

Assume
that x —
the # of
daughters per each
mother follows
a Poisson
distribution

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Population does not grow $\rightarrow \lambda=1$

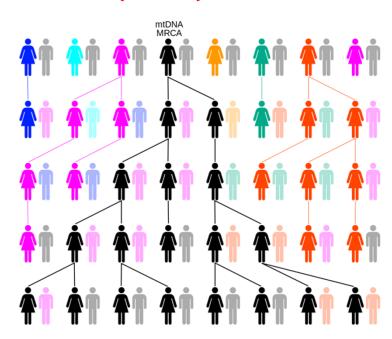
Prob(merge)=
=E[x(x-1)]/N=
=
$$\lambda^2/N=1/N$$

 $P(T=t)=(1-1/N)^{t-1}(1/N) \approx (1/N) \exp(-(t-1)/N)$

Most Recent Common Ancestor (MRCA)

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

merger is
$$\frac{2}{(N-1)(N-2)}$$



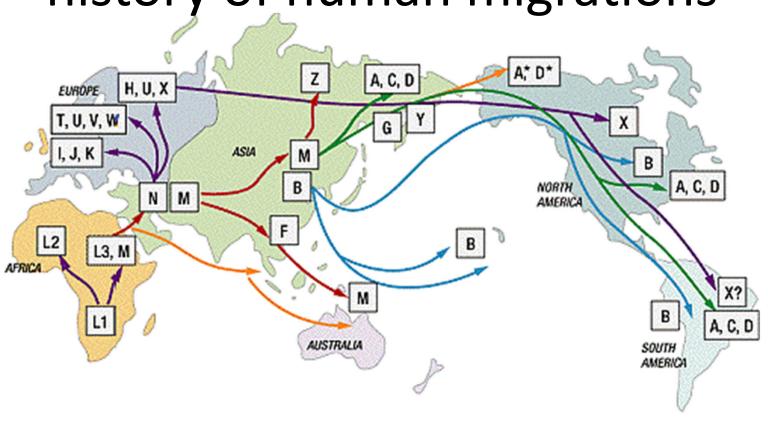
Most Recent Common Ancestor (MRCA)

