Poisson processes in biology.

https://github.com/quinlan-lab/applied-computational-genomics

Aaron Quinlan

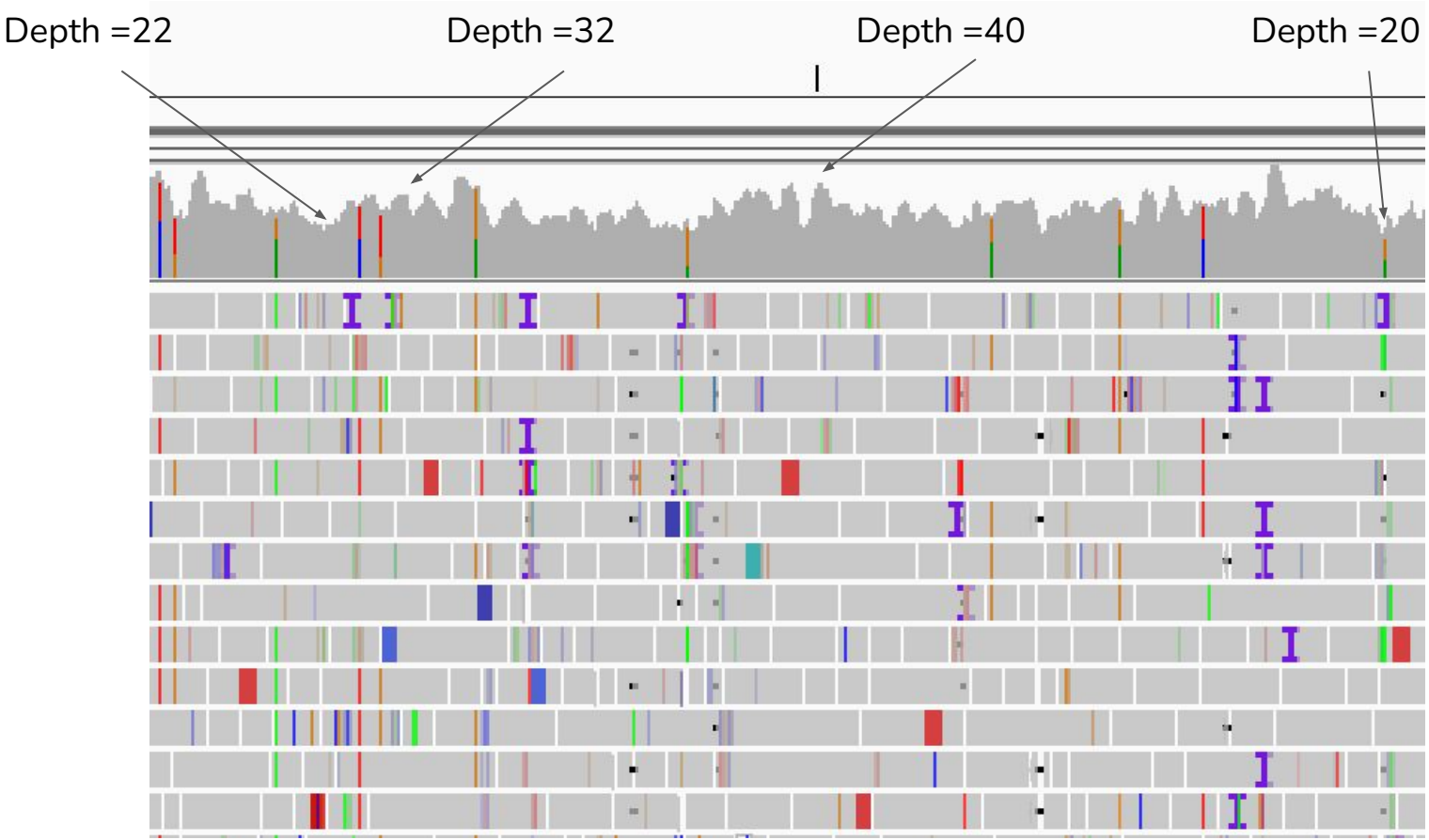
Departments of Human Genetics and Biomedical Informatics

