Reference Databases for mtDNA Casework: Examples from Central Asia

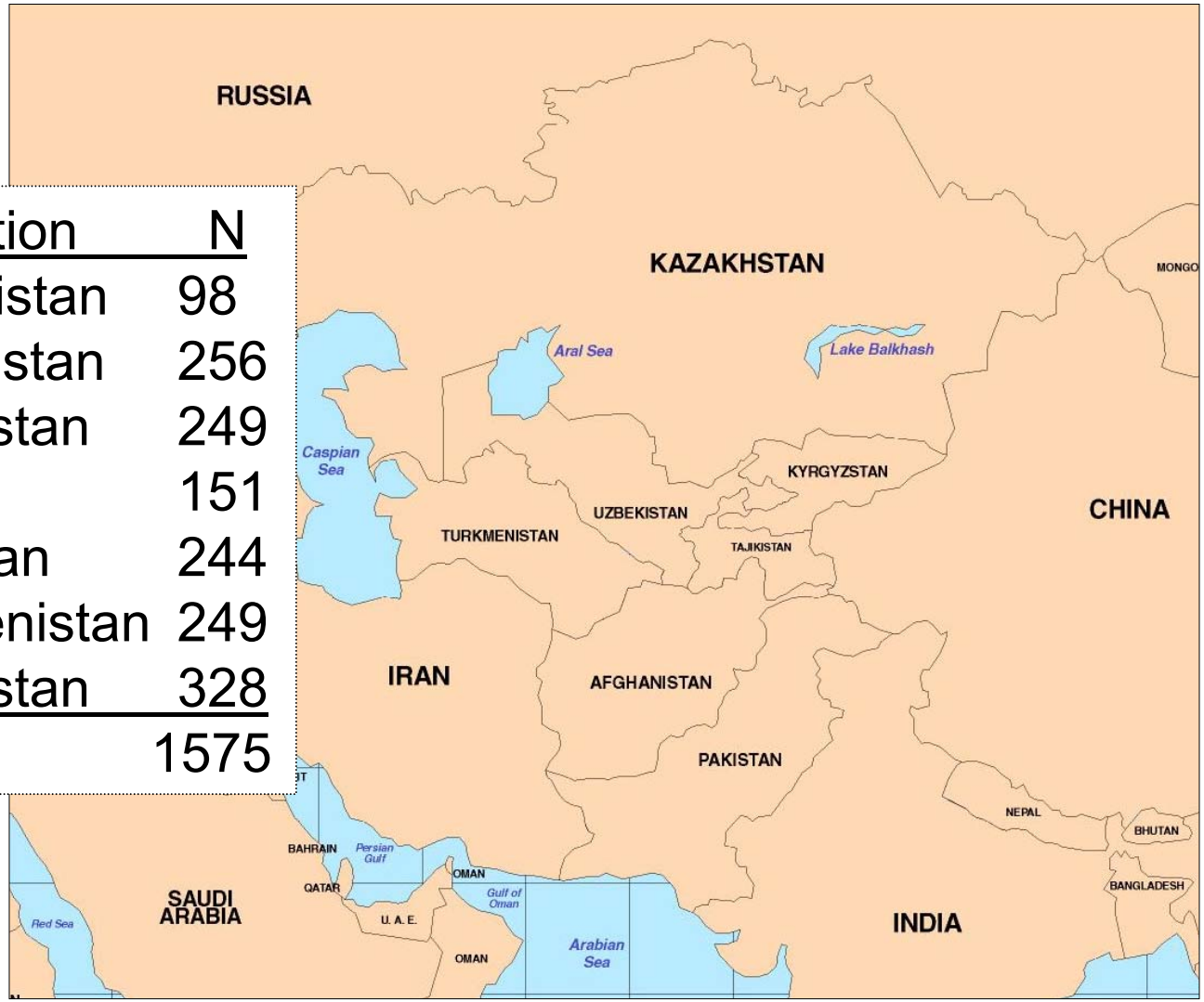
Jodi A. Irwin

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

Population	N
Afghanistan	98
Kazakhstan	256
Kyrgyzstan	249
Russia	151
Tajikistan	244
Turkmenistan	249
Uzbekistan	328
Total	1575



Turkmenistan Tajikistan Kazakhstan

Afghanistan Russia Kyrgyzstan

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

Carracedo A. et al. (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing Forensic Science International 110:79-85

Tully G. et al. (2001) Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. Forensic Science International 124:83-91

mtDNA Control Region Databases

- 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 θ.”

- 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

Population/Region	Sub-Population/ Region	Total # Samples
Afghanistan		98
Bahrain		218
China	Hong Kong	377
Cyprus		91
UAE		191
Egypt		278
Greece		319
Indonesia	Sulawesi	279
Hungary	Budapest Caucasian Baranya Roma	215 211
Iraq		189
Jordan		210
Kazakhstan		256
Kyrgyzstan		249
Kenya	Nairobi	103
Lebanon		198
Pakistan		433
Russia		151
Tajikistan		244
Turkmenistan		249
Uzbekistan		328
Vietnam		187
TOTAL		**5074**

Global populations databased since late 2004

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*

Int J Legal Med (2001) 115 : 64–69

ORIGINAL ARTICLE

H.-J. Bandelt · P. Lahermo · M. Richards · V. Macaulay

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

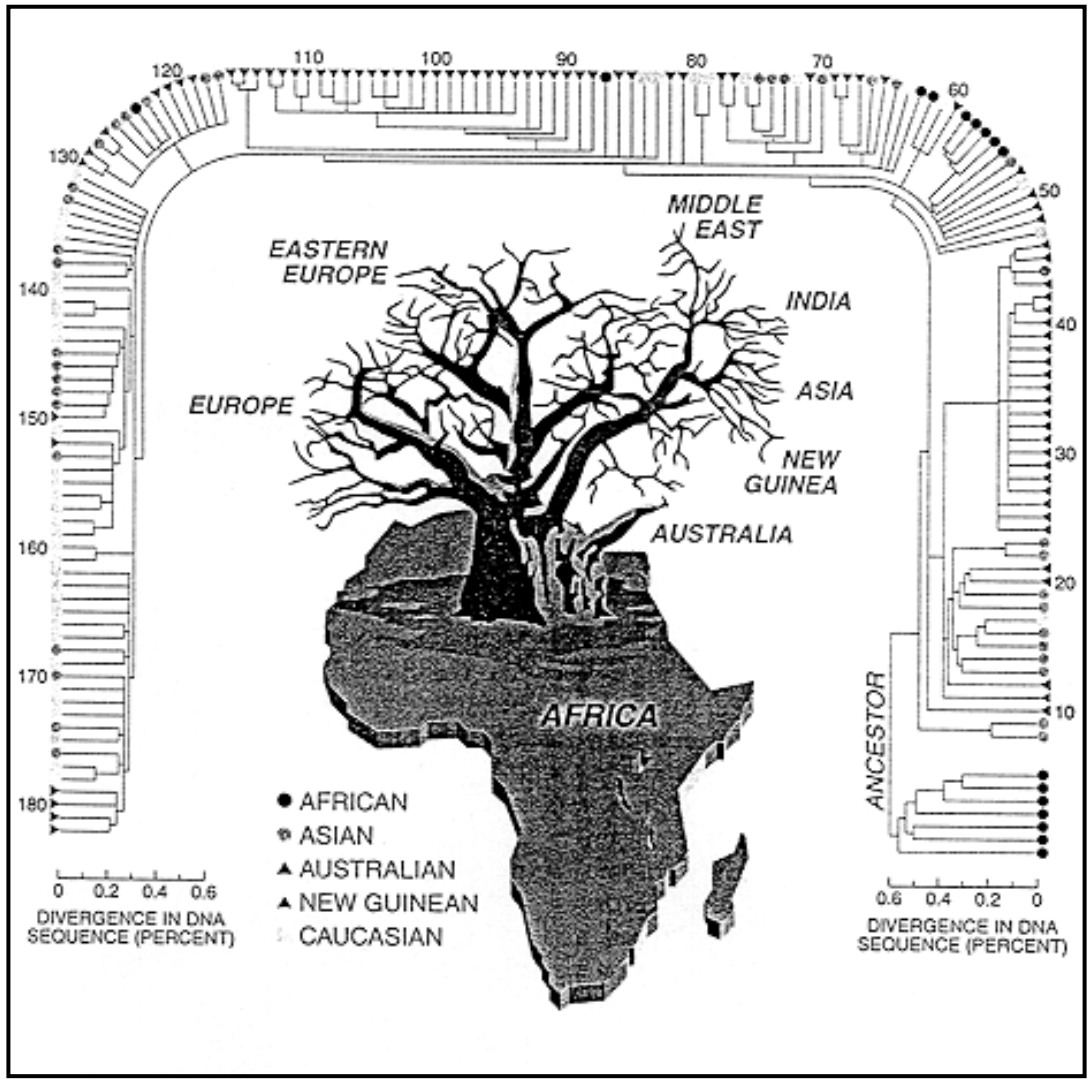
"more than half of the mtDNA sequencing studies ever published contain obvious errors…"

