The historical record tells us stories of migrations, population expansions and colonization events in the last few thousand years, but what was their demographic impact? Genetics can throw light on this issue, and has mostly done so through the maternally inherited mitochondrial DNA (mtDNA) and the male-specific Y chromosome. However, there are a number of problems, including marker ascertainment bias, possible influences of natural selection, and the obscuring layers of the palimpsest of historical and prehistorical events. Y-chromosomal lineages are particularly affected by genetic drift, which can be accentuated by recent social selection. A diversity of approaches to expansions in Europe is yielding insights into the histories of Phoenicians, Roma, Anglo-Saxons and Vikings, and new methods for producing and analysing genome-wide data hold much promise. The field would benefit from more consensus on appropriate methods, and better communication between geneticists and experts in other disciplines, such as history, archaeology and linguistics.

Keywords: human; population history; genetic drift; natural selection; genetic diversity

1. INTRODUCTION

In this review, I ask how genetics can help us to understand the relatively recent past, and in particular, the impact of historical events on human genetic diversity. The word 'history' refers specifically to written records of events, and the time depth of this varies considerably in different regions of the world, but for the purposes of this discussion I will focus on the events of the last 2–3 millennia. I ask what genetic tools are available that can help us understand this period, and what problems there are in interpreting genetic evidence, give some examples of what has been found, and finally discuss the importance of interdisciplinary efforts, and how this kind of collaborative work can be difficult to accomplish.

2. DNA: A MESSAGE FROM OUR ANCESTORS

As DNA passes down to us through the generations, it changes through mutation, so that it varies among individuals. It is important to realize that this variation is slight, and that the DNA sequences of all modern humans that can be aligned and compared are essentially 99.9 per cent the same [1]. The pattern of DNA variation among individuals can tell us about the demographic history of populations—migrations, expansions and colonizations [2]. For population geneticists interested in such things, natural selection, the selective survival of genotypes due to their influence on fitness, is regarded as a nuisance that obscures the demographic history. By contrast, for many of the other authors of this issue, natural selection is the absorbing issue—for example, selection for resistance to infectious disease—and demographic history the nuisance. Reconciling these two views can sometimes be difficult.

There are two ways to read DNA’s message. One is the relatively direct means of studying ancient samples, such as bones, themselves. This has advantages—in a sense it provides ‘real’ information about the past—but also problems: sample sizes are small; there are difficulties with modern DNA contamination; even skeletons had ancestors, so it is still necessary to consider the ancient specimens’ own demographic histories; and it is possible that the bearers of these ancient DNA sequences have no living descendants, and therefore limited relevance to today’s genetic diversity. Novel sequencing technologies [3] are being applied to ancient DNA [4], and this will change the way we think about the past. For example, they have yielded draft genome sequences for Neanderthal [5] and other archaic hominin [6] specimens; a Neanderthal sequence track is now available in the UCSC (University of California at Santa Cruz) Genome Browser (genome.ucsc.edu/cgi-bin/hgGateway), a fact to which researchers have become rapidly accustomed, but which would have seemed a fantasy a few years ago. These ancient sequences are illuminating the way in which admixture between archaic and modern humans might have shaped our genomes, and our immune systems [7].

The second method to address the past, which I will focus on here, is the indirect means of analysing the genomes of modern populations. This has the advantage that it is relatively straightforward to obtain samples, but the disadvantage that statistical methods of inference are required to determine what processes gave rise to the patterns of diversity that we see in these samples.

A major difficulty in inferring the recent past from the present patterns of genetic diversity is that recent events
are laid on top of a series of more ancient events: the African origin of modern humans about 150 Ka, giving a high level of genetic diversity in Africa [8]; and a set of serial founder events with the migration out of Africa and the peopling of the Old and New Worlds, leading to a cline of decreasing genetic diversity with distance from East Africa [9,10]. These ancient events still have a very strong echo in modern diversity. Then there are later events that were also important: the development of agriculture, starting about 10 Ka in the Near East, precipitated a massive increase in human population size [11] that continues today. This revolution also had important consequences for infectious disease, since the large populations allowed epidemic diseases to persist, and animal domestication and storage of food facilitated the zoonotic transfer of diseases from animals (whether domesticates or rodents) to humans [12].

