

# “Errors wanted for demonstration” – Quasi-mediane Netzwerke und Biotools in EMPOP zur Identifikation von Fehlern in mtDNA Daten am Beispiel GEDNAP 38 und 39

**31. Spurenworkshop**  
**Hamburg, 25.-26. Februar 2011**

*Bettina Zimmermann<sup>1</sup>, Alexander Röck<sup>1,2</sup>, Carsten Hohoff<sup>3</sup>, Walther Parson<sup>1</sup>*

1 Institut für Gerichtliche Medizin, Medizinische Universität Innsbruck, Österreich

2 Institut für Mathematik, Leopold Franzens Universität Innsbruck, Österreich

3 Institut für Forensische Genetik, Münster, Deutschland

# *A posteriori* Qualitätskontrolle von mtDNA Daten

Quasi-medianes (QM) Netzwerk

- grafische Darstellung einer mtDNA Datentabelle
- Erkennen potenzieller Fehler und Artefakte im Datensatz

QM Netzwerk-Software (*Parson und Dür 2007, Zimmermann et al. 2011*)

- [www.empop.org](http://www.empop.org)

Die Komplexität eines QM Netzwerks beruht (nach Filterung) auf

- der **Qualität** der untersuchten Proben, wenn
- die Probenanzahl nicht zu hoch ist ( $< 300$ ) und
- die Haplotypen die selbe phylogenetische Herkunft teilen

# Beispiel: Westeurasischer Etalon

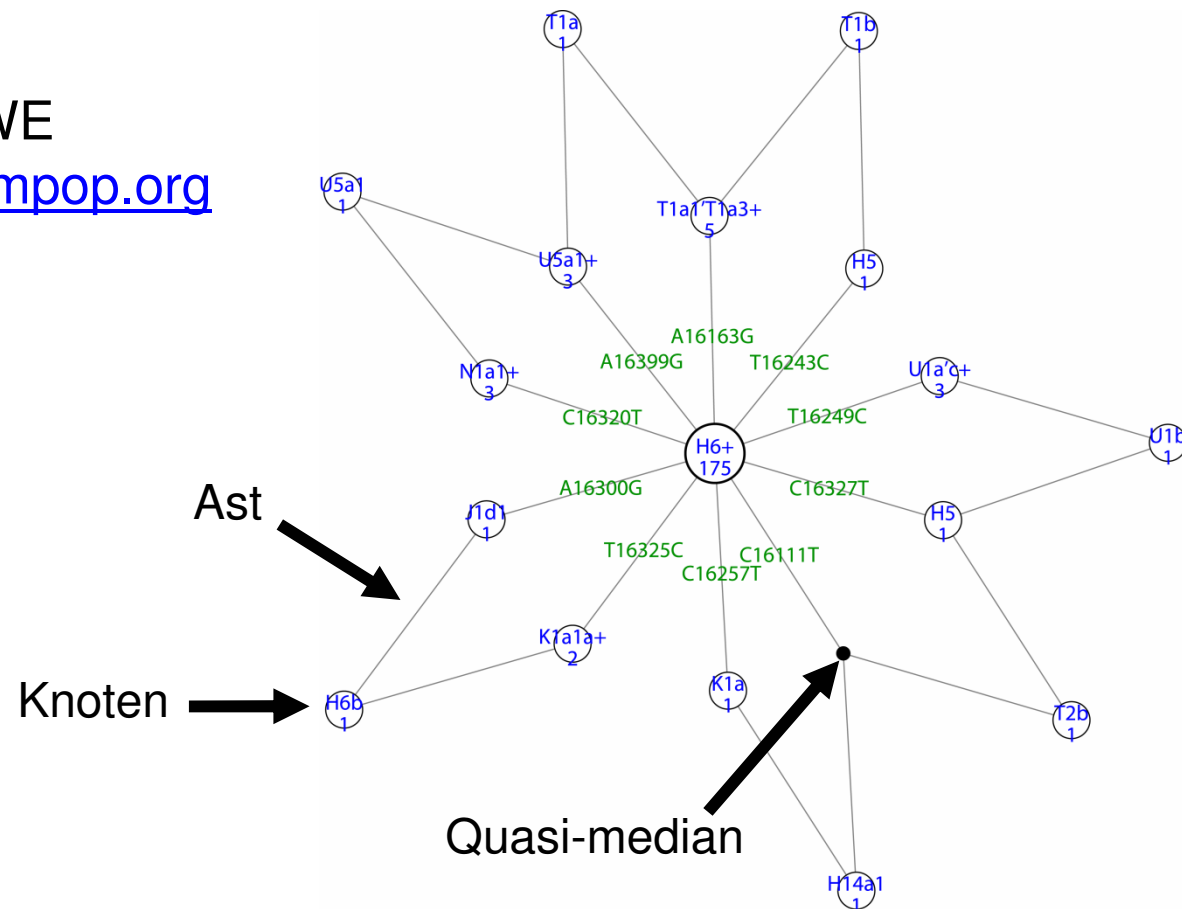
QM Netzwerk (Torso)

HVS-I: 16024-16569

Filter: EMPOPspeedyWE

WE-Etalon aus [www.empop.org](http://www.empop.org)

