Discrete Probability Distributions Geometric and Negative Binomial illustrated by Mitochondrial Eve and Cancer Driver/Passenger Genes

Binomial Distribution

- Number of successes in n independent Bernoulli trials
- The probability mass function is:

$$P\left(X = x\right) = C_x^n p^x \left(1 - p\right)^{n-x} \text{ for } x = 0, 1, \ldots n \qquad \text{(3-7)}$$

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.
  - Random number of successes = x.

$$P\left(X=x\right)=C_x^n p^x\left(1-p\right)^{n-x}$$

- 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, … , 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*,

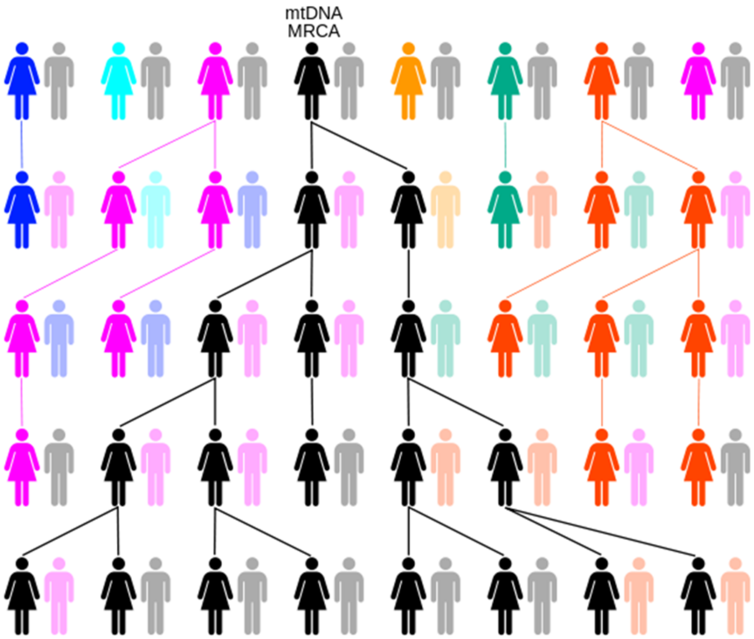
$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \qquad (3\text{-}10)$$

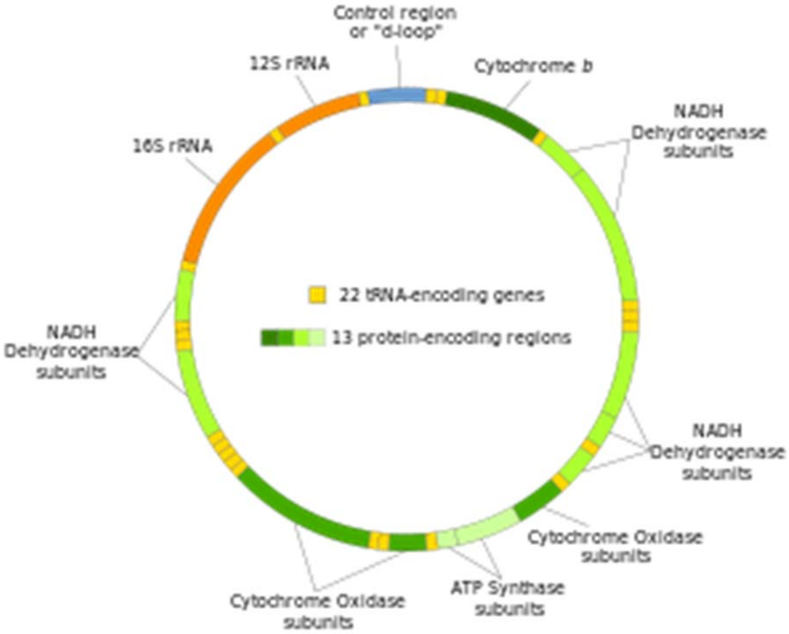
- For small *p* the standard deviation ~= mean

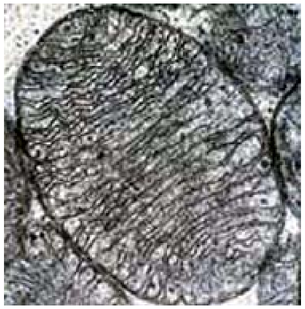
- Very different from Poisson, where it is

variance = mean and standard deviation = mean$^{1/2}$

Geometric distribution in biology



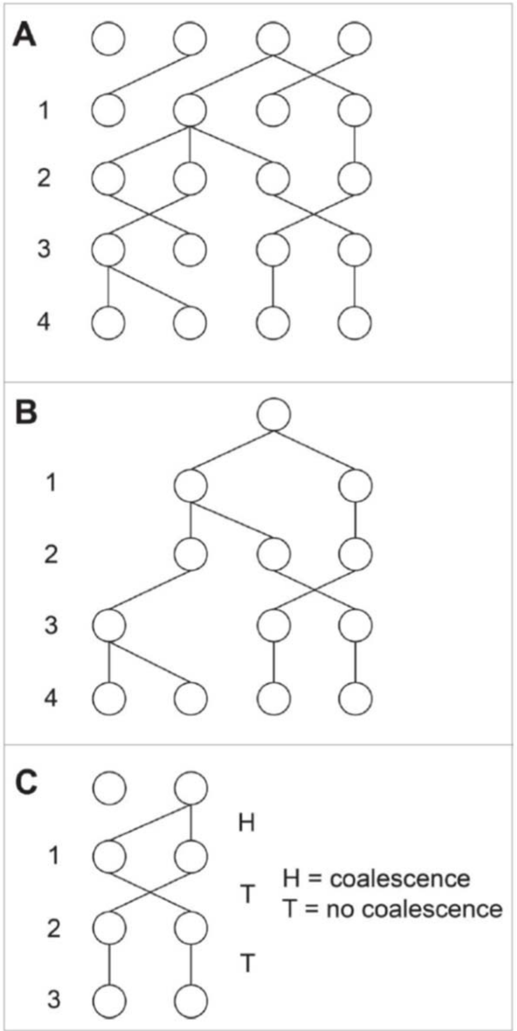




- Each of our cells has mitochondria with 16.5kb of mtDNA **inherited only from our mother**
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC’s Carl R. Woese fame)
- Since then most mitochondrial genes were transferred to 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.
- $P(T=t)=(1-1/N)^{t-1}(1/N) \approx (1/N)\exp(-T/N)$
- T can be inferred from the density of differences on mtDNA =2μT



Maddamsetti R, MOBILE GENETIC ELEMENTS, 6, e1137380(2016)

- There are about $N=3.5x10^{9}$ women living today
- For a random pair of women the average number of generations to the last common maternal ancestor is:

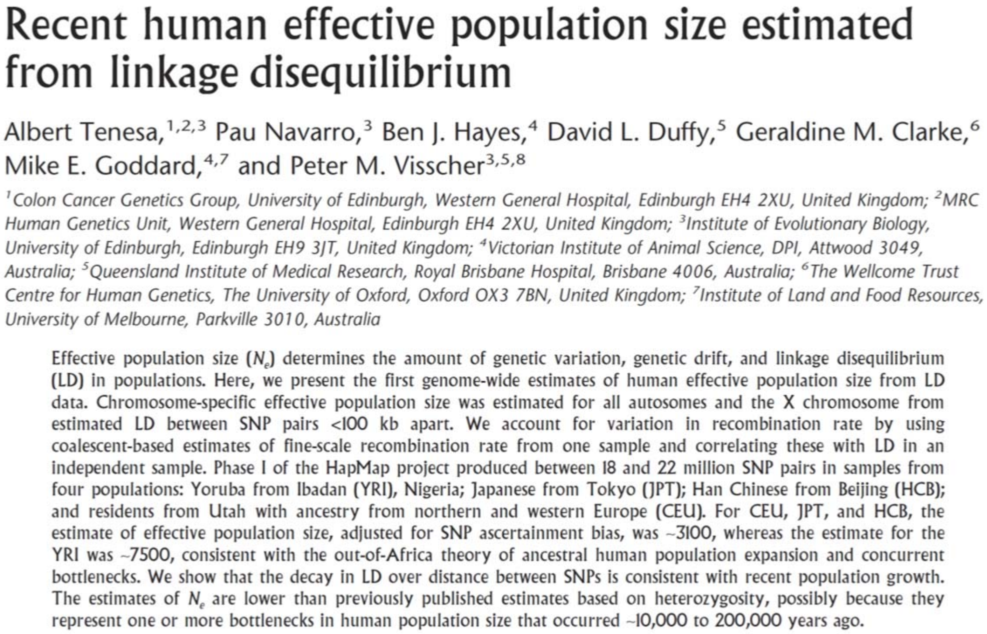
$$E(T)=\sum_{T=1}^{\infty} T \cdot exp(-T/N)=1/p=N$$

- Most Recent maternal Common Ancestor (MRCA) of all people living today lived $T_{MRCA}=2N$ generations ago
- $T_{MRCA} = 2 \cdot 3.5x10^{9}$ generations
- If the generation time 20 years it is 140 billion years > 10 times the time since the Big Bang.
- Something is wrong here!

