What ancient viral sequences* can and will teach us about virus evolution

Terry Jones
tcj25@cam.ac.uk

Zoology, University of Cambridge
Institut für Virologie, Charité

* With a little guidance from Charles Darwin

Part I: A zoological / geological perspective on studying long-term virus evolution

- We have been trying to understand long-term virus evolution with no ancient sequences!
- Imagine studying zoology with no fossils.
  - The fact of evolution, combined with only modern evidence (and, sometimes, endogenous viruses), requires adding possibly arbitrary or implicit assumptions. The outcome is likely inaccurate!
  - Things will now begin to change, but only slowly.

“I look at the geological record as a history of the world imperfectly kept and written in a changing dialect. Of this history we possess the last volume alone, relating only to two or three countries. Of this volume, only here and there a short chapter has been preserved, and of each page, only here and there a few lines. Each word of the slowly-changing language, more or less different in the successive chapters, may represent the forms of life, which are entombed in our consecutive formations, and which falsely appear to have been abruptly introduced.”

Charles Darwin

On the Origin of Species
Chapter 10: On the Imperfection of the Geological Record

Ancient viral sequences to date

- Previous ancient sequences have typically come from permafrost, glaciers, mummified organs, etc.
- These are rare finds, and we have just a handful of them. Most are just one sequence, not definitely confirmed to be ancient, and not very old.
- Anellovirus, Avipoxvirus, Barley stripe mosaic virus, Hepatitis B, Simian T-cell leukemia virus, Human T-cell leukemia virus, Influenza A, Monkeypox, Papillomavirus, Human Parvovirus B19, Pithovirus, Smallpox, Tomato mosaic tobamovirus, Caribou feces associated virus, and Northwest Territories cripavirus.

What could make a virus more likely to be preserved and found?

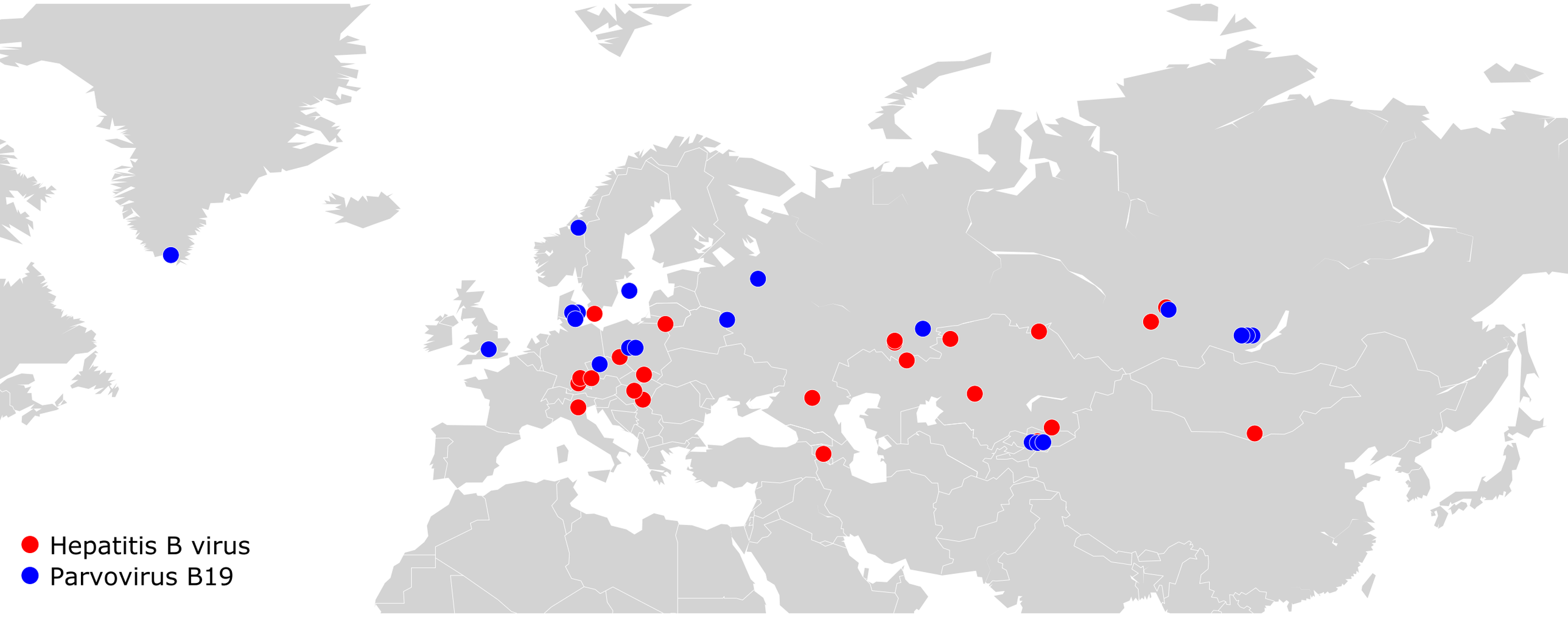
- Double-stranded DNA.
- Virion stability.
- People have to die with the virus. So chronic infections or lethal viruses.
- High viral titres.
- Viraemic (assuming organs weren't preserved).
- And there are non-viral factors, such as climate.

Part II - Ancient virus discovery & analysis

304 ancient human samples were screened from Eurasia to Western Europe, aged up to ~7000 years.

Individuals positive for viruses

- 25 with reads matching Hepatitis B virus.
- 20 with reads matching human parvovirus B19.

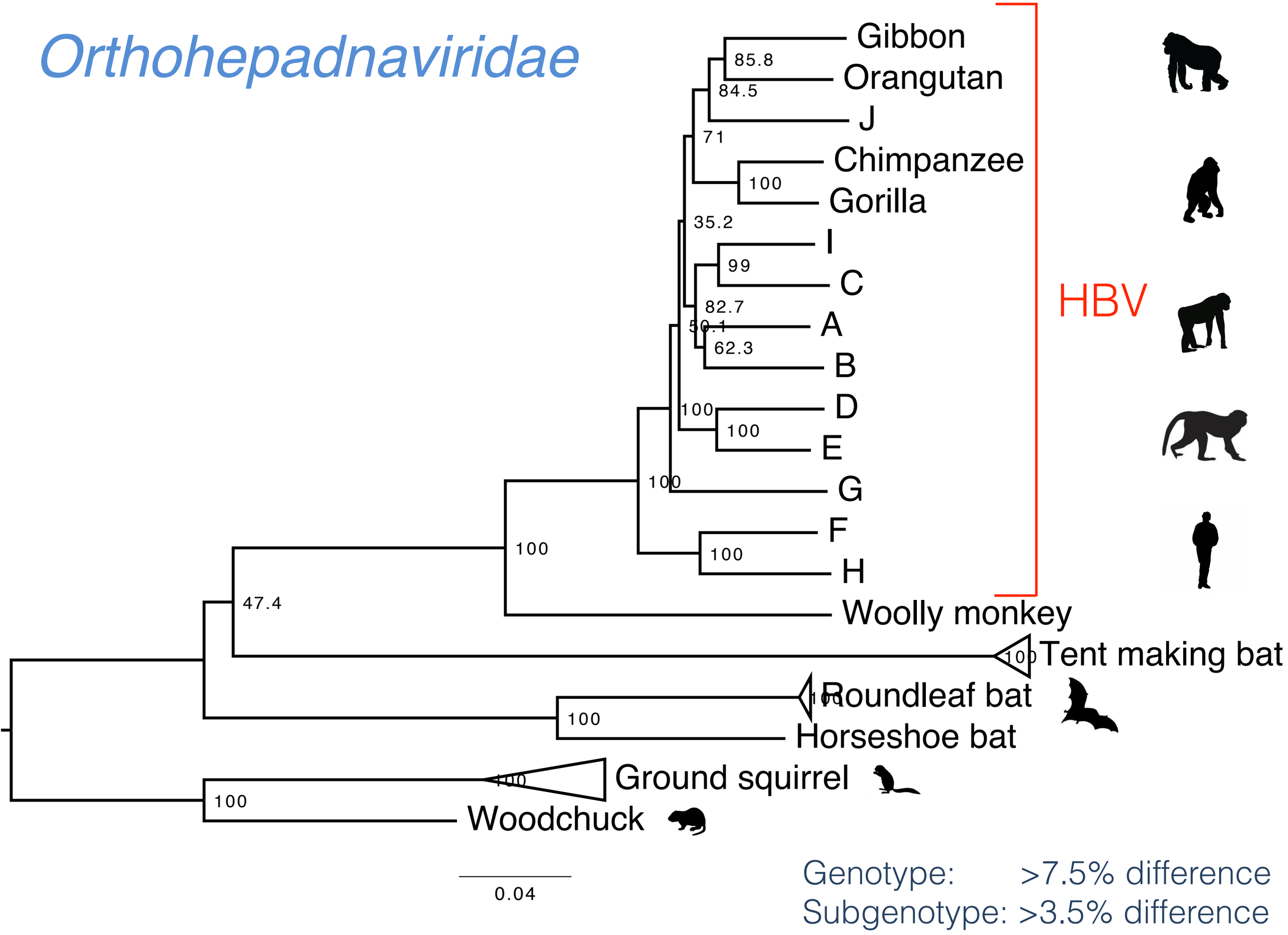


Hepatitis B virus

Hepatitis B Virus (HBV)

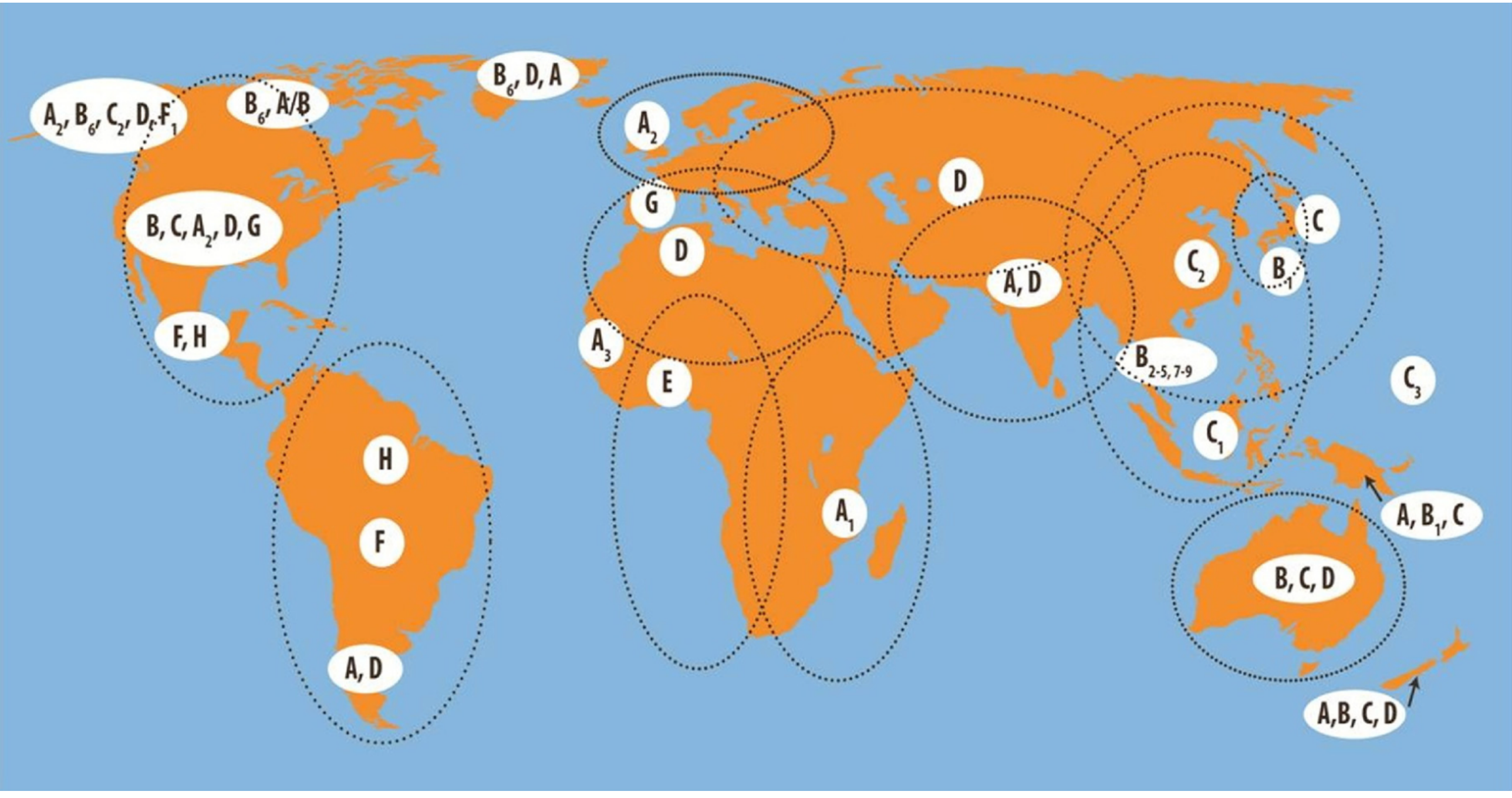
- Partially double stranded DNA virus, ~3200bp.
- Causes liver inflammation and liver cancer.
- Transmitted perinatally or horizontally via blood or genital fluids.
- Chronic infections result in high virus titres for years or decades.
- 257 million chronic carriers, 887,000 deaths annually.

