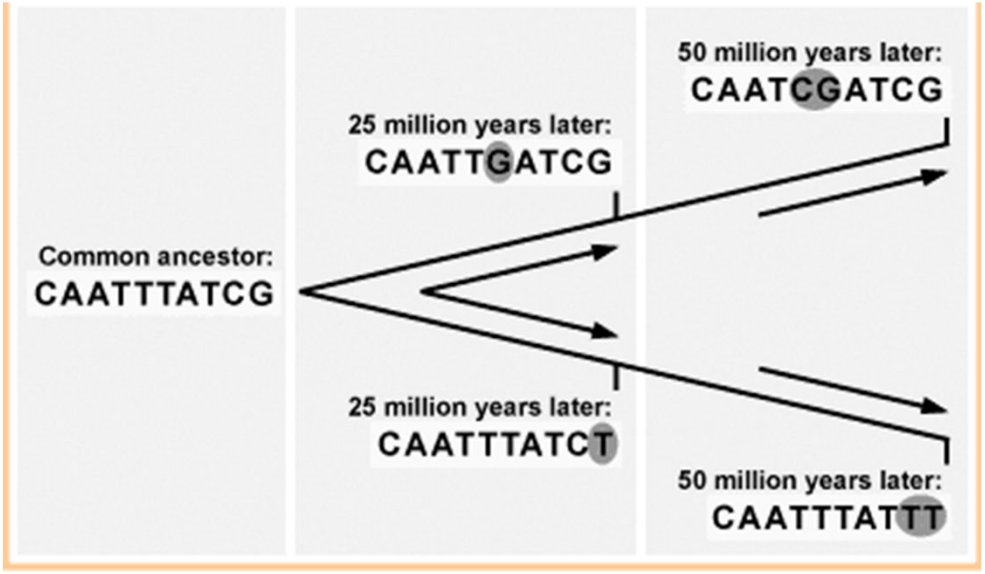
Poisson and Geometric Distributions

Molecular clock hypothesis

- Conjectured by Émile Zuckerkandl and Linus Pauling in 1962

- Justified by the neutral theory of molecular evolution developed by Motoo Kimura in 1968



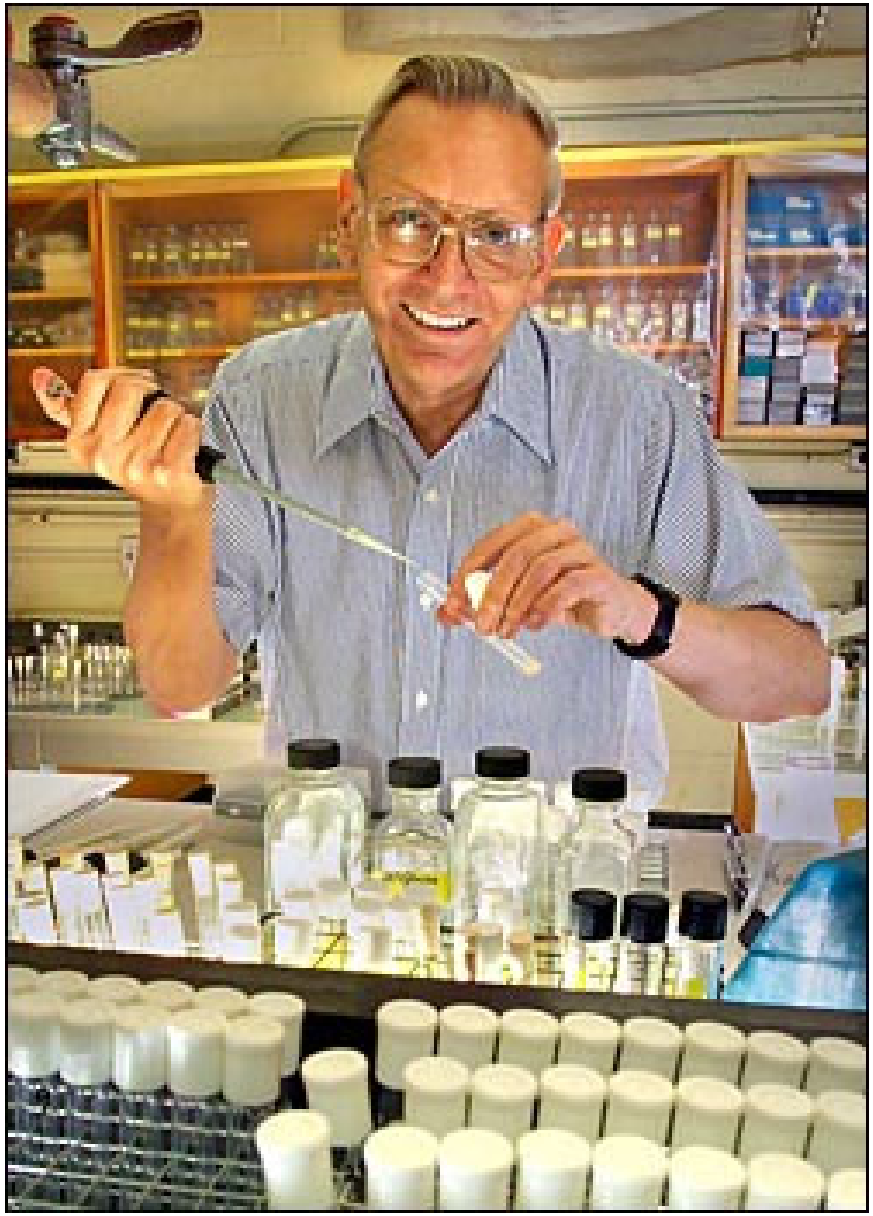
ρ - density of differences between two sequences

$$\rho = 2\,\mu\,T$$

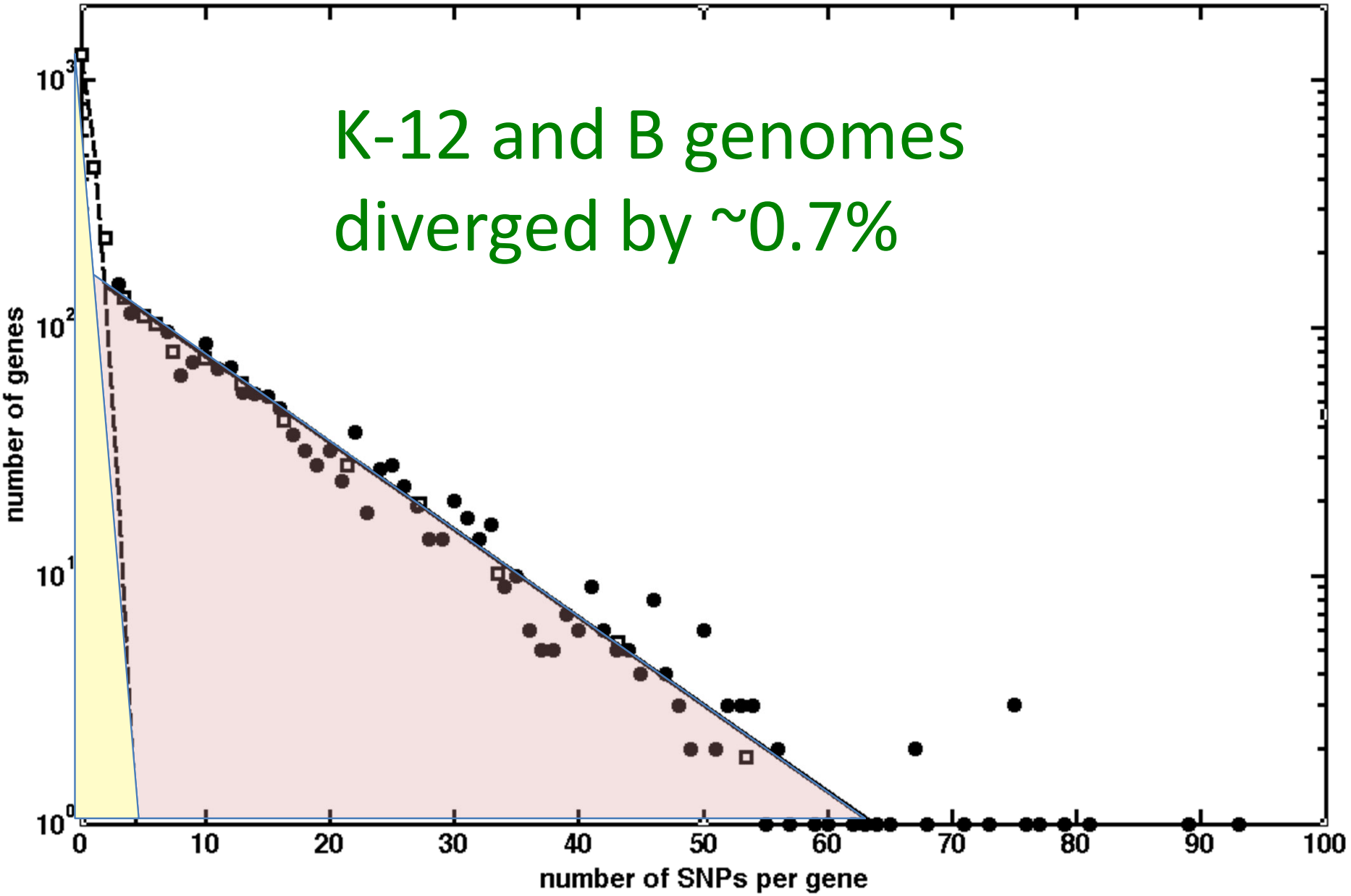
Where μ is the mutation rate per base pair per generation and T is the number of generations since last common ancestor