- Population is not constant and for a long time was very low
- Change N to “effective” size N_e
- Current thinking is that for all of us including people of African ancestry N_e~7500 people
- For humans of European + Asian ancestry N_e~ 3100 people

- Mito Eve lived ~

2*Ne*20 years=
=2*7500*20 years=
300,000 years ago



Recent human effective population size estimated from linkage disequilibrium

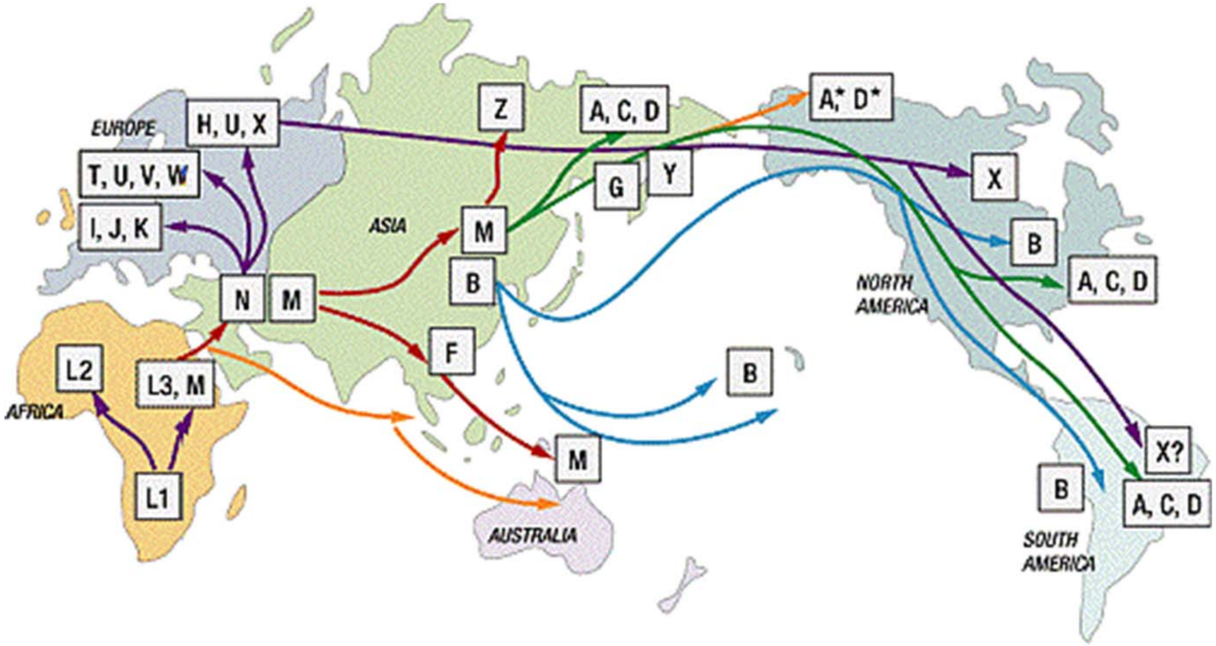
Albert Tenesa,[1,2,3] Pau Navarro,[3] Ben J. Hayes,[4] David L. Duffy,[5] Geraldine M. Clarke,[6] Mike E. Goddard,[4,7] and Peter M. Visscher[3,5,8]

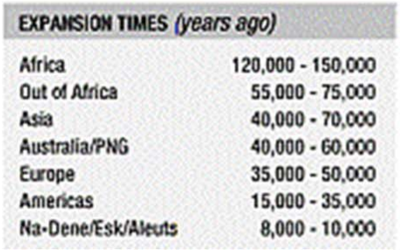
[1]Colon Cancer Genetics Group, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; [2]MRC Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; [3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom; [4]Victorian Institute of Animal Science, DPI, Attwood 3049, Australia; [5]Queensland Institute of Medical Research, Royal Brisbane Hospital, Brisbane 4006, Australia; [6]The Wellcome Trust Centre for Human Genetics, The University of Oxford, Oxford OX3 7BN, United Kingdom; [7]Institute of Land and Food Resources, University of Melbourne, Parkville 3010, Australia

Effective population size (N_e) determines the amount of genetic variation, genetic drift, and linkage disequilibrium (LD) in populations. Here, we present the first genome-wide estimates of human effective population size from LD data. Chromosome-specific effective population size was estimated for all autosomes and the X chromosome from estimated LD between SNP pairs <100 kb apart. We account for variation in recombination rate by using coalescent-based estimates of fine-scale recombination rate from one sample and correlating these with LD in an independent sample. Phase I of the HapMap project produced between 18 and 22 million SNP pairs in samples from four populations: Yoruba from Ibadan (YRI), Nigeria; Japanese from Tokyo (JPT); Han Chinese from Beijing (HCB); and residents from Utah with ancestry from northern and western Europe (CEU). For CEU, JPT, and HCB, the estimate of effective population size, adjusted for SNP ascertainment bias, was ~3100, whereas the estimate for the YRI was ~7500, consistent with the out-of-Africa theory of ancestral human population expansion and concurrent bottlenecks. We show that the decay in LD over distance between SNPs is consistent with recent population growth. The estimates of N_e are lower than previously published estimates based on heterozygosity, possibly because they represent one or more bottlenecks in human population size that occurred ~10,000 to 200,000 years ago.

Tenesa, Albert, et al. *Genome research* 17.4 (2007): 520-526.