USTAR Center for Genetic Discovery

University of Utah

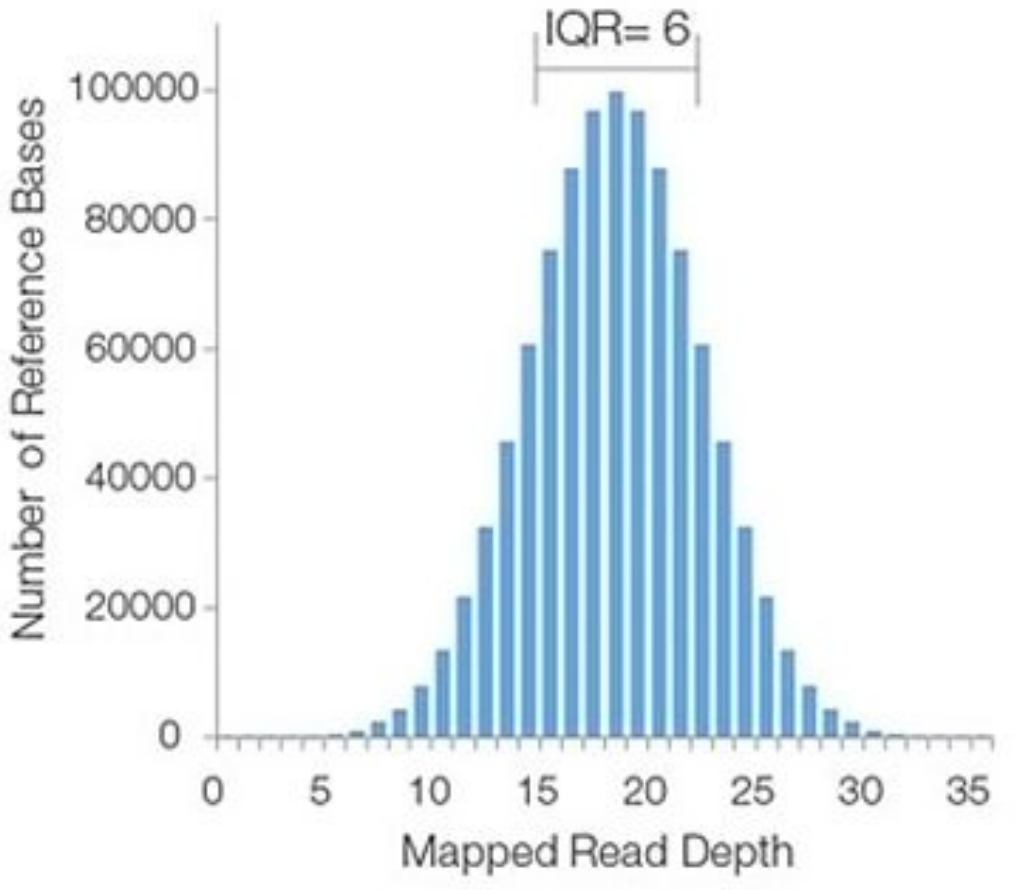
quinlanlab.org

Depth of aligned sequence is variable



What accounts for variability in depth?

What accounts for variability in depth?



Siméon-Denis Poisson

French mathematician, engineer, physicist (1781 – 1840)

- One of 72 scientists whose name is on the Eiffel tower.
- >300 publications on math, physics, and astronomy
- Creator of the **Poisson distribution**. Incredibly useful in science for cases where we need to count and model random events

The Poisson distribution is used to describe the distribution of rare events in a large population. For example, at any particular time, there is a certain probability that a particular cell within a large population of cells will acquire a mutation.

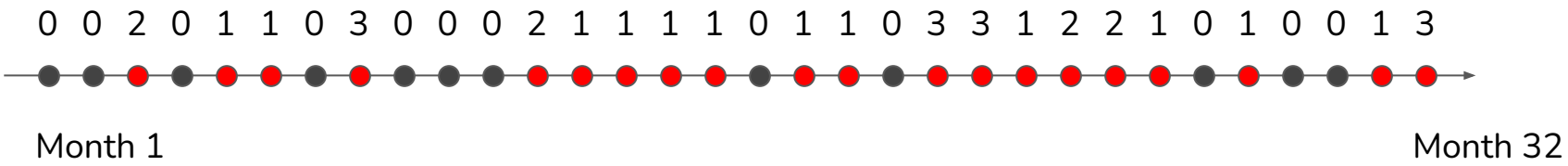
A Poisson process is appropriate here because mutation acquisition is a rare event, and each mutation event is independent of one another.

https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/poisson-distribution

Poisson Process

Useful model when counting the occurrences of events that appear to happen at a certain rate, but at random.

Let's say that there is 1 earthquake above 6.0 worldwide per month, on average.

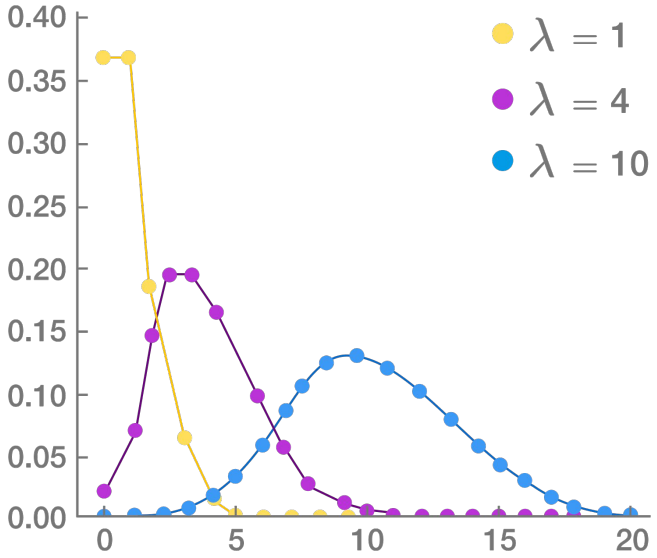


Important requirement: the number of events per time period are **independent**.

Poisson Distribution. One parameter. Lambda

Expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate

$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

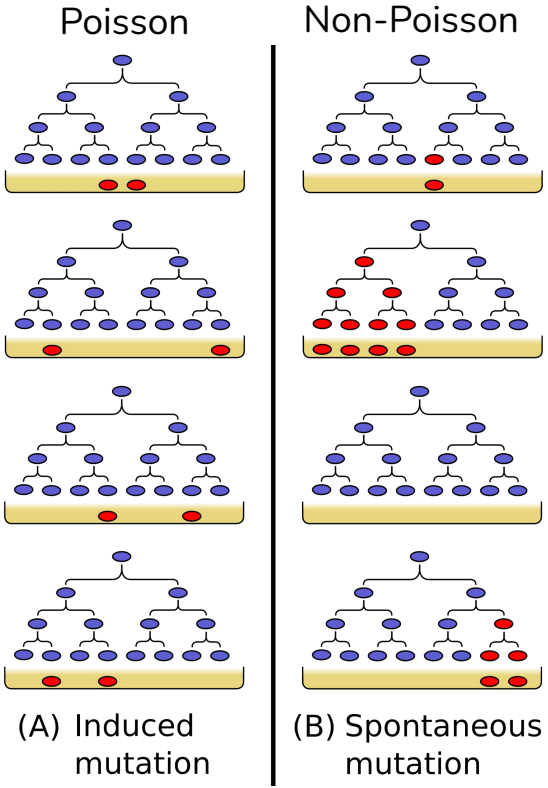


P(2 earthquakes in one month) = $e^{-2}(1^2/2!)$

```
# Plug lambda and k into the equation
1^2 * exp(-1) / factorial(2)
[1] 0.1839397

# Shortcut (dpois)
dpois(x=2,lambda=1)
[1] 0.1839397
```

Luria-Delbruck experiment

How do phage-resistant bacteria colonies arise?



