Reference Databases for mtDNA Casework: Examples from Central Asia

Jodi A. Irwin
Forensically relevant mtDNA issues apparent in the Central Asian dataset

- mtDNA data quality
- Nomenclature
- Heteroplasmy
- Reference population databases
  - appropriate and “representative”
  - Population mtDNA variation and sub-structure
## Central Asian Dataset

<table>
<thead>
<tr>
<th>Country</th>
<th>Population N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>98</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>256</td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td>249</td>
</tr>
<tr>
<td>Russia</td>
<td>151</td>
</tr>
<tr>
<td>Tajikistan</td>
<td>244</td>
</tr>
<tr>
<td>Turkmenistan</td>
<td>249</td>
</tr>
<tr>
<td>Uzbekistan</td>
<td>328</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1575</strong></td>
</tr>
</tbody>
</table>
Uzbekistan sub-populations:
Karakalpakstan, Tashkent, Qashkadarya, Fergana, Xorezm

Total Sample Number = 328
Reporting Statistics

- When “cannot exclude” is the interpretation, then a statistical estimate is needed in order to weigh the significance of the observed match.

- Counting method is most common approach used and involves counting the number of times that a particular mtDNA haplotype has been observed in a database.

- Estimated mtDNA haplotype frequencies should be interpreted in the context of mtDNA distributions among, and potential substructure of, relevant populations (Carracedo et al. 2000).

- Tully et al. (2001) go on to suggest that ‘small, relatively isolated European populations need to be analysed in order to improve understanding of the population genetics of mtDNA at the local level.


Since the vast majority of mtDNA haplotypes are unique, larger databases tend to increase the strength of the evidence in most cases. However, the degree to which separate mtDNA databases must be maintained is still not well understood for many “populations”.

Diverse sampling is required to determine the magnitude and significance of inter-population differentiation, and the level at which separate databases should be maintained.

Buckleton, Triggs, Walsh “…further investigation into how to compensate for population subdivision at the mtDNA locus is warranted urgently. In the absence of new theory, it is imperative that every effort should be made to use appropriate local databases and hence no correction or a low value for $\theta$.”

One of the biggest issues in forensics presently concerns the size, sampling, and quality of forensic mtDNA databases.
AFDIL Control Region Databasing

- International Collaborators
- Specifically target populations that are not well represented in available databases
- Provide entire control region data
- Generate consistent, high quality data (*EMPOP collaboration*)
- Adhere to a consistent nomenclature scheme
- Make data publicly available, via publications, GenBank and EMPOP.
- Describe and better understand the mtDNA diversity of local and underrepresented populations
AFDIL’s Recent Global DB Efforts

Global populations databased since late 2004

<table>
<thead>
<tr>
<th>Population/Region</th>
<th>Sub-Population/Region</th>
<th>Total # Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td></td>
<td>98</td>
</tr>
<tr>
<td>Bahrain</td>
<td></td>
<td>218</td>
</tr>
<tr>
<td>China</td>
<td>Hong Kong</td>
<td>377</td>
</tr>
<tr>
<td>Cyprus</td>
<td></td>
<td>91</td>
</tr>
<tr>
<td>UAE</td>
<td></td>
<td>191</td>
</tr>
<tr>
<td>Egypt</td>
<td></td>
<td>278</td>
</tr>
<tr>
<td>Greece</td>
<td></td>
<td>319</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Sulawesi</td>
<td>279</td>
</tr>
<tr>
<td>Hungary</td>
<td>Budapest Caucasian Baranya Roma</td>
<td>215</td>
</tr>
<tr>
<td>Iraq</td>
<td></td>
<td>189</td>
</tr>
<tr>
<td>Jordan</td>
<td></td>
<td>210</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td></td>
<td>256</td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td></td>
<td>249</td>
</tr>
<tr>
<td>Kenya</td>
<td>Nairobi</td>
<td>103</td>
</tr>
<tr>
<td>Lebanon</td>
<td></td>
<td>198</td>
</tr>
<tr>
<td>Pakistan</td>
<td></td>
<td>433</td>
</tr>
<tr>
<td>Russia</td>
<td></td>
<td>151</td>
</tr>
<tr>
<td>Tajikistan</td>
<td></td>
<td>244</td>
</tr>
<tr>
<td>Turkmenistan</td>
<td></td>
<td>249</td>
</tr>
<tr>
<td>Uzbekistan</td>
<td></td>
<td>328</td>
</tr>
<tr>
<td>Vietnam</td>
<td></td>
<td>187</td>
</tr>
</tbody>
</table>

