SISSA, July 7th 2016

Language, languages, genes and human diversity

Giuseppe Longobardi

UNIVERSITY of York

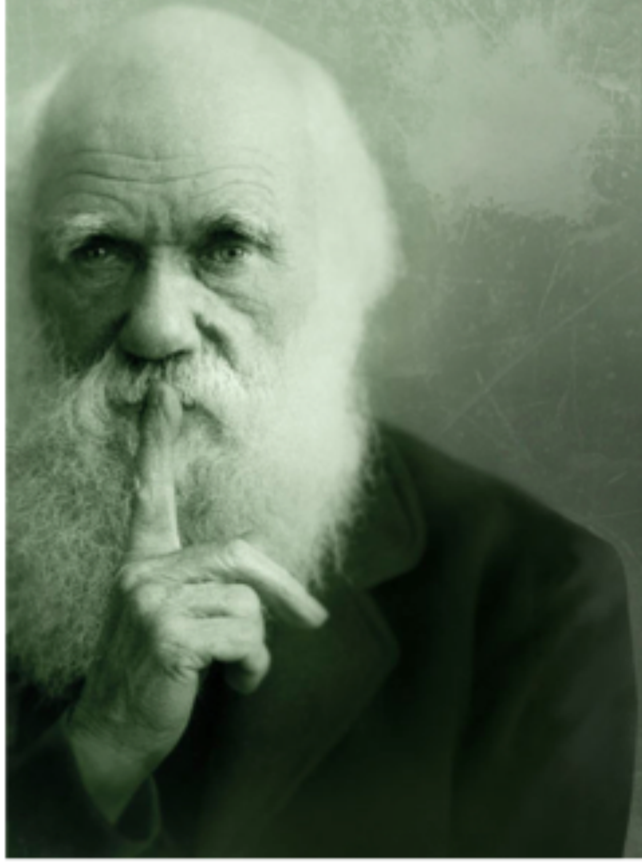
with the *Langelin* project team

LANGELIN - LANGUAGES GENES LINEAGES
ERC ADVANCED GRANT N. 295733
MEETING DARWIN'S LAST CHALLENGE

erc

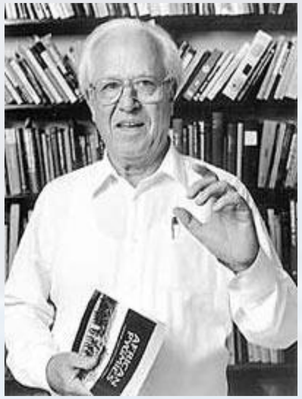
Funded by the European Union

Darwin's last challenge (*The origin of the species, ch. 14*)

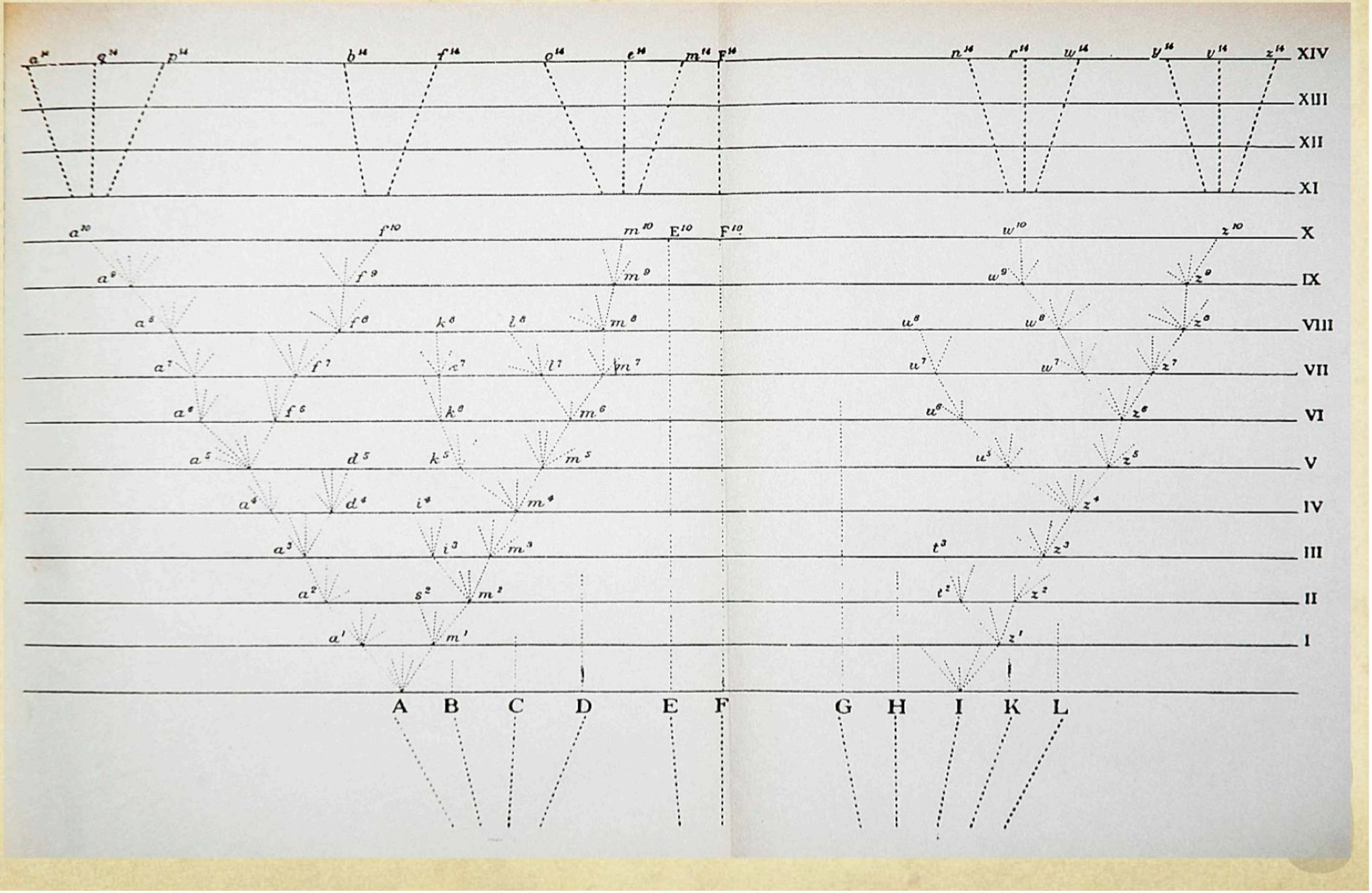


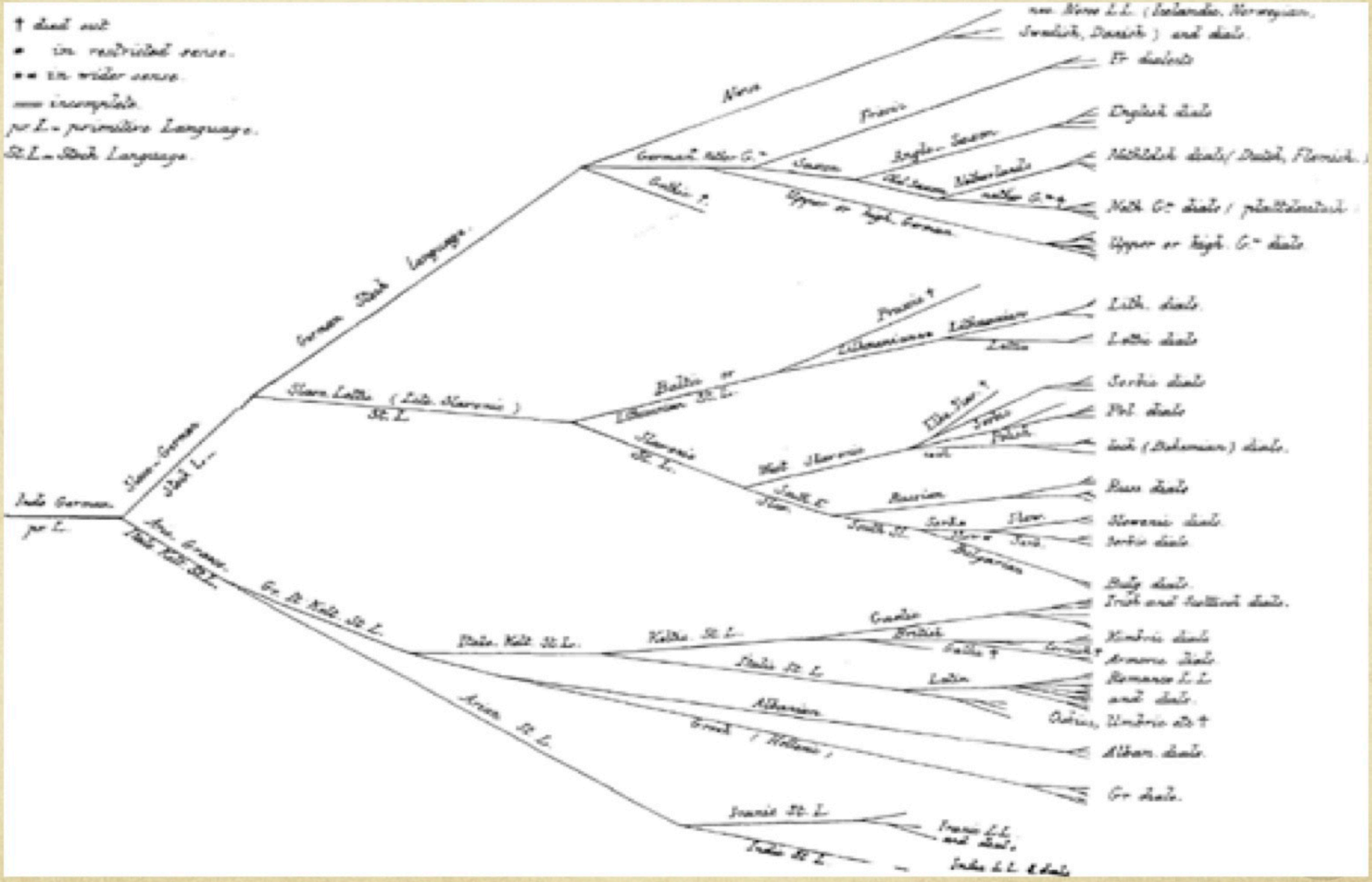
If we possessed a perfect pedigree of mankind,
a genealogical arrangement of the races of man,
we would afford
the best classification of the various languages
now spoken throughout the world;
*and if all **extinct** languages,*
*and all intermediate and slowly changing **dialects,***
were to be included,
such an arrangement would be
the only possible one

