MtDNA profiles and associated haplogroups

A systematic approach to an old problem

Anita Brandstätter[1], Alexander Röck[2], Arne Dür[2], Walther Parson[1]

[1]Institute of Legal Medicine
Innsbruck Medical University

[2]Institute of Mathematics
University of Innsbruck

22nd Congress of the International Society for Forensic Genetics
Copenhagen 2007

Outline

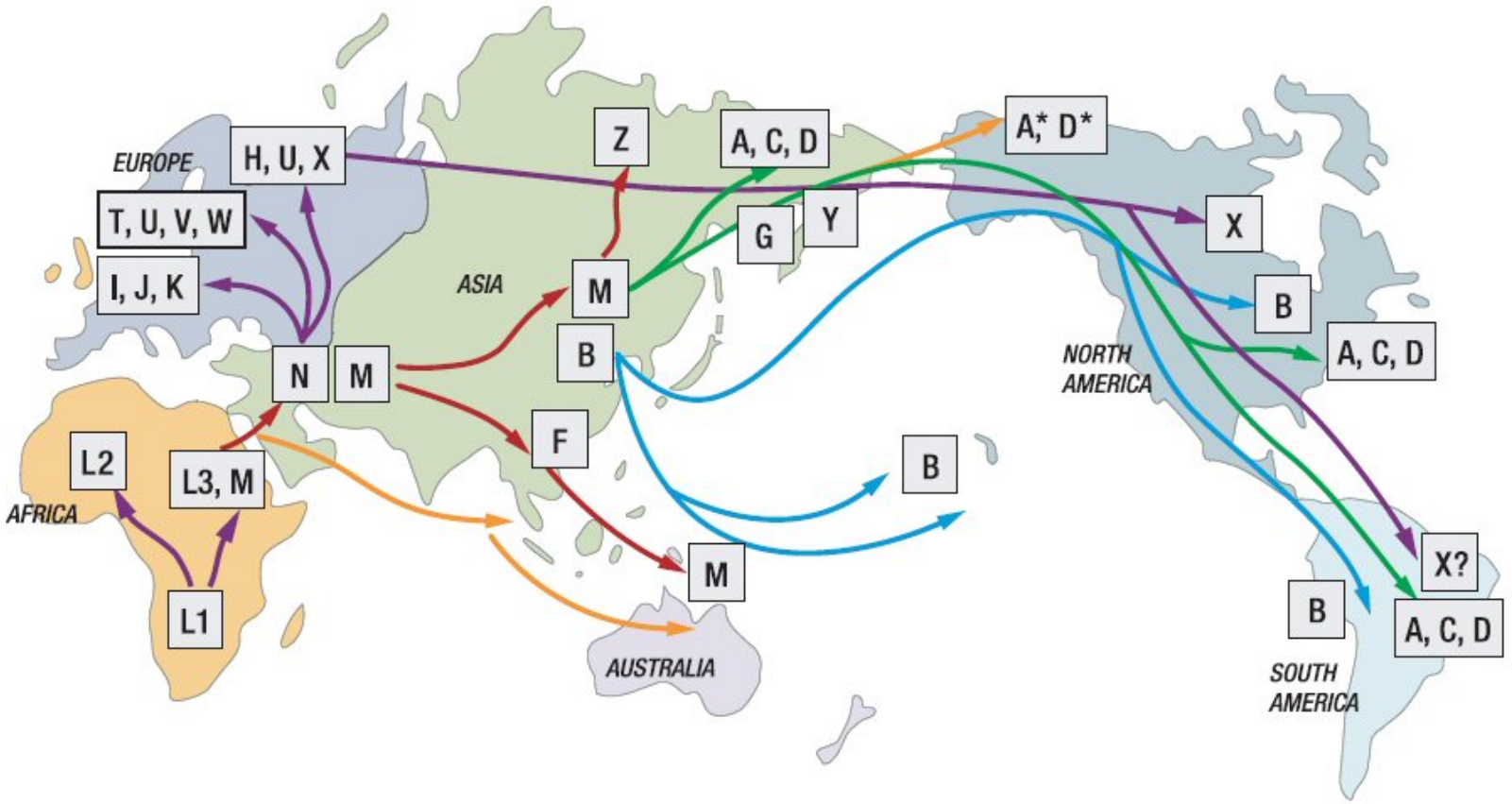
Outline

Human mitochondrial haplogroups

Characterization

- Human phylogeny $\Leftrightarrow$ Emergence of distinct lineages
- Haplogroups:
  - Clusters of evolutionary closely related haplotypes
  - Defined by the presence of specific polymorphisms in the entire mitochondrial genome that are identical by descent
  - Result of the emigration of human populations out of Africa
  - Reflect human migration routes over the different continents
  - Determine an association of distinct mitochondrial polymorphisms to ethnic populations

Human migration routes (rough overview)



Human mitochondrial haplogroups

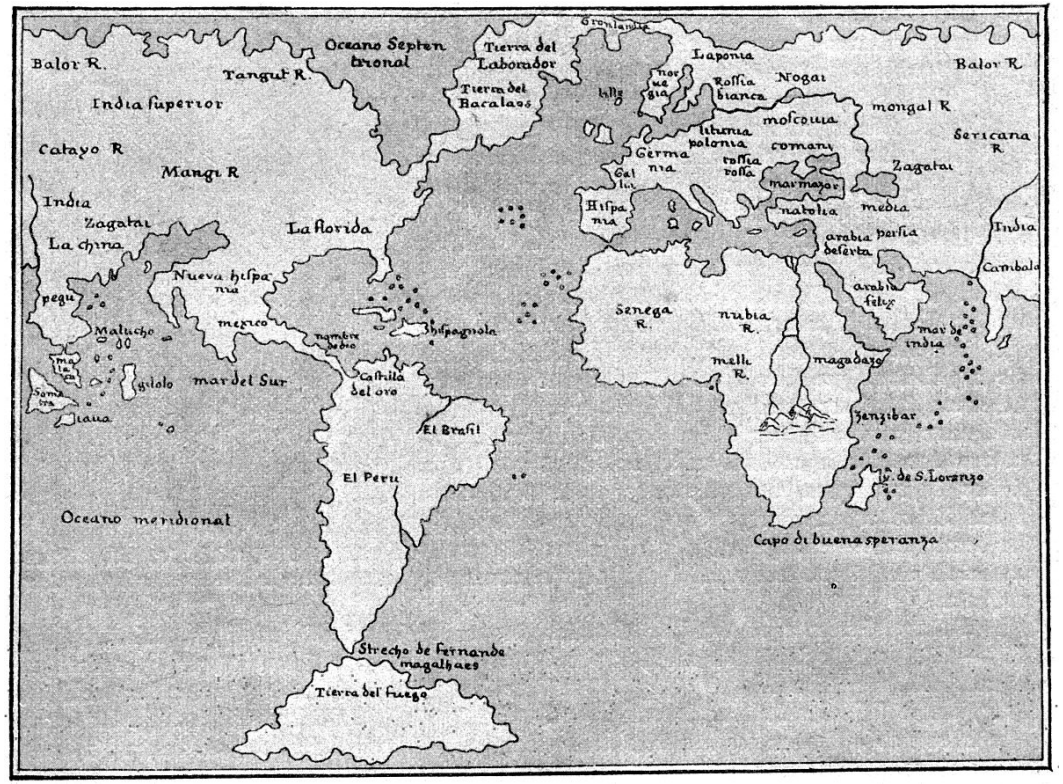
Problems

- Backbone of the human mitochondrial tree is well defined
- New haplogroups are being found continuously