Delbruck and Luria at CSHL

MUTATIONS OF BACTERIA FROM VIRUS SENSITIVITY TO VIRUS RESISTANCE[1,2]

S. E. LURIA[3] AND M. DELBRÜCK
Indiana University, Bloomington, Indiana, and Vanderbilt University, Nashville, Tennessee

Received May 29, 1943

"The distribution has been studied experimentally and has been found to conform with the conclusions drawn from the hypothesis that the **resistant bacteria arise by mutations of sensitive cells independently of the action of virus.**"

In other words, genetic mutations arise in the absence of selection, rather than being a response to selection.

https://en.wikipedia.org/wiki/Luria%E2%80%93Delbr%C3%BCck_experiment

Mutations at autosomal nucleotide sites are roughly 10^{-9} per year.

Consider a position in your genome. If you could trace its ancestry back across the last 10^9 years, **what is the probability that you would find no mutations?**

The expected number of mutations is $\lambda = ut$, where $u = 10^{-9}$ and $t = 10^9$. Thus, $\lambda = 1$. The probability of no mutations at the site you chose is:

$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

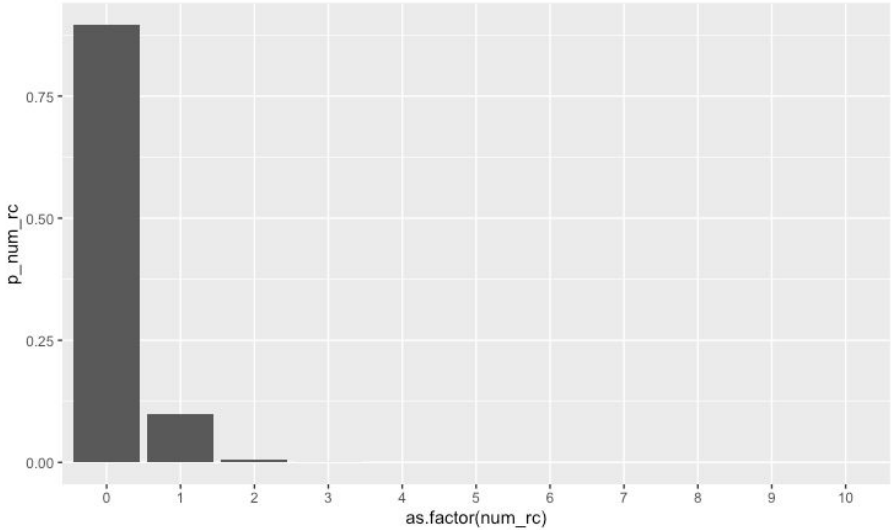
```
1^0 * exp(-1) / factorial(0)
[1] 0.3678794

dpois(x=0,lambda=1)
[1] 0.3678794
```

What about the probability of one or more mutations?

Alan Rogers

How many red cards do we expect in an English Premier League game?

Average of 0.11 bookings per game in the 2018/2019 season.




```
# install.packages("ggplot2")
library(ggplot2)
num_rc <- 0:10
p_num_rc <- dpois(0:10, lambda=0.11)
rc_prob = data.frame(num_rc, p_num_rc)
ggplot(rc_prob, aes(x=as.factor(num_rc), y=p_num_rc)) +
    geom_col()
```

https://www.gambling.com/online-betting/strategy/8-essential-premier-league-stats-10400

An intro to ggplot2

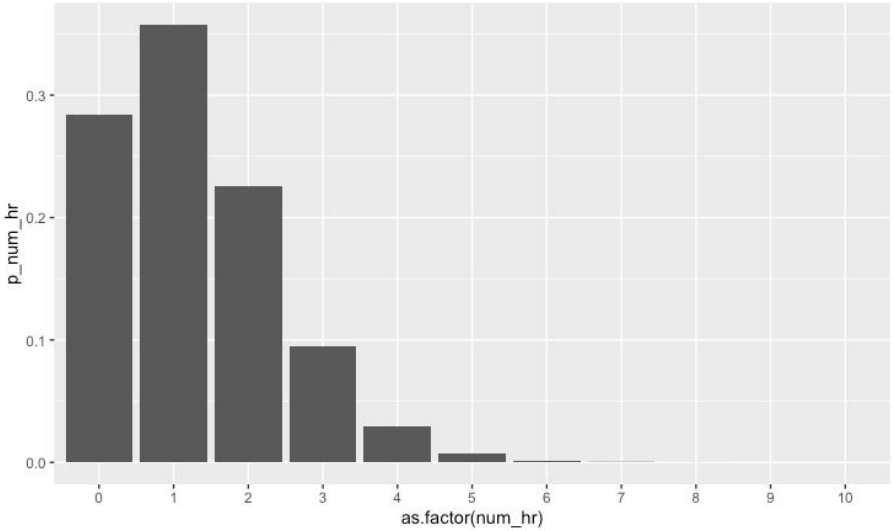
Slides:
https://docs.google.com/presentation/d/1T2KTEdk1h18oxic728BqI4UvW6P0OZyvP7VUZKXsM8E/edit

Video:

https://youtu.be/4G4mvSWvwWo

How many home runs do we expect in a Major League Baseball game?

Average #home runs in 2017 was 1.26 (a record).