Total time until the MRCA

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

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

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		



Y-chromosomal Adam also lived in Africa

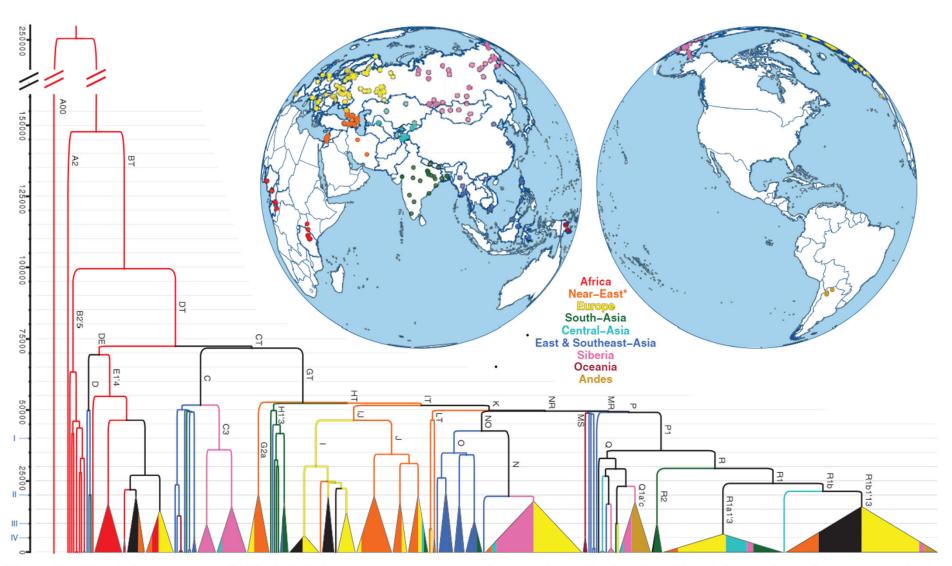
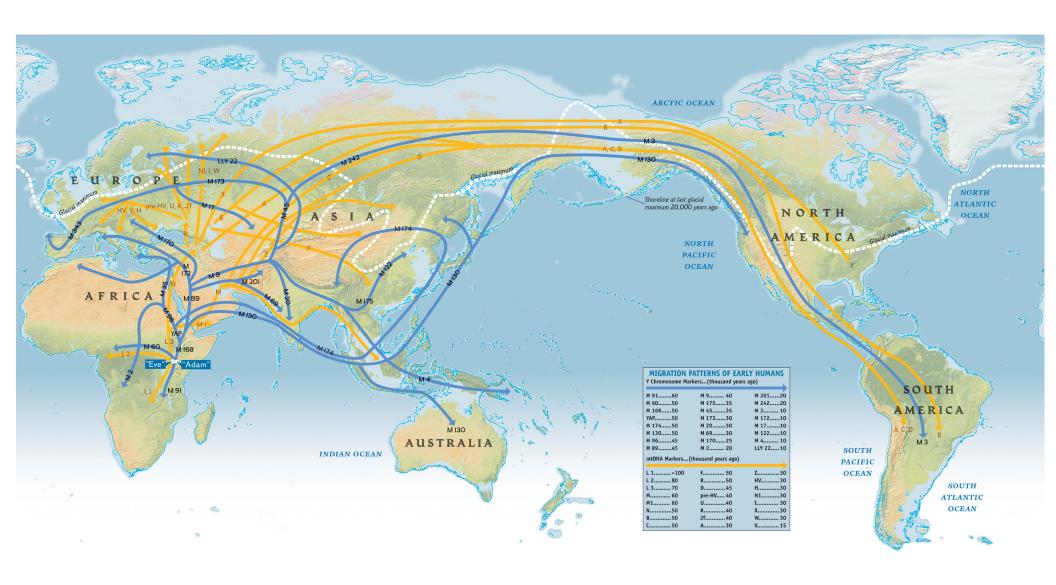


Figure 1. The phylogenetic tree of 456 whole Y chromosome sequences and a map of sampling locations. The phylogenetic tree is reconstructed using BEAST. Clades coalescing within 10% of the overall depth of the tree have been collapsed. Only main haplogroup labels are shown (details are provided in Supplemental 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. 660 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. 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

- Unlike Mito or Y-chromosome, nuclear DNA gets mixed with every generation
 - Each of us gets 50% of nuclear DNA from the father & 50% 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 until recently)
 - Constant size population (not true: much smaller in the past)
 - In 33 generations the number of ancestors:
 2³³ =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 NATURE | VOL 431 | 30 SEPTEMBER 2004 | and rare migrations

562

Modelling the recent common ancestry of all living humans

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

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.

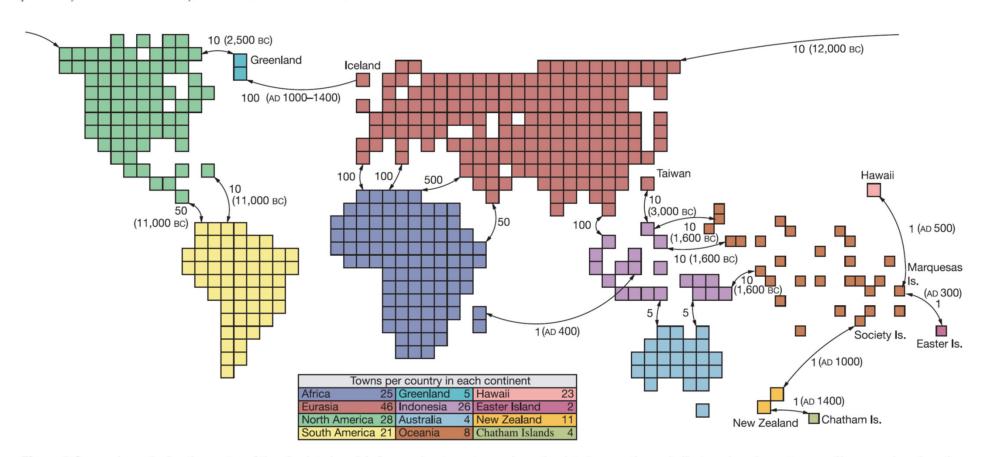


Figure 2 Geography and migration routes of the simulated model. Arrows denote ports and the adjacent numbers are their steady migration rates, in individuals per generation. If

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.

