

# Fathers and sons: the Y chromosome and human evolution

MARK A. JOBLING AND CHRIS TYLER-SMITH

People have always been curious about their origins. Traditionally, the past has been investigated by historians, archaeologists and paleontologists, while indirect evidence from modern human populations has been sought by linguists and, increasingly today, molecular biologists. This is because our DNA has been passed down to us from our ancestors, accumulating mutations on the way. The DNAs of modern humans are, therefore, different from each other, and these differences, or polymorphisms, provide a record of our relatedness and genetic history. Modern techniques now allow this record to be read, and have led to an upsurge in interest in human genetic diversity and evolution.

When a sperm fertilizes an egg, half of the DNA of the mother and half of that of the father are brought together in the zygote. The genetic contribution of the sperm differs from that of the egg in two complementary ways: it fails to contribute any mitochondria, which contain their own small genomes – these are, therefore, maternally inherited; and in half of fertilizations it alone contributes a Y chromosome, which is, therefore, paternally inherited. Neither of these segments of DNA recombines at meiosis, and this means that they each contain a particularly simple record of their past (Fig. 1). There is an exception to this: a ‘pseudoautosomal’ region at each end of the Y chromosome does recombine, but for the remainder of this review ‘Y chromosome’ will refer to the Y-specific region that makes up most of the chromosome.

While X chromosomes and autosomes each have multiple ancestors because of recombination, all modern mitochondrial genomes (mtDNAs) have a single maternal ancestor, and all modern Y chromosomes have a single paternal ancestor. This is inevitable, although it cannot be predicted *a priori* whether the ancestor was a recent human or an ancient pre-human. It does not imply that there was ever only one man in the population, or that there was anything special about him (Fig. 2). This genetic simplicity has been best exploited in studies of human mtDNA<sup>1</sup> because it is technically easier to analyse. The Y chromosome is now becoming amenable to similar forms of analysis. The purpose of this review is to ask what issues Y chromosome research can address, to summarize the progress that has been made, and to suggest the future directions that this research should take.

## Aims of Y chromosome research

### *Current issues in recent human evolution*

Most geneticists and many paleontologists support the ‘Out of Africa’ hypothesis, which holds that there was a single origin for modern humans in Africa less than 200 000 years ago, after which these people dispersed throughout the world without mixing with existing populations, such as the Neanderthals<sup>2</sup>. However, the details of these dispersals are far from clear and many questions remain under active investigation. These include the peopling of southern Asia and Australia, perhaps by migration out of Africa approximately 60 000–70 000 years ago via a southern route around the shores of the Indian Ocean, and the possibility of a separate, more northerly, migration out of Africa approximately 40 000–50 000 years ago, giving rise to the people of Europe and northern Asia<sup>2</sup>. There

***It should be possible to use Y chromosome DNA polymorphisms to trace paternal lineages for evolutionary and other studies, but progress in these areas has been slow because it has been difficult to find suitable markers. However, it is now possible to use selected, slowly evolving polymorphisms to draw a rudimentary Y chromosome tree, while more rapidly evolving polymorphisms allow most independent Y chromosomes to be distinguished. Different populations often have characteristically different Y chromosomes, and Y chromosome studies are soon likely to make a major contribution to our understanding of the origins of modern humans.***

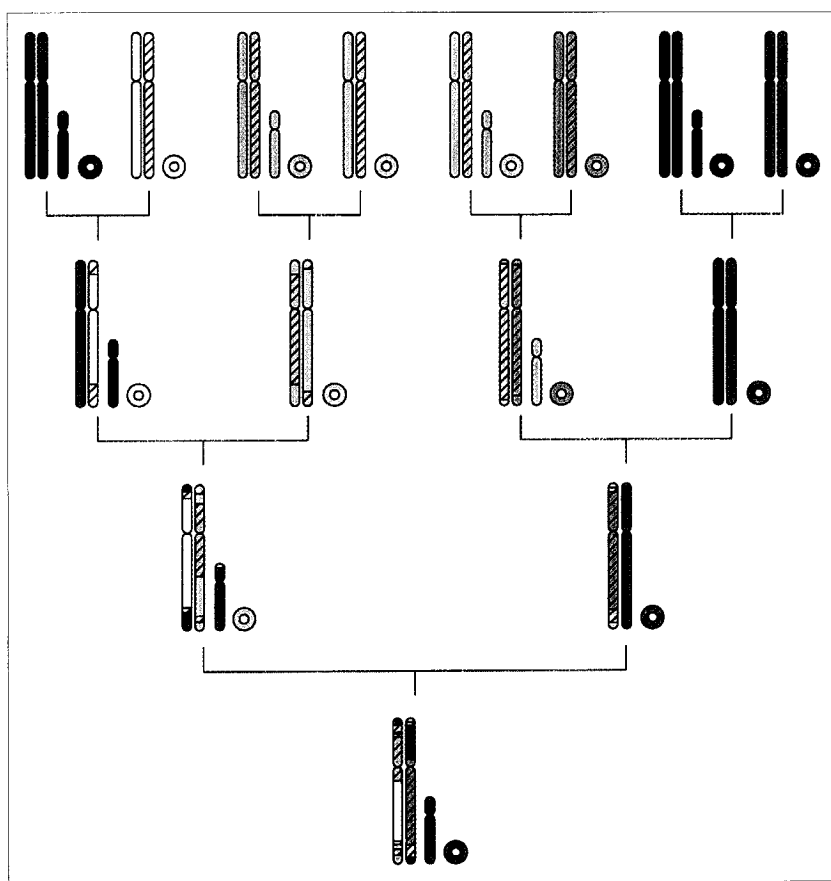
are opposing models for the origin of the European Neolithic farmers less than 10 000 years ago, involving either the spread of the farmers themselves, or the spread only of farming technology, while the timing of the colonization of the Americas and the geographical origin of the Pacific populations remain unusually controversial areas of research. More general issues are the genetic affiliations of linguistically related populations, and the history of population admixture. The traditional studies of the past have not given clear answers to these questions.

