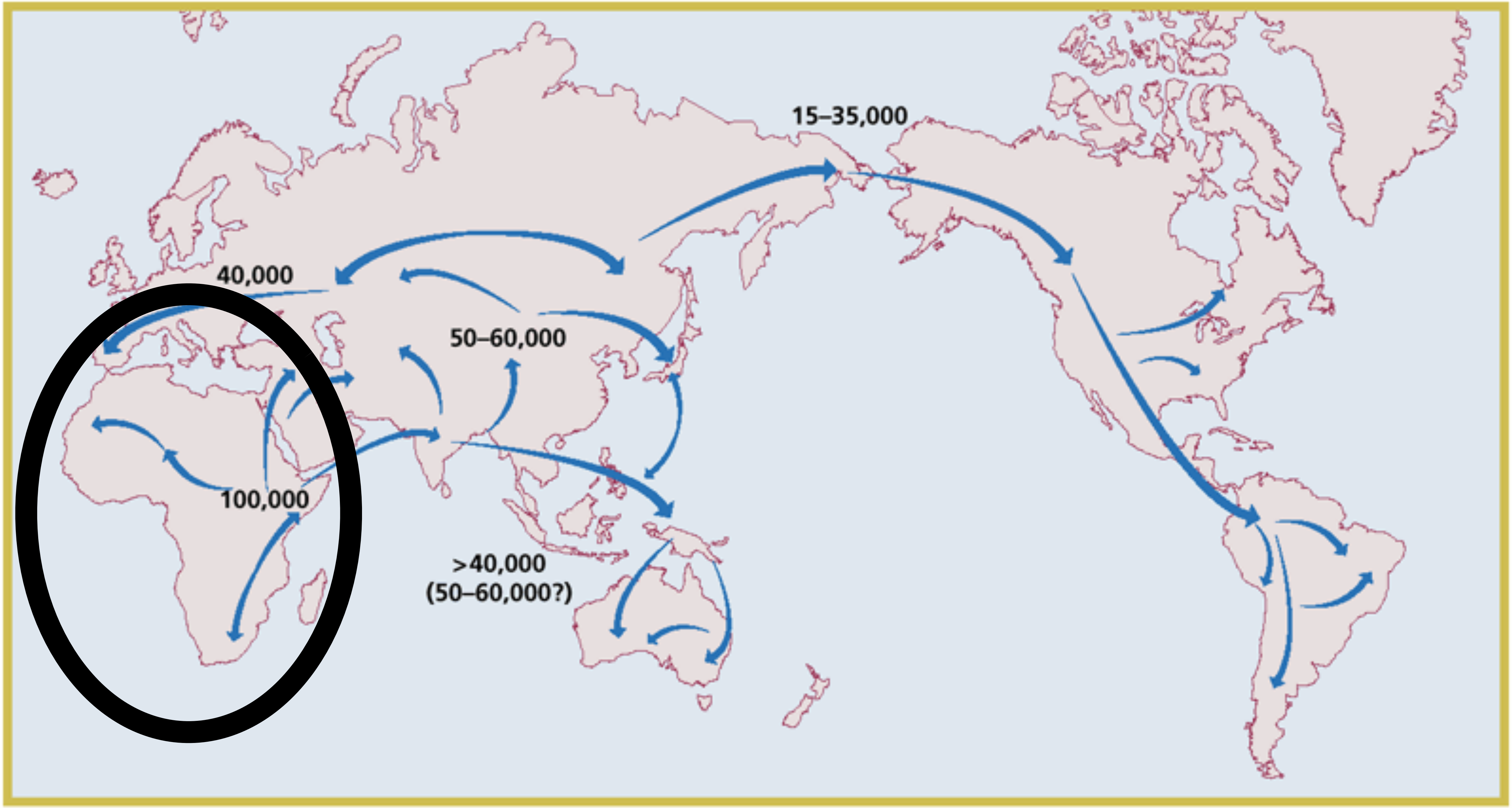
Application 2: Human history in Africa



Cavalli-Sforza and Feldman (2003)

Diversity of Africa is suggestive of major population upheavals

Diversity of Africa is suggestive of major population upheavals

Hadza + Sandawe

Khoisan

(Erstwhile) Hunter-gatherers in both eastern and southern Africa

Is there a deep genetic relationship between hunter-gatherers in southern and eastern Africa?

Previous genetic work (Tishkoff et al. 2007, Tishkoff et al. 2009, Henn et al. 2012) inconclusive though suggestive

Could settle a long-standing debate in anthropology about the makeup of Africa ~10kya

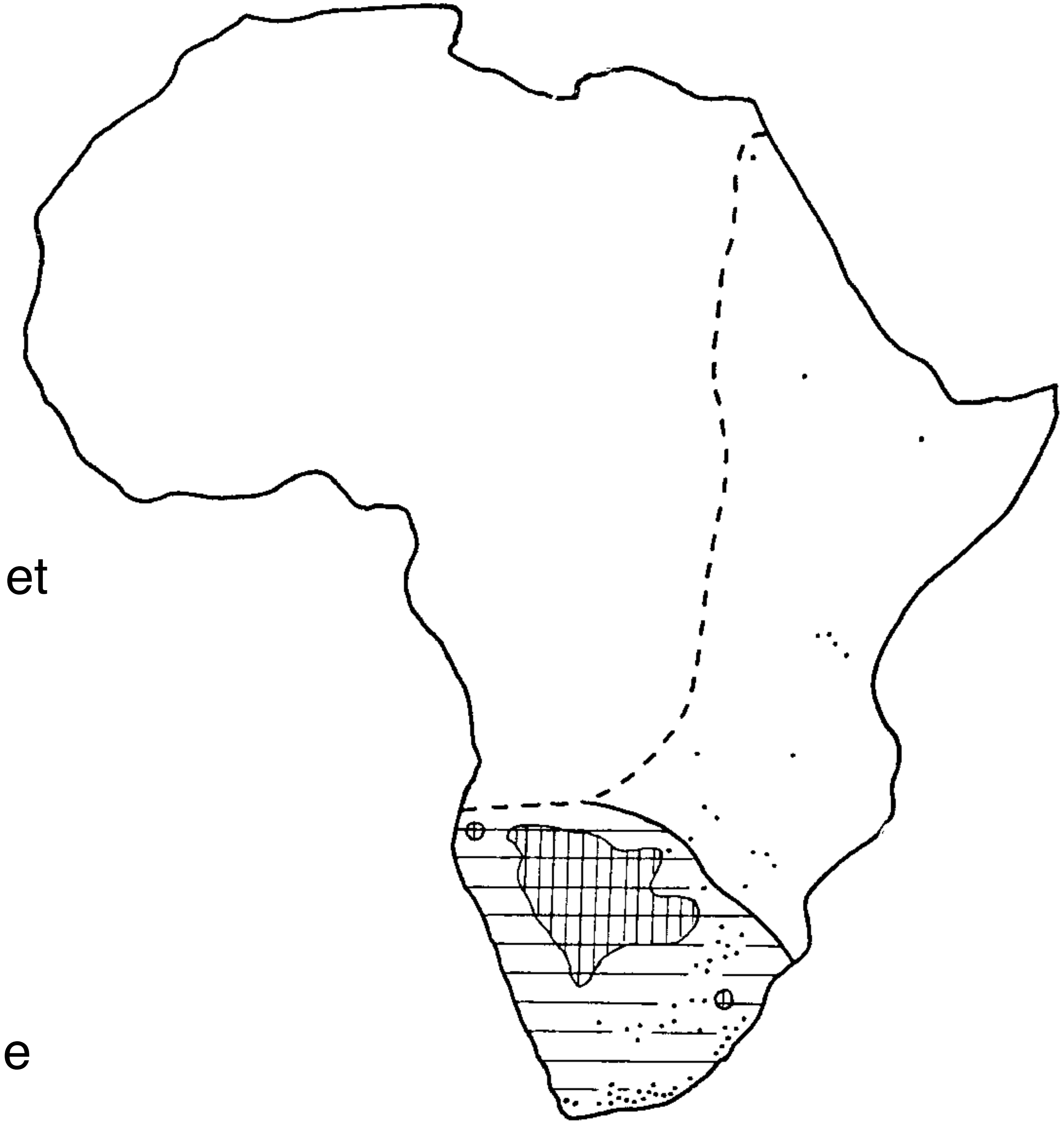


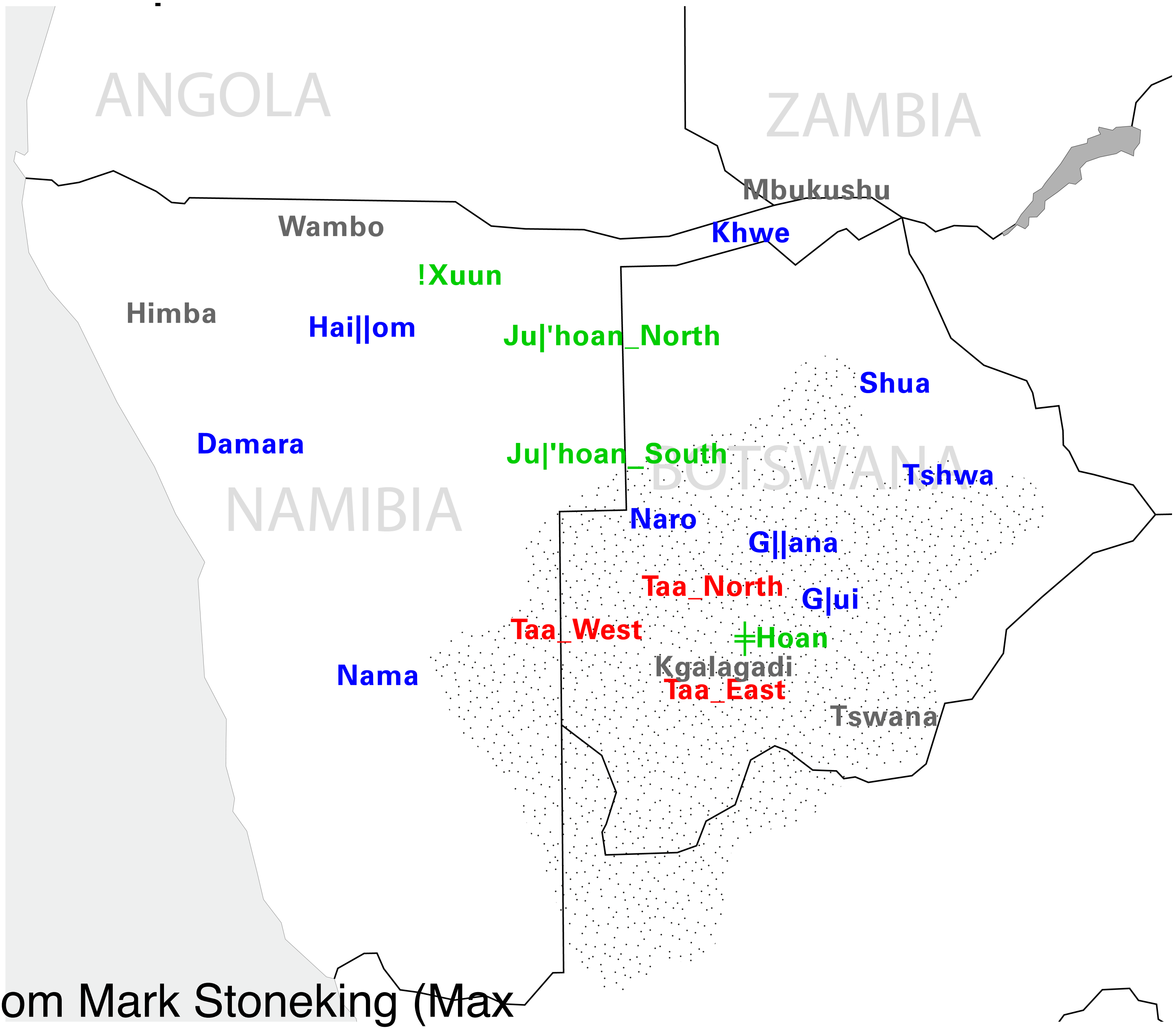
Fig. 1. Past and Present Distribution of Bushmen.