$\Rightarrow$ Tips of the terminal branches are still under refinement

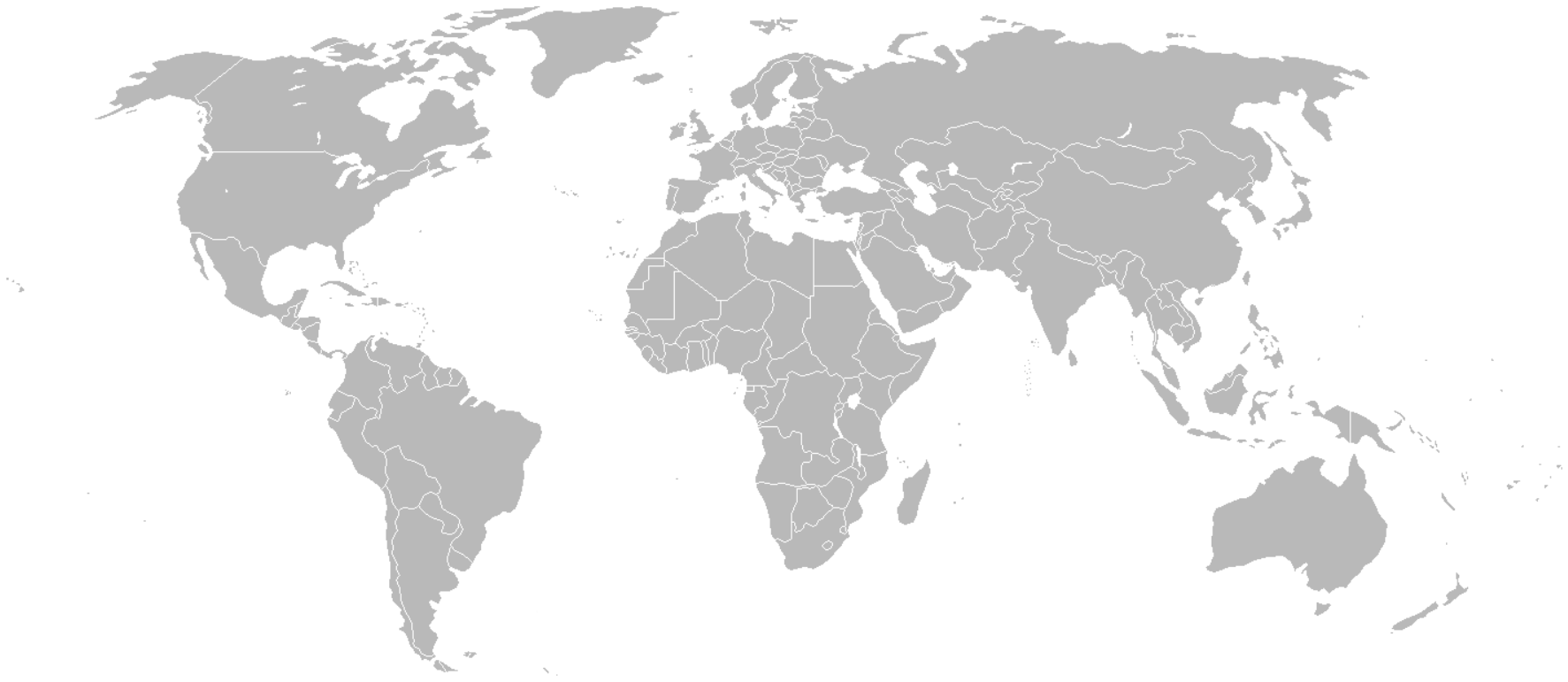
History of the geographical perception of the world...

Around 1550



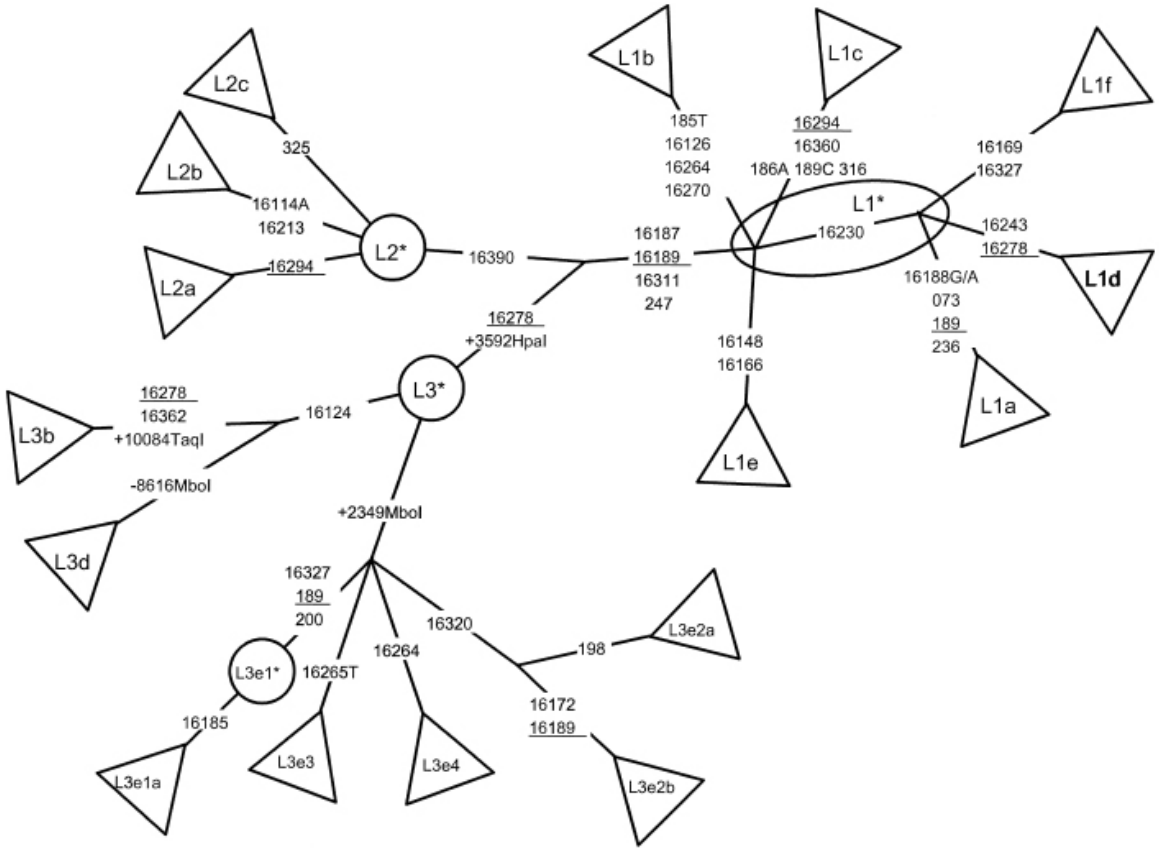
History of the geographical perception of the world...

Today



History of haplogroup characterization...