Orthohepadnaviridae



Genotype: >7.5% difference
Subgenotype: >3.5% difference

Modern HBV genotype distribution

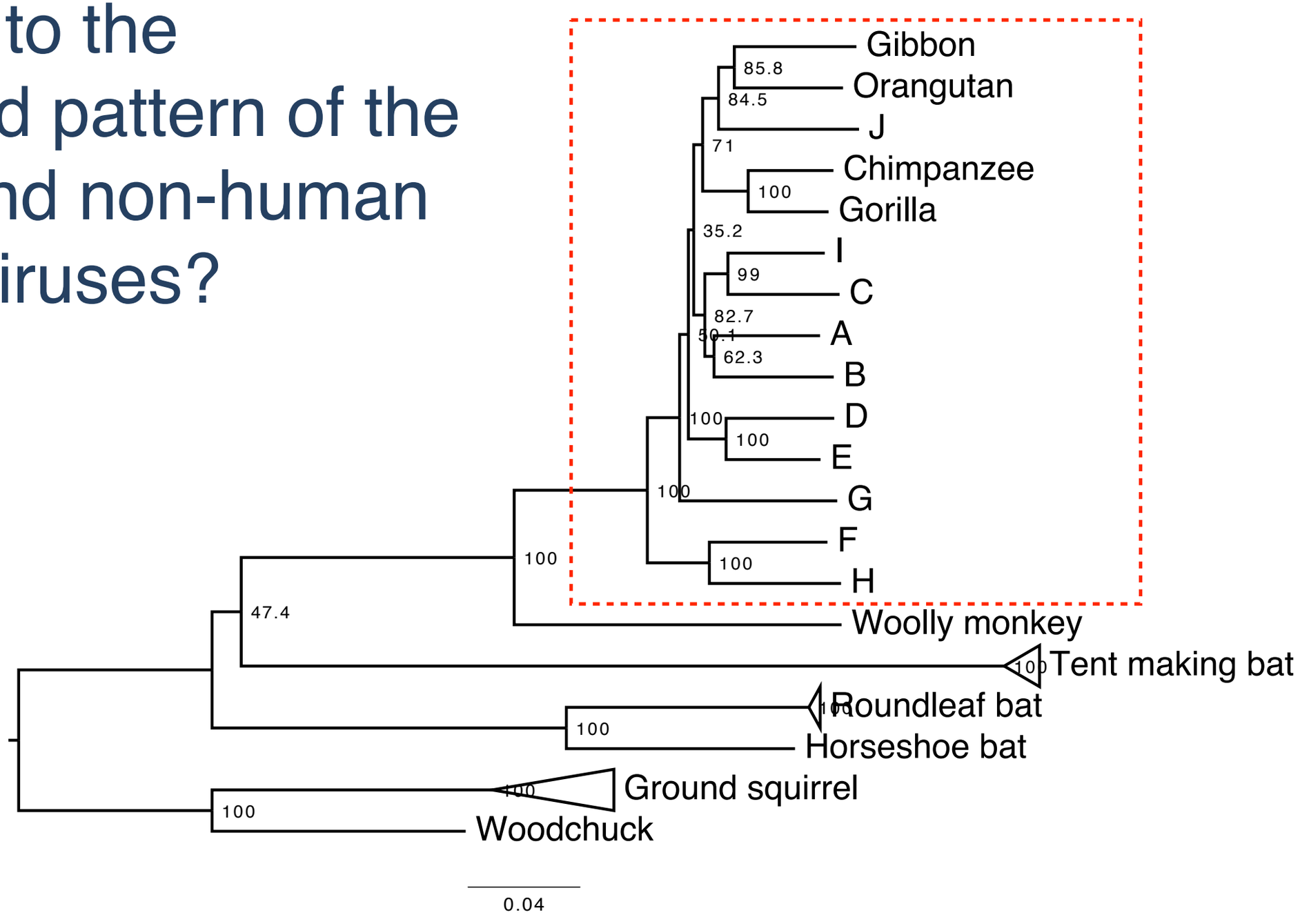


Genotype: >7.5% difference
Subgenotype: >3.5% difference

Locarnini et al., 2013

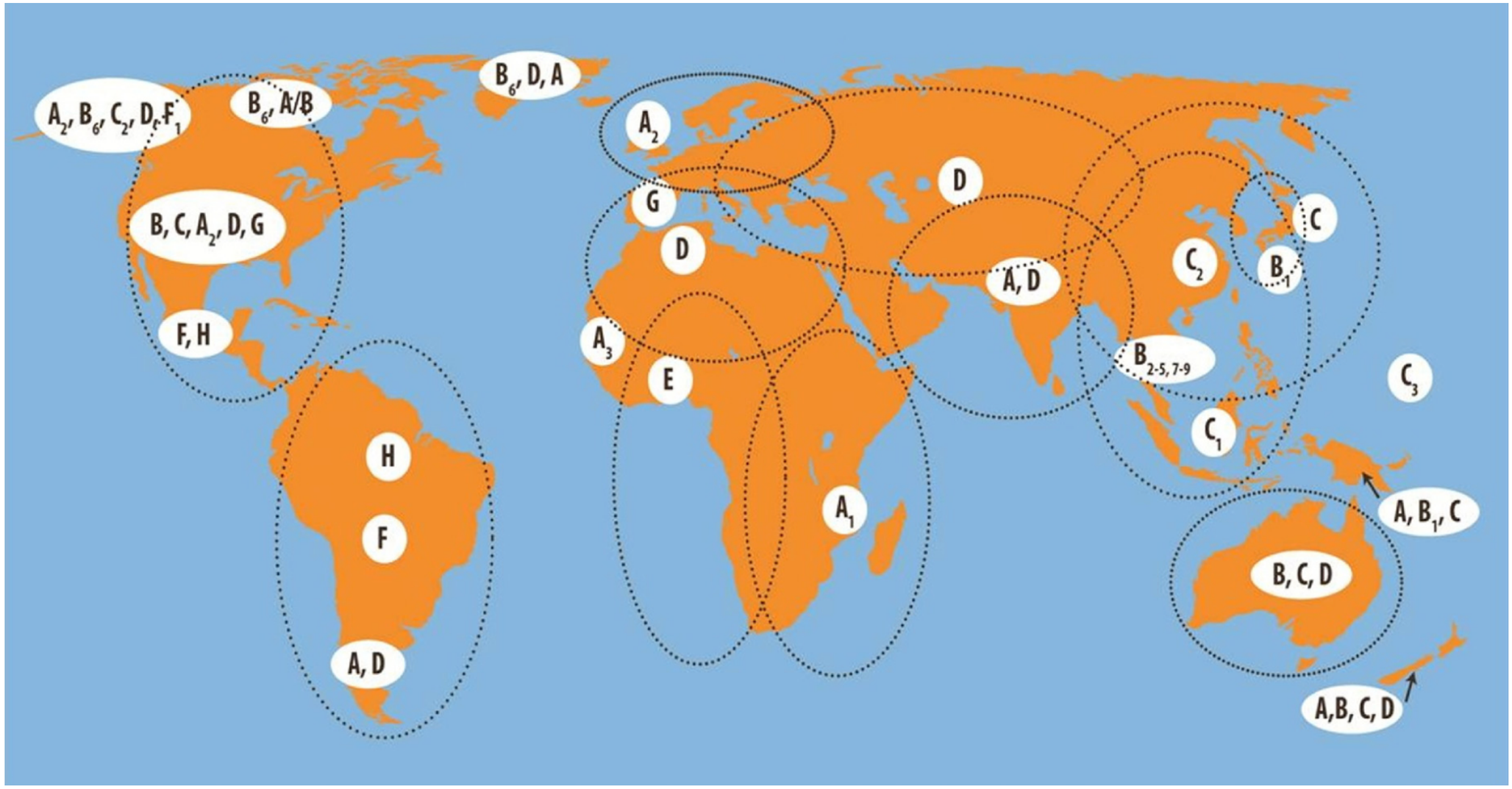
Open questions in HBV evolution

What led to the intermixed pattern of the human and non-human primate viruses?



Open questions in HBV evolution

What led to the current genotype distribution?



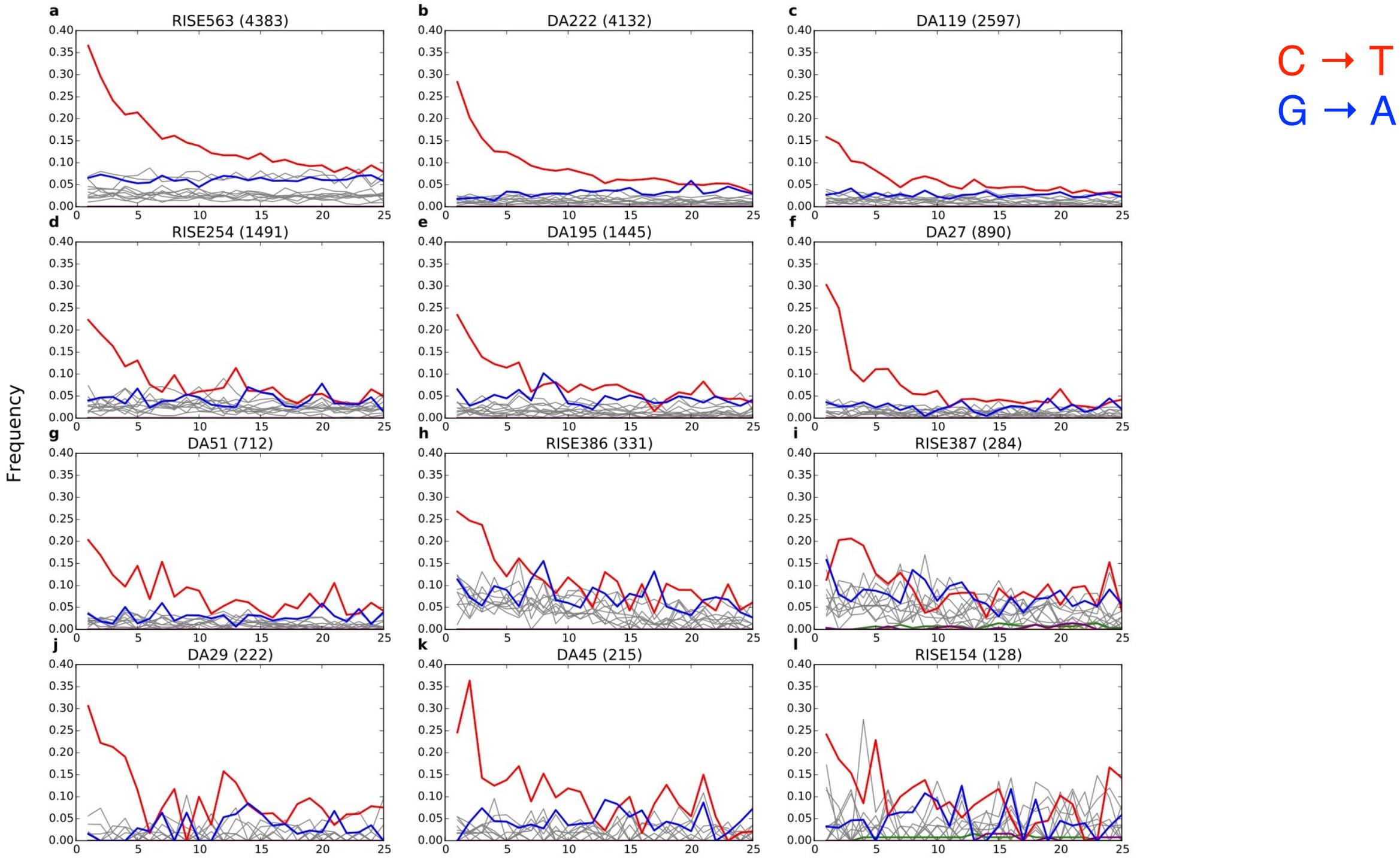
Open questions in HBV evolution

- When did HBV get into humans, or has it always been there?
- How did HBV come to be diversified in humans?
- What is the substitution rate?
- And other mysteries…

Authentic, ancient, and exogenous?

- DNA damage patterns.
- Basal phylogenetic position of many sequences.
- Sequences match multiple genotypes.
- HBV is blood-borne, so contamination unlikely.
- Coverage depth consistent with single- versus double-stranded regions of HBV genome.
- Distinct human/virus coverage levels.
- No reads with mixed human/virus DNA → exogenous virus.

