doi 10.4436/jass.90020

JASs Do it yourself
Journal of Anthropological Sciences
Vol. 90 (2012), pp. 201-215

Online databases for mtDNA and Y chromosome polymorphisms in human populations

Alessandra Congiu[1,2,§], Paolo Anagnostou[1,4,§], Nicola Milia[1,2,§], Marco Capocasa[3,4], Francesco Montinaro[4,5] & Giovanni Destro Bisol[1,4]

1) Università di Roma "La Sapienza", Dipartimento di Biologia Ambientale, Roma, Italy
e-mail: destrobisol@uniroma1.it

2) Università di Cagliari, Dipartimento di Scienze della Vita e dell'Ambiente, Cagliari, Italy

3) Università di Roma "La Sapienza", Dipartimento di Biologia e Biotecnologie "Charles Darwin", Roma, Italy

4) Istituto Italiano di Antropologia, Roma, Italy

5) Università Cattolica, Facoltà di Medicina, Istituto di Medicina Legale, Roma, Italy

Summary – *This study presents an overview of online databases for mtDNA and Y chromosome polymorphisms in human populations. In order to provide readers with information which may help optimize their use, we focus on: (i) type, quantity and source of data contained; (ii) possibilities of downloading and uploading; (iii) availability of data filters and population genetics tools. We show that some of these databases offer a useful complement to the primary databases by giving access to additional data and making it possible to perform queries which exploit some specific metadata. Thereafter, we evaluate the state of the art from an evolutionary anthropologist's point of view. We suggest that online databases could become even more useful research tools by combining an easier data retrieval with quality control and by making a more extensive use of metadata regarding populations and individuals. Making population data on mtDNA and Y chromosome polymorphisms more complete, well ordered and easily accessible, we could better exploit the potential of new generation sequencing techniques for advancements in human evolutionary genetics.*

Keywords – *Human genetic variation, Secondary databases, Data sharing, Data downloading, Data uploading, Population genetics tools, Anthropological metadata, Data quality control.*

mtDNA and Y chromosome polymorphisms as tools in Evolutionary Anthropology

Since their introduction in the 80's, unilinearly transmitted polymorphisms of mitochondrial DNA (mtDNA) and Y chromosome have been increasingly used in different fields of biological and medical research (Brown *et al.,* 1980; Casanova *et al.,* 1985). In the early 90s, they become more frequently used due to the introduction of Polymerase chain reaction (PCR) and automated sequencing (Sullivan *et al.,* 1991; Ehrlich & Arnheim, 1992). In fact, these methods made it possible to reduce the time and costs of genotyping and carry out surveys of mtDNA and Y chromosome variation at population level. As shown by a PUBMED search, mtDNA and Y chromosome polymorphisms have largely been used in the last decade (Fig. 1). The data reported illustrate that, today, they still represent a widespread source of genetic information for research work, despite the recent introduction of more powerful genome-wide approaches.

[§] These authors made equal contributions to this work.

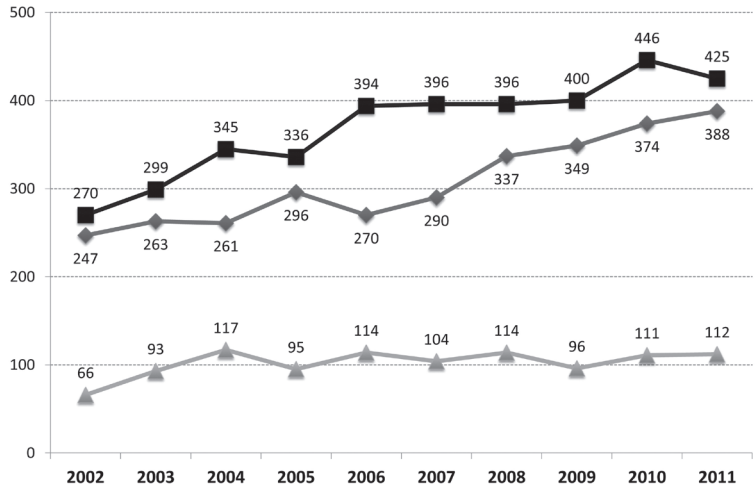

Fig. 1 – Counts on a yearly basis of entries retrieved from Pubmed combining the key words "mtDNA" and "polymorphisms" (squares) or "Y-chromosome" and "polymorphisms" (triangles). Entries retrieved using "autosomal" and "polymorphisms" (diamonds) are shown for comparison. The database was accessed on May 20th, 2012. The color version of this figure is available at the JASs website.