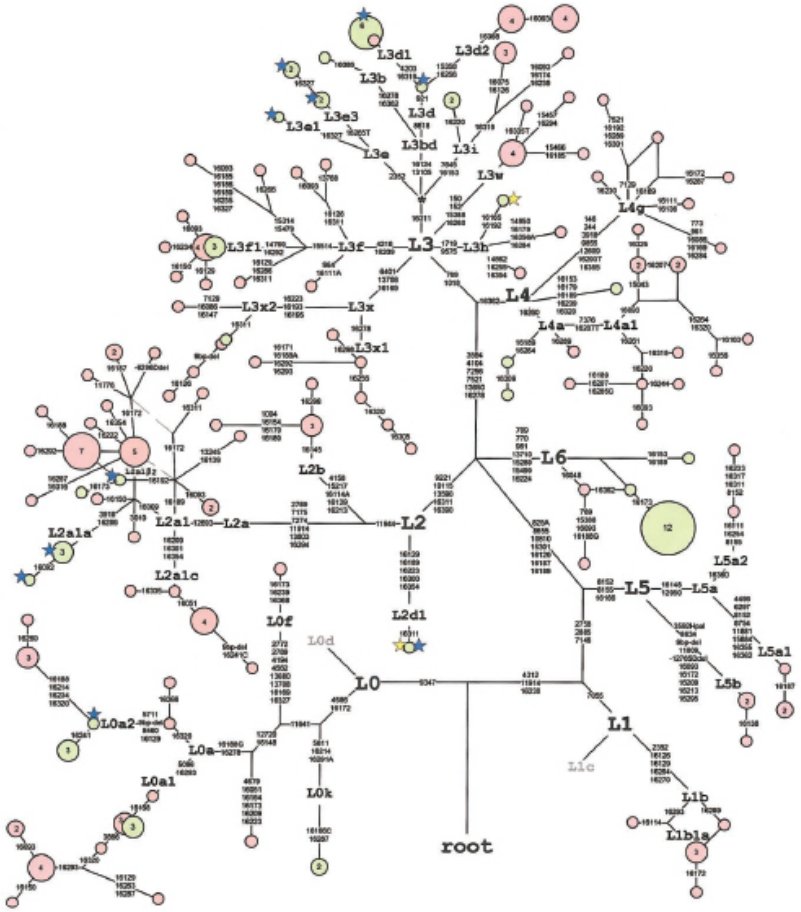
2001



Pereira et al. *AHG* 2001

History of haplogroup characterization...

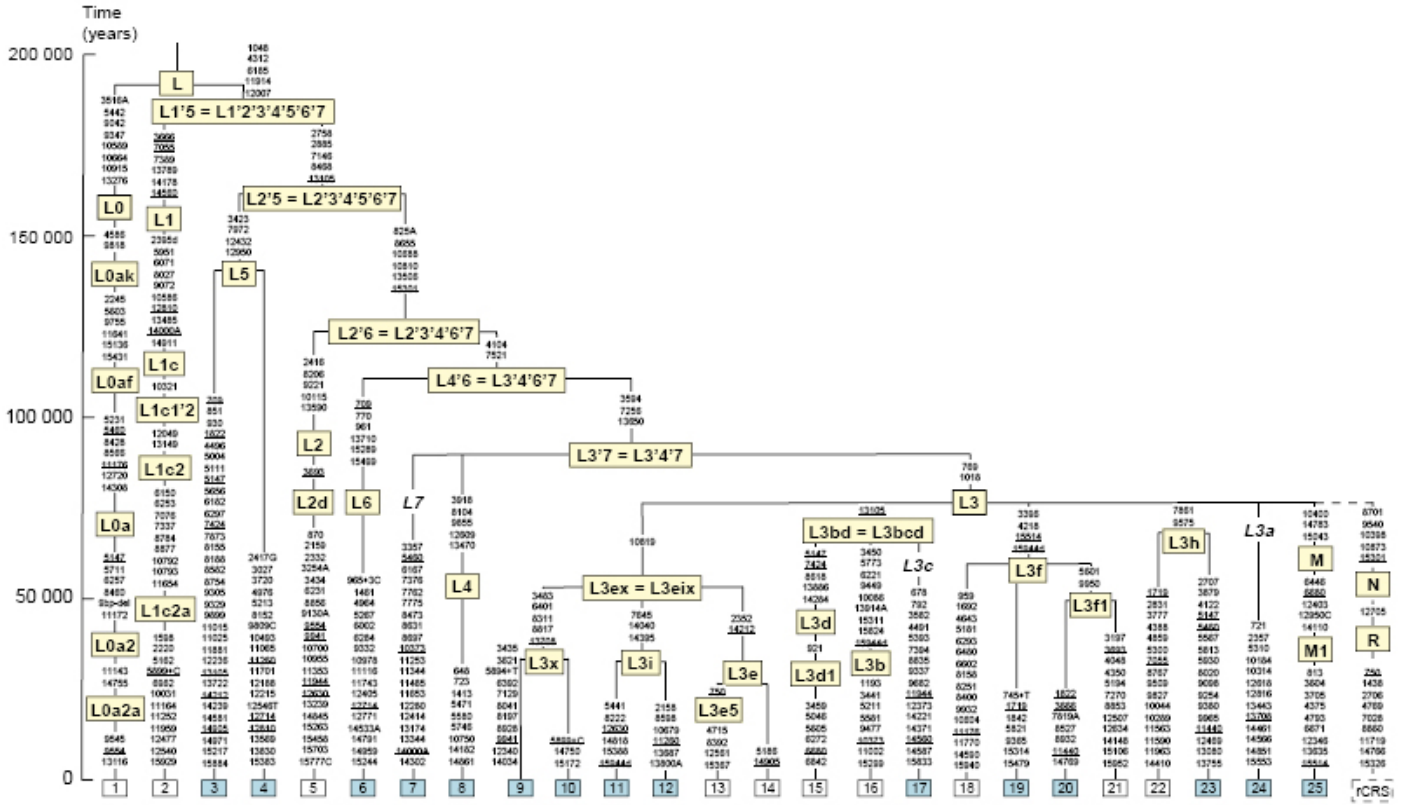
2004



Kivisild et al. *AJHG* 2004

History of haplogroup characterization...

2006



Torroni et al. *Trends Genetics* 2006

Publications for haplogroup characterization

L

Gonder *MBE* 2007
Kivisild *Genetics* 2006
Torroni *Genetics* 2006
Kivisild *AJHG* 2004
Mishmar *HumMut* 2004
Salas *AJHG* 2002
Pereira *AHG* 2001
Graven *MBE* 1995

. . .

M

Hill *AJHG* 2007
Kong *HumMolGen* 2006
Sun *MBE* 2006
Thangaraj *BMCGen* 2006
Pierson *MBE* 2006
Friedlaender *MBE* 2005
Kong *AJHG* 2003
Kivisild *MBE* 2002

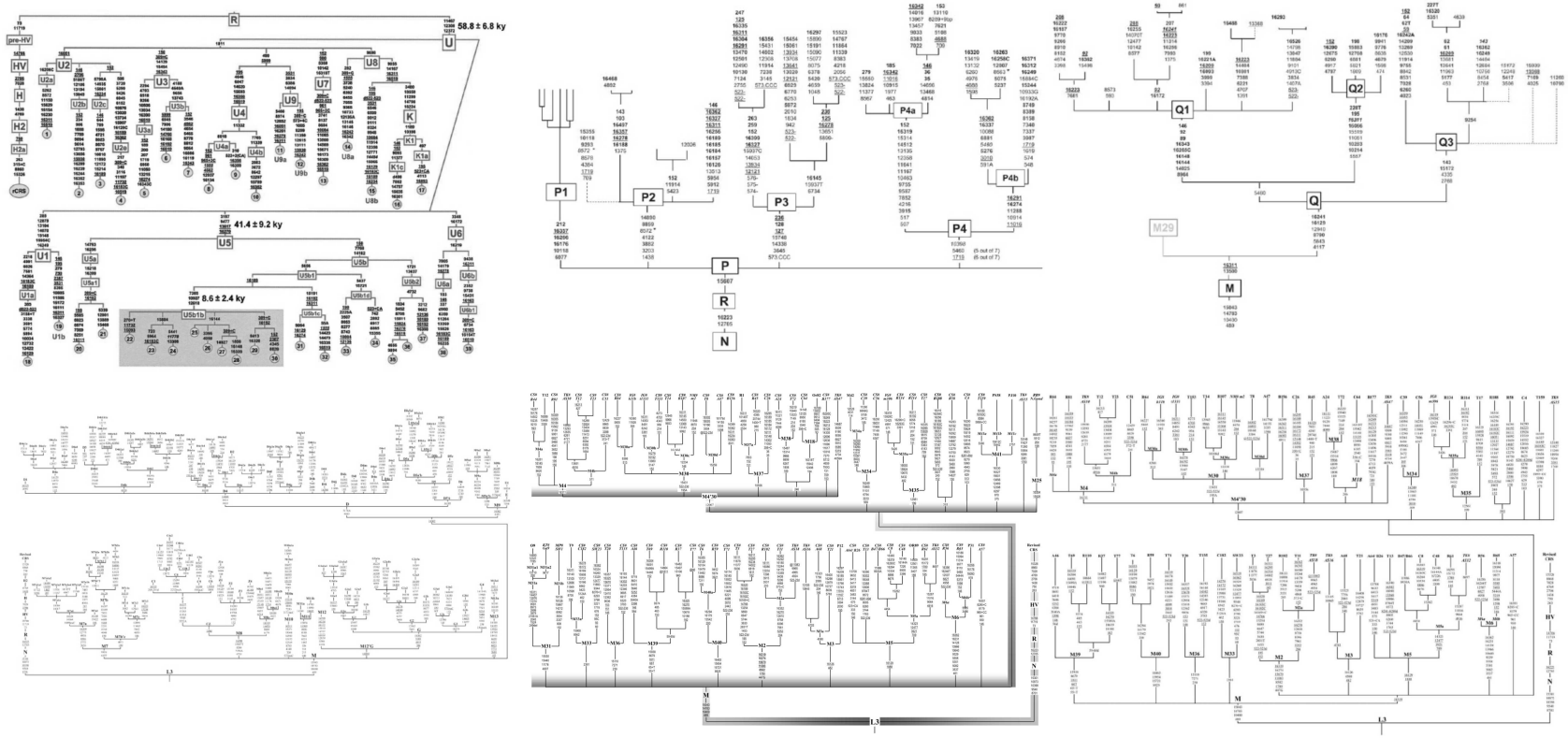
. . .