The analogy is often made to a palimpsest—a piece of parchment in which text was overwritten in different directions. We look at the past through a series of different intervening events, and when we try to learn about a particular period in the past this causes difficulties that are sometimes intractable. An additional problem is that geneticists who observe a pattern in their data and seek an explanation for it tend to visit a library, take out a history book and read about a past event that seems to explain the pattern they see. This kind of historical cherry-picking leads to a lack of objectivity in asking what kinds of past events could have given rise to modern genetic diversity.

3. THE RAW MATERIAL: GENOME VARIATION

Clearly, the ideal, unbiased and comprehensive way to assay genome variation within a population is to sequence the entire genomes of all individuals. Until recently, this was completely impractical because of the cost and time involved, but the revolutionary development of next-generation sequencing technologies [3] now permits the sequencing of whole human genomes [1] for a relatively low price, which continues to decrease. So far, the focus of the large genome sequencing consortia has been on medical genetic applications, but within the next few years it seems inevitable that population genome sequencing will become commonplace. This promises to open a new vista in our understanding of population history, but also poses big methodological and computational challenges in analysing and interpreting data on such a large scale [13]. Currently, published population studies interrogate specific sites of variation within genomes, and it is these approaches that will be addressed in this review.

Ninety-eight per cent of human DNA is contained within our set of 22 pairs of autosomes, and the X chromosome, present in one copy in males and two in females. Much focus has been on the remaining 2 per cent [14], which comprises two peculiar segments of DNA that are inherited from one parent only (uniparentally): one is the mitochondrial DNA (mtDNA), a circular molecule present in both males and females, but passed on only from mothers to their children; the other is the Y chromosome, which is passed down from fathers to sons because it is sex-determining.

Studying the biparentally inherited majority of our genome (the autosomes and X chromosome) has some advantages. It provides a picture of both of our parents’ ancestries—there is no particular sex bias. The development of genome-wide single nucleotide polymorphism (SNP) chips [15] allows hundreds of thousands of variable sites to be surveyed readily. Such data are also in a sense unbiased as many independently inherited loci are being surveyed simultaneously, and, although natural selection may be acting on individual loci, it is unlikely to be acting on all of them in the same way. So the overall picture obtained can be regarded as selectively neutral.

Autosomal diversity is geographically structured: application of clustering algorithms, such as STRUCTURE [16], to high-density genotype data [17,18] automatically partitions global sample sets into continental clusters that reflect their true origins with no prior geographical information. Within Europe, analysis of multilocus genotypes finds a close correspondence between genetic and geographical distances [19,20]; a two-dimensional summary of genetic variation through principal components analysis effectively recapitulates a geographical map of Europe [19]. One problem is that these geographical structures may emerge, but not with any idea of when they were established—a temporal aspect to the pattern is lacking. This is changing with the development of methods that consider not only variable sites in the genome, but also the way in which such sites are associated with each other (linkage disequilibrium; LD), and how this association breaks down with time through recombination. The pattern and signal of decay of LD may allow the timing of past admixture events to be estimated [21,22]. Other methods have been developed that do not rely on LD [23], instead identifying recombination breakpoints occurring since admixture. Research into methods to interpret genome-wide data in terms of population history is very active [13], and seems set to yield many new insights in the near future.