Tobias (1964)

Genetics of southern African Khoisan

Genotyped 187 individuals on the Affymetrix Human Origins Array (~500k SNPs in panels of known ascertainment, 150,000 ascertained in the HGDP “San”)

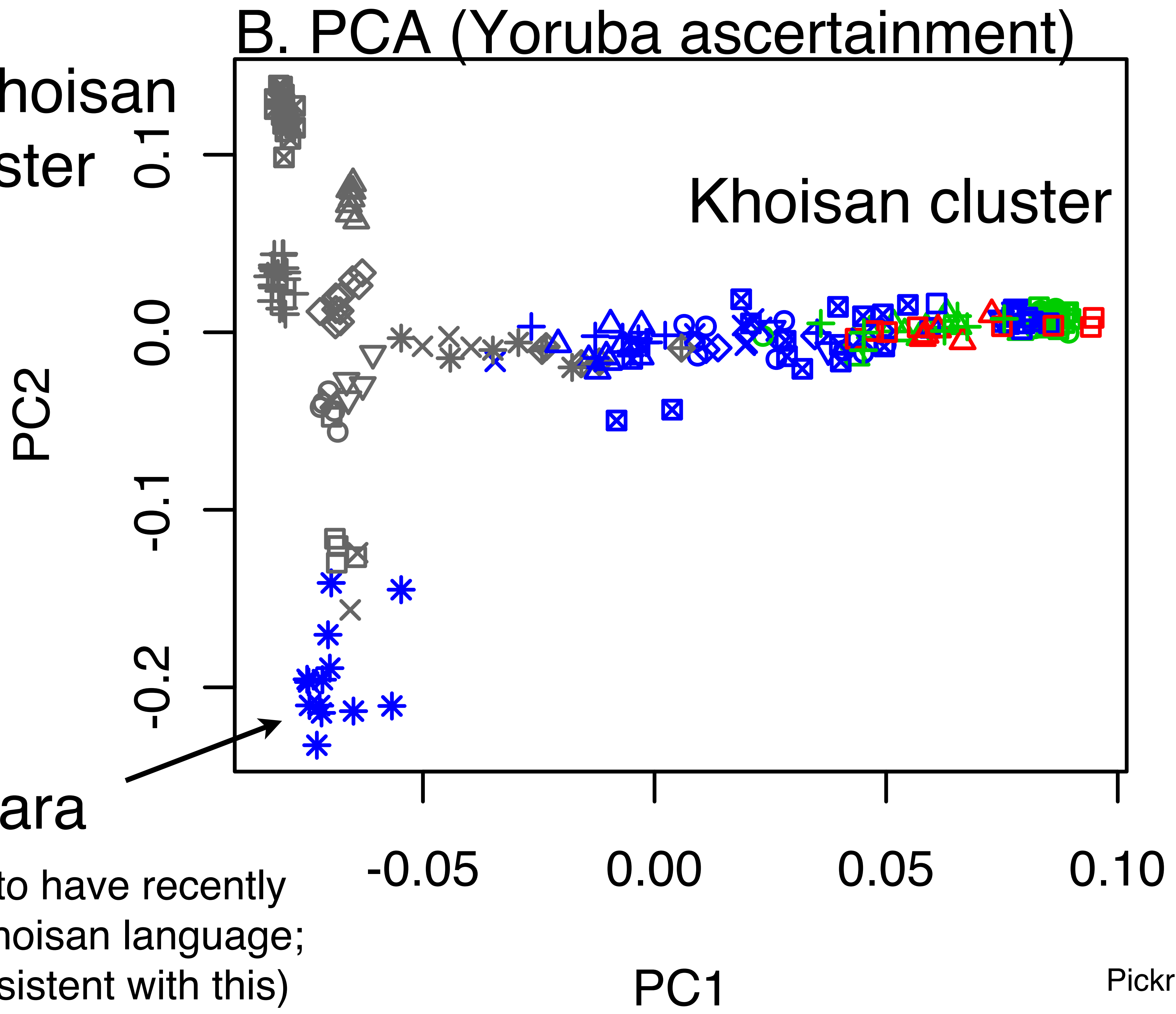
Hadza and Sandawe not shown on the map; populations colored according to linguistic affiliaton



Khoisan samples from Mark Stoneking (Max Plank Institute) and Brigitte Pakendorf (Lyon); Hadza from Brenna Henn (Stanford); Sandawe from Lachance et al. (2012)

Pickrell et al. (2012)

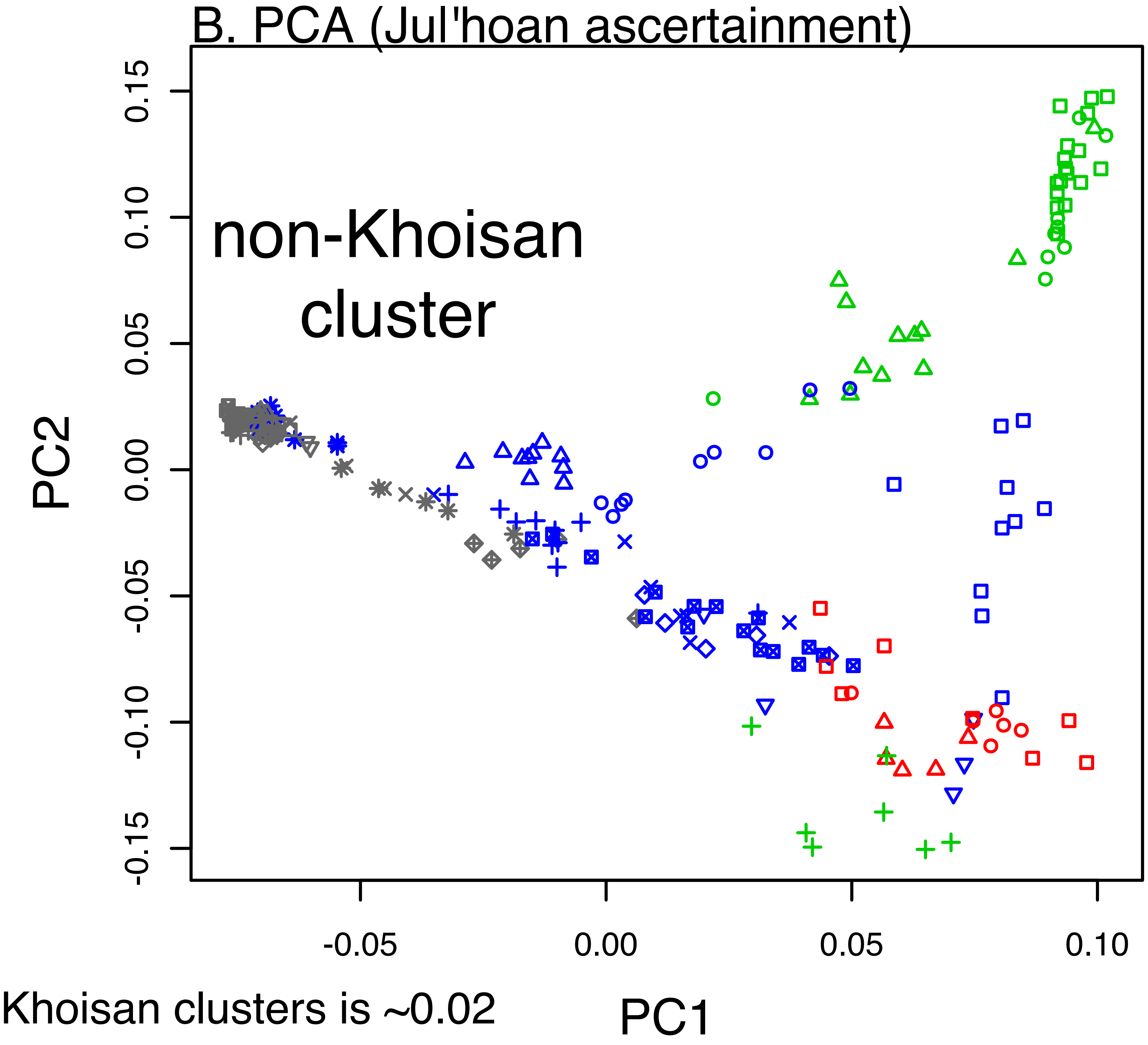
Different SNP panels highlight different aspects of the data



(hypothesized to have recently switched to a Khoisan language; genetics is consistent with this)

Pickrell et al. (2012)