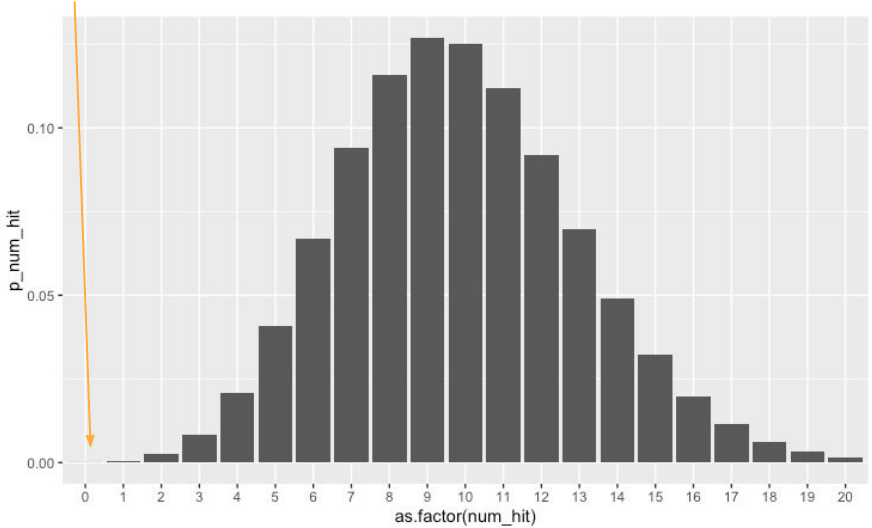

```
library(ggplot2)
num_hr <- 0:10
p_num_hr <- dpois(0:10, lambda=1.26)
home_run_prob = data.frame(num_hr, p_num_hr)
ggplot(home_run_prob, aes(x=as.factor(num_hr), y=p_num_hr))
    + geom_col()
```

https://www.boston.com/sports/mlb/2019/05/02/mlb-home-runs-strikeouts-record

How many hits do we expect a team to have in a Major League Baseball game?

Highest average hits per game in 2019 is 9.85 (as of 8/21/2019).

"No hitters" are rare

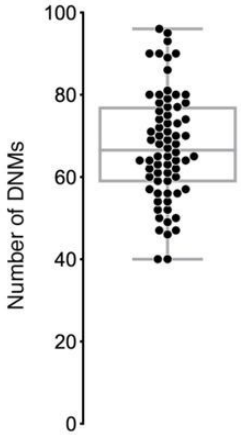



```
library(ggplot2)
num_hit <- 0:20
p_num_hit <- dpois(0:20, lambda=9.85)
hit_prob = data.frame(num_hit, p_num_hit)
ggplot(hit_prob, aes(x=as.factor(num_hit), y=p_num_hit))
    + geom_col()
```

https://www.teamrankings.com/mlb/stat/hits-per-game

How many coding, de novo mutations do we expect per exome?

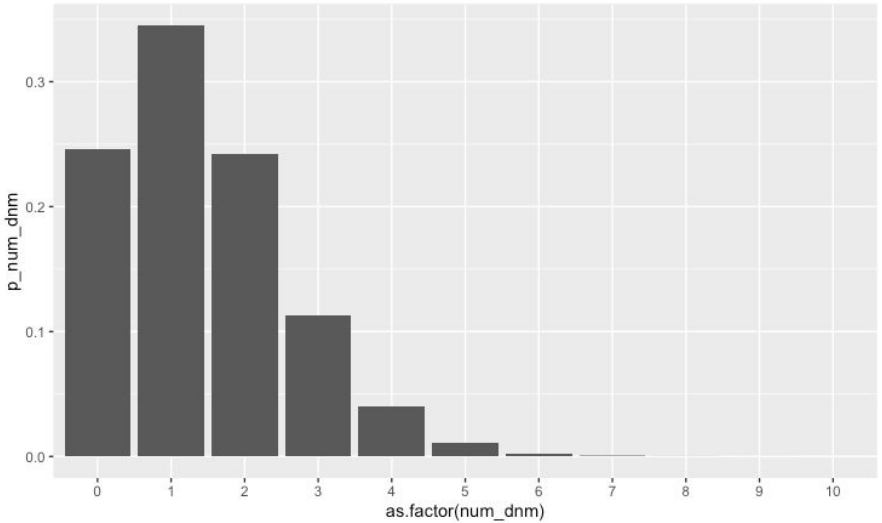
70.1 de novo mutations / offspring genome-wide



X 0.02 (exome is ~2% of genome)

= **1.402**

biorxiv.org/content/10.1101/552117v2.full

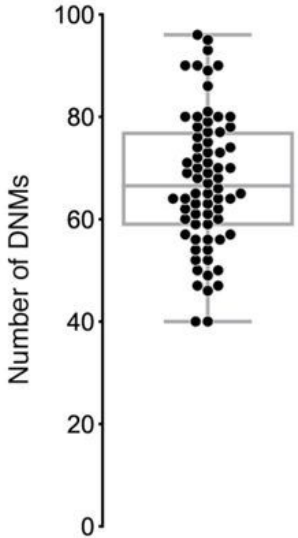



```
library(ggplot2)
num_dnm <- 0:10
p_num_dnm <- dpois(0:10, lambda=1.402)
dnm_prob = data.frame(num_dnm, p_num_dnm)
ggplot(dnm_prob, aes(x=as.factor(num_dnm), y=p_num_dnm)) +
    geom_col()
```

What is the probability of 1 or more coding mutations?

What is the expected number of de novo mutations in randomly chosen, 1 megabase chunks of a human genome?

70.1 de novo mutations / offspring genome-wide



biorxiv.org/content/10.1101/552117v2.full

Is the number of chocolate chunks in a cookie a Poisson random variable?

Count the visible chunks. Enter the count here.

Chip count data

Load our observed data into R.