Since Darwin (1859), research in human biology has tried to address empirically the parallels between genetic and linguistic diversification: Cavalli Sforza et al. (1988), Sokal (1988)



A positive answer would allow researchers to support (pre-)historical hypotheses using evidence coming from **two different domains**


a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} e^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
a^{10} f^{10} m^{10} E^{10} F^{10} w^{10} z^{10}
X
a^{9} f^{9} m^{9} w^{9} z^{9}
IX
a^{8} f^{8} k^{8} l^{8} m^{8} u^{8} w^{8} z^{8}
VIII
a^{7} f^{7} k^{7} l^{7} m^{7} u^{7} w^{7} z^{7}
VII
a^{6} f^{6} k^{6} m^{6} u^{6} z^{6}
VI
a^{5} d^{5} k^{5} m^{5} u^{5} z^{5}
V
a^{4} d^{4} i^{4} m^{4} z^{4}
IV
a^{3} i^{3} m^{3} t^{3} z^{3}
III
a^{2} s^{2} m^{2} t^{2} z^{2}
II
a^{1} m^{1} z^{1}
I
A B C D E F G H I K L

† dead out
* in restricted sense.
** in wider sense.
—— incomplete.
pr L = primitive Language.
St L = Stock Language.
Indo German
pr L.
Slavo-German Stock L.
German Stock Language.
Slavo Lettic (Lito Slavonic) St. L.
Ario Graeco Italo Kelt. St. L.
Gr. It. Kelt. St. L.
Norse
German. Nether G.
Gothic †.
Frisic
Saxon
Anglo-Saxon
Old Saxon
Netherlandic
nether G.**
Upper or high German
new Norse L.L. (Icelandic, Norwegian,
Swedish, Danish) and dials.
Fr dialects
English dials
Netherlandic dials (Dutch, Flemish.)
Neth. G. dials / plattdeutsch.
Upper or high. G. dials
Baltic or Lithuanian St. L.
Prussic †
Lithuanianic
Lithuanian
Lettic
Lith. dials.
Lettic dials
Slavonic St. L.
West Slavonic
Elbe Slav. †
Sorbic
Polish
tech.
Sorbic dials
Pol. dials
tech (Bohemian) dials.
South E. Slav.
Russian
South SI.
Serbo
Slov.*
Slov.
Serb.
Bulgarian
Russ dials
Slowenic dials
Serbic dials
Bulg dials.
Italo. Kelt. St. L.
Keltic. St. L.
Gaelic
British
Gallic †
Cornish †
Irish and Scottish dials.
Kimbric dials
Armoric dials
Italic St. L.
Latin
Romance L.L. and dials.
Oskic, Umbric etc †
Albanian
Greek (Hellenic)
Alban. dials.
Gr dials.
Arian St. L.
Iranic St. L.
Iranic L.L. and dials.
Indic St. L.
Indic L.L. & dials

Lexical cognates

Distances for cognate words (lexical etymologies) are:

time shallow

and

hardly quantifiable

because of

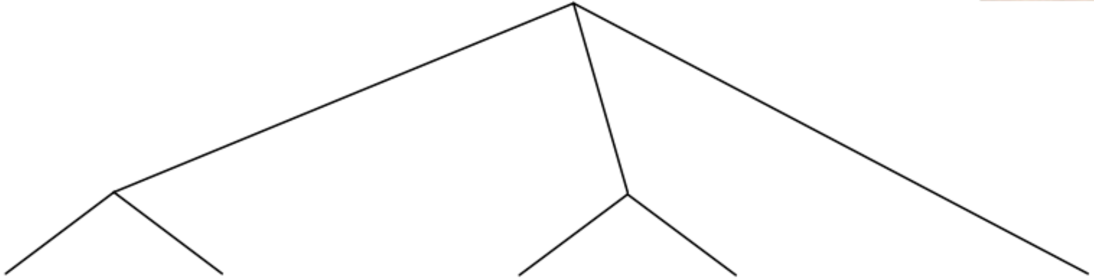
Vagueness and instability of meaning

Complexity of word structure also in form

Vagueness of lexical *comparanda*

- ⇨ **partial identity of** *form*: ***prendo*** **vs.** ***get***
- ⇨ **(or *haemorrhoid* and *serpent*!)**
- ⇨ **partial identity of** *meaning*: ***Hund*** **vs.** ***hound/dog***
- ⇨ **identity of** *form*, **not of** *meaning*: ***klein*** **vs.** ***clean***
- ⇨ **similarity *of*** *meaning shifts*, **no (real) correspondence of form: *fegato* vs. συκώτι**
- ⇨ **difficulty of measuring relative distances: *(je) fonds, (ich) giesse, juhomi***

Meaning ‘DOG’:



Italian *cane*

French *chien*

English *hound*

German *Hund*

Class. Greek *κύων*

Abkhaz = ***ala***

Ainu = ***seta***

Algonquian = ***athemwa***

Amharic = ***wäshsha***

Apache = ***góshé***

Arabic = ***kalb***

Aragonian = ***gos***

Assamese = ***kukur***

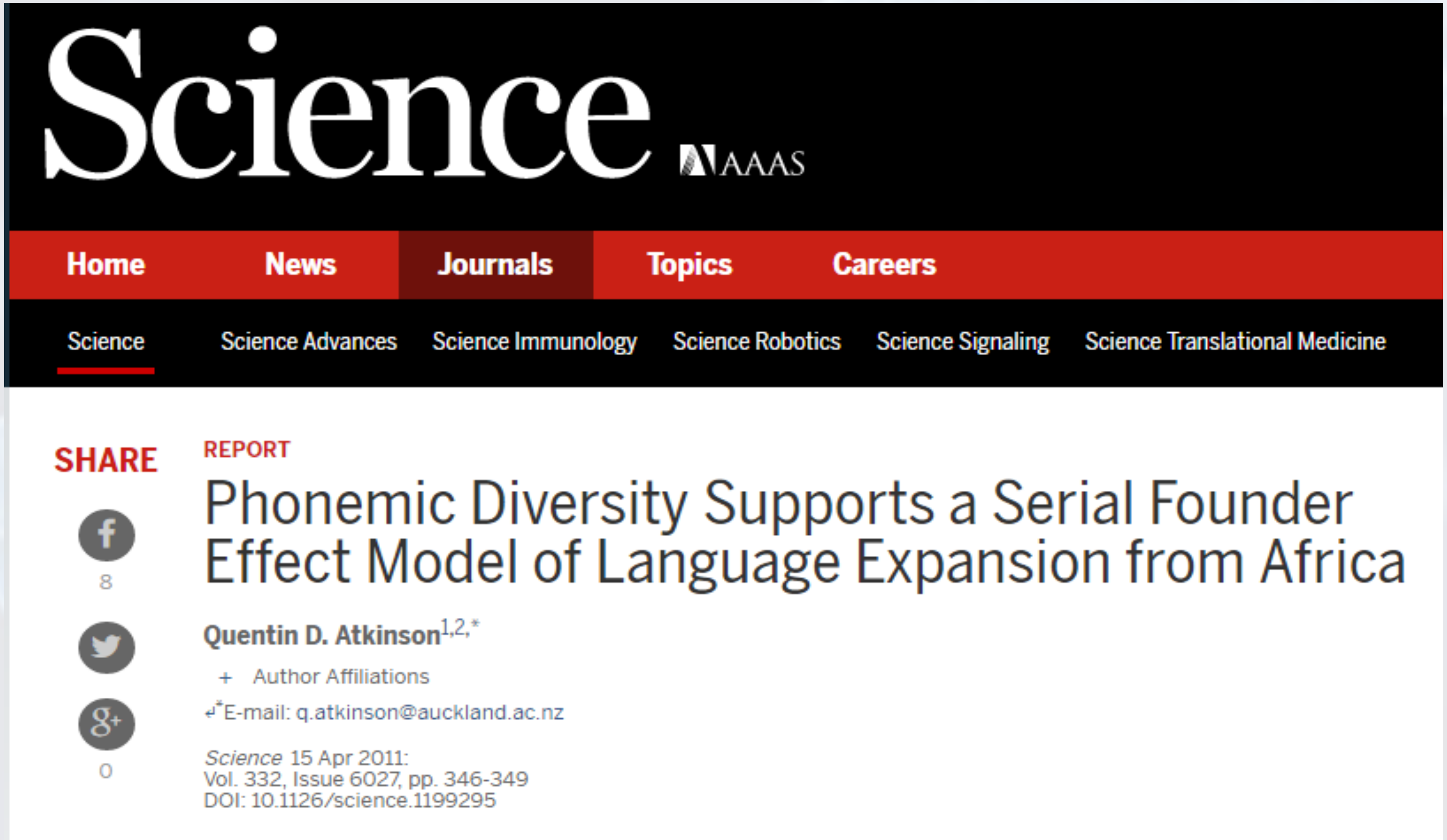
Asturian = ***perru***

Atayal = ***huzil***

Aymara = **anu**

♉

Atkinson (2011)



REPORT

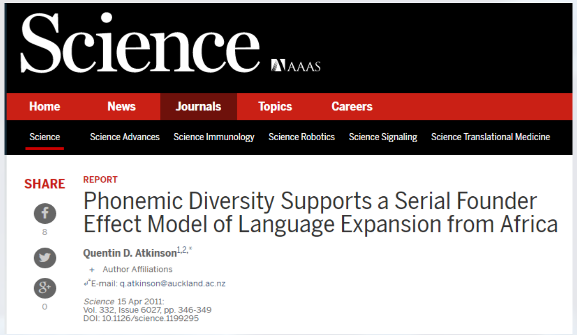
Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa

Quentin D. Atkinson[1,2,*]

+ Author Affiliations

*E-mail: q.atkinson@auckland.ac.nz

Science 15 Apr 2011:
Vol. 332, Issue 6027, pp. 346-349
DOI: 10.1126/science.1199295



REPORT

Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa

Quentin D. Atkinson[1,2,*]

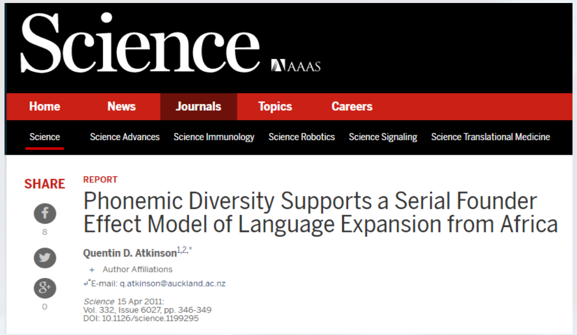
+ Author Affiliations

[*]E-mail: q.atkinson@auckland.ac.nz

Science 15 Apr 2011:
Vol. 332, Issue 6027, pp. 346-349
DOI: 10.1126/science.1199295

