

A gallery of useful discrete probability distributions

Geometric Distribution

- A series of **Bernoulli trials** with **probability of success = p** . continued **until the first success**. X is the number of trials.
- Compare to: Binomial distribution has:
 - Fixed number of trials = n . $P(X = x) = C_x^n p^x (1 - p)^{n-x}$
 - Random number of successes = x .
- Geometric distribution has reversed roles:
 - Random number of trials, x
 - Fixed number of successes, in this case 1.
 - Success always comes in the end: so no combinatorial factor C_x^n
 - $P(X=x) = p(1-p)^{x-1}$ where:
 $x-1 = 0, 1, 2, \dots$, the number of failures until the 1st success.
- **NOTE OF CAUTION:** Matlab, Mathematica, and many other sources use x to denote the **number of failures until the first success**. We stick with **Montgomery-Runger notation**

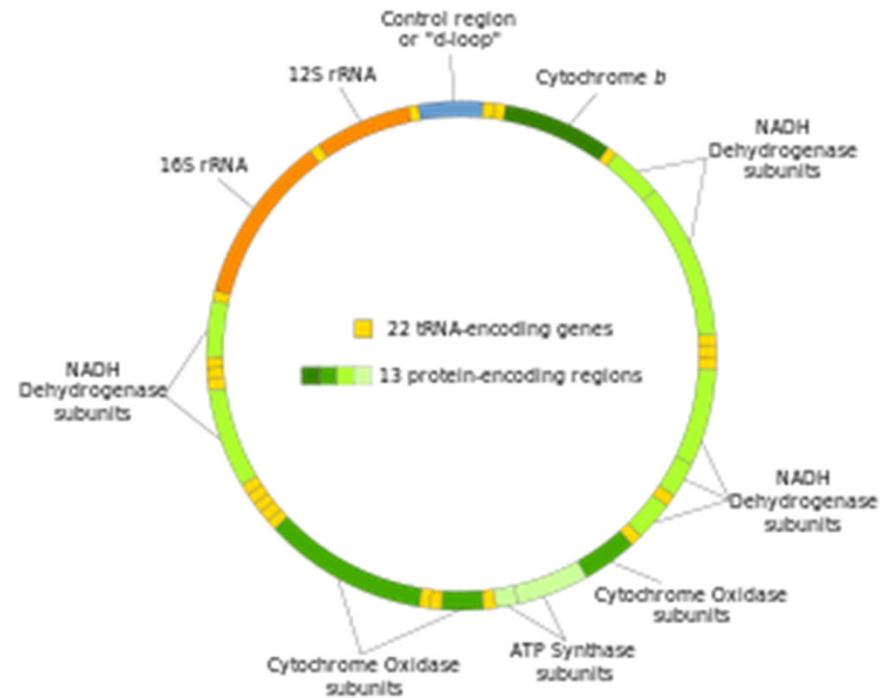
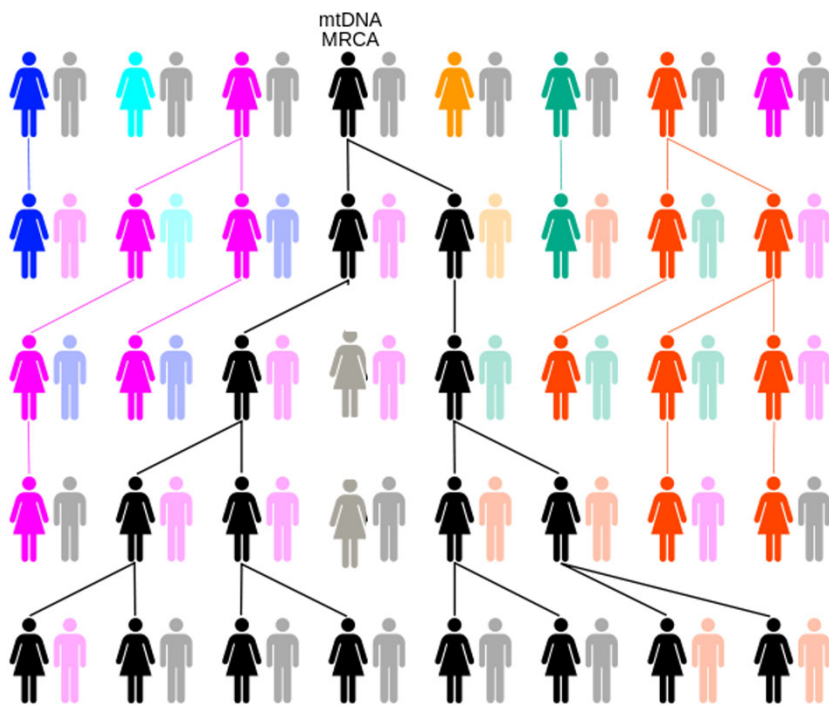
Geometric Mean & Variance

- If X is a geometric random variable (**according to Montgomery-Bulmer**) with parameter p ,

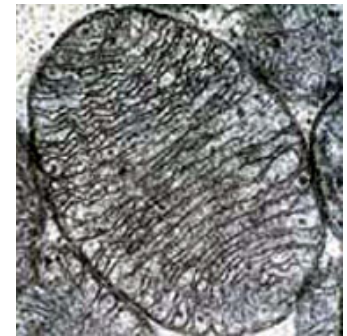
$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \quad (3-10)$$

- For small p the **standard deviation** $= (1-p)^{0.5}/p \approx$
mean $= 1/p$
- Very different from Binomial and Poisson, where
variance $=$ **mean** and **standard deviation** $=$ **mean**^{1/2}

Geometric distribution in biology

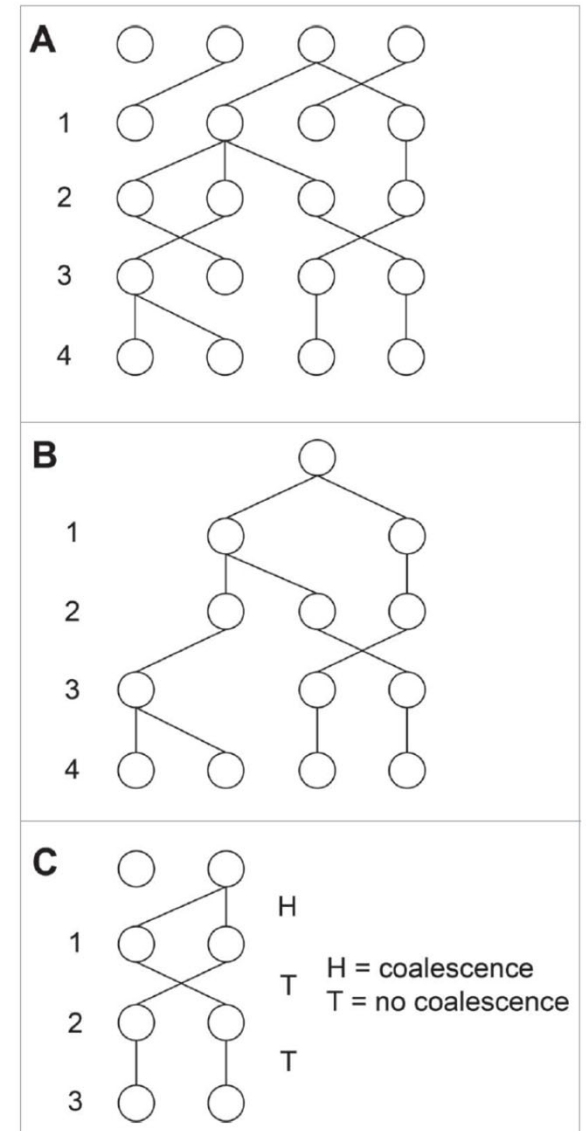


- Each of our cells has mitochondria with 16.5kb of mtDNA **inherited only from our mother**
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeon (of UIUC's Carl R. Woese fame)
- Since that time most mitochondrial genes were transferred into the nucleus
- Plants also have plastids with genomes related to cyanobacteria



Time to the last common (maternal) ancestor follows geometric distribution

- **Constant population** of N women
- **Random number** of (female) **offsprings**. Average is 1 (but can be 0 or 2)
- **Randomly** pick **two women**.
Question: how many **generations** T since their **last maternal ancestor**?
- T is a random variable What is its PMF: $P(T=t)$?
Answer: $P(T=t)$ follows a **geometric distribution**
- Do these two women have **the same mother**? Yes: **“success”** in finding their last common ancestor ($p=1/N$). $P(T=1)=1/N$.
- No? “failure” ($1-p=1-1/N$). Go to their mothers and repeat the same question.
- $P(T=t)=(1-1/N)^{t-1}(1/N) \approx (1/N) \exp(-(t-1)/N)$
- t can be inferred from **the density of differences on mtDNA** $=2\mu t$



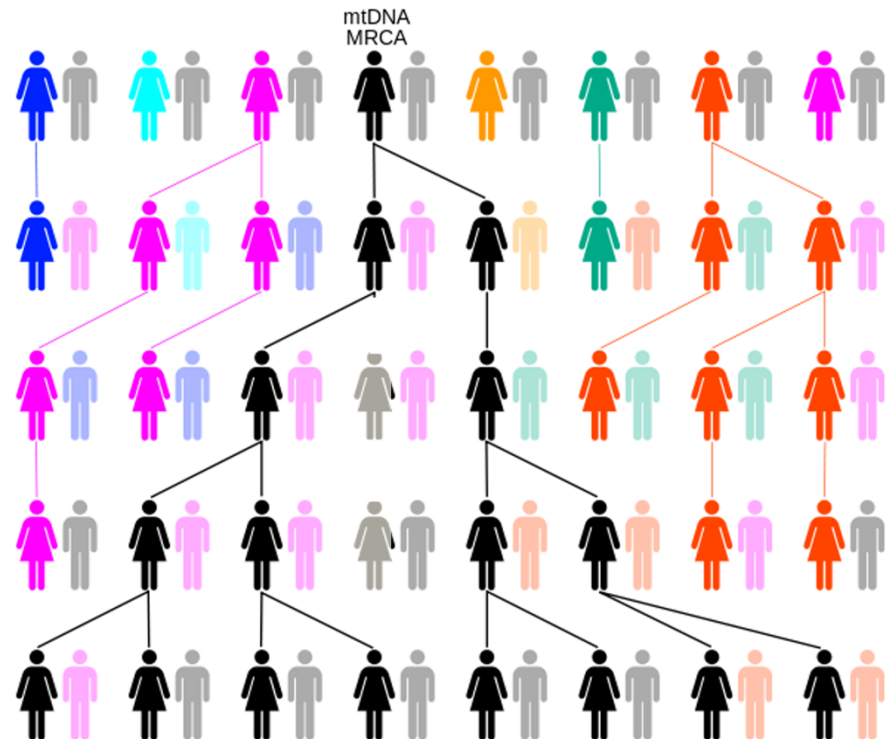
Most Recent Common Ancestor (MRCA)

- Consider N women living today. Let $A(t)$ be the number of maternal ancestors t generations before present connected to all modern women
- At $t=0$, $A(0) = N$. Any of $\frac{A(0)(A(0)-1)}{2}$ pairs of woman can coalesce with probability $\frac{1}{N}$

- We expect $\frac{1}{N} \frac{A(0)(A(0)-1)}{2}$ pairs to coalesce

$$A(1) = A(0) - \frac{1}{N} \frac{A(0)(A(0)-1)}{2}$$

- Now the expected number of coalescing pairs $\frac{1}{N} \frac{A(1)(A(1)-1)}{2}$



