THE X CHROMOSOME IN POPULATION GENETICS

Stephen F. Schaffner

Genetic variation records a large amount of information about the history of a species and about the processes that create and shape that variation. Owing to the way in which it is inherited, the X chromosome is a rich resource of easily accessible genetic data, and therefore provides a unique tool for population-genetic studies. The potential of the human X chromosome, which rivals that of the more traditional mtDNA and Y chromosome, has only just begun to be tapped.

PHYLOGENETIC TREE
A graph that depicts the ancestor–descendant relationships between organisms or gene sequences. The sequences are the tips of the tree. Branches of the tree connect the tips to their (unobservable) ancestral sequences.

The X chromosome has a curious role in population genetics. It is present in a single copy in males, which makes it easier to study than the autosomes. This trait, which it shares with the mitochondrial genome (mtDNA) and the Y chromosome, explains its use in an increasing number of studies, especially those that address the history of the human population. Nevertheless, the X chromosome continues to be overshadowed in these studies by the Y chromosome and mtDNA, despite containing far more genetic information than either. One reason for the lagging position of the X chromosome has been a delay in the generation of useful data. The first PHYLOGENETIC TREE for human mtDNA was published in 1987 (REF. 1), and the first for the Y chromosome was published in 1989 (REF. 2). By contrast, population-genetic studies of the X chromosome had to wait until the advent of practical DNA-sequencing technology: the first detailed sequence studies of the X chromosome appeared in 1997 and 1998 (REFS 3–6), and the first phylogenetic trees only in 1999 (REFS 7–9). Since then, X-chromosome-based studies have slowly accumulated; with the maturation of sequencing and genotyping technologies, a tremendous increase in all kinds of genetic data looms. It is therefore an appropriate time to reflect on what the X chromosome can offer to population-genetic studies.

In this article, I describe some of the distinctive features of the X chromosome, and how these allow us to address several biological questions, such as how mutation rates and recombination vary between the sexes, and how natural selection has influenced the history of our species. I then discuss the historical studies that have used the X chromosome, particularly for investigating the origin of non-African populations, and the much larger potential of the X chromosome for addressing the history of populations and anthropological questions about historical differences between males and females. Throughout the article, I focus primarily on human genetics, as this is the area to which the X chromosome has contributed most, although many of the issues raised here will no doubt arise in the study of other mammals.

Whitehead/MIT Center for Genome Research, Cambridge, Massachusetts 02139, USA.
e-mail: sfs@genome.wi.mit.edu
doi:10.1038/nrg1247

The X chromosome as a marker system

Vestige of an autosomal past. The usefulness of a marker for population-genetic studies depends both on its intrinsic characteristics and on those of the population in which it is being studied (BOX 1; TABLE 1). In this context, what does the mammalian X chromosome look like? In many respects, it looks very similar to an autosome, which is not surprising given its history. The two sex chromosomes, the X and the Y, diverged from a single autosome ~300 million years ago[10]; indeed, they are still homologous and recombine with each other near their ends, in the two pseudoautosomal regions. (Note that although the pseudoautosomal regions are interesting in themselves — for example, the recombination rate in the major pseudoautosomal region is 20 times the genome average[11] — their unique character puts them outside the scope of this paper.) Elsewhere, however, they have taken different evolutionary paths. The Y chromosome has lost the bulk of both its sequence and its genes, and has developed a unique pattern of repeated sequence[12,13]. By contrast, the X chromosome

Box 1 | **What's in a locus?**

The usefulness of a locus for population-genetic studies depends on three main characteristics: its age, its mutation rate and its recombination rate. These charactersitics are described below, and the different markers used in population-genetic studies are compared in TABLE 1 with respect to these characteristics and their consequences.

Age

The age of a locus is the time to the most recent common ancestor (MRCA) of all extant copies of it. The age of a locus defines the period from which genetic variation has been preserved, and therefore delimits the historical period that can be investigated using that locus.

For human population-genetic studies, interesting timescales include: a few thousand years for recent historical events; ~50,000 years for the expansion of *Homo sapiens* out of Africa; and 100,000–200,000 years for the emergence of modern *Homo sapiens*.

Mutation rate

The mutation rate can make a locus uninformative by being either too high or too low. If the rate is too low (compared with the age of the locus), there will be too little genetic variation to study, but if the rate is too high, recurrent mutations will occur at every site and will obscure the process under study. In practice, single-base substitution rates (those producing single nucleotide polymorphisms, or SNPs) are low enough that recurrent mutation can be neglected, at least when studying a single species. The exception is mtDNA, which mutates at a much higher rate than the rest of the genome. Mutation rates at MICROSATELLITE markers, on the other hand, are high enough that recurrent mutations become problematic after a few tens of thousands of years.

Recombination rate

The recombination rate controls the size of a chromosomal region that shares a single genealogical history. In the absence of recombination, every chromosome falls on a unique branch of the phylogenetic tree, whereas recombination combines different branches onto a single physical chromosome. Too much recombination, therefore, makes reconstruction of phylogenetic trees impossible, as individual markers represent different trees.

Mutation and recombination rates are biologically determined, whereas the age of a locus is primarily a function of the size of the population — larger populations tend to have older loci. Together, the recombination rate and the age determine the amount of LINKAGE DISEQUILIBRIUM (LD) that is observed because recombination over time breaks down LD.

has not lost its autosomal character: the X chromosome is physically the most stable nuclear chromosome, at least among placental mammals, and is the only one to retain complete large-scale SYNTENY between mouse and human[14]. The size of the X chromosome — 150 million base pairs (Mb) in humans — is consistently ~5% of the genome among mammals. Its gene density is low, ranking about seventeenth among the 24 human nuclear chromosomes[15].