N=202



Einzelne/wenige Haplotypen können mit dem Etalon verknüpft werden

QM Netzwerk (Torso)  
HVS-I: 16024-16569  
Filter: EMPOPspeedyWE

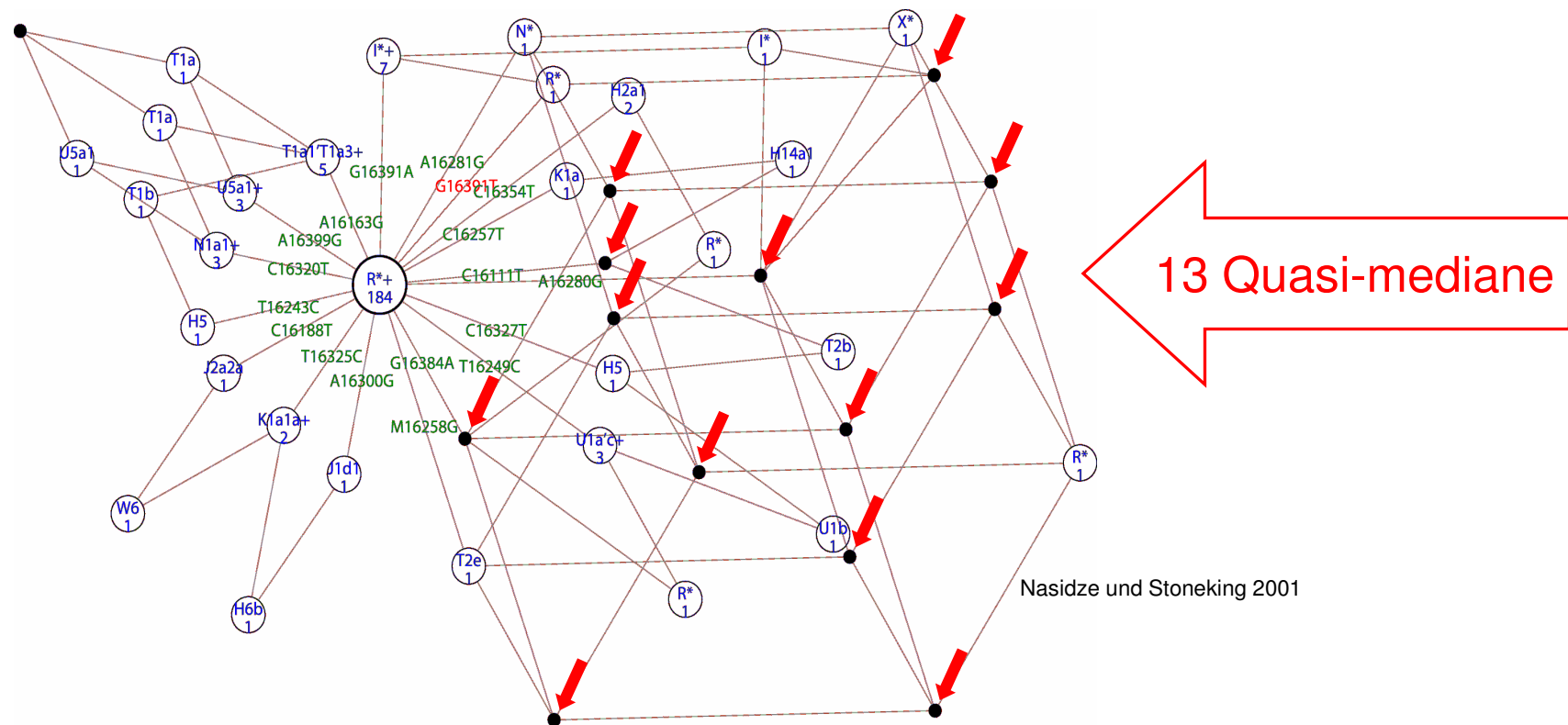


# Fragwürdige Sequenzdaten (N=29) + Etalon (N=202)

QM Netzwerk (Torso)