### *Potential contributions from Y chromosome studies*

The ‘Out of Africa’ model predicts that the paternal ancestor of all modern Y chromosomes lived recently in Africa. One of the aims of Y chromosome research is, therefore, to produce a tree showing the evolutionary relationships of modern Y chromosomes, to identify its origin (root) and to date its branchpoints. Another aim is to measure the frequencies of different Y chromosome types in different populations. Closely related populations are likely to have similar frequencies of each type; more distantly related populations are expected to show progressively greater differences. Migration of a self-contained population might lead to the establishment of a new population containing a subset of the original Y chromosomes, while the movement of smaller numbers of people who mix with an existing population will lead to the admixture of Y chromosomes. These events should be recognizable from the distribution of modern Y chromosomes.

### *Analogy with mtDNA*

These aims are the same as those of mtDNA research in human evolution. Although the Y and mtDNA have similarities in their mode of inheritance and lack of recombination, the reasons for the comparative lack of progress in Y chromosome research lie in the many differences between the systems. However, some of these differences give the Y chromosome inherent advantages.



**FIGURE 1.** Inheritance of autosomal, Y-chromosomal and mtDNA sequences. Autosomal sequences (large pair of chromosomes) recombine each generation, while Y-specific (small chromosome) and mtDNA sequences (circle) do not. Consequently, an individual (bottom) can trace his/her autosomal sequences back to multiple ancestors (top – each is a different colour), but the Y chromosome (excluding the pseudoautosomal regions) and mtDNA have only a single ancestor.

The mtDNA is a small (16.5 kb) circular molecule whose entire sequence is known. It has a high density of genes, with little noncoding DNA, and a base-substitutional mutation rate about ten times higher than that of nuclear DNA, leading to very high diversity<sup>1</sup>. In contrast, the Y is a large (approximately 60 Mb) linear molecule whose sequence is still largely unknown. Unlike mtDNA, it is located in the nucleus and does not pass through female mitosis or meiosis. Although about 4000 times the size of the mtDNA, it has fewer known genes and contains many different kinds of sequence, including tandem and dispersed repeat families in addition to unique sequences<sup>3</sup>.

Although the high complexity and low mutation rate of the Y have been obstacles to progress, they are also advantages that should ultimately allow it to yield more informative data on human evolution than mtDNA. The large and complex Y chromosome can carry a much more diverse spectrum of polymorphisms: small- and large-scale rearrangements, such as insertions, deletions, duplications and inversions; and polymorphisms associated with tandemly repeated DNA sequences ranging from large satellite loci, down to mini- and microsatellites. These different loci have different mutation rates; consequently it should be possible to select appropriate Y polymorphisms for studying evolution over different

time scales. The high base substitution rate of mtDNA, although it leads to high diversity, also creates problems: it is not possible to determine ancestral states by analysis of, for example, chimpanzee mtDNA; and the rate in some parts of the molecule is so high that particular substitutions can revert or recur within human lineages. Base substitutions on the Y chromosome are rare enough to be regarded as unique events and, thus, ancestral states can usually be determined.

The history of human Y chromosomes is likely to be different from that of mtDNAs; this will reflect aspects of population histories, such as cultural practices governing mating structures, and the differential behaviour of males and females in migrations, wars and colonizations.

#### What do we know? The current position

The first Y chromosome DNA polymorphisms<sup>4,5</sup> were reported ten years ago, amid optimism about the potential of the Y in evolutionary studies. However, rather little progress has been made since then. This is because conventional DNA polymorphisms have been difficult to find, and those discovered have often not proved very useful for evolutionary

purposes; also, few populations have been adequately surveyed.

#### Scarcity of polymorphisms

A number of systematic searches<sup>6-8</sup> have amply demonstrated the relative lack of conventional polymorphisms on the Y: two large restriction fragment length polymorphism (RFLP) studies screened, on average, 833 bp (Ref. 6) and 2215 bp (Ref. 7) in each of 22 Y chromosomes and found between them only three polymorphisms. Two more recent studies have used DNA sequencing. In one about 1.4 kb of DNA was sequenced in 12–16 chromosomes of diverse geographical origins and a single polymorphic nucleotide substitution was found<sup>9</sup>, while in the other a 729 bp intron in the *ZFY* gene was sequenced in 38 males and no variation was found<sup>10</sup>. This compares with a base substitution frequency in noncoding regions elsewhere in the nuclear genome of about one in 235 bp for 12–20 chromosomes<sup>11</sup>. To many, this reduced sequence variation has seemed puzzling: the Y is seen as a gene-poor chromosome, which should be freer than other parts of the genome to accumulate mutations and 'junk' sequences. Also, the exclusive passage of the Y through male gametogenesis has led to the prediction that the mutation rate of the Y should be elevated over that of the rest of the nuclear genome. This is because

## REVIEWS

the Y spends proportionally more of its time in the male germline than do the X and autosomes (100% compared with 33% and 50%, respectively); the production of sperm involves many more cell divisions and, hence, potentially mutagenic DNA replications, than that of eggs<sup>12</sup>.

There are a number of possible explanations for the reduced diversity. The simplest is a numerical explanation: for each Y chromosome in the population, there are four of each autosome and three X chromosomes. This reduced chromosome population size will be reflected in correspondingly reduced diversity<sup>13</sup>. This factor alone could explain much of the observed reduction. A related explanation, which is likely to be particularly important in specific human populations, is to do with mating patterns: if a small number of males are very successful and have many offspring, while most males have none or only very few, then the effective population size of the Y chromosome will be reduced even further and this will lead to restricted Y chromosome diversity. Striking examples of this behaviour are found in villages of the Yanomamö Indians in South America<sup>14</sup> and are accompanied by very low Y diversity<sup>15</sup>. An alternative explanation invokes the lack of recombination on the Y: if a selectively advantageous mutation arose in a Y-specific gene, this would spread through the population, bringing with it, as a particularly bulky 'hitchhiker' carrying little or no variation, the rest of the chromosome. Variants on other Y chromosomes cannot recombine onto the selected Y and consequently are lost.