Low diversity. The distinctive characteristics of the X chromosome largely derive from how it is inherited. Males have only one copy of the X chromosome, so every existing X chromosome has spent two-thirds of its history in females. Consequently, mutations occur less frequently on the X chromosome than on autosomes because the nucleotide mutation rate in females is several-fold lower than in males[16–19]; the result is lower genetic diversity, as well as smaller interspecies

Table 1 | **Comparison of population-genetic markers**

	Marker type				References
	mtDNA	**Y chromosome**	**X chromosome**	**Autosomes**	
Size (Mb)	0.017	60	150	3,000	15,16,59
Number of usable loci	1	1	Hundreds	Thousands	–
Mutation rate (mutations per Mb per generation)	Very high (1–300)	High (0.033)	Low (0.015)	Moderate (0.020)	19,60,61
Recombination rate (cM/Mb)	0	0	0.8	1.1	21
Diversity (fraction of discordant base pairs)	Very high (0.4%)	Low (0.02%)	Moderate (0.04%)	High (0.08%)	42,62
Accessible haplotypes*	Yes	Yes	Yes	No	–
Genetic drift‡§	High	High	Moderate	Low	–
Age§	100,000 years	100,000 years	750,000 years	1,000,000 years	–
Effective population size (relative to autosomes)	1/4	1/4	3/4	1	–

*A haplotype is a set of genetic markers that is present on one chromosome; ‡genetic drift describes the random changes in allele frequency that occur because genes that appear in offspring are not a perfectly representative sample of the parental genes (for example, as occurs in small populations); §these entries are approximate population genetics inferences, based on the consensus estimate for the effective population size in humans. cM, centiMorgan; Mb, megabase; mt, mitochondrial.

MICROSATELLITE
A class of repetitive DNA that is made up of repeats that are 2 to 8 nucleotides in length. They can be highly polymorphic and are frequently used as molecular markers in population-genetic studies.

LINKAGE DISEQUILIBRIUM
A nonrandom correlation between alleles of physically linked loci.

SYNTENY
Collinearity in the order of genes (or of other DNA sequences) in a chromosomal region of two species.

Box 2 | Genetic diversity of the X chromosome and autosomes

The lower mutation rate and the smaller population size of the X chromosome, compared with autosomes, lead to an unambiguous prediction that genetic diversity should also be lower there. As measurements of the diversity at individual loci on the X chromosome and on autosomes vary widely (as discussed in BOX 3), a direct comparison of the overall diversity between the X chromosome and autosomes is difficult. Despite experimental difficulty, however, several human studies support this prediction. An analysis[56] of 47 kb of sequence on the X chromosome found diversity to be 80% of that on the autosomes — which is surprisingly high. Larger studies of the X chromosome have found values more in line with expectation: a study of variation in genes[23] in a total of ~70 kb found the ratio of the diversity between the X chromosome and autosomes to be 48%, and the large SNP-discovery project carried out by THE SNP CONSORTIUM (TSC), which analysed 50 million base pairs (Mb), found the ratio to be 59%, which corresponds to diversity on the chromosome of one difference in every 2,000 bp[42].

In non-African populations, which are genetically less diverse than African ones, the diversity on the X chromosome can be markedly lower than on the autosomes. In one study (S. F. S. *et al.*, unpublished observations, listed in TABLE 2), 40% of all polymorphisms showed no variation in a sample of East Asian chromosomes, whereas only 10% showed no variation in a West African sample. In a similar vein, extensive 'SNP deserts' — regions (some of which are more than a million base pairs long) with very few SNPs — were found on European X chromosomes[57].

divergence, on the X chromosome. According to one study, for example, the divergence between human and chimpanzee X chromosomes is 83% of that observed between the autosomes of these two species[20]. Diversity is further reduced by the EFFECTIVE POPULATION SIZE (N_e) of the X chromosome, which, because males lack a second copy, is three-quarters that of the autosomes. The net effect is that diversity on the X chromosome (most often measured as HETEROZYGOSITY, or π) should be about half of that on the autosomes. This expectation has been borne out by several experiments (see BOX 2). Diversity on the X chromosome might be low compared with the autosomes, but it is still twice that on the Y chromosome, and it provides enough variant sites (~1–10 per 1,000 base pairs (bp)) for reasonably sized regions to be informative. In terms of age, autosomes record slightly older time periods than the X chromosome, but both record substantially older histories than either the Y chromosome or mtDNA (TABLE 1 and BOX 3).

Pronounced population structure. Another consequence of the smaller population size of the X chromosome is that genetic drift is faster for the X chromosome than for the autosomes. As a result, POPULATION STRUCTURE should be more pronounced on the X chromosome; that is, populations should differ more in their X chromosomes than in their autosomes. Indeed, as shown in TABLE 2, GENETIC DISTANCES between populations are significantly larger on the X chromosome.

Long linkage disequilibrium intervals. X chromosomes recombine only in females as males have a single copy; therefore, only two-thirds of X chromosomes recombine in each generation. The measured recombination rate for the X chromosome is, in fact, almost exactly two-thirds of the genome average[21]. As a result, we can expect linkage disequilibrium (LD) to be greater on the X chromosome and the size of regions with a single genetic history to be larger. This effect is reinforced by the younger age of the X chromosome, as younger loci have had less time for recombination to break down LD. Systematic comparisons of LD on the X chromosome and the autosomes have not been published for mammals; relevant data sets do exist for humans[22,23], however, and it is expected that their analysis will support the prediction that LD is greater on the X chromosome.

Biological studies

The biological questions that can be studied with the X chromosome flow from the characteristics described above. The female-dominated history of the X chromosome makes it an ideal system for studying population-genetic differences between males and females, particularly the differences in mutation rate and in patterns of recombination. The presence of a single copy of the X chromosome in males means that X-linked alleles are more exposed to natural selection, making the X chromosome an attractive place to examine the role of selection in human history.