N

Roostalu *MBE* 2007
Behar *AJHG* 2006
Achilli *AJHG* 2005
Palanichamy *AJHG* 2004
Reidla *AJHG* 2003
Herrnstadt *AJHG* 2002
Finnilä *AJHG* 2001
Richards *AHG* 1998

. . .

Haplogroup definition trees



Haplogroup assignment: a demanding task

Requirements

- Profound knowledge of past and recent publications
- Background in human phylogenetics
- Good estimation of mutation rates of different polymorphisms

Haplogroup assignment: a demanding task

Determination of haplogroup affiliation

- Thorough forensic investigation includes the determination of the haplogroup affiliation of samples
- Annotation of indel positions relies on the phylogenetic background of the sample (Bandelt and Parson *IJLM* 2007)
- Potential errors can be uncovered

Phylogenetic annotation of indel positions

African haplogroup L5a1

Brandstätter et al. *IJLM* 2004:

Nai023 **16129A, 16148T, 16166G,** 16183C, 16186T, **16189C, 16223T, 16278T, 16311C, 16355T, 16362C**

Nai068 16093C, **16129A, 16148T, 16166G,** 16183DEL, 16187T, **16189C, 16223T, 16278T, 16311C, 16355T, 16362C**

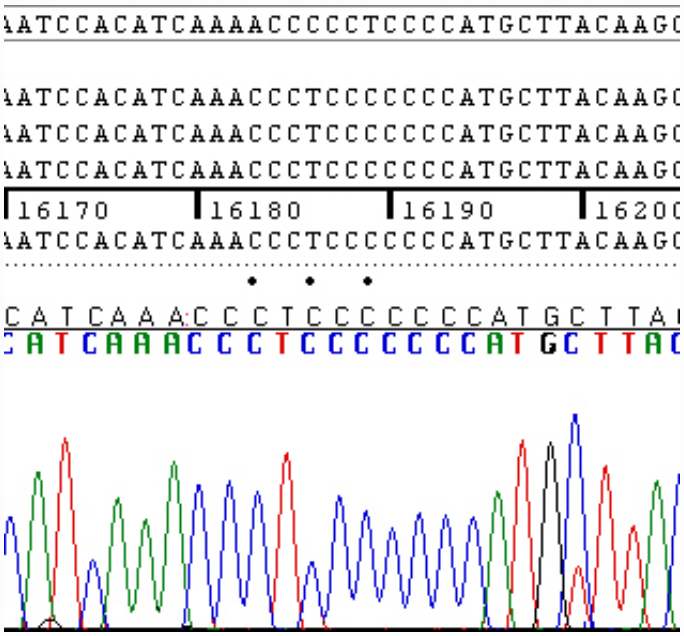
- Samples belong to L5a1
- Kivisild et al. *AJHG* 2004

