Y-Chromosome and Mitochondrial DNA Analysis

mitochondrial DNA

NEAFS 2006 Workshop
Rye Brook, NY
November 1, 2006

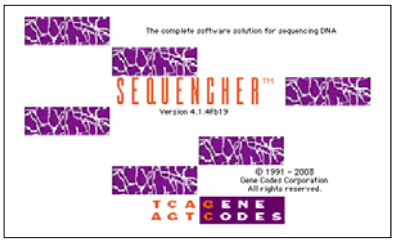
Northeastern Association of Forensic Scientists

Dr. John M. Butler
Dr. Michael D. Coble

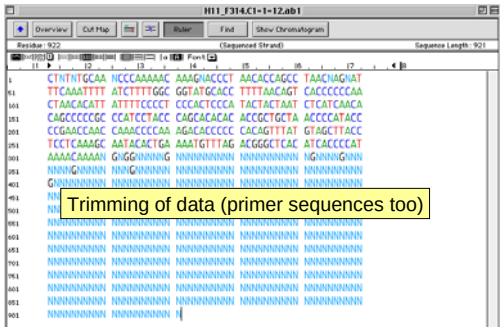
john.butler@nist.gov
Michael.Coble@afip.osd.mil

Interpreting and Reporting mtDNA Results

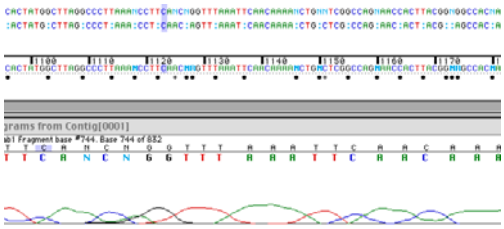
Data Review and Editing



Data Review and Editing

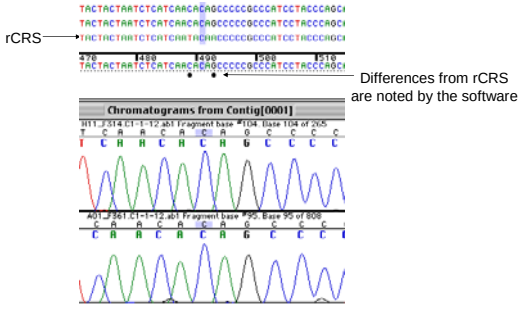


Data Review and Editing

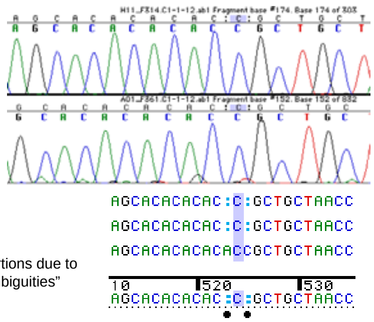


Poor resolution at the end of long fragments

Data Review and Editing

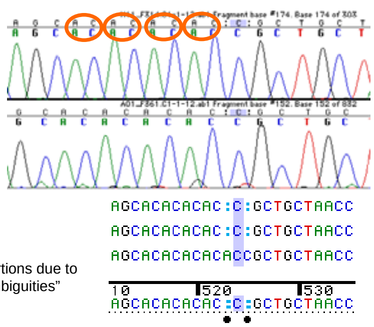


Data Review and Editing



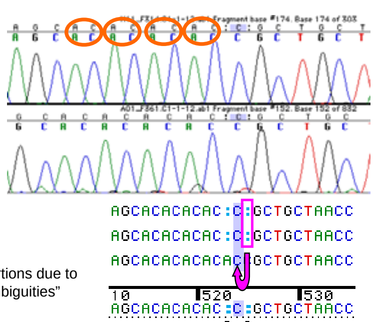
"Insertions due to ambiguities"

Data Review and Editing



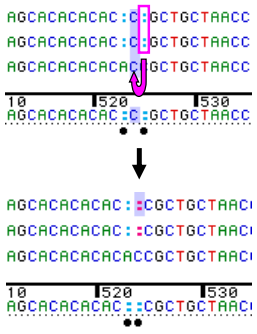
"Insertions due to ambiguities"

Data Review and Editing



"Insertions due to ambiguities"

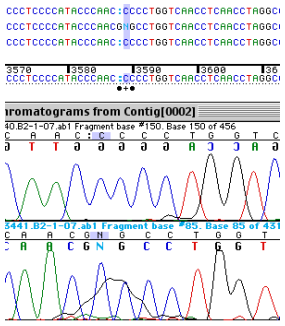
Data Review and Editing



Data Review and Editing

R primer – great data!

Small dye blob ~80bp into F primer creates ambiguities

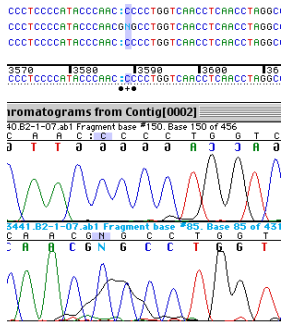


Data Review and Editing

R AAC : CCCCT

F AACGNGCCT

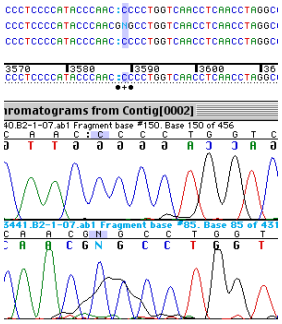
rCRS AAC : CCCCT



Data Review and Editing

R AAC : CCCCT
F AACGNGCCT
rCRS AAC : CCCCT

Delete without affecting the sequence

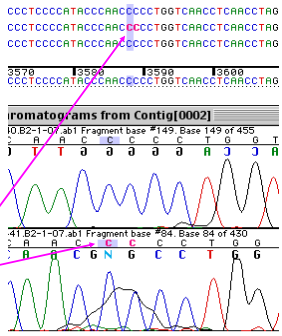


Data Review and Editing

R AACCCCCT
F AACNGCCT
rCRS AACCCCCT

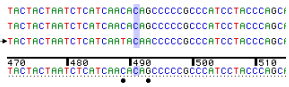
Change to "C"

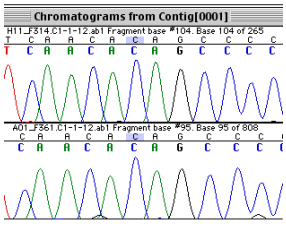
Changes recorded in pink



Reporting Differences from rCRS

rCRS



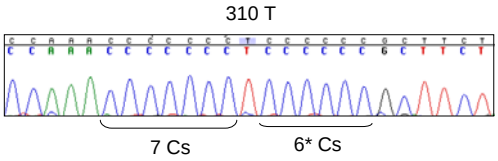


Point Mutations are listed as differences from the rCRS

489 T-C
493 A-G

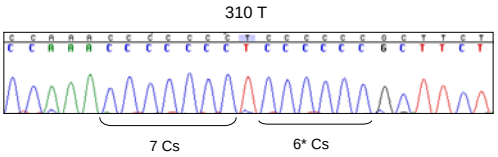
Reporting Differences from rCRS

- Insertions and Deletions (Indels)



*The rCRS has 5 Cs in this region

Reporting Differences from rCRS



Sequence AAACCCCCCCCTCCCCCCGCTT
rCRS AAACCCCCCCCTCCCCCGCTTC
303 310 315

Reporting Differences from rCRS

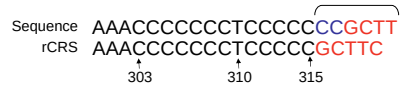
Sequence AAACCCCCCCCTCCCCCCGCTT
rCRS AAACCCCCCCCTCCCCCGCTTC
303 310 315

Sequence AAACCCCCCCCTCCCCCCGCTT
rCRS AAACCCCCCCCTCCCCC: GCTTC
303 310 315

315.1 C

Reporting Differences from rCRS

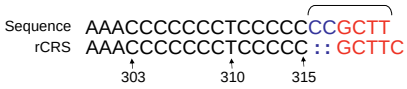
- What if….



