Int J Legal Med (2010) 124:505–509
DOI 10.1007/s00414-010-0492-y

SHORT COMMUNICATION

Publication of population data of linearly inherited DNA markers in the International Journal of Legal Medicine

Walther Parson · Lutz Roewer

Received: 31 May 2010 / Accepted: 2 July 2010 / Published online: 22 July 2010

Abstract This manuscript extends on earlier recommendations of the editor of the *International Journal of Legal Medicine* on short tandem repeat population data and provides details on specific criteria relevant for the analysis and publication of population studies on haploid DNA markers, i.e. Y-chromosomal polymorphisms and mitochondrial DNA. The proposed concept is based on review experience with the two forensic haploid markers databases YHRD and EMPOP, which are both endorsed by the International Society for Forensic Genetics. The intention is to provide guidance with the preparation of population studies and their results to improve the reviewing process and the quality of published data. We also suggest a minimal set of required information to be presented in the publication to increase understanding and use of the data. The outlined procedure has in part been elaborated with the editors of the journal *Forensic Science International Genetics*.

Keywords mtDNA · Mitochondrial DNA · Y-chromosome · STR · SNP · Guidelines

Introduction

In 1997, the editor of the *International Journal of Legal Medicine* has formulated guidelines for the submission of manuscripts on short tandem repeat (STR) population data [3]. The current letter extends on the earlier recommendations and details on specific criteria relevant for the analysis and publication of population studies on haploid DNA markers, i.e. Y-chromosomal polymorphisms and mitochondrial DNA (mtDNA). The proposed concept is based on experience with review of Y-chromosomal and mtDNA data by aid of analysis tools available on Y chromosome haplotype reference database (YHRD; yhrd.org; [14] and EDNAP forensic mtDNA population database (EMPOP) (empop.org; [12]), which are both endorsed by the International Society for Forensic Genetics (isfg.org). The intention of this letter is to provide guidance with the preparation of population studies and their results to improve the reviewing process and the quality of published data. We also suggest a minimal set of required information to be presented in the publication to increase understanding and use of the data. The outlined procedure has in part been elaborated with the editors of the journal *Forensic Science International Genetics* [4].

The procedure of Y-STR/mtDNA data review is organised as follows:

Step 1) Prior to submission to the journal data table(s), one for each studied population, must be sent to YHRD/EMPOP (see details at http://www.yhrd.org/Contribute and http://www.empop.org/modules/contribute);

Step 2) Data review by YHRD/EMPOP, communication with author(s), provision of YHRD/EMPOP accession numbers if data quality is acceptable;

Step 3) Authors submit manuscript and population data with reference to YHRD/EMPOP accession numbers; and

Step 4) Peer review and editorial decision.

W. Parson (✉)
Institute of Legal Medicine, Innsbruck Medical University,
Innsbruck, Austria
e-mail: walther.parson@i-med.ac.at

L. Roewer (✉)
Institute of Legal Medicine and Forensic Sciences,
Charité - Universitätsmedizin Berlin,
Berlin, Germany
e-mail: lutz.roewer@charite.de

Preparation of the manuscript

Sampling

Sample collections arise from convenience sampling (blood banks, paternity testing, immigration cases, laboratory staff, etc.) or targeted sampling (project specific sample recruitment). This may have implications for further use of the data in the forensic context and therefore, this information is a requirement. Sample acquisition should be performed following the defined ethical considerations and under informed consent. This procedure should be stated explicitly. Sample numbers depend on the size of the investigated population. Urban and cosmopolitan population studies should involve at least 200 individuals, whereas, regional datasets should contain a minimum of 100 samples. Exceptions are valid for extremely small population sizes.

Description of the investigated populations

Haploid lineages show much more pronounced stratification compared to autosomal markers, which has consequences for their application in forensic genetics. It is therefore of utmost importance to carefully describe the sampled population correctly and in detail with respect to geographic origin and demographic background applying termini from molecular anthropology and population genetics. This includes the use of a correct ethonym (e.g. "Roma" instead of "Gypsy", "Europeans" instead of "Caucasians", etc.), the definition of the linguistic, and (if applicable) cultural groups (e.g. casts) and subgroups. The authors should follow the concept of metapopulations (as used in YHRD and EMPOP), and further ethnic backgrounds should be indicated (e.g. 'North America, Idaho, Blackfoot Indians'). Both databases supply inventories of included populations that serve as guidance for the assignment of the relevant termini.