Annals of Human Genetics 2003 67:2-4

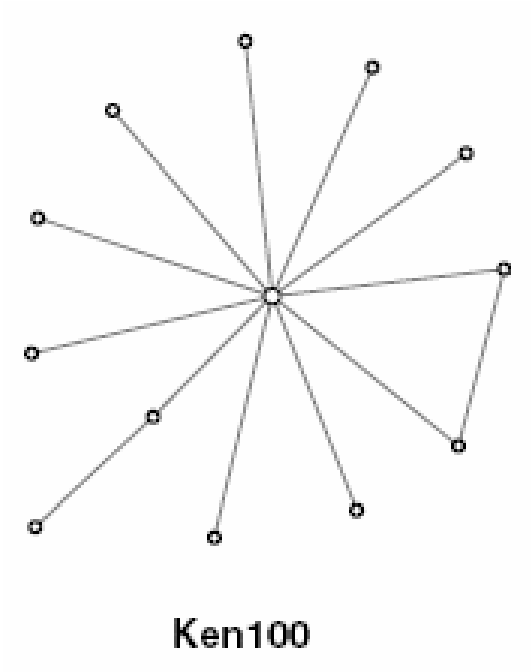
How are errors detected *a posteriori*?

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

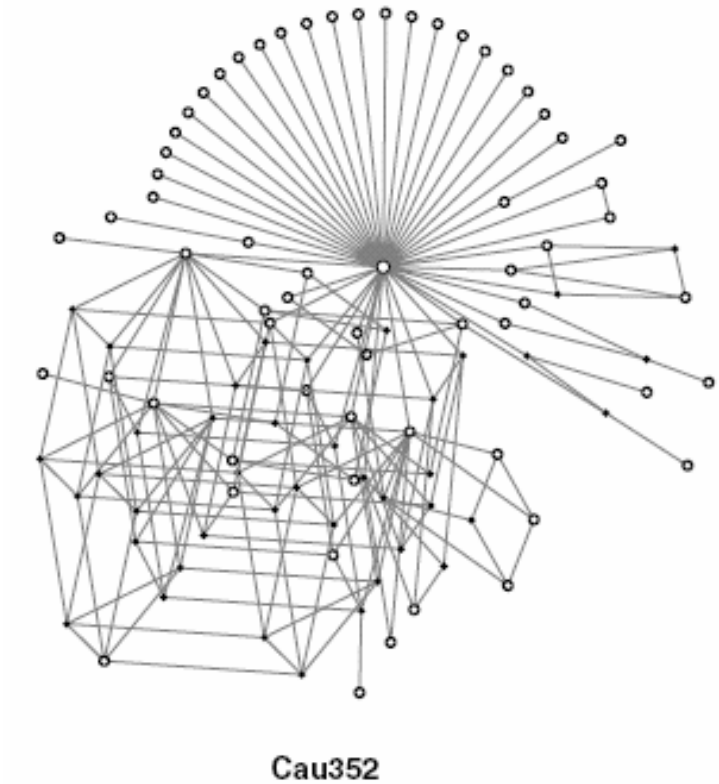
Highlights polymorphisms that are either rare or incompatible with the mtDNA phylogeny



Reduced Network with Filtering



Good data



Poor data

Bandelt HJ, Dür A (2006) Translating DNA data tables into quasi-median networks for parsimony analysis and error detection. Molec Phyl Evol doi:10.1016/j.ympev.2006.07.013

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

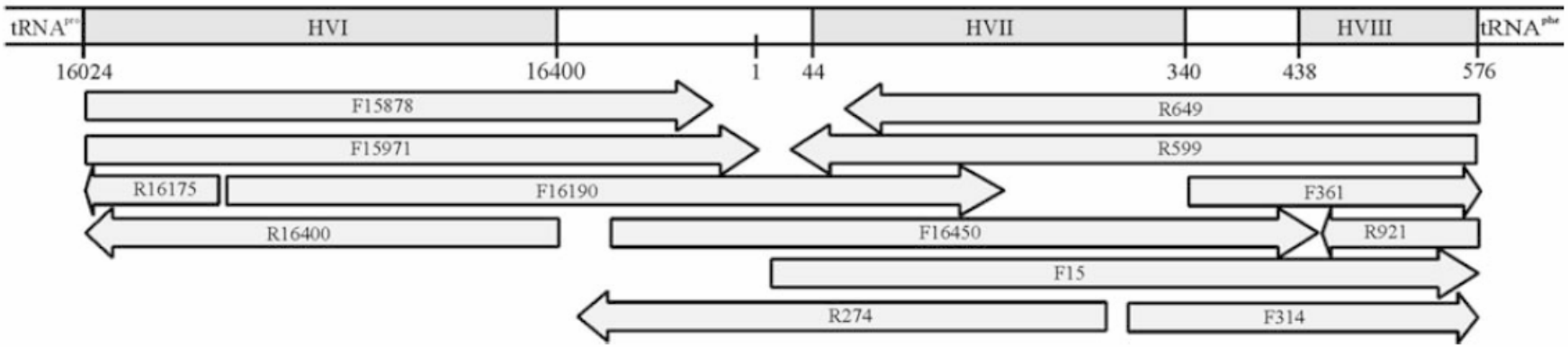
Int J Legal Med (2004) 118: 294–306
DOI 10.1007/s00414-004-0466-z

ORIGINAL ARTICLE

Anita Brandstätter · Christine T. Peterson ·
Jodi A. Irwin · Solomon Mpoke · Davy K. Koech ·
Walther Parson · Thomas J. Parsons

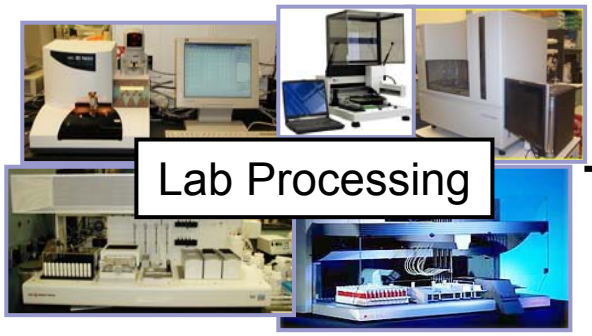
Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database

Received: 16 February 2004 / Accepted: 4 June 2004 / Published online: 10 July 2004

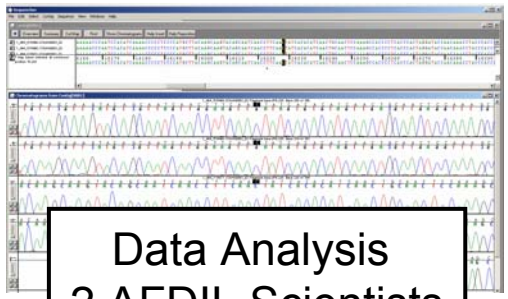


Highly redundant sequencing strategy avoids “phantom mutations”