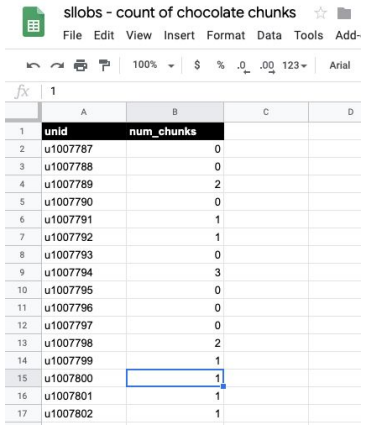
1

Install the "datapasta" package

```
install.packages("datapasta")
library(datapasta)
```

2

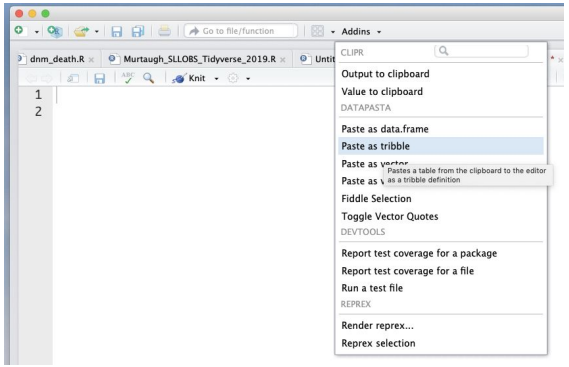
Copy the data from google sheets



	A	B	C	D
1	unid	num_chunks		
2	u1007787	0		
3	u1007788	0		
4	u1007789	2		
5	u1007790	0		
6	u1007791	1		
7	u1007792	1		
8	u1007793	0		
9	u1007794	3		
10	u1007795	0		
11	u1007796	0		
12	u1007797	0		
13	u1007798	2		
14	u1007799	1		
15	u1007800	1		
16	u1007801	1		
17	u1007802	1		

3

Use the "Addins" dropdown to paste as tribble



4

Create a data frame from the generated code

Plot our observed data into R.

4 Create a tibble from the generated code (below is an example)

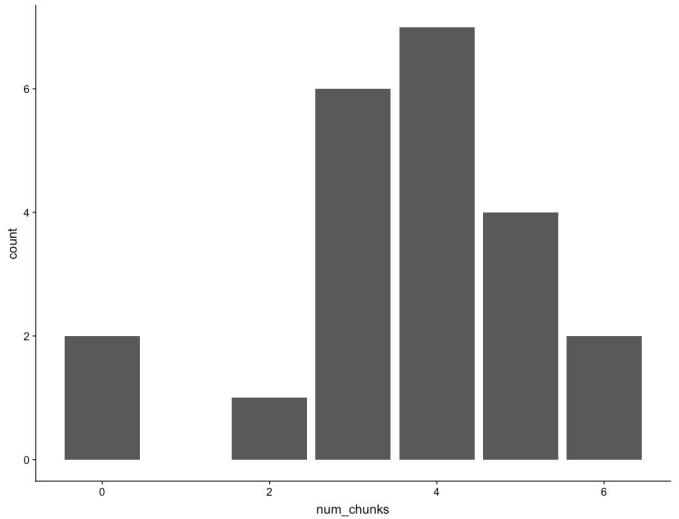
```
chunks <- tibble::tribble(
       ~unid, ~num_chunks,
  "u1007787",           0,
  "u1007788",           0,
  "u1007789",           2,
  "u1007790",           0,
  "u1007791",           1,
  "u1007792",           1,
  "u1007793",           0,
  "u1007794",           3,
  "u1007795",           0,
  "u1007796",           0,
  "u1007797",           0,
  "u1007798",           2,
  "u1007799",           1,
  "u1007800",           1,
  "u1007801",           1,
  "u1007802",           1,
  "u1007803",           0,
  "u1007804",           1,
  "u1007805",           1
  )
```

5 Look at the freqs with table()

```
table(chunks$num_chunks)
```

6 Plot the number of chips per cookie

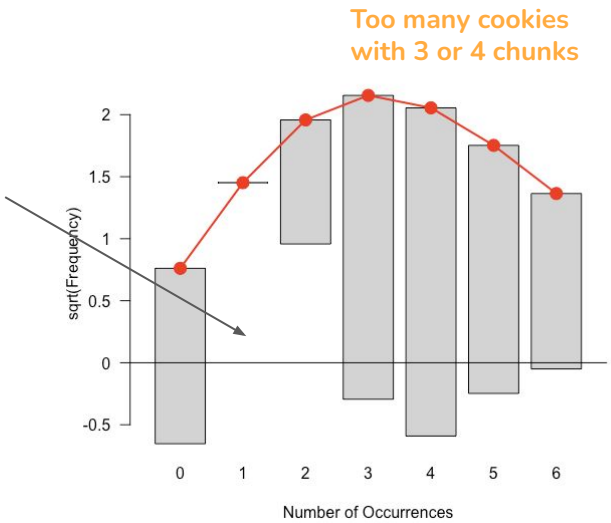
```
library(ggplot2)
library(cowplot)
ggplot(chunks,aes(x=num_chunks)) +
  geom_bar()
```