“Mitochondrial Eve” lived in Africa



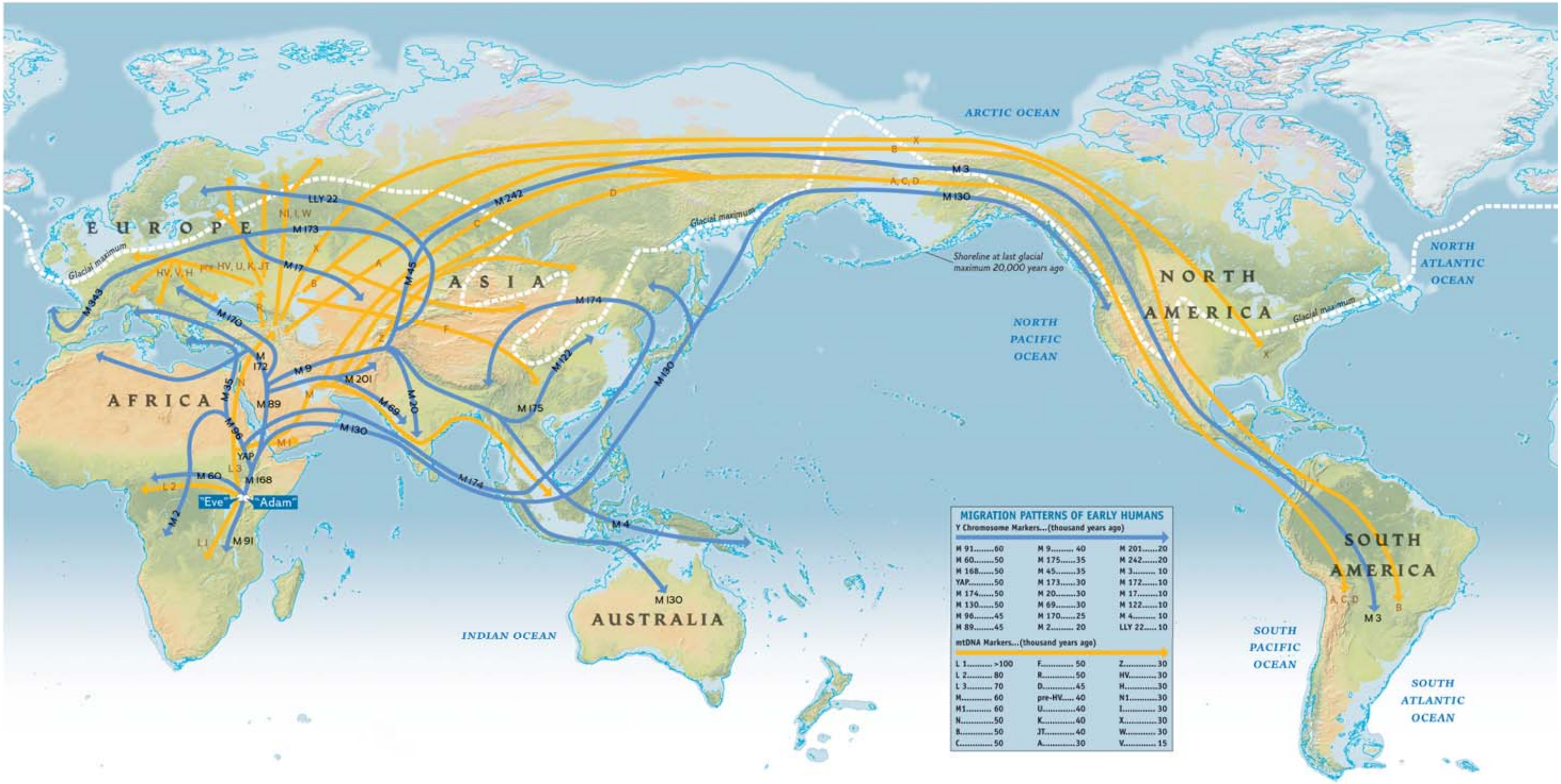


EXPANSION TIMES *(years ago)*	
Africa	120,000 - 150,000
Out of Africa	55,000 - 75,000
Asia	40,000 - 70,000
Australia/PNG	40,000 - 60,000
Europe	35,000 - 50,000
Americas	15,000 - 35,000
Na-Dene/Esk/Aleuts	8,000 - 10,000

FamilyTreeDNA
mtDNA Migrations Map

- “Mitochondrial Eve” lived in Africa between 100,000 and 150,000 years ago (or 240,000?)
- *Poznik GD, et al (Carlos Bustamante lab in Stanford), Science **341**: 562 (August 2013).*

“Adam” and “Eve” are both out of Africa



MIGRATION PATTERNS OF EARLY HUMANS

Y Chromosome Markers...(thousand years ago)

M 91	60	M 9	40	M 201	20
M 60	50	M 175	35	M 242	20
M 168	50	M 45	35	M 3	10
YAP	50	M 173	30	M 172	10
M 174	50	M 20	30	M 17	10
M 130	50	M 69	30	M 122	10
M 96	45	M 170	25	M 4	10
M 89	45	M 2	20	LLY 22	10

mtDNA Markers...(thousand years ago)

L 1	>100	F	50	Z	30
L 2	80	R	50	HV	30
L 3	70	D	45	H	30
M	60	pre-HV	40	N1	30
M1	60	U	40	I	30
N	50	K	40	X	30
B	50	JT	40	W	30
C	50	A	30	V	15

- “Mitochondrial Eve” lived in Africa between 100,000 and 150,000 years ago (or 240,000?)
- “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?

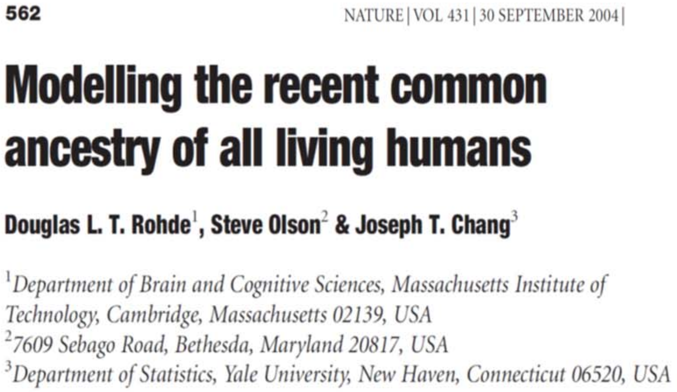
A. 1 million years ago
B. 200,000 years ago
C. 3400 years ago
D. 660 years ago
E. Yesterday, I really have no clue

Get your i-clickers

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

- Nuclear DNA gets mixed with every generation
  - Each of us gets 50% of nuclear DNA from father & 50% from mother
  - Each has 2 parents, 4 grandparents, 8 great-grand parents ….
- If one assumes:
  - Well-mixed marriages (not true: mostly local until recently)
  - Constant size population (not true: much smaller)
  - In 33 generations the number of ancestors:
    2^{33} =8 billion > 7 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 ~1300 AD

Corrected for mostly local marriages

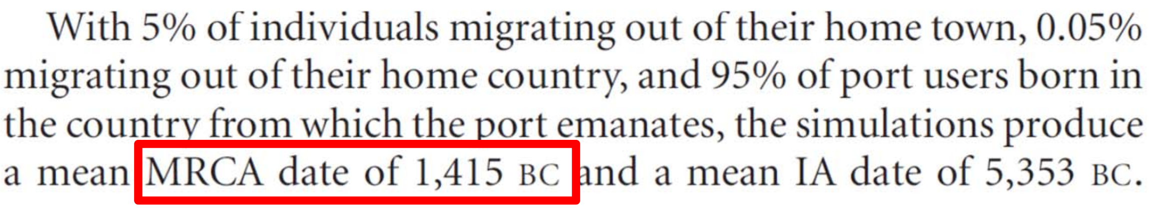


562 NATURE | VOL 431 | 30 SEPTEMBER 2004 |

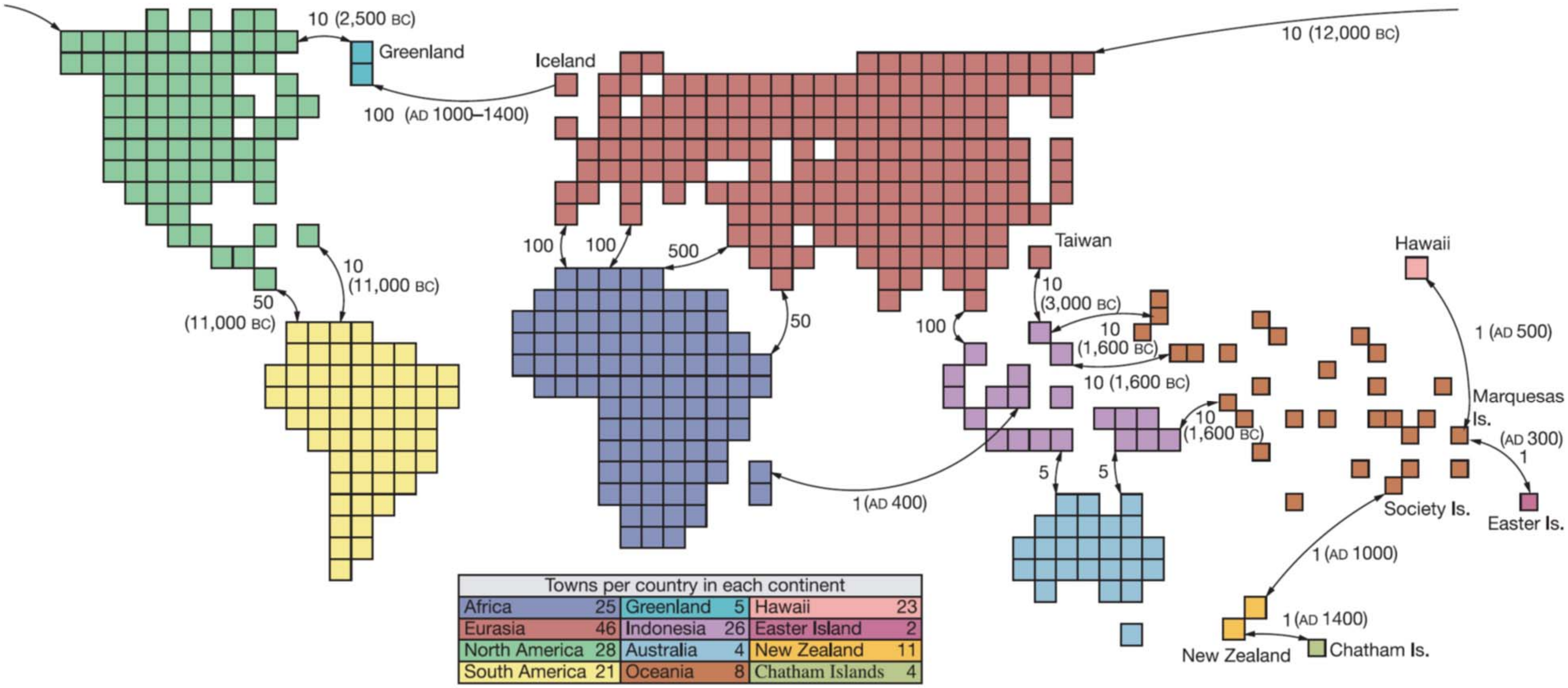
Modelling the recent common ancestry of all living humans