General Overview



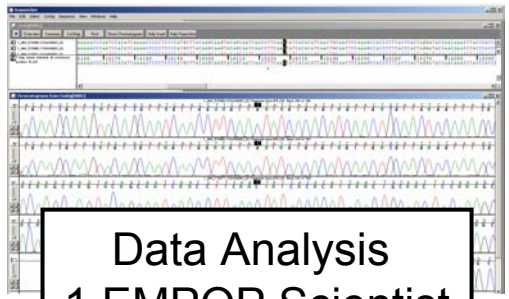
Lab Processing



Data Analysis
2 AFDIL Scientists

Local DB entry
- electronic data transfer
- data confirmation
- phantom mutation x-check

Data Transfer to EMPOP
- raw sequence data
- 'finalized' haplotypes stored in local DB



Data Analysis
1 EMPOP Scientist

Local DB entry
- electronic data transfer
- 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

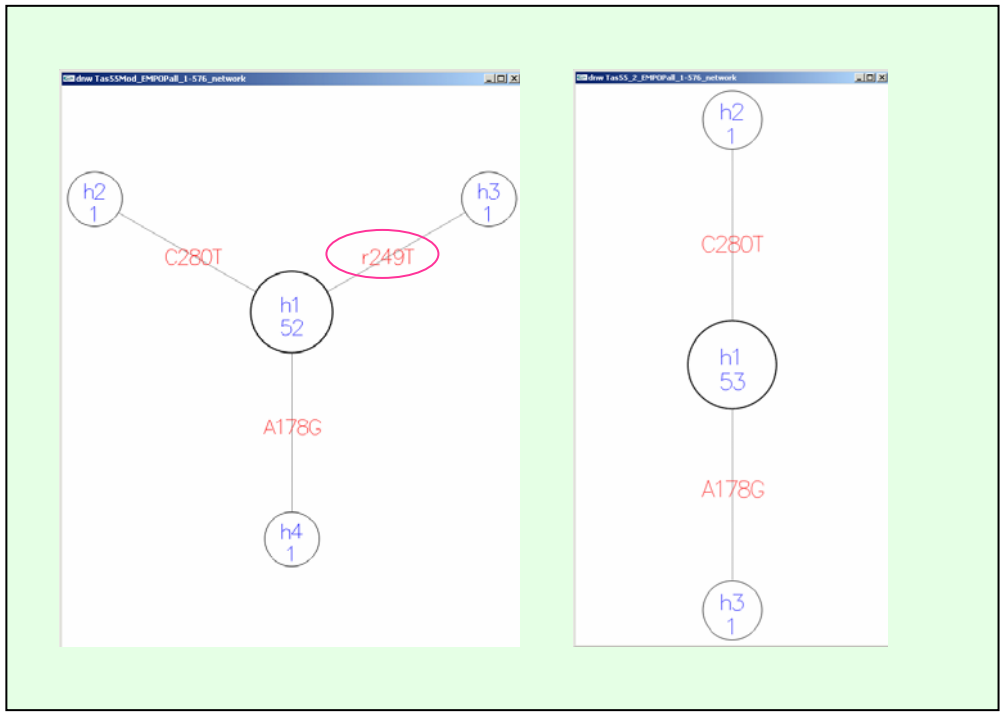
Table 2. Positions hit by phantom mutations at least three times in the experimental study (average QVs in brackets)

Frequency	Position (average QV)
15	527 (14.5), 16197 (14.9)
14	320 (14.8)
11	441 (15.1), 481 (11.2)
10	16036 (18.3), 16242 (14), 16271 (13.8), 16303 (16.2), 16366 (24.3)
9	(16.8), 16384 (18.7)
8	(21.5)
7	), 513 (20.9), 16095 (20.9), 16155 (19.1), 16248 (14.4), 16270
6	), 544 (18.2), 16024 (15.3), 16274 (29.3), 16380 (21.5)
5	08 (19.6), 531 (5), 539 (7), 545 (14.4), 16046 (17), 16085 (18), 278 (24.4), 16346 (15), 16351 (23.2), 16358 (17.2), 16487 (16.2)
4	(24.3), 345 (17.3), 404 (16), 420 (15), 431 (14.5), 462 (27.8), 483 038 (17.5), 16064 (18.3), 16213 (17.8), 16221 (36.5), 16228 (15), 16311 (18.8), 16313 (16.5), 16355 (22), 16360 (17.8), 16361 6.3), 16460 (22.5)
3	4 (16.7), 106 (13.7), 137 (18.7), 149 (23), 165 (27), 186 (34), 215 (17.7), 332 (14.3), 371 (15), 388 (14), 409 (23), 414 (22), 445 (21.7), 518 (25.3), 525 (16), 563 (17.3), 564 (8.3), 16029 (24.3), 16145 (15.3), 16154 (13), 16204 (15), 16208 (15.3), 16212 6.3), 16368 (19.7), 16370 (13.7), 16381 (21.7), 16386 (15.3), 16427 (24), 16446 (11), 16477 (15), 16482 (33.3), 16540 (27.3)

Brandstätter et al. Electrophoresis 2005

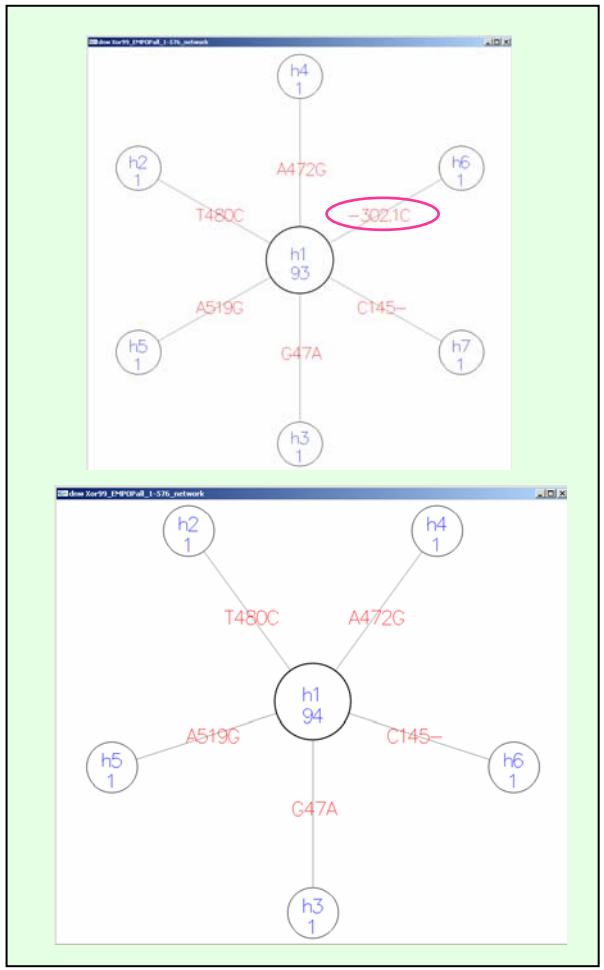
- Had these errors made it beyond the data cross-check and been used in the phylogenetic checks, this is what we would have seen….

