Model-based approaches to inferring population history

Uma Ramakrishnan

Stanford University, NCBS

Model-based approaches to inferring population history

- Understanding population history
- Methods
- Examples: Ancient DNA and Etruscans
- Examples: Y chromosome STRs and sub-saharan africa
- Examples: STR data and Common Ancestry Profiles
- Conclusion

Reconstructing population history

Genetic variation: shaped by Micro-evolutionary processes

- Drift (effective population size)
- Mutation
- Gene flow
- Selection

Population history: biotic and abiotic environment

- changes in population size ⇒ change in effective size
- changes in movement ⇒ changes in gene flow
- changes in survival of certain types ⇒ mutation, selection

Methods to reconstruct population history

- Frequentist

  summary statistic based methods

  Hypothesis-testing using simulations

  Likelihood

- Bayesian

Frequentist approaches: Summary statistics

Statistics calculated from observed genetic data. e.g. Heterozygosity, F_{st}, number of segregating sites

Equilibrium between mutation, drift and gene flow results in predictable summary statistic value.

Use summary statistic to estimate parameter of interest

e.g. calculate effective population size from heterozygosity

Disadvantage: summary statistic and population parameter relationship based on equilibrium models

Frequentist approaches: Hypothesis testing

Are the observed data consistent with a given hypothesis of population history?

Use computer-based simulations to model genetic data. Calculate summary statistics for simulated data

Repeat to get distribution of simulated data

Determine whether observed data fall within expected distributions

Repeat for different hypotheses

Disadvantage: What if observed data are consistent with different hypotheses?

Frequentist approaches: Likelihood

Likelihood (population parameter/obs data)

e.g. Likelihood (effective size/heterozygosity)

Maximize likelihood: most likely population historic parameter value

Ex FLUCTUATE, IM, MIGRATE

Disadvantage: Must explicitly work out likelihood function, difficult for complex models

