

# Statistics of cancer incidence vs age

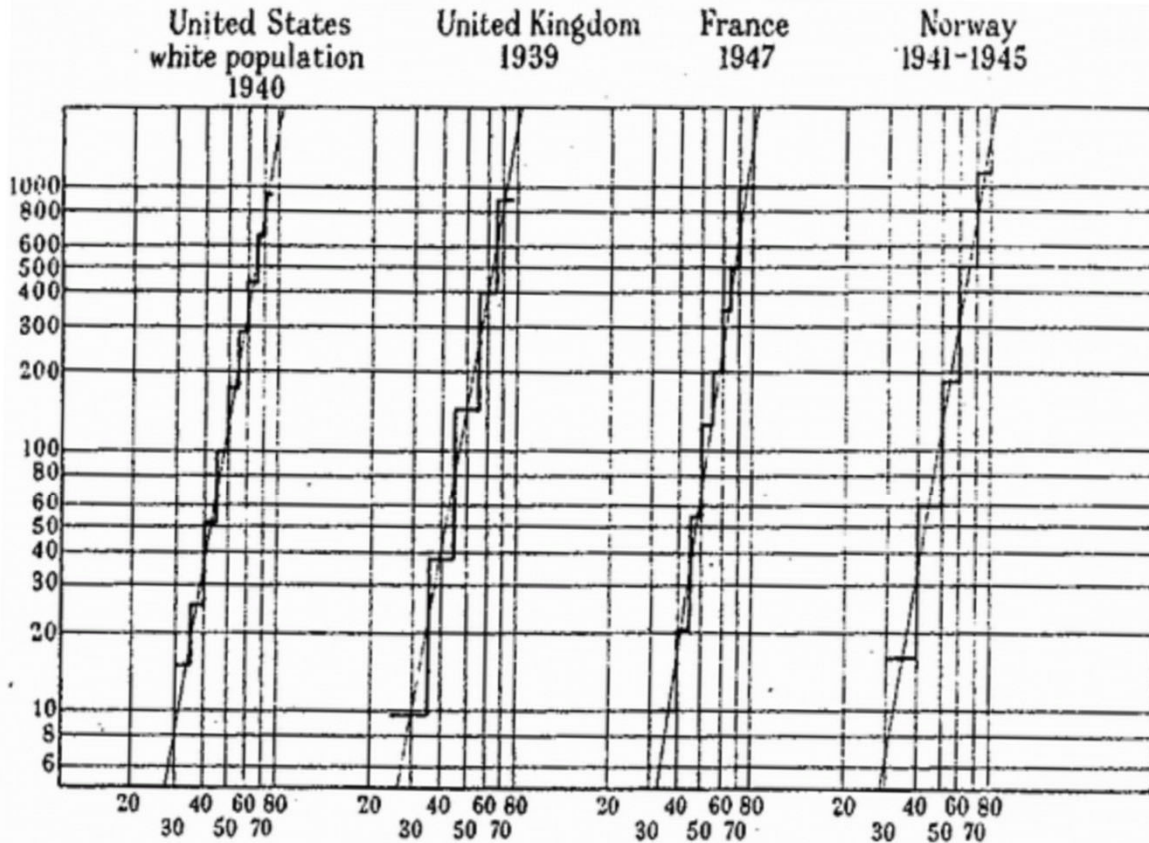


FIG. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

Multi-mutation theory of cancer:  
 Carl O. Nordling (British J. of  
 Cancer, March 1953):

Cancer death rate  
 $\sim (\text{patient age})^6$

It suggests the  
 existence of  
 $k=7$  driver genes

$$P(T_{\text{cancer}} \leq t) \sim (u_1 t)(u_2 t) \dots (u_k t) \sim u_1 u_2 \dots u_k t^k$$

$$P(T_{\text{cancer}} = t) \sim \frac{d}{dt} (u_1 t)(u_2 t) \dots (u_k t) \sim k u_1 u_2 \dots u_k t^{k-1}$$

# How many driver gene mutations for different types of cancer?

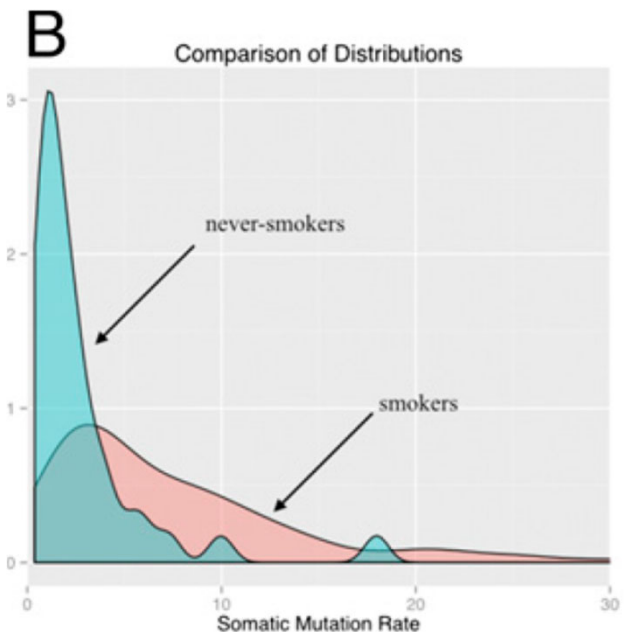
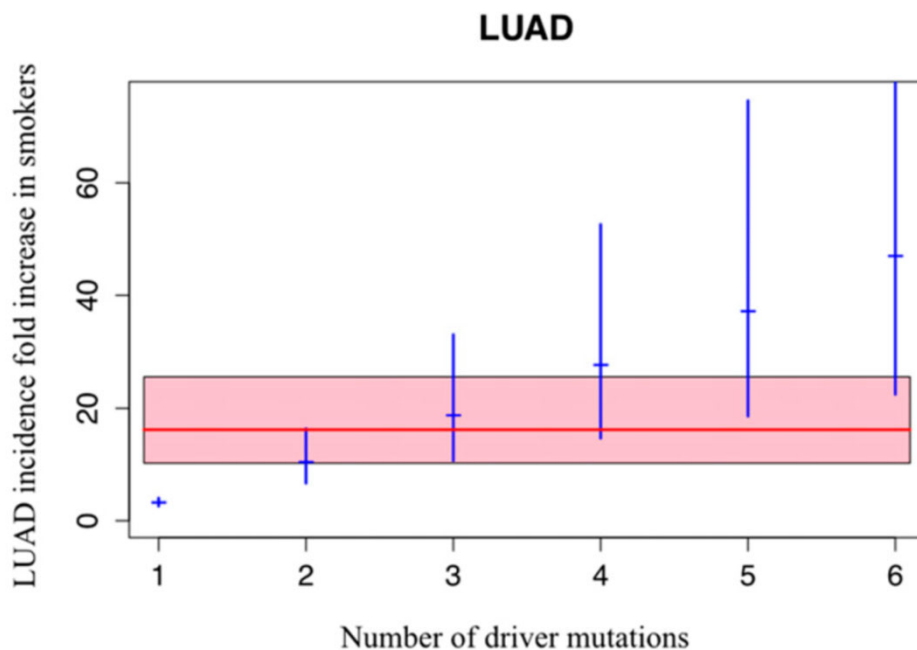
## Only three driver gene mutations are required for the development of lung and colorectal cancers

