Human Population Structure and Demographic History using Genetic Markers

James Flett Wilson

D.Phil.
Board of the Faculty of Biological Sciences
University of Oxford

2002

To my parents.

‘light will be thrown on the origin of man and his history’

Charles Darwin, *The Origin of Species*, 1859

Human Population Structure and Demographic History using Genetic Markers

James Flett Wilson, Department of Zoology, D.Phil., Hilary 2002

The evolutionary history of the human species has generated complex patterns of population structure and linkage disequilibrium (non-random associations of alleles at different loci or LD). The understanding of these patterns is crucial to two of the most important challenges facing biomedical science today: the identification of disease predisposing genes and prediction of variable drug reactions. The genetic variation revealed by these endeavours can also illuminate the underlying population historical processes. Here, I illustrate each of these applications: first, by assessing the demographic context of cultural change in the British Isles. Y chromosome variation indicates that the Viking age invasions left a significant paternal legacy (at least in Orkney), while the Neolithic and Iron Age cultural transitions did not. In contrast, mitochondrial DNA and X chromosome variation indicate that one or more of these pre-Anglo-Saxon revolutions had a major effect on the maternal genetic heritage of the British Isles. Second, I provide conclusive evidence that diverse demographic histories produce strikingly different patterns of association. Elevated LD extends an order of magnitude further in the Lemba, a Bantu-Semitic hybrid population, than in the putative parental populations. A significant relationship between allele-frequency differentials in the parental populations and the Lemba LD demonstrates that it is admixture-generated. Third, I demonstrate that the genetic structure inferred in a heterogeneous sample using neutral markers (a) shows ethnic labels to be inaccurate descriptions of human population structure, and (b) predicts drug metabolising profiles, defined by the distribution of drug metabolising enzyme variants. Thus the trade-off between therapeutic response and adverse drug reactions will differ between different sub-clusters. Assessment of genetic structure during drug trials is therefore, like the empirical evaluation of each population's pattern of LD, a necessity.

Contents

This thesis is made up of the following three papers, as permitted by the rules, to which I have added an introduction and discussion.

Ch. 2 is based on:

Wilson JF *et al*. (2001) *Proc. Natl Acad. Sci. USA* **98**: 5078-5083.

Ch. 3 is based on:

Wilson JF & Goldstein DB (2000) *Am. J. Hum. Genet*. **67**: 926-935.

Ch 4 is based on:

Wilson JF *et al*. (2001) *Nat. Genet*. **29**: 265-269.

Chapter 1

Introduction

Human genetic variation is central to the study of genetic history, medical genetics and pharmacogenetics. Understanding the structure of this variation has occupied population geneticists interested in the evolutionary history of our species since the discovery of the first cross-population differences in polymorphism frequencies – that in the ABO blood groups – almost a century ago (Hirszfeld and Hirszfeld 1919). With the completion of the draft human genome sequence last year (Lander et al. 2001; Venter et al. 2001), more than a million polymorphisms are now available (Sachidanandam et al. 2001). The unparalleled opportunities thereby opened up for the study of human population structure and demographic history are only beginning to be exploited.

The use of such genetic markers to answer questions in human evolution has both academic and applied motivation because the demographic history of our species is in part responsible for the present day population structuring of genetic variation - including the variation responsible for disease susceptibility and differential response to drugs. Furthermore, the demographic history of a population determines its pattern of association or linkage disequilibrium (LD), that is the non-random co-occurrence or association of alleles at different loci in a gamete. The mapping of genes predisposing to complex diseases such as diabetes and heart disease – the major challenge of contemporary medical genetics – requires knowledge of the pattern of these associations in the genome. For instance, association between a condition and a marker allele implicates nearby genes as candidates for harbouring susceptibility alleles, however the extent of LD in this region will determine the distance which should be screened for such variants. Thus, efficient mapping will also require an understanding of how our evolutionary history has shaped the present day distribution of variation.

Population genetic data have allowed diverse hypotheses regarding human history to be tested: from the origin of our species to the origin of Pacific island populations less than one thousand years ago. Biomedically, the most important such hypothesis is that anatomically modern humans did not evolve across the old world from *Homo erectus*, but instead originated recently in Africa and spread from there to the rest of the world: the out of Africa theory (Cann et al. 1987). Many lines of evidence support the theory, such as the higher diversity observed in African populations at many genetic loci and the fact that the deepest split in population trees is between Africans and the rest of the world (Cavalli-Sforza et al. 1994). However, a larger effective population size in Africa could also explain the increased diversity as it depends on both the size and age of a population. As branch lengths in population trees are shortened by admixture, increased gene flow among populations outside of Africa could explain the increased African divergence (Templeton 1997). Populations outside of Africa also appear to have more extensive LD than those in Africa and so perhaps share a more recent LD-generating event, such as a bottleneck upon exit from the continent (Tishkoff et al. 1996).

The most persuasive evidence, however, is that provided by haplotypes, which allow both the genealogical history of a locus to be reconstructed and the ages of lineages to be estimated. For all genealogical systems in which placement of the root is possible, from mitochondrial DNA (mtDNA) (Cann et al. 1987) and Y chromosomes (Underhill et al. 2000) to multiple autosomal regions (Alonso and Armour 2001; Harding et al. 1997), it falls in Africa; convincingly pointing to our origin there. But it is the time to most recent common ancestor of the haploid systems, estimated to be around 50,000-150,000 years (Hammer et al. 1998; Horai et al. 1995; Thomson et al. 2000; Watson et al. 1997),

that most surely points to an origin much more recently than the first emergence of *H. erectus* from Africa ~1.7 million years ago (Gabunia and Vekua 1995; Swisher et al. 1994).

Such population genetic structuring of variation may have important functional consequences. The virtually fixed difference in allele frequency between Sub-Saharan Africa and the rest of the world at the pyruvate dehydrogenase E1 alpha subunit locus (Harris and Hey 1999; Yu and Li 2000) demonstrates the scope for substantial population subdivision. Functional variants can also show large allele frequency differentials between populations in Africa and those elsewhere. African and European populations, for example, differ by more than 30% in the frequency of an allele which results in the absence of cytochrome P450 monooxygenase (CYP) 3A5, an enzyme involved in the metabolism of around half of all oxidatively-metabolised drugs (Kuehl et al. 2001). Structuring of variation can also be due to differential selection – such as that exerted by malaria on glucose-6-phosphate dehydrogenase (Tishkoff et al. 2001) and on the Duffy blood group locus (Hamblin and Di Rienzo 2000).

Heterogeneity is also found at finer scales, reflecting the subsequent history of migration of peoples across the globe. For example, within Europe, many loci show a Southeast to Northwest cline in allele frequencies (Cavalli-Sforza et al. 1994; Menozzi et al. 1978), considered by many to be a consequence of a Neolithic wave of advance of genetically differentiated farmers into the indigenous hunter-gatherer populations (Ammerman and Cavalli-Sforza 1984; Barbujani et al. 1994). The distribution of duplicated alleles of the CYP2D6 locus, which result in an 'ultra-metaboliser' phenotype and thus to the therapeutic failure of more than 40 drugs (Weber 1997), shows a similar

pattern. Ultra-metaboliser alleles vary in frequency from 10% in Northern Spain (Bernal et al. 1999) and 6% in Turkey (Aynacioglu et al. 1999) to 1-2% in Germany, Denmark and Sweden (Bathum et al. 1998; Dahl et al. 1995; Sachse et al. 1997), highlighting the clinical relevance of population history even within Europe.