Two genetic clusters of southern African Khoisan



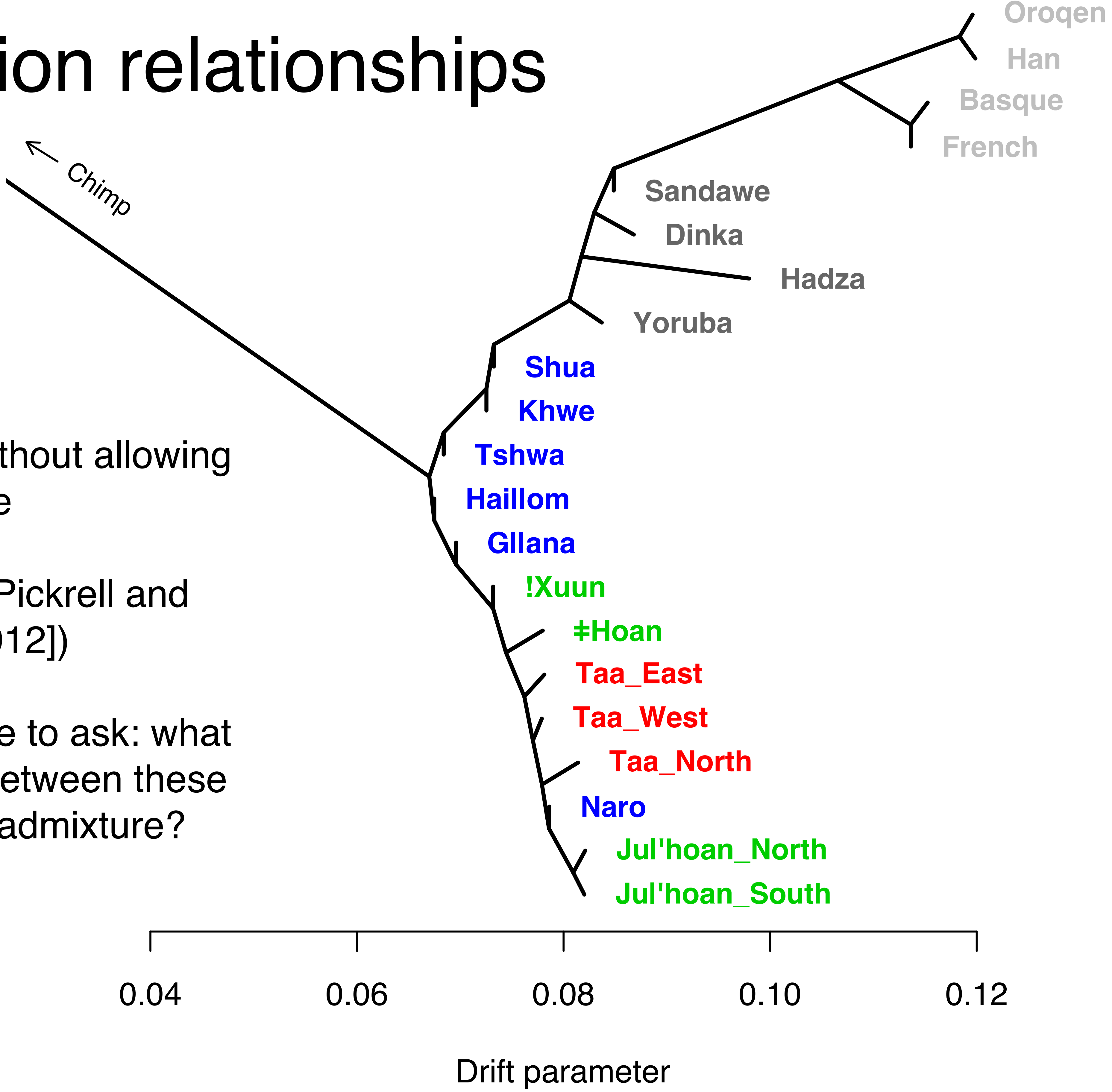
NW Kalahari Khoisan

SE Kalahari Khoisan

Fst between Khoisan clusters is ~0.02 (non trivial!), all Khoisan populations are admixed with non-Khoisan

Pickrell et al. (2012)

Admixture confounds inference of population relationships

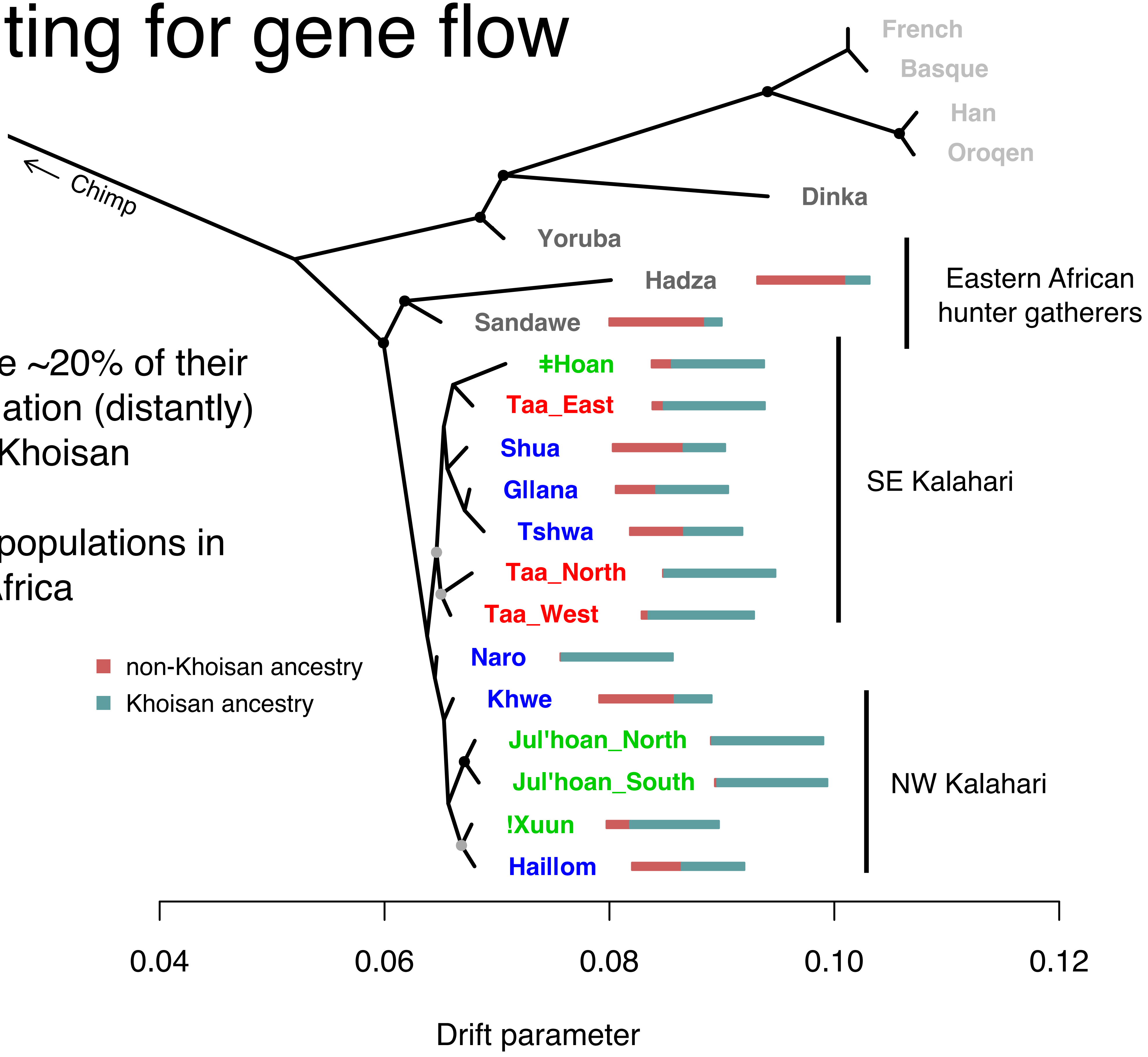


Tree of populations without allowing admixture

Built with *TreeMix* (Pickrell and Pritchard [2012])

Question we would like to ask: what was the relationship between these populations before admixture?

Inferring population relationships accounting for gene flow



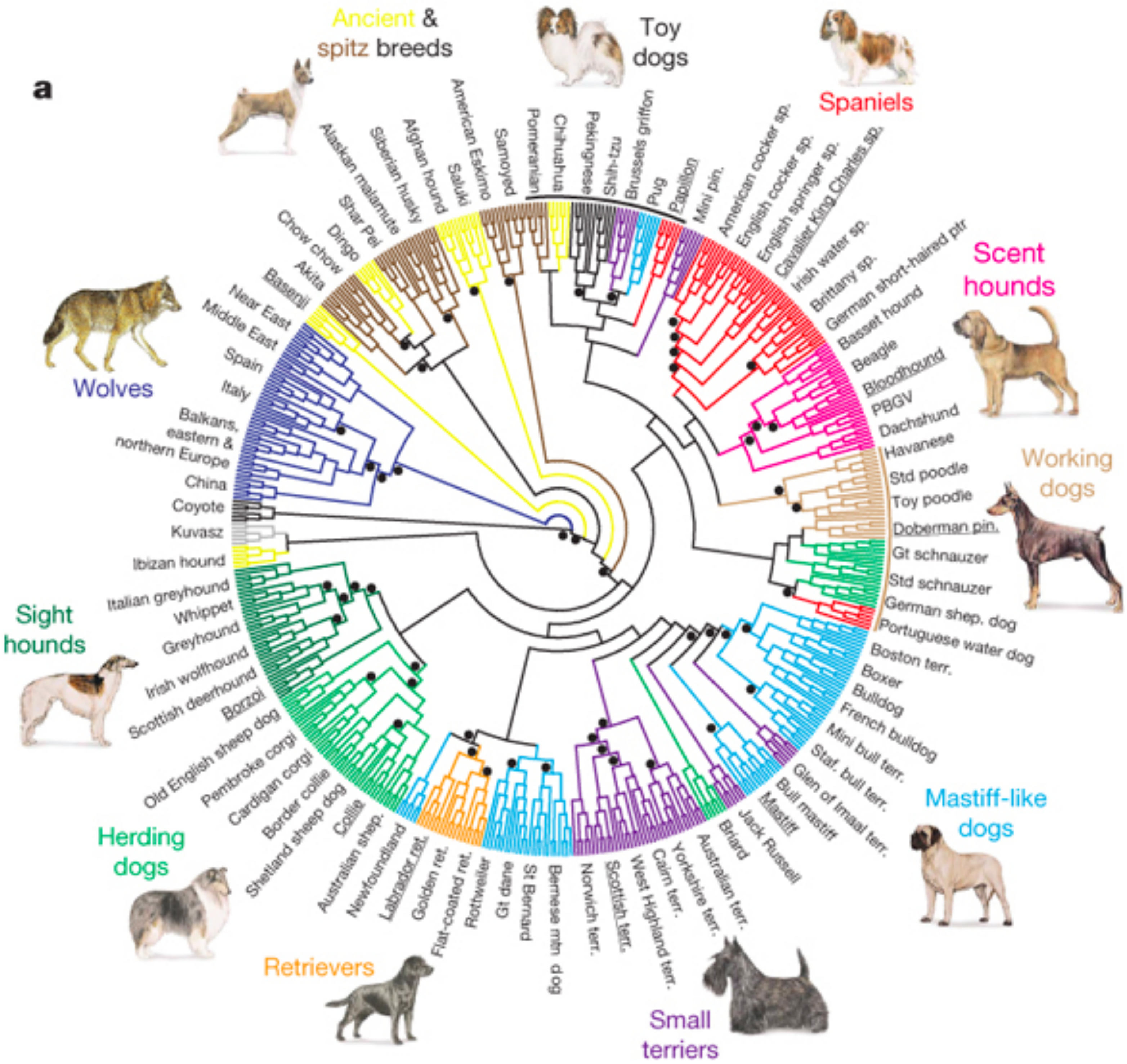
Hadza/Sandawe trace ~20% of their ancestry from a population (distantly) related to the Khoisan

