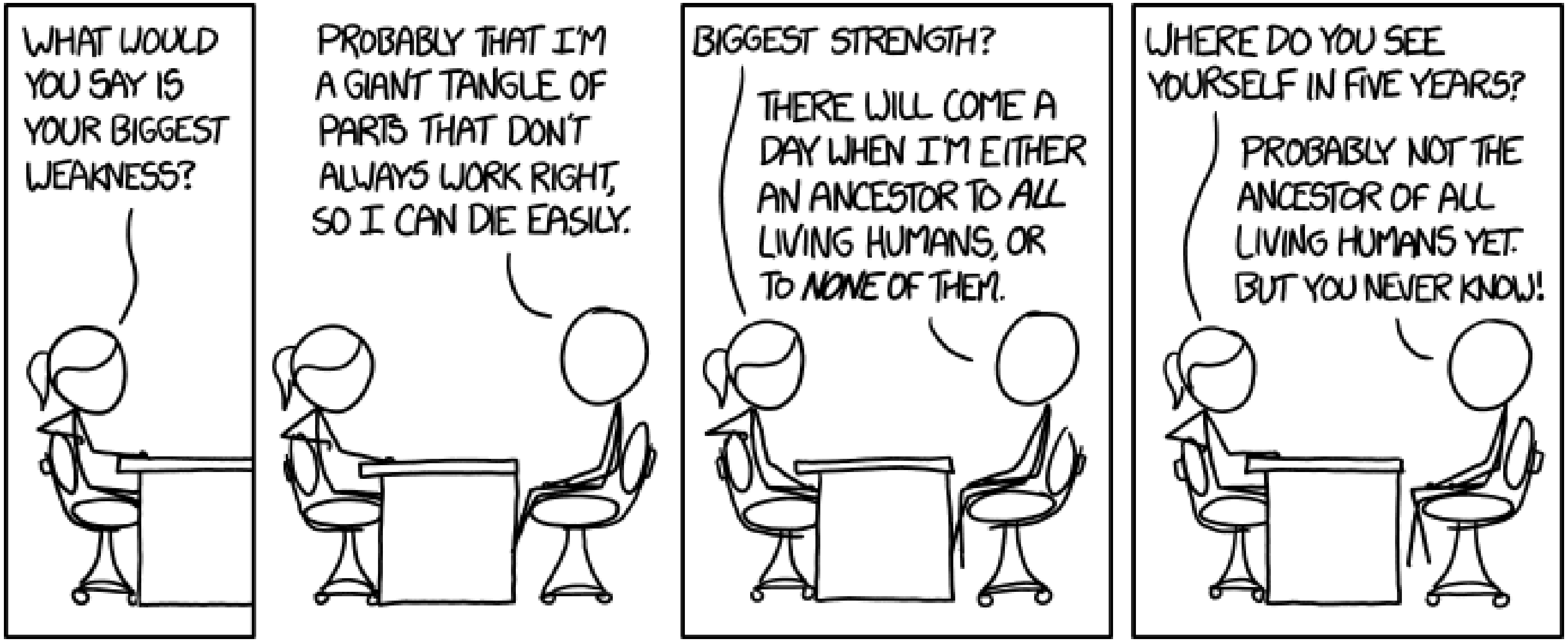
MIMM4750G
The coalescent

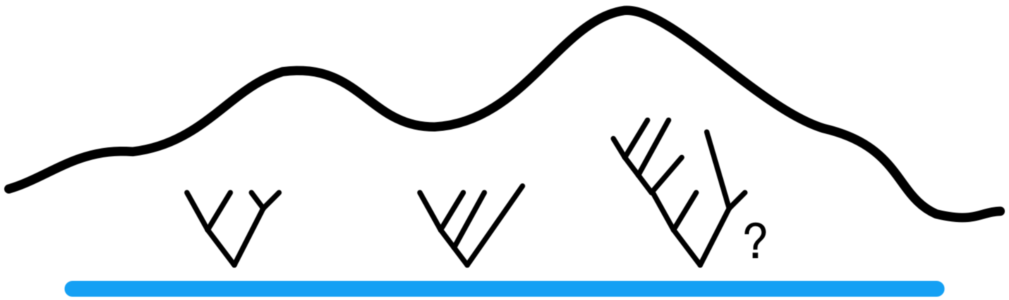

WHAT WOULD YOU SAY IS YOUR BIGGEST WEAKNESS?
PROBABLY THAT I'M A GIANT TANGLE OF PARTS THAT DON'T ALWAYS WORK RIGHT, SO I CAN DIE EASILY.
BIGGEST STRENGTH?
THERE WILL COME A DAY WHEN I'M EITHER AN ANCESTOR TO *ALL* LIVING HUMANS, OR TO *NONE* OF THEM.
WHERE DO YOU SEE YOURSELF IN FIVE YEARS?
PROBABLY NOT THE ANCESTOR OF ALL LIVING HUMANS YET. BUT YOU NEVER KNOW!

Bayesian inference

- We spent the previous lecture and lab talking about and learning work with Bayesian methods.
- Bayesian inference is extremely useful for dealing with complex models like *trees*.
