Genetic Influences on Tonogenesis and the Geographical Distribution of Tone Languages

MRI Scans of a microcephalic (left, 13 yo) and a normal (right, 11 yo) brains (Bond et al., Nature Genetics 32: 316-320, 2002)

World map of tonality: no tone, simple and complex tones. Generated with WALS (Haspelmath et al., 2005)

D. Robert Ladd
Linguistics and English Language
University of Edinburgh
bob@ling.ed.ac.uk

Dan Dediu
Language Evolution and Computation Research Unit
University of Edinburgh
d.dediu@sms.ed.ac.uk
The two genes

Legend:
- frequency of haplogroup D chromosomes
- frequency of non-D chromosomes

ASPM-D
Mekel-Bobrov et al. (2005:1721)

MCPH-D
Evans et al. (2005:1719)
The two genes: details

**ASPM** (Abnormal SPindle-like, Microcephaly-associated/MCPH5, OMIM 605481):
- **Chromosome 1** (1q31);
- haplogroup D ("derived") **ASPM-D**:
  - haplotypes with the derived allele $G$ at site A44871G;
  - age of MRCA: **5.8 ky** (95% CI 0.5-14.1 ky).

**Microcephalin** (**MCPH1**, OMIM 607117):
- **Chromosome 8** (8p23);
- haplogroup D ("derived") **MCPH-D**:
  - haplotypes with the derived allele $C$ at site G37995C;
  - age of MRCA: **37 ky** (95% CI 14-60 ky).

**Both ASPM & Microcephalin**:
- probably involved in brain growth and development;
- deleterious mutations (**not these**) are associated with high functioning Microcephaly;
- accelerated evolution of these genes in the human lineage (~2 favorable mutations/myr);
- haplogroups **D** ("derived"): show signs of natural selection;
- have a strong geographic structure in human populations.

**Notations**: MRCA = Most Common Recent Ancestor; ky = thousands of years; CI = confidence interval; OMIM = Online Mendelian In Man.
The hypothesis and its testing

**Hypothesis:**
there is a correlation between the frequencies of ASPM-D and MCPH-D in a population and the usage of *tone* contrasts in the language(s) spoken by it. This correlation is non-spurious, in the sense that it cannot be explained by other factors.

**Testing:**
- statistical in nature;
- assess the strength of the correlation between *tone*, ASPM-D and MCPH-D when controlling for other plausible explanatory factors (geography and history);
- compare this correlation with other correlations involving other genes and linguistic features (is it “special”?).

**Results:**
this correlation is strong, significant and “special”, even when controlling for geographical and historical explanations.
The data: populations

The two original papers (Mekel-Bobrov et al., 2005; Evans et al., 2005) use 59 poorly characterized world-wide populations, which we managed to identify (see handout). We excluded 10 populations from the analysis:

- **5 American populations** (*Karitiana, Surui, Colombian, Pima & Maya*) due to poor sampling of the Americas – used as a test case;
- **4 African populations** (*Masai, Sandawe, Burunge & Zime*) due to unavailability of genetic data;
- **Papuan** due to its vagueness and very high suspicions of European contamination of the sample.

Our sample: **49 Old World populations**.

**Australia**: unfortunately missing but very important for our hypothesis – Australian languages are strongly non-tonal.

**Caveats**: difficulty of obtaining genetic samples without European and East Asian admixture.
The data: genetics & linguistics

For each of our 49 populations we determined the language(s) spoken (Gordon, 2005; Campbell, 2000; Haspelmath et al., 2005).
The 141 linguistic features in Haspelmath et al. (2005) were carefully inspected and 24 met the selection criteria:
1. best coverage of the 49 populations;
2. meaningful coding as a binary variable: due to statistical considerations and uniformity of coding.
2 new features added by us → a total of 26 binary linguistic features.