Later split between populations in southern Africa

Pickrell et al. (2012)

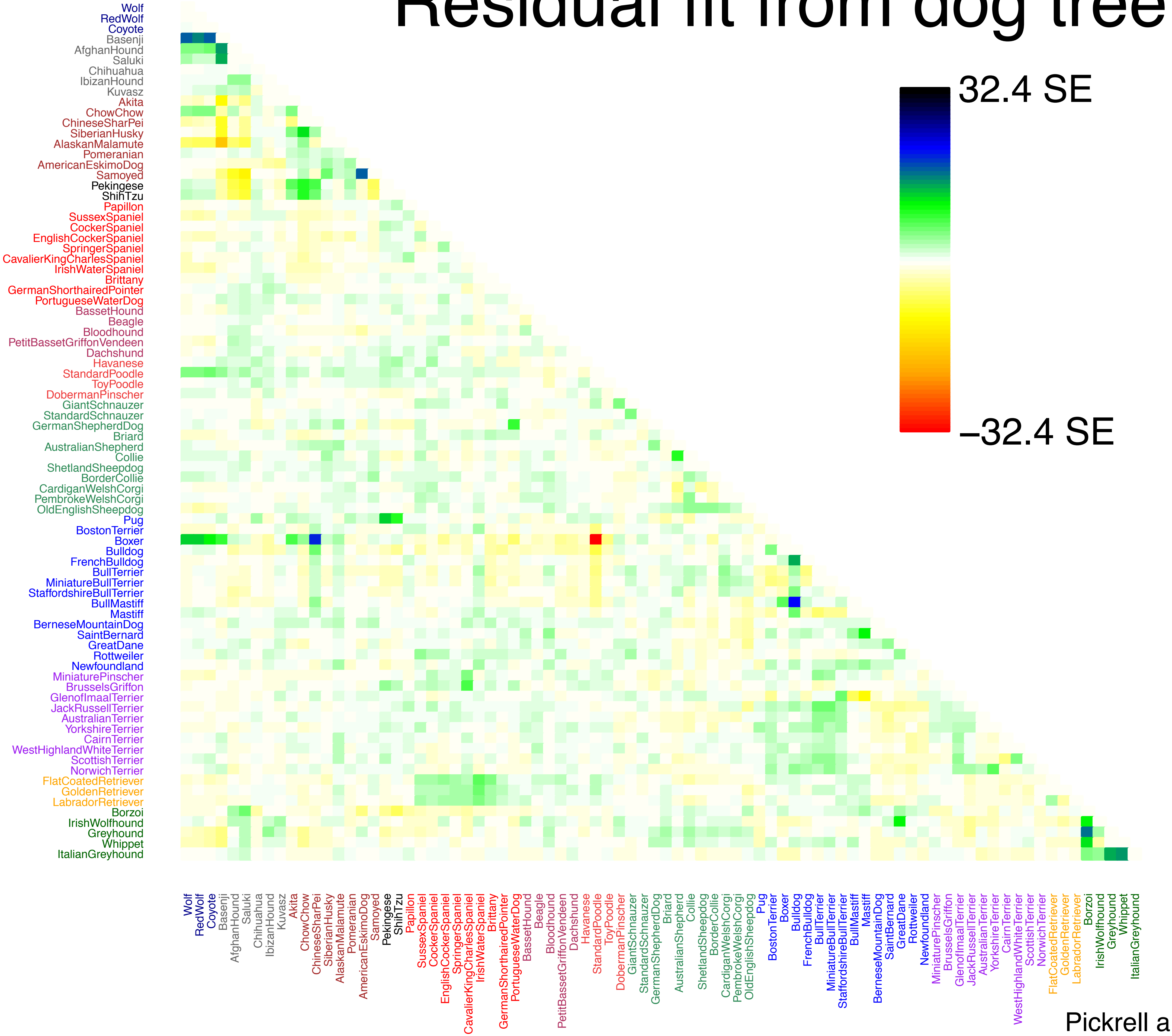
Application 3: dogs and wild canids

- Data from vonHolt et al. (2010): ~60,000 sites genotyped in wolves, coyotes, and 82 dog breeds



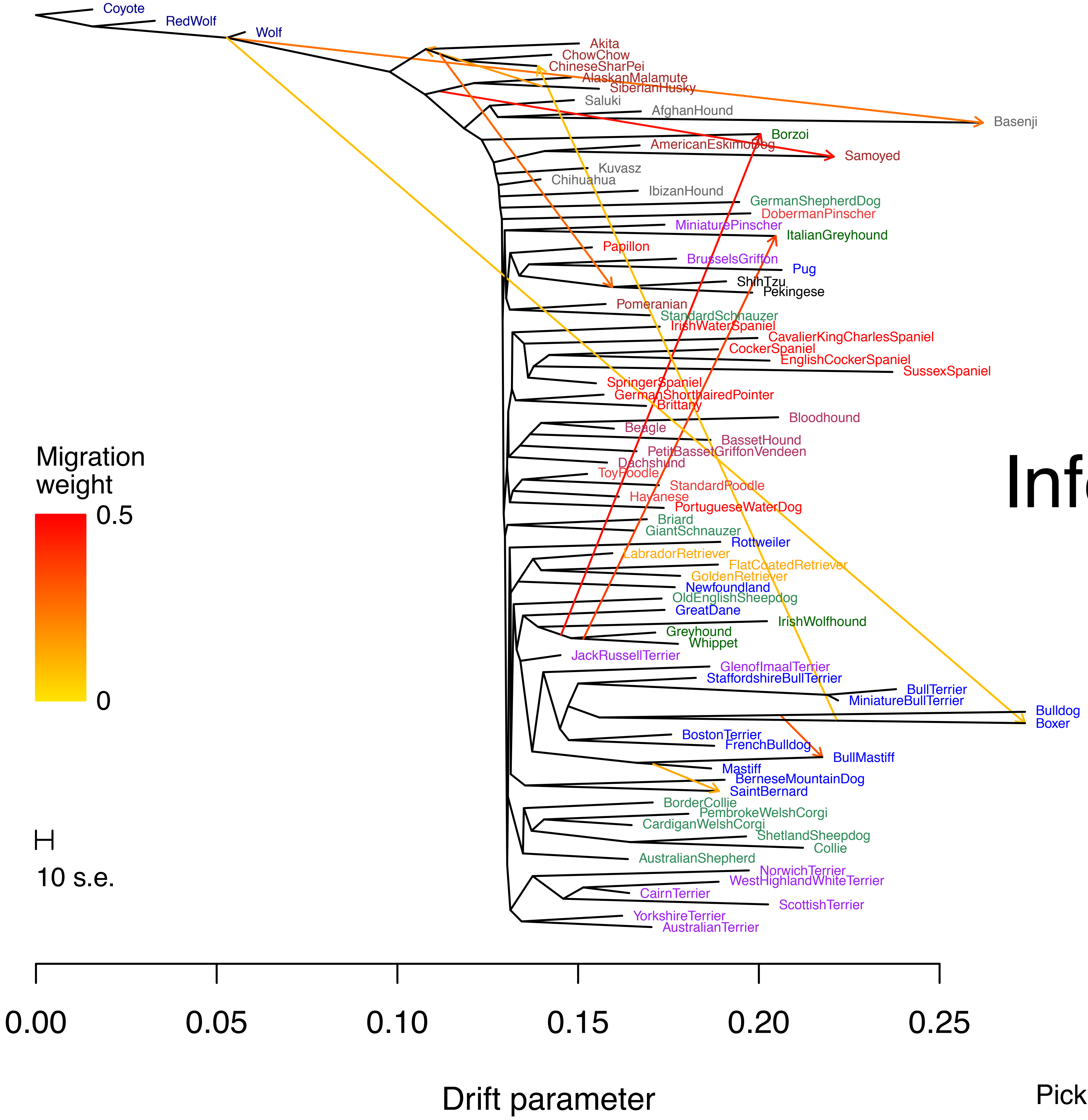
vonHolt et al. (2010)

Residual fit from dog tree



Pickrell and Pritchard (2012)

