

# EMPOP and Consistent MtDNA Alignments

Dixie Peters, MS

UNTHSC Center for Human Identification

National Technical Leader Summit

CODIS Missing Persons Working Group

Thursday, December 6, 2018

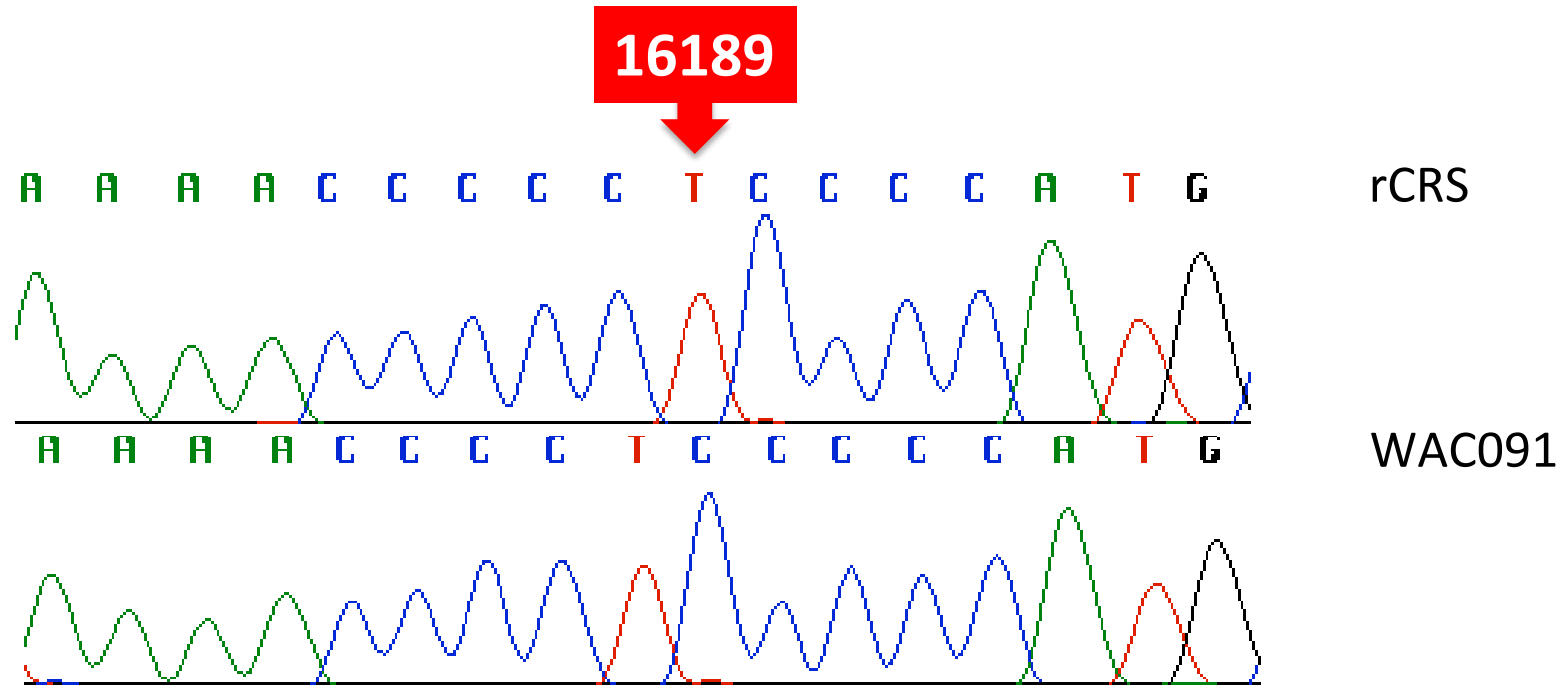
# SWGDM Update

- NGS Committee revised Interpretation Guidelines for MtDNA Analysis by Forensic DNA Testing Laboratories
  - Specific to NGS related items
  - Separated NGS from SS
- LM Committee will revise guidelines again
- LM Committee will finalize revisions for Y-STR guidelines in 2019