Population history has also become important to gene mapping in the last decade. Prior to this, Mendelian diseases were mapped successfully by assessing the co-inheritance of marker alleles and diseases in pedigrees: linkage analysis. With the failure of linkage analyses to map complex diseases, LD-based approaches, most commonly in a case-control design, are now considered the most promising method in attempts to do so (Risch and Merikangas 1996). Genetic variation between subjects presenting with the condition and control individuals from the same population is assayed in an attempt to identify variants influencing disease susceptibility. As it is not possible to catalogue every difference, this approach relies on LD between the unknown causal variant(s) and the genotyped markers. The much larger number of recombination events in the history of a population compared to that in even a large family should allow much finer localisation. The pattern of LD, however, depends on many factors as well as recombination – mutation; selection; population admixture, bottlenecking, growth and subdivision. If association-based approaches are to be successful in their aims, an improved understanding is therefore required of the distribution of LD in the human genome as well as of the forces that generated it.

The genetic distance over which LD is observed will determine the marker density required in genome scans for association, while the consistency at a given distance will determine how many markers need be placed in an interval. Empirical data

indicate that LD typically extends to between five and seventy-five kilobases (kb) in human populations (Kidd et al. 2000; Reich et al. 2001), although the variance in LD at a given distance has attracted less attention. A simulation study based on a simplified model of population growth predicted useful LD would only extend to 3 kb (Kruglyak 1999), contrasting sharply with these figures. However, data on human variation suggest much more complex demographic histories for most populations involving expansions, bottlenecks and admixture (Cavalli-Sforza et al. 1994; Collins et al. 1999; Reich et al. 2001).

The diverse demographic histories of human populations predict each will show different patterns of LD which will affect their utility for mapping. For example, LD extends over shorter distances in African populations in accordance with the greater time depth of human habitation there. In a study of 19 genomic regions, Reich *et al.* (2001) found the distance at which the average LD is half-maximal to be only ~5 kb in the Nigerian Yoruba, near the prediction of Kruglyak (1999). LD tends to extend over much longer distances in Eurasian populations – this 'half-length' of LD was ~60 kb in Europeans (Reich et al. 2001). Special populations that were founded relatively recently, e.g. Finns (Peterson et al. 1995) and Sardinian sub-isolates (Angius et al. 2001; Wright et al. 1999; Zavattari et al. 2000), appear to show LD over even longer distances, although data are contradictory in some cases (Eaves et al. 2000; Lonjou et al. 1999; Taillon-Miller et al. 2000). As it is not clear whether isolates will prove to be more suitable for mapping, empirical work describing patterns of LD systematically in both younger and older populations is urgently required. The recent discovery that, at least in parts of the genome, the fine-scale structure of LD is block-like – due to the presence of

recombination hotspots interspersed among low recombination regions (Daly et al. 2001) – exemplifies the need for more experimental investigation. The fact that there appears to be three classes of population with respect to the extent of LD suggests a three-tiered approach to mapping. Coarse mapping is first carried out in a young population with long range LD, followed by finer-scale mapping in an older, out-bred population and finally localisation is performed in populations such as Africans showing LD only over very short distances.

Understanding the patterns of population structure and allelic association is thus critical to the advance of biomedical science and will also cast light on our history and origins. Here, I illustrate each of these applications of data on human variation: an investigation into the history of the British Isles in chapter two, a description of the pattern of LD in an admixed population in chapter three and a demonstration of the importance of genetic structuring in variable drug reaction in chapter four. Much of the previous work in these areas has concentrated on single loci or on unlinked markers. In this thesis I have consistently used multiple genetic regions which, by incorporating independent runs of the evolutionary process, allow for much more robust interpretation than single-locus studies do. I have also used haplotype systems which enable reconstruction of the genealogical history of a locus. There are many advantages of such a genealogical approach, one being that gene genealogies provide the means to distinguish divergence due to ancient population fission from that due to restricted gene flow. The ease of interpretation of haplotypic data in the face of admixture is not approached using allelic data.

Chapter 2

Genetic evidence for different male and female roles during cultural transitions in the British Isles

Human history is punctuated by periods of rapid cultural change. Although archaeologists have developed a range of models to describe cultural transitions, in most real examples we do not know whether the processes involved the movement of people or of culture only. With a series of relatively well-defined cultural transitions, the British Isles present an ideal opportunity to assess the demographic context of cultural change. Important transitions following the first Palaeolithic settlements include the Neolithic, the development of Iron Age cultures, and various historical invasions from Continental Europe. Here we show that patterns of Y chromosome variation indicate that the Neolithic and Iron Age transitions in the British Isles occurred without large-scale male movements. The more recent invasions from Scandinavia, on the other hand, appear to have left a significant paternal genetic legacy. In contrast, patterns of mtDNA and X chromosome variation indicate that one or more of these pre-Anglo-Saxon cultural revolutions had a major effect on the maternal genetic heritage of the British Isles.

Introduction

Archaeologists once assumed that the British Isles were settled by successive waves of continental invaders, from Neolithic times onwards (Hawkes 1931). Today the pendulum has swung the other way, with archaeologists tending to postulate considerable cultural exchange, such as the establishment of trading networks, with little or no movement of people (Cunliffe 1997; Renfrew 1987). It is likely, however, that the extent of genetic continuity in the face of cultural change has varied from case to case.

We have utilised a number of genetic marker systems to determine the genetic legacy of cultural change. Analyses of the non-recombining part of the Y chromosome are becoming increasingly important in uncovering paternal heritage in human evolutionary studies due to the recent development of a highly informative combination of different genetic markers (Jobling and Tyler-Smith 1995; Karafet et al. 1999; Zerjal et al. 1997). Slowly evolving biallelic markers are used to define distinct genealogical groups (haplogroups), while rapidly evolving microsatellites are used to distinguish more closely related chromosomes within haplogroups. Together, the two sets of markers identify well-defined haplotypes which have proven powerful tools in identifying relationships among populations. Occurrences of particular haplotypes, for example, have been suggested as population specific signatures, as in the case of a high frequency haplotype that appears to mark Jewish populations (Thomas et al. 1998). Here we contrast the pattern observed on the Y chromosome with that observed using multiple genetic systems influenced by female migration (mtDNA and unlinked X chromosome

systems) in order to evaluate whether cultural changes in the British Isles have involved different demographic roles for males and females.

Identification of genetic changes associated with these transitions requires that the source populations be distinguished with respect to some genetic marker. There are numerous candidate source populations for the British Isles from the pre-Anglo-Saxon British, to the Romans, Anglo-Saxons, Scandinavians and Normans. For tractability, we have focussed mainly on two, the pre-Anglo-Saxon British and the Scandinavians. We have achieved this by concentrating on the Celtic-speaking populations and on Orkney, a Northern Scottish archipelago with Viking and pre-Anglo-Saxon British heritage.

Materials and Methods

Samples and Genotyping. Buccal swabs were scraped on the inside of the cheek by each subject and replaced in collection tubes to which 0.05M EDTA/0.5% SDS had been added. In all cases informed consent was obtained before samples were collected. Standard phenol/chloroform DNA extractions were then performed. Three Y chromosome multiplex PCR kits were used as described (Thomas et al. 1999). The products of each kit were subjected to electrophoresis on ABI 377 or ABI 310 automated sequencers and analysed by Genescan™ software. Conversions to repeat lengths were standardised using control individuals sequenced by P. de Knijff. To assess the reliability of our data, 876 Y chromosome microsatellite genotypes were retyped blindly, 6 were found to differ, an error rate of 0.7%. The Irish data do not include DYS388, so this locus was dropped from comparisons involving Ireland. The first hypervariable segment of the

mitochondrial control region was amplified and sequenced from nucleotide positions 16090 to 16365 (Richards et al. 1996). 34 X-linked microsatellites were genotyped using multiplex PCR kits (Ch 2).