HVS-I: 16024-16400

Filter: EMPOPspeedyWE

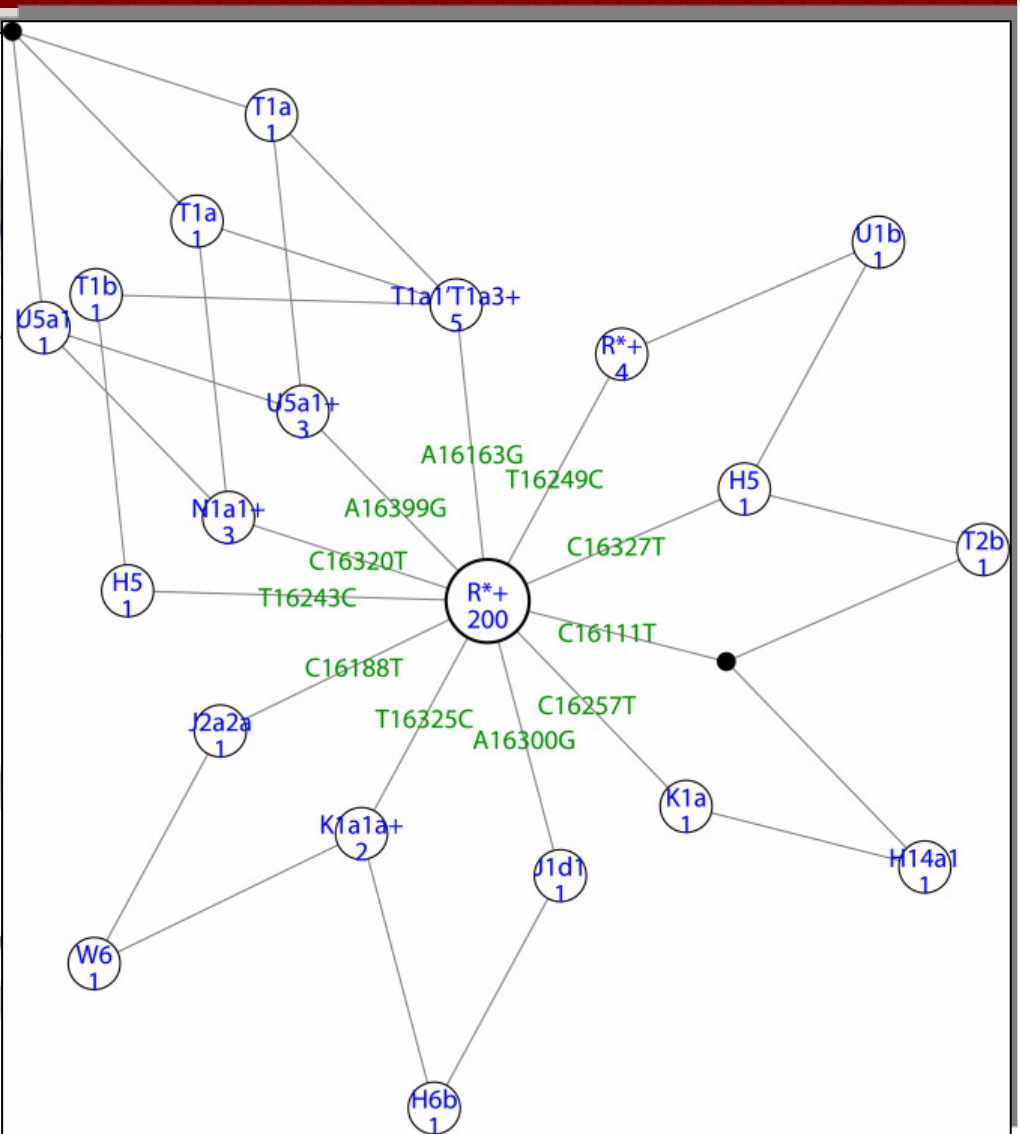


# Korrekturmaßnahmen

	A	B	C	D	E
1	# Nasidze & Stoneking PRSLB 2001 Darg				
2	# Mitochondrial DNA variation and language				
3	# 29 samples				
4	#! 16024-16400				
5	DAR04_Darginian	R0	1	16093C	
6	DAR012_Darginian	R0	1	16189C	
7	DAR16_Darginian	R0	1	16189C	
8	DAR49_Darginian	R0	1	16093C	
9	DAR014_Darginian	J*	1	16069T	1612
10	DAR42_Darginian	J*	1	16126C	1619
11	DAR06_Darginian	J*	1	16126C	1621
12	DAR31_Darginian	K*	1	16093C	1622
13	DAR03_Darginian	T1a	1	16126C	1616
14	DAR013_Darginian	U4	1	16261T	1635
15	DAR02_Darginian	U5	1	16189C	1619
16	DAR3_Darginian	U5	1	16256T	1627
17	DAR011_Darginian	U5	1	16256T	1627
18	DAR37_Darginian	U5	1	16256T	1627
19	DAR05_Darginian	N*	1	16129A	1622
20	DAR015_Darginian	R*	1	16234T	1638
21	DAR016_Darginian	R*	1	16150T	1635
22	DAR07_Darginian	R*	1	16189C	1624
23	DAR35_Darginian	R*	1	16294T	1639
24	DAR36_Darginian	R*	1	16258G	1628
25	DAR01_Darginian	R*	1	16292T	
26	DAR08_Darginian	R*	1	16223T	
27	DAR010_Darginian	R*	1	16292T	
28	DAR51_Darginian	R*	1	16234T	
29	DAR14_Darginian	I*	1	16114T	1612
30	DAR09_Darginian	I*	1	16129A	1622
31	DAR018_Darginian	I*	1	16129A	1622
32	DAR23_Darginian	X*	1	16129A	1622
33	DAR46_Darginian	W6	1	16188T	1619



16279 C->G,



HVS-I Torso:16024-16569  
Filter: EMPOPspeedyWE  
WE-Etalon + 1 afrikanischer Ht  
N=202 + 1



# Qualitätskontrolle von GEDNAP - Daten



## Ringversuch GEDNAP (GN) 38 mtDNA

Leserahmen: 16024-576

Person 1 R0 16519C 263G 315.1C

Person 2 R0 195C 263G 315.1C 523del 524del

Person 3 J2a1a1 16069T 16126C 16145A 16231C 73G 150T 152C 195C 215G 263G 295T 310.1T 315.1C 319C 489C 513A

