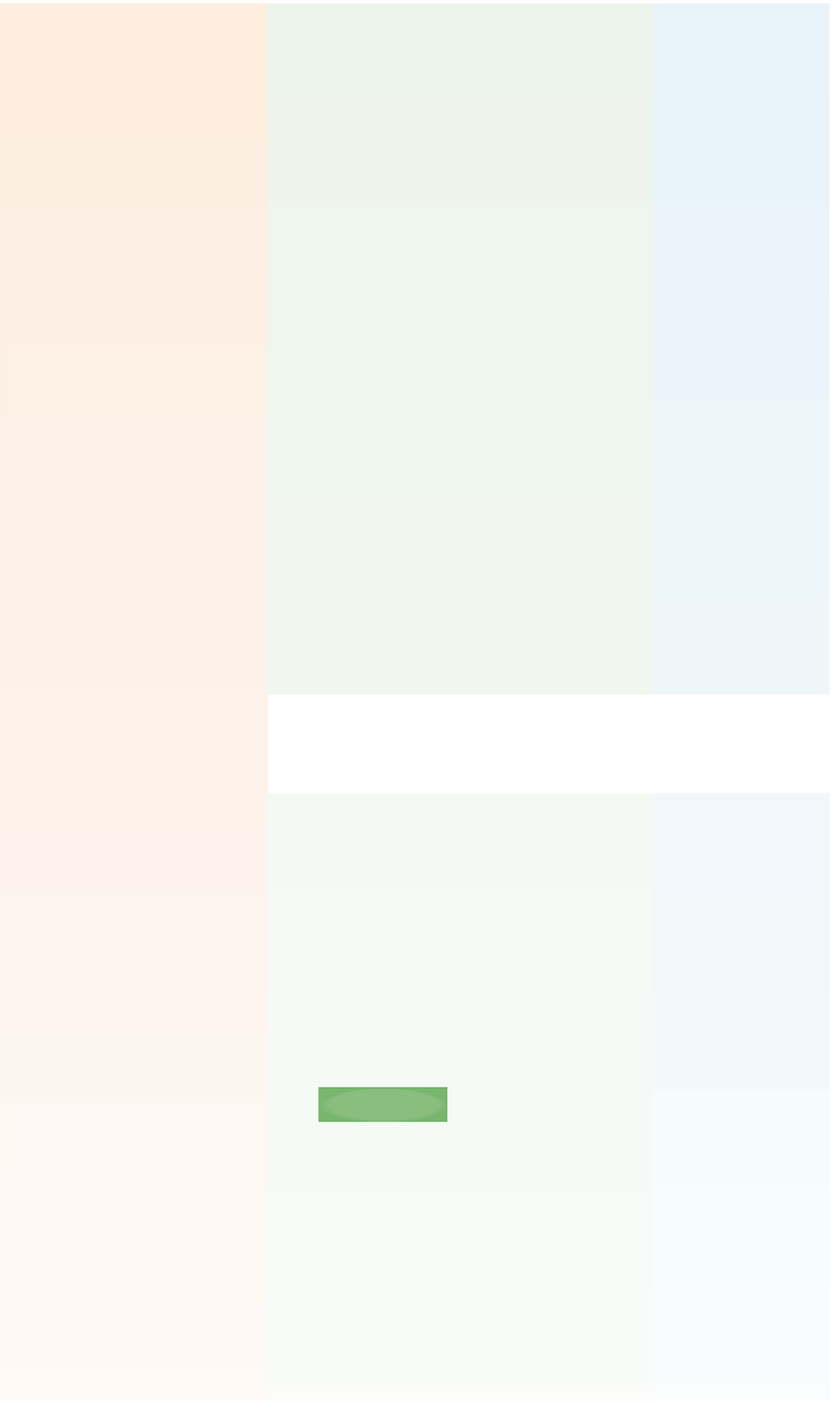
The properties of the Y chromosome read like a list of violations of the rulebook of human genetics: it is not essential for the life of an individual (males have it, but females do well without it), one-half consists of tandemly repeated SATELLITE DNA

genes (FIG. 1): but how complete is this catalogue, what proportion of genes can be identified and what do they do? Efforts to discover genome-wide sequence variation have identified vast numbers of Y-specific single nucleotide polymorphisms (SNPs): the Ensembl database lists 28,650 at the time of writing, which might seem enough to provide an extremely detailed PHYLOGENETIC TREE of Y-chromosomal lineages. But how many of these SNPs are real, and how many are artefacts that are produced by unknowingly comparing true Y-chromosomal sequences with similar sequences (PARALOGUES) elsewhere on the same or other chromosomes[2]? Also, are these SNPs a representative set of sequence variants from the human population as a whole? The answer is no, because of ascertainment bias (BOX 1) in the range of populations that were surveyed for variation. As well as this possible treasure trove of unverified SNPs, the availability of Y-chromosome sequence means that there are now more than 200 binary polymorphisms that are well characterized, and 100–200 potentially useful new MICROSATELLITES, as well as the ~30 published polymorphic tri-, tetra- and pentanucleotide repeat markers. Finally, there is a robust and developing phylogeny of Y-chromosomal haplotypes (FIG. 3) that are defined by binary polymorphisms (haplogroups), and a unified nomenclature system[3] that allows diversity data from different research groups to be readily integrated.



