

Sep 18th, 4:15 PM - 5:15 PM

Applying a historical science as a practical science: big data, evolution, medicine, and public health

Jeffrey Townsend
Yale University

Follow this and additional works at: <http://elischolar.library.yale.edu/dayofdata>

Jeffrey Townsend, "Applying a historical science as a practical science: big data, evolution, medicine, and public health" (September 18, 2015). *Yale Day of Data*. Paper 1.
<http://elischolar.library.yale.edu/dayofdata/2015/Schedule/1>

This Event is brought to you for free and open access by EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Yale Day of Data by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

Applying a historical science as a practical science: big data, evolution, medicine, and public health

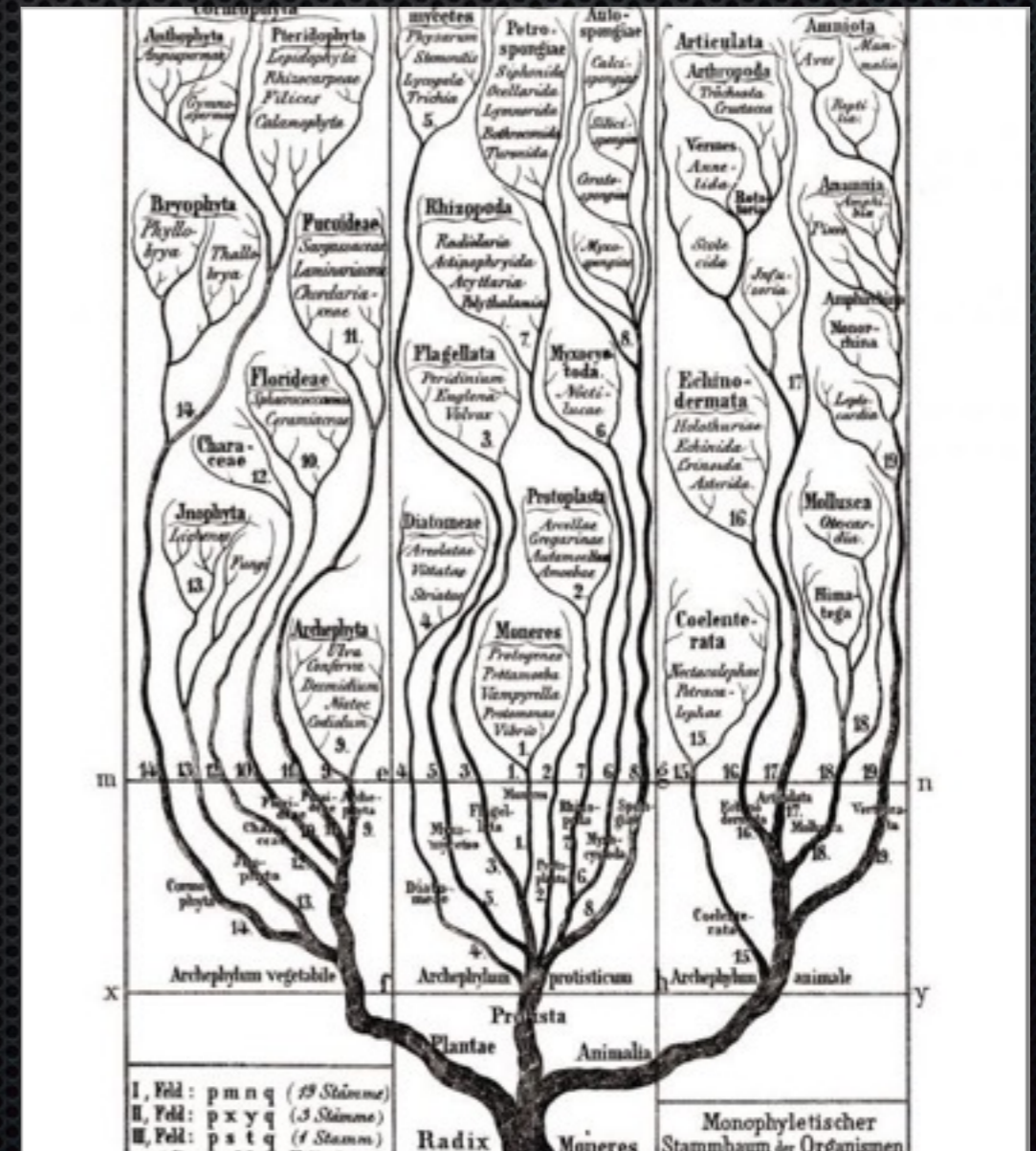
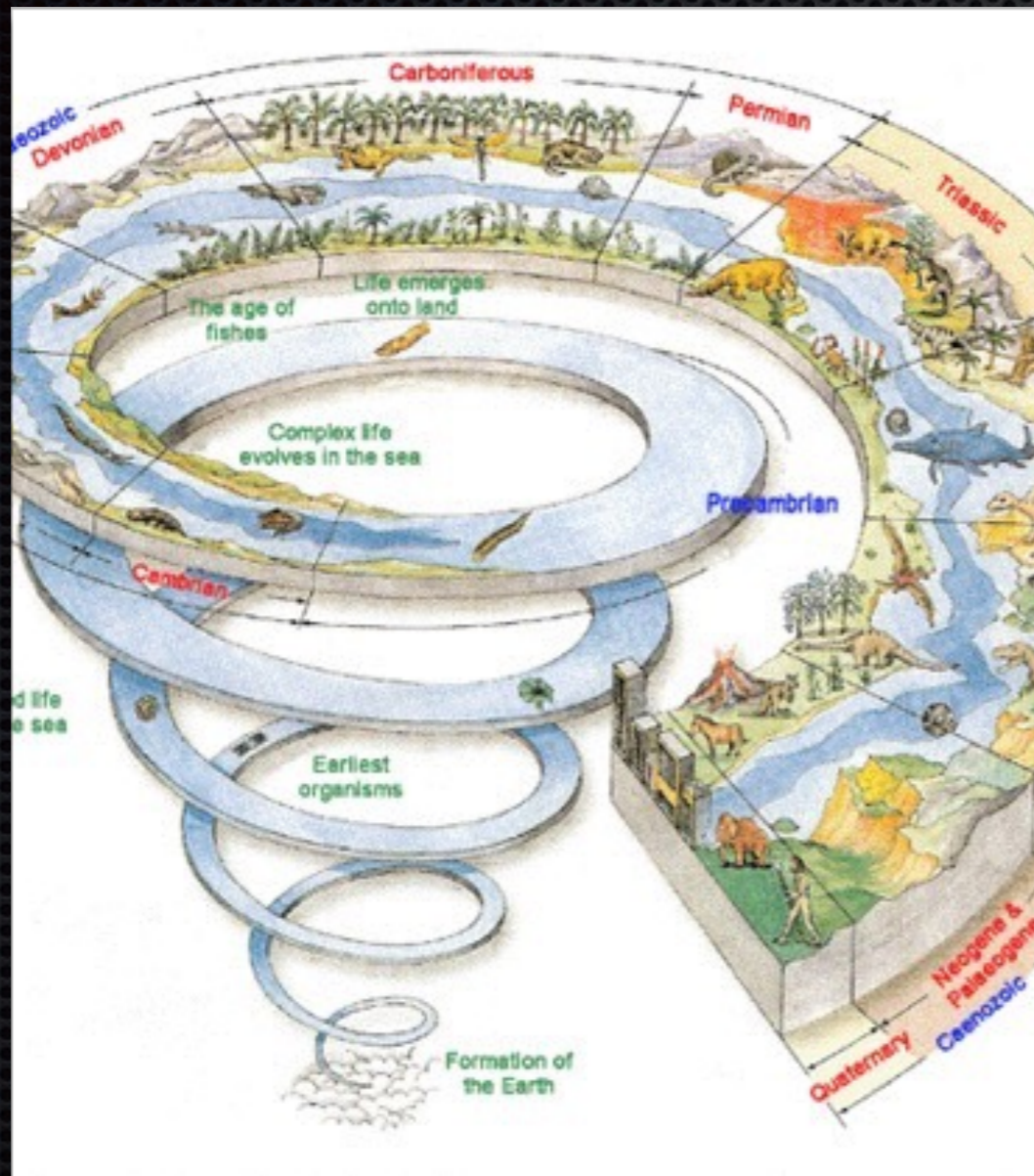
Jeffrey P. Townsend, Ph.D.

Associate Professor of Biostatistics and Ecology & Evolutionary Biology

Director of Bioinformatics, Yale Center for Analytical Sciences

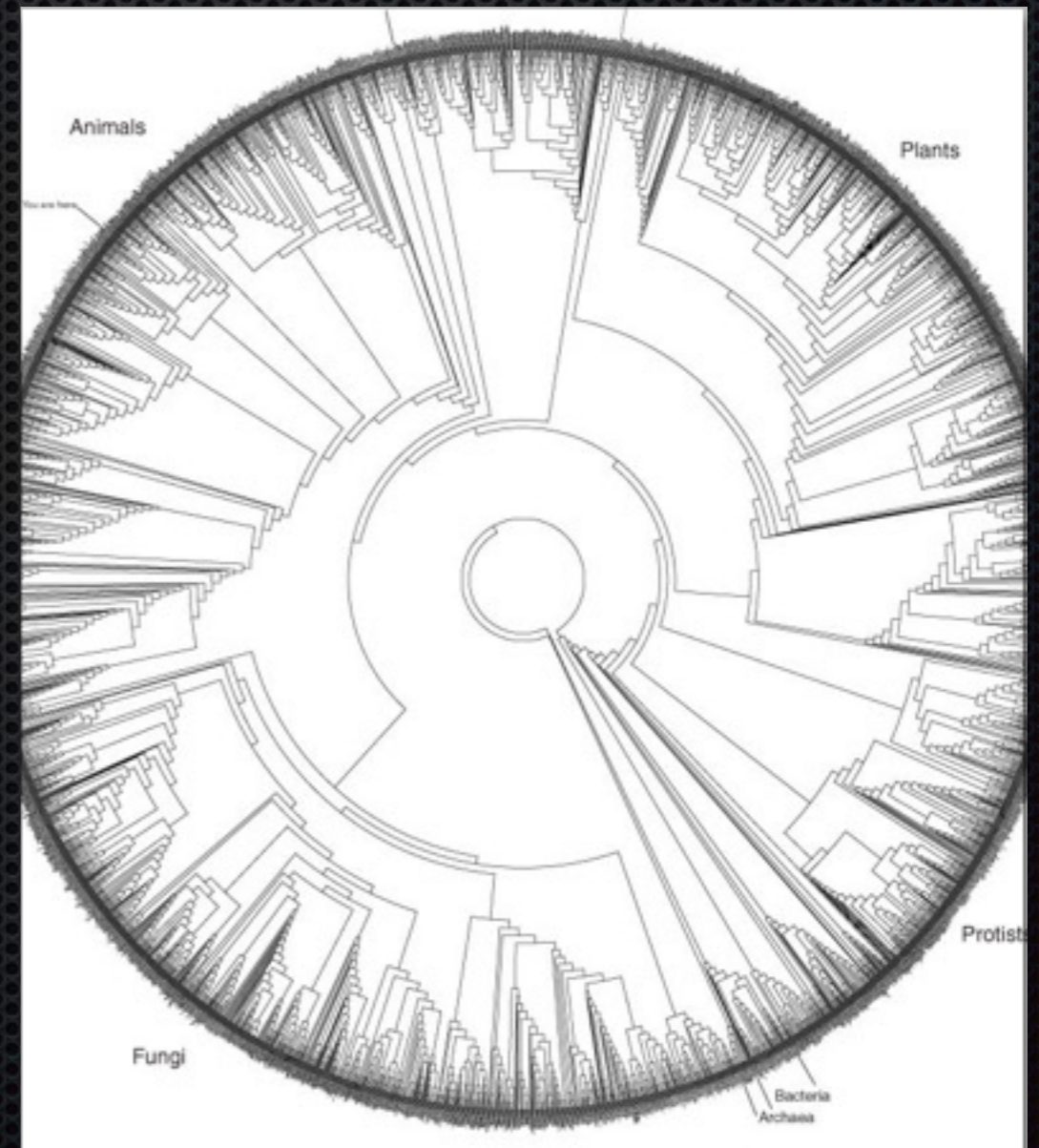
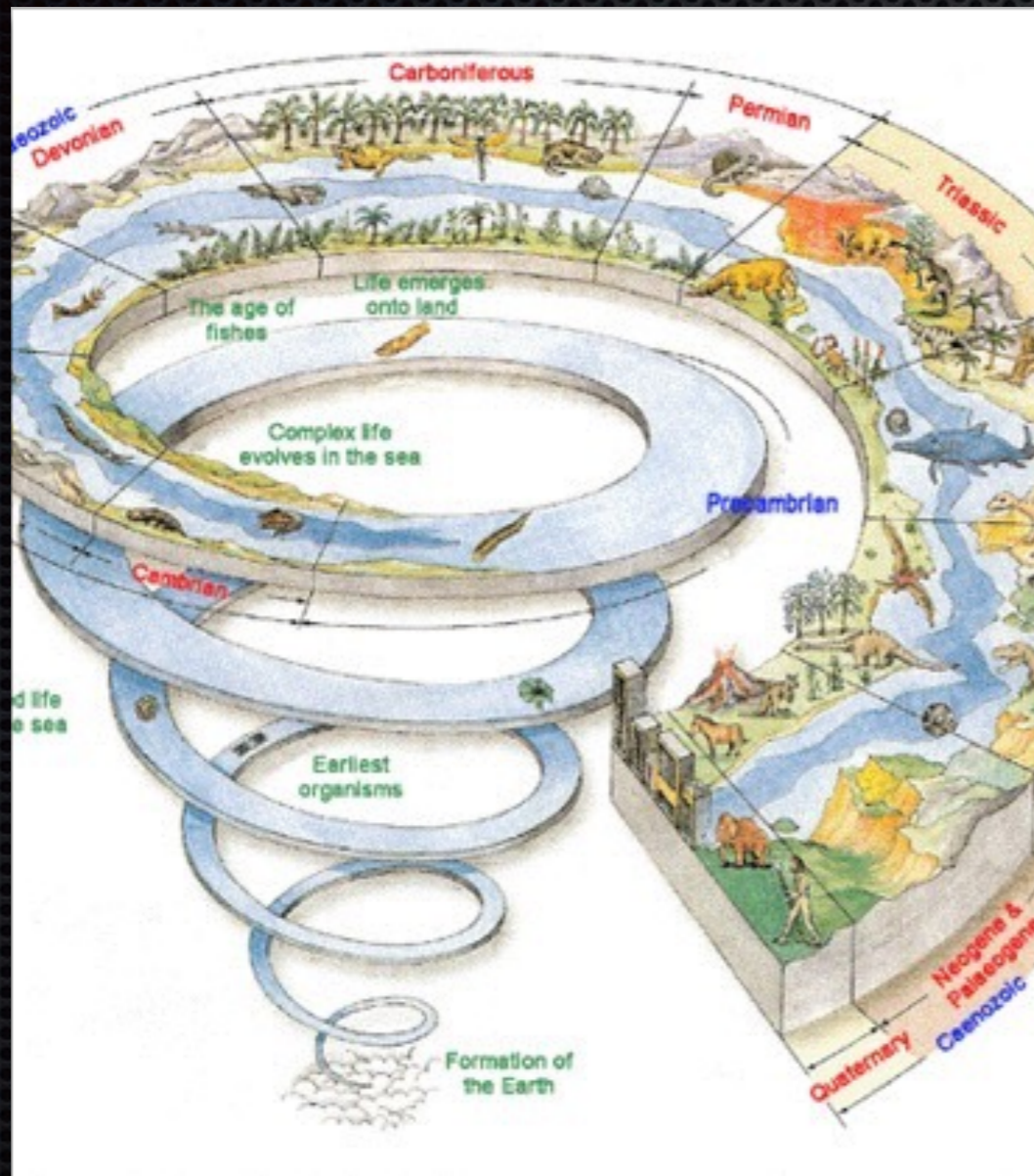
Yale University