Presentation of the population data

Individual haplotypes have to be listed in the publication, usually as electronic supplement. Derived analytical data (statistical summaries) do not recompense for the presentation of the haplotypes. The haplotypes will also be included in the respective databases upon publication; note that contents between publication and database may differ due to continued database maintenance efforts.

Specific recommendations for Y-chromosomal STR systems

Population studies on new or known Y-STR markers are aimed to establish or to widen databases for biostatistical applications in a forensic genetic context. Generally spoken, such submissions are not hypothesis-driven and do not fulfil the requirements of an original article, but rather serve as an important documentary of locus- or haplotype-specific frequencies. Major population databases as the YHRD endorse a policy of publication of local databases prior to its upload and refer to the published material as the original, quality proven source of information.

To facilitate and streamline the reviewing and the editing process and to avoid redundancy within and between articles, we propose an article format presenting the frequency data as follows:

Following the abstract (~100 words) the communication should be reduced to the following topics.

Introduction

Description of the population: a short, but detailed description of the population is essential (a map might be included here), as well as a description of the interest of that population for population genetics and forensic purposes. Previous population genetic studies should be referred, as well as the geographic location, ethnicity, method of sampling, and characteristics of the population.

Materials and methods

This should be reduced to a skeleton of references describing standard methods and commercial kits; exceptions from this should be reported in an abbreviated form.

The statement on informed consent/approval of an ethical committee should be included here.

The accession numbers issued by the YHRD curators after review of the raw data should be included in this section (see http://www.yhrd.org/Contribute for details).

Results

- the Y-STR haplotype data (extended by Y chromosome single nucleotidepolymorphisms (Y-SNPs) if possible) should be represented in an obligatory Table 1 published as electronic supplementary material (ESM).
- this table is a standard spreadsheet file (e.g. Microsoft Excel, the example file and the format description are available at http://www.yhrd.org/Contribute). The first two columns specify an identification number and the origin of the samples. For the latter, we request a ternary identifier in the form "region, country [ethnic group]" e.g. "Berlin, Germany [German]". The geographic background of the samples should be further detailed in an accompanying text or by maps.
- the following columns list the common Y-STR loci; the allele nomenclature should closely follow the updated

and very detailed recommendations of the International Society of Forensic Genetics (ISFG) on the use of Y-STRs in forensic analysis [6]; according to these recommendations, the nomenclature of ten widely used Y-STRs ("minimal haplotype or minHt") should not be altered (e.g. DYS389AB is still DYS389II in a forensic context)

- Please note these additional format rules
  (1) alleles at multiplied loci are separated by a comma e.g. "11,14"
  (2) intermediate alleles are indicated by a dot (e.g. 11.2)
  (3) "Null" alleles as resulting from molecular mechanisms (primer site mutation, deletion) and confirmed by appropriate techniques are indicated by a "0"
- to transcribe different published locus and allele nomenclatures to one widely accepted standard is of utmost importance (e.g. DYSA7.2 is DYS461 according to Genbank nomenclature); if an ambiguous allele nomenclature is chosen, the transcription to other published nomenclatures should be made transparent (for example, allele 29 at DYS389II could be called allele 16 if the repeat 13 of DYS389I is subtracted from DYS389II)
- locus-specific allele frequencies should not be reported, with the exception of (1) the locus has never been described before, or (2) the allele frequency of a given allele, or the frequency distribution curve as a whole, points to an underlying mechanism which needs a further detailed description in the "Remarks" section
- the samples should be given unique identifiers, rather than haplotype numbers, to trace them back if additional information is published in the future
- biostatistical parameters to describe the population sample(s) may include: discrimination capacity, haplotype diversity, frequency of the most frequent haplotype(s)
- informative comparisons to relevant reference populations (e.g. from the YHRD) should be included; values as Fst, Φst or Rst and a significance test (*p values*) should be calculated, an MDS plot or STR network to illustrate genetic distances between populations could be included (tools to perform such analyses e.g. AMOVA are available at http://www.yhrd.org/Analyse); the outcome of these comparisons should be discussed in short

