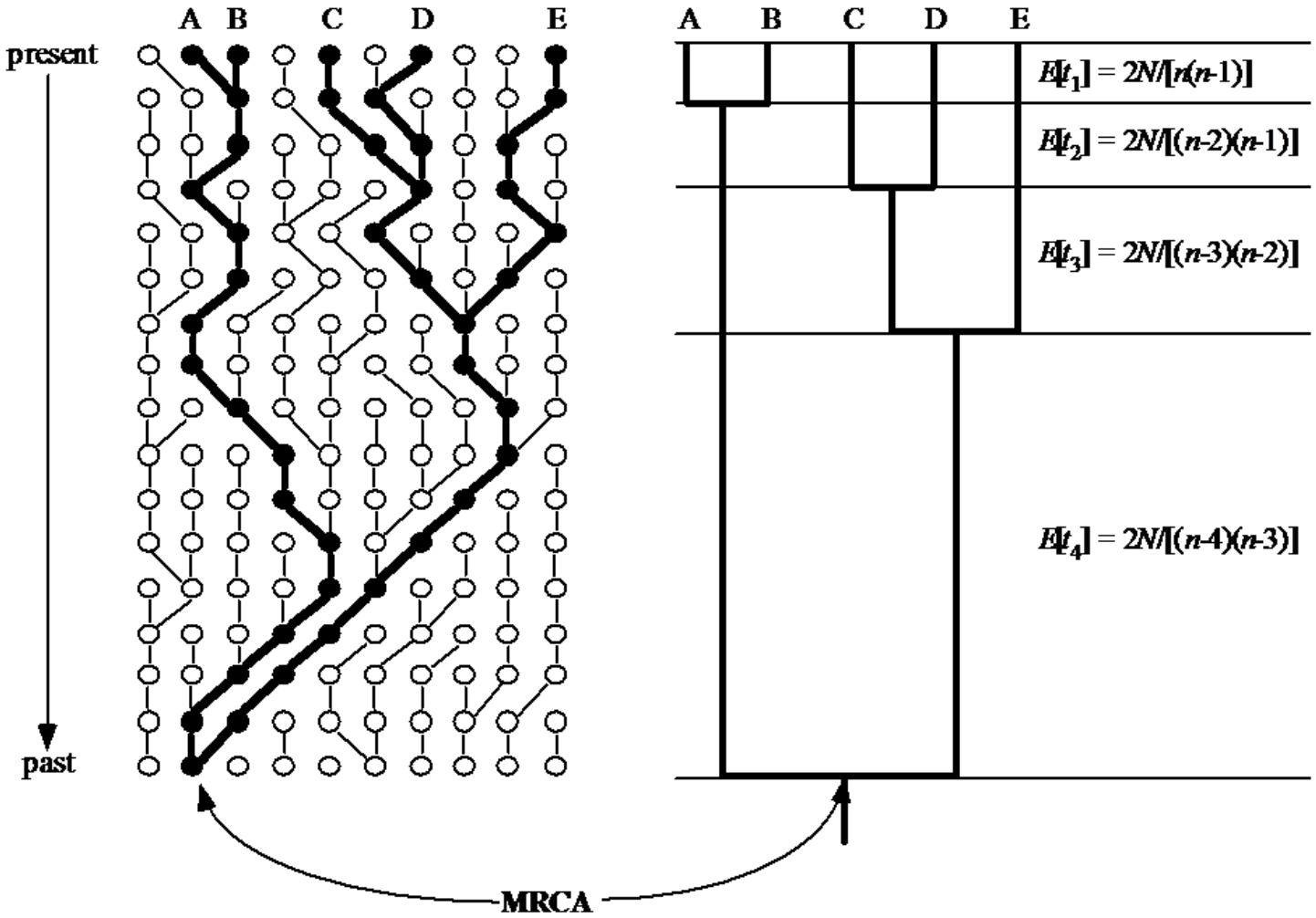
The Coalescent: Genealogies and their Properties

Steven Wu, Katia Koelle, and Allen Rodrigo
Duke University and NESCent

Genealogies

- Genealogies describe the ancestral histories of individuals in a population
- The times to most recent common ancestry (tMRCA) can be used as a proxy for relatedness.
- Superimposed on genealogies are mutations that may be associated with phenotypic traits.
- Consequently, genealogies have been used to model the distributions of mutations, and their associations to such traits.