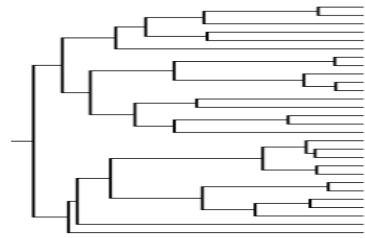
Evolutionary biology is usually considered a historical science



Ernst Haeckel (1866)

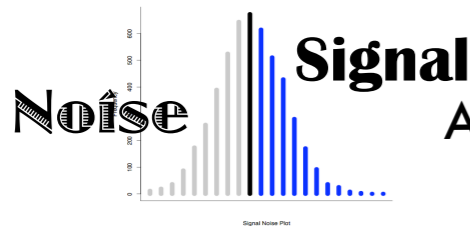
Increases in sequencing data have provided enormous illumination regarding the disciplinary mission of illuminating the deep history of life





ACGTTGCAACGT

$$\prod_{i \in N, j \in M} (p_{ik} q_{jk})^{c_{ijk}} (1 - p_{ik} q_{jk})^{S_{ij} - c_{ijk}}$$



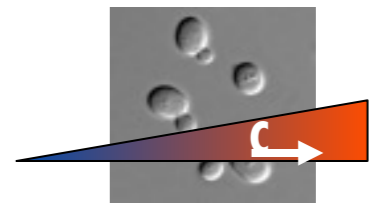
ACGTTGCAACGT



ACGTTGCAACGTT

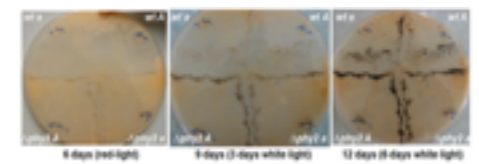
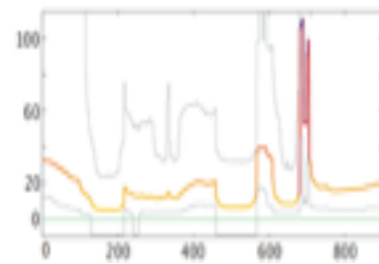
ACGTTGCAACGT

ACGTTGCAACGT

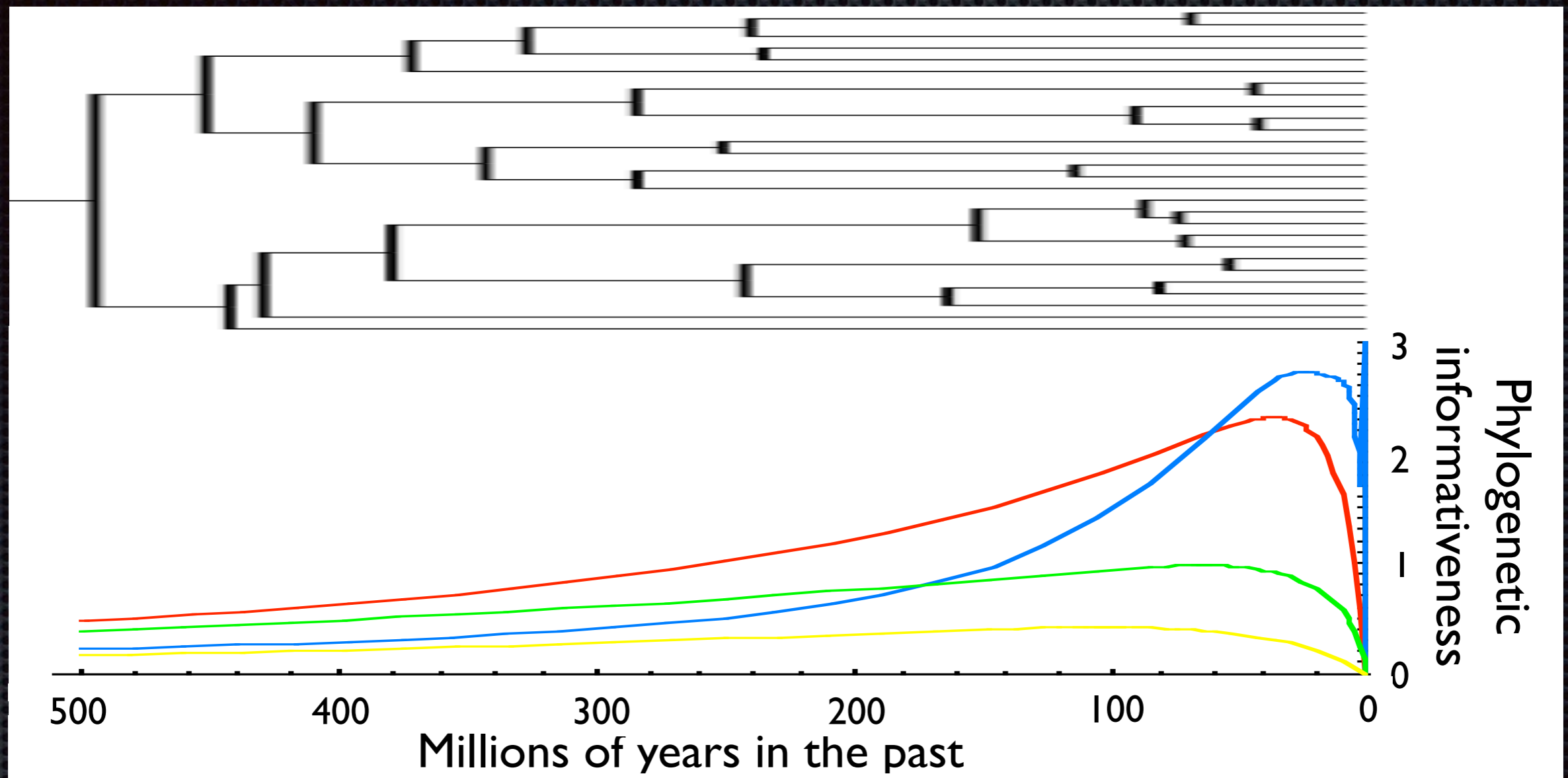


ACGTTGCAACGT

ACGTTGCAACGT

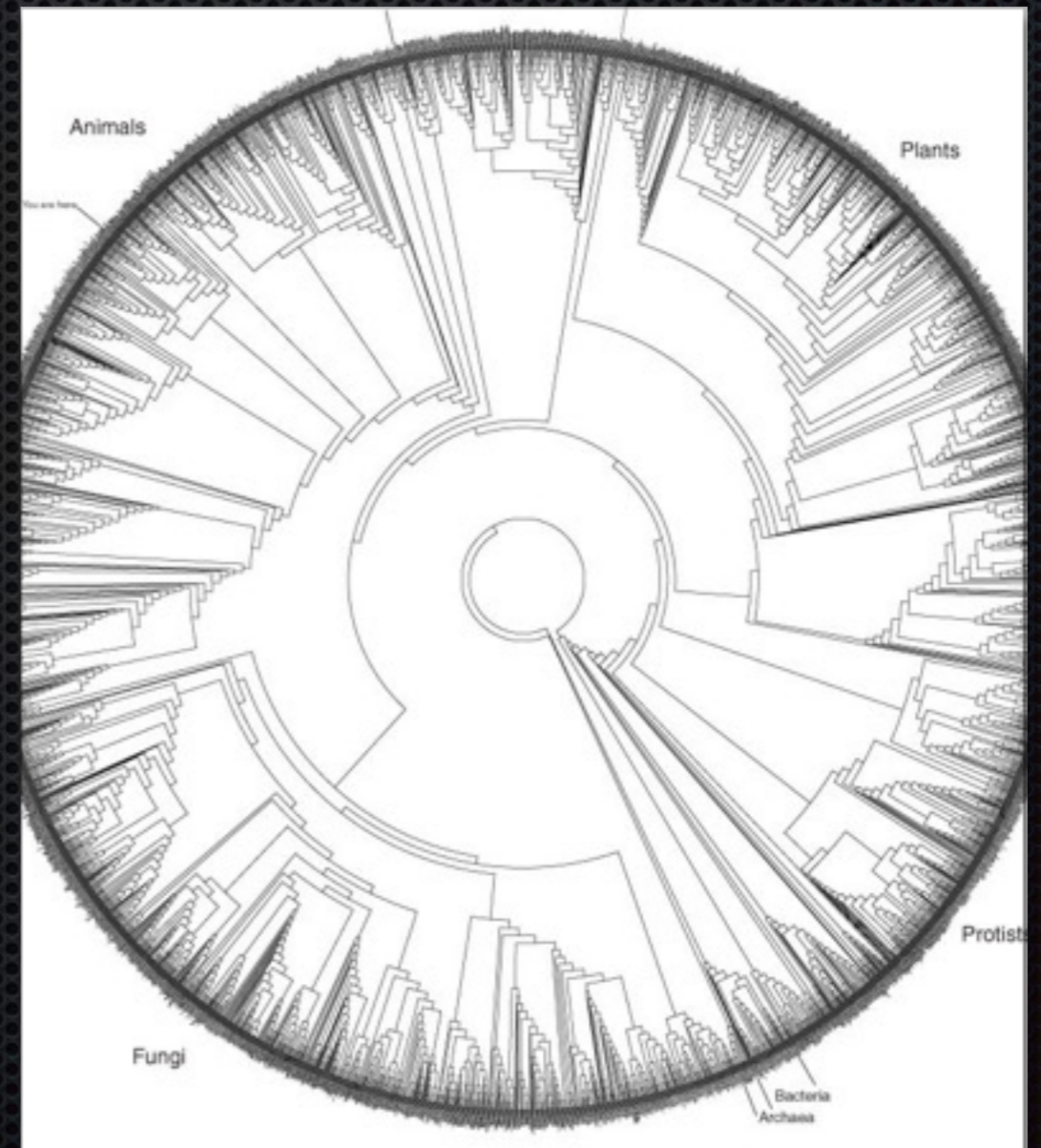
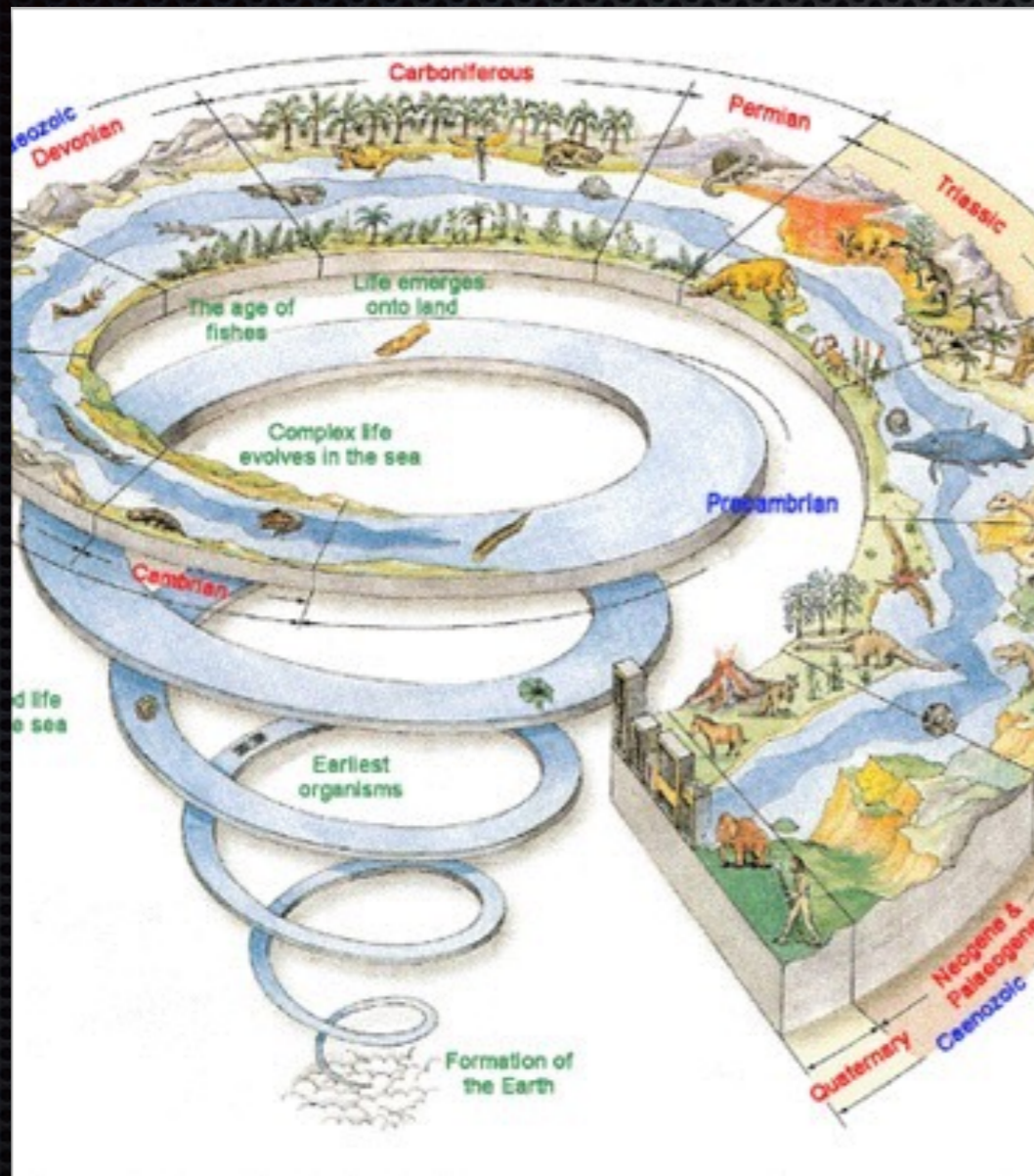


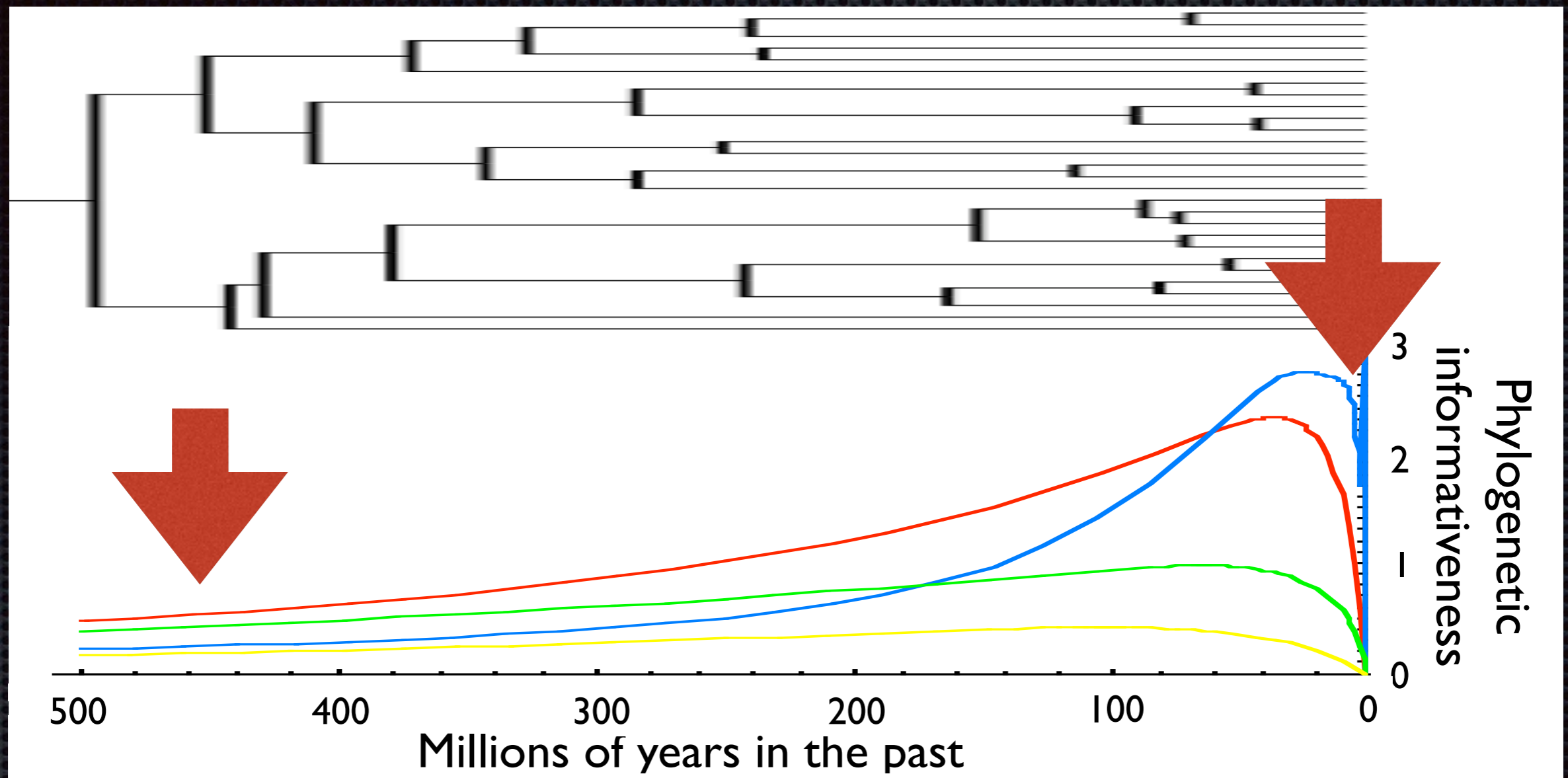
The production of increasing amounts of sequence data has provided increasing resolution



