An Updated World-Wide Characterization of the Cohen Modal Haplotype

J.E. Ekins[1], E.N. Tinah[2], N.M. Myres[1], K.H. Ritchie[1], U.A. Perego[1], J.B. Ekins[1], L.A.D. Hutchison[1], L. Layton[1], M.L. Lunt[1], S.S. Masek[1], A.A. Nelson[1], M.E. Nelson[1], K.L. Pennington[1], J.L. Peterson[1], T. Tolley[1], S.R. Woodward[1] 1) Sorenson Molecular Genealogy Foundation (SMGF), Salt Lake City, UT, USA; 2) Department of Micro and Molecular Biology, Brigham Young University, Provo, UT, USA

Abstract

Since the definition of the Cohen Modal Haplotype (CMH) in 1998, the 6 SNP-6 STR genetic motif has been utilized to infer connections of contemporary individuals and communities to the ancient Hebrew population. The elucidation of the YCC SNP Phylogeny has allowed cataloguing of chromosomes compatible with the original CMH definition into several different Y-SNP subclades. Haplogroup membership was determined for 283 samples matching at ≥5 of the CMH STR alleles, defined as the Cohen Modal Haplogroup (CMHg). The bulk of the CMHg chromosomes were observed in J1 (54.1%) and J2 (41.7%), with a small portion falling outside of haplogroup J (4.2%). Members of the CMHg were observed throughout the world, with significant frequencies in various Arab populations: Yemen (34.2%), Oman (22.8%), Iraq (19.2%), Palestine (8.0%). Coalescent simulations were performed for CMH chromosomes within each SNP haplogroup using 11 STR loci. Estimates within J1 [8.7ky(4.1-21.4ky)] and J2 [17.9ky(8.8-39.6ky)] were substantially deeper than previous figures obtained from a heavily weighted Jewish sampling, indicating a likely origin of the compound haplotype prior to the establishment of the Hebrew population. The significant presence of CMH chromosomes in deeply divergent clades J1 and J2 (>25ky), indicates the present CMH definition is not sufficient to distinguish lineages that likely arose by parallel IBS mutations. An expanded STR definition is proposed which allows differentiation between CMH-compatible chromosomes in J1 and J2. The inference of Jewish ancestry based on the original CMH definition should be performed with caution as subjects may be falsely categorized into the eponymous CMH lineage when the true origin is in the deeply divergent IBS branch. These observations underscore the importance of using updated SNP classifications when utilizing the CMH to infer ancestry in Jewish populations, or the use of the expanded STR definition.

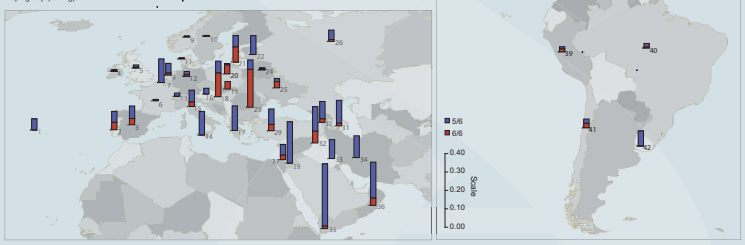
Geographic and Ethno-Religious Survey

The originally defined 6 SNP-6 STR CMH (Table 1) was first observed in Jewish populations, and has since been identified elsewhere in Africa, various parts of the Middle East, and Europe (Thomas et al. 2000; Nebel et al. 2000, 2001; Behar et al. 2003; Bonne-Tamir 2003; Shen 2004). It has since been used to infer relation to the ancient Hebrew population in individuals and groups possessing it. A world-wide sampling of 12,372 individuals with known paternal geographic origin was surveyed for occurrence of chromosomes compatible with at least 5 of the 6 STR alleles defined as members of the CMHg (Thomas et al. 1998). Data was gathered from both newly produced, and previously published sources (Nebel 2000, 2001; Thomas 2000; Weale 2001; Behar 2003; Cinnioglu 2004; Shen 2004; Goncalves 2005), comprising a total geographic coverage of 128 countries from all 6 inhabited continents. CMHg-compatible chromosomes were observed in 49 countries in Eurasia, Africa, America, and Australia. All samplings were reduced for known relatedness such that a known familial group was represented by a single individual. The random geographic sampling was not overly biased by religious or ethnic selection.

Locus	Allele
STR	
DYS19/394	14
DYS388	16
DYS390	23
DYS391	10
DYS392	11
DYS393	12
SNP	
YAP	-
SRY+4064	G
sY81	A
SRY+465	C
92R7	C
Tat	T

Table 1. Original CMH Definition. The originally defined haplotype includes 6 SNP and 6 STR markers. Individuals matching at all SNP markers and only 5 of the 6 STR alleles are termed as belonging to the Cohen Modal Haplogroup (CMHg).

Of note in the geographic distribution of CMHg chromosomes is the high occurrence in Middle Eastern regions that are not traditionally considered admixed with mainstream Jewish populations. CMHg types are seen in high frequency in Yemen (34.2%), Oman (22.8%), Negev (21.9%), and Iraq (19.2%) populations with a presence also in various parts of the Levant and Anatolia, although with lower incidence. The great majority of the observations of CMHg-compatible chromosomes in these geographically defined Middle Eastern populations occurs in individuals harboring only 5/6 CMH STR alleles. In contrast, individuals matching at 6/6 of the CMH markers are found in the greatest proportions in Eastern European populations. This observation is compatible with known Jewish diasporan movements into that area. CMHg-compatible chromosomes, both 5/6 and 6/6, were observed at lower levels throughout Europe and South America.



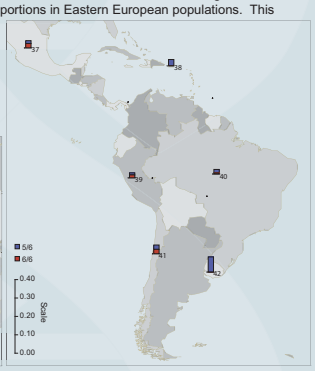


Figure 1. Frequency of CMH by Geography: Eurasia and America. Each stacked bar represents the proportion of the population exhibiting the CMH (5/6 or 6/6). Data are from published and newly generated sources. Population samplings were chosen that were not biased by religious or ethnic selection, and in general represent a random sampling of the geographic population. Populations are numbered as follows: [1]Acores, [2]Portugal, [3]Spain, [4]Ireland, [5]United Kingdom, [6]France, [7]Belgium, [8]Netherlands, [9]Norway, [10]Sweden, [11]Denmark, [12]Germany, [13]Switzerland, [14]Sicily, [15]Italy, [16]Austria, [17]Greece, [18]Hungary, [19]Slovakia, [20]Poland, [21]Lithuania, [22]Latvia, [23]Romania, [24]Belarus, [25]Ukraine, [26]Russia, [27]Palestine, [28]Negev, [29]Turkey, [30]Armenia, [31]Azerbaijan, [32]North Iraq, [33]Iraq, [34]Iran, [35]Yemen, [36]Oman, [37]Mexico, [38]Puerto Rico, [39]Peru, [40]Brazil, [41]Chile, [42]Uruguay. The scale is directly comparable to Figure 2.

Note: The CMH was also observed in Bolivia, Colombia, Samaria, Syria, and Turkmenistan, but with insufficient sampling to approximate a frequency in the population. Frequencies observed in Australia (0.012), and Canada (0.014) were not illustrated.

In agreement with previous observations (Thomas et al. 1998), both Sephardic and Ashkenazi self-identified Cohanim exhibited the highest frequency of the CMHg (65.2%-67.1%, 47.4%-52.2% 6/6) of any other subgrouping surveyed, ethnic or geographic. Other Jewish designations also present CMHg individuals, although at lower levels (<25%). Many other non-Jewish groupings have a comparable occurrence of the type: Muslim Kurd (22.1%), Bedouin (21.9%), and Armenian (12.7%). However the frequency of 6/6 CMH chromosomes is markedly higher in Jewish than in non-Jewish populations, where individuals matching at only 5/6 markers are most commonly observed. Note that in comparing estimates from Figures 1 and 2, the most frequent occurrence of CMHg-compatible chromosomes following the Cohanim occurs in Yemen, which is currently inhabited by a Muslim Arab population.

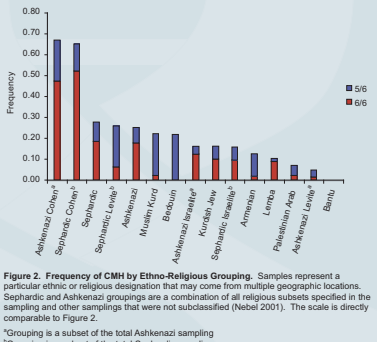


Figure 2. Frequency of CMH by Ethno-Religious Grouping. Samples represent a particular ethnic or religious designation that may come from multiple geographic locations. Sephardic and Ashkenazi groupings are a combination of all religious subsets specified in the sampling and other samplings that were not subclassified (Nebel 2001). The scale is directly comparable to Figure 2.

[a]Grouping is a subset of the total Ashkenazi sampling
[b]Grouping is a subset of the total Sephardic sampling

References

Behar D, Thomas MG, Skorecki K, Hammer MF, Bulygina E, Rosengarten D, Jones AL, Held K, Moses V, Goldstein D, Bradman N, Weale ME (2003) Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. Am J Hum Genet 73:768-779

Cinnioglu C, King R, Kivilsid T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA (2004) Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet 114:127-148

Di Giacomo F, Luca F, Popa LO, Akar N, Anagnou N, Banyko J, Brdicka R, Barbujani G, Papola F, Ciavarella G, Cucci F, Di Stasi L, Gavrila L, Kerimova MG, Kovatchev D, Kozlov AI, Loutradis A, Mandarino V, Mammi' C, Michalodimitrakis EN, Paoli G, Pappa KI, Pedicini G, Terrenato L, Tofanelli S, Malaspina P, Novelletto A (2004) Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet (2004) 115:357-371

Goncalves R, Freitas A, Branco M, Rosa A, Fernandes AT, Zhivotovsky LA, Underhill PA, Kivilsid T, Brehm A (2005) Y-chromosome lineages from Portugal, Madeira and Acores record elements of Sephardim and Berber Ancestry. Ann Hum Genet 69:1-12

Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598-612

Nebel A, Filon D, Weiss DA, Weale M, Faerman M, Oppenheim A, Thomas MG (2000) High-resolution Y chromosome haplotypes of Israeli and Palestinian Arabs reveal geographic substructure and substantial overlap with haplotypes of Jews. Hum Genet 107:630-641

Nebel A, Filon D, Brinkmann B, Majumder PP, Faerman M, Oppenheim A (2001) The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. Am J Hum Genet 69:1095-1112

Raymond M, Rousset F (1995) An exact test for population differentiation. Evolution 49:1280-1283

Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner P, Zhivotovsky LA, King R, Torroni A, Cavalli-Sforza LL, Underhill PA, Santachiara-Benerecetti AS (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet 74:1023-1034

Shen P, Lavi T, Kivilsid T, Chou V, Sengun D, Gefel D, Shpirer I, Woolf E, Hillel J, Feldman MW, Oefner PJ (2004) Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial sequence variation. Hum Mut 24:248-260

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament Priests. Nature 394:138-140

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N, Goldstein DB (2000) Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba-the "Black Jews of South Africa." Am J Hum Genet 66:674-686

Weale ME, Yepiskoposyan L, Jager RF, Hovhannisyan N, Khudoyan A, Burbage-Hall O, Bradman N, Thomas MG (2001) Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group. Hum Genet 109:659-674

Wilson IJ, Weale ME, Balding D (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J Roy Stat Soc: Ser A 166:155-188

Y-SNP Haplogroup Membership

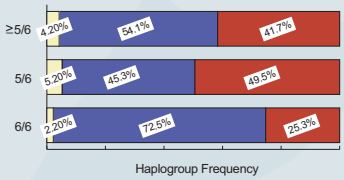


Figure 3. SNP haplogroup distribution of CMH. Test samples matching the original STR and SNP definition were found in both J1 and J2 subclades in similar frequency. When the STR definition was restricted to 6/6, matching individuals had a more frequent occurrence in J1. The presence of CMH-compatible chromosomes in two deeply divergent SNP clades suggests that the original CMH definition is not as specific to the Hebrew population as previously believed.

By applying the SNP markers of the original CMH definition (Table 1) to current YCC SNP Phylogeny nomenclature, the 'UEP group' in which the original CMH was found would now be termed as Y(xDE, N3, O2b, PR). The updated SNP Phylogeny allows selection of markers that provide more meaningful SNP designations for CMHg-chromosomes. A random sampling of 283 individuals matching the CMHg STR definition was assessed for SNP haplogroup affiliation. When examining all individuals having ≥5/6 CMH STR alleles, SNP haplogroup membership was observed in both J1 and J2 with near equal frequency. Examining individuals with 6/6 and 5/6 CMH matches separately produced fluctuations in these proportions (Figure 3). Test samples were of either European (51.6%) or Middle Eastern (48.4%) origin. An exact test for population differentiation (Raymond and Rousset 1995) indicated insignificant differences in SNP haplogroup frequencies among the geographies.

The original estimates for the coalescence of the CMH were based on a heavily weighted Jewish sampling, and placed the common origin point at 2.1-3.25ky (Thomas et al. 1998). J1 and J2 that have an estimated coalescence of 31.7ky (Semino et al. 2004), vastly predating the establishment of the ancient Hebrew population. The similarly frequent observations of CMH chromosomes in these two deeply divergent subclades suggest that the haplotype has less specificity to the Hebrew population than supposed initially.

As suggested by the coalescence time (2.1ky) of the CMH within Jews, it is likely that the original Cohanim in which this haplotype was detected belonged to a single SNP haplogroup, either J1 or J2. By increasing the stringency of CMH-compatibility to 6/6 STR matches, observed most frequently in the Cohanim (Figure 2), a greater proportion of individuals (72.5%) are observed in J1. This indicates that perhaps the eponymous CMH clade occurs within J1, although more testing is necessary to confirm this presumption. Despite the determination of the placement of the true CMH clade, it is evident from the similar proportions of CMHg-compatible individuals in J1 and J2, that using the current CMH definition to infer relation of individuals or groups to the Cohen or ancient Hebrew populations would produce many false-positive results.

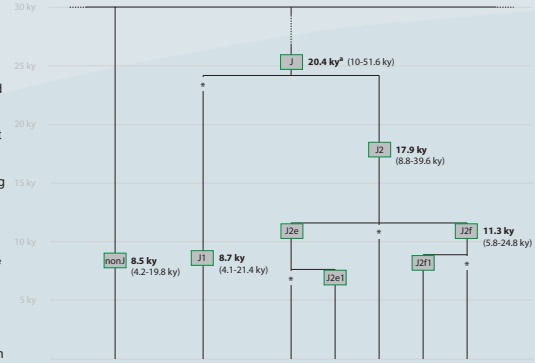


Figure 4. Coalescent times of CMH-compatible chromosomes within Y-SNP subclades. The subclade phylogeny and coalescent times of haplogroup J have been presented elsewhere [Jobling and Tyler-Smith (2003), Di Giacomo et al. (2004), Semino et al. (2004)]. Subhaplogroups in which CMH-compatible chromosomes were observed are shown on the backbone of the previously proposed coalescent time-scale of haplogroup J. The TMRCA of CMH-compatible chromosomes within each haplogroup is indicated where there was sufficient number of observations to perform the analysis.

[a]Coalescent time of haplogroup J is underestimated in comparison to Semino et al. (2004), likely due to the under-representation of the total diversity of J1 found in CMH-compatible chromosomes.

BATWING software (Wilson et al. 2003) and 11 non-CMH STR loci were used to estimate coalescence for CMHg individuals within SNP haplogroups (Figure 4). TMRCA for CMHg chromosomes within J2f and J2 are essentially identical to previous estimates using random samplings unrestricted to CMHg individuals (Di Giacomo et al. 2004, Semino et al.2004). In contrast, the approximation for CMHg chromosomes within J1 is significantly more shallow than others determined using unrestricted representatives of J1 (8.6ky vs. 24.1ky), but are deeper than estimates made using a biased Jewish sampling. The presence of the CMHg in J1 and J2, but with a comparatively shallow history in J1 suggests independent converging of types to the CMH (IBS). It is possible that the originally defined CMH represents a slight permutation of a more general Middle Eastern type that was established early on in the population prior to the divergence of haplogroup J. Under such conditions, parallel convergence in divergent clades to the same STR haplotype would be possible.

Extended STR Definition

18 additional STR loci were typed for individuals matching the original CMHg STR definition in J1 or J2. Each locus was examined separately for presence of distinguishing modal alleles to separate J1 CMHg chromosomes from J2. In loci with distinct modal alleles for J1 and J2, the allele was subjected to a chi-square test for dependence to test for association of the STR allele to SNP state within CMHg chromosomes. STR alleles that produced a significant probability of dependence on these conditions were determined to be valuable in distinguishing J1 CMHg chromosomes from J2. 7 loci were identified in this manner and are highlighted in Table 2. Employing the alleles identified as specific to J1 or J2 within CMHg individuals, as a test for distinguishing chromosomes belonging to the two different SNP clades produced quantifiable type I and type II errors. The null hypothesis in this case was that the individuals bearing the alleles identified as CMHg SNP-state specific, belonged to the corresponding SNP haplogroup. Type I and type II errors are quantified in Figure 5.

For alleles identified as specific to CMHg-J1 chromosomes, type I and type II errors were minimal when ≥5/7 alleles were used to identify individuals. For CMHg-J2 specific alleles, errors were minimized when ≥3/7 alleles were used to make inference for unknown individuals. These observations form the basis for an extended definition of the Cohen Modal Haplotype where CMHg-J1 individuals can be distinguished with accuracy from CMHg-J2 individuals by inference using 7 additional STR alleles.

Locus	J1 modal allele	J1 p-value	J2 modal allele	J2 p-value
DYS460	11	3.50E-08	10	5.69E-05
DYS459 haplotype	8-9	1.08E-09	9-9	1.08E-12
YGATAC4	≤21	1.77E-12	≥22	1.77E-12
DYS437	14	1.14E-16	15	1.82E-16
DYS461	10	2.11E-17	11	2.47E-07
DYS445	11	1.54E-22	12	2.35E-18
DYS458 insertion	+	3.07E-29	-	3.07E-29
DYS389B	17	2.14E-01	16	5.93E-04
YGATAA10	13	7.01E-02	12	3.08E-01
DYS385 haplotype	13-18	1.25E-02	13-16	1.44E-04
DYS459 allele	8	4.63E-08	9	5.30E-08
DYS385 allele	13	n/a	13	n/a
DYS389A	13	n/a	13	n/a
DYS439	11	n/a	11	n/a
DYS448	23	n/a	23	n/a
DYS454	11	n/a	11	n/a
DYS456	15	n/a	15	n/a
DYS458	17	n/a	17	n/a
DYS462	11	n/a	11	n/a
GGAAT1B07	11	n/a	11	n/a
YGATAC4	21	n/a	21	n/a
YGATAH4	11	n/a	11	n/a

Table 2. Assessment of alleles useful for distinguishing CMHg-J1 chromosomes from CMHg-J2. Loci with distinct modal alleles for CMHg individuals in J1 and J2, and also demonstrating significant dependence on the desired SNP state within CMHg chromosomes are highlighted.

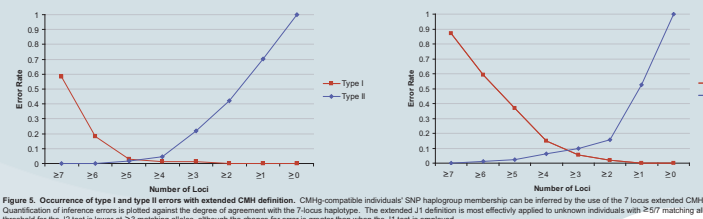


Figure 5. Occurrence of type I and type II errors with extended CMH definition. CMHg-compatible individuals' SNP haplogroup membership can be inferred by the use of the 7 locus extended CMH definition. Quantification of inference errors is plotted against the degree of agreement with the 7-locus haplotype. The extended J1 definition is most effectively applied to unknown individuals with ≥5/7 matching alleles. The threshold for the J2 test is lower at ≥3 matching alleles, although the chance for error is greater than when the J1 test is employed.

Conclusion

The Cohen Modal Haplotype is observed in high frequency within the Cohanim, but also presents with significant incidence in other non-Jewish populations. The occurrence of the CMH in deeply divergent SNP haplogroups also indicates a lack of specificity of the CMH to the ancient Hebrew population. As such, inference of relation to Jewish populations for individuals or groups should be performed with caution when using the original CMH definition, as a false-positive result is likely. The original definition in tandem with SNP testing focusing on the subhaplogroups of J, or the use of the extended STR definition will reduce the occurrence of error in such inferences.

SORENSON MOLECULAR GENEALOGY FOUNDATION

www.smgf.org