Atkinson (2011) shows that «the number of phonemes used in a global sample of 504 languages […] fits a serial founder–effect model of expansion from an inferred origin in Africa»

Data refer to the size of vowel inventories, consonant inventories, and tone inventories taken from WALS (Dryer and Haspelmath 2013)



REPORT

Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa

Quentin D. Atkinson[1,2,*]

[*] E-mail: q.atkinson@auckland.ac.nz

Science 15 Apr 2011:
Vol. 332, Issue 6027, pp. 346-349
DOI: 10.1126/science.1199295

Languages are represented through three **equally weighted** multi-state characters

Consonant Inventory

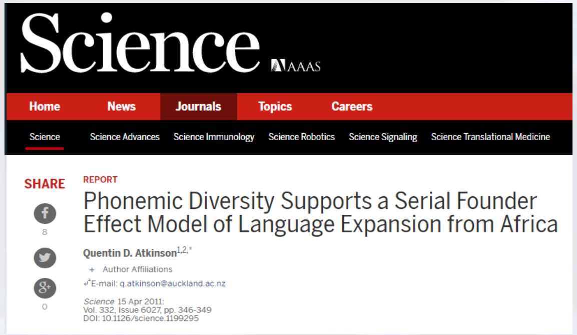
1 = Small (6-14)
2 = Moderately Small (15-18)
3 = Average (19-25)
4 = Moderately Large (26-33)
5 = Large (33+)

Tone Inventory

1 = No Tones
2 = Simple Tone system
3 = Complex Tone system

Vowel Inventory

1 = Small (2-4)
2 = Average (5-6)
3 = Large (7-14)



REPORT

Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa

Quentin D. Atkinson[1,2,*]

+ Author Affiliations

*E-mail: q.atkinson@auckland.ac.nz

Science 15 Apr 2011:
Vol. 332, Issue 6027, pp. 346-349
DOI: 10.1126/science.1199295

Which kind of information do phonemic inventories provide about language history?

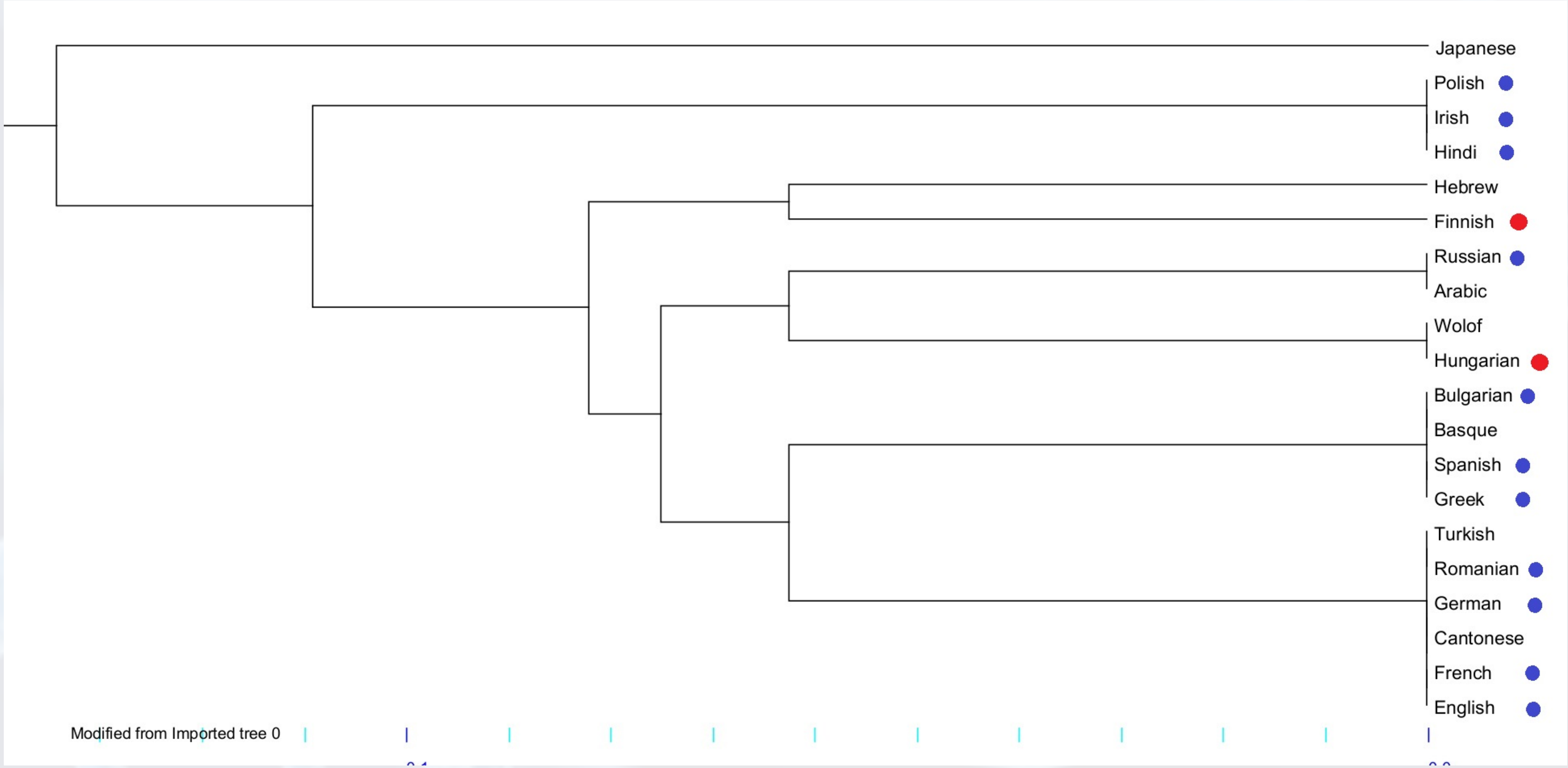
Empirical Test: Eurasia (different language families)

We can compute phylogenetic trees to check if phonemic inventories contain a historical signal

Distance-based trees

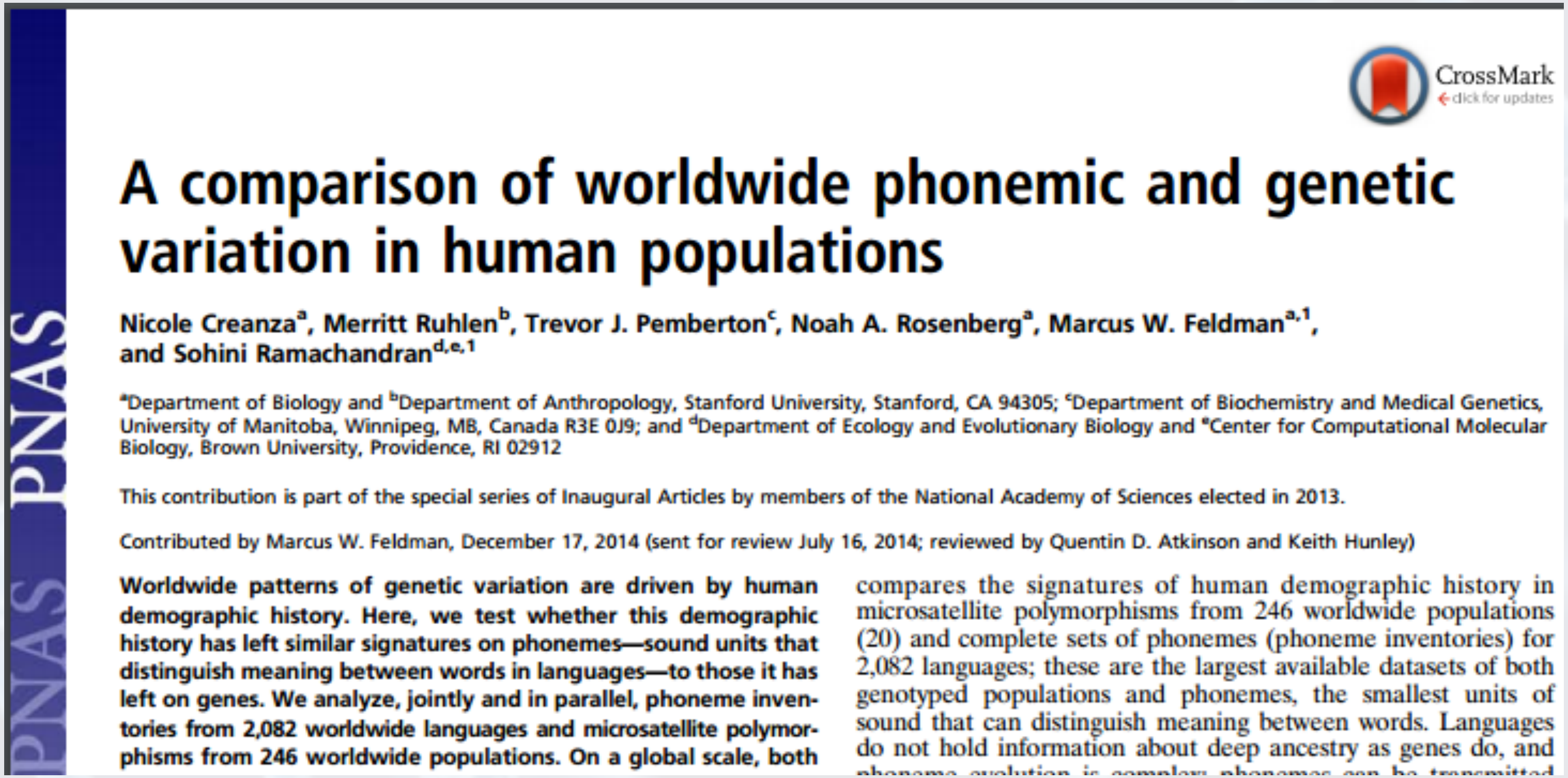
KITSCH (Phylip package)

Felsenstein (2004)



Tree calculated from the data in WALS
employed by Atkinson (2011)

Creanza et al. (2015)



A comparison of worldwide phonemic and genetic variation in human populations

Nicole Creanza[a], Merritt Ruhlen[b], Trevor J. Pemberton[c], Noah A. Rosenberg[a], Marcus W. Feldman[a,1], and Sohini Ramachandran[d,e,1]

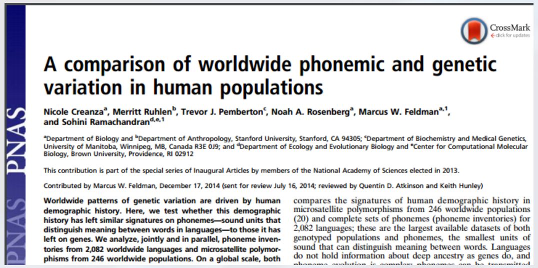
[a]Department of Biology and [b]Department of Anthropology, Stanford University, Stanford, CA 94305; [c]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, Canada R3E 0J9; and [d]Department of Ecology and Evolutionary Biology and [e]Center for Computational Molecular Biology, Brown University, Providence, RI 02912

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2013.