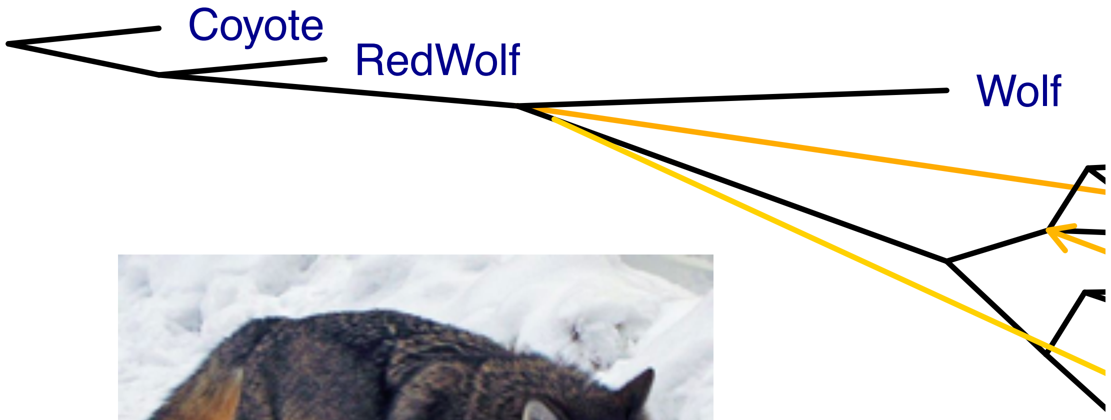
Inferred dog graph

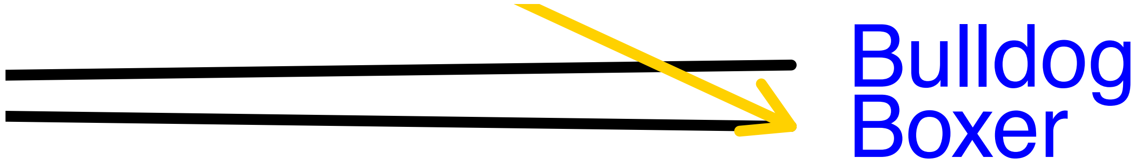


Pickrell and Pritchard (2012)

Wolf contribution to dog breeds since domestication

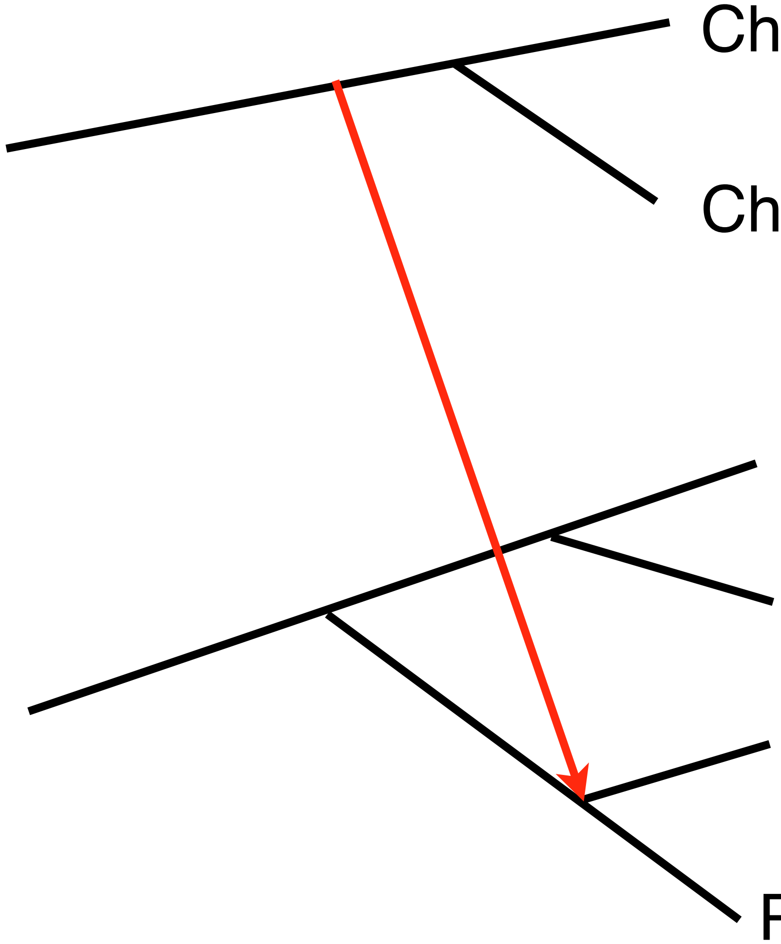
~25% wolf ancestry





~9% wolf ancestry

Origin of East Asian toy breeds



Chow Chow

Chinese shar-pei

Brussels griffon

Pug

ShihTzu

Pekingese

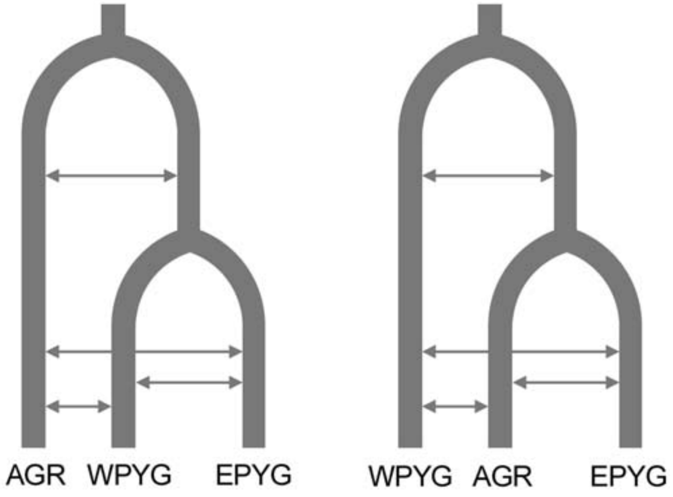
~28% "ancient" East Asian breed

Conclusions (inferring topology)

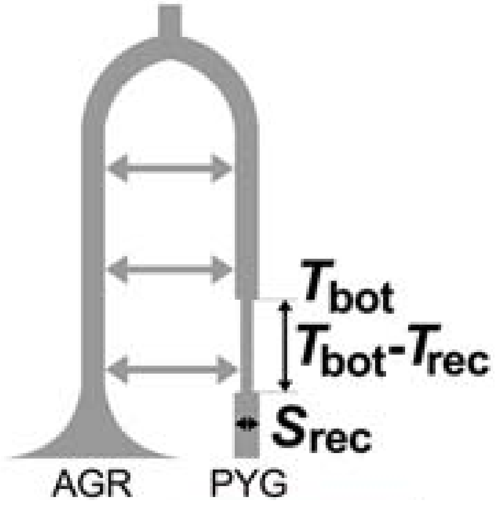
- Using normal approximation to genetic drift, can build trees of many populations
- Tests of “treeness” indicate populations where a tree is a poor model
- Graphs of populations incorporate both population splits and mixtures

What can we hope to learn?

- Topology
  - what is the branching structure of populations?



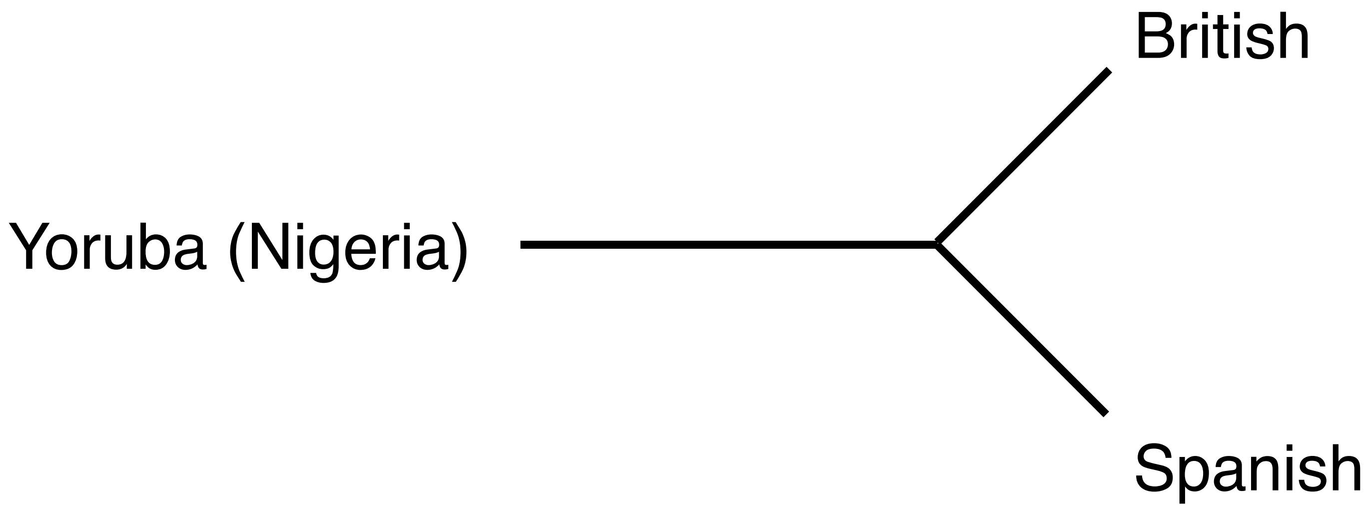
- Demography
  - when did demographic events occur?



Patin et al. (2009)

Returning to human example

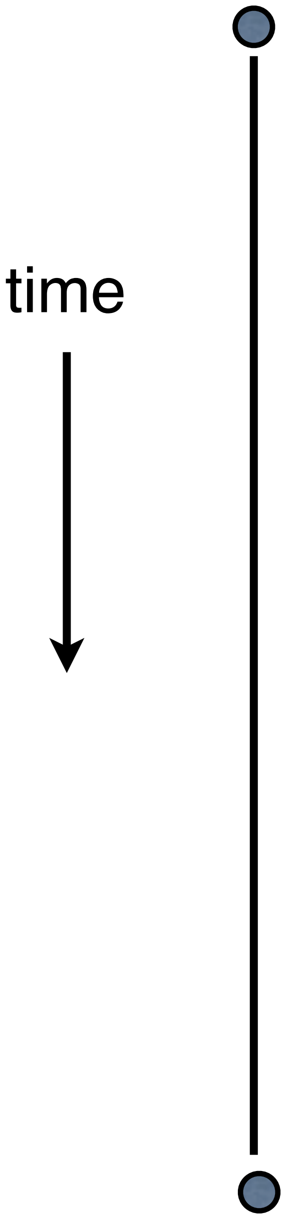
- Consider the following unrooted tree



- Test Spanish for admixture; ie. calculate (S-Y)(S-B). Negative three-population test; Z-score around -20 (p-value negligible)
- What is the interpretation?
- When did admixture occur?