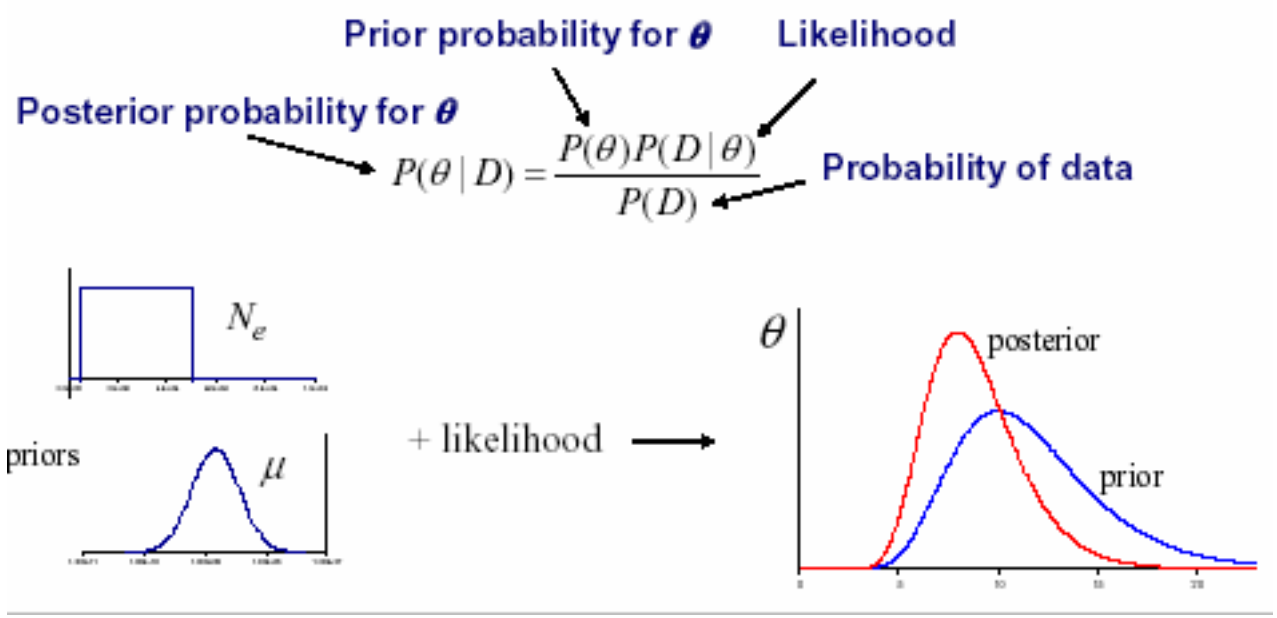
Biased for small sample sizes

Computationally intensive, Model comparison is difficult

Bayesian approaches

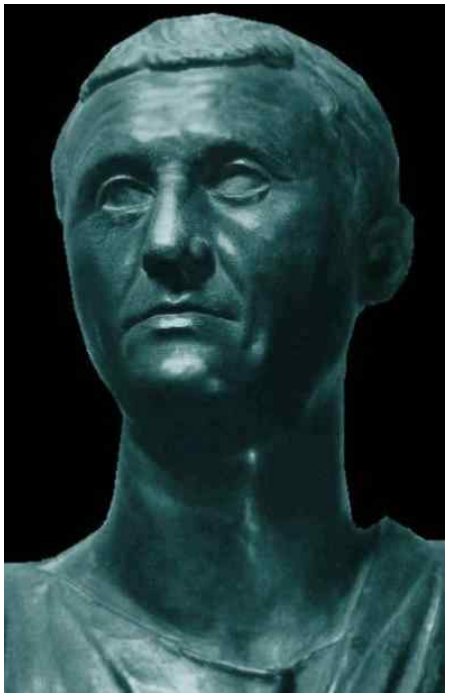
- Use prior data to influence estimate

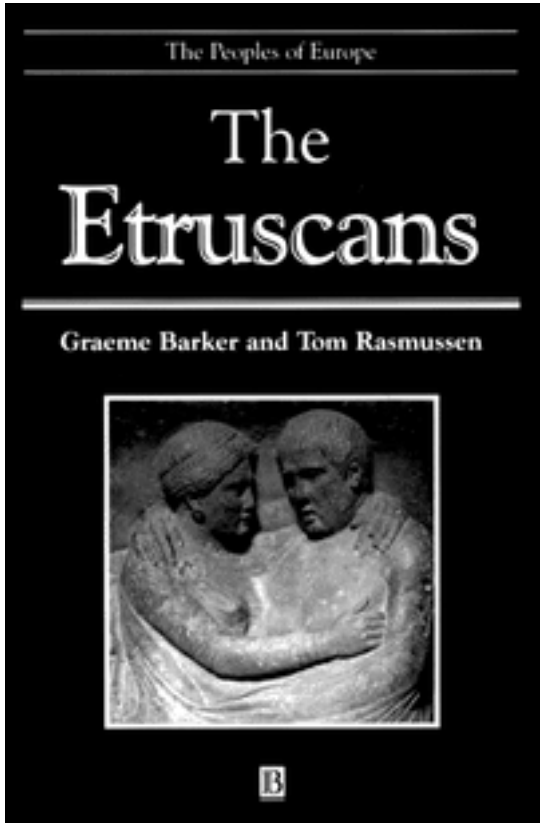
  Ex GENETREE, BATWING



Disadvantage: Not enough model checking
Convergence problems
Computationally intensive methods

Mysterious Etruscans





Etruscan cities established in 1 BC in central Italy
Flourished between 7th and 5th century A.D.
Disappear close to Roman expansion

Figure 1 Map of Italy showing the area of Etruscan influence (*gray*) in the 7th and 6th centuries B.C., from Barker and Rasmussen (1998). A solid line identifies the boundaries of Etruria proper. Solid circles are sampling locations: A, Adria (17 samples, 5 DNA sequences used for statistical analyses); V, Volterra (6, 3); S, Castelfranco di Sotto (2, 1); P, Castelluccio di Pienza (1, 1); M, Magliano and Marsiliana (25, 6); T, Tarquinia (18, 5); C, Capua (8, 6). Additional samples that yielded no amplifiable DNA were from Castelnuovo Berardenga (1, 0) and Pitigliano (2, 0).

Vernesi et al., 2004

Table 1

Consensus HVR-I Mitochondrial Sequences in 28 Etruscan Individuals

Site	Century (B.C.)	Haplotype Label	HVR-I Motif (16024–16384)	14766 *Mse*I	N_{SH}
Volterra	6th–5th	1V	193-219	−	0
Volterra[a]	3rd–2nd	2V	069-186-189-223-319-362	−	0
Volterra	2nd–1st	3V	189-274-334-356	−	0
Volterra	6th–5th	4V	261	+	7
Adria	5th–4th	5AM	CRS	−	32
Adria	5th–4th	6AM	126	+	8
Adria	5th–4th	7AC	126-193-278	+	0
Adria	5th–4th	8A	129	−	10
Adria	5th–4th	9A	223	NA	9
Capua[a]	3rd	10C	189-311-356	−	0
Capua	3rd	11C	069-095-223-261	−	0
Capua	3rd	12C	126-274-356	−	0
Capua	3rd	13C	193-219-356	+	0
Capua	3rd	7AC	126-193-278	+	0
Capua	3rd	14CMT	126-193	+	0
Castelluccio di Pienza	?	15P	193-219-256-270-291	−	0
Castelfranco di Sotto	?	16S	189-356	−	4
Magliano/Marsiliana	6th	17M	***095G***-126-189	−	0
Magliano/Marsiliana	7th	18M	066-126-193-219	−	0
Magliano/Marsiliana	6th	19M	311	−	26
Magliano/Marsiliana	6th	6AM	126	−	0
Magliano/Marsiliana	6th	14CMT	126-193	+	0
Magliano/Marsiliana[a]	7th–6th	5AM	CRS	NA	0
Tarquinia	3rd	20T	126-229-362	+	0
Tarquinia	5th	14CMT	126-193	+	0
Tarquinia	3rd	21T	126-193-228-229-278	+	0
Tarquinia	5th	22T	278-334	+	0
Tarquinia	3rd	23T	098-311-327	+	0

NOTE.—CRS is the Cambridge reference sequence (Anderson et al. 1981). The HVR-I motif is the position (−16,000) where substitutions were observed, with respect to the CRS; the only observed transversion is in boldface italic type. In the haplotype labels, capital letters indicate the site(s) where the haplotype was observed: A, Adria; C, Capua; M, Magliano and Marsiliana; P, Castelluccio di Pienza; S, Castelfranco di Sotto; T, Tarquinia; V, Volterra. The designation "14766 *Mse*I" indicates the presence (+) or absence (−) of a diagnostic restriction cut. N_{SH} is the number of modern populations sharing that haplotype, among the 34 in the database. Haplotype 2V was excluded from the statistical analyses. NA = not available.

[a] Samples for which DNA was independently reextracted and retyped in Barcelona.

Vernesi et al., 2004

mitochondrial and ancient DNA

Maternally inherited
Present in large numbers in cells
No recombination
High mutation rate
Used extensively to reconstruct human population history.