Most Recent Common Ancestor (MRCA)

$$\frac{dA(t)}{dt} = - \overbrace{\frac{1}{N}}^{\text{probability of coalescence}} \times \overbrace{\frac{A(t)(A(t)-1)}{2}}^{\text{number of pairs}} \approx \frac{A(t)^2}{2N}$$

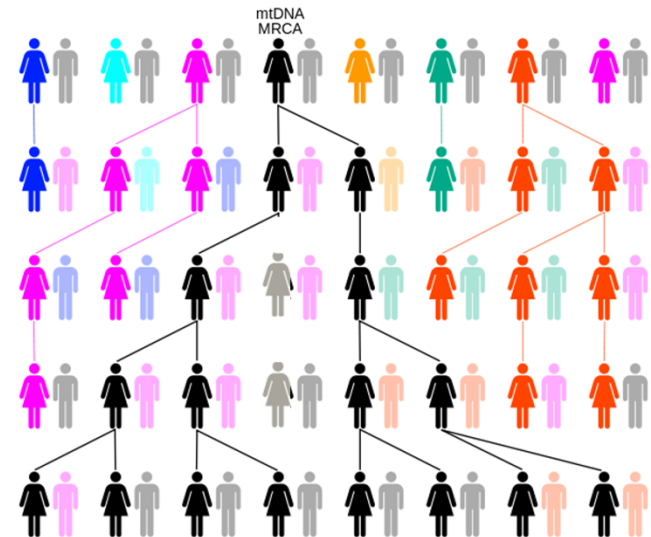
Solution: $A(t) = \frac{C}{t+2}$

Let's check: $\frac{dA}{dt} = -\frac{C}{(t+2)^2} = \frac{A^2}{C} \rightarrow C = 2N$

$$T_{MRCA} = 2(N-1) \approx 2N$$

Most Recent Common Ancestor (MRCA)

- Start with N individuals. Unit of time is N generations (time for one pair to merge) since $E(T) = \sum_{t=1}^{\infty} t \cdot (1/N) \exp(-t/N) = N$
- Any of $\frac{N(N-1)}{2}$ pairs can merge first. The average time for the first pair to merge is $\frac{2}{N(N-1)}$
- After the merger $N \rightarrow N - 1$,
- So, the time until the next merger is longer $\frac{2}{(N-1)(N-2)}$



Most Recent Common Ancestor (MRCA)

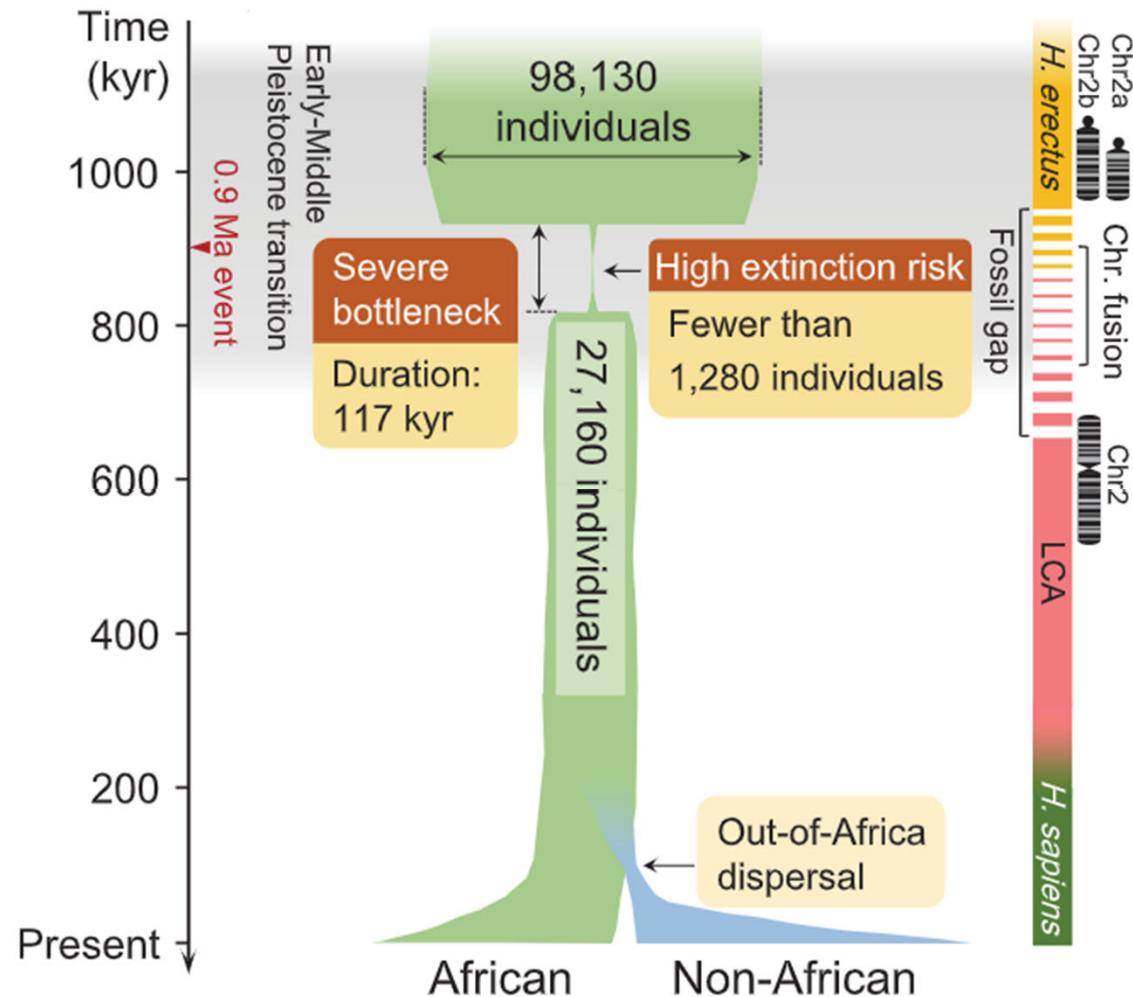
Total time until the MRCA

$$T_{MRCA} = N \cdot \sum_{k=2}^N \frac{2}{k(k-1)}$$

$$= 2N \sum_{k=2}^N \left(\frac{1}{k-1} - \frac{1}{k} \right) = 2N \left(1 - \frac{1}{N} \right) \approx 2N$$

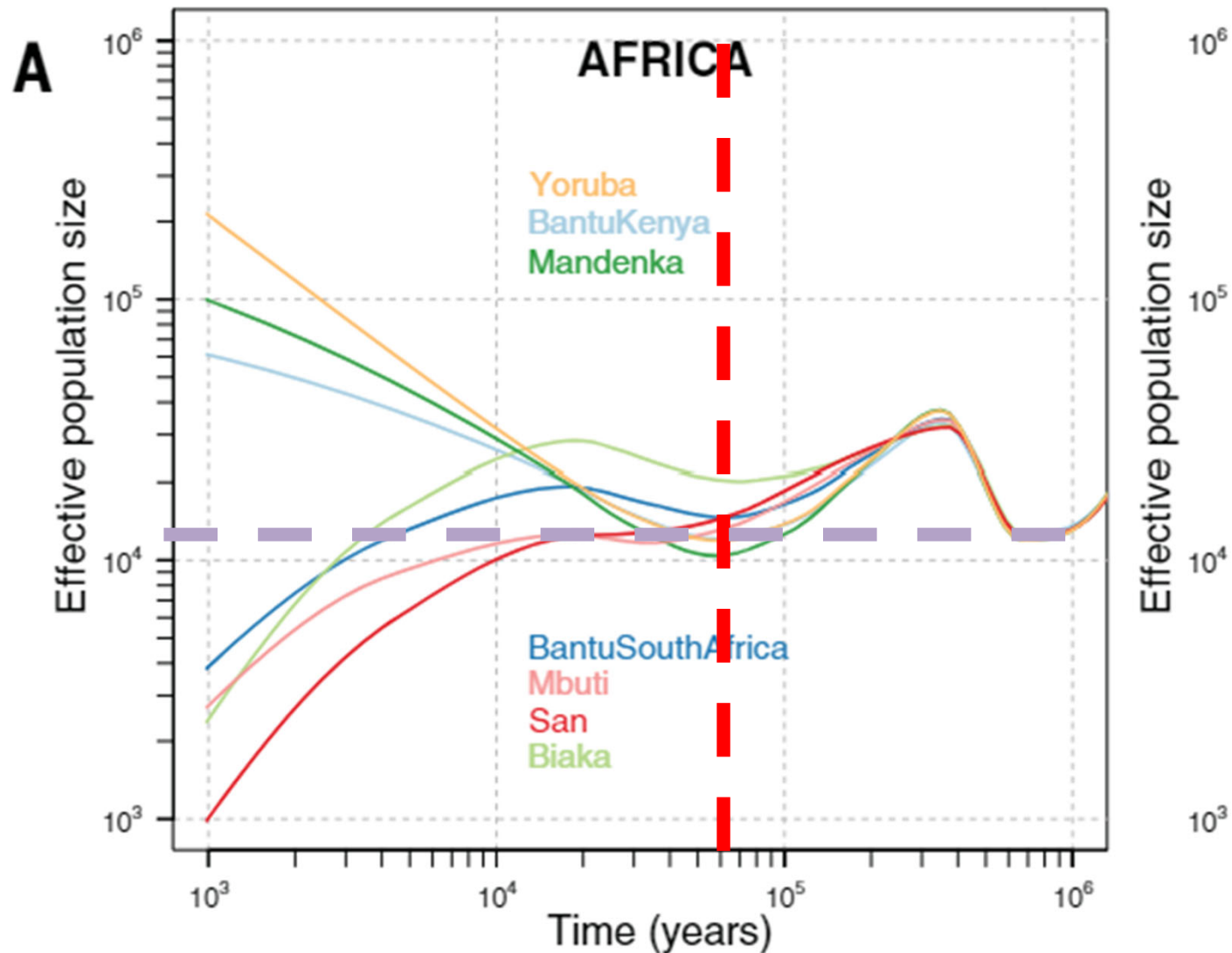
- There are about $N = 8 \times 10^9 / 2 = 4 \times 10^9$ women living today
- **M**ost **R**ecent maternal **C**ommon **A**ncessor (**MRCA**)
of all people living today lived $T_{MRCA} = 2N$ generations ago
- $T_{MRCA} = 2 \cdot 4 \times 10^9$ generations
- If the generation time 20 years it is 160 billion years > **10 times the time since the Big Bang.**
- Something is wrong here!