The ease of determination, uniparental inheritance and lack of recombination (except for the pseudo-recombining portion of the Y chromosome), along with the possibility of combining markers with different rates of evolution make these polymorphisms of particular usefulness for the study of human evolutionary processes (Pakendorf & Stoneking, 2005; Lahn *et al.,* 2001; Destro Bisol *et al.,* 2010). Furthermore, due to its high copy number in each cell (hundreds to thousands of copies), mtDNA is the molecule of choice for analyzing ancient DNA (Rizzi *et al.,* 2012), while recently there have also been developments regarding Y chromosome (e.g. Lacan *et al.,* 2011). As also shown by some papers recently published in this Journal, mtDNA and Y chromosome studies have provided useful insights into key issues in human evolution, shedding light on the genetic history of human populations in various geographic contexts and timescales (Cann *et al.,* 1987; Comas, 2010; Francalacci *et al.,* 2010; Rosa & Brehm, 2011). In a current perspective, unilinear polymorphisms provide a data basis for a number of computational tools developed to investigate complex evolutionary scenarios, an approach which is expected to lead to important advances in the near future (Destro Bisol *et al.,* 2010; Tofanelli *et al.,* 2011; Hoban *et al.,* 2012).

An overview of online databases

Sharing new experimental results and building more comprehensive datasets are basic tasks for studies of human genetic variation, as well as in other fields of biological research (Milia *et al.,* 2012). The interoperating *GenBank*, *European Nucleotide Archive* and *DNA Data Bank of Japan*, usually referred to as "primary databases", may be used for these purposes. Giving free access to nucleotide sequences for more than 250,00 species, these tools represent a comprehensive (and sustainable) data source for human population genetics and many other fields of genetic research (Benson *et al.,* 2012).

An additional resource is represented by the "secondary" databases of mtDNA and Y chromosome polymorphisms, which have been developed since the mid 90's. Naturally, they

may complement primary databases and provide additional results. From the point of view of an evolutionary anthropologist, it would be important that such tools made use of metadata which make it possible to filter data for populations, areas or haplogroups and describe the spatial distribution, cultural factors and social structure of surveyed groups. This series of information is probably too specific and demanding to be incorporated in a comprehensive and non-specialized database such as *GenBank*, although a step forward has been made with the introduction of the *PopSet* database. Explicitly developed to "contain sequences collected to analyze the evolutionary relatedness of populations", *PopSet* makes downloading of population data easier.

Here we present an inspection and comparison of mtDNA and Y chromosome online databases. This overview has two main objectives. We intend to provide JASs readers with practical information which may help them optimize the use of these tools. To make a follow up possible and share information more easily, the results of our overview are accessible online (*AnthrodigIt*^{data})[1]. Thereafter, we comment on some limits which impair an exhaustive exploitation of the mtDNA and Y chromosome online databases and discuss how they could be overcome.

Four practical questions[2]

How many databases are presently accessible online?

We have been able to find 14 **mtDNA** online databases, three of which also contain Y chromosome data (*DNA-Fingerprint, Family Tree DNA* and *SMGF*) (Tab. 1). Online publication ranges from 1996 for *GOBASE* to 2012 for *mtDNA community*. Of the 12 databases for which this information was obtainable, only six have been updated in the course of 2012. A reference paper and online help is available only for 10 databases. We were able to list 7 **Y chromosome** databases, including the three cited above, which have been published online between 2001 (*YHRD*) and 2006 (*US Y-STR Database*) (Tab. 2). Only three databases were found to have been updated in 2012. A reference paper is available for 2 databases, and all provide online help.

What types and quantity of data are contained therein?

Family Tree DNA is the largest archive for **mtDNA** sequences (mainly unpublished) both at low (HVR-1 and II) and high resolution (complete mtDNA or coding region) (see Appendix 1A). *Phylotree* and *mtDNA Community* provide the largest wealth of published whole genome sequences, with figures (14508 and 13492, respectively) not far from *GenBank* (16414). Nine databases are based only on literature data. *Empop*, a database of HVR-1 and HVR-2 data, provides an accurate quality control through the use of post-laboratory procedures[3]. The largest number of **Y chromosome** STR haplotypes is available in *Family Tree DNA* (236302), *Ysearch* (112513) and *YHRD* databases (101055) (Appendix 2A). The former is also the greatest source of SNP/STR combined haplotypes (62795). Data from scientific literature are used in *YHRD*. By contrast, *US Y-STR database* seems to contains most, if not only, haplotypes submitted from forensic laboratories and institutions. It is noteworthy that, unlike with mtDNA, *GenBank* does not give access to Y chromosome population data in the haplotypic form.

What databases make it possible to retrieve/share primary datasets?

Unrestricted downloading is possible from 9 **mtDNA** databases, whereas three of them (*Family Tree DNA, DNA-Fingerprint* and *mtDNA manager*) make it possible to retrieve only a part of the data (Appendix 1B). The most used file format for data download (in 7 out of 15) is FASTA (Pearson

[1] http://www.isita-org.com/Antro-Digit/data.htm; integrations, updates and corrections to the information reported therein may be sent to isita@isita-org.com.

[2] Data reported here were obtained on July 13th 2012.

[3] Raw data are analyzed - and electronically compared - by at least 3 experienced scientists. Differently from other databases, *Empop* provides information on length and point heteroplasmy and creates consensus sequences based on multiple coverage (W. Parson, personal communication).

Tab. 1 - Basic information on online databases for mtDNA polymorphisms in human populations. To avoid the risk of overloading the analysis with redundant and not validated data, we have not taken into account databases hosting data produced by single laboratories. Databases were accessed on July 13th 2012.