Ancient DNA: tends to be degraded
Best results with high copy number genes like mtDNA
Many factors involved in DNA preservation: temperature, precipitation etc.
Reliable DNA extracted from upto 100,000 year old

Results from genetic comparisons

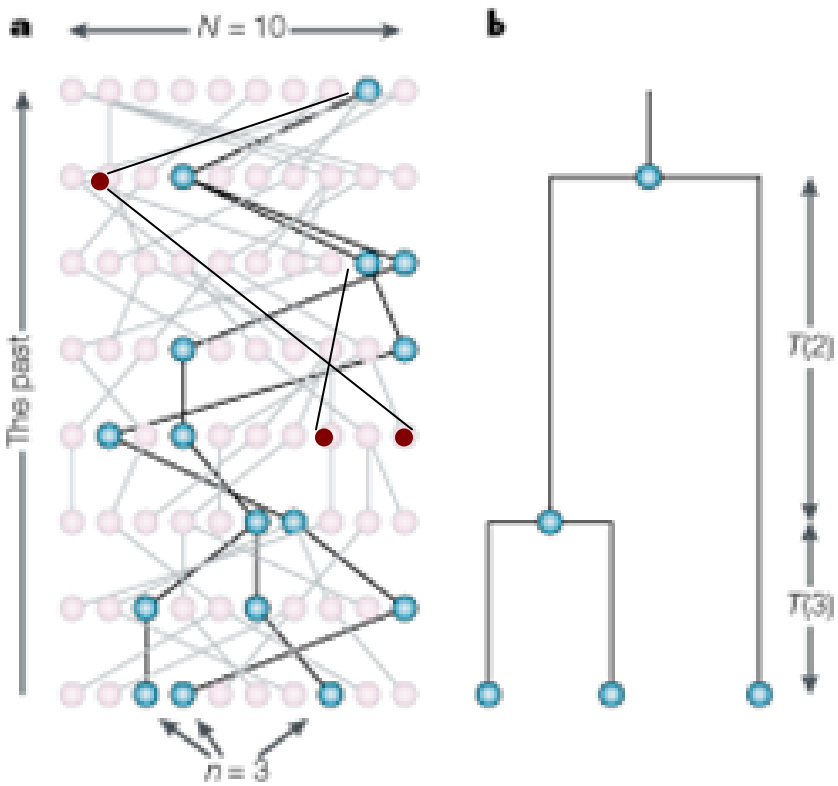
Sequenced 260bp of control region for 27 Etruscans:
Etruscans are as variable as other European groups

Compared Etruscans to other European groups:
Etruscans-European genetic distance > any European-European comparison

Q) Are the Etruscans a distinct population, or ancestral to present-day Tuscans?

Vernesi et al., 2004

Modeling temporal data



Serial coalescent process

Observed Statistics

	Etruscans	Tuscans
Sample size	27	49
Haplotype number	22	40
Haplotypic diversity	0.9465 ± 0.0148	0.9487 ± 0.0185
Nucleotidic diversity	0.0109 ± 0.0063	0.0140 ± 0.0077
Average pairwise difference	3.91 ± 2.02	5.03 ± 2.49
Allele sharing *	9.1%	5.0%

- Combined allele sharing: 3.3%
- Nei's genetic distance: 0.19

Belle, Ramakrishnan, Mountain & Barbujani, in pr

Single population models

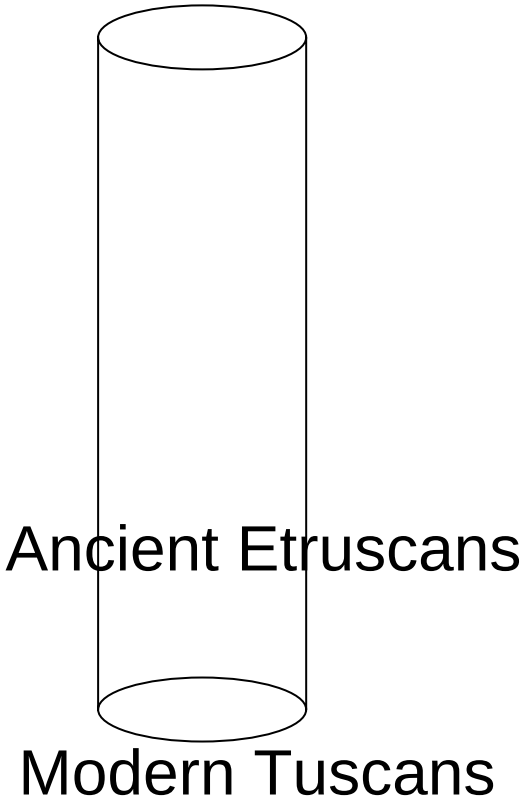
Model 1:A large population of constant size

Model 2: A small population of constant size

Model 3: An expanding population

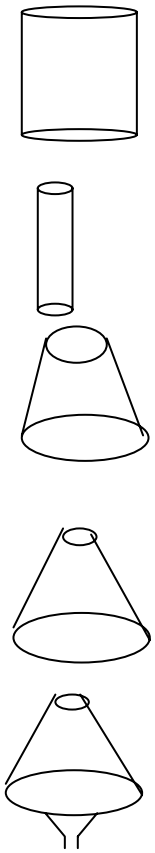
Model 4: Expansion from a small population size

Model 5: Expansion from a small population size followed by a recent population reduction (or selection)