Douglas L. T. Rohde[1], Steve Olson[2] & Joseph T. Chang[3]

[1]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[2]*7609 Sebago Road, Bethesda, Maryland 20817, USA*
[3]*Department 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.



Towns per country in each continent					
Africa	25	Greenland	5	Hawaii	23
Eurasia	46	Indonesia	26	Easter Island	2
North America	28	Australia	4	New Zealand	11
South America	21	Oceania	8	Chatham Islands	4

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.

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?

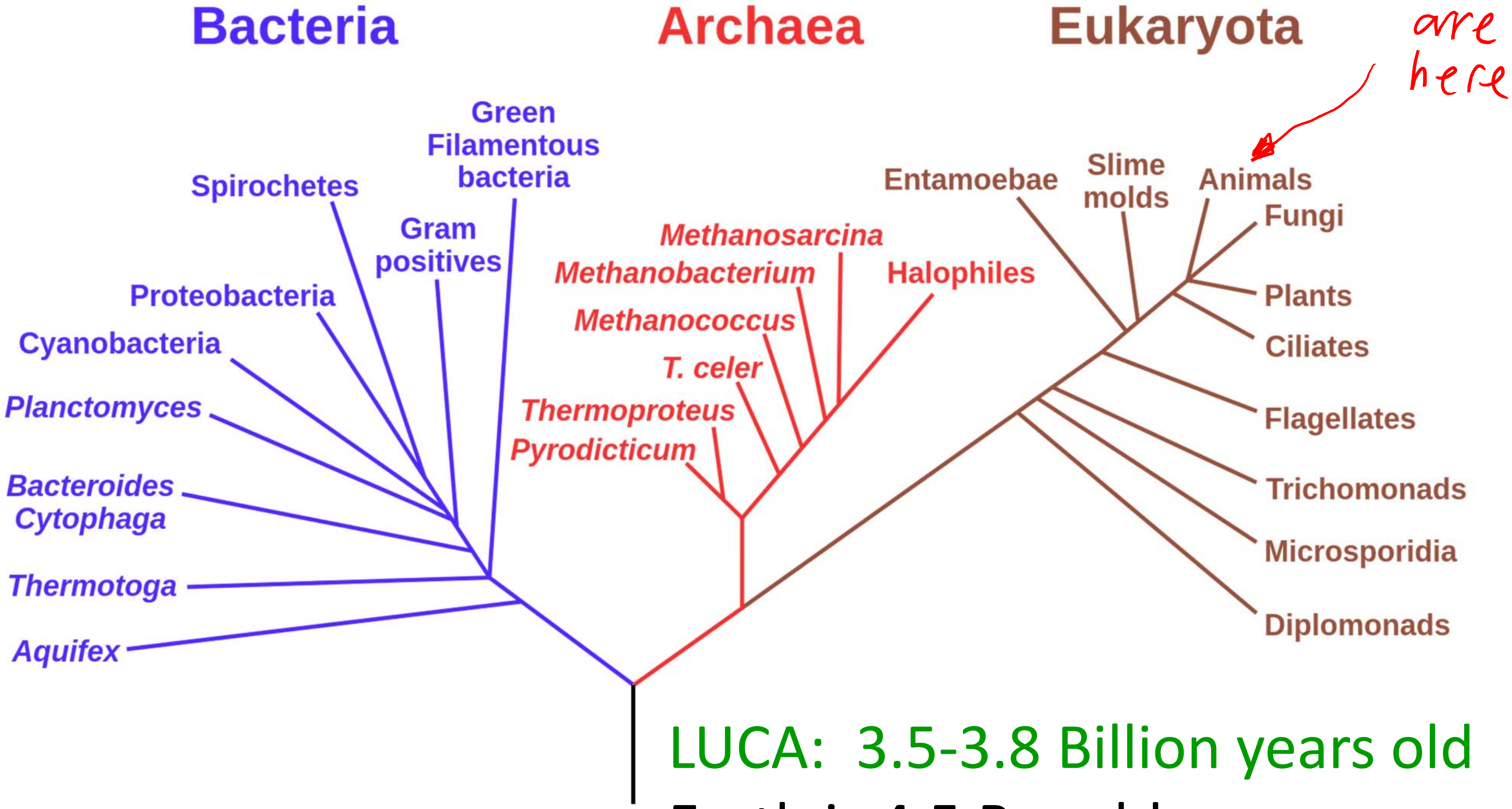
A. 1 million years ago
B. 200,000 years ago
[✓ C. 3400 years ago]
D. 660 years ago
E. Yesterday, I really have no clue

Get your i-clickers

Last Universal Common Ancestor (LUCA)

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

You are here



LUCA: 3.5-3.8 Billion years old

Earth is 4.5 Byr old

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:
  $$f\left(x\right)=C_{r-1}^{x-1}p^{r}\left(1-p\right)^{x-r}\quad \text{for } x=r,r+1,r+2\ldots \qquad (3\text{-}11)$$
- Compare it to binomial
  $$f\left(x\right)=C_{x}^{n}p^{x}\left(1-p\right)^{n-x}\quad \text{for } x=1,2,\ldots \text{ n}$$

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:

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

Cancer is scary!

- It hit my family twice last year
- Approximately 39.6 percent of men and women will be diagnosed with cancer at some point during their lifetimes (source: NCI website)

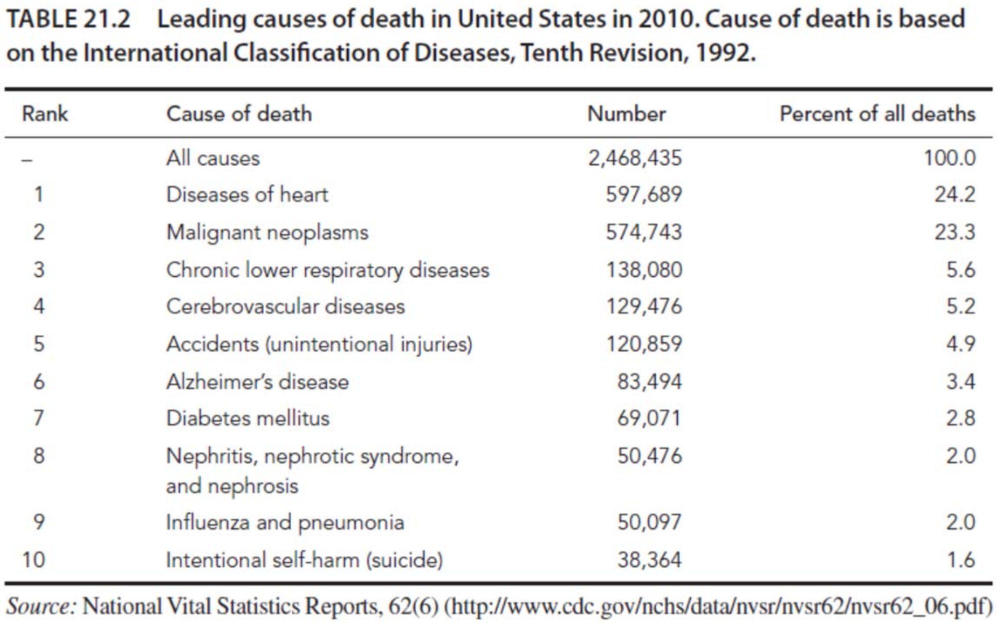


TABLE 21.2 Leading causes of death in United States in 2010. Cause of death is based on the International Classification of Diseases, Tenth Revision, 1992.