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

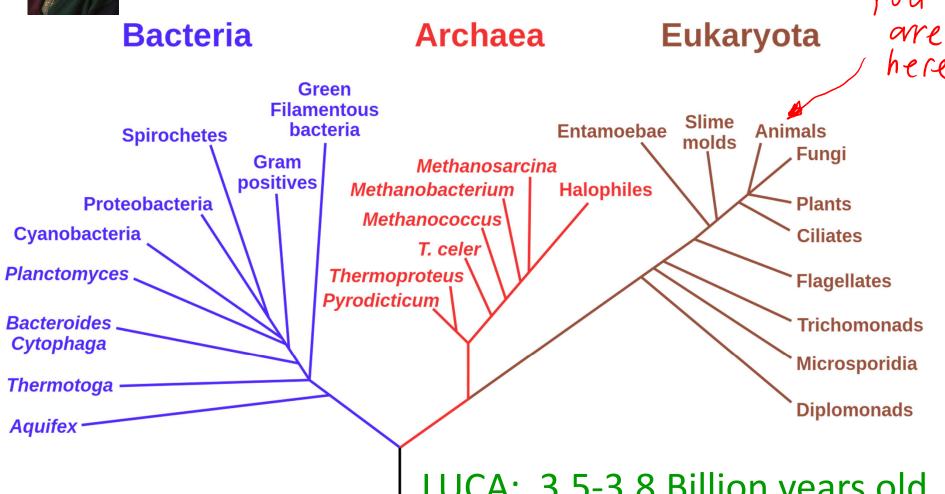
²7609 Sebago Road, Bethesda, Maryland 20817, USA

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

Last Universal Common Ancestor (LUCA)



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



LUCA: 3.5-3.8 Billion years old Earth is 4.5 Billion years 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:

$$0 and $r = 1, 2, 3,$$$

The probability mass function is:

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

Compare it to binomial

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

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,

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

Compare to geometric distribution:

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

Matlab exercise

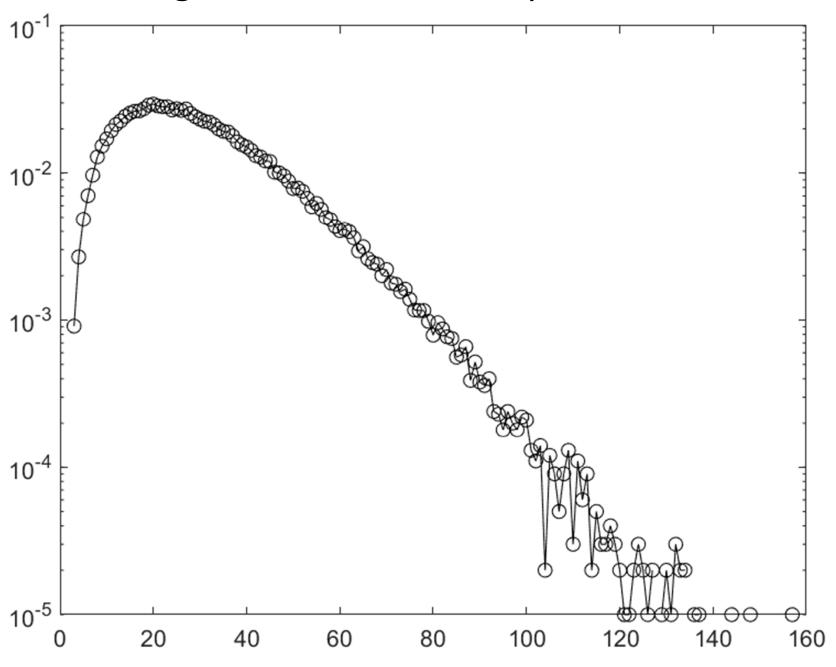
 Estimate mean, variance, and PMF based on 100,000 random variables drawn from a negative binomial distribution with p=0.1, r=3