The **Kingman coalescent** describes the genealogy of a sample of sequences from a large population.

The expected time for each coalescent interval is exponentially distributed with mean $E[t_{n \to n-1}] = 2N / n(n-1)$ generations, where $n << N$

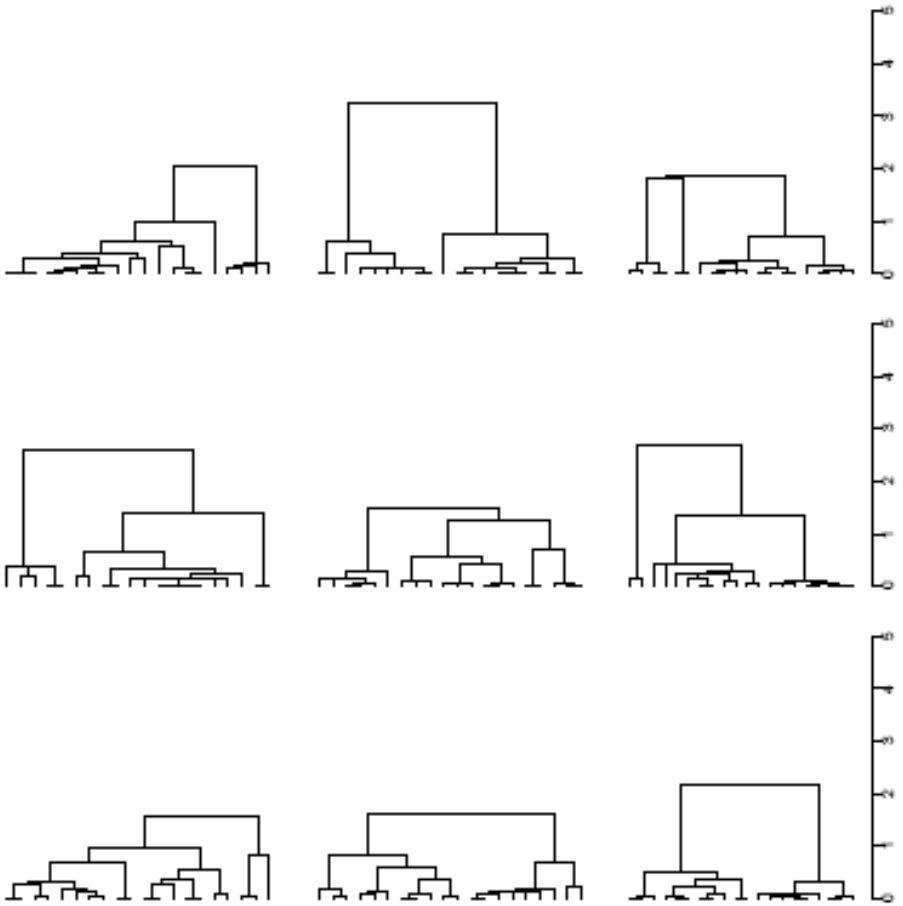
Simulating from the Coalescent

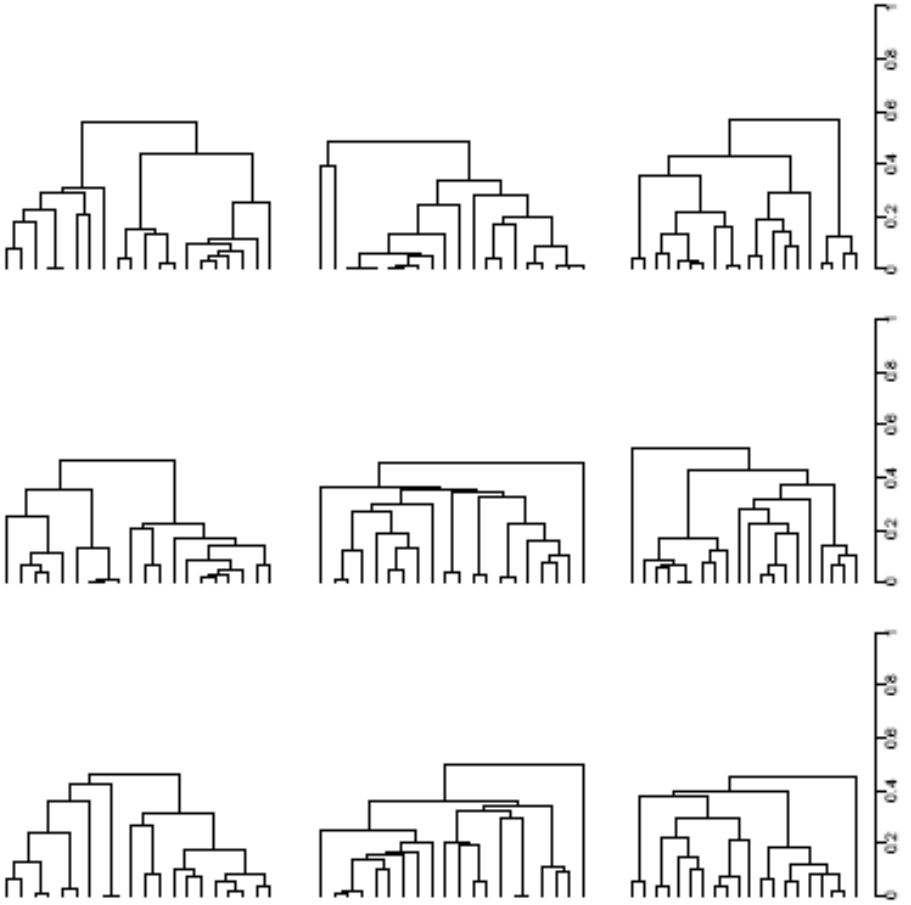
1. Select two individuals at random from the number of individual lineages, *n*, that have yet to coalesce.
2. Draw a random variable from the exponential distribution with mean $2N/n(n-1)$
3. join the two lineages drawn in Step 1, after the interval has elapsed.
4. If $n=2$, stop; otherwise, set n to $n-1$ and return to Step 1.

.