- A tree is a model that is made up of branching patterns and branch lengths. This is a lot of parameters!
- The MLE of the tree is only one point in a *massive* space of all possible trees.

Priors on trees

- We've learned about priors on continuous variables. There are also priors for discrete variables.
- What about trees?
- We'd need some way of arranging trees along a meaningful axis in order to draw a prior distribution the usual way.



Common ancestry

- A tree is a model of common ancestry.
- How far back in time do we have to go to reach the common ancestor of two people we pick at random?
- Assumptions:

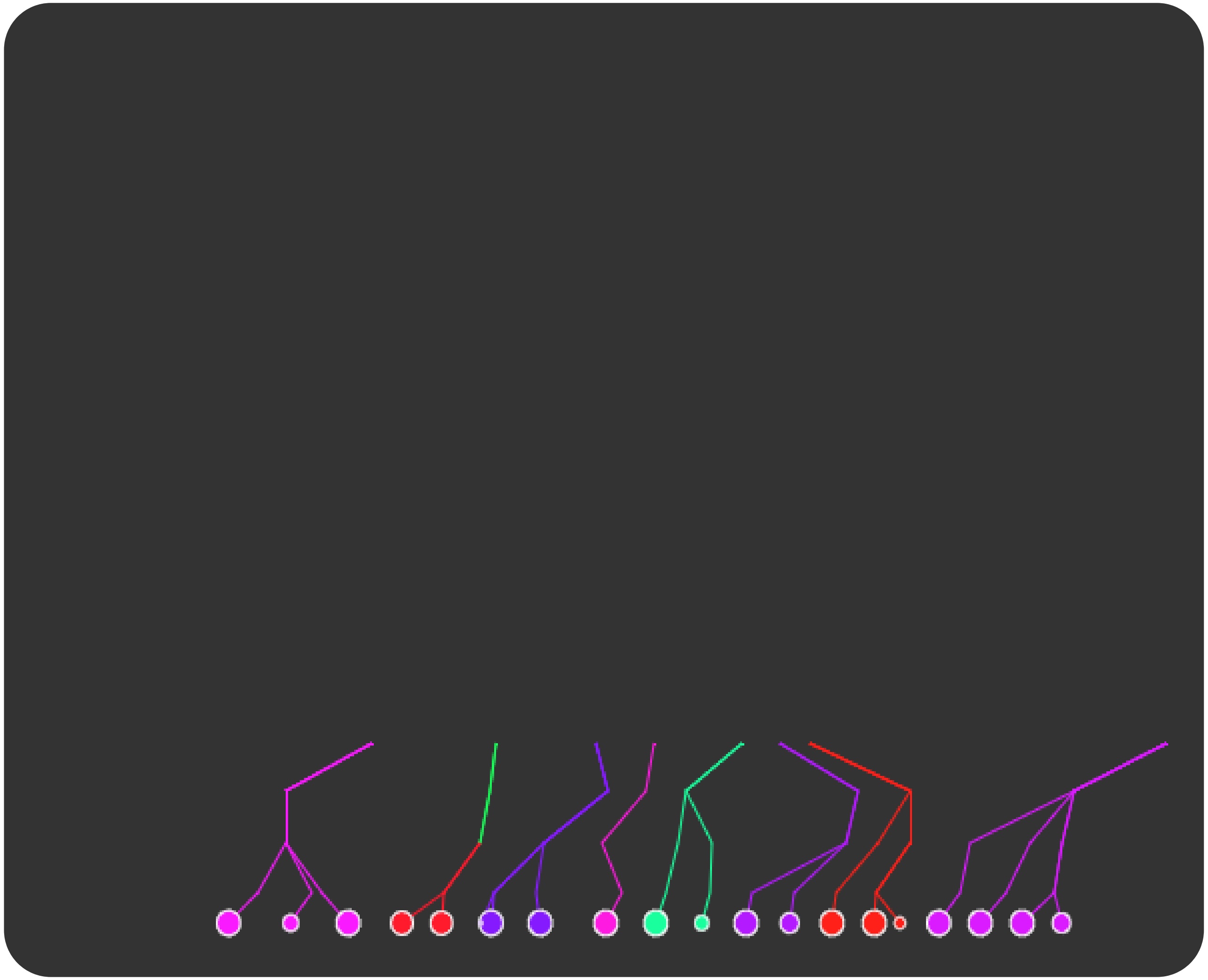
  1. Every individual has the same chance of contributing offspring (no selection).