Hot off the press: human ancestors almost got extinct about 1M years ago



Hu W, et al. Science. 2023;381: 979–984

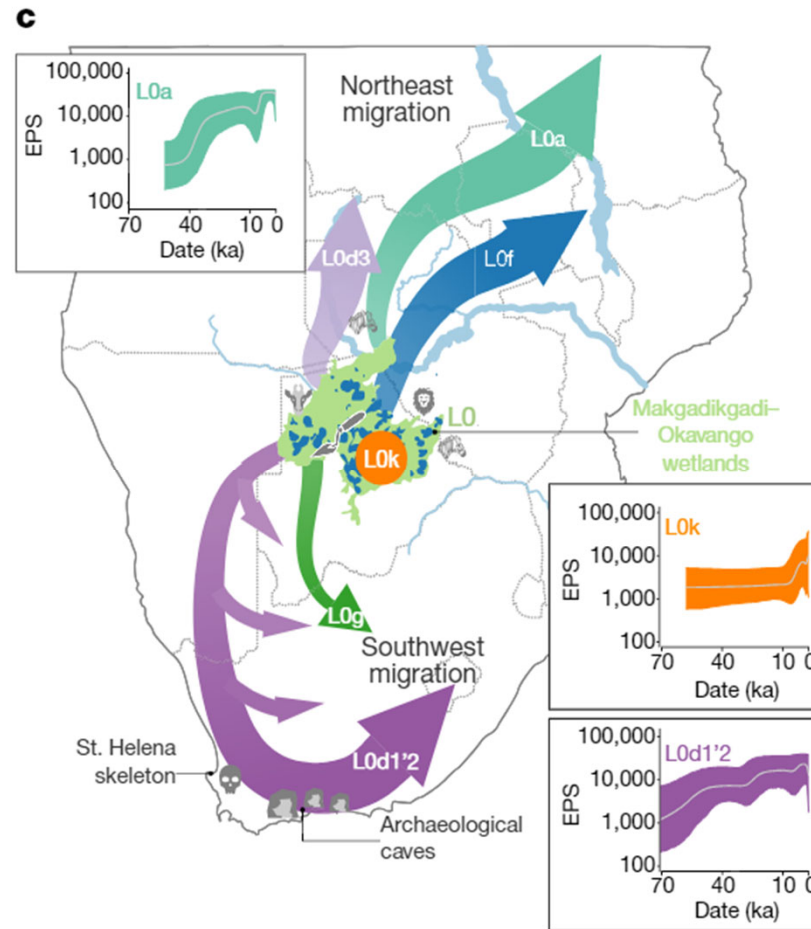
Bottleneck for human population in Africa around 10,000 individuals reached around 60,000 years ago



From ~1000 modern genomes: Bergström A, et al. Science. 2020;367

- Population is **not constant** and for a long time was very low
- Change N to the “effective” size N_e reached during **the bottleneck**
- Current thinking is that for all of humankind $N_e \sim 10,000$ people
- **Mito Eve lived in Africa** $\sim 2 * (N_e / 2) * 20$
years = $10,000 * 20$ years = **200,000 years ago**

“Mitochondrial Eve” lived in Africa



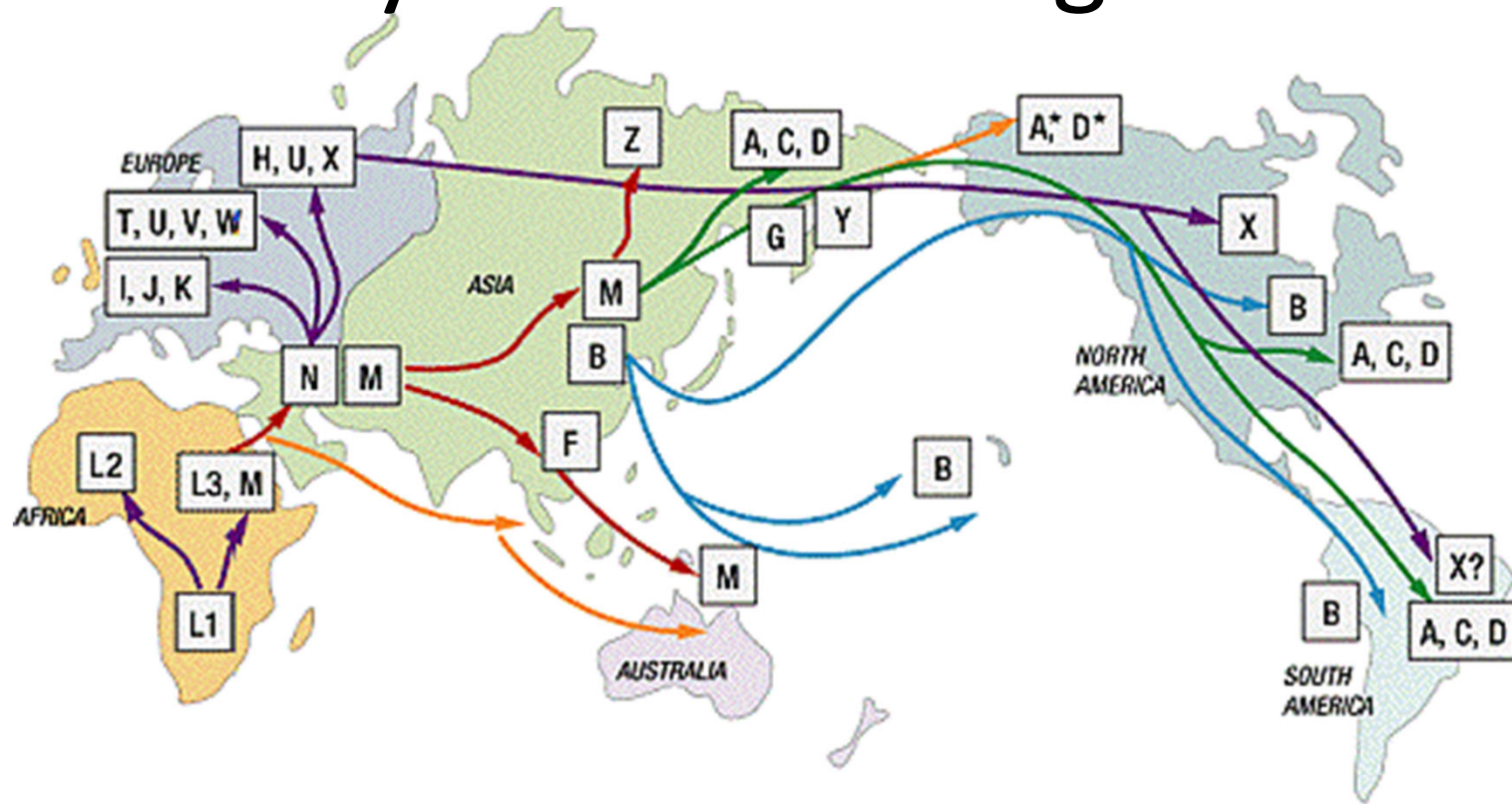
“Mitochondrial Eve” lived in Makgadikgadi–Okavango paleo-wetland of southern Africa ~200,000 years ago (between 165,000 and 240,000 years ago)

Chan EKF, et al. Nature. 2019; 575: 185–189.

Okavango Delta now



Modern mitochondrial DNA contains history of human migrations

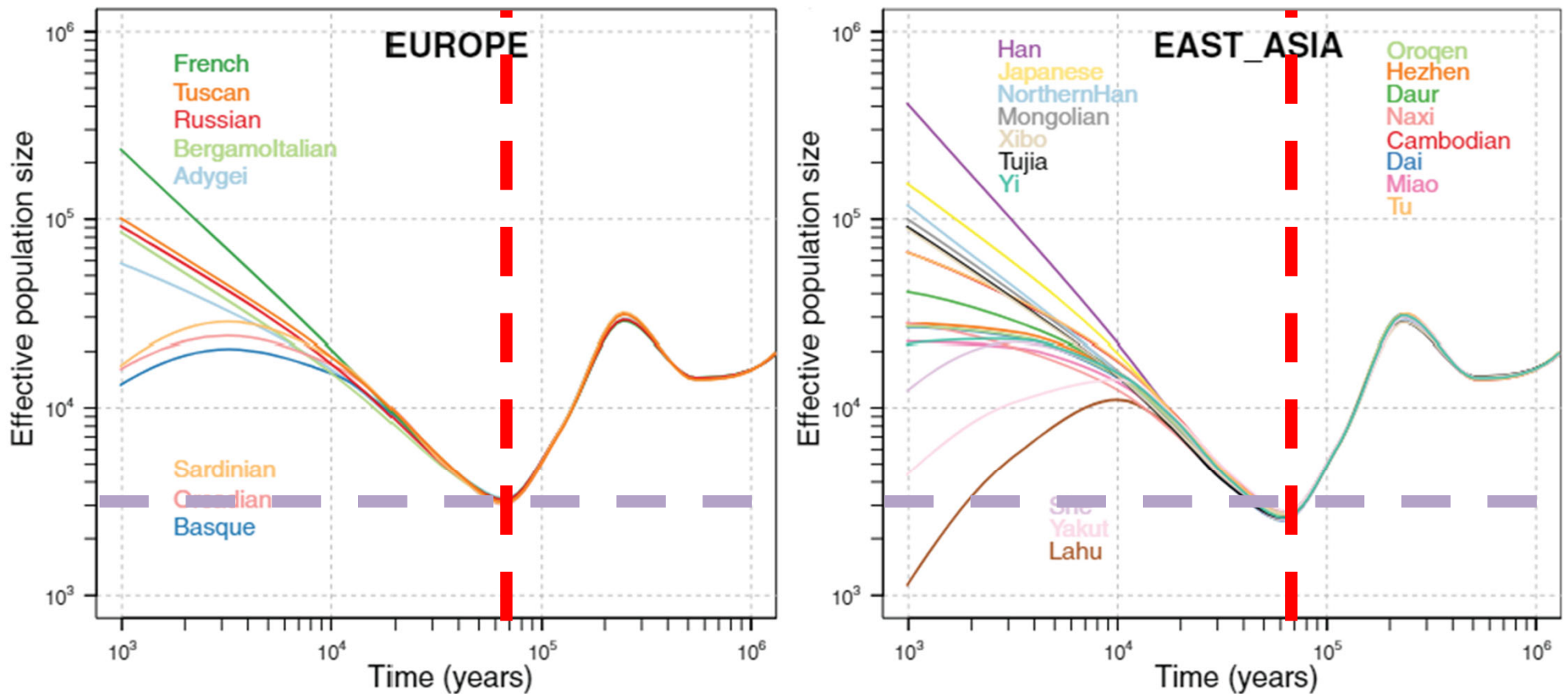


EXPANSION TIMES (years ago)	
Africa	120,000 - 150,000
Out of Africa	55,000 - 75,000
Asia	40,000 - 70,000
Australia/PNG	40,000 - 60,000
Europe	35,000 - 50,000
Americas	15,000 - 35,000
Na-Dene/Esk/Aleuts	8,000 - 10,000



Poznik GD, et al (Carlos Bustamante lab in Stanford), *Science* **341**: 562 (August 2013).

- Bottleneck for human population in Europe and Asia was around 3,000 individuals reached around 70,000 years ago
- Non-African Eve lived ~60,000 years ago



From ~1000 modern genomes: Bergström A, et al. Science. 2020;367

What about men?

- Y-chromosome is transferred from father to son
- Like mitochondria it can be used to trace ancestry of all men to the “Y-chromosome Adam”
- Where did “Adam” live? Did he meet the “mitochondrial Eve”?