mtDNA provides a matrilineal picture of diversity, while the Y chromosome provides a patrilineal picture. These pieces of DNA escape from the process of reshuffling, or recombination, that occurs in the rest of the genome whenever sperm and eggs are produced. This gives them a relatively simple history, based on the accumulation of mutation events without their reassortment via recombination. mtDNA and the Y chromosome also give us pictures of sex-specific processes in the past—males and females do not behave the same way today, and probably never have, and these differences can be illuminated by comparing patterns of DNA diversity on these two molecules [24]. Because of the absence of recombination, it is simple to build phylogenetic trees using sequence variants [14]. It is possible to superimpose such trees upon maps, and thereby make inferences about migration history; this is the field of phylogeography, which because of its ad hoc nature is regarded by some as unreliable, though it has yielded some useful insights. It is also possible to estimate the ages of particular lineages if a ‘molecular clock’ (the mutation rate of DNA markers)
is available, but such estimates are often controversial, particularly for the Y-chromosomal analyses. There is disagreement about how best dating should be done, and about the appropriate mutation rates to use [25], which is a serious current problem. The disadvantage of using uniparentally inherited systems is that they provide only two biased snapshots of evolutionary history, and are particularly susceptible to the influences of genetic drift and natural selection.

The most commonly used variable DNA markers are SNPs. These are present throughout the genome, and have a low mutation rate—typically approximately $10^{-8}$ per nucleotide per generation [26], although in mtDNA there are some nucleotides that mutate much more rapidly [27]. SNPs are usually binary (one base or another) and their low mutation rate means that the chance of finding an SNP that has mutated independently in two genomes is small, so they tend to show identity by descent. For these reasons, they are relatively simple to understand and to incorporate into evolutionary models. A major problem until recently in using SNPs to examine human diversity has been ascertainment bias: SNPs were ascertained in one population, and then typed in others, which overestimates the diversity in the discovery population. Technological advances now allow access to very large amounts of sequence, which is eliminating this bias. For example, many studies sequence the entire 16.5 kb mtDNA [28], rather than just a small hypervariable segment, and next-generation sequencing technologies now allow the sequencing of large segments of the Y chromosome [29] (1400 times larger than mtDNA), and of whole human genomes [1].

To give an example of the impact of these methods, the current standard Y-chromosome phylogeny [30] contains about 600 SNPs ascertained over a long period of time using many different approaches. These SNPs define 311 lineages (called haplogroups), each of which is carried by many males. More recently, the 1000 Genomes Project [1] produced a preliminary phylogeny constructed from low-coverage Y data generated from whole-genome sequencing. Two striking features emerge: all of the 77 Y chromosomes included in this tree are different, so the 2870 SNPs identified from only four populations provide great discriminatory power; and the branch lengths now gain significance because of the unbiased SNP ascertainment—they become proportional to time, and promise much better dating methods for the future.

Natural selection is a potential problem, particularly for the non-recombining mtDNA and Y chromosome, and there are many tales of possible selective effects. One recent example is the finding that mtDNA haplogroup H, common today throughout Europe [31], is strongly underrepresented among British skeletal samples that are greater than 1000 years old [32], suggesting that it may have undergone a major increase since that time. Together with the observation that sepsis patients carrying haplogroup H are more likely to survive their illness [33], this leads to the speculation that a major disease episode in the last millennium might have acted to select individuals carrying haplogroup H. As with many such speculations, the usual suspect is the plague. Similar examples for the Y chromosome are not so clear, but the chromosome is not without variable phenotypic effects—for example, carriers of the European-specific haplogroup I show a significantly faster progression to HIV-AIDS following infection [34].

4. GENETIC DRIFT AND SOCIAL SELECTION

Genetic drift, arising through stochastic variation in the number of offspring, is a particularly important factor for uniparentally inherited segments of DNA. The effect is stronger for these parts of the genome than for autosomal segments, because, population-wide, they are fewer in number—for each individual Y chromosome or mtDNA type that can be passed to the next generation, there are four copies of each autosome in the population, so the opportunity for change through stochastic variation differs correspondingly. But because of the sex-specific pattern of inheritance of the Y, in particular, drift can be greatly accentuated by social selection.