Diverse types of information on evolutionary history follow a characteristic function

Illumination of the deep history of life has mostly disciplinary outcomes





More data is especially helpful not only for deep evolutionary questions, but also extremely recent ones

Two examples of interdisciplinary collaboration using the techniques of evolutionary biology facilitated by next-generation sequencing data

- ✦ a collaboration with modelers in the School of Public Health, where only now are sequencing and analysis are getting rapid enough and sophisticated enough to provide revealing information about ongoing epidemics



Two examples of interdisciplinary collaboration using the techniques of evolutionary biology facilitated by next-generation sequencing data

- with pathologists, geneticists, and pharmacologists in the School of Medicine and an industrial partner, we sequenced, processed, stored, managed, and shared results from a massive tumor sequencing program



Two examples of interdisciplinary collaboration using the techniques of evolutionary biology facilitated by next-generation sequencing data

- ✦ a collaboration with modelers in the School of Public Health, where only now are sequencing and analysis are getting rapid enough and sophisticated enough to provide revealing information about ongoing epidemics



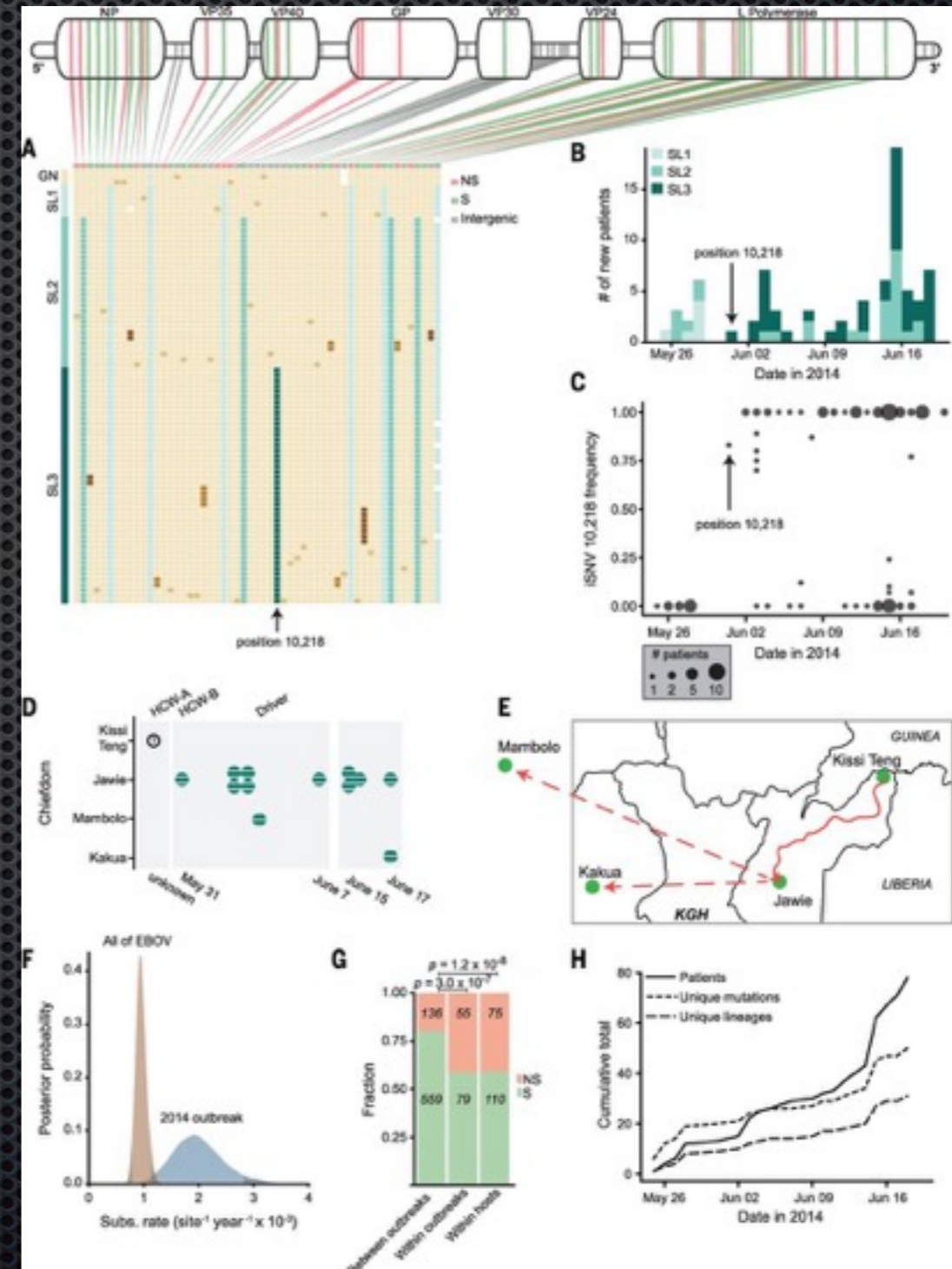
Genomic surveillance data was quickly released for the 2014 West African Ebola outbreak

VIRAL EVOLUTION

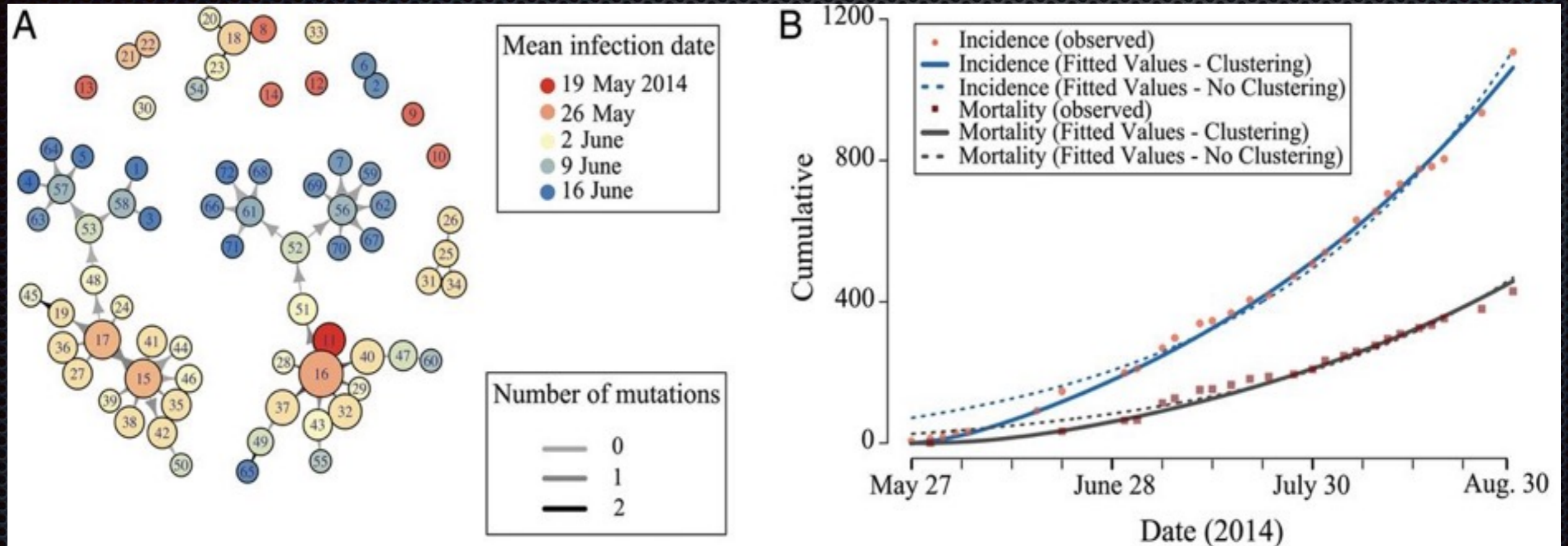
Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak

Stephen K. Gire,^{1,2*} Augustine Goba,^{3*†} Kristian G. Andersen,^{1,2*†} Rachel S. G. Sealton,^{2,4*} Daniel J. Park,^{2*} Lansana Kanneh,³ Simbirie Jalloh,³ Mambu Momoh,^{3,5} Mohamed Fullah,^{3,5†} Gytis Dudas,⁶ Shirlee Wohl,^{1,2,7} Lina M. Moses,⁸ Nathan L. Yozwiak,^{1,2} Sarah Winnicki,^{1,2} Christian B. Matranga,² Christine M. Malboeuf,² James Qu,² Adrienne D. Gladden,² Stephen F. Schaffner,^{1,2} Xiao Yang,² Pan-Pan Jiang,^{1,2} Mahan Nekoui,^{1,2} Andres Colubri,¹ Moinya Ruth Coomber,³ Mbalu Fonnle,^{3†} Alex Moigboi,^{3†} Michael Gbakie,³ Fatima K. Kamara,³ Veronica Tucker,³ Edwin Konuwa,³ Sidiki Saffa,^{3†} Josephine Sellu,³ Abdul Azziz Jalloh,³ Alice Kovoma,^{3†} James Koninga,³ Ibrahim Mustapha,³ Kande Kargbo,³ Momoh Foday,³ Mohamed Yillah,³ Franklyn Kanneh,³ Willie Robert,³ James L. B. Massally,³ Sinéad B. Chapman,² James Bochicchio,² Cheryl Murphy,² Chad Nusbaum,² Sarah Young,² Bruce W. Birren,² Donald S. Grant,³ John S. Scheffelin,⁸ Eric S. Lander,^{2,7,9} Christian Happi,¹⁰ Sahr M. Gevao,¹¹ Andreas Gnirke,^{2§} Andrew Rambaut,^{6,12,13§} Robert F. Garry,^{8§} S. Humarr Khan,^{3†§} Pardis C. Sabeti^{1,2†§}

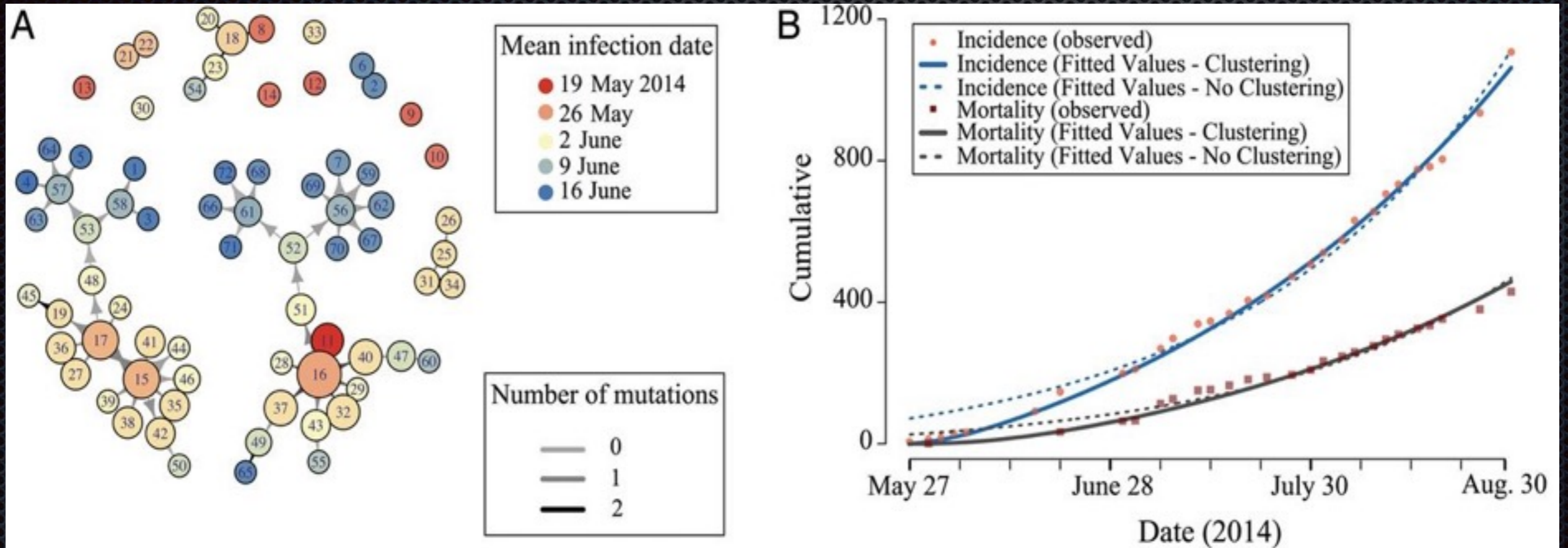
In its largest outbreak, Ebola virus disease is spreading through Guinea, Liberia, Sierra Leone, and Nigeria. We sequenced 99 Ebola virus genomes from 78 patients in Sierra Leone to ~2000x coverage. We observed a rapid accumulation of interhost and intrahost genetic variation, allowing us to characterize patterns of viral transmission over the initial weeks of the epidemic. This West African variant likely diverged from central African lineages around 2004, crossed from Guinea to Sierra Leone in May 2014, and has exhibited sustained human-to-human transmission subsequently, with no evidence of additional zoonotic sources. Because many of the mutations alter protein sequences and other biologically meaningful targets, they should be monitored for impact on diagnostics, vaccines, and therapies critical to outbreak response.



Samuel V. Scarpino,^{1,a} Atila Iamarino,^{2,3,a} Chad Wells,^{4,5} Dan Yamin,^{4,5}
Martial Ndeffo-Mbah,^{4,5} Natasha S. Wenzel,⁴ Spencer J. Fox,⁶
Tolbert Nyenswah,⁷ Frederick L. Altice,^{5,8} Alison P. Galvani,^{4,5,9,10}
Lauren Ancel Meyers,^{1,6} and Jeffrey P. Townsend^{2,9,10}



Epidemiological and viral genomic
sequence analysis of the 2014 Ebola
outbreak reveals clustered transmission



Our analysis revealed something of key importance that had no other means for evaluation: underreporting

Fewer Ebola Cases Go Unreported Than Thought, Study Finds

By DONALD G. McNEIL Jr. DEC. 16, 2014

Email

Share

Tweet

Save

More



Transmission of the Ebola virus occurs mostly within families, in hospitals and at funerals, not randomly like the flu, Yale scientists said Tuesday, and far fewer cases go unreported than has previously been estimated.