DNA damage patterns in HBV reads

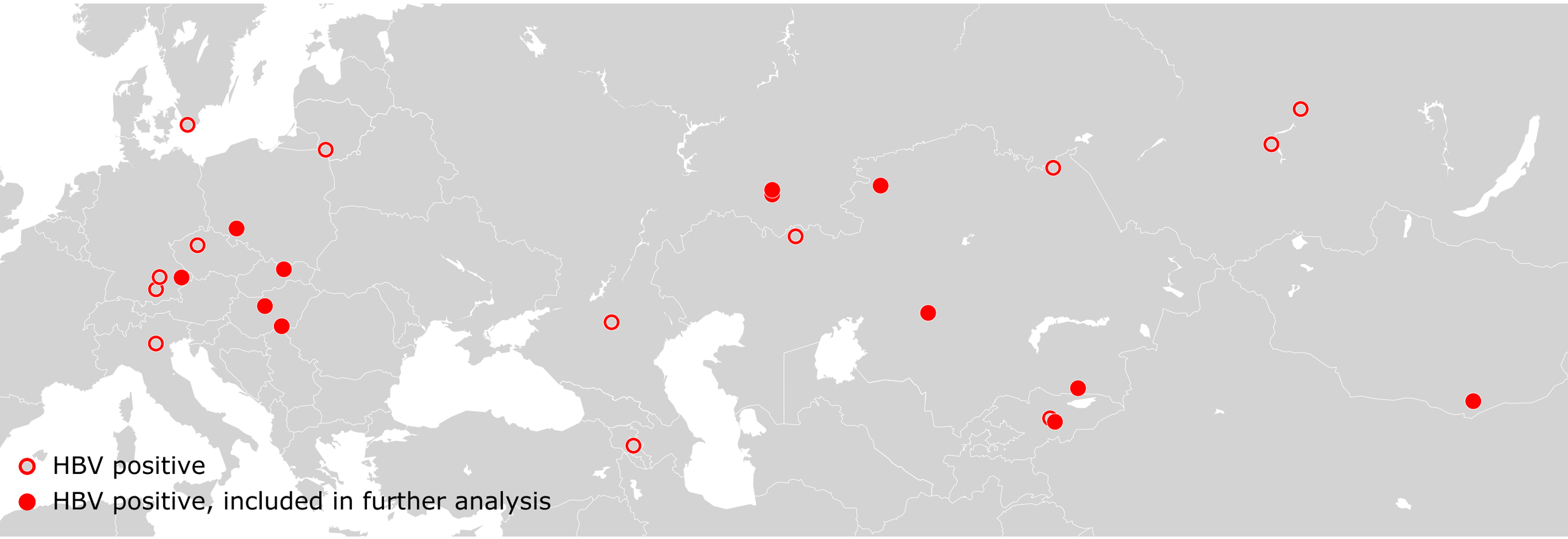


C → T

G → A

25 individuals had reads matching HBV

12 sequences, ~800 to 4500 years old, with >50% coverage were used for phylogenetic analysis

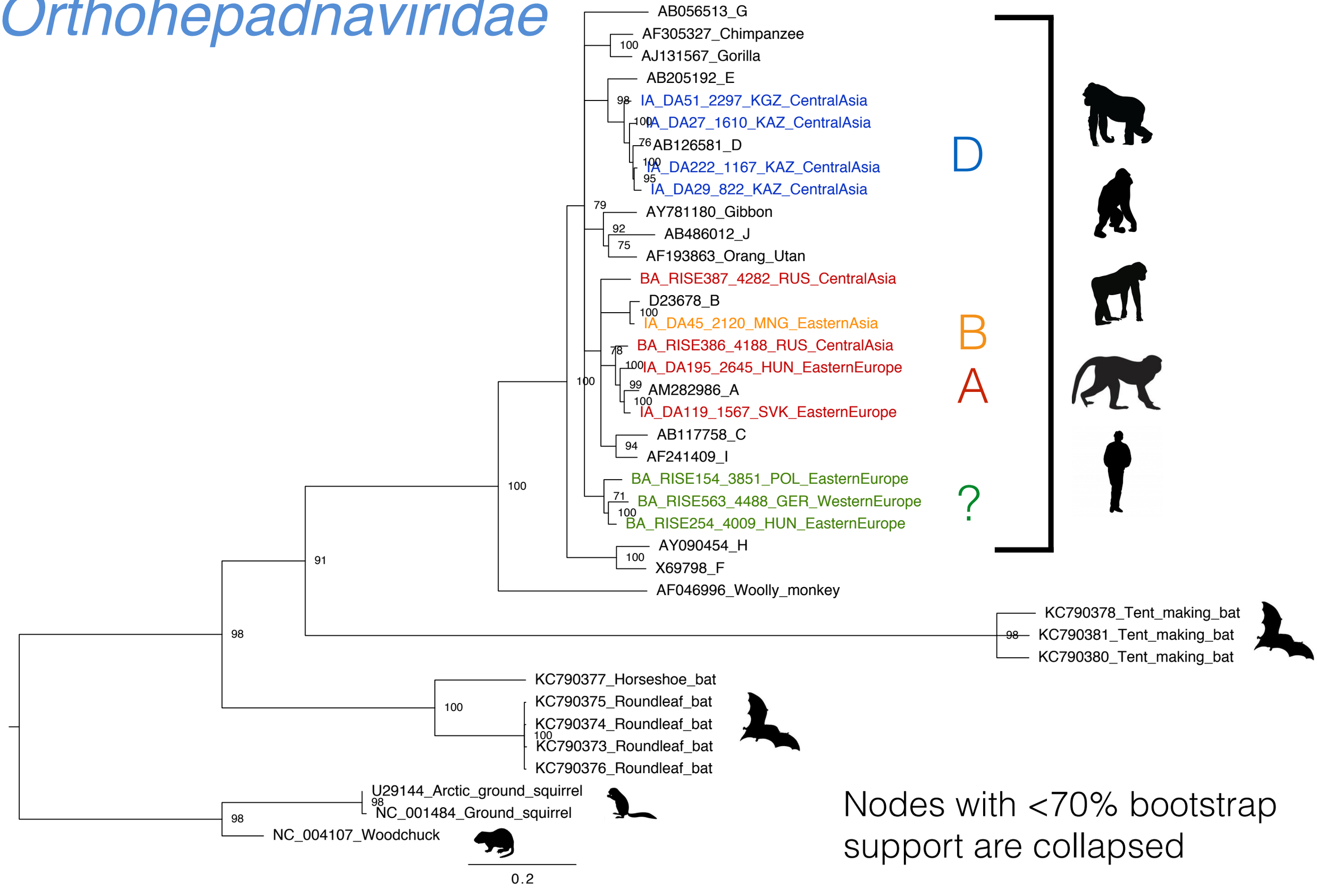


Mühlemann, Jones, Damgaard, Allentoft et al. 2018, Nature.

HBV results

- Phylogenetics
- Novel genome properties
- Recombination event
- Genotype distribution

Orthohepadnaviridae



Nodes with <70% bootstrap support are collapsed

Orthohepadnaviridae

Old World ape HBV

AB056513_G
AF305327_Chimpanzee
100
AJ131567_Gorilla
AB205192_E
98
IA_DA51_2297_KGZ_CentralAsia
100
IA_DA27_1610_KAZ_CentralAsia
76
AB126581_D
100
IA_DA222_1167_KAZ_CentralAsia
95
IA_DA29_822_KAZ_CentralAsia
D
79
AY781180_Gibbon
92
AB486012_J
75
AF193863_Orang_Utan
BA_RISE387_4282_RUS_CentralAsia
D23678_B
100
IA_DA45_2120_MNG_EasternAsia
B
78
BA_RISE386_4188_RUS_CentralAsia
100
IA_DA195_2645_HUN_EasternEurope
100
99
AM282986_A
100
IA_DA119_1567_SVK_EasternEurope
A
AB117758_C
94
AF241409_I
BA_RISE154_3851_POL_EasternEurope
71
BA_RISE563_4488_GER_WesternEurope
100
BA_RISE254_4009_HUN_EasternEurope
?
100
AY090454_H
100
X69798_F
AF046996_Woolly_monkey
91
98
KC790378_Tent_making_bat
98
KC790381_Tent_making_bat
KC790380_Tent_making_bat
KC790377_Horseshoe_bat
100
KC790375_Roundleaf_bat
KC790374_Roundleaf_bat
100
KC790373_Roundleaf_bat
KC790376_Roundleaf_bat
U29144_Arctic_ground_squirrel
98
NC_001484_Ground_squirrel
98
NC_004107_Woodchuck
0.2

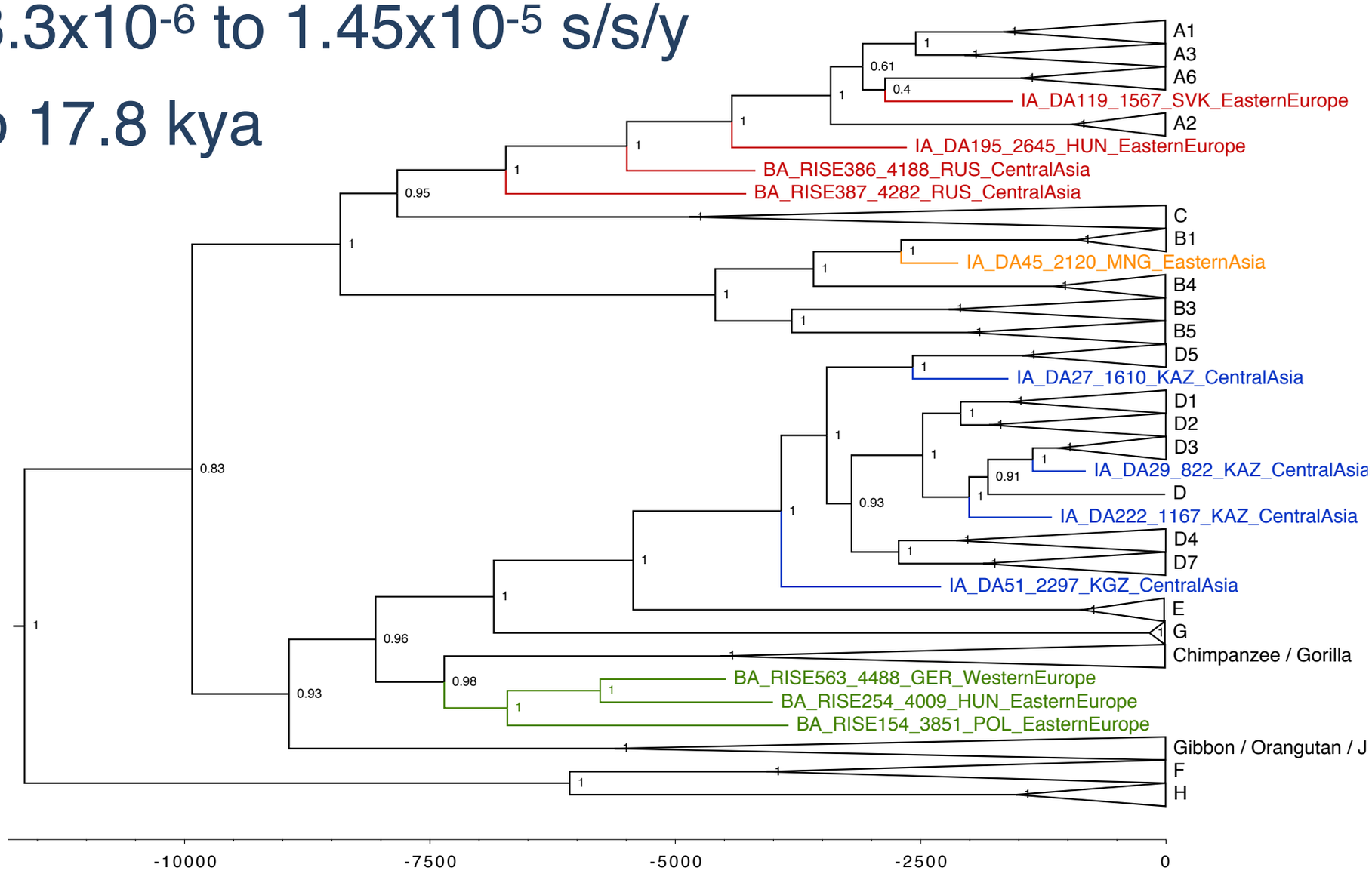
Nodes with <70% bootstrap support are collapsed

Substitution rates are lower than rates inferred using modern sequences

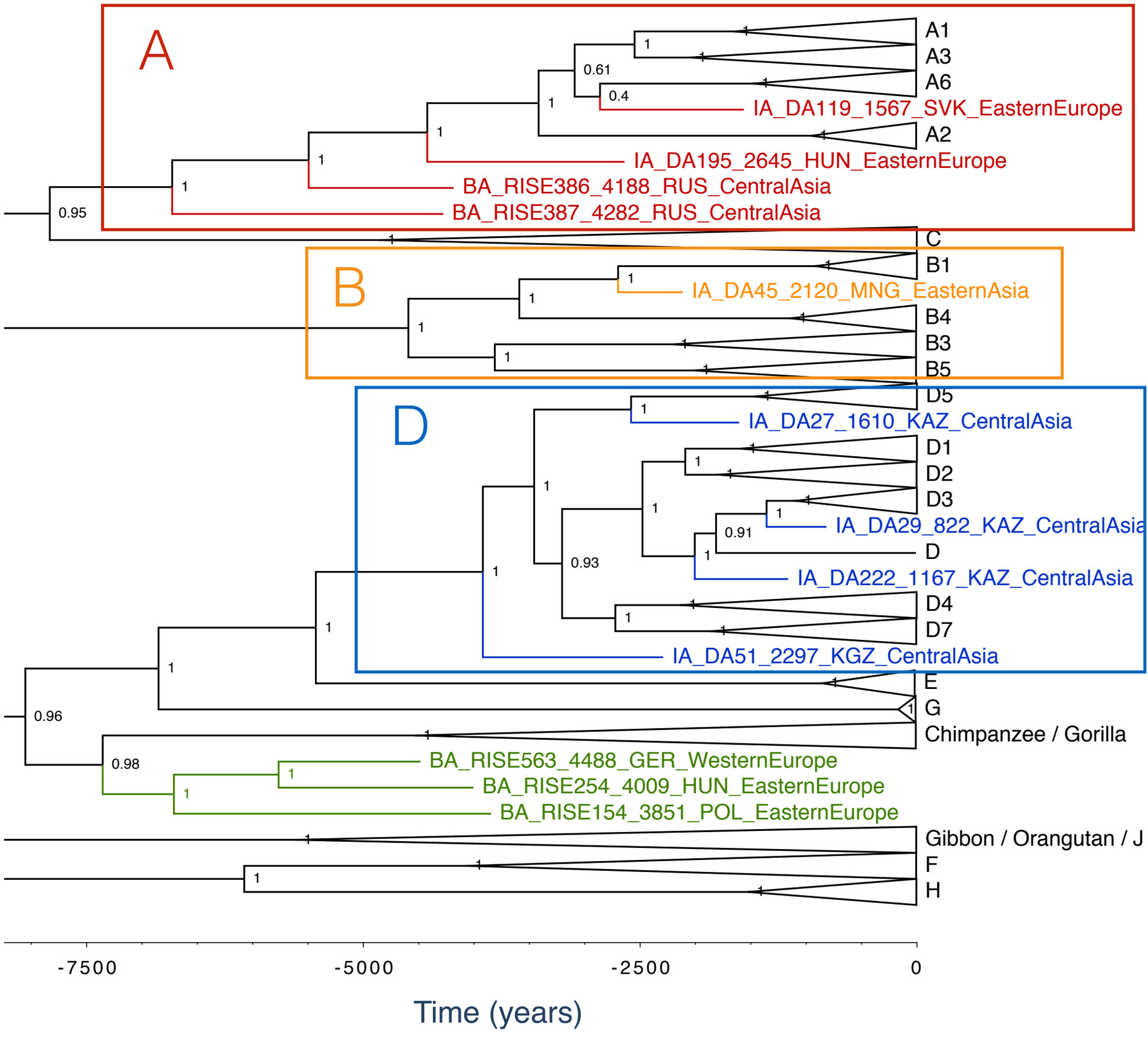
Dated coalescent tree (BEAST2)