For each of our 49 populations frequency information for 981 new genetic variants was gathered from public databases:
- ALFRED (ALlele FREquency Database; Rajeevan et al., 2003; Osier et al., 2002);
- Human Diversity Panel Genotypes (Rosenberg et al., 2002).
Selection criterion: to provide information for at least 44 of the 49 populations.
Analyzing the linguistic data

General obs.: Holm's (1979) multiple comparisons correction was systematically applied in all appropriate cases.

**Missing data analysis**: OK. Linguistic features' distribution: mostly in the 20%-80% bounds, except NumClassifiers, TenseAspect, FrontRdV & SVWO.

**Correlations between all pairs of linguistic features** (Pearson's $r \equiv \Phi$ correlation coefficient):

$N = 325$, mean = 0.012 & sd = 0.274.

Only 23 correlations are significant at the 0.05 level (see handout).

Alternative codings for Papuan (0, 1 or missing data) suggest that it can be safely neglected.
Analyzing the genetic data

The correlations (Pearson's $r$) between all pairs of genetic variants fit a normal distribution: $N = 482,653$, mean $= 0.024$ & sd $= 0.225$. 
Correlations between genetic variants & linguistic features

Pearson's $r$ (≡ Point-biserial correlation coefficient) for all pairs of genetic variants and linguistic features fit a normal distribution: $N = 25,558$, mean $= -0.006$ & sd $= 0.218$.

The correlations:
- **Tone & ASPM**: $r = -0.53$;
- **Tone & MCPH**: $r = -0.54$,
  both significant at $p < 0.05$ (two-tailed) and in the top 1.4% of the empirical distribution.
Correlations between linguistic features & pairs of genetic variants

Logistic regressions of all linguistic features on all pairs of genetic variants ($N = 11,582,690$ converged): heavily skewed towards poor fits.

The logistic regression of tone on ($ASPM$ & $MCPH$) is very good:
- Nagelkerke's $R^2 = 0.528$,
- correct classification: 73% and in the top 2.7% of the empirical distribution.

Estimates:
- Intercept = 4.478, $p = 0.015$
- $B_{ASPM} = -7.170$, $p = 0.010$
- $B_{MCPH} = -4.952$, $p = 0.026$
Controlling for geography and history: general considerations

| General method: | compute the correlation between distance/similarity matrices | Mantel (1967) corrects the p-level for matrices (still Pearson's $r$, though).
<table>
<thead>
<tr>
<th>1st &amp; 2nd order partial correlations</th>
<th>controls for one (or more) variables.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographic distances</td>
<td>estimate distances on land (handout):</td>
</tr>
<tr>
<td>• great circle distances between locations on the same continent;</td>
<td></td>
</tr>
<tr>
<td>• intercontinental paths pass through connection points: Damascus, Bangkok and Fairbanks.</td>
<td></td>
</tr>
<tr>
<td>Genetic distances</td>
<td>how genetically different are two populations (handout):</td>
</tr>
<tr>
<td>• Nei's $D$ (1972), depends on genetic variants' frequencies in the two populations.</td>
<td></td>
</tr>
<tr>
<td>Linguistic (feature) distances</td>
<td>how different are two languages (handout):</td>
</tr>
<tr>
<td>• weighted Euclidean distances on the linguistic features' space;</td>
<td></td>
</tr>
<tr>
<td>• three methods: (1) equal weighting of features, (2) weighting proportional to the feature's information entropy and (3) inversely proportional $\rightarrow$ equivalent.</td>
<td></td>
</tr>
<tr>
<td>Linguistic (historical) distances</td>
<td>how related (historically) are two languages (handout):</td>
</tr>
<tr>
<td>• the method of Nettle &amp; Harriss (2003), compatible also with Poloni et al. (1997)</td>
<td></td>
</tr>
<tr>
<td>• distance is: 1, same language; 2, languages in the same branch of a family; 3, languages in different branches of same family; or 4, languages not demonstrably related (using Gordon, 2005).</td>
<td></td>
</tr>
</tbody>
</table>
## Controlling for geography and history: results

The (partial) Mantel correlations (all genetic variants & linguistic features):

- Geography & genetics: \( r = 0.509, p = 0.000 \)
- Geography & linguistics: \( r = 0.283, p = 0.000 \)
- Genetics & linguistics: \( r = 0.162, p = 0.011 \)
- Genetics & linguistics, controlling geography: \( r = 0.021, p = 0.407 \)

→ in general, the correlation between genes & languages is explained by geography.