Alignment/transcription errors



249T

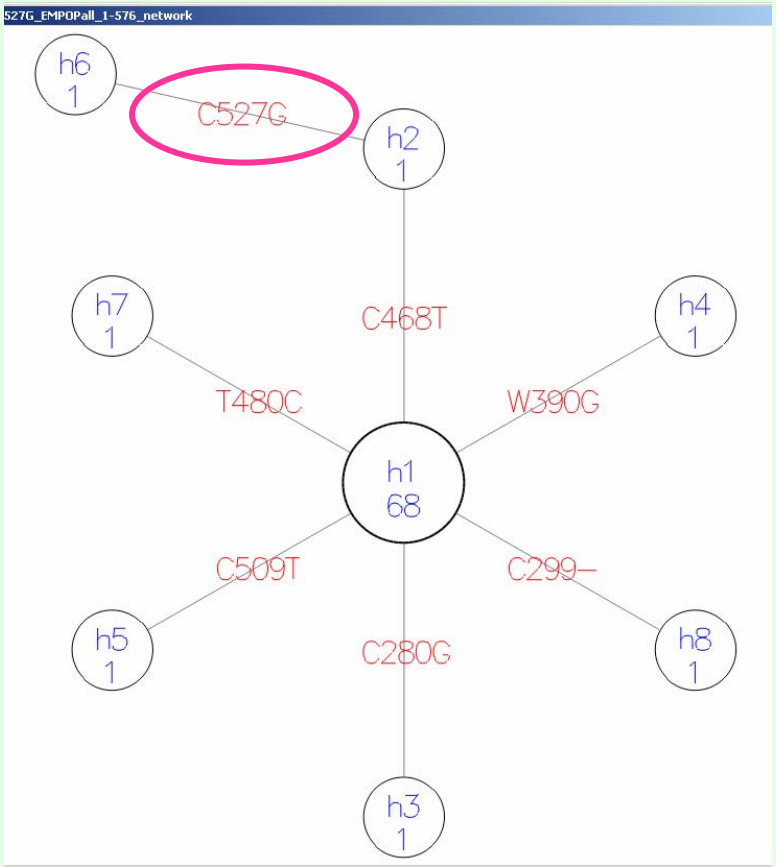
Uzbek/Tashkent network



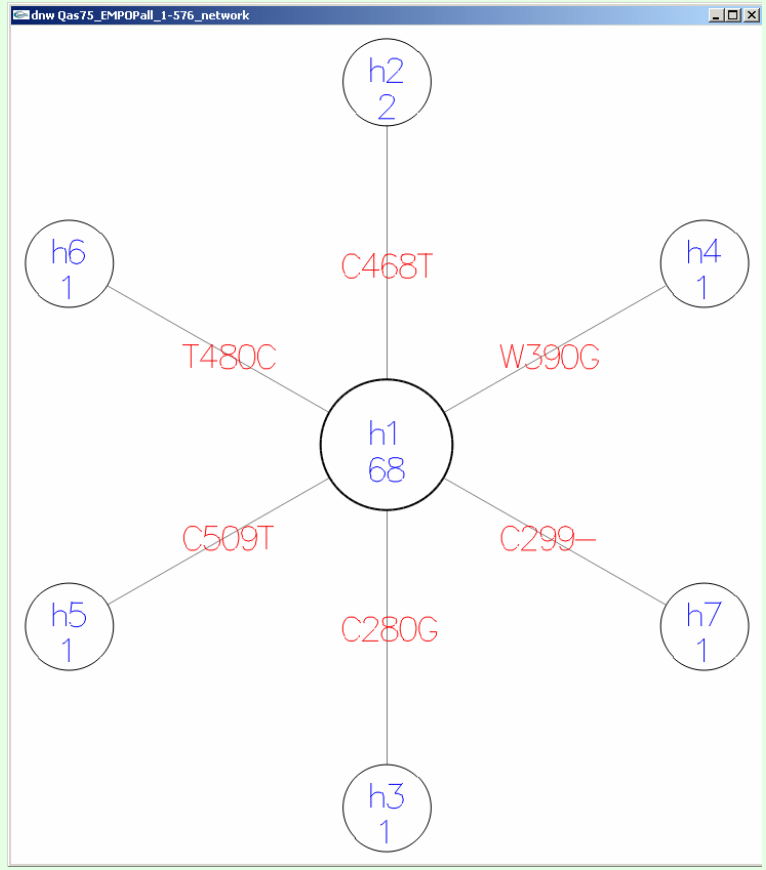
302.1C

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

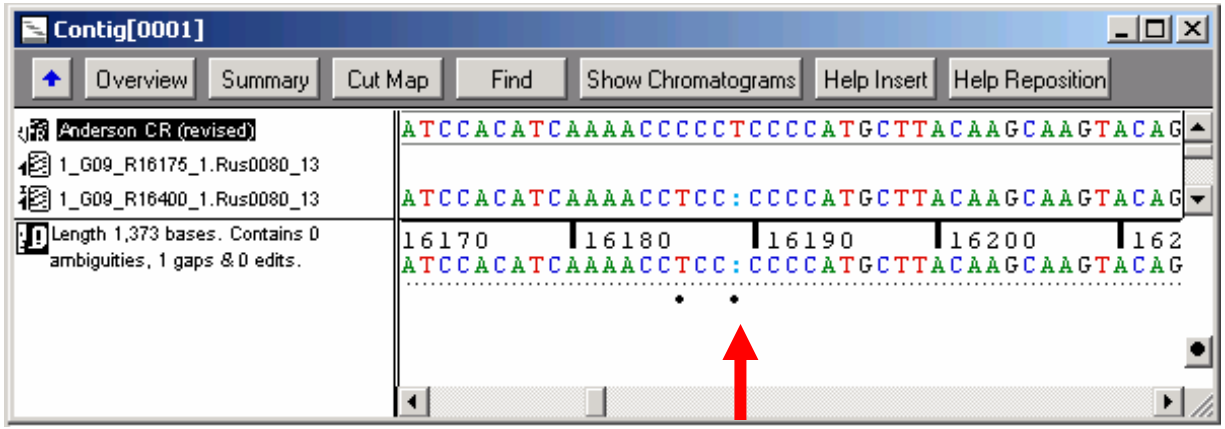
5 – Nomenclature differences – *we'll re-visit*

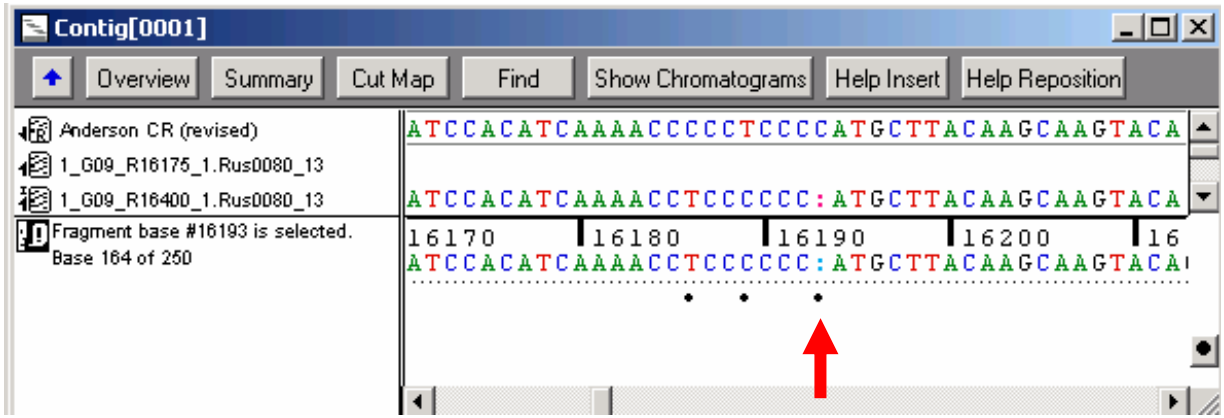
3 – Mistakes *pre-phantom mutation screen*
pre-phylogenetic check

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.