Y chromosome haplogroups. The haplogroup (hg) and haplotype cluster designations are as follows with UEP genotypes in the order sY81, SRY4064, YAP, SRY10831, M13, M9, Tat, M20, SRY+465, 92R7, M17 and microsatellite genotypes in the order DYS388-DYS393-DYD392-DYS19-DYS390-DYS391: **hg 1** – AG-GGGTACT+G not including the 1.15+ cluster (in the Basque data the subclade of haplogroup 1 defined by a mutation at SRY-2627 was included in haplogroup 1 as we did not genotype this polymorphism); **haplotype cluster 1.15+** – haplogroup 1 chromosome with microsatellite genotype 12-13-13-14-24-11 and one-step mutational neighbours, note DYS388 was not typed in the Irish, but is almost monomorphic at 12 repeats in haplogroup 1, only ~3% of the haplogroup 1 chromosomes in this study have different alleles; **hg 2** - AG-GGCTACC+G not including the 2.47+ cluster; **haplotype cluster 2.47+** – haplogroup 2 chromosome with microsatellite alleles 14-13-11-14-22-10 and one-step network; **hg 3** – AG-AGGTACT-G not including the 3.65+ cluster; **haplotype cluster 3.65+** – haplogroup 3 chromosome with microsatellite alleles 12-13-11-16-25-11 and one-step network; **hg 7** – AG-ACCTACC+G; **hg 8** – GA+GGCTACC+G; **hg 9** – haplogroup 2 chromosome with microsatellite haplotypes found only in 12f2 deleted chromosomes: either DYS388*14, DYS393*12, DYS392*11 or 15-12-11, 15-13-11, 17-12-11 or 16-12-11 in the same order; **hg 16** – AG-GGGCACC+G; **hg 21** – AA+GGCTACC+G; **hg 26** – AG-GGGTACC+G; **hg 28** – AG-GGGTGCC+G. A tree

presenting the genealogical relationships of these haplogroups (except haplogroup 28, which branches from haplogroup 26) is presented in (Zerjal et al. 1999)

mtDNA haplogroups. Haplotypes were assigned to haplogroups according to the West Eurasian mtDNA genealogy (Macaulay et al. 1999). Haplogroup assignment proceeded using the following algorithm. All numbering is according to Anderson et al. (1981) less 16000 in the control region for brevity: 069T 126C 223C assigned to haplogroup **J** (note in all but four cases 069 information was available); 126C 223C 294T to **T**; 129A 223T 391A to **I** (391 information was available); 223T 292T to **W**; 189C 223T 278T to **X**; 223C 224C 311C to **K**; 223C 249C and either 189C or 327T to **U1**; 129C 223C to **U2** (051G, if information available); 223C 343G to **U3**; 223C 356C to **U4**; 223C 270T to **U5**; 172C 219G 223C to **U6**; 223C 318T to **U7**; 223C 298C to **V**; 067T 223C to **HV1** (067 information usually available); 126C 223C 362C to **pre-HV**; 145A 176G 223T to **N1b**; 223T 278T 390A to **L2**; 187T 189C 223T 278T 311C to **L1**. For sequences not matching any of those above the algorithm used was: if 223T, test for +10397 *Alu*I (where + indicates restriction site presence and – absence) for **M**; –10871 *Mnl*I and –10397 *Alu*I for **L1**, **L2** or **L3**; if 223C, test for –7025 *Alu*I for **H**; –14766 *Mse*I, +7025 *Alu*I, –4577 *Nla*III for **HV***; +12308 *Hin*fI for **U***; otherwise assign to **R***. HVS-1 sequences were also checked for matches with common East Asian haplogroup motifs. Recurrent mutations may cause ambiguities by eliminating part of a diagnostic motif or recreating it in another part of the tree. In many cases the presence of substitutions defining subclades within the major haplogroups allowed sequences to be assigned even when reversion had occurred at a haplogroup motif site. In the case of hybrid motifs, PCR-RFLP testing was used to assign haplogroup (Torroni et al. 1996). Particularly in

the Welsh and Irish data HVS1 sequences matching a unique haplotype in an RFLP-defined haplogroup were assigned to that haplogroup.

Analysis. Exact tests and AMOVAs were calculated using ARLEQUIN (Schneider et al. 1997). Principal components analyses were performed on haplogroup and allele frequencies using POPSTR (H. Harpending, *pers. comm.*). Population structure was assessed using the model-based clustering method implemented in STRUCTURE (Pritchard et al. 2000), using the admixture model with a burn in of 50,000 steps and a run length of 10^6 steps. All loci within 2 centimorgans of another locus were excluded from the STRUCTURE analysis, leaving 23 loci.

Results and Discussion

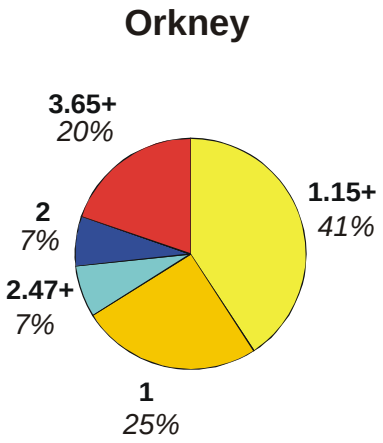
Genetic history of Orkney. When the Norsemen invaded (*c*. 800 AD), Orkney was populated by the Picts, little understood pre-Anglo-Saxon inhabitants. Orkney remained a Norse colony while an increasing number of Scottish settlers arrived in the islands, which were pledged to Scotland in 1468 (Thomson 1986). Archaeological interpretations of the Pict-Viking transition have often suggested continuity in both artefacts and lifestyle, more compatible with considerable integration between native Picts and incoming Norsemen (Ritchie 1993). However, as the place-names of Orkney are almost entirely Old Norse in origin (Lamb 1993) and a Nordic language replaced the earlier tongue, linguists and others argue that the Viking invaders completely replaced the native population (Barnes 1998; Smith 2001). To investigate whether Orkney's Viking heritage is genetic as well as cultural, we sampled 71 adult males claiming at least three

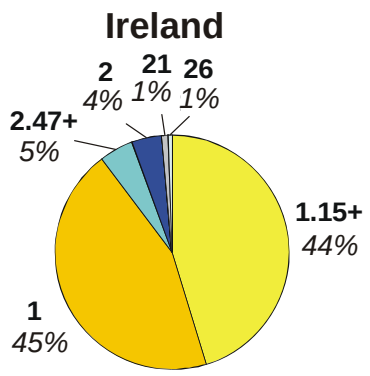
unrelated paternal generations in Orkney and all with surnames found on the islands before 1700 (Lamb 1981). For comparison, we used analogous criteria to sample 78, 88 and 94 individuals from Norway, Anglesey (North Wales) and West Friesland (Netherlands), respectively. Data on 146 Irish males with Irish Gaelic surnames were also included (Hill et al. 2000).

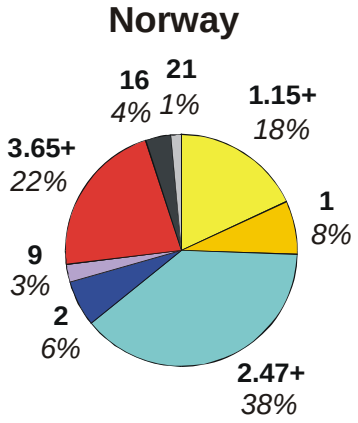
The Irish and Welsh are not significantly differentiated from each other at the haplogroup level ($P=0.16$, (Raymond and Rousset 1995)) and will hereafter be called 'Celtic'. However, Celtic, Frisian, Norwegian and Orcadian Y chromosomes are all highly differentiated at the haplogroup level ($P<0.0001$), (Fig. 1, Table 1). The Orkney sample appears intermediate between the Celtic samples and Norway, consistent with an origin by admixture between two such populations: haplogroup 1 decreases from the Celtic populations through Orkney, to Norway, while haplogroups 2 and 3 show the opposite trend. With respect to microsatellite variation within haplogroups, the Celtic samples are very homogeneous: the modal haplotype (microsatellite haplotype 15 within haplogroup 1, or haplotype 1.15) has a frequency of 26% in Wales and 18% in Ireland, and along with its one-mutational-step neighbours (Thomas et al. 1998) constitutes 70% of the Welsh and 44% of the Irish chromosomes (as well as 56% of a Scottish sample (Helgason et al. 2000)). Frequencies of haplotype 1.15 (and its neighbours) in Orkney and Norway are 11% (41%) and 6% (18%), respectively. Although haplogroup 2 is much more diverse than 1, there is a subcluster found at high frequency only in Norway (network not shown): haplotype 2.47 and one-step network constitutes 38% of the sample. This mini-network, however, also occurs at a frequency of 16% in the Frisians, who may be similar to an Anglo-Saxon source population. The appearance of this cluster