Y-chromosomal Adam also lived in Africa

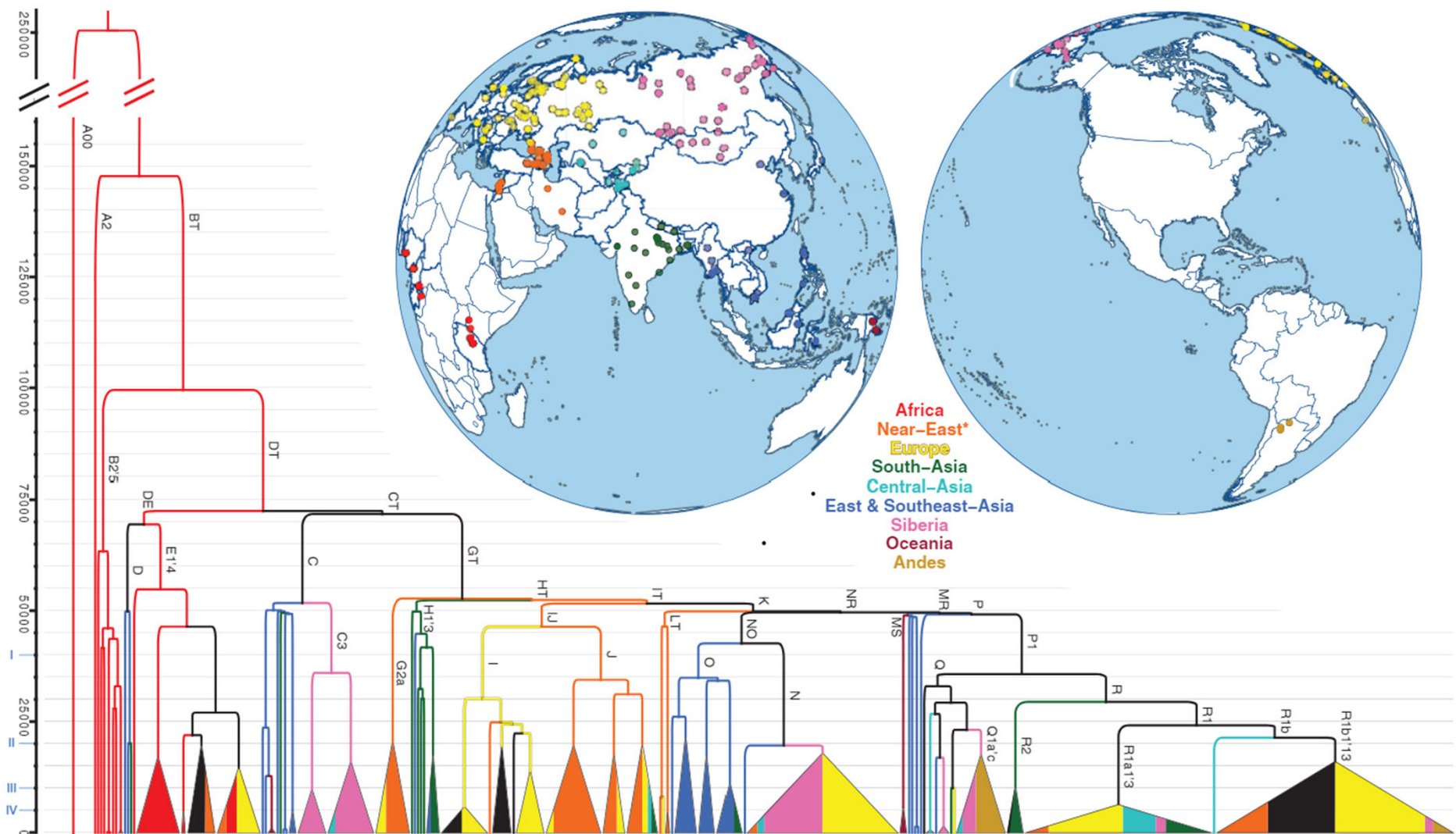
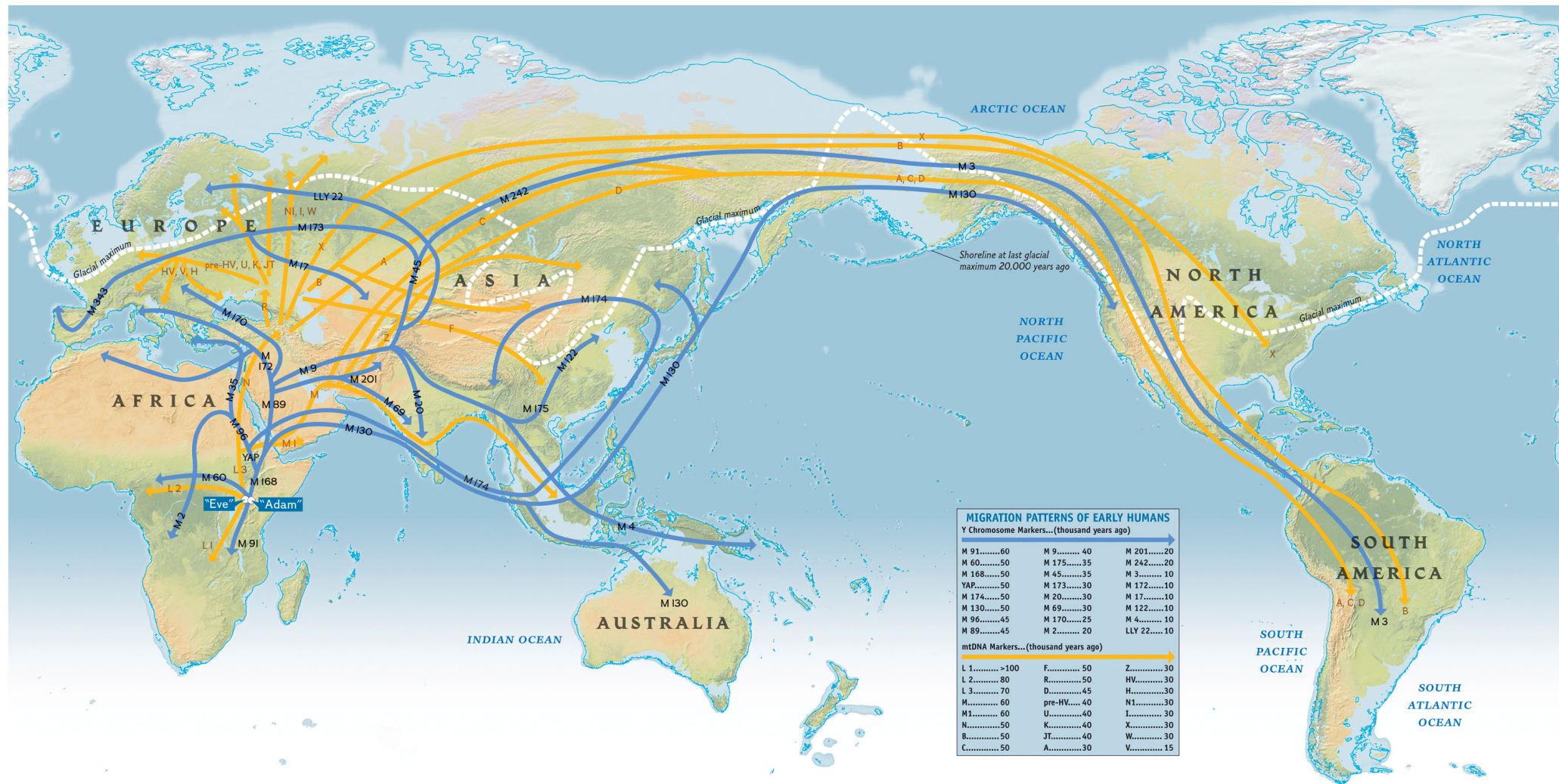


Figure 1. The phylogenetic tree of 456 whole Y chromosome sequences and a map of sampling locations. The phylogenetic tree is reconstructed using BEAST. Clades coalescing within 10% of the overall depth of the tree have been collapsed. Only main haplogroup labels are shown (details are provided in Supplemental Information 6). Colors indicate geographic origin of samples (Supplemental Table S1), and fill proportions of the collapsed clades represent the proportion of samples from a given region. Asterisk (*) marks the inclusion of samples from Caucasus area. Personal Genomes Project (<http://www.personalgenomes.org>) samples of unknown and mixed geographic/ethnic origin are shown in black. The proposed structure of Y chromosome haplogroup naming (Supplemental Table S5) is given in Roman numbers on the y-axis.

Karmin M, Saag L, Vicente M, Sayres MAW, Järve M, Talas UG, et al. *Genome Res.* 2015;25: 459–466.

“Adam” and “Eve” both lived in Africa



- “Mitochondrial Eve” lived in Africa between 100,000 and 240,000 years ago
- “Y-chromosome Adam” also lived in Africa between 120,000 and 160,000 years ago
- *Poznik GD, et al (Carlos Bustamante lab in Stanford), Science* **341**: 562 (August 2013).

Mitochondrial Eve (maternally transmitted ancestry)
Y-chromosome Adam (paternally transmitted ancestry)
lived ~200,000 years ago.

**When lived the latest common ancestor
shared by all of us based on nuclear DNA?**

- A. 1 million years ago
- B. 200,000 years ago
- C. 3400 years ago
- D. 1320 years ago
- E. Yesterday, I really have no clue

Get your i-clickers

Mitochondrial Eve (maternally transmitted ancestry)
Y-chromosome Adam (paternally transmitted ancestry)
lived ~200,000 years ago.

When lived the latest common ancestor
shared by all of us based on nuclear DNA?

- A. 1 million years ago
- B. 200,000 years ago
- C. 3400 years ago
- D. 1320 years ago
- E. Yesterday, I really have no clue

Get your i-clickers

Last common ancestor in nuclear (non Y-chr) DNA is another matter

- Unlike Mito or Y-chromosome, **nuclear DNA gets mixed with every generation**
 - Each of us gets 1/2 of nuclear DNA from the father and 1/2 from the mother
 - Each of us has 2 parents, 4 grandparents, 8 great-grand parents ...
- If one assumes:
 - Well-mixed marriages (not true: mostly local marriages)
 - Constant size population (not true: much smaller in the past)
 - In 33 generations the number of ancestors:
 $2^{33} = 8 \text{ billion}$ = 8 billion people living today
- Every pair of us living today should have at least one shared ancestor who lived
 - 33 generations * 20 years/generation = 660 years ago ~1360 AD
- **Assuming $T_{MRCA} = 2 T_{\text{average pairwise ancestor}} = 2 \cdot 660 \text{ years ago} = 700 \text{ AD}$**

Corrected for (mostly) local marriages and rare migrations

562

NATURE | VOL 431 | 30 SEPTEMBER 2004 |

Modelling the recent common ancestry of all living humans

Douglas L. T. Rohde¹, Steve Olson² & Joseph T. Chang³

¹Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

²7609 Sebago Road, Bethesda, Maryland 20817, USA

³Department of Statistics, Yale University, New Haven, Connecticut 06520, USA

With 5% of individuals migrating out of their home town, 0.05% migrating out of their home country, and 95% of port users born in the country from which the port emanates, the simulations produce a mean **MRCA date of 1,415 BC** and a mean **IA date of 5,353 BC**.