Mutation rates in males and females. The higher mutation rate in male mammals, including humans, is generally attributed to the larger number of mitoses that germline cells go through in males. Direct measurement of the difference in rates is difficult, however, as the absolute rates are low (~2 x 10^{-8} mutations/bp/generation). The easiest way to obtain an accurate measurement has been to examine homologous regions on the X and Y chromosomes and to compare their divergence from the inferred ancestral sequence; a higher male mutation rate will be reflected in a higher rate of substitutions on the Y chromosome copy. Such studies have been done for humans[16,17,19,24,25] and for several other organisms (other primates[19,25], sheep and goats[26], and rodents[27,28]). The consistent finding has been that males do have higher mutation rates than females; most studies have also concluded that in humans and apes, the ratio is particularly high, presumably because of the long generation time and unusually large number of male mitoses in those species. By comparing all three systems (the X and Y chromosomes, and the autosomes), it has been possible to test an alternative hypothesis for the lower mutation rate on the X chromosome[29], namely that, independent of whether it is in males or females, the X chromosome has evolved an unusually low mutation rate to protect itself against deleterious mutations. The available data, however, do not support this hypothesis[19,20].

Despite the apparent simplicity of these measurements, controversy continues about the magnitude of the difference between the mutation rate in males and females[18]. Most studies have estimated about a fivefold higher rate in males, but two studies have yielded ratios close to two[16,17]. This lower value might be caused by a failure to take into account polymorphism present in the ancestral population, a possibility that is supported by the observation that divergences over longer evolutionary times yield higher ratios[19]; on the other hand, a

EFFECTIVE POPULATION SIZE (N_e). The size of the ideal population in which the effects of random drift would be the same as those seen in the actual population.

HETEROZYGOSITY
A measure of the genetic variation in a population: the mean number of differences found when comparing two copies of a sequence. Usually expressed as the number of differences per base pair.

THE SNP CONSORTIUM (TSC). A public–private effort that mapped approximately 1 million SNPs across the human genome.

POPULATION STRUCTURE
A departure from random mating as a consequence of factors such as inbreeding, overlapping generations, finite population size and geographical subdivision.

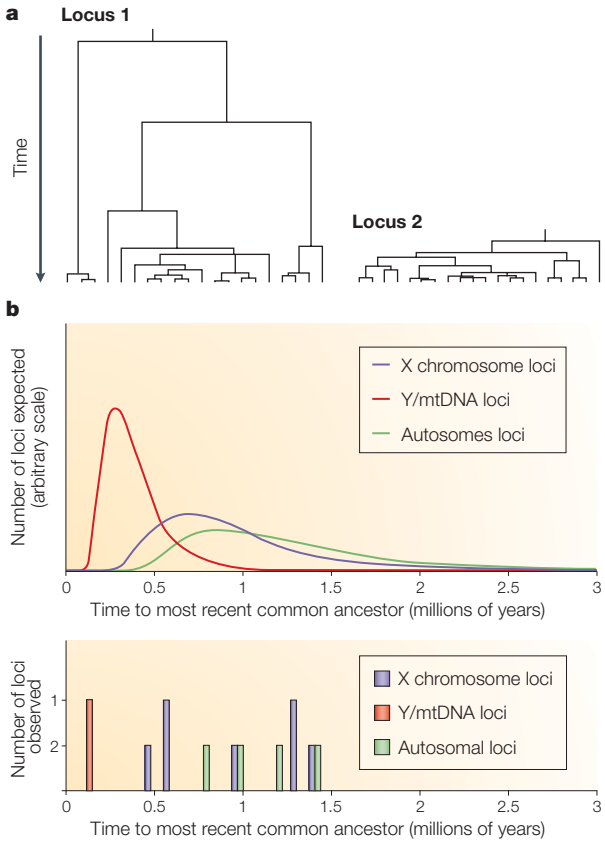
GENETIC DISTANCE
The degree of genetic differentiation between two populations. It is measured by comparing allele frequencies (and in the case of microsatellite markers, by comparing allele sizes) between populations.

Box 3 | **Variation between loci**

Unlike the Y chromosome and mtDNA, the X chromosome contains many independent loci, each with its own phylogenetic tree. It is a characteristic of genealogies, whatever chromosome they occur on, that they vary randomly; that is, under identical circumstances, the phylogenetic trees for two loci can be very different, both in shape and in depth.

The two trees shown in panel a are the result of simulations of a constant-sized population for two loci, and are typical of the amount of variation observed. Although the two simulated loci share an identical population history, the age (and therefore the diversity) of locus 1 is many times that of locus 2; inferring the characteristics of the population from either tree alone will therefore give a badly skewed result.

Panel b shows the full range of ages expected for the three types of chromosome, on the basis of an OUT OF AFRICA MODEL **of human origins. As the X chromosome has three times the effective population size of the Y chromosome or mtDNA, loci on the X chromosome can be expected to be much older; the same is true for autosomes, which have four times the effective population of the Y chromosome. Note the broad age range expected for different loci from the same type of chromosome. The histogram shows published estimates of the age of various loci; all have large uncertainties (not shown)[58]. Similar variation from locus to locus occurs in other inferences, such as those concerning** BOTTLENECKS **in population size, or about the source of migrations into a region.**



recent study[20] that accounted for this effect still reported a fairly low value for the ratio (~threefold). Studies into this unresolved question will probably continue.