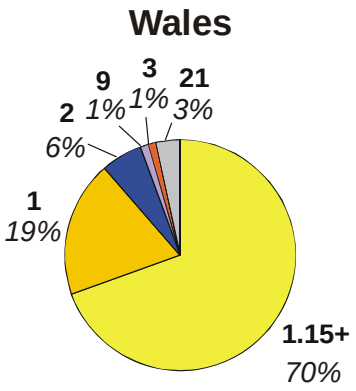
in mainland Britain, therefore, could be explained by either Scandinavian or Anglo-Saxon influence. Another one-step network, within haplogroup 3 (centred around haplotype 3.65), however is also present at high frequency in Norway (22%) and in Orkney (20%), but is rare in Friesland (and in the Netherlands (de Knijff 2000)). These frequency distributions suggest that both haplotypes 2.47 and 3.65 are diagnostic of Viking invaders in parts of Britain in which the only candidate parental populations are Celtic and Scandinavian, such as Orkney. In mainland Britain, however, it appears that only haplotype 3.65 would distinguish Norse and Anglo-Saxon contributions.

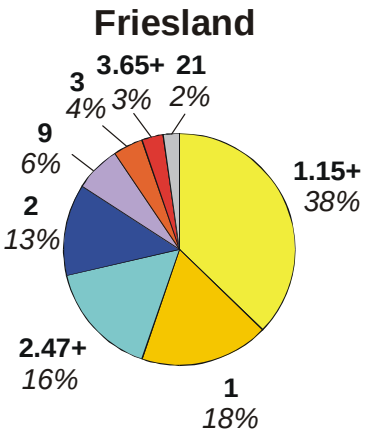
Figure 1. Charts of Y chromosome haplogroup and haplotype cluster frequencies in each population. Haplotype clusters 1.15+, 2.47+ and 3.65+ are within haplogroups 1, 2 and 3, respectively.

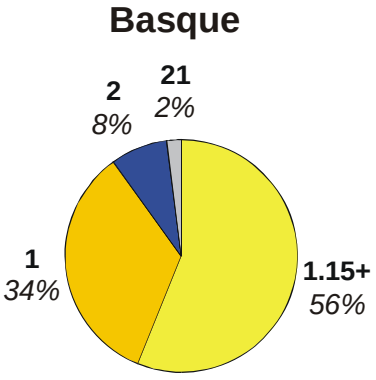

Orkney
3.65+
20%
1.15+
41%
2
7%
2.47+
7%
1
25%

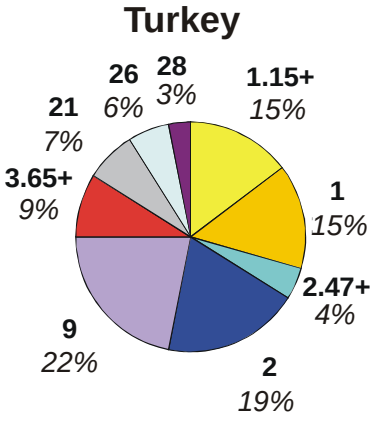

Ireland
2
4%
21
1%
26
1%
2.47+
5%
1.15+
44%
1
45%

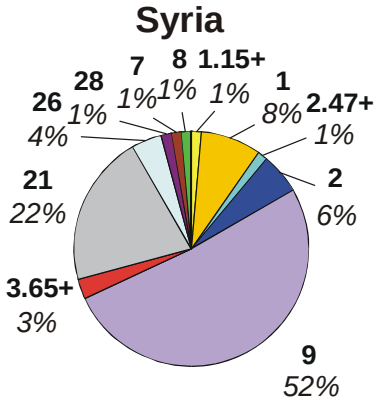

Norway
16
4%
21
1%
1.15+
18%
3.65+
22%
1
8%
9
3%
2
6%
2.47+
38%


Wales
9
1%
3
1%
21
3%
2
6%
1
19%
1.15+
70%


Friesland
3.65+
3%
21
2%
3
4%
9
6%
1.15+
38%
2
13%
2.47+
16%
1
18%


Basque
2
8%
21
2%
1
34%
1.15+
56%


Turkey
26
6%
28
3%
1.15+
15%
21
7%
3.65+
9%
1
15%
2.47+
4%
9
22%
2
19%


Syria
7
1%
8
1%
1.15+
1%
1
8%
2.47+
1%
28
1%
26
4%
2
6%
21
22%
3.65+
3%
9
52%

Correlation between Y chromosomes and surnames in Orkney. Orcadian surnames present a second method for identifying markers of Scandinavian contributions in the British Isles. Orcadian names can be divided into two classes: indigenous names endemic to the islands and those brought to the islands with Scottish settlers (Lamb 1981). As Y chromosomes co-segregate with surnames, haplotypes might be expected to reflect this partition to the extent that Norwegian and Scottish Y chromosome types can be distinguished. In fact, the distribution of chromosome types between the surname classes is significantly different both at the haplogroup (P=0.029, Table 1) and the haplotype (P=0.035) (Raymond and Rousset 1995) levels. Moreover, the putative Norse (2.47, 3.65) and pre-Anglo-Saxon British (1.15) types are clearly concentrated in the expected classes (indigenous and Scottish, respectively, data not shown). This confirms the heavier Viking component in the indigenous surname class and the increased pre-Anglo-Saxon British contribution to the Scottish surname class.

The frequency distribution in the indigenous Orkney chromosomes is consistent with a substantial Scandinavian contribution to the Orcadian Y chromosome pool. As the Scottish Orkney surname class is not statistically differentiated from the Welsh and Irish (P>0.2), it cannot have a significant Norse component. Considering, therefore, only the indigenous surnames, 38% of the Y chromosomes can be identified as Scandinavian in origin (haplogroup 3 and the 2.47 cluster), while those in haplogroup 1 are not of obvious provenance. Thus, the legacy of the Viking age in Orkney was both cultural and genetic.

Genetic continuity in the British Isles. Given the similarity of the Irish and Welsh samples (Table 1), the Y chromosome distributions shown in Fig. 1 appear to represent the pre-Anglo-Saxon population of the British Isles and Ireland. If extensive

genetic drift had occurred, there is no reason why these communities would remain so similar, especially as Wales and Ireland represent two different branches of the Celtic languages, P-Celtic and Q-Celtic, respectively. Two extreme possibilities regarding the demographic nature of early cultural transitions in the British Isles can be contrasted: (a) demic diffusion models such as the wave of advance model (Ammerman and Cavalli-Sforza 1984) proposed for the arrival of farming in Europe (Renfrew 1987) which predicts considerable genetic discontinuity, and (b) cultural diffusion models involving little or no movement of people, only the diffusion of technology, which predict genetic continuity. Similarly, the arrival of a Celtic material culture including Hallstatt and La Tène elite goods and skills in the late Bronze Age and early Iron Age was once interpreted as reflecting waves of immigrants, but is now usually explained without invoking folk migrations (Cunliffe 1997). As with the Neolithic, however, no solid evidence is available.

Table 1. Pairwise comparisons of Y chromosome haplogroup distributions. P values are from a test (Raymond and Rousset 1995) on RxC contingency tables analogous to Fisher's exact test for a 2 x 2 table. Scot Ork, Scottish Orkney Surnames; Indig Ork, Indigenous Orkney surnames; P>0.05 in bold; N/A not applicable.