Other remarks

- here, additional data, e.g. on observed mutations, haplogroup affiliation, etc. can be reported in short
- if mutations are observed (affecting the repeat length, the flanking site or the number of amplified sequences), the repeat type should be verified, usually by sequencing, and mutated alleles should be reported to the YHRD as recommended by the ISFG [6]
- the father's age distribution in the cases with and without a mutation should be reported
- if a mutation at one or more STR is caused by a known chromosomal rearrangement (e.g., DYS464), the observation should be related to the appropriate reference
- when haplotypes are extended by phylogenetically informative Y-SNPs (as recommended), the most recent nomenclature [7] should be followed; the full panel of typed Y-SNPs, specified by "+" for the derived state, "–" for the ancestral state at a given locus ("−1" for "not typed") should be included in the Table 1 as an extension to each individual Y-STR haplotype. The last two columns of the Table 1 should include the final branch marker used for haplogroup assignment, e.g. M3 and the assigned branch name (Q1a3a).

The following statement should conclude this topic: "This paper follows the recommendations of the ISFG on the use of Y-STRs in forensic analysis [6] and the guidelines for publication of population data requested by the journal."

A statement that the Y-STR population data have been submitted to the YHRD should be included in the manuscript.

These rules apply in principle also to the communication of data on Y-STR mutation rates, although a combined submission of haplotype and mutation data are recommended [6]. The minimum number of allele transmissions should be 100 per locus.

Specific recommendations for mtDNA population data

MtDNA population studies are aimed to establish or to widen databases for biostatistical applications in a forensic genetic context. In a similar way, as outlined for YHRD, the mtDNA database EMPOP endorses a policy of publication of local databases prior to its upload and refers to the published material as the original source of information. In contrast to Y-STR data that are generally established under well-defined conditions (commercial kits, allelic ladders, international positive controls), mtDNA data rely on the scientific and technical expertise of the individual laboratory. Therefore, and due to the complexity of mtDNA data, it has been demonstrated that quality control plays a more important role compared to other genetic markers [9, 13]. The following criteria help in improving the quality of mtDNA studies:

Subjects

It is imperative that the authors detail where the investigated samples originate from (see chapter "Description of the investigated polymorphism" above for further details).

Methods

The methods of DNA extraction, the selected amplification and sequencing primers, and the version of the sequencing chemistry play an important role for the footprint of the sequence raw data [10] and should be clearly defined.

The analysis of the individual hypervariable segments HVS-I and HVS-II within the mtDNA control region (CR) is no more state of the art. First, independent amplification and sequencing of multiple mtDNA segments increase the risk of sample mix-up (artificial recombination, [11]). Second, many useful polymorphisms residing at CR positions outside HVS-I/HVS-II are missed, which results in decreased discrimination power and the loss of relevant haplogroup-specific sites. It has been shown that sequencing of the entire CR substantially adds to our knowledge of the mtDNA phylogeny and variation, and provides a sound basis for forensic mtDNA analysis. Streamlined and optimised protocols for amplification and sequence analysis of the entire CR are presented in [11] and references therein.

Single-stranded sequence data have shown to harbour an unacceptable rate of phantom mutations resulting in artificial mtDNA haplotypes. The consensus haplotype reported for a sample must originate from at least full double-stranded sequence information. Although this was demanded earlier, it has not yet become a routine practise. This is most importantly true for samples exhibiting length heteroplasmy, which require additional sequencing in order to achieve confirmed base call assignment. In particular, we recommend sequencing length heteroplasmic regions with additional primers to unambiguously define the dominant variant, as suggested in [2, 5] and references therein.

The statement on informed consent/approval of an ethical committee should be included here.

The accession numbers issued by the EMPOP curators after review of the raw data should be included in this section (see http://www.empop.org/modules/contribute for details).