Recombination in males and females. A related issue centres on the differences between males and females in the pattern of recombination. The overall rate of recombination is higher in women, and the rate varies along chromosomes differently for the two sexes; for example, recombination in males tends to be higher near the telomere, whereas in females it is higher near the centromere[21]. In males, the distribution of recombination can be studied directly at a very fine scale by typing DNA from a large number of sperm[11,30–33]. So far, these studies have shown that recombination is highly clustered in 'hot spots', with little recombination occurring elsewhere. It is not known how universal or variable this phenomenon is across the genome, and little is known about what determines where recombination occurs on either fine or coarse scales. The hope is that further sperm studies will identify these determinants. No comparable studies of female recombination will be possible, however, so the only access to fine-scale patterns of female recombination will be through the LD patterns that they leave behind. The obvious platform for such a study will be the X chromosome, because it is only there that female recombination occurs in the absence of male recombination.

Natural selection. Natural selection has two important functions in evolution: it allows adaptation (positive selection) and it suppresses deleterious mutations (background selection). The direct effect of the latter can be easily measured by observing the reduced polymorphism and interspecies divergence at coding sites. The extent of positive selection is less clear. It is not known how often episodes of positive selection occur or how important they have been in shaping our genome; nor is it clear what effect background selection has on overall neutral variation. The reason for considering the X chromosome in this context is that selection might be more visible there than on the autosomes. Two explanations could account for this; first, as recessive alleles on the X chromosome are exposed to selection in males regardless of their frequency, a larger fraction of mutations undergo selection; and second, higher LD on the X chromosome means that when selection does act on a locus, it is likely to affect a larger region than it would on

OUT OF AFRICA MODELS
Models for the origin of modern human populations in which anatomically modern *Homo sapiens* evolved ~100,000–150,000 years ago in Africa and expanded from there to the rest of the world.

BOTTLENECK
A marked reduction in population size followed by the survival and expansion of a small random sample of the original population.

Table 2 | **Measuring genetic distance using X-linked loci***

Human X-linked marker or gene	Number of loci	Number of markers	Size of DNA sequence analysed (kb)	Mean F_{ST}‡ (African versus non-African)	Reference
PDHA1	1	25	4	0.62	8
ZFX	1	10	1	0.08	7
dystrophin	1	36	8	0.04	5
Xq13.3	1	33	10	0.09	9
DACH2	1	44	10	0.02	44
Multilocus survey	16	250	110	0.26	Unpublished observations (S.F.S. *et al.*)
Total	21	398	143	0.24	–

*Values for *DACH2* and Xq13.3 were calculated from the published data; other values are quoted from the original publications; ‡ F_{ST} is a measure of genetic distance, based on allele frequencies, that indicates the proportion of genetic diversity found between populations relative to the amount within populations. Values range from 0.0 to 1.0. Because studies sample different populations, it is often difficult to compare directly genetic distance measurements from different studies; for this reason, only the large-scale measurement between African and non-African populations is given. On average, the divergence values for X-chromosome-linked loci, shown here, are significantly higher than those of autosomal loci, for which the F_{ST} value is typically between 0.1 and 0.15 (REF. 47); this indicates that the X chromosome has a more pronounced population structure than that of autosomes. Note also the large variation in genetic distance among the X-chromosome loci. *DACH2*, *dachshund* homologue 2; kb, kilobase; *PDHA1*, pyruvate dehydrogenase (lipoamide) α 1; *ZFX*, zinc finger protein, X-linked.

the autosomes. So, a SELECTIVE SWEEP will typically cover more sequence on the X chromosome. A mildly deleterious mutation, which reduces the number of viable chromosomes in the population, has a similar effect of reducing variation in a larger region than on autosomes.

Traces of selective sweeps (or at least good candidates for them) have indeed been found on the X chromosome[34,35]. Of more general interest is the relative importance of sweeps, and of selection generally, in the two types of chromosome. Several tests have been applied in the search for traces of selection. Perhaps the most robust of these involves detecting whether diversity varies in step with recombination. Because selective sweeps lower diversity, and because they extend further where there is little recombination, we expect less diversity where the recombination rate is low. (A similar effect also occurs for background selection[36,37].) A positive correlation between heterozygosity and recombination rate is, therefore, a signature of extensive selection. Such a correlation has been observed on human autosomes, but it is weak and can probably better be explained by factors other than selection[38,39] (FIG. 1a). Several studies have indicated that the correlation is larger on the X chromosome[6,40,41]; here, also, the evidence is weak, but is consistent with a greater role for selection on this chromosome. The large data set from the SNP consortium (TSC)[42], however, shows no correlation between heterozygosity and recombination rate at all on the X chromosome (FIG. 1b). A second indicator of selection would be a difference in diversity between the X chromosome and the autosomes, even after adjusting for their different N_e and mutation rates. Positive selection (but not background selection) would lower the diversity of the X chromosome compared with autosomes[43]. Here, too, supporting evidence of selection seems to be missing. The relative X-to-autosome diversity cited in BOX 2 (59% for the TSC data) agrees almost perfectly with the expected value of 58% that is obtained by assuming a ratio of male-to-female mutation rate of fivefold. Both the observed and estimated values have uncertainties attached to them, so a discrepancy — albeit probably only a small one — might still emerge. At present, therefore, there is no compelling evidence for a larger role for natural selection on the X chromosome. Other avenues remain to be explored, however, such as comparing the X chromosome and the autosomes in terms of genetic distance between populations, or comparing the frequencies of coding polymorphisms for the two types of chromosome.

SELECTIVE SWEEP
The process by which new, favourable mutations become fixed so quickly that physically linked alleles also become fixed by 'hitchhiking'.