Spur A HV0 16298C 72C 121A 263G 309.1C 309.2C 315.1C 466C 508G

Spur D R\* 16519C 73G 150T 198T 260A 263G 309.1C 309.2C 315.1C



## Ringversuch GEDNAP (GN) 39 mtDNA

Leserahmen: 16024-576

Person 1 K1a 16048A 16129A 16224C 16311C 16519C 73G 189G 263G 309.1C 315.1C 497T

Person 2 HV0 16298C 72C 195C 263G 309.1C 315.1C

Person 3 H1c 16271C 16519C 195C 257G 263G 309.1C 309.2C 315.1C 477C

Spur A U2e 16051G 16092C 16129C 16182C 16183C 16189C 16362A 16519C 73G 152C 217C 263G 315.1C 508G

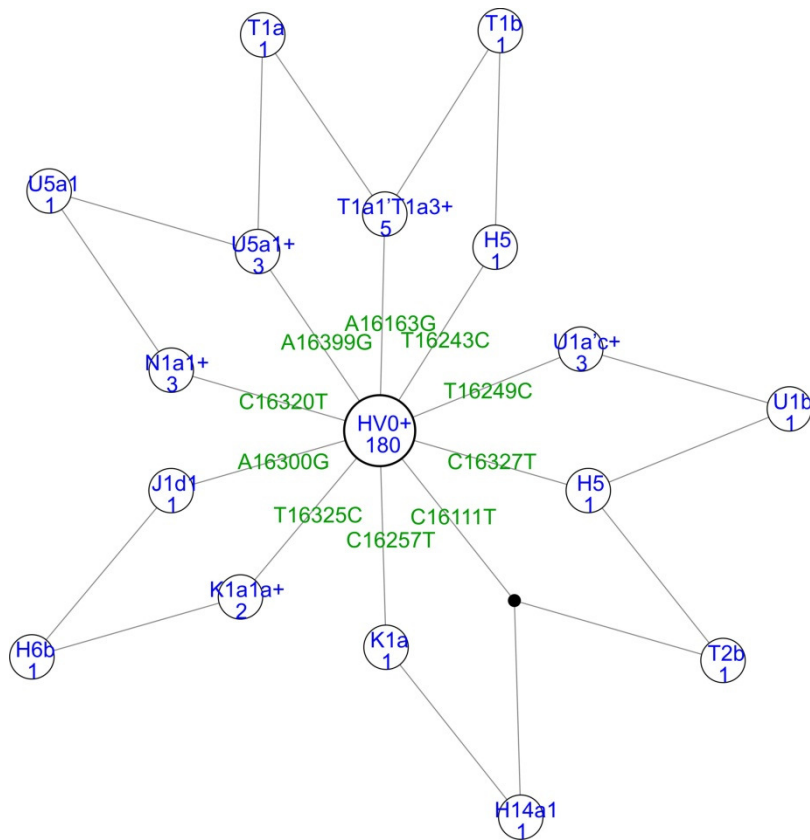
Spur B U4 16365C 73G 146C 195C 228A 263G 315.1C 499A

Spur C R0 16519C 152T 263G 309.1C 309.2C 315.1C 524.1A 524.2C

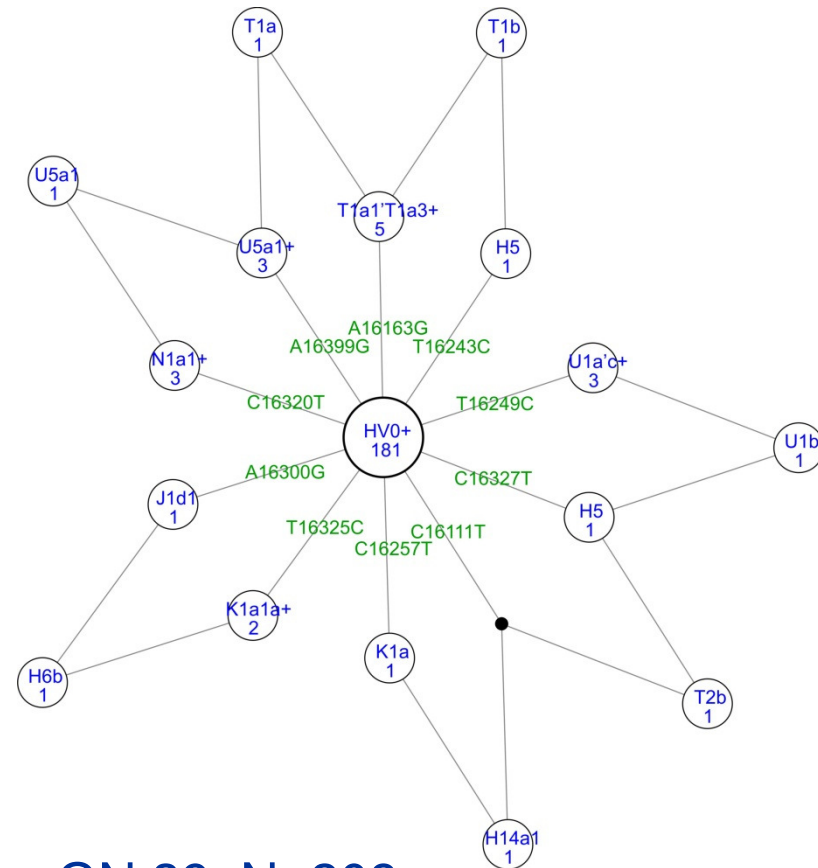


# Korrekte Typisierung – unauffällige QM-Netzwerke

In Kombination mit dem westeurasischen Etalon Datensatz (N=202)