DATABASE	PUB. ONLINE	LAST UPDATE	USER HELP	REFERENCE	CONTACT
DNA-Fingerprint[1,2]	2004	n.a.	x[3]	-	info@familytreedna.com; thomas@dna-fingerprint.com
Empop	2006	04/2012	x	Parson & Dür, 2007	walther.parson@i-med.ac.at
Family Tree Dna	2000	07/2012	x	-	info@familytreedna.com
GOBASE	1996	06/2010	x	O'Brien *et al.*, 2009	gobase@BCH.UMontreal.CA
HmtDB	n.a.	06/2012[4]	x	Attimonelli *et al.*, 2005	marcella.attimonelli@uniba.it
HVRbase++	1998	06/2005	x	Handt *et al.*, 1998 Burckhardt *et al.*, 1999 Kohl J *et al.*, 2006	arndt.von.haeseler@univie.ac.at
Mitosearch[2,5]	2004	n.a.	x	-	info@familytreedna.com
MitoVariome	2009	07/2009	x	Lee *et al.*, 2009	yslee@kribb.re.kr
mtDB	2000	03/2007	-	Ingman & Gyllensten, 2006	Max.Ingman@genpat.uu.se
mtDNA Community[6]	2012	09/2012	-	Behar *et al.*, 2012	media@familytreedna.com
mtDNA manager	n.a.	10/2011	x	Lee *et al.*, 2008	kjshin@yuhs.ac
mtSNP database	2003	01/2006	-	Tanaka *et al.*, 2004	mtsnp@giib.or.jp
PhyloTree	2008	05/2012	-	van Oven & Kayser, 2009	m.vanoven@erasmusmc.nl
SMGF[7]	1999	07/2012	x	-	info@smgf.org

[1] users may modify their own entries.
[2] exchanges data with Family Tree DNA.
[3] x stands for presence of a specific feature.
[4] Personal communication, M. Attimonelli.
[5] accepts data from Genographic users.
[6] accessed on 20/09/2012
[7] Sorenson Molecular Genealogy Foundation

et al., 1988). *MtDNA Community* is the only one which uses the recently proposed Reconstructed Sapiens Reference Sequence (RSRS) (Behar *et al.,* 2012) to identify mutation motifs. Researchers may upload population data, both published and unpublished, by using *Empop* and *mtDNA Community*. Data can be downloaded from only one **Y chromosome** database (*Ysearch*), whereas another two allow a partial retrieval (*Family Tree DNA* and *DNA-Fingerprint*) (Appendix 2B). Data may be freely uploaded in *Ysearch* with the only limit being having to use at least 8 STR markers. *YHRD* behaves differently, requiring that contributing laboratories pass a "quality test", proving they are able to type correctly 5 blind DNA samples for the Y-STR markers they are submitting. Differently from mtDNA, no specific data format is available for Y chromosome haplotypes.

What tools and/or filters are available?

A "map" of available tools and filters is reported in Appendices 1C, 1D, 2C and 2D. Matching

Tab. 2 - Basic information on online databases for Y chromosome polymorphisms in human populations. To avoid the risk of overloading the analysis with redundant and not validated data, we have not taken into account databases hosting data produced by single laboratories. Databases were accessed on July 13th 2012.

DATABASE	PUB. ONLINE	LAST UPDATE	USER HELP	REFERENCE	CONTACT
DNA-Fingerprint[1,2]	2004	n.a.	x[3]	-	info@familytreedna.com; thomas@dna-fingerprint.com
Family Tree Dna	2000	07/2012	x	-	info@familytreedna.com
SMGF[4]	2004	03/2011	x	-	smgfsupport@ancestry.com
US Y-STR Database	2006	01/2012	x	Ge *et al*. 2010	lfatolit@mail.ucf.edu
Y-Filer Haplotype Database	n.a.	n.a.	x	-	n.a.
YHRD[5]	2001	02/2012	x	Willuweit & Roewer, 2007	s@rprojekt.org
Ysearch	2003	n.a.	x	-	ysearch@usernet.com

[1] users may modify their own entries.
[2] exchanges data with Family Tree DNA.
[3] x stands for presence of a specific feature.
[4] Sorenson Molecular Genealogy Foundation.
[5] Y chromosome Haplotype Reference Database.

tools are implemented in 6 **mtDNA** databases, 3 of which (*Empop, mtDNA manager* and *SMGF*) allows you to choose type (complete or partial) and nucleotide range of match. Other population genetic tools (e.g. software, either online or downloadable) are hosted by 6 databases, two of which (*HmtDB* and *mtDNA Manager*) give the opportunity to perform haplogroup prediction/estimation (but see Bandelt *et al.,* 2012 for an evaluation of the reliability of haplogroup predictors). Data to be downloaded or to be used for matching analysis can be filtered with most databases (10 out of 15), but multiple queries may be performed with only 9 of them. Browsing data for parameters of interest for evolutionary studies, such as geographical location or haplogroup, is possible with 8 and 7 databases, respectively. Analysis of complete match may be performed in 6 **Y chromosome** databases, two of which them can also identify partial matches (*SMGF* and *Ysearch*). Population genetics tools are provided by *US-YSTR database*, *Ysearch* and *YHRD*. Filters are implemented in 5 databases, 3 of which allowing multiple queries. Browsing data for geographical origin or haplogroup is possible with 3 (*DNA-Fingerprint*, *YHRD* and *Ysearch*) and 2 (*DNA-Fingerprint* and *Ysearch*) databases, respectively.