  2. The size of the population does not change.

  3. Everyone reproduces at the same time and dies (discrete generations).

  4. The size of the population is large.

The coalescent



coaltrace by Trevor Bedford

The coalescent

- To make things simple, assume everyone has one and only one ancestor (pathogens!).
- The probability that two *randomly sampled* individuals have a common ancestor in the previous generation is $1/N$, where N is the population size.
- Every individual is equally likely to be the ancestor.

Geometric distribution

- The expected number of generations we go back is described by a geometric distribution:
$$P(t) = (1-p)^{t-1}p$$
- $t-1$ "failures" until a success.
- $p = 1/N$, gives us:
$$P(t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}$$

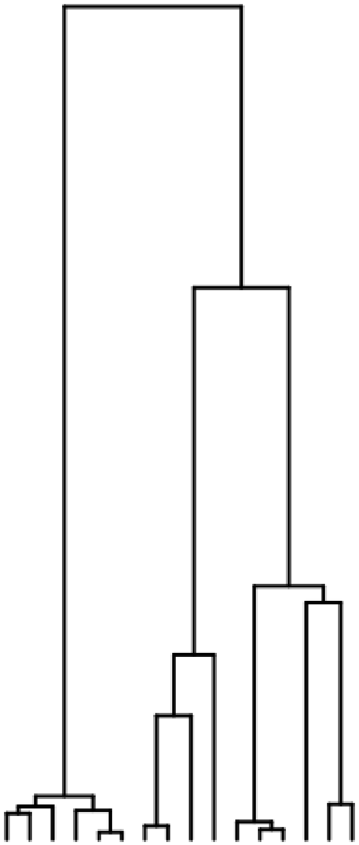
Q1. The Leafs' penalty kill percentage is currently 80% (19th in the league). What is the probability that they kill four penalties in a row before giving up a PP goal?

Mean time to ancestor

- The average for the geometric distribution is simply $1/p$.
- Since $p = 1/N$, we expect to go N generations until we reach the common ancestor of two random individuals.
- *We can learn something about the entire population from sampling a much smaller number of individuals.*

Sampling more than two

- If we sample n individuals at random, then there are
$$\binom{n}{2} = \frac{n!}{(n-2)!2!} = \frac{n(n-1)}{2} \text{ pairs.}$$
- We multiply the p by this amount (go back in time until *any* pair coalesces).
- After each coalescence, we subtract 1 from n. The total coalescence rate gets slower!



The MRCA

- Most recent common ancestor.
- The "most recent" recognizes the idea that the ancestor of the MRCA is *also* a common ancestor of the entire sample.
- The *total* mean time to get to the MRCA of all *n* individuals:

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

  generations.

Effective population size

- The coalescent model required a lot of assumptions.
- None of these assumptions are very good.
- The *effective* population size (N_e) is what N *would* be if the population actually met all of the assumptions.