2 additional Cs after 315?

Reporting Differences from rCRS

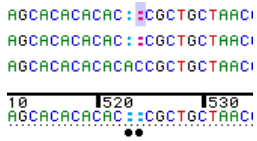
- What if….



315.1 C
315.2 C

Deletions

- Deletions – report the position and bases deleted…



523 A-del
524 C-del

Nomenclature Issues

Wilson, MR et al. (2002) "Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region." ***Forensic Science International*** **129(1): 35-42.**

1. Use the least number of differences
2. Prioritization of indels > transitions > transversions

Transitions	Transversions
A-G	A-T
C-T	A-C
	G-T
	G-C
(Purine-Purine) (Pyrimidine-Pyrimidine)	(Purine-Pyrimidine) (Pyrimidine-Purine)

3. Indels placed at the 3' end with respect to the light strand

Nomenclature Issues

Wilson, MR et al. (2002) "Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region." ***Forensic Science International*** **129(1): 35-42.**

1. Use the least number of differences
2. Prioritization of indels > transitions > transversions

Sample TTTACCCAT
rCRS TTTTGCCCAT
1 10

Sample TTTA:CCCAT 4 T-A
rCRS TTTTGCCCAT 5 G-del

Sample TTT:ACCCAT 4 T-del
rCRS TTTTGCCCAT 5 G-A

3. Indels placed at the 3' end with respect to the light strand

Nomenclature Issues

Wilson, MR et al. (2002) "Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region." ***Forensic Science International*** **129(1): 35-42.**

1. Use the least number of differences
2. Prioritization of indels > transitions > transversions

Sample TTTACCCAT
rCRS TTTTGCCCAT
1 10

Sample TTTA:CCCAT 4 T-A
rCRS TTTTGCCCAT 5 G-del

Sample TTT:ACCCAT 4 T-del
rCRS TTTTGCCCAT 5 A-G

3. Indels placed at the 3' end with respect to the light strand

Nomenclature Issues

Wilson, MR *et al.* (2002) "Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region." *Forensic Science International* 129(1): 35-42.


Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region

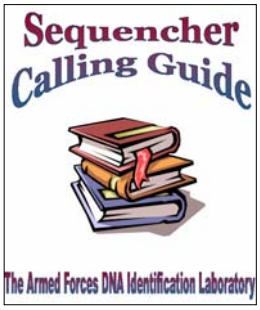

Forensic Science Communications

Research and Technology

Further Discussion of the Consistent Treatment of Length Variants in the Human Mitochondrial DNA Control Region

Wilson, MR *et al.* (2002) *Forensic Science Communications*

Nomenclature Issues



Initiated by the Research Section at AFDIL. Presented at the Promega HID Symposium, 2004.

Kim Sturk will present a MS Access version at the AAFS meeting in San Antonio, TX in Feb. 2007.

Gives examples along with EPGs

POP QUIZ!!!

ATACAACCCCACCCAT
rCRS ATACAACCCCCACCCAT
488 504

ATTGATGTC
rCRS ATTGAATGTC
244 253

ATTGAAATGTC
rCRS ATTGAATGTC
244 253

CATAACAAAATTT
rCRS CATAACAAAAAATTT
280 294

POP QUIZ!!!

TGGCACTTTTCGTCT
rCRS TGGTATTTTCGTCT
52 65

TATCTTTCGT
rCRS TATTTTCGT
55 63

TATTTTTCGTCT
rCRS TATTTTCGTCT
55 65

AAACCCCCCCCCTCCCCCCGCT
rCRS AAACCCCCCCTCCCCCGCT
300 318

Nomenclature Issues

- Consistency is needed – especially for database searches.

513
Lab 01 Sample GCACACACACACACCGCT
rCRS GCACACACACACCGCT

Lab 02 Sample GCACACACACACACCGCT
rCRS GCACACACACACCGCT

Lab 03 Sample GCACACACACACACCGCT
rCRS GCACACACACACCGCT

Nomenclature Issues

- Consistency is needed – especially for database searches.

513
Lab 01 Sample GCACACACACACACCGCT 524.1 A
rCRS GCACACACACAC::CGCT 524.2 C
1 2 3 4 5

Lab 02 Sample GCACACACACACACCGCT 523.1 C
rCRS GCACACACACA::CCGCT 523.2 A
1 2 3 4 5

Lab 03 Sample GCACACACACACACCGCT 514.1 A
rCRS GC::ACACACACACCGCT 514.2 C
1 2 3 4 5

Nomenclature Issues

Lab 01		Lab 02		Lab 03
16519 T-C		16519 T-C		16519 T-C
263 A-G		263 A-G		263 A-G
315.1 C	≠	315.1 C	≠	315.1 C
524.1 A		**523.1 C**		**514.1 A**
524.2 C		**523.2 A**		**514.2 C**

Each lab submits 20 sequences above into the population DB (N=1000)

Lab 04 → Will match the 20 samples submitted by Lab 01

16519 T-C
263 A-G
315.1 C
524.1 A
524.2 C

Apparent Frequency = 20/1000 (0.02)
True Frequency = 60/1000 (0.06)

Underestimation of the true frequency

Interpretational Issues - Heteroplasmy

- Heteroplasmy – the presence of more than one mtDNA type in an individual (Melton 2004).
- Once thought to be rare, heteroplasmy exists (at some level) in all tissues (Melton 2004).
- Especially important in hair analysis (semi-clonal).

Heteroplasmy

- Some interesting papers (forensic focus)…

- Melton, T. (2004) Mitochondrial DNA heteroplasmy. *Forensic Science Reviews* **16:**1-20.
- Calloway *et al.* (2000) The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age. *Am J Hum Genet.* **66(4):**1384-1397.
- Stewart *et al.* (2001) Length variation in HV2 of the human mitochondrial DNA control region. *Journal of Forensic Science* **46(4):**862-870.
- Sekiguchi *et al.* (2003) Inter- and intragenerational transmission of a human mitochondrial DNA heteroplasmy among 13 maternally-related individuals and differences between and within tissues in two family members. *Mitochondrion* **2(6):**401-414.
- Salas *et al.* (2001) Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: a case report. *Int J Legal Med.* **114(3):**186-190.
- Tully, L *et al.* (2000) A sensitive denaturing gradient-Gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region. *Am J Hum Genet.* **67(2):**432-443.

Interpretational Issues - Heteroplasmy

- Two types: Length (most common) and Point Heteroplasmy.