GN 38; N=207



GN 39; N=208

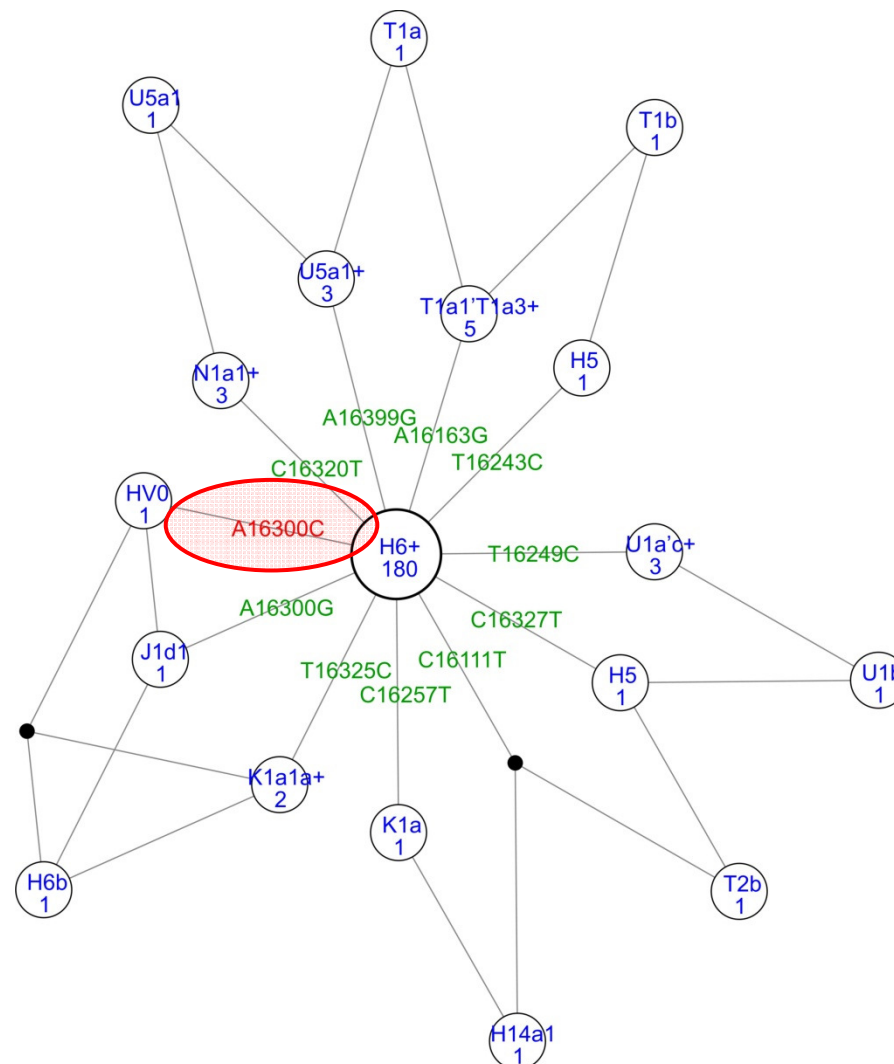
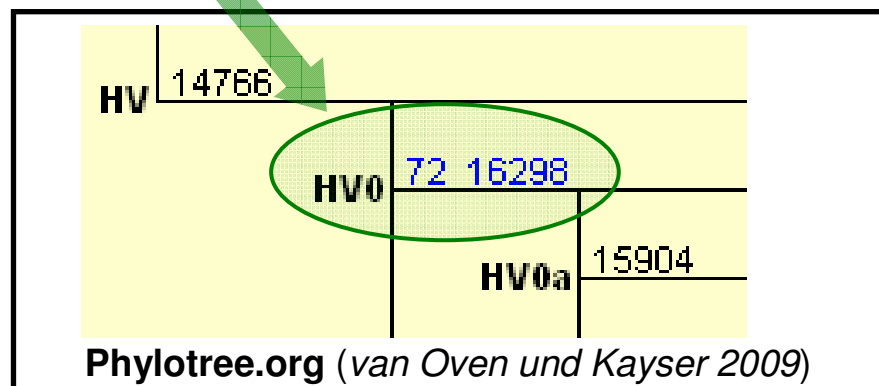
# Falschtypisierung – zusätzlicher QM



GN 39 (6 Proben) + Etalon

Person 2, Labor XY: 16039-16354 51-426  
**16300C** 72C 195C 263G 309.1C 315.1C

Person 2, korrekt: 16024-576  
 16298C 72C 195C 263G 309.1C 315.1C



# *A posteriori* Qualitätskontrolle von GEDNAP Daten

Home :: EMPOP.org - Mitochondrial DNA Control Region Database - Windows Internet Explorer

http://empop.org/

Version: 2.1, Release: 3

Welcome bettina! Logout

**EMPOP**

Home  
Contribute  
Help  
Imprint  
Contact  
Terms of use

► Query  
► Tools  
► Account

**News** Introduction Concept Alignment Users Collaborators

**Release 3 is online (2010-12-27)**  
We are happy to announce EMPOP release 3. A total of 1277 haplotypes from Brazil, China, Germany, Guinea-Bissau, Pakistan, Portugal, and Spain have been added to the database. A major portion of the sequences were supplied through the successful collaboration with the [Spanish and Portuguese Speaking Working Group of the ISFG](#) motivated by the untiring activity of [Lourdes Prieto Solla](#). Haplotypes of previous releases have been evaluated in the course of establishing R3, no changes were necessary.

**Update on search engine to Version 2.1 (2010-08-30)**  
After the launch of EMPOP 2 in April 2010 we are happy to provide an update with a faster query engine that significantly reduces search times. [...](#)

**Change of polymorphisms (2010-08-30)**  
Thorough review and feedback by the authors brought 4 changes in the HV2 C-stretch: [...](#)

**Welcome to EMPOP 2 (2010-04-16)**  
We are happy to announce that the 2<sup>nd</sup> release of EMPOP is now online, along with some new features and tools! [...](#)

**Information for EMPOP 1 users - registration for EMPOP 2 (2010-04-16)**  
Changes from EMPOP 1 to EMPOP 2 make it necessary to newly register for EMPOP 2. This also accounts for users that are already registered for EMPOP 1. Follow this [link](#) to register for EMPOP 2.