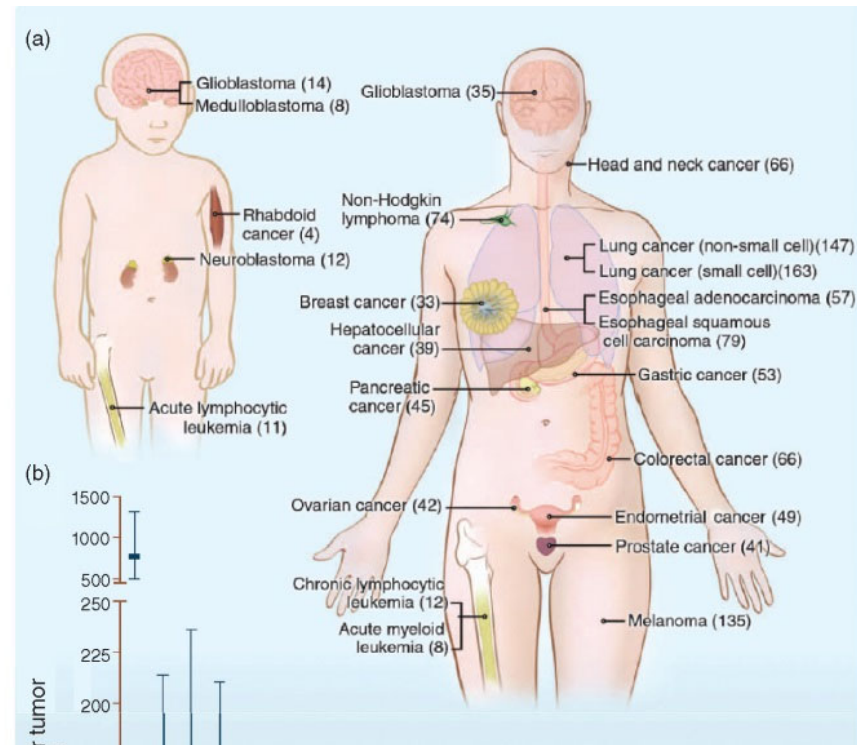
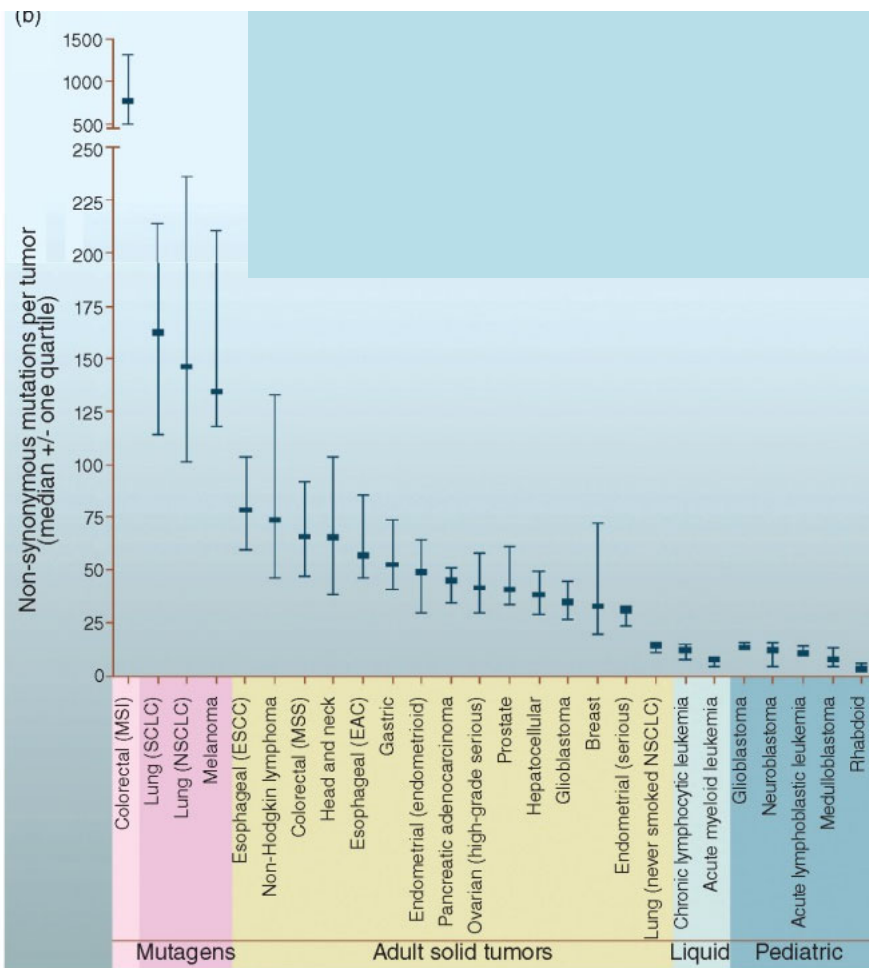
Cristian Tomasetti<sup>a,b,1</sup>, Luigi Marchionni<sup>c</sup>, Martin A. Nowak<sup>d</sup>, Giovanni Parmigiani<sup>e</sup>, and Bert Vogelstein<sup>f,g,1</sup>

<sup>a</sup>Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, and <sup>b</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; <sup>c</sup>Cancer Biology Program, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; <sup>d</sup>Program for Evolutionary Dynamics, Department of Mathematics, Harvard University, Cambridge, MA 02138; <sup>e</sup>Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02215; and <sup>f</sup>Ludwig Center for Cancer Genetics and Therapeutics and <sup>g</sup>Howard Hughes Medical Institute, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205

Contributed by Bert Vogelstein, November 21, 2014 (sent for review July 31, 2014; reviewed by Zvia Agur)

Smokers have 3.23 times more mutations in lungs





**FIGURE 21.10** Somatic mutations in representative human cancers, based on genome-wide sequencing studies. (a) The genomes of adult (right) and pediatric (left) cancers are represented. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein *et al.* (2013). Reproduced with permission from AAAS. (b) Median number of nonsynonymous substitutions per tumor. Horizontal bars indicate the 25% and 75% quartiles. MSI: microsatellite instability; SCLC: small cell lung cancers; NSCLC: non-small cell lung cancers; ESCC: esophageal squamous cell carcinomas; MSS: microsatellite stable; EAC: esophageal adenocarcinomas.

*Bioinformatics and Functional Genomics*, Third Edition, Jonathan Pevsner.  
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.  
 Companion Website: [www.wiley.com/go/pevsnerbioinformatics](http://www.wiley.com/go/pevsnerbioinformatics)

- Cancer cells carry both **“Driver”** and **“Passengers”** mutations
- **Passenger** mutations cause **little to no harm** (see later for how even little harm matters)
- Both are common as **cancers** **elevate mutation rate**

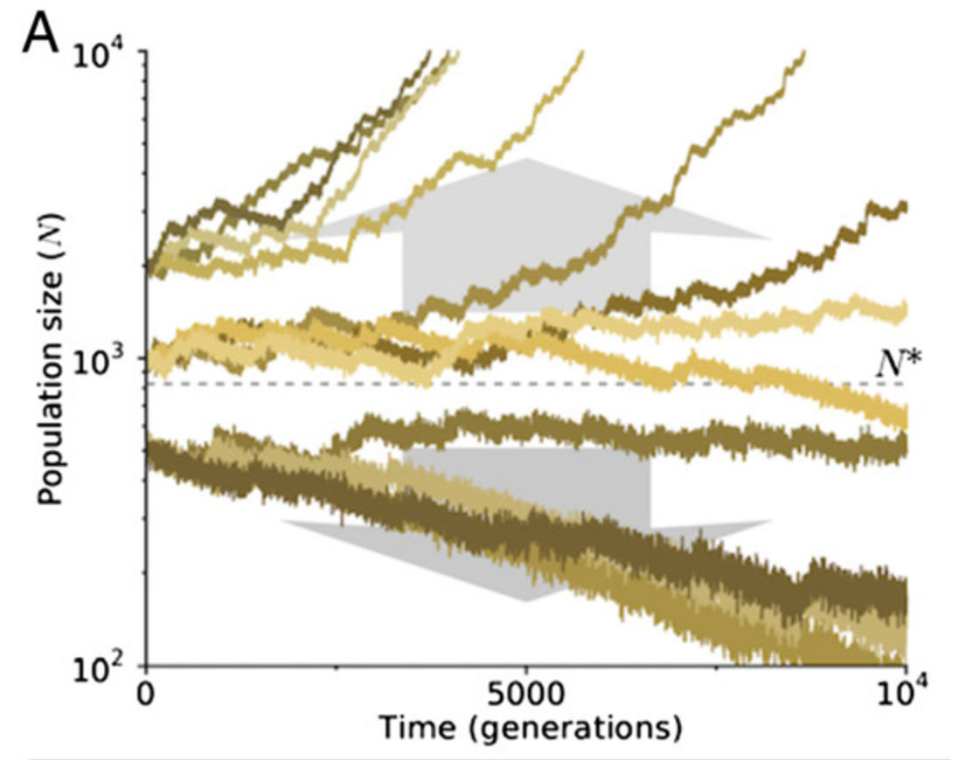
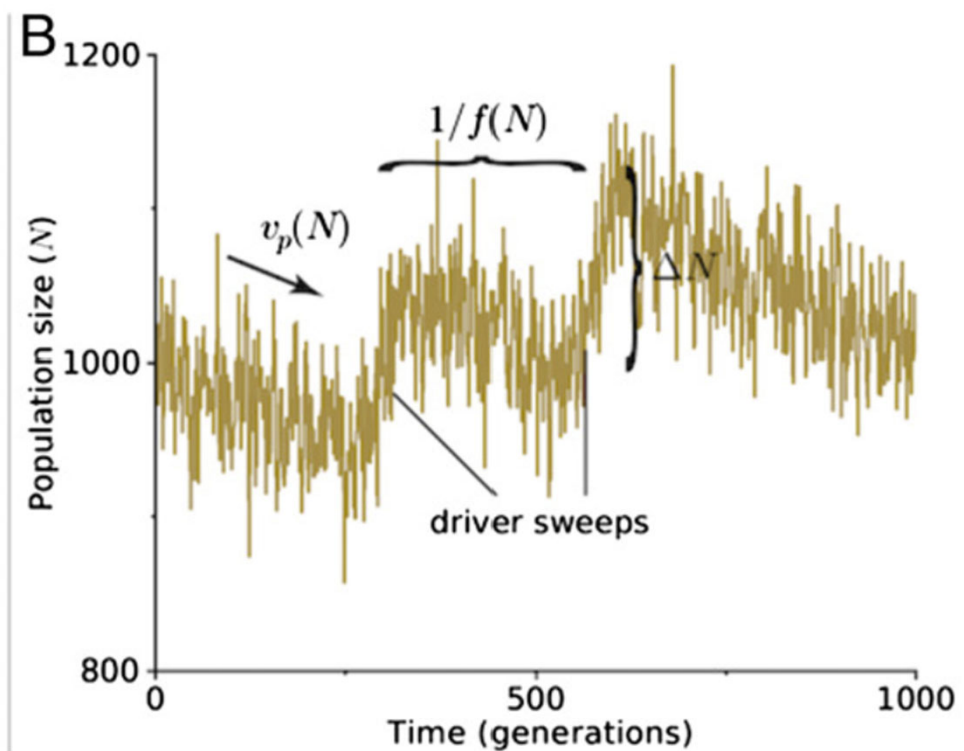
# Number of passenger+driver mutations follows negative binomial distribution