Studies of this phenomenon require a more fine-grained distinction of Y-chromosome types than can be achieved with the widely typed SNPs. Sets of short tandem repeat (STR) markers are more suitable, as their rapid mutation rates can lead to high haplotype diversity even among Y chromosomes that share a recent common ancestor [35]. Furthermore, with an estimate of the mutation rate of STRs available, it is possible to estimate how much time has elapsed since this common ancestor lived. However, this is an area of controversy because of a lack of agreement over methodology, and even over the appropriate mutation rate to use. ‘Pedigree’ rates measured directly by comparing the haplotypes of sons with those of their fathers (typically, an average rate of about 0.2% per STR per generation [35,36]) are claimed by some to be unrealistically high for application to long time-spans, since the back and forth mutational process of the STRs leads to ‘saturation’ of mutation. This has led to the introduction of a so-called ‘evolutionary rate’ [25], calibrated on the basis of migration events of ‘known’ time depth, and is threefold lower than the pedigree rate, leading to a corresponding difference in the estimated dates. Such a wide range of possible dates (even neglecting other sources of error) causes considerable dismay among archaeologists and historians.

Despite these problems, the use of STRs has provided useful insights, particularly about the influence of social selection and male-specific behaviour on haplotype diversity. For example, a study in a large set of Asian males [37] revealed a remarkably common 15-locus STR haplotype, which, together with close neighbours arising from it by STR mutation, made up 8 per cent of the sample, and perhaps as much as 0.5 per cent of Y chromosomes globally. The most remarkable feature of this ‘star cluster’ was its geographical distribution—extremely widespread, and present in many different populations speaking many different languages. The published estimate for the age (though see the caveats above) is about 1000 years, and this, together with its distribution and its likely origin based on greatest diversity in Mongolia, led to the suggestion that it descended from Genghis
Khan. This founder of a patrilineal dynasty had many wives (polygyny), and therefore many sons, and importantly these sons in turn went on to have many sons themselves. This could lead to a rapid amplification of a specific haplotype.

Additional studies have revealed two further examples of ‘star clusters’ that have been ascribed to patrilineal dynasties: another in Asia, the so-called Manchu cluster [38], dating back about 500 years; and one in Ireland [39], claimed to descend from the mediaeval Ui Neill dynasty 1000–1700 years ago. These male-driven expansions can have a major effect on Y-chromosome diversity; whether more examples exist, and the extent to which they affect the diversity of the rest of the genome, are questions that await investigation.

5. LONG-RANGE MIGRATIONS AND LINEAGES OUT OF PLACE

Y-chromosomal diversity of indigenous populations on a global scale is very geographically structured [40], so each continent or major region has its characteristic set of haplogroups. Originally, this differentiation was ascribed to patrilocality [41], the mating practice adopted by about three-quarters of indigenous populations in which men, rather than women, tend to remain close to their birthplaces upon marriage [42]. However, though patrilocality is still thought to be important at a more local scale, the global patterns are now thought to arise from migration, population expansion and genetic drift [43]. mtDNA diversity, too, is strongly geographically differentiated at a global scale [14].

Recent long-range population movement within a framework of long-established patterns can often be recognized readily. A very clear case is the impact of the events of the last 500 years on South America, where different contributions from indigenous native American populations, colonizing Europeans, and Africans brought via the transatlantic slave trade mixed to form current populations (another important European contribution was the collection of microbes which the indigenous peoples had never encountered before, leading to massive population decline). As an example [44], mtDNA admixture proportions for a Colombian population were 90 per cent Native American, 4 per cent European and 6 per cent African; however, in stark contrast the Y-chromosomal proportions were dominated by incoming European lineages (79%), with Native American and African lineages making up only 12 per cent and 9 per cent, respectively. This is an example of sex-biased admixture, in which, because of the sexual politics of colonization and slavery, liaisons between European men and non-European women were particularly over-represented.

The events in South America gave rise to lineages distributed out of their normal ranges, which can be readily recognized using uniparentally inherited markers. The same can be seen in populations which, despite a widespread geographical distribution, remain isolated from the surrounding populations, largely through the social practice of endogamy.

These have been referred to as ‘transnational isolates’, and a good example is that of the European Roma [45]. The genetic coherence in this group results from recent migration from one place of origin, and linguistic and genetic evidence point to northern India, about 1000 years ago [46]. The Roma carry a typically Indian Y-chromosomal lineage, haplogroup H1a, ranging from 17 per cent in Portugal [47] to 45 per cent in Bulgaria [48], together with others indicating variable degrees of admixture with host populations. Maternal lineages, too, reveal the presence of founding lineages from northern India [49].