Whatever the explanation for this reduced diversity, its significance is that it tells us that the modern population of Y chromosomes has a recent common ancestor.

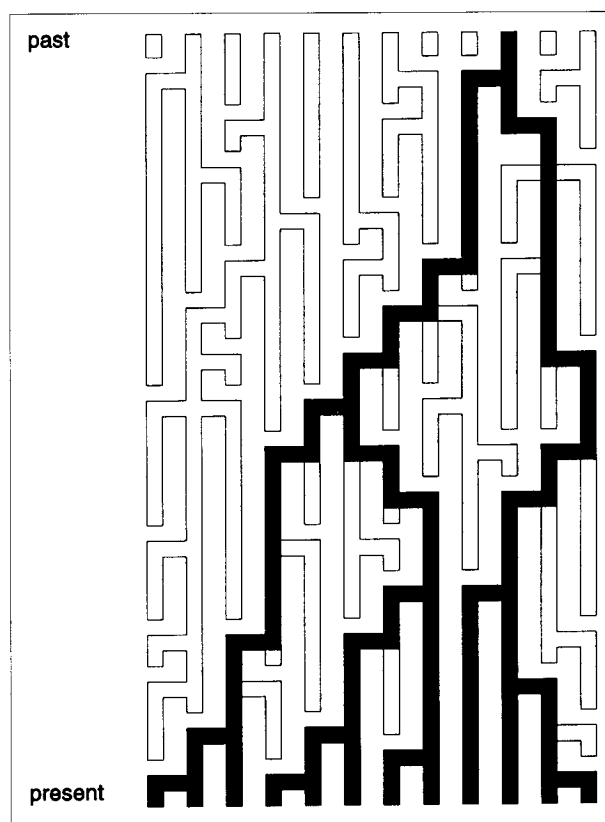
### Known polymorphisms

A variety of polymorphisms has been described on the Y chromosome (Fig. 3). The kinds of polymorphisms sought and studied have been driven by the advent of new technologies: to the fruits of the initial searches for conventional RFLPs<sup>6-8</sup> have been added long-range polymorphisms<sup>16,17</sup> detected by pulsed-field gel electrophoresis (PFGE), a range of different polymorphisms detectable by the polymerase chain reaction (PCR)<sup>9,18</sup>, and now DNA sequence variation can be identified in the course of large-scale sequencing projects.

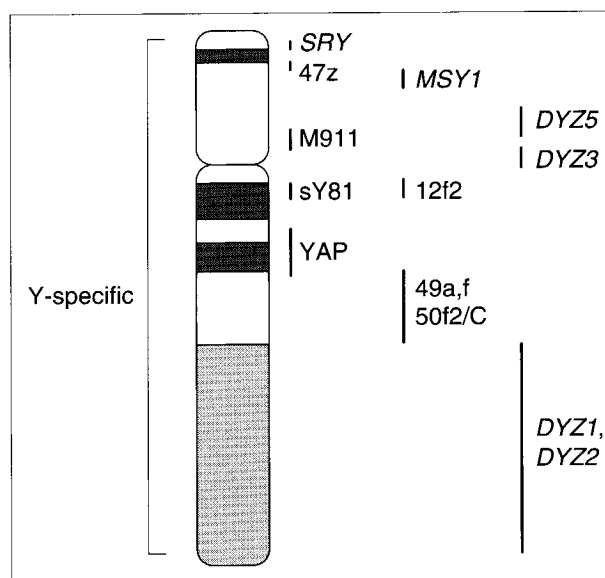
Comprehensive lists of Y-specific polymorphisms are given elsewhere (see Box 1) and Table 1 contains examples showing the kinds of polymorphisms, their usefulness and their disadvantages. Table 2 lists all available Y polymorphisms that can be typed by PCR.

### Compound haplotypes

The states of multiple Y polymorphisms of diverse kinds can be combined to give compound haplotypes. The most variable polymorphisms distinguish between all independent Y chromosomes; the less variable allow groups of related chromosomes to be defined. In one set of studies, which used both highly variable and binary markers, four haplotypic groups were identified<sup>16,17,19</sup>. Following on from this and additional work<sup>20</sup>, we have now used all the available polymorphisms that appear to represent unique mutations to construct compound haplotypes. There are nine such polymorphisms and together they define nine compound haplotypes (Fig. 4a).



**FIGURE 2.** Y chromosome lineages. In each generation (row) some of the Y chromosomes in the population are transmitted to the next generation and some are not. Consequently, after a number of generations all surviving Y chromosomes can be traced back to a single chromosome (black lineages).



**FIGURE 3.** Location of selected polymorphic loci on the Y chromosome. Most of the chromosome is Y-specific (left) and the positions of some polymorphic loci are shown (right). The ideogram of the chromosome shows the light giemsa-stained (G) bands (white), the dark G bands (dark grey) and the heterochromatic region (grey).

# REVIEWS

## Box 1. The Y Chromosome Consortium

The Y Chromosome Consortium is a group involved in a collaborative effort to study genetic variation on the human Y chromosome. Its aims are: (1) to establish a repository of lymphoblastoid cell lines derived from individuals representing indigenous populations from all inhabited continents; (2) to provide DNA isolated from these cell lines to investigators searching for polymorphisms on the Y chromosome; and (3) to establish a common database containing the results of typing DNAs from the Repository's cell lines at as many Y-specific polymorphic loci as possible. Communication is by the Y Chromosome Consortium Newsletter, which also contains comprehensive lists of known Y polymorphisms and relevant references. Additional blood samples or lymphoblastoid cell lines are required from several populations, particularly from the Pacific and Australian regions. Details can be obtained from:

**Michael F. Hammer**

hammer@brahms.biosci.arizona.edu  
Department of EEB, Biosciences West, University of  
Arizona, Tucson, AZ 85721, USA.

