A Comparison of Publicly Accessible Y Chromosome Databases for Ancestral and Population Studies

Ugo A. Perego, Jayne E. Ekins, Robert Hughes, Ann Turner, Scott R. Woodward
Sorenson Molecular Genealogy Foundation, Salt Lake City, Utah, USA (www.SMGF.org)

Abstract

Y chromosome testing has been heavily utilized in academic settings for ancestral and population studies in recent years. As molecular technologies have become more affordable and publicly available, a number of searchable online genetic databases appeared to serve interested individuals. Among the largest and most comprehensive Y chromosome databases utilized by the public for personal ancestral research are the Y Chromosome Haplotype Reference Database (www.YHRD.org), the Sorenson Molecular Genealogy Foundation Y Chromosome Database (www.SMGF.org), and the Family Tree DNA Ysearch Database (www.YSEARCH.org), each containing thousands of haplotypes correlated to genealogical, historical and geographic data. This study provides a comparison of the unique array of applications available to the end user, often at no cost, from each database. All have the common goal of making powerful genetic tools available to the public, in order to promote discovery of recent and ancient paternal connections of the human family that are significant to the individual.

Introduction

In recent years, Y chromosome research has been the focus of many publications and studies, including population [1,2], historical [3] and genealogical studies [4,5,6]. Straightforward paternal inheritance (Figure 1) and the absence of recombination make the Y chromosome a valuable source for these types of studies. The vast amount of new knowledge that has become available through these investigations and the availability of consumer-oriented genetic testing offered by a number of laboratories in the USA and Europe [7] has generated a wave of interest towards DNA testing among non-academic circles. A large array of current DNA surname studies [8] has produced thousands of Y chromosome haplotypes available to individuals on the internet on personal websites or in publicly accessible databases. Moreover, two large projects are currently underway to collect hundreds of thousands of DNA samples representing populations and lineages around the world to further our understanding of the recent and deep history of mankind [9,10]. Personal genetic testing and the availability of searchable databases of correlated genetic and ancestral/geographical data complement each other, as they are both needed to further the understanding of Man’s own history. The paradox associated with consumer-based genetic testing is that in order to understand one’s own genetic history, one is required to know something about the genetic past of everyone else. Thus, once a person obtains a copy of his Y chromosome haplotype, a comparison to other haplotypes, each correlated with additional historical and geographical data, will often produce previously unavailable information about the individual’s paternal lineage. Thousands of searchable haplotypes are currently available in three large online Y chromosome databases available to both academic and non-academic settings. Together with other smaller databases, they are growing both in size and number of loci and are contributing significantly to the discovery of previously unknown paternal related lineages.

Y Chromosome Haplotype Reference Database*

The Y Chromosome Haplotype Research Database (YHRD) was established in 2000 by the joint efforts of 31 forensic and anthropological institutions [11] and offers the largest assortment of Y chromosomal STR minimal haplotypes (minHt) currently available online (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS395a,b). The latest release of YHRD contains 41,965 haplotypes tested at a minimum of 7 loci (minHt less DYS385a,b) representing 357 populations, including 24,497 Y chromosome profiles from males with European ancestry (Table 1). The database is composed of anonymous records contributed by a number of forensic laboratories worldwide who are required to pass a quality-control exercise. Users may query YHRD by entering the 9 loci from the minimal haplotype (40,108 records) or by adding two more markers (DYS438 and DYS439, 14,835 records) from the Scientific Working Group on DNA Analysis Methods (SWGDAM) core set (Figure 2). Distributions of haplotypes in various populations and metapopulations are presented as tables or maps (Figure 3), thus linking unique haplotypes to specific geographical areas and providing some indications of the spatial origin of paternal lineages.

Table 1: YHRD haplotype statistics

<table>
<thead>
<tr>
<th>Population</th>
<th>Haplotypes (minHt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>24,497</td>
</tr>
<tr>
<td>Worldwide</td>
<td>41,965</td>
</tr>
</tbody>
</table>

FTDNA Ysearch Database*

Ysearch is a publicly available Y chromosome repository containing approximately 32,000 haplotypes between 0 to 100 loci in length, with the majority tested at 12, 25, and 37 markers (Figure 4). Customers of Houston-based Family Tree DNA and participants in the National Geographic’s Genographic Project are given the option to include their haplotypes in the database [12]. Additionally, individuals who have been tested with other laboratories may upload their Y chromosome data. Those who decide to have their genetic information included in this database are given the option to be contacted by those who may match their haplotype. The database offers a feature for uploading GEDCOM files containing genealogical data linked to the donor’s Y chromosome profile (3,482 records). Comparison between haplotypes is displayed in the form of tables (Figure 5). Features also include haplogroup comparison, genetic distance reports, and search by surname. However, since the database contains large amounts of self-reported genetic and genealogical data, assessing the accuracy of the information retrieved is a difficult process.

Sorenson Molecular Genealogy Foundation Y Chromosome Database*

The Sorenson Molecular Genealogy Foundation (SMGF) is a non-profit organization dedicated to building free databases of DNA corresponding pedigree data. In 2000, SMGF began collecting genetic and genealogical data from volunteer participants worldwide who could supply a minimum of four generations of biological genealogical information. A subset of the DNA samples collected by SMGF was released on March 2004 in the form of a searchable database of Y chromosome haplotypes and family histories. The current version of the SMGF Y Chromosome Database include 15,569 haplotypes [13] of which nearly 80% have been typed at 37 loci (multipoly c有序 loci such as DYS385, DYS395, DYS464 and YCAI are counted as one marker). The SMGF haplotypes represent 106 countries and are linked to 3,348,541 ancestral records (Figure 6). Individuals querying the SMGF Y Chromosome Database have the option to do a search by haplotype or by surname. With each match displayed in the result page, Time to the Most Recent Common Ancestor (TMRCA) and pedigree data can also be viewed. The database includes additional features such as Place/Time Analysis (Figure 7), database statistics, molecular genealogy tutorials, and instructions on how to request the no-cost participation kit for those interested in being included in the database. In addition to the Y chromosome, each DNA sample submitted to SMGF is genotyped for inclusion in mtDNA and autosomal DNA databases.

Figure 1: Y chromosome haplotype count

Figure 2: YHRD haplotype count

Figure 3: Ysearch example of haplotype distribution

Figure 4: Ysearch haplotype count

Figure 5: Display of results from Ysearch.org

Figure 6: SMGF haplotype count

Figure 7: SMGF Place/Time Analysis

* Data from each database was accessed during the month of September 2006.

References

12. DNA databases for autosomal DNA databases.
14. DNA databases for Y chromosomes are generated by the National Geographic’s Genographic Project.
15. DNA databases for the SMGF Y Chromosome Database are generated by Sorenson Molecular Genealogical, Salt Lake City, Utah (www.SorensonGenealogy.com).

Table: YHRD haplotype statistics

<table>
<thead>
<tr>
<th>Population</th>
<th>Haplotypes (minHt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>24,497</td>
</tr>
<tr>
<td>Worldwide</td>
<td>41,965</td>
</tr>
</tbody>
</table>

Figure 1: Y chromosome haplotype count

Figure 2: YHRD haplotype count

Figure 3: Ysearch example of haplotype distribution

Figure 4: Ysearch haplotype count

Figure 5: Display of results from Ysearch.org

Figure 6: SMGF haplotype count

Figure 7: SMGF Place/Time Analysis