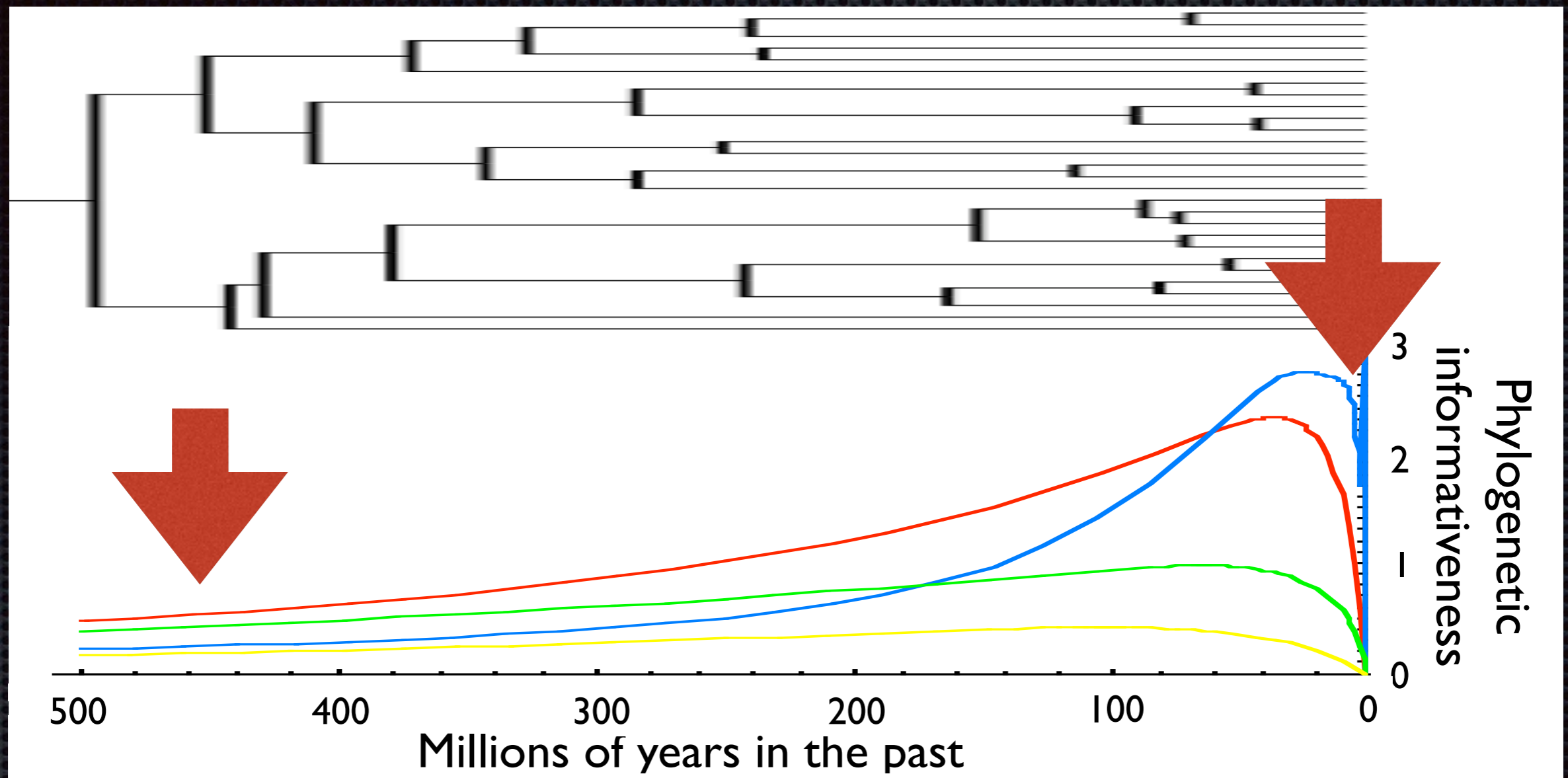
That implies, they said, that the epidemic is unlikely to reach the gloomy scenarios of [hundreds of thousands of cases](#) that studies released in September had forecast were possible; the most pessimistic one, from the [Centers for Disease Control and Prevention](#), had predicted [up to 1.4 million cases](#) by late January.

As of Monday, there were 18,464 confirmed cases in Liberia, Sierra Leone and Guinea, with 6,841 deaths, according to the [World Health Organization](#), far more than from all the



A woman in Monrovia, Liberia, spoke to children who were quarantined after their parents showed signs of Ebola; the father died. Daniel Berehulak for The New York Times

Our analysis revealed something of key importance that had no other means for evaluation: underreporting



More data is especially helpful not only for deep evolutionary questions, but also extremely recent ones

Two examples of interdisciplinary collaboration using the techniques of evolutionary biology facilitated by next-generation sequencing data

- with pathologists, geneticists, and pharmacologists in the School of Medicine and an industrial partner, we sequenced, processed, stored, managed, and shared results from a massive tumor sequencing program

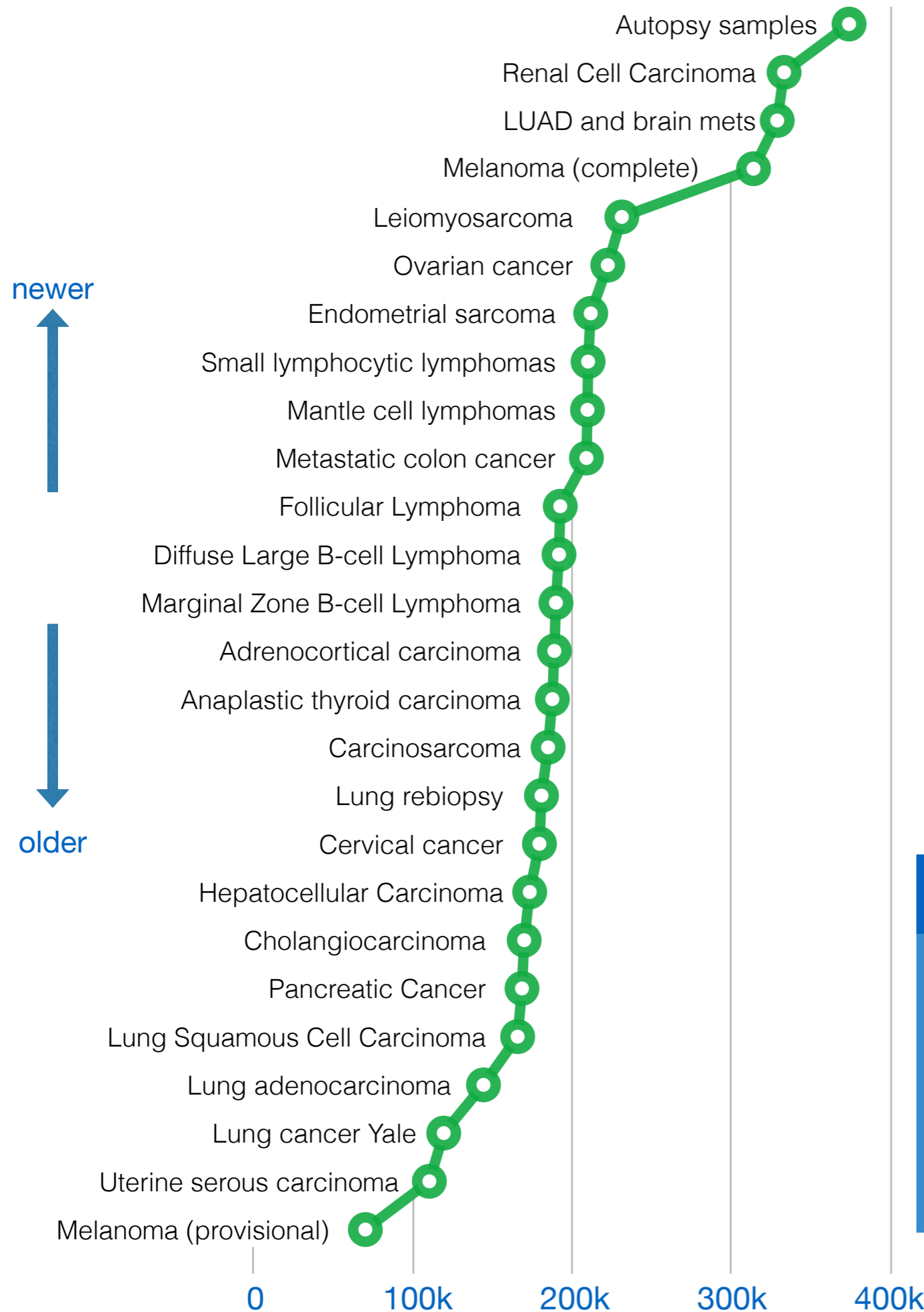


Nearly 400k somatic mutations have been uploaded to 26 Yale-Gilead projects.

- 1668 matched tumors
- 216 unmatched tumors
- additional 12 TCGA datasets
- and 2 projects with raw paths only

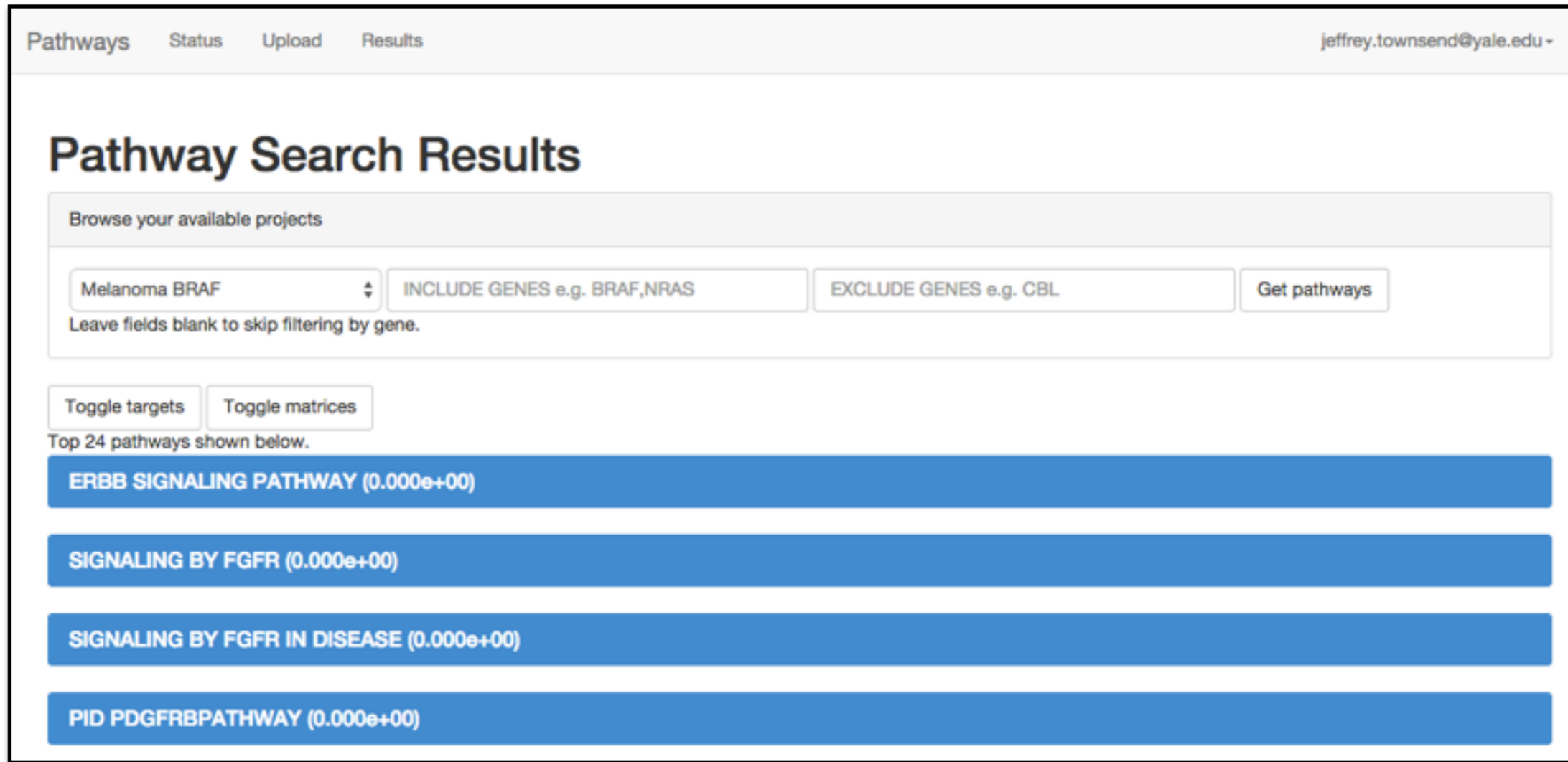
Largest projects by sample count

Project	Sample count	Somatic mutation count
Melanoma (complete)	354	82712
Metastatic colon cancer	177	16444
Autopsy samples	171	40821
Ovarian cancer	161	10715
Lung adenocarcinoma	108	24971
Lung squamous cell carcinoma	108	21461

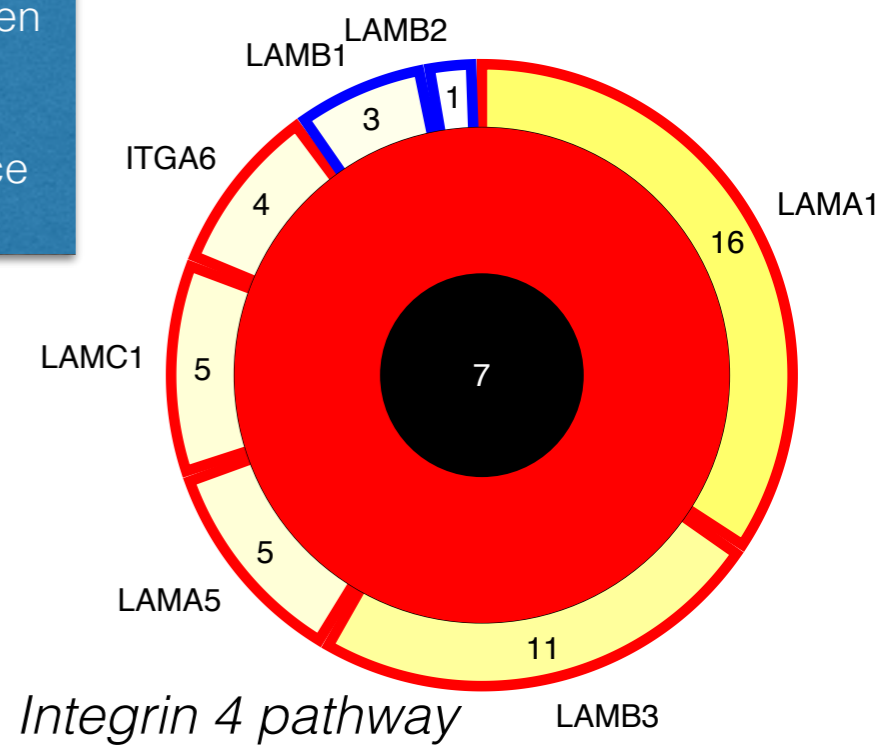
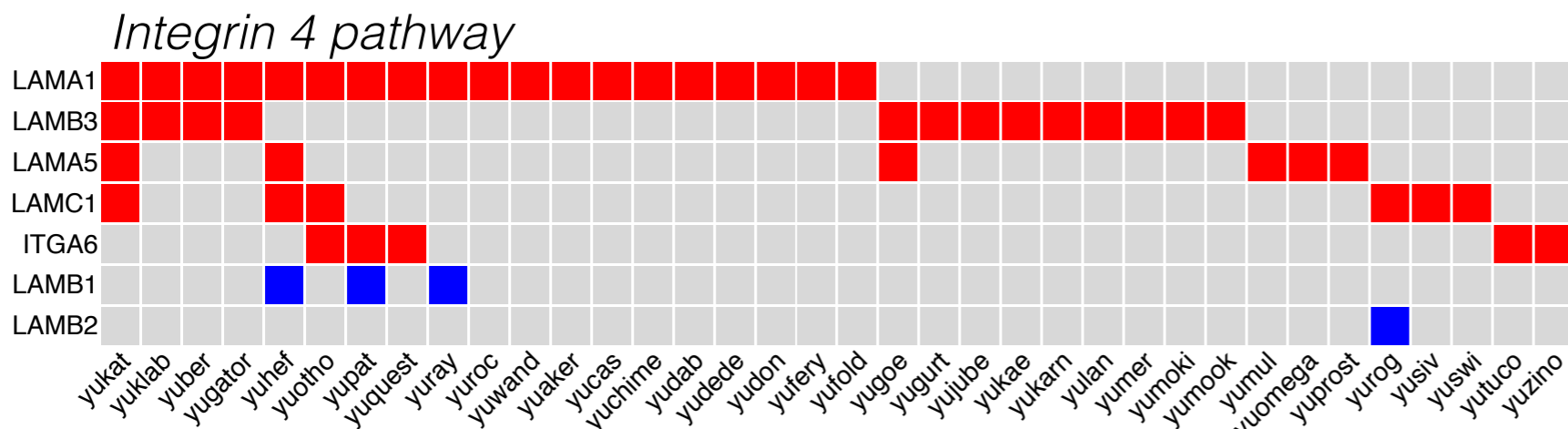


<https://cancerdb.eeb.yale.edu/progress.jsp>

Significantly altered pathways can clarify “common pathway” origins of cancers.