**Nathan A. Ellis**

nellis@server.nybc.org  
Laboratory of Human Genetics, New York Blood Center,  
310 East 67th Street, New York, NY 10021, USA.

(Contributed by Michael F. Hammer and Nathan A. Ellis.)

### Tree drawing

If it is assumed that the Y is evolving in a haploid fashion, then tree drawing based on binary polymorphisms, which have unique origins for each allele and known ancestral states, should be simple and computer-free. Using only a pencil and paper, and the principle of parsimony (the assumption of a minimum number of

mutational events), a single tree can indeed be drawn with this information (Fig. 4c). In this tree, each haplotype is shown as a node and the position of each mutation is shown on a line connecting two nodes. The tree is not rooted because the ancestral state of each polymorphism is not known. The ancestral states of some loci have been inferred from an analysis of ape DNA [sY81

**TABLE 1. Examples of Y polymorphisms**

Type	Discovery or characteristics	Usefulness	Example(s) <sup>a</sup>	How detected	Refs
Base substitution	RFLP searches: detected with one enzyme only; sequencing	Probable unique event defining a monophyletic group	47z ( <i>DXYS5Y</i> )/ <i>StuI</i> 92R7/ <i>HindIII</i> sY81 ( <i>DYS271</i> )/ <i>NlaIII</i> , (A→G transition)	Filter hybridization Filter hybridization PCR-RFLP	22 19 9
Duplication or deletion	Identical change in fragment size detected with more than one enzyme	Some (e.g. 12f2) are apparently unique events; others (e.g. <i>DYZ7/C</i> ) are not	12f2 ( <i>DYS11</i> )/ <i>TaqI</i> or <i>EcoRI</i> 50f2 ( <i>DYS7/C</i> )/ <i>EcoRI</i>	Filter hybridization Filter hybridization PCR	4 30
Insertion	Identical change in fragment size detected with more than one enzyme; sequencing	Probable unique event defining a monophyletic group	YAP ( <i>DYS287</i> )/ <i>TaqI</i> or <i>EcoRV</i>	Filter hybridization, PCR length polymorphism	21, 25
Complex rearrangement	Diverse; absent or additional bands; band size variation	Difficult to interpret, molecular basis unclear; apparently identical alleles can arise independently	49a/f( <i>DYS1</i> )/ <i>TaqI</i> and others (>100 different haplotypes)	Filter hybridization	5
<b>Tandem repeats:</b>					
Major satellite	Large hypervariable arrays; some contain polymorphic restriction sites	Very variable. PFGE a major disadvantage; internal sites	YαI ( <i>DYZ3</i> )/e.g. <i>BglII</i> (array length polymorphism); <i>AvaII</i> , <i>EcoO1091</i>	Filter hybridization after PFGE	16
Minisatellite	Hypervariable arrays of 10–50 bp repeat units; internal repeat unit variation leads to very high diversity	High level of discrimination; mutation rate measurable in principle, so might be useful for dating trees	MSY1 ( <i>DYF155S1</i> )	MVR-PCR <sup>b</sup>	28
Microsatellite	Length-variable arrays of 2–5 bp repeat units	Alleles identifiable unambiguously, but isoallelism a problem	27H39LR ( <i>DYS19</i> ) tetranucleotide	PCR	18

<sup>a</sup>The examples are given as: probe or primer pair name (locus name, where available)/restriction enzyme (where applicable) (other comments).

<sup>b</sup>Minisatellite variant repeat PCR.

Abbreviations: RFLP, restriction fragment length polymorphism; PFGE, pulsed-field gel electrophoresis.

# REVIEWS

TABLE 2. Y polymorphisms that can be typed using PCR<sup>a</sup>

Locus	Name	Primers (5'→3')		Comments	Ref.
<i>DYS287</i>	YAP	CAGGGGAAGATAAAGAAATA	ACTGCTAAAAGGGGATGGAT	Alu insertion <sup>b</sup>	25
<i>DYS271</i>	sY81	AGGCACTGGTCAGAATGAAG	AATGGAAAATACAGCTCCCC	<i>Nla</i> III detects point mutation <sup>b</sup>	9
<i>SRY</i>		TCCTTAGCAACCATTAATCTGG	AAATAGCAAAAATGACACAAGGC	Point mutation	<sup>c</sup>
<i>YRRM2</i>		CTTTGAAAACAATTCCTTTTCC	AGAGATGCACCTTCAGAGG	Product present or absent	31
<i>DYZ3</i>	Y $\alpha$ I	TCTGAGACACTTCTTTGTGGTA	CGCTCAAAATATCCACTTTCAC	<i>Hind</i> III cleavage (partial) indicates 6.0 kb unit	<sup>d</sup>
<i>DYF155S2</i>	50f2/C	CTCAAGCTAGGACAAAGGGAAAGG	GAGGTAGATGCTGAAGCGGTATAG	196 bp fragment present or absent	<sup>e</sup>
<i>DYS288</i>		CATTACAATACCTGGACACTG	TTGCTTTGCTTGTTCATTTCAGA	Dinucleotide repeat, 1 Y locus <sup>f</sup>	GDB <sup>g</sup>
–	YCAI	CCCATGCCTGTTCTCCAGATT	GAGAGTGTGACACATCAGGTA	CA repeat, 2 Y loci	19 <sup>g</sup>
–	YCAII	TATATFAAATAGAAAGTAGTGA	TATCGATGTAATGTTATATTA	CA repeat, 2 Y loci	19
–	YCAIII	CCACATTTGTGTAATGTGTGA	TCCTCAGAGAAGGAGAAACTA	CA repeat, 2 Y loci	19
<i>DYS388</i>		GTGAGTTAGCCGTTTAGCGA	CAGATCGCAACCACTGCG	Trinucleotide repeat, 1 Y locus <sup>f</sup>	GDB <sup>g</sup>
<i>DYS392</i>		TCATTAATCTAGCTTTTAAAAACAA	AGACCCAGTTGATGCAATGT	Trinucleotide repeat, 1 Y locus <sup>f</sup>	GDB <sup>g</sup>
<i>DYS19</i>	27H39LR	CTACTGAGTTTCTGTTATAGT	ATGGCATGTAGTGAGGACA	GATA repeat, 1 Y locus <sup>f</sup>	18
<i>DYS390</i>		TATATTTTACACATTTTGGGCC	TGACAGTAAAATGAACACATTGC	Tetranucleotide repeat, 1 Y locus <sup>f</sup>	GDB <sup>g</sup>
<i>DYS391</i>		CTATTCATTCAATCATAACCCCA	GATTCTTTGTGGTGGGCTCTG	Tetranucleotide repeat, 1 Y locus <sup>f</sup>	GDB <sup>g</sup>
<i>DYS393</i>		GTGGTCTTCTACTTGTGTCAATAC	AACTCAAGTCCAAAAATGAGG	Tetranucleotide repeat, 1 Y locus <sup>f</sup>	GDB <sup>g</sup>
<i>DYS385</i>		AGCATGGGTGACAGAGCTA	TGGGATGCTAGGTAAAGCTG	Tetranucleotide repeat, 2 Y loci	GDB <sup>g</sup>
<i>DYS389</i>		CCAACTCTCATCTGTATTATCTATG	TCTTATCTCCACCCACCAGA	Tetranucleotide repeat, 2 Y loci	GDB <sup>g</sup>
<i>DXYS156Y</i>		GTAGTGGTCTTTTGCCTCC	CAGATACCAAGGTGAGAATC	TAAAA repeat on X and Y	32