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 givernment)
- 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

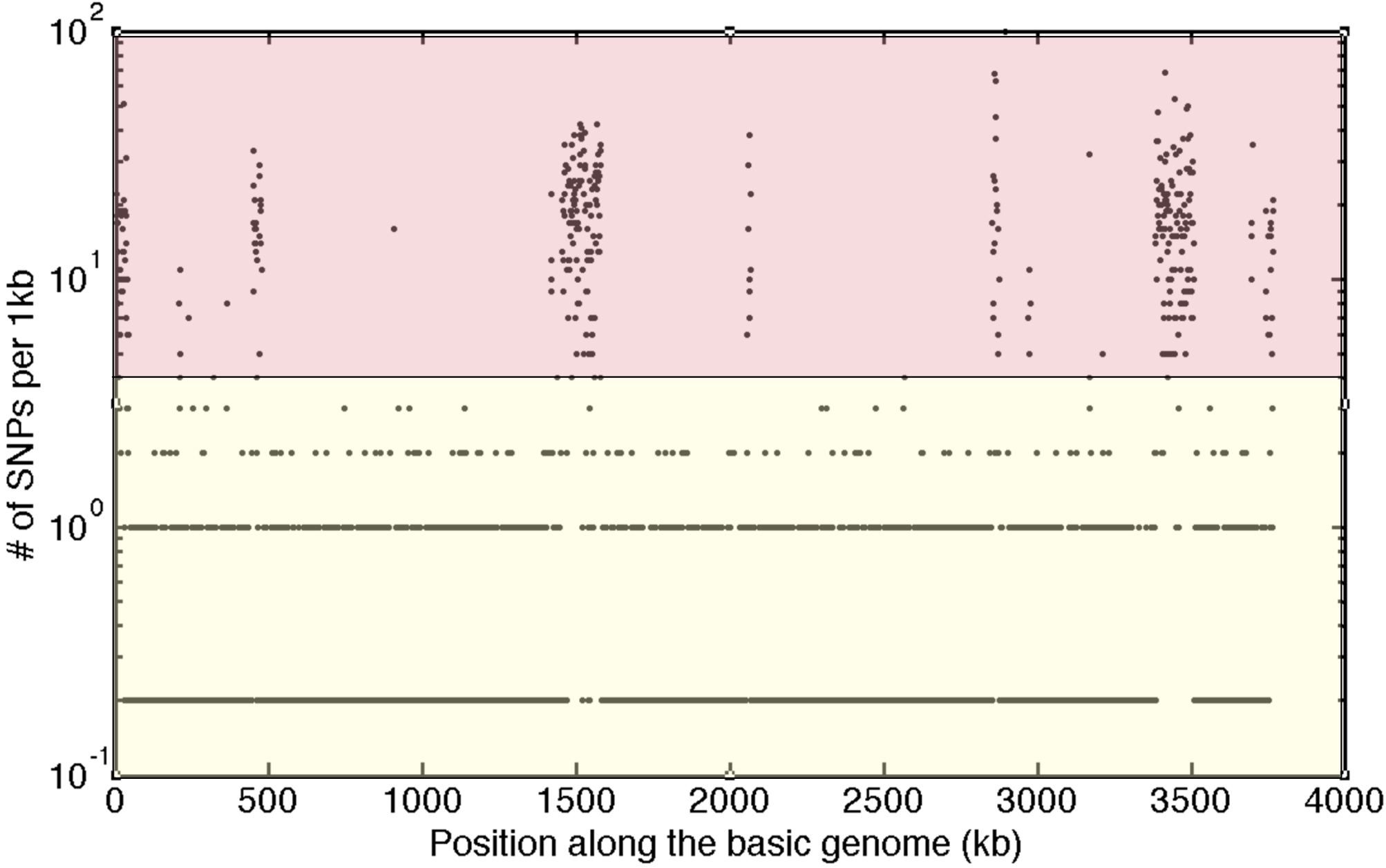


K-12 vs BL21(DE3) strains of E. coli



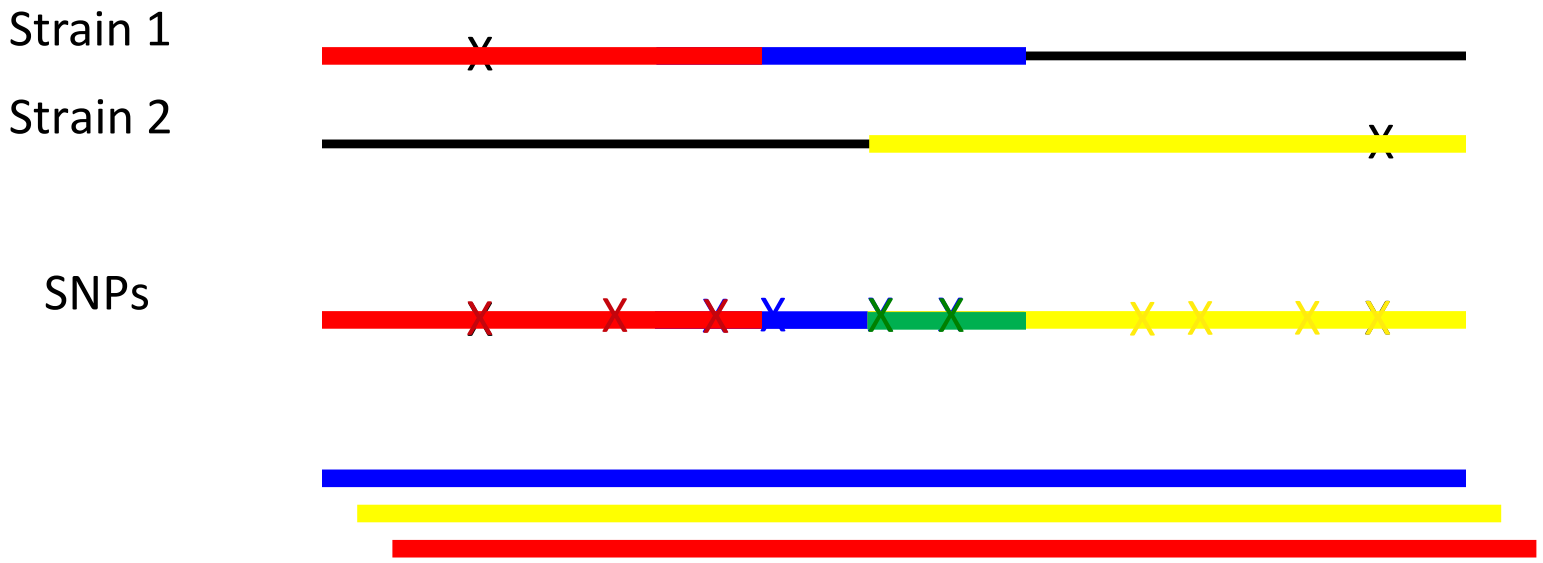
Studier FW, Daegelen P, Lenski RE, **Maslov S**, Kim JF, J. Mol Biol. (2009)