Phylogenetic annotation of indel positions

Nai023: 16129A, 16148T, 16166G,. . ., 16223T, 16278T, 16311C, 16355T, 16362C

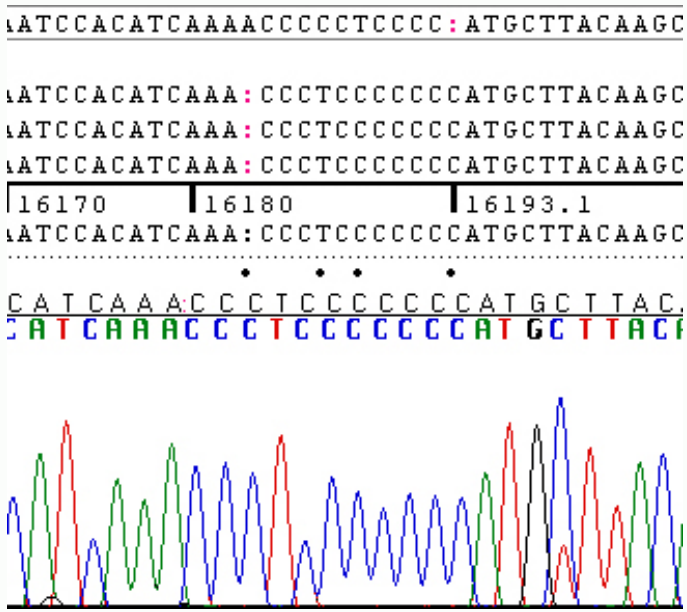
Formal alignment rules

16183C, 16186T, 16189C



Phylogenetic alignment

16183DEL, 16187T, 16189C, 16193.1C

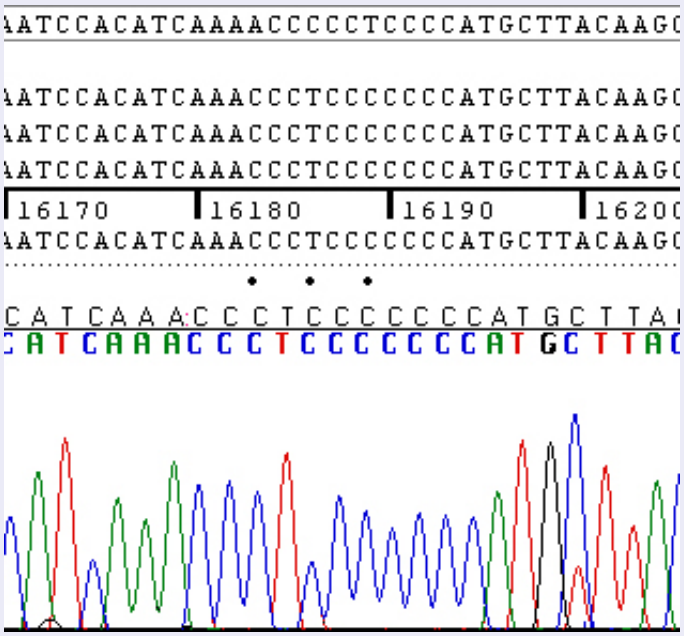


Phylogenetic annotation of indel positions

Nai023: 16129A, 16148T, 16166G,. . ., 16223T, 16278T, 16311C, 16355T, 16362C

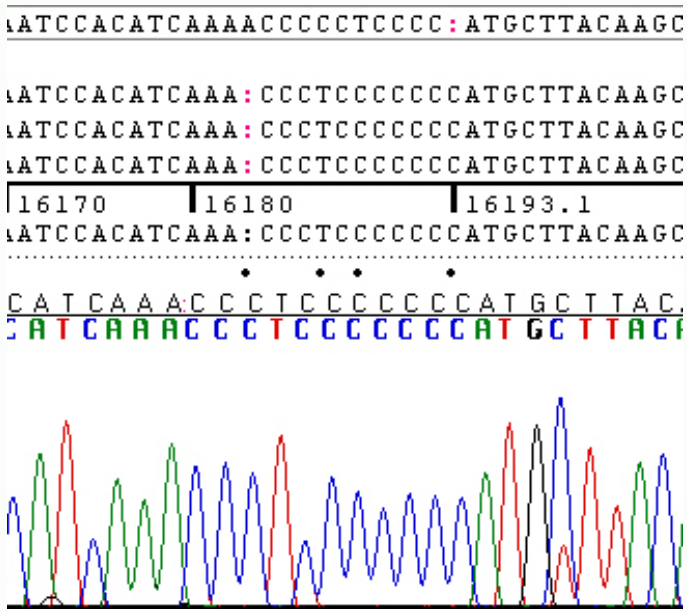
Formal alignment rules

16183C, 16186T, 16189C



Phylogenetic alignment

16183DEL, 16187T, 16189C, 16193.1C

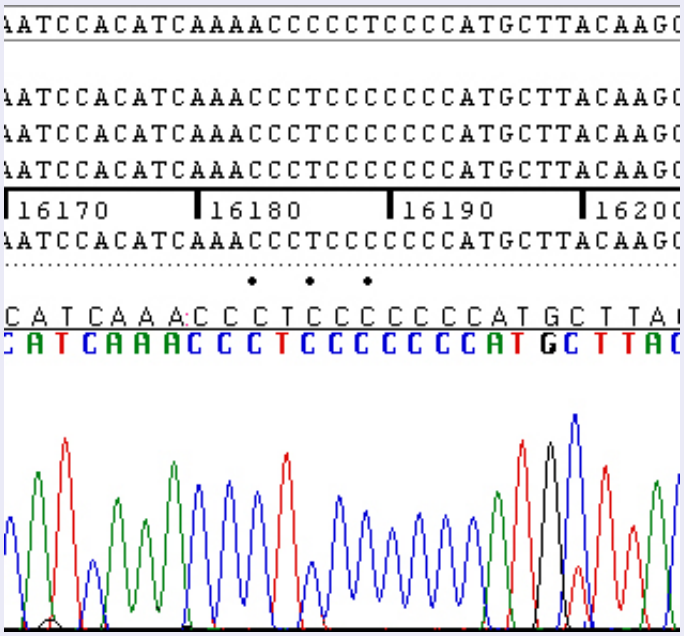


Phylogenetic annotation of indel positions

Nai023: 16129A, 16148T, 16166G,. . ., 16223T, 16278T, 16311C, 16355T, 16362C

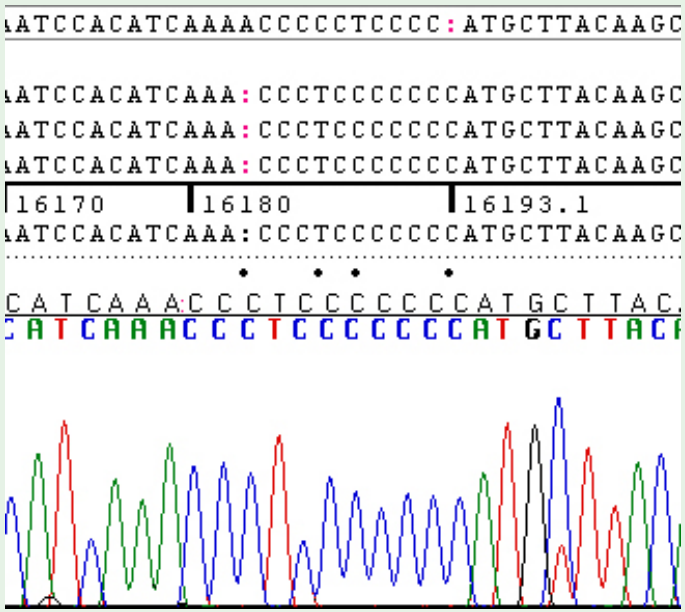
Formal alignment rules

16183C, 16186T, 16189C



Phylogenetic alignment

16183DEL, 16187T, 16189C, 16193.1C



Phylogenetic annotation of indel positions

Problem with formal alignment

Sample from Ruanda:

Ru2A5 16129A, 16148T, 16166G, 16183DEL, 16187T, 16189C, 16223T, 16278T, 16311C, 16355T, 16362C

- 1 difference to Nai068 on 16093 (hotspot)
- 3 differences to Nai023 on 16183, 16186, 16187

Phylogenetic annotation of indel positions

If Nai023 was phylogenetically aligned...

Sample from Ruanda:

Ru2A5 16129A, 16148T, 16166G, 16183DEL, 16187T, 16189C, 16223T, 16278T, 16311C, 16355T, 16362C

- 1 difference to Nai068 on 16093 (hotspot)
- 1 difference to Nai023 on 16193.1C (length heteroplasmic hotspot)

Problems with haplogroup identification

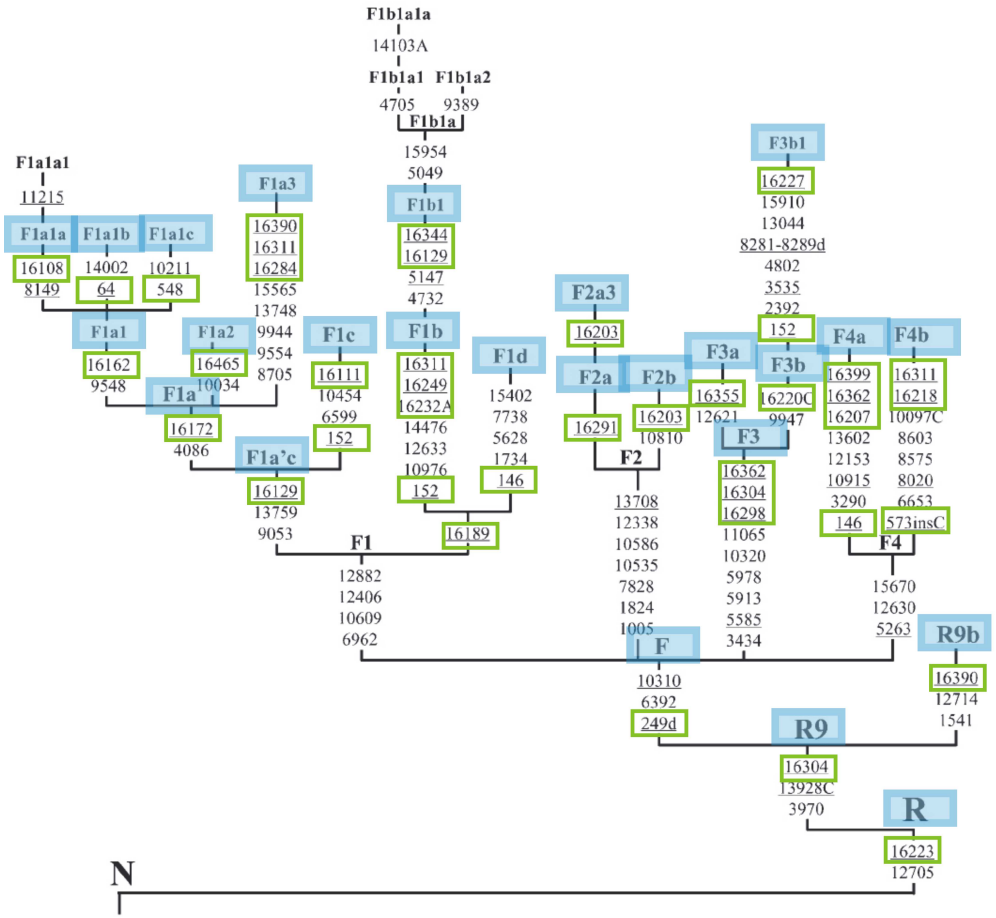
Definition of haplogroups

- Haplogroup definitions mainly rely on phylogenetically stable positions within the coding region
  - However, also there are some hotspots (e.g. 3010)
- Associations between haplogroups and polymorphisms within the control region are less stable
  - However, some polymorphisms are very stable (e.g. 497C in K1a)
- Mutational hotspots lead to homoplasy
- Most hotspots reside within HVS-I and HVS-II