<sup>a</sup>Some loci listed in the Genome Data Base (GDB) have been omitted from this table. *DYS384* is not Y-specific (M. Kayser, pers. commun.), *DYS394* is *DYS19* (A. Nystuen, pers. commun.) and the *DYS395* primers are very similar to the *DYS393* primers.

<sup>b</sup>Unique mutation with known ancestral state.

<sup>c</sup>C. Kwok and R. Hawkins, pers. commun.

<sup>d</sup>F. Santos, S. Pena and C. Tyler-Smith, unpublished. Polymorphic heteroduplex bands are also seen.

<sup>e</sup>*DYF155S2* lies 4 kb away from 50f2/C (*DYS7/C*); the primers also amplify *DYF155S1* (MSY1); M.A. Jobling, unpublished.

<sup>f</sup>Microsatellite with single polymorphic Y locus.

<sup>g</sup>Unpublished information on the Y specificity, polymorphism and copy number of these loci was provided by M. Kayser, L. Roewer, P. de Knijff, A. Nystuen and S. Gerken.

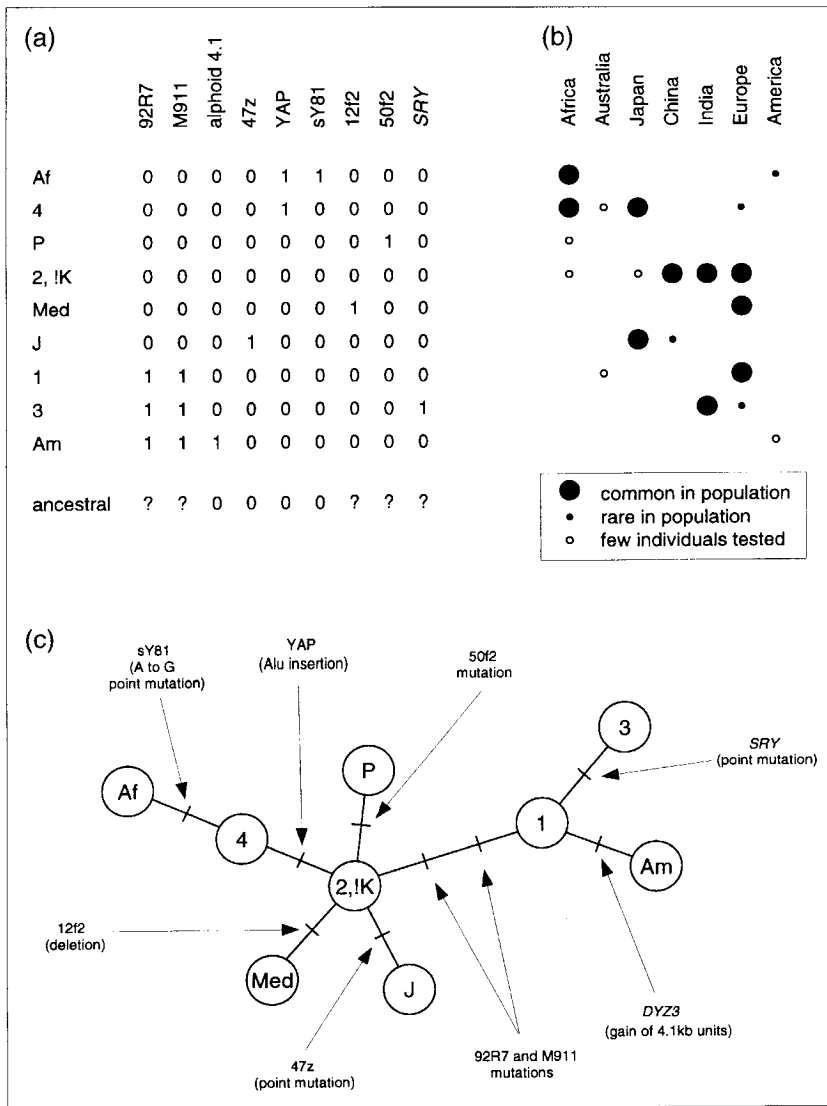
(Ref. 9), YAP (Ref. 21); Table 1] or a consideration of the mechanisms of DNA change [47z (Ref. 22; Table 1), alphoid 4.1 kb units<sup>19</sup>]. Haplotypes Af, 4, J and Am (Fig. 4a) cannot be ancestral because they lack the ancestral alleles at these loci. Among the other five haplotypes, 2,K is the best candidate for the ancestral haplotype because it contains the commonest allele at each locus and is the most geographically widespread (Fig. 4b). According to this line of reasoning, the root would be placed at the 2,K node.