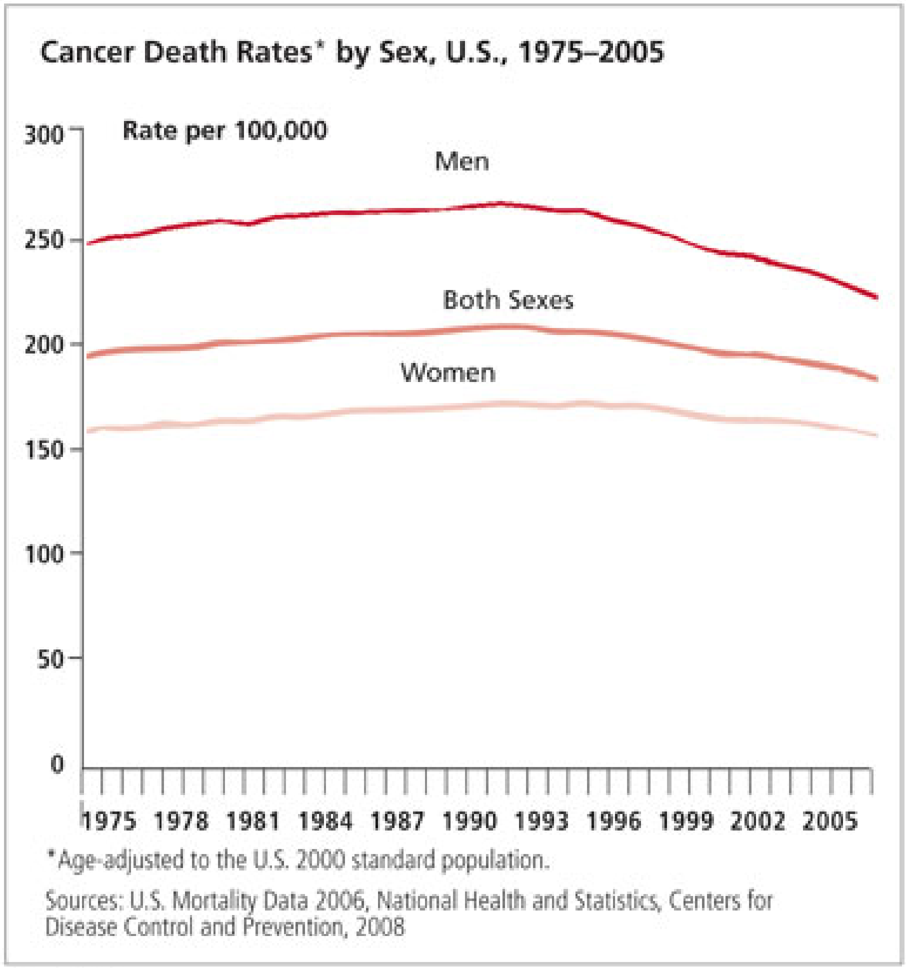
Rank	Cause of death	Number	Percent of all deaths
–	All causes	2,468,435	100.0
1	Diseases of heart	597,689	24.2
2	Malignant neoplasms	574,743	23.3
3	Chronic lower respiratory diseases	138,080	5.6
4	Cerebrovascular diseases	129,476	5.2
5	Accidents (unintentional injuries)	120,859	4.9
6	Alzheimer's disease	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

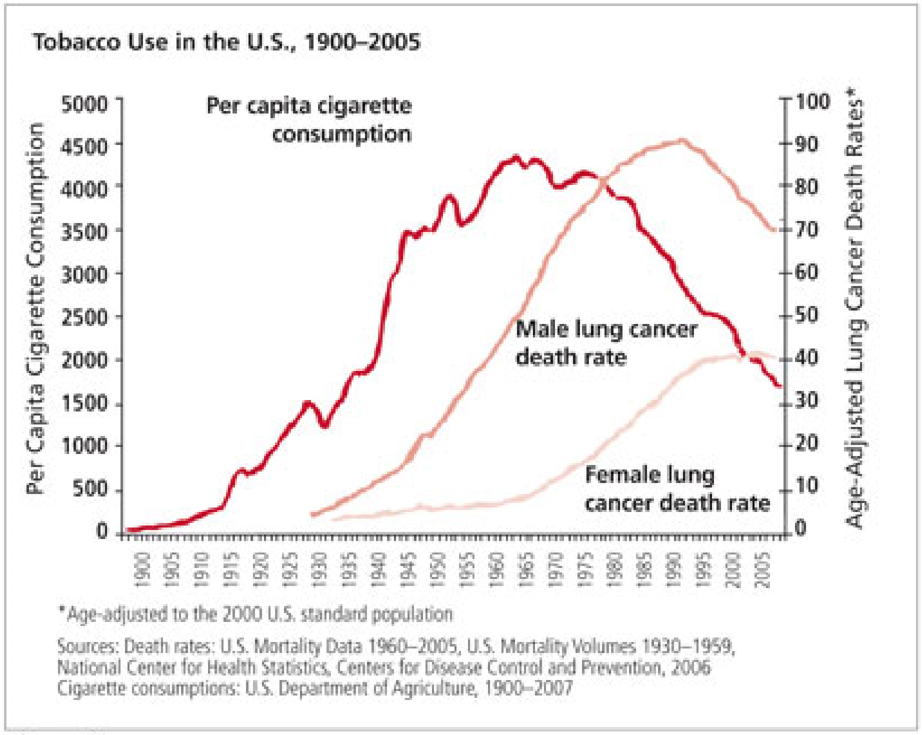
“War on Cancer” progress report


Cancer Death Rates* by Sex, U.S., 1975–2005

*Age-adjusted to the U.S. 2000 standard population.

Sources: U.S. Mortality Data 2006, National Health and Statistics, Centers for Disease Control and Prevention, 2008

Figure 2


Tobacco Use in the U.S., 1900–2005

*Age-adjusted to the 2000 U.S. standard population

Sources: Death rates: U.S. Mortality Data 1960–2005, U.S. Mortality Volumes 1930–1959, National Center for Health Statistics, Centers for Disease Control and Prevention, 2006
Cigarette consumptions: U.S. Department of Agriculture, 1900–2007

Figure 3

Statistics of Cancer

- Bert Vogelstein et al: Cancer is caused by accumulation of “driver” gene mutations
  - Oncogenes: ↑
  - Tumor suppressors: ↓ (may need 2 mutations)
  - 7 strikes and you are out