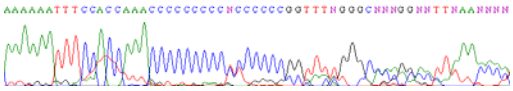
"Out of phase!"

Sequence 1 AAACCCCCCCCCTCCCCCCGCTTC
Sequence 2 AAACCCCCCCCCCTCCCCCCGCTTC
rCRS AAACCCCCCC : : :TCCCCCGCTTC
303 310 315

Sequence 1 has 9 Cs before 310T
Sequence 2 has 10 Cs before 310T

HV2 Length Heteroplasmy

Dr. Cintia Friedman, São Paulo, Brasil



"Out of phase!"

Sequence 1 AAACCCCCCCCCTCCCCCCGCTTC
Sequence 2 AAACCCCCCCCCCTCCCCCCGCTTC

Double coverage is important to determine sequences surrounding HV1, HV2, HV3 C-stretches.

Point Heteroplasmy

"Y" Pyrimidine

"Hotspot" for heteroplasmy

16093 (C/T)

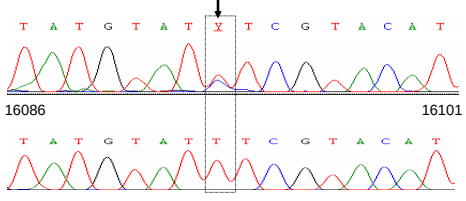


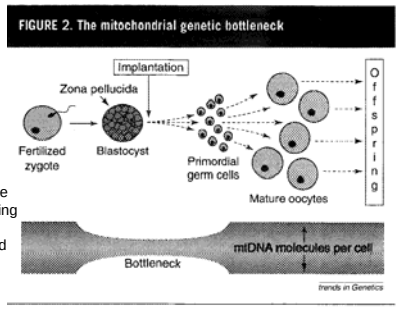
Figure 10.9, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

Origination of Heteroplasmy

Ovum – 100K mitochondria

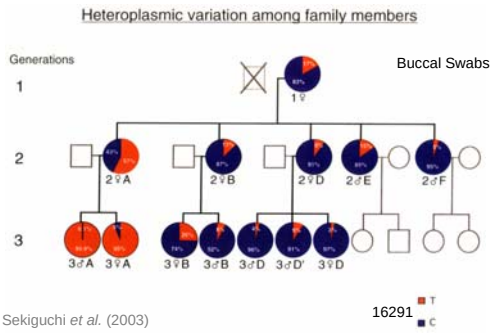
Very little mito growth until implantation

Females – produce ~7 million ova during fetal development only a few hundred become mature oocytes



Chinnery *et al.* (2000) Trends in Genetics

Heteroplasmic Variation



Sekiguchi *et al.* (2003)

Heteroplasmy in Hairs

Individual	Sample	Cytosines Before Position 310 T: Major Component	Minor Component	Type
Z	Blood sample 1 hair	8 Cs	9 Cs	C_8TC_6 C_9TC_6
	1 hair	8 and 9 Cs	10 Cs	C_8TC_6 C_9TC_6 $C_{10}TC_6$
	1 hair	7 and 8 Cs		C_7TC_6 C_8TC_6
	1 hair	7 Cs		C_7TC_6

Length variants in HV2 alone should not be used to support an interpretation of exclusion.

!

Blood + 1 hair	C_8TC_6 C_9TC_6
1 hair	C_7TC_6
1 hair	C_8TC_6 C_9TC_6 $C_{10}TC_6$

Stewart *et al.* JFS (2001)

Heteroplasmy in Hairs

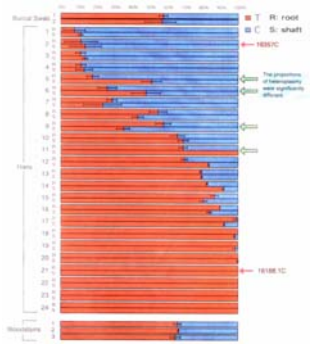
TABLE 2—*Seventy-eight independent observations of sequence heteroplasmy in 691 hairs.*

Heteroplasmy Position*	Other Matched K or Q Sample in Case for Comparison?	Nucleotide Present in Comparison Sample
70[†]	7 Q hairs	CRS in all 7
94	3 Q and 1 K hairs	CRS in all 4
120[†]	No	...
150	No	...
152	No	...
152	2 Q hairs and 1 K hair	All 3 hairs have T/C heteroplasmy
185	1 Q hair and 1 K blood	Both have A (substitution from CRS)

78 observations of point heteroplasmy in 691 hairs (11.4%)

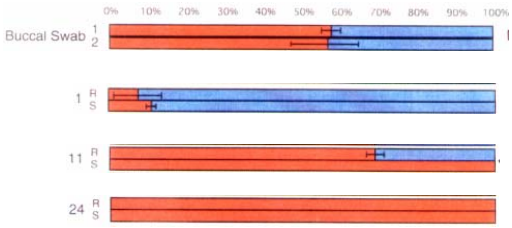
Melton *et al.* JFS (2005)

Heteroplasmy in Hairs



Sekiguchi *et al.* (2003)

Heteroplasmy in Hairs

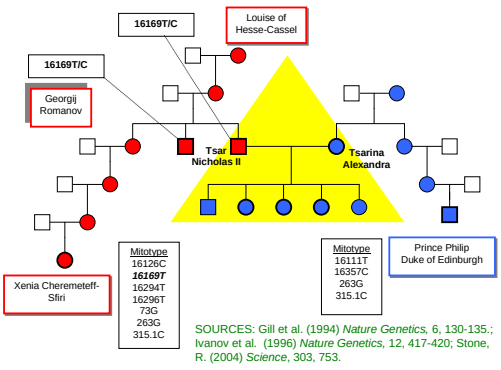


Heteroplasmy Detection

- Detection of heteroplasmy – sequencing can detect only to ~10% level.
- Other methods (e.g. Denaturing Gradient Gel Electrophoresis) are much more sensitive.

Famous Case Involving Heteroplasmy

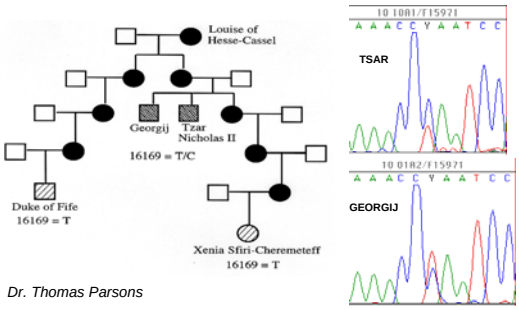
Identification of the Romanov Remains
(the Last Russian Czar)



SOURCES: Gill et al. (1994) *Nature Genetics,* 6, 130-135.; Ivanov et al. (1996) *Nature Genetics,* 12, 417-420; Stone, R. (2004) *Science*, 303, 753.

D.N.A. Box 10.2, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

AFDIL – Confirmation of FSS



Dr. Thomas Parsons

Interpretation of mtDNA Results

- Once the sequence has been generated (Q and K), and the differences from the rCRS are noted, what next?