Subst. rate: 8.3×10^{-6} to 1.45×10^{-5} s/s/y

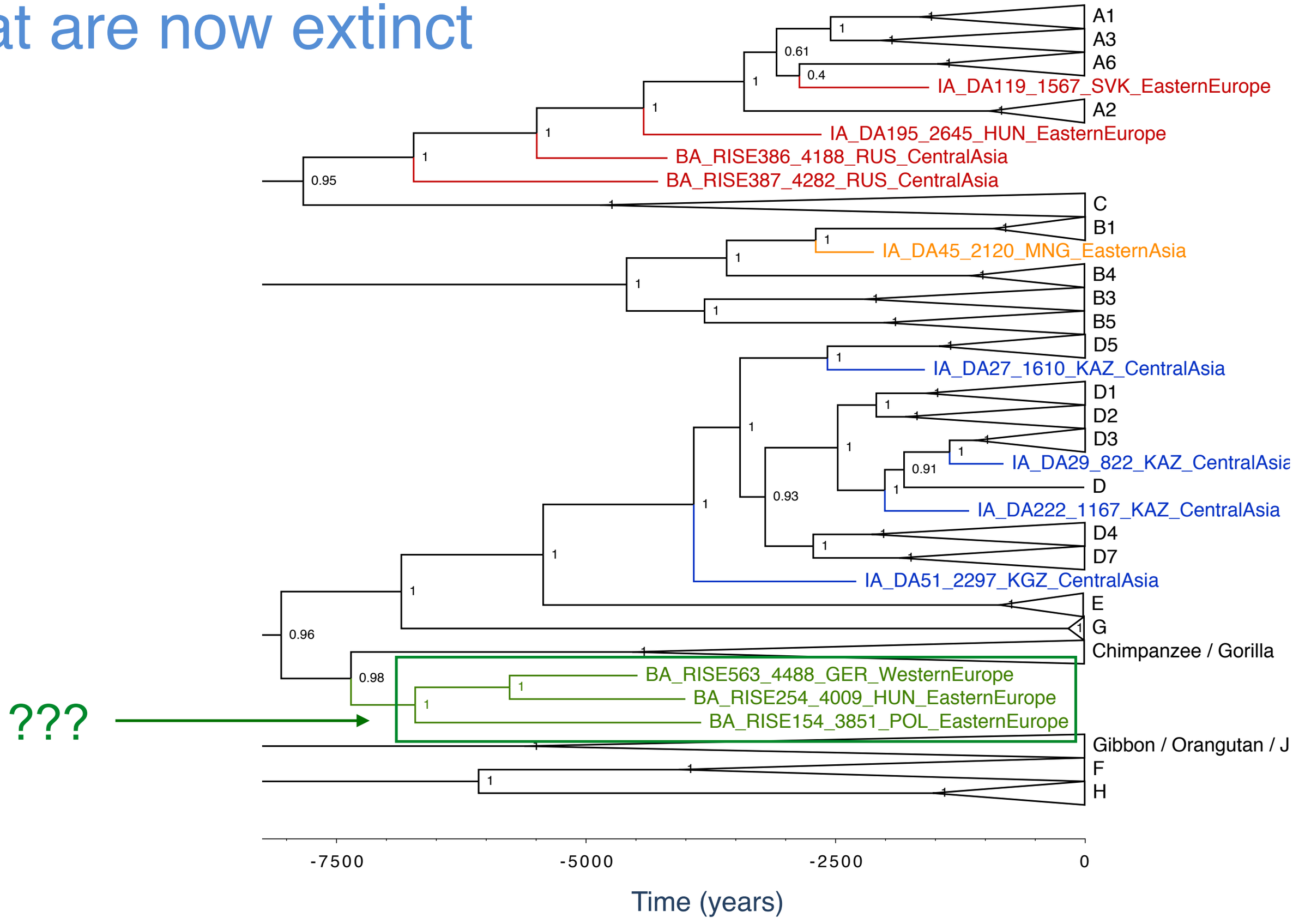
MRCA: 8.6 to 17.8 kya



Nine ancient sequences can be assigned to modern human genotypes



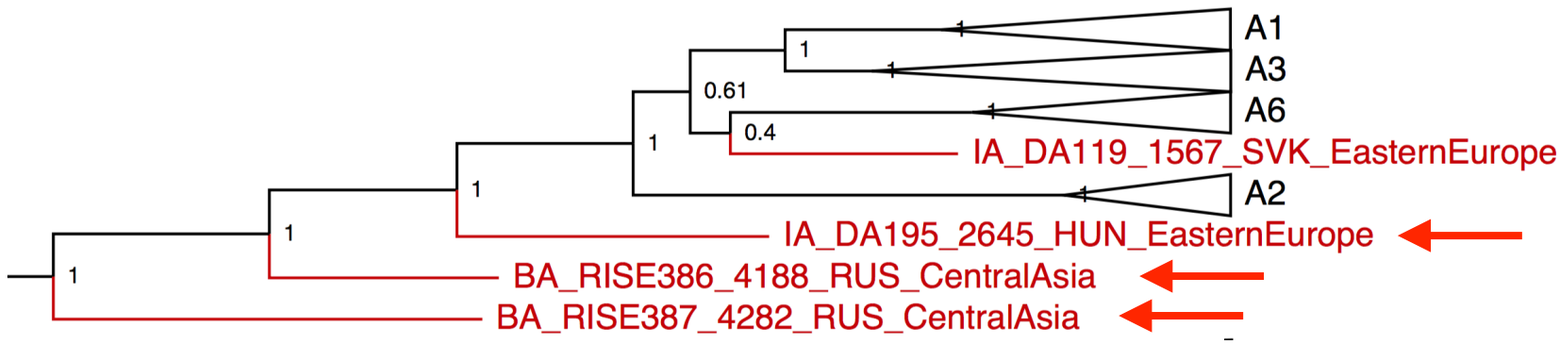
Three represent genotypes that are now extinct

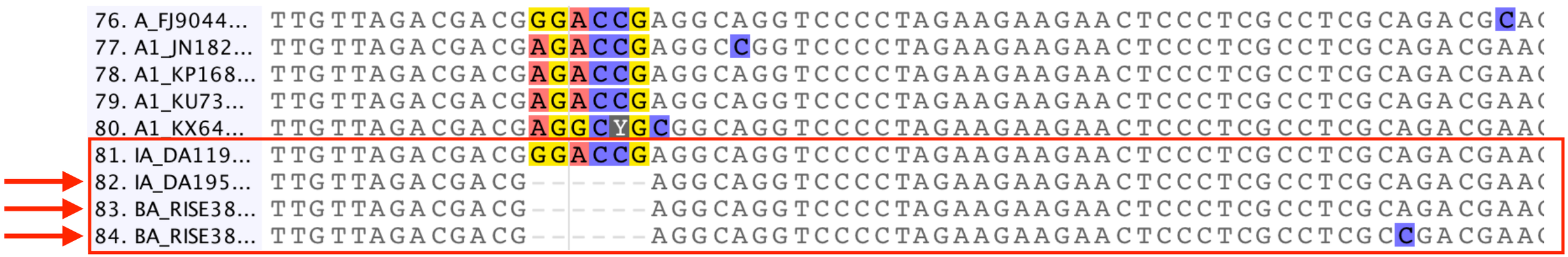


???

Novel genome properties

The three oldest ancient genotype A sequences lack a 6nt insertion in the C-terminus of the core region relative to modern genotype A.



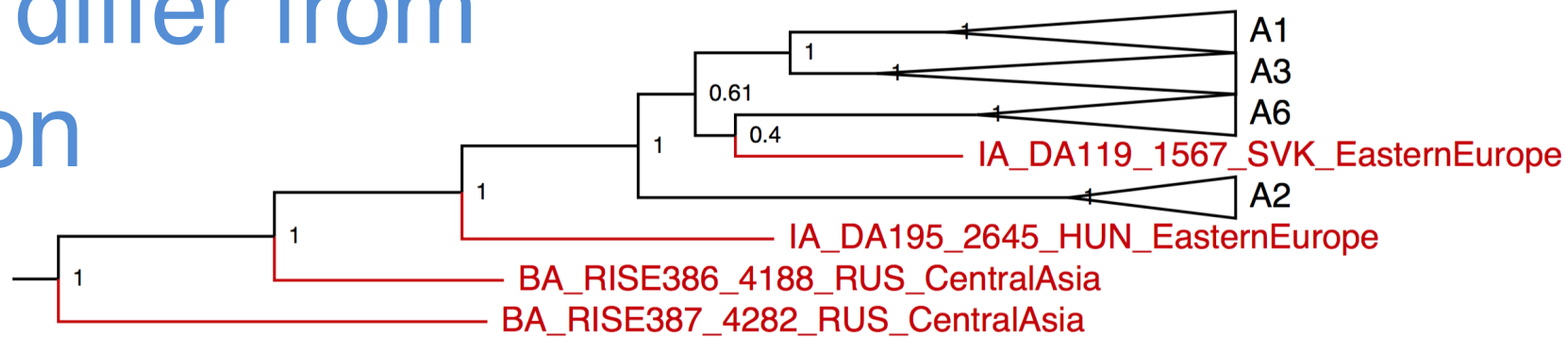


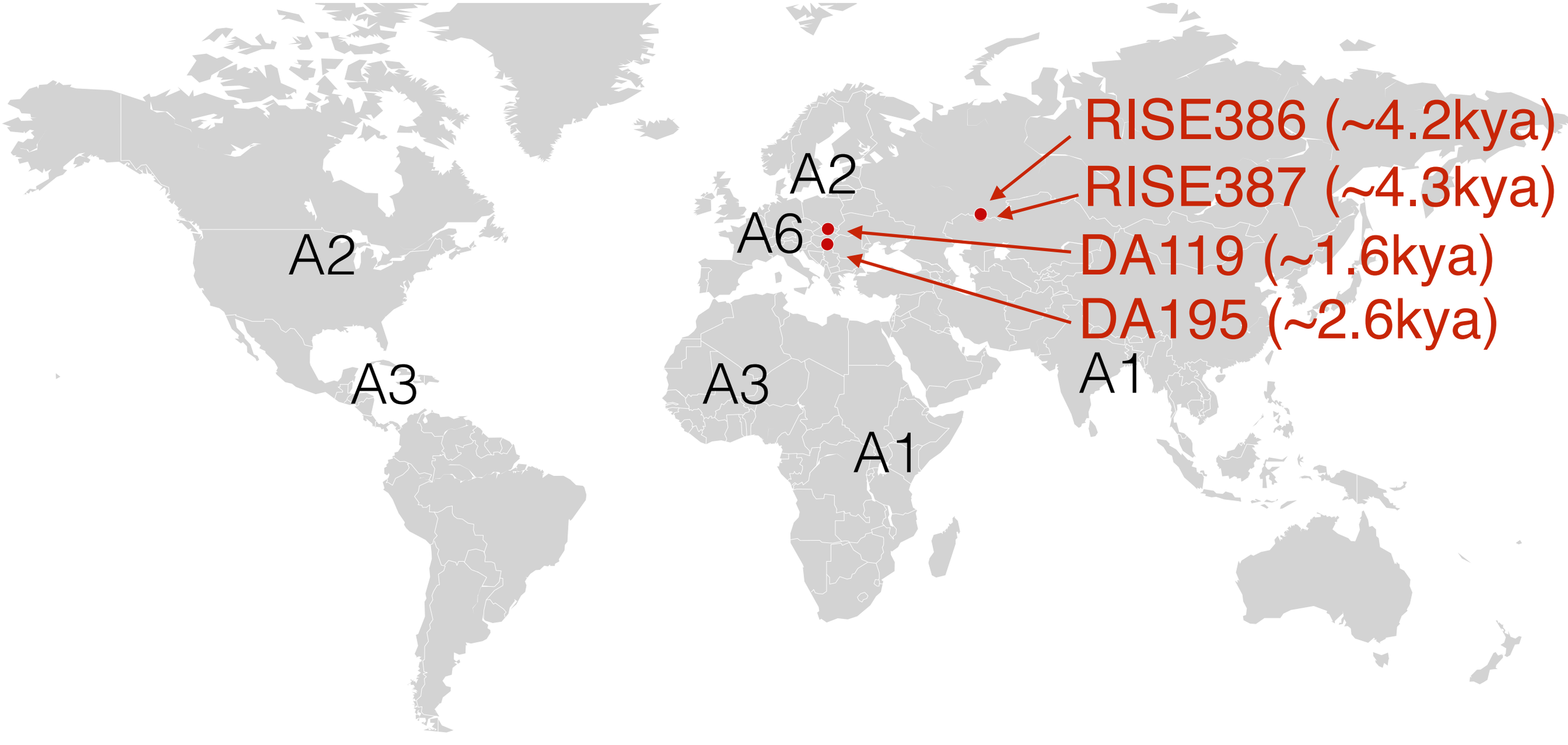
Recombination between genotype A and ancient genotype D

- Recombination breakpoints match polymerase.
- Genotype A likely formed by recombination with a genotype D ancestor and an unknown parent.
- At least two lineages of genotype D circulated in the past, one of which recombined with genotype A and has now gone extinct.