**Release history (2010-04-16)**  
Launch of EMPOP release 2 comes along with a thorough check of all data of release 1. The following changes were introduced. [...](#)

EMPOP is a collaborative project. Its quality increases with your comments and suggestions. Please [contact us](#).

[Show all news](#)


endorsed by  
**ISFG**  
© 1999-2011  
**UMI**

Fertig Internet 100%

# *A posteriori* Qualitätskontrolle von GEDNAP Daten

[www.empop.org/modules/downloads/](http://www.empop.org/modules/downloads/)

Version: 2.1, Release: 3 Welcome bettina! [Logout](#)

► 

► Query

▼ **Tools**

- Network
- EMP Tool
- Statistics
- Downloads**

► Account

## Downloads

### Network

File	Info
<a href="#">DNWsetup.zip</a>	Tool for drawing quasi-median networks (dnw.exe)
<a href="#">AUT273_spec.emp</a>	emp example file
<a href="#">Etalon_WE.emp</a>	Etalon dataset for Westeurasia


### EMP format

File	Info
<a href="#">EMPOPtemplate.zip</a>	Template and example for emp file
<a href="#">EMPOPformatdescription.pdf</a>	Description of emp file format


### Documents

File	Info
<a href="#">Parson_Berlin2010.pdf</a>	Oral presentation within Haploid DNA markers in forensic genetics, Berlin 2010; Title: EMPOP 2
<a href="#">Roeck_Berlin2010.pdf</a>	Oral presentation within Haploid DNA markers in forensic genetics, Berlin 2010; Title: Never miss a profile - String-based search using EMPOP 2 and application to phylogenetic alignment

endorsed by



© 1999-2011



# *A posteriori* Qualitätskontrolle von GEDNAP Daten

[www.empop.org/modules/network/](http://www.empop.org/modules/network/)

The screenshot shows the EMPOP web interface. At the top, it says 'Version: 2.1, Release: 3' and 'Welcome bettina! Logout'. On the left is a navigation menu with 'Query', 'Tools' (highlighted with a red circle), and 'Account'. Under 'Tools' are 'Network', 'EMP Tool', 'Statistics', and 'Downloads'. The main area is titled 'Input' and contains four rows of input fields: 'Sample Info' with a text box; 'Input File' with a text box and a 'Durchsuchen...' button; 'Filter' with a dropdown menu set to 'EMPOspeedy' and a small icon; and 'Range' with a text box. A 'PROCEED' button is at the bottom right.

- Detaillierte Anleitung zur QM Netzwerkanalyse unter *help*
- Spezifische Anleitung für **GEDNAP** Proben ist derzeit in Ausarbeitung und wird in **EMPOP** zur Verfügung gestellt

# Beispiel – Formatfehler, unvollständige Angaben



GN 38 Person 3 Labor XX

16024-16365: 16069\_ 16126C 16145A 16231C

73-340: 73G 150T 152C 195C 215G 263G 295T 310.1T 315.1C 319C

Version: 2.1, Release: 3 Welcome bettina! Logout

EMPQ

Query

Tools

Network

EMP To

Statist

Downlo

Account

Input

Input File:

- object G38\_P3: invalid modification in column 4: 16069

Input File:

- object G38\_P3: invalid modification in column 4: 16069

PROCEED

# Beispiel – Formatfehler, falsche Positionsangabe



GN 38 Person 3 Labor XY

16039-16354:      1607T 16126C 16145A 16231C

51-426:            73G 150T 152C 195C 215G 263G 295T 310.1T 315.1C 319C

Version: 2.1, Release: 3 Welcome bettina! [Logout](#)

EMPOP

Query

Tools

- Network
- EMP Tool
- Statistics
- Download

Account

Input

Input File:

- object G38\_P3: rCRS symbol in column 4: 1607T

Input File:

- object G38\_P3: rCRS symbol in column 4: 1607T

# Beispiel – Verletzung des Leserahmens



GN 38 Person 3 Labor XZ

16039-16354: 16069T 16126C 16145A 16231C

51-426: 73G 150T 152C 195C 215G 263G 295T 310.1T 315.1C 319C 541T

Version: 2.1, Release: 3 Welcome bettina! Logout

EMPOP

Query

Tools

Net

EM

Sta

Dov

Acc

Input

Input File:

- object G38\_P3: modification 541T outside reading frame

Input File:

- object G38\_P3: modification 541T outside reading frame

PROCEED



# Beispiel – Alignierung von Insertionen



GN 38 Person 3 HVS-II C-Stretch

**korrekte Alignierung** ...<sup>310</sup>T<sup>315</sup>.1T<sup>315</sup>.1C.  
(Bär et al. 2000, Bandelt und Parson 2008)

zusätzlich

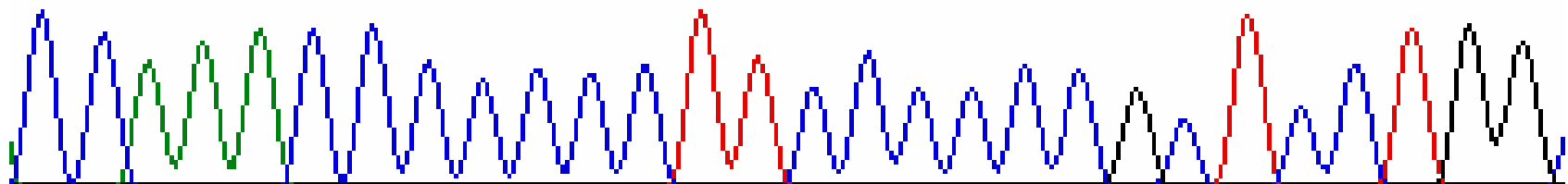
C C A A A C C C C C C C C T T C C C C C C  
C C A A A C C C C C C C C T T C C C C C C G C T C C T G G I