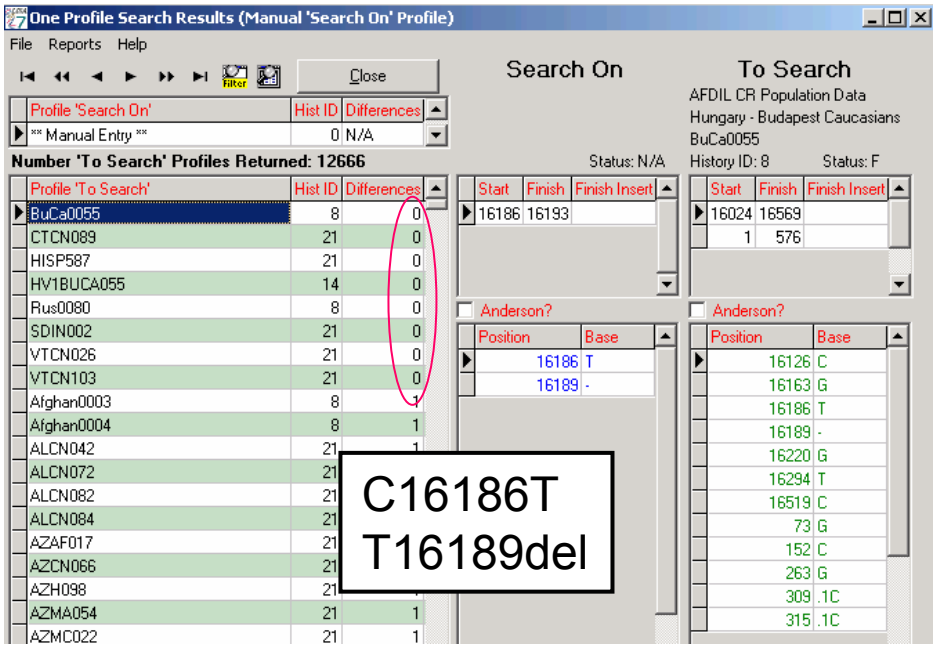
Nomenclature problems

- 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*)

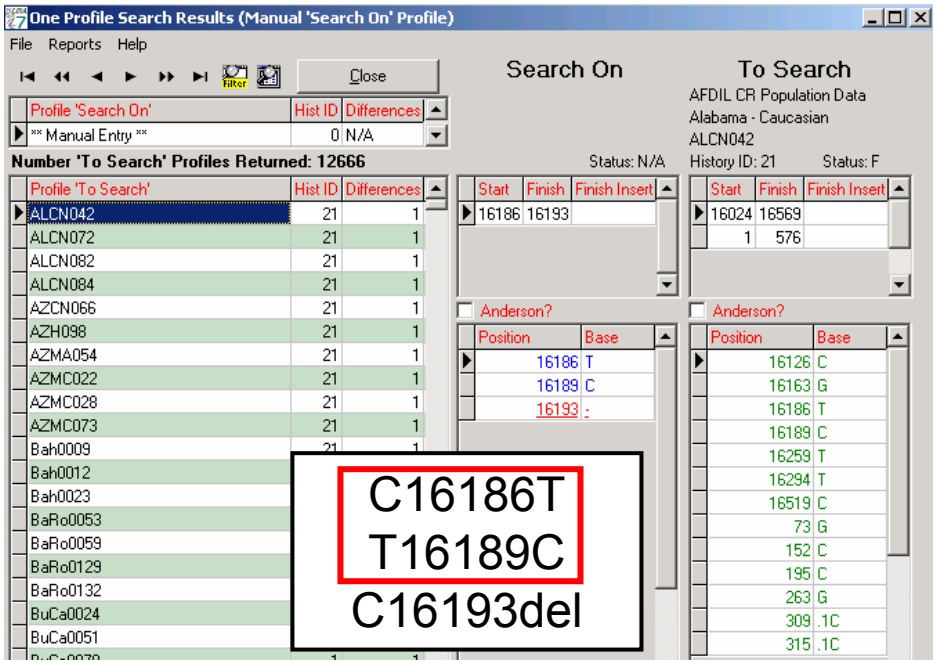
Nomenclature problems

- 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



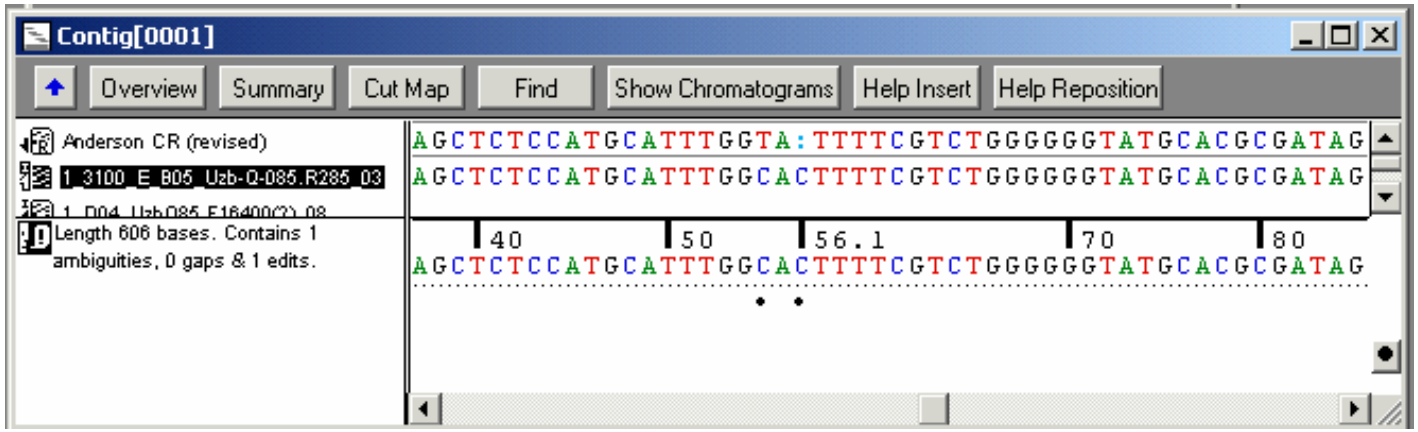
0 consistent haplotypes in database

Variants associated with hg T1a

Nomenclature Issues

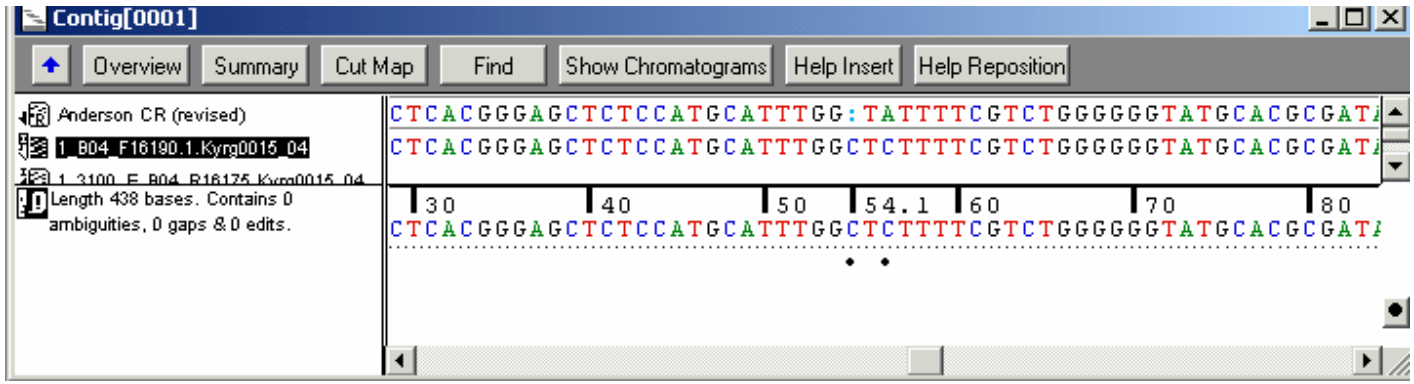
- Recent suggestions by Parson and Bandelt suggest using phylogenetic information to guide indel placement.
- In our experience, these guidelines resolve the vast majority of cases. However, an intimate knowledge of mtDNA evolution and the mtDNA literature is required.
- No matter which guidelines are 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…