SWGDAM Guidelines for Mitochondrial DNA (mtDNA)
Nucleotide Sequence Interpretation

(1) Exclusion

(2) Inconclusive

(3) Cannot Exclude (Failure to Exclude)

Interpretation of mtDNA Results

- Exclusion – if there are **two or more** nucleotide differences between the questioned and known samples, the samples can be excluded as originating from the same person or maternal lineage.

Q TATTG**C**AC**A**G
K TATTGTACGG

Sample Q	Sample K
6 T-C	263 A-G
9 G-A	315.1 C
263 A-G	
315.1 C	

Exclusion

Interpretation of mtDNA Results

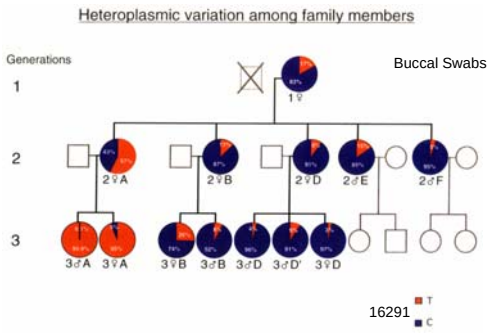
- Inconclusive – if there is **one** nucleotide difference between the questioned and the known samples, the result will be inconclusive.

Q TATTG**C**ACGG
K TATTG**T**ACGG

Sample Q	Sample K
6 T-C	263 A-G
263 A-G	315.1 C
315.1 C	

Inconclusive

Heteroplasmic Variation



Sekiguchi *et al.* (2003)

Interpretation of mtDNA Results

- Cannot Exclude – if the sequences from questioned and known samples under comparison have a common base at each position or a common length variant in the HV2 C-stretch, the samples cannot be excluded as originating from the same person or the same maternal lineage.

Q TATTGTACGG
K TATTGTACGG

Sample Q	Sample K
152 T-C	152 T-C
263 A-G	263 A-G
315.1 C	315.1 C

Cannot Exclude

POP QUIZ!!!

- How would you interpret these results?

Q TATTGTAC***A/G***G
K TATTGTAC ***G*** G

Sample Q	Sample K
9 G-R	263 A-G
263 A-G	315.1 C
315.1 C	

Q TATTGTAC***A/G***G
K TATTGTAC***G/A***G

Sample Q	Sample K
9 G-R	9 G-R
152 T-C	152 T-C
263 A-G	263 A-G
315.1 C	315.1 C

POP QUIZ!!!

Q TA***T***TGTAC***A***G
K TA***C***TGTAC***G***G

Sample Q	Sample K
16519 T-C	16519 T-C
9 G-A	3 T-C
263 A-G	263 A-G
315.1 C	315.1 C

Q TATTGTACGG
K TATTGTACGG

Sample Q	Sample K
16519 T-C	16519 T-C
152 T-C	152 T-C
263 A-G	263 A-G
309.1 C	315.1 C
315.1 C	

POP QUIZ!!!

Q TATTGTACGG
K TATTG***C***ACGG

Sample Q	Sample K
16519 T-C	16519 T-C
263 A-G	152 T-C
309.1 C	263 A-G
315.1 C	315.1 C

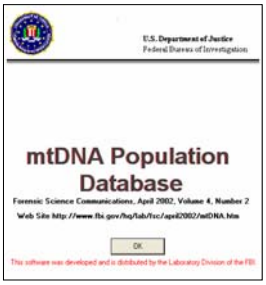
Q TATTGT***T***ACGG
K TAT **:** GT **:** ACGG

Sample Q	Sample K
16519 T-C	16519 T-C
6.1 T	4 T-del
263 A-G	263 A-G
315.1 C	315.1 C

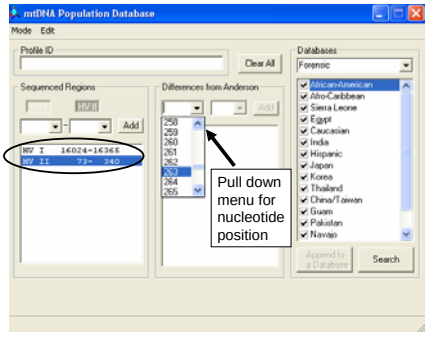
Reporting Statistics

- 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 (sequence) is seen in a database
- The larger the number of unrelated individuals in the database, the better the statistics will be for a random match frequency estimate.

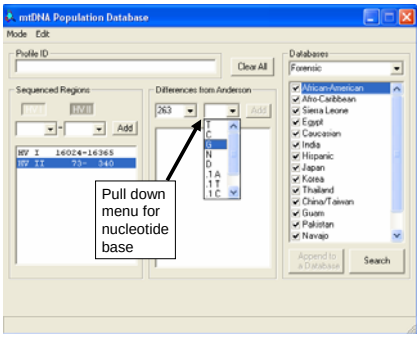
FBI mtDNA Population DB



FBI mtDNA Population DB

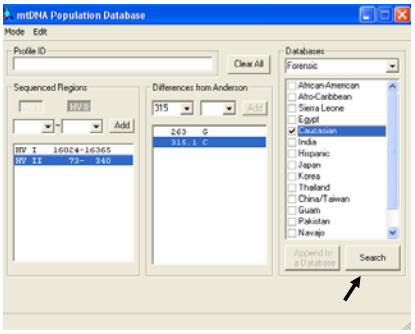


FBI mtDNA Population DB



FBI mtDNA Population DB

Enter all PMs from the haplotype



FBI mtDNA Population DB

Summary of Databases

Database	Classification	# Profiles
Caucasian	Caucasian origin	1655
Total		1655

Overall Match Results

Number of Differences From Search Profile	Number	Freq.	Cum. Num.	Cum. Freq.
0	70	0.0423	70	0.0423
1	171	0.1033	241	0.1456
2	218	0.1317	459	0.2773
3	173	0.1045	632	0.3819
4	195	0.1178	827	0.4997
5	198	0.1196	1025	0.6193
>5	630	0.3807	1655	1.0000

Average Number of Differences = 4.845

FBI mtDNA Population DB

Summary of Databases

Database	Classification	# Profiles
Caucasian	Caucasian origin	1655
Total		1655

~10% of the database is 1 mutation away from the searched haplotype

Number of Differences From Search Profile	Number	Freq.	Cum. Num.	Cum. Freq.
0	70	0.0423	70	0.0423
1	171	0.1033	241	0.1456
2	218	0.1317	459	0.2773
3	173	0.1045	632	0.3819
4	195	0.1178	827	0.4997
5	198	0.1196	1025	0.6193
>5	630	0.3807	1655	1.0000

Average Number of Differences = 4.845

Example Calculation of mtDNA Profile Frequency Estimate

The frequency (p) of observing a mtDNA profile (X) times in a database having a size of (N) is…

$p = X/N$

$p = 70/1665 = 0.042$

A 95% confidence interval can be determined by using a normal approximation of the binomial

$$p \pm 1.96 \sqrt{\frac{(p)\,(1-p)}{N}}$$

Upper Bound

$$p \pm 1.96 \sqrt{\frac{(0.042)\,(0.958)}{1665}} = p \pm 1.96\,(0.0049) = 0.032 \text{ and } 0.052$$