Genotype A

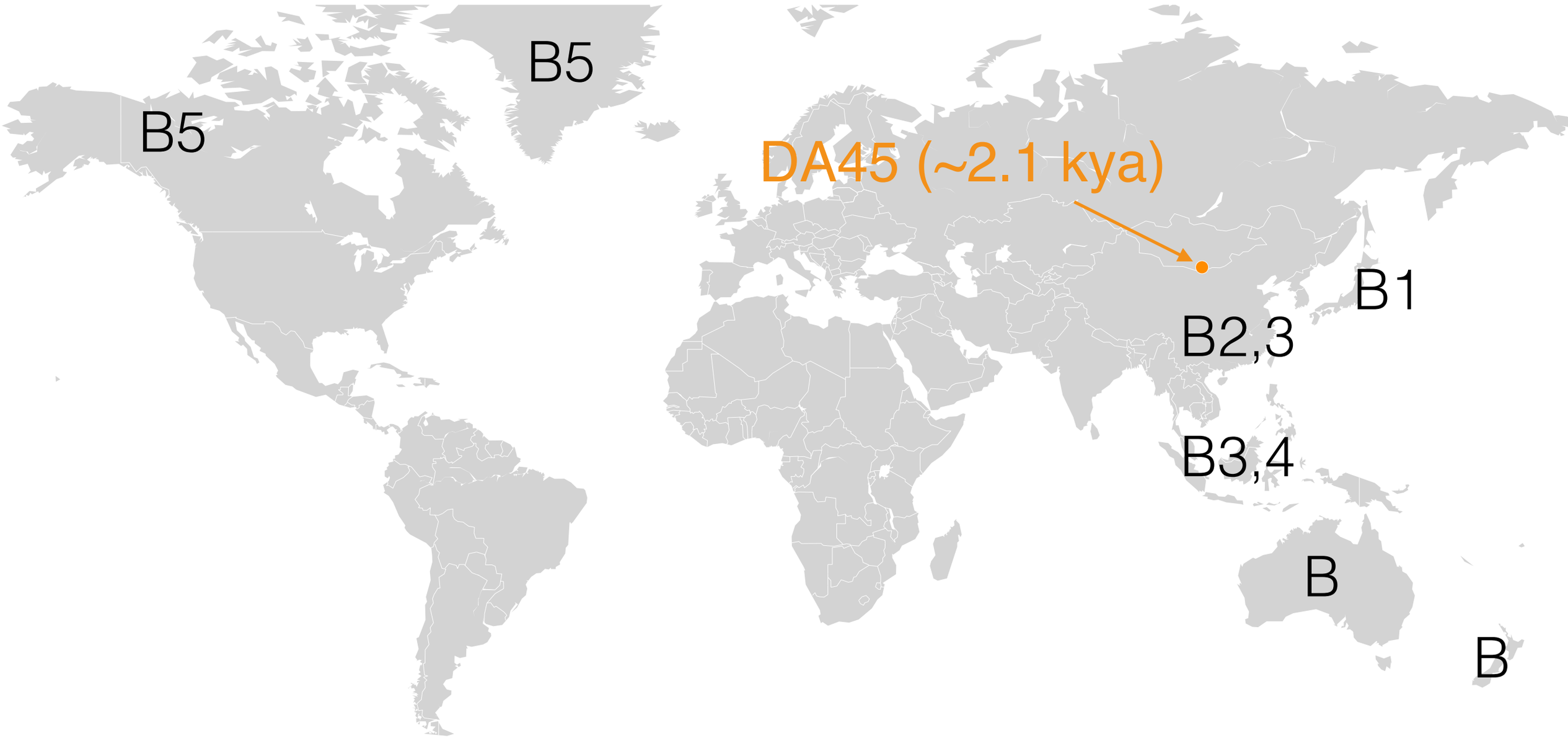
Ancient locations differ from modern distribution





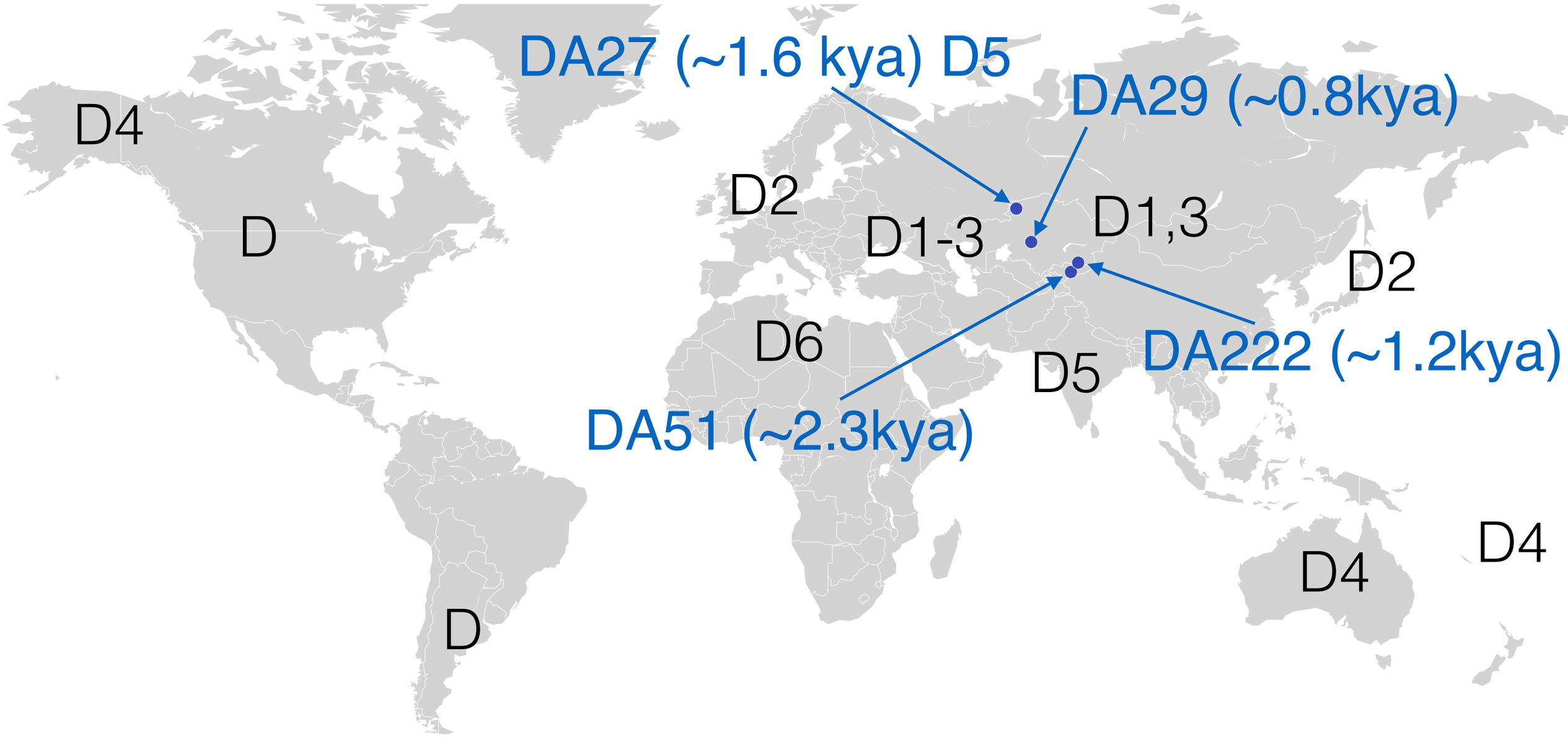
Genotype B

Ancient location matches modern distribution

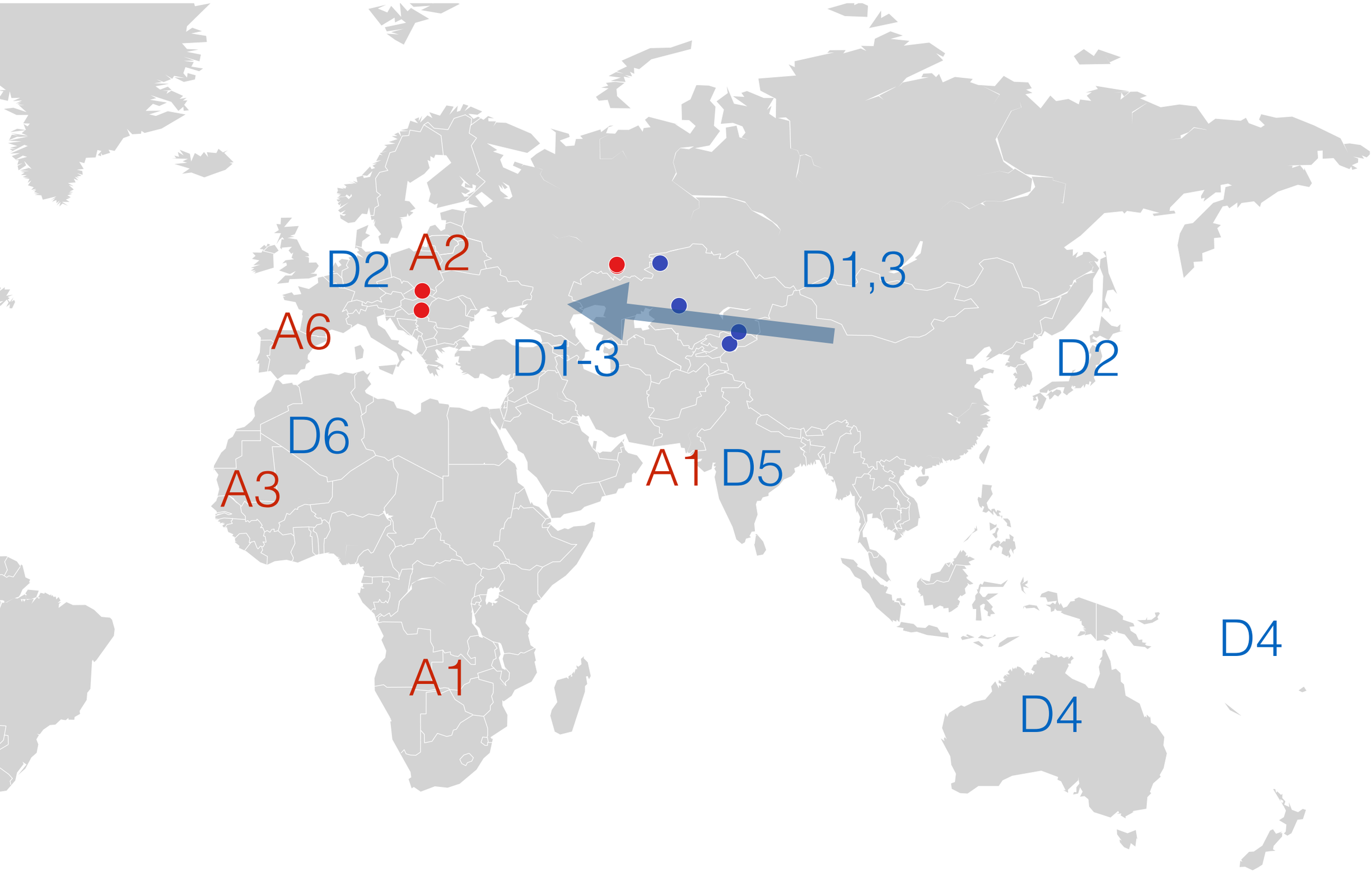


Genotype D

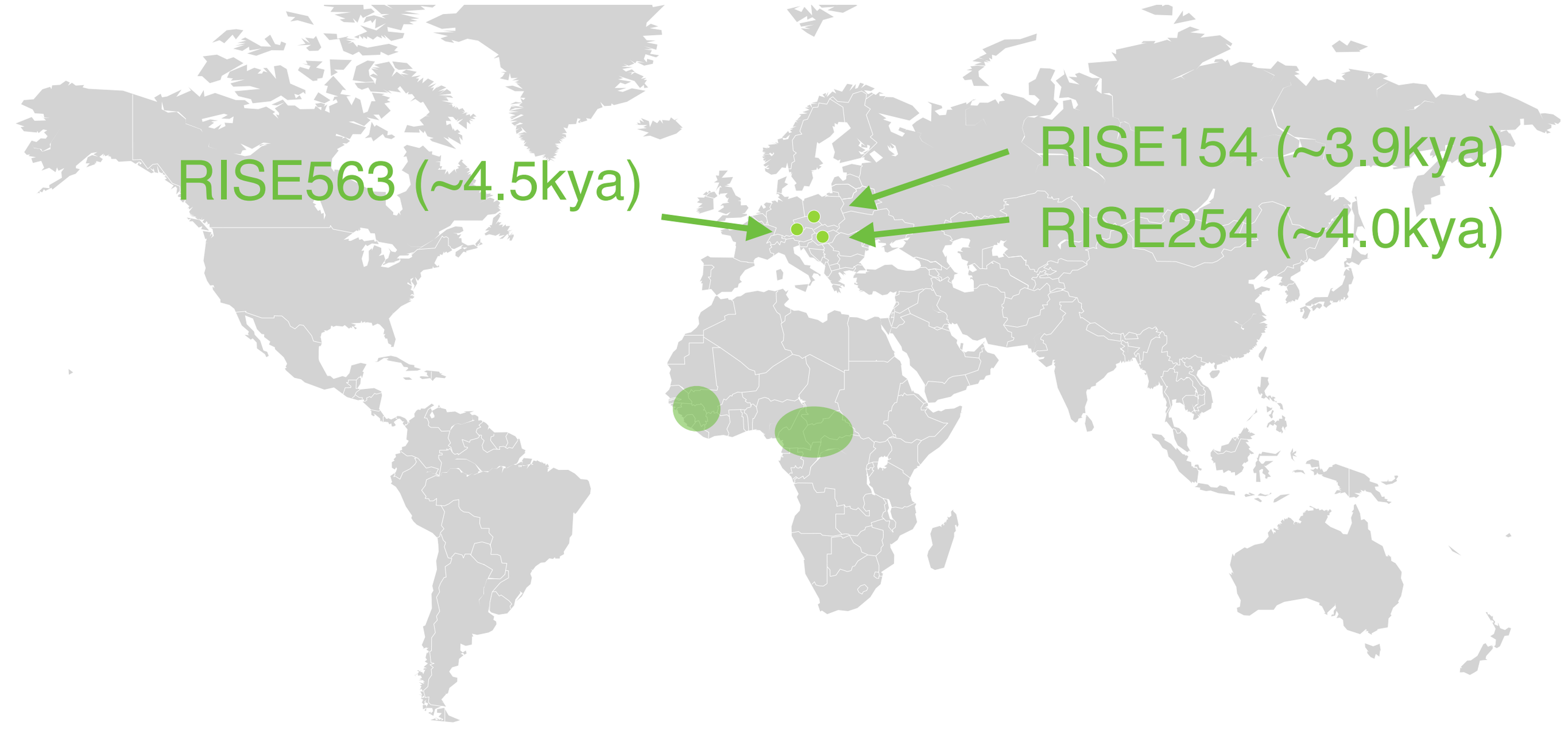
Ancient D5 location differs from modern distribution



Did genotype D replace genotype A in Central Asia?