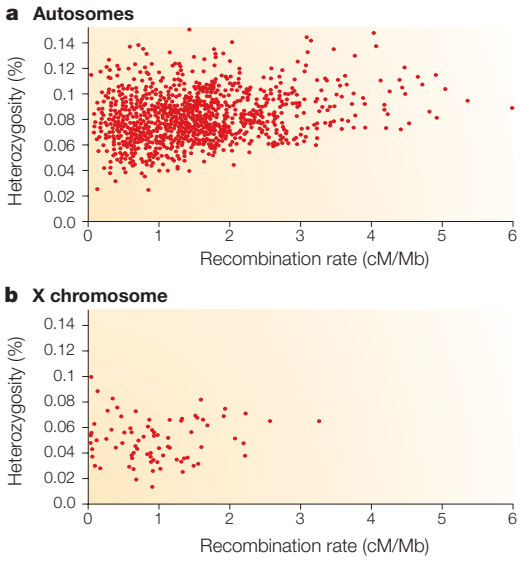


Figure 1 | **Signal for selection?** Natural selection produces a correlation between diversity (shown here as heterozygosity) and recombination rate[38–39]. **a** | Heterozygosity measurements from The SNP Consortium SNP-discovery survey[42] show a correlation on the autosomes, but the correlation is weak and probably results from factors other than selection[38–39]. **b** | A much stronger correlation on the X chromosome would indicate a larger effect of natural selection there, but the same data show no correlation between diversity and recombination rate on this chromosome.

Studies of human history

It is a different set of features that makes the X chromosome particularly valuable for historical research. First, there is a matter of practicality. Because there is a single copy of the X chromosome in males, it is easy to determine haplotypes (BOX 4), a feature that is not present in the autosomes. As haplotypes are needed to infer the phylogeny of a region, the X and Y chromosomes, and mtDNA are the natural choices for studies that use phylogenies, an advantage that explains their dominant role in historical studies. Until technology to read single molecules of DNA becomes practical, their pre-eminence will probably continue, or even increase: development of high-throughput genotyping and sequencing technology has markedly expanded the possibilities for large-scale haplotype studies, whereas extracting haplotypes from autosomal loci remains labour intensive.

Second, there is the presence of recombination. If ease of haplotyping distinguishes the X chromosome from the autosomes, it is the occurrence of recombination that distinguishes it from the Y chromosome and mtDNA. In the latter two systems, the entire chromosome acts as a single locus and shares a single genealogical history. The X chromosome and the autosomes, on the other hand, are broken up by recombination at every generation, so that different regions have different histories. This difference has important implications for historical investigations. In practical terms, recombination makes it harder to study phylogenetic trees and so X chromosome-based phylogenies must be restricted to regions with very strong LD. In broader terms, however, recombination creates a tremendous resource: it means that the X chromosome records hundreds or thousands of different snapshots of the population's history, whereas the Y chromosome and mtDNA each record only a single one. Because the history of any single locus only crudely records the history of the population in which it lived (see BOX 3), information from a recombining system is crucial for providing as complete a view as possible of the history of human populations. It is the combination of accessible haplotypes and multiple genetic histories, therefore, that makes the X chromosome a uniquely powerful tool for historical studies.

Historical studies can be divided into two types: those that reconstruct gene phylogenies on the basis of haplotypes and those that use SUMMARY STATISTICS to infer properties of ancestral populations. Owing to their dependence on haplotypes, it is the phylogenetic studies that can be expected to have a particular focus on the X and Y chromosomes, and mtDNA.

Studies can also be classified by their focus. The broadest studies look at global characteristics of a population, typically using summary statistics as their basis. A second class looks at the geographical history of subpopulations and the phylogenetic relationships between them. In studies of human history, an approximate but useful distinction can be made, within the second class, between studies that focus on large-scale patterns, especially those that aim to distinguish between Out of Africa and MULTIREGIONAL MODELS, and those that analyse the history of smaller geographical areas or ethnic groups. As discussed below, these three categories — global population characteristics, large-scale history and local history — have varied considerably in the extent to which they have used the X chromosome.

Global population characteristics. Global features of a population include N_e, the occurrence of expansions or bottlenecks and the extent of subdivision within the population. All three features can be inferred from statistical information about the population. For example, because genetic diversity and population size are correlated, N_e can be inferred from the measured amount of diversity. Studies of this kind that have used X-chromosome loci have yielded estimates of N_e that range from 4,900 to 37,000 for humans[5–7,44]; this wide range is expected because of the large stochastic variation between loci (BOX 3). A global estimate based on resequencing a third of the X chromosome yields an N_e of 12,000 (REF. 42). Similar studies can be done using other genetic systems, and the autosomes, Y chromosome and mtDNA have all been used for them. The only reason the X chromosome is over-represented here is that X-chromosome loci are often chosen for haplotype analysis, and because summary statistics can conveniently be studied at the same time, using the same data.

The X chromosome could provide a unique resource for one topic to which it has so far not contributed: the way in which population-genetic parameters differ between males and females. Numerous mechanisms might contribute to this phenomenon. For example, different breeding and mortality patterns can produce differences in the average length of a generation, and polygamy increases female N_e relative to that of males. PATRILOCALITY produces larger genetic distances between groups for males than for females, whereas military conquest has the opposite effect. The question is not whether these processes have operated in human history — they all have — but how important the different effects have been. Comparisons between mtDNA and the Y chromosome have shown that genetic distances tend to be larger for males than for females; the cause of the difference is very controversial[45–49]. As the X chromosome spends twice as much time in females as in males, differences between males and females will also be reflected in differences between the X chromosome and the autosomes. Given the complexity of the questions, it would be useful to bring the large body of genetic information available on the X chromosome to bear on them. Until now, however, the power of any given X chromosome-linked locus to address these questions has been too low, simply because of the mixture of male and female histories on the X chromosome. For example, a parameter (such as diversity) that is (hypothetically) 50% higher in a purely female lineage than in a male one will only be 7% higher on the X chromosome than on the autosomes. In the long run, however, the volume of data from the X chromosome promises to provide an independent perspective on questions that neither the Y chromosome nor mtDNA is able to resolve.