Holland and Parsons (1999) *Forensic Sci. Rev.*

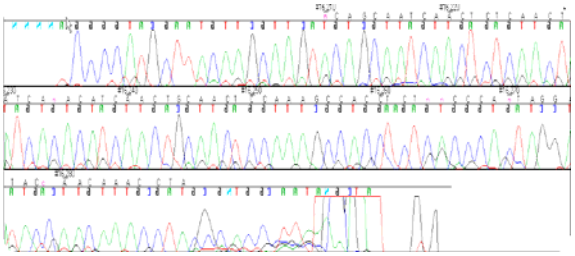
Issues Impacting mtDNA Interpretation

Challenges with mtDNA

- Data Interpretation
  - Heteroplasmy, Mixtures, Taq Error, and other Issues (Pseudogenes, etc…)
- DNA Database Sizes
  - Similar issues to Y-STRs but takes longer to generate mtDNA data than Y-STR haplotypes
- DNA Database Quality

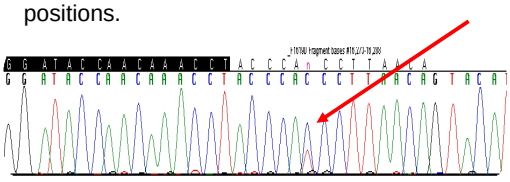
Poor Clean-Up

Poor clean-up of your amplification or sequencing product can cause pull-up, dye blobs, or other high background noise to appear in the EPGs.



Taq Error

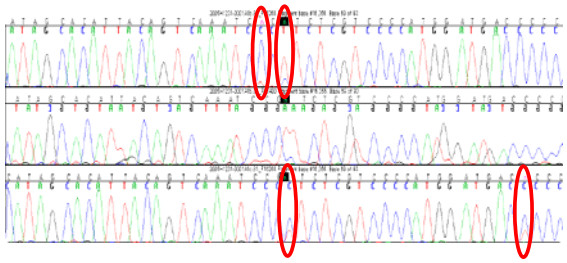
- Degraded template DNA and non-proofreading can cause mis-incorporation of bases at single positions.



- Re-amplification can correct the issue or show possible contamination.

Mixtures

- Samples can be truly mixed, either at collection, extraction or amplification.

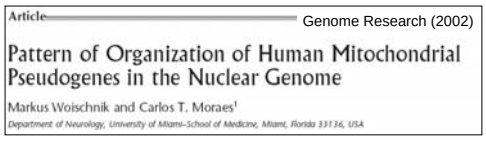


Heteroplasmy

- Heteroplasmy can look a lot like a mixture, but is *typically* only present at one position in the CR.
- Verification and authenticity of heteroplasmy by a second extraction of the sample is required.

Nuclear Pseudogenes

- Thoughout history – movement of mtDNA genes into the nucleus.
- Nuclear Pseudogenes (nuclear-mitochondrial like sequence ***numts***) – could potentially be amplified, confounding interpretation. “Molecular Fossils”

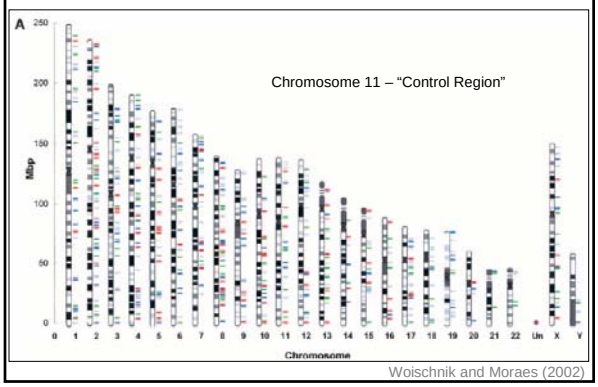

Article Genome Research (2002)

Pattern of Organization of Human Mitochondrial Pseudogenes in the Nuclear Genome

Markus Woischnik and Carlos T. Moraes[1]

Department of Neurology, University of Miami–School of Medicine, Miami, Florida 33136, USA

Nuclear Pseudogenes



Woischnik and Moraes (2002)

Nuclear Pseudogenes

- Typically – numts are not a problem for forensics – mtDNA high copy number
- "Mitochondrial DNA pseudogenes in the nuclear genome as possible sources of contamination" - *Goios A, Amorim A, Pereira L.* ISFG meeting in the Azores, 2005.
- Extraordinary measures to observe a numt with forensic mtDNA primers (Possibly seen by Grzbowski 2000 – nested PCR ~ 60 cycles).

mtDNA Recombination

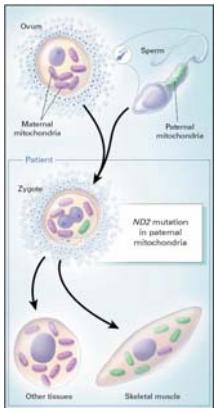
- Adam Eyre-Walker and colleagues - proposed that paternal contribution of mtDNA has caused recombination.
- Some of their assumptions along with the data that was analyzed have been wrong (more on that later).

Brief Report

PATERNAL INHERITANCE OF MITOCHONDRIAL DNA

MARIANNE SCHWARTZ, PH.D., AND JOHN VISSING, M.D., PH.D.

NEJM, 347(8) August 22, 2002



Williams - *NEJM,* 347(8) August 22, 2002

TABLE 2. SEQUENCE DIFFERENCES BETWEEN THE TWO MITOCHONDRIAL DNA (mtDNA) HAPLOTYPES FOUND IN THE PATIENT'S BLOOD AND MUSCLE AND IN HIS PARENTS' BLOOD.

NUCLEOTIDE POSITION	PATIENT'S BLOOD*	PATIENT'S MUSCLE	MOTHER'S BLOOD	FATHER'S BLOOD
477	C	T	C	T
1303	G	A	G	A
3192	C	T	C	T
3197†	T	C	T	C
3591	G	A	G	A
4592	T	C	T	C
5132–5134	AAA	delAA	AAA	AAA
11296	C	T	C	T
11467†	A	G	A	G
11719‡	G	A	G	A
11938	C	T	C	T
12308†	A	G	A	G
12372†	G	A	G	A
12618	G	A	G	A
13617†	T	C	T	C
14766‡	C	T	C	T
14793	A	G	A	G
15218	A	G	A	G

Paternal Leakage/Recombination

- Appears to be exceptionally rare – other studies of mtDNA diseases have not observed such a phenomenon.
- Kraytsberg *et al.* (2004) used DNA from the same affected patient to screen for recombination. These authors cite the presence of recombinants
- Single cell PCR is prone to generating artifacts (Bandelt 2005)

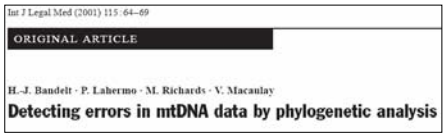
mtDNA Population Database: Size and Quality of Information

- Population databases are critical for estimating expected frequencies. The more, the better.

Database	# Profiles
African-American	1148
Afro-Caribbean	0
Sierra Leone	109
Caucasian	1655
Hispanic	686
Japan	163
Korea	182
Thailand	52
Navajo	146
Apache	180
Egypt	48
China/Taiwan	356
Guam	87
India	19 ←
Pakistan	8 ←
Total	4839

mtDNA Population Database: Size and Quality of Information

- Recently – mtDNA database quality has become an issue…


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

Artificial recombinations and phantom mutations plague the quality of mtDNA data in population genetics, forensics, and clinical studies

mtDNA Population Database: Size and Quality of Information

- Bandelt et al. (2001)

"In order to meet high-quality standards in forensics, sequencing should be performed in both directions (Bär et al. 2000). It is then important to read the two series of outputs separately (against the CRS) and to transform either series into a data table independently, preferably of different formats (motif vs dot table); finally, the two tables should be compared by computer."