PHYLOGENETIC TREE
A diagram that represents the evolutionary relationships between a set of taxa (lineages).

PARALOGUES
Sequences, or genes, that have originated from a common ancestral sequence, or gene, by a duplication event.

MICROSATELLITE
A class of repetitive DNA sequences that are made up of tandemly organized repeats that are 2-8 nucleotides in

polymorphic and are frequently used as molecular markers in population genetics studies.

HAPLOGROUP
A haplotype that is defined by binary markers, wOich is more stable but less detailed than one defined by microsatellites.

Y chromosome than elsewhere in the nuclear genome, which is indeed observed[4,5]. We also expect it to be more susceptible to genetic drift, which involves random changes in the frequency of haplotypes owing to sampling from one generation to the next. Drift accelerates the differentiation between groups of Y chromosomes in different populations — a useful property for investigating past events. However, because of drift, the frequencies

NRY, NRPY AND MSY
Several neologisms have been introduced to refer to the portion of the Y chromosome that excludes the pseudoautosomal regions, for example, non-recombining Y (NRY), non-recombining portion Y (NRPY) and male-specific Y (MSY), but none has achieved wide acceptance.

EFFECTIVE POPULATION SIZE
The size of an idealized population that shows the same amount of genetic drift as the population studied. This is approximately 10,000 individuals for humans, in contrast to the census population size of $>6 \times 10^9$.

PARAGROUP

As with all regions of human DNA — except for the mtDNA control region — base substitutional mutation occurs at too low a rate to be analysed directly. However, the secure phylogenetic framework and haploidy of the Y chromosome mean that recurrent mutations can be identified unambiguously, and data that accumulate from resequencing will provide information about the mutational properties of individual bases. The human population is so large that, even given the low average mutation rate of $\sim 2 \times 10^{-8}$ per base per generation[11], we expect recurrent mutations to occur at every base of the Y chromosome in each global generation. However, these modern recurrences will usually go undetected. At present, the number of recurrent base substitutions in the Y Chromosome Consortium (YCC) tree is only five (REF. 3), although this is likely to increase as more chromosomes are typed.

Studies of genetic diseases show a strong bias towards fathers as the source of new mutations, and also show increasing mutation rate with paternal age (reviewed in REF. 12). The explanations generally used for these two observations are, respectively, the larger number of cell divisions (and hence DNA replications) in male than in female gametogenesis, and the increase of mutation rate with time through continuing divisions of spermatogenic stem cells. As Y chromosomes pass only through the male germline, its mutagenic properties affect the Y chromosome more than any other. The ratio of male to female mutation rates — the α-factor (α) — can be estimated by comparing the number of mutations that have accumulated in homologous autosomal Y-chromosomal and X-chromosomal sequences over a given time period. Estimates for α vary considerably between studies, but all show a significantly higher mutation rate in the male

microsatellites in different lineages (for example, REF. 23) that there are marker- and allele-specific differences in mutation rates. In principle, direct analysis of sperm DNA (which is ideal for Y chromosomes) offers access to these rates. The only study to attempt this gave mutation rat

does not imply its absence in other populations in which different lineages predominate, so further studies on a larger scale are warranted to take advantage of improved genotyping methods[36,37] and phylogenetic r

populations remains unclear, but the use of 35 years

so frequencies are known approximately and rare lineages often remain undetected. Evidence of the presence of a lineage is usually reliable, but a lack of evidence does not prove absence. Present distributions are the culmination of many past events. These include some relatively recent ones: intercontinental travel is now common, and will be of great interest to future evolutionary geneticists, but does not concern us today. Similarly, migrations during the past 500 years, although of profound modern epidemiological and forensic significance, are not usually the main focus of attention. It is assumed that recent events can be iden-

haplogroups originated too recently to have been present at the time of these initial migrations, but their present distributions could reflect the earlier movements of their precursors. mtDNA, the only other locus for which comparable phylogenetic data are available, also shows a general distinction between southeastern Asia/Australia where mtDNA haplogroup M and its derivatives predominate, and

number, the patterns of diversity that were established largely by migration and drift during the Palaeolithic period have been 'frozen', and large-scale changes have become less frequent. Nevertheless, recent events can still occasionally have a significant influence on Y-chromosome diversity.

Comparisons of Y-chromosome data with mtDNA data have been particularly revealing about the sex-specific gene flow that accompanied the expansion of Europeans into the Americas and Oceania in the past 500 years. A typical pattern of strong introgression of European Y chromosomes with retention of indigenous mtDNA lineages is seen in Polynesia[81], Greenland[82,83] and South America[84–86], which reflects the sexual politics of colonial activity.

Conclusions

Development of the Y chromosome as an informative system for evolutionary studies in the 18 years0.1ion ofmD2-

identification of regions of low recombination on the autosomes (HAPLOTYPE BLOCKS) now indicates that this might be possible (BOX 6).

Sequencing of the chimpanzee genome is underway, and promises a cornucopia of information about the evolution of our own genome. Assembly of a chimp genome sequence using the human sequence as a framework