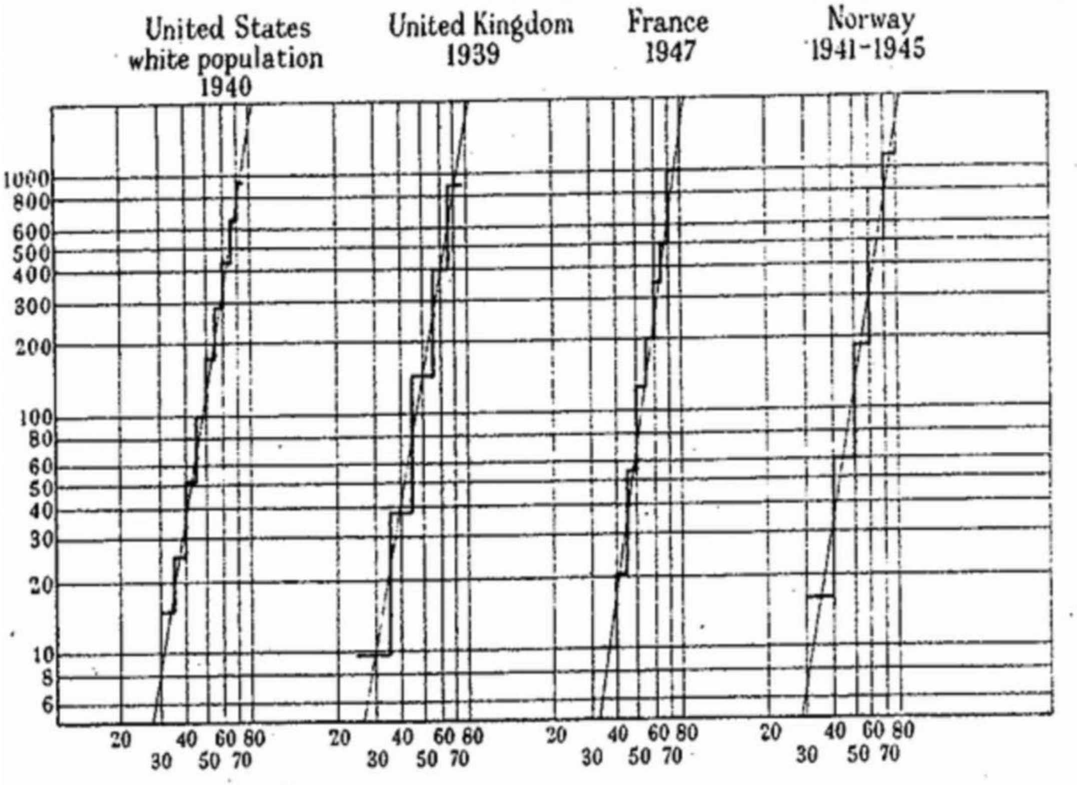
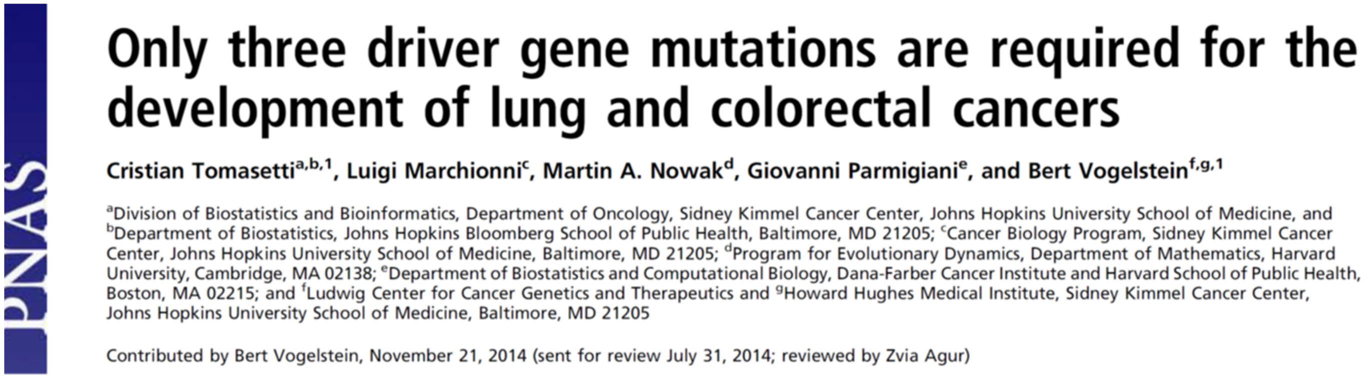


Fig. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

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

Cancer death rate ~ (patient age)^6

Ongoing discussion: how many strikes?



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

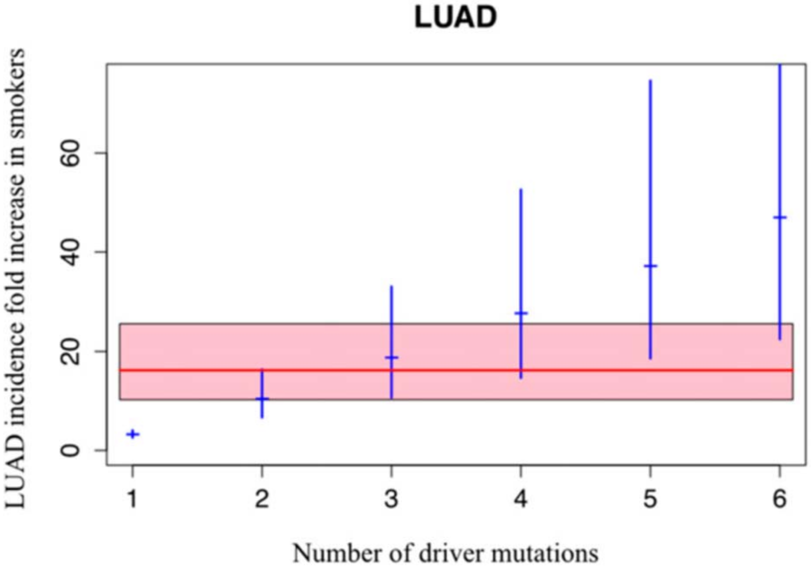
Cristian Tomasetti[a,b,1], Luigi Marchionni[c], Martin A. Nowak[d], Giovanni Parmigiani[e], and Bert Vogelstein[f,g,1]

[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

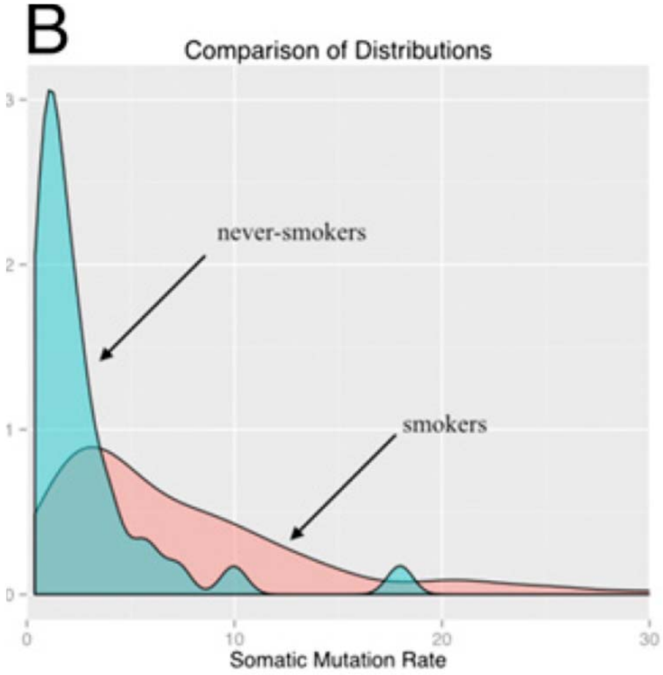
Contributed by Bert Vogelstein, November 21, 2014 (sent for review July 31, 2014; reviewed by Zvia Agur)