Highly variable segments are clustered



K-12 vs UMNF18 diverged by ~0.18%

Model of bacterial evolution by mutations and homologous recombination



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

Dixit P, Pang TY, Studier FW, **Maslov S,** PNAS (2015); arXiv:1405.2548

Why exponential tail?

- Empirical data for E. coli: Prob(δ)=exp(-δ/0.01)
  Similar slopes in other species as distant as B. subtilis
- Theory 1: PopGen 101 coalescence time distribution:
  - Prob(T) ~ exp(-T/N_e) →
    Prob(δ) ~ exp(- δ/ 2μN_e) = exp(- δ/θ)
    θ =2μN_e~0.01, μ~10^{-10} → N_e~10^8
- Theory 2: biophysics of homologous recombination:
  - Requires perfect matches of L=30bp on each side →
    Prob(δ)=(1- δ)^{2L}=exp(-60•δ)=exp(-δ/0.016)=exp (-δ/δ_{TE})
- Both mechanisms likely to work together:
  biophysics of recombination affects the effective population size

Negative Binomial Definition

- In a series of independent trials with constant probability of success, p, let the random variable X denote the number of trials until r successes occur. Then X is a negative binomial random variable with parameters:
  $0 < p < 1$ and r = 1, 2, 3, ...
- The probability mass function is:

$$f(x) = C_{r-1}^{x-1} p^r (1-p)^{x-r} \quad \text{for } x = r, r+1, r+2... \qquad (3\text{-}11)$$

- Compare it to binomial

$$f(x) = C_x^n p^x (1-p)^{n-x} \quad \text{for } x = 1, 2, ... \text{ n}$$

NOTE OF CAUTION: Matlab, Mathematica, and many other sources use x to denote the number of failures until one gets r successes.
We stick with Montgomery-Runger.

Negative Binomial Mean & Variance

- If *X* is a negative binomial random variable with parameters *p* and *r*,

$$\mu = E(X) = \frac{r}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{r(1-p)}{p^2} \qquad (3\text{-}12)$$

- Compare to geometric distribution:

$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \qquad (3\text{-}10)$$

Matlab exercise

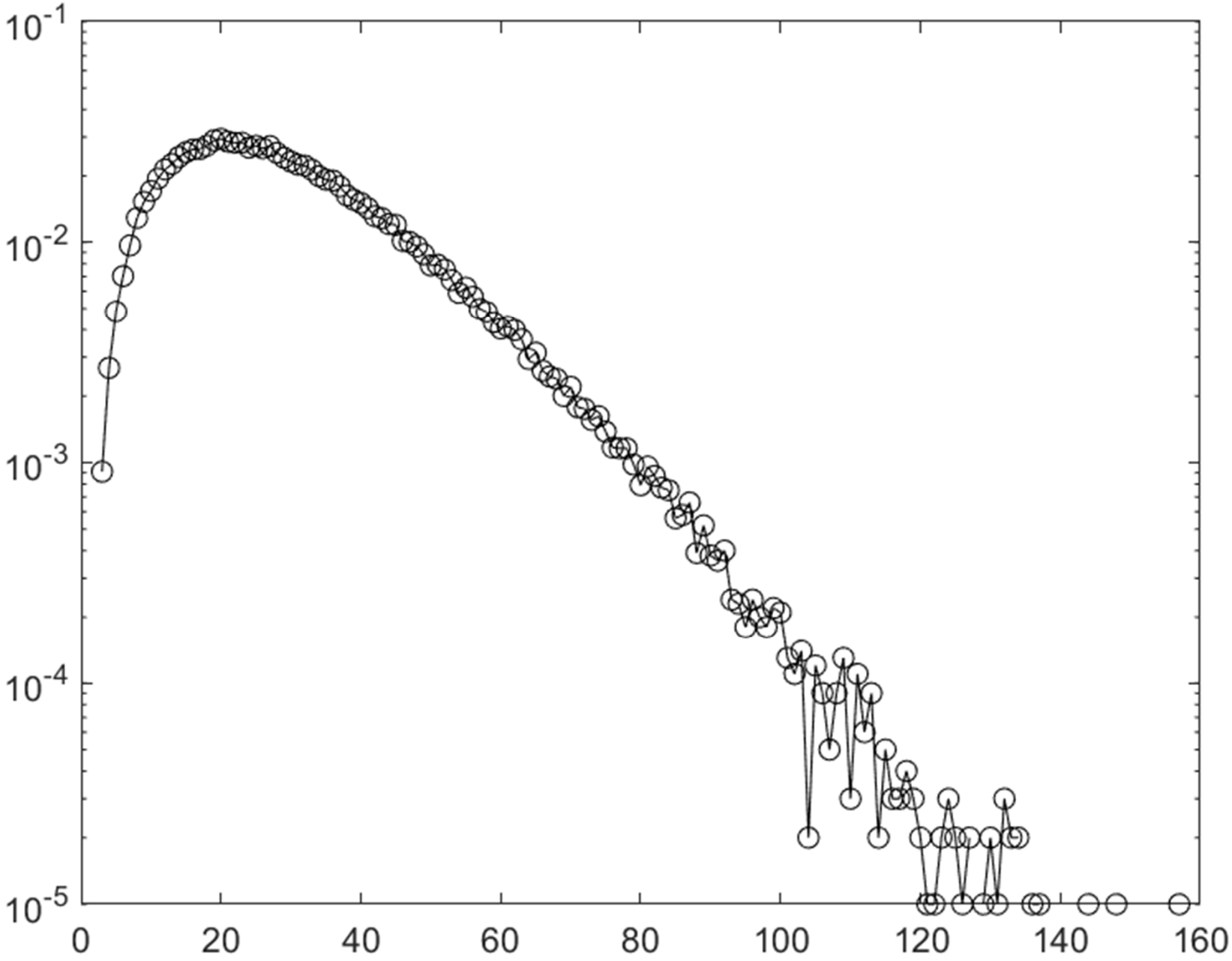
- Estimate mean, variance, and PMF based on 100,000 random variables drawn from a negative binomial distribution with p=0.1, r=3