A striking feature of the tree is the star-like pattern with five haplotypes radiating from the putative ancestral haplotype. Such 'star phylogenies' arise when each new mutation has a high probability of surviving<sup>23</sup>. This can happen when selection is favouring a particular haplotype (and the variants that arise on it) or when a population is expanding. In the absence of selection in a constant-sized population, a more even distribution of branches would be seen with few nodes containing multiple branches. Starlike patterns have been seen in mtDNA trees and it has not been easy to choose between the two explanations<sup>24</sup>. If more detailed Y trees that are

more representative of world populations maintain this appearance, one goal of Y chromosome research will be to date the possible expansion. Coincidence of the dates of mtDNA and Y chromosome expansion would then suggest an expansion of the human population at that time.

### Population distribution of haplotypes

Too little information is available to establish reliably the frequencies of the haplotypes in a representative sample of world populations, but we have summarized the limited information available in Fig. 4b. Two features emerge from this analysis. Firstly, several haplotypes are largely confined to particular populations. For example, the Af haplotype has been reported mainly from Africans<sup>9</sup>, while the J haplotype is common in Japan<sup>22,25</sup> and has been detected in Koreans and a few Chinese individuals<sup>26</sup>. Each population has a unique combination of Y haplotypes. This population-specificity of Y-chromosomal polymorphisms contrasts with the lack of specificity of many polymorphisms on other chromosomes. Secondly, individual African populations appear



**FIGURE 4.** Y chromosome haplotypes. (a) Compound haplotypes are named after the groups identified in references<sup>16,17,19</sup> (where they were given the arbitrary numbers 1, 2, 3 and 4 according to the order in which they were discovered), or after the populations or locations in which they are largely found<sup>20</sup>. Abbreviations: Af, African; P, Pygmy; !K, !Kung; Med, Mediterranean; J, Japanese; Am, Amerindian. Haplotypes (rows) consist of the polymorphic states for 92R7 (Ref. 19; *Hind*III, 0 = 4.6 kb, 1 = 6.7 kb), M911 (Ref. 19; *Xba*I, 0 = 55 kb, 1 = 33 kb), alphoid 4.1 (Ref. 19; 0 = 4.1 kb units absent, 1 = present), 47z (Ref. 22; *Stu*I, 0 = 17 kb, 1 = 5.3 kb), YAP (Ref. 21; 0 = absent, 1 = present), sY81 (Ref. 9; 0 = A, 1 = G), 12f2 (Ref. 4; *Taq*I, 0 = 10 kb, 1 = 8 kb), 50f2 (Ref. 17; *Taq*I, 0 = 8.5 kb, 1 = 3.1 kb) and SRY (0 = G, 1 = A). Note that in some cases the assignment of 0 and 1 to alleles differs from that used previously. (b) Distribution of haplotypes in different geographical areas (based on the information in Refs 4, 9, 16, 17, 19, 21, 22, 25, 26). These data exclude occurrences due to known migrations within historical times. (c) Unrooted evolutionary tree showing the suggested relationships between the compound haplotypes in part (a). Each haplotype is present at a node and each mutation is shown on a line joining two nodes<sup>33</sup>.

to have a low level of Y diversity<sup>19,27</sup>, contrasting with their high levels of mtDNA and autosomal sequence diversity<sup>1</sup>. This might reflect the small effective population size of the Y chromosome in these societies.

**The future**

*New polymorphisms*

A major requirement is for additional polymorphisms and it is important that these be sought in

diverse populations to counter the Eurocentric ascertainment bias of many diversity studies. The ideal marker for many evolutionary purposes would represent a unique mutational event with a known ancestral state; for technical reasons it should be testable using PCR. Base substitutions meet these requirements and it should be possible to carry out systematic searches for more of these using direct sequencing, or scanning methods [e.g. single-strand conformation polymorphism (SSCP) analysis]. How many polymorphisms are required? It is difficult to give a definite answer to this question, but an additional 20 would clarify the topology of the tree. The present haplotypic groups would probably be subdivided and groups such as 2,!K would disappear entirely if derived mutations were found in all chromosomes. The resolution in the tree obtained using base substitutions will be low. Much higher resolution can be obtained using more variable tandemly repeated sequences, such as the minisatellite MSY1 probe (Ref. 28), the major satellites or combinations of microsatellites (which are likely to be numerous - Table 2). These markers will probably have undergone recurrent mutations and so will not be useful for generating trees using parsimony, but they will provide information about the diversity within each haplotypic group and it might be possible to use them in other methods, such as pairwise difference comparisons<sup>23</sup>. They might be the markers of choice for microevolutionary purposes. A study using multiple autosomal microsatellites has shown their potential as phylogenetically informative markers<sup>29</sup>.

*Dating*

In the work on mtDNA, dating has been one of the most important and controversial issues<sup>1</sup>. How

could the branchpoints of a Y-chromosomal tree be dated? It would be possible to estimate dates by comparing human Y chromosome sequence diversity with human-ape Y chromosome differences and using estimates of the time of the human-ape divergence to calibrate the measurement. Alternatively, well-dated events in human prehistory, such as the colonization of Papua New Guinea about 60 000 years ago<sup>1</sup>, could be used for calibration. It should also be possible to detect new

# REVIEWS

mutations at the most variable loci, either by studying their transmission in families or by analysing single molecules from sperm. With an understanding of the mutational processes, a measurement of their rates and assumptions about population structure, it should then be possible to model the evolution of these sequences and to obtain independent estimates of dates in the Y chromosome tree. However, if selection is acting on the Y chromosome, the significance of all these estimates would be questionable. It might be possible to examine ancient DNA directly, although the analysis of ancient nuclear DNA is even more technically problematic than ancient mtDNA. Amplification of very short regions spanning known point mutations could provide dates by which these mutations had occurred and information about their geographical distribution.

