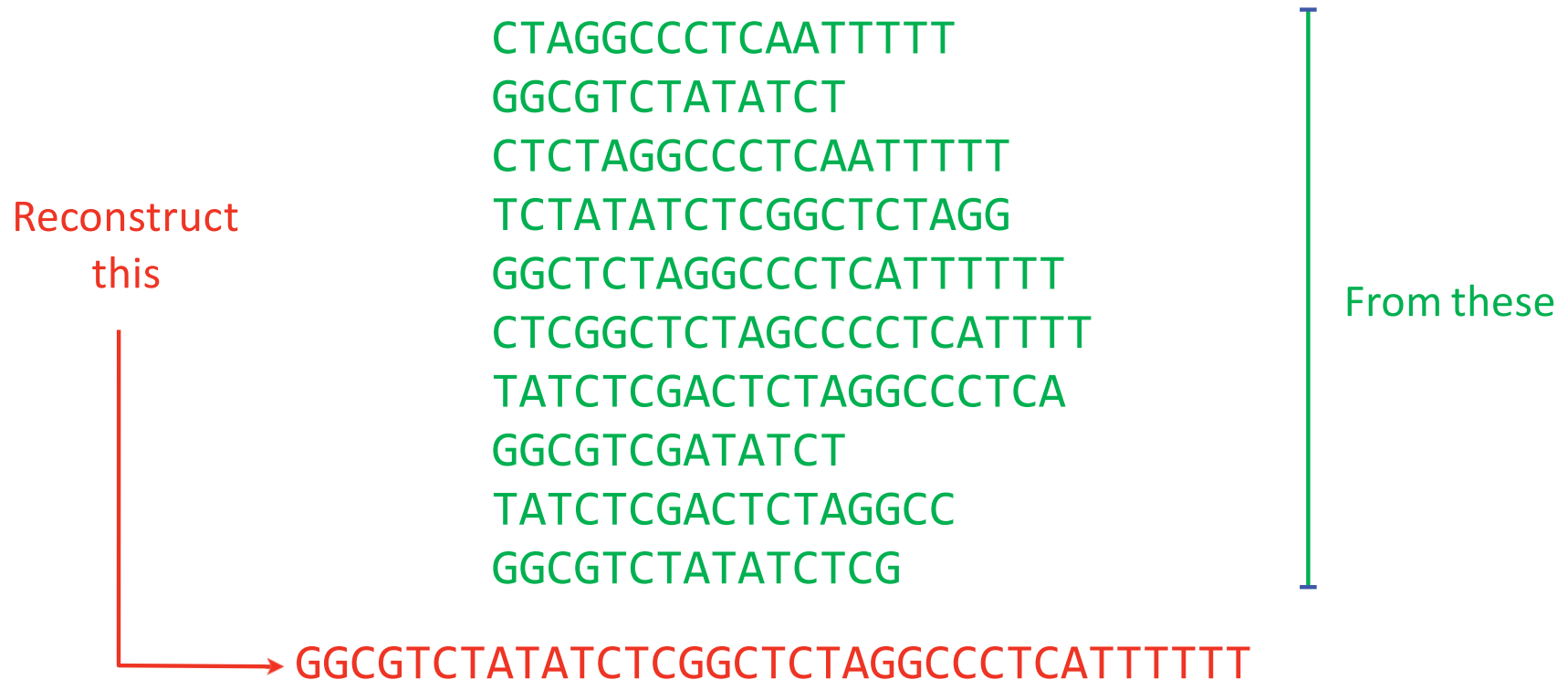
Courtesy of [Ben Langmead](http://www.langmead-lab.org). Used with permission.

<http://www.langmead-lab.org/teaching-materials/>

Assembly

Assume sequencing produces such a large # fragments that almost all genome positions are *covered* by many fragments...

...but we don't know what came from where



Courtesy of [Ben Langmead](http://www.langmead-lab.org/teaching-materials/). Used with permission.

Assembly

Overlaps between short reads help to put them together

```
          CTAGGCCCTCAATTTTT
        CTCTAGGCCCTCAATTTTT
       GGCTCTAGGCCCTCATTTTT
      CTCGGCTCTAGCCCCTCATTTT
     TATCTCGACTCTAGGCCCTCA
    TATCTCGACTCTAGGCC
   TCTATATCTCGGCTCTAGG
  GCGTCTATATCTCG
 GCGTCGATATCT
 GCGTCTATATCT
 GCGTCTATATCTCGGCTCTAGGCCCTCATTTTT
```

177 nucleotides

35 nucleotides

Where is the Poisson?

- G - genome length (in bp)
- L - short read average length
- N - number of short read sequenced
- λ - sequencing coverage redundancy = LN/G
- x - number of short reads covering a given site on the genome

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Poisson as a limit of Binomial: For a given site on the genome for each short read Prob(site covered): $p=L/G$ is very small. Number of attempts (short reads): N is very large. Their product (sequencing redundancy): $\lambda = NL/G$ is $O(1)$.



What fraction of the genome is missing?

What fraction of genome is covered?

- Coverage: $\lambda = NL/G$,
X – random variable equal to the number of times a given site is covered by short reads.
Poisson: $P(X=x) = \lambda^x \exp(-\lambda) / x!$
 $P(X=0) = \exp(-\lambda)$, $P(X>0) = 1 - \exp(-\lambda)$
- Total length covered: $G * [1 - \exp(-\lambda)]$

| λ | 2 | 4 | 6 | 8 | 10 | 12 |
|-----------------------------------|---------|---------|---------|---------|---------|---------|
| Mean proportion of genome covered | .864665 | .981684 | .997521 | .999665 | .999955 | .999994 |

Table 5.1. The mean proportion of the genome covered for different values of λ

How many contigs?

How many contigs?

- A given short read is the right end of a contig if and only if no left ends of other short reads fall within it.
- The left end of another short read has the probability $p=(L-1)/G$ to fall within a given read. There are $N-1$ other reads. Hence the expected number of left ends inside a given short read is $p \cdot (N-1)=(N-1) \cdot (L-1)/G \approx \lambda$
- If significant overlap required to merge two short reads is L_{ov} , modified λ is given by $(N-1) \cdot (L - L_{ov})/G$
- Probability that no left ends fall inside a short read is $exp(-\lambda)$. Thus the Number of contigs is $N_{contigs}=Ne^{-\lambda}$:

| λ | 0.5 | 0.75 | 1 | 1.5 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------------------|------|------|------|------|------|------|------|-----|-----|-----|
| Mean number of contigs | 60.7 | 70.8 | 73.6 | 66.9 | 54.1 | 29.9 | 14.7 | 6.7 | 3.0 | 1.3 |

Table 5.2. The mean number of contigs for different levels of coverage, with $G = 100,000$ and $L = 500$.