- Repeat with negative binomial distribution with p=0.1, r=100

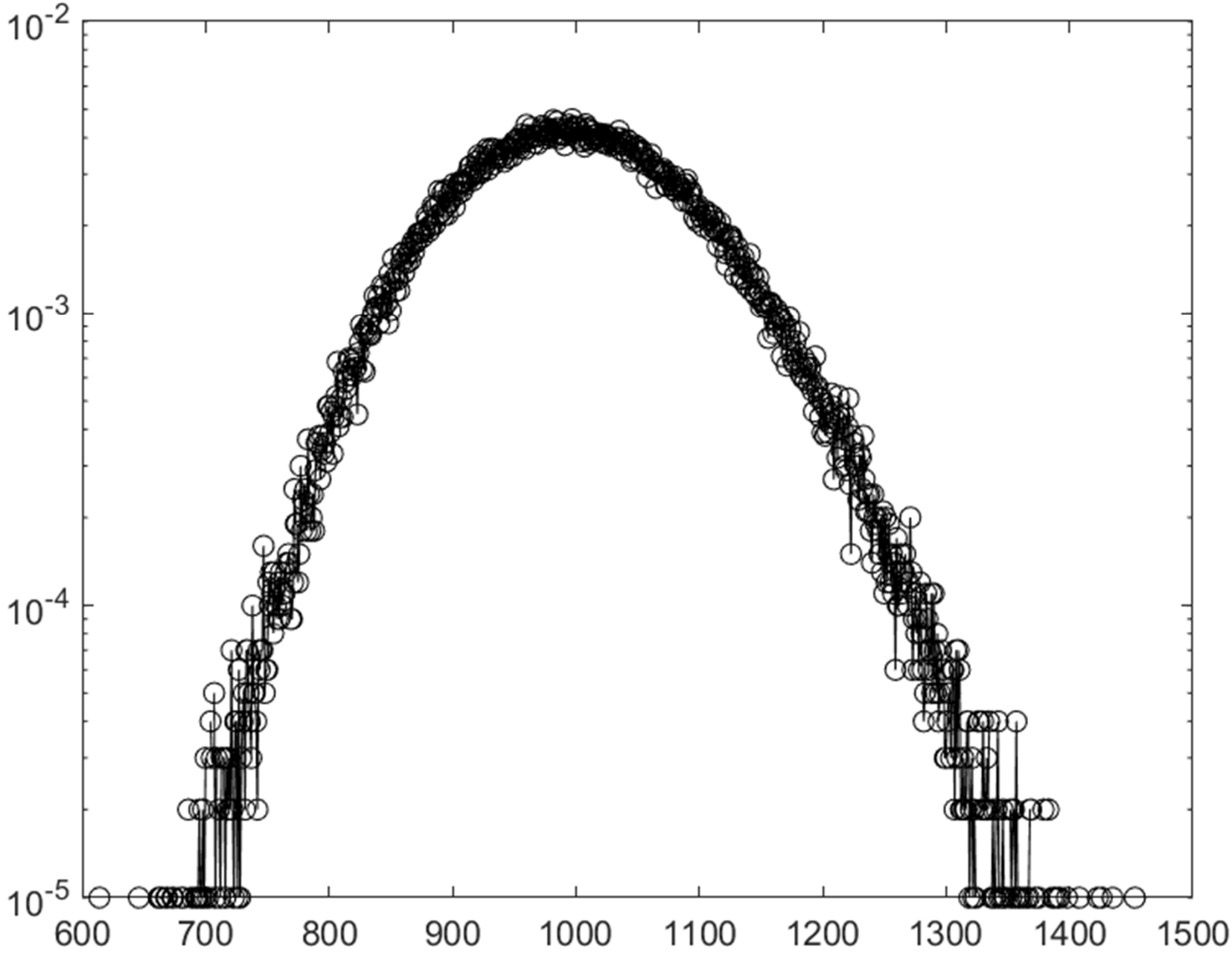
Matlab: Negative binomial distribution

```
Stats=100000;
r=3; p=0.1;
r2=zeros(Stats,1);
for k=1:Stats
    n_trials=0;
    n_successes=0;
    while n_successes<r
        if rand<p
            n_successes=n_successes+1;
        end;
        n_trials=n_trials+1;
    end;
    r2(k)=n_trials;
end;
disp('Observed average value’); disp(sum(r2)./Stats);
disp('Expected average value’); disp(r./p);
disp('Observed variance’); disp(sum(r2.^2)./Stats-(sum(r2)./Stats).^2);
disp('Expected variance’); disp(r.*(1-p)./p^2);
[a,b]=hist(r2, 1:max(r2));
p_nb=a./sum(a);
figure; semilogy(b,p_nb,'ko-');
```

Negative binomial PMF, p=0,1 r=3



Negative binomial PMF, p=0,1 r=100



Cancer is scary!

- Approximately 40% of men and women will be diagnosed with cancer at some point during their lifetimes (source: NCI website)

TABLE 21.2 Leading causes of death in United States in 2010. Cause of death is based on the International Classification of Diseases, Tenth Revision, 1992.

Rank	Cause of death	Number	Percent of all deaths
–	All causes	2,468,435	100.0
1	Diseases of heart	597,689	24.2
2	Malignant neoplasms	574,743	23.3
3	Chronic lower respiratory diseases	138,080	5.6
4	Cerebrovascular diseases	129,476	5.2
5	Accidents (unintentional injuries)	120,859	4.9
6	Alzheimer's disease	83,494	3.4
7	Diabetes mellitus	69,071	2.8
8	Nephritis, nephrotic syndrome, and nephrosis	50,476	2.0
9	Influenza and pneumonia	50,097	2.0
10	Intentional self-harm (suicide)	38,364	1.6

Source: National Vital Statistics Reports, 62(6) (http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_06.pdf)

Table from J. Pevsner 3rd edition

- "War on Cancer" – president Nixon 1971.
  "Moonshot to Cure Cancer" – vice-president Joe Biden 2016