Random Trees

Exponentially Growing Populations





NESCent
National Evolutionary Synthesis Center

www.nescent.org

Phylodynamics

Grenfell et al. 2004 *Science* 303

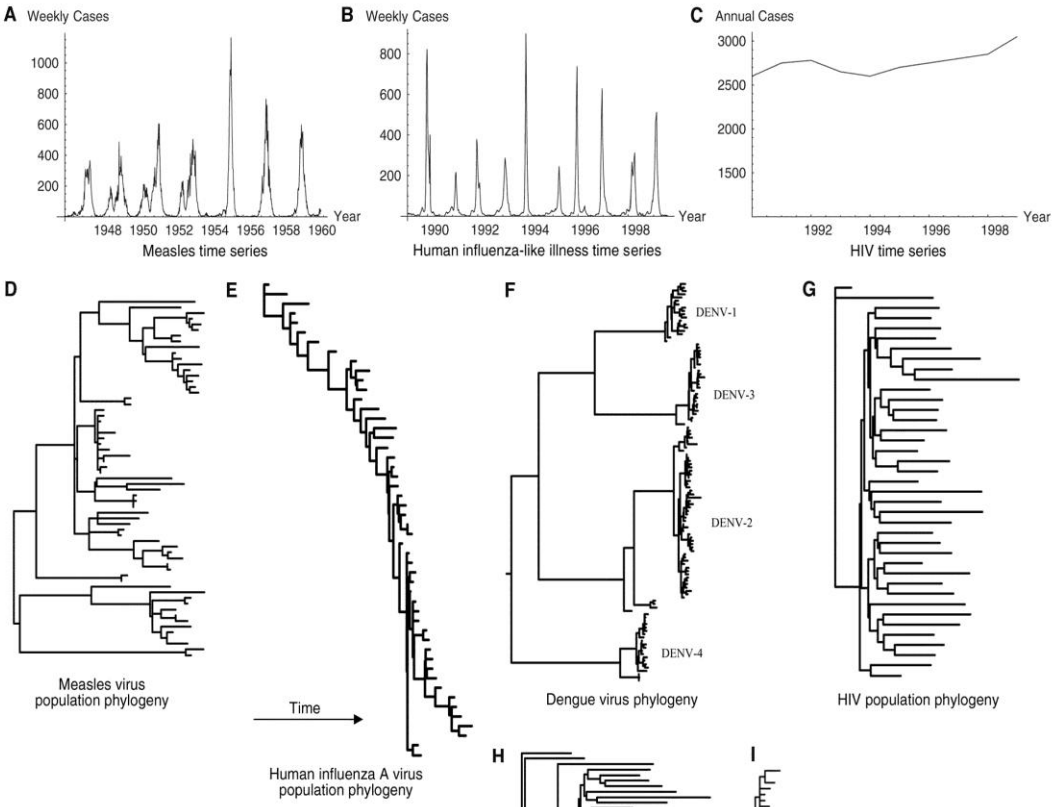


Fig. 1. (**A**) Prevaccination measles dynamics: weekly case reports for Leeds, UK (*7*). (**B**) Weekly reports of influenza-like illness for France (*44*). (**C**) Annual diagnosed cases of HIV in the United Kingdom (*45*). (**D**) Measles phylogeny: the measles virus nucleocapsid gene [63 sequences, 1575 base pairs (bp)]. (**E**) Influenza phylogeny: the human influenza A virus (subtype H3N2) hemagglutinin (*HA1*) gene longitudinally sampled over a period of 32 years (50 sequences, 1080 bp). (**F**) Dengue phylogeny: the dengue virus envelope gene from all four serotypes (DENV-1 to DENV-4, 120 sequences, 1485 bp). (**G**) HIV-1 population phylogeny: the subtype B envelope (E) gene sampled from different patients (39 sequences, 2979 bp). (**H**) HCV population phylogeny: the virus genotype 1b E1E2 gene sampled from different patients (65 sequences, 1677 bp). (**I**) HIV-1 within-host phylogeny: the partial envelope (E) gene longitudinally sampled from a single patient over 5.8 years [58 sequences, 627 bp; patient 6 from (*26*)]. All sequences were collected from GenBank and trees were constructed with maximum likelihood in PAUP* (*46*). Horizontal branch lengths are proportional to substitutions per site. Further details are available from the authors on request.