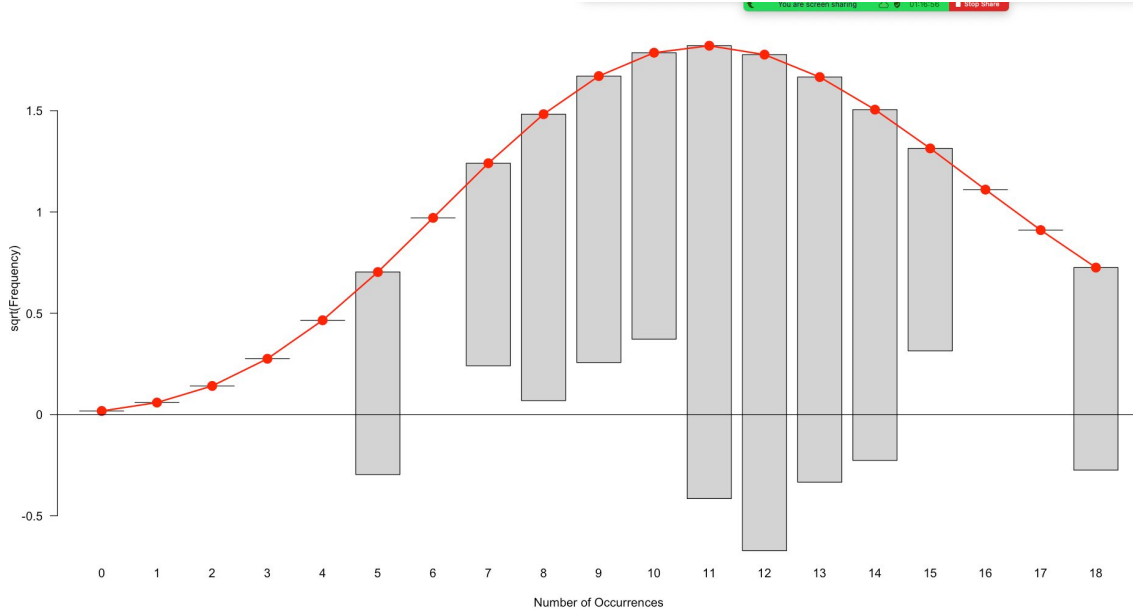
Is the number of chocolate chunks in a cookie a Poisson random variable?

- We want to know the "goodness of fit". That is, which theoretical distribution fits our data best? Is it Poisson?
- The "vcd" (Visualizing Categorical Data) R package has two very helpful functions: "goodfit" and "rootogram" for testing this.

```
install.packages("vcd")
library("vcd")
gf <- goodfit(chunks$num_chunks, "poisson")
rootogram(gf)
```