	Basque	**Wales**	**Ireland**	**Scot Ork**	**Friesland**	**Orkney**	**Norway**	**Indig Ork**	**Turkey**
Wales	**0.96**								
Ireland	**0.75**	**0.16**							
Scot Orkney	**0.50**	**0.45**	**0.21**						
Friesland	0.00	0.00	0.00	**0.19**					
Orkney	0.00	0.00	0.00	N/A	0.00				
Norway	0.00	0.00	0.00	0.00	0.00	0.00			
Indig Orkney	0.00	0.00	0.00	0.03	0.00	N/A	0.00		
Turkey	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Syria	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Basque population history. To investigate the degree of paternal genetic continuity in the British Isles through the Neolithic and the development of Iron Age cultures, we compared the Welsh and Irish samples with 50 Basques (Bosch et al. 1999; Perez-Lezaun et al. 1997). The Basques are widely believed to be descended from the Palaeolithic inhabitants of Europe for reasons including: (a) Basque is a non-Indo-European language with some similarities suggesting a distant relationship with the North Caucasian language family (Bengtson 1991; Gamkrelidze and Ivanov 1990) and thus likely predates the arrival of the Indo-European languages. (b) Analyses of classical markers consistently place the Basques as genetic outliers in Europe, in particular the Basques have the highest frequency in Europe of the blood group O and of Rhesus *cde*, thought to represent the contribution of Palaeolithic Europeans (Cavalli-Sforza et al. 1994). (c) An analysis of European mtDNA estimates the Neolithic component in the Basques to be the lowest for any region in Europe. Although the criteria used to identify Near Eastern founder types are somewhat arbitrary and involve many assumptions, the relative number of types in different European populations should still be informative, and the Basque component, estimated at 6%, clearly lies outside the distribution within the rest of Europe estimated to range between 9% and 19% (Richards et al. 2000). We also sampled 68 and 72 unrelated, adult male Anatolian Turks and Syrians, respectively, the former as representatives of the source population for the European Neolithic and the latter as a representative of the Near East more generally. If the pre-Anglo-Saxon British, therefore, trace genetically to the European Palaeolithic, we might expect a similarity between the Irish and Welsh Y chromosomes and those of the Basques.

Basque and Celtic Y chromosomes. The Y chromosome complements of Basque and Celtic-speaking populations are strikingly similar (Fig. 1). Haplotype 1.15 is also modal in the Basques, and constitutes 41% of the sample, rising to 56% for the cluster of one-step neighbours. We call this the Atlantic Modal Haplotype (AMH). In each of the Basque, Welsh and Irish populations, a total of 89-90% of the chromosomes are in haplogroup 1 (which contains the M173-defined Eu18 haplogroup in Semino *et al.* (Semino et al. 2000)), with the majority of the remainder in haplogroup 2. The Turkish sample, however, is much more diverse at the haplogroup level (Fig. 1). The AMH and one-step neighbours are present (15%), but only one chromosome from this group is found in the Syrian sample (Fig. 1), and it is absent in India (unpublished) and Central Asia (Perez-Lezaun et al. 1999). There is no evidence, therefore, that incoming Neolithics or later immigrants originating in the Near East carried the AMH at frequencies as high as those characterising the Atlantic populations.

Other studies have suggested the possibility of a Basque-Celtic connection, most notably the synthetic maps of Cavalli-Sforza et al. (1994) that show Irish and Basque populations falling very near one another on the first principal component which is thought to reflect the spread of Neolithic farmers from the Near East. The relative proximity of the Basque and Irish on this axis may therefore reflect the relatively small Neolithic component in these populations. More recently, Hill *et al.* (Hill et al. 2000) have used a NW to SE cline through Europe of p49a, f haplotype XV (Semino et al. 1996) (which forms a subclade of haplogroup one (Jobling 1994)) to argue that haplogroup 1 in Ireland must be old. The present study, however, is the first to provide direct evidence of a close relationship in the paternal heritage of the Basque and the

Celtic-speaking populations of Britain. In fact, treating Orkney as a single population, all pairwise comparisons of haplogroup distributions between the populations included here are significantly different (Table 1) except for those within the Atlantic group – the Welsh, Irish and Basque – showing that they form a Y chromosome community with members more closely related to one another than any is to the other European populations. It should be noted that Basque-Celtic similarity not only implies that Basque and Celtic-speaking populations derive from common paternal ancestors, but that genetic drift in these communities has not been sufficiently great to differentiate them.

Analysis of molecular variance (AMOVA) was used to apportion Y chromosome genetic diversity among individuals within populations, among populations within groups and between groups in a hierarchical manner. When the Atlantic community (Irish, Welsh and Basques) form one group and the Frisians and Norwegians, the other, the within groups variance component is lowest (4.3%) and the between groups component highest (12.1%), consistent with the pattern of differentiation seen in Table 1. Moving the Basques to the Frisian-Norwegian group almost doubles the within groups variance component (to 8.0%) at the expense of the between groups component. Swapping the Irish or Welsh across groups increases the within groups component even more (to 10.6% and 12.4%, respectively).

The signal of Basque-Celtic similarity depends to a large degree on the AMH, which has much higher frequency in these populations than in other European populations. With one-step neighbours the AMH comprises only 38% of the Frisian sample (significantly different, $P<0.05$), consistent with the view that the Basques are genetically distinguishable from Continental populations generally. As three alleles

within this six locus haplotype are known to follow a SE to NW cline in Europe (Quintana-Murci et al. 1999), it is likely that most other European populations will have even lower frequencies than the Frisians. Both the Basque and the Celtic populations show high frequencies of the AMH. Because the former are generally considered to have received a very limited input of Near Eastern genes in the Neolithic, that similarity suggests that also in the British Isles the Neolithic transition did not entail a major demographic shift. Accordingly, farming may have spread in Britain more through cultural transmission than through some form of gene flow.

Coalescent times. Genealogical depths in haplogroup one were estimated to investigate whether coalescent times are consistent with its presence in Britain since the Palaeolithic. We used the average squared distance (ASD), that is the average across loci of the squared difference in the microsatellite repeat numbers between two haplotypes (Goldstein and Pollock 1997; Slatkin 1995). Under the single stepwise mutation model, the expectation of the ASD calculated between the ancestral and all observed haplotypes is equal to the product of the mutation rate and the average coalescence time in generations (Slatkin 1995). In haplogroup 1, we designated haplotype 1.15 as ancestral as it is modal and has modal alleles at all of its constituent loci as well as being the haplotype connected to the most other haplotypes in a network. Using a mutation rate of 1.2×10^{-3} (Bianchi et al. 1998), and a generation time of 27 years, the average coalescent times in the British Isles and the Continent are 6,800 and 7,100 years, respectively. The similarity of these values indicates that the populations in the Isles have not undergone extensive drift during colonisation or afterwards. However, the confidence intervals on the mutation rate alone (Bianchi et al. 1998) widen both estimates to between

approximately 2,900 and 18,400 years. This uncertainty associated with the estimated mutation rate is compounded by a likely systematic bias due to the mis-specification of the mutation model (Goldstein et al. in press). Finally, it should be noted that the real uncertainty is influenced not only by these factors, but also by the variation associated with the stochastic distribution of mutations through the haplogroup one genealogy. Because we do not know the shape of the haplogroup one genealogy it is difficult to assess this source of error (Goldstein et al. 1999). For these reasons the coalescent calculations are consistent with almost any historical scenario. Unfortunately it is not possible to calculate coalescence times for haplogroup 2, as it may be made up of divergent genealogical clades, but the availability of a biallelic marker defining the common European haplotype 2.47 cluster would allow comparison to the genealogical depth of haplogroup 1.