A tool user perspective

In this overview, we have considered online databases owned by genealogical companies and others which are more oriented towards scientific research. Our study shows that the former (*Family Tree DNA* and related databases, *SMGF*) provide the largest quantity of mtDNA and Y chromosome data, often with high resolution. However, differently from research databases which are based mainly on published population data, the genealogical ones count on individual submissions of company costumers. This feature may lead to an overrepresentation of geographical areas where customers are most easily found. Additionally, in some cases results may be freely uploaded (*Mitosearch* and *Ysearch*) or modified (*DNA-Fingerprint*) by customers. This makes it particularly important to select data carefully.

Finally, metadata are not always adequate to reconstruct samples of unrelated individuals or extract subpopulations. However, despite these limitations, results from genealogical databases represent a potential source of information in evolutionary studies (e.g. Balaresque *et al.,* 2010; but see Busby *et al.,* 2012).

Focusing on research databases, *Phylotree* contains the largest number of complete mtDNA genomes, with an appreciable proportion of entries (~4%) which is unshared with primary databases (see Appendix 1A). Notes regarding sequencing errors are available and some results are made available after correction. Unfortunately, the simple structure of the database does not permit data to be browsed using specific queries. A slightly lower number of mtDNA genomes is available in the recently published *mtDNA Community* (679 not available in *GenBank*), which uses a mtDNA phylogeny as reassessed by Behar *et al.,* (2012) and allows data filtering for haplogroup. A smaller number of complete sequences (8813) is obtainable from *HmtDB*, which, on the other hand, makes it possible to perform a number of simple and multiple queries. The number of sequences available in *GenBank* outnumbers these databases. However, with primary databases, queries for specific populations, geographical areas and haplogroups may be cumbersome and prone to errors. This is due to the fact that it is not mandatory to supply the corresponding metadata when submitting the data. *Empop* is the database which contains the largest body of mtDNA data for HVR-1 and HVR-2, with several tools and filters for the analysis of match between sequences. Notably, it also provides an accurate quality control for both published and unpublished data (Prieto *et al.,* 2012), which is essential to minimize incorrect haplotype or haplogroup assignment due to sequencing errors (Parson & Bandelt, 2007; Bandelt *et al.,*2007). The above feature of *Empop* is congruent with the attention of forensic geneticists to data quality, demonstrated by the numerous collaborative exercises of DNA genotyping they have carried out since 1994 (Gill *et al.,* 1994, Carracedo *et al.,* 1998), initiatives which may be beneficial also to other fields of research. It is worth noting that mtDNA sequences deposited in *GenBank* have been shown not to be exempt from errors (Yao *et al.,* 2009; but see Pereira & Samuels 2009). On the other hand, *Empop* data (as it happens with *YHRD*) cannot be downloaded, a feature which may be explained in some cases by potential risks for privacy violation and specific forensic needs. Unfortunately, retrieving data from the relevant papers or (for unpublished data) obtaining them from corresponding authors is not always an easy task. A recent study by our group shows that raw datasets are not immediately available in ~20% of mtDNA studies, while only 28.6% of email requests to make withheld datasets available are positively answered (Milia *et al.,* 2012). *HVRBase++* permit immediate downloading of partial mtDNA sequences, but its dataset is substantially smaller than in *Empop*.

Concerning Y chromosome databases, *YHRD* contains a large number of high quality data for both STR and SNP loci. However, it cannot be directly accessed and permits less data filtering than its mtDNA counterpart (*Empop*). On the whole, tools and filters are scantier than in mtDNA sites (Appendices 2C and 2D). In practice, there are two not mutually exclusive ways to build a Y chromosome population dataset through the exploitation of these online databases. Relevant papers can be selected by querying *YHRD* (by SNP or geography). Then, datasets might be manually retrieved from them or requested to corresponding authors when results are not available in their complete form. Otherwise, one can use genealogical databases, some of which facilitate search by haplogroup and geographical origin.

Concluding Remarks

In our overview, we inspected numerous secondary databases for human mtDNA and Y chromosome polymorphisms. We have shown that these tools may offer a useful complement to primary databases. In fact, some of them give

access to types of polymorphisms that are not considered by primary databases (i.e. Y chromosome haplotypes), while others contain additional data (e.g. *Phylotree, mtDNA Community* and genealogical databases) or carry out a specific data quality control (*Empop* and *YHRD*). However, the number of available databases is not entirely proportional to their actual usefulness. In fact, in several cases, their datasets overlap, databases for complete mtDNA sequences being the best example. Furthermore, the features of most tools, even those developed for specific research purposes, do not entirely adapt to the needs of human evolutionary genetics. This is due to the impossibility of immediate data downloading and lack of filters linked to exhaustive and well organized metadata which may really help construct datasets tailored for anthropological studies.