Phylodynamics

Grenfell et al. 2004 *Science* 303

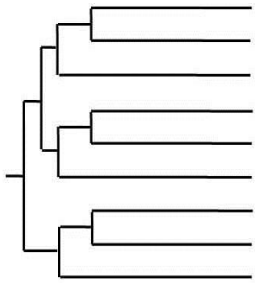
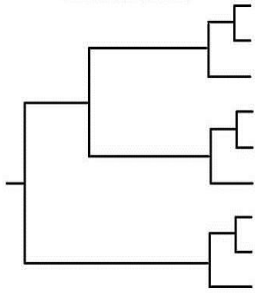
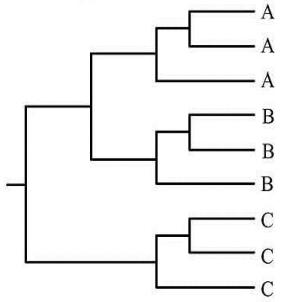
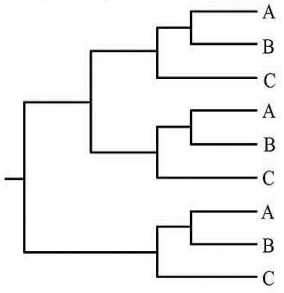
	Continual Immune Selection	Weak or Absent Immune Selection	
		Tree shape controlled by non-selective population dynamic processes	
		Population size dynamics	Spatial dynamics
Idealized Phylogeny Shapes	Time →	**Exponential growth**  **Constant size** 	**Strong spatial structure**  **Weak spatial structure** 
Examples	Human influenza A virus intra-host HIV	inter-host HIV inter-host HCV	Measles, rabies inter-host HIV
Tree Inferences	Detection of antigenic escape mutations	Estimation of population growth rates	Estimation of population migration rates

Fig. 3. Idealized tree shapes under different phylodynamic processes. The main division is between those viruses subject to continual immune-driven selection (such as human influenza A virus and intra-host HIV), in which trees have a strong temporal structure, and viruses where immune selection is absent or weak (such as many RNA viruses), in which the trees depict population size and spatial dynamics. The types of evolutionary inference that can be made from the various phylogenies are also indicated. (A, B, and C represent three subpopulations from which viruses have been sampled.)