# Interp Guidelines for MtDNA Analysis

- Recommends use of EMPOP:
  - For those rare sequences containing peculiar insertions and deletions, it is recommended that the forensic analyst use the EDNAP mtDNA Population Database (EMPOP) to help determine a consistent mtDNA haplotype for entry into forensic databases.
- Same examples included
- EMPOP version updated
- **Rule 3b** - Homopolymeric C-Stretches in Hypervariable Region II (HVII): C-stretches in HV2 should be interpreted with a 310C when the otherwise anchored T at position 310 is not present. **C-stretches should be interpreted with a 311T when the anchored T at position 310 is followed by a second T.**
- Rule 9 added (outside control region)
- Nomenclature Examples Documented-not updated

# Alignment can be ambiguous



**16188T 16189C**  
 phylogenetic alignment (Bandelt & Parson 2008)

=

**16188- 16193+C**  
 Formal alignment rules (Wilson et al 2002)

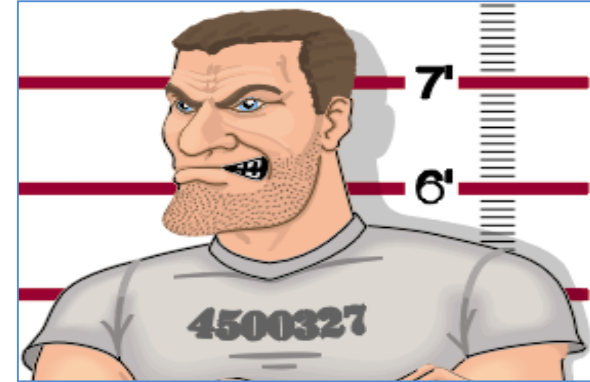
- Phylogenetic rule
- Anchor 16189 and 310
- 3' Alignment

- 1 Apply Max Parsimony
- 2 Indels > Transversions > Transitions
- 3 3' Alignment

# Consequences of alignment ambiguity

1. Forensic interpretation
2. Database searches

# Effect of alignment on forensic interpretation\*



Exclusion	Inconclusive	Inclusion
two or more differences	one difference	identical (+Het)
16188T 16189C phylogenetic alignment (Bandelt & Parson 2008)		16188- 16193+C Formal alignment rules (Wilson et al 2002)

**3 differences**

\* Carracedo et al *FSI* 2000, SWGDAM 2013, Parson et al *FSIG* 2014

# Effect of alignment on database searches

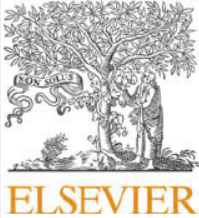

Search method	16188T 16189C	16188- 16193+C
rCRS-coded	28 matches	0 matches

EMPOP, N = 34,617

# Alignment-immune searches in EMPOP

Forensic Science International: Genetics 5 (2011) 126–132

Contents lists available at ScienceDirect

 Forensic Science International: Genetics 

journal homepage: [www.elsevier.com/locate/fsig](http://www.elsevier.com/locate/fsig)

---

## SAM: String-based sequence search algorithm for mitochondrial DNA database queries

Alexander Röck<sup>a</sup>, Jodi Irwin<sup>b</sup>, Arne Dür<sup>a</sup>, Thomas Parsons<sup>c</sup>, Walther Parson<sup>d,\*</sup>

<sup>a</sup>Institute of Mathematics, University of Innsbruck, Technikerstrasse 13, 6020 Innsbruck, Austria  
<sup>b</sup>The Armed Forces DNA Identification Laboratory, 1413 Research Blvd., Rockville, MD 20850, USA  
<sup>c</sup>The International Commission on Missing Persons, Alipašina 45 A, 71000 Sarajevo, Bosnia and Herzegovina  
<sup>d</sup>Institute of Legal Medicine, Innsbruck Medical University, Müllerstrasse 44, 6020 Innsbruck, Austria

Alignment-immune sequence queries guarantee that a haplotype is not missed in a database search.  
What about reported haplotypes?

SAM on EMPOP since v3 (12-27-2010)



# 2013 SWGDAM mtDNA IG Nomenclature Rules

## 2.3.3 SWGDAM Nomenclature Rules

Variants from the rCRS should be coded in accordance with the following nomenclature rules:

**Rule 1** – Maintain known patterns of polymorphisms (a.k.a. known phylogenetic alignments). Most violations to known patterns of polymorphisms involve insertions and deletions. |

Example: Maintain deletions at positions 249, 290 and/or 291 when present. See other examples in the SWGDAM mtDNA Nomenclature Examples document.