Non-human (modern-)primate-like sequences



Reconstructing ancient HBV viruses

- Dieter Glebe, Felix Lehmann, and Nora Goldmann (all at Justus Liebig University, Giessen) have done experimental work on these ancient viruses.
- HDV made with the ancient HBV surface proteins are infective.
- HDV infectivity is neutralized by the modern vaccine.
- Surface protein antigenicity has been preserved. Modern diagnostic test kits detect the ancient antigens.

HBV conclusions

- HBV widespread in humans since the Bronze Age.
- Modern diversity arose after the split of the Old and New World genotypes.
- Complexity of HBV evolution not seen from modern sequences.
    - Genotypes have gone extinct
    - Lower substitution rate estimate
    - Geographic movement revealed
    - Evidence for recombination
- Earlier (and simpler) patterns of HBV evolution have likely been overwritten.

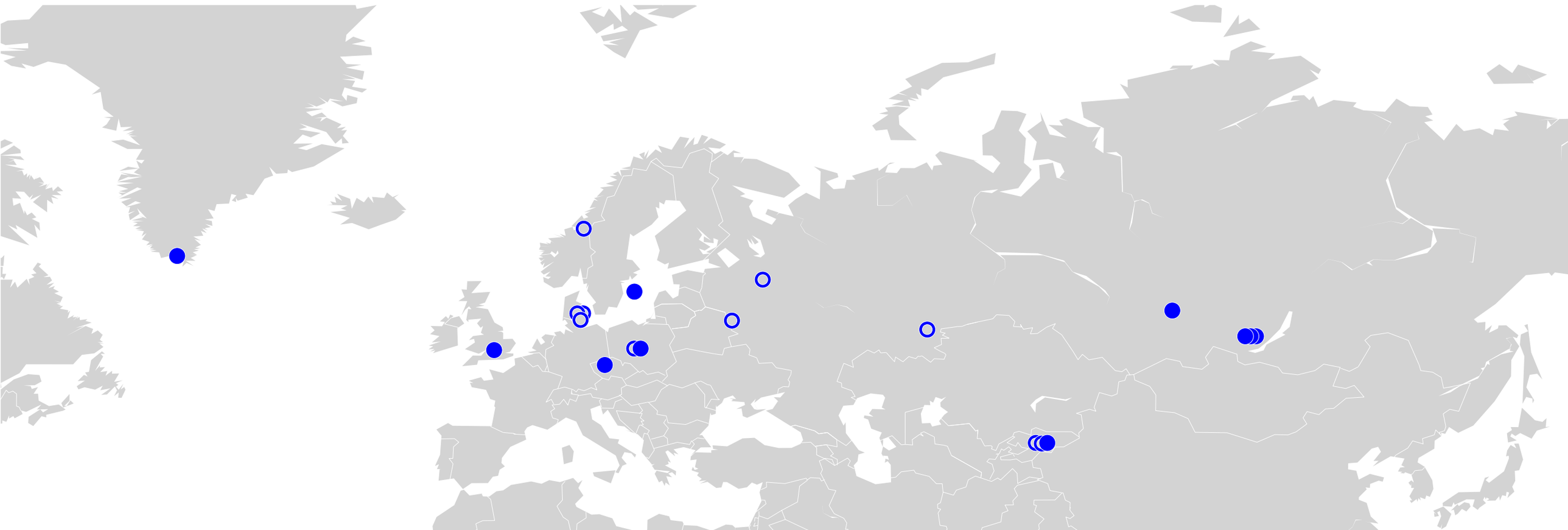
Human Parvovirus B19

Human Parvovirus B19 (B19)

- Single stranded DNA virus.
- Fifth disease in children, Hydrops fetalis, transient or persistent erythroid aplasia, and aplastic crisis.
- Transmitted via respiratory or blood borne route.
- Life-long persistence.
- Three genotypes.

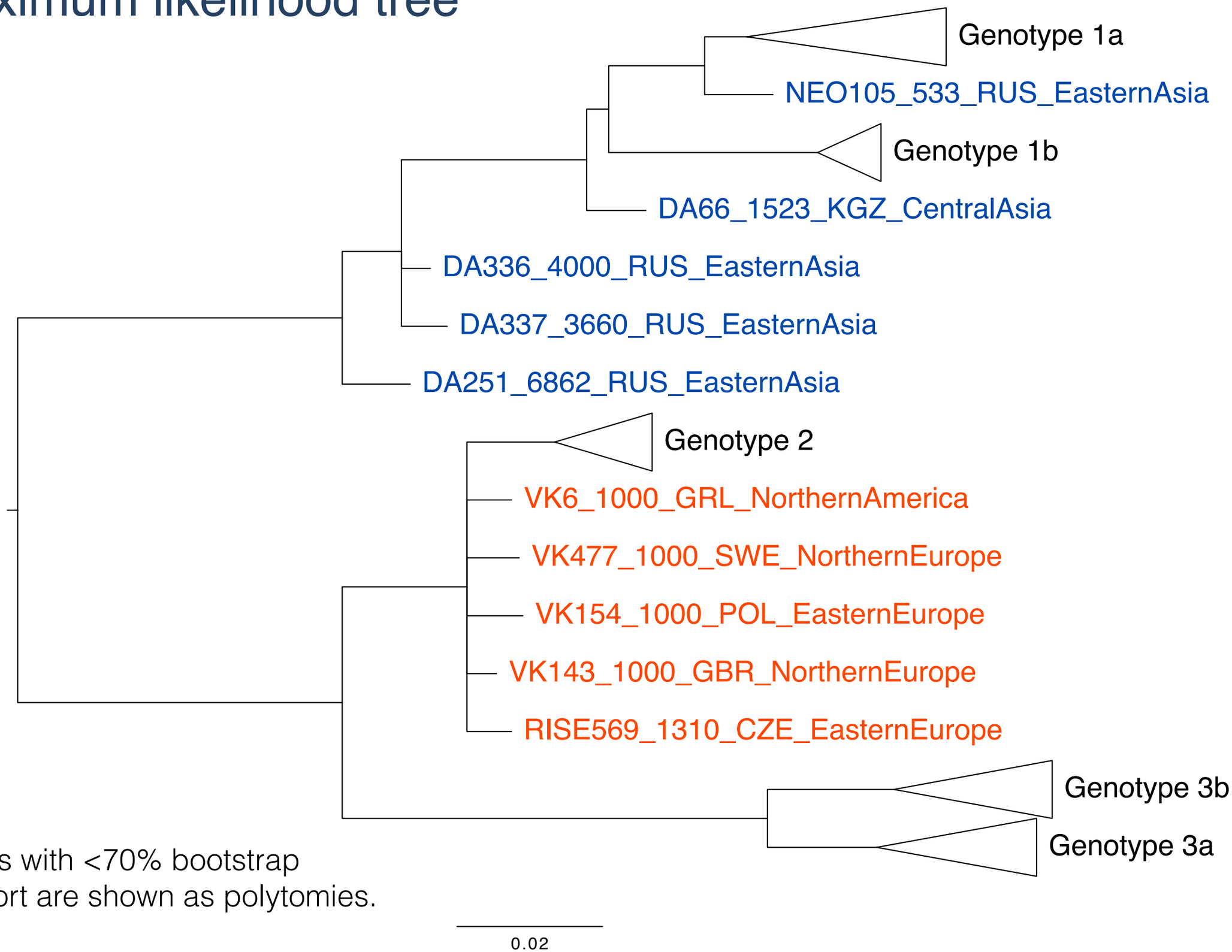
20 individuals had reads matching B19

10 individuals (coverage >50%) were included in further phylogenetic analysis (~0.5 to ~6.9 kya).



Mühlemann, Margaryan, Damgaard, Allentoft et al. 2018, PNAS.

Maximum likelihood tree



Nodes with <70% bootstrap support are shown as polytomies.