Uzb-Q-085

T55C
56.1C
263G
315.1C



Kyrg-015

54.1C
A56C
215G
263G
315.1C

- If we remove rCRS from the alignment and align the two samples to eachother:

T55C; 56.1C

54.1C; A56C

- If we look at the entire CR haplotype and not just that variable region
  - Both of these samples are on a particular haplotypic background

Uzb-Q-085
T55C
56.1C
263G
315.1C

Kyrg-015
54.1C
A56C
215G
263G
315.1C

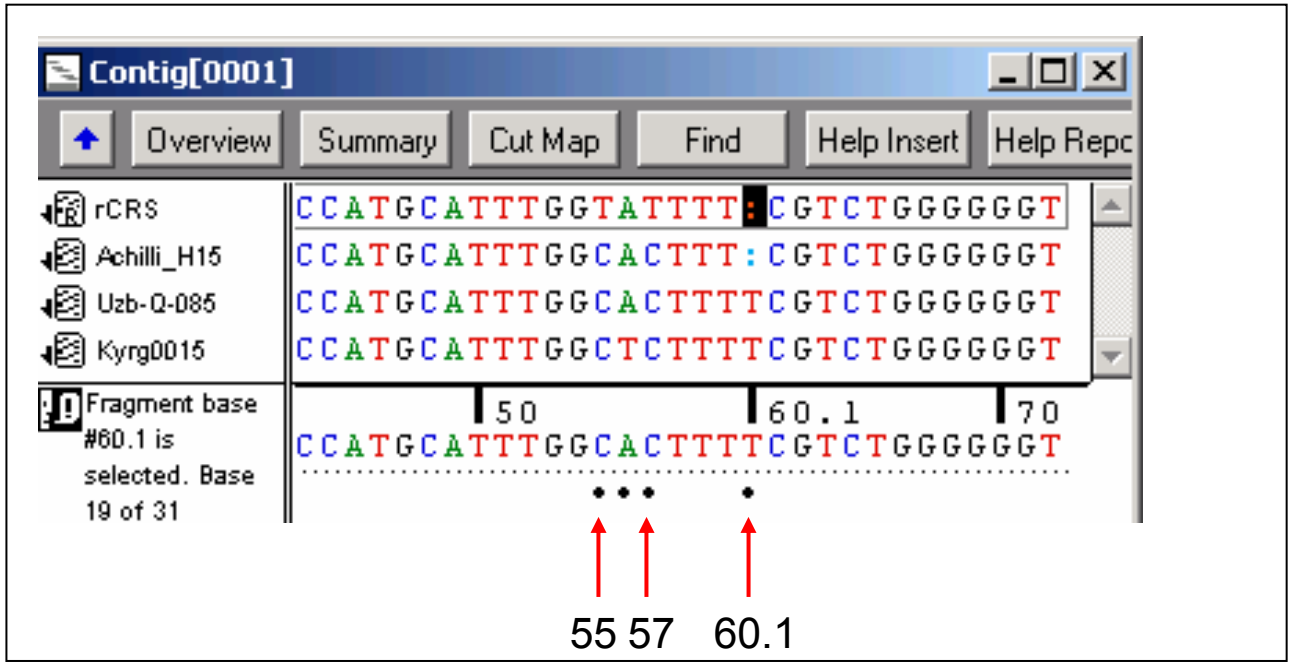
And finally, a review of the literature reveals…

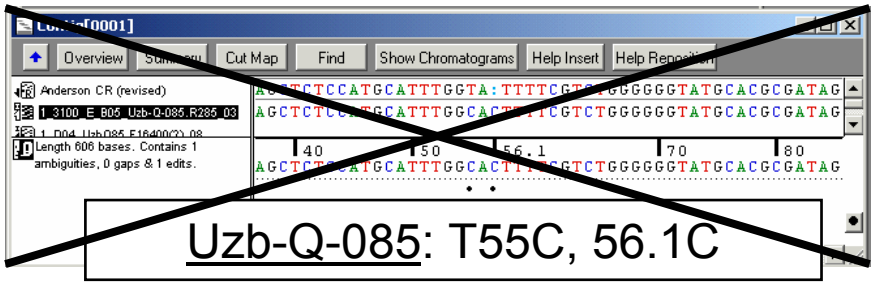
- A similar haplotype described by Achilli et al. (*AJHG 2004*) with no insertion in the region between 54-60