Belle, Ramakrishnan, Mountain & Barbujani, in pr

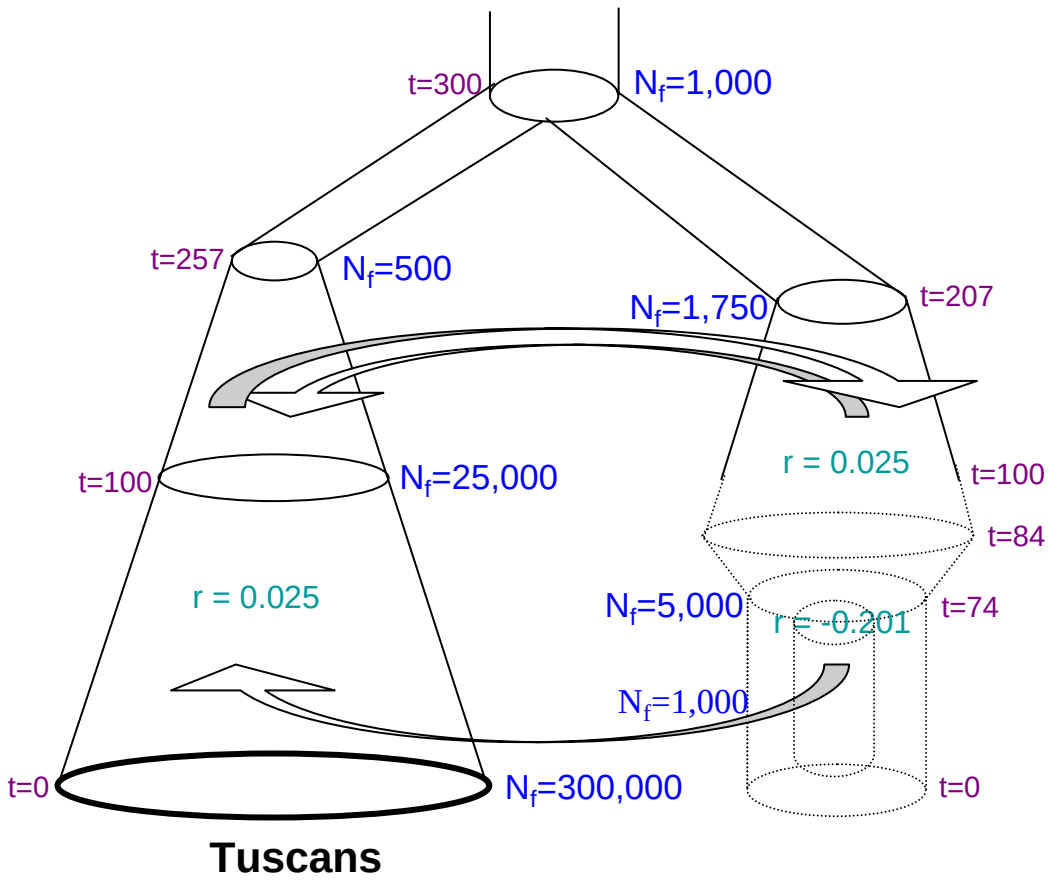
Single population models: Results



Number of Haplotypes		Gene diversity		Nucleotide diversity		Average pairwise difference		Percentage of shared haplotypes		
T	E	T	E	T	E	T	E	T	E	C
47	26	0.9779	0.9602	0.3079	0.3042	110.86	109.52	0.0	0.0	0.0
49	27	0.9796	0.9630	0.3934	0.3956	141.62	142.43	2.1	3.8	1.4
40	23	0.9705	**0.952 0**	0.1176	0.1170	42.32	42.11	**2.2**	**3.7**	**1.4**
46	27	0.9774	**0.963 0**	0.2046	0.2085	73.65	75.07	**12.8**	**21.7**	**8.6**
43	23	0.9738	**0.952 0**	0.1180	0.1138	42.48	40.99	**2.1**	**3.7**	**1.4**
48	27	0.9788	**0.963 0**	0.2019	0.2030	72.70	73.09	**13.3**	**24.0**	**9.2**
29	12	**0.908 7**	**0.792 9**	**0.008 5**	**0.006 0**	**3.05**	**2.15**	10.5	22.2	7.8
41	20	**0.968 0**	**0.935 5**	**0.026 1**	**0.024 0**	**9.38**	**8.65**	26.9	57.1	21.7
31	**15**	**0.941 8**	**0.883 4**	**0.019 1**	**0.016 2**	**6.88**	**5.85**	11.9	23.8	9.1
42	**23**	**0.973 1**	**0.949 2**	**0.054 6**	**0.054 8**	**19.66**	**19.74**	27.9	56.2	22.2
31	**15**	**0.937 1**	**0.883 4**	**0.018 4**	**0.016 1**	**6.62**	**5.79**	12.1	23.5	9.0
42	**22**	**0.972 1**	**0.949 2**	**0.053 7**	**0.054 3**	**19.32**	**19.54**	28.6	56.2	22.7