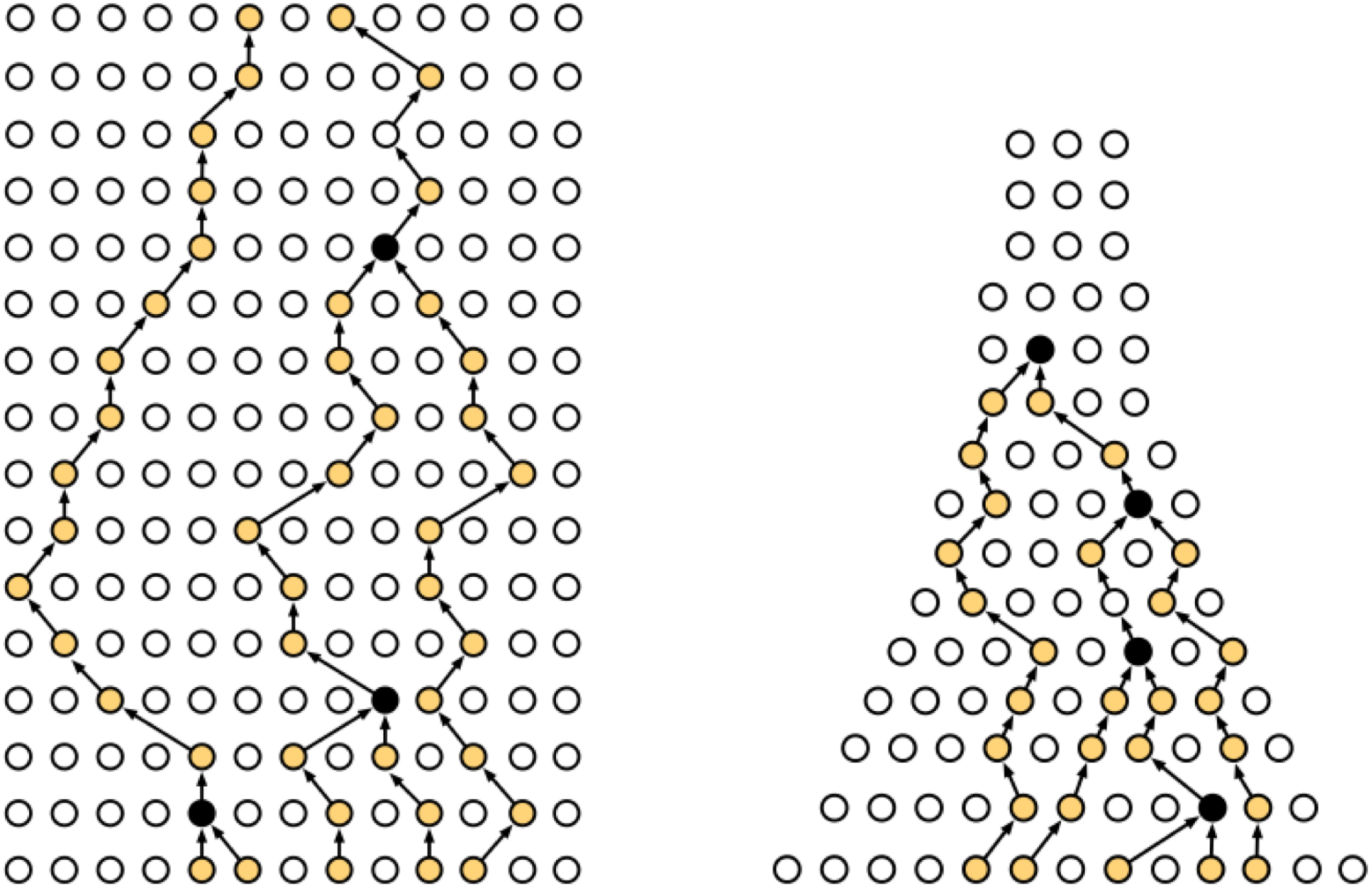
Q2. Name two reasons why N_e won't equal N.

N_e for pathogens

- The model is even worse for pathogens!
- Pathogen populations are even more structured than "large organism" populations.
- Each host is a population.
- We usually use one bulk (average) sequence from each host!
- We often talk about *effective number of infections*, but it is a lot more complicated than that!