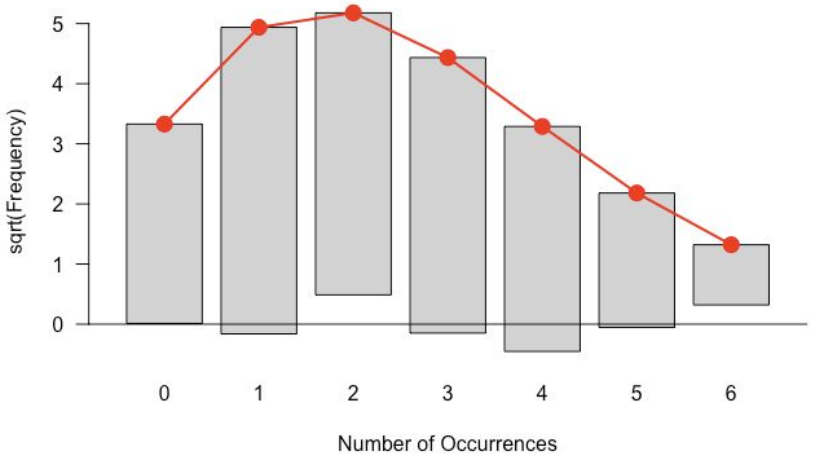


https://www.huber.embl.de/msmb/Chap-Models.html

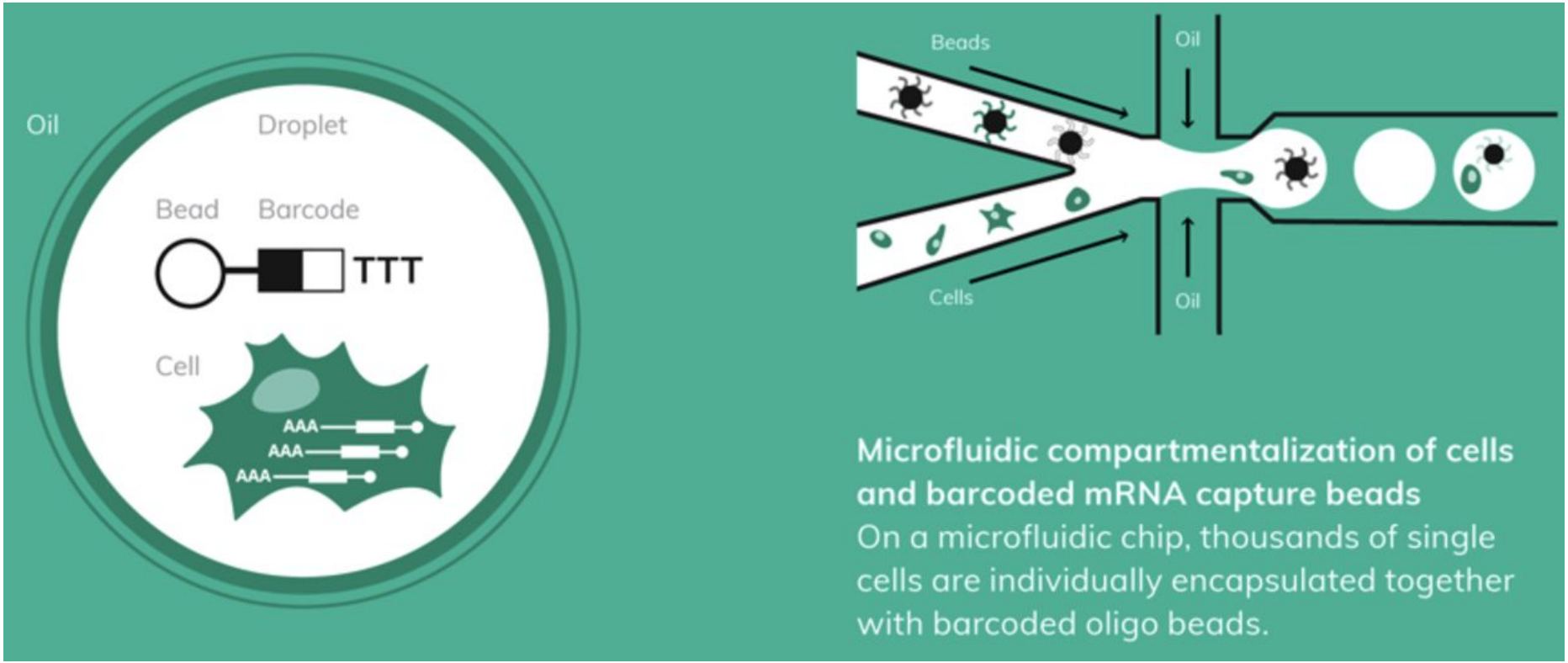

sqrt(Frequency)
1.5
1
0.5
0
-0.5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Number of Occurrences

Control: what does the fit look like with data simulated from a Poisson random variable?

```
# simulate 1000 cookies with a Poisson
# distributed number of chunks. The mean
# is 3 chunks per cookie
sim_chunks <- rpois(1000, lambda=3)
gf2 <- goodfit(sim_chunks, "poisson")
rootogram(gf2)
```

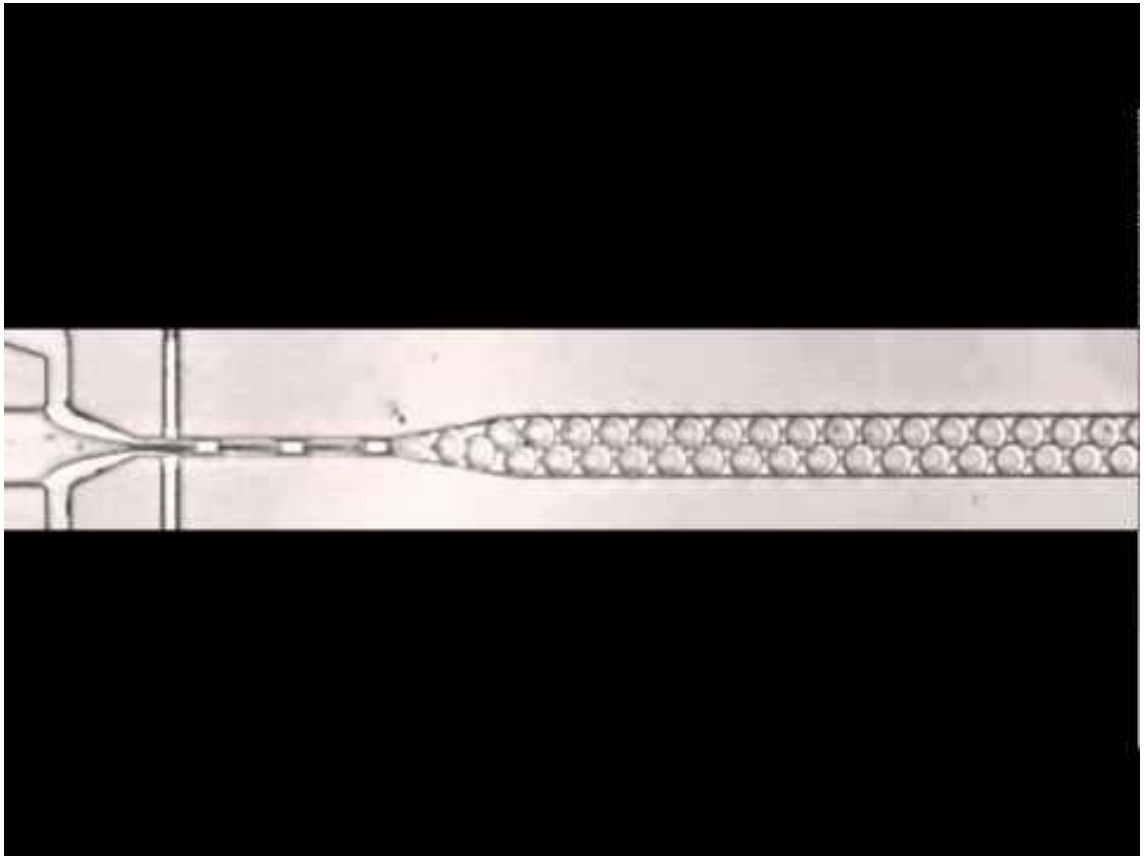


Drop-seq (single cell sequencing in droplets)



https://www.dolomite-bio.com/wp-content/uploads/Drop-Seq-Workflow.png

Drop-seq (single cell sequencing in droplets)

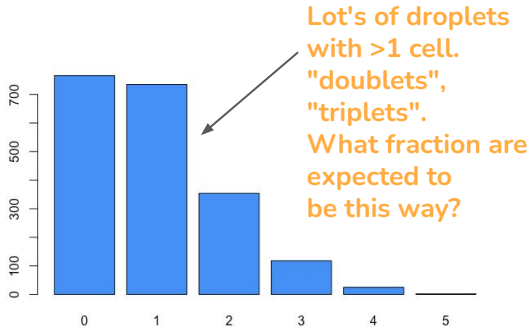


https://www.youtube.com/watch?time_continue=3&v=usK71SG30t0

Goal: one cell per droplet

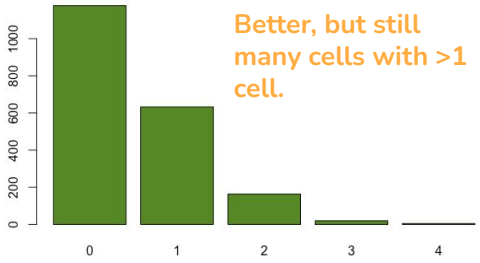
Should one load shooting for an average of one cell per droplet (lambda = 1)? Let's simulate.

```
sc_sim <- rpois(2000, lambda=1)
barplot(table(sc_sim),
col="dodgerblue")
```