Belle, Ramakrishnan, Mountain & Barbujani, in pr

Two-population models



No gene flow

Ancient gene flow

Etruscans &

Recent gene flow

Samples: social elite

Belle, Ramakrishnan, Mountain & Barbujani, in pr

Two-population models: Results

	Number of haplotypes		Haplotypes		Nei's		Gene diversity		Nucleotide diversity		Pairwise difference		Percent of shared		distance	
	T	C	E		T	C	E		T	C	E		T	C	E	
No gene flow	30	41	15	23	0.918 0	0.970 4	0.877 9	0.949 2	0.010 68	0.040 7	0.010 9	0.041 3	3.84	14.65	3.91	14.86	0.0	11.4	0.0	21.4	0.0	8.0	0.093	2.83
Ancient gene flow	31	42	15	23	0.928 8	0.972 1	0.880 7	0.952 0	0.012 9	0.039 7	0.011 6	0.039 1	4.63	14.29	4.17	14.07	2.9	16.1	5.6	31.2	2.0	11.8	0.096	3.41
Recent gene flow	31	42	15	23	0.929 6	0.972 1	0.875 0	0.949 2	0.012 8	0.038 6	0.011 6	0.039 0	4.61	13.90	4.19	14.03	2.5	14.3	4.7	28.6	1.7	10.0	0.088	2.79
Continuous gene flow	32	43	15	23	0.928 8	0.972 9	0.883 4	0.952 0	0.016 8	0.071 2	0.016 4	0.070 8	6.03	25.65	5.91	25.48	3.0	16.2	5.6	31.6	2.0	11.8	0.086	3.53
Social elite	32	43	15	23	0.935 4	0.972 9	0.877 9	0.952 0	0.013 5	0.043 2	0.011 0	0.041 3	4.87	15.55	3.95	14.86	4.8	17.1	9.1	33.3	3.3	12.5	0.059	2.59

Belle, Ramakrishnan, Mountain & Barbujani, in pr

Conclusions: Etruscans

- Ancient sampled Etruscans were not the ancestors of the modern Tuscans
- Two population models needed to explain ancient and modern data

Q) How to distinguish between two population models?

Belle, Ramakrishnan, Mountain & Barbujani, in pr

Reconstructing population history in sub-Saharan Africa

- All genetic data point to relatively ancient origin of African groups
- Regions like Tanzania include very high linguistic diversity
- What are the relationships between groups?

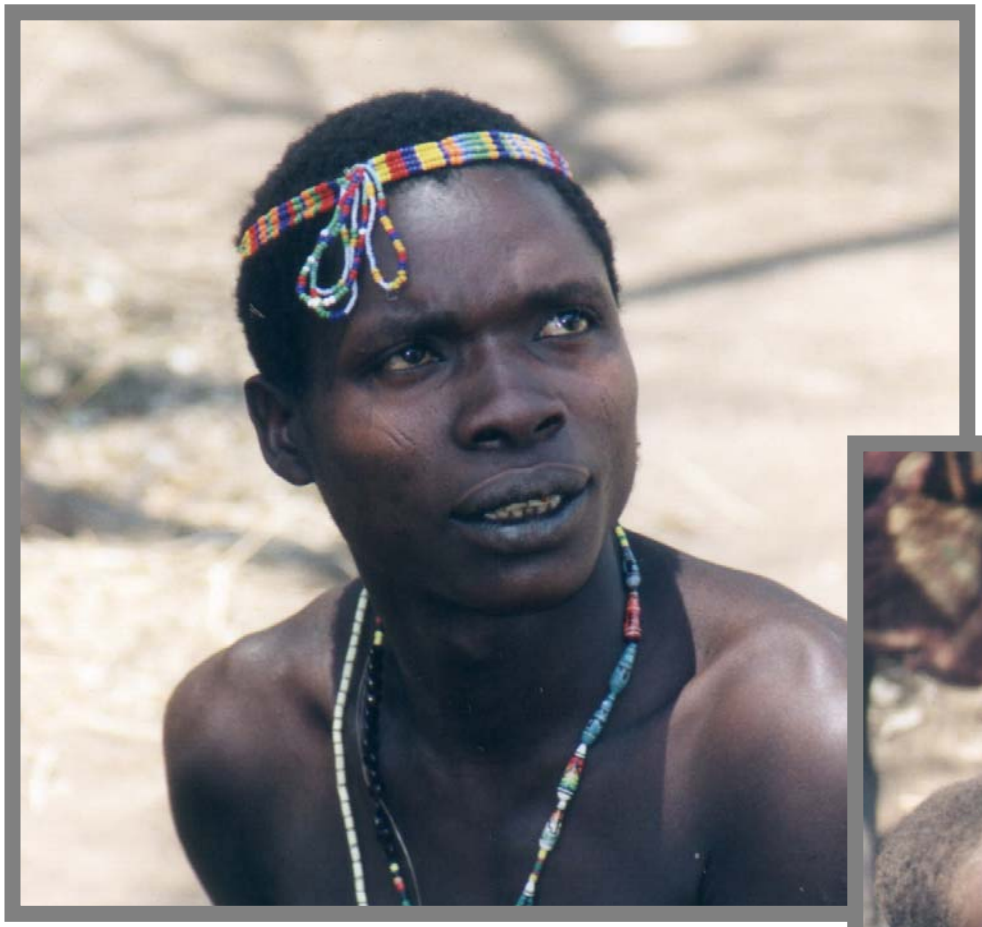
Click speaking vs Bantu speakers

Populations: Click-speakers: Hadzabe, Sandawe

Bantu-speakers: Yoruba

Data: Non-recombining region of Y

Hadzabe (Hadza)



- Foragers of north-central Tanzania
- Small population
- Language includes click consonants

Hadzabe (Hadza)

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Sandawe

Click-speakers

Dodoma region, Tanzania

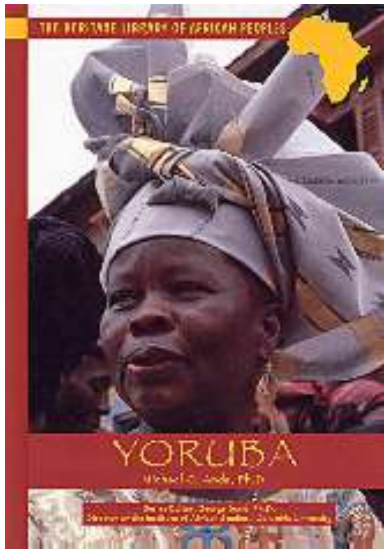
QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