- What is the **probability** to have  $n_p$  **passenger mutations** or  $(n_p+k)$  **total mutations** by the time you are diagnosed with cancer requiring  $k$  **driver mutations**?
- Let  $p$  is the probability that a mutation is **a driver** ( $p = \text{Genome\_target\_of\_driv} / (\text{Genome\_target\_of\_driv} + \text{Genome\_target\_of\_pass})$ )  
**(1-p)** – it is **a passenger mutation**

$$P(n_p + k | p, k) = \binom{n_p + k - 1}{n_p} (1 - p)^{n_p} p^k$$

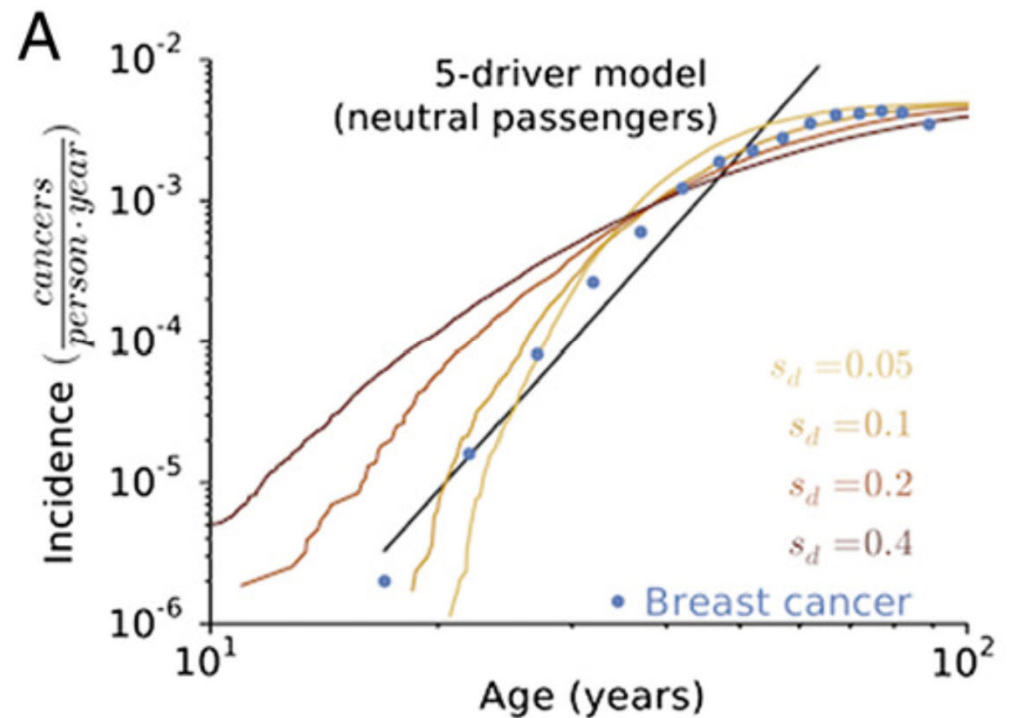
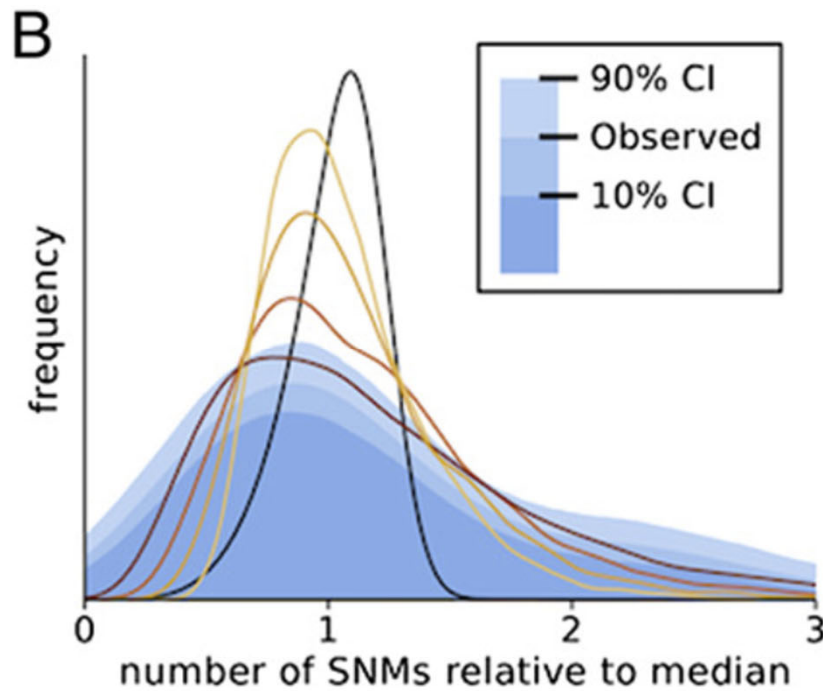


# What if passenger mutations slow down the growth of cancer tumors?



McFarland CD, Mirny L, Korolev KS, PNAS 2014

# Can we prove/quantify it using statistics?



Assume: growth rate of cancer =  $(1+s_d)^{N_d} / (1+s_p)^{N_p}$

$\mu = 10^{-8}$ ,  $\text{Target}_d = 1,400$ ,  $\text{Target}_p = 10^7$ ,  $s_d = 0.05$  to  $0.4$ ,  $s_p = 0.001$

$s_p/s_d$  for breast:  $0.0060 \pm 0.0010$ ;

melanoma:  $0.016 \pm 0.003$ ; lung:  $0.0094 \pm 0.0093$ ;

Blue - data on breast cancer: incidence; non-synonymous mutations



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 AUTOCOMLETE



WHY ARE THERE WEEKS  
WHY DO I FEEL DIZZY

WHY AREN'T ECONOMISTS RICH  
WHY DO AMERICANS CALL IT SOCCER  
WHY ARE MY EARS RINGING  
WHY ARE THERE SO MANY AVENGERS  
WHY ARE THE AVENGERS FIGHTING THE X MEN  
WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE SWARMS OF GNATS  
WHY IS THERE PHLEGM  
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN  
WHY IS PSYCHIC WEAK TO BUG  
WHY DO CHILDREN GET CANCER  
WHY IS POSEIDON ANGRY WITH ODYSSEUS  
WHY IS THERE ICE IN SPACE

# WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED  
WHY IS SPACE BLACK  
WHY IS OUTER SPACE SO COLD  
WHY ARE THERE PYRAMIDS ON THE MOON  
WHY IS NASA SHUTTING DOWN



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

WHY DO WHALES JUMP  
WHY ARE WITCHES GREEN  
WHY ARE THERE MIRRORS ABOVE BEDS  
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 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 ARE THERE BRIDESMAIDS  
WHY DO DYING PEOPLE REACH UP  
WHY AREN'T THERE VARICOSE ARTERIES  
WHY ARE OLD KUNGONS DIFFERENT  
WHY IS THERE HELL IF GOD FORGIVES  
WHY IS THERE NO GPS IN LAPTOPS  
WHY DO KNEES CLICK  
WHY AREN'T THERE E GRADES  
WHY IS ISOLATION BAD  
WHY DO BOYS LIKE ME  
WHY DON'T BOYS LIKE ME  
WHY IS THERE ALWAYS A JAVA UPDATE  
WHY ARE THERE RED DOTS ON MY THIGHS  
WHY IS LYING GOOD



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 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 ARE ULTRASOUNDS IMPORTANT  
WHY ARE ULTRASOUND MACHINES EXPENSIVE  
WHY IS STEALING WRONG

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

# Important terms & concepts for discrete random variables

- Probability Mass Function (PMF)
- Cumulative Distribution Function (CDF)
- Complementary Cumulative Distribution Function (CCDF)
- Expected value
- Mean
- Variance
- Standard deviation

**Boldface and underlined** are the same for continuous distributions



Which distribution is this?

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

- A. Uniform
- B. Binomial
- C. Geometric
- D. Negative Binomial
- E. Poisson

Get your i-clickers

Which distribution is this?

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

A. Uniform

B. Binomial

C. Geometric

D. Negative Binomial

E. Poisson

Get your i-clickers

Which distribution is this?

$$\binom{x-1}{r-1} (1-p)^{x-r} p^r$$

- A. Uniform
- B. Binomial
- C. Geometric
- D. Negative Binomial
- E. Poisson

Get your i-clickers



Which distribution is this?

$$\binom{x-1}{r-1} (1-p)^{x-r} p^r$$

- A. Uniform
- B. Binomial
- C. Geometric
- D. Negative Binomial**
- E. Poisson

Get your i-clickers

Which distribution is this?

$$\frac{e^{-\lambda} \lambda^x}{x!}$$

- A. Uniform
- B. Binomial
- C. Geometric
- D. Negative Binomial
- E. Poisson

Get your i-clickers

Which distribution is this?

$$\frac{e^{-\lambda} \lambda^x}{x!}$$

- A. Uniform
- B. Binomial
- C. Geometric
- D. Negative Binomial
- E. Poisson

Get your i-clickers



Name	Probability Distribution	Mean	Variance
<b>Discrete</b>			
Uniform	$\frac{1}{n}, a \leq b$	$\frac{(b + a)}{2}$	$\frac{(b - a + 1)^2 - 1}{12}$
Binomial	$\binom{n}{x} p^x (1 - p)^{n-x},$ $x = 0, 1, \dots, n, 0 \leq p \leq 1$	$np$	$np(1 - p)$
Geometric	$(1 - p)^{x-1} p,$ $x = 1, 2, \dots, 0 \leq p \leq 1$	$1/p$	$(1 - p)/p^2$
Negative binomial	$\binom{x - 1}{r - 1} (1 - p)^{x-r} p^r$ $x = r, r + 1, r + 2, \dots, 0 \leq p \leq 1$	$r/p$	$r(1 - p)/p^2$
Poisson	$\frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots, 0 < \lambda$	$\lambda$	$\lambda$

# What distributions we learn

- Uniform distribution
- Bernoulli distribution/trial
- Binomial distribution
- Poisson distribution
- Geometric distribution
- Negative binomial distribution

Why do we need to know  
these simple distributions?

