Pairing gene-specific and language-specific evidence for population contacts—towards a typology

Nicolas Brucato

Leiden University

Søren Wichmann

MPI-EVA

Structure of presentation

- Questions
- Methods
  - genetics
  - linguistics
- Data
  - genetics
  - linguistics
- Results

Initial questions

- Overall correlations among languages, different genetic markers, and geography
- Are there ways of spotting language shift?
  - (1) greater-than-expected genetic distances given the linguistic distances
  - (2) perhaps detectable linguistic substrate effects

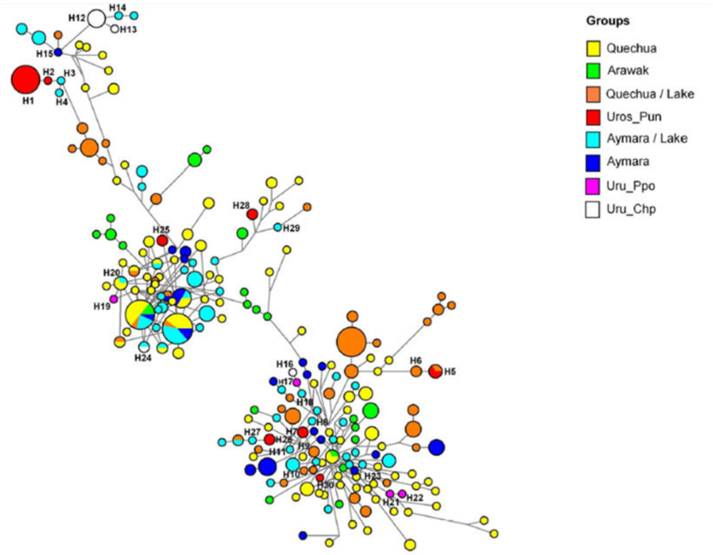
Methods: genetics

Genetic markers

Uniparental markers

Mitochondrial DNA: maternal lineages

Non recombinant Y-chromosome: paternal lineages



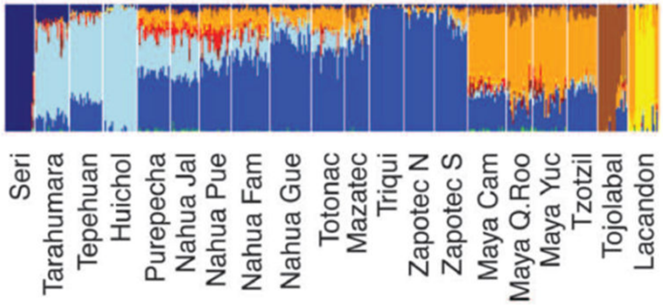
Median Joining network for STR haplotypes of one haplogroup among 22 Peruvian and Bolivian populations. (Sandoval et al., 2013)

Biparental markers

Unique mutations (ex: SNP) on the autosomes

Genome-wide data

Millions of mutations across the genome



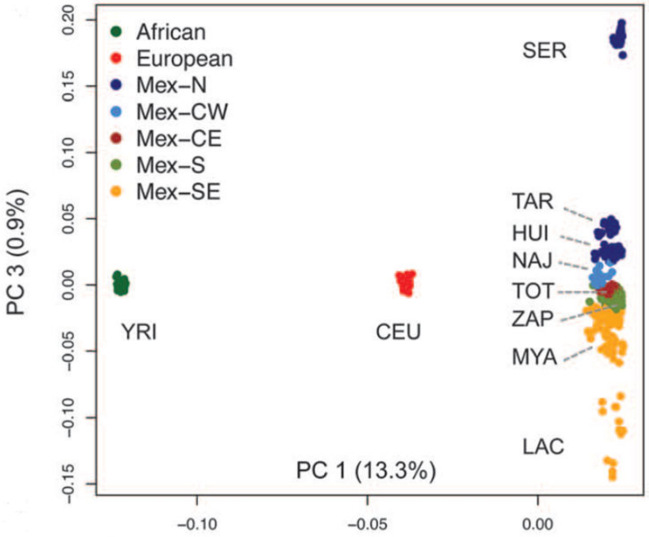
Global ancestry proportions in Native Mexicans at K = 9. (Moreno-Estrada et al. 2014)

Statistical analyses

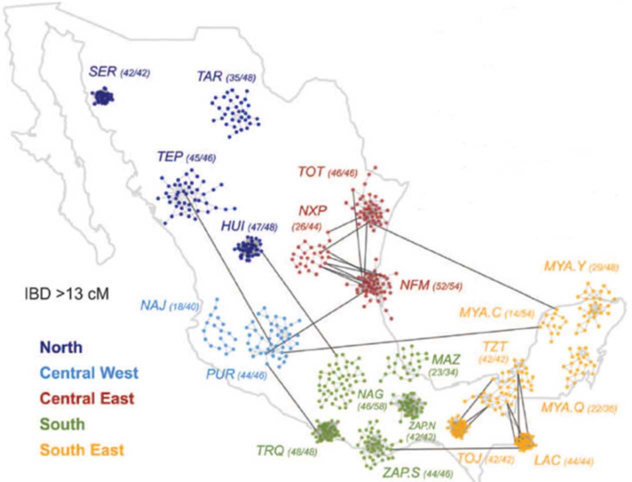
Diversity
Closely related populations have similar genetic diversities (ex: Fst index)

Sharing
Closely related individuals share long genetic fragments (ex: Identity-by-descent)

Admixture
Each population is a mixture of at least two parental groups (ex: percentage of admixture)



Principal component Analysis of Native Mexicans with African and European samples. (Moreno-Estrada et al. 2014)