NESCent

National Evolutionary Synthesis Center

NESCent
National Evolutionary Synthesis Center

www.nescent.org

$tMRCA_{AB} \sim \exp(1/N)$

$E[tMRCA_{AB}] = N$

NESCent
National Evolutionary Synthesis Center

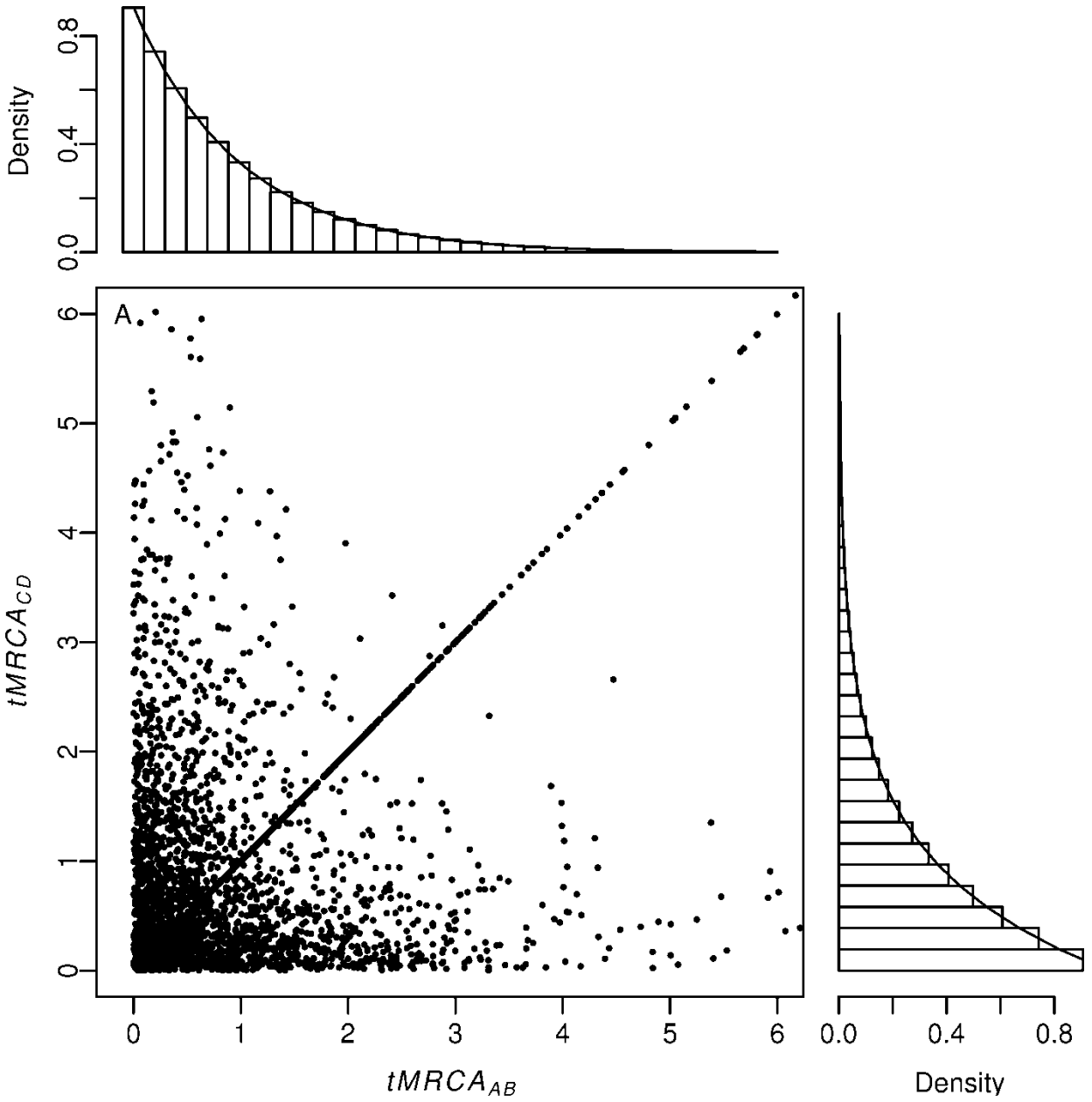
Given that we know *tMRCA*$_{AB}$, what do we know about *tMRCA*$_{CD}$?