**TOTAL** 5074

In addition, we are databasing regional populations of the U.S. – over 6000 regional U.S. samples sequenced since 2004
AFDIL Control Region Databasing

- International Collaborators
- Specifically target populations that are not well represented in available databases
- Provide entire control region data
- Generate consistent, high quality data
- Adhere to a consistent nomenclature scheme
- Make data publicly available, via publications, GenBank and EMPOP.
- Describe and better understand the mtDNA diversity of local and underrepresented populations
Detecting errors in mtDNA data by phylogenetic analysis

Errors primarily result from sequence data artifacts and transcription mistakes

Commentary

To Err is Human

P. Forster

“more than half of the mtDNA sequencing studies ever published contain obvious errors…”

Annals of Human Genetics 2003 67:2-4
How are errors detected \textit{a posteriori}?

Phylogenetic analysis that evaluates the data within the context of known mtDNA variation

Highlights polymorphisms that are either rare or incompatible with the mtDNA phylogeny
Reduced Network with Filtering

Good data

Poor data

Safeguards against DB Errors

- **Multiple scientists at key laboratory steps** – initial sample placement, cherry-picking for re-dos.

- **Robust robotics** - standard placement of samples, reagent blanks, negative controls; elimination of sample switches at every step.

- **Redundant data review** – At least 3 scientists review the RAW sequence data for every sample. Conducted in collaboration with scientists at EMPOP

- **Electronic data transfer** – No manual transcription of data. Electronic transfer both into and out of master database.

- **All data cross-checked against common phantom mutations (sequencing) artifacts** - recently implemented at AFDIL

- **Phylogenetic data checking and review** - EMPOP
Highly redundant sequencing strategy avoids “phantom mutations”
General Overview

Lab Processing

Data Analysis
- 2 AFDIL Scientists
  - electronic data transfer
  - data confirmation
  - phantom mutation x-check

Local DB entry
- electronic data transfer
- data confirmation
- ‘finalized’ haplotypes stored in local DB

Data Transfer to EMPOP
- raw sequence data

1 EMPOP Scientist
- data confirmation
- haplogroup assignment
- phylogenetic data check

AFDIL/EMPOP data comparisons
Reconciliation of any differences

phantom mutation x-check
phylogenetic data check
Data Analysis Differences between AFDIL/EMPOP

- Of the 1575 sequences in the Central Asian Dataset...

  17 - # of samples that differed between AFDIL and EMPOP’s analyses
  9 – Length heteroplasmy interpretation
  5 – Nomenclature differences – we’ll re-visit
  3 – Mistakes

  pre-phantom mutation screen
  pre-phylogenetic check
Data cross-check discrepancies

- **Two samples** – alignment/transcription errors
  - 309.1C mis-scored as 302.1C
  - 249del mis-scored as 249T *(note – hmmm… not at alignment, I don’t think)*

- **One sample** – phantom mutation
  - 527G

---

**Brandstätter et al.**
Electrophoresis 2005
Had these errors made it beyond the data cross-check and been used in the phylogenetic checks, this is what we would have seen....
Alignment/transcription errors

249T
Uzbek/Tashkent network

302.1C
Uzbek/Xorezm network
C527G - Phantom Mutation

Network with error

Network without error

Uzbek/Qashkadarya network
Generation of high-quality data

- With multiple safeguards in place at all steps of the process, and an additional phantom mutation screen recently implemented, the final Network analysis will hopefully be superfluous.

- However, these examples also demonstrate that even given numerous safeguards, mistakes can happen.
Data Analysis Differences between AFDIL/EMPOP

- Of the 1575 sequences in the Central Asian Dataset…

17 - # of samples that differed between AFDIL and EMPOP’s analyses

9 – Length heteroplasmy interpretation

5 – Nomenclature differences – we’ll re-visit

3 – Mistakes

pre-phantom mutation screen
pre-phylogenetic check
Nomenclature

- Representation of the sequence as a list of differences from the rCRS
- Treatment of insertions and deletions
- How to place them relative to the rCRS?
- This affects database searches –
  - Problems can arise if the nomenclature of the queried sequence differs from the nomenclature of the database searched
Nomenclature: An example...