Therefore, how could we make online databases of mtDNA and Y chromosome polymorphisms even more useful than they are today? From what has already been said, it is clear that the best strategy could be to combine easier data retrieval with quality control and a more extensive use of metadata for populations (e.g. geographical coordinates, language/s spoken, ancestry and matrimonial behaviour) and individuals (place of birth, age, gender and parentage relationships with other donors) (Destro Bisol *et al.,* 2012). Using a common metadata format, which could be obtained by modifying one of those already available (e.g. gedcom, see *mtDNA Community*), would make it possible to ensure homogeneity across datasets. Certainly, these improvements are not easy to achieve. The use of detailed metadata for individuals may be a threat for privacy and probably needs to be adjusted in relation to the discriminating power of polymorphisms used and the size of population under study. Recruiting economic resources which would make the effort sustainable is another critical point. Avoiding partial duplicates which do not really add any substantial advantages to the existing tools and which are often abandoned after several years could help us better direct our efforts on a few but well managed databases. Optimizing the ratio between costs and benefits seems to be even more appropriate in the light of the increasing diffusion of new generation sequencing data, which is inevitably destined to attract a substantial fraction of resources for research on human genetic variation. Last but by no means least, by making preexisting sources of genetic information more complete, well ordered and easily accessible, we could better exploit the potential of this ongoing transition for advancements in human evolutionary genetics.

Acknowledgements

We would like to thank Mannis van Oven, Walther Parson, Bennett Greenspan, Arndt von Haeseler, Masashi Tanaka and Thomas Krahn for giving us a useful feedback. However, we would like to underline that the authors accept full responsibility for any errors or factual inaccuracy. This work was supported by the Ministero dell'Istruzione, dell'Università e della Ricerca (PRIN 2009-2011, prot.n. 200975T9EW) *and the* Istituto Italiano di Antropologia.

References

Attimonelli M., Accetturo M., Santamaria M., Lascaro D., Scioscia G., Pappadà G., Russo L., Zanchetta L. & Tommaseo-Ponzetta M. 2005. HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. *BMC Bioinformatics*, 6 Suppl. 4:S4.

Balaresque P., Bowden G.R., Adams S.M., Leung H.Y., King T.E., Rosser Z.H., Goodwin J., Moisan J.P., Richard C., Millward A., Demaine A.G., Barbujani G., Previderè C., Wilson I.J., Tyler-Smith C. & Jobling M.A. 2010. A predominantly Neolithic origin for European paternal lineages. *PLoS Biol.*, 8: e1000285.

Bandelt H.J., van Oven M. & Salas A. 2012. Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. *Int. J. Legal Med.* 126: 901-916.

Bandelt H.J., Yao Y.G., Salas A., Kivisild T. & Bravi C.M. 2006. High penetrance of sequencing errors and interpretative shortcomings in mtDNA sequence analysis of LHON patients. *Biochem. Biophys. Res. Commun.*, 352: 283-291.

Behar D.M., van Oven M., Rosset S., Metspalu M., Loogväli E.L., Silva N.M., Kivisild T., Torroni A. & Villems R. 2012. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.*, 90: 675-684.

Benson D.A., Karsch-Mizrachi I., Clark K., Lipman D.J., Ostell J. & Sayers E.W. 2012. GenBank. *Nucleic. Acids Res.*, 40: 48-53.

Brown W.M. 1980. Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 77:3605-3609.

Burckhardt F., von Haeseler A. & Meyer S. 1999. HvrBase: compilation of mtDNA control region sequences from primates. *Nucleic Acids Res.*, 27:138-142.

Busby G.B., Brisighelli F., Sánchez-Diz P., Ramos-Luis E., Martinez-Cadenas C., Thomas M.G., Bradley D.G., Gusmão L., Winney B., Bodmer W., Vennemann M., Coia V., Scarnicci F., Tofanelli S., Vona G., Ploski R., Vecchiotti C., Zemunik T., Rudan I., Karachanak S., Toncheva D., Anagnostou P., Ferri G., Rapone C., Hervig T., Moen T., Wilson J.F. & Capelli C. 2012. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proc. Biol. Sci.*, 279: 884-892.

Cann R.L., Stoneking M. & Wilson A.C. 1987. Mitochondrial DNA and human evolution. *Nature*, 325: 31-36.

Carracedo A., D'Aloja E., Dupuy B., Jangblad A., Karjalainen M., Lambert C., Parson W., Pfeiffer H., Pfitzinger H., Sabatier M., Syndercombe Court D. & Vide C. 1998. Reproducibility of mtDNA analysis between laboratories: a report of the European DNA Profiling Group (EDNAP). *Forensic Sci. Int.*, 97: 165-170.

Casanova M., Leroy P., Boucekkine C., Weissenbach J., Bishop C., Fellous M., Purrello M., Fiori G. & Siniscalco M. 1985. A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science*, 230: 1403-1406.

Comas D. 2010. The Genographic Project: insights into Western/Central European variation. *J. Anthropol. Sci.*, 88: 243-244.