SUMMARY STATISTIC
A single number that summarizes complex data; examples include mean and variance.

MULTIREGIONAL MODELS
Models for the origin of modern human populations in which anatomically modern *Homo sapiens* evolved simultaneously throughout Africa, Europe and Asia.

PATRILOCALITY
A residential pattern in which a married couple settles in the husband's home or community.

Large-scale geographical studies: Africa and beyond. The second class of study provides the best illustration of the potential of the X chromosome among existing phylogenetic studies. The central issue in these studies has been the debate about the origin of non-African populations, between multiregional and Out of Africa models, a debate that has largely been settled in favour of the latter. Much of the evidence invoked in the course of the debate has been in the form of phylogenetic studies that give clues to the source of all modern populations. Naturally, the Y chromosome and mtDNA have played a part, each contributing one phylogeny; several autosomal studies have also been done[50–53]. The largest source of evidence, however, has been the X chromosome[5,7–9,44,54].

Taken together, these studies indicate an Out of Africa origin for modern humans. Individually, however, the studies point to diverse conclusions. For example, two studies reconstructed haplotypes in introns of X chromosome-linked genes, using chromosomes drawn from many populations around the world. The first study[7] found a pattern that was strongly indicative of a recent expansion (and subsequent isolation by distance) of humans from a single geographical source. The chromosomes showed two common, old (200,000–800,000 years) haplotypes with worldwide distribution, as well as many younger, derived haplotypes (~100,000 years old) with limited geographical distribution. Africa was identified as the probable source for the expansion on the basis of the higher genetic diversity there, and by the continued presence of the ancestral human haplotype (as determined by comparison with other primates). The second study[44], on the other hand, found a pattern that was more consistent with multiregionalism. At this locus, most haplotypes were not shared between Africans and non-Africans, indicating independent histories. Moreover, within the phylogenetic tree, haplotypes from different geographical areas were often intermingled, indicating no single source. The exception was one group of related haplotypes that were exclusively non-African, and that were estimated to be old enough (66,000–264,000 years) to make it likely that they predated the putative Out of Africa migration (thought to have occurred ~50–60,000 years ago). These two studies, therefore, point to opposite conclusions about the basic structure of human populations. It is for this reason that many loci have been (and will continue to be) needed.

Fine-scale geographical studies. Studies on smaller geographical and ethnic units abound, with the focus ranging from the colonization of the Americas and the Pacific to the history of small, isolated populations such as the Andaman Islanders. The studies that use the X chromosome have so far been virtually absent from these efforts, not because they would not be valuable but because the bulk of the work, and all of the phylogenetic effort, has been done with the Y chromosome and mtDNA. The reason for this dominance is partly the amount of work that has already been done using these systems; the existing phylogenetic trees and geographical information about Y chromosome and mtDNA haplotypes provide a context for all new studies. In addition, the Y chromosome and mtDNA are peculiarly suited to tracking recent local history. They have higher mutation rates and much higher rates of genetic drift than the X chromosome, which mean that populations will tend to differ more at these loci than they do at an X-chromosome locus, allowing finer resolution for local studies. On the other hand, this last characteristic might not always be an advantage: faster drift also means that relationships between populations are more easily erased, sometimes making it harder to identify source populations. (It is interesting to speculate, for instance, that X-chromosome loci will identify the nearest Asian relatives of native Americans more successfully than the Y chromosome and mtDNA have done.)

Many more loci are needed, however, to unravel the complexities of the history of populations. This is why, despite its minimal contribution so far, the X chromosome has tremendous potential for this kind of work. This potential will only be realized if there actually are many suitable loci on the X chromosome — that is, regions with markers in strong LD and containing enough markers to construct phylogenetic trees. Fortunately, this seems likely to be the case. It has become clear in the past several years that haplotypes in much of the genome form block-like structures, which are a few kilobases to many tens of kilobases in length, with little recombination visible within them[31,55]. A block on the X chromosome that is a few tens of kilobases long can be expected to contain ~50 to 100 SNPs. On the basis of observations on the autosomes, we can expect to find hundreds of them on this chromosome. Furthermore, many of these blocks will also contain one or more microsatellite markers. This is useful because microsatellites, with their much higher mutation rate, make it possible to distinguish between closely related haplotypes, therefore improving the resolution for local studies. Shorter regions can also be used — no region studied so far has been much longer than 10 kb — but to achieve the resolution needed for localized studies, greater length is preferable. The amount of information that is available on the X chromosome is large enough to be daunting: duplicating the efforts that have already been made for the Y chromosome and mtDNA for each

Box 4 | **Haplotypes**

To reconstruct phylogenetic trees, it is first necessary to determine which haplotypes are present in the population. This is straightforward for mtDNA, the Y chromosome and (in males) the X chromosome, because in all three cases, an individual possesses a single haplotype. For the autosomes, however, the task is harder: as both copies of the locus are examined together, it is difficult to distinguish between alleles that belong to one haplotype and those that belong to the other. It is technically possible to examine only a single haplotype (for example, by using cloning or allele-specific PCR), but such methods add significantly to the cost and effort of a study. An alternative solution is to reconstruct the haplotypes statistically — that is, to estimate which alleles typically go together in the population as a whole. This technique works well for common haplotypes, but becomes increasingly unreliable for rarer ones, to the point that it is unworkable for haplotypes that combine SNPs and microsatellites.

X-chromosome locus in turn will be a major task, but one that will become easier as high-throughput screening and genotyping methods become more efficient.