Russia0080
C16186T
T16189del

Or...
C16186T
T16189C
C16193del
Nomenclature

- Guidelines suggested by the FBI in 2002
  *FSI (2002) 129:35*
  - Hierarchical model based on
    - minimizing the number of differences between the questioned haplotype and the rCRS,
    - differentially weighting indels, transitions and transversions.
Nomenclature problems

- Not all laboratories are following these guidelines
- These guidelines do not (and cannot) encompass all of the unique situations encountered
- There are particular situations for which the evolutionary history of a length variant haplotype provides additional information upon which interpretation can be based (Bandelt and Parson, *IJLM 2006*)
Nomenclature problems

- At best, these nomenclature differences will be reflected in hypervariable C-stretch regions that are generally ignored in evidence interpretation.
- At worst, these differences will underestimate the frequency of particular haplotypes.
- “In the case of an unusual/complex sample with indel variation, the practitioner must be able to conceive of all possible calling alternatives to search correctly” – Terry Melton, Mitotyping Technologies
Russia0080

8 consistent haplotypes in database

C16186T
T16189del

0 consistent haplotypes in database

C16186T
T16189C
C16193del

Variants associated with hg T1a
Nomenclature Issues

- Recent suggestions by Parson and Bandelt suggest using phylogenetic information to guide indel placement.

- In our experience, these guidelines resolve the vast majority of cases. However, an intimate knowledge of mtDNA evolution and the mtDNA literature is required.

- No matter which guidelines are followed, some of these variants are so tricky that they may slip through even the most careful evaluations.

- In our own hands, despite attempts to maintain ‘consistency’, we are encountering samples with inconsistent nomenclature.
A couple of tricky examples…
If we remove rCRS from the alignment and align the two samples to each other:

If we look at the entire CR haplotype and not just that variable region

- Both of these samples are on a particular haplotypic background
And finally, a review of the literature reveals...

- A similar haplotype described by Achilli et al. (AJHG 2004) with no insertion in the region between 54-60
Uzb-Q-085: T55C, 56.1C

Kyrg-015: 54.1C, A56C

Uzb-Q-085: T55C, T57C, 60.1T

Kyrg-015: T55C, A56T, T57C, 60.1T
Following a global alignment that considers similar haplotypes…

2 diffs

3 diffs
Nomenclature – bottom line

Consistency is difficult to maintain due to the extreme variability of the mtDNA control region

Be aware of unusual length variants and potential alternate ‘calls’
Central Asia

- Region with an extremely rich history in terms of human demographics
  - Major corridor for different population migrations between Asia, the Middle East, India and Europe
    - Comas et al. EJHG 2004 detected a high proportion of sequences originating elsewhere; suggesting intense gene flow
  - More recent political changes in and around the area have occurred and may have contributed to the extreme genetic heterogeneity of the region

- This history and molecular diversity must be taken into consideration when genetic markers are used for forensic purposes

- Unique aspects of the populations in this data set introduce additional considerations
Diversity Indices for Sub-Populations of Uzbekistan

<table>
<thead>
<tr>
<th>Population Statistic</th>
<th>Fergana (n = 53)</th>
<th>Karakalpakstan (n = 46)</th>
<th>Qashkadarya (n = 75)</th>
<th>Tashkent (n = 55)</th>
<th>Xorezm (n = 99)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pairwise Random Match Prob.</td>
<td>1.60%</td>
<td>0.30%</td>
<td>0.18%</td>
<td>0.61%</td>
<td>0.18%</td>
</tr>
<tr>
<td>Haplotypes</td>
<td>40 (6)</td>
<td>43 (3)</td>
<td>71 (4)</td>
<td>50 (3)</td>
<td>94 (3)</td>
</tr>
<tr>
<td>Mean Pairwise Differences</td>
<td>12.9</td>
<td>12.2</td>
<td>11.9</td>
<td>12.5</td>
<td>12.1</td>
</tr>
<tr>
<td>Genetic Diversity</td>
<td>0.987</td>
<td>0.997</td>
<td>0.998</td>
<td>0.994</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Population statistics for five sub-populations of Uzbekistan. Random match probabilities were generated empirically. Polymorphic sites do not include C insertions at 16193, 309, or 573. Haplotype numbers in parentheses indicate the subset of total haplotypes shared among individuals.
Haplogroup Distributions of Uzbekistan sub-populations: Karakalpakstan, Tashkent, Qashkadarya, Fergana, Xorezm
Φst Values based on Haplotype Data in Various Sub-populations of Uzbekistan