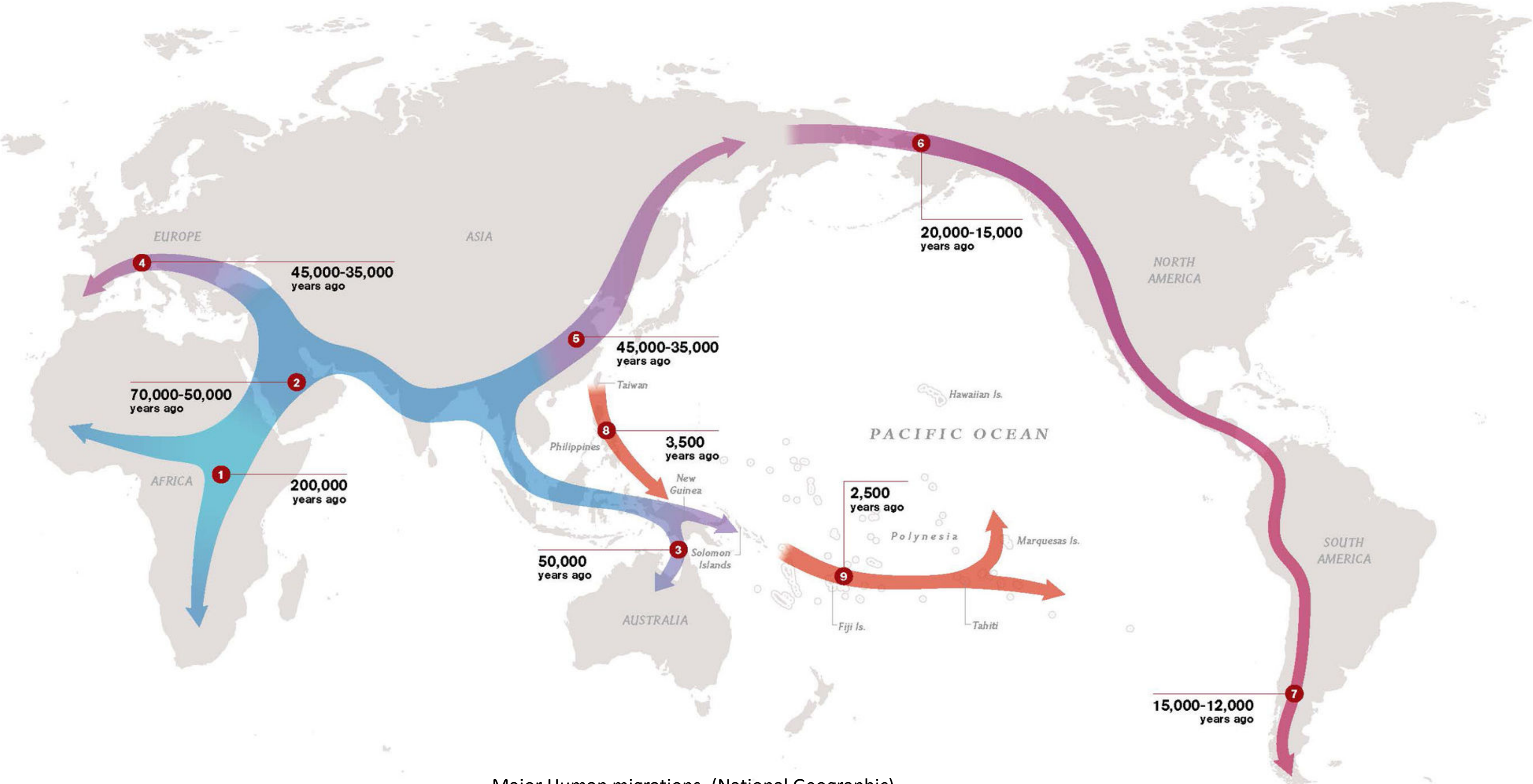


Relatedness graph of individuals sharing more than 13 cM of the genome. (Moreno-Estrada et al. 2014)

Genetic Diversity

Migrations and isolations created genetic differentiation.

Genetic diversity decreases according to the distance from Africa.

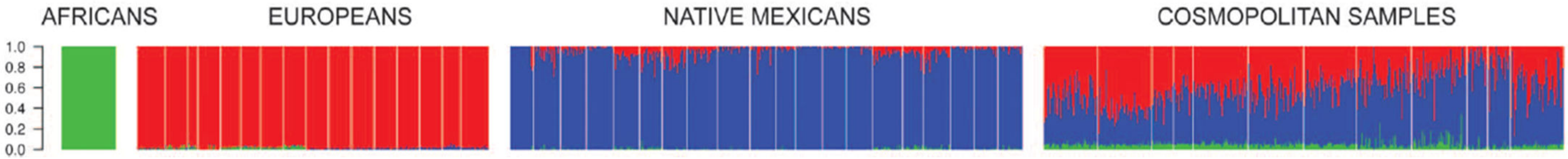


Major Human migrations. (National Geographic)

Genetic Diversity

Isolation of parental populations creates a genetic signature that can be retrieved in admixed groups.

Gene flow between human groups have always occurred.

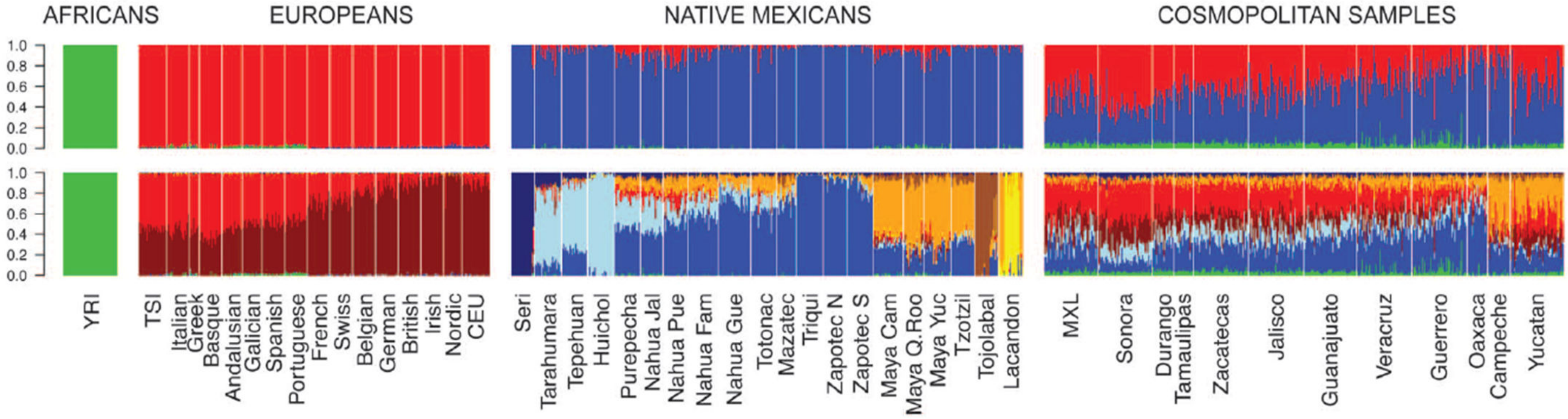


Global ancestry proportions at K = 3 (top) and K = 9 (bottom). (Moreno-Estrada et al. 2014)

Genetic Diversity

Isolation of parental populations creates a genetic signature that can be retrieved in admixed groups.

Gene flow between human groups have always occurred.