Conclusions

The use of the X chromosome in population genetics is still in its infancy. It has already proved its worth in studies of the early history of modern *Homo sapiens*, but in most research areas its potential remains largely untapped. That potential is needed — the Y chromosome and mtDNA, despite their enormously fruitful contributions, are not very informative about some questions (such as the size of ancestral populations), and the information that they can provide about others (such as population history before the Out of Africa migration) has largely already been mined. The X chromosome is therefore the logical place to turn for more information. Many of the same questions can be addressed by either the X chromosome or the autosomes, but the X chromosome has a clear advantage in allowing easy access to haplotypes; the cost of extracting haplotypes from autosomes remains high, even as sequencing and genotyping become much faster and cheaper.

Much work is needed to realize the promise of the X chromosome. Most important, of course, will be more data: more markers, from more loci, studied in more populations. We will also need a better understanding of patterns of recombination and their causes and of mutation rates, as well as improvements in modelling the history of populations and in extracting information from summary statistics. Two continuing projects promise to provide some assistance. The first is the sequencing of the chimpanzee genome, which is due to be completed before the end of 2003. This will provide useful information of at least two kinds. The chimpanzee sequence usually records the ancestral state of sites that are variable in humans, and so provides a root to phylogenetic trees. More importantly, the comparison of the two genomes will give a detailed map of the variation in mutation rate across the chromosome, allowing better calibration of ages, at least for SNPs.

The second resource will be one generated by the International HapMap Project (see online links box), an effort to describe the haplotype and LD structure for every region of the human genome; initial data from this project are due in 2003. For the studies described in this article, the map will provide an easy way to identify regions of strong LD — regions with a single genealogical history — and therefore candidates for phylogenetic investigation. Although the map will also provide frequency information (in three populations) about more than 30,000 SNPs on the X chromosome, this information will be of uncertain value, partly because there has been a strong bias in the selection of markers towards those present at high frequency. Discovery of markers (SNPs and microsatellites) within the regions will still require extensive screening, but the identification of the regions themselves will be a considerable benefit.

In the long term, for investigations that rely on intrinsic features of the X chromosome (such as studies of differences between males and females in mutation, recombination and history, and studies into the behaviour of exposed recessives under natural selection), the role of this chromosome will remain unaltered. By contrast, for investigations that rely on X chromosome loci because they are easy to haplotype, including most historical studies, the situation might change. If technology becomes available that can cheaply and reliably extract haplotype information directly from diploid chromosomes (for example, by sequencing individual molecules of DNA), there will no longer be a reason to prefer the X chromosome to the autosomes. This does not mean that the role I have described for the X chromosome will be eliminated; it simply means that it will be broadened to include the 20-fold larger store of data that is present on the autosomes.

1. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
2. Lucotte, G., Guerin, P., Halle, L., Loirat, F. & Hazout, S. Y chromosome DNA polymorphisms in two African populations. *Am. J. Hum. Genet.* **45**, 16–20 (1989).
3. Hey, J. Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol.* **14**, 166–172 (1997).
4. Zietkiewicz, E. *et al*. Nuclear DNA diversity in worldwide distributed human populations. *Gene* **205**, 161–171 (1997).
5. Zietkiewicz, E. *et al*. Genetic structure of the ancestral population of modern humans. *J. Mol. Evol.* **47**, 146–155 (1998).
   An early summary-statistic study of a human X-chromosome locus (8 kb of intronic sequence in the dystrophin gene). It is distinguished by the very large data samples that were used (250 chromosomes were used for SNP ascertainment and 860 chromosomes were genotyped).
6. Nachman, M. W., Bauer, V. L., Crowell, S. L. & Aquadro, C. F. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141 (1998).
   A study of diversity at seven X-chromosome loci in humans that focuses on evidence for natural selection.
7. Jaruzelska, J. *et al*. Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy. *Genetics* **152**, 1091–1101 (1999).
   An early haplotype-based study of a human X-chromosome locus (1.1 kb) that supported a classic Out of Africa model of human migration.
8. Harris, E. E. & Hey, J. X chromosome evidence for ancient human histories. *Proc. Natl Acad. Sci. USA* **96**, 3320–3324 (1999).
   A study of haplotypes in a 4.2-kb X-chromosome locus, showing surprisingly large differences in variation between African and non-African human populations.
9. Kaessmann, H., Heissig, F., von Haeseler, A. & Pääbo, S. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nature Genet.* **22**, 78–81 (1999).
   An important haplotype study that analysed a 10-kb non-coding region on the human X chromosome.
10. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human Y chromosome. *Science* **286**, 964–967 (1999).
11. May, C. A., Shone, A. C., Kalaydjieva, L., Sajantila, A. & Jeffreys A. J. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nature Genet.* **31**, 272–275 (2002).
12. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome: a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
13. Jobling, M. A. & Tyler-Smith, C. The human Y chromosome: an evolutionary marker comes of age. *Nature Rev. Genet.* **4**, 598–612 (2003).
   A recent comprehensive review of the Y chromosome.
14. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
15. Venter, J. C. *et al*. The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).
16. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
17. Bohossian, H. B., Skaletsky, H. & Page, D. C. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* **406**, 622–625 (2000).
18. Li, W. H., Yi, S. & Makova, K. Male-driven evolution. *Curr. Opin. Genet. Dev.* **12**, 650–656 (2002).
   A report that summarizes many of the issues involved in estimating male and female mutation rates.
19. Makova, K. D. & Li, W. H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624–626 (2002).
20. Ebersberger, I., Metzler, D., Schwarz, C. & Pääbo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.,* **70**, 1490–1497 (2002).
   A comprehensive study of genetic divergence between humans and chimpanzees, using autosomes and sex chromosomes.
21. Kong, A. *et al*. A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
22. Lin, S., Cutler, D. J., Zwick, M. E. & Chakravarti, A. Haplotype inference in random population samples. *Am. J. Hum. Genet.* **71**, 1129–1137 (2002).