Phylotree.org

(van Oven und Kayser 2009)

J2a1a1

310.1T 1850



# Beispiel – Alignierung von Insertionen



GN 38 Person 3 HVS-II C-Stretch Labor YY

falsche Alignierung: ... **309.1T** 315.1C 319C ...

CCAAACCCCCC : TCCCCC : GCTTCTGGI

## Input File:

- object G38\_P3: insertion T at position 309 not known to EMPOP
- object G38\_P3: irregular insertion T 309.1 before match 310 T

Verletzung der 3'-Insertions-Regel nach *Bär et al. 2000*

# Beispiel – Alignierung von Insertionen



GN 38 Person 3 HVS-II C-Stretch Labor YX

falsche Alignierung: ... 310.1T **311.1C** 319C ...

CCAAACCCCCCT:C:CCCCGCTTCTGGI

## Input File:

- object G38\_P3: insertion C at position 311 never observed in EMPOP yet
- object G38\_P3: irregular insertion C 311.1 before match 312 C



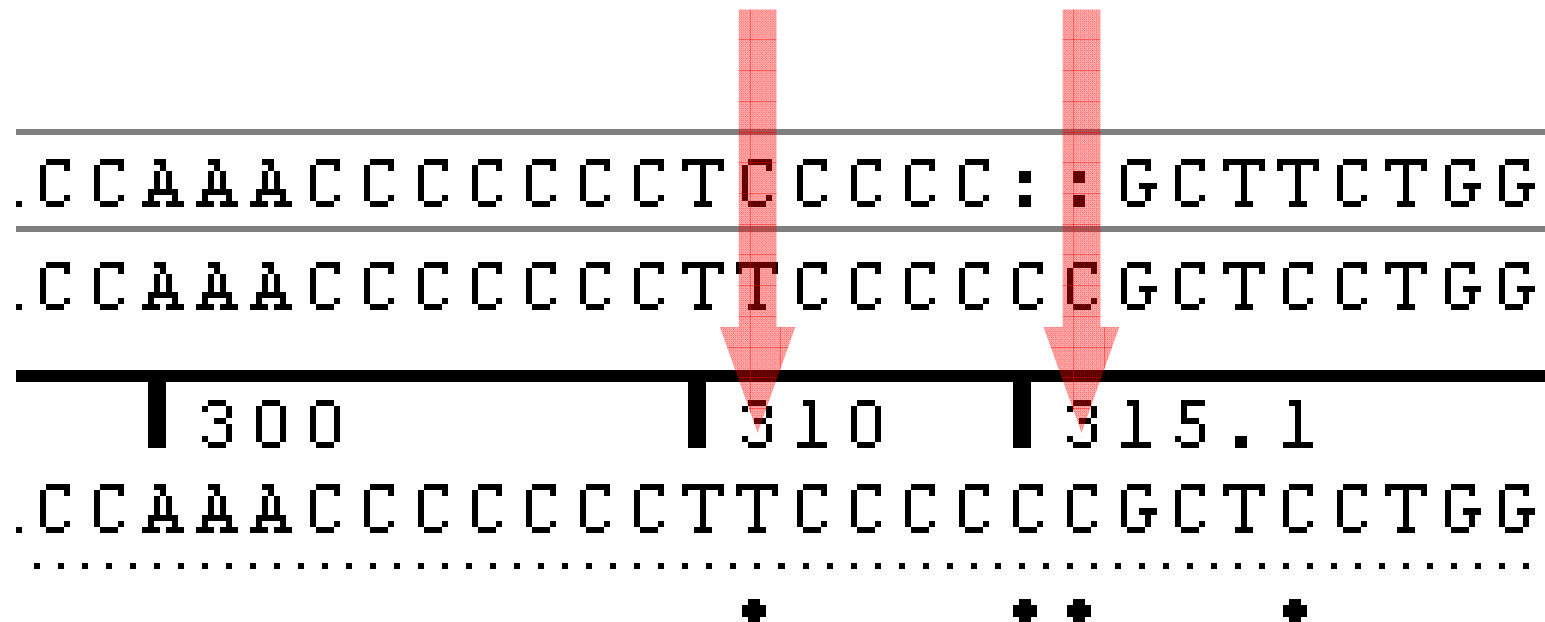
Verletzung der 3'-Insertions-Regel nach *Bär et al. 2000*

# Beispiel – Alignierung von Insertionen



GN 38 Person 3 HVS-II C-Stretch Labor YZ

alternative Alignierung: ... **311T** 315.1C **315.2C** 319C ...



entspricht nicht der phylogenetischen Alignierung nach *Bandelt und Parson 2008*

# Beispiel – Redundante Positionsangabe



GN38 Person 3 HVS-II C-Stretch Labor ZX

16024-16365: 16069T 16126C 16145A 16231C

72-340: 73G 150T 152C 195C 215G 263G 295T 315.1T 315.1C 319C

Version: 2.1, Release: 3 Welcome bettina! [Logout](#)