Beyond their similarity, the lack of variation within the Atlantic populations is also remarkable. The Basque, Welsh and Irish samples have mean microsatellite repeat count variances of 0.39-0.42, less than half that of Turkey (0.92) and much lower than Friesland, Norway, Syria and Orkney (0.62-0.72). The similarity and homogeneity together suggest one of two explanations. Either pre-agricultural European Y chromosomes were homogeneous, or there was a specific connection between the Basques, the pre-Anglo-Saxon British and Irish. With regard to the latter hypothesis it is interesting that a Northward expansion from a glacial refugium in Iberia has been postulated from the diffusion of Magdalenian industries (Otte 1990) and patterns of Y chromosome (Semino et al. 2000) and mtDNA variation (Torroni et al. 1998), but see

Simoni *et al.* (2000). More detailed investigation of the diversity present in and around Europe may allow these hypotheses to be distinguished.

Maternal and biparental genetic systems. Given the extraordinary similarity of the Atlantic Y chromosomes compared to those in other European populations, it is important to assess whether a similar pattern is observed in other genomic regions. In particular, we shall use a comparison of Y chromosome and mtDNA patterns of variation in order to evaluate whether cultural change in the British Isles has differentially affected male and female patterns of movement. In order to assess whether any differences are due to demographic factors as opposed to other differences between the two uniparental systems, we also include X chromosome markers influenced by both male and female patterns of movement.

Mitochondrial DNA. To investigate whether mtDNA variation showed the same patterns as the Y chromosome data, we sequenced the first hypervariable section of the control region (HVS1) and genotyped coding region variants as necessary to assign haplogroup (see Materials and Methods) in 90 Frisians and 59 Orcadians and compared these with 231 Norwegians (Opdal et al. 1998; Richards et al. 2000), 92 Welsh (Richards et al. 1996), 156 Basques (Bertranpetit et al. 1995; Richards et al. 2000), 101 Irish (Richards et al. 2000), 218 Turks (Richards et al. 2000) and 69 Syrians (Richards et al. 2000). Slowly evolving coding region variants and control region sites are used to assign mtDNAs into genealogical clades or haplogroups, while more quickly evolving control region sites define haplotypes within haplogroups. The haplogroup distributions in all the European populations are very similar (Pult et al. 1994; Richards et al. 1996). Turkey and Syria, however are distinct with much lower frequencies of the most common European

haplogroup (H) and large proportions of haplogroups not present or extremely rare in the European samples. The lack of structure is also evident at the haplotype level of resolution: AMOVA apportions 99% of the variance in our European populations between individuals within populations, regardless of the grouping scheme.

Principal components analysis. PC analyses were performed on both Y chromosome and mtDNA haplogroup frequencies (Fig. 2). In each case the first PC (explaining 65% and 54% of the variation, respectively) depicts a general East-West population gradient; a pattern usually interpreted as indicating the Neolithic component (Cavalli-Sforza et al. 1994; Cavalli-Sforza and Minch 1997). In line with this interpretation, the poles of the first PC of both systems are defined, on the one hand, by the Basques and, on the other, by Turkey and Syria. As may be expected, in the Y chromosome plot, the Celtic-speaking populations fall extremely close to the Basques and Orkney falls mid way between the Atlantic cluster and Norway. This is in sharp contrast with the mtDNA pattern in which the Celtic-speaking populations are closer to the centre of the plot, indicating that they have undergone more female-mediated gene flow from other European populations than the Basques have. Thus, at least one of the cultural transitions in the British Isles since the Upper Palaeolithic must have involved a demic component on the female side. The similarity of the non-Basque European populations means that there is no power to apportion the Orcadian maternal heritage into Scandinavian and pre-Anglo-Saxon British components using the available mtDNA data.

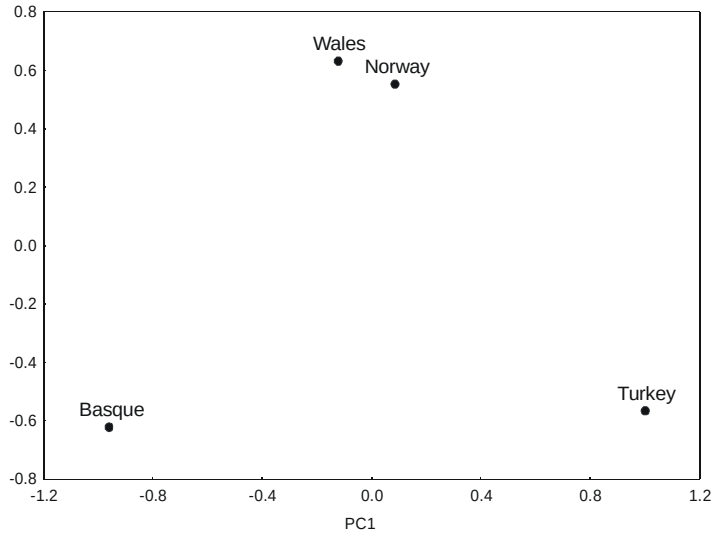
X chromosome microsatellites. To assess which of the two uniparentally inherited genetic systems more closely reflects the history of the genome more widely and to check that the lack of differentiation among the British and non-Basque European

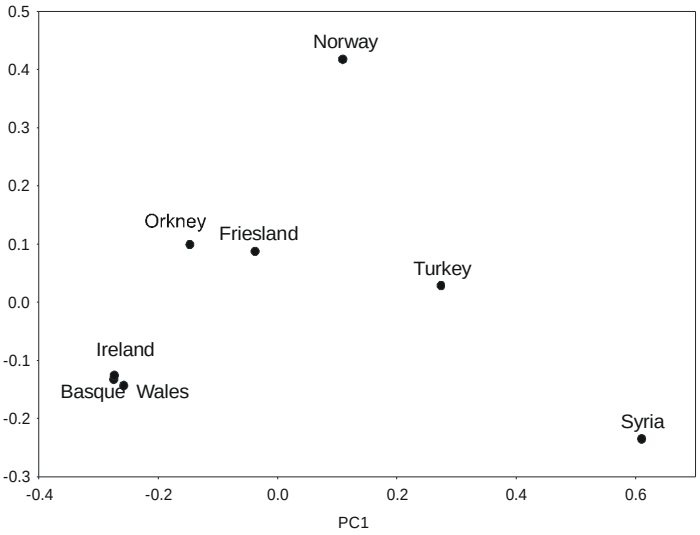
populations is not due to a lack of resolution in the mtDNA data, we analysed microsatellites on the X chromosome. Although having far less genealogical information at each genetic locus than is available for completely linked systems such as mtDNA and the Y chromosome, multilocus genotypes are known to provide a sensitive test of population structure (Goldstein et al. 1999; Pritchard et al. 2000). Thirty-four dinucleotide markers located across the length of the X chromosome were genotyped in the Basques, Norwegians, Welsh and Turks. Population structure was assessed using a model-based clustering approach implemented in the STRUCTURE program (Pritchard et al. 2000). Briefly, the model assumes K populations, each characterised by a set of allele frequencies at each locus and individuals are assigned to these populations on the basis of their genotypes. We estimated $\Pr(X|K)$, where X is the data, for $K = \{1,2,3,4\}$. Using Bayes' theorem and assuming a uniform prior on K between 1 and 4, we can then approximate the posterior distribution, $\Pr(K|X)$. For the Basque, Welsh, Norwegian and Turkish data, all the posterior probability is on $K = 1$, i.e. there is no detectable genetic structure (Table 2).

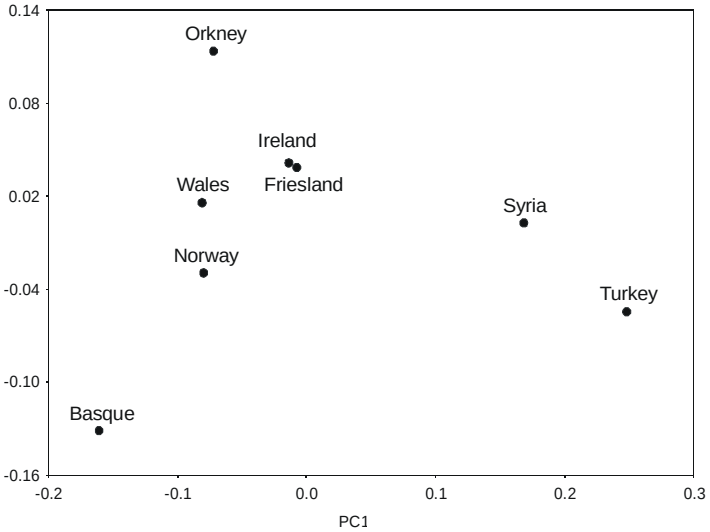
Table 2. Inferring the number of clusters.