Contributed by Marcus W. Feldman, December 17, 2014 (sent for review July 16, 2014; reviewed by Quentin D. Atkinson and Keith Hunley)

Worldwide patterns of genetic variation are driven by human demographic history. Here, we test whether this demographic history has left similar signatures on phonemes—sound units that distinguish meaning between words in languages—to those it has left on genes. We analyze, jointly and in parallel, phoneme inventories from 2,082 worldwide languages and microsatellite polymorphisms from 246 worldwide populations. On a global scale, both

compares the signatures of human demographic history in microsatellite polymorphisms from 246 worldwide populations (20) and complete sets of phonemes (phoneme inventories) for 2,082 languages; these are the largest available datasets of both genotyped populations and phonemes, the smallest units of sound that can distinguish meaning between words. Languages do not hold information about deep ancestry as genes do, and phoneme evolution is complex; phonemes can be transmitted



A comparison of worldwide phonemic and genetic variation in human populations

Nicole Creanza[a], Merritt Ruhlen[b], Trevor J. Pemberton[c], Noah A. Rosenberg[a], Marcus W. Feldman[a,1], and Sohini Ramachandran[d,e,1]

[a]Department of Biology and [b]Department of Anthropology, Stanford University, Stanford, CA 94305; [c]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, Canada R3E 0J9; and [d]Department of Ecology and Evolutionary Biology and [e]Center for Computational Molecular Biology, Brown University, Providence, RI 02912

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2013.

Contributed by Marcus W. Feldman, December 17, 2014 (sent for review July 16, 2014; reviewed by Quentin D. Atkinson and Keith Hunley)

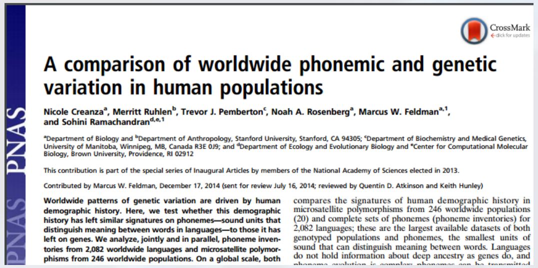
Worldwide patterns of genetic variation are driven by human demographic history. Here, we test whether this demographic history has left similar signatures on phonemes—sound units that distinguish meaning between words in languages—to those it has left on genes. We analyze, jointly and in parallel, phoneme inventories from 2,082 worldwide languages and microsatellite polymorphisms from 246 worldwide populations. On a global scale, both

compares the signatures of human demographic history in microsatellite polymorphisms from 246 worldwide populations (20) and complete sets of phonemes (phoneme inventories) for 2,082 languages; these are the largest available datasets of both genotyped populations and phonemes, the smallest units of sound that can distinguish meaning between words. Languages do not hold information about deep ancestry as genes do, and phoneme evolution is complex; phonemes can be transmitted

Genes are represented through microsatellite polymorphisms

Languages are represented through binary characters, which code for the absence/presence of phonemes:

	English	French	Japanese
/x/	1	0	0
/h/	1	0	1
/p/	0	1	1



PNAS PNAS

CrossMark

A comparison of worldwide phonemic and genetic variation in human populations

Nicole Creanza[a], Merritt Ruhlen[b], Trevor J. Pemberton[c], Noah A. Rosenberg[a], Marcus W. Feldman[a,1], and Sohini Ramachandran[d,e,1]

[a]Department of Biology and [b]Department of Anthropology, Stanford University, Stanford, CA 94305; [c]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, Canada R3E 0J9; and [d]Department of Ecology and Evolutionary Biology and [e]Center for Computational Molecular Biology, Brown University, Providence, RI 02912

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2013.

Contributed by Marcus W. Feldman, December 17, 2014 (sent for review July 16, 2014; reviewed by Quentin D. Atkinson and Keith Hunley)

Worldwide patterns of genetic variation are driven by human demographic history. Here, we test whether this demographic history has left similar signatures on phonemes—sound units that distinguish meaning between words in languages—to those it has left on genes. We analyze, jointly and in parallel, phoneme inventories from 2,082 worldwide languages and microsatellite polymorphisms from 246 worldwide populations. On a global scale, both

compares the signatures of human demographic history in microsatellite polymorphisms from 246 worldwide populations (20) and complete sets of phonemes (phoneme inventories) for 2,082 languages; these are the largest available datasets of both genotyped populations and phonemes, the smallest units of sound that can distinguish meaning between words. Languages do not hold information about deep ancestry as genes do, and

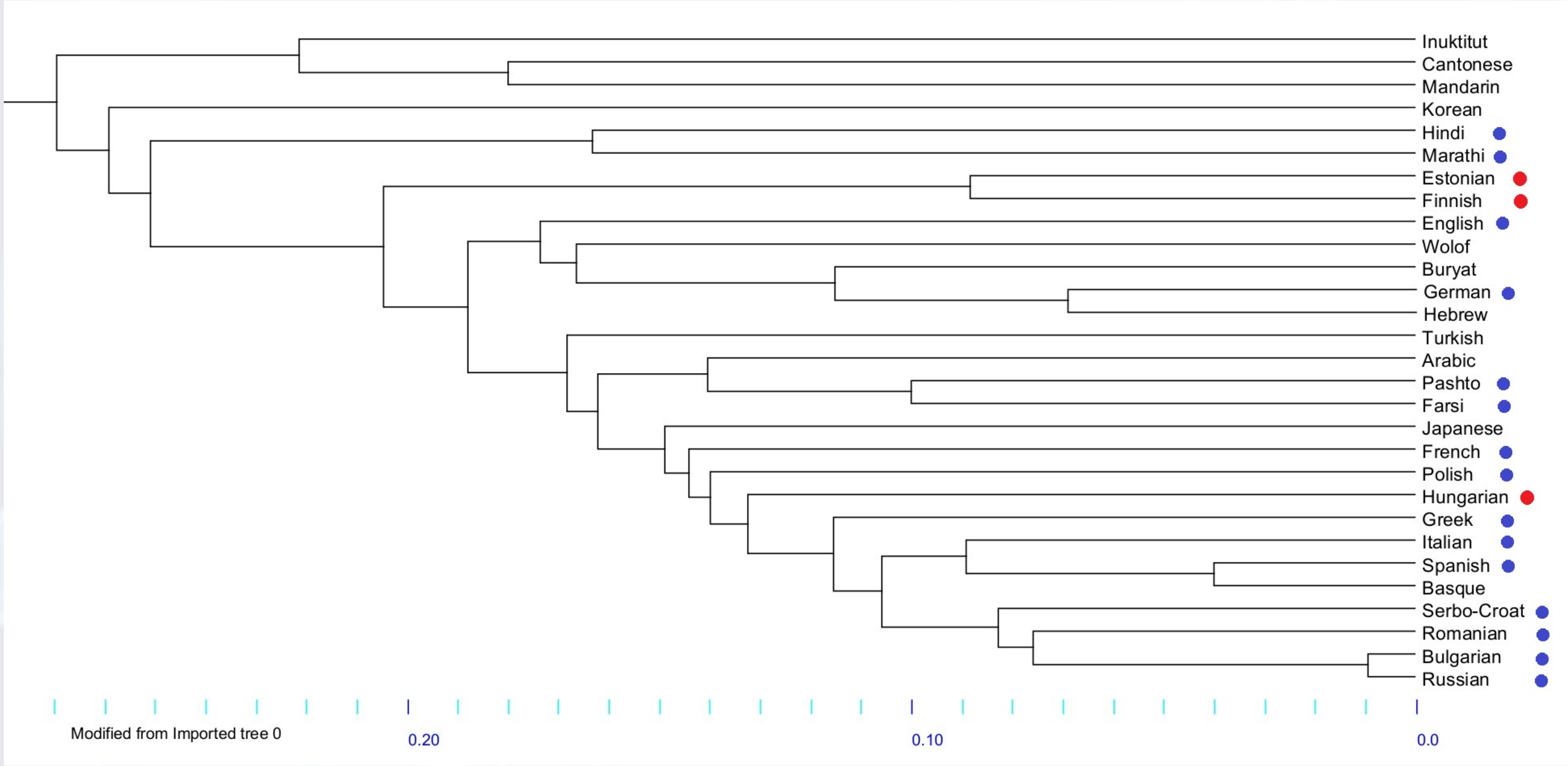
The major conclusions of the paper are:

1) No serial founder-effect out of Africa
2) Correlation between genes and languages is strong worldwide, but it is entirely predictable from geography (Partial Mantel Test: R= 0.05, p=0.16-0.17)
3) Geographical isolation leads to an increase in phonemic inventory sizes (vs. genetic drift)
4) The correlation between geography and phonemic distances ignores family boundaries (and saturates after 10.000 km)

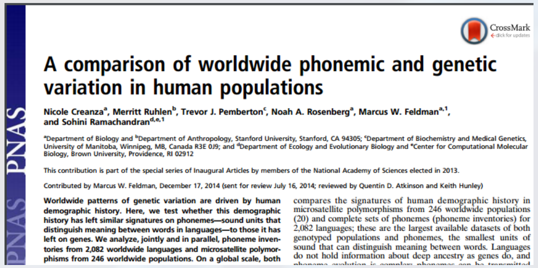
Distance-based trees

KITSCH (Phylip package)

Felsenstein (2004)



Tree calculated from a sample of the Ruhlen phonemic database



PNAS PNAS

CrossMark

A comparison of worldwide phonemic and genetic variation in human populations

Nicole Creanza[a], Merritt Ruhlen[b], Trevor J. Pemberton[c], Noah A. Rosenberg[a], Marcus W. Feldman[a,1], and Sohini Ramachandran[d,e,1]

[a]Department of Biology and [b]Department of Anthropology, Stanford University, Stanford, CA 94305; [c]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, Canada R3E 0J9; and [d]Department of Ecology and Evolutionary Biology and [e]Center for Computational Molecular Biology, Brown University, Providence, RI 02912

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2013.