Simulation Analysis

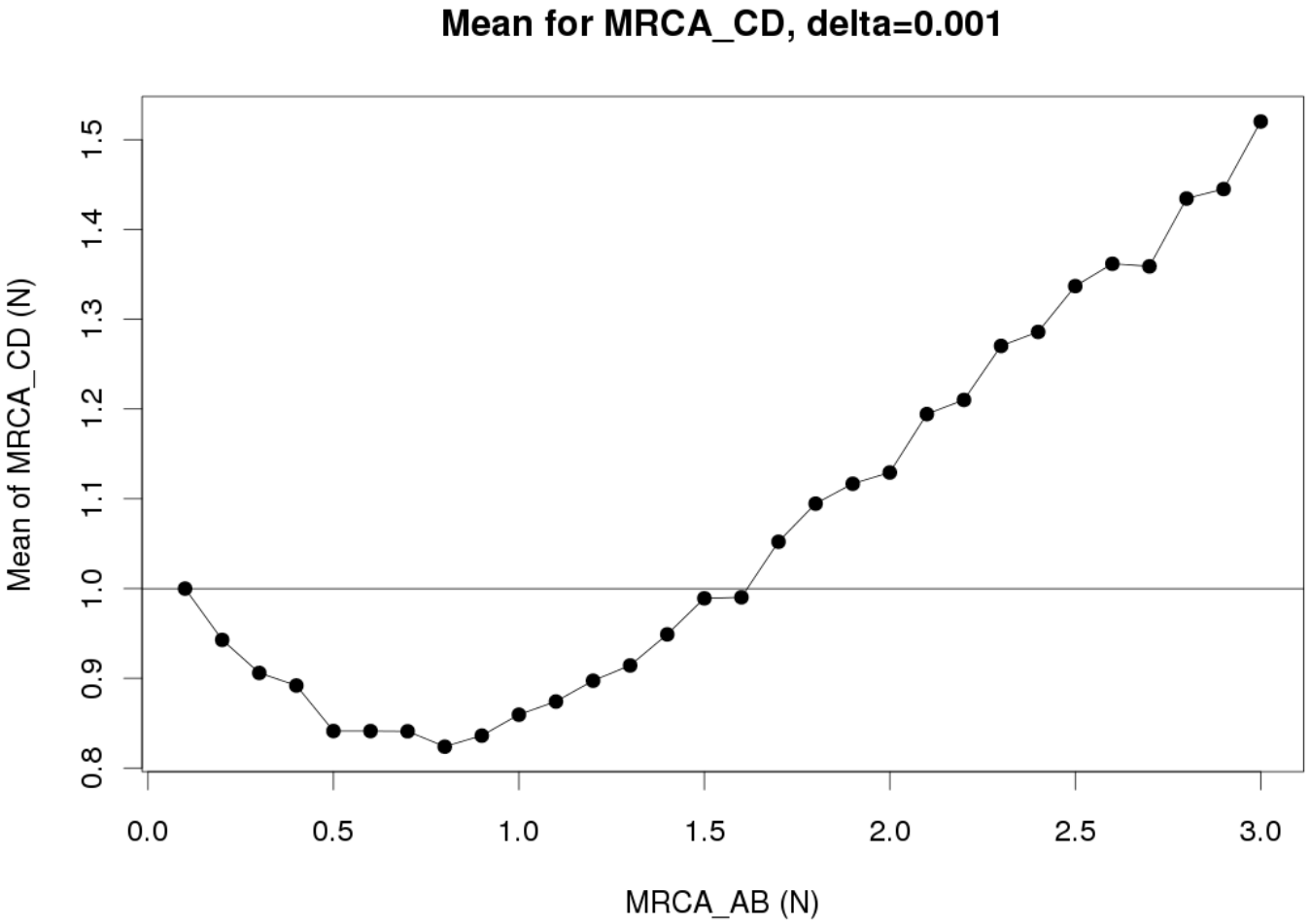
- Simulate genealogies of *A, B, C* and *D*
  - Population size N = 10,000
  - No. of iterations = 10,000,000
  - Scale all times to common ancestry in units of N
- For each genealogy, plot $tMRCA_{AB}$ against $tMRCA_{CD}$

NESCent

National Evolutionary Synthesis Center




NESCent
National Evolutionary Synthesis Center
A
B
C
D
E
www.nescent.org


Mean for MRCA_CD, delta=0.001
Mean of MRCA_CD (N)
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
MRCA_AB (N)

Coalescent Entanglement

- The **conditional dependence** of times to common ancestry of randomly sampled pairs of individuals.