Destro Bisol G., Capocasa M. & Anagnostou P. 2012. When gender matters: new insights into the relationships between social systems and the genetic structure of human populations. *Mol. Ecol.*, 21: 4917-4920.

Destro-Bisol G., Jobling M.A., Rocha J., Novembre J., Richards M.B., Mulligan C., Batini C. & Manni F. 2010. Molecular anthropology in the genomic era. *J. Anthropol. Sci.*, 88: 93-112..

Erlich H.A. & Arnheim N. 1992. Genetic Analysis Using the Polymerase Chain Reaction. *Annu. Rev. Genet.*, 26: 479-506

Francalacci P., Morelli L., Useli A. & Sanna D. 2010. The history and geography of the Y chromosome SNPs in Europe: an update. *J. Anthropol. Sci.*, 88: 207-214.

Ge J., Budowle B., Planz J.V., Eisenberg A.J., Ballantyne J. & Chakraborty R. 2010. US forensic Y chromosome short tandem repeats database. *Leg. Med. (Tokyo)*, 12: 289-295.

Gill P., Kimpton C., D'Aloja E., Andersen J.F., Bar W., Brinkmann B., Holgersson S., Johnsson V., Kloosterman A.D., Lareu M.V. *et al.,* 1994. Report of the European DNA profiling group (EDNAP)--towards standardisation of short tandem repeat (STR) loci. *Forensic Sci. Int.*, 65:51-59.

Hoban S., Bertorelle G. & Gaggiotti O.E. 2012. Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.*, 10: 110-122.

Ingman M. & Gyllensten U. 2006. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Res.*, 34: D749-751.

Kohl J., Paulsen I., Laubach T., Radtke A. & von Haeseler A. HvrBase++: a phylogenetic database for primate species. *Nucleic Acids Res.*, 34: D700-704.

Lacan M., Keyser C., Ricaut F.X., Brucato N., Tarrús J., Bosch A., Guilaine J., Crubézy E. & Ludes B. 2011. Ancient DNA suggests the leading role played by men in the Neolithic

dissemination *Proc. Natl. Acad. Sci. U.S.A.*, 108: 18255-18259.

Lahn B.T., Pearson N.M. & Jegalian K. 2001. The human Y chromosome, in the light of evolution. *Nat. Rev. Genet.*, 2: 207-216.

Lee H.Y., Song I., Ha E., Cho S.B., Yang W.I. & Shin K.J. 2008. mtDNAmanager: a Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences. *BMC Bioinformatics*, 9: 483.

Lee Y.S., Kim W.Y., Ji M., Kim J.H. & Bhak J. 2009. MitoVariome: a variome database of human mitochondrial DNA. *BMC Genomics*, 10 Suppl 3:S12.

Milia N., Congiu A., Anagnostou P., Montinaro F., Capocasa M., Sanna E. & Destro Bisol G. 2012. Mine, yours, ours? Sharing data on human genetic variation. *PLoS One*, 7:e37552.

O'Brien E.A., Zhang Y., Wang E., Marie V., Badejoko W., Lang B.F. & Burger G. 2009. GOBASE: an organelle genome database. *Nucleic Acids Res.*, 37: D946-950.

Pakendorf B. & Stoneking M. 2005. Mitochondrial DNA and human evolution. *Annu. Rev. Genomics. Hum. Genet.*, 6: 165-183.

Parson W. & Bandelt H.J. 2007. Extended guidelines for mtDNA typing of population data in forensic science. *Forensic Sci. Int. Genet.*, 1: 13-19.

Parson W. & Dür A. 2007. EMPOP--a forensic mtDNA database. *Forensic Sci. Int. Genet.*, 1: 88-92.

Parson W., Brandstätter A., Alonso A., Brandt N., Brinkmann B., Carracedo A., Corach D., Froment O., Furac I., Grzybowski T., Hedberg K., Keyser-Tracqui C., Kupiec T., Lutz-Bonengel S., Mevag B., Ploski R., Schmitter H., Schneider P., Syndercombe-Court D., Sørensen E., Thew H., Tully G. & Scheithauer R. 2004. The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises:organisation, results and perspectives. *Forensic Sci. Int.*, 139: 215-226.

Pearson W.R. & Lipman D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.*, 85: 2444-2448.

Pereira L. & Samuels D.C. 2009. Response to Yao et al.. *Am. J. Hum. Genet.*, 85: 933.

Prieto L., Alves C., Zimmermann B., Tagliabracci A., Prieto V., Montesino M., Whittle M.R., Anjos M.J., Cardoso S., Heinrichs B., Hernandez A., López-Parra A.M., Sala A., Saragoni V.G., Burgos G., Marino M., Paredes M., Mora-Torres C.A., Angulo R., Chemale G., Vullo C., Sánchez-Simón M., Comas D., Puente J., López-Cubría C.M., Modesti N., Aler M., Merigioli S., Betancor E., Pedrosa S., Plaza G., Masciovecchio M.V., Schneider P.M. & Parson W. 2012. GHEP-ISFG proficiency test 2011: Paper challenge on evaluation of mitochondrial DNA results. *Forensic. Sci. Int. Genet.* (in press) DOI: 10.1016/j.fsigen.2012.04.006.