<table>
<thead>
<tr>
<th></th>
<th>Fergana</th>
<th>Karakalpakstan</th>
<th>Qashkadarya</th>
<th>Tashkent</th>
<th>Xorezm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fergana</td>
<td>-0.00324</td>
<td>0.00779</td>
<td>0.00937*</td>
<td>0.00816*</td>
<td></td>
</tr>
<tr>
<td>Karakalpakstan</td>
<td>0.00201</td>
<td>-0.00178</td>
<td>-0.00075</td>
<td>0.00075</td>
<td></td>
</tr>
<tr>
<td>Qashkadarya</td>
<td>0.00201</td>
<td>0.00919*</td>
<td>0.0004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tashkent</td>
<td></td>
<td></td>
<td>0.00704*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Values are significant at the 0.05 level.

- While some of these values are statistically significant, but the magnitude of the inter-population differences is marginal in each case
- Furthermore, if the Bonferroni correction is applied, the differences are no longer statistically significant
Diversity Indices for Seven Central Asian Populations

<table>
<thead>
<tr>
<th>Population Statistic</th>
<th>Afghanistan (n = 98)</th>
<th>Kazakhstan (n = 256)</th>
<th>Kyrgyzstan (n = 249)</th>
<th>Russia (n = 151)</th>
<th>Tajikistan (n = 244)</th>
<th>Turkmenistan (n = 249)</th>
<th>Uzbekistan (n = 328)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pairwise Random Match Prob.</td>
<td>5.50%</td>
<td>0.13%</td>
<td>0.35%</td>
<td>0.70%</td>
<td>2.30%</td>
<td>0.90%</td>
<td>0.15%</td>
</tr>
<tr>
<td>Polymorphic Sites</td>
<td>106</td>
<td>239</td>
<td>206</td>
<td>136</td>
<td>154</td>
<td>187</td>
<td>266</td>
</tr>
<tr>
<td>Haplotypes</td>
<td>46 (14)</td>
<td>223 (28)</td>
<td>184 (44)</td>
<td>121 (15)</td>
<td>102 (38)</td>
<td>136 (51)</td>
<td>279 (30)</td>
</tr>
<tr>
<td>Mean Pairwise Differences</td>
<td>11.3</td>
<td>12.4</td>
<td>11.9</td>
<td>9.3</td>
<td>13.0</td>
<td>11.4</td>
<td>12.3</td>
</tr>
<tr>
<td>Genetic Diversity</td>
<td>0.9460</td>
<td>0.9990</td>
<td>0.9970</td>
<td>0.9932</td>
<td>0.9826</td>
<td>0.9916</td>
<td>0.9990</td>
</tr>
</tbody>
</table>

Population statistics for seven Central Asian Populations. Random match probabilities were generated empirically. Polymorphic sites do not include C insertions at 16193, 309, or 573. Numbers in parentheses indicate the subset of total haplotypes shared among individuals.
Haplogroup Distributions Among Central Asian Populations
### Genetic Differentiation between Central Asian Populations

<table>
<thead>
<tr>
<th></th>
<th>Afghanistan</th>
<th>Kyrgyzstan</th>
<th>Kazakhstan</th>
<th>Turkmenistan</th>
<th>Russia</th>
<th>Uzbekistan</th>
<th>Tajikistan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>0.067*</td>
<td>0.052*</td>
<td>0.044*</td>
<td>0.058*</td>
<td>0.035*</td>
<td>0.046*</td>
<td></td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td>0.006*</td>
<td>0.015*</td>
<td>0.063*</td>
<td>0.005*</td>
<td>0.009*</td>
<td>0.024*</td>
<td></td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>0.009*</td>
<td>0.015*</td>
<td>0.039*</td>
<td>0.003*</td>
<td>0.009*</td>
<td>0.023*</td>
<td></td>
</tr>
<tr>
<td>Turkmenistan</td>
<td></td>
<td></td>
<td></td>
<td>0.040*</td>
<td></td>
<td></td>
<td>0.056*</td>
</tr>
<tr>
<td>Russia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.013*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uzbekistan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tajikistan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Values are significant at the 0.05 level

Genetic differentiation between any two Central Asian populations comprised between 0.6% (Kazakhstan and Kyrgyzstan) and 8.4% (Kyrgyzstan and Russia) of the total genetic variation among the respective population pairs.