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

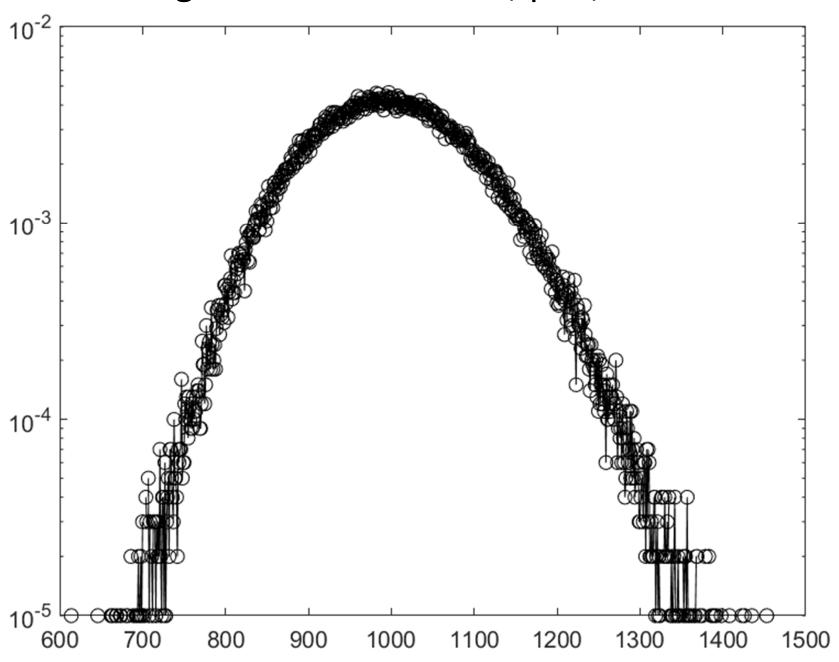
Matlab: Negative binomial distribution

```
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

Table from J. Pevsner 3rd edition

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

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

"War on Cancer" progress report

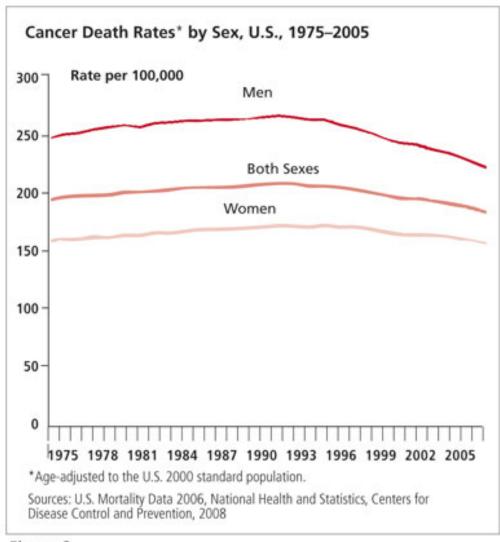


Figure 2

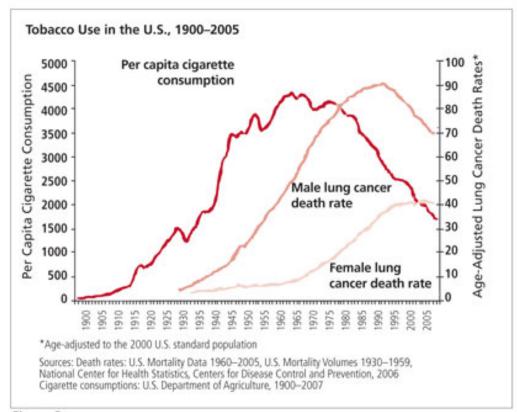


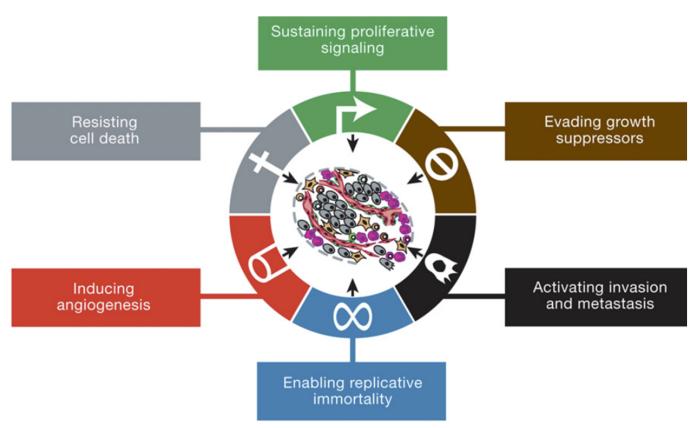
Figure 3

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 <u>Hallmarks of Cancer</u>: The Next Generation Cell 144, 2011



Statistics of cancer incidence vs age

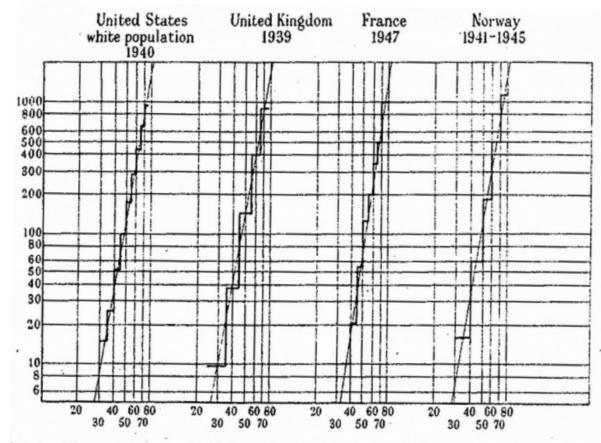


Fig. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

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

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

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