23. Stephens, J. C. *et al*. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
24. Anagnostopoulos, T., Green, P. M., Rowley, G., Lewis, C. M. & Giannelli, F. DNA variation in a 5-Mb region of the X chromosome and estimates of sex-specific/type-specific mutation rates. *Am. J. Hum. Genet.* **64**, 508–517 (1999).
A summary-statistic study that focuses on differing mutation rates between the sexes.
25. Huang, W. *et al*. Sex differences in mutation rate in higher primates estimated from AMG intron sequences. *J. Mol. Evol.* **44**, 463–465 (1997).
26. Lawson, L. J. & Hewitt G. M. Comparison of substitution rates in ZFX and ZFY introns of sheep and goat related species supports the hypothesis of male-biased mutation rates. *J. Mol. Evol.* **54**, 54–61 (2002).
27. Chang, B. H. *et al*. Weak male-driven molecular evolution in rodents. *Proc. Natl Acad. Sci. USA* **18**, 827–831 (1994).
28. Chang, B. H. & Li, W. H. Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked Ube 1 genes and pseudogenes. *J. Mol. Evol.* **40**, 70–77 (1995).
29. McVean, G. T. & Hurst L. D. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386**, 388–392 (1997).
30. Hubert, R., MacDonald, M., Gusella, J. & Arnheim, N. High resolution localization of recombination hot spots using sperm typing. *Nature Genet.* **7**, 420–424 (1994).
31. Jeffreys, A. J., Ritchie, A. & Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum. Mol. Genet.* **9**, 725–733 (2000).
32. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
33. Schneider, J. A., Peto, T. E., Boone, R. A., Boyce, A. J. & Clegg, J. B. Direct measurement of the male recombination fraction in the human β-globin hot spot. *Hum. Mol. Genet.* **11**, 207–215 (2002).
34. Ruwende, C. & Hill, A. Glucose-6-phosphate dehydrogenase deficiency and malaria. *J. Mol. Med.* **76**, 581–588 (1998).
35. Sabeti, P. C. *et al*. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
36. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
37. Nordborg, M., Charlesworth, B. & Charlesworth, D. The effect of recombination on background selection. *Genet. Res.* **67**, 159–174 (1996).
38. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends. Genet.* **18**, 337–340 (2002).
39. Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S. & Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**, 1527–1535 (2003).
40. Payseur, B. A., Cutter, A. D. & Nachman, M. W. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* **19**, 1143–1153 (2002).
A broad search for evidence of selection in the human genome that includes differences in selection between the X chromosome and the autosomes.
41. Nachman, M. W. Single nucleotide polymorphisms and recombination rate in humans.*Trends Genet.* **17**, 481–485 (2001).
42. Sachidanandam, R. *et al*. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
43. Begun, D. J. & Whitley, P. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. *Proc. Natl Acad. Sci. USA* **97**, 5960–5965 (2000).
44. Yu, N. *et al*. DNA Polymorphism in a worldwide sample of human X chromosomes. *Mol. Biol. Evol.* **19**, 2131–2141 (2002).
A large haplotype-based study of a 10-kb region on the X chromosome that indicates a multiregional model of human origins.
45. Seielstad, M. T., Minch, E. & Cavalli-Sforza, L. L. Genetic evidence for a higher female migration rate in humans. *Nature Genet.* **20**, 278–280 (1998).
46. Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T. & Stoneking, M. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genet.* **29**, 20–21 (2001).
47. Laporte, V. & Charlesworth, B. Effective population size and population subdivision in demographically structured populations. *Genetics* **162**, 501–519 (2002).
48. Oota, H. *et al*. Extreme mtDNA homogeneity in continental Asian populations. *Am. J. Phys. Anthropol.* **118**, 146–153 (2002).
49. Pennisi, E. Tracking the sexes by their genes. *Science* **291**, 1733–1734 (2001).
50. Harding, R. M. *et al*. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**, 772–789 (1997).
51. Rana, B. K. *et al*. High polymorphism at the human melanocortin 1 receptor locus. *Genetics* **151**, 1547–1557 (1999).
52. Jin, L. *et al*. Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. *Proc. Natl Acad. Sci. USA* **96**, 3796–3800 (1999).
53. Rogers, E. J. *et al*. Integrated analysis of sequence evolution and population history using hypervariable compound haplotypes. *Hum. Mol. Genet.* **9**, 2675–2681 (2000).
54. Yu, N. & Li, W. H. No fixed nucleotide difference between Africans and Non Africans at the pyruvate dehydrogenase e1 α-subunit locus *Genetics* **155**, 1481–1483 (2000).
55. Gabriel, S. B. *et al*. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
56. Yu, N. *et al*. Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* **161**, 269–274 (2002).
A summary of a large amount of data on genetic variation within and between human populations.
57. Miller, R. D., Taillon-Miller, P. & Kwok, P. Y. Regions of low single-nucleotide polymorphism incidence in human and orangutan Xq: deserts and recent coalescences. *Genomics* **71**, 78–88 (2001).
58. Takahata, N., Lee, S. H. & Satta, Y. Testing multiregionality of modern human origins. *Mol. Biol. Evol.* **18**, 172–183 (2001).
59. Anderson, S. *et al*. Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
60. Heyer, E. *et al*. Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am. J. Hum. Genet.* **69**, 1113–1126 (2001).
61. Nachman, M. W. & Crowell, S., L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
62. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).

Acknowledgements

I wish to thank D. Reich, M. Daly and two anonymous reviewers for helpful comments on the manuscript, and G. Thorisson for providing remapped TSC data.

Competing interests statement

The authors declare that they have no competing financial interests.

Online links

FURTHER INFORMATION
International HapMap Project: http://www.hapmap.org
MIT Center for Genome Research:
http://www.genome.wi.mit.edu
Medical and Population Genetics Group:
http://www.genome.wi.mit.edu/mpg
Access to this interactive links box is free online.