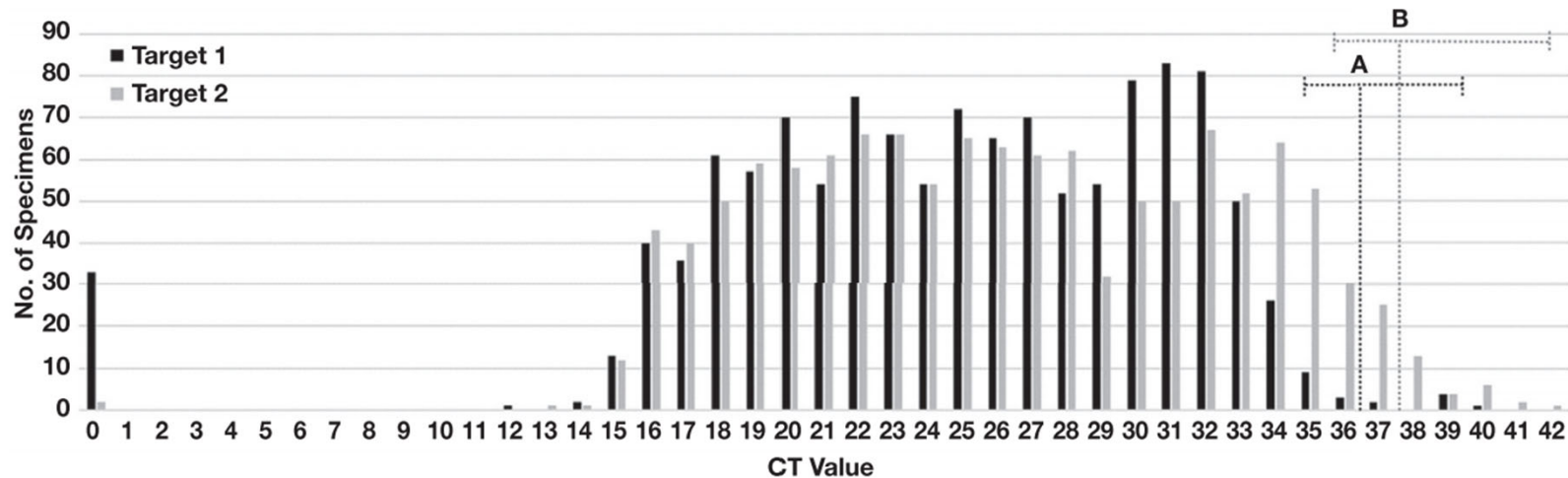


# Ways to use statistics

- **To process your experimental data**
  - What do you need? Mean, Variance, Standard deviation. **No need to know any textbook distributions**
- **To plan experiments**
  - **Need to know distributions**, e.g., Poisson to plan how much redundancy to use for genome assembly
- **To learn biological processes behind your data**
  - **Need to know distributions** to compare empirical distributions in your data to what you expect based on a simple hypothesis

# Uniform distribution

# Examples of uniform distribution: Ct value of PCR test of a virus



**Figure 3** Distribution of cycle threshold (CT) values. The total number of specimens with indicated CT values for Target 1 and 2 are plotted. The estimated limit of detection for (A) Target 1 and (B) Target 2 are indicated by vertical dotted lines. Horizontal dotted lines encompass specimens with CT values less than 3x the LoD for which sensitivity of detection may be less than 100%. This included 19/1,180 (1.6%) reported CT values for Target 1 and 81/1,211 (6.7%) reported CT values for Target 2. Specimens with Target 1 or 2 reported as “not detected” are denoted as a CT value of “0.”

## Distribution of SARS-CoV-2 PCR Cycle Threshold Values Provide Practical Insight Into Overall and Target-Specific Sensitivity Among Symptomatic Patients

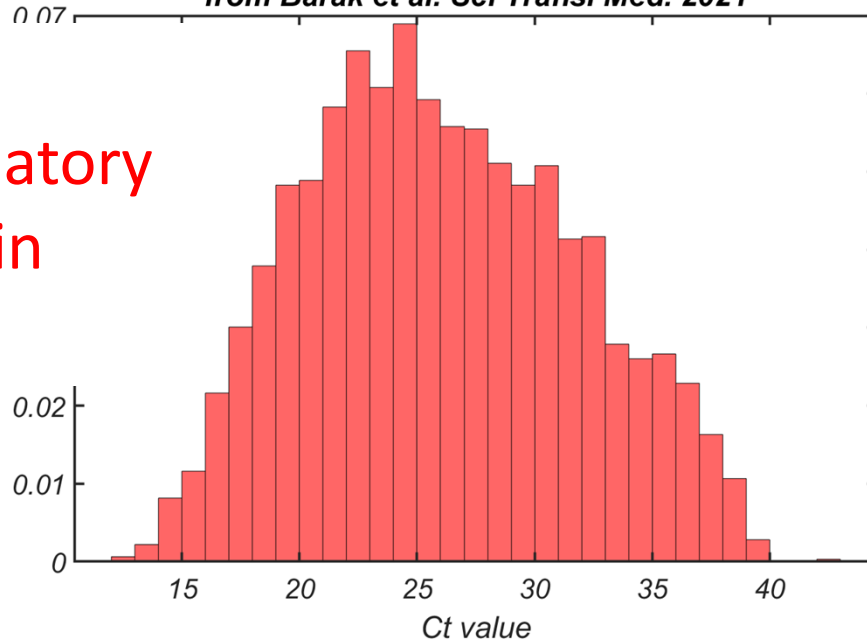
Blake W Buchan, PhD, Jessica S Hoff, PhD, Cameron G Gmehlin, Adriana Perez, Matthew L Faron, PhD, L Silvia Munoz-Price, MD, PhD, Nathan A Ledebor, PhD *American Journal of Clinical Pathology*, Volume 154, Issue 4, 1 October 2020,

<https://academic.oup.com/ajcp/article/154/4/479/5873820>

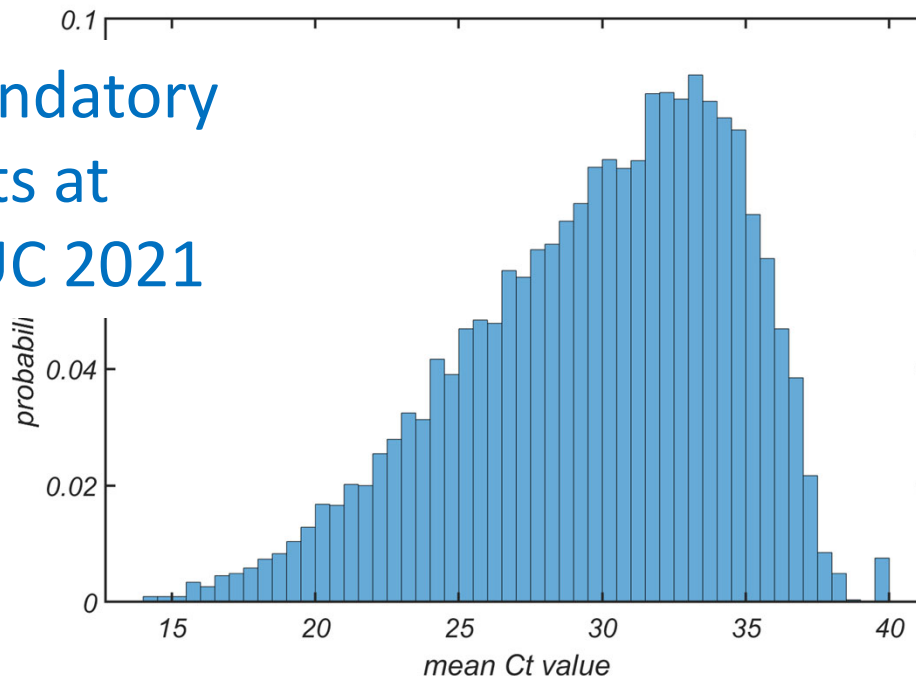
# Why should we care?

3191 individual positive tests  
from Barak et al. *Sci Transl Med.* 2021

Non-  
mandatory  
tests in  
Israel



Mandatory  
tests at  
UIUC 2021



- High Ct value means we identified the infected individual early, hopefully before transmission to others
- When testing is mandatory, and people are tested frequently – Ct value is skewed towards high values