Population dynamics

- If the population changes, the coalescent tree should as well:

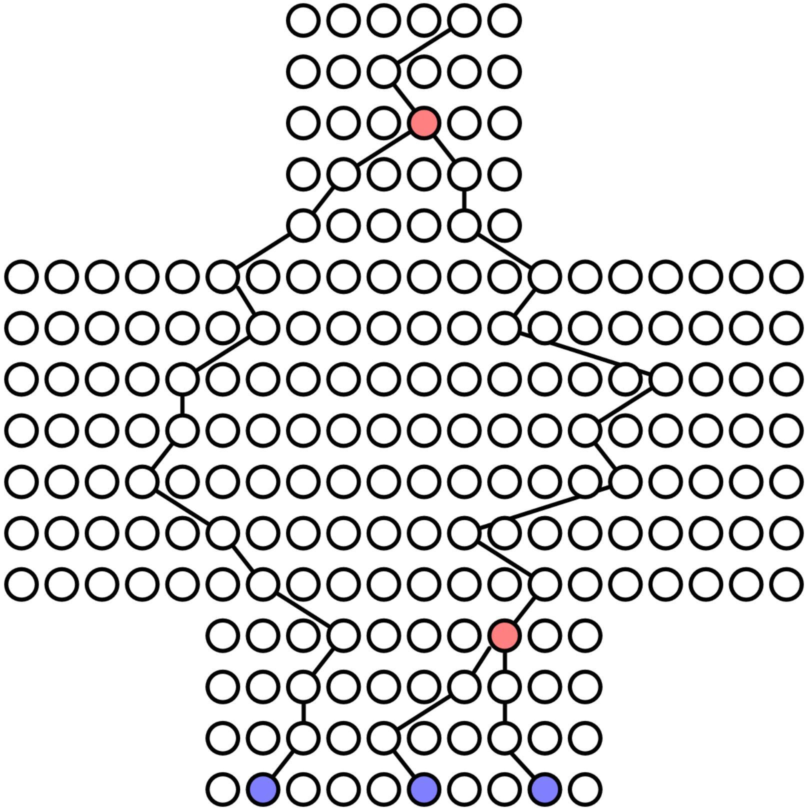


Tree shapes

- An exponentially growing population should have a "star-like" tree with most coalescent events in the past.
- A constant-size population should have most coalescent events in the recent past.
- A collapsing population should have many more coalescent events in the recent past.