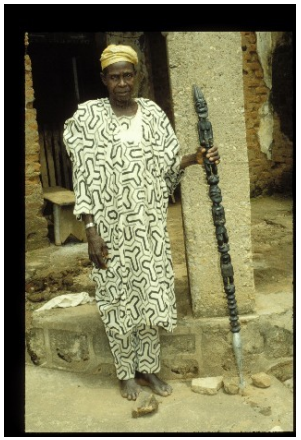
QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Yoruba



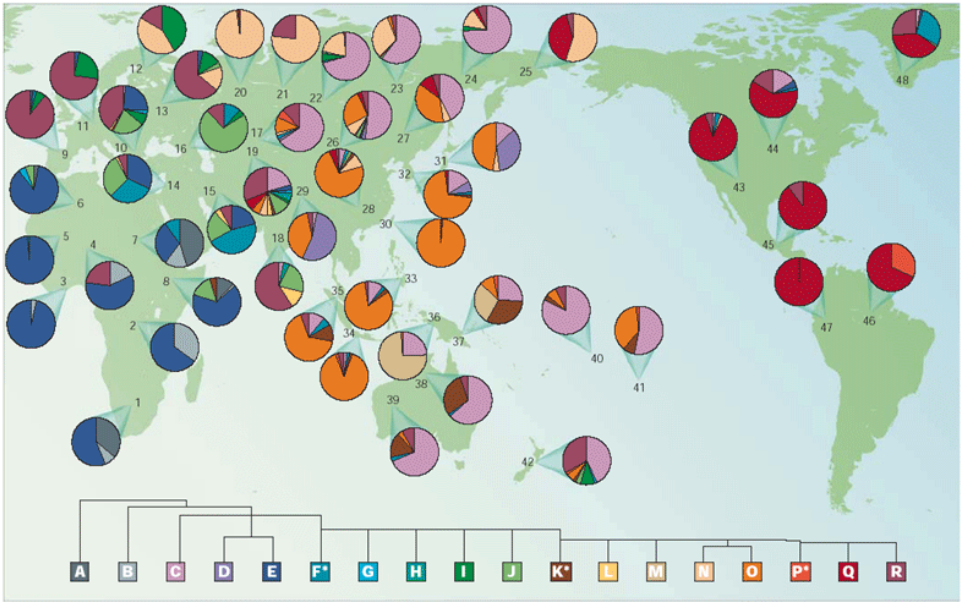


Study populations in Africa

Bantu-speakers: Yoruba Expansion from West Africa

Hadzabe and Sandawe: Lake Eyasi region of north-central Tanzania

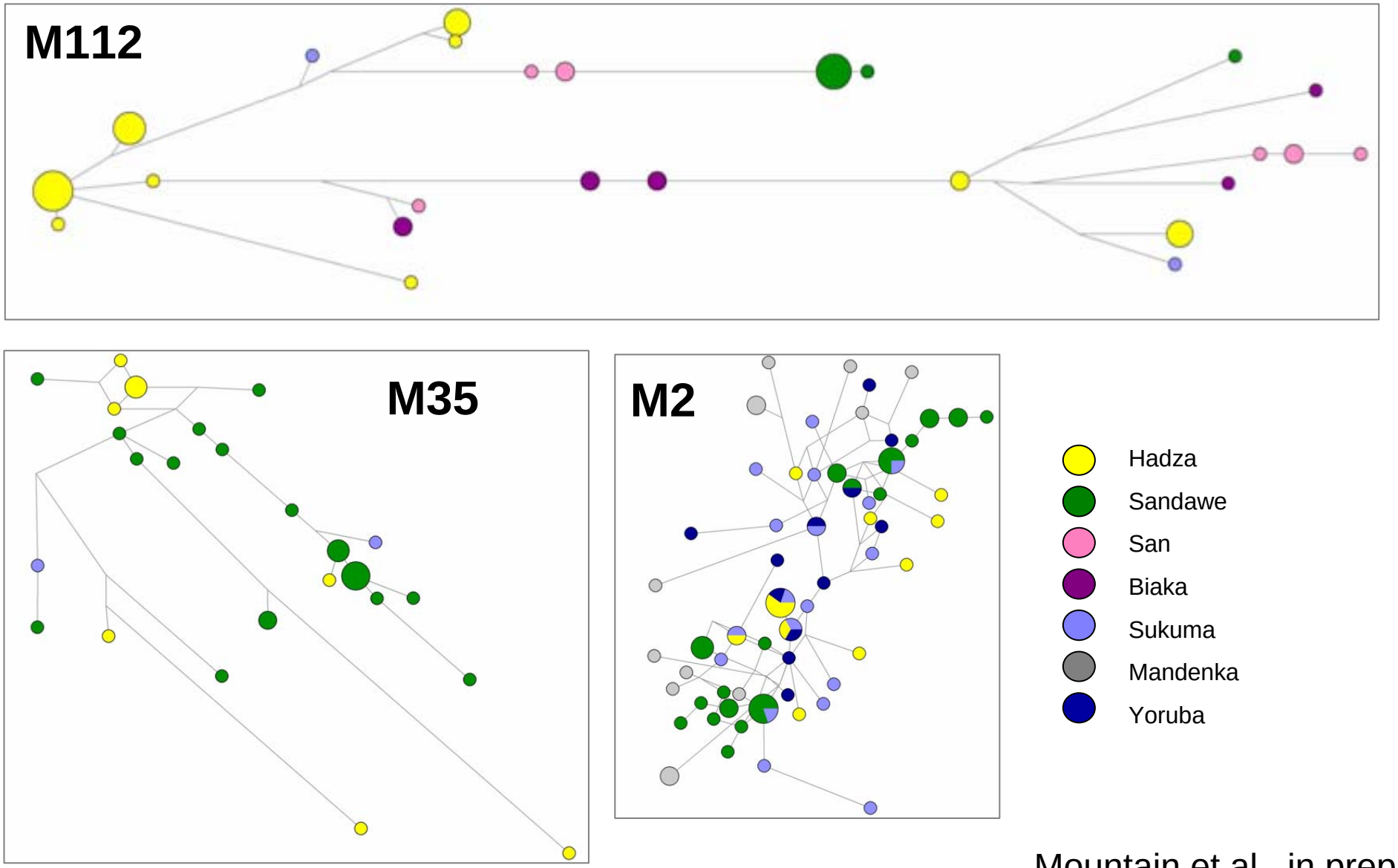
Y chromosome



Males inherit from father as a single, non-recombining unit
Consists of linked UEPs and STRs
UEPs define haplogroups, different ages
Very useful tool to investigate human history