Contributed by Marcus W. Feldman, December 17, 2014 (sent for review July 16, 2014; reviewed by Quentin D. Atkinson and Keith Hunley)

Worldwide patterns of genetic variation are driven by human demographic history. Here, we test whether this demographic history has left similar signatures on phonemes—sound units that distinguish meaning between words in languages—to those it has left on genes. We analyze, jointly and in parallel, phoneme inventories from 2,082 worldwide languages and microsatellite polymorphisms from 246 worldwide populations. On a global scale, both

compares the signatures of human demographic history in microsatellite polymorphisms from 246 worldwide populations (20) and complete sets of phonemes (phoneme inventories) for 2,082 languages; these are the largest available datasets of both genotyped populations and phonemes, the smallest units of sound that can distinguish meaning between words. Languages do not hold information about deep ancestry as genes do, and

“This suggests that phoneme inventories are affected by recent population processes and thus carry little information about the distant past”

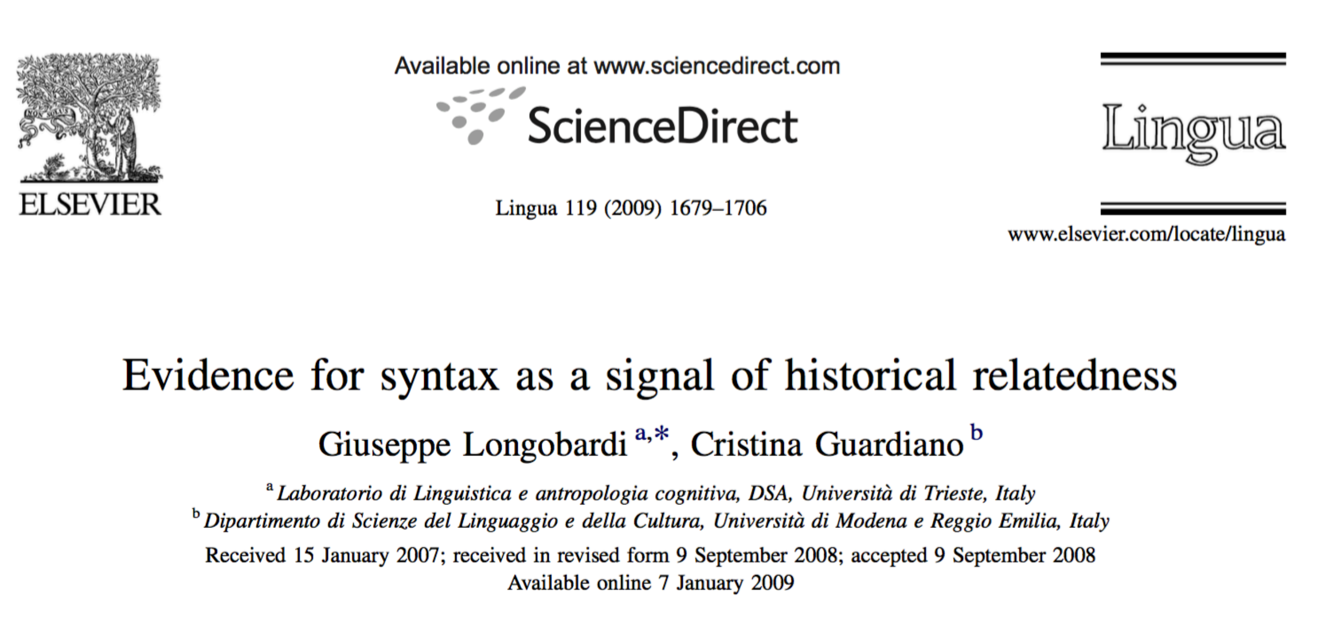
(Creanza et al. 2015:1269)

Language as an epiphenomenal (somewhat misleading) notion:

Speech (vocalization)
Symbolic function (vocabulary)
Grammar (syntax…)

Cognitive sciences, primatology: *they may have **co-evolved** and may historically **co-vary**, but this an empirical hypothesis*

Is there any other linguistic domain that can be used to study historical relationships at a global scale?



ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Lingua 119 (2009) 1679–1706

Lingua

www.elsevier.com/locate/lingua

Evidence for syntax as a signal of historical relatedness

Giuseppe Longobardi [a,*], Cristina Guardiano [b]

[a] *Laboratorio di Linguistica e antropologia cognitiva, DSA, Università di Trieste, Italy*

[b] *Dipartimento di Scienze del Linguaggio e della Cultura, Università di Modena e Reggio Emilia, Italy*

Received 15 January 2007; received in revised form 9 September 2008; accepted 9 September 2008
Available online 7 January 2009

Guardiano and Longobardi (2005) and Longobardi and Guardiano (2009) propose to look at generative **Syntax (PCM)**

Parametric Comparison Method

Longobardi (2003),

Guardiano and Longobardi (2005),

Longobardi and Guardiano (2009),

Longobardi, Guardiano, et al. (2013)

Parameter values may appropriately act as ***comparanda*** for historical reconstruction

It becomes possible:

- to **precisely calculate** the syntactic distance between any two languages
- to **assess** the **probabilistic value** of such distances

The syntax of the Nominal Domain (DP): 75 binary parameters (Guardiano and Longobardi 2016)

Crosslinguistic morphosyntactic difference > parameter

if and only if it entails

(A) the presence of **obligatory formal expression** for a semantic or morphological distinction (*grammaticalisation,* i.e. the obligatory presence of a feature in the computation to obtain the relevant interpretation and its coupling with an uninterpretable counterpart)

(B) the **variable form of a category** depending on the syntactic context (selection and feature agreement)

(C) the **position of a category** (movement, ±overt attraction triggered by grammaticalised features)

(D) The **availability** in the lexicon of certain functional categories (e.g. functional genitive projections)

Crossparametric Implications

Languages are encoded as lists of binary parameters (+,-)

Grammaticalized Person (FGP) and Strong Person (NSD)

	English	French	Chinese
FGP: gramm. person	+	+	-
NSD: strong person	-	+	?

Crossparametric Implications

Languages are encoded as lists of binary parameters (+,-)

Grammaticalized Person (FGP) and Strong Person (NSD)

	Conditions	English	French	Chinese
FGP: gramm. person		+	+	-
NSD: strong person	**(+FGP)**	-	+	**0**

TableA

```
It  -+---00+-+++++---0--+-+0--0000+++0+0+----+000-+--0--0+-+---00-+--
Sp  ++---00+-+++++---0--+-+0+++-+++++0+0+----+000-+--0--0+-+---00+++-
Fr  ++---00+-+-+++0--0--+-00-+-0+0+++0+0+----+000-+--0--0+-+---00++--
Ptg ++---00+-+++++---0--+-+0+-0000+++0+0+----+000-+--0--0+-+---00?+?-
Rm  ++---00+-+++++--+0--+-+00++-+++++0+0+----+000-+--0--0--+0--00--0-
Grk ++---00+-+++++---0--+-+0-++---+++0+0+------+0+---0--0-++0--+0--0+
E   ++---00+-++++----0--+-+00-0000+--0+0----------+--0--++-+0--0-0-0-
D   ++---00+-++++----0--+-+00-0000++-++0---------0+--0--++++---+00+0-
Da  ++---00+-++++---+---+-+00-0000+++++0---------0+--0--++-+0--0-0-0-
Ice ++---00+-++++---++--+--00-0000+++++0---------0+--0---+++0--+0+-0-
Nor ++---00+-++++---++--+-+00-0000+++++0---------0+--0--++-+0--0-0-0-
Blg ++---00+-+++++--+0--+--00-0000+++0+0---------0---0--0+-++--0+-+?-
SC  ++---00+-++-00-0000-+-0000++00+++0+0---------0+--0----+++--+00+0-
Slo ++---00+-++-00-0000-+-0000++00+++0+0---------0---0----+++--+00+0-
Po  ++---00+-++-00-0000-+-0000++00+++0+0--------+0---0----++---+00+0-
Rus ++---00+-++-00-0000-+-0000++00+++0+0--------+0---0----+++--+00+0-
Ir  ++---00+-++++----0--+--00+-0+-++--+0----+0000-0--0---+++0--00+-0-
Wel ++---00+-++++----0-----00+-0+-++--+0----+0000-0--0---+++0--00+-0-
Ma  ++---00+-++-00-0000-++0-00++00+++0-++--------00--0---+-00--0+0-0-
Hi  ++---00+-++-00-0000-++0-00++00+++0-++--------00--0---+-00--0+0-0-
Pas ++---00+-++-00-0000---0000++00+--0-+---------0+--0---+-00--0+0-0-
Man -0--+++00000000000000+0-0+++0-+000-----------00-00--+--00+00-0-0-
Can -0--++-00000000000000+0-0+++0-+000-----------00-00--+--00+00-0-0-
Ar  ++---00+-+++++---0+++--0-++-+-+++0+0+++000000+0--0+-0+++0--00--0-
Heb ++---00+-+++++---0--+--0-+-0+++++0+0-++000000+0+-0+-0+++0--00--0-
Hu  ++---00--+++++---0---+0--+++00+-+0+0---------0+-+-00000-0--0+0000
Est ++---00--++-00+0000---0000++00+++0+0---------00--0-+000-0---+0-0-
Fin ++---00--++-00+0000---0000++00+++0+0---------0+-+-00000-0--0+0000
Tur ++---00--++-00-0000--+0-00++00+--0-----------00++-0000000--0+0000
Bur ++---00--++-00-0000--+0-00++00+--0-----------00+-0---+-00--0+--0+
cB  ++---00-+00-0+0-00000-00+-0000+0+0---+--000000+--0--0+-00--000-0-
wB  ++---00++00-0+0+00000-00++-0-0+0+0---+--000000+--0--0+-00--000-0-
Wo  ++---00++00++-0+-0+00+0-+-0000-0+0+000--00000-0--0----++0+0+0?-0+
```

Distances

How to choose a distance measure?

Since we have a lot of '0' values, we cannot rely on a simple Hamming distance.

We can use a **Jaccard-Tanimoto distance** between “comparable” values:

$\delta(A,B) = d(A,B) / [d(A,B) + i(A,B)]$

= differences / identities + differences

E.g.: Italian-English: (35 id., 6 diff.) $\delta = 6 / 41 = 0.146$

Macro- and micro-classification

Indo-European:

Longobardi, G., Guardiano, C., Silvestri, G., Boattini, A., & Ceolin, A. (2013). Toward a syntactic phylogeny of modern Indo-European languages. *Journal of Historical Linguistics*, *3*(1), 122-152.

Greek and Romance micro-variation

Guardiano, C., D. Michelioudakis, A. Ceolin, M. Irimia, G. Longobardi, N. Radkevic, G. Silvestri, A. Sitaridou (2015) South by SouthEast. A syntactic approach to Greek and Romance micro- variation. *L'Italia Dialettale*.

