Resource

A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups

The Y Chromosome Consortium[1]

The Y chromosome contains the largest nonrecombining block in the human genome. By virtue of its many polymorphisms, it is now the most informative haplotyping system, with applications in evolutionary studies, forensics, medical genetics, and genealogical reconstruction. However, the emergence of several unrelated and nonsystematic nomenclatures for Y-chromosomal binary haplogroups is an increasing source of confusion. To resolve this issue, 245 markers were genotyped in a globally representative set of samples, 74 of which were males from the Y Chromosome Consortium cell line repository. A single most parsimonious phylogeny was constructed for the 153 binary haplogroups observed. A simple set of rules was developed to unambiguously label the different clades nested within this tree. This hierarchical nomenclature system supersedes and unifies past nomenclatures and allows the inclusion of additional mutations and haplogroups yet to be discovered.

[Supplementary Table 1, available as an online supplement at www.genome.org, lists all published markers included in this survey and primer information.]

In recent years, an explosion in data from the nonrecombining portion of the Y chromosome (NRY) in human populations has been witnessed. This explosion has been driven, in part, by the many recently discovered polymorphisms on the NRY. There has been a keen interest in using polymorphisms on the NRY to examine questions about paternal genetic relationships among human populations since the mid-1980s (Casanova et al. 1985). In more recent years, a use has been found for these polymorphisms in DNA forensics (Jobling et al. 1997), genealogical reconstruction (Jobling 2001), medical genetics (Jobling and Tyler-Smith 2000) and human evolutionary studies (Hammer and Zegura 1996). The low level of polymorphism on the NRY hindered research for many years. By the end of 1996, there were fewer than 60 known polymorphisms on the NRY. Most of these (~80%) were long-range polymorphisms (detectable by pulsed-field electrophoresis), conventional restriction fragment length polymorphisms (RFLPs), or short tandem repeats (STRs). Until 1997, there were only 11 known binary polymorphisms that could be genotyped by PCR-based methods (Hammer 1994; Seielstad et al. 1994; Hammer and Horai 1995; Whitfield et al. 1995; Santos et al. 1995; Jobling et al. 1996; Underhill et al. 1996). These included single nucleotide polymorphisms (SNPs), an Alu insertion polymorphism, and a deletion. Then, in 1997, Underhill et al. (1997) published 19 new PCR-based binary polymorphisms that were discovered by a novel and efficient mutation detection method known as denaturing high-performance liquid chromatography (DHPLC). This method has since been used to discover more than 200 SNPs and small insertions/deletions (indels) on the NRY (Shen et al. 2000, Underhill et al. 2000; Hammer et al. 2001). These polymorphisms are particularly useful because of their low rate of parallel and back mutation, which makes them suitable for identifying stable paternal lineages that can be traced back in time over thousands of years. As the number of known binary polymorphisms increased, so did the number of publications and the number of different systems used to name these binary haplogroups. Currently, there are at least seven different nomenclature systems in use, making it very difficult to compare results from one publication to the next. Our purpose here is twofold: (1) to construct a highly resolved tree of NRY binary haplogroups by genotyping most published PCR-based markers on a common set of samples, and (2) to describe a new nomenclature system that is flexible enough to allow the inevitable changes that will result from the discovery of new mutations and NRY lineages. We hope that the nomenclature presented here will be adopted by the community at large and will improve communication in this highly interdisciplinary field.

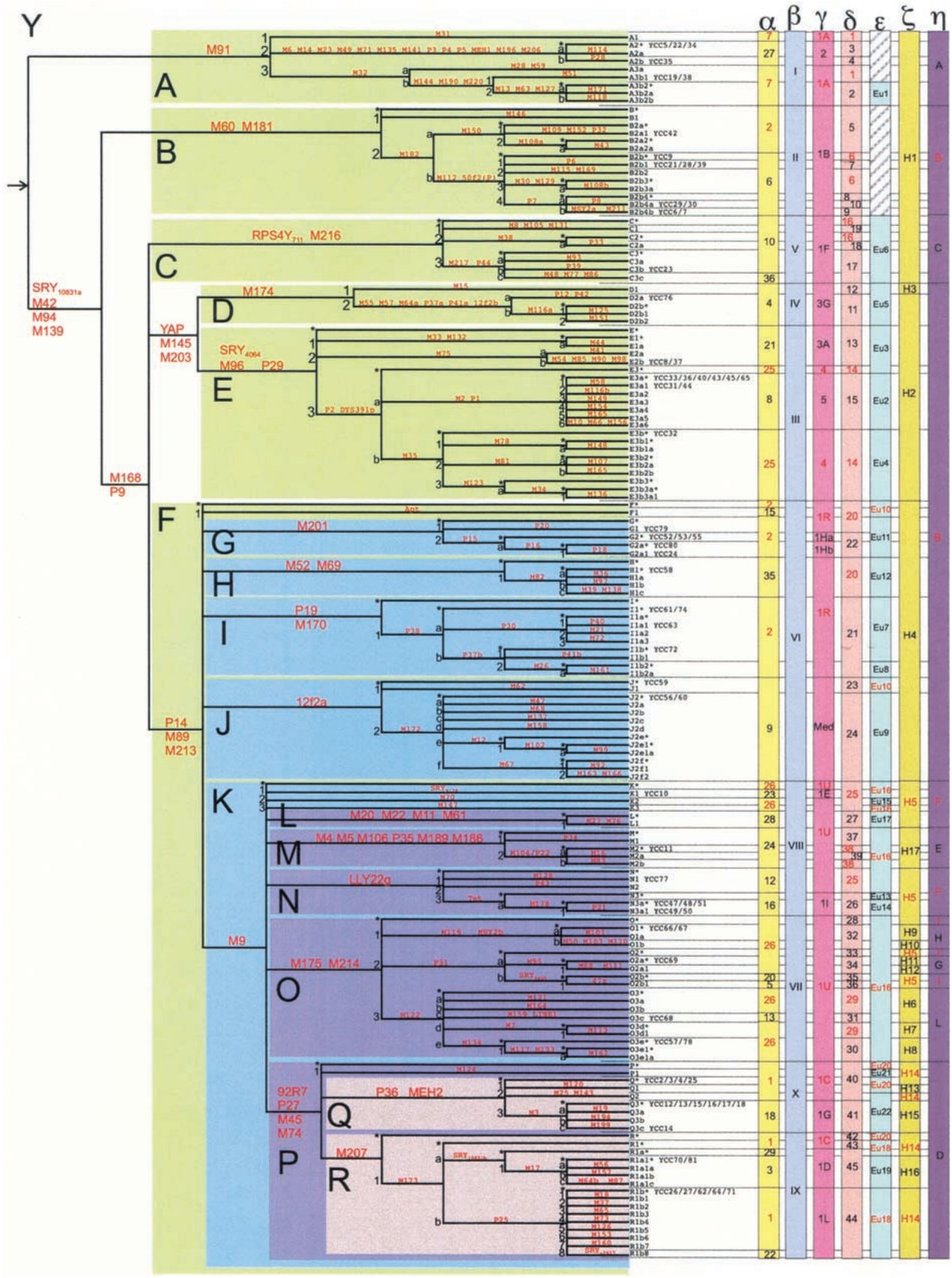
[1]See Acknowledgments for list of Consortium members.
Corresponding author: Michael Hammer, Department EEB, Biosciences West, University of Arizona, Tucson, Arizona 85721, USA.
E-MAIL mhammer@u.arizona.edu; FAX (520) 626-8050.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.217602.

RESULTS AND DISCUSSION

NRY Haplogroup Tree and Haplogroup Nomenclature

We constructed a comprehensive haplogroup tree for the human NRY by genotyping most of the known polymorphisms on the NRY in a single set of samples (74 male Y Chromosome Consortium [YCC] cell lines). Some polymorphisms known to be variable in other DNAs showed no variation in the YCC panel; therefore, additional samples were included to improve the resolution of the phylogeny. This served to increase the number of polymorphic sites mapped onto the haplogroup tree to 237. Two mutational events occurred at each of eight sites. However, these recurrent mutations were found on different haplogroup backgrounds and thus were distinguishable events. The 245 mutational events gave rise to 153 NRY haplogroups. The single most parsimonious tree for these 153 NRY haplogroups is shown in Figure 1, with mutational events shown along the branches.

The tree was drawn as asymmetrically as possible by sorting the descendants of each interior node so that the bottommost descendant had the greatest number of immediate descendants. The position of the root in Figure 1 (indicated by an arrow) was determined by outgroup comparisons. In other words, whenever possible, homologous regions on the NRY of closely related species (e.g., chimpanzees, gorillas, and oran-



(legend on facing page)

Figure 1 The single most parsimonious tree of 153 haplogroups (*left*) showing correspondences with prior nomenclatures (*right*). The root of the tree is denoted with an arrow. Haplogroup names and Y Chromosome Consortium (YCC) sample numbers are given at the tips of the tree, and major clades are labeled with large capital letters and shaded in color (the entire cladogram is designated haplogroup Y). The "*" symbol indicates an internal node on the tree or paragroup (see text). For space reasons, subclade labels are entered to the left of the corresponding links. Mutation names are given along the branches; major clades are labeled with a larger font than are their subclades. The length of each branch is not proportional to the number of mutations or the age of the mutation; each subclade is given a unit of depth in the tree. Some of the branches were elongated artificially to make room for a number of phylogenetically equivalent markers on a single branch. The order of phylogenetically equivalent markers shown on each branch is arbitrary. Prior nomenclatures are named according to author and are taken from the following publications: (α) Jobling and Tyler-Smith (2000) and Kaladjieva et al. (2001); (β) Underhill et al. (2000); (γ) Hammer et al. (2001); (δ) Karafet et al. (2001); (ε) Semino et al. (2000); (ζ) Su et al. (1999); and (η) Capelli et al. (2001). Noncontiguous naming systems in prior nomenclatures result either from the use of non-PCR markers that have not been typed on the YCC panel or unpublished lineage definitions. Prior haplogroup names shown in red are found in more than one position in the phylogeny. Cross-hatching within the "Semino" nomenclature indicates lineages that cannot be named according to their system. Mutations M104 and P22 on lineage M2 are independent discoveries of the same polymorphic marker.

gutans) were sequenced to determine the ancestral states at human polymorphic sites (Underhill et al. 2000, Hammer et al. 2001). The root of the tree falls between a clade defined by M91 and a clade defined by a set of markers: SRY_{10831a}, M42, M94, and M139. The NRY tree in Figure 1 can be seen as a series of nested monophyletic clades (i.e., a set of lineages related by a shared, derived state at a single or set of sites). To devise a nomenclature system at a reasonable scale, we assigned a capital letter to several of the major clades, beginning with the letter A (for the haplogroup above the position of the root in Fig. 1) and continuing through the alphabet to the letter R. The letter Y was assigned to the most inclusive haplogroup comprising haplogroups A–R. Deciding which clades are to receive the highest labeling level can only be, to some extent, arbitrary. Here, we label with single capital letters those clades that seem to us to represent the major divisions of human NRY diversity. Only 19 letters have been assigned to clades to allow for the possible expansion and further resolution of this phylogeny (the implications of which are discussed below).

We propose here two complementary nomenclatures. The first is hierarchical and uses selected aspects of set theory to enable clades at all levels to be named unambiguously. The capital letters (A–R) used to identify the major clades constitute the front symbols of all subsequent subclades (Fig. 1). Unlabeled clades can be named as the "join" of two subclades; for example, clade CR includes all chromosomes that share the derived state of the M168 and P9 polymorphisms. Note that this is distinct from the set theoretic "union," which, in the above example, would not define a monophyletic clade. Lineages that are not defined on the basis of a derived character represent interior nodes of the haplogroup tree and are potentially paraphyletic (i.e., they are comprised of basal lineages and monophyletic subclades). Thus, we suggest the term "paragroup" rather than haplogroup to describe these lineages. Paragroups are distinguished from haplogroups (i.e., monophyletic groupings) by using the * (star) symbol, which represents chromosomes belonging to a clade but not its sub-

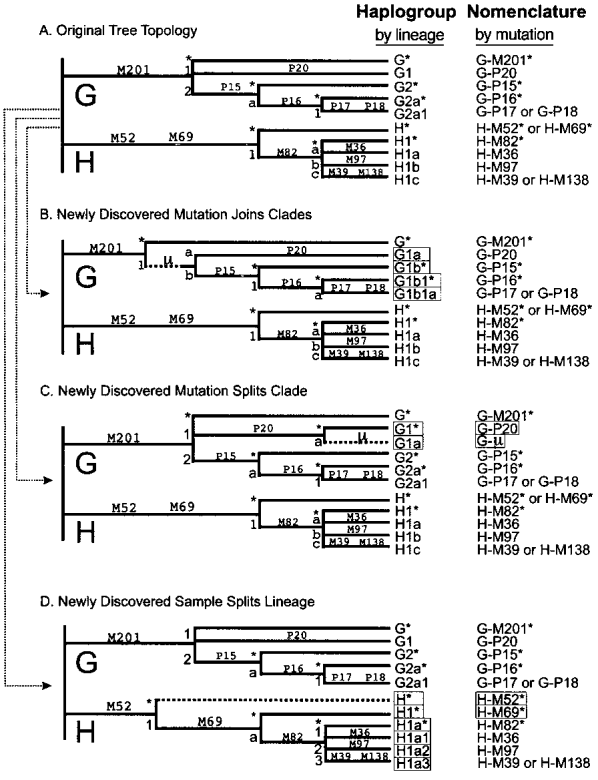


Figure 2 Potential examples of revisions in topology necessitated by the discovery of new mutations and new samples with intermediate haplogroups. Haplogroup nomenclature systems are shown to the right of the tree. (*A*) The G and H haplogroups are as shown in Figure 1. (*B*) Case of a newly discovered marker that joins haplogroups within haplogroup G. (*C*) Newly discovered mutation (μ) that splits clades within haplogroup G. (*D*) Case of a newly discovered sample with the derived state at M52 and the ancestral state at M69. Names shown in boxes indicate haplogroup names that require changes from those shown in *A*. Dotted lines indicate newly created lineages.

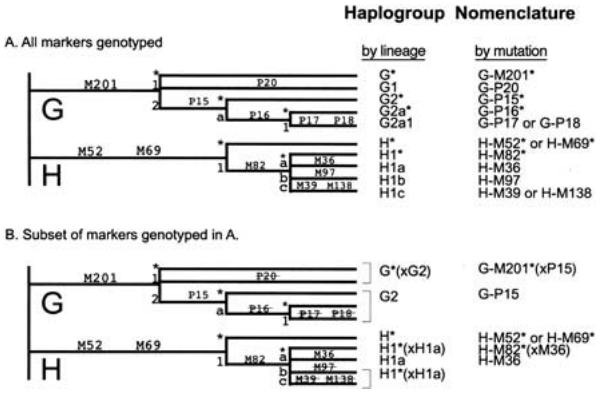


Figure 3 Examples of haplogroup names for cases in which subsets of markers in Figure 1 are genotyped. Markers that were not genotyped are shown with a strikethrough. The lineage- and mutation-based full nomenclature systems are shown to the right of the tree.

Table 1. Details of the Markers Incorporated within Six Published Prior Nomenclature Systems, Illustrated in Figure 1

System	Name	Derived state at	Ancestral state at	Name by lineage
Tyler-Smith & Jobling (2000)	1	92R7	M3, SRY_{10831b}, SRY_{-2627}	P*(xR1b8,R1a,Q3)
	2	SRY_{10831a}	50f2(P), $RPS4Y_{711}$, YAP Apt, M52, 12f2a, M9	BR*(xB2b,CE,F1,H,JK)
	3	SRY_{10831b}		R1a
	4	YAP	SRY_{4064}	DE*(xE)
	5	47z		O2b1
	6	50f2(P)		B2b
	7		SRY_{10831a}, MEH1	Y*(xBR,A2)
	8	M2		E3a
	9	12f2a		J
	10	$RPS4Y_{711}$		C
	12	LLY22g	Tat	N*(xN3)
	13	LINE1		O3c
	15	Apt		F1
	16	Tat		N3
	18	M3		Q3
	20	SRY_{+465}	47z	O2b*
	21	SRY_{4064}	P2	E*(xE3)
	22	SRY_{-2627}		R1b8
	23	SRY_{9138}		K1
	24	M4		M
	25	P2	M2	E3*(xE3a)
	26	M9	SRY_{9138}, M20, M4, LLY22g, SRY_{+465}, LINE1, 92R7	K*(xK1,LN,O2b,O3c,P)
	27	MEH1		A2
	28	M20		L
	35	M52		H
Underhill (2000)	I	M91		A
	II	M60		B
	III	M96		E
	IV	M174		D
	V	$RPS4Y_{711}$		C
	VI	M89	M9	F*(xK)
	VII	M175		O
	VIII	M9	M175, M45	K*(xO,P)
	IX	M173		R1
	X	M45	M173	P*(R1)
Hammer (2001)	1A		P3, SRY_{10831a}	Y*(xBR,A2)
	2	P3		A2
	1B	SRY_{10831a}	$RPS4Y_{711}$, YAP, P14	BR*(xF,DE,C)
	1C	P27	SRY_{10831b}, P25, M3	P*(xR1a,R1b,Q3)
	1D	SRY_{10831b}		R1a
	1E	SRY_{9138}		K1
	1F	$RPS4Y_{711}$		C
	1G	M3		Q3
	1Ha	P15	P16	G2*
	1Hb	P16		G2a
	1I	Tat		N3
	1L	P25		R1b
	1U	M9	P27, Tat	K*(xP,N3)
	1R	P14	12f2a, P15, M9	F*(xJ,G2,K)
	3G	YAP	SRY_{4064}	DE*(xE)
	3A	SRY_{4064}	P2	E*(xE3)
	4	P2	P1	E3*(xE3a)
	5	P1		E3a
	Med	12f2a		J
Karafet (2001)	1		SRY_{10831a}, M13, P3, P4, M6, M14	Y*(xBR,A2,A3b2)
	2	M13		A3b2
	3	P4, P3, M6, M14	SRY_{10831a}	A2*(xA2b)
	4	P28	SRY_{10831a}	A2b
	5	SRY_{10831a}	P9, 50f2(P)	BR*(xCR,B2b)
	6	50f2(P)	P6, P7, P8, MSY2a	B2b*(xB2b1,B2b4)
	7	P6		B2b1
	8	P7	P8, MSY2a	B2b4*
	9	MSY2a		B2b4b
	10	P8		B2b4a
	11	M174	M15	D*(xD1)
	12	M15		D1
	13	SRY_{4064}	P2, P1	E*(xE3)
	14	P2	P1	E3*(xE3a)

Table 1. (*Continued*)

System	Name	Derived state at	Ancestral state at	Name by lineage
	15	P1		E3a
	16	RPS4Y$_{711}$, M216	M8, M217, P33	C*(xC1,C2a,C3)
	17	M217		C3
	18	P33		C2a
	19	M8		C1
	20	P14	P15, P19, 12f2, M9	F*(xG2,I,J,K)
	21	P19		I
	22	P15		G2
	23	12f2a	M172	J*(xJ2)
	24	M172		J2
	25	M9	M20, M4, Tat, M175, P27	K*(xL,M,N3,O,P)
	26	Tat		N3
	27	M20		L
	28	M175	M119, P31, M122	O*
	29	M122	LINE-1, M134	O3*(xO3c,O3e)
	30	M134		O3e
	31	LINE-1		O3c
	32	M119, MSY2b		O1
	33	P31	M95, SRY$_{+465}$	O2*
	34	M95		O2a
	35	SRY$_{+465}$	47z	O2b*
	36	47z		O2b1
	37	M4, M5	P22	M*(xM2)
	38	P22	M16	M2*(xM2a)
	39	M16		M2a
	40	P27	M207	P*(xQ3,R)
	41	M3		Q3
	42	M207	M173	R*
	43	M173	SRY$_{10831b}$, P25	R1*
	44	P25		R1b*
	45	SRY$_{10831b}$		R1a*
Semino	Eu1	M13		A3b2
(2000)	Eu2	M2		E3a
	Eu3	SRY$_{4064}$	M2, M35	E*(xE3a,E3b)
	Eu4	M35		E3b
	Eu5	YAP	SRY$_{4064}$	DE*(xE)
	Eu6	RPS4Y$_{711}$		C
	Eu7	M170	M26	I*(xI1b2)
	Eu8	M26		I1b2
	Eu9	M172		J2
	Eu10	M89	M170, M172, M201, M69, M9	F*(xI,J2,G,H,K)
	Eu11	M201		G
	Eu12	M69		H
	Eu13	Tat	M178	N3*
	Eu14	M178		N3a
	Eu15	M70		K2
	Eu16	M9	M70, Tat, M11, M45	K*(xK2,N3,L,P)
	Eu17	M11		L
	Eu18	M173	M17	R1*(xR1a1)
	Eu19	M17		R1a1
	Eu20	M45	M173, M124, M3	P*(xR1,Q3,P1)
	Eu21	M124		P1
	Eu22	M3		Q3
Su	H1	M89,	YAP	Y*(xDE,F)
(1999)	H2	YAP	M15	DE*(xD1)
	H3	M15		D1
	H4	M89	M9	F*(xK)
	H5	M9	M122, M119, M95, M45, M5	K*(xO3,O1,O2a,P,M)
	H6	M122	M7, M134	O3*(xO3d,O3e)
	H7	M7		O3d
	H8	M134		O3e
	H9	M119	M50	O1*(xO1b)
	H10	M50		O1b
	H11	M95	M88	O2a*
	H12	M88		O2a1
	H13	M120		Q1
	H14	M45	M120, M3, M17	P*(xQ1,Q3,R1a1)
	H15	M3		Q3
	H16	M17		R1a1
	H17	M5		M

Table 1. (*Continued*)

System	Name	Derived state at	Ancestral state at	Name by lineage
Capelli	A		SRY_{10831a}	Y*(xBR)
(2001)	B	SRY_{10831a}	$RPS4Y_{711}$, M9	BR*(xC,K)
	C	$RPS4Y_{711}$		C
	D	92R7		P
	E	M4		M
	F	M9	92R7, M4, M175	K*(xP,M,O)
	G	M95		O2a
	H	M119		O1
	I	M175	M95, M119, M122	O*(xO1,O2a,O3)
	L	M122		O3

Lineages defined solely by the presence of a derived marker are monophyletic. Lineages defined by the presence of a derived marker and the absence of the derived state at other markers are potentially paraphyletic and under the present nomenclature are differentiated by an asterisk (*), unless they are known to be monophyletic.

clades. For example, paragroup B* belongs to the B clade; however, it does not fall into haplogroup B1 or B2. As illustrated in Figure 2, internal nodes are highly sensitive to changes in tree topology. Thus, the * symbol cautions that a given paragroup name may refer to different sets of chromosomes in succeeding versions of the phylogeny.

Subclades nested within each major haplogroup defined by a capital letter are named using an alternating alphanumeric system. For example, within haplogroup E, there are three basal haplogroups that are named E1, E2, and E3, and the underived paragroup becomes E*. Nested clades within each of these haplogroups are named in a similar way, except that lower-case letters are used instead of numerals. Again, paragroups are labeled with an * symbol, and the remaining haplogroups are labeled with an "a," "b," "c," etc. This naming system continues to alternate between numerals and lower-case letters until the most terminal branches are labeled (tip haplogroups). Therefore, the name of each haplogroup contains the information needed to find its location on the tree.

Alternatively, haplogroups can be named by the "mutations" that define lineages rather than by the "lineages" themselves. Thus, we propose a second nomenclature that retains the major haplogroup information (i.e., 19 capital letters) followed by the name of the terminal mutation that defines a given haplogroup. We distinguish haplogroup names identified "by mutation" from those identified "by lineage" by including a dash between the capital letter and the mutation name. For example, haplogroup H1a would be called H-M36 (Fig. 2). When multiple phylogenetically equivalent markers define a haplogroup, the one typed is used. For example, if M39 but not M138 were typed within haplogroup H1, then H1c becomes H-M39. If multiple equivalent markers were typed, this notation system omits some marker information, and a statement of which additional markers were typed should be included in the Methods section. Note that the mutation-based nomenclature has the important property of being more robust to changes in topology (Fig. 2).

While it is straightforward to name monophyletic clades, it is more challenging to devise a simple and flexible system to name underived interior nodes. This is especially important to facilitate the naming of haplogroups in studies where not all markers are typed, and to provide a standard set of names for previously described haplogroups (and paragroups). For instances where not all markers within a clade are typed, we introduce a bracketing system that encloses an "x" (for "excluding") and the lineages that have been shown to be absent. This system can be applied equally well to the lineage-based and mutation-based nomenclatures. The following examples portray the lineage-based nomenclature first, followed by the mutation-based nomenclature. Lineages (or markers) excluded from a haplogroup are listed within parentheses after the name of the haplogroup (or the last derived marker in the case of the mutation-based nomenclature). For example, if M82-derived chromosomes are typed with all downstream markers, then the underived chromosomes belong to H1* or H-M82* (Fig. 3A). However, if M82-derived chromosomes are typed only with M36, then the underived chromosomes belong to H1*(xH1a) or H-M82*(xM36) (Fig. 3B). If we apply this bracketing method to the naming of Underhill et al.'s (2000) paraphyletic haplogroup VI, then its label becomes F*(xK) or F-M89*(xM9) (Table 1). In the more extreme case of a study genotyping only the YAP and M3 markers, chromosomes ancestral for both markers would be named Y*(xDE,Q3) or Y*(xYAP,M3), where Y refers to the most inclusive haplogroup encompassing the total cladogram. See Table 1 for application of this bracketing system to lineage-based names of previously published haplogroups. When using the mutation-based nomenclature, the adoption of this bracketing system is optional, as long as full lineage-based names of haplogroups have been given elsewhere in the manuscript (e.g., in the form of a table or a tree). The lineage- and mutation-based nomenclatures each has advantages and disadvantages, and each can be used where most appropriate.

Cross-Referencing to Previous Nomenclatures

A number of investigators have developed nomenclature systems based on overlapping subsets of the markers typed here. To facilitate comparisons among seven previously published nomenclatures and our present proposed nomenclature, Figure 1 and Table 1 illustrate direct comparisons among these different systems. These nomenclature systems are extremely inconsistent (i.e., nonisomorphic) in how they define haplogroups. Moreover, when there is consistency between two systems (e.g., between Underhill et al.'s [2000] haplogroup V and Hammer et al.'s [2000] haplogroup 1F), different names are used for the same haplogroups. All of the major human NRY nomenclature schemes used thus far have included paraphyletic groupings (see Fig. 1), and these paragroups can be

misinterpreted as being necessarily ancestral to "downstream" haplogroups containing derived characters. Three major benefits of the proposed system are (1) its ability to distinguish between underived interior nodes (paragroups) and monophyletic clades (haplogroups), (2) its flexibility in naming haplogroups at different levels of the phylogenetic hierarchy, and (3) its ability to accommodate new haplogroups as new mutations are discovered (see below). If broadly accepted and utilized, this system also will serve to standardize the names of NRY haplogroups in the literature.

Caveats and Changes in Nomenclature

In addition to the long-term challenges posed by any attempt to form a stable nomenclature system, there are several caveats that should be raised relating to the way the current tree topology was inferred. First, it is important to point out that not all polymorphisms were genotyped in all individuals. Indeed, continued genotyping of these polymorphisms may result in slight changes in the topology of the tree in Figure 1. It is also possible that some mutational events that were assumed to be unique actually are recurrent on the tree (i.e., there are undetected multiple hits at some additional sites). More importantly, because it is extremely difficult to devise a nomenclature system that is both informative in a phylogenetic sense and impervious to the need for renaming groups as new polymorphisms are discovered, a set of guidelines is needed to minimize the impact of future structural changes in the tree.

To facilitate the evolution of the present nomenclature, we make a number of proposals. Firstly, a nomenclature committee comprising some of the current participants in the YCC will receive requests from investigators who wish new binary markers or haplogroups to be incorporated into the nomenclature, and will decide on the changes to be made to the existing system. At any one time, the current nomenclature and the committee's contact details will be made available on the following URL: http://ycc.biosci.arizona.edu. Consequently, we recommend that if investigators wish to use new markers prior to their incorporation into the nomenclature, they distinguish between consensus and novel parts of the clade labels by use of a forward slash. For example, a new mutation (μ) that divides clade D1 in two creates D1/-μ and D1/-M15*. This makes it clear to the reader which parts of the label are specific to that study and which can be cross-referenced to other publications. This will minimize confusion should two contemporaneous papers introduce novel markers within the same clade. In this manner, information from VNTR and STR haplotypes also can be incorporated; a standard nomenclature for Y-STRs already is available (Gill et al. 2001). Because new versions of the YCC nomenclature will be published annually to reflect changes in the tree topology resulting from newly discovered mutations, we suggest that each paper cite the particular version of the YCC NRY tree that was used (e.g., YCC NRY Tree 2002).

Summary

The cladistic nomenclature of human mtDNA diversity adopted by many groups some years ago has greatly advanced studies of maternal lineages and the communication of their conclusions (Richards et al. 1998). By contrast, recent dramatic advances in the resolution of paternal lineages have resulted in multiple nomenclature systems that have hampered communication among NRY researchers and the scientific community at large. Here, we introduce a strictly phylogenetic (cladistic) nomenclature for human NRY variation based on the phylogeny of 153 paternal lineages. This system is flexible in its ability to assign haplogroup names at different levels of the phylogenetic hierarchy. The phylogeny of the human NRY lies at the heart of a multidisciplinary enterprise in which unambiguous communication is vital. The nomenclature proposed here along with guidelines for revisions, represent an important resource to those interested in medical, forensic, and evolutionary genetics alike.

METHODS

YCC Cell Lines

The YCC is a collaborative group involved in an effort to detect and study genetic variation on the human NRY. The YCC was initiated in 1991 by Michael Hammer and Nathan Ellis with the following goals: (1) to establish a repository of lymphoblastoid cell lines (YCC cell line repository) derived from a sample of males representing worldwide populations, (2) to provide DNA isolated from these cell lines to investigators searching for polymorphisms on the NRY, and (3) to establish a common database containing the results of typing DNAs from the Repository cell lines at as many Y-specific polymorphic sites as possible (YCC Newsletter: http://www.ycc.biosci.arizona.edu/ycc1.html). Lymphoblastoid cell lines were established at the New York Blood Center from blood donated by volunteers who gave informed consent. Additional cell lines were donated by Luca Cavalli-Sforza, Trefor Jenkins, Judy Kidd, and Ken Kidd; or were purchased from the Coriell Institute. See Table 2 for a list of the YCC cell lines, as well as associated geographic, ethnic, and linguistic information.

Other DNA Samples

In constructing the tree, a great deal of phylogenetic information was retained from previous studies. When markers from different laboratories mapped on the same branch of the tree, an attempt was made to determine the order of mutational events. Toward this end, a variety of samples was provided by each of the participating laboratories, all of which were obtained with informed consent. These samples represented known haplogroups that were not present in the YCC cell line DNAs and thus served to map many additional markers on the haplogroup tree.

Genotyping SNPs and Indels

The protocols for genotyping many of the 237 polymorphic sites analyzed have been published (see Underhill et al. 2000, 2001; Hammer et al. 2001, and references therein); some of these assays were converted from conventional RFLPs and DNA sequence data (e.g., Jobling 1994; Hammer et al. 1997; Pandya et al. 1998; Bergen et al. 1999; Shinka et al. 1999; Bao et al. 2000). The remainder will be published in future manuscripts. Recurrent mutations, observed at SRY_{10831}, 12f2, MSY2, M116, M64, M108, P37, and P41 are counted as distinct polymorphisms. Supplementary Table 1 (available as an online supplement at http://www.genome.org) lists all published markers included in this survey and primer information.

Terminology

The terms "haplogroup" and "haplotype" have various, overlapping definitions in the literature. Here, we use the terminology of de Knijff (2000) in which "haplogroup" refers to NRY lineages defined by binary polymorphisms. The term "haplotype" is reserved for all sublineages of haplogroups that are defined by variation at STRs on the NRY (Y-STRs). Muta-

Table 2. Geographic/Ethnic Origins and Language Affiliations of YCC Cell Line Donors

YCC#	Geographic/ ethnic origin	Language affiliation	Cladistic Name	
			by lineage[a]	by mutation[b]
2	North America/Amerindian	Amerind	Q*	Q-P36*
3	North America/Amerindian	Amerind	Q*	Q-P36*
4	North America/Amerindian	Amerind	Q*	Q-P36*
5	Namibia/Tsumkwe San	!Kung	A2*	A-M6*
6	Banandu, CAR/Biaka	Aka	B2b4b	B-MSY2a
7	Banandu, CAR/Biaka	Aka	B2b4b	B-MSY2a
8	Ituri, Zaire/Mbuti	Niger/Kordofanian	E2b	E-M54
9	Ituri, Zaire/Mbuti	Niger/Kordofanian	B2b*	B-50f2(P)*
10	Solomon Islands/Melanesian	Nasioi	K1	K-SRY_{9138}
11	Solomon Islands/Melanesian	Nasioi	M2*	M-P22*
12	Rondonia, Brazil/Karitiana	Tupi	Q3*	Q-M3*
13	Rondonia, Brazil/Karitiana	Tupi	Q3*	Q-M3*
14	Rondonia, Brazil/Surui	Tpui	Q3c	Q-M199
15	Rondonia, Brazil/Surui	Tupi	Q3*	Q-M3*
16	Rondonia, Brazil/Surui	Tupi	Q3*	Q-M3*
17	Campeche, Yucatan/Mayan	Yucatec	Q3*	Q-M3*
18	Campeche, Yucatan/Mayan	Yucatec	Q3*	Q-M3*
19	Namibia/Tsumkwe San	!Kung	A3b1	A-M51
21	Namibia/Tsumkwe San	!Kung	B2b1	B-P6
22	Namibia/Tsumkwe San	!Kung	A2*	A-M6*
23	Arizona, US/Navajo	Navajo	C3b	C-P39
24	US/Ashkenazi Jew	English	G2a1	G-P18
25	Arizona, US/Tohono O'odham	Pima	Q*	Q-P36*
26	UK/English	English	R1b*	R-P25*
27	S. Carolina, US/Porch Creek	Porch Creek	R1b*	R-P25*
28	Namibia/Tsumkwe San	!Kung	B2b1	B-P6
29	Namibia/Tsumkwe San	!Kung	B2b4a	B-P8
30	Namibia/Tsumkwe San	!Kung	B2b4a	B-P8
31	South Africa/Herero	W. Bantu	E3a1	E-M58
32	South Africa/Sotho	E. Bantu	E3*	E-P2*
33	South Africa/Pedi	E. Bantu	E3a*	E-M2*
34	Namibia/Tsumkwe San	!Kung	A2*	A-M6*
35	Namibia/Tsumkwe San	!Kung	A2b	A-P28
36	South Africa/Tswana	E. Bantu	E3a*	E-M2*
37	South Africa/Ovambo	W. Bantu	E2b	E-M54
38	Namibia/Tsumkwe San	!Kung	A3b1	A-M51
39	Namibia/Tsumkwe San	!Kung	B2b1	B-P6
40	South Africa/Herero	W. Bantu	E3a*	E-M2*
42	South Africa/Zulu	E. Bantu	B2a1	B-P32
43	South Africa/Tswana	E. Bantu	E3a*	E-M2*
44	South Africa/Herero	W. Bantu	E3a1	E-M58
45	South Africa/Herero	W. Bantu	E3a*	E-M2*
47	Siberia/Yakut	Turkic	N3a*	N-M178*
48	Siberia/Yakut	Turkic	N3a*	N-M178*
49	Siberia/Yakut	Turkic	N3a1	N-P21
50	Siberia/Yakut	Turkic	N3a1	N-P21
51	Siberia/Yakut	Turkic	N3a*	N-M178*
52	Krasnador/Adygean	N. Caucasian	G2*	G-P15*
53	Krasnador/Adygean	N. Caucasian	G2*	G-P15*
55	Krasnador/Adygean	N. Caucasian	G2*	G-P15*
56	Krasnador/Adygean	N. Caucasian	J2*	J-M172*
57	Kashmir/Pakistani	Urdu	O3e*	O-M134*
58	Lahore/Pakistani	Punjabi	H1*	H-M82*
59	US/Ashkenazi Jew	English	J*	J-12f2a*
60	Multan/Pakistani	Punjabi	J2*	J-M172*
61	Germany/German	German	I1*	I-P38*
62	Germany/German	German	R1b*	R-P25*
63	Germany/German	German	I1a1	I-P40
64	Germany/German	German	R1b*	R-P25*
65	Ituri, Zaire/Mbuti	Niger/Kordofanian	E3a*	E-M2*
66	China/Han	Sino-Tibetan	O1*	O-M119*
67	China/Han	Sino-Tibetan	O1*	O-M119*
68	China-Han	Sino-Tibetan	O3c	O-LINE1
69	Cambodia/Khmer	Mon-Khmer	O2a*	O-M95*
70	Russia/Russian	Russian	R1a1*	R-M17*
71	Russia/Russian	Russian	R1b*	R-P25*
72	Russia/Russian	Russian	I1b*	I-P37b*

Table 2. (*Continued*)

YCC#	Geographic/ ethnic origin	Language affiliation	Cladistic Name	
			by lineage[a]	by mutation[b]
74	Russia/Russian	Russian	I1*	I-P38*
76	Japan/Japanese	Japanese	D2a	D-P42
77	Noto Peninsula/Japanese	Japanese	N1	N-M128
78	Gifu/Japanese	Japanese	O3e*	O-M134*
79	Turkey/Turkish	Turkic	G1	G-P20
80	US/Ashkenazi Jew	English	G2a*	G-P16*
81	US/Ashkenazi Jew	English	R1a1*	R-M17*

An asterisk (*) indicates an internal node on the tree or paragroup (see text).
[a]See Figure 1.
[b]See text and Figure 2.

tions labeled with the prefix "M" (standing for "mutation") were published by Underhill et al. (2000, 2001). Many of the mutations with the prefix "P" (standing for "polymorphism") were described by Hammer et al. (1998, 2001). The eight recurrent mutational events are indicated by their mutation name followed by a or b.

ACKNOWLEDGMENTS

The YCC wishes to thank the many people involved in this collaborative project. Following is a list of many of the contributors to this project and sources of funding.

YCC Organizers

Nathan Ellis (Memorial Sloan-Kettering Cancer Center), Michael Hammer (University of Arizona).

Genotyping

Michael Hammer (University of Arizona), Matthew E. Hurles (McDonald Institute for Archaeological Research), Mark A. Jobling (University of Leicester), Tatiana Karafet (University of Arizona), Turi E. King (University of Leicester), Peter de Knijff (Leiden University), Arpita Pandya (University of Oxford), Alan Redd (University of Arizona), Fabrício R. Santos (University of Oxford and Universidade Federal de Minas Gerais), Chris Tyler-Smith (University of Oxford), Peter Underhill (Stanford University), and Elizabeth Wood (University of Arizona). Mark Thomas (University College London) provided information on the order of the M17/SRY_{10831b} mutations.

Cell Lines

Luca Cavalli-Sforza (Stanford University), Nathan Ellis (Memorial Sloan-Kettering Cancer Center), Michael Hammer (University of Arizona), Trefor Jenkins (University of Witwatersrand), Judy Kidd (Yale University), Ken Kidd (Yale University).

Nomenclature Committee

Peter Forster (McDonald Institute for Archaeological Research), Michael Hammer (University of Arizona), Matthew E. Hurles (McDonald Institute for Archaeological Research), Mark A. Jobling (University of Leicester), Peter de Knijff (Leiden University), Chris Tyler-Smith (University of Oxford), Peter Underhill (Stanford University).

Helpful Discussions

Stephen Zegura (University of Arizona), Matthew Kaplan (University of Arizona).

This work was supported by grants from the National Science Foundation (OPP-9806759) and the National Institute of General Medical Sciences (GM53566) to MH; from the NIH (GM28428 and GM55273) to PAU and LCS, from the BBSRC to AP; from the Leverhulme Trust to FRS; and from the CRC to CTS. MAJ is a Wellcome Trust Senior Fellow in Basic Biomedical Science (grant number 057559). The Y Chromosome Consortium thank Colin Renfrew and the McDonald Institute for Archaeological Research for running a workshop attended by the members of the nomenclature committee at which many issues were resolved in a collaborative spirit.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bao, W., Zhu, S., Pandya, A., Zerjal, T., Xu, J., Shu, Q., Du, R., Yang, H., and Tyler-Smith, C. 2000. MSY2: A slowly evolving minisatellite on the human Y chromosome which provides a useful polymorphic marker in Chinese populations. *Gene* **244:** 29–33.

Bergen, A.W., Wang, C.Y., Tsai, J., Jefferson, K., Dey, C., Smith, K.D., Park, S.C., Tsai, S.J., and Goldman, D. 1999. An Asian-Native American paternal lineage identified by RPS4Y resequencing and by microsatellite haplotyping. *Ann. Hum. Genet.* **63:** 63–80.

Capelli, C., Wilson, J.F., Richards, M., Stumpf, M.P., Gratrix, F., Oppenheimer, S., Underhill, P., Pascali, V.L., Ko, T.M., and Goldstein, D.B. 2001. A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. *Am. J. Hum. Genet.* **68:** 432–443.

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

Gill, P., Brenner, C., Brinkmann, B., Budowle, B., Carracedo, A., Jobling, M.A., de Knijff, P., Kayser, M., Krawczak, M., Mayr, W.R., et al. 2001. DNA commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome STRs. *Int. J. Legal. Med.* **114:** 305–309.

Hammer, M.F. 1994. A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol.* **11:** 749–761.

Hammer, M.F. 1995. A recent common ancestry for human Y chromosomes. *Nature* **378:** 376–378.

Hammer, M.F. and Horai, S. 1995. Y chromosomal DNA variation and the peopling of Japan. *Am. J. Hum. Genet.* **56:** 951–962.

Hammer, M.F. and Zegura, S.L. 1996. The role of the Y chromosome in human evolutionary studies. *Evol. Anthropol.* **5:** 116–134.

Hammer, M.F., Spurdle, A.B., Karafet, T., Bonner, M.R., Wood, E.T., Novelletto, A., Malaspina, P., Mitchell, R.J., Horai, S., Jenkins, T., et al. 1997. The geographic distribution of human Y chromosome variation. *Genetics* **145:** 787–805.

Hammer, M.F., Karafet, T., Rasanayagam, A., Wood, E.T., Altheide, T.K., Jenkins, T., Griffiths, R.C., Templeton, A.R., and Zegura, S.L. 1998. Out of Africa and back again: Nested cladistic analysis

of human Y chromosome variation. *Mol. Biol. Evol.* **15:** 427–441.
Hammer, M.F., Karafet, T.M., Redd, A.J., Jarjanazi, H., Santachiara-Benerecetti, S., Soodyall, H., and Zegura, S.L. 2001. Hierarchical patterns of global human y-chromosome diversity. *Mol. Biol. Evol.* **18:** 1189–1203.
Jobling, M. 1994. A survey of long-range DNA polymorphisms on the human Y chromosome. *Hum. Mol. Genet.* **3:** 107–114.
Jobling, M.A. 1997. In the name of the father: Surnames and genetics. *Trends Genet.* **17:** 353–357.
Jobling, M.A., Samara, V., Pandya, A., Fretwell, N., Bernasconi, B., Mitchell, R.J., Gerelsaikhan, T., Dashnyam, B., Sajantila, A., Salo, P.J., et al. 1996. Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. *Hum. Mol. Genet.* **5:** 1767–1775.
Jobling, M.A., Pandya, A., and Tyler-Smith, C. 1997. The Y chromosome in forensic analysis and paternity testing. *Int. J. Legal Med.* **110:** 118–124.
Jobling, M.A. and Tyler-Smith, C. 2000. New uses for new haplotypes the human Y chromosome, disease and selection. *Trends Genet.* **16:** 356–362.
Jobling, M.A. 2001. In the name of the father: Surnames and genetics. *Trends Genet.* **17:** 353–357.
Kalaydjieva, L., Calafell, F., Jobling, M.A., Angelicheva, D., de Knijff, P., Rosser, Z.H., Hurles, M.E., Underhill, P., Tournev, I., Marushiakova, E., et al. 2001. Patterns of inter- and intra-group genetic diversity in the Vlax Roma as revealed by Y chromosome and mitochondrial DNA lineages. *Eur. J. Hum. Genet.* **9:** 97–104.
Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R.S., Redd, A.J., Zegura, S.L., and Hammer, M.F. 2001. Paternal population history of East Asia: Sources, patterns, and microevolutionary processes. *Am. J. Hum. Genet.* **69:** 615–628.
de Knijff, P. 2000. Messages through bottlenecks: On the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am. J. Hum. Genet.* **67:** 1055–1061.
Pandya, A., King, T.E., Santos, F.R., Taylor, P.G., Thangaraj, K., Singh, L., Jobling, M.A., and Tyler-Smith, C. 1998. A polymorphic human Y-chromosomal G to A transition found in India. *Ind. J. Hum. Genet.* **4:** 52–61.
Richards, M.B., Macaulay, V.A., Bandelt, H.-J., and Sykes, B.C. 1998. Phylogeography of mitochondrial DNA in western Eruope. *Ann. Hum. Genet.* **62:** 241–260.
Santos, F.R., Pena, S.D.J., and Tyler-Smith, C. 1995. PCR haplotypes for the human Y chromosome based on alphoid satellite variants and heteroduplex analysis. *Gene* **165:** 191–198.
Seielstad, M.T., Hebert, J.M., Lin, A.A., Underhill, P.A., Ibrahim, M., Vollrath, D., and Cavalli-Sforza, L.L. 1994. Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. *Hum. Mol. Genet.* **3:** 2159–2161.
Semino, O., Passarino, G., Oefner, P.J., Lin, A.A., Arbuzova, S., Beckman, L.E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., et al. 2000. The genetic legacy of paleolithic homo sapiens in extant Europeans: A Y chromosome perspective. *Science* **290:** 1155–1159.
Shen, P., Wang, F., Underhill, P.A., Franco, C., Yang, W.H., Roxas, A., Sung, R., Lin, A.A., Hyman, R.W., Vollrath, D., et al. 2000. Population genetic implications from sequence variation in four Y chromosome genes. *Proc. Natl. Acad. Sci.* **97:** 7354–7359.
Shinka, T., Tomita, K., Toda, T., Kotliarova, S.E., Lee, J., Kuroki, Y., Jin, D.K., Tokunaga, K., Nakamura, H., and Nakahori, Y. 1999. Genetic variations on the Y chromosome in the Japanese population and implications for modern human Y chromosome lineage. *J. Hum. Genet.* **44:** 240–245.
Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., Akey, J., Huang, W., Shen, D., Lu, D., Luo, J., et al. 1999. Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am. J. Hum. Genet.* **65:** 1718–1724.
Underhill, P.A., Jin, L., Zemans, R., Oefner, P.J., and Cavalli-Sforza, L.L. 1996. A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proc. Natl. Acad. Sci.* **93:** 196–200.
Underhill, P.A., Jin, L., Lin, A.A., Mehdi, S.Q., Jenkins, T., Vollrath, D., Davis, R.W., Cavalli-Sforza, L.L., and Oefner, P.J. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* **7:** 996–1005.
Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., et al. 2000. Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26:** 358–361.
Underhill, P.A., Passarino, G., Lin, A.A., Shen, P., Mirazon Lahr, M., Foley, R.A., Oefner, P.J., and Cavalli-Sforza, L.L. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65:** 43–62.
Whitfield, L.S., Sulston, J.E., and Goodfellow, P.N. 1995. Sequence variation of the human Y chromosome. *Nature* **378:** 379–380.

Received October 4, 2001; accepted in revised form December 4, 2001.