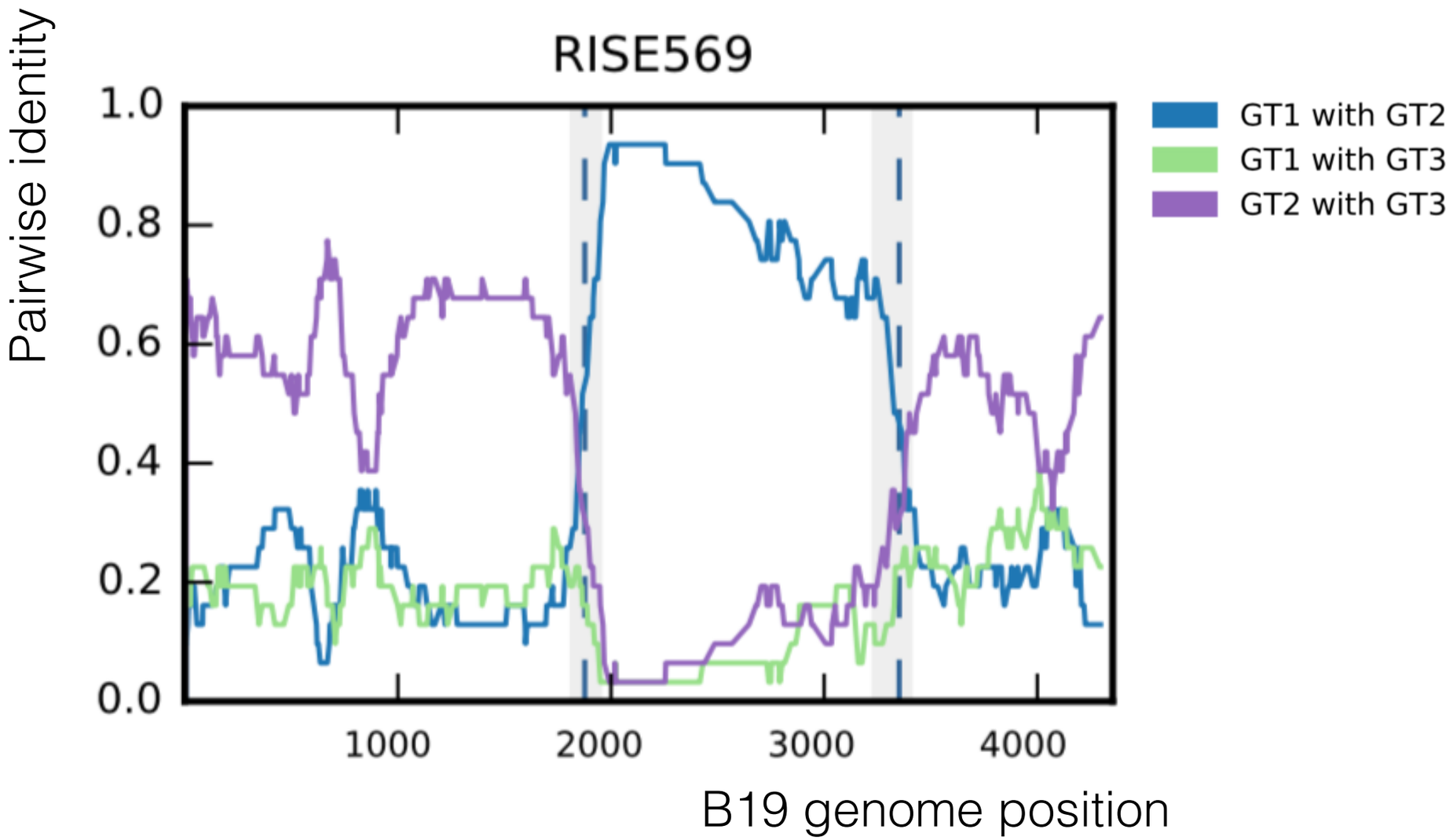
1

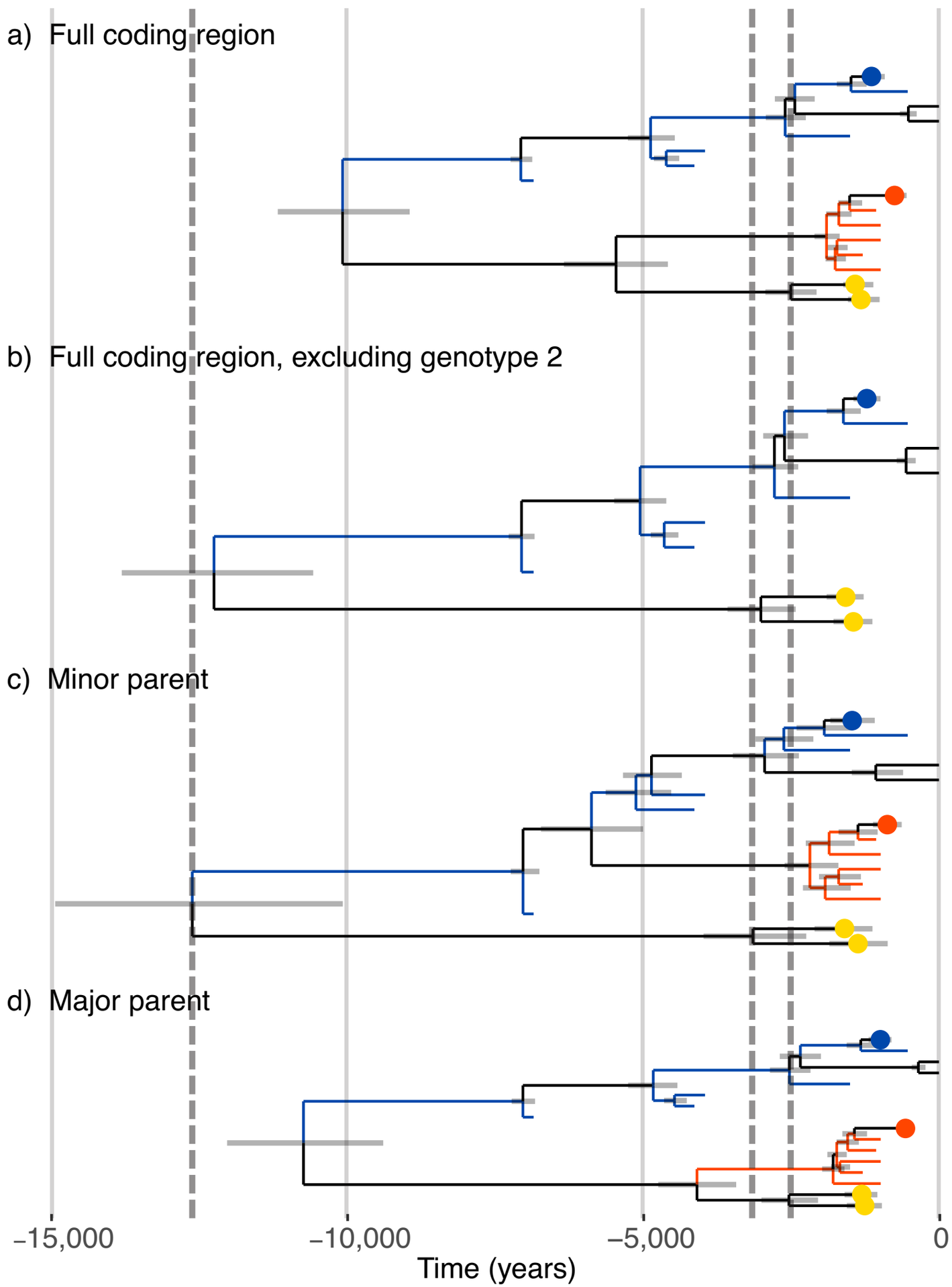
2

3

Genotype 2 is a recombinant

Genotype 1 and 3 recombined to form genotype 2.





Dated coalescent trees (BEAST2)

MRCA: ~12.6 kya

1.22×10^{-5} (1.04×10^{-5} - 1.40×10^{-5}) s/s/y

Gt 2 was formed by recombination, ~5.0-6.8 kya.

Conclusions - B19

- Recovered B19 sequences up to 6900 years old.
- Revised (slower) substitution rate. B19 was thought to have a rate more like an RNA virus.
- MRCA adjusted from ~1800 to ~12,000 years ago.
- Dating of recombination event.

Part III - conclusions and keynoting

The ancient HBV and B19 work has revealed interesting information:

- Genotype extinction.
- Facts (lower bounds) about virus age in human.
- Altered substitution rate estimates.
- Evidence for recombination.
- Geographic movement.
- Early evolutionary patterns may have been overwritten.
- Library of possible (future) genetic variation.

But… a more general and more important issue is revealed

- Modern data, taken alone, can be very misleading.
- An initial handful of ancient sequences may not make things clearer.
- They give us more facts but also (at this point) raise more questions than they can answer.

Let’s resist the urge to speculate too early

- For the viruses we have some ancient data for, the evidence is clear: we must be very cautious when speculating on long-term virus evolution.

- That’s true even if the analysis is solid, done by experts, and all available ancient data is also included. The data sets are still miniscule!

- Resist the temptation to draw conclusions based on tiny sample sizes (sometimes as low as 1).

- Reliable ancient HBV data is fewer than 20 samples from tens of millions of km^2, over 7000 years. What could possibly go wrong?

- Darwin was clear on that, too.

“From these considerations, from our ignorance of the geology of other countries beyond the confines of Europe and the United States, and from the revolution in our palaeontological knowledge effected by the discoveries of the last dozen years, it seems to me to be about as rash to dogmatize on the succession of organic forms throughout the world, as it would be for a naturalist to land for five minutes on a barren point in Australia, and then to discuss the number and range of its productions.”

Charles Darwin
On the Origin of Species (ch. 10)

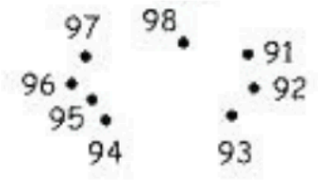
"From these considerations, from our ignorance of the geology of other countries beyond the confines of Europe and the United States, and from the revolution in our palaeontological knowledge effected by the discoveries of the last dozen years, it seems to me to be about as rash to dogmatize on the succession of organic forms throughout the world, as it would be for a naturalist to land for five minutes on a barren point in Australia, and then to discuss the number and range of its productions."

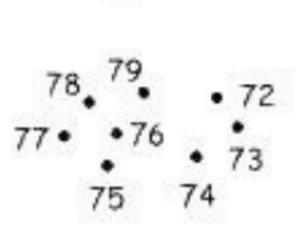
Charles Darwin
On the Origin of Species

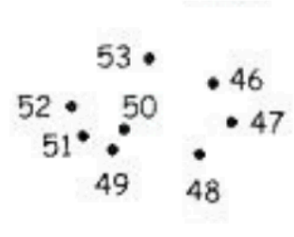
Connect the dots!

"It's a few circular things, right?

Or maybe they're actually trees, seen from above?"



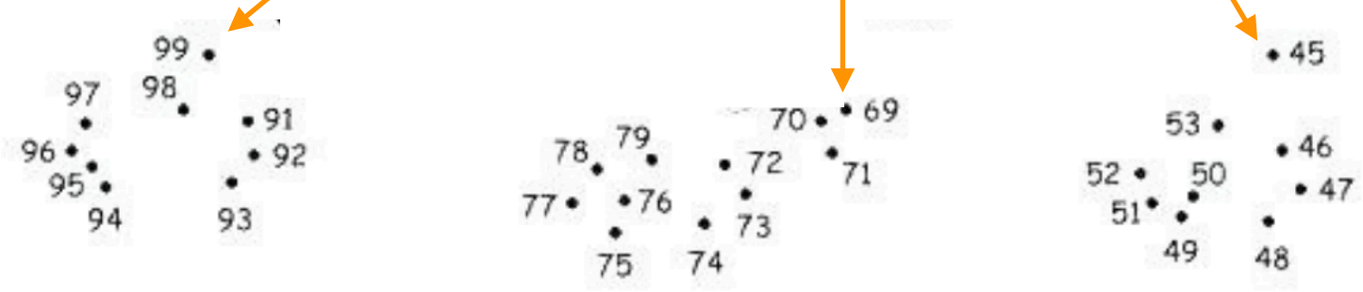




Connect the dots!

"A few circular things, but, you know, with bumps on them"

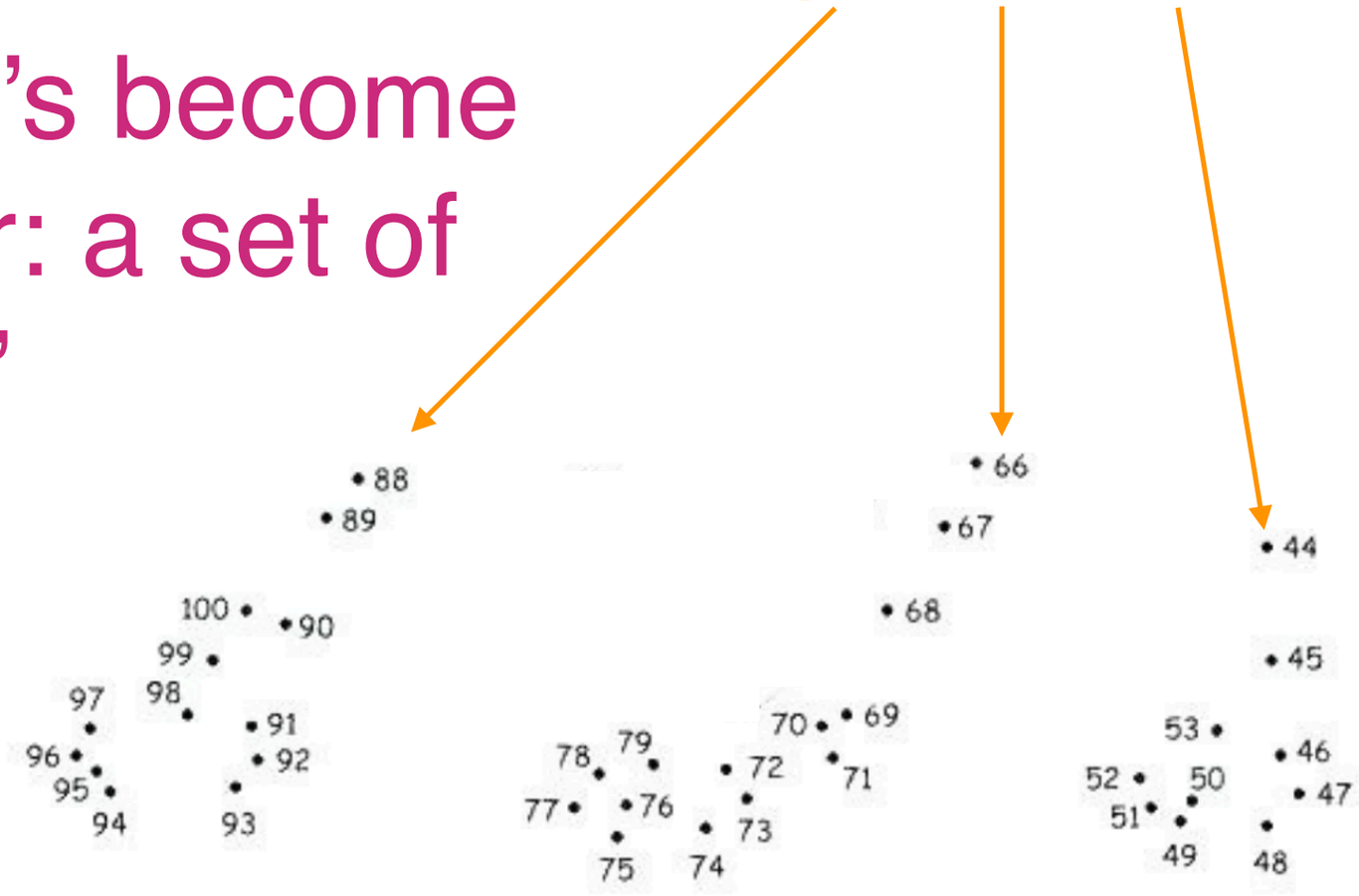
Amazing ancient samples



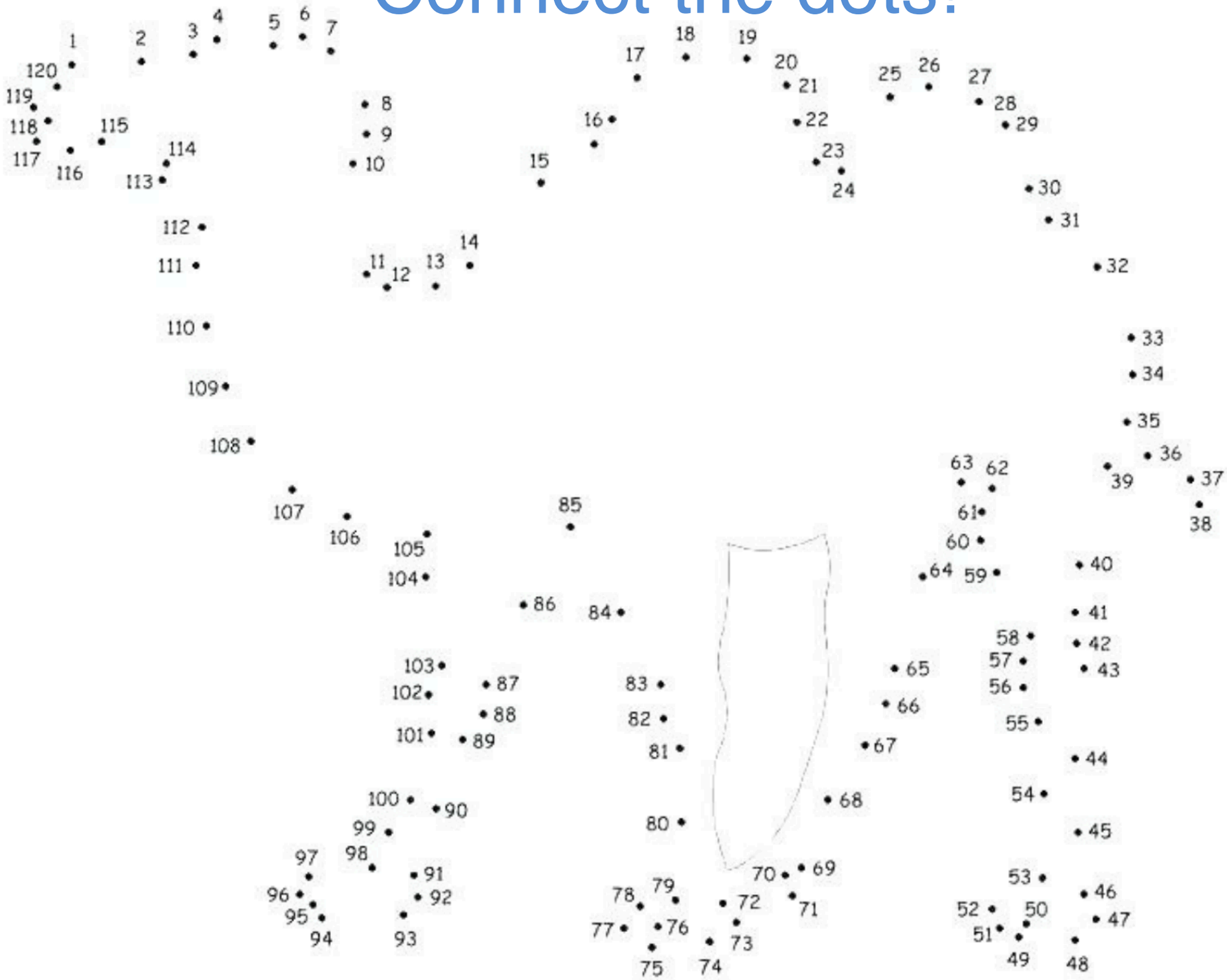
Connect the dots!