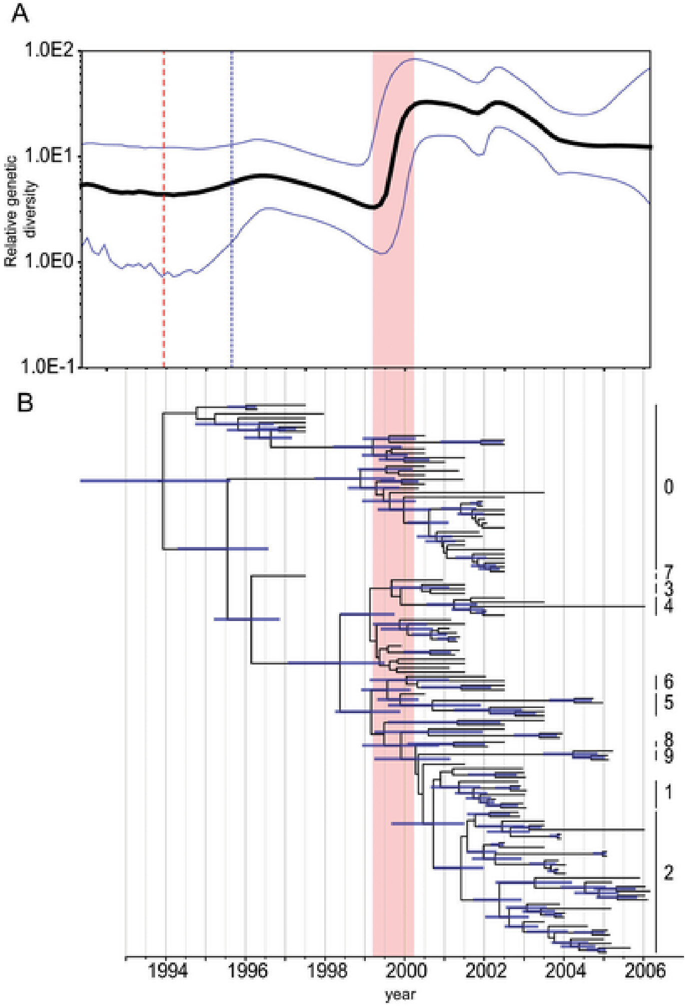
Skyline models

- What if the population dynamics don't follow any simple model?
- A skyline estimates when the population size changes, and the new size.



Smooth skylines

- "Averaging" the skyline over a sample from the posterior distribution.
- (right) Population dynamics of genetic diversity of H5N1 viruses isolated from poultry in China (Vijaykrishna *et al*, PLOS Pathogens 2008:4)



Hepatitis C virus in Egypt

- About 15% of adult population infected by HCV genotype 4
- Coalescent reconstructed found epidemic growth associated with massive public health campaign against snail fever.

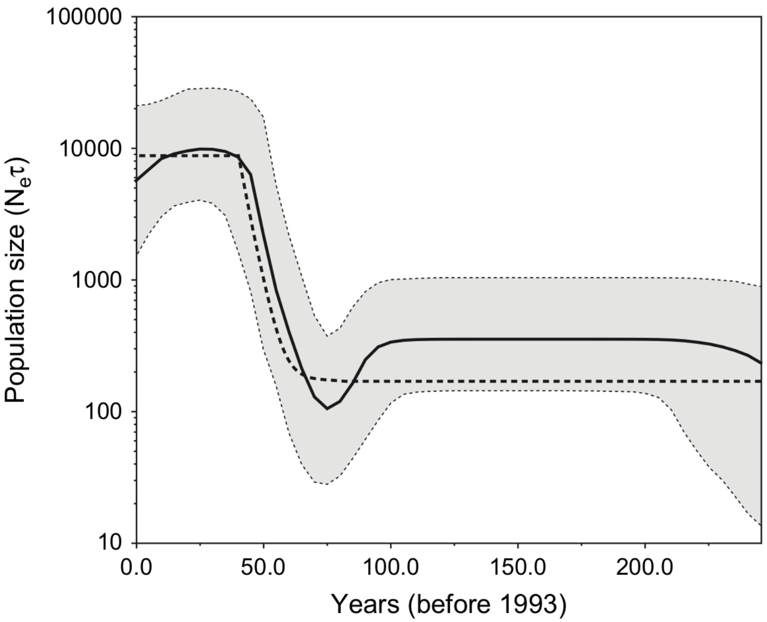
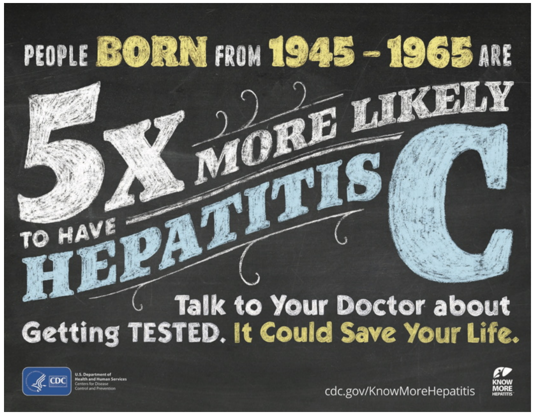


Figure from Drummond *et al*, 2005. *Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences.* Mol Biol Evol 22: 1185-1192.

Hepatitis C virus in North America

- HCV is highly prevalent in the "baby boomer" generation
- Why? Unsafe sex practices, experimenting with drugs?
- Who will pay for new treatments that cost $10,000's of dollars?