## Population sampling

A second requirement is a more adequate sampling of Y chromosomes. Ideally, samples should be cell lines, which will provide an unlimited source of material for many different kinds of analysis. A cell repository has been established by the Y Chromosome Consortium for this purpose. Details are given in Box 1 and readers are urged to make suitable blood samples and cell lines available to the repository. There has been a debate in genome diversity circles about the best strategy for sampling. One suggestion has been to sample populations on the basis of a regular grid, for example every 100 km; the alternative is to use all available information, including language and culture, to select the most diverse populations. Whatever the strategy, it is crucial that adequate numbers of males are included in the sample. Population sampling has raised serious ethical issues: is it right for scientists (usually from rich Western countries) to take blood samples from people who do not have access to the scientific information obtained and who will gain nothing from it? Although these questions have been discussed, they remain unresolved. In our experience, people in many countries are themselves interested in analysing their origins and genetic relationships. Mechanisms for providing scientific funding to such people would solve some of the ethical problems and would greatly enrich the field. It is to be hoped that some funding agencies will have the imagination to do this.

## Genealogical and forensic studies

The availability of Y chromosome DNA polymorphisms will also allow other issues to be addressed. In many societies, surnames are coinherited with Y chromosomes; powerful markers for discriminating between Y chromosomes could be used as genealogical tools. They would be most useful where surnames are expected to have a single or few origins but would be complicated by non-paternity. Such markers would also be useful in forensic studies, particularly in mixed DNA samples. They are unlikely to be variable enough to specifically identify individual males, but will be valuable as exclusion tools. Their nonrandom distribution in different populations could provide clues about the population of origin of a sample, though admixture (and socio-political considerations) will make this problematic.

## Human evolution

With suitable Y chromosome DNA polymorphisms and population samples, it will be possible to address questions about human origins from a Y chromosome perspective. The low level of base substitution polymorphisms already suggests that Y chromosomes have a recent common ancestor, as predicted by the 'Out of Africa' hypothesis, but at present it is impossible to obtain a reliable date or tell where this ancestor lived. The identification of the sources of ancient migrations will require the study of larger numbers of chromosomes, but already a start can be made. The evidence so far available from South American Y chromosomes<sup>19</sup> favours an origin from the postulated northern migration out of Africa, in contrast to the southern route sometimes suggested<sup>2</sup>. The geographical distribution of loci, such as YAP, focuses attention on specific novel questions about population origins. Although the *Alu* insertion is found at highest frequency in Africa, it is found at moderate frequency in Japan and at low frequency in some other areas, such as Europe<sup>21</sup>. Its presence outside Africa is often not due to recent admixture, but does it represent, for example, admixture a few centuries ago or the presence of YAP<sup>+</sup> chromosomes in the founding populations? These questions cannot be answered yet, but more detailed analysis of the Y chromosomes should allow such possibilities to be distinguished.

Work of this kind should eventually identify unambiguous paternal lineages running through the human population. It will complement work on maternal lineages. However, the Y chromosome ancestor need not have been the ancestor of any other sequences in the genome, and it is possible that he lived at a different time in a different place to the mtDNA ancestor. Neither the Y chromosome nor mtDNA seems likely to provide evidence about events dating back more than a few hundred thousand years. Genetics provides a fresh approach to the problems of human evolution, but if reliable conclusions are to be drawn it will be necessary to combine data from many parts of the genome with the more traditional studies.

## Acknowledgements

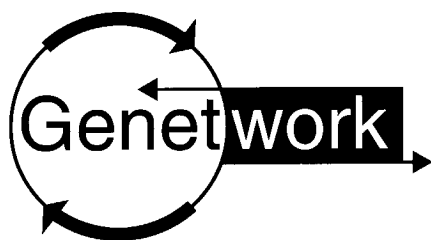
We thank all the individuals who have made this work possible by donating their DNA or collecting samples; Steven Gerken, Ross Hawkins, Manfred Kayser, Peter de Knijff, Cheni Kwok, Arne Nystuen, Sergio Pena, Lutz Roewer and Fabricio Santos for providing unpublished information; and John Armour, William Brown, Gabby Dover, Nathan Ellis, Michael Hammer, Rosalind Harding, Alec Jeffreys, Lutz Roewer and Ed Southern for advice and comments on the manuscript. M.A.J.'s work has been supported by the MRC and the EC; C.T.S.'s work has been supported by the CRC, the MRC, the BBSRC and The Royal Society.

## References

- 1 Stoneking, M. (1993) *Evol. Anthropol.* 2, 60–73
- 2 Lahr, M.M. and Foley, R. (1994) *Evol. Anthropol.* 3, 48–60
- 3 Affara, N.A. *et al.* (1994) *Cytogenet. Cell Genet.* 67, 360–380
- 4 Casanova, M. *et al.* (1985) *Science* 230, 1403–1406
- 5 Lucotte, G. and Ngo, N.Y. (1985) *Nucleic Acids Res.* 13, 8285
- 6 Jakubiczka, S. *et al.* (1989) *Hum. Genet.* 84, 86–88
- 7 Malaspina, P. *et al.* (1990) *Ann. Hum. Genet.* 54, 297–305
- 8 Spurdle, A. and Jenkins, T. (1992) *Hum. Mol. Genet.* 1, 169–170