"Ah yes, it's become quite clear: a set of golf clubs"

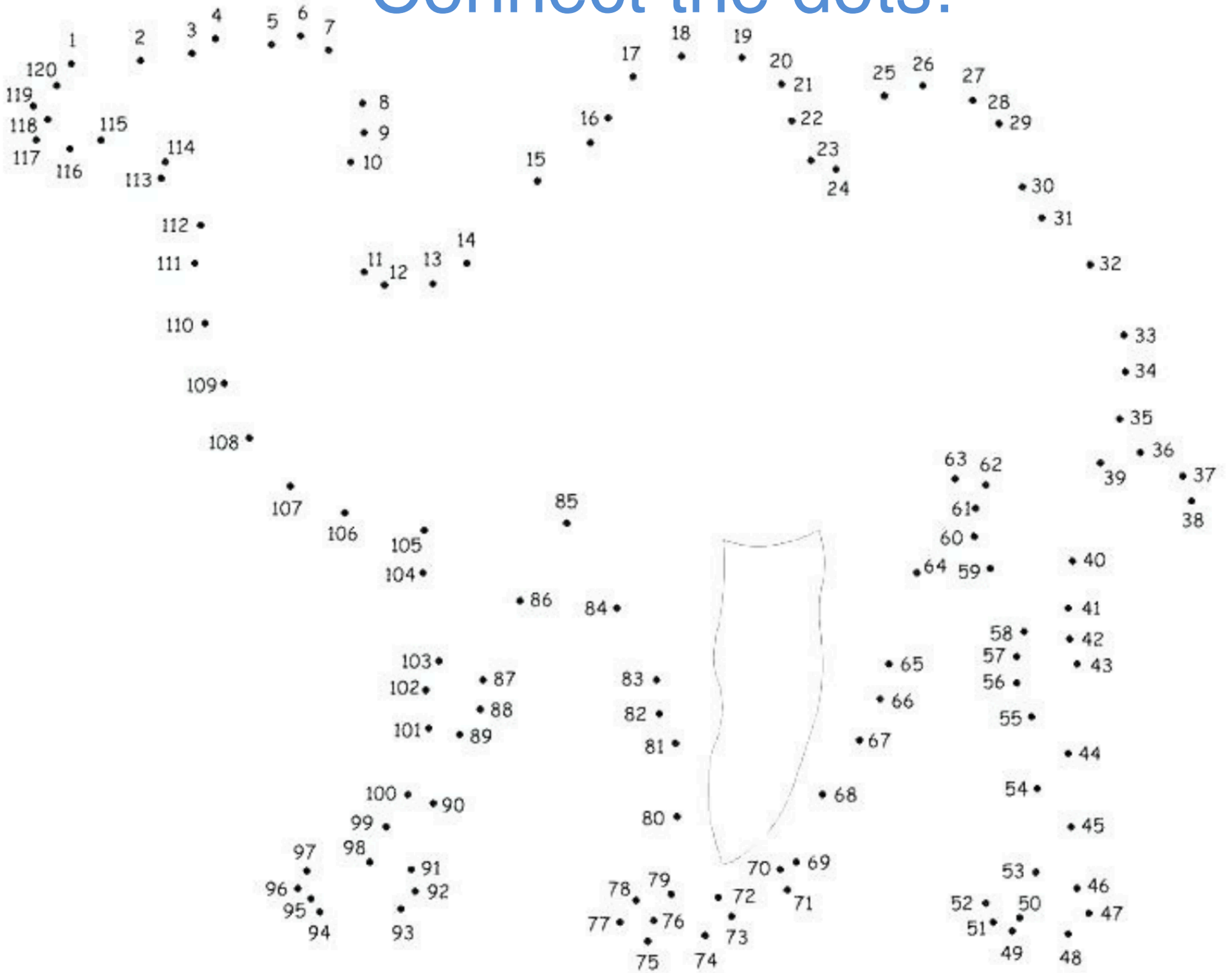
Even more amazing ancient samples!



Connect the dots!



Connect the dots!



10^8

10^7

10^6

10^5

10^4

10^3

10^2

10^1

Time (years)

Actually, let's be more cautious in general

- Let's do parsimony right. Refrain from speculating instead of feeling like we must come up with an over-arching storyline to explain a couple of isolated facts.
- The tree model isn't right to begin with (at least for viruses). Maximum parsimony = no opinion (yet).
- The broad outlines of human migrations give a highly simplified, discretized picture. Plus, humans could be moving in one direction while a virus goes in the other.
- While we have little data, let's prefer our brains to underpowered and possibly inappropriate models.

In considering ancient sequences, take care not to conflate the ancient with the modern

Even Darwin had trouble with that:

“In the first place, it should always be borne in mind what sort of intermediate forms must, on the theory, have formerly existed. I have found it difficult, when looking at any two species, to avoid picturing to myself forms DIRECTLY intermediate between them. But this is a wholly false view; we should always look for forms intermediate between each species and a common but unknown progenitor; and the progenitor will generally have differed in some respects from all its modified descendants. ”

Charles Darwin
On the Origin of Species (ch. 10)

"In the first place, it should always be borne in mind what sort of intermediate forms must, on the theory, have formerly existed. I have found it difficult, when looking at any two species, to avoid picturing to myself forms DIRECTLY intermediate between them. But this is a wholly false view; we should always look for forms intermediate between each species and a common but unknown progenitor; and the progenitor will generally have differed in some respects from all its modified descendants."

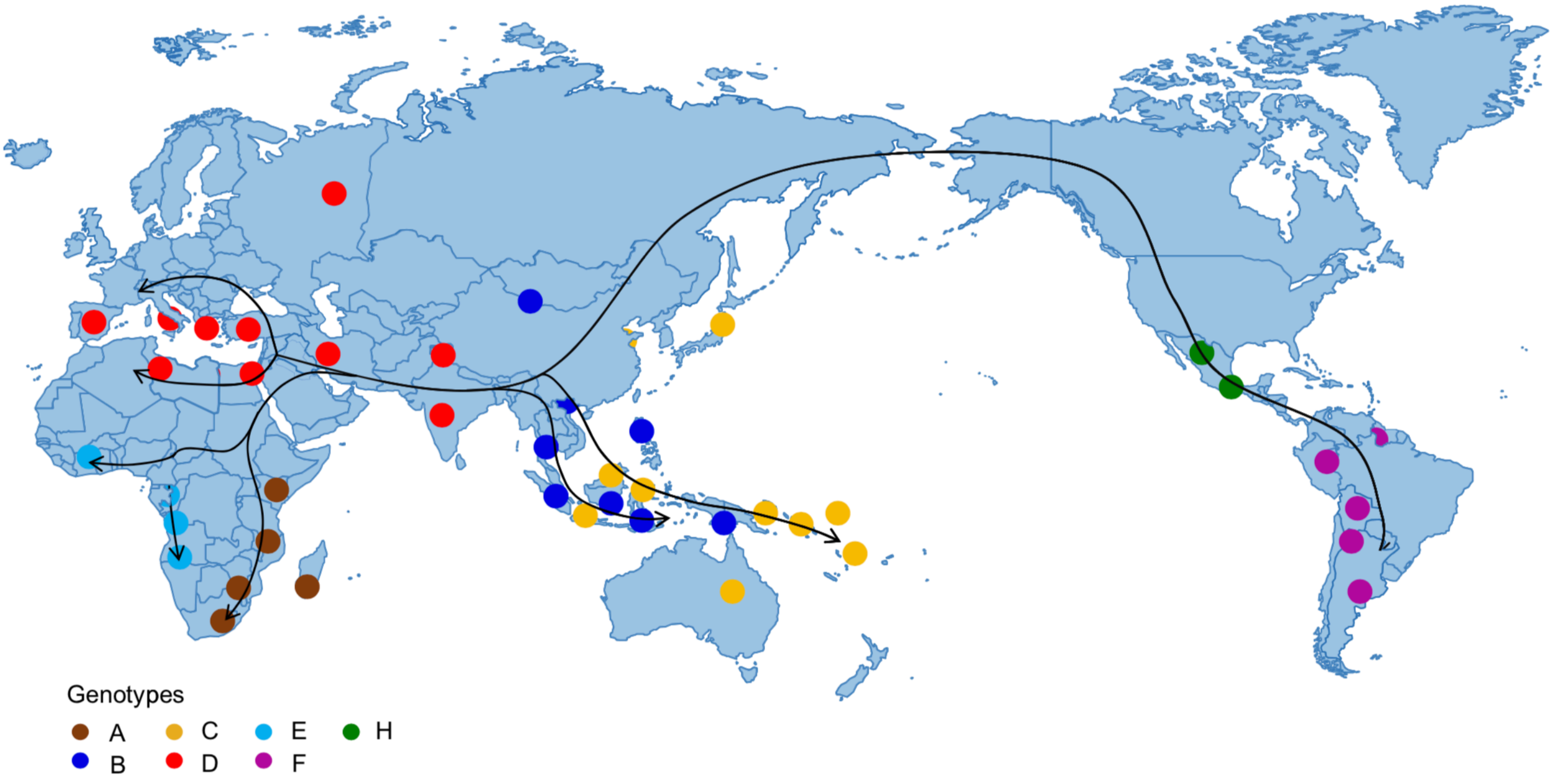
Charles Darwin
On the Origin of Species (ch. 10)


ARBITRARY
ASSUMPTIONS
NEXT 10km

Hold your horses!

- HBV is 400 years old and originated in South America?
- Human Parvovirus B19 genotype 1 arose in the 1960s?
- HBV genotype A originated in Africa?
- Inferred rates and MRCA from modern data alone.
- Simple guesses at human/NHP transfer.
- Phylogeography tools will now tell you genotype A originated in Eurasia!
- Geographic pattern of the 6 nucleotide insertion in HBV genotype A?

Proposed dispersal pattern of HBV geographic expansion



Paraskevis et al. 2015

Thought experiment

- Build a computational model including: time, genetic change, cross-species transmission, and geographic movement.
- Repeatedly simulate 10^4 years; make 10^{13} sequences each time.
- *Severely* under-sample (10^2) the data from each run.
- Infer trees, MRCA, and rates for each sample.
- Would the results in any way resemble the truth?
- How much variation in the inferences from the tiny samples?
- What might it appear to tell us about zoonoses?
- Would it teach us to be more cautious?

“We continually forget how large the world is, compared with the area over which our geological formations have been carefully examined; we forget that groups of species may elsewhere have long existed, and have slowly multiplied, before they invaded the ancient archipelagoes of Europe and the United States. We do not make due allowance for the enormous intervals of time which have elapsed between our consecutive formations, longer perhaps in many cases than the time required for the accumulation of each formation. These intervals will have given time for the multiplication of species from some one parent-form: and in the succeeding formation, such groups or species will appear as if suddenly created.”

Charles Darwin
On the Origin of Species (ch. 10)

"We continually forget how large the world is, compared with the area over which our geological formations have been carefully examined; we forget that groups of species may elsewhere have long existed, and have slowly multiplied, before they invaded the ancient archipelagoes of Europe and the United States. We do not make due allowance for the enormous intervals of time which have elapsed between our consecutive formations, longer perhaps in many cases than the time required for the accumulation of each formation. These intervals will have given time for the multiplication of species from some one parent-form: and in the succeeding formation, such groups or species will appear as if suddenly created."

Charles Darwin
On the Origin of Species (ch. 10)

Many thanks to many people!

Bioinformatics & Virology

Barbara Mühlemann

Christian Drosten

Derek Smith

Dieter Glebe

Ron Fouchier

Ab Osterhaus

Population genetics, GeoGenetics, Copenhagen

Eske Willerslev

Peter de Barros Damgaard

Morten Allentoft

Ashot Margaryan

Lasse Vinner

Anders Hansen

Martin Sikora

Simon Rasmussen

Karl-Göran Sjögren

Kristian Kristiansen

Archaeologists

Irina Shevnina

Andrey Logvin

Emma Usmanova

Irina P. Panyushkina

Bazartseren Boldgiv

Tsevel Bazartseren

Kadicha Tashbaeva

Victor Merz

Nina Lau

Andrzej Weber

Vladimir I. Bazaliiskii

Sabine Sten

Niels Lynnerup

Václav Smrčka

Dmitry Voyakin

Egor Kitov

Andrey Epimakhov

Dalia Pokutta

Magdolna Vicze

T. Douglas Price

Vyacheslav Moiseyev

Martyna Molak

Jette Arneborg

Wieslaw Bogdanowicz

Ceri Falys

Mikhail Sablin

UNIVERSITY OF CAMBRIDGE
Department of Zoology

compare

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 643476.