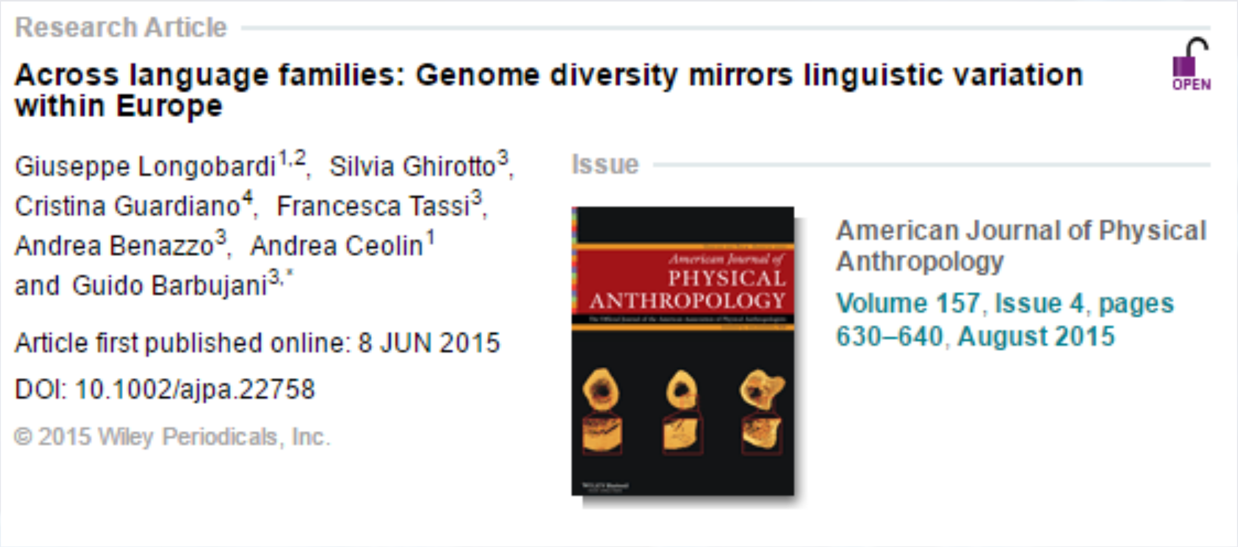
The classifications so obtained largely match the results of well-established and sophisticated methods

- ➢ extremely high correlation with distributions of etymological distances

Going beyond well-established families and beyond the historical depth of PIE, no other linguistic tools, e.g. etymology, can be used as benchmarks/standards of comparison.

- ➢ Do cross-family syntactic distances correlate with genetic distances? Is the correlation comparable to that of within-family distances?

Genes and Languages in Europe (15 populations)



Research Article

Across language families: Genome diversity mirrors linguistic variation within Europe

Giuseppe Longobardi[1,2], Silvia Ghirotto[3], Cristina Guardiano[4], Francesca Tassi[3], Andrea Benazzo[3], Andrea Ceolin[1] and Guido Barbujani[3,*]

Article first published online: 8 JUN 2015

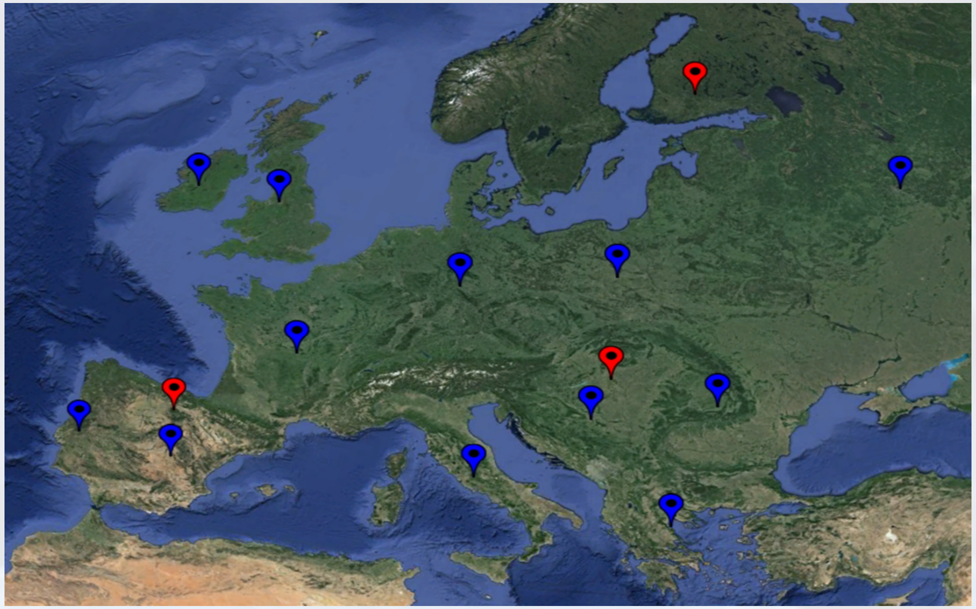
DOI: 10.1002/ajpa.22758

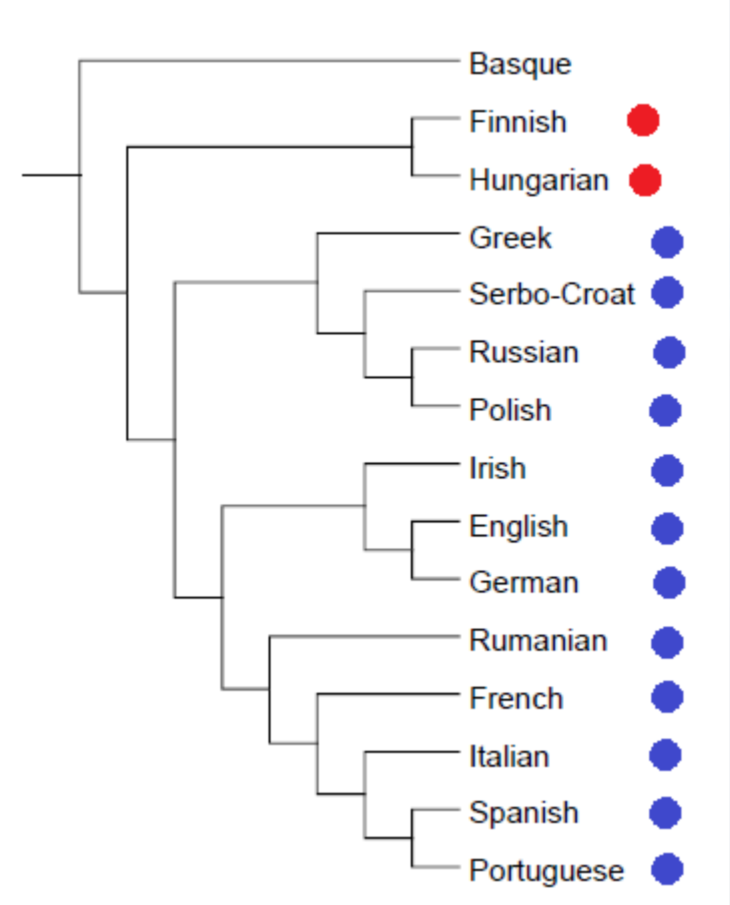
Issue

American Journal of Physical Anthropology
Volume 157, Issue 4, pages 630–640, August 2015

12 IE populations and 3 non-IE populations (Basque, Hungarians, and Finns) are analyzed from the viewpoint of their syntactic, lexical genetic and geographic distances.

Genes and Languages in Europe (15 populations)

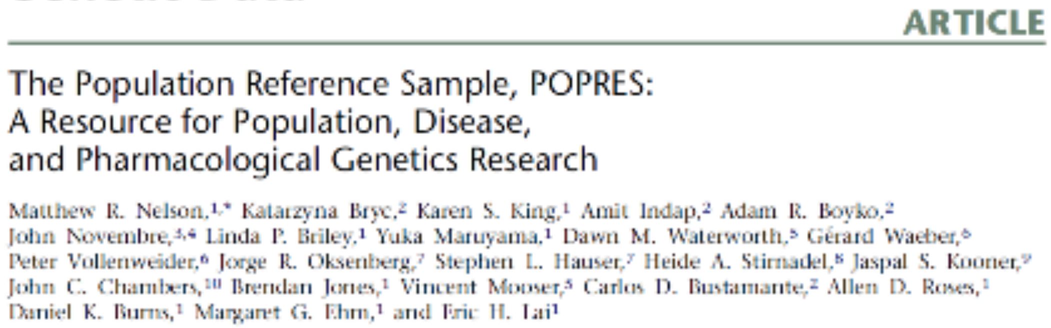




Syntactic differences are distributed following historical patterns (contrary to the phonological data in Atkinson 2011 and Creanza 2015)

Tree from **Longobardi et al. 2015**

Genetic Data

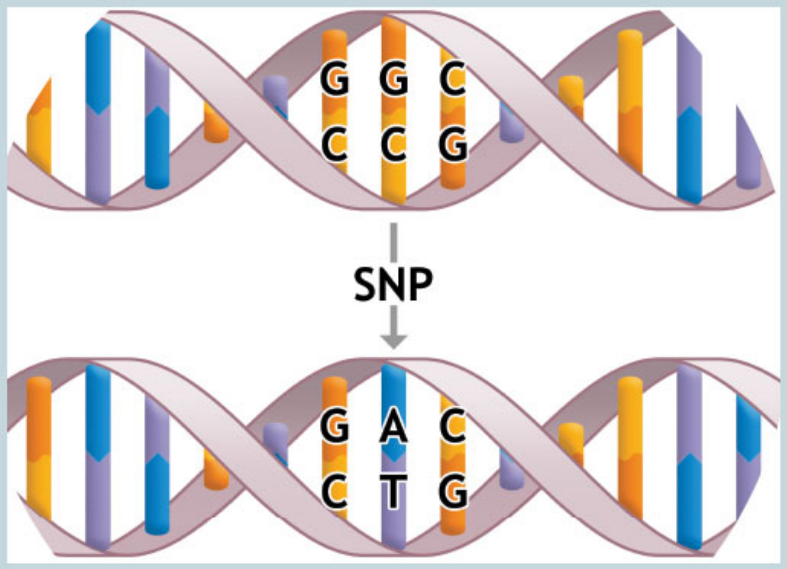


ARTICLE

The Population Reference Sample, POPRES:
A Resource for Population, Disease,
and Pharmacological Genetics Research