How many driver gene mutations for different types of cancer?

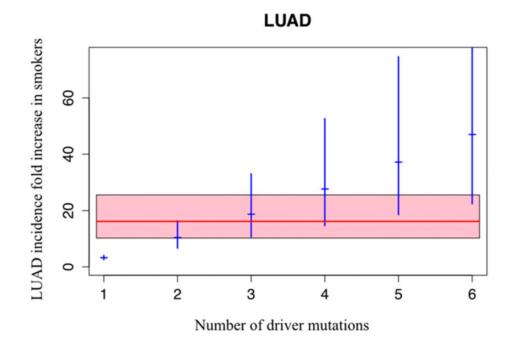
Only three driver gene mutations are required for the development of lung and colorectal cancers

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

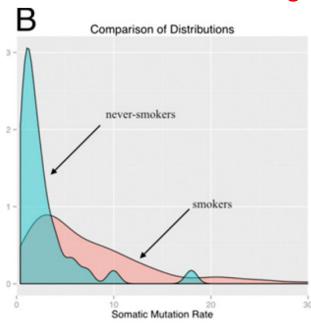
^aDivision of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, and ^bDepartment of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; ^cCancer Biology Program, Sidney Kimmel Cancer Center, 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



SANG



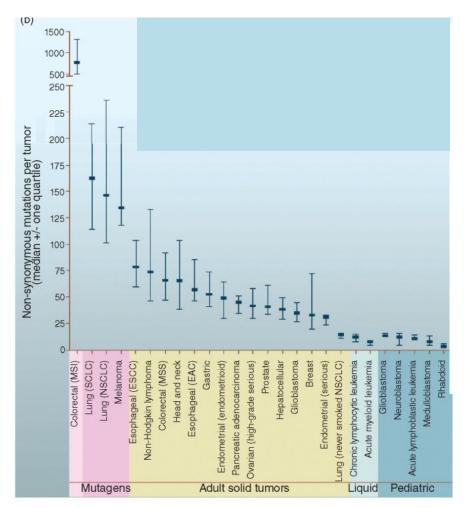
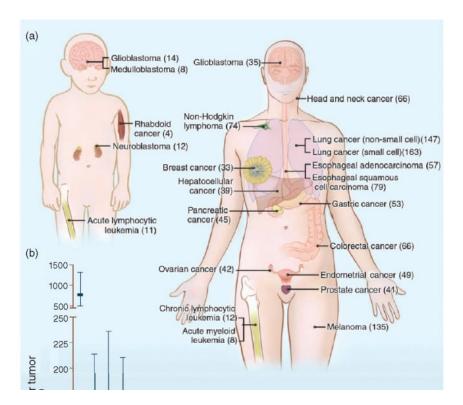


FIGURE 21.10 Somatic mutations in representative human cancers, based on genome-wide sequencing studies. (a) The genomes of adult (right) and pediatric (left) cancers are represented. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein et al. (2013). 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. © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd. Companion Website: www.wiley.com/go/pevsnerbioinformatics