# Negative binomial distribution

# Statistics of cancer incidence vs age

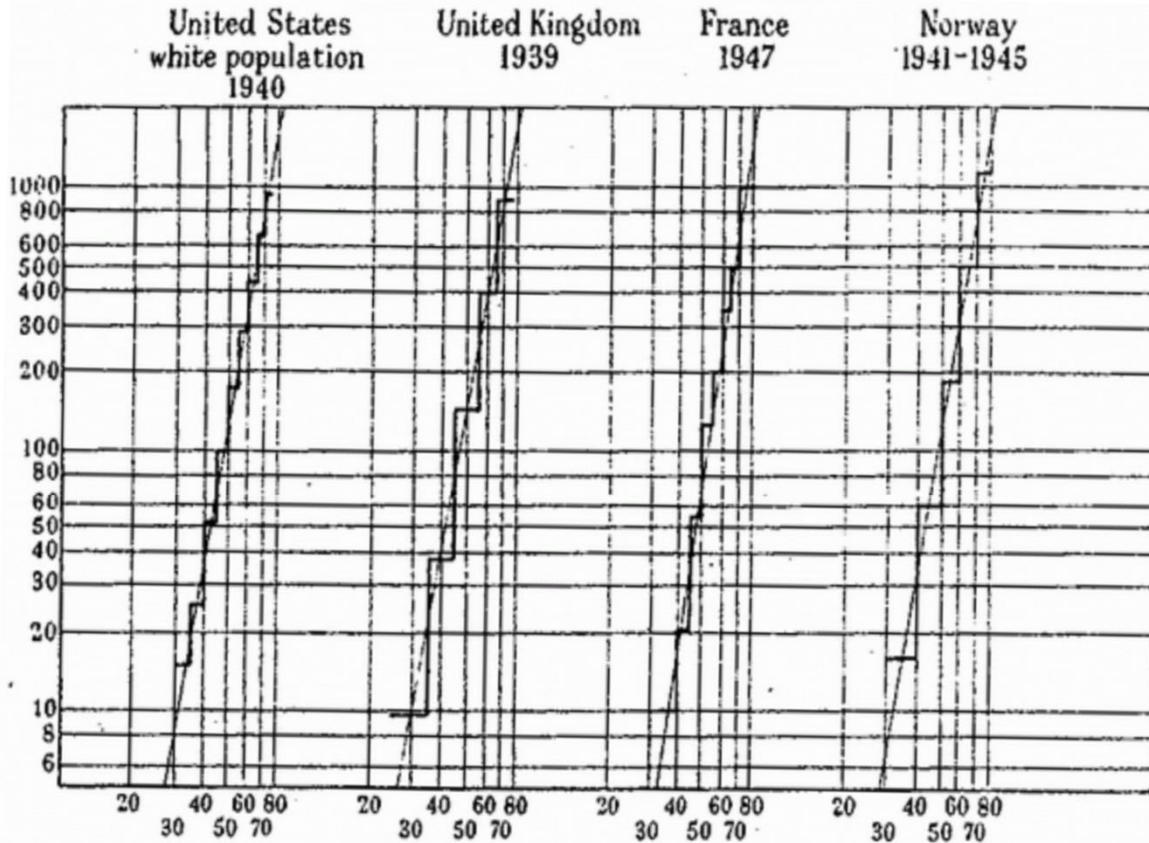


FIG. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

Multi-mutation theory of cancer:  
Carl O. Nordling (British J. of  
Cancer, March 1953):

Cancer death rate  
 $\sim (\text{patient age})^6$

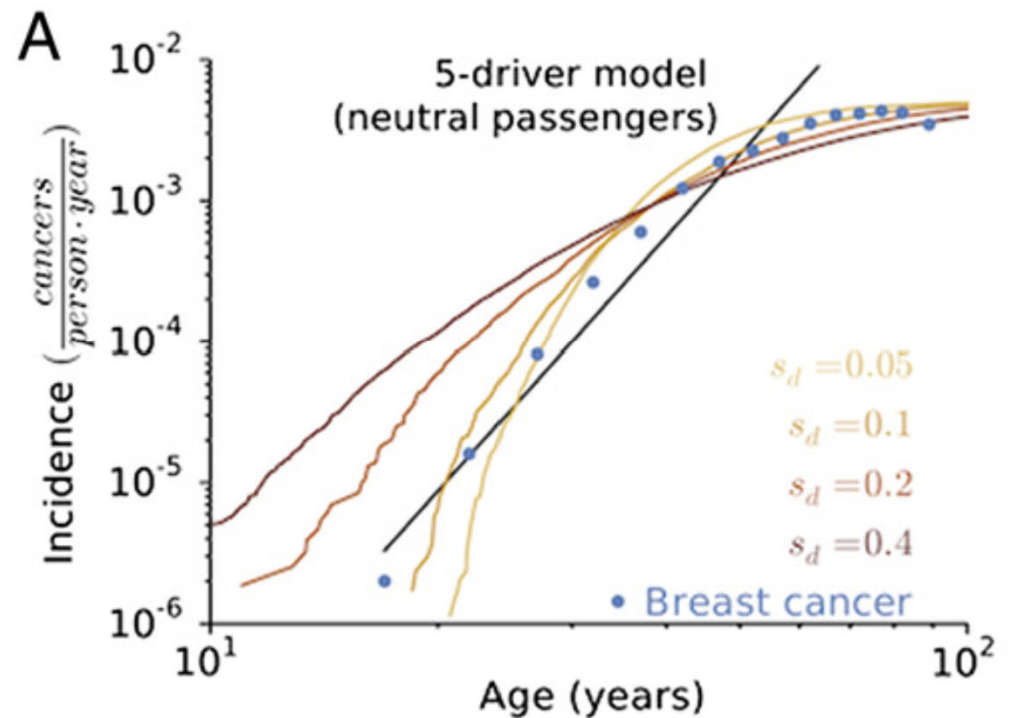
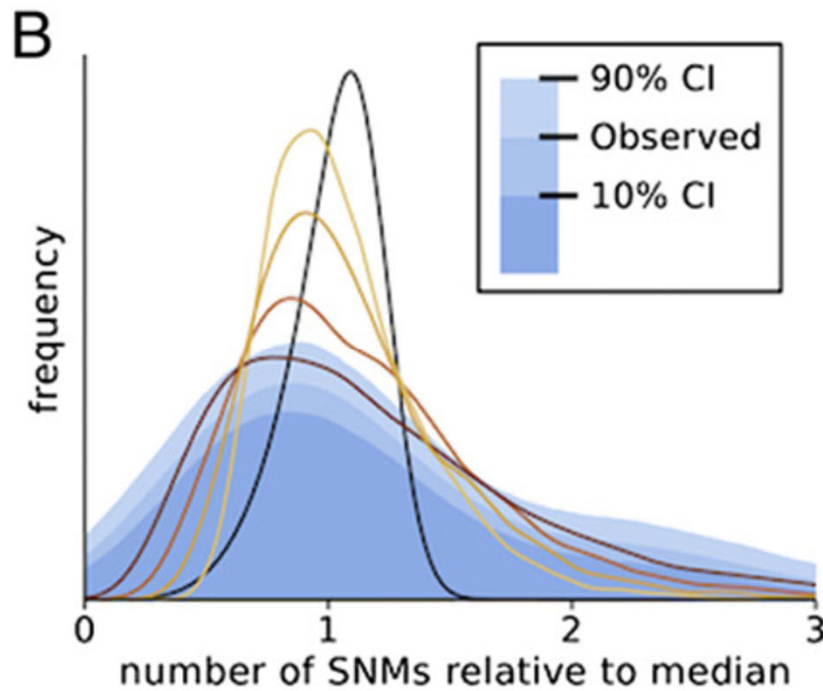
It suggests the  
existence of  
 $k=7$  driver genes

$$P(T_{\text{cancer}} \leq t) \sim (u_1 t)(u_2 t) \dots (u_k t) \sim u_1 u_2 \dots u_k t^k$$

$$P(T_{\text{cancer}} = t) \sim \frac{d}{dt} (u_1 t)(u_2 t) \dots (u_k t) \sim k u_1 u_2 \dots u_k t^{k-1}$$



# Can we prove/quantify it using statistics?



Assume: growth rate of cancer =  $(1+s_d)^{N_d} / (1+s_p)^{N_p}$

$\mu = 10^{-8}$ ,  $\text{Target}_d = 1,400$ ,  $\text{Target}_p = 10^7$ ,  $s_d = 0.05$  to  $0.4$ ,  $s_p = 0.001$

$s_p/s_d$  for breast:  $0.0060 \pm 0.0010$ ;

melanoma:  $0.016 \pm 0.003$ ; lung:  $0.0094 \pm 0.0093$ ;