The large genetic distance estimated for Kyrgyzstan and Russia can largely be explained by the disparity in representation of western Eurasian and eastern Eurasian/South Asian lineages between the two populations.

All values are still significant even after application of the Bonferroni correction.
Observations of Each Central Asian Population’s Most Common Haplotype in the Other Central Asian Samples

<table>
<thead>
<tr>
<th></th>
<th>Afghanistan (n = 98)</th>
<th>Kazakhstan (n = 256)</th>
<th>Kyrgyzstan (n = 249)</th>
<th>Russia (n = 151)</th>
<th>Tajikistan (n = 244)</th>
<th>Turkmenistan (n = 249)</th>
<th>Uzbekistan (n = 338)</th>
<th>Total # in Pooled Pop.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>14 (15.2%)*</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14 (0.9%)</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>0</td>
<td>4 (1.9%)*</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4 (0.3%)</td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td>0</td>
<td>1</td>
<td>5 (2.4%)*</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6 (0.4%)</td>
</tr>
<tr>
<td>Russia</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>10 (7.2%)*</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>26 (1.7%)</td>
</tr>
<tr>
<td>Tajikistan</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>16 (6.9%)*</td>
<td>0</td>
<td>0</td>
<td>17 (1.1%)</td>
</tr>
<tr>
<td>Turkmenistan</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>11 (4.8%)*</td>
<td>0</td>
<td>12 (0.8%)</td>
</tr>
<tr>
<td>Uzbekistan</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5 (1.8%)*</td>
<td>5 (0.4%)</td>
</tr>
</tbody>
</table>

Population specific haplotype frequencies that are statistically different from the pooled population frequency are denoted by asterisks. In all cases, the p-value < 0.01

- No two populations share the same most common haplotype.
- The most common haplotype in each of the populations was rarely seen in other populations
- *In all cases, the use of a pooled Central Asian population underestimated the frequency of each individual population’s most common haplotype.*
Central Asian Populations

- Sub-populations of Uzbekistan did not exhibit a high degree of population substructure
  - Uzbekistani sub-populations can likely be pooled together for forensic purposes

- The ethnic subpopulations of Uzbekistan *did* exhibit significant substructure
  - mtDNA frequency estimates would likely be most conservative if the populations were considered separately
Forensically relevant mtDNA issues addressed with the Central Asian dataset

- **mtDNA data quality**
  - A highly redundant laboratory and analysis strategy, as well as post-sequencing phylogenetic tools, will help to improve the quality of mtDNA sequences

- **Nomenclature**
  - The interpretation of unusual length variation can affect database searches and should be carefully evaluated

- **Heteroplasmy**
  - The general incidence of point heteroplasmy among the Central Asian dataset is consistent with what we’re observing in our large scale databasing effort and is higher than previous reports

- **Reference population databases**
  - We are increasing the size and quality of global mtDNA data for the forensic community, with specific emphasis on poorly characterized populations and local mtDNA variation at the local level
Global databasing effort is ongoing

If you are interested in participating:

- Collaborative effort between Labs, with AFDIL funding and conducting the control region sequencing

- Please share samples if:
  - They are anonymous, non-related, and collected with correct geographic data.

- We will make data available to all: via Genbank, SWGDAM and EMPOP
Acknowledgements

- **ICMP**: Thomas Parsons

- **AFDIL**: Jessica Saunier, Jennifer O’Callaghan, Rebecca Just, Mike Coble

- **Uzbekistan**: Abror Ikramov, Abdurakhmon Nuritdinov, Rustam Mukhamedov

- **EMPOP**: Walther Parson, Anita Brandstätter