- Identifies pathways with a greater mutation burden than would be expected by chance, given pathway size.
- Users can browse plots for each pathway to aid identification of important:
 - matrix plots show all pathway mutations and can indicate exclusivity and co-occurrence
 - ‘target’ plots show genes scaled according to mutation frequencies

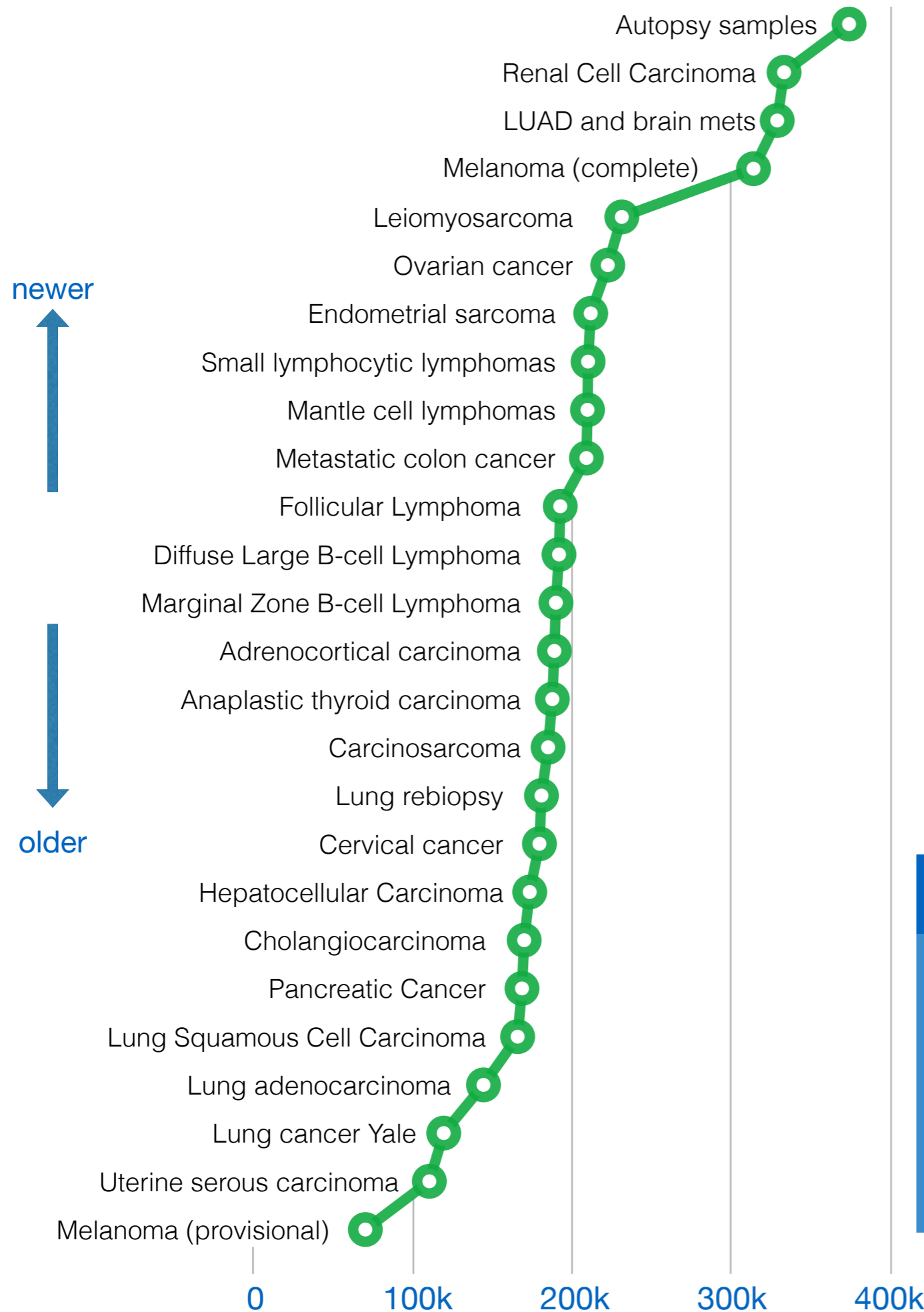


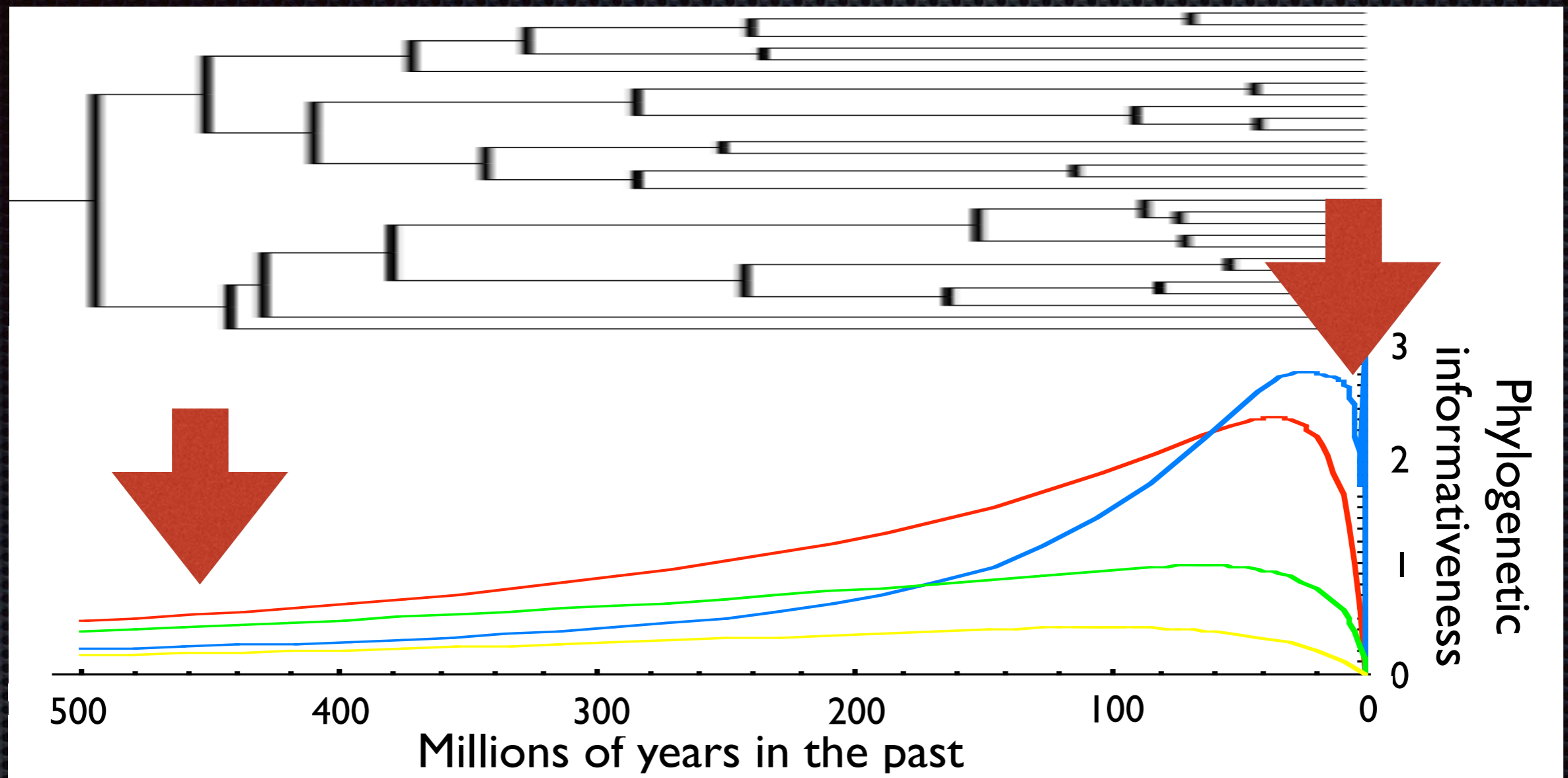
Nearly 400k somatic mutations have been uploaded to 26 Yale-Gilead projects.

- 1668 matched tumors
- 216 unmatched tumors
- additional 12 TCGA datasets
- and 2 projects with raw paths only

Largest projects by sample count

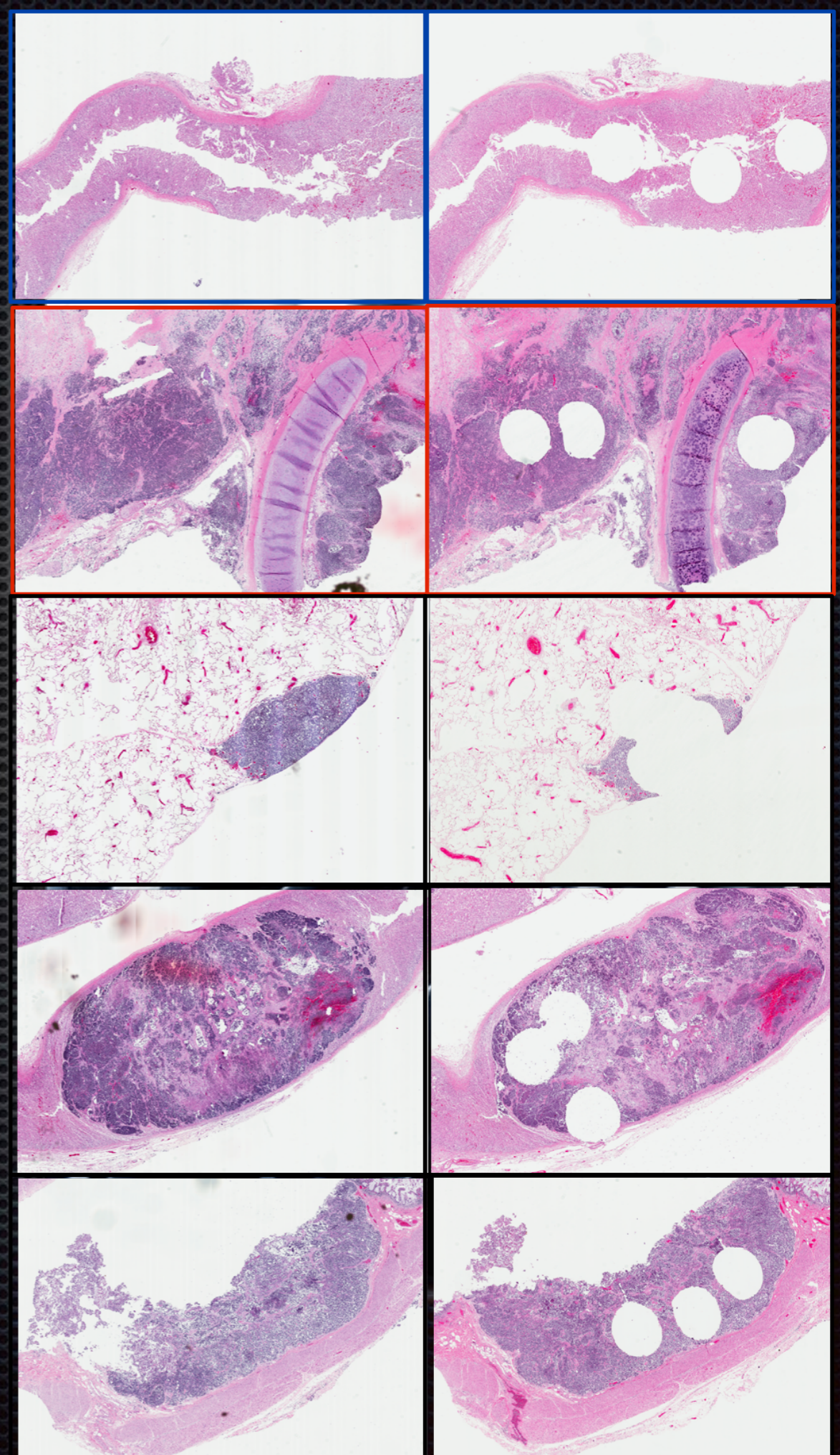
Project	Sample count	Somatic mutation count
Melanoma (complete)	354	82712
Metastatic colon cancer	177	16444
Autopsy samples	171	40821
Ovarian cancer	161	10715
Lung adenocarcinoma	108	24971
Lung squamous cell carcinoma	108	21461





More data is especially helpful not only for deep evolutionary questions, but also extremely recent ones

These evolutionary methodologies can be applied to cancer



Tumor tissues:

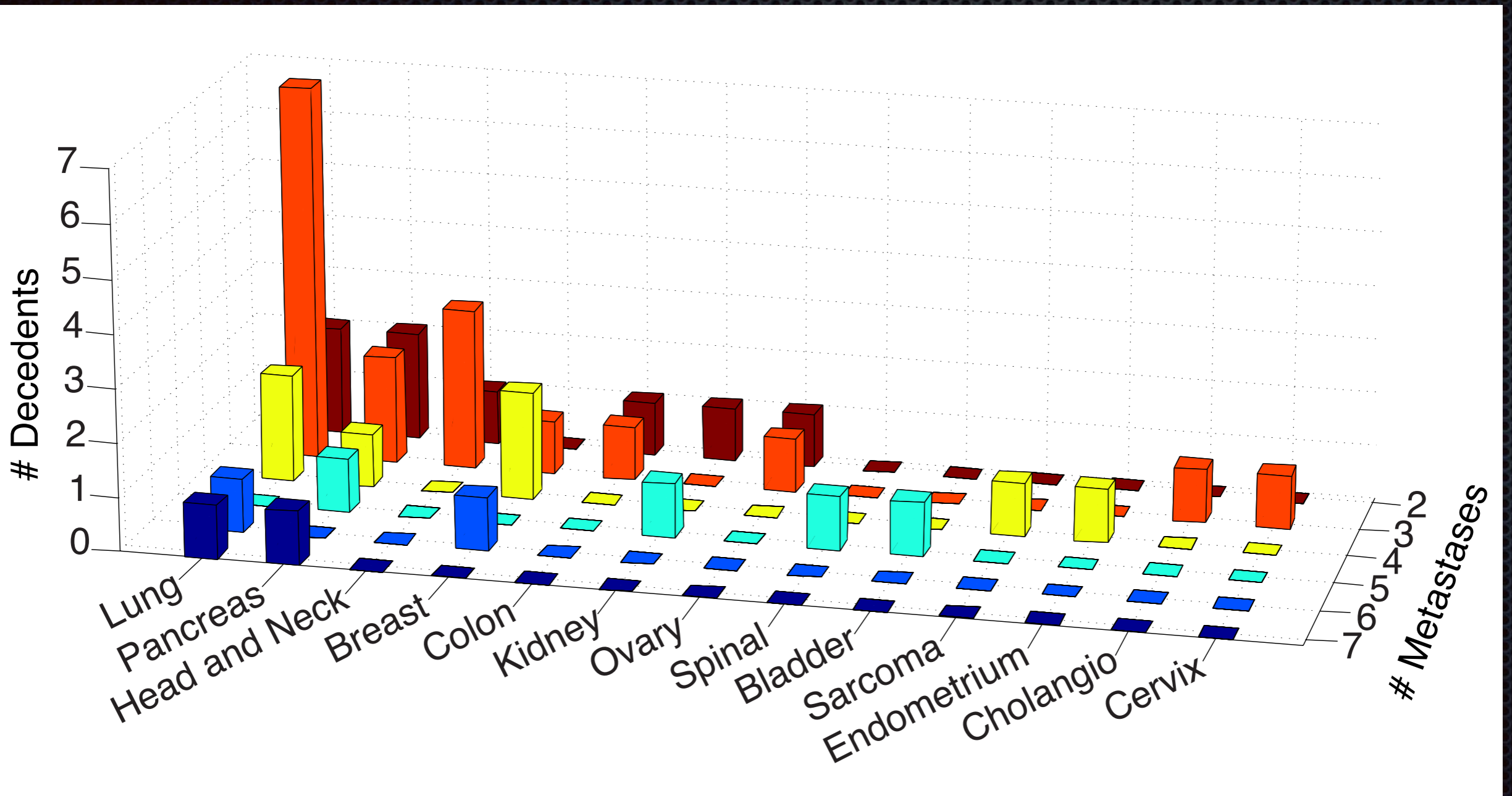
Normal (Adrenal)

Right Lung (Primary)

Left lung metastasis

Adrenal metastasis

Small bowel metastasis

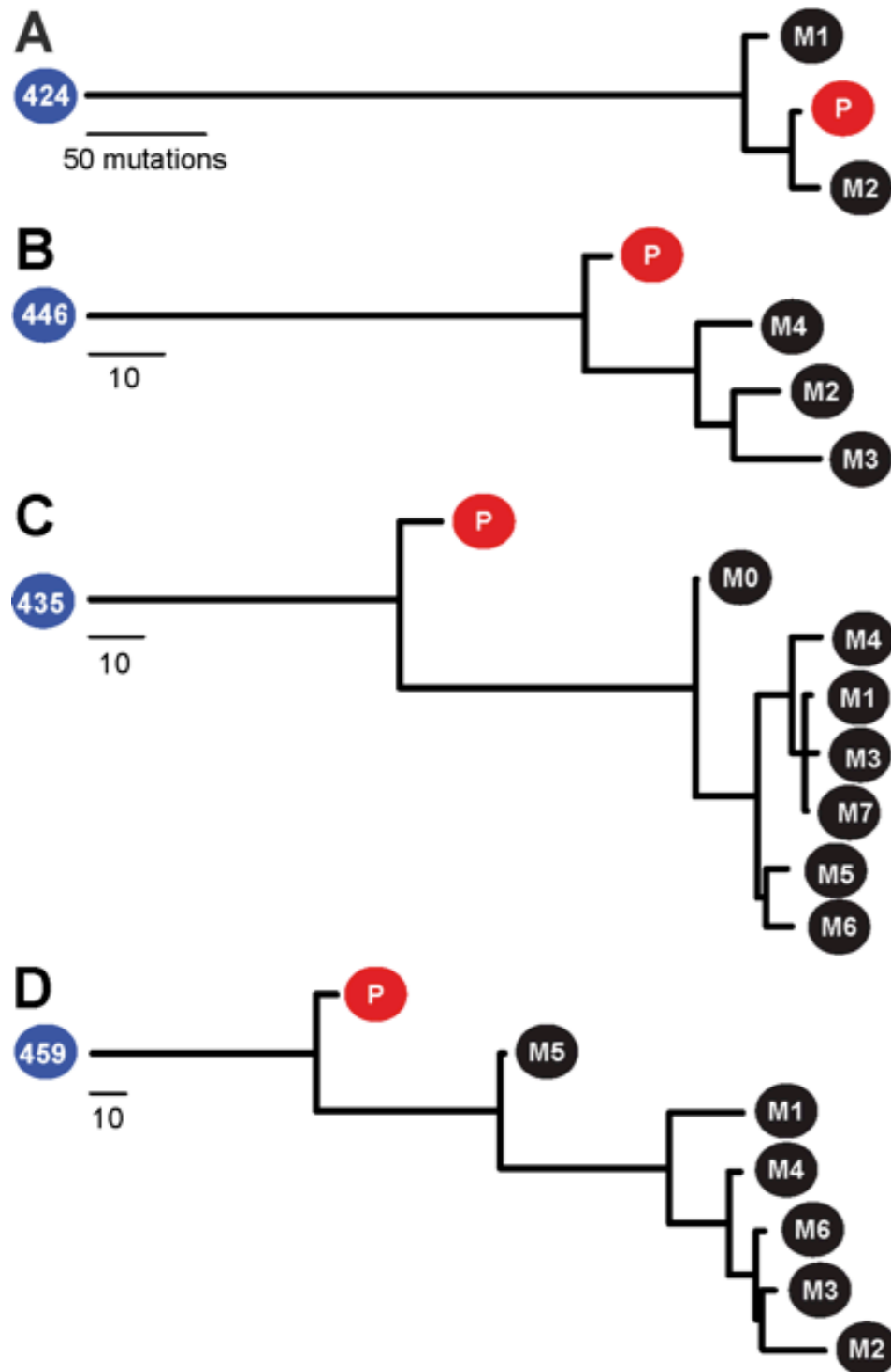


Tissue cores were extracted from normal tissue, primary tumors, and metastases

A phylogenetic analysis can address questions about the timing and relationships of metastases

- Are there single or multiple genetic origins of metastases within the primary tumor?
- How early do metastases genetically diverge from primary tumors?
- What is the chronology of cancerous tissue origination?
- Do driver mutations occur early or late in cancer?

A linear model for all cancer is not supported. Some genetic, epigenetic, or physiological disposition toward metastasis is.

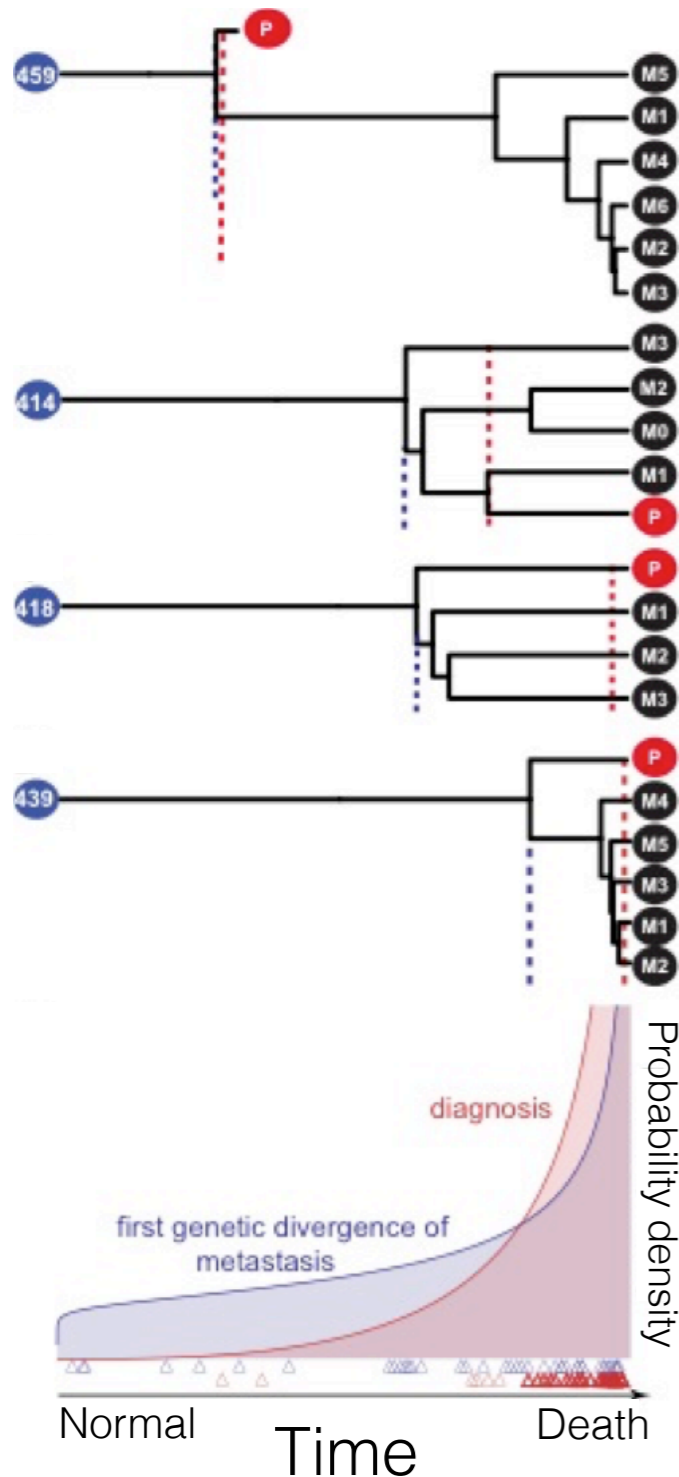


- Of 16 cancer phylogenies featuring a well-supported topological position of the primary tumor, 6 (38%) exhibited a most likely topology in which metastatic tumor lineages were not monophyletic and the primary tumor was not the outgroup to all metastases

- Integration over Bayesian posteriors for all 32 phylogenies yielded **45%** (CI 31%–56%)

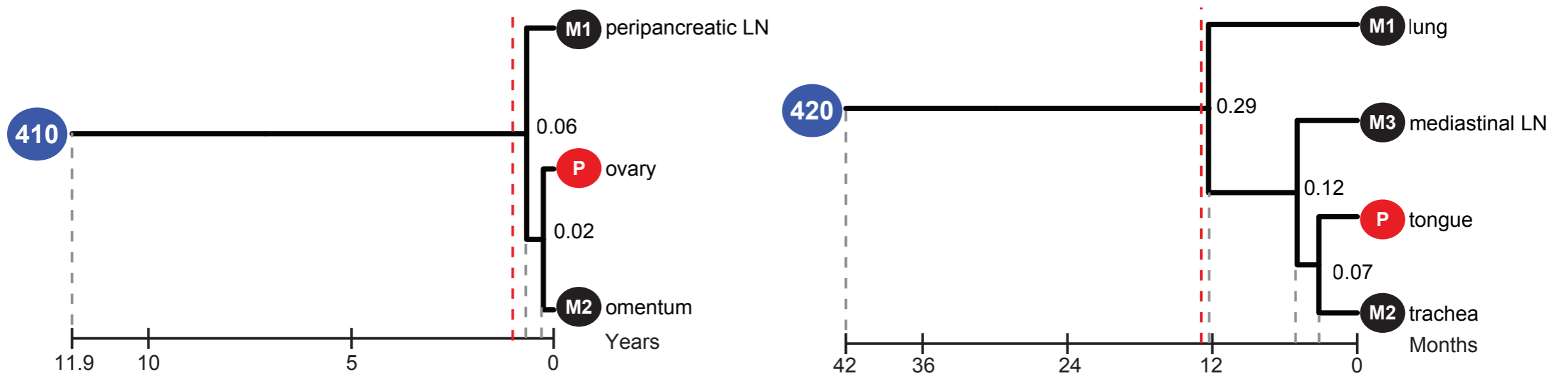
- This value, **significantly higher than the random expectation of 21%**, demonstrates that heritable genetic, epigenetic, or other lineage-specific events can contribute a proclivity within lineages toward metastasis of the primary tumor. However, the lineage-specific effect is not so strong as to universally lead to monophyletic metastases as predicted by the linear model ($P < 10^{-11}$)

Genetic divergence of metastatic lineages from primary tumors can occur early in tumor evolution

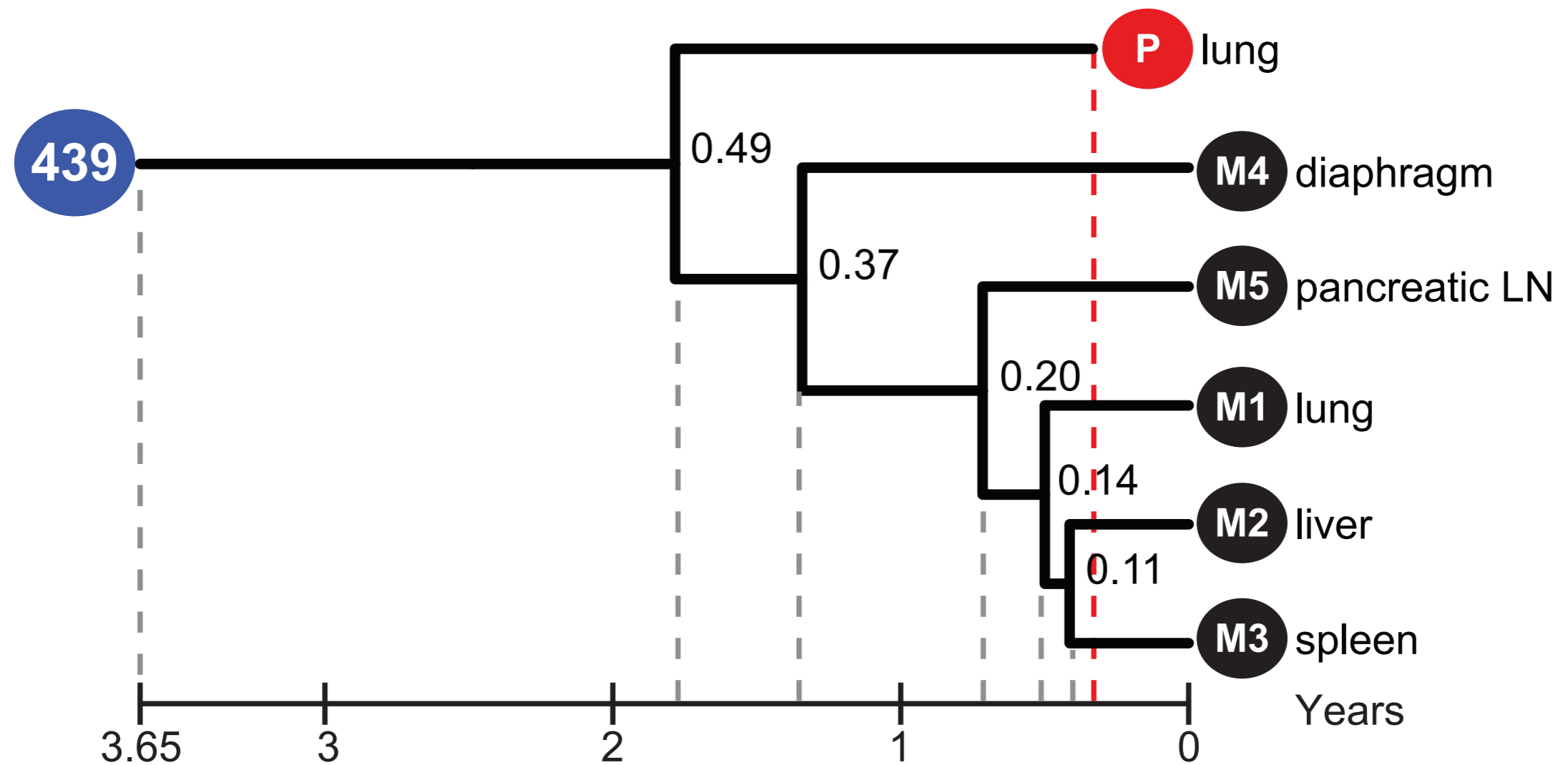


- In the cancer molecular evolutionary trees for 11 out of 40 subjects, the shared ancestral lineage of all tumors was shorter than the subsequent branch lengths leading to a metastatic tissue sampled at autopsy.
- We inferred cancer chronograms by applying a relaxed clock calibrated with the timings of diagnosis, biopsy, surgical resection, and autopsy, and parameterized by cell division times of primary tumor cells
- The first genetic divergence of metastasis usually predated diagnosis.