- 9 Seielstad, M.T. *et al.* (1994) *Hum. Mol. Genet.* 3, 2159–2161
- 10 Dorit, R.L., Akashi, H. and Gilbert, W. (1995) *Science* 268, 1183–1185
- 11 Nickerson, D.A. *et al.* (1992) *Genomics* 12, 377–387
- 12 Miyata, T. *et al.* (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 863–867
- 13 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* pp. 40–50, Cambridge University Press
- 14 Chagnon, N.A. (1972) in *The Structure of Human Populations* (Harrison, G.A. and Boyce, A.J., eds), pp. 252–282, Clarendon Press
- 15 Roewer, L. *et al.* (1993) in *DNA Fingerprinting: State of the Science* (Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J., eds), pp. 221–230, Birkhäuser Verlag
- 16 Oakey, R. and Tyler-Smith, C. (1990) *Genomics* 7, 325–330
- 17 Jobling, M.A. (1994) *Hum. Mol. Genet.* 3, 107–114
- 18 Roewer, L. *et al.* (1992) *Hum. Genet.* 89, 389–394
- 19 Mathias, N., Bayés, M. and Tyler-Smith, C. (1994) *Hum. Mol. Genet.* 3, 115–123
- 20 Tyler-Smith, C. and Hammer, M. (1995) *Y Chromosome Consortium Newsletter* 2, 4–5
- 21 Hammer, M.F. (1994) *Mol. Biol. Evol.* 11, 749–761
- 22 Nakahori, Y., Tamura, T., Yamada, M. and Nakagome, Y. (1989) *Nucleic Acids Res.* 17, 2152
- 23 Slatkin, M. and Hudson, R.R. (1991) *Genetics* 129, 555–562
- 24 Di Rienzo, A. and Wilson, A.C. (1991) *Proc. Natl Acad. Sci. USA* 88, 1597–1601
- 25 Hammer, M.F. and Horai, S. (1995) *Am. J. Hum. Genet.* 56, 951–962
- 26 Lin, S.J. *et al.* (1994) *Jpn. J. Hum. Genet.* 39, 299–304
- 27 Torroni, A. *et al.* (1990) *Ann. Hum. Genet.* 54, 287–296
- 28 Jobling, M.A., Fretwell, N., Dover, G.A. and Jeffreys, A.J. (1994) *Cytogenet. Cell Genet.* 67, 390
- 29 Bowcock, A.M. *et al.* (1994) *Nature* 368, 455–457
- 30 Distche, C.M. *et al.* (1986) *Proc. Natl Acad. Sci. USA* 83, 7841–7844
- 31 Nakahori, Y. *et al.* (1994) *Hum. Mol. Genet.* 3, 1709
- 32 Chen, H., Lowther, W., Avramopoulos, D. and Antonarakis, S.E. (1994) *Hum. Mut.* 4, 208–211
- 33 Griffiths, R.C. and Tavaré, S. (1994) *Statistical Sci.* 9, 307–319

**M.A. JOBLING IS IN THE DEPARTMENT OF GENETICS, UNIVERSITY OF LEICESTER, UNIVERSITY ROAD, LEICESTER, UK LE1 7RH; C. TYLER-SMITH IS IN THE CRC CHROMOSOME MOLECULAR BIOLOGY GROUP, DEPARTMENT OF BIOCHEMISTRY, UNIVERSITY OF OXFORD, SOUTH PARKS ROAD, OXFORD, UK OX1 3QU.**



## FlyBase: a virtual *Drosophila* cornucopia

Less than ten years ago there was only a single source of information that compiled all our knowledge on the fruitfly: *Genetic Variations of Drosophila melanogaster*, affectionately known as 'The Redbook', written by Dan Lindsley and Ed Grell<sup>1</sup>. One of us (H.B.) clearly remembers complaining to his former advisor for several years (1983–1985) that *The Redbook* was really out of date and that there was an urgent need for an updated version. First, relief came at the end of 1985 with *Genes from A–K* by Lindsley and Zimm<sup>2</sup>, rapidly followed by a series of useful *Drosophila*

Information Service publications, culminating in Ashburner's *Drosophila Genetic Maps*<sup>3</sup> and *The New Redbook*<sup>4</sup>. The plethora of organized knowledge that is now available on the fly is flabbergasting, and the speed by which it is updated is remarkable and very laudable. Here, we discuss how to access FlyBase<sup>5</sup>, what is available, and comment on some of the useful and less useful aspects of the database. Interested readers who want to know more about the history of FlyBase and those who are involved are referred to Ashburner and Drysdale<sup>6</sup>.

FlyBase is a comprehensive database, which contains information on genetics, molecular and cell biology of *Drosophila melanogaster*. The database is expanding daily as new information is added on, for example, stock lists, bibliographical references, genes and alleles, cloned regions. As of September 1995, FlyBase contained information on 25 000 alleles and 9000 genes, distilled from over 73 000 publications. This information can be browsed in many different ways and can be accessed through the WWW, ftp, Gopher and GopherMail (see Box 1). Once in Flybase the reader will find an extensive list of information as well as a list of the search tools that can be used to access these data.

The continuous stream of information, published as well as unpublished (the latter is mainly generated by the genome projects and the stock centers), that must be entered in the database presents a major challenge for the FlyBase project. Efficient data entry is quintessential as approximately 3000 publications appear yearly on *Drosophila*. Data generated and curated by the European and Berkeley *Drosophila* Genome Projects are also integrated into the database. The data of the Berkeley *Drosophila* Genome Project (BDGP) are so far only partially available in FlyBase, though all Berkeley data are available on CD-ROM, which is produced jointly by the BDGP and FlyBase under the name *Encyclopaedia of Drosophila*. This is a database program derived from ACEDB, the *Caenorhabditis elegans* database, which contains much of the bibliographic and genetic data, as well as the BDGP data, of FlyBase. This complicates matters at the present time as, for example, not all the mapped *P*-elements and clones shown on the CD-ROM are available in FlyBase, necessitating double searches. It should be noted, however, that it is the goal of the BDGP and FlyBase to have all the information now available on CD-ROM integrated in FlyBase in the near future. A welcome move!