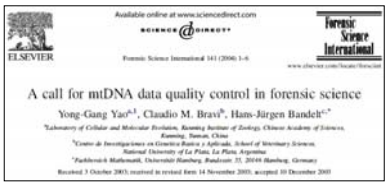
mtDNA Population Database: Size and Quality of Information

Problems in FBI mtDNA Database

Bandelt, Salas, and Bravi (2004) *Science*

Found 5 examples of artificial recombination among the 1148 African Americans in the database

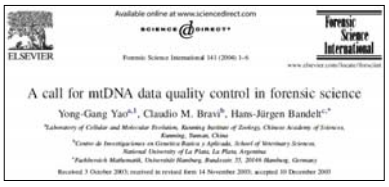
mtDNA Population Database: Size and Quality of Information



mtDNA database - USA.AFR.000942

HV1	HV2
16126-16187-16189-16223-16264	73-249d-263-290d-291d
16270-16278-16293-16311-16519	309.1C-315.1C-489

mtDNA Population Database: Size and Quality of Information



mtDNA database - USA.AFR.000942

HV1	HV2
16126-16187-16189-16223-16264	73-249d-263-290d-291d
16270-16278-16293-16311-16519	309.1C-315.1C-489
African haplogroup L1b	Asian haplogroup C1

mtDNA Population Database: Size and Quality of Information

- Phantom mutations – Bandelt et al. (2002); Brandstatter et al. (2005).
- Phantom mutations are systematic artifacts generated during cycle sequencing. These can be created by either the sequencing chemistry, the automated sequencer, or lab procedures.
- Single-strand sequencing (e.g. F only) is highly susceptible to generating phantom mutations.

mtDNA Population Database: Size and Quality of Information

Table 1. Combinations of sequencing instruments, PCR-purification procedures, and sequencing chemistries in the experimental study

Sequencer	Sequencing chemistries[a]	Post-PCR treatment	Number of sequences	Artificial deletions	Artificial insertions	Artificial substitutions	Artifacts per sequence	Sample score (average QV)	Source[b]
ABI310	BD1	Enzymatic	23	10	2	13	1.1	30.6	1
ABI310	BD1	*Via* columns	20	5	1	10	0.8	28.9	1
ABI310	BD2	*Via* columns	30	0	1	6	0.2	28.4	2
ABI310	DR	Enzymatic	48	14	4	57	1.6	29.7	1
ABI310	DR	*Via* columns	45	16	3	51	1.6	29.0	1
ABI377	BD2	*Via* columns	45	0	6	35	0.9	31.6	2
ABI377	BD3	*Via* columns	18	0	7	48	3.1	26.6	2
ABI377	DP	*Via* columns	34	5	2	56	1.9	28.1	2
ABI377	DR	*Via* columns	31	0	1	88	2.9	24.9	2
ABI3100	BD1	Enzymatic	50	20	2	65	1.7	31.6	1
ABI3100	BD2	*Via* columns	46	3	5	49	1.2	28.2	2
ABI3100	BD3	*Via* columns	38	2	5	52	1.6	31.6	1
ABI3100	DR	Enzymatic	21	29	3	181	10.1	27.7	1
ABI3100	DR	*Via* columns	16	23	0	169	12.0	28.9	1
ABI3700	BD2	Enzymatic	72	0	12	43	0.8	31.3	2

dRhodamine Terminator Cycle Sequencing Kit

Brandstatter *et al.* (2005).

mtDNA Population Database: Size and Quality of Information

Table 5. Haplotypes truncated to positions 317, 320, 330, 343, and 345 as observed in the SWGDAM database and in complete mtDNA sequences

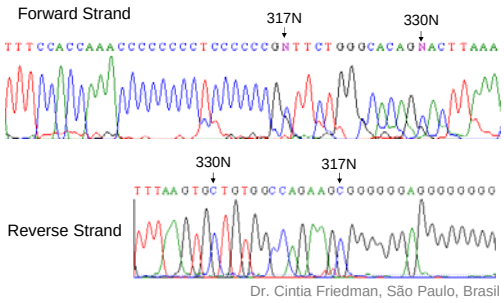
317	320	330	343	345	SWGDAM[c]	Complete mtDNA
C	C	C	C	C	$N = 4839$	$N > 1700$
A	.	.	.	.	1	1[g]
G	.	.	.	.	0	2[d,f]
T					1	0
N	.	.	.	.	14	0
del	.	.	.	.	1	0
.	T	.	.	.	0	2[d]
.	N	.	.	.	29	0
.	.	G	.	.	0	1[d]
.	.	N	.	.	10	0
.	.	.	N	.	3	0
.	.	.	.	N	4	0
.	.	.	.	T	1	1[e]
N	N	.	.	.	25	0
N	.	N	.	.	1	0
N	N	N	.	.	1	0
.	N	N	.	.	2	0
.	N	.	.	N	3	0
.	N	.	N	N	1	0
.	.	N	.	N	1	0
.	.	.	N	N	1	0
					99	7

HV2 phantom mutations

Post C-stretch ambiguities

Brandstatter *et al.* (2005).