- Analogy to “quantum entanglement” which Einstein described as “spooky action at a distance”.

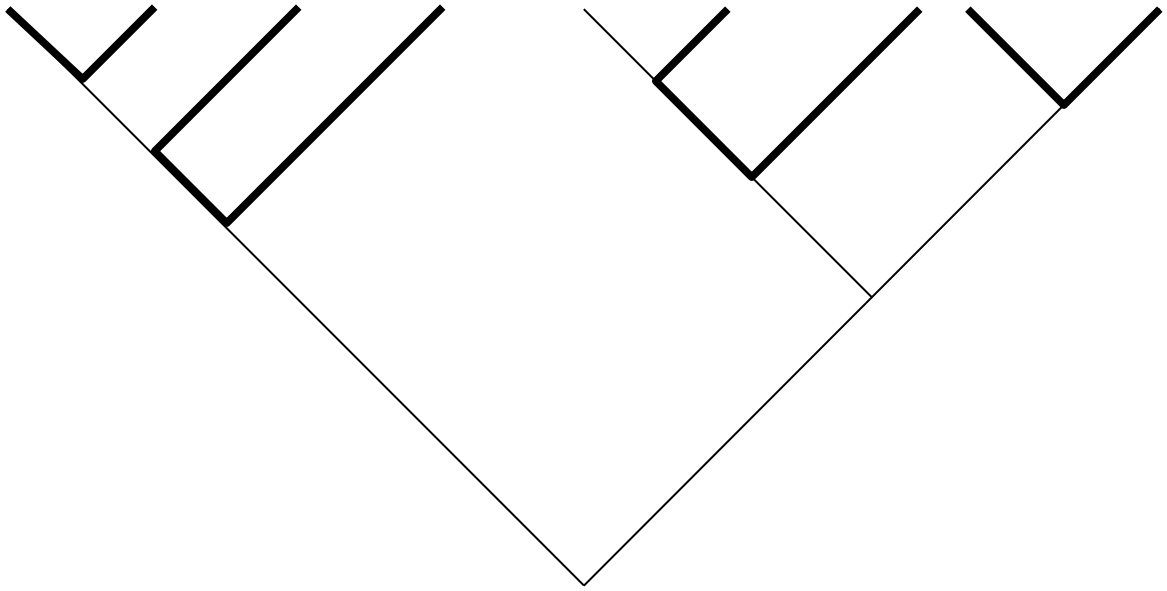
NESCent

National Evolutionary Synthesis Center

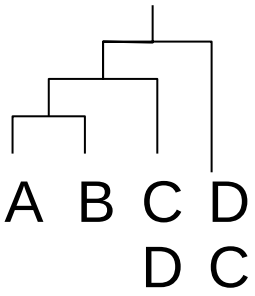
What about phylogenetically independent pairs?

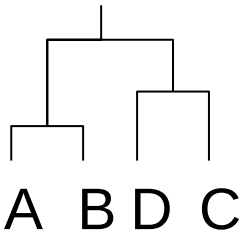
- "Independent" pairwise comparison – Identify pairs of taxa with no common lineages in the connecting paths.

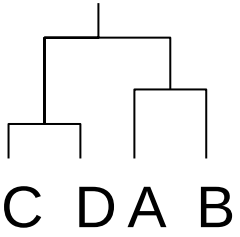
Height	5'8	5'9	5'11	5'6	5'10	5'8	5'10	5'7	5'11
Weight	150	155	180	145	170	165	165	170	175

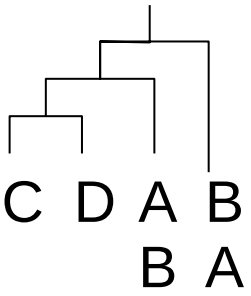


Phylogenetically independent pairs (A,B) and (C,D)