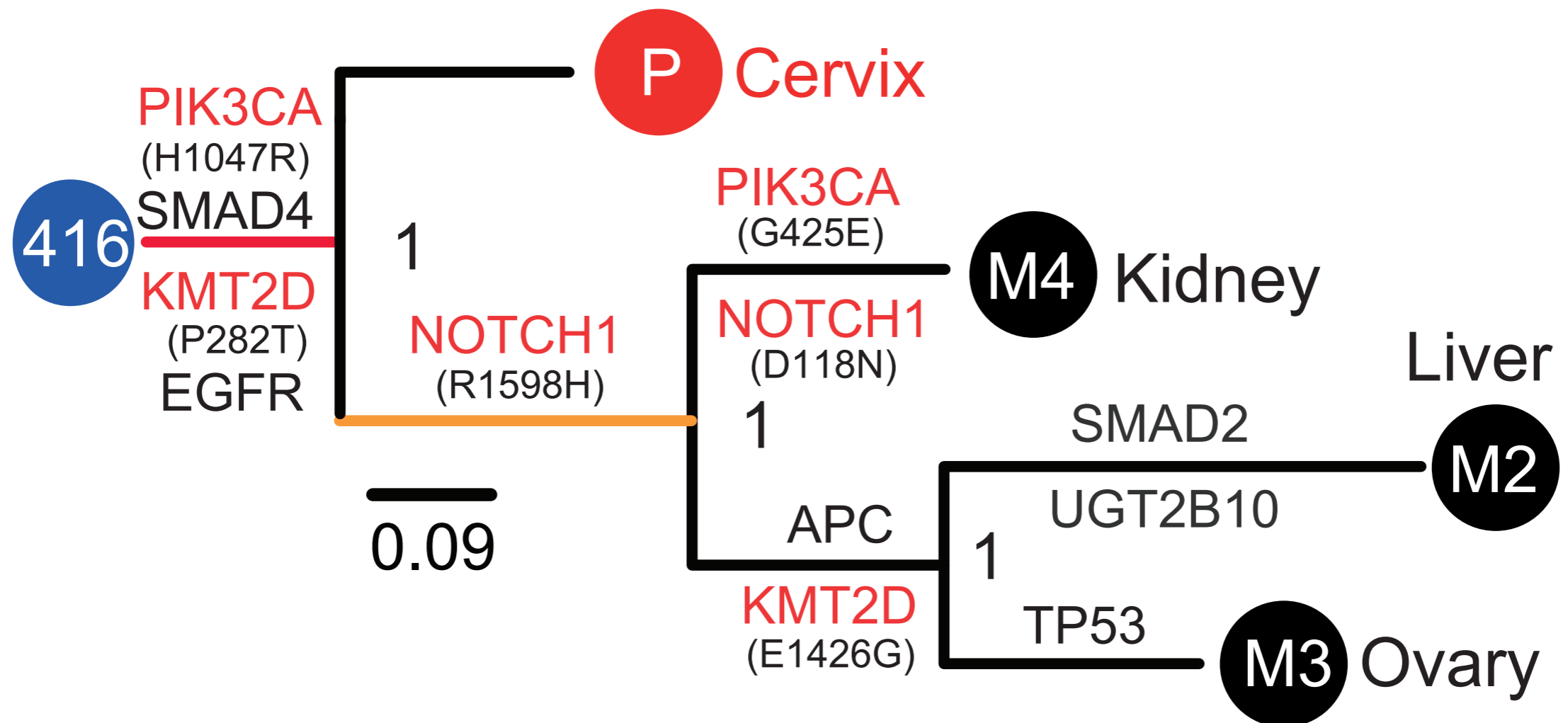
Metastatic lineages sampled sometimes arose subsequent to diagnosis



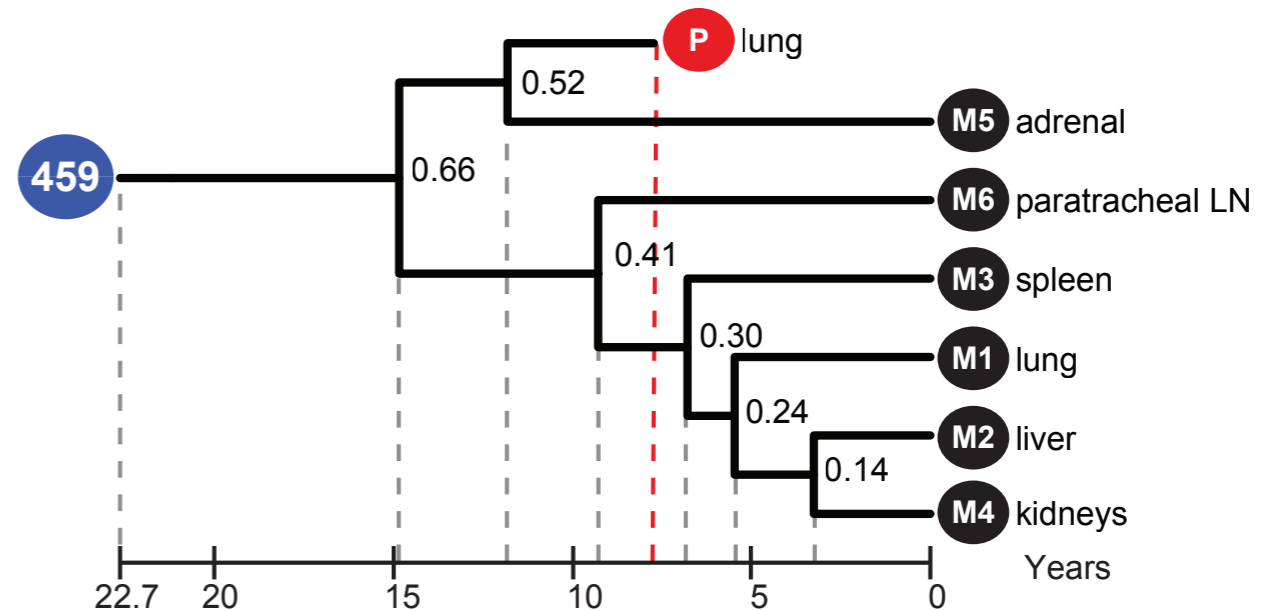
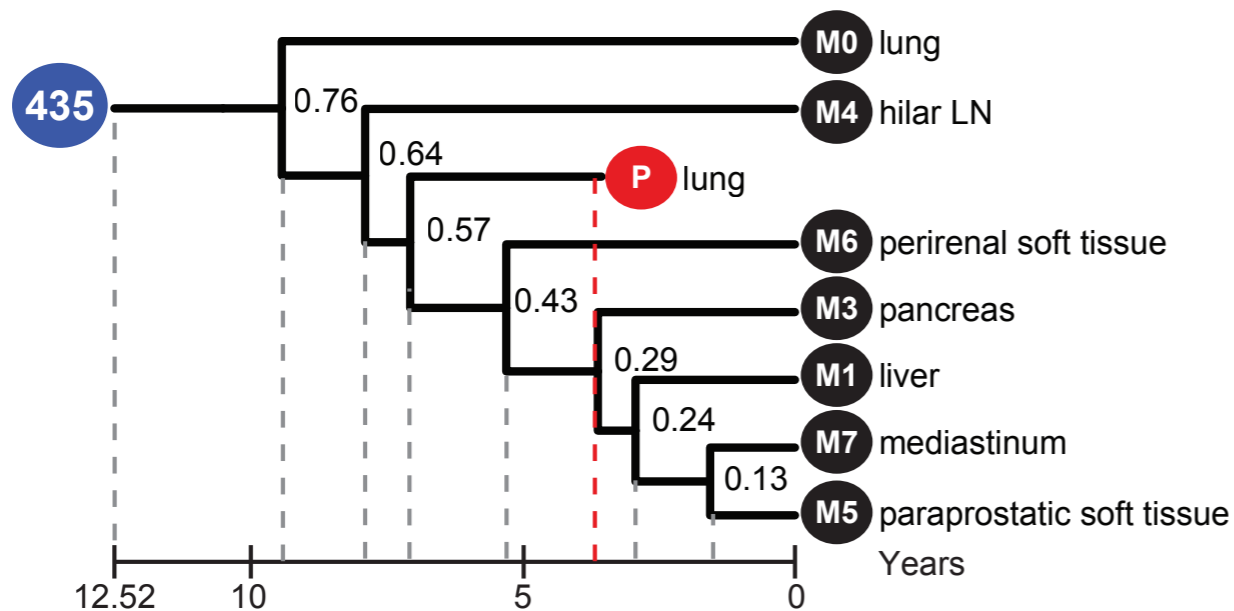
Metastatic lineages sampled often arose prior to diagnosis & resection

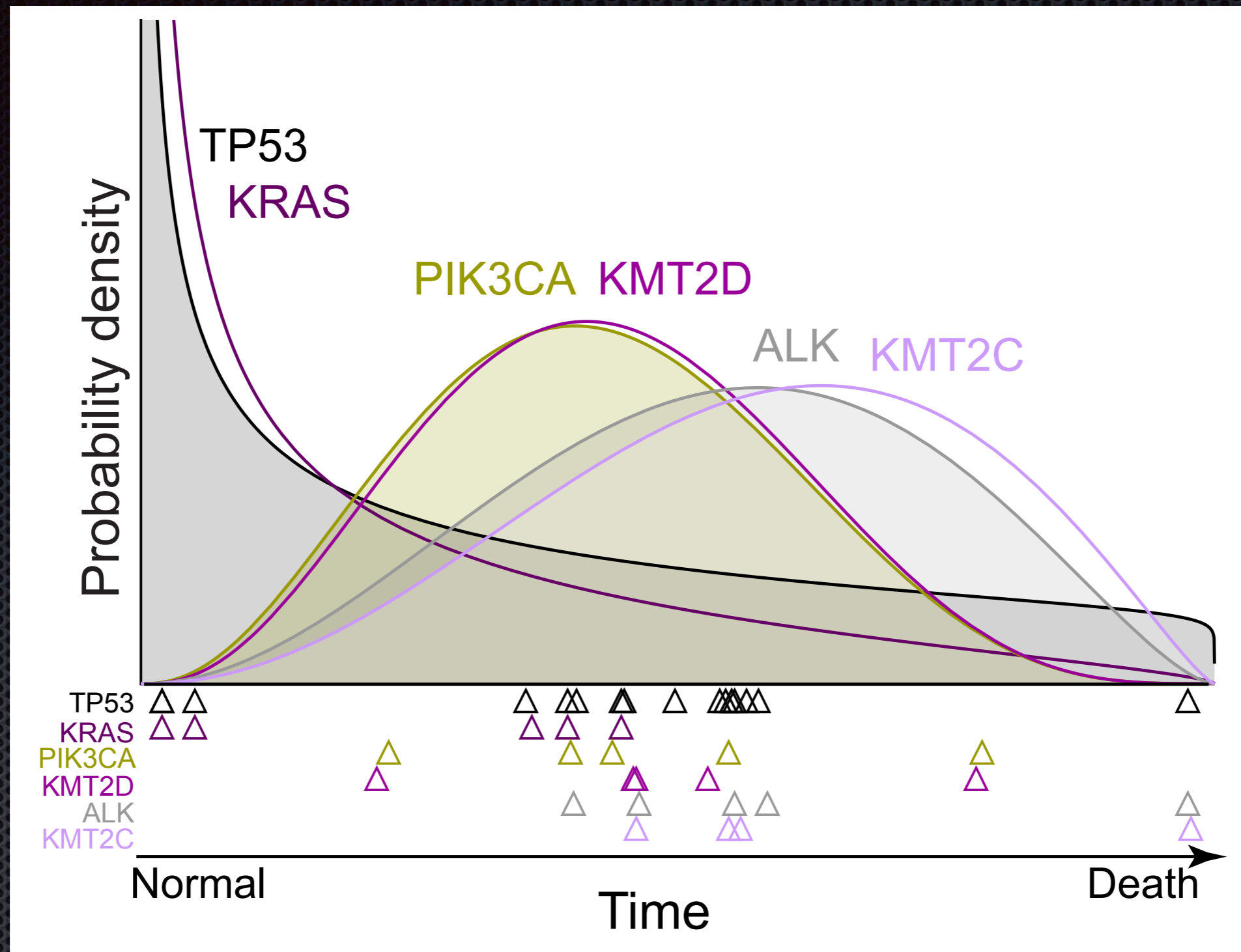


“Known driver mutations” could be mapped to the tumorigenesis lineage prior to all tumor tissue, or to the premetastatic lineage prior to all metastatic tissue.



Metastatic lineages sampled often arise prior to diagnosis & resection



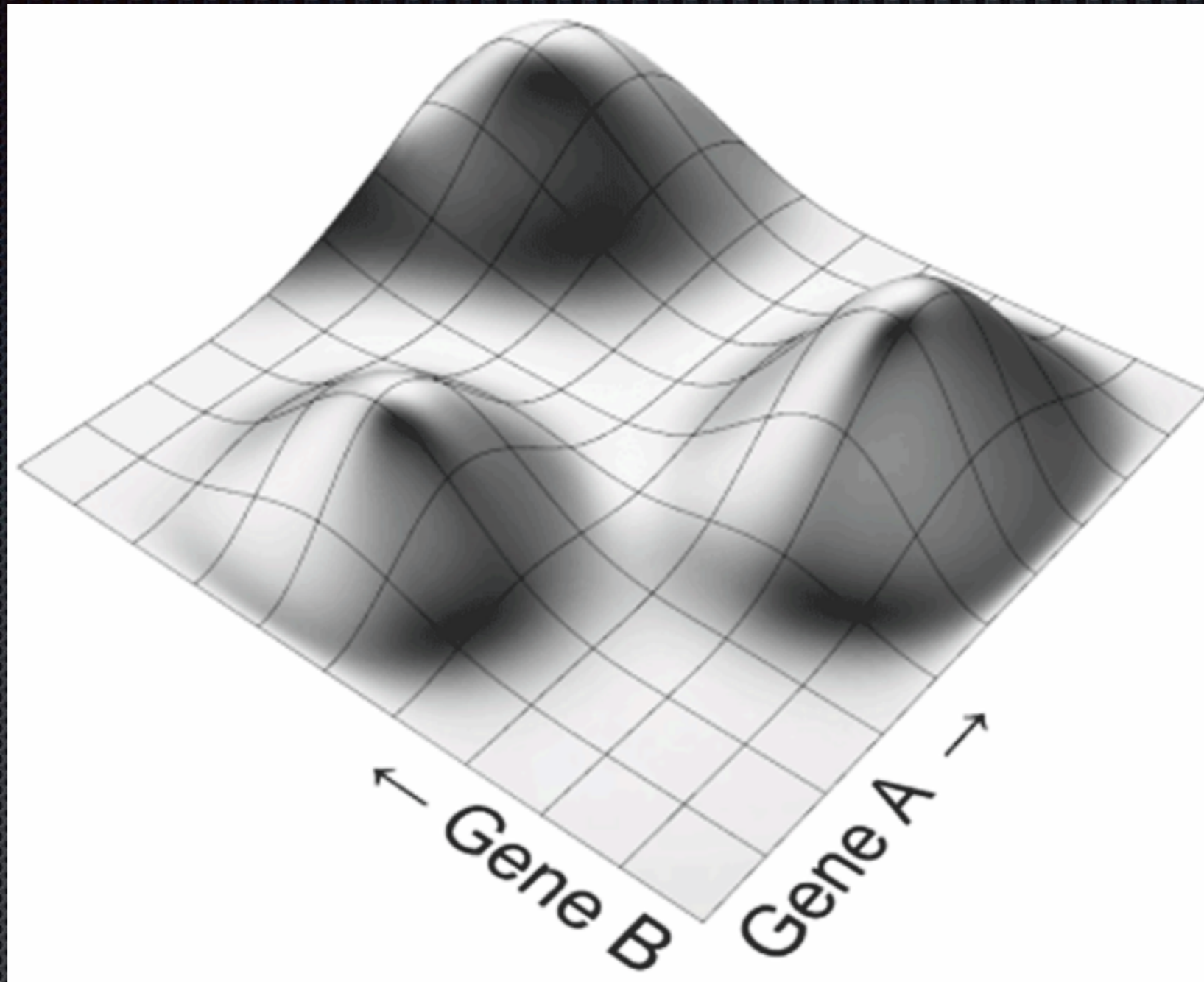


Integration over uncertainty in timing yielded probability distributions across patients for timing of driver gene mutation