Uzb-Q-085
T55C
56.1C
263G
315.1C

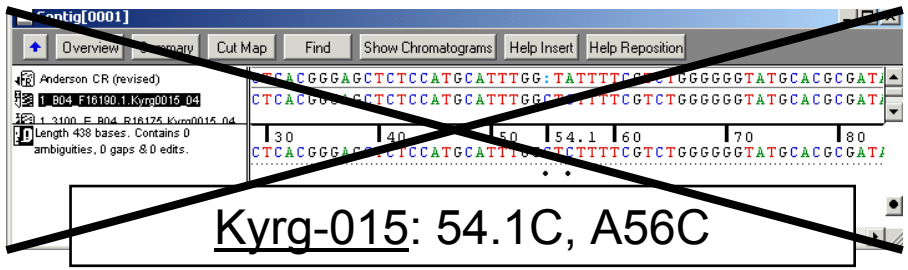
Achilli H15
T55C
T57C
263G
315.1C

Kyrg-015
54.1C
A56C
215G
263G
315.1C

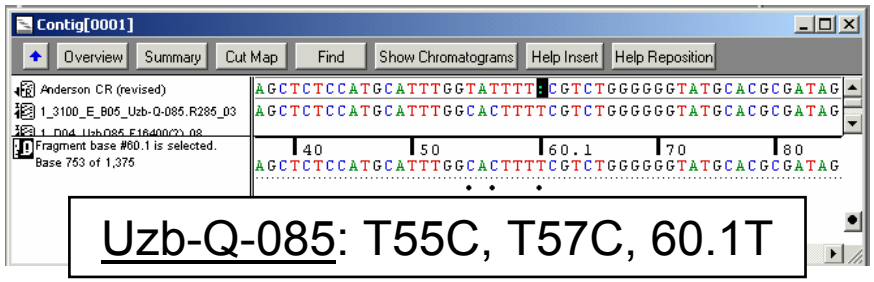




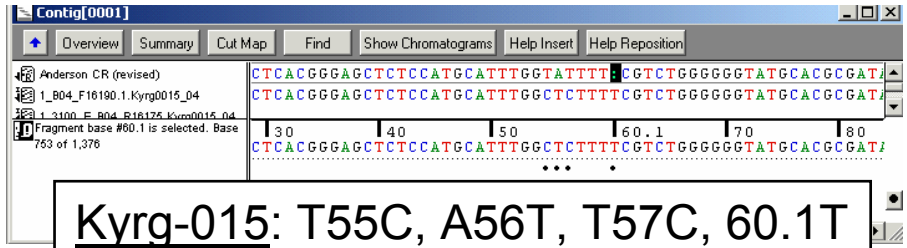
Uzb-Q-085: T55C, 56.1C



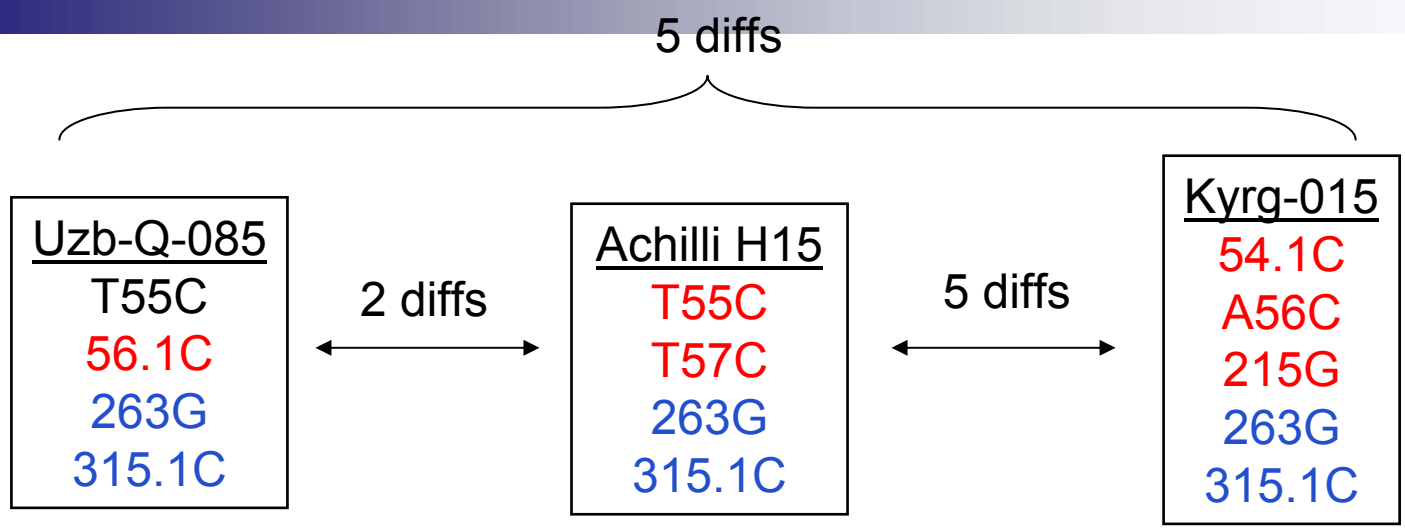
Kyrg-015: 54.1C, A56C



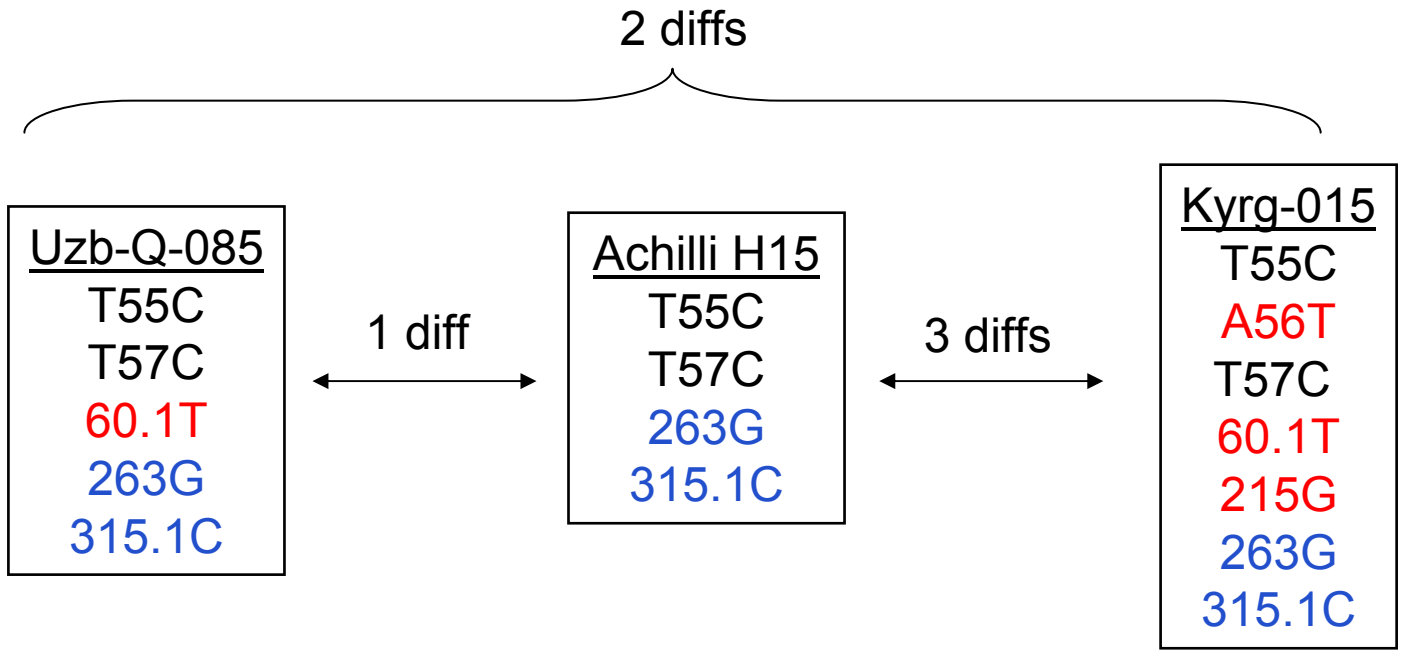
Uzb-Q-085: T55C, T57C, 60.1T



Kyrg-015: T55C, A56T, T57C, 60.1T



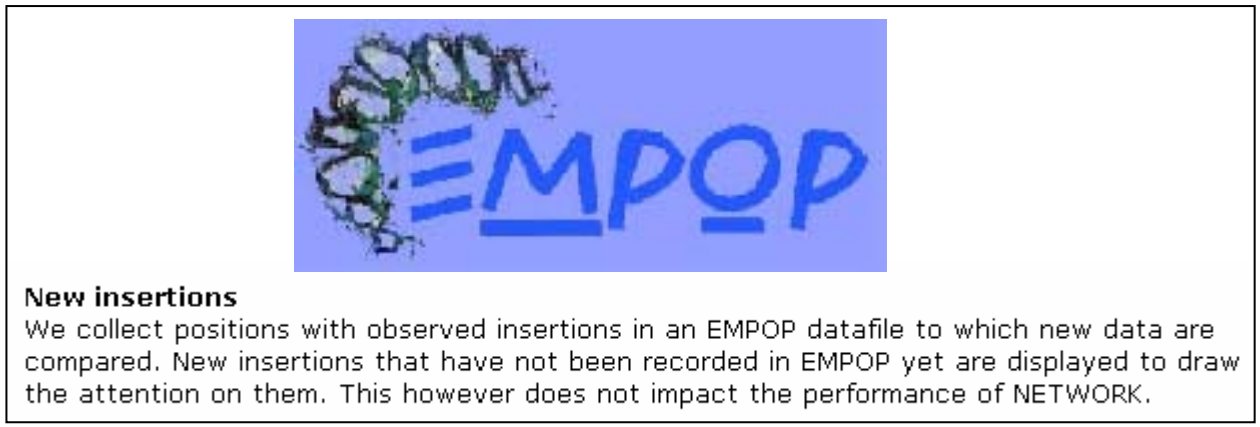
Following a global alignment that considers similar haplotypes…