**Rule 2** - Use nomenclature with the least number of differences unless it violates known patterns of polymorphisms.

**Rule 3a** – Homopolymeric C-Stretches in Hypervariable Region I (HVI): C- stretches in HV1 should be interpreted with a 16189C when the otherwise anchored T at position 16189 is not present. Length variation in the short A- tract preceding 16184 should be noted as transversions.

**Rule 3b** - Homopolymeric C-Stretches in Hypervariable Region II (HVII): C- stretches in HV2 should be interpreted with a 310 C when the otherwise anchored T at position 310 is not present.

**Rule 4** – Maintain the AC Repeat Motif in the HVIII region from np 515-525.

**Rule 5** – Prefer substitutions to insertions/deletions (indels).

**Rule 6** – Prefer transitions to transversions unless this is in conflict with Rule 1.

**Rule 7** – Place indels contiguously when possible.

**Rule 8** - Place indels on the 3' end of the light strand.

# What is the phylogenetic rule?

- Ewans and Grant (2001) Statistical methods in bioinformatics, Springer, pg. 184
  - “Good alignments of related sequences are ones that better reflect the evolutionary relationship between them.”
- According to **the phylogenetic rule** the preferred alignment is based relative to the closest evolutionary neighbors and not relative to the rCRS

**Alignment** relative to the closest evolutionary neighbor  
**Reporting** relative to rCRS

# SAM2



- SAM lacked the features to harmonize alignment in reported haplotypes
- SAM2 provides users with unbiased database search results and harmonizes haplotype alignment

# Alignment-immune searches in EMPOP

Forensic Science International: Genetics 37 (2018) 204–214

---

Contents lists available at [ScienceDirect](#)

 Forensic Science International: Genetics 

journal homepage: [www.elsevier.com/locate/fsigen](http://www.elsevier.com/locate/fsigen)

---

Research paper

## Next generation database search algorithm for forensic mitogenome analyses

Nicole Huber<sup>a</sup>, Walther Parson<sup>a,b,\*</sup>, Arne Dür<sup>c</sup>

<sup>a</sup> *Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria*  
<sup>b</sup> *Forensic Science Program, The Pennsylvania State University, University Park, PA, USA*  
<sup>c</sup> *Institute of Mathematics, University of Innsbruck, Austria*

---

<p>ARTICLE INFO</p> <hr/> <p><b>Keywords:</b> Mitochondrial DNA</p>	<p>ABSTRACT</p> <hr/> <p>Mitochondrial DNA (mtDNA) variation is being reported relative to the corrected version of the first sequenced human mitochondrial genome. A review of the existing literature across disciplines that employ mtDNA de-</p>
---	--

FSIG: 37 (2018) 204-234

**SAM2** on EMPOP since v4/R11 (9-10-2018)

# Hierarchy of EMPOP searches

1. Determines the subset of EMPOP database samples includes the range of the query haplotype
2. Converts difference-coded format to strings and compares
3. Computes the best transcript for all neighbors and determines the minimum cost
  - Transcript shows the difference between query sample and the database sample(s) the query sample hit matched

# Pattern vs. Literal

- Pattern mode matches to multiple options
  - 152Y will match to database samples with 152C, 152T, and 152Y
  - Standard choice
- Literal mode will match only to the designated difference
  - 152Y will only match to database samples with 152Y
  - Used for investigating occurrence of point heteroplasmy

# Extended IUPAC rules

Forensic Science International: Genetics 13 (2014) 134–142



Contents lists available at [ScienceDirect](#)

Forensic Science International: Genetics

journal homepage: [www.elsevier.com/locate/fsig](http://www.elsevier.com/locate/fsig)



## DNA Commission of the International Society for Forensic Genetics: Revised and extended guidelines for mitochondrial DNA typing