Blue - data on breast cancer: incidence; non-synonymous mutations

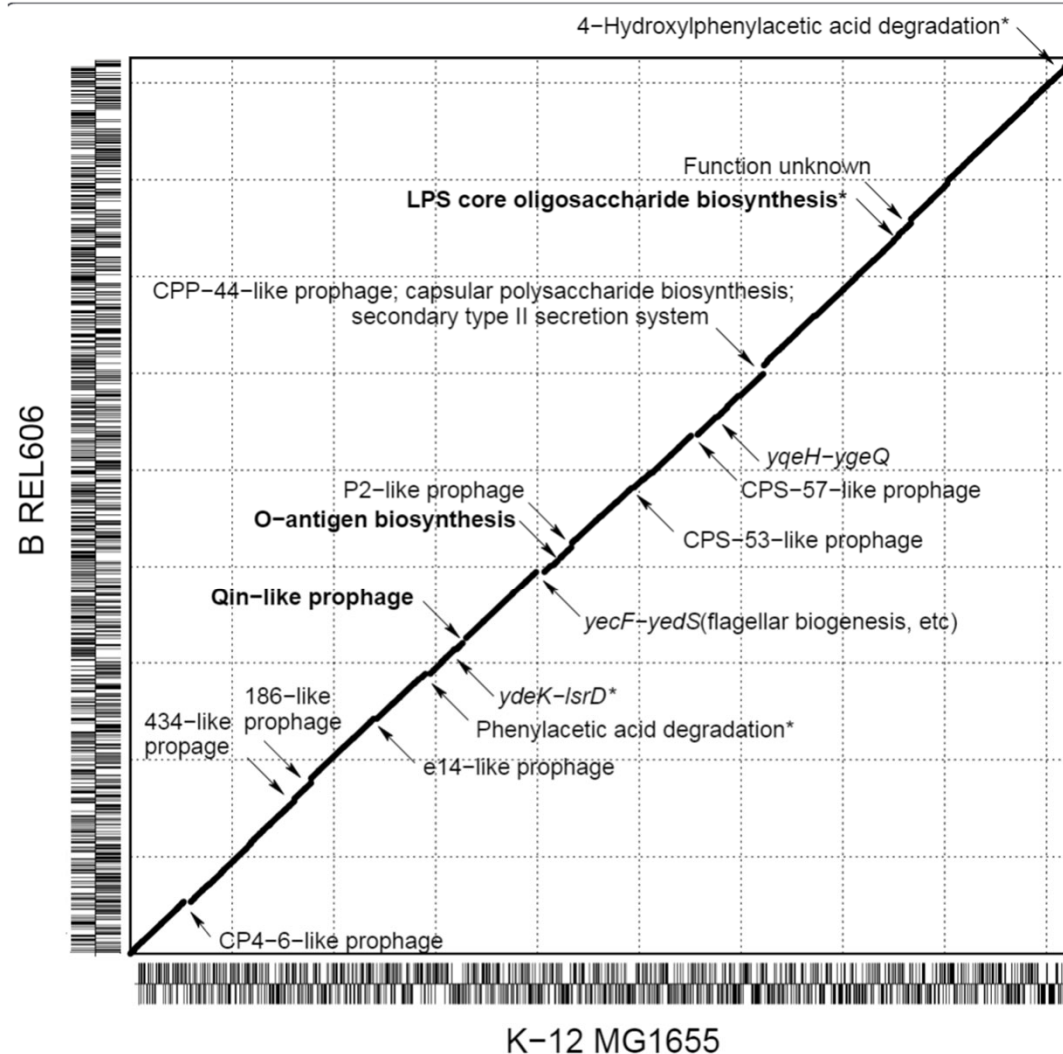
# Poisson and Exponential Distributions

# F. William Studier

- Worked at Brookhaven National Laboratory, Long Island, NY since 1964
- Inventor of **slab gel electrophoresis in 1970** (not patented- back then no incentive to patent work if you are supported by the US government)
- Inventor of **T7 phage expression system for fast production of proteins**. Licensed by over 900 companies, generated over \$55 million for the lab  
[https://en.wikipedia.org/wiki/T7\\_expression\\_system](https://en.wikipedia.org/wiki/T7_expression_system)
- In 2024 he was awarded \$400,000 Merkin Prize in Biomedical Technology (Broad Institute of MIT and Harvard) “for development of widely used protein- and RNA-production platform”  
<https://merkinprize.org/2024-pr>



# K-12 and B are the most common lab strains of *E. coli*



- B strain can be traced back to d'Hérelle(Lambda-phage), Pasteur Institute, Delbruck, Rockefeller U., Studier and Rich Lenski LTEE
- K-12 was isolated in 1922 at Palo Alto hospital. First sequenced genome
- Both have ~4000 genes in roughly the same order

Yoon et al. *Genome Biology* 2012, **13**:R37  
<http://genomebiology.com/2012/13/5/R37>



RESEARCH

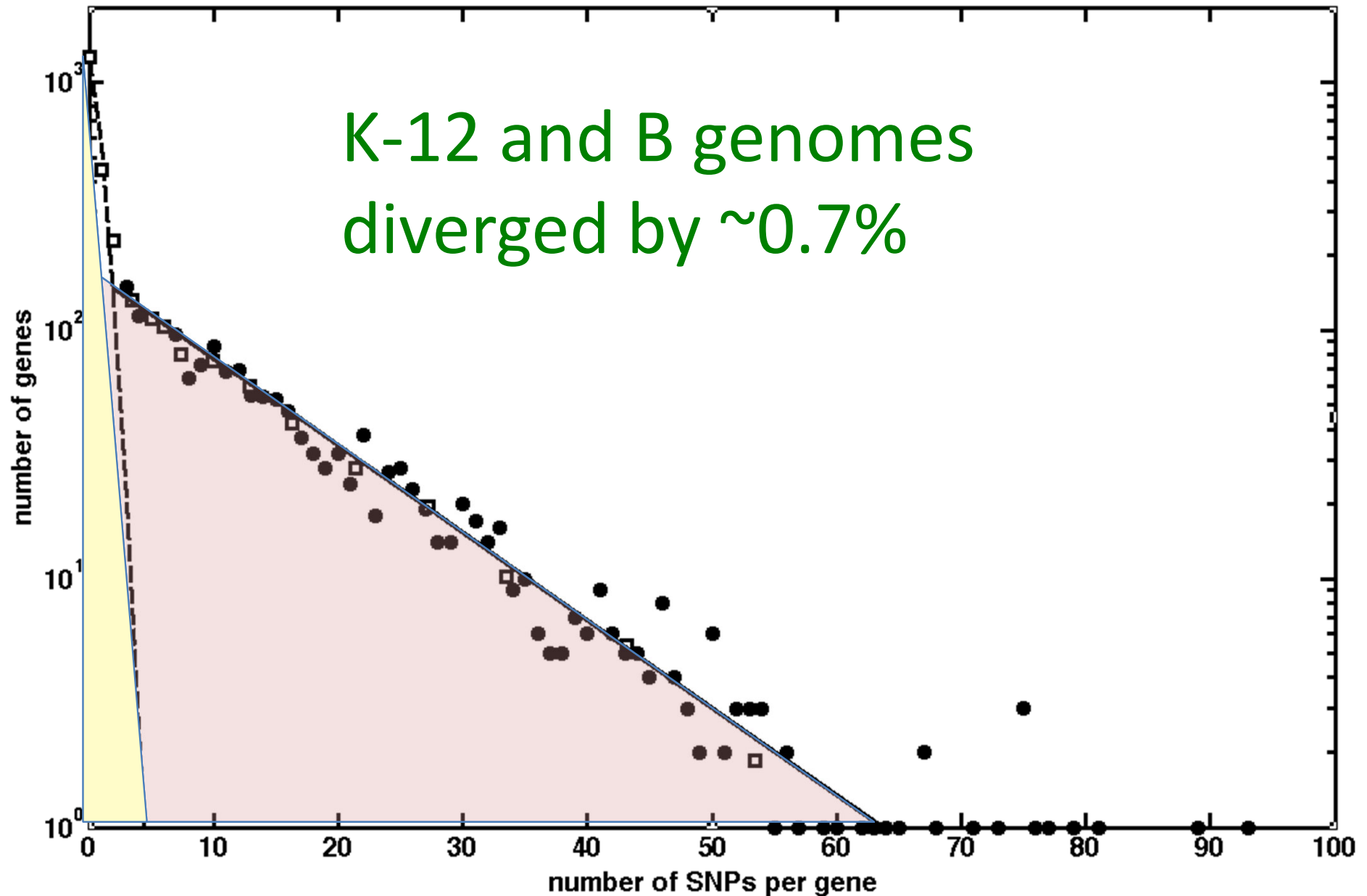
Open Access

## Comparative multi-omics systems analysis of *Escherichia coli* strains B and K-12

Sung Ho Yoon<sup>1</sup>, Mee-Jung Han<sup>2,3</sup>, Haeyoung Jeong<sup>1</sup>, Choong Hoon Lee<sup>1,4,5</sup>, Xiao-Xia Xia<sup>2</sup>, Dae-Hee Lee<sup>1</sup>, Ji Hoon Shim<sup>1</sup>, Sang Yup Lee<sup>2,6</sup>, Tae Kwang Oh<sup>7</sup> and Jihyun F Kim<sup>1,5\*</sup>

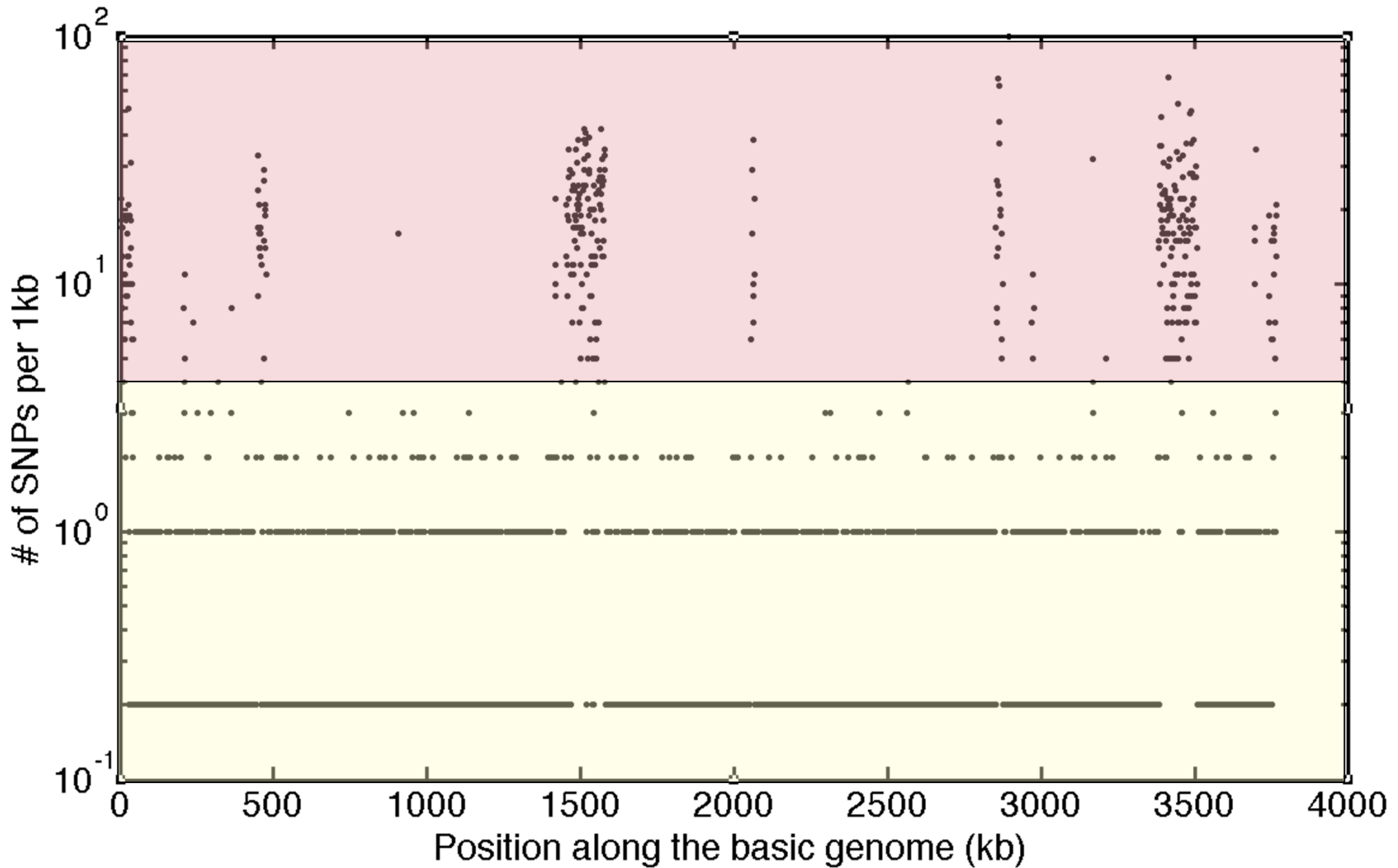
**Figure 1** Whole genome comparison of *E. coli* B REL606 and *E. coli* K-12 MG1655. Strain-specific regions are indicated by discontinuities or the diagonal line (those >10 kb are marked by arrows). Short vertical lines on each axis represent coding sequences that reside on the forward or reverse strand. Segments that occupy the same location on each genome and encode equivalent functions but are highly dissimilar are shown in bold. Except for those marked with asterisks, all the strain-specific regions coincide with genomic islands that were identified by genomic anomalies. Ticks are marked every 500 kb.