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

Turkmenistan Tajikistan Kazakhstan

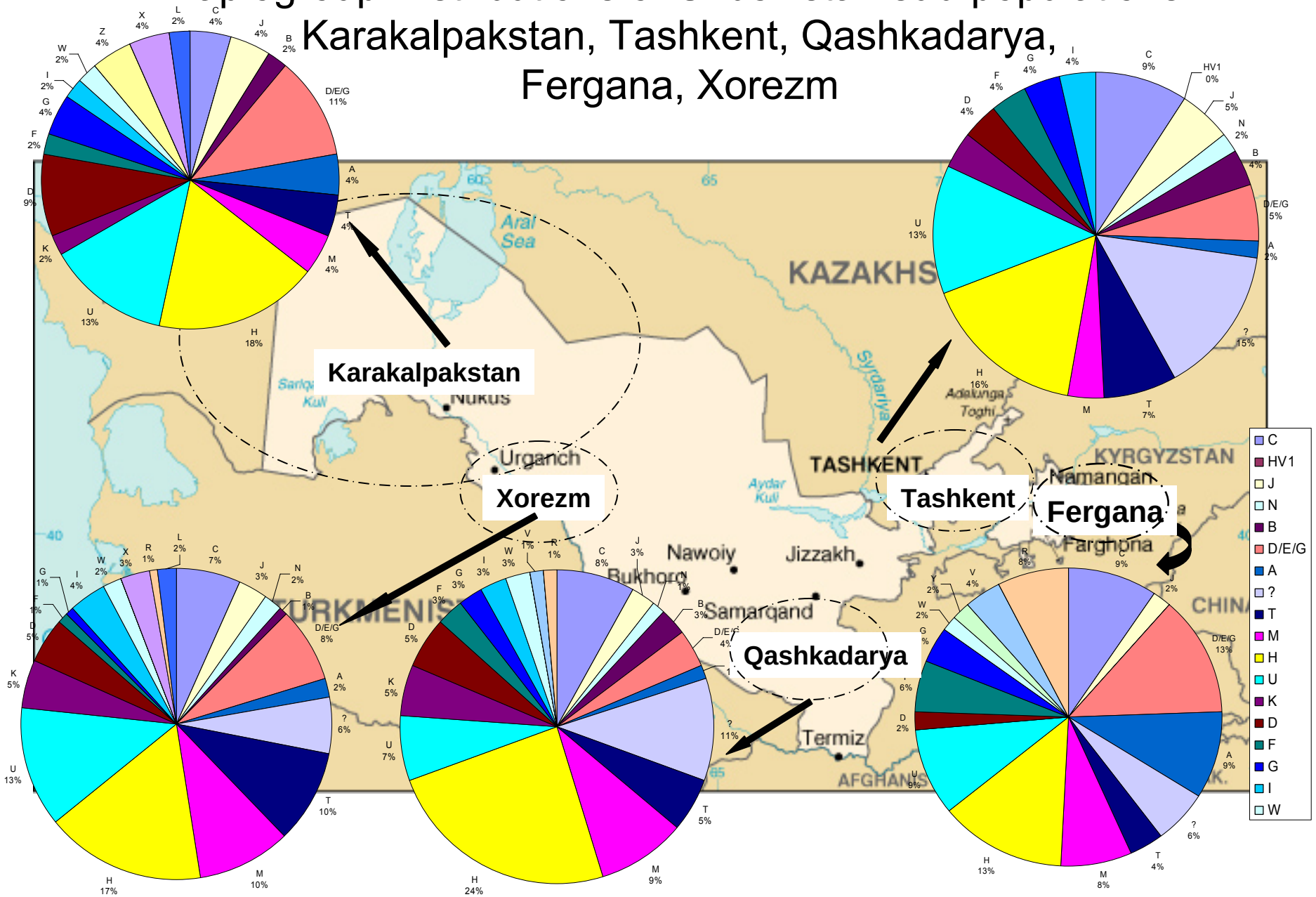
Afghanistan Russia Kyrgyzstan

Diversity Indices for Sub-Populations of Uzbekistan

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

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

	Fergana	Karakalpakstan	Qashkadarya	Tashkent	Xorezm
Fergana		-0.00324	0.00779	0.00937*	0.00816*
Karakalpakstan			0.00201	-0.00178	-0.00075
Qashkadarya				0.00919*	0.0004
Tashkent					0.00704*

* Values are significant at the 0.05 level.

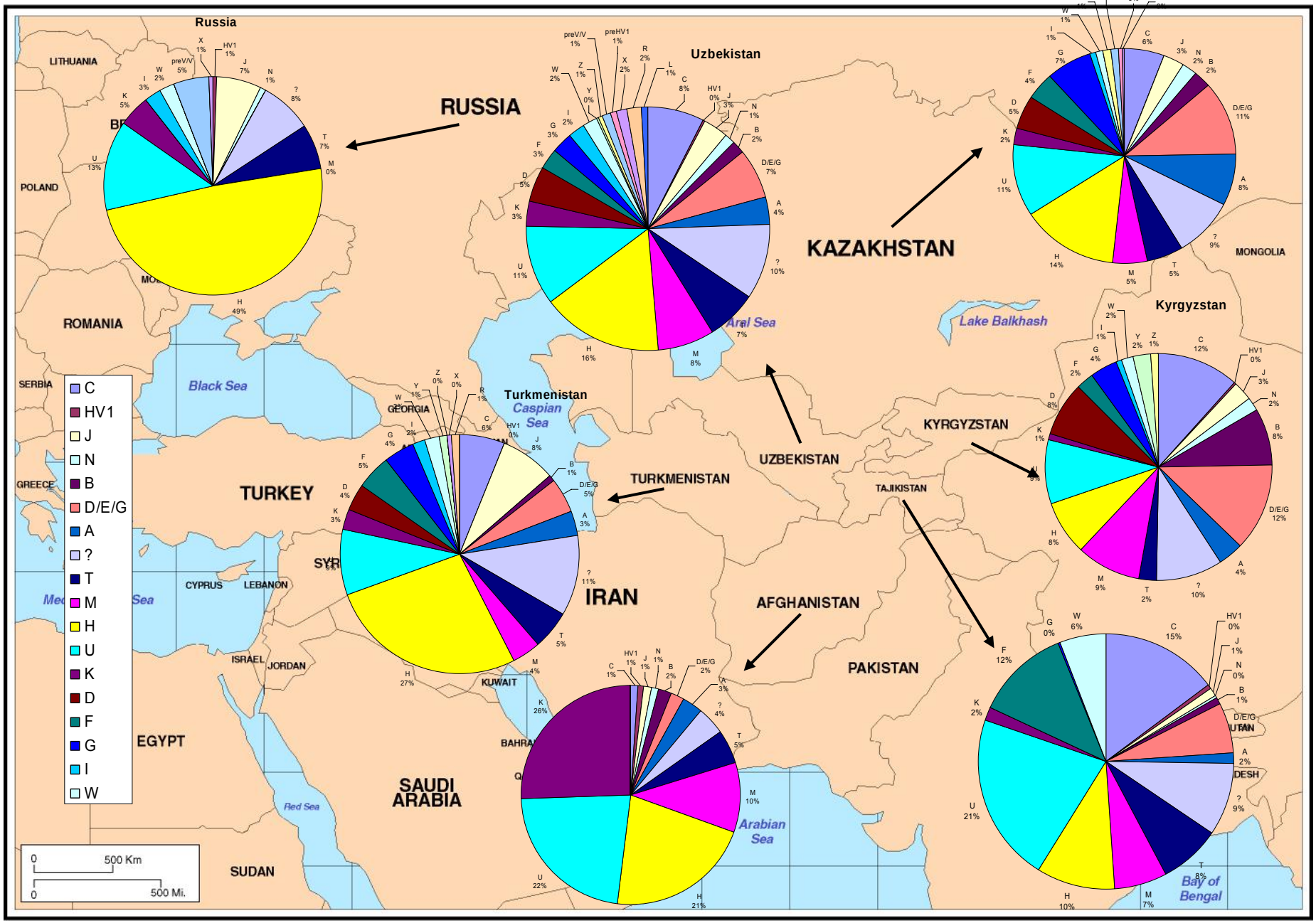
- While some of these values are statistically significant, but the magnitude of the inter-population differences is marginal in each case
- Furthermore, if the Bonferroni correction is applied, the differences are no longer statistically significant

Diversity Indices for Seven Central Asian Populations

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

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



Φ_{st} Values based on Haplotype Data In Various Central Asian Populations

	Afghanistan	Kyrgyzstan	Kazakhstan	Turkmenistan	Russia	Uzbekistan	Tajikistan
Afghanistan		0.067*	0.052*	0.044*	0.058*	0.035*	0.046*
Kyrgyzstan			0.006*	0.015*	0.084*	0.009*	0.024*
Kazakhstan				0.009*	0.063*	0.005*	0.023*
Turkmenistan					0.039*	0.003*	0.018*
Russia						0.040*	0.056*
Uzbekistan							0.013*

* 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

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

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