W. Parson<sup>a,b,\*</sup>, L. Gusmão<sup>c,d</sup>, D.R. Hares<sup>e</sup>, J.A. Irwin<sup>e</sup>, W.R. Mayr<sup>f</sup>, N. Morling<sup>g</sup>, E. Pokorak<sup>e</sup>,  
M. Prinz<sup>h</sup>, A. Salas<sup>i</sup>, P.M. Schneider<sup>j</sup>, T.J. Parsons<sup>k</sup>

<sup>a</sup> Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria

<sup>b</sup> Penn State Eberly College of Science, University Park, PA, USA

<sup>c</sup> DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Brazil

<sup>d</sup> IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto, Portugal

<sup>e</sup> FBI Laboratory, Quantico, VA, USA

<sup>f</sup> Division of Blood Group Serology, Medical University of Vienna, Austria

<sup>g</sup> Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

<sup>h</sup> Department of Sciences, John Jay College for Criminal Justice, New York, NY, USA

<sup>i</sup> Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, and Instituto de Ciencias Forenses, Grupo de Medicina Xenómica (GMX),  
Facultade de Medicina, Universidade de Santiago de Compostela, 15872 Galicia, Spain

<sup>j</sup> Institute of Legal Medicine, Medical Faculty, University of Cologne, Cologne, Germany

<sup>k</sup> International Commission on Missing Persons, Alipasina 45a, 71000 Sarajevo, Bosnia and Herzegovina

## Mixture of C299 and 299DEL

Suggested notation: c299



**Fig. 1.** Example showing a mixture of a deletion at 299 (HVS-II) and the undeleted variant C (=rCRS). The IUPAC code does not provide acronyms for mixtures of deleted/undeleted (inserted/non-inserted) bases. We suggest extending the IUPAC code with lower case letters to describe such mixtures. In this example the assignment would be c299. The upper epg shows the forward sequencing result where the heteroplasmic C/- position is indicated by a C/A base call, following by subsequent base overlaps. The lower epg shows the reverse sequencing result where base overlaps occur upstream from c299.



## IUPAC code

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N
-	Gap	-

## Extended IUPAC code

a ... A/del

g ... G/del

y ... C/T/del

d ... A/G/T/del

n ... A/G/C/T/del

# Parson et al 2014 *FSIG*

## **Recommendation #8**

**IUPAC conventions using capital letters shall be used to describe differences to the rCRS and (point heteroplasmic) mixtures. Lower case letters should be used to indicate mixtures between deleted and non-deleted (inserted and non-inserted) bases. N-designations should only be used when all four bases are observed at a single position (or if no base call can be made at a given position). For the representation of deletions, “DEL”, “del” or “–” shall be used.**


# EMPOP is case sensitive

Lower case letters are turned into upper case letters except when explicitly disabled by ticking the “Use extended IUPAC code” option

Sample ID

Ranges

Profile

Use extended IUPAC code 

Release

When using the extended IUPAC notation EMPOP discerns between upper and lower case letters. See updated ISFG guidelines, Parson et al. 2014 FSIG

# Acceptable Annotations

Updated ISFG recommendations (Parson et al 2014 *FSIG*)

**Table 1 – Notation Guidelines:**

<b>Type</b>	<b>Possible annotations</b>	<b>Comment</b>
Base changes	73G, A73G	If preceding bases are used they must match rCRS base at the given position
Insertions	315.1C -315.1C 315+C	For multiple insertions all preceding insertions need to be stated, i.e. annotating 309.2C is not possible without annotating 309.1C
Deletions	249- A249- 249delA 249del	'del' is treated case insensitive, e.g. Del, DEL, dEL, deL etc is accepted. Please note that the single character 'D' is considered a mixture of A, G, and T (IUB code). The single character 'd' is considered a mixture of A, G, T, and deletion (see <a href="#">Röck et al. 2013</a> for details).