The (partial) Mantel correlations (only ASPM, MCPH & tone):

- Geography & tone: \( r = 0.169, p = 0.015 \)
- Geography & ASPM: \( r = 0.074, p = 1.000 \)
- Geography & MCPH: \( r = 0.543, p = 0.000 \)
- \( (ASPM, MCPH) \) & tone: \( r = 0.333, p = 0.000 \)
- \( (ASPM, MCPH) \) & tone, controlling geography: \( r = 0.291, p = 0.003 \)

→ ASPM, MCPH and tone correlate even after controlling for geography.

**Tone, ASPM & MCPH** – more similar inside than across linguistic families (independent samples t-test). The partial Mantel correlations between (ASPM, MCPH) and tone:

- controlling historical linguistics: \( r = 0.271, p = 0.000 \)
- controlling historical linguistics & geography: \( r = 0.283, p = 0.000 \)

The correlation between ASPM, MCPH and tone is significant and important even after controlling for geographical and historical linguistic factors.
The hypothesis of a correlation between tone, ASPM and MCPH holds, even after controlling geographical and historical linguistic factors.

Supports the claim that it is non-spurious.

Approximate threshold values:
- 0.293 (ASPM)
- 0.425 (MCPH).

Americas as predicted:
- low ASPM & high MCPH
tonal & non-tonal.

Australia – very important.
India – also potentially interesting.
PNG – probably contaminated.
Outline of Second Part of Talk

What could be the mechanism linking genes and tone?

What does it tell us about tonogenesis (and vice-versa)?

Where do we go from here?
A proposed mechanism: biasing the language transmission

The mechanism:
- basic idea: acquisition and/or processing biases may have an effect on the cultural transmission of language, influencing the trajectory of language change (e.g., Kirby & Hurford, 2001; Smith, 2003; etc.);
- increasing evidence for certain kinds of acquisition and processing biases.

The plausible types of biases:
- biases relevant to tone could be of various kinds:
  - related to perception (e.g., pitch tracking) or
  - related to the construction of cognitive categories and linguistic structures.

Here we focus on the difference between sequential and simultaneous structure.

Sequential structure: \[ C V C V C \]

Simultaneous structure: \[ T T C V C V C \]

Different individuals might have a bias toward one or the other type of structure.
Why a preference for sequential or simultaneous structure?

**Functional pressures** on structure as vocabulary increases

- If **sequential** structure becomes more complex (e.g. longer words, more complex syllable structure), there is an increase in demands on **phonological short-term memory**.
- If **simultaneous** structure becomes more complex (e.g. more suprasegmental phonemic distinctions), there is an increase in demands on **perceptual acuity**.

**Phonological memory and perceptual acuity:**

- are **variable across individuals** and may therefore be heritable
- may be related to **brain structure** (e.g. Golestani et al., *forthcoming*).
- may affect **language processing** (e.g. Swets et al., *in press*).
- could plausibly **vary across populations** as well as individuals
- could plausibly be **involved in acquisition biases**
Example:
Some Diary Data

From a child acquiring English, around the “50-word stage” (handout).

Minimal pairs/sets involving pitch and/or voice quality, ~16 months (persisted 3-4 months):

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Description</th>
<th>Word(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ka]</td>
<td>(high-pitched, breathy)</td>
<td>‘cat’</td>
</tr>
<tr>
<td>[ka]</td>
<td>(modal)</td>
<td>‘bottle/cup/milk’</td>
</tr>
<tr>
<td>[ka]</td>
<td>(low-pitched, pressed)</td>
<td>‘duck’</td>
</tr>
<tr>
<td>[a]</td>
<td>(modal)</td>
<td>‘rabbit’ or ‘hummus’</td>
</tr>
<tr>
<td>[a]</td>
<td>(low-pitched, pressed)</td>
<td>‘dog’, later also ‘tiger’</td>
</tr>
<tr>
<td>[a]</td>
<td>(voiceless)</td>
<td>‘hot’</td>
</tr>
</tbody>
</table>