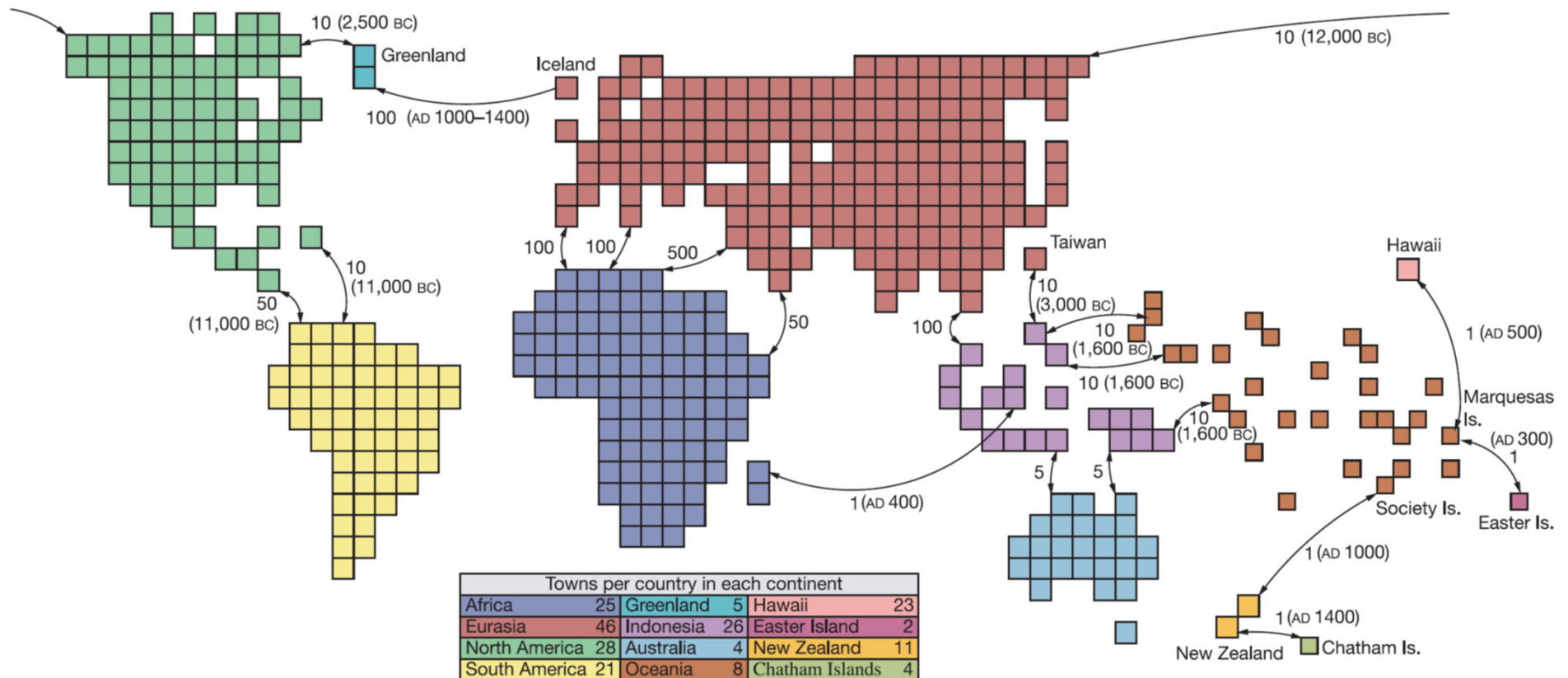
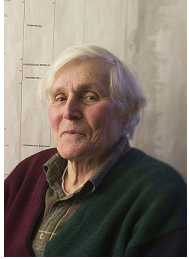


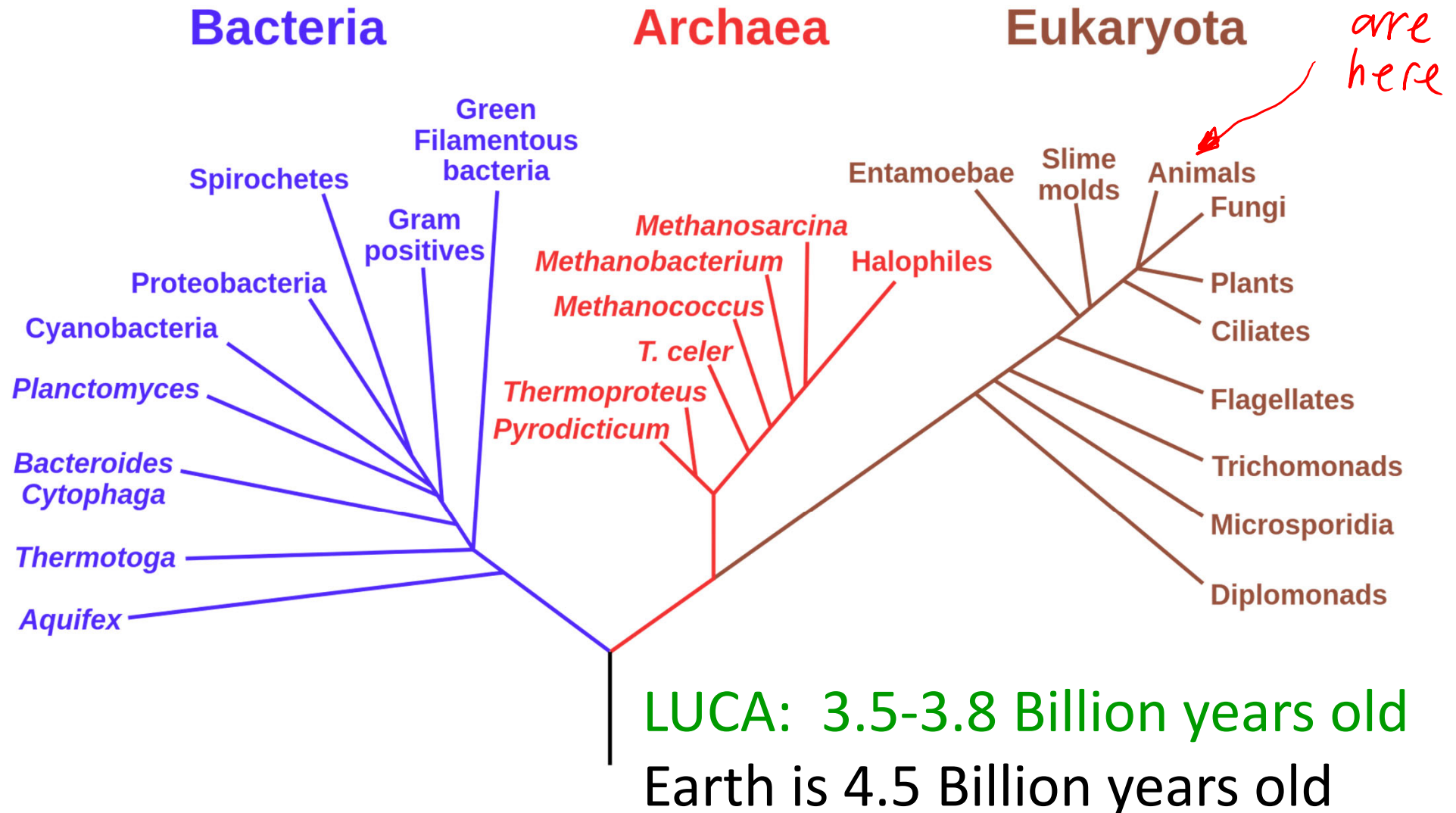
Figure 2 Geography and migration routes of the simulated model. Arrows denote ports and the adjacent numbers are their steady migration rates, in individuals per generation. If

given, the date in parentheses indicates when the port opens. Upon opening, there is usually a first-wave migration burst at a higher rate, lasting one generation.

Last Universal Common Ancestor (LUCA)



Archaea were discovered here at **UIUC** in 1977
by **Carl R. Woese (1928-2012)** and **George E. Fox**



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

WHY AREN'T MY
ARMS GROWING



WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

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 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 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 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 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 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 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 BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KINGDOMS 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 IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE

WHY ARE THERE FEMALE MR NIMES

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:

$$0 < p < 1 \text{ and } r = 1, 2, 3, \dots$$

- The probability mass function is:

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

- Compare it to binomial

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

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 ,

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

- Compare to **geometric** distribution:

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

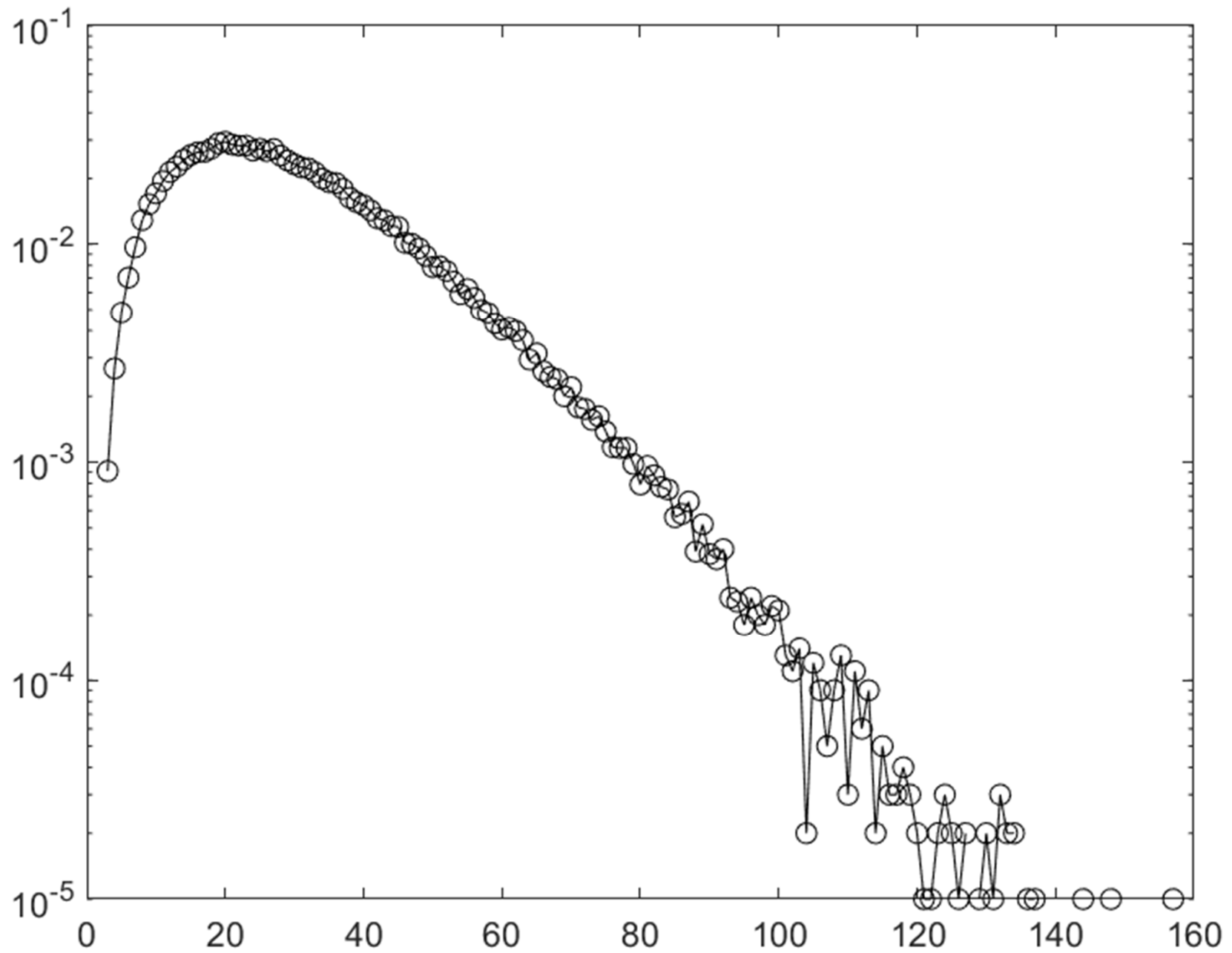
Matlab exercise

- 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 $p=0.1$, $r=100$

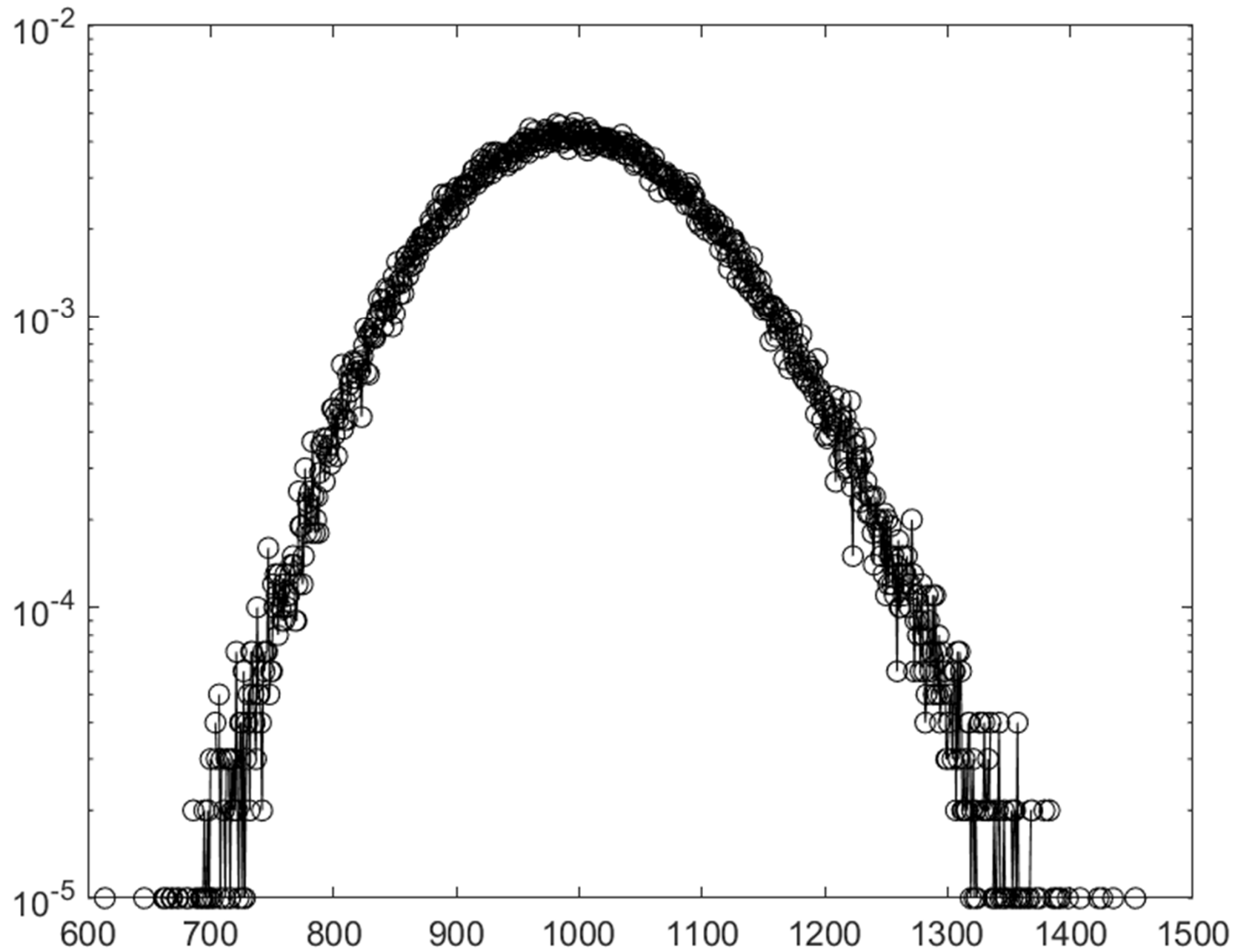
Matlab: Negative binomial distribution

- 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

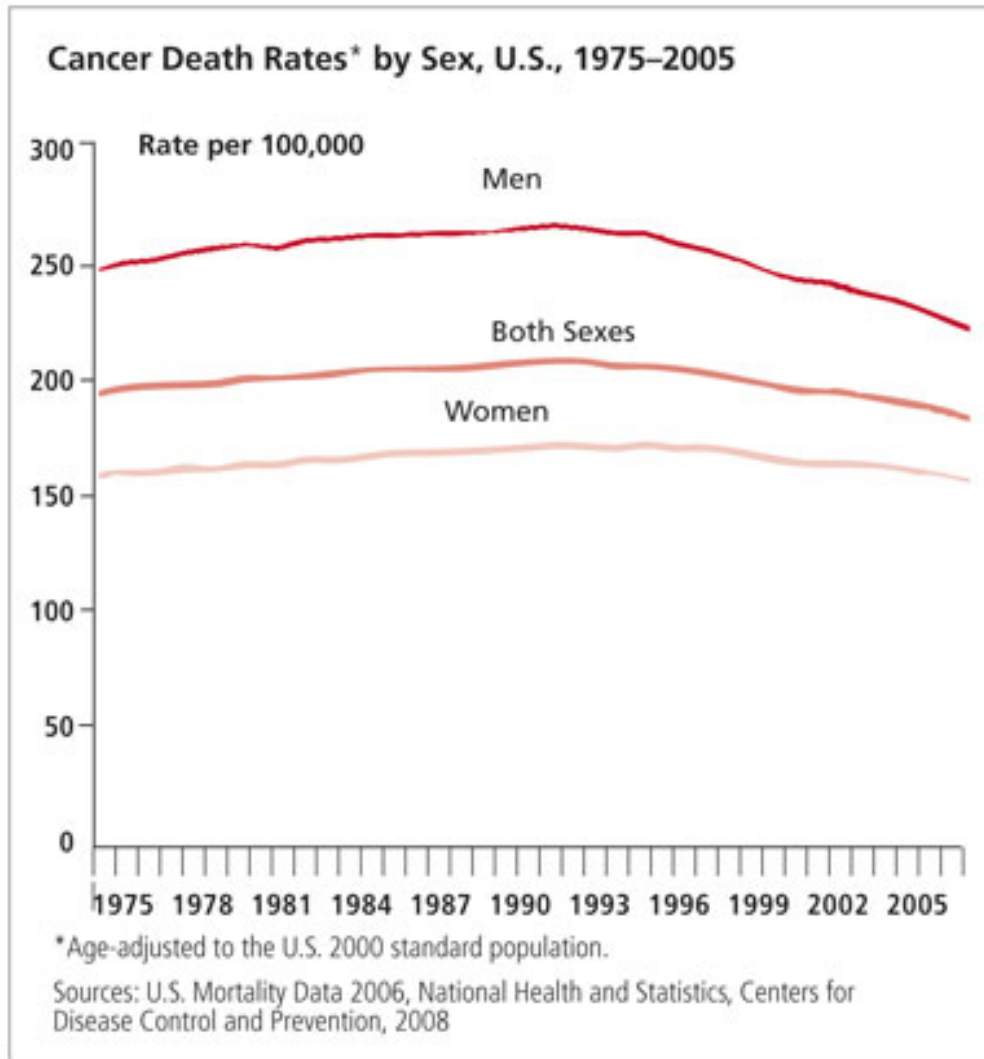


Figure 2

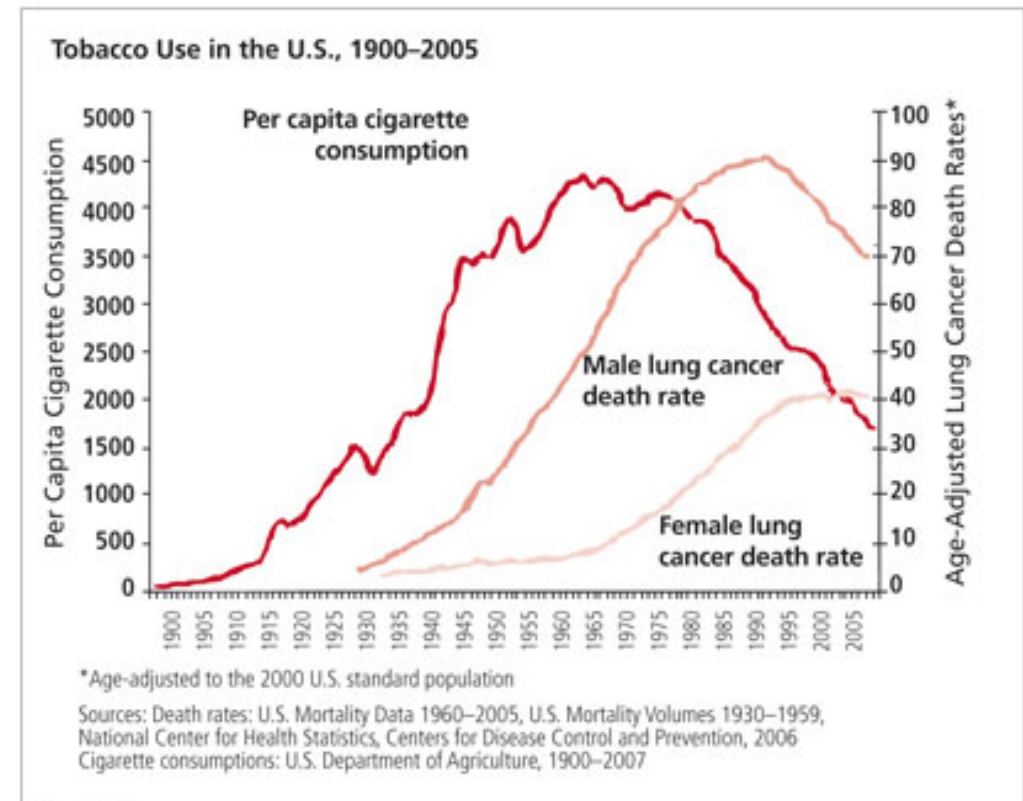


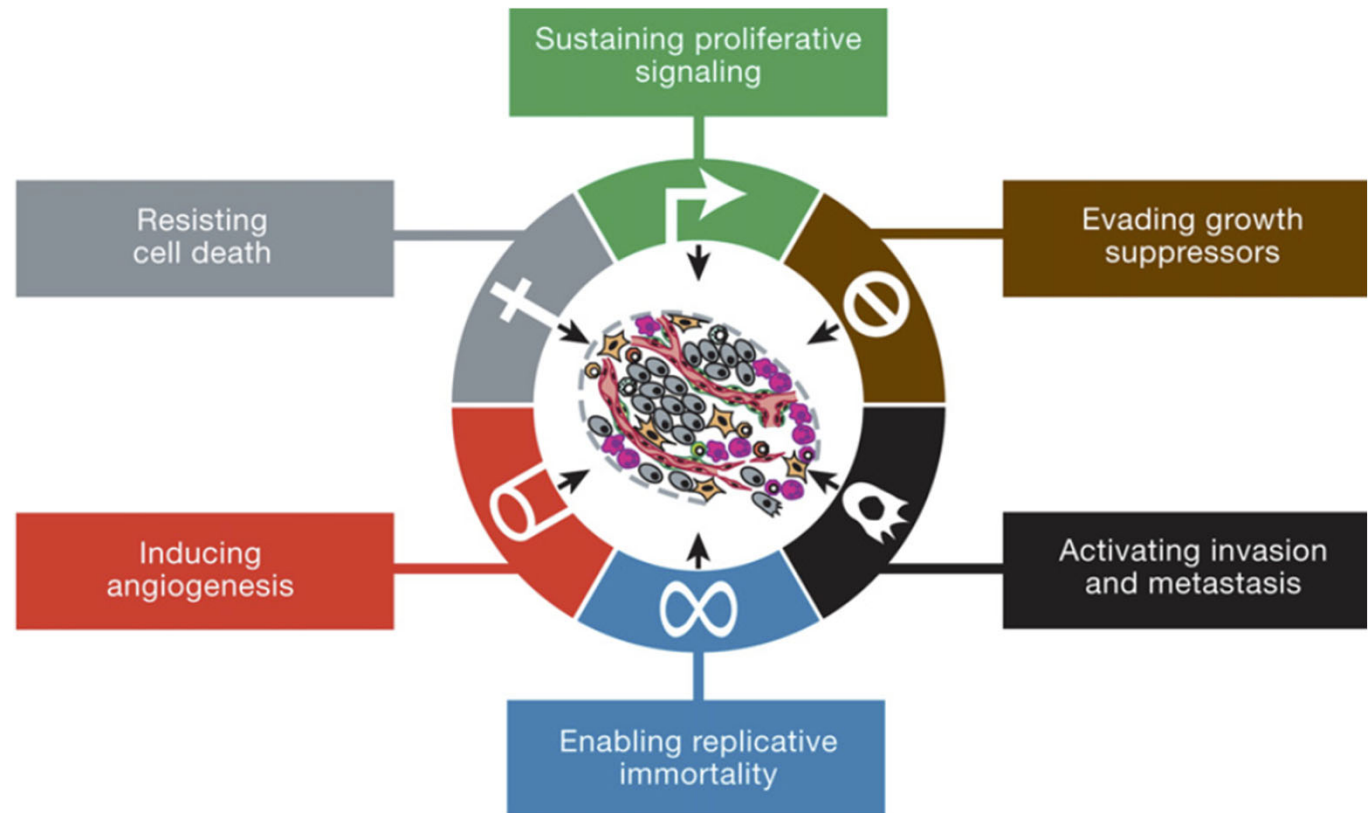
Figure 3

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

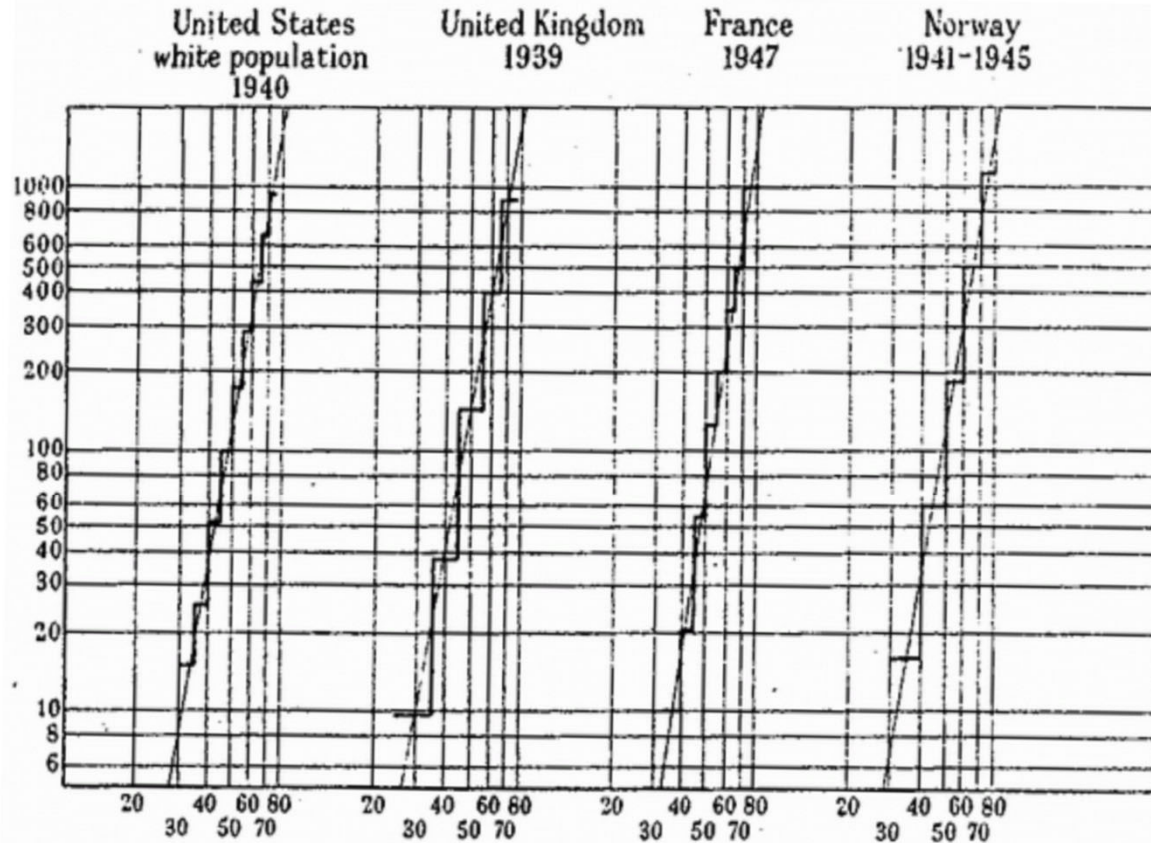


FIG. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

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

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

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?

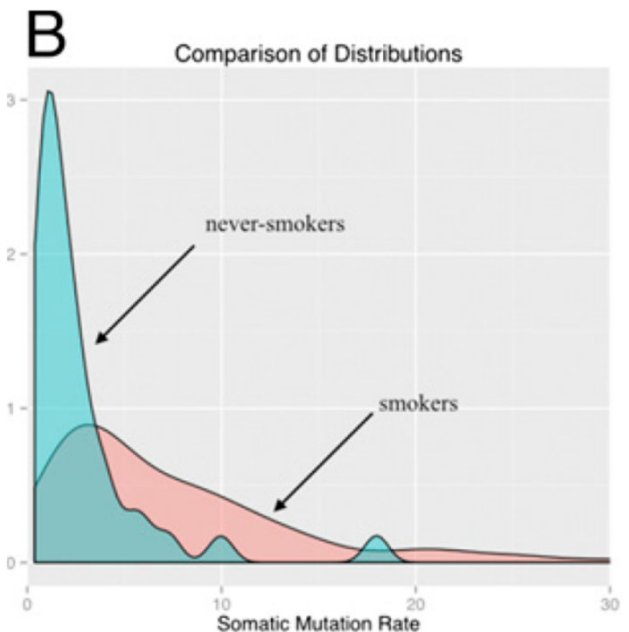