African Y chromosome diversity

Networks of three SNP-defined lineages (11 STR markers)



Mountain et al., in prep

Click-speaking groups in Africa

M112 (oldest): Hadza maintain high diversity

M35 (younger): Sandawe maintain large diversity

M2 (youngest): Bantu-speaking groups high diversity; evidence of population growth

Relationship between click-speaking groups:

Recent common ancestry or deep common ancestry?

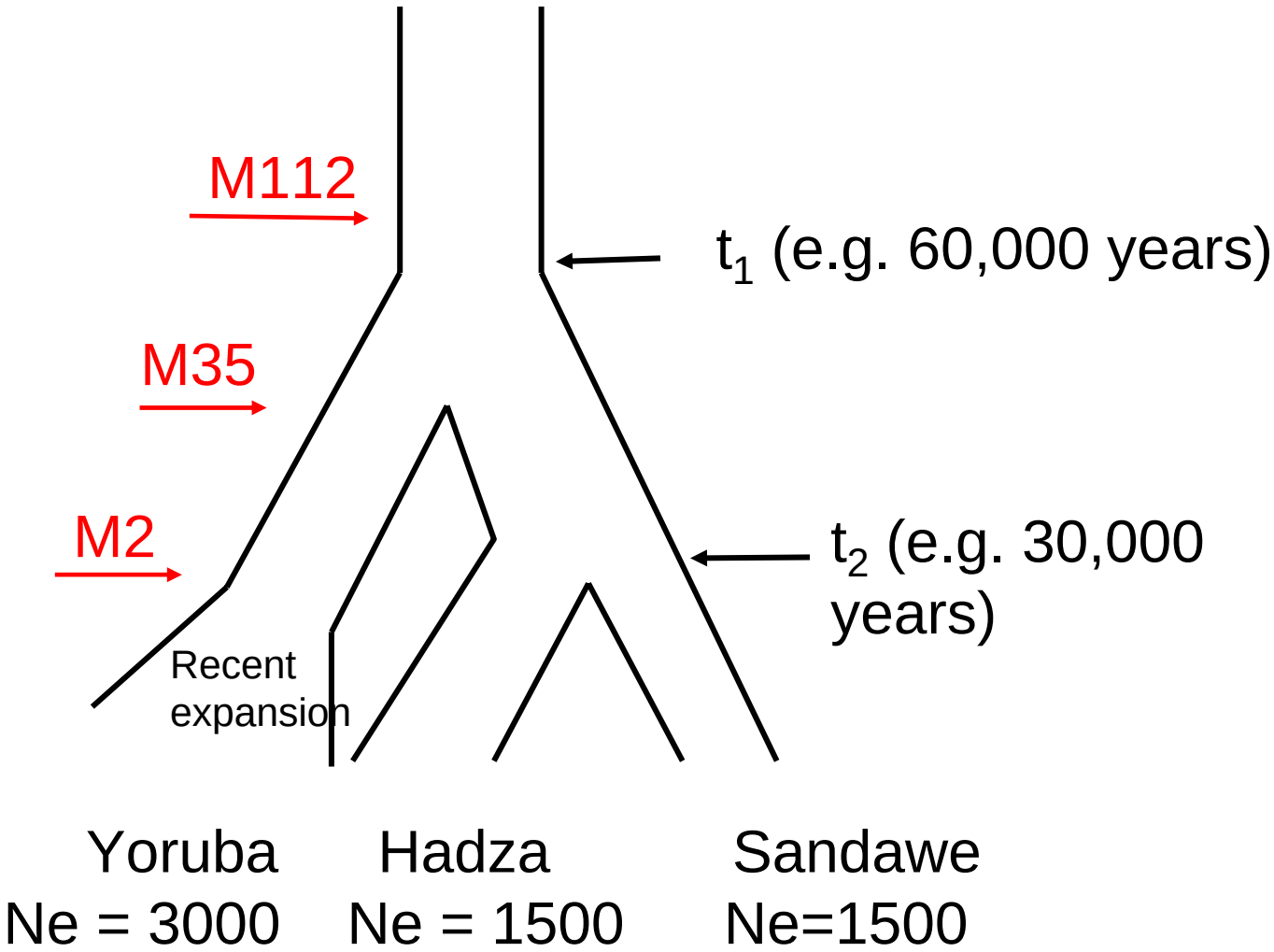
Gene flow between click-speaking groups?