EMPOP

Query

Tools

Ne

EM

Sta

Do

Account

Input

Input File:

- object G38\_P3: double specification in column 16: 315.1C

**Input File:**

- object G38\_P3: double specification in column 16: 315.1C**

PROCEED

# Beispiel – Notierung von Insertionen



GN 39 Spur C

16024-16365: 16519C

73-340: 152C 263G 309.1C 309.2C 315.1C


# Beispiel – Notierung von Insertionen



GN 39 Spur C Labor ZY

16024-16365:	16519C
73-340:	152C 263G <u>      </u> 309.2C 315.1C

Version: 2.1, Release: 3 Welcome bettina! [Logout](#)

▶  EMPOP

▶ Query

▼ Tools

- Network
- EMP Tool
- Statistics
- Downloads

▶ Account

**Input**

**Input File:**

- object G39\_SC: invalid position 0

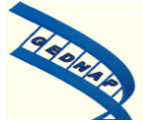
**Input File:**

- object G39\_SC: invalid position 0

[PROCEED](#)

*Tully et al. 2001*

# Beispiel – Rückfall auf rCRS (Reference Bias)



GN 39 Spur B

16024-16365:

16356C

72-340:

73G 146C 309.1C 309.2C 315.1C 524.1C 524.2C



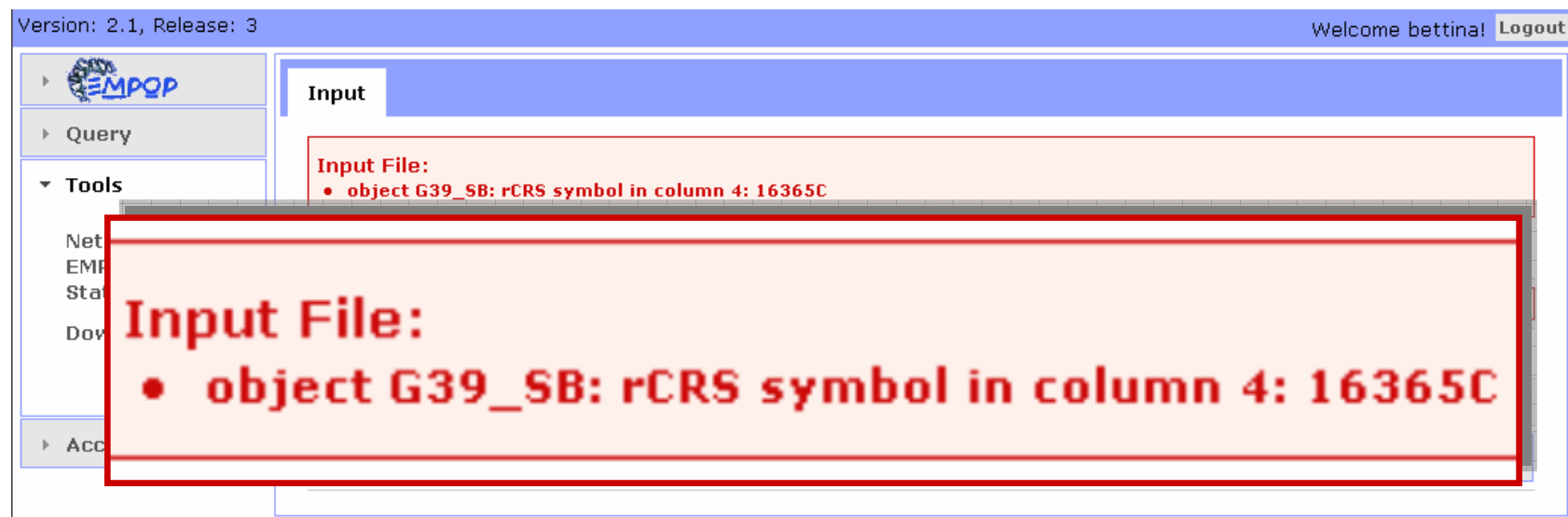
# Beispiel – Rückfall auf rCRS (Reference Bias)



GN 39 Spur B Labor ZZ

16024-16365: 16365C

72-340: 73G 146C 309.1C 309.2C 315.1C 524.1C 524.2C



# *A posteriori* Qualitätskontrolle von mtDNA Daten

Erkennung durch **EMPOP**:

- Unvollständige Angaben (Formatfehler; 16069\_)
- falsche Positionsangabe (1607T)
- Verletzung des Leserahmens (541T bei 51-426)
- Alignierung von Insertionen (309.1T)
- Notierung von Insertionen (309.2C)
- Redundante Angaben (315.1C und 315.1T)
- Rückfall auf Referenzsequenz (Reference bias; 16365C)
- Phantommutation (16300C) → **Fehler-Erkennung im Netzwerk**

**Fehler-Erkennung**

**bereits vor der**

**Netzwerk-Erstellung**

## 8 Fehlerklassen in 11 Proben!

# Limitierungen

Nicht erkannt werden

→ nicht eingetragene Mutationen

→ Mutationen auf gefilterten Positionen (im Netzwerk)

**EMP Tool unterstützt die Qualitätskontrolle, kann aber**

**KEIN GARANT**

**für fehlerfreie Daten sein!**

# Vielen Dank für die Aufmerksamkeit!