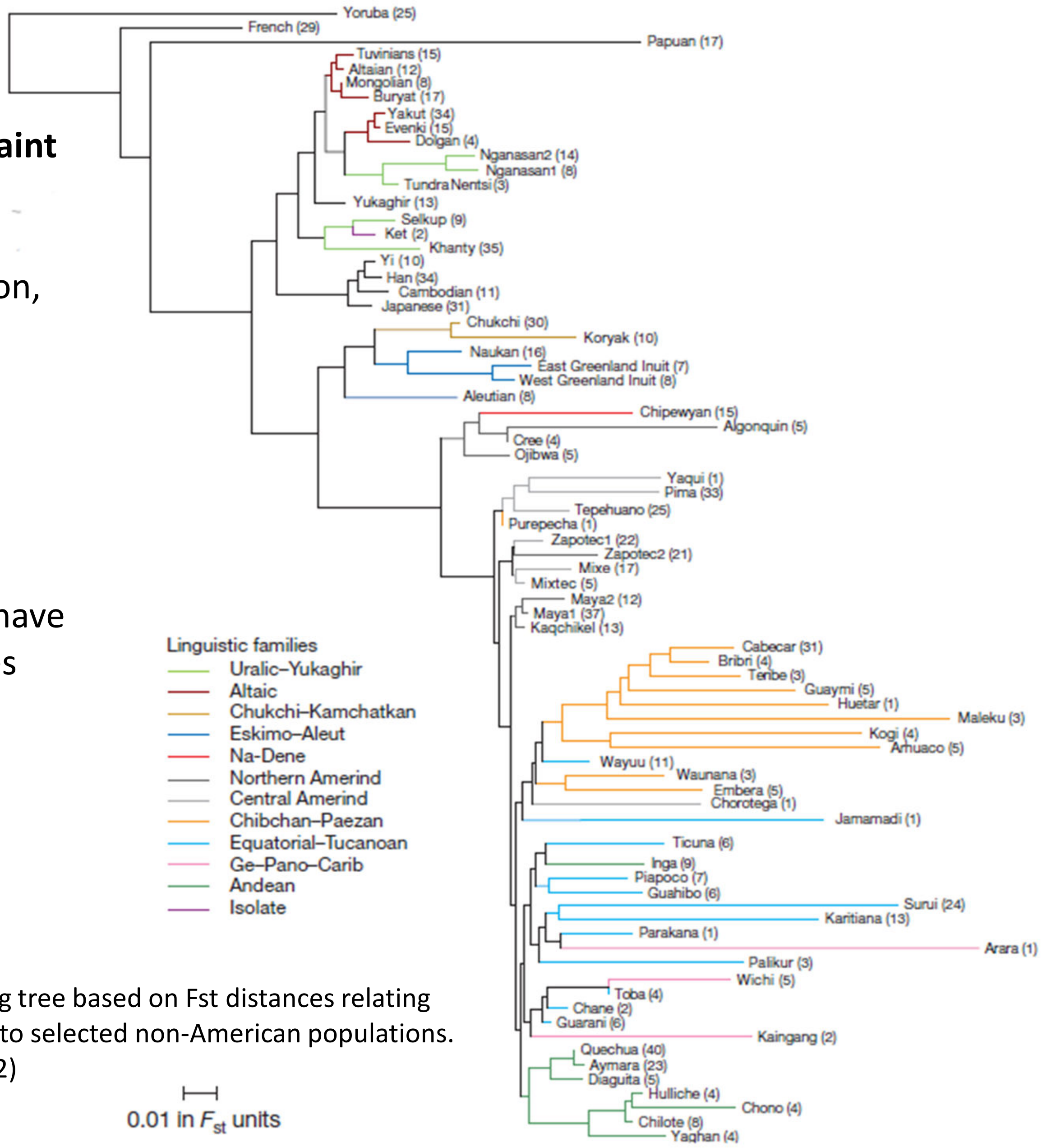
Global ancestry proportions at K = 3 (top) and K = 9 (bottom). (Moreno-Estrada et al. 2014)

Genetic and Linguistic correlation

Language is a strong cultural constraint to gene flow.

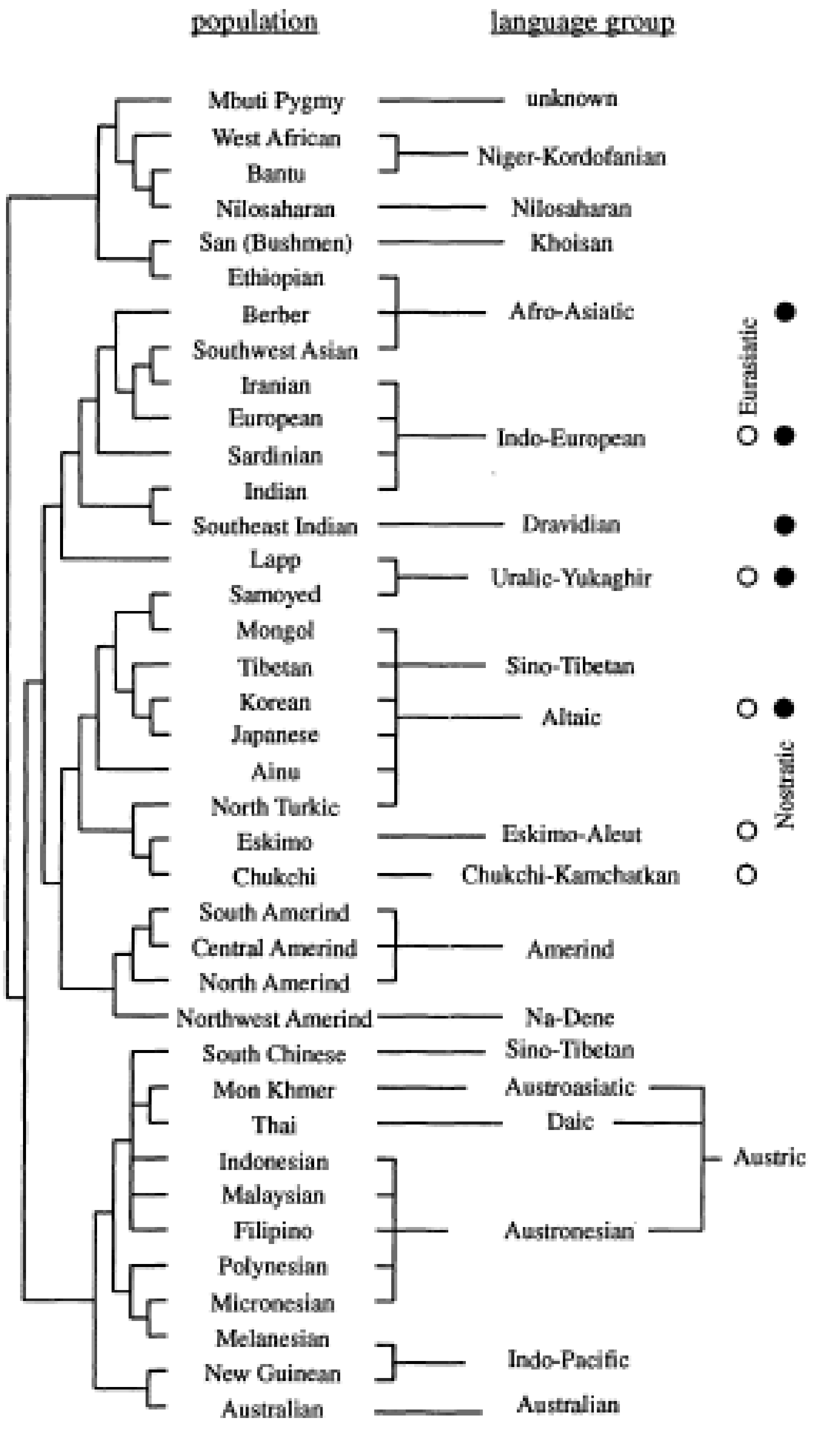
BUT so are geography, history, religion, etc.