# V4/R11

- <https://empop.online/>
- EMPOP Stats
  - Total: 34,617
    - Full genomes: 256
    - 33,691: 16024-16365, 73-340
    - 26,127: 16024-570
  - US: 10,799
    - Full genomes: 0

# HV1 Examples

### Phylogenetic alignment

Input Profile	64T	73G	146C	153R	235G	263G	309.1C	315.1C	16111T	16188.1C	16192Y	16223T	16290T	16319A	16362C			
Phylogenetic alignment	64T	73G	146C	153R	235G	263G	309.1C	315.1C	16111T		16189C	16190T	16191.1C	16192Y	16223T	16290T	16319A	16362C

Alignment was estimated using SAM 2.0 on the basis of 5,440 haplogroup motifs (Phylotree, Build 17) following the phylogenetic concept and the recommendations of the ISFG and was derived from haplogroup

AZ+(64)+16189 | AZap

in range 16024-16366 52-407 by the following transcript with cost 2.04:

[M16183A(0.00)] C16190T(1.03) 16191insC(0.25) [Y64T(0.00)] G153R(0.36) 309insC(0.40)

### Warnings produced by your query

query profile range 16024-16366 52-407 has smaller length than CR (phylogenetic alignment unsure for small ranges)

Reference		...
16,111	C	I
16,188.1	:	C
16,192	C	Y
16,223	C	I
16,290	C	I
16,319	G	A
16,362	T	C
64	C	I
73	A	G
146	T	C
153	A	R
235	A	G
263	A	G
309.1	:	C
315.1	:	C



0 matches/28078

16024-16366 52-407

## Phylogenetic alignment

Input Profile	73G	210G	263G	315.1C	16183C	16189C	16193.1C	16195-	16217C	16258C
Phylogenetic alignment	73G	210G	263G	315.1C	16183C	16189C	16193.1C	16195-	16217C	16258C

Alignment was estimated using SAM 2.0 on the basis of 5,440 haplogroup motifs (Phylotree, Build 17) following the phylogenetic concept and the recommendations of the ISFG and was derived from haplogroup

B2 | B2b | B2b3 | B2c | B2c1 | B2c1a | B2c1b | B2c1c | B2d | B2e | B2f | B2h | B2i | B2i2 | B2l | B2n | B2p | B2q | B2r | B4 | B4b | B4b'd'e'j | B4c | B4c1 | B4c1a | B4c1a'b | B4c1a2 | B4d | B4d1 | B4d1'2'3 | B4d1a | B4d2

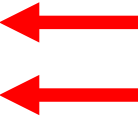
in range 16024-16385 50-407 by the following transcript with cost 3.46:

[M16183C(0.00)] 16193insC(0.20) 16195delT(1.39) A16258C(0.96) A210G(0.91)

0 matches/28066

16024-16385 50-407

Reference		08...
16,183	A	<u>C</u>
16,189	T	<u>C</u>
16,193.1	:	<u>C</u>
16,195	T	:
16,217	T	<u>C</u>
16,258	A	<u>C</u>
73	A	<u>G</u>
210	A	<u>G</u>
263	A	<u>G</u>
315.1	:	<u>C</u>
⚡	Total	10



Original: 16194C, 16195A



HV2 examples (50-72)

### Phylogenetic alignment

Input Profile	58C	60.1T	60.2T	64A	65-	71-	73G	189G	194T	195C	204C	207A	263G	309.1C	315.1C	16223T	16292T	16311C
Phylogenetic alignment	58C	60.1T	60.2T	64-	65-	66A	73G	189G	194T	195C	204C	207A	263G	309.1C	315.1C	16223T	16292T	16311C

Alignment was estimated using SAM 2.0 on the basis of 5,440 haplogroup motifs (Phylotree, Build 17) following the phylogenetic concept and the recommendations of the ISFG and was derived from haplogroup

W5b1

in range 16024-16386 51-407 by the following transcript with cost 2.43:

T16311C(0.45) C66A(1.59) 309insC(0.40)

Reference		08-...
16,223	C	I
16,292	C	I
16,311	T	C
58	T	<u>C</u>
60.1	:	I
60.2	:	I
64	C	<u>A</u>
65	T	:
71	G	:
73	A	<u>G</u>
189	A	<u>G</u>
194	C	I
195	T	<u>C</u>
204	T	<u>C</u>
207	G	<u>A</u>
263	A	<u>G</u>
309.1	:	<u>C</u>
315.1	:	<u>C</u>

0 matches/28068

16024-16386 51-407

## Phylogenetic alignment

Input Profile	55.1T	59-	60-	71.1G	73G	143A	152C	263G	308-	309-	315.1C
Phylogenetic alignment	55.1T	59-	60-	71.1G	73G	143A	152C	263G	308-	309-	315.1C