Problems with haplogroup identification

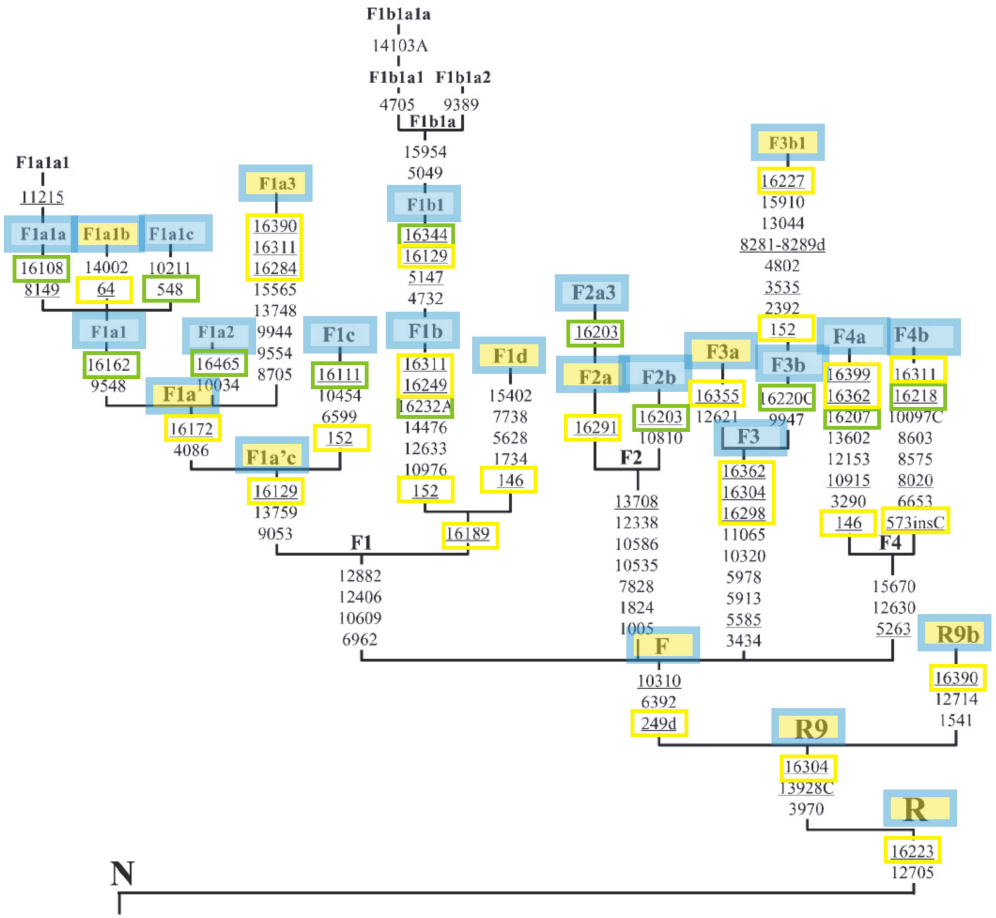
Do CR polymorphisms provide sufficient information for haplogroup determination?