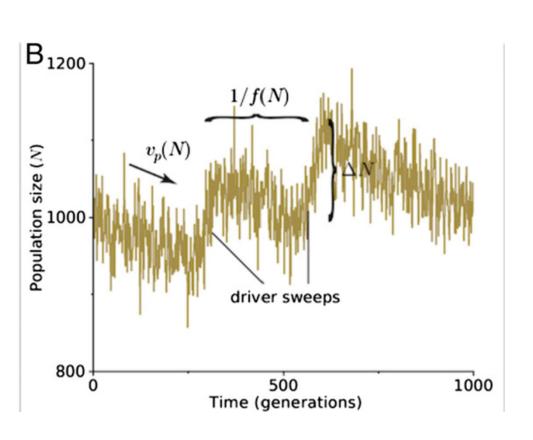
- Cancer cells carry both "Driver" and "Passengers" mutations
- Passenger mutations cause little to no harm (see later for how even little harm matters)
- Both are common as cancers elevate mutation rate

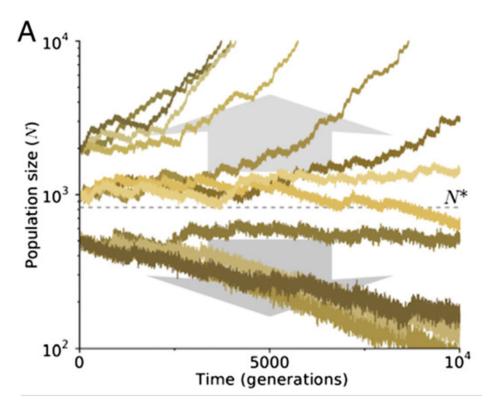
Number of passenger+driver mutations follows negative binomial distribution

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

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

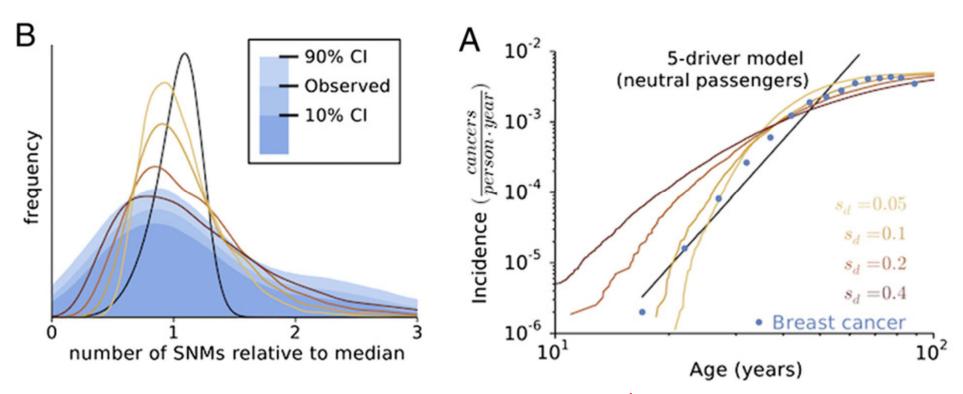
What if passenger mutations slow down the growth of cancer tumors?





McFarland CD, Mirny L, Korolev KS, PNAS 2014

Can we prove/quantify it using statistics?



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

 μ =10⁻⁸, Target_d=1,400, Target_p=10⁷, 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