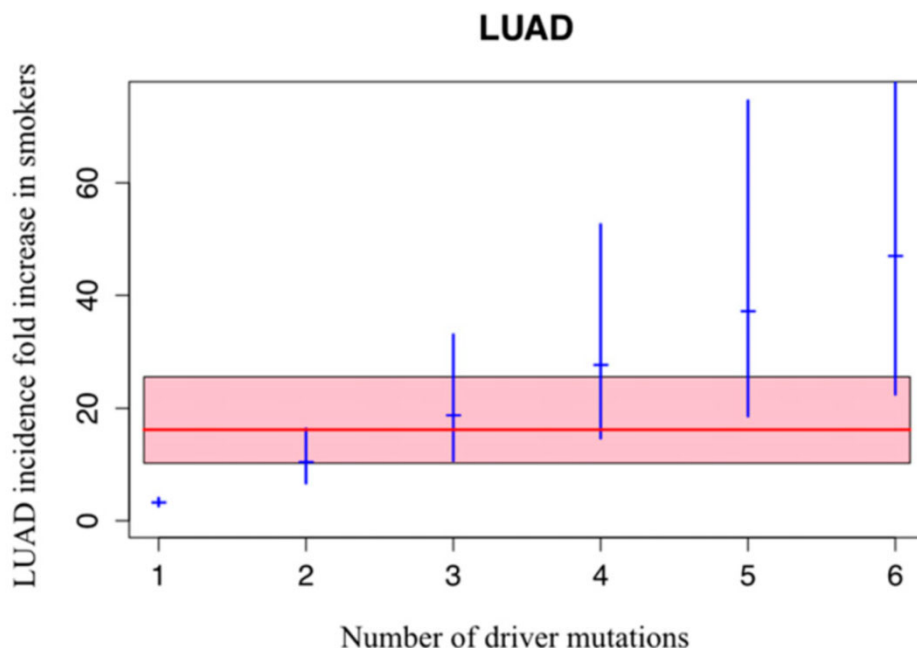
Only three driver gene mutations are required for the development of lung and colorectal cancers

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

^aDivision of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, and ^bDepartment of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; ^cCancer Biology Program, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; ^dProgram for Evolutionary Dynamics, Department of Mathematics, Harvard University, Cambridge, MA 02138; ^eDepartment of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02215; and ^fLudwig Center for Cancer Genetics and Therapeutics and ^gHoward 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



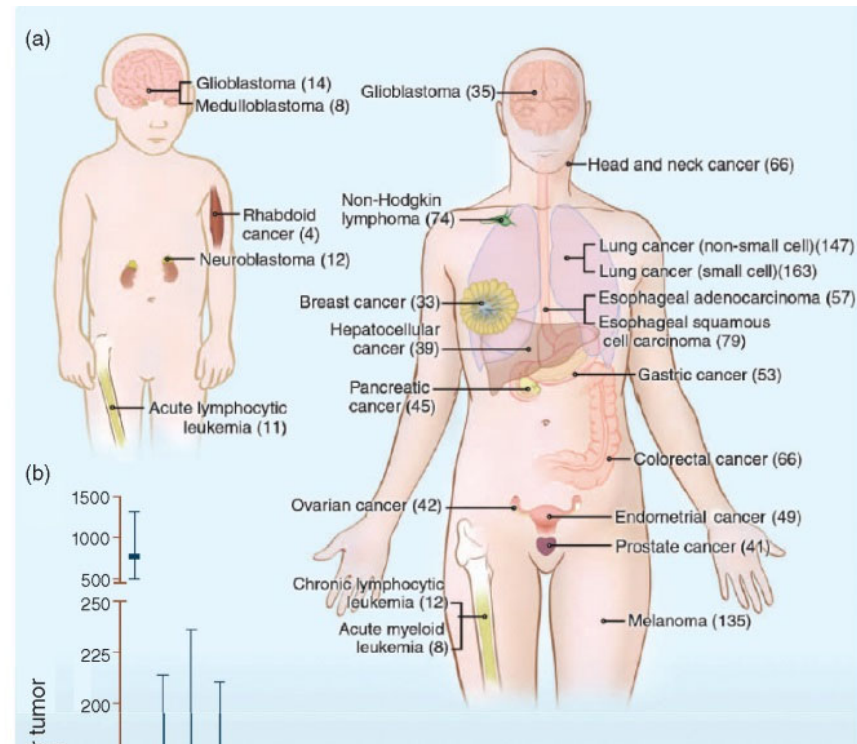
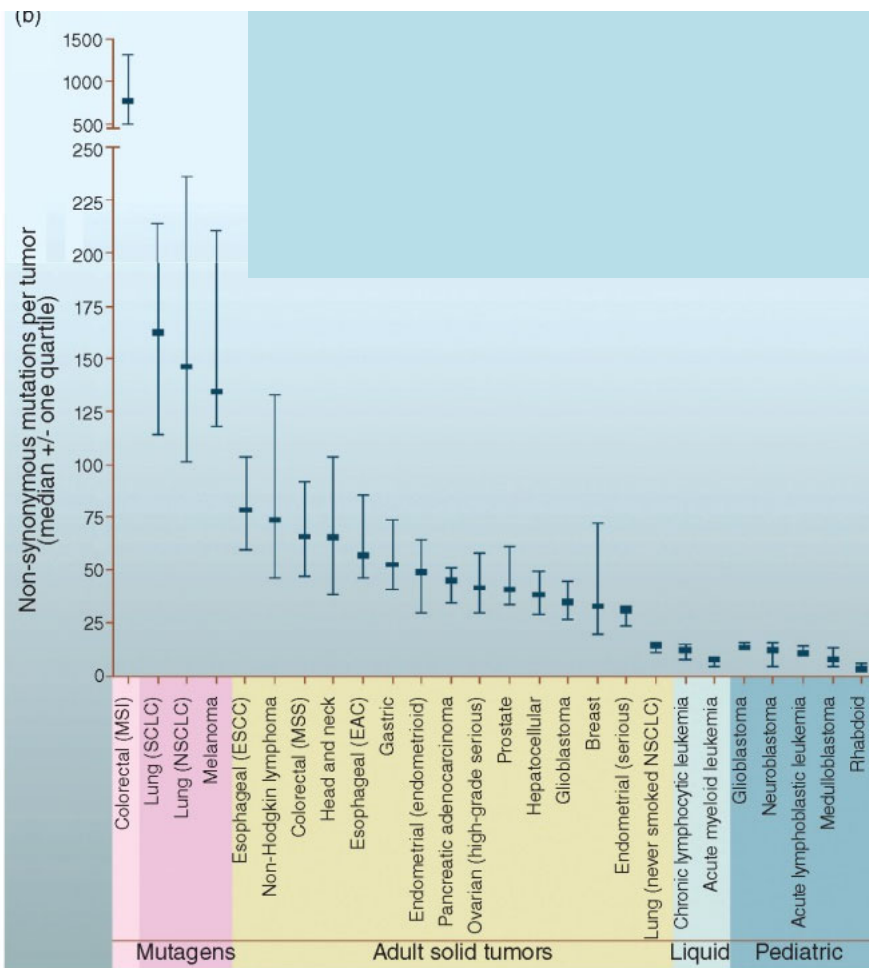