“War on Cancer” progress report

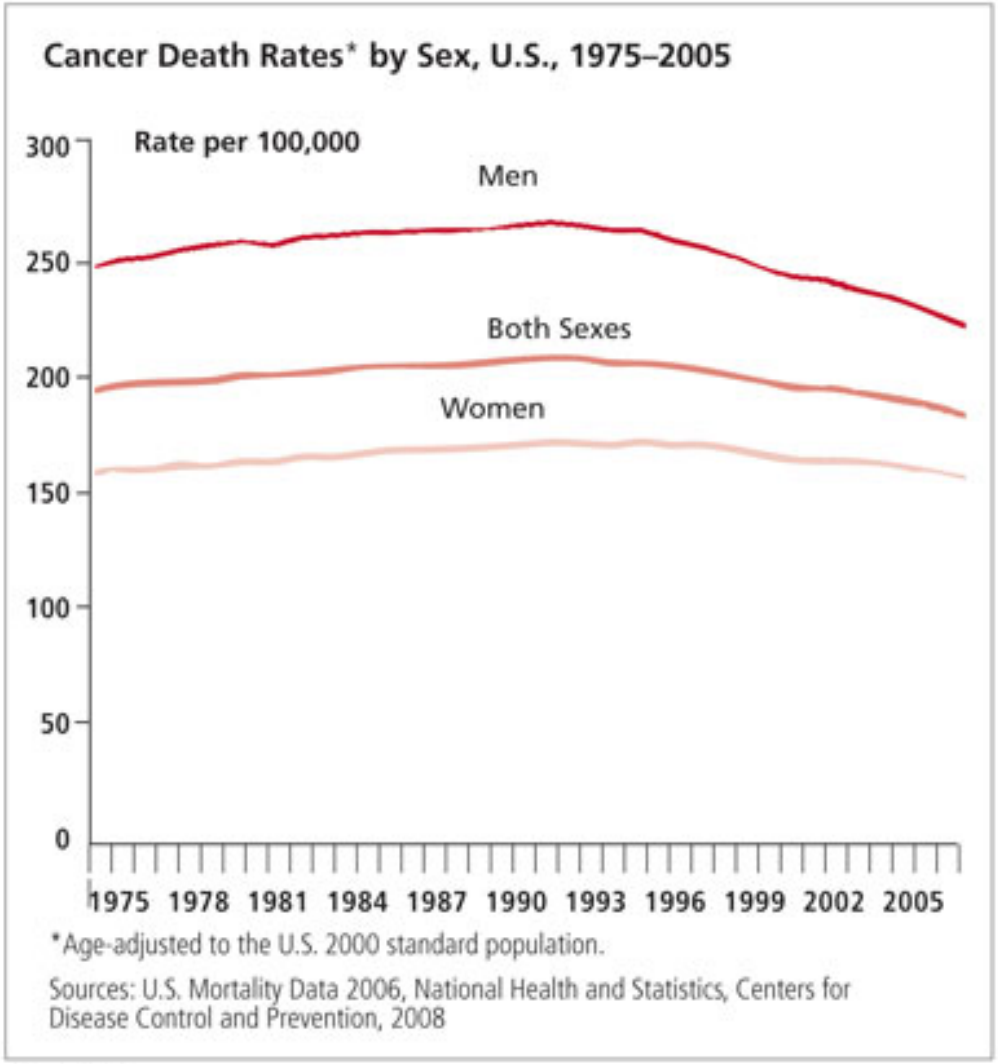


Figure 2

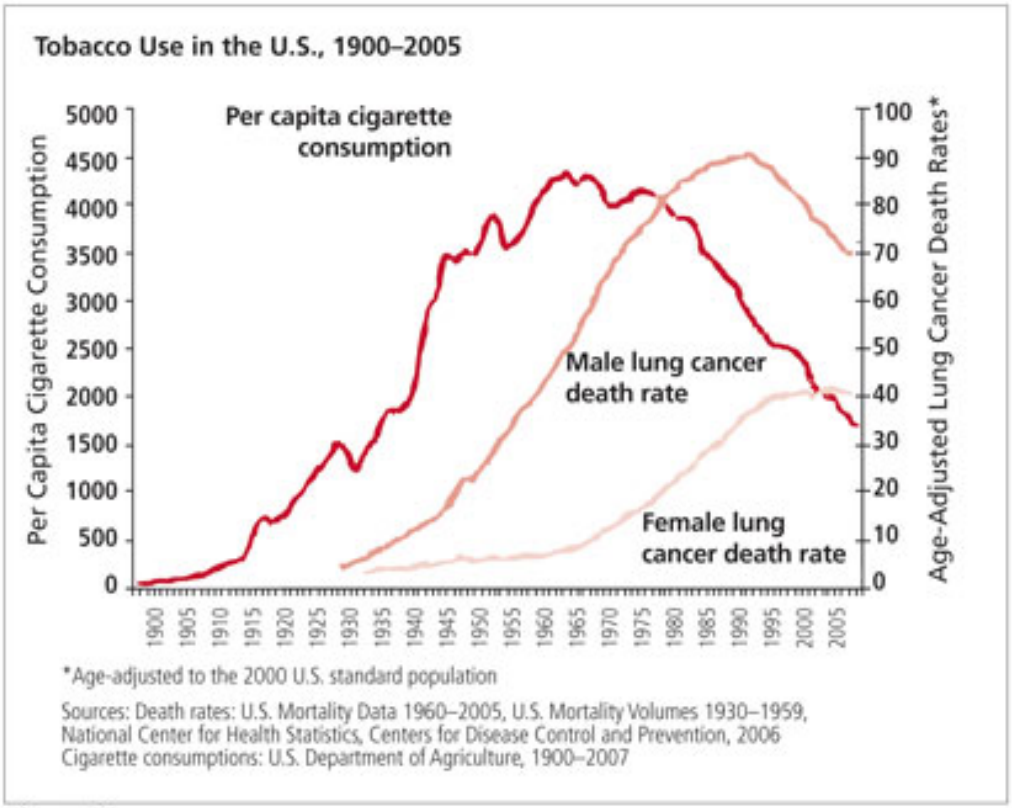


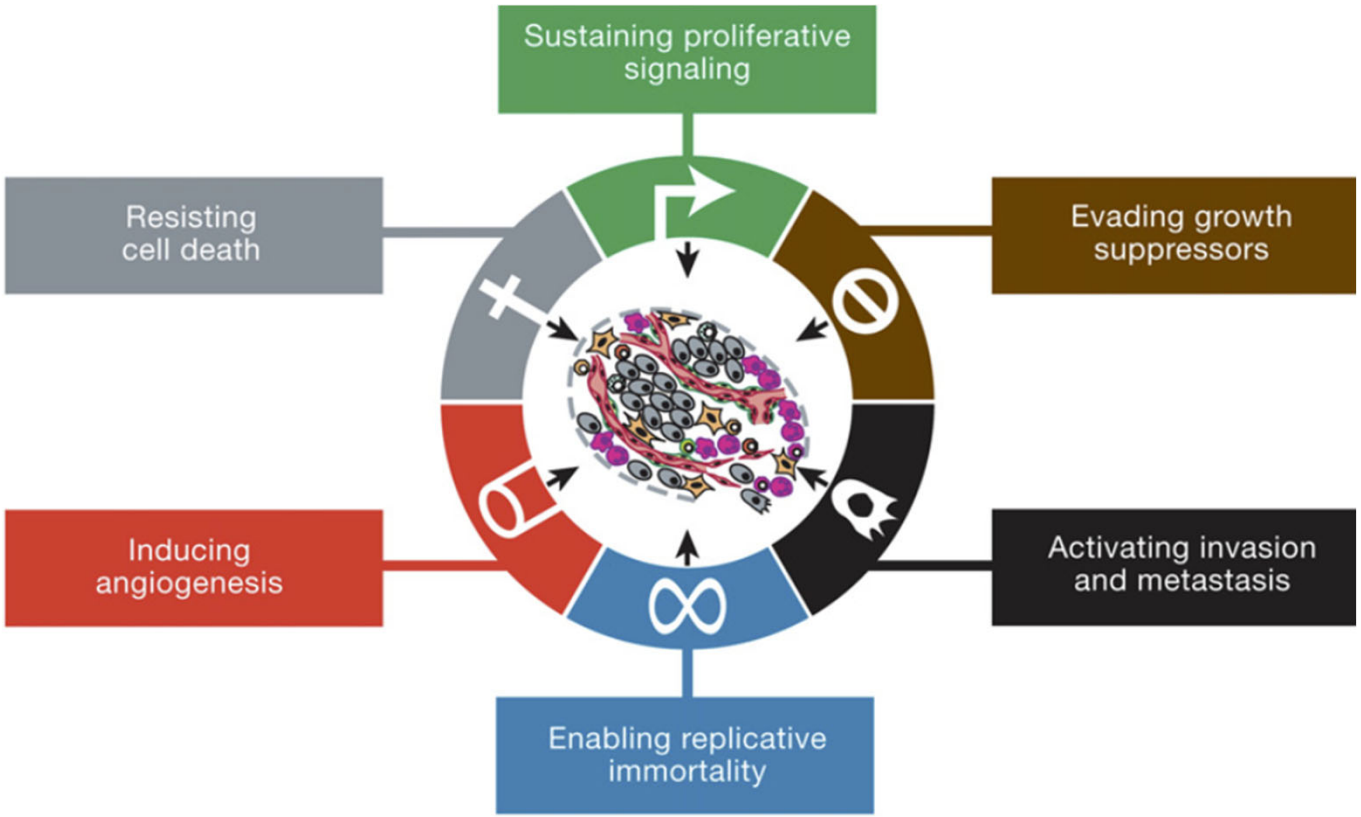
Figure 3

Probability theory and statistics
is a powerful tool to
learn new cancer biology

“Driver genes” theory

- Progression of cancer is caused by accumulation of mutations in a handful of “driver” genes
- Mutations in driver genes boost the growth of a tumor
- Oncogenes: expression needs to be elevated for cancer
- Tumor suppressors (e.g. p53) need to be turned off in cancer