All studies on correlations between genetic data and linguistic diversity have relied on linguistic phylogenetic trees (ex : Greenberg, Ruhlen) from which classes or distances were infered.



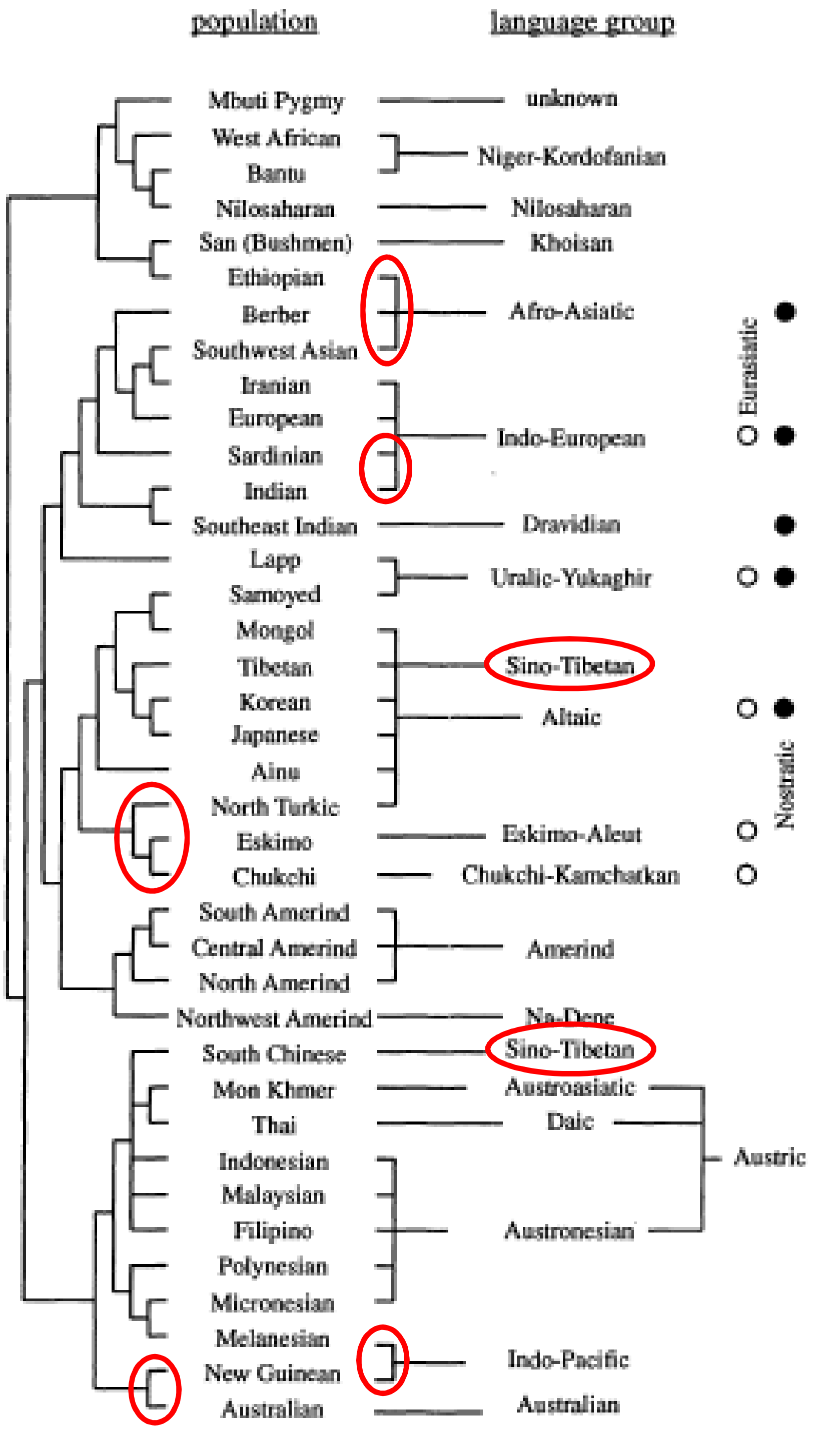
Neighbour-joining tree based on Fst distances relating Native American to selected non-American populations. (Reich et al., 2012)

Earlier approaches



Cavalli-Sforza et al. (1992)