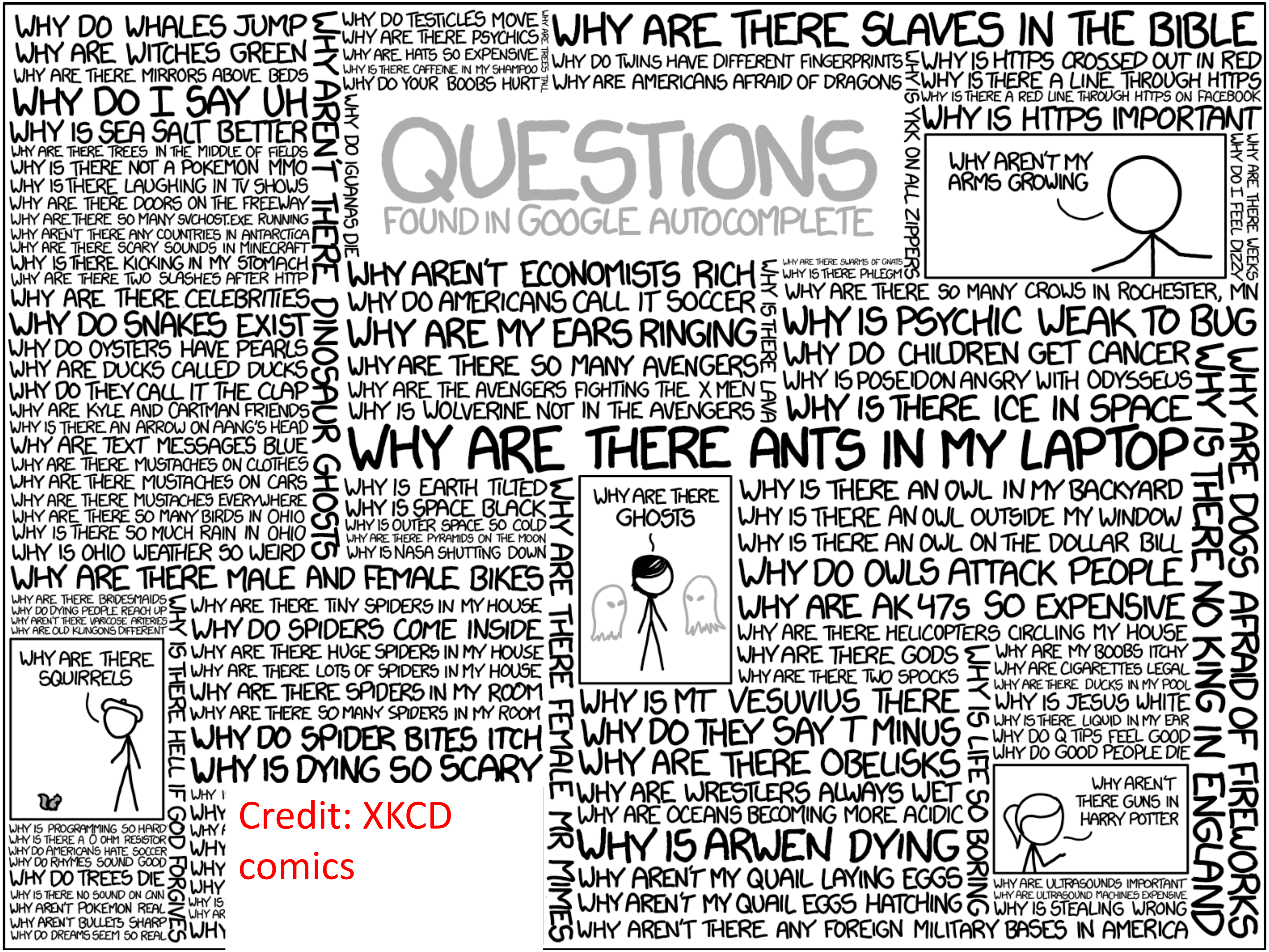
$$P(T_{cancer} \leq t) \sim (u_1 t)(u_2 t)..(u_n t) = \sim u_1 u_2 .. u_n t^n$$

$$P(T_{cancer} = t) \sim (u_1 t)(u_2 t)..(u_n t) = \sim u_1 u_2 .. u_n t^{n-1}$$

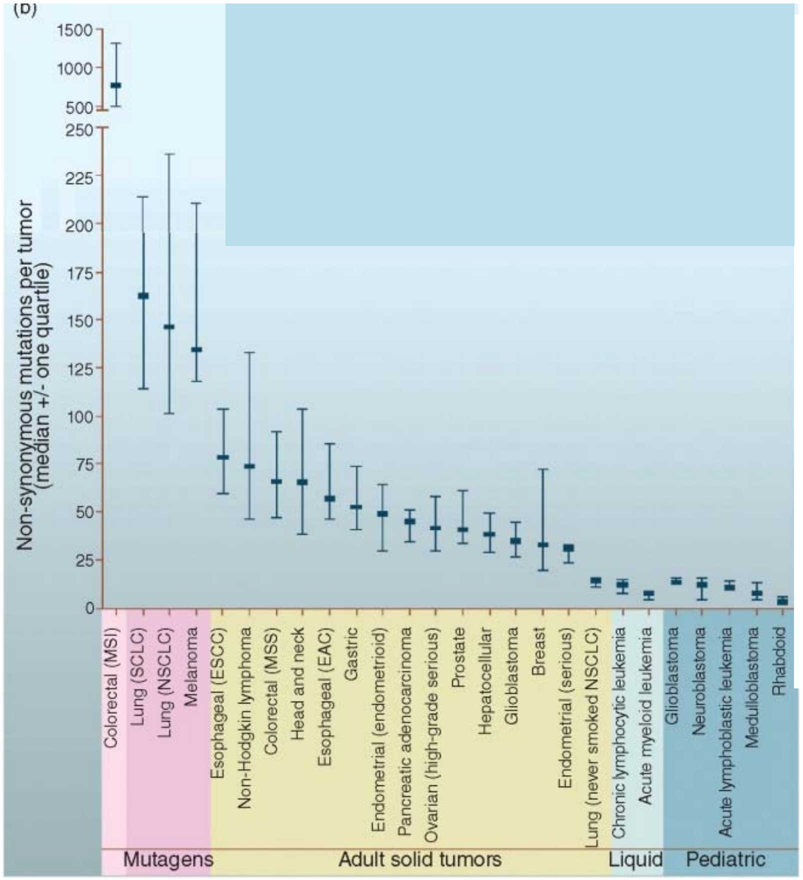


Smokers have 3.23 times more mutations




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 AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY DO TESTICLES MOVE
WHY ARE THERE PSYCHICS
WHY ARE HATS SO EXPENSIVE
WHY IS THERE CAFFEINE IN MY SHAMPOO
WHY DO YOUR BOOBS HURT
WHY ARE TREES TALL
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 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 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 IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
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 FEMALE MR MIMES
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 LIFE SO BORING
WHY ARE MY BOOBS ITCHY
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 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 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 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 I
WHY
WHY A
WHY
WHY
WHY
WHY IS
WHY AR
WHY
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 AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Credit: XKCD comics



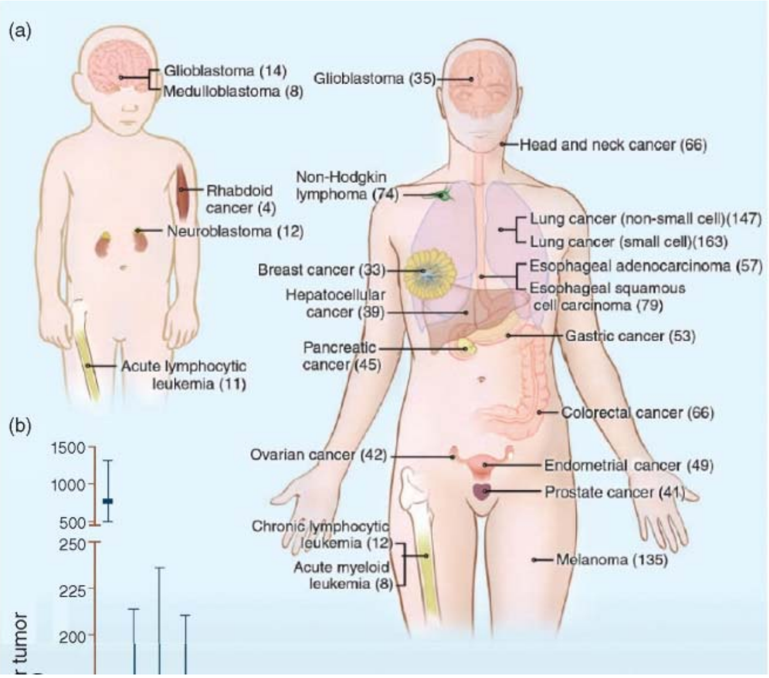

(a)
Glioblastoma (14)
Medulloblastoma (8)
Rhabdoid cancer (4)
Neuroblastoma (12)
Acute lymphocytic leukemia (11)
Glioblastoma (35)
Head and neck cancer (66)
Non-Hodgkin lymphoma (74)
Lung cancer (non-small cell)(147)
Lung cancer (small cell)(163)
Breast cancer (33)
Esophageal adenocarcinoma (57)
Esophageal squamous cell carcinoma (79)
Hepatocellular cancer (39)
Gastric cancer (53)
Pancreatic cancer (45)
Colorectal cancer (66)
Ovarian cancer (42)
Endometrial cancer (49)
Prostate cancer (41)
Chronic lymphocytic leukemia (12)
Acute myeloid leukemia (8)
Melanoma (135)
(b)
1500
1000
500
250
225
200
r tumor