Alignment was estimated using SAM 2.0 on the basis of 5,440 haplogroup motifs (Phylotree, Build 17) following the phylogenetic concept and the recommendations of the ISFG and was derived from haplogroup

M39

in range 48-407 by the following transcript with cost 4.37:

65delT(1.49) [K66G(0.00)] 71insG(1.06) G143A(0.64) T152C(0.38) 308-309delCC(0.80)

Reference		F-2...
55.1	:	T
59	T	:
60	T	:
71.1	:	G
73	A	G
143	G	A
152	T	C
263	A	G
308	C	:
309	C	:
315.1	:	C

4 matches/28058

48-407

Original: 56- 58A 71.1G

# HV2 Examples

### Phylogenetic alignment

Input Profile	73G	198T	210G	247-	263G	290-	291-	309.1C	315.1C	16223T	16298C	16325C	16327T	
Phylogenetic alignment	73G	198T	210G	247A	249-	263G	290-	291-	309.1C	315.1C	16223T	16298C	16325C	16327T

Alignment was estimated using SAM 2.0 on the basis of 5,440 haplogroup motifs (Phylotree, Build 17) following the phylogenetic concept and the recommendations of the ISFG and was derived from haplogroup

C1f

in range 16024-16383 51-407 by the following transcript with cost 2.04:

C198T(0.73) A210G(0.91) 309insC(0.40)

Reference		Contig...
16,223	C	<u>I</u>
16,298	T	<u>C</u>
16,325	T	<u>C</u>
16,327	C	<u>I</u>
73	A	<u>G</u>
198	C	<u>I</u>
210	A	<u>G</u>
247	G	<u>A</u>
249	A	:
263	A	<u>G</u>
290	A	:
291	A	:
309.1	:	<u>C</u>
315.1	:	<u>C</u>



0 matches/28068

16024-16383 51-407

Entered as 247-

## Phylogenetic alignment

Input Profile	73G	152C	263G	310-	315-	16182C	16183C	16189C	16217C	16295T	
Phylogenetic alignment	73G	152C	263G	310C	314-	315-	16182C	16183C	16189C	16217C	16295T

Alignment was estimated using SAM 2.0 on the basis of 5,440 haplogroup motifs (Phylotree, Build 17) following the phylogenetic concept and the recommendations of the ISFG and was derived from haplogroup

**B2c2b**

in range 16024-16379 51-340 by the following transcript with cost 3.04:

1 matches/29036

16024-16379 51-340

Reference		URP1...
16,182	A	<u>C</u>
16,183	A	<u>C</u>
16,189	T	<u>C</u>
16,217	T	<u>C</u>
16,295	C	<u>I</u>
73	A	<u>G</u>
152	T	<u>C</u>
263	A	<u>G</u>
310	T	<u>C</u>
314	C	:
315	C	:



Entered as 310- 315-

## Phylogenetic alignment

Input Profile	73G	263G	310.1T	315.1C	
Phylogenetic alignment	73G	263G	311T	315.1C	315.2C

Alignment was estimated using SAM 2.0 on the basis of 5,440 haplogroup motifs (Phylotree, Build 17) following the phylogenetic concept and the recommendations of the ISFG and was derived from haplogroup

0 matches/29040

51-399

73G

263G

310.1T

315.1C



# Know your weird spots!

- HV1 C-stretch
- HV2 C-stretch
- HV2 anything before 73G (50-71)
- 247- (should be 247A, 249-)
- Any odd insertion/deletion combination



# Going Forward

- Update interp guidelines to reflect the 2013/2019 SWGDAM IG nomenclature rules
  - Include the use of EMPOP
- Train analysts to be aware of trouble spots
  - Include training on use of EMPOP
- Decide when and how EMPOP will be used (all haplotypes?)
  - May require new or less experienced analysts to run all haplotypes through EMPOP

# Going Backward

- May receive QC list from NDIS
- Laboratory review QC list (TL, analysts)
- Laboratory will need to
  - Devise procedure for EMPOP alignment check
  - Realign sequence data to ensure fit
  - Tech review of EMPOP alignment check and realignment of data
  - Update CODIS entry