Moorjani et al. (2011)

Fundamental problem when using allele frequency differences to infer times of demographic events



x_A

- Drift depends on both time and effective population size

- True even without normal approximation

$X_t \sim N(x_A, cx_A[1 - x_A])$

$$c \approx \frac{t}{2N_e}$$

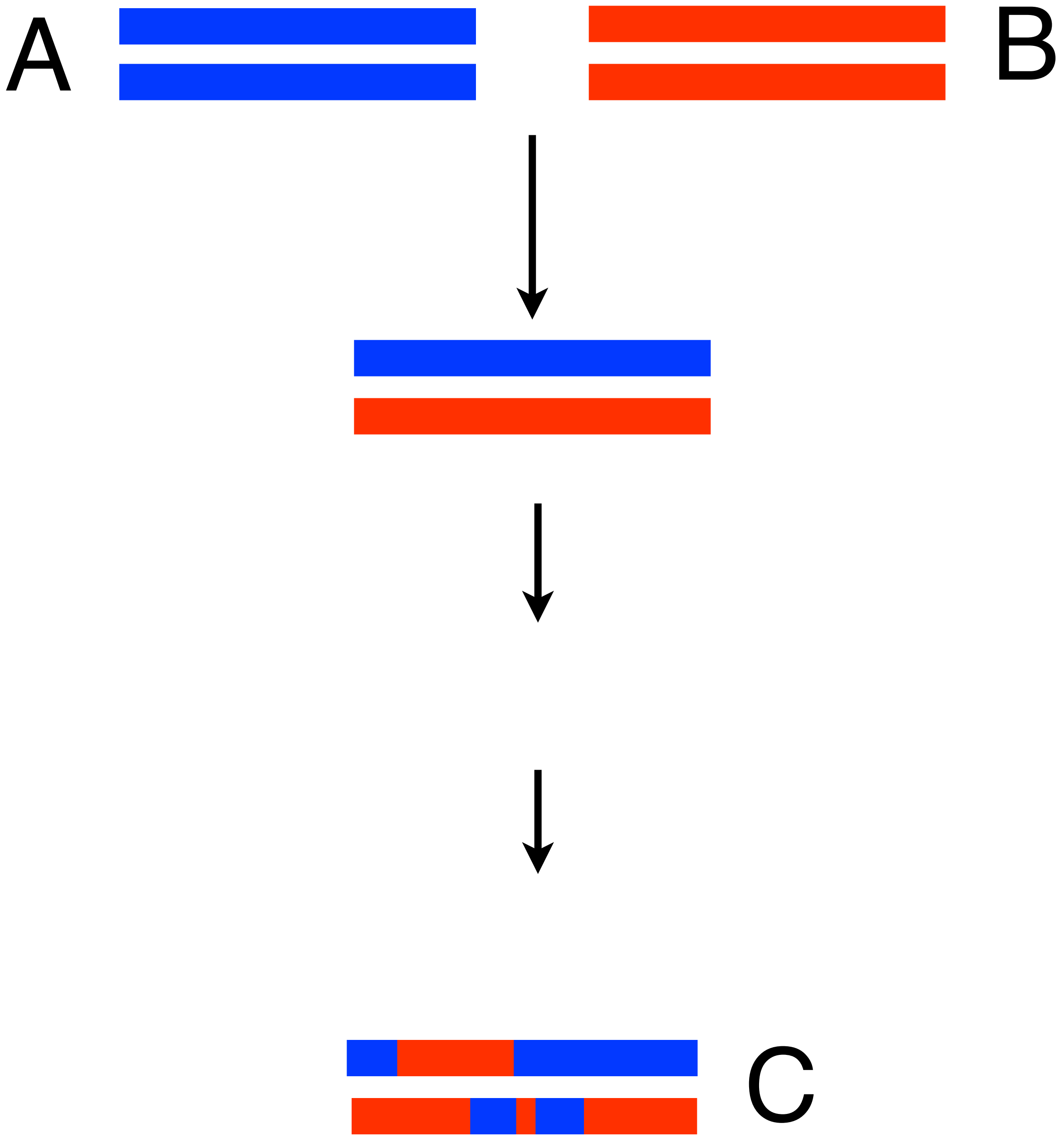
What sources of information do we have?

- Allele frequencies--more closely related population have more similar allele frequencies. E.g. clustering algorithms (STRUCTURE/PCA), tree-building algorithms

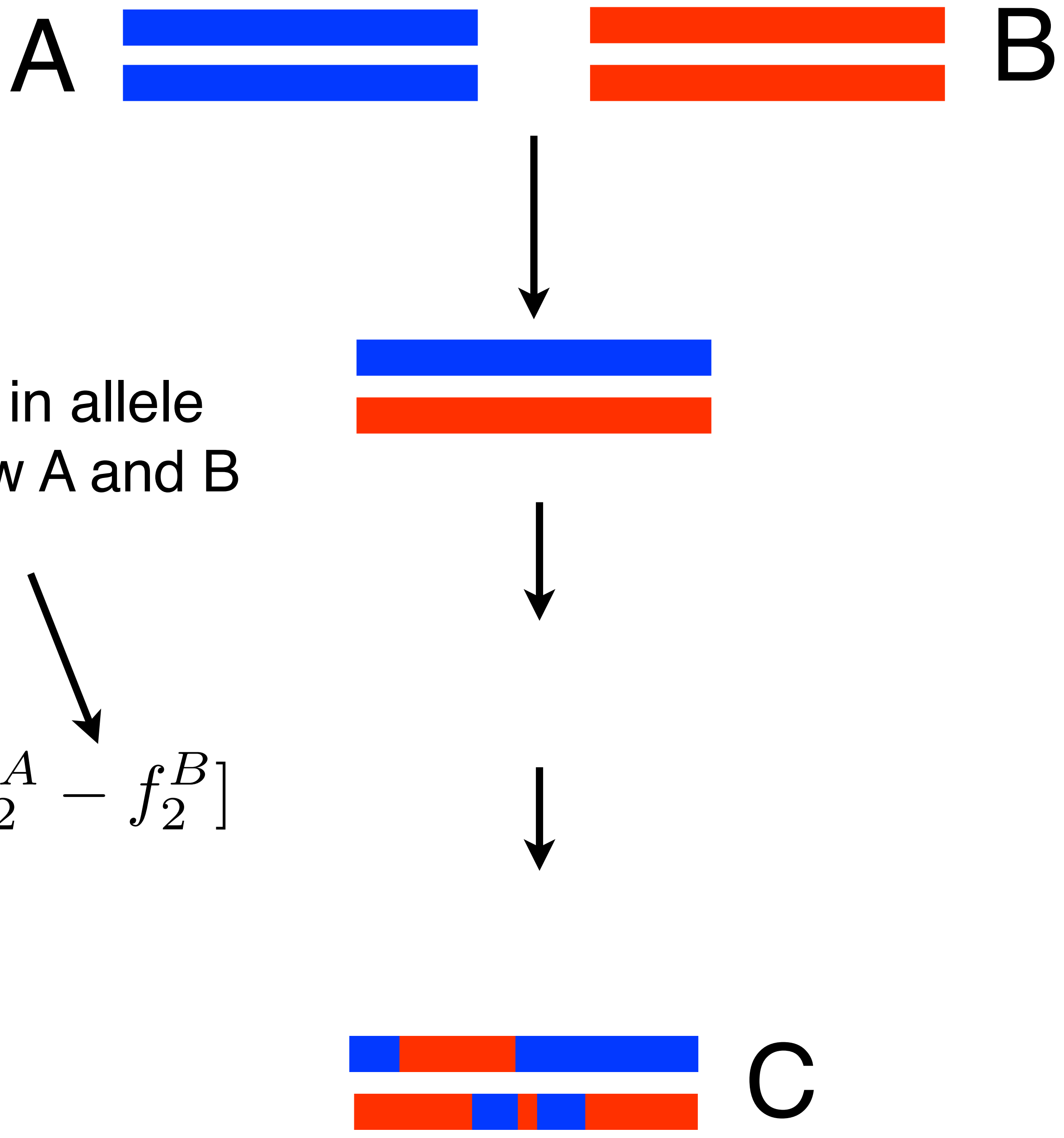
- Linkage disequilibrium--influenced by mixture between populations. E.g. local ancestry inference, ROLLOFF

- Mutations--shared rare mutations between populations indicate shared history. E.g. mtDNA trees (not going to cover this)

Learning about history from genetics (admixture times)



Learning about history from genetics (admixture times)