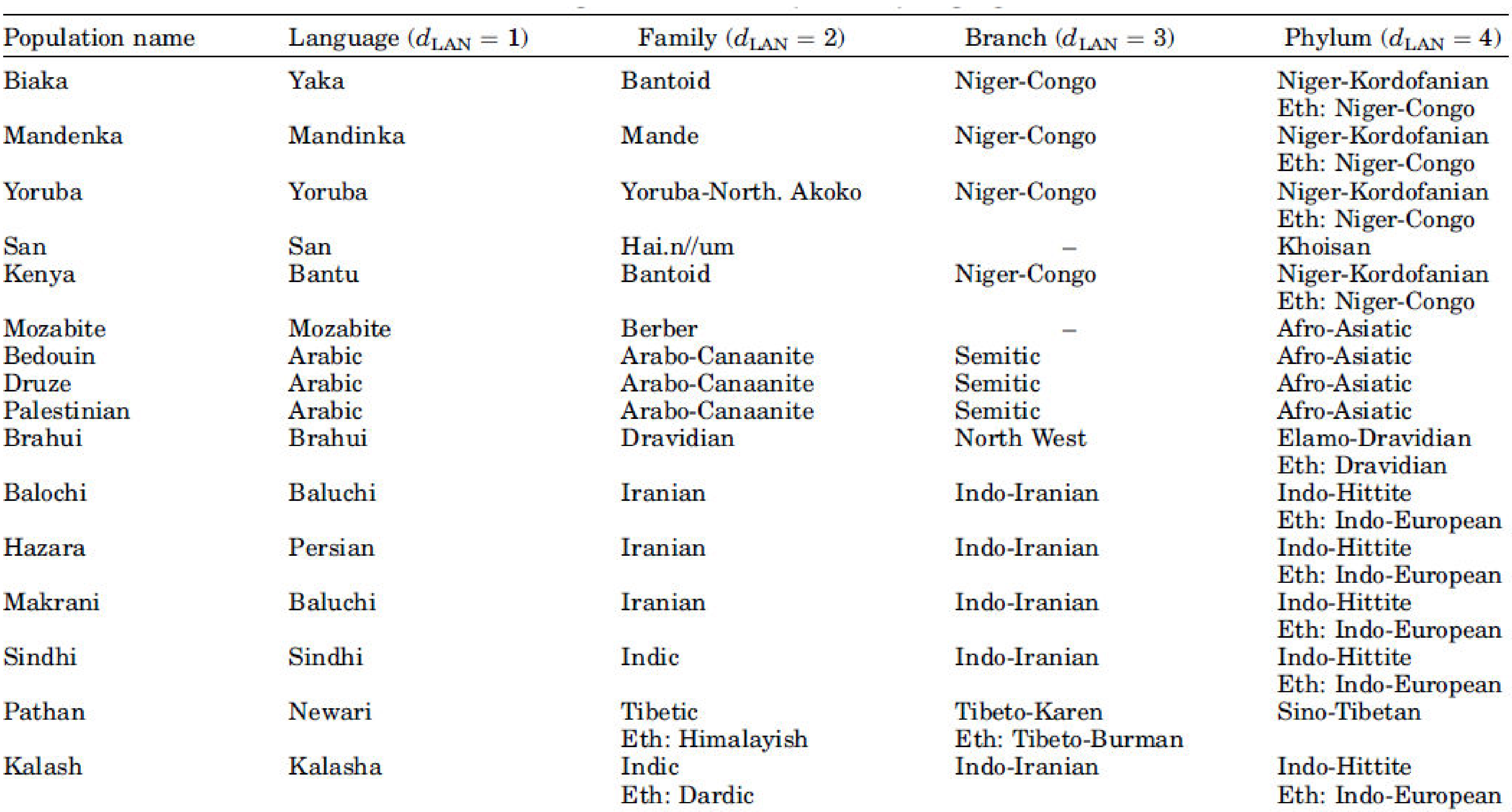
Earlier approaches



Cavalli-Sforza et al. (1992)

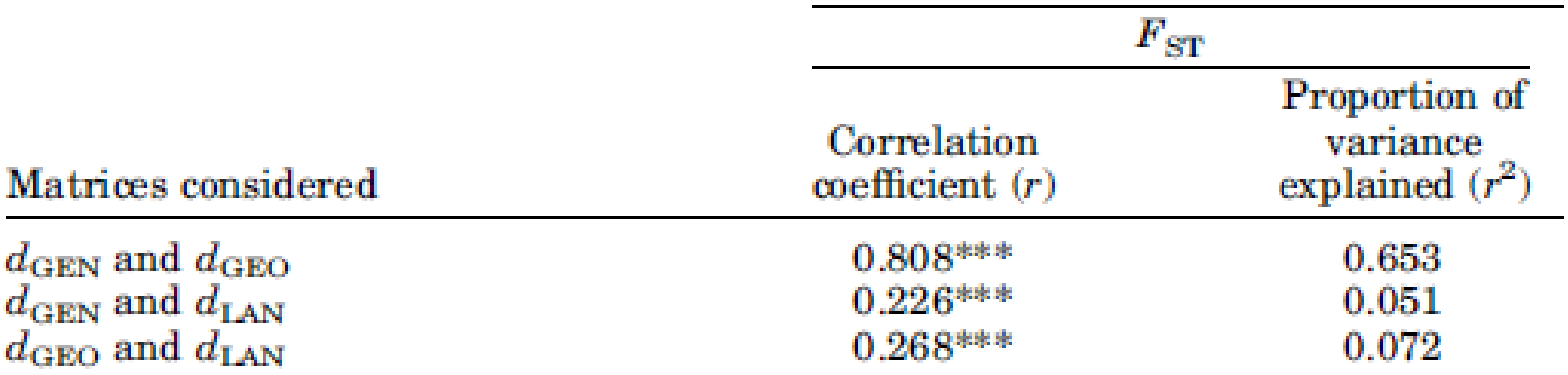
- Belle and Barbujani (2007) define four linguistic distances:
  - d_{LAN} = 1: different language, same family
  - d_{LAN} = 2: different family, same branch
  - d_{LAN} = 3: different branch, same phylum
  - d_{LAN} = 4: different phyla (using two alternative classifications: Ruhlen and *Ethnologue*)

Examples



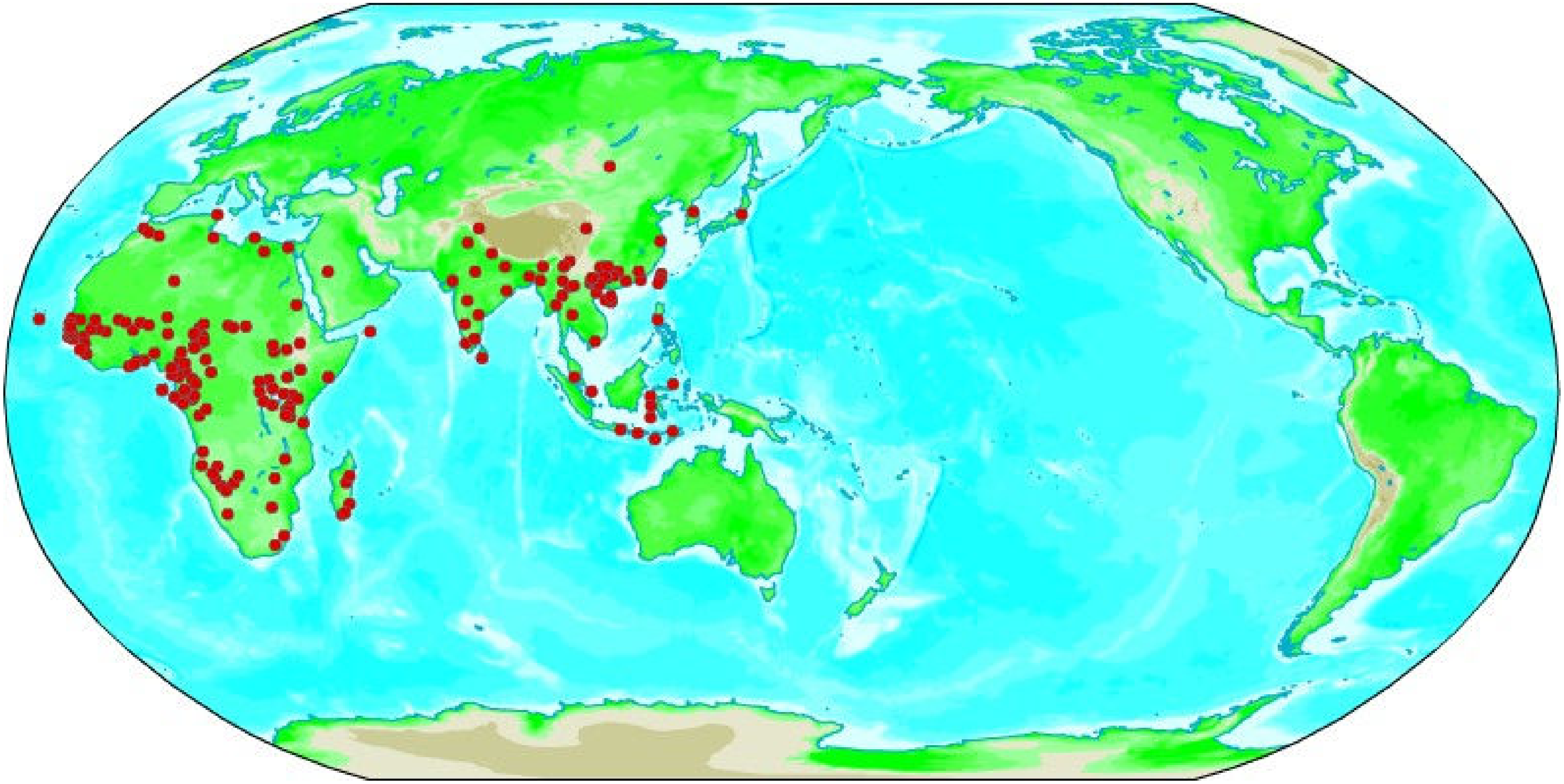
Population name	Language ($d_{\mathrm{LAN}} = 1$)	Family ($d_{\mathrm{LAN}} = 2$)	Branch ($d_{\mathrm{LAN}} = 3$)	Phylum ($d_{\mathrm{LAN}} = 4$)
Biaka	Yaka	Bantoid	Niger-Congo	Niger-Kordofanian Eth: Niger-Congo
Mandenka	Mandinka	Mande	Niger-Congo	Niger-Kordofanian Eth: Niger-Congo
Yoruba	Yoruba	Yoruba-North. Akoko	Niger-Congo	Niger-Kordofanian Eth: Niger-Congo
San	San	Hai.n//um	–	Khoisan
Kenya	Bantu	Bantoid	Niger-Congo	Niger-Kordofanian Eth: Niger-Congo
Mozabite	Mozabite	Berber	–	Afro-Asiatic
Bedouin	Arabic	Arabo-Canaanite	Semitic	Afro-Asiatic
Druze	Arabic	Arabo-Canaanite	Semitic	Afro-Asiatic
Palestinian	Arabic	Arabo-Canaanite	Semitic	Afro-Asiatic
Brahui	Brahui	Dravidian	North West	Elamo-Dravidian Eth: Dravidian
Balochi	Baluchi	Iranian	Indo-Iranian	Indo-Hittite Eth: Indo-European
Hazara	Persian	Iranian	Indo-Iranian	Indo-Hittite Eth: Indo-European
Makrani	Baluchi	Iranian	Indo-Iranian	Indo-Hittite Eth: Indo-European
Sindhi	Sindhi	Indic	Indo-Iranian	Indo-Hittite Eth: Indo-European
Pathan	Newari	Tibetic Eth: Himalayish	Tibeto-Karen Eth: Tibeto-Burman	Sino-Tibetan
Kalash	Kalasha	Indic Eth: Dardic	Indo-Iranian	Indo-Hittite Eth: Indo-European