At ~17 months a few other such contrasts were added (lasted for 1-2 months):

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Description</th>
<th>Word(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ti]</td>
<td>(high-pitched, breathy)</td>
<td>‘teddy’</td>
</tr>
<tr>
<td>[ti]</td>
<td>(modal)</td>
<td>‘teeth’ or ‘chair’ or ‘juice’</td>
</tr>
<tr>
<td>[kika]</td>
<td>(modal)</td>
<td>‘scaffolding’</td>
</tr>
<tr>
<td>[kika]</td>
<td>(voiceless)</td>
<td>‘leopard’</td>
</tr>
</tbody>
</table>

Note also isolated words:

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Description</th>
<th>Word(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[nənənə]</td>
<td>(high-pitched, breathy, throughout)</td>
<td>‘penguin’ (ca. 16 months)</td>
</tr>
<tr>
<td>[kuka]</td>
<td>(modal 1st syll., low pressed 2nd syll.)</td>
<td>‘duck book’ (ca. 16 months)</td>
</tr>
</tbody>
</table>

There is a similar anecdotal report in Crystal (1975: 148f).
Could it be that children who use non-segmental features in their early lexicon this way are biased to expect simultaneous phonemes?
Population bias for/against tone should manifest itself in patterns of tonogenesis and tone loss. Both processes are known to have happened in widely separated (and genetically different) areas of the globe: bad news for our claim?

We propose to distinguish two types of tonogenesis:
- secondary split
- spontaneous reanalysis.

Tonogenesis is usually linked to the loss of some other phonological distinction:
- loss of final syllables (e.g. Scandinavia)
- loss of voicing contrasts (e.g. SE Asia)
- loss of laryngeal contrasts (e.g. Athabaskan)

In secondary split, tonogenesis is a compensatory mechanism: the loss of another distinction “forces” speakers to rely on a previously allophonic difference to preserve the cue to some important grammatical or lexical distinction. Genetic biases are at least partially irrelevant.

In spontaneous reanalysis, tonogenesis involves a reinterpretation of the relation between acoustic cues and phonological categories. Genetic biases could be active.
Voicing-based tonogenesis as spontaneous reanalysis

The cues to the distinction between “[ba]” and “[pa]” are distributed throughout most of the syllable.

A sequential bias will favor treating the burst+VOT as one phonological event and the voiced vowel as the next.
A simultaneous bias will favor treating the combination of VOT and higher/lower F0 as a phonological event simultaneous with the vowel event.

The same acoustic signal can be interpreted phonologically in two different ways; there is no need to invoke secondary split as an explanation for this kind of tonogenesis.
Kera shows three types of stop-vowel sequences, conventionally notated as e.g. [pá], [pā] and [bà] (conventional notation devised by European linguists based on impressionistic transcription.)

But VOT and F0 covary (and substantially overlap). Tone is primary basis of 3-way distinction for most native speakers and “voicing” not relevant (or not even perceived?) except that speakers living in town and exposed to French seem to be developing a voicing distinction and reducing the tonal distinction to [±High], with subtle but measurable effects on their production.
Kera: Diagrams

Kera village men: normalized F0 and VOT

Town Kera men: normalized F0 and VOT. 5 speakers
Where do we go from here?

Others’ work:
- Much ongoing work on acquisition and processing biases
- Much ongoing work on relation between brain structure and linguistic behaviour
- Some ongoing work on relation between genes and brain structure

Our work:
- Theoretical work on distinction between sequential and simultaneous structure (Ladd, Leverhume Individual Research Fellowship, 2007-2008)
- Computational and mathematical modeling in order to explore the conditions and consequences of this type of relationships (Dediu)
- Experimental studies: a series of experimental studies, attempting to test the existence and nature of this bias at the individual level (projects under development).