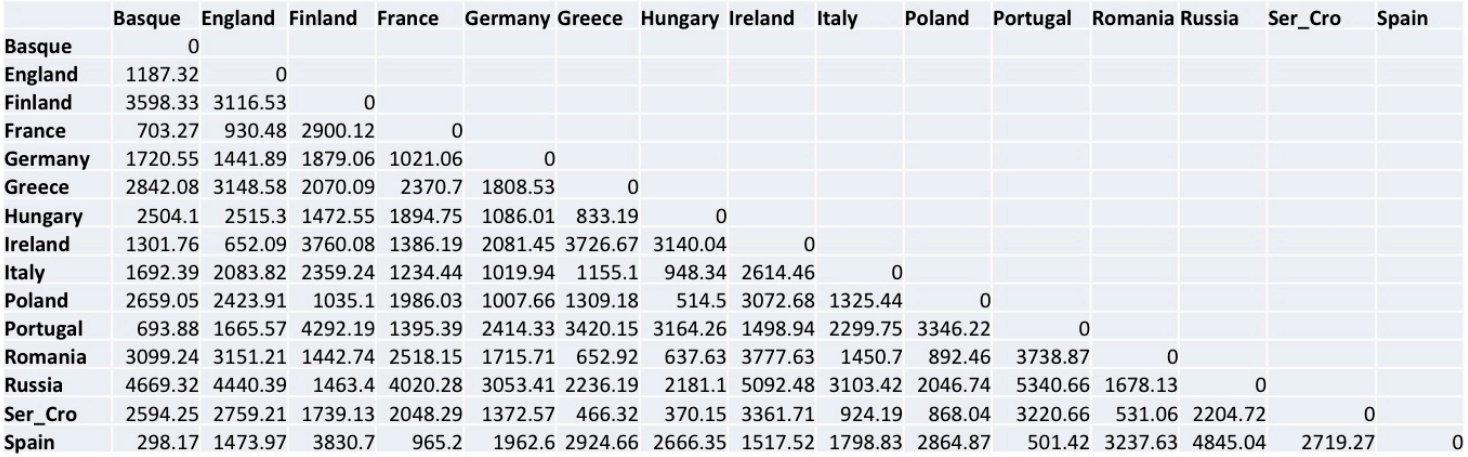
Matthew R. Nelson,[1,*] Katarzyna Bryc,[2] Karen S. King,[1] Amit Indap,[2] Adam R. Boyko,[2] John Novembre,[3,4] Linda P. Briley,[1] Yuka Maruyama,[1] Dawn M. Waterworth,[5] Gérard Waeber,[6] Peter Vollenweider,[6] Jorge R. Oksenberg,[7] Stephen L. Hauser,[7] Heide A. Stirnadel,[8] Jaspal S. Kooner,[9] John C. Chambers,[10] Brendan Jones,[1] Vincent Mooser,[5] Carlos D. Bustamante,[2] Allen D. Roses,[1] Daniel K. Burns,[1] Margaret G. Ehm,[1] and Eric H. Lai[1]

5,886 subjects genotyped at 500,568 loci using the Affymetrix 500K single nucleotide polymorphism

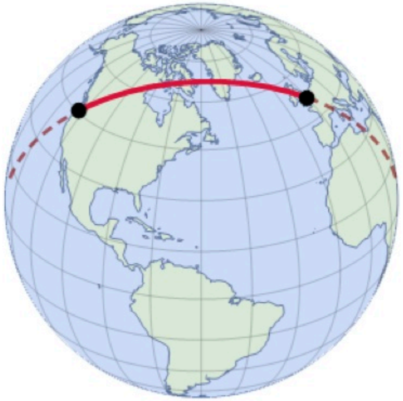


Geographic distances

Great Circle Distances (the shortest distance between two points on the surface of a sphere)



	Basque	England	Finland	France	Germany	Greece	Hungary	Ireland	Italy	Poland	Portugal	Romania	Russia	Ser_Cro	Spain
Basque	0														
England	1187.32	0													
Finland	3598.33	3116.53	0												
France	703.27	930.48	2900.12	0											
Germany	1720.55	1441.89	1879.06	1021.06	0										
Greece	2842.08	3148.58	2070.09	2370.7	1808.53	0									
Hungary	2504.1	2515.3	1472.55	1894.75	1086.01	833.19	0								
Ireland	1301.76	652.09	3760.08	1386.19	2081.45	3726.67	3140.04	0							
Italy	1692.39	2083.82	2359.24	1234.44	1019.94	1155.1	948.34	2614.46	0						
Poland	2659.05	2423.91	1035.1	1986.03	1007.66	1309.18	514.5	3072.68	1325.44	0					
Portugal	693.88	1665.57	4292.19	1395.39	2414.33	3420.15	3164.26	1498.94	2299.75	3346.22	0				
Romania	3099.24	3151.21	1442.74	2518.15	1715.71	652.92	637.63	3777.63	1450.7	892.46	3738.87	0			
Russia	4669.32	4440.39	1463.4	4020.28	3053.41	2236.19	2181.1	5092.48	3103.42	2046.74	5340.66	1678.13	0		
Ser_Cro	2594.25	2759.21	1739.13	2048.29	1372.57	466.32	370.15	3361.71	924.19	868.04	3220.66	531.06	2204.72	0	
Spain	298.17	1473.97	3830.7	965.2	1962.6	2924.66	2666.35	1517.52	1798.83	2864.87	501.42	3237.63	4845.04	2719.27	0



12 IE populations + Basque, Hungarians, and Finns

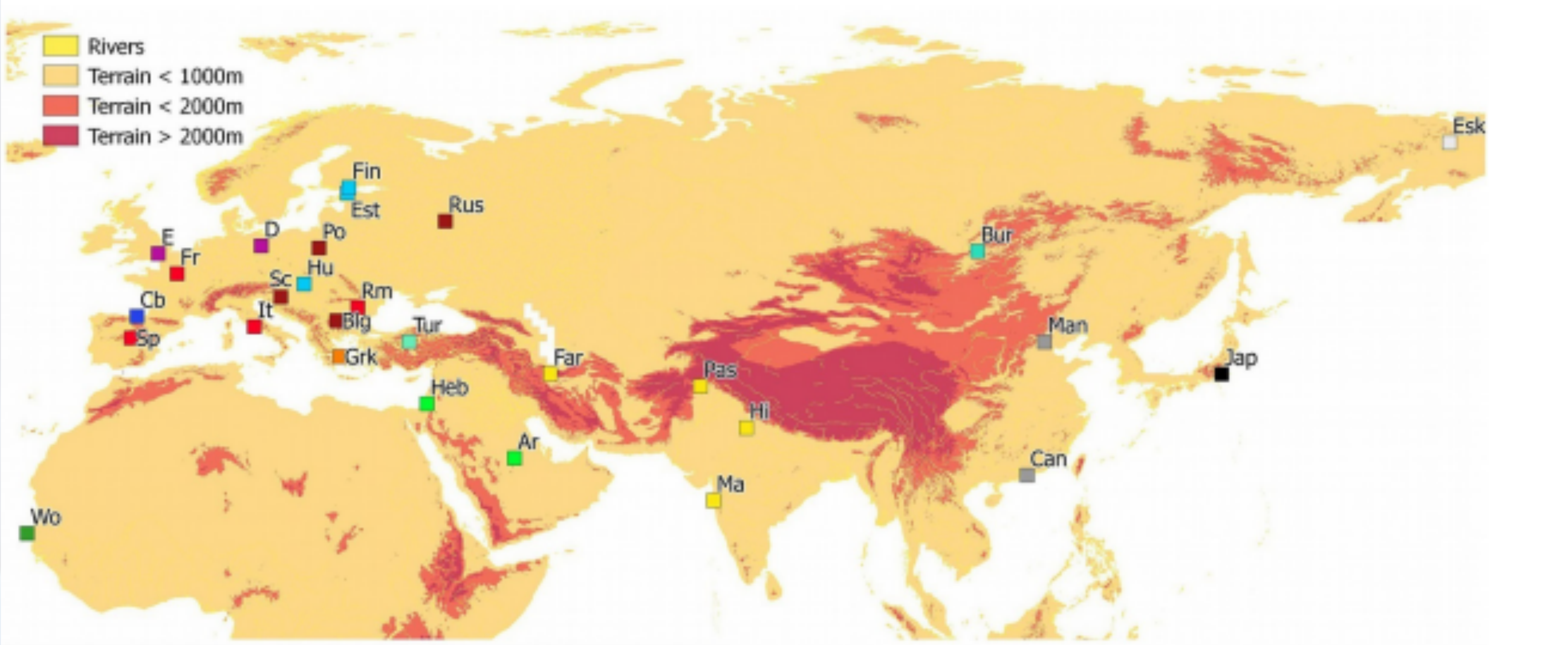
Distance matrices	***r***	***P***
d_{GEN} d_{GEO} Genetic - Geographic	0.299	0.030
d_{SYN} d_{GEO} Syntactic - Geographic	0.240	0.039
d_{SYN} d_{GEN} Syntactic - Genetic	**0.599**	0.001
d_{SYN} d_{GEN} (d_{GEO}) Syntactic - Genetic (Geography held constant)	**0.570**	0.002

An exception: Hungarian

“Careful analyses of 10th century ancient DNA in Hungary showed a predominance of European mitochondrial haplotypes in burials attributed to the lower classes, and a high incidence of Asian haplotypes in high-status individuals of that period (Tömöry et al. 2007), which points to the Asian immigrants as representing a social élite […]

[…] when a Finno-Ugric language was introduced in Hungary, the genetic buildup of the population changed only in part, thus retaining similarities with its geographic neighbors, an example of the process called **élite dominance** by Renfrew (1992).”