Results of Belle and Barbujani (2007)



	F_{ST}	
Matrices considered	Correlation coefficient (r)	Proportion of variance explained (r^2)
d_{GEN} and d_{GEO}	0.808***	0.653
d_{GEN} and d_{LAN}	0.226***	0.051
d_{GEO} and d_{LAN}	0.268***	0.072

Our data



Map of current population-language matches
(more to be added from Eurasia soon, and eventually from other areas)

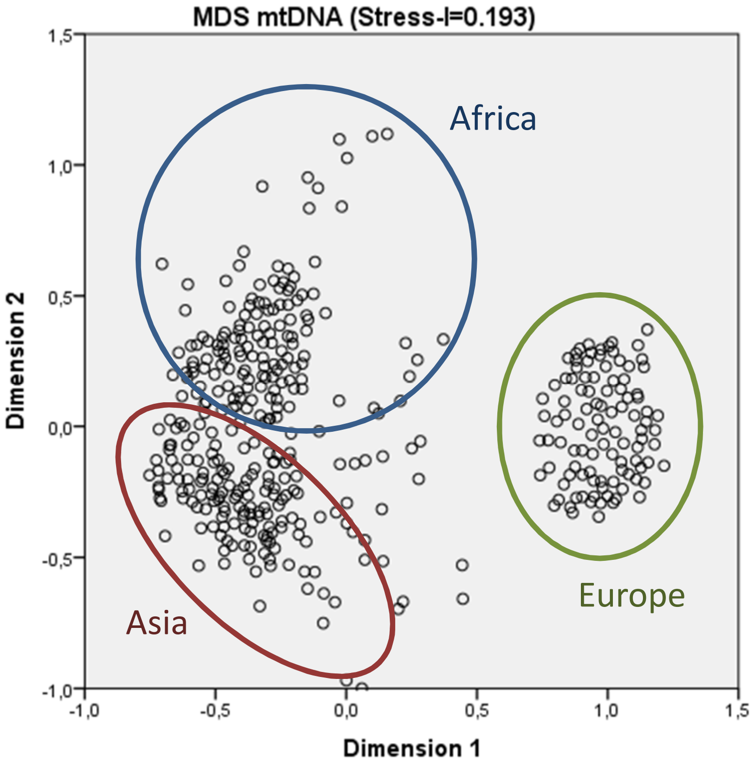
Our methods

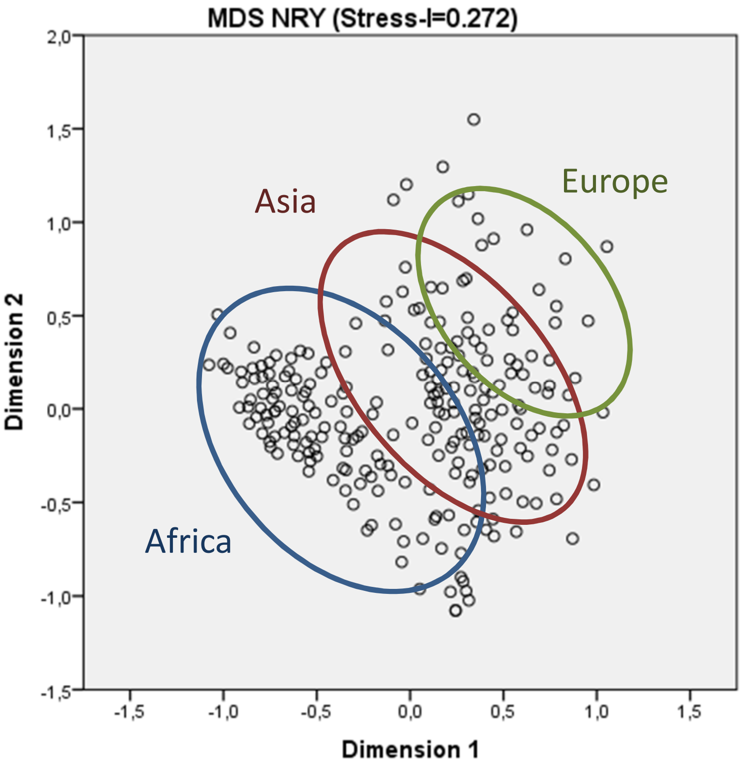
- Distance measures based on average Levenshtein distances (more precisely so-called LDND, cf. Wichmann et al 2010) from 40-item word lists from the database of the Automated Similarity Judgment Program

Our approach

Compile a large worldwide genetic and linguistic dataset.
So far:

- mtDNA: Npop =220; Nind = 13,414
- Y-chromosome: Npop = 90; Nind = 5,192
- Linguistic distances ASJP: Npop =220





Preliminary results

Linguistic distances explain 7.5% of the male genetic diversity but only 2.5% of the female genetic diversity.
In contrast, geographical distances explain female better than male genetic diversity.