Amount of LD in C after t generations between two sites at genetic distance x

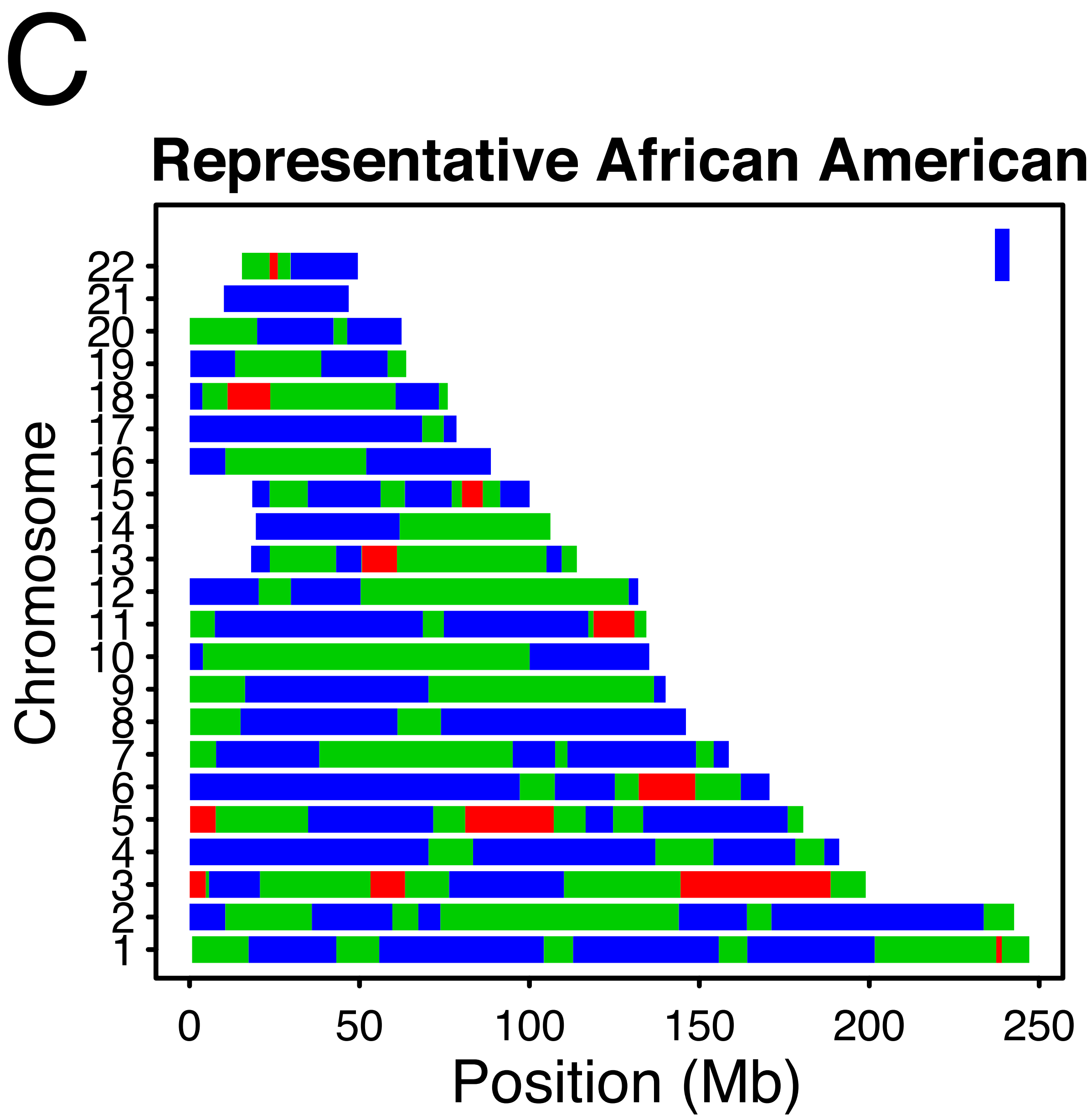
Differences in allele frequency btw A and B

$$D_{12}^{C}(t) \approx \alpha(1-\alpha)e^{-tx}[f_1^A - f_1^B][f_2^A - f_2^B]$$

number of generations since mixture

Chakraborty and Weiss (1988)

Recently admixed populations



Bryc et al. (2009)

Variation in ancestry within a population can cause problems in association analysis (like population structure)

Often useful to know which sections of an individual's genome have recent ancestry from different populations

Called "local" ancestry inference (as opposed to "global" ancestry)

Software for local ancestry inference

- HapMix (http://www.stats.ox.ac.uk/~myers/software.html)
- LAMP (http://lamp.icsi.berkeley.edu/lamp/)
- StepPCO (https://bioinf.eva.mpg.de/download/StepPCO/)

HapMix

- Goal: identify which regions of a person’s genome have recent ancestry from one of two reference populations

- Need to write down the likelihood of observed data

- Use Li and Stephens approximation

Li and Stephens (2003)

- Imagine you have data from a number of haplotypes.
- What is the probability of this set of data, given (say, known) mutation rates, recombination rates, etc.?
- Impossible to calculate except in very specific situations

- Li and Stephens approximation:

$$L_{\mathrm{PAC}}(\rho) = \hat{\pi}(h_1|\rho)\hat{\pi}(h_2|h_1;\ \rho)\ \ldots\ \hat{\pi}(h_n|h_1,\ \ldots,\ h_{n-1};\ \rho).$$

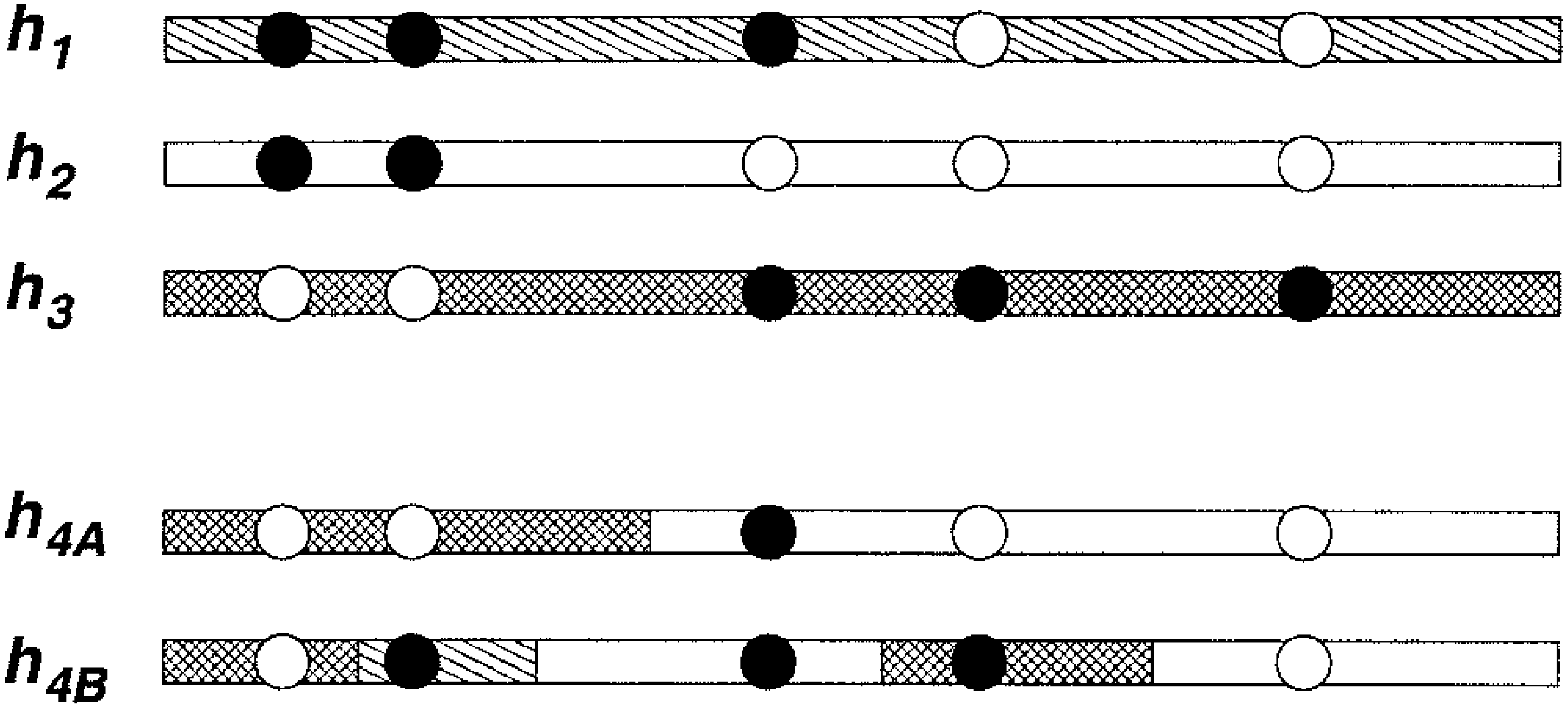
Li and Stephens (2003)

- Imagine you have data from a number of haplotypes.
- What is the probability of this set of data, given (say, known) mutation rates, recombination rates, etc.?
- Impossible to calculate except in very specific situations

- Li and Stephens approximation:

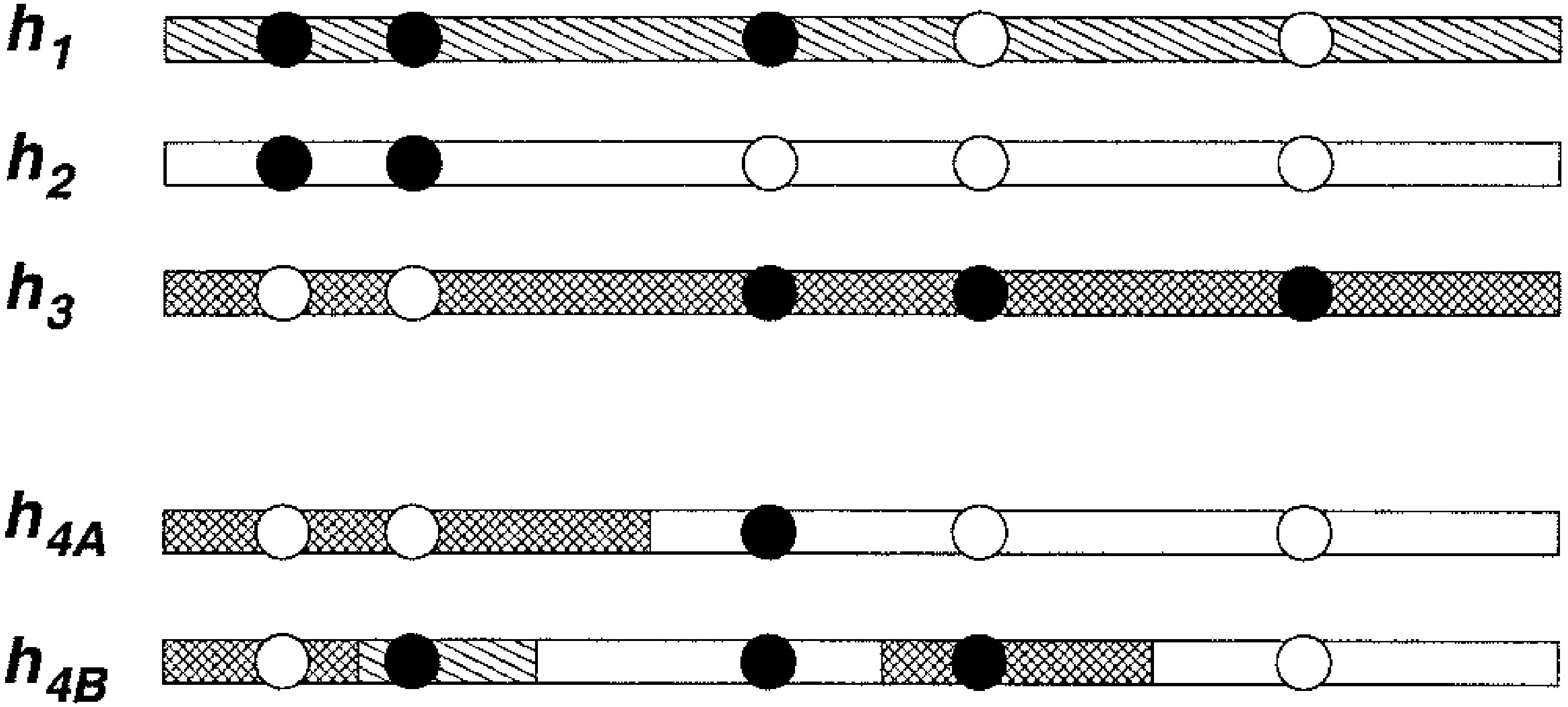
$$L_{\mathrm{PAC}}(\rho) = \hat{\pi}(h_1|\rho)\hat{\pi}(h_2|h_1;\rho)\ldots\boxed{\hat{\pi}(h_n|h_1,\ldots,h_{n-1};\rho).}$$