6. DISENTANGLING LOCAL MIGRATION

When admixture involves geographically well-separated parental populations, with lineages well outside their natural places, it is relatively easy to recognize and understand; however, trying to use genetic diversity to infer the past in landscapes where this long-range migration has not occurred is much more difficult. The British Isles provides a good example of this. For example, there are Y-chromosome data with good geographical coverage [50], but these do not show anything obviously interesting: the same lineages are present in most places, at slightly different frequencies.

The story of the British Isles is often told as a series of invasions—the Romans, the Anglo-Saxons, the Vikings and the Normans [51]. The extent to which the corresponding cultural transitions were actually associated with mass migration, rather than with small numbers of culturally influential individuals, is a matter of considerable debate. Genetic evidence on this subject is clouded by the fact that any lineages imported by Anglo-Saxons, Danish Vikings and Normans (themselves descendants of Vikings) derive from the same gene pool sampled at different times, and may be difficult or impossible to disentangle when admixture analysis is carried out.

The influence of Norwegian Vikings, who settled the western side of the British Isles, is less difficult to recognize and interpret because there were no recurrent migrations, and because modern Norwegians (used as a proxy in admixture studies for Vikings) are more distinct from modern British samples than are modern Danish samples [50,52]. One study [53], examining both Y and mtDNA lineages, produced the interesting finding that close to Norway (in Shetland, Orkney and the northwest Scottish coast) proportions of Norse admixture were similar for both maternal and paternal lineages, suggesting family-based settlement. Further afield, however, in the Western Scottish Isles and in Iceland, the proportion for paternal lineages was greater, suggesting that settlement here may have been male-biased.

Sampling within Britain, an urbanized island in which the industrial revolution caused much population movement, risks missing the signals of 1000 year-old Viking migrations. To address this, the traditional sampling procedure [50] is to seek individuals who can at least vouch for the place of birth of their grandparents—and, for Y studies, their paternal grandfather in particular. Sampling on the basis of surnames provides a means to assemble proxies for older populations [54]. This is because patrilineal surnames
and Y chromosomes are passed down in the same way, and there is a strong relationship between Y type and surname, particularly for the rarer names [55,56]. Armed with lists of surnames that were present in a particular place in medieval times, we can sample modern males who have known ancestry in these places, but also the relevant surnames. Comparisons of such samples with samples based only on grandparental place of birth [57] show a significant difference in Y haplogroup composition, which can be explained by increased Scandinavian admixture proportions in the surname-ascertained samples. In effect, these samples seem to be more like populations from before the industrial revolution, and depleted in lineages that arrived subsequently through migration from other parts of Britain.

The regions focused upon in the study described above [57] were West Lancashire and the Wirral peninsula in northwest England, where the high density of Norse Viking place-names, as well as archaeological and written evidence [58], suggests a strong Norse influence. The approach is now being extended to other parts of northern Britain. Well-ascertained samples across the British Isles promise to add genome-wide data to the picture [59].

A different approach to the palimpsest problem of population movement in different eras has been taken to the study of the expansion of the Phoenicians, settlers and traders who expanded across the Mediterranean region in the first millennium BC. To distinguish signatures of Phoenician expansion from those of earlier expansions of Neolithic people and Greeks, as well as the Jewish diaspora, this study [60] sought pairs of sites that were matched for their distance from the source of Phoenician expansion, but differed in that one of each pair showed good evidence for Phoenician contact, and the other one no evidence for such contact. This allowed a weak but systematic signal of a specific Y-chromosome type to be discerned, and suggested a greater than 6 per cent overall contribution of Phoenician lineages to the sampled populations. Such methods could be applied to other expansions.