K	ln Pr(X|K)	Pr(K|X)
1	-9281.6	1.000
2	-9522.1	~0
3	-9673.2	~0
4	-9776.9	~0

Fig. 2 Plots of the first and second principal components from top to bottom of (i) X microsatellite allele, (ii) Y chromosome haplogroup and (iii) mtDNA haplogroup frequency distributions. All Y haplogroups were included while the following common European mtDNA haplogroups were included: H, V, J, T, I, W, X, U3, U4, U5 and K. In the Y chromosome data, the proportion of the variation explained by PC1 is 65% and by PC2 is 28%, while in the mtDNA and X microsatellite data, these figures are 54% and 15%, and 46% and 33%, respectively.


Wales
Norway
Basque
Turkey
PC1


Norway
Orkney
Friesland
Turkey
Ireland
Basque
Wales
Syria
PC1


Orkney
Ireland
Wales
Friesland
Syria
Norway
Turkey
Basque
PC1

However, when we performed a PC analysis on the allele frequencies at these 34 X-linked microsatellites, we observe a pattern essentially identical to that seen for mtDNA (Fig. 2). Once more the Basques and Turks occupy opposite poles of PC1 and the Welsh and Norwegians fall in the centre of the plot. Despite their being no statistical support for genetic structuring in the X microsatellite data considered on their own, the similarity of the patterns observed across different genetic systems provides robust evidence that the Basques are differentiated from the other European populations, specifically in having a lower input from the Near East.

Female-mediated gene flow between the Celtic-speaking populations and other North European populations has thus homogenised the variation, not only for mtDNA but also for other parts of the genome affected by female migration. There are two extreme scenarios that could account for the sharp differences observed between the genetic systems that are (mtDNA, X chromosome) and are not (Y chromosome) affected by female movement. First, the pre-Anglo-Saxon British source populations may have been geographically structured for the Y chromosome but less so for other regions of the genome, which would indicate that the female gene flow occurred prior to the founding of the Welsh and Irish populations. This explanation is inconsistent with the position of the Basques, however, which is distinctive for both the Y chromosome and the systems affected by female migration. The second explanation is that the European Palaeolithic populations were originally distinct from the current European population for both the Y chromosome and other parts of the genome, but this distinctiveness was subsequently eroded by the female movements between the Celtic-speaking and non-Basque European

populations. In other words at least one of the Neolithic or Iron Age cultural transitions in the British Isles involved some female immigration.

Population parameters such as estimates of divergence times inferred from one locus systems always have a high variance because information is only incorporated from one realisation of the evolutionary process. Certain evolutionary questions, however, are less subject to this source of variation, and can be profitably addressed with only a single genetic locus. For example, identification of related lineages in different populations could be taken as secure evidence of some kind of connection between the populations such as gene flow or common ancestry, even though genetic drift at a single locus would make it impossible to accurately estimate parameters reflecting the quantitative relationship (e.g. migration rate or population separation time). Despite these problems, in cases where female migrations have homogenised the variation in other parts of the genome, the Y chromosome may be the only signal of certain historical relationships.

In summary, we have identified markers of paternal Scandinavian influence in the British Isles which suggest that the Viking settlement of Orkney involved substantial genetic as well as cultural replacement. Accepting the widely-held view that the Basques are representative of pre-Neolithic European Y chromosomes (Cavalli-Sforza et al. 1994), we have also shown that the Neolithic, the Iron Age and subsequent cultural revolutions had little effect on the paternal genetic landscape of the Celtic-speaking populations: there has been continuity from the Upper Palaeolithic to the present. However, comparison with mtDNA and X-linked microsatellites revealed that at least one of these cultural revolutions had a major effect on the maternal genetic heritage of the Celtic-speaking populations.

Chapter 3

Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population

The optimal marker density for genome scans in case control association studies, and the appropriate study design for testing candidate genes, both depend on the genomic pattern of linkage disequilibrium (LD). Here we provide the first conclusive demonstration that the diverse demographic histories of human populations have produced dramatic differences in *genome-wide* patterns of LD. Using a panel of 66 markers spanning the X chromosome we show that in the Lemba, a Bantu-Semitic hybrid population, markers up to about 21 cM apart have a greater tendency to show LD than unlinked markers. In three populations with less evidence of admixture, however, excess LD disappears beyond 2 cM. Moreover, analysis of Bantu and Ashkenazi populations as putative parental populations of the Lemba shows a significant relationship between allele frequency differentials and the LD observed in the Lemba, demonstrating that much of the excess LD is due to admixture. Our results suggest that demographic history has such a profound effect on LD that it will not be possible to predict patterns *a priori*, but will be necessary to empirically evaluate the patterns in all populations of interest.

Introduction

Case-control association studies are increasingly the method of choice in efforts to map genes underlying complex traits (Risch and Merikangas 1996). It remains unclear, however, how densely markers must be distributed throughout the genome in order to reliably detect association with causal variants (Collins et al. 1999; Kruglyak 1999). In particular, considering variants relevant to common disease, the genetic distance over which significant linkage disequilibrium (LD) occurs will determine the appropriate spacing of markers, while the consistency of LD at a given genetic distance will determine the number of markers of a given type required within each interval.

Uncertainty about optimal marker spacing is due in part to our ignorance of the nature of the genetic variation influencing common disease. In particular, recent mutations will tend to have more LD than older ones, and it is not clear whether the common diseases are influenced more by common or by rare alleles (Kruglyak 1999; Wright et al. 1999). Uncertainty is greatly compounded, however, by our ignorance of the distribution of LD in human populations for any class of variants. It should be noted that even when exhaustive single nucleotide polymorphism (SNP) maps are developed, the interpretation of association studies will still require detailed knowledge of LD. In this case the pattern of LD will determine the genetic distance over which false signals of causation may be generated by the association of a candidate SNP with a linked causal variant. This complication would apply equally to tests of the role of specific variants in candidate genes using case-control designs. For these reasons interest has focused on measuring and predicting levels of LD in human populations. These patterns of association, however, are influenced both by

demographic factors, which affect the entire genome, and by genetic factors such as mutation rates and selection influencing particular genomic regions (Freimer et al. 1997; Wright et al. 1999). Uncertainties concerning the demographic histories of human populations, however, make it difficult to accurately predict patterns of LD even in the case of neutrality. In a recent theoretical study, for example, Kruglyak (1999) noted that under simplified demographic assumptions useful LD would not extend beyond about 3 kb in most human populations. The little evidence that is available, however, indicates much more complicated demographic scenarios for most populations, involving both rapid expansions and bottlenecks (Collins et al. 1999; Reich and Goldstein 1998).