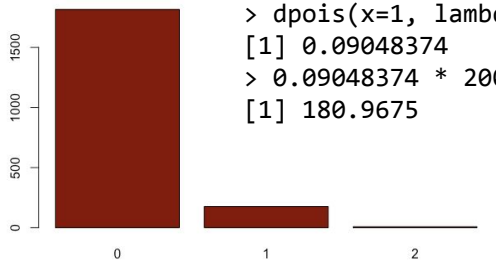
What about (lambda = 0.5)?

```
sc_sim <- rpois(2000, lambda=0.5)
barplot(table(sc_sim), col="chartreuse4")
```



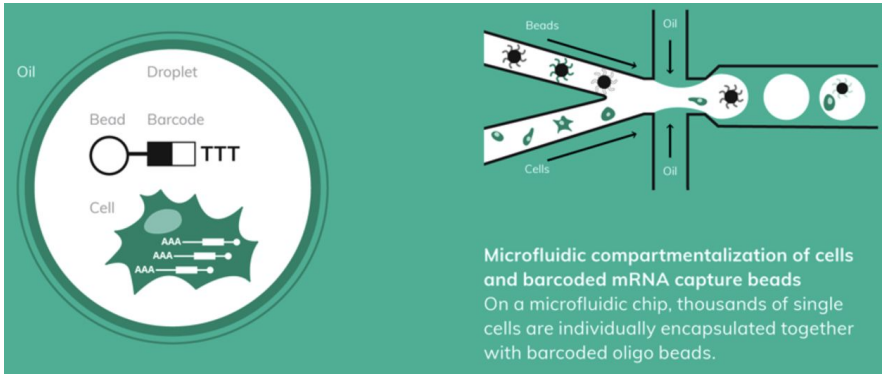
What about (lambda = 0.1)?

```
sc_sim <- rpois(2000, lambda=0.1)
barplot(table(sc_sim), col="darkred")
```

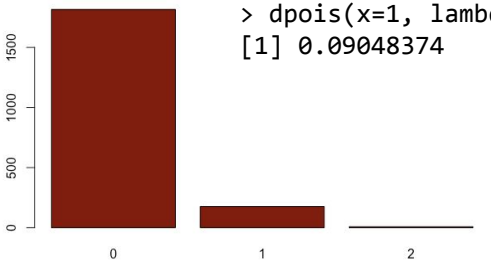



```
> dpois(x=1, lambda=0.1)
[1] 0.09048374
> 0.09048374 * 2000
[1] 180.9675
```

Goal: one cell and one bead per droplet



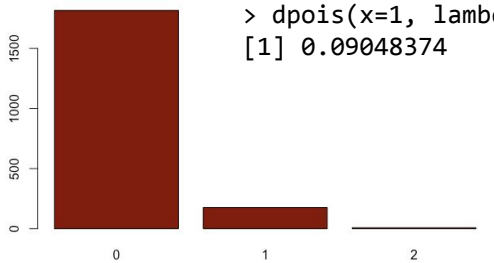
Lamdba = 0.1 to minimize
>1 **cell** per droplet




```
> dpois(x=1, lambda=0.1)
[1] 0.09048374
```

X

Lamdba = 0.1 to minimize
>1 **bead** per droplet

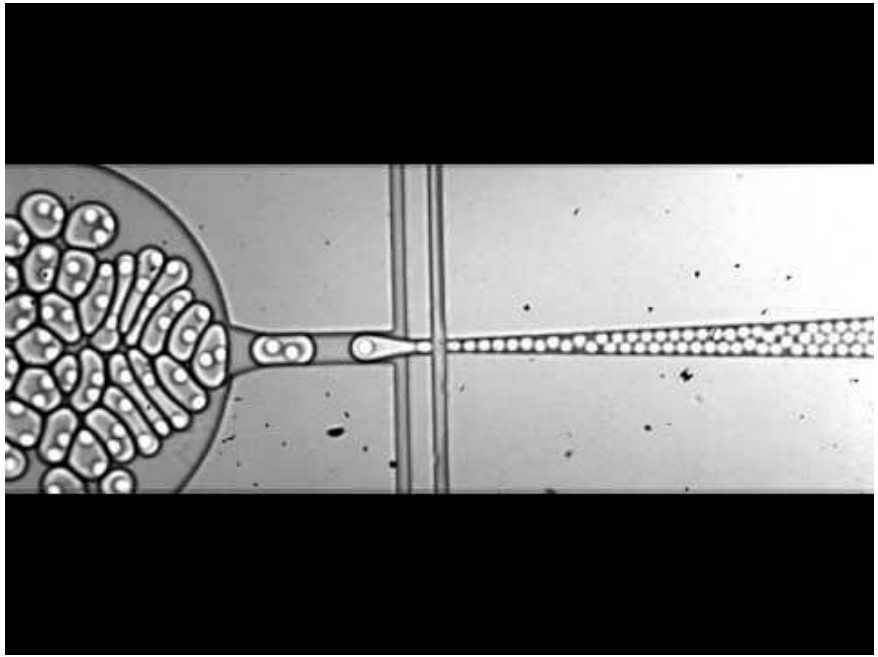



```
> dpois(x=1, lambda=0.1)
[1] 0.09048374
```

= 0.0081

Why is single-cell "sub-Poisson"?

https://liorpachter.wordpress.com/2019/02/07/sub-poisson-loading-for-single-cell-rna-seq/



Bulk RNA expression

Single-cell RNAseq

Bulk RNAseq