mtDNA Population Database: Size and Quality of Information

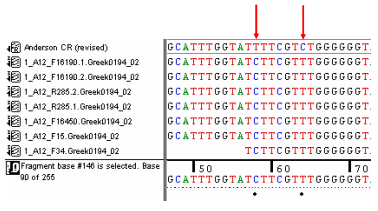


Dr. Cintia Friedman, São Paulo, Brasil

mtDNA Population Database: Size and Quality of Information

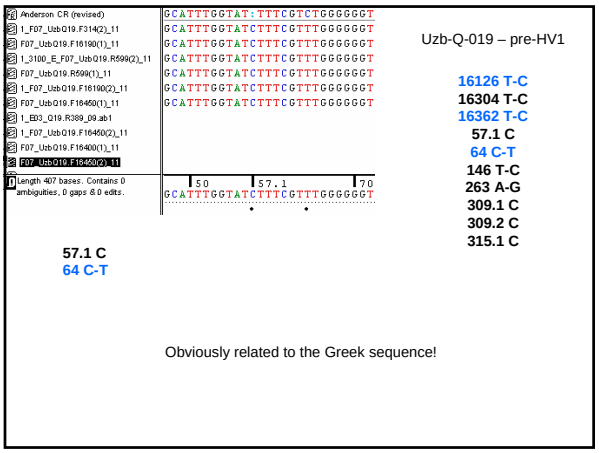
Additional Nomenclature Issues

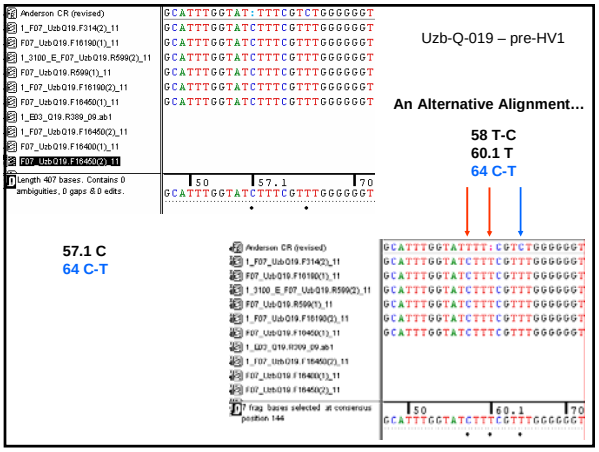
Greek0194 – pre-HV1

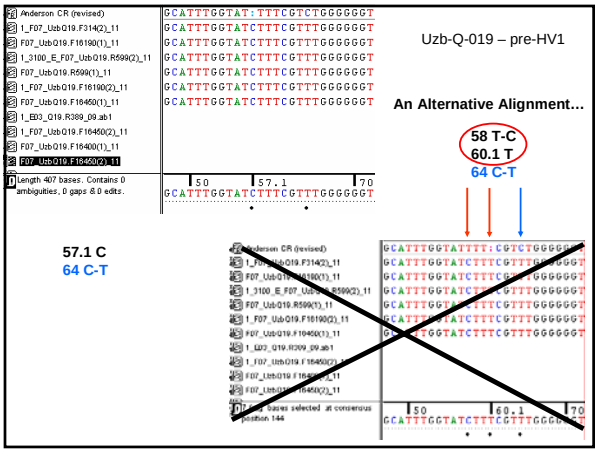


58 T-C
64 C-T

16126 T-C
16355 C-T
16362 T-C
58 T-C
64 C-T
146 T-C
152 T-C
263 A-G
309.1 C
309.2 C
315.1 C


Uzb-Q-019 – pre-HV1
16126 T-C
16304 T-C
16362 T-C
57.1 C
64 C-T
146 T-C
263 A-G
309.1 C
309.2 C
315.1 C
57.1 C
64 C-T
Obviously related to the Greek sequence!

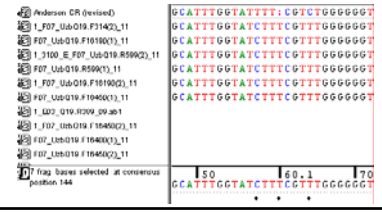

Uzb-Q-019 – pre-HV1
An Alternative Alignment…
58 T-C
60.1 T
64 C-T
57.1 C
64 C-T


Uzb-Q-019 – pre-HV1
An Alternative Alignment…
58 T-C
60.1 T
64 C-T
57.1 C
64 C-T

Align the "wrong" sequence with the Greek sequence….

Greek0194 – pre-HV1

58 T-C
64 C-T

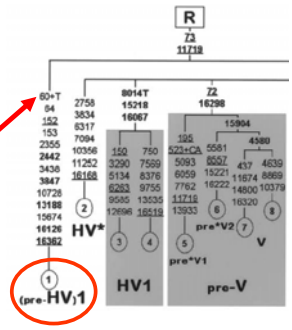


Uzb-Q-019 – pre-HV1

58 T-C
60.1 T
64 C-T

"Greek background + 60.1T"

Achilli *et al.* (2005) *AJHG*



T-insertion at np 60 has been observed before…

Uzb-Q-019 – pre-HV1 (wrong alignment)

58 T-C
60.1 T
64 C-T

Greek0194 – pre-HV1	Uzb-Q-019 – pre-HV1	
16126 T-C	**16126 T-C**	**16126 T-C**
16355 C-T	**16304 T-C**	**16304 T-C**
16362 T-C	**16362 T-C**	**16362 T-C**
58 T-C	**57.1 C**	**58 T-C**
64 C-T	**64 C-T**	**60.1 T**
146 T-C	**146 T-C**	**64 C-T**
152 T-C	**263 A-G**	**146 T-C**
263 A-G	**309.1 C**	**263 A-G**
309.1 C	**309.2 C**	**309.1 C**
309.2 C	**315.1 C**	**309.2 C**
315.1 C		**315.1 C**
	FBI method	"Global" Alignment

(3 differences between the two alignments)

mtDNA Population Database: Size and Quality of Information

- Recent efforts to increase DB sizes and quality have been undertaken by the NIJ (Grant to AFDIL Research Section) for entire control region sequences.
- EDNAP Mitochondrial Population Database (EMPOP) – developing QC tools to check sequences, including the ability to see electropherograms of all polymorphisms.

Control Region Databasing Goals

- Sequence **7500** individuals in two years.
- Focus on U.S. populations for now (W.European Caucasians, African Americans, Hispanics, Native Americans, and Asians).
- Provide entire control region data.
- Generate consistent, high quality databases.
- Adhere to a consistent nomenclature scheme
- Make data publicly available, via publications, GenBank and **EMPOP**.
- Conduct population genetic analyses of regional mtDNA substructure.

Wallac DBS Puncher

- positions plate
- intuitive
- bar-code ready
- THREE scientists

Qiagen 9604

•4 probe instrument
•extracts multiple sample types
•96 samples in 2-3 hrs

Corbett CAS-1200

•fully enclosed
•single probe, disposable tips
•amplification set-up in 30 minutes

Tecan Genesis

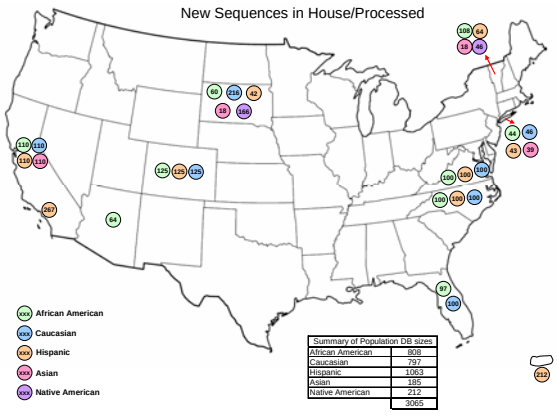
•8 probe instrument
•integrated plate sealer and thermal cyclers
•multiple plates in a day

Applied Biosystems 3730

•50-cm, 48-capillary array
•POP-7
•Holds 16 plates
•Barcode reader links plate to run sheet
•2 hrs. per plate

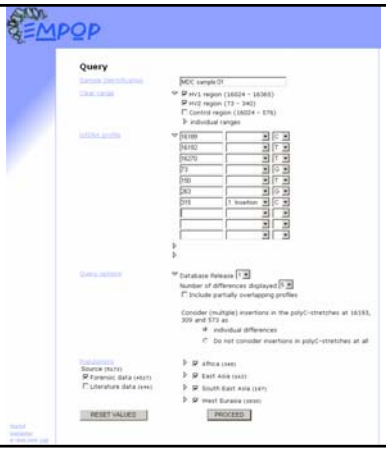
Safeguards against DB Errors

- **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 4 scientists review the RAW sequence data for every sample.
- **Electronic data transfer** – No manual transcription of data into master database.
- **Phylogenetic data checking and review** - EMPOP



Walther Parson
Institute of Legal Medicine, Innsbruck Medical University

Version 1.0 is online (16 October 2006)