The kinds of approaches described above provide plausible and interesting insights into the past, but can be criticized because of their ad hoc nature, and because they fail to address other possible explanations for their observations. An alternative approach is to ask if hypotheses about demographic history and social structure are compatible with the observed genetic data by using computer simulations, with parameters informed by historical information. This allows systematic testing of assumptions, and can also address the issue of ‘knowability’—how much we can know about a particular era in the past from modern genetic diversity, given the complexity of intervening events. Simulation approaches have been taken to the question of Anglo-Saxon influence in Britain, supporting initially a mass migration model from Y-chromosome data [63], and subsequently elaborating this into an ‘apartheid’ model in which differential reproductive success of individuals with indigenous or Anglo-Saxon ancestry played a major role [64].

7. ESCAPING THE BIOSCIENCE GHETTO

Until recently, studies in human population genetics were driven purely by academics supported by funding agencies. This is changing through a combination of cheaper technologies for genotyping, easy mass communication via the Internet, and the widespread general interest in genealogy, ancestry and history [65]. With ‘recreational genomics’ [66] the general public has entered the arena, and is now contributing not only DNA samples, but also data and curiosity to the mix. Part of the activity of the Genographic Consortium, responsible for the Phoenician study mentioned above, is driven by public interest and contribution [67]. There are large databases containing genotypes and haplotypes of the customers of genealogical testing companies, and some academics have taken advantage of this to design studies [68]. Some companies now offer targeted or general genome sequencing, and there is online discussion about the sharing and mining of individuals’ genome-wide data (www.genomesunzipped.org).

For the academic geneticist, more challenging than collaborations with the general public are collaborations with other academics from different disciplines. And yet, if the contribution of genetics to understanding the past is to be truly useful, then this interdisciplinary collaboration with historians, linguists, archaeologists and demographers is essential. To avoid the problem of ‘cherry-picking’, the collaboration needs to be meaningful, with a genuine dialogue and attempt to understand each other’s disciplines.

There is, however, a cultural divide between arts and humanities subjects and bioscience that makes this task difficult. The scientist inhabits the world of the journal article, the impact factor and the h-index, while the historian values the book, the monograph and the festschrift. For the geneticist, the methods of archaeology and history are often unclear—there are few testable hypotheses, and no $p$-values. Journals tend to be subject-specific, with peer-review often similarly subject-specific, so some aspects of an interdisciplinary study can remain unexamined by an independent expert. The exigencies of research quality assessment (e.g. the UK’s Research Excellence Framework (www.hefce.ac.uk/research/ref)) actually militate against interdisciplinarity because outputs are assessed within subject-specific units. More fundamentally, perhaps, there can be an attitude of distrust: one eminent historian colleague once opined to me that ‘Geneticists tell us either something we knew already, or something that can’t possibly be true’.

In an attempt to address the history of the British Isles over the last three millennia, we have joined with local colleagues in history, archaeology, linguistics, place-names studies and social psychology in an interdisciplinary programme of research, ‘The impact of diasporas in the making of Britain’ (www2.le.ac.uk/projects/impact-of-diasporas). Funding comes from the Leverhulme Trust, a body that is unusual in seeking interdisciplinary and not normally supported by other funders. The programme will employ a group of post-docs from different disciplines working in close proximity. However, not everyone is pleased. An inevitable side-effect of re-examining the evidence for our pictures of the past is that some
people get upset. When our project was announced through a press release, comments appeared on the website of the British National Party, a far-right British political organization that opposes immigration and favours ‘voluntary repatriation’. One contributor described our project as: ‘Another government-funded (sic) justification for immigration which will say there is no such thing as indigenous British’, and our research group as ‘a bunch of Marxist parasites who have never done a proper day’s work’. Well, proper or not, there is much work to do, and we Marxist parasites have high hopes that it will produce new insights.

My work is supported by a Wellcome Trust Senior Fellowship in Basic Biomedical Science (grant no. 087576).

REFERENCES

31 Richards, M. et al. 2000 Tracing European founder lineages on April 30, 2017http://rstb.royalsocietypublishing.org/Downloaded from
Review. Genetic impact of recent events  M. A. Jobling 799