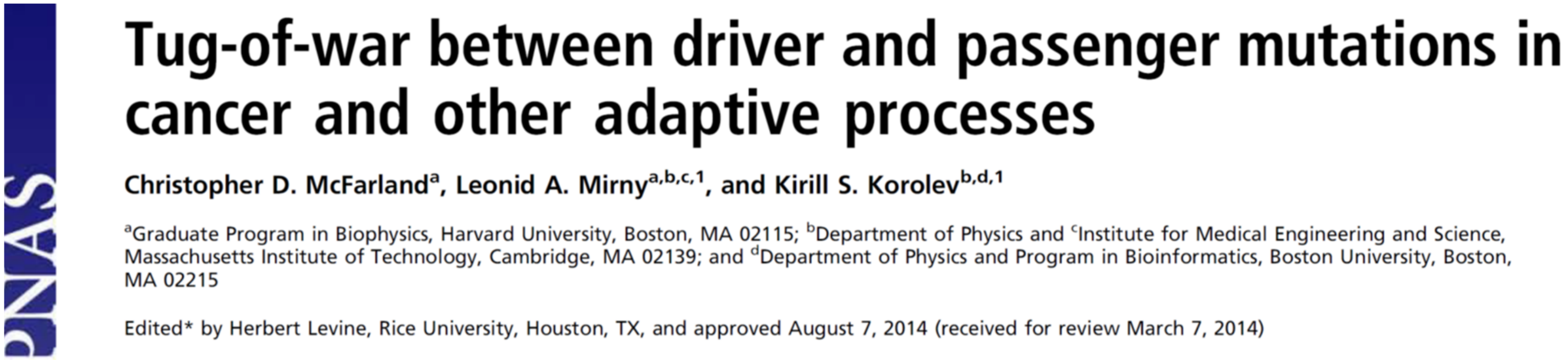
FIGURE 21.10 Somatic mutations in representative human cancers, based on genome-wide sequencing studies. (a) The genomes of adult (right) and pediatric (left) cancers are represented. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein *et al*. (2013).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

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.

Companion Website: www.wiley.com/go/pevsnerbioinformatics

- "Drivers" carry "Passengers"
- "Passenger" mutations cause little to no harm
- "Passengers" are common as cancers elevate mutation rate

Passenger mutations: negative binomial distribution



Tug-of-war between driver and passenger mutations in cancer and other adaptive processes

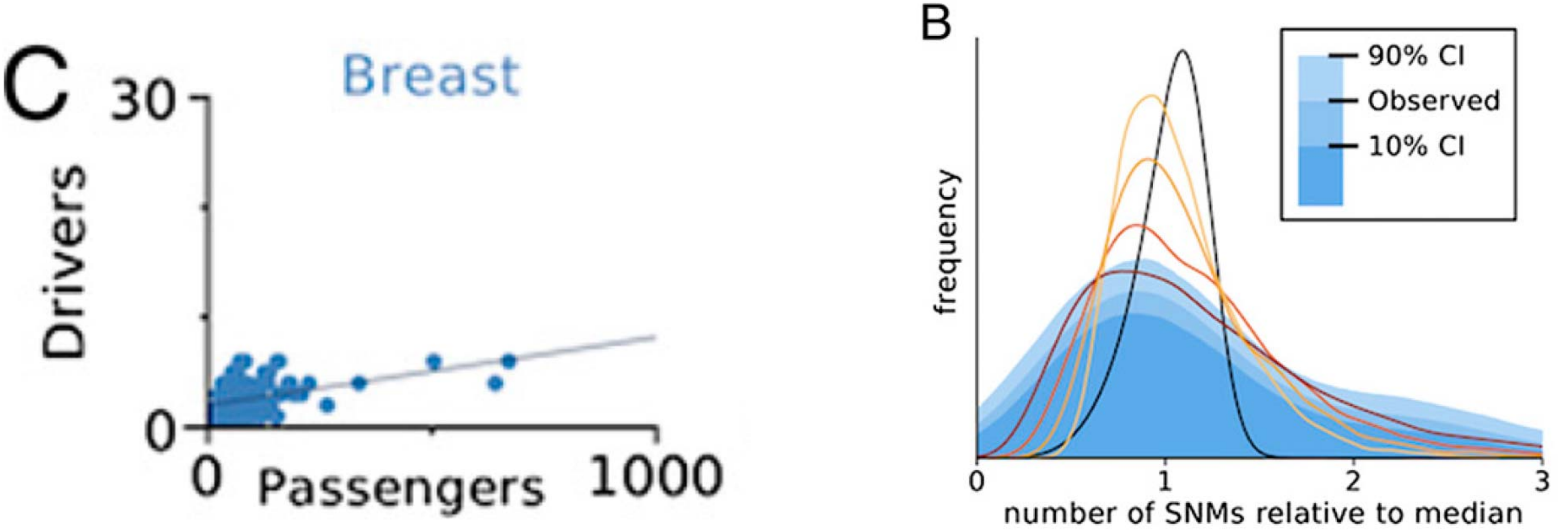
Christopher D. McFarland[a], Leonid A. Mirny[a,b,c,1], and Kirill S. Korolev[b,d,1]

[a]Graduate Program in Biophysics, Harvard University, Boston, MA 02115; [b]Department of Physics and [c]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139; and [d]Department of Physics and Program in Bioinformatics, Boston University, Boston, MA 02215

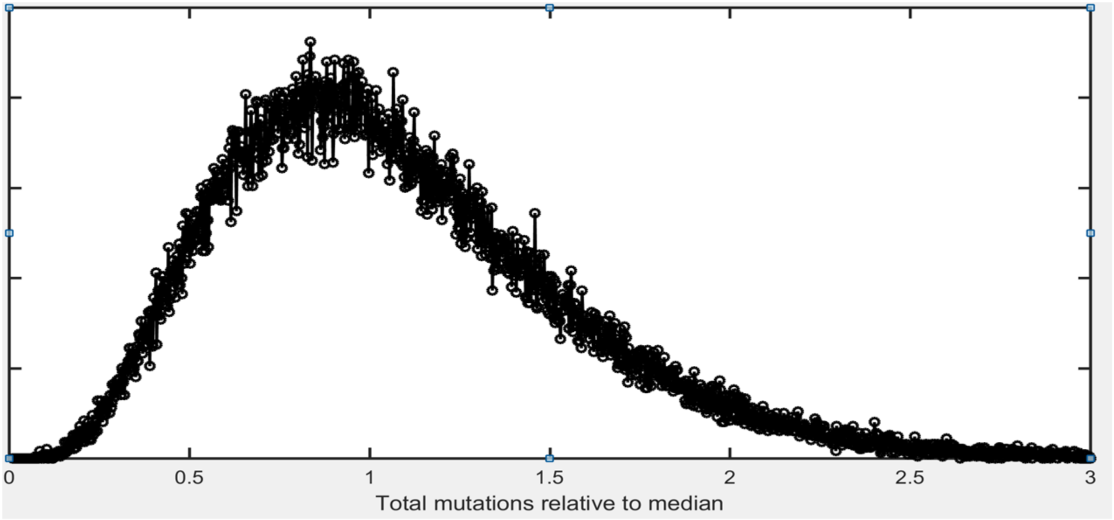
Edited* by Herbert Levine, Rice University, Houston, TX, and approved August 7, 2014 (received for review March 7, 2014)

- 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 (1-p) – it is a passenger

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



McFarland CD, Mirny L, Korolev KS, PNAS 2014



Matlab exercise

- Find mean, variance, and histogram of 100,000 geometrically-distributed numbers with p=0.1
- Hint: Use help page for **random** command on how to generate geometrically-distributed random numbers

Matlab: Geometric distributions

```
Stats=100000;
p=0.1;
r2=random('Geometric',p,Stats,1);
r2=r2+1;
disp(mean(r2));
disp(var(r2));
disp(std(r2));
[a,b]=hist(r2, 1:max(r2));
p_g=a./sum(a);
figure; semilogy(b,p_g,'ko-');
```

Matlab: Negative binomial distributions

```
Stats=100000;
r=3; p=0.1;
r2=random('Negative Binomial',r,p,Stats,1);
r2=r2+r;
disp(mean(r2));
disp(var(r2));
disp(std(r2));
[a,b]=hist(r2, 1:max(r2));
p_nb=a./sum(a);
figure; semilogy(b,p_nb,'ko-');
```