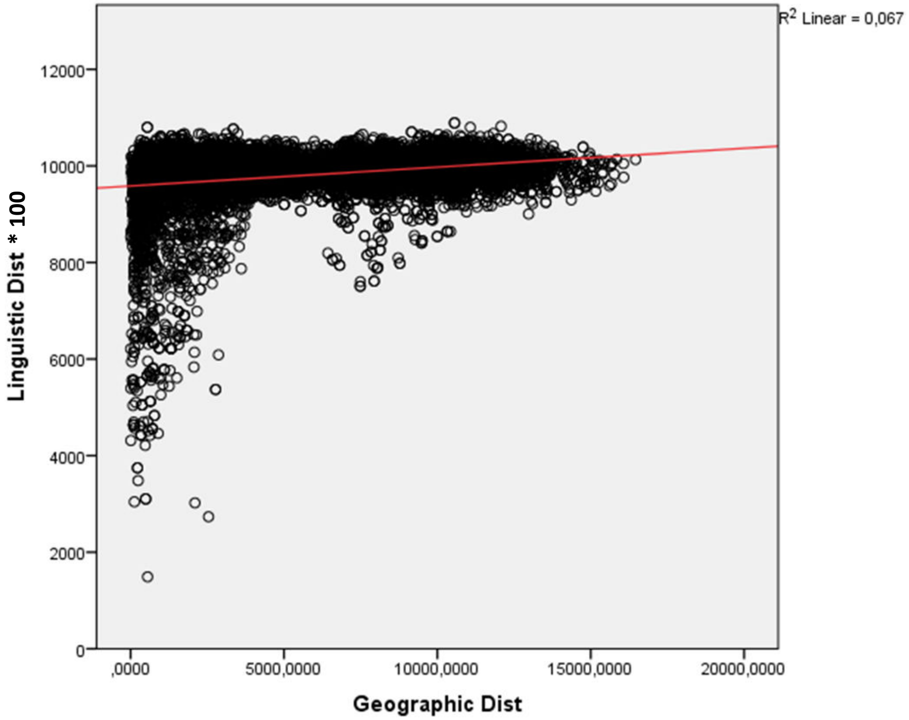
Mantel tests (1000 iterations)	Npop	r^2	p
Language-mtDNA	175	0,025	$< 10^{-6}$
Language-NRY	90	0,075	$< 10^{-6}$
Geography-mtDNA	175	0,133	$< 10^{-6}$
Geography-NRY	90	0,018	$< 10^{-6}$
Geography-Language	220	0,059	$< 10^{-6}$

These analyses confirm the previously reported findings on correlations between linguistic and genetic data as well as the potential sex-biased dispersal on a worldwide scale.

Validates the approach to detect divergences from the observed global correlations.

Preliminary results—language and geography

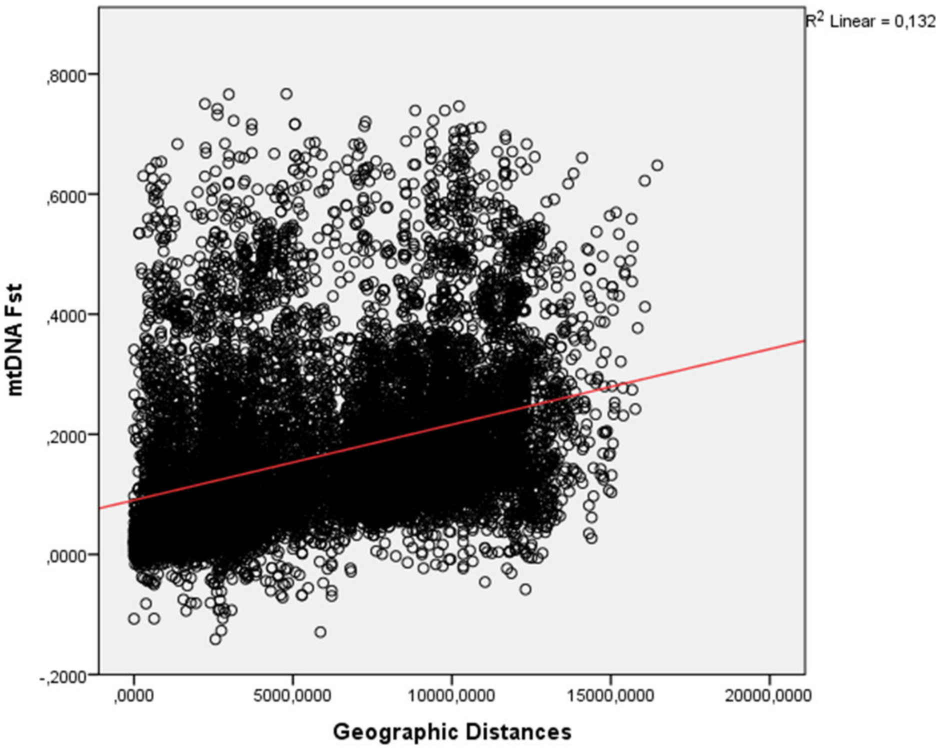
Distribution of linguistic distances with geographic distances

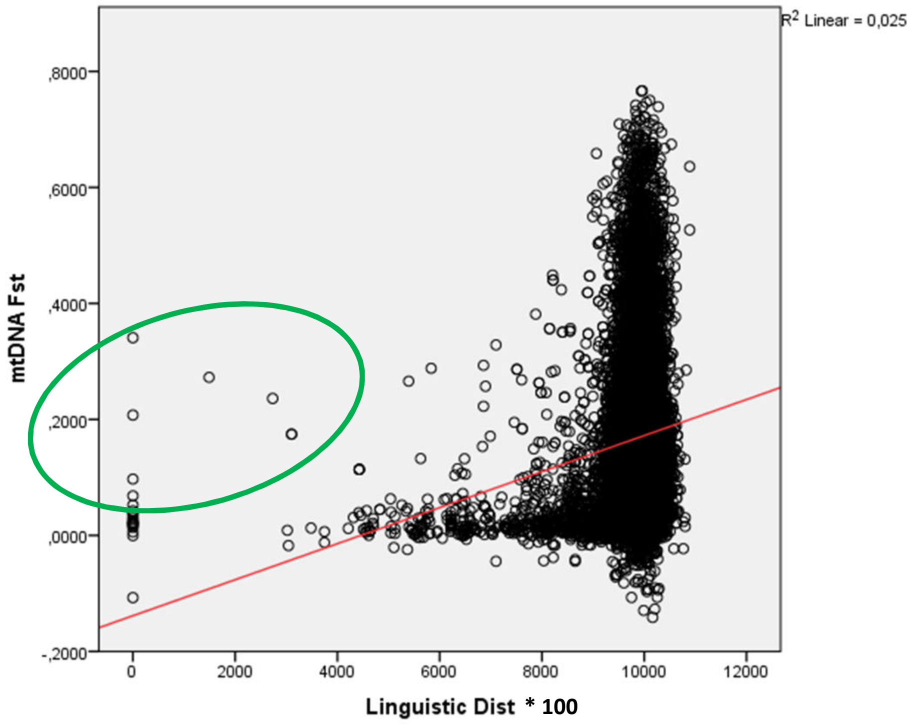


Preliminary results—mtDNA

Distances diverging from the correlation line can be identified from the global distribution.