Input check

Sample Identification	MDC sample 01
Database Release	1
Maximum differences displayed	5
Include partially overlapping profiles	no
Consider (multiple) insertions in the polyC-stretches at 16193, 309 and 573 as	Do not consider insertions in polyC-stretches at all
Selected ranges	73 - 340 16024 - 16365
mtDNA query profile	73A>G 150C>T 263A>G 315.1C 16189T>C 16192C>T 16270C>T
Source	Forensic data
Selected populations	Africa: ALL East Asia: ALL South East Asia: ALL West Eurasia: ALL

EDIT INPUT SUBMIT QUERY

Output

Selected ranges	73 - 340 16024 - 16365
mtDNA query profile	73A>G, 150C>T, 263A>G, 315.1C, 16189T>C, 16192C>T, 16270C>T

EDIT INPUT

OVERALL RESULT SUMMARY (4527/4527)

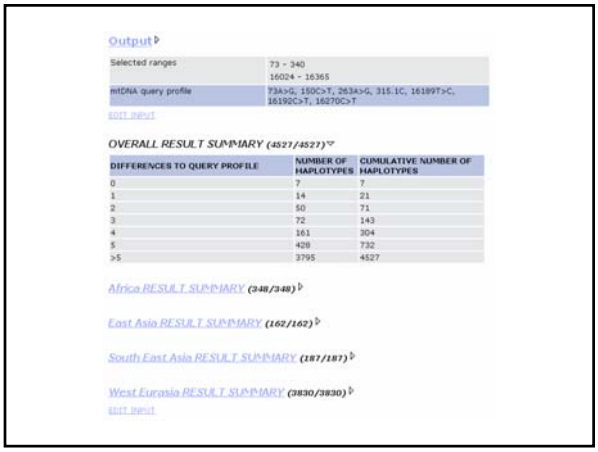
Africa RESULT SUMMARY (348/348)

East Asia RESULT SUMMARY (162/162)

South East Asia RESULT SUMMARY (187/187)

West Eurasia RESULT SUMMARY (3830/3830)

EDIT INPUT



Output

Selected ranges	73 - 340 16024 - 16365
mtDNA query profile	73A>G, 150C>T, 263A>G, 315.1C, 16189T>C, 16192C>T, 16270C>T

EDIT INPUT

OVERALL RESULT SUMMARY (4527/4527)

DIFFERENCES TO QUERY PROFILE	NUMBER OF HAPLOTYPES	CUMULATIVE NUMBER OF HAPLOTYPES
0	7	7
1	14	21
2	50	71
3	72	143
4	161	304
5	428	732
>5	3795	4527

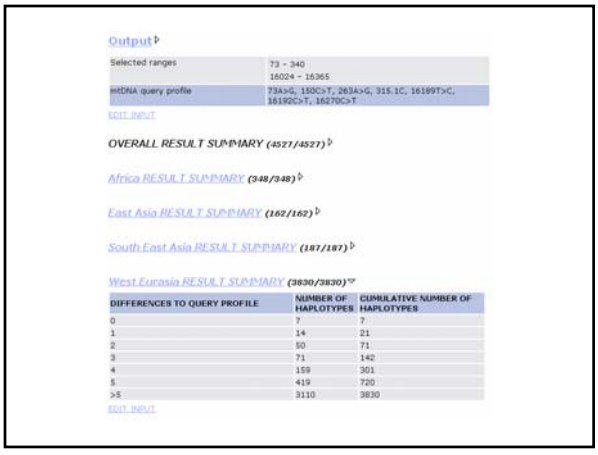
Africa RESULT SUMMARY (348/348)

East Asia RESULT SUMMARY (162/162)

South East Asia RESULT SUMMARY (187/187)

West Eurasia RESULT SUMMARY (3830/3830)

EDIT INPUT



Output

Selected ranges	73 - 340 16024 - 16365
mtDNA query profile	73A>G, 150C>T, 263A>G, 315.1C, 16189T>C, 16192C>T, 16270C>T

EDIT INPUT

OVERALL RESULT SUMMARY (4527/4527)

Africa RESULT SUMMARY (348/348)

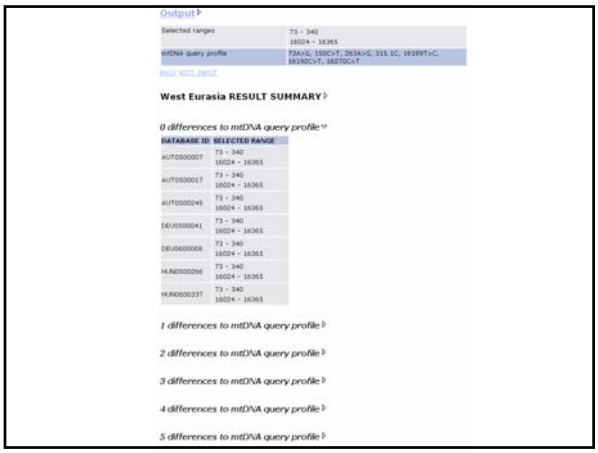
East Asia RESULT SUMMARY (162/162)

South East Asia RESULT SUMMARY (187/187)

West Eurasia RESULT SUMMARY (3830/3830)

DIFFERENCES TO QUERY PROFILE	NUMBER OF HAPLOTYPES	CUMULATIVE NUMBER OF HAPLOTYPES
0	7	7
1	14	21
2	50	71
3	71	142
4	159	301
5	419	720
>5	3110	3830

EDIT INPUT



Output

Selected ranges	73 - 340 16024 - 16365
mtDNA query profile	73A>G, 150C>T, 263A>G, 315.1C, 16189T>C, 16189C>T, 16270C>T

BACK PRINT INPUT

West Eurasia RESULT SUMMARY

0 differences to mtDNA query profile

DATABASE ID	SELECTED RANGE
AUT0500007	73 - 340 16024 - 16365
AUT0500017	73 - 340 16024 - 16365
AUT0500245	73 - 340 16024 - 16365
DEU0500041	73 - 340 16024 - 16365
DEU0600068	73 - 340 16024 - 16365
HUN0500266	73 - 340 16024 - 16365
HUN0600337	73 - 340 16024 - 16365

1 differences to mtDNA query profile

2 differences to mtDNA query profile

3 differences to mtDNA query profile

4 differences to mtDNA query profile

5 differences to mtDNA query profile

EMPOP tool for data checking

Translating DNA data tables into quasi-median networks for parsimony analysis and error detection

Hans-Jürgen Bandelt [a,*], Arne Dür [b]

[a] *Department of Mathematics, University of Hamburg, 20146 Hamburg, Germany*
[b] *Institute of Mathematics, University of Innsbruck, 6020 Innsbruck, Austria*

Received 4 June 2006; revised 19 July 2006; accepted 19 July 2006

Molecular Phylogenetics and Evolution (in press)

EMPOP tool for data checking

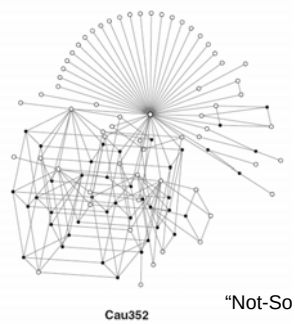
ASG373

Ken100

"Good Data"

EMPOP tool for data checking

Repetitious Artifacts



"Not-So Good Data"