Haplogroup F



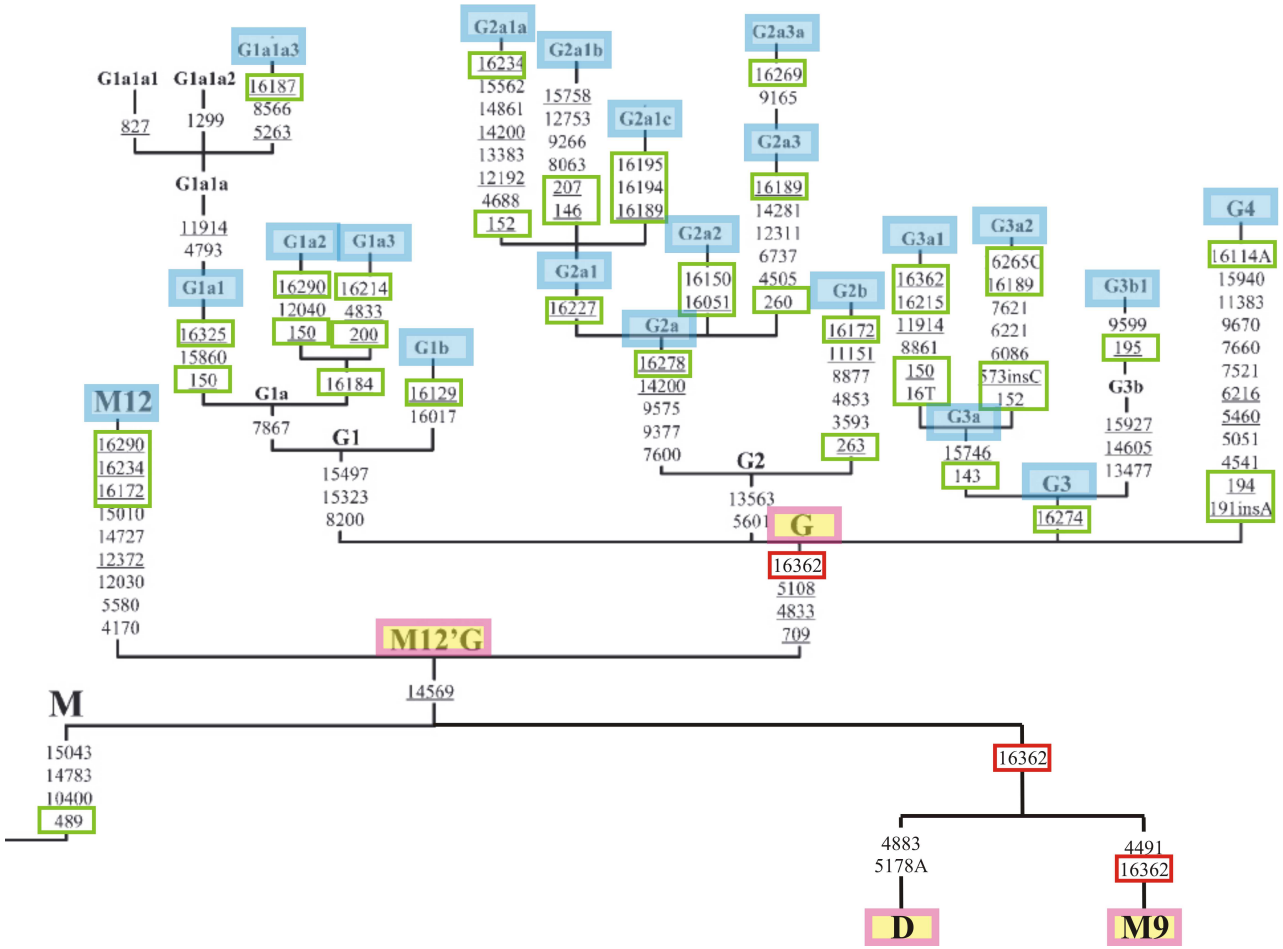
Kong et al. *Hum Mol Gen* 2006

Haplogroup F



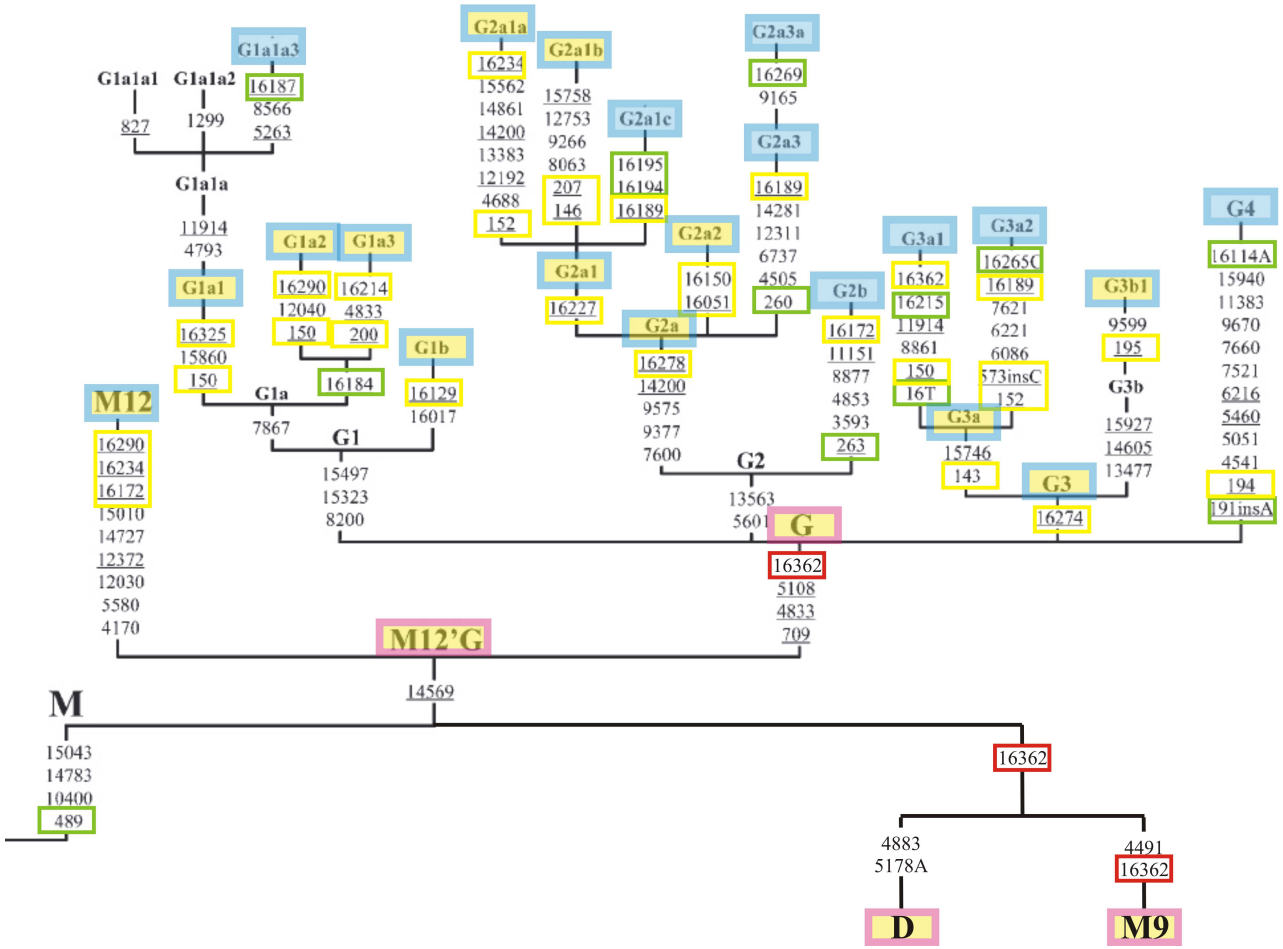
Kong et al. *Hum Mol Gen* 2006

Haplogroups M9 - M12 - D - G



Kong et al. *Hum Mol Gen* 2006 (with modifications)

Haplogroups M9 - M12 - D - G



Kong et al. *Hum Mol Gen* 2006 (with modifications)

Forensic haplogroups

Definition of haplogroups

- Only a fraction of all worldwide haplogroups can be distinguished by control region polymorphisms
- These polymorphisms need to be phylogenetically stable
- The estimation should rather be conservative than tentative

Confident haplogroup determinations can only be reached by typing selected coding region SNPs.

Forensic haplogroups

Definition of haplogroups

- Only a fraction of all worldwide haplogroups can be distinguished by control region polymorphisms
- These polymorphisms need to be phylogenetically stable
- The estimation should rather be conservative than tentative

Confident haplogroup determinations can only be reached by typing selected coding region SNPs.

Methods for haplogroup-identification

Brandstätter, Parsons, Parson.
Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups.
Int J Legal Med (2003)

Quintans, Alvarez-Iglesias, Salas, Phillips, Lareu, Carracedo.
Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing.
FSI (2004)

Lee, Yoo, Park, Chung, Kim, Shin.
East Asian mtDNA haplogroup determination in Koreans: haplogroup-level coding region SNP analysis and subhaplogroup-level control region sequence analysis.
Electrophoresis (2006)

Outline

Software solutions for haplogroup identification

Requirements

- Support haplogroup assignment of mtDNA profiles
- Handle large population samples
- Deal with different reading frames
- Incorporate all available information
- Model phylogenetic instability

First approach: Haplogroup-ID

Concept

- Haplogroups were summarized from the literature
  - Listed as virtual haplotypes
- **Linearized approach**:
  - Sum of weights of matching polymorphisms between virtual haplogroup and sample in question is calculated
  - Weights refer roughly to estimated mutation rates
  - Highest scoring haplogroups are chosen

First approach: Haplogroup-ID

Problems

- Virtual haplotypes are difficult to define
  - Are to some extend arbitrary
  - Reflect reality only to a certain extent
- Linearized approach: sum was not influenced by
  - Number of unexplained (private) polymorphisms
  - Number of missing defining polymorphisms
- Weights were estimated empirically

$\Rightarrow$ 5% of classifications were outside the correct super-haplogroups.

Second approach: Phylocheck

Concept

- Different HVS-I + HVS-II sequences with classified haplogroups were collected and stored in two datafiles:
  - One datafile containing only HVS-I sequences, the other one containing only HVS-II sequences
- By nearest neighbor search, the genetically closest sequence to a new sample was determined in HVS-I and HVS-II separately
  - Model of nucleotide substitution: $GTR + I + \Gamma$ on non-indel sites
  - Hamming distance on indel sites

Second approach: Phylocheck

Concept

- The haplogroup affiliations of 'closest' sequences were compared
- **Match**: hint for haplogroup affiliation of new sample
- **Mismatch**: hint to artificial recombination in the new sample

Second approach: Phylocheck

Problems

- Backbone datafile needs continuous care and updating
- $GTR + I + \Gamma$ is O.K. for phylogenetic reconstructions, but does not take site-specific variation of sites into account
- Some haplogroups are identical in HVS-II

$\Rightarrow$ More applicable for detection of artificial recombination than for haplogroup assignment.