Results

- The manual transcription of mtDNA profiles into tables of publications constitutes the highest risk of introducing errors. IT-based conversion of the data is recommended. In any case, examination of the compiled data should be performed with scrutiny. This involves confirmation of the population data in the table by an independent scientist and/or double reading of the data under the four-eye principle.
- It is important to present and evaluate population data on the basis of the established mtDNA phylogeny. Alignment and annotation of mtDNA sequences should be performed according to the recommendations by the ISFG [1]; refined guidelines for the notation of haplotypes with more than one possible alignment are available in [2].
- Haplotypes should be assigned to mtDNA haplogroups using the relevant literature (a continuously updated mtDNA tree is provided by ([8]; www.phylotree.org). Especially for samples with unclear haplogroup affiliation, it has proven useful to confirm their assignment using coding region information.
- The "phylogenetic check" is a very probative means for the evaluation of the data as yet unobserved variation becomes apparent and can be inspected by means of the raw data. This can be performed using the relevant literature or the NETWORK software provided via the EMPOP database (www.empop.org; [12]).
- All electropherograms that are necessary to reconstruct the submitted data should be stored by the authors and provided upon particular request (check for questionable base calls).

Format of result tables (ESM)

- The representation of mtDNA variation in matrix-based "dot-tables" is error-prone due to the fact that columns and rows can be mistakenly re-arranged during the publication process. Also, these tables are usually difficult to read.
- It is required that mtDNA haplotypes are reported one by one in difference-based format (with respect to the rCRS) using a unique identifier as sample information that can further be used to make reference to a sequence.
- Also, the interpretation range of the haplotypes needs to be specified. This range usually complies with the sequenced fragment, primer sequences disregarded (e.g. 16024-576 for the entire control region), but can also be restricted to smaller regions.
- EMP-format: EMPOP offers software for quality control of mtDNA sequences which requires a standard format to enable solid performance. Specific information on the file-format and an example file can be found on the EMPOP website (see Tools and Download sections at www.empop.org). This format can also be used for publication of the data as ESM in the journal.

Statistical parameters

- The report of frequency values of single mtDNA polymorphisms is meaningless, as transitions and transversions are linked within haplotypes. Values to

describe the population sample biostatistically should include: number of haplotypes, number of unique haplotypes, haplotype diversity, random match probability, and frequency of the most frequent haplotype(s).
- Informative comparisons to relevant populations should include values as Φst or Rst and a significance test (*p values*).

General issues

- A statement that the mtDNA population data have been submitted to the EMPOP database should be included in the manuscript.
- The following statement should conclude this topic: "This paper follows the recommendations of the ISFG on the use of mtDNA in forensic analysis and the guidelines for publication of population data requested by the journal."

Acknowledgements Some of the mtDNA recommendations are based on research supported by the FWF Austrian Science Fund Translational Research Programme (L397). We would like to thank Sascha Willuweit and Alexander Röck for valuable assistance and discussion.

References

1. Bär W, Brinkmann B, Budowle B et al (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Int J Legal Med 113:193–196
2. Bandelt H-J, Parson W (2008) Consistent treatment of length variants in the human mtDNA control region: a reappraisal. Int J Legal Med 122:11–21
3. Brinkmann B (1997) Editorial. Int J Legal Med 110:117
4. Carracedo A, Butler JM, Gusmao L et al (2010) Publication of population data for forensic purposes. Forensic Sci Int Genet 4:145–147
5. Forster L, Forster P, Gurney SM et al (2010) Evaluating length heteroplasmy in the human mitochondrial DNA control region. Int J Legal Med 124:133–142
6. Gusmao L, Butler JM, Carracedo A et al (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. Int J Legal Med 120:191–200
7. Karafet TM, Mendez FL, Meilerman MB et al (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18:830–838
8. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386–E394
9. Parson W, Brandstätter A, Alonso A et al (2004) The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results and perspectives. Forensic Sci Int 139:215–226
10. Parson W (2007) The art of reading sequence electropherograms. Ann Hum Genet 71:276–278
11. Parson W, Bandelt H-J (2007) Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet 1:13–19
12. Parson W, Dür A (2007) EMPOP—a forensic mtDNA database. Forensic Sci Int Genet 1:88–92
13. Röhl A, Brinkmann B, Forster L, Forster P (2001) An annotated mtDNA database. Int J Legal Med 115:29–39
14. Willuweit S, Roewer L, on behalf of the International Forensic Y chromosome User Group (2007) Y Chromosome Haplotype Reference Database (YHRD): update. Forensic Sci Int Genet 1:83–87