Douglas Hanahan and Robert A. Weinberg **Hallmarks of Cancer**: The Next Generation Cell 144, 2011



Statistics of cancer incidence vs age

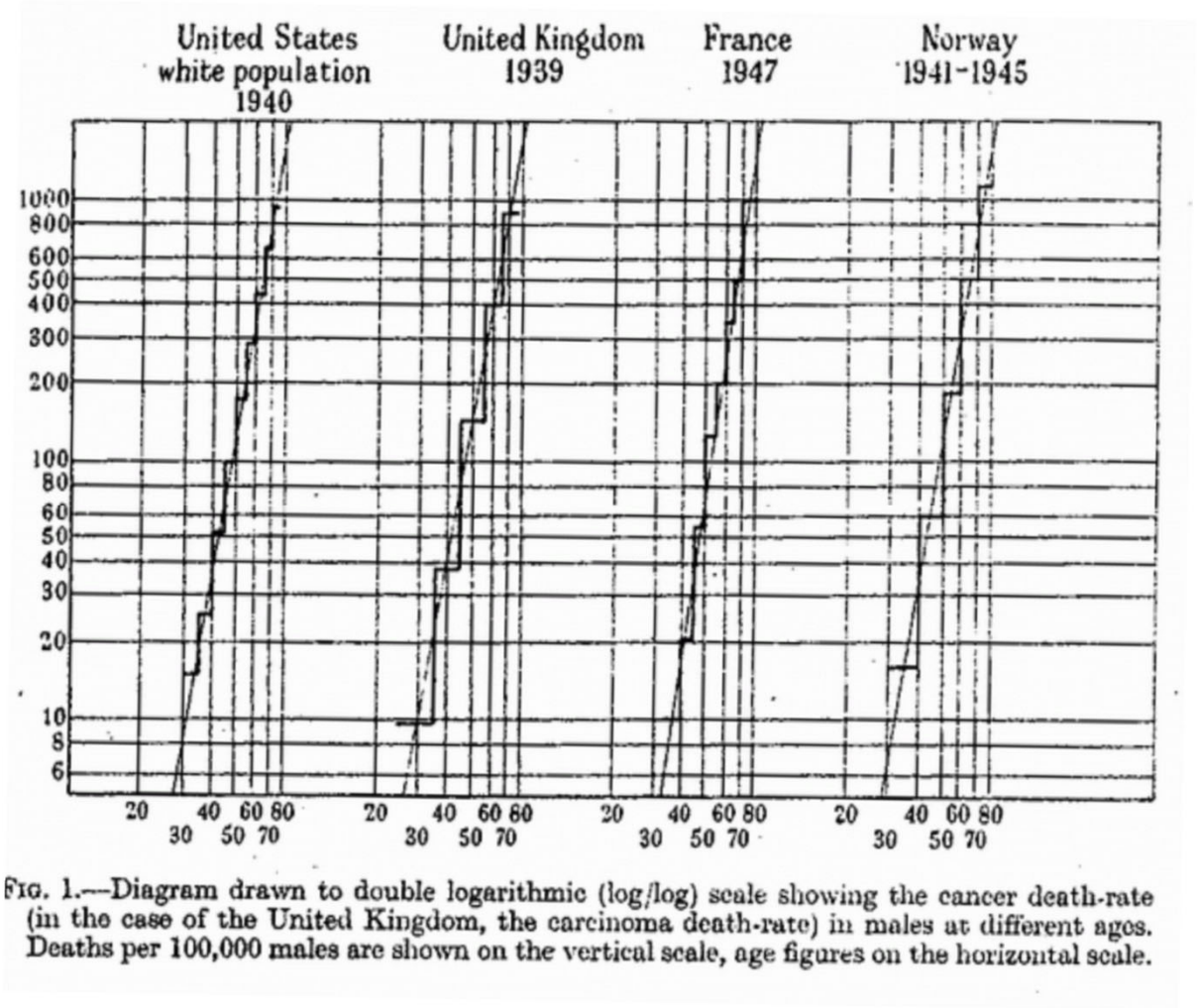


FIG. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

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

Cancer death rate
~ (patient age)^6

It suggests the existence of
k=7 driver genes

$$P(T_{cancer} \le t) \sim (u_1 t)(u_2 t)..(u_k t) \sim u_1 u_2 .. u_k \, t^k$$

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

How many driver gene mutations for different types of cancer?

PNAS

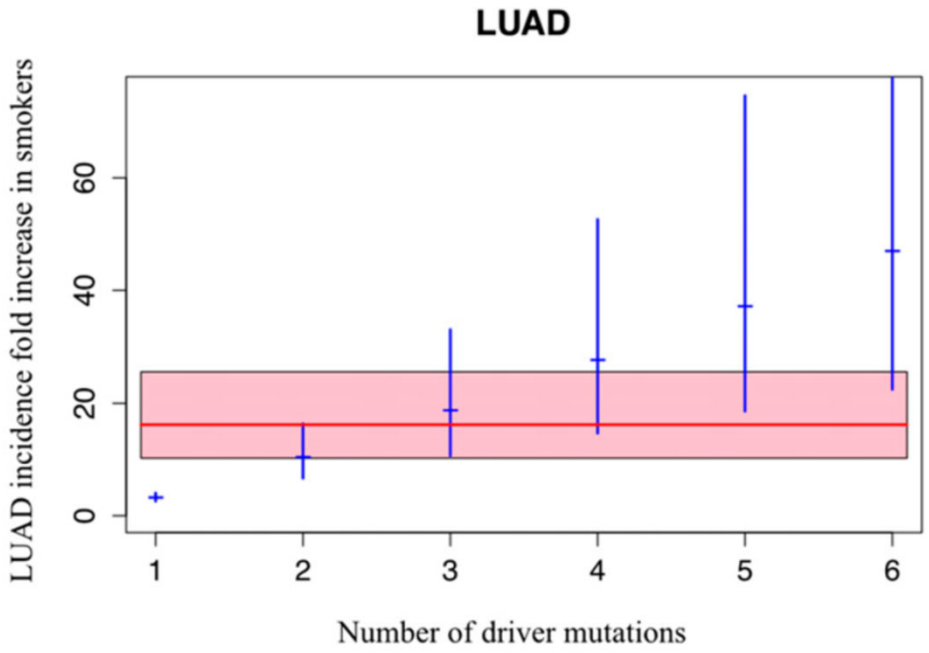
Only three driver gene mutations are required for the development of lung and colorectal cancers

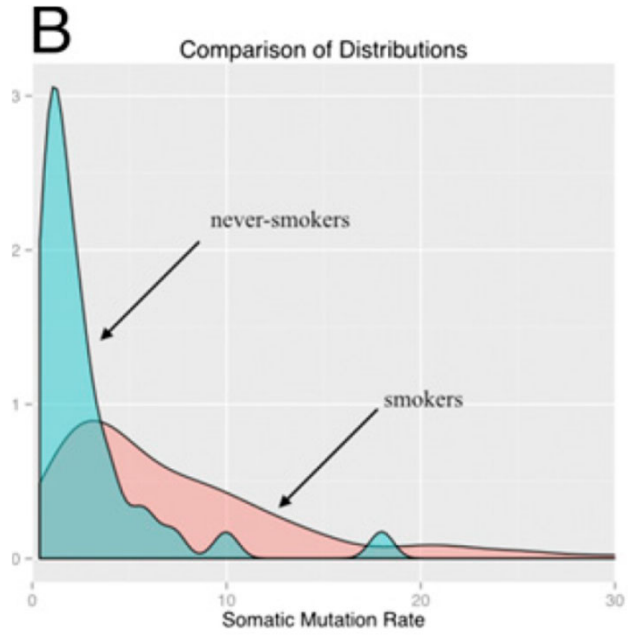
Cristian Tomasetti[a,b,1], Luigi Marchionni[c], Martin A. Nowak[d], Giovanni Parmigiani[e], and Bert Vogelstein[f,g,1]

