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)



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.

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 over the internet on personal websites or in publicly accessible databases. Moreover, two large projects are currently underway to collect hundred 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 paternally related lineages.

Metapopulation	# Haplotypes	# distinct Haplotypes	# Populations	lvg. # haplotypes per population
Westdwide	41,965	111,560	357	117,54901
Eurasian MP	29,781	12,296	254	122,053276
Eurapian MP / European MP	24,497	0.901	178	137,5236
Eurasian MP / Altaic MP	2,038	1,376	26	78.38461
Eurasian MP / Caucasian MP	502	395	13	38.615383
Eurasian MP / Uralic MP	399	145	-1	399
Eurasian MP / Indo Iranian MP	1,183	633	19	67.213157
Eurasian MP / Indian MP	1,162	823	7	166
East Asian MP	6,698	4,106	49	136,09389
East Asian MP / Korean MP	384	336		120
East Asian MP / Japanese MP	545	449	5	109
East Asian MP / Sino Tibetan MP	4050	2,289	24	188.75
East Asian MP / Austroasiatic MP	313	246	4	18.25
East Asian MP / Thai MP	112	109	1	117
East Asian MP / Austronesian MP	1,142	918	10	114.7
East Asian MP / Indo Pacific MP	29	26	1	21
East Asian MP / Dravidian MP	122	112	1	127
Australian Aboriginal MP	0	0	0	NaN
African MP	2.019	1,549	25	102.70
African MP / Subsaharan MP	1112	633	9	123,65556
African MP / Afro-American MP	1,310	881	14	93.9142
African MP / Afro-Caribbean MP	397	328	2	1985
Amerindian MP	283	215	10	28.7
Eskimo Aleut MP	70	46	1	76
Admixed MP	1,296	842	18	72
Afroeuranian MP	1,018	662	10	10 8.6
Afrogurasian MP / Semitic MP	787	578	0	98.375
Afroeurasian MP / Berber MP	30	20	1	30
Afternoonies ME / Cushitis MD	204	97		201

Table 1: YHRD haplotype statistics

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



Figure 5: Display of results from Ysearch.org

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, DYS389I, DYS390, DYS391, DYS392, DYS393, DYS385a,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



Figure 2: YHRD haplotype count

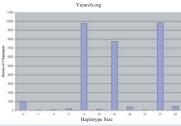


Figure 4: Ysearch haplotype count



Figure 6: SMGF haplotype count

- I. P.A. Underhill, et al. 2000. C Thomosome Squence Variation and the History of Human Populations. Nature Genetics. 26, 358-361.

 2. M.A. Johling and C. Dyler-Smith. 2003. The Human V Chromosome An Evolutionary Marker Comes to Age. Nature Reviews. Genetics. 4, 598-612.

 2. M.A. Johling and C. Dyler-Smith. 2003. The Human V Chromosome: An Evolutionary Marker Comes to Age. Nature Reviews. Genetics. 4, 598-612.

 2. H. E. Syless and C. Liven. 2000. Surnames and the Y Chromosome. American Journal of Human Genetics. 66, 1417-1419.

 3. M.A. Johling, 2010. In the Name of the Fathers Summass and Genetics. T. 873-535.

 5. T.E. King et al. 2006. Genetic Signatures of Coancestry within Summaness. Current Biology. 16, 384-388.

 7. H. E. King et al. 2006. Genetic Signatures of Coancestry within Summanes. Current Biology. 16, 384-388.

 7. H. E. King et al. 2006. Genetic Signatures of Coancestry within Summanes. Current Biology. 16, 384-388.

 7. H. E. King et al. 2006. Genetic Signatures of Coancestry within Summanes. Current Biology. 16, 384-388.

 7. H. E. King et al. 2006. Genetic Signatures of Coancestry within Summanes. Current Biology. 16, 384-388.

 7. H. E. King et al. 2006. Genetic Signatures of Coancestry within Summanes. Current Biology. 16, 384-388.

 7. H. E. King et al. 2006. Genetic Signatures of Coancestry within Summanes. Current Biology. 16, 384-388.

 8. Cyndis Last, http://www.cyndislist.com/sum-dan/summanes. Current Biology. 16, 384-388.

 9. National Geographic. woman contained geographic Conference of Coancestry within Summanes. Current Biology. 16, 384-388.

 10. Some Signature of Coancestry within Summanes. Current Biology. 16, 384-388.

 11. R. E. King et al. 2006. Genetic Signatures of Coancestry within Summanes. Current Biology. 16, 384-388.

 12. H. L. Rower, et al. 2001. Children Reference Database of European Y-Chromosomal Short Tandem Repeat (STR) Haplotypes. Forensis Science International. 118, 106-113.

 12. DNA haplotypes for te SNGC Y Chromosomal Short Tandem Repeat (STR) Haplotypes. Forensis Scienc

Figure 3: YHRD example of haplotype distribution

Sorenson Molecular Genealogy Foundation Y Chromosome Database*

The Sorenson Molecular Genealogy Foundation (SMGF) is a non-profit organization dedicated to building the largest collection of DNA and 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 correlated 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 (multicopy loci such as DYS385, DYS459, DYS464 and YCAII 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.