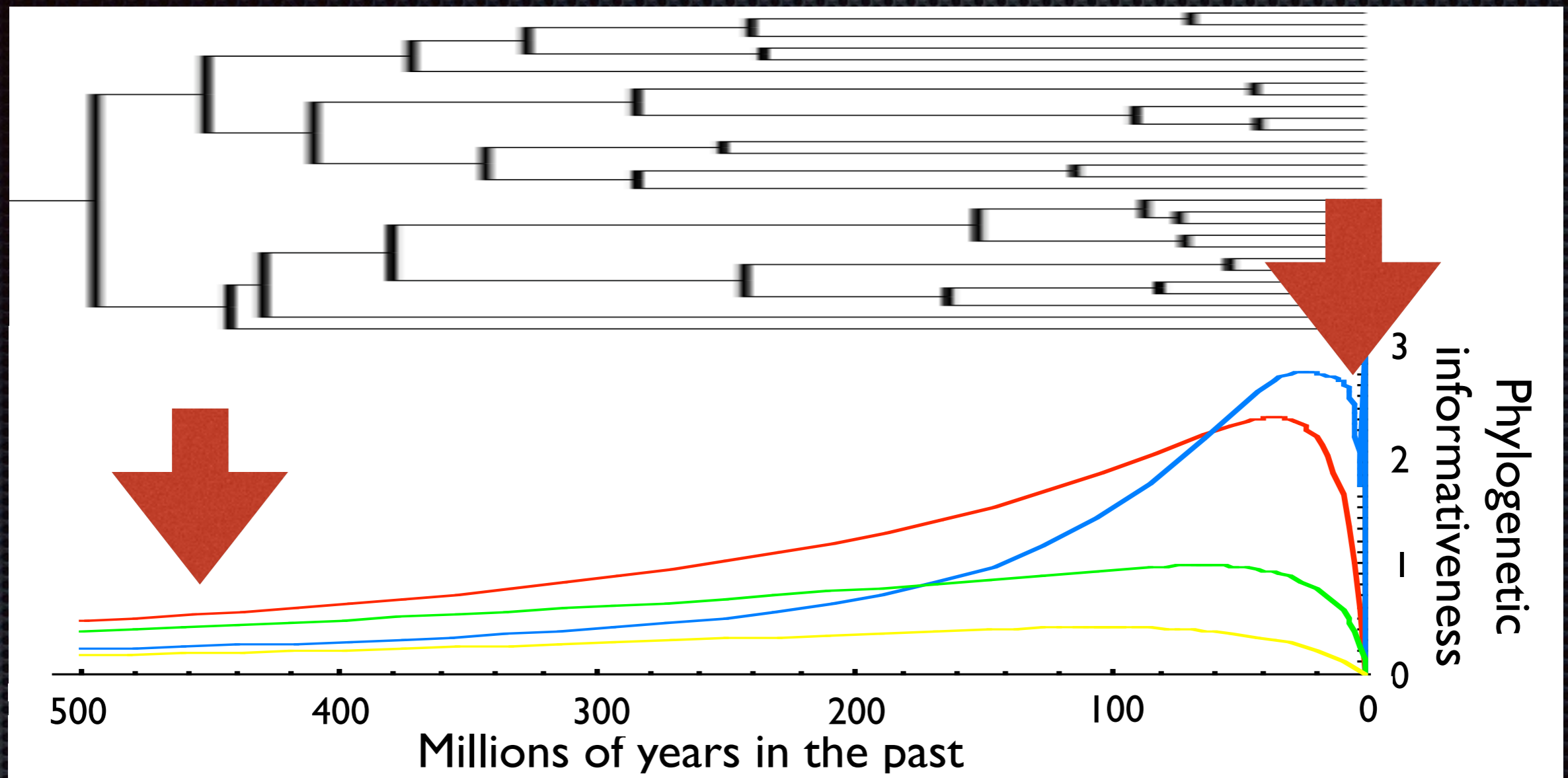
Of 104 genes mutated, 13 have known drugs that target them, with 3 of these being anti-neoplastic.

Table 5: Patient's Mutated Genes with Known Drug Targets				
CUBN	S1231N	Hydroxocobalamin	No	9.27
AASS	E537D	NADH, L-Glutamic acid	No	4.44
PTEN	I303S	1372540-25-4	Yes	4.03
MAN2A1	G1033D	Ghavamiol, swainsonine	No	3.63
TEP1	A1974T	GRN163L	No	3.23
PAPSS1	R445W	ADP, Glycerol	No	2.82
NRP2	R421W	Daunorubicin	No	2.02
MAP2K2	C125S	Trametinib, Mek162, Selumetinib	Yes	2.02
ARAF	G322S	LGX818, Sorafenib, Regorafenib, XL281	Yes	1.61
ANXA11	P83S	Bevacizumab	No	1.21
CASK	P239S	Formic acid	No	1.21
CXCL12	M1T	Tinzaparin (binder)	No	0.81
G6PD	D194E	Doxorubicin, Aspirin, Chloroquine,	No	0.40

Cancer treatment by targeted drug therapies is growing increasingly complex: an example of a patient with melanoma



Analyses will increasingly have to consider the fitness landscape of cancer mutations across multiple genes



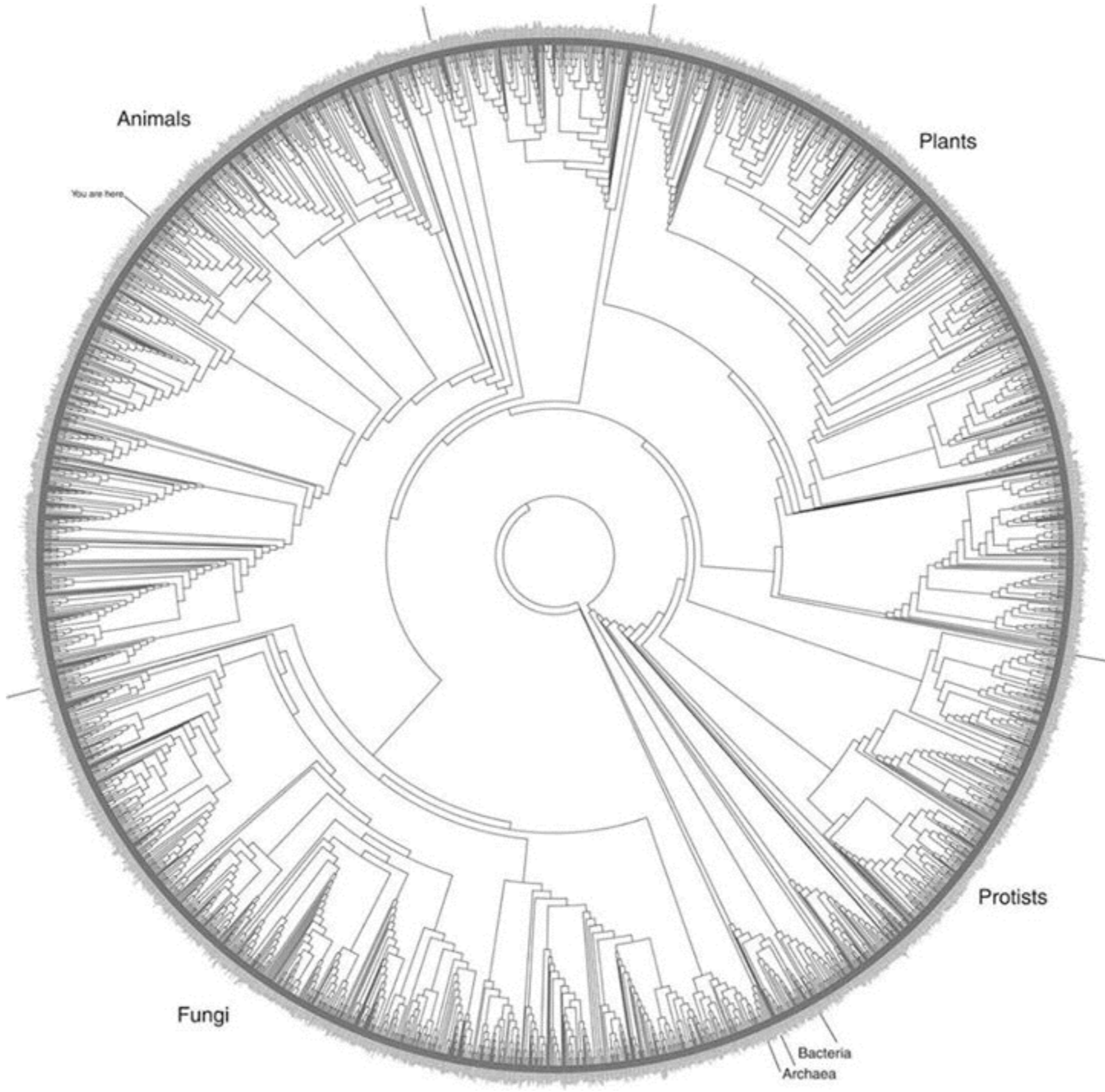
More data is especially helpful not only for deep evolutionary questions, but also extremely recent ones

Two examples of interdisciplinary collaboration using the techniques of evolutionary biology facilitated by next-generation sequencing data



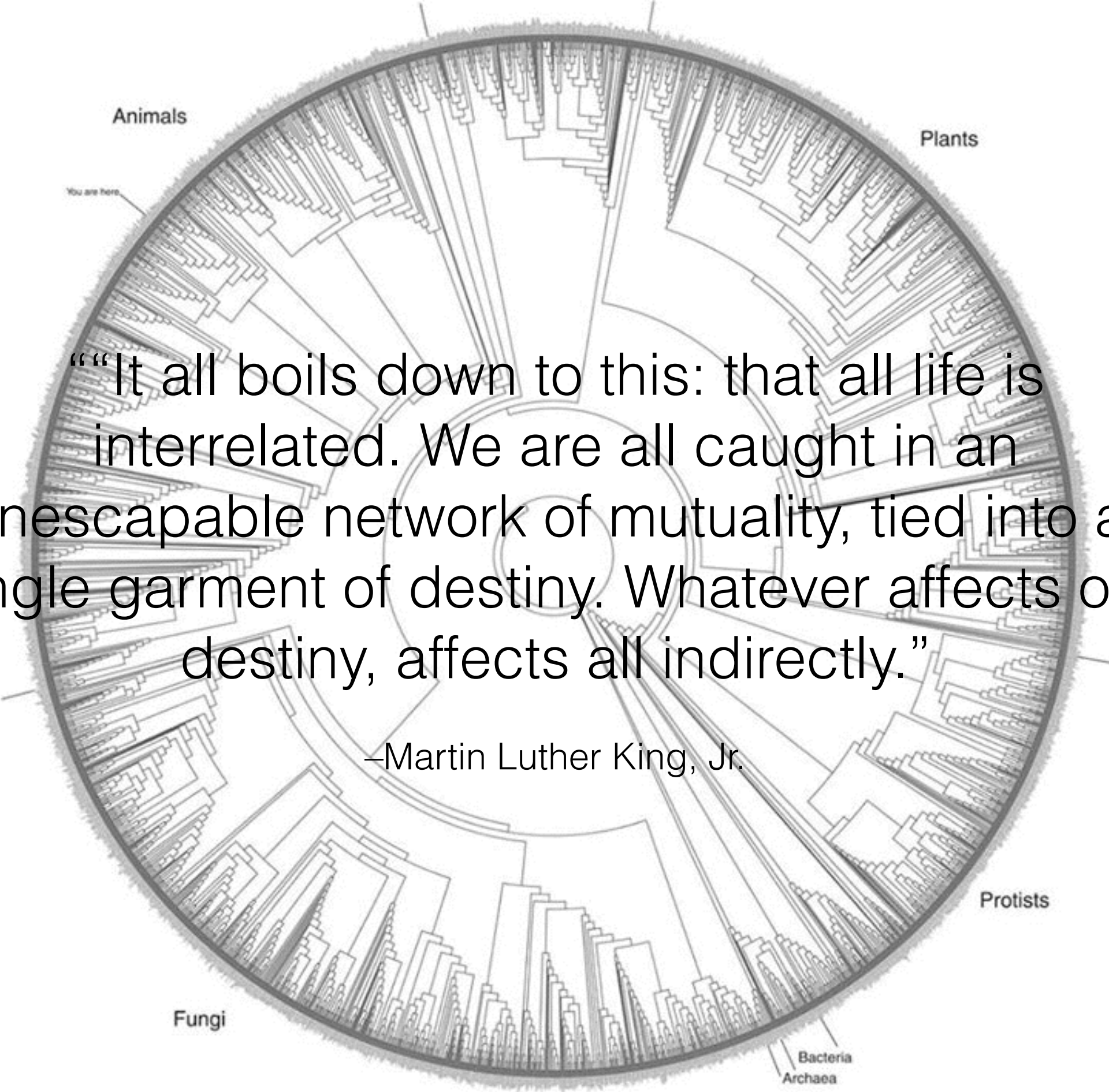
Notsew Orm
Sands
Foundation





““It all boils down to this: that all life is interrelated. We are all caught in an inescapable network of mutuality, tied into a single garment of destiny. Whatever affects one destiny, affects all indirectly.”

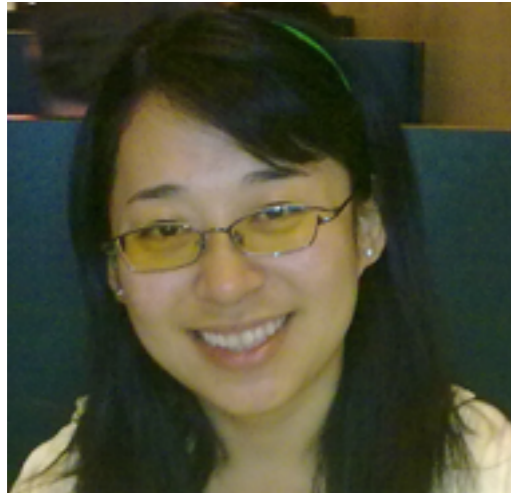
–Martin Luther King, Jr.



“It all boils down to this: that all life is interrelated. We are all caught in an inescapable network of mutuality, tied into a single garment of destiny. Whatever affects one destiny, affects all indirectly.”

—Martin Luther King, Jr.

Acknowledgements—Cancer Research



Ziming Zhao



Stephen Gaffney



Atila Iamarino

Collaborators: Brian Xiao & Richard Lifton
Yalai Bai & David Rimm
Joseph Schlessinger



Funding:

Gilead Sciences