Li and Stephens (2003)



- Treat each new haplotype as an "imperfect mosaic" of the previously observed haplotypes

Li and Stephens (2003)



Prob. of being in state x’ given you used to be in state x

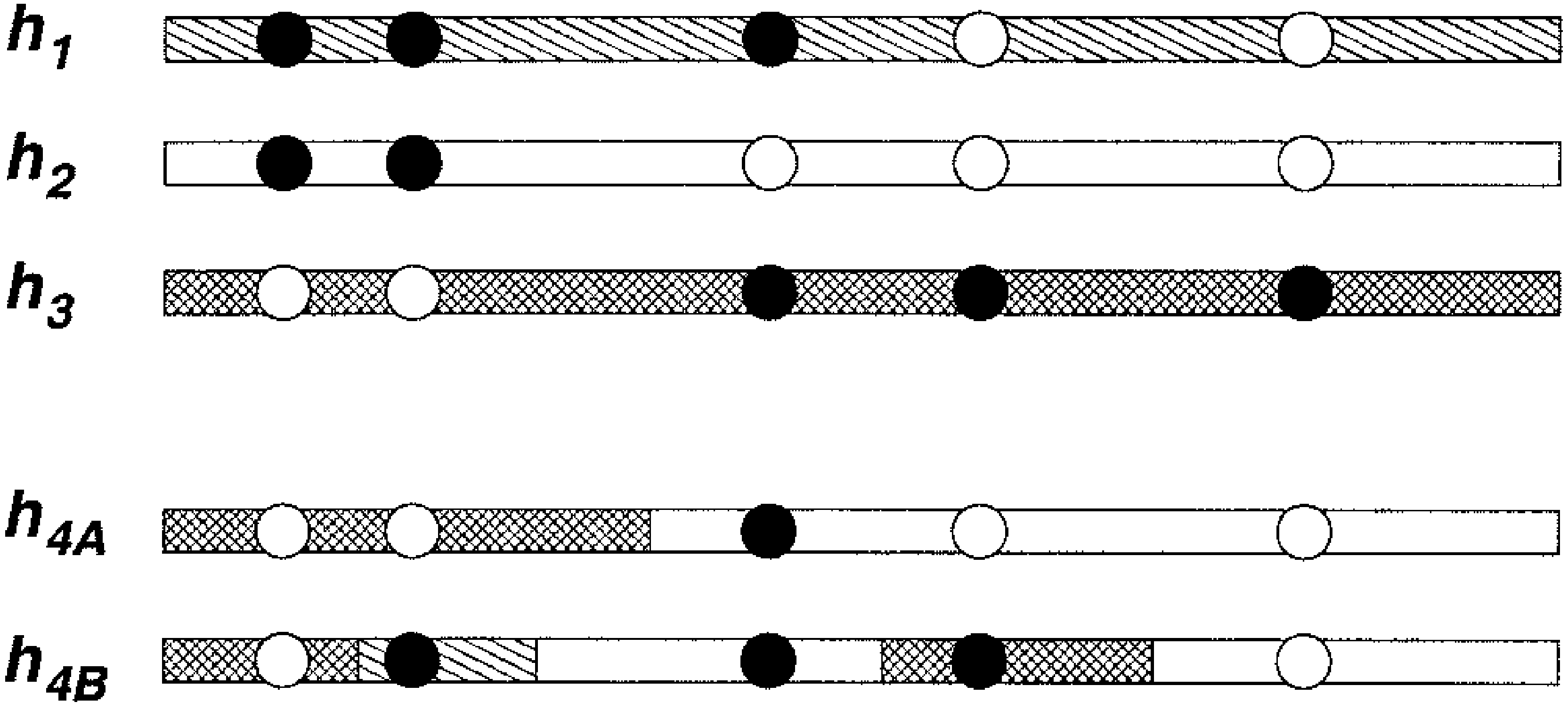
recombination rate

number of samples

$$\Pr(X_{j+1} = x' \mid X_j = x)$$

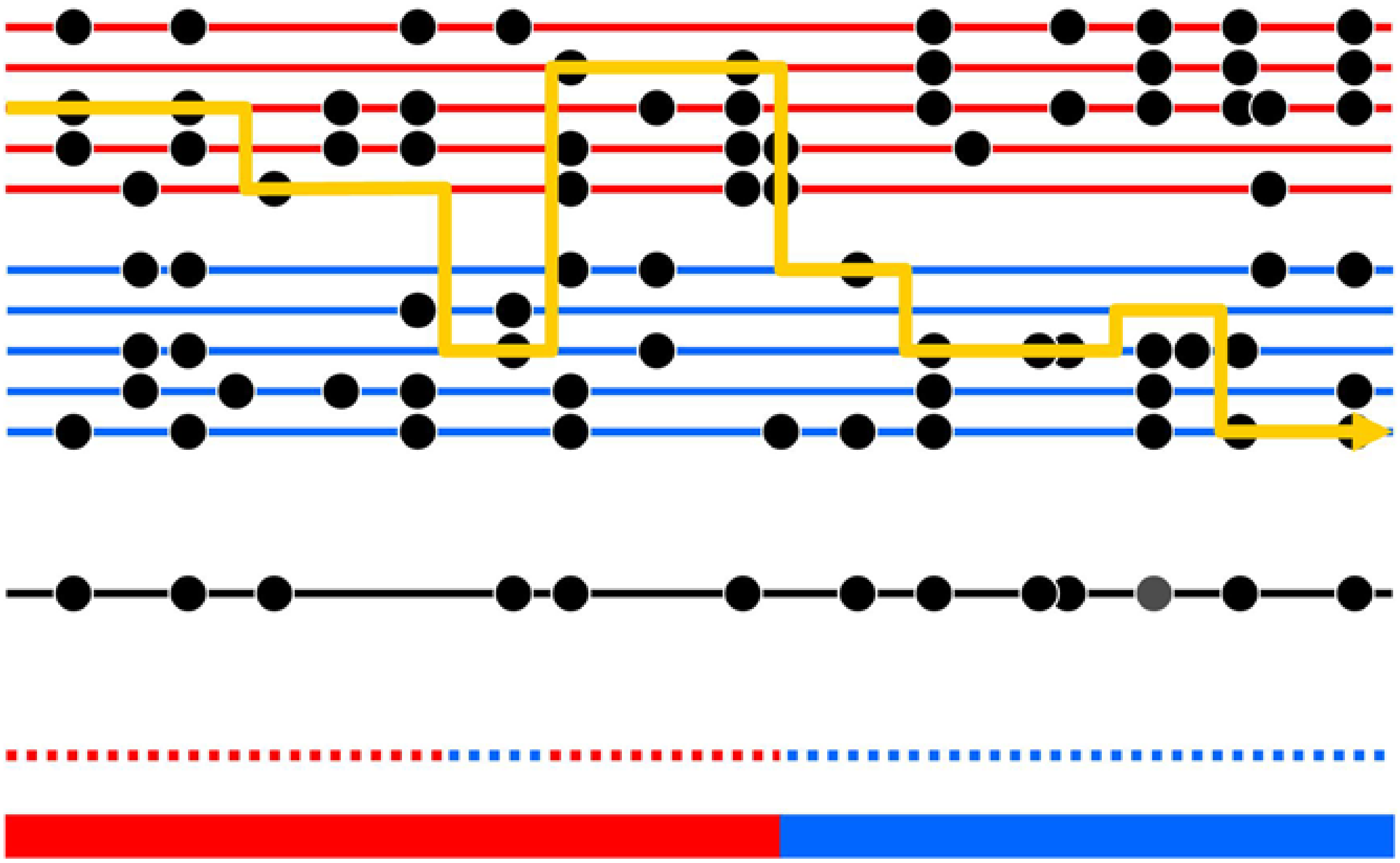
$$= \begin{cases} \exp(-\rho_j d_j/k) + (1 - \exp(-\rho_j d_j/k))(1/k) & \text{if } x' = x; \\ (1 - \exp(-\rho_j d_j/k))(1/k) & \text{otherwise}, \end{cases} \quad \text{(A1)}$$

Li and Stephens (2003)



- Can use hidden Markov model techniques to compute the likelihood of a haplotype given all the previous ones

HapMix (Price et al. 2009)



- Treat each haplotype in the admixed population as a mosaic of haplotypes in two reference populations
- Likelihood depends on the number of generations since admixture, can in principle infer this parameter

Application to the Mozabite

- Mozabite are a mixture of European and African chromosomes
- Short ancestry chunks = older admixture

Price et al. (2009)

Population	No. of samples	Estimated percent European ancestry from HAPMIX	Estimated generations since mixture from HAPMIX
Yoruba	21	0%	N/A
Mandenka	21	2%	120
Mozabite	26	77%	100
Bedouin	45	91%	90
Palestinian	41	93%	75
Druze	39	97%	60
Adygei	16	100%	N/A
Basque	24	100%	N/A
French	28	100%	N/A
Italian	12	100%	N/A
Orcadian	14	100%	N/A
Russian	25	100%	N/A
Tuscan	8	100%	N/A

Population	No. of samples	Estimated percent European ancestry from HAPMIX	Estimated generations since mixture from HAPMIX
Yoruba	21	0%	N/A
Mandenka	21	2%	120
Mozabite	26	77%	100
Bedouin	45	91%	90
Palestinian	41	93%	75
Druze	39	97%	60
Adygei	16	100%	N/A
Basque	24	100%	N/A
French	28	100%	N/A
Italian	12	100%	N/A
Orcadian	14	100%	N/A
Russian	25	100%	N/A
Tuscan	8	100%	N/A

Some of these dates are “too” old. Why?