Rizzi E., Lari M., Gigli E., De Bellis G. & Caramelli D. 2012. Ancient DNA studies: new perspectives on old samples. *Genet. Sel. Evol.*, 44:21.

Rosa A. & Brehem A. 2011. African human mtDNA phylogeography at-a-glance. *J. Anthropol. Sci.*, 89: 25-58.

Sullivan K.M., Hopgood R., Lang, B. & Gill P. 1991. Automated amplification and sequencing of human mitochondrial DNA. *Electrophoresis*, 12: 17–21.

Tanaka M., Takeyasu T., Fuku N., Li-Jun G. & Kurata M. 2004. Mitochondrial genome single nucleotide polymorphisms and their phenotypes in the Japanese. *Ann. N.Y. Acad. Sci.*, 1011: 7-20.

Tofanelli S., Taglioli L., Merlitti D. & Paoli G. 2011. Tools which simulate the evolution of uni-parentally transmitted elements of the human genome. *J. Anthropol. Sci.*, 89: 201-219.

van Oven M. & Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.*, 30: E386-394.

Willuweit S. & Roewer L.2007. Y chromosome haplotype reference database (YHRD): update. *Forensic Sci. Int. Genet.*, 1:83-87.

Yao Y.G., Salas A., Logan I. & Bandelt H.J. 2009. mtDNA data mining in GenBank needs surveying. *Am. J. Hum. Genet.*, 85: 929-933.

Appendix 1A - Information on polymorphisms and data source available in online databases for mtDNA polymorphisms in human populations.

DATABASE	POLYMORPHISM				DATA SOURCE	OVERLAP WITH NCBI[1]
	HVRI	HVRII	CODING REGION	WHOLE GENOME		
DNA-Fingerprint	1303	1303	-	-	Samples sent from individual donors	-
Empop	17321	17321	-	-	Literature and unpublished data[2]	p[3]
Family Tree Dna	146153	64221[4]	19324[4]	-	Samples sent from individual donors[5]	p
GOBASE	-	-	-	2714	Literature	p
HmtDB	-	-	975	8813	Literature	c[6]
HVRbase++	13873	4940	-	1376	Literature	p
Mitosearch	44741	15555[4]	-	-	From various companies[6]	n.a.
MitoVariome	-	-	-	5344	Literature	c
mtDB	-	-	839	1865	Literature	c
mtDNA Community	-	-	-	13492	Literature and unpublished data	p
mtDNA manager	9294	9294	-	-	Literature	p
mtSNP database	-	-	-	1736	Literature	c
PhyloTree	-	-	943	14508	Literature	p
SMGF	75406	75406	-	-	Samples sent from individual donors	n.a.

[1] Using an ad hoc mitomap link, GenBank was found to include a total 16,414 complete DNA sequences (Database accessed on 20/09/2012). The number of HVR-1 and HVR-2 data available cannot be reliably calculated .

[2] Most sequences are based on raw data, not (only) literature (W. Parson, personal communication).

[3] p stands for partial overlap.

[4] Personal communication, B. Greenspan.

[5] Includes individual data from the Genographic project.

[6] c stands for complete overlap.

[7] SMGF, Family Tree, Ancestry, Genographic and others unspecified.

Appendix 1B - Downloading and uploading from online databases for mtDNA polymorphisms in human populations.

DATABASE	DOWNLOADING	FILE FORMAT	UPLOADING POPULATION DATA	FILE FORMAT
DNA-Fingerprint	p[1]	ns[2]	-	
Empop	-	-	x[3]	Empop
Family Tree Dna[4]	p	ns	-	
GOBASE	x	Fasta	-	
HmtDB	x	Fasta	-	
HVRbase++	x	Various formats[5]	-	
Mitosearch	x	ns	-	
MitoVariome	x	Fasta	-	
mtDB	x	Fasta	-	
mtDNA Community	x	ns	x	ns
mtDNA manager	p	ns	-	
mtSNP database	x	Fasta	-	
PhyloTree	x	Fasta	-	
SMGF	-	-	-	

[1] p stands for the possibility to download only partial data.
[2] ns, non specific; no formally defined format was used.
[3] x stands for presence of a specific feature.
[4] Data from Genographic project may be uploaded.
[5] Fasta, Genbank, HTML, Metadata, Phylip, TSV, XML

Appendix 1C - Matching and other tools available in online databases for mtDNA polymorphisms in human populations.

DATABASE	MATCH			OTHER POPULATION GENETICS TOOLS
	COMPLETE	PARTIAL	RANGE	
DNA-Fingerprint	x[1]	-	-	-
Empop	x	x	x	Drawing quasi-median network
Family Tree Dna	-	-	-	-
GOBASE	-	-	-	-
HmtDB	-	-	-	Haplogroup predictor
HVRbase++	-	-	-	Newick-tree viwer, convert numbering
Mitosearch	x	-	-	-
MitoVariome	-	-	-	-
mtDB	-	-	-	-
mtDNA Community	x	-	-	FastmtDNA, mtDNAble
mtDNA manager	x	x	x	Haplogroup estimation, match probability
mtSNP database	-	-	-	mtSNP statistics
PhyloTree	-	-	-	-
SMGF	x	x	x	-

[1] x stands for presence of a specific feature.