Next step: Eurasia (28 languages)



Indo-European (15)

Finno-Ugric (3)

Altaic (2)

Semitic (2)

Sinitic (2)

Niger-Congo (1)

Basque (1)

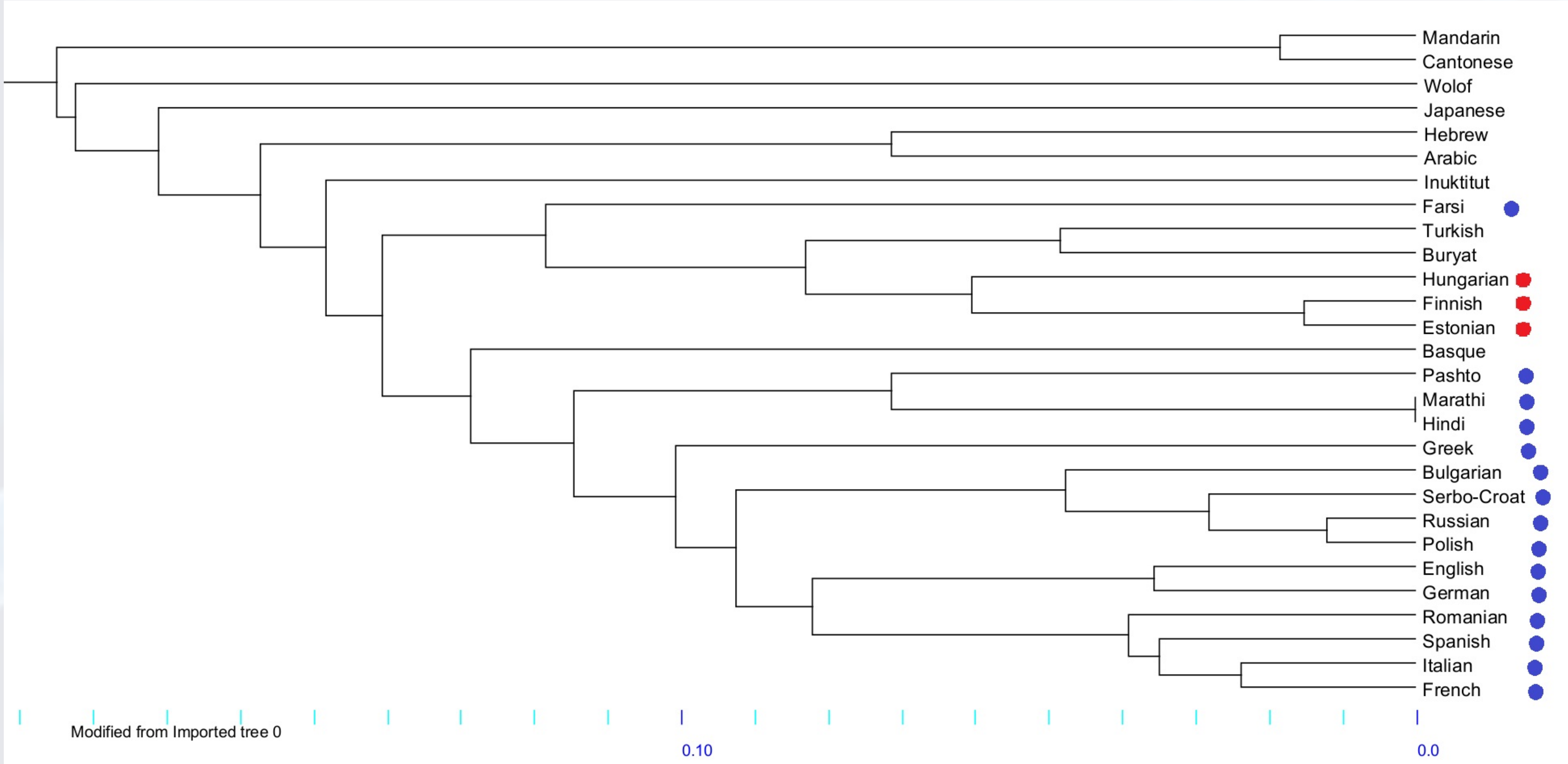
Japanese (1)

Inuit (1)

Distance-based trees

KITSCH (Phylip package)

Felsenstein (2004)



Correlations in Eurasia: 28 populations

Distance matrices	***r***	***P***
d_{GEN} d_{GEO} Genetic - Geographic	0.8319	0.0001
d_{SYN} d_{GEO} Syntactic - Geographic	0.4669	0.0001
d_{SYN} d_{GEN} Syntactic - Genetic	**0.5286**	0.0001
d_{SYN} d_{GEN} (d_{GEO}) Syntactic - Genetic (Geography held constant)	**0.2857**	0.0036

Syntactic and phonetic evidence in correlation with genes

	This study	**Creanza et al. (PNAS 2015)**	
	Syntactic	*Phonetic* (Ruhlen)	*Phonetic* (PHOIBLE)
d_{Gen}-d_{Lin}	**0.529** p=0.0001	**0.157** p=0.002	**0.240** p=0.0002
d_{Gen}-$d_{Lin(Geo)}$	**0.2857** p=0.0036	**0.05** p=0.16	**0.05** p=0.17

Syntactic and phonetic evidence in correlation with genes (Eurasia)

	This study	**Creanza et al. (PNAS 2015)**
	Syntactic	*Phonetic* (Ruhlen)
d_{Gen}-d_{Lin}	**0.529** p=0.0001	**0.4232** p=0.005
d_{Gen}-$d_{Lin(Geo)}$	**0.2857** p=0.0036	**0.0359** p=0.3344

Modeling geography

Great Circle Distances (GCD) are the standard measures in correlation studies

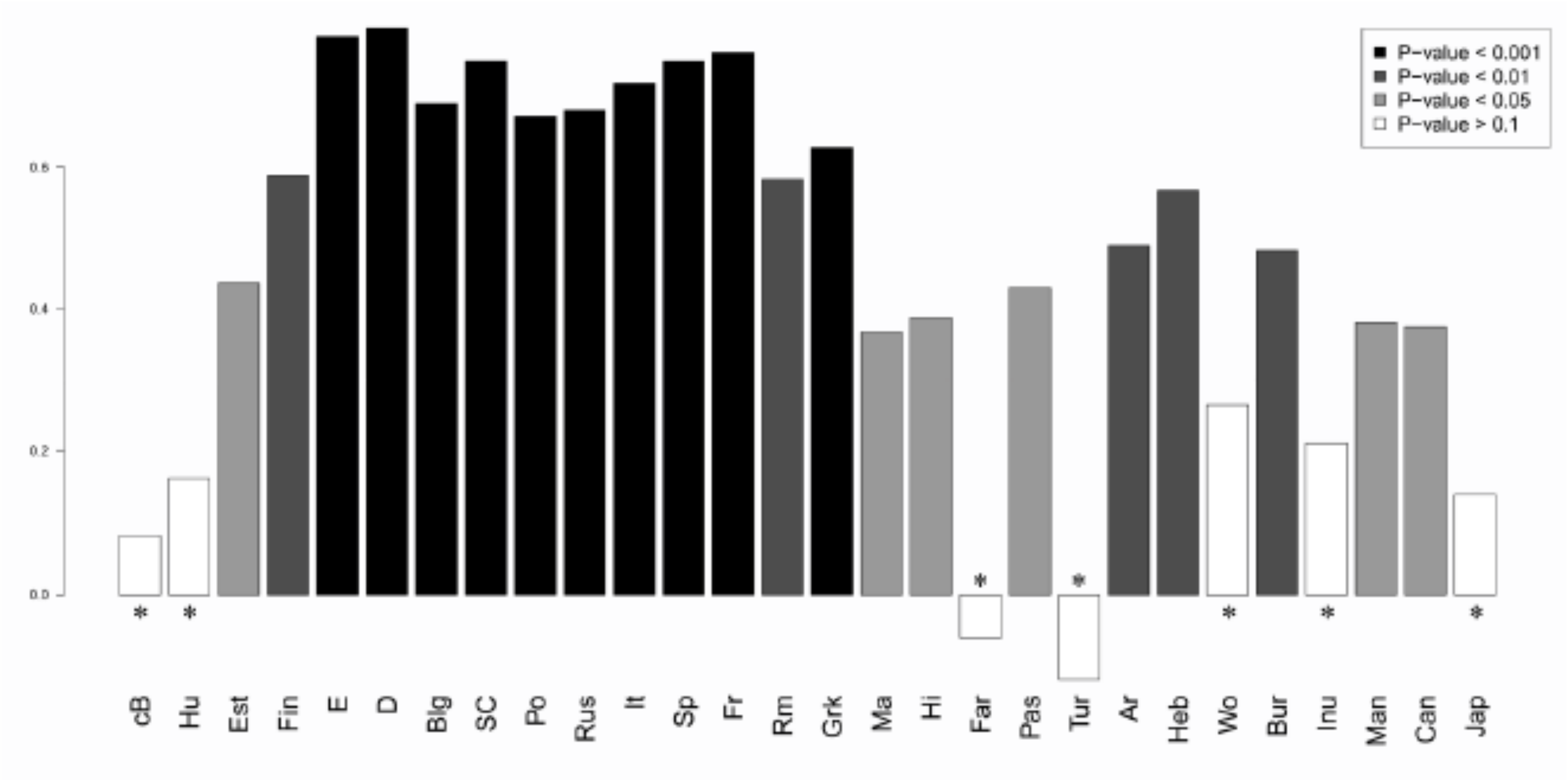
Can we test models closer to reality? We have four different models:

1- GCD with WayPoints
2- Road Maps
3- Least Cost Path
4- Resistance

Correlations in Eurasia: 28 populations

Distance matrices	***r***	***P***
d_{SYN} d_{GEN} (d_{GEO}) (GCD WayPoints)	**0.2770**	0.0063
d_{SYN} d_{GEN} (d_{GEO}) (RoadMaps)	**0.2641**	0.0082
d_{SYN} d_{GEN} (d_{GEO}) (Least Cost Path)	**0.3049**	0.0030
d_{SYN} d_{GEN} (d_{GEO}) (Resistance)	**0.3508**	0.0011

Partial correlations



Exceptions

Wolof: it is more salient as an outlier genetically rather than linguistically. This can derive from grammatical variation being more constrained (by UG?)

Inuktitut: likely to be an insufficient sampling approximation: the language is spoken in Eastern Canada, while the nearest genetic proxy available was in North-Eastern Asia

Exceptions

Hungarian is still an exception, as it was in Longobardi et al. (2015)

Turkish, Farsi, Basque, Japanese can all be explained in terms of **élite dominance** (like Hungarian) and related demographic processes

Conclusions

- An abstract deductive model of language structure/transmission/acquisition (based on a theory of UG) is surprisingly well reflected in the history of languages. It is only marginally affected by horizontal transmission and it can allow the investigation of macro-families

- Languages (modeled as cognitive objects at that abstract level) and genes seem to follow the same axes of variation **independently** of geography (vs. Creanza et al. 2015)

- A single process (**élite dominance**, Renfrew 1992) can explain a few cases of mismatch between linguistic and genetic variation

- Tools provided in the cognitive sciences might provide new insights for the historical study of human migrations across the world

THANKS!

Selected references:

Atkinson, Q. D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science, 332(6027), 346-349.

Creanza, N., Ruhlen, M., Pemberton, T. J., Rosenberg, N. A., Feldman, M. W., & Ramachandran, S. 2015. A comparison of worldwide phonemic and genetic variation in human populations. Proceedings of the National Academy of Sciences, 112(5), 1265-1272.

Longobardi, G. et al., 2015. Across Language Families: Genome diversity mirrors linguistic variation within Europe, American Journal of Physical Anthropology, 157(4):630-640.

Renfrew, C. 1992. Archaeology, genetics and linguistic diversity. Man, 445-478.