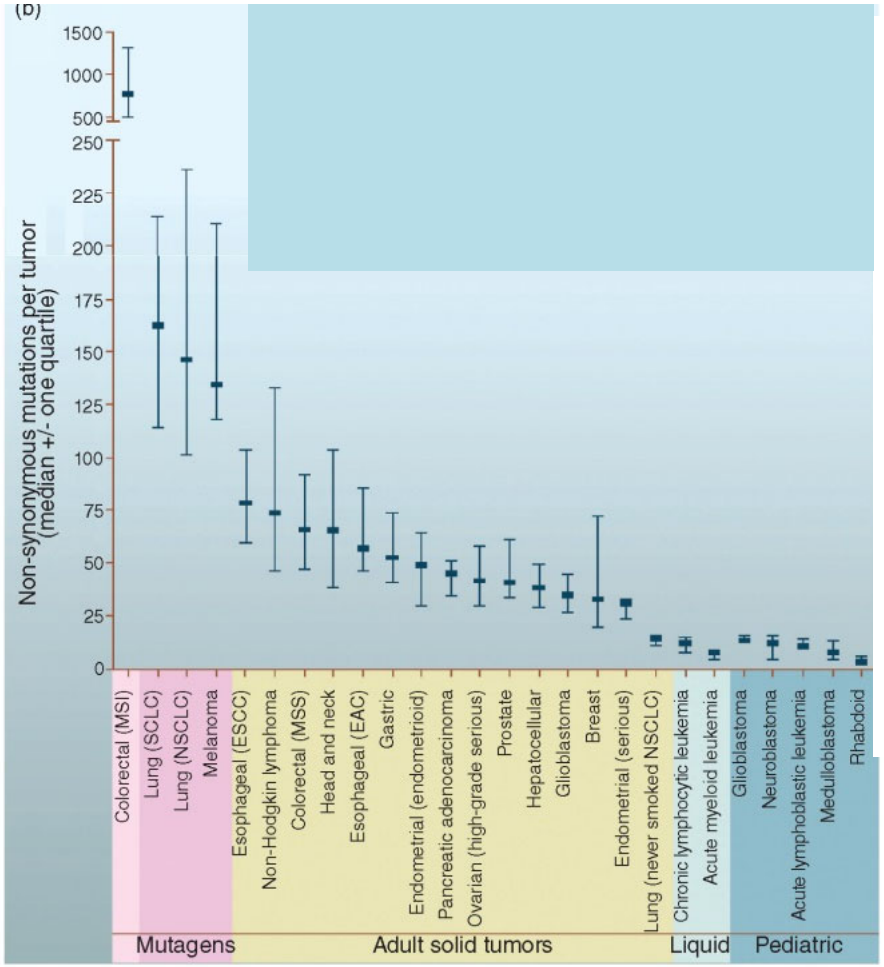
[a]Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, and [b]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; [c]Cancer Biology Program, Sidney Kimmel Cancer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







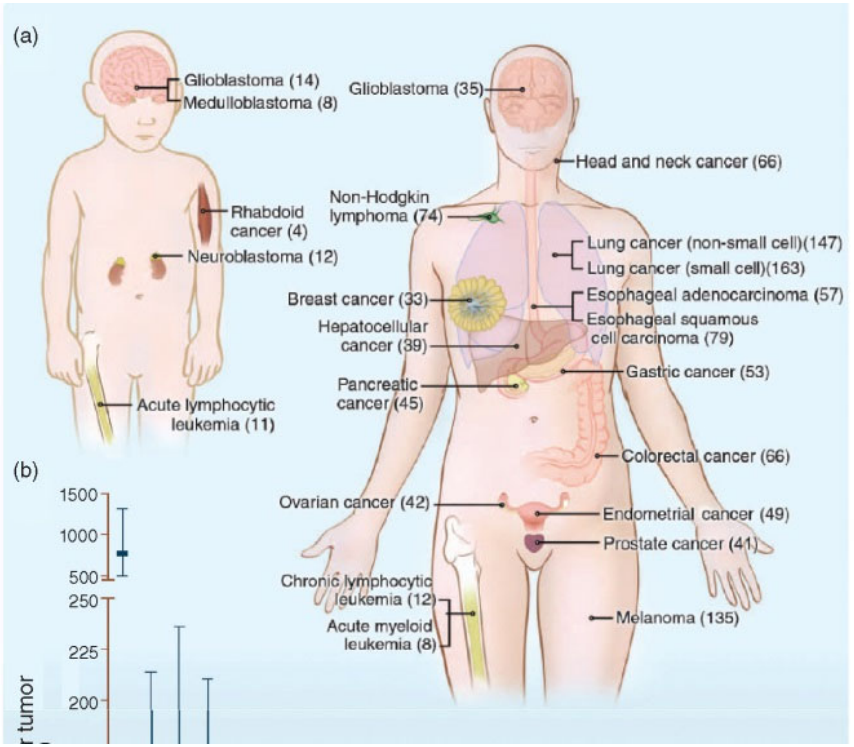


FIGURE 21.10 Somatic mutations in representative human cancers, based on genome-wide sequencing studies. (a) The genomes of adult (right) and pediatric (left) cancers are represented. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein *et al.* (2013). Reproduced with permission from AAAS. (b) Median number of nonsynonymous substitutions per tumor. Horizonal 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.

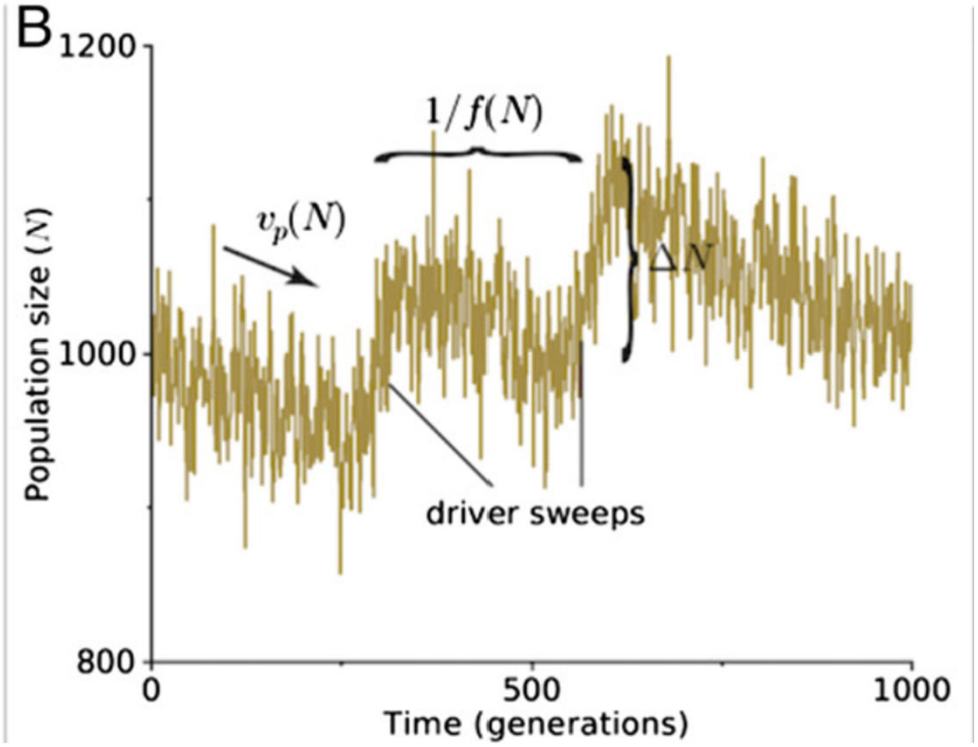
- Cancer cells carry both "Driver" and "Passengers" mutations
- Passenger mutations cause little to no harm (see later for how even little harm matters)
- Both are common as cancers elevate mutation rate