PEOPLE BORN FROM 1945 – 1965 ARE
5X MORE LIKELY
TO HAVE HEPATITIS C
Talk to Your Doctor about Getting TESTED. It Could Save Your Life.
CDC
cdc.gov/KnowMoreHepatitis
KNOW MORE HEPATITIS

Born between
1945 & 1965?
We are generation HEP
Over 75% of British Columbians who have hepatitis C belong to our generation.
Get tested. It's easy. There's a cure.
generationHEP.com

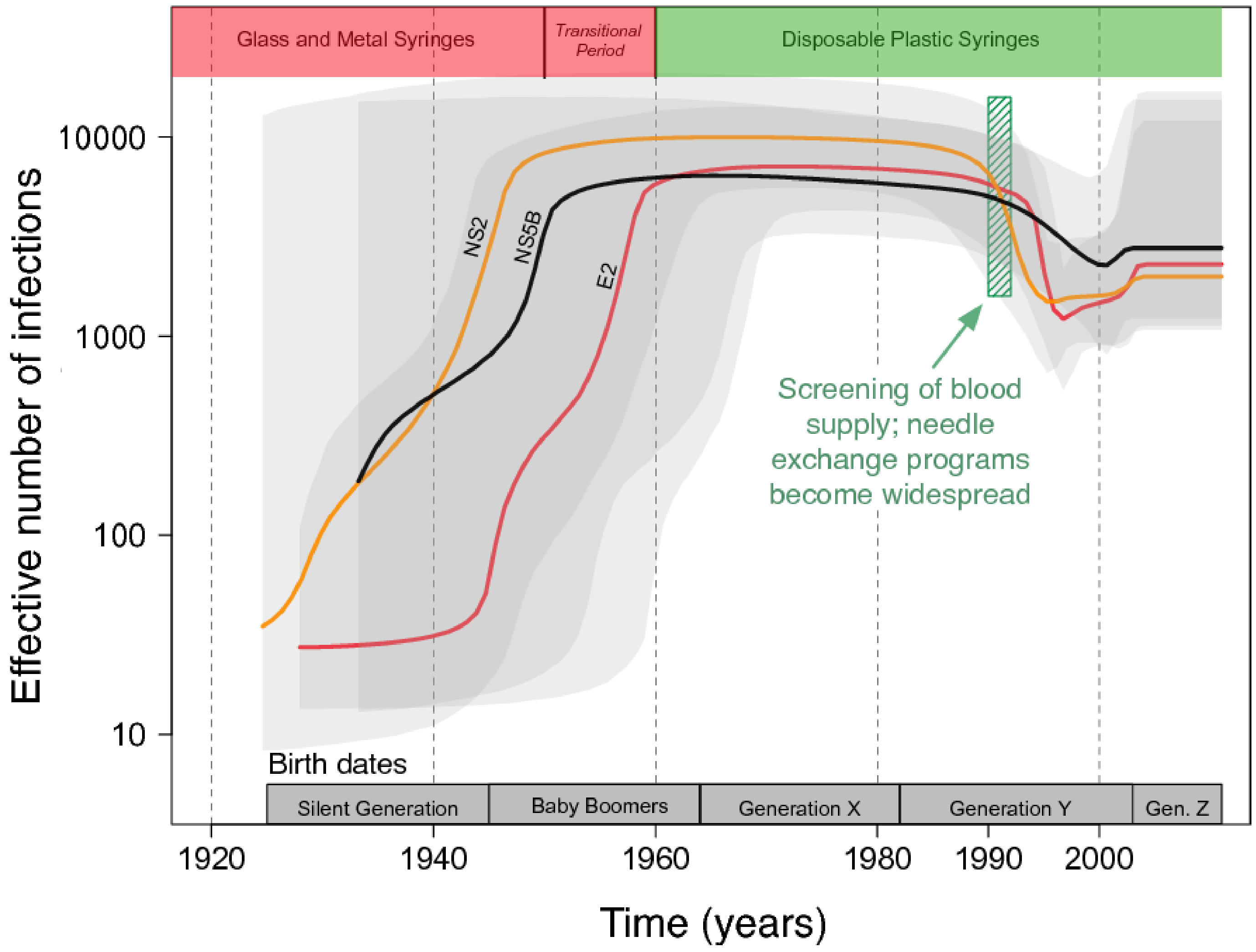


Figure from Joy *et al*. Lancet Inf Dis 16(6): 698-702.