# K-12 vs BL21(DE3) strains of E. coli



Studier FW, Daegelen P, Lenski RE, Maslov S, Kim JF, J. Mol Biol. (2009)

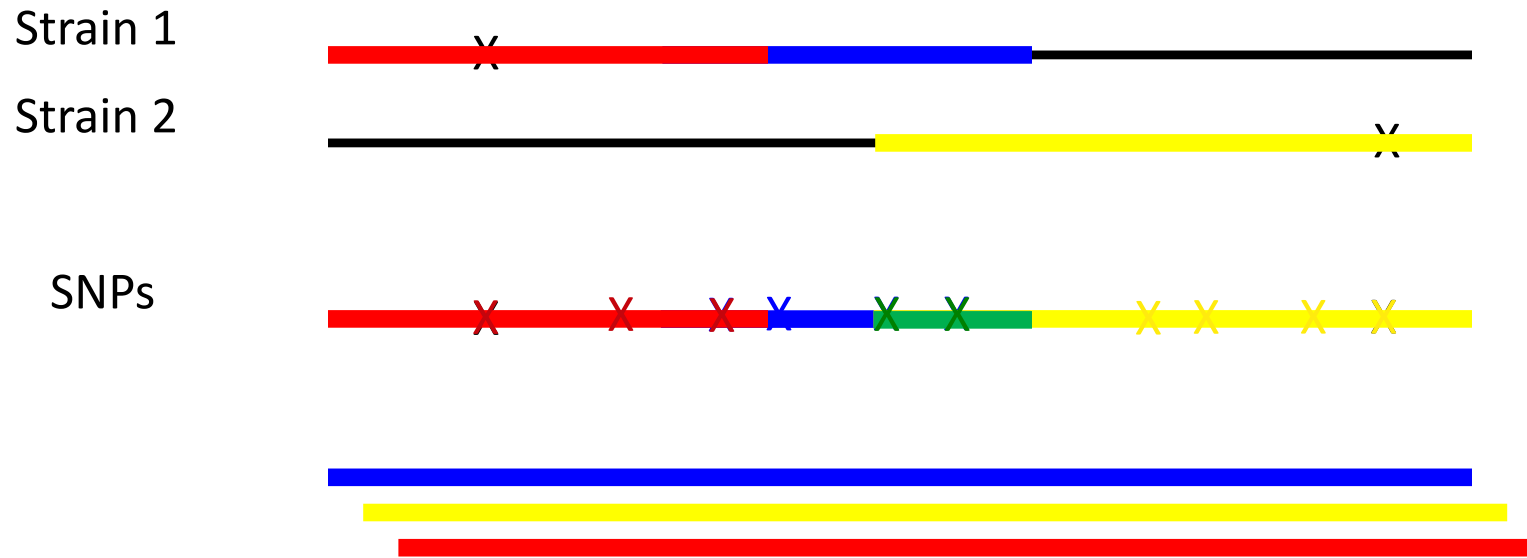
# Highly variable segments are clustered



K-12 vs UMN18 diverged by  $\sim 0.18\%$



# Model of bacterial evolution by mutations and homologous recombination



- Mutation rate  $\mu$  (bp/generation)
- Recombination rate  $\rho$  (bp/generation)
- $l_R$ - average length of recombined segments
- $\theta=2\mu N_e$  depending on  $N_e$  – (effective) population size
- $\delta_{TE}$  transfer efficiency: Prob(successful transfer + recombination):  $\sim \exp(-\delta/\delta_{TE})$

# Why exponential tail?

- Empirical data for E. coli:  $\text{Prob}(\delta) = \exp(-\delta/0.01)$   
Similar slopes in other species as distant as B. subtilis
- Theory 1: PopGen 101 coalescence time distribution:
  - $\text{Prob}(T) \sim \exp(-T/N_e) \rightarrow$   
 $\text{Prob}(\delta) \sim \exp(-\delta/2\mu N_e) = \underline{\exp(-\delta/\theta)}$   
 $\theta = 2\mu N_e \sim 0.01, \mu \sim 10^{-10} \rightarrow N_e \sim 10^8$
- Theory 2: biophysics of homologous recombination:
  - Requires perfect matches of  $L=30\text{bp}$  on each side  $\rightarrow$   
 $\text{Prob}(\delta) = (1 - \delta)^{2L} = \exp(-60 \cdot \delta) = \exp(-\delta/0.016) = \underline{\exp(-\delta/\delta_{TE})}$
- Both mechanisms likely to work together:  
biophysics of recombination affects the effective population size

# Continuous Probability Distributions

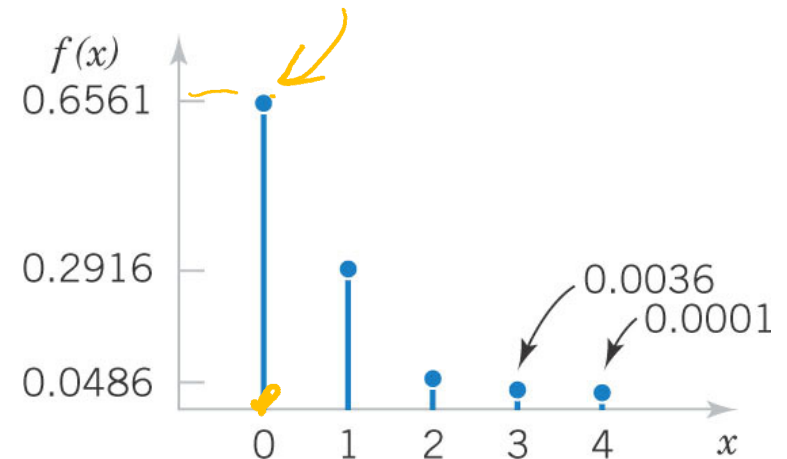
## Uniform Distribution

# Continuous & Discrete Random Variables

- A **discrete random variable** is usually integer number
  - $N$  – the number of proteins in a cell
  - $D$  - number of nucleotides different between two sequences
- A **continuous random variable** is a real number
  - $C=N/V$  – the concentration of proteins in a cell of volume  $V$
  - Percentage  $D/L * 100\%$  of different nucleotides in protein sequences of different lengths  $L$   
(depending on set of  $L$ 's may be discrete but dense)

# Probability Mass Function (PMF)

- $X$  – discrete random variable
- Probability Mass Function:  $f(x) = P(X=x)$   
– the probability that  $X$  is exactly equal to  $x$



Probability Mass Function for the # of mismatches in 4-mers

$P(X=0) =$	0.6561
$P(X=1) =$	0.2916
$P(X=2) =$	0.0486
$P(X=3) =$	0.0036
$P(X=4) =$	0.0001
$\sum_x P(X=x) =$	1.0000

# Probability Density Function (PDF)

Density functions, in contrast to mass functions, distribute probability continuously along an interval