Gene flow between click-speakers and Bantu speakers?

Explore population historic scenarios using simulations

Mountain et al., in prep

Y chromosome simulation: 3 UEPs+11 STRs

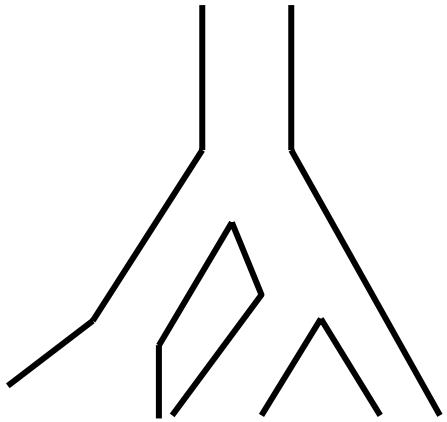


Three-population models

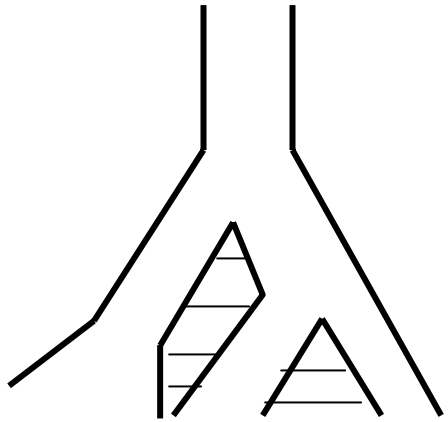
Mountain et al., in prep

Models: Uni- and bi-directional gene flow

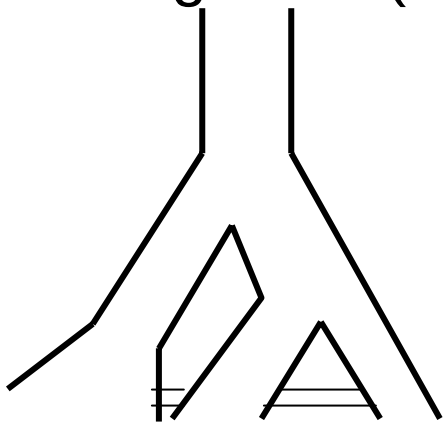
Complete isolation (CI)



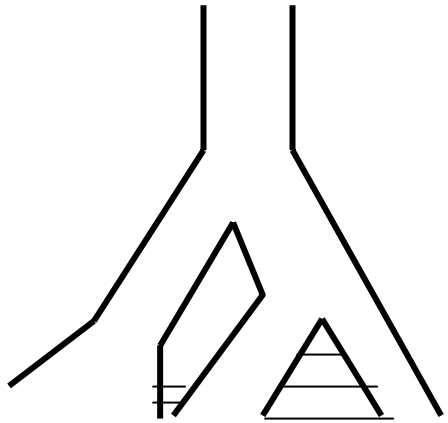
Isolation migration (IM)



Complete Isolation
Recent Migration (CIRM)



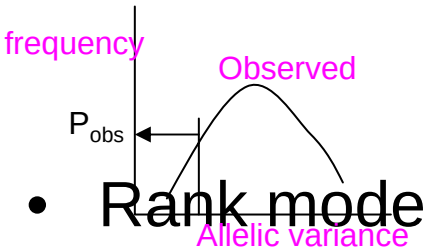
CIRM/IM



Mountain et al., in prep

Methods

- Run simulations for particular model
- Ascertainment based on UEP frequencies
- Calculate summary statistics
- Calculate simulated likelihood (L_{sim})



- Rank models by L_{sim}

$$L_{sim} = \prod_{allstats} p_{obs}$$

Mountain et al., in prep

Results: Top 5 models

Model	Parameters	L_{sim} ($x10^{-18}$)
CIRM7	recent gene flow over the last 3,000 years; unidirectional gene flow from the Yoruba into the Hadza and Sandawe populations (5 migrants per generation), more recent divergence between Hadza and Sandawe (15,000 years before present).	0.938
CIRM/ IM4	recent unidirectional gene flow over the last 3,000 years from the Yoruba into the Hadza and Sandawe populations (5 migrants per generation); continuous unidirectional gene flow following population divergence from the Sandawe to the Hadza (2 migrants per generations)	0.105
CIRM 5	recent gene flow over the last 3,000 years; bidirectional gene flow between the Hadza and the Sandawe (2 migrants per generation) and unidirectional gene flow from the Yoruba into the Hadza and Sandawe populations (5 migrants per generation).	0.035
CIRM/ IM3	recent unidirectional gene flow over the last 3,000 years from the Yoruba into the Hadza and Sandawe populations (5 migrants per generation); continuous unidirectional gene flow following population divergence from the Sandawe to the Hadza (1 migrants per generations)	0.028
CIRM8	recent gene flow over the last 3,000 years; unidirectional gene flow from the Yoruba into the Hadza and Sandawe populations (5 migrants per generation), more recent divergence between Hadza and Sandawe (10,000 years before present).	0.015

Mountain et al., in prep

Conclusions

- We can reject complete isolation and isolation migration models
- Accept more complex versions of history

  click-speaking groups isolated or geneflow from Sandawe into Hadza

  received migrants from Bantu-speakers
- Method provides a set of possible histories
  - Test with STR data?

Mountain et al., in prep