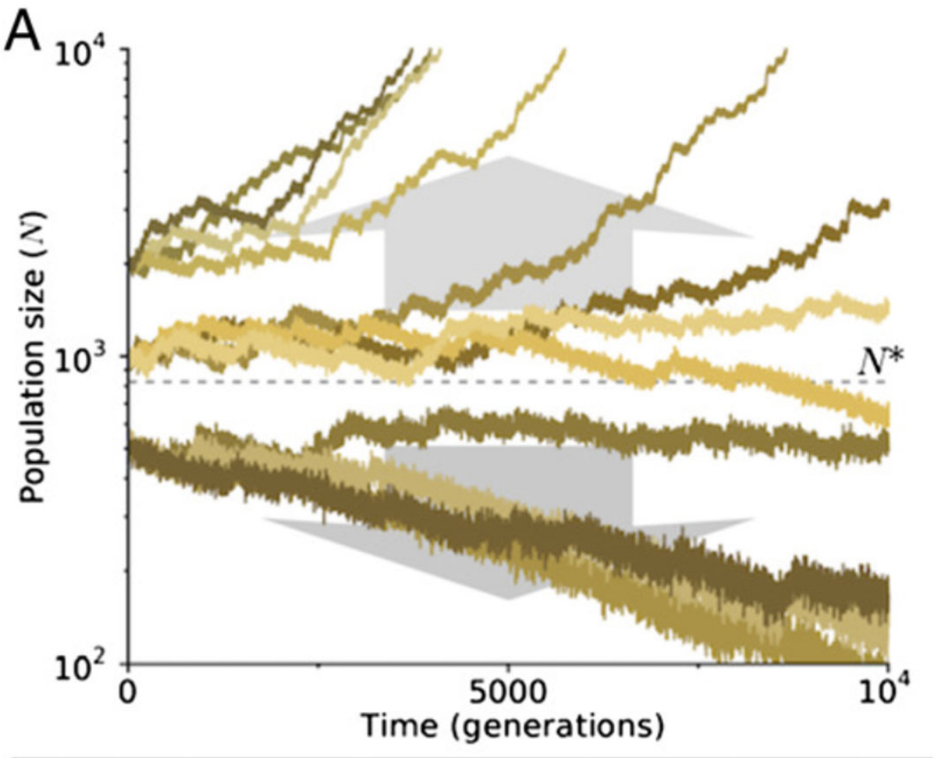
Number of passenger+driver mutations follows negative binomial distribution

- 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= Genome_target_of_driv/ (Genome_target_of_driv+Genome_target_of_pass)
  (1-p) – it is a passenger mutation

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

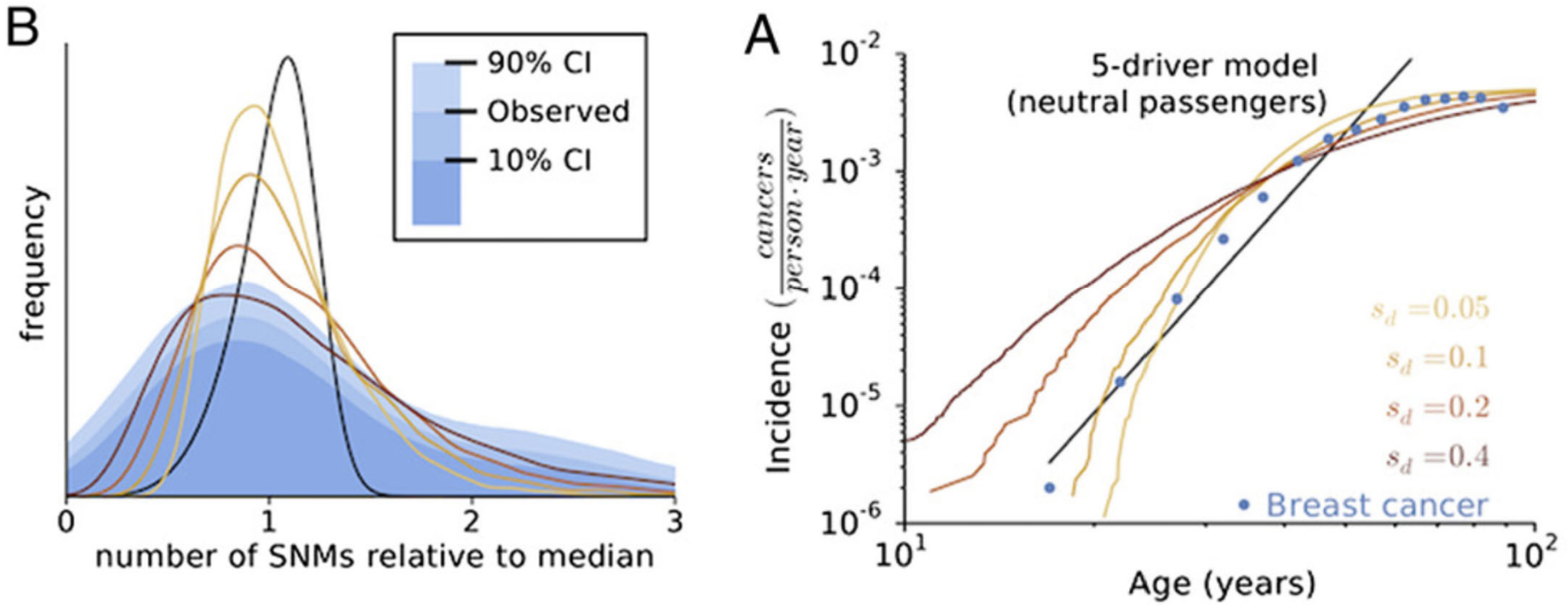
What if passenger mutations slow down the growth of cancer tumors?





McFarland CD, Mirny L, Korolev KS, PNAS 2014

Can we prove/quantify it using statistics?



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

$\mu=10^{-8}$, $Target_d=1{,}400$, $Target_p=10^7$, $s_d=0.05$ to 0.4, $s_p=0.001$

s_p/s_d for breast: 0.0060±0.0010;

melanoma: 0.016±0.003; lung: 0.0094±0.0093;

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

Credit: XKCD comics

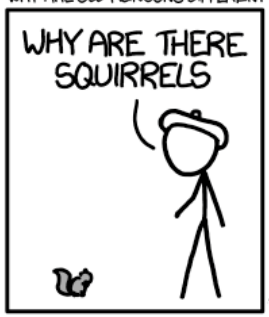
QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

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

WHY AREN'T THERE DINOSAUR GHOSTS

WHY DO IGUANAS DIE

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 IS THERE HELL IF GOD FORGIVES

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 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 GPS FREE

WHY IS SEX SO IMPORTANT

WHY ARE THERE FEMALE MR MIMES

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 ARE THERE GHOSTS

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 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 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 IS THERE NO KING IN ENGLAND

WHY ARE DOGS AFRAID OF FIREWORKS

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS

WHY IS YKK ON ALL ZIPPERS

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

WHY AREN'T MY ARMS GROWING

WHY DO I FEEL DIZZY

WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM

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 IS THERE LAVA

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