FIGURE 21.10 Somatic mutations in representative human cancers, based on genome-wide sequencing studies. (a) The genomes of adult (right) and pediatric (left) cancers are represented. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein *et al.* (2013).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

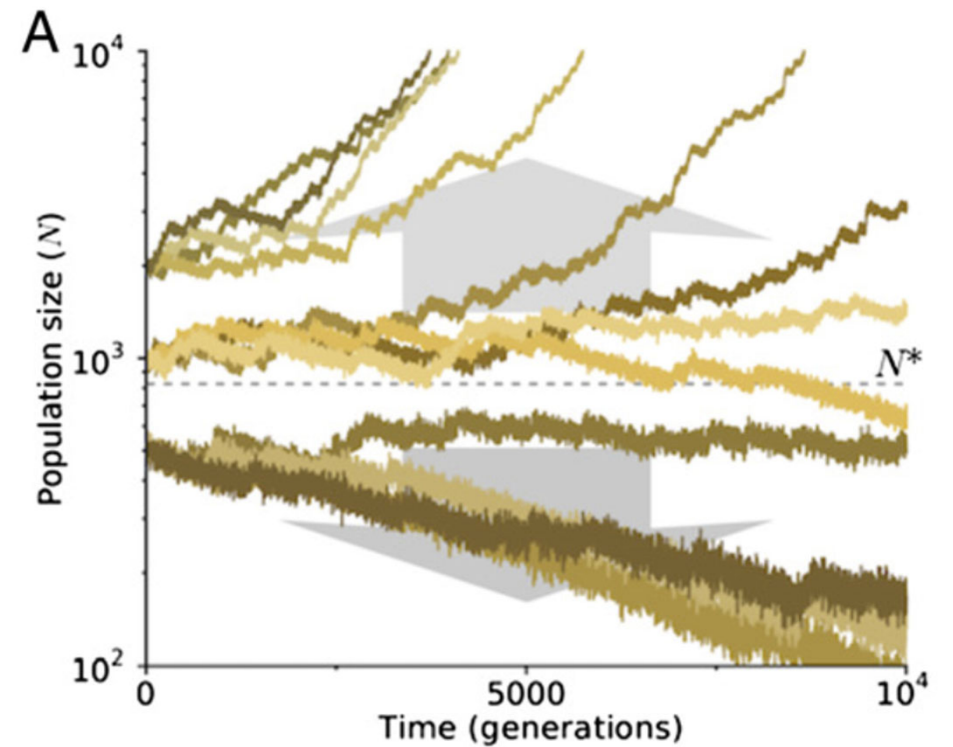
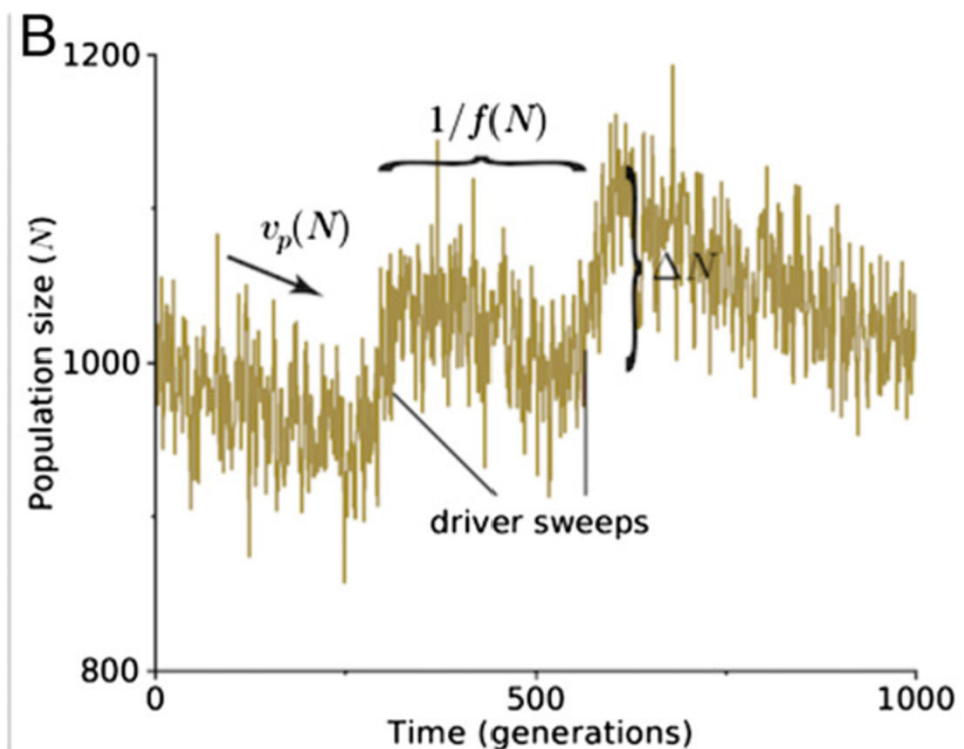
- Cancer cells carry both “Driver” and “Passengers” mutations
- Passenger mutations cause little to no harm (see later for how even little harm matters)
- Both are common as cancers elevate mutation rate

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**

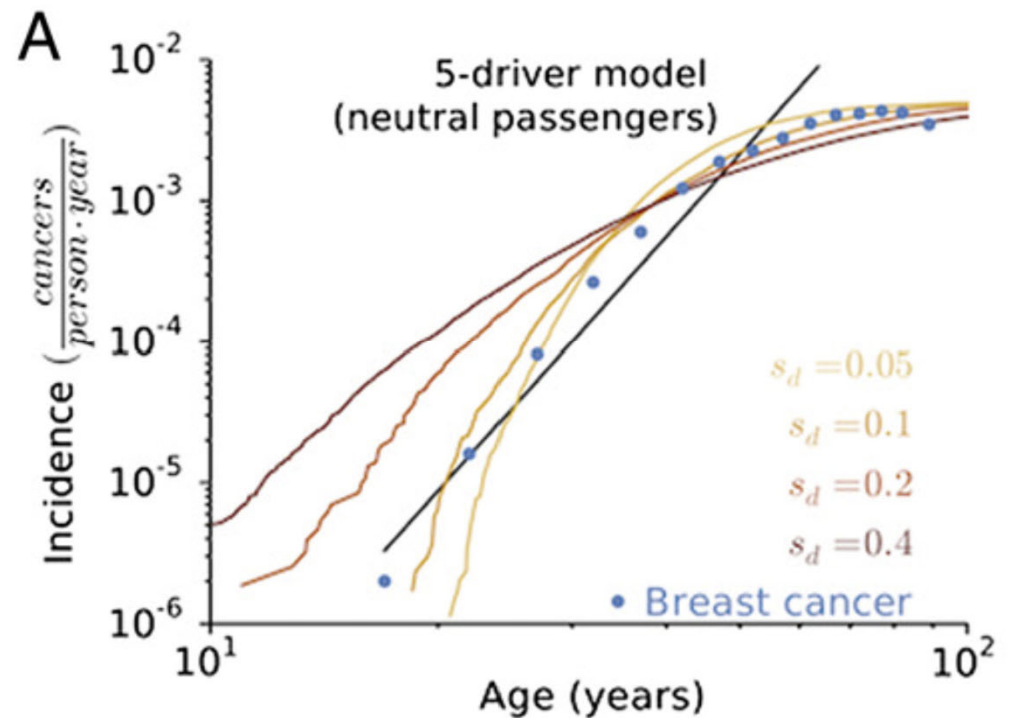
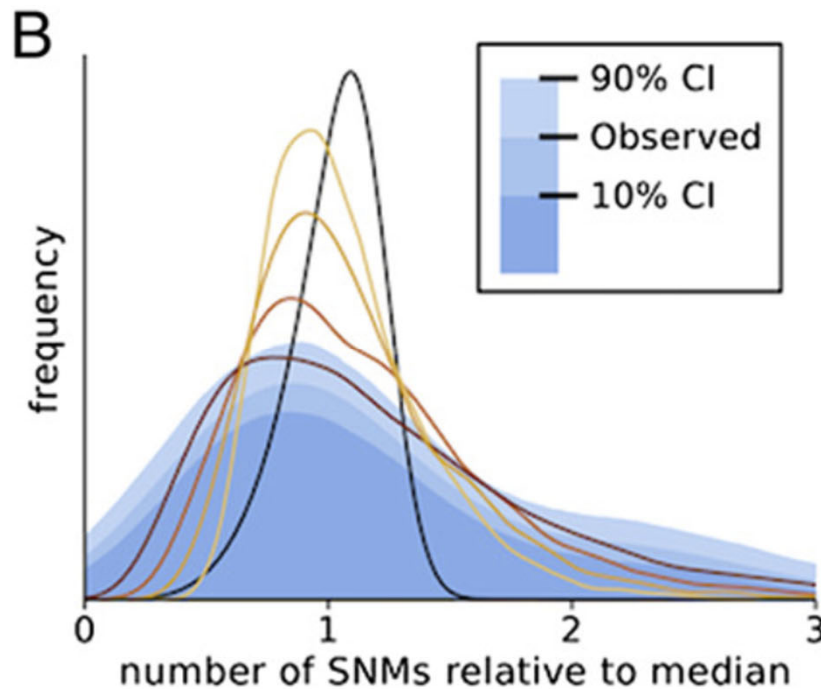
$$P(n_p + k \mid p, k) = \binom{n_p + k - 1}{n_p} (1-p)^{n_p} p^k$$

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 = $(1+s_d)^{N_d} / (1+s_p)^{N_p}$

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

s_p/s_d for breast: 0.0060 ± 0.0010 ;

melanoma: 0.016 ± 0.003 ; lung: 0.0094 ± 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

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

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 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 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 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 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 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 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 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 BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KINGDOMS 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 IS SEX
SO IMPORTANT



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 DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND