

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

The Y Chromosome Consortium¹

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

Figure 1 The single most parsimonious tree of 153 haplogroups () showing correspondences with prior nomenclatures (). 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-

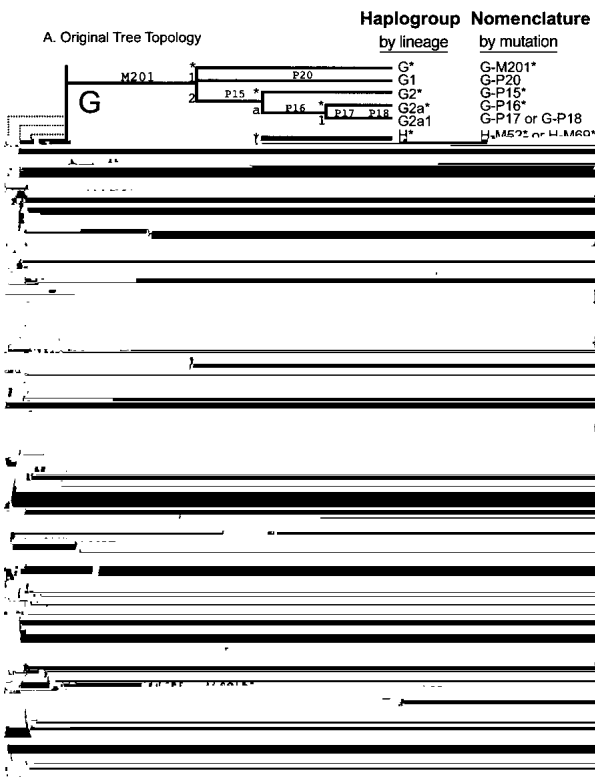


Figure 2 Potential examples of revisions in topology necessitated by the discovery of new mutations and new samples with intermediate haplogroups. Haplogroup nomenclature systems are shown to the right of the tree. (A) The G and H haplogroups are as shown in Figure 1. (B) Case of a newly discovered marker that joins haplogroups within haplogroup G. () Newly discovered mutation (μ) that splits clades within haplogroup G. () 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.

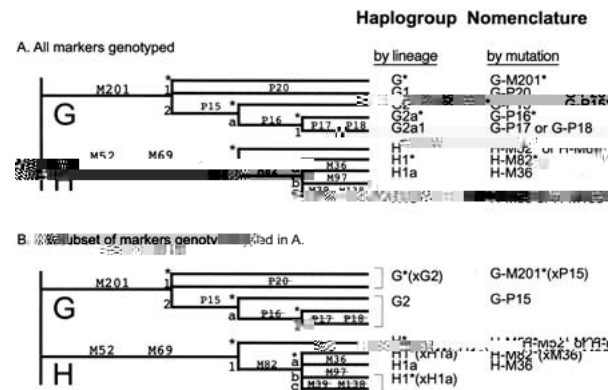


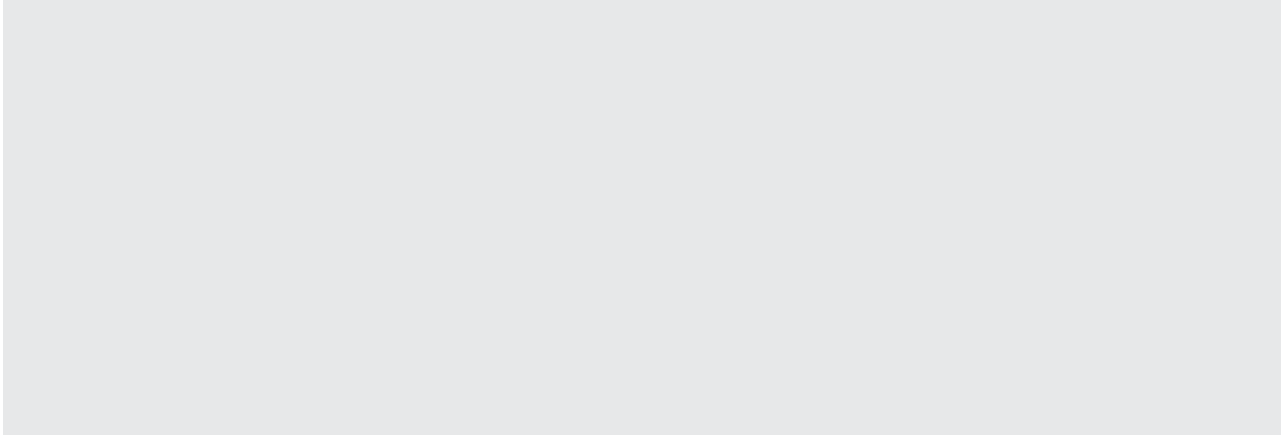
Figure 3 Examples of haplogroup names for cases in which subsets of markers in Figure 1 are genotyped. Markers that were not genotyped are shown with a strikethrough. The lineage- and mutation-based full nomenclature systems are shown to the right of the tree.

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 _{10831br} , SRY ₋₂₆₂₇	P*(xR1b8,R1a,Q3)	
	2	SRY _{10831a}	50f2(P), RPS4Y ₇₁₁ , YAP	R1a	
	3	SRY _{10831b}	Apt, M52, 12f2a, M9	BR*(xB2b,CE,F1,H,J,K)	
	4	YAP		R1a	
	5	47z		DE*(xE)	
	6	50f2(P)		O2b1	
	7		SRY _{10831ar} , MEH1	B2b	
	8	M2		Y*(xBR,A2)	
	9	12f2a		E3a	
	10	RPS4Y ₇₁₁		J	
	12	LLY22g	Tat	C	
	13	LINE1		N*(xN3)	
	15	Apt		O3c	
	16	Tat		F1	
	18	M3		N3	
	20	SRY ₊₄₆₅	47z	Q3	
	21	SRY ₄₀₆₄	P2	O2b*	
	22	SRY ₋₂₆₂₇		E*(xE3)	
	23	SRY ₉₁₃₈		R1b8	
	24	M4		K1	
	25	P2	M2	M	
	26	M9	SRY _{9138r} , M20, M4, LLY22g, SRY _{+465r} , LINE1, 92R7	E3*(xE3a)	
	27	MEH1		K*(xK1,LN,O2b,O3c,P)	
	28	M20		A2	
	35	M52		L	
	Underhill (2000)	I	M91		H
		II	M60		A
		III	M96		B
		IV	M174		E
		V	RPS4Y ₇₁₁		D
		VI	M89	M9	C
		VII	M175		F*(xK)
		VIII	M9	M175, M45	O
		IX	M173		K*(xO,P)
		X	M45		R1
Hammer (2001)	1A		M173	P*(R1)	
	2	P3	P3, SRY _{10831a}	Y*(xBR,A2)	
	1B	SRY _{10831a}	RPS4Y ₇₁₁ , YAP, P14	A2	
	1C	P27	SRY _{10831br} , P25, M3	BR*(xF,DE,C)	
	1D	SRY _{10831b}		P*(xR1a,R1b,Q3)	
	1E	SRY ₉₁₃₈		R1a	
	1F	RPS4Y ₇₁₁		K1	
	1G	M3		C	
	1Ha	P15	P16	Q3	
	1Hb	P16		G2*	
	1I	Tat		G2a	
	1L	P25		N3	
	1U	M9	P27, Tat	R1b	
	1R	P14	12f2a, P15, M9	K*(xP,N3)	
	3G	YAP	SRY ₄₀₆₄	F*(xJ,G2,K)	
	3A	SRY ₄₀₆₄	P2	DE*(xE)	
	4	P2	P1	E*(xE3)	
5	P1		E3*(xE3a)		
Karafet (2001)	Med	12f2a		E3a	
	1		SRY _{10831ar} , M13, P3, P4, M6, M14	J	
	2	M13		Y*(xBR,A2,A3b2)	
	3	P4, P3, M6, M14	SRY _{10831a}	A3b2	
	4	P28	SRY _{10831a}	A2*(xA2b)	
	5	SRY _{10831a}	P9, 50f2(P)	A2b	
	6	50f2(P)	P6, P7, P8, MSY2a	BR*(xCR,B2b)	
	7	P6		B2b*(xB2b1,B2b4)	
	8	P7	P8, MSY2a	B2b1	
	9	MSY2a		B2b4*	
	10	P8		B2b4b	
	11	M174	M15	B2b4a	
	12	M15		D*(xD1)	
	13	SRY ₄₀₆₄	P2, P1	D1	
14	P2	P1	E*(xE3)		
			E3*(xE3a)		

Table 1. (. . .)

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		



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 let-

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 undervived 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 scien-

tific 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

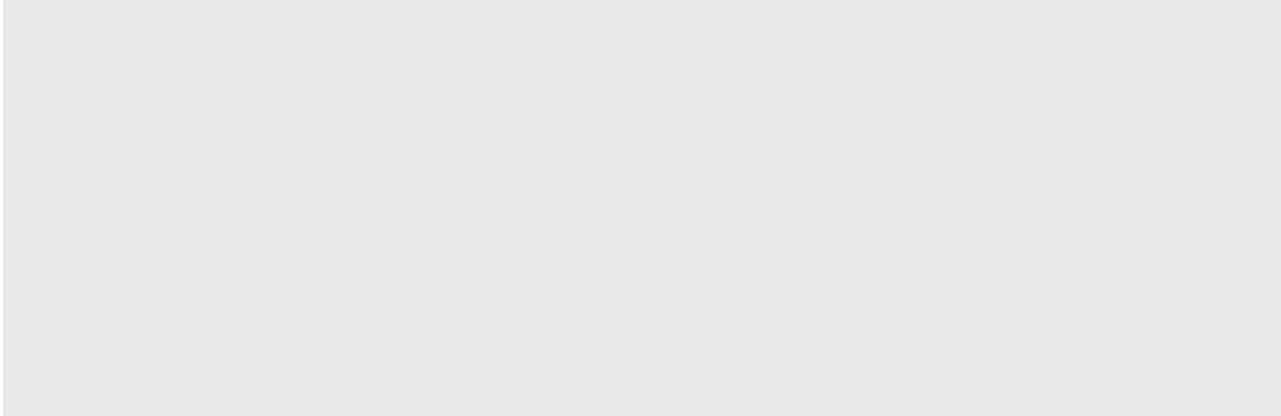
In constructing the tree, a great deal of phylogenetic information was retained from previous studies. When markers from different laboratories mapped on the same branch of the tree, an attempt was made to determine the order of mutational events. Toward this end, a variety of samples was provided by 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 `assa5amplomj01,D2MULSU5anoj01nstrrvSU5anBoj01Z4ZZLJaeanoj01`

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*



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

- 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 KnijP