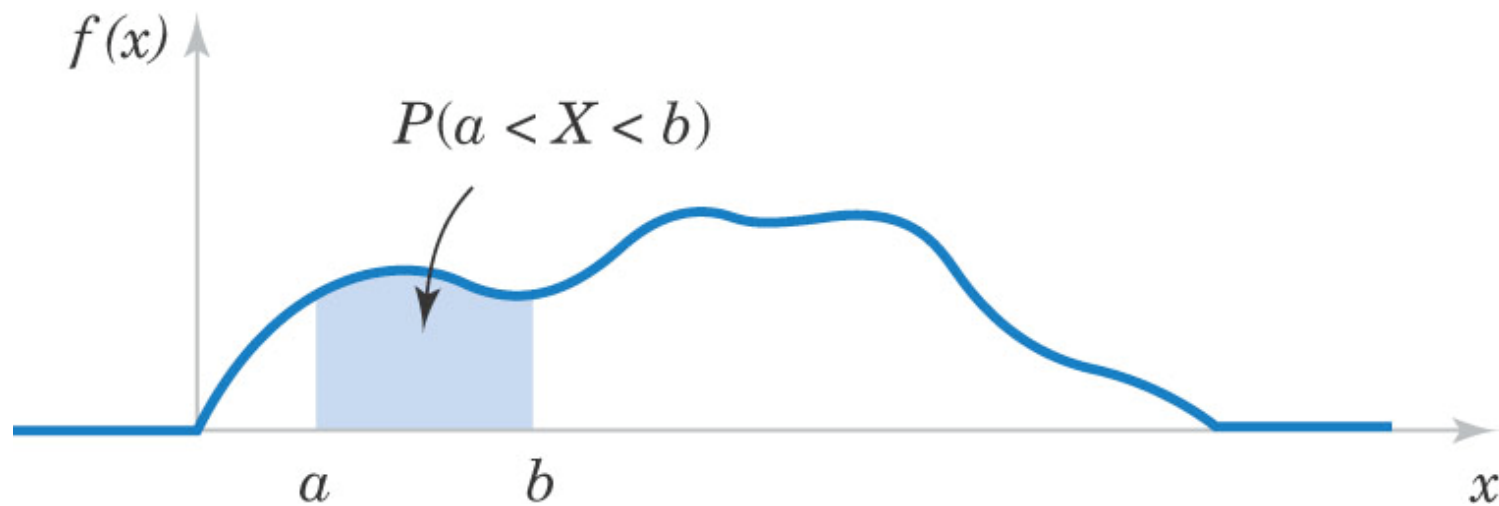


Figure 4-2 Probability is determined from the area under  $f(x)$  from  $a$  to  $b$ .



# Probability Density Function

For a continuous random variable  $X$ ,  
a **probability density function** is a function such that

(1)  $f(x) \geq 0$  means that the function is always non-negative.

$$(2) \int_{-\infty}^{\infty} f(x) dx = 1$$

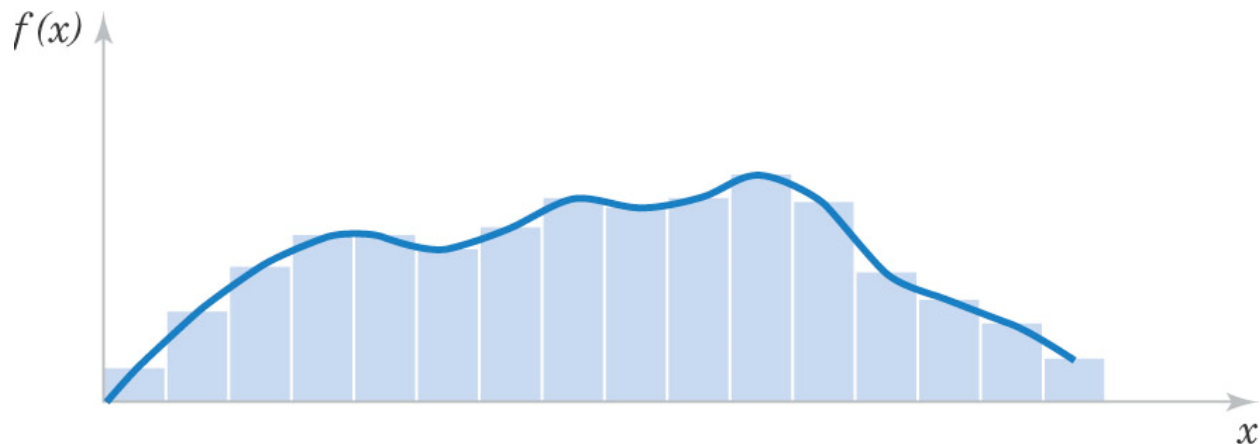
$$(3) P(a \leq X \leq b) = \int_a^b f(x) dx = \text{area under } f(x) dx \text{ from } a \text{ to } b$$

# Normalized histogram approximates PDF

A **histogram** is graphical display of data showing a series of adjacent rectangles. Each rectangle has a **base** which represents an **interval of data values**. The height of the rectangle is a **number of events** in the sample **within the base**.

When base length is narrow, the histogram could be normalized to approximate PDF ( $f(x)$ ):

**height of each rectangle =  
=(# of events within base)/(total # of events)/width of its base.**



Normalized histogram approximates a probability density function.

# Cumulative Distribution Functions (CDF & CCDF)

The **cumulative distribution function (CDF)** of a continuous random variable  $X$  is,

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du \text{ for } -\infty < x < \infty \quad (4-3)$$

One can also use the **inverse cumulative distribution function** or **complementary cumulative distribution function (CCDF)**

$$F_{>}(x) = P(X > x) = \int_x^{\infty} f(u)du \text{ for } -\infty < x < \infty$$

**Definition of CDF for a continuous variable is the same as for a discrete variable**

# Density vs. Cumulative Functions

- The probability density function (PDF) is the derivative of the cumulative distribution function (CDF).

$$f(x) = \frac{dF(x)}{dx} = -\frac{dF_{>}(x)}{dx}$$

as long as the derivative exists.

# Mean & Variance

Suppose  $X$  is a continuous random variable with probability density function  $f(x)$ . The **mean** or **expected value** of  $X$ , denoted as  $\mu$  or  $E(X)$ , is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx \quad (4-4)$$

The **variance** of  $X$ , denoted as  $V(X)$  or  $\sigma^2$ , is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

The **standard deviation** of  $X$  is  $\sigma = \sqrt{\sigma^2}$ .

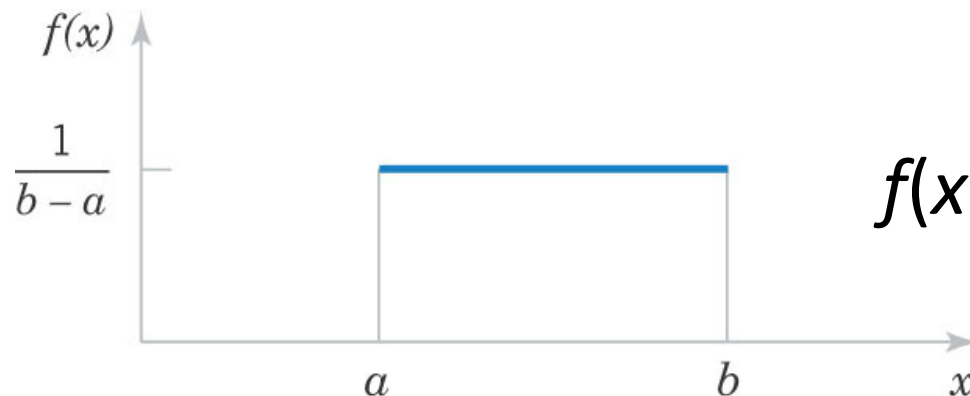
# Gallery of Useful Continuous Probability Distributions



# Continuous Uniform Distribution

- This is the simplest continuous distribution and analogous to its discrete counterpart.
- A continuous random variable  $X$  with probability density function

$$f(x) = 1 / (b-a) \text{ for } a \leq x \leq b \quad (4-6)$$



*Compare to  
discrete*

$$f(x) = 1/(b-a+1)$$

Figure 4-8 Continuous uniform PDF

# Comparison between Discrete & Continuous Uniform Distributions

## Discrete:

- PMF:  $f(x) = 1/(b-a+1)$
- Mean and Variance:  
 $\mu = E(x) = (b+a)/2$   
 $\sigma^2 = V(x) = [(b-a+1)^2-1]/12$

## Continuous:

- PMF:  $f(x) = 1/(b-a)$
- Mean and Variance:  
 $\mu = E(x) = (b+a)/2$   
 $\sigma^2 = V(x) = (b-a)^2/12$

X is a **continuous** random variable  
with a uniform distribution  
between 0 and 3.

What is Probability( $X=1$ )?

- A.  $1/4$
- B.  $1/3$
- C. 0
- D. Infinity
- E. I have no idea

Get your i-clickers

X is a **continuous** random variable  
with a uniform distribution  
between 0 and 3.

What is  $P(X=1)$ ?

A.  $1/4$

B.  $1/3$

C. 0

D. Infinity

E. I have no idea

Get your i-clickers

X is a **continuous** random variable  
with a uniform distribution  
between 0 and 3.

What is  $P(X < 1)$ ?

- A.  $1/4$
- B.  $1/3$
- C. 0
- D. Infinity
- E. I have no idea

Get your i-clickers

X is a **continuous** random variable  
with a uniform distribution  
between 0 and 3.

What is  $P(X < 1)$ ?

A.  $1/4$

B.  $1/3$

C. 0

D. Infinity

E. I have no idea

Get your i-clickers

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 DO WHALES JUMP  
WHY ARE WITCHES GREEN  
WHY ARE THERE MIRRORS ABOVE BEDS  
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 DO IGUANAS DIE  
WHY AREN'T THERE DINOSAUR GHOSTS

WHY AREN'T ECONOMISTS RICH  
WHY DO AMERICANS CALL IT SOCCER  
WHY ARE MY EARS RINGING  
WHY ARE THERE SO MANY AVENGERS  
WHY ARE THE AVENGERS FIGHTING THE X MEN  
WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE SWARMS OF GNATS  
WHY IS THERE PHLEGM  
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN  
WHY IS PSYCHIC WEAK TO BUG  
WHY DO CHILDREN GET CANCER  
WHY IS POSEIDON ANGRY WITH ODYSSEUS  
WHY IS THERE ICE IN SPACE

# WHY ARE THERE ANTS IN MY LAPTOP

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



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

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 IS SEX SO IMPORTANT



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 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 ARE ULTRASOUNDS IMPORTANT  
WHY ARE ULTRASOUND MACHINES EXPENSIVE  
WHY IS STEALING WRONG