Appendix 1D - Filters in online databases for mtDNA polymorphisms available in human populations.

DATABASE	MUTATION	IN/DEL	HG	REFERENCE	GEOGRAPHICAL LOCATION	ACCESSION NUMBER	SAMPLE INFORMATION				MULTIPLE QUERIES
							DISEASE STATE	GENDER	AGE	TISSUE	
DNA-Fingerprint	-	-	x[1]	-	x	-	-	-	-	-	x
Empop	x	x	-	x	x	x	-	-	-	-	x
Family Tree Dna	-	-	t[2]	-	t	-	-	-	-	-	-
GOBASE	-	-	x	x	-	x	x	-	-	-	x
HmtDB	x	x	x	x	x	x	x	x	x	x	x
HVRbase++	x	-	x	x	x	-	-	-	-	-	x
Mitosearch	-	-	x	-	x	-	-	-	-	-	x
MitoVariome	-	-	x	-	x	x	-	-	-	-	x
mtDB	-	-	-	t	t	t	-	-	-	-	-
mtDNA Community	x	-	x	-	-	x	-	-	-	-	-
mtDNA manager	-	x	-	-	x	-	-	-	-	-	x
mtSNP database	x	x	-	x	x	x	-	x	x	x	x
PhyloTree	-	-	-	t	-	-	-	-	-	-	-
SMGF	-	-	-	-	-	-	-	-	-	-	-

[1] x stands for presence of a specific feature.

[2] t stands for reference table.

Appendix 2A - Information on polymorphisms and data source available in online databases for Y chromosome polymorphisms in human populations.

DATABASE	POLYMORPHISM		DATA SOURCE
	STR HAPLOTYPES	STR/SNP HAPLOTYPES	
DNA-Fingerprint	1305	114	Samples sent from individual donors
Family Tree DNA	236302	62795[1]	Samples sent from individual donors[2]
SMGF	38447	-	Samples sent from individual donors
US Y-STR Database	18719	-	From various laboratories[3]
Y-Filer Haplotype Database	11393	-	n.a.
YHRD	101055	9039	Literature and unpublished data
Ysearch	112513	-	From various companies

[1] Personal communication, B. Greenspan.

[2] Includes individual data from the Genographic project.

[3] National Center for Forensic Science, ReliaGene, Promega, Applied Biosystems, University of Arizona, Illinois State Police, Orange County CA Coroner, Santa Clara County CA Crime Laboratory, California Department of Justice Sacramento Crime Lab, Marshall University Forensic Science Center, Washington State Patrol Crime Lab in Vancouver, Richland County Sheriff's DNA and Trace Department, San Diego Sheriff's Department.

Appendix 2B - Downloading and uploading from online databases for Y chromosome polymorphisms in human populations.

DATABASE	DOWNLOADING	FILE FORMAT	UPLOADING POPULATION DATA	FILE FORMAT
DNA-Fingerprint[1]	p[2]	ns[3]	-	ns
Family Tree DNA[4]	p	ns	-	ns
SMGF	-	ns	-	ns
US Y-STR Database	-	ns	-	ns
Y-Filer Haplotype Database	-	ns	-	ns
YHRD	-	ns	x[5]	ns
Ysearch[6]	x	ns	-	ns

[1] Some data are deleted from users.

[2] p stands for the possibility to download only partial data.

[3] ns, non specific; no formally defined format was used.

[4] Data from Genographic project may be uploaded.

[5] x stands for presence of a specific feature.

[6] Accepts everyone's data, providing a minimum of 8 STR markers (B. Greenspan, personal comunication 02/08/2012).

Appendix 2C - Matching and other tools available in online databases for Y chromosome polymorphisms in human populations.

DATABASE	MATCH		OTHER POPULATION GENETICS TOOLS
	COMPLETE	PARTIAL	
DNA-Fingerprint	x[1]	-	-
Family Tree DNA	-	-	-
SMGF	x	x	-
US Y-STR Database	x	-	Mixture analysis tools
Y-Filer Haplotype Database	x	-	-
YHRD	x	-	AMOVA, Mixture tool
Ysearch	x	x	Genetic distances,TMRCA calculator

[1] x stands for presence of a specific feature.

Appendix 2D - Filters available in online databases for Y chromosome polymorphisms in human populations.

DATABASE	SNPS	HG	REFERENCE	ANCESTRY	GEOGRAPHICAL LOCATION	ACCESSION NUMBER	SURNAME	MULTIPLE QUERIES
DNA-Fingerprint	x[1]	x	-	-	x	-	-	x
Family Tree DNA	-	-	-	-	t[2]	-	t	-
SMGF	-	-	-	-	-	-	x	-
US Y-STR Database	-	-	-	x	-	-	-	-
Y-Filer Haplotype Database	-	-	-	-	-	-	-	-
YHRD	x	-	x	-	x	x	-	x
Ysearch	-	x	-	-	x	-	x	x

[1] x stands for presence of a specific feature.

[2] t stands for reference table.