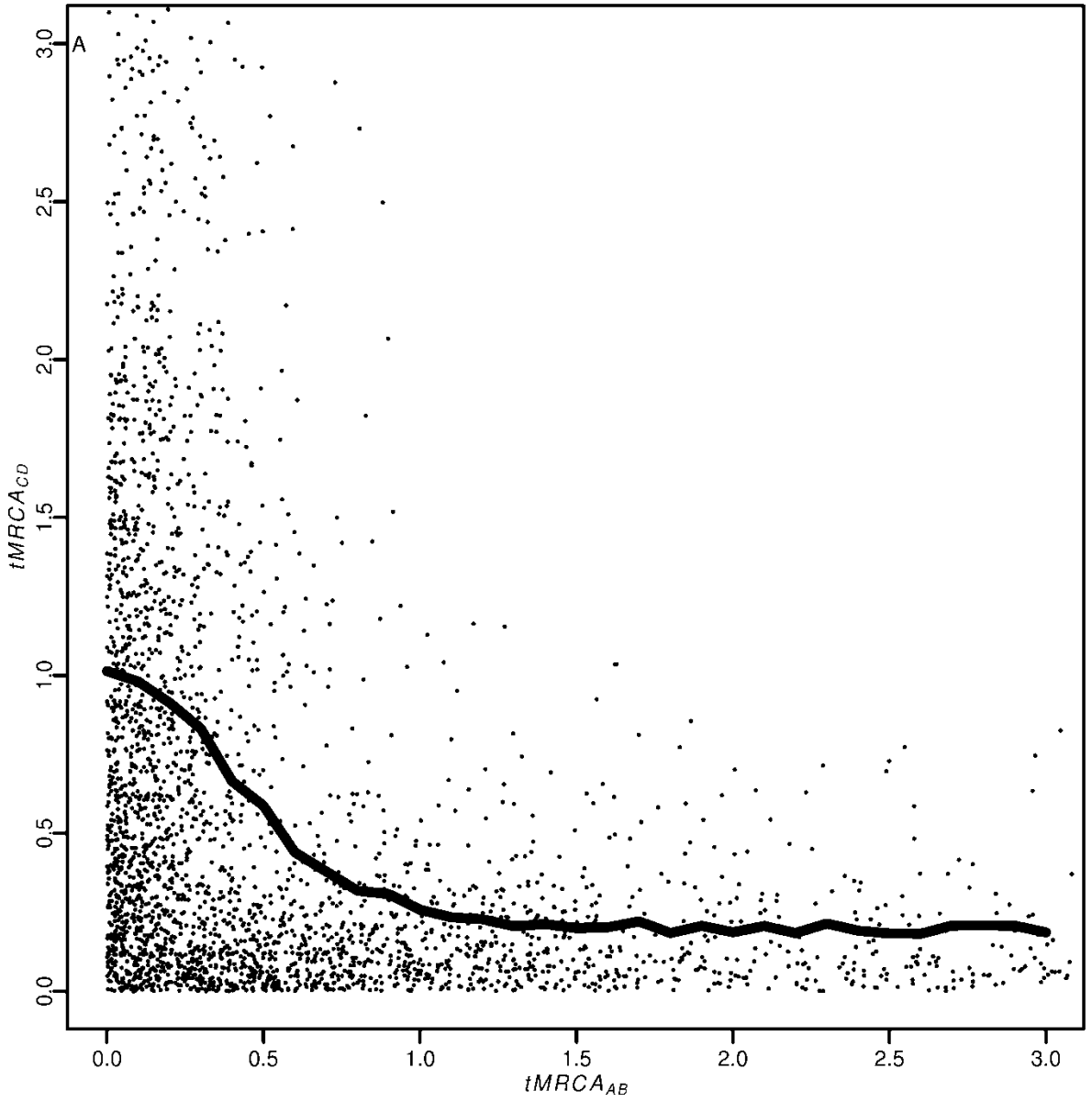






NESCent

National Evolutionary Synthesis Center



Summary

- The coalescent provides a model to understand the properties of genealogies.
- In turn, we can make inferences about quantitative traits of individuals in a population.
- The coalescent exposes counterintuitive dependencies of the traits of apparently unrelated individuals.