New approach: Maximum likelihood

Reference database

- Correctly classified sequences were collected from the literature
- Antiquated nomenclature was updated using contemporary articles
- Haplogroup classification confirmed with coding region information in the form of either whole genome sequences or SNPs from the coding region

New approach: Maximum likelihood

Reference database

- Backbone database comprises >5000 sequences with different reading frames
- Condensation to control region and reduction of duplicates yields >3000 different haplotypes

New approach: Maximum likelihood

Advantages of the new reference database

- All of the CR-information is used in the classification decision, rather than simply counting on rule-defining sites
- Any mutations within haplogroups that have appeared in backbone samples are useful for classification

$\Rightarrow$ Phenomena such as homoplasy or back mutations in haplogroup-defining sites exert less influence on the estimation

- Other loci support correct classification

New approach: Maximum likelihood

Concept

- Initial cost of a profile is calculated as sum of weights of all its polymorphisms
- Weights refer to phylogenetic stability of each single mutation
- Weights of matching polymorphisms in haplotypes with known hg-affiliation are subtracted
- Haplogroups of haplotypes with lowest costs are chosen

Maximum likelihood: costs

Weights of polymorphisms

- Weights reflect phylogenetic stability of mutations
- Mutational hotspots have low phylogenetic stability
- Low reliability $\Rightarrow$ low weights
- High stability of mutations $\Rightarrow$ high weights
- Floating-point numbers $\in [0, \infty]$

Maximum likelihood: costs

Estimation of weights

- Weights reflect mutability or stability of certain positions in certain haplogroups
- Polymorphisms with high frequency (>95%) in certain haplogroups and low frequency in other haplogroups (<5%) are assigned a weight near **1**
- Polymorphisms with an average frequency (30-70%) in many different haplogroups are assigned a weight near **0**
- Hypervariable polymorphisms, i.e. length heteroplasmic C-insertions are assigned weight **0**

Weights: Example

Estimate from a collection of 5223 entire CR profiles

T16217C

- B4: 97%
- HV2: 100%
- rest: < 5%

$\Rightarrow$ weight near **1**

A16343G

- U3: 100%
- rest: < 5%

$\Rightarrow$ weight near **1**

T152C

- A: 58%
- B4: 15%
- L0a: 50%
- L2b: 100%
- M7b: 15%

. . .

$\Rightarrow$ weight near **0**

Haplogroup assignment process: Example

Profile and costs

- Profile: 73G(1.00); 152C(0.33); 263G(1.00); 295T(1.00); 315.1C(0.00); 462T(1.00); T489C(1.00); 16069T(1.00); 16126C(0.33); 16193T(0.33); 16519C(0.33)
- Analyzed frame: 1-576 16024-16569
- Costs of profile = **7.33**

Haplogroup assignment process: Example

Results

Feasible haplotype with lowest cost **1.00** and minimum number **3** of differences:

- Palanichamy AJHG04 (**J1**)
- Observed mutations: 73G(-1.00); 263G(-1.00); 295T(-1.00); 315.1C(0.00); 462T(-1.00); 489C(-1.00); 16069T(-1.00); 16126C(-0.33)
- Missing mutations: none
- Extra mutations: 152C(0.33); 16193T(0.33); 16519C(0.33)
- Differences: **3** mutations in **1162** positions

ML Estimator for Independent Positions

Likelihood function

$$L_x(h) = \prod_i p_i(a_i \to a_i) \cdot \prod_{i \in \bar{x}} \frac{p_i(a_i \to x_i)}{p_i(a_i \to a_i)} \cdot \prod_{i \in \bar{h} \cap \bar{x}} \frac{p_i(h_i \to x_i)}{p_i(a_i \to x_i)} \cdot \prod_{i \in \bar{h} \setminus \bar{x}} \frac{p_i(h_i \to a_i)}{p_i(a_i \to a_i)}$$

where

- x is the given profile,
- h is the haplotype of the haplogroup,
- a is the haplotype of the rCRS,
- $\bar{h}$ and $\bar{x}$ denote the sets of positions of h or x different to the rCRS,
- i runs over the specified positions, and
- $p_i(h_i \to x_i), \ldots, p_i(h_i \to a_i)$ are the transition probabilities.

ML Estimator for Independent Positions

To maximize the likelihood and scale according to the reading frame, we minimize the

Cost function

$$
\begin{aligned}
C_x(h) &= \log\left(\frac{\prod_i p_i(a_i \to a_i)}{L_x(h)}\right) \\
&= \sum_{i\in\bar{x}} \log\left(\frac{p_i(a_i \to a_i)}{p_i(a_i \to x_i)}\right) + \sum_{i\in\bar{h}\cap\bar{x}} \log\left(\frac{p_i(a_i \to x_i)}{p_i(h_i \to x_i)}\right) \\
&+ \sum_{i\in\bar{h}\setminus\bar{x}} \log\left(\frac{p_i(a_i \to a_i)}{p_i(h_i \to a_i)}\right).
\end{aligned}
$$

ML Estimator for Independent Positions

Scaling

Within the cost function, the value of the logarithm is scaled in a way that non-speedy transitions are assigned weight **1**.

ML Estimator for Independent Positions

For a perfect match $h = x$ the

Total costs

$$C_x(h) = \sum_{i \in \bar{x}} \log \left(\frac{p_i(a_i \to a_i)}{p_i(h_i \to h_i)} \right)$$

approximate 0.

Problems with the ML approach

Reading frames

- Haplotypes from the literature were typed for a variety of different reading frames
  - Entire genomes
  - HVS-I + HVS-II + SNPs from the coding region
  - HVS-I + SNPs from the coding region
- Definition of standard reading frame for haplogroup estimation:
  - 16024-16181 16184-16193 16194-16518 16520-16569 1-309 310-315 316-522 525-573.1 574-576
- Problem: partially overlapping profiles

Problems with the ML approach

Solution

Only polymorphisms in overlapping regions between the profile in question and haplotypes with known hg-affiliation are taken into consideration for the minimization of the cost function

Problems with the ML approach

Weights of mutations

- Sound estimation of the reliability of different mutations, which correspond to the transition probabilities in the ML-model, is crucial for the performance of the estimator
- Problem:
  - Representative sample of entire CR population data from all different haplogroups
  - Idea how to model the reliability of different mutations

Problems with the ML approach

Solution

- Compilation of >5000 CR profiles from worldwide populations
- Randomized population samples
- When possible, full CR + coding region SNPs
- Estimation of reliability by determining the frequency of every mutation in the CR in every (sub-)haplogroup

Problems with the ML approach

Independency of mutations

- Our model requires independency of the reliability of the positions from each other and from the phylogenetic background.
- Only then, weights can be summarized after logarithmizing.
- Problem: Do mutations occur independently from each other?

Problems with the ML approach

Solution

- Compilation of >5000 CR profiles from worldwide populations
- Inference of all partitions, which are based on 2 or more polymorphisms
- 'Linked' polymorphisms, such as the AC-repeat around 523-524 in HVS-III or the Chibcha-deletion in HVS-II (bp 106-111) are treated as one single locus.

Outline

Conclusions

Problems with control region typing

- CR polymorphisms have a limited reliability for haplogroup determination
- Mutational hotspots vs. diagnostic sites
- Modeling this uncertainty is a difficult task

Reliability of mutations

- Estimating the stability of markers in haplogroups as inferred from their frequencies in different population samples yields a good approximation of the reliability

Conclusions

Problems with control region typing

- CR polymorphisms have a limited reliability for haplogroup determination
- Mutational hotspots vs. diagnostic sites
- Modeling this uncertainty is a difficult task

Reliability of mutations

- Estimating the stability of markers in haplogroups as inferred from their frequencies in different population samples yields a good approximation of the reliability

Conclusions

Maximum likelihood

- Estimation relies on real mtDNA control region profiles with haplogroup-defining coding region SNPs
- Maximum likelihood approach yields better results than linear approach
- Taking reading frames into consideration improved the estimation considerably
- Site-specific reliability values further enhanced the accuracy of the haplogroup appraisal

Thank you very much for your attention!