For mtDNA:





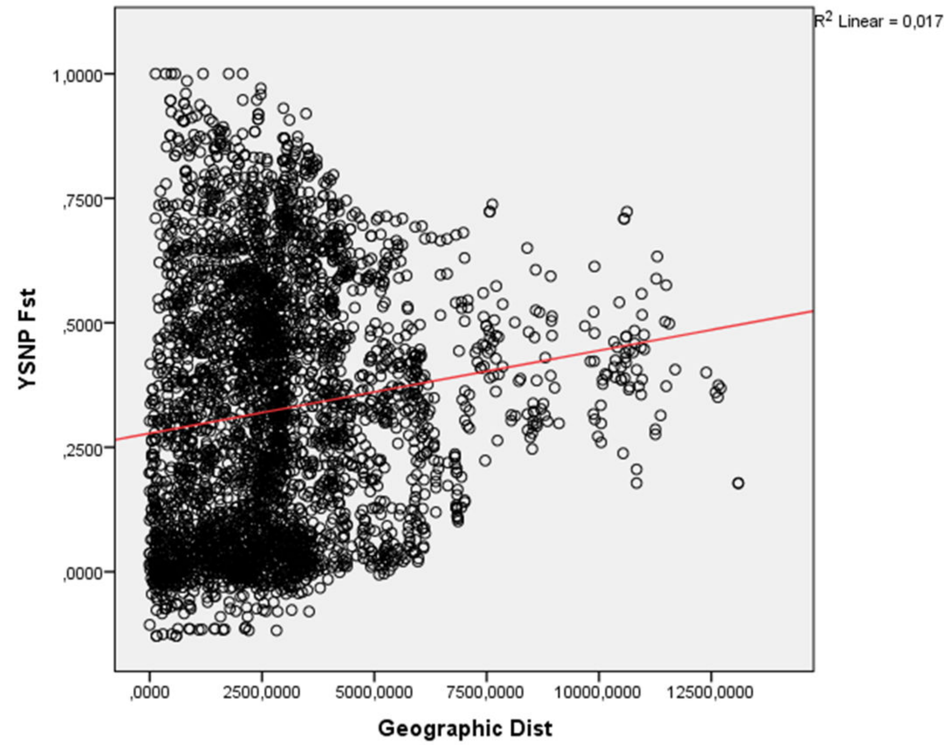
Outliers with large genetic distances and small linguistic distances

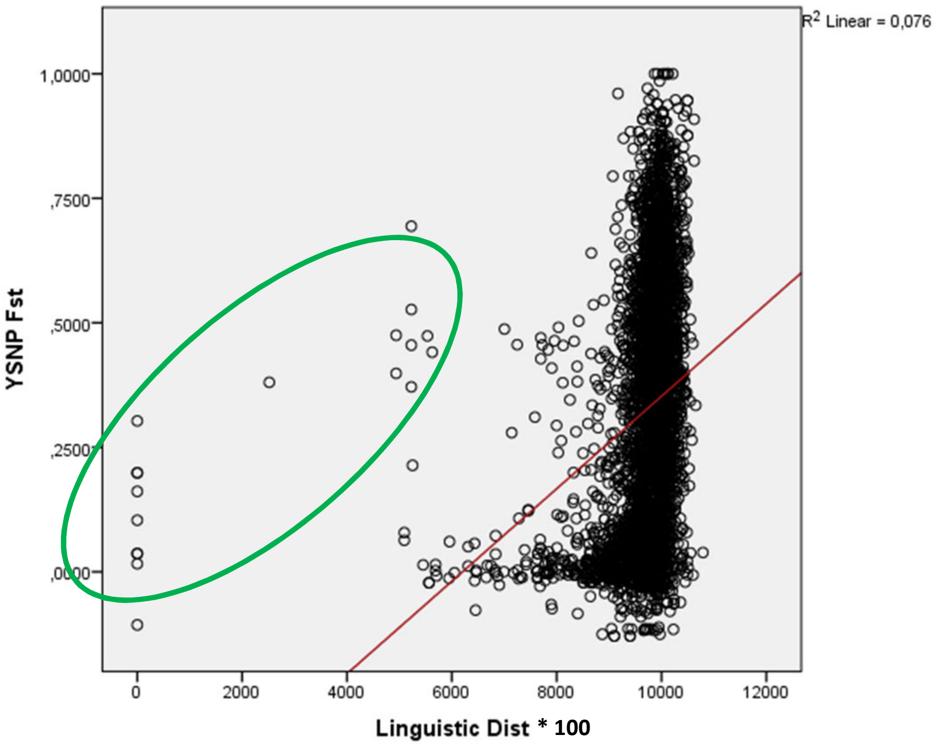
Pop A	Pop B	mtDNA dist	ling dist
Thailand_Lahu	China_Lahu	0.34081	0%
Mongolia_Mongolian	China_Mongolian	0.20745	0%
Somalia_Somali	Ethiopia_Somali	0.09718	0%
Kenya_Somali	Ethiopia_Somali	0.05241	0%

Preliminary results—Y-SNP

Distances diverging from the correlation line can be identified from the global distribution.

For Y-SNP:





Outlier with large genetic distances and small linguistic distances

Pop A	Pop B	ySNP dist	ling dist
Morocco_Arab	Egypt_Arab	0.21373	52.51%

Future work

- Expand the database so that it will be easier to identify directions of gene flow and language contact
- Include typological data (from WALS and perhaps other sources)
- Systematically address the issue of a typology of contact situations
- Model processes of gene flow, language contact, and migration

References

- Belle, Elise M. S. and Guido Barbujani. 2007. Worldwide analysis of multiple microsatellites: Language diversity has a detectable influence on DNA diversity. *American Journal of Physical Anthropology* 133:1137-1146.
- Cavalli-Sforza, L. L., Erich Minch, and J. L. Mountain. 1992. Coevolution of genes and languages revisited. *Proceedings of the National Academy of Sciences of the U.S.A.* 89(12): 5620-5624*.*
- Roewer et al. 2014 Continent-wide decoupling of Y-chromosomal genetic variation from language and geography in Native South Americans. *Plos Genetics* 9(4):e1003460.
- Wichmann, Søren, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A.* 389: 3632-3639.