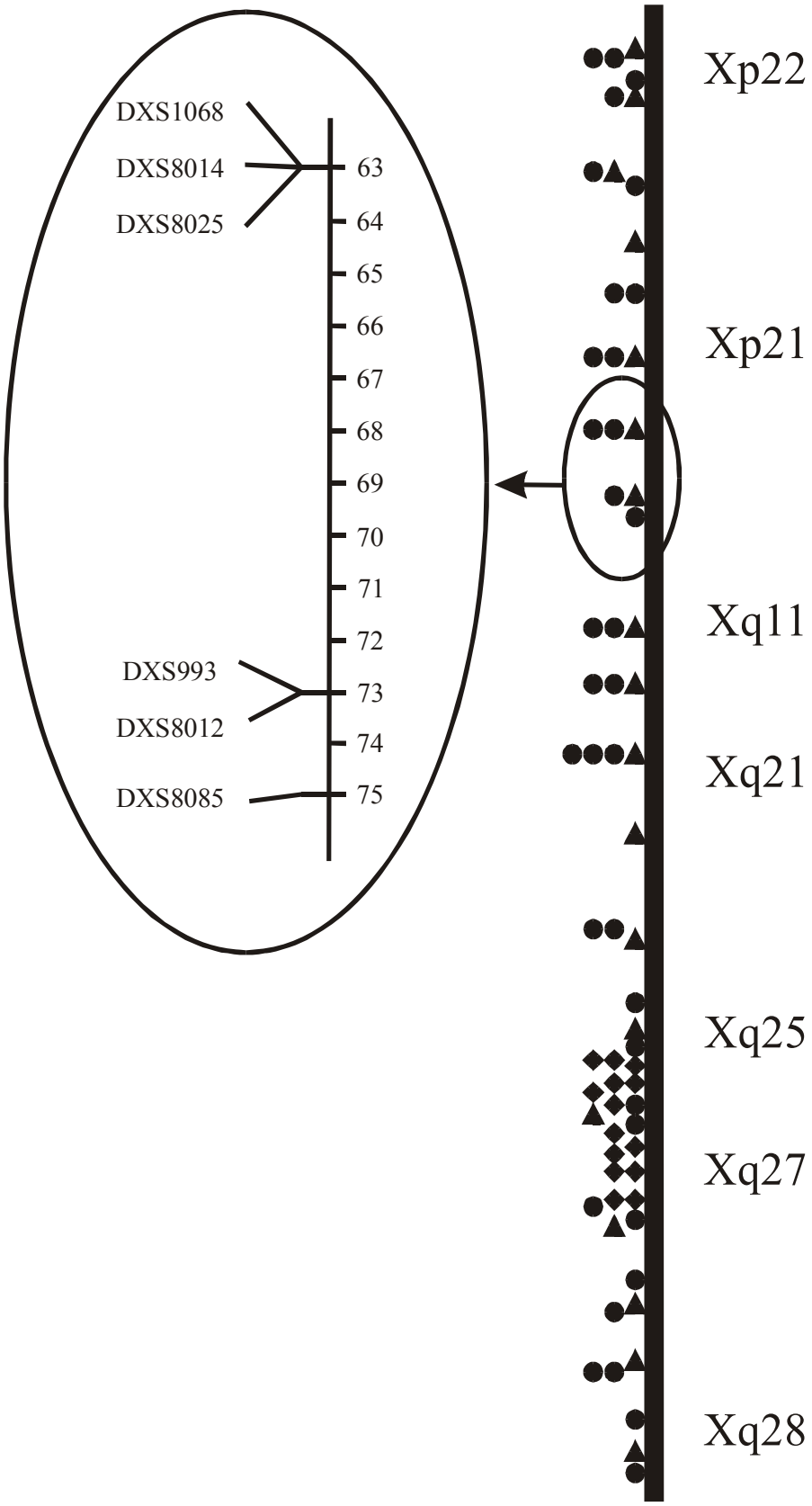
Optimal design and interpretation of association studies therefore requires empirical study of LD between pairs of markers widely distributed in multiple genomic regions. A description of such "background" LD would allow observed associations to be evaluated in the context of the distribution of LD in the relevant population. Most studies of background LD to date, however, are limited either in the number of genomic regions considered (Laan and Pääbo 1997), the number and type of populations studied (Huttley et al. 1999), or the range of genetic distances represented (Goddard et al. 2000), but see Taillon-Miller *et al.* (2000). It is noteworthy that even in the model populations targeted in many association studies (e.g. Ashkenazi Jews and Finns) the assumptions of relative homogeneity and limited disequilibrium beyond half a centimorgan (cM) or so have not been confirmed through systematic genetic evaluation. In fact, in the case of Finland, preliminary data suggest both genetic heterogeneity (Kittles et al. 1998) and the presence of some long-range LD (Peterson et al. 1995). What is critically needed is a consistent

experimental framework for measuring LD across large genomic regions that can be applied to multiple populations with different demographic histories.

Here we concentrate on admixture, arguably the most important demographic factor that is ignored in theoretical treatments seeking to predict patterns of LD in human populations. It is well known that population substructuring (or stratification) can generate spurious signals in association studies, and various methods have been proposed for identifying it (Pritchard and Rosenberg 1999) and for taking it into account in statistical analyses (Reich and Goldstein 2001). Even if no stratification is present, however, historical admixture between differentiated populations can result in significantly elevated LD over large genomic regions for many generations (Chakraborty and Weiss 1988; McKeigue 1998; Stephens et al. 1994). In fact, it has even been suggested that in some populations the LD generated by admixture would allow genome scans with only one marker every 10 cM or so, more than three orders of magnitude less than the theoretical suggestion of Kruglyak (1999). These predictions about patterns of LD in admixed populations, however, have yet to be validated empirically.

Here we report on the analysis of 66 microsatellite markers specifically designed to allow formal assessment of the effect of admixture on linkage disequilibrium. The markers were selected in order to provide dense, and so far as possible uniform coverage of a broad range of genetic distances. As this panel was developed to assess the effect of recent admixture we have concentrated attention on relatively large genetic distances, and the coverage within the interval 0-1 cM is therefore very uneven. Collectively these markers provide 2145 pairwise observations of LD in multiple regions throughout the X chromosome (Fig. 1).

Figure 1. Distribution of the 67 markers in the admixture linkage disequilibrium panel across the X chromosome. Note DXS1047 is not reported on here. Approximate cytogenetic locations are indicated. Genetic distances are in centimorgans (cM) according to the Genethon map. The X chromosome ABI Prism linkage mapping panel markers (▲) are spaced approximately every 10 cM across the chromosome, flanking these are usually two microsatellites within 1 cM (●) (see detail). These together provide multiple pairwise comparisons over genetic distances of 0-1 cM and multiples of 10 cM. Distances between 1-10 cM and their multiples are similarly covered by a series of microsatellites (♦) at Xq25-27.


DXS1068
DXS8014
DXS8025
63
64
65
66
67
68
69
70
71
72
DXS993
73
DXS8012
74
DXS8085
75
Xp22
Xp21
Xq11
Xq21
Xq25
Xq27
Xq28

Materials and Methods

Samples. All DNA samples had been collected from paternally unrelated males. Lemba subjects originated from Sekhukuneland in Mpumalanga, South Africa, Bantu-speakers from various chieftainships in the Pretoria area and mixed caste Ashkenazi Jews from Tel Aviv, Israel (Thomas et al. 2000). Ethiopians were sampled in Addis Ababa and consisted mainly of Amharic and Oromo speakers from the Wollo and Shewa provinces.

Genotyping. The admixture LD marker panel was chosen as described in Fig. 1. Markers additional to the X chromosome ABI Prism linkage mapping panel (PE Biosystems, Foster City, CA) were chosen from the Genome Database (http://gdbwww.gdb.org/) and some primers were redesigned from GenBank sequences as were primers for a novel marker 7.5 kb from DXS1203 in PAC 455H14. Multiplex PCR was performed on 65 dinucleotide loci (from the Genethon linkage map (Dib et al. 1996)), one tetranucleotide ($GATA_n$) locus (from the Marshfield map (Broman et al. 1998)) and the novel (CA_n) microsatellite in 13 kits, details of which are given in Table 1.

Briefly, a 10X primer mastermix was made in advance, the PCR mastermix was then made up of primer mastermix (between 1.5-5.0 pmol each primer), 10X Super Taq PCR buffer 1 (HT Biotechnol., Cambridge, England), 0.2 mM dNTPs (Advanced Biotechnologies, Epsom, England), Super Taq (HT Biotechnol., Cambridge, England) and TaqStart (Clontech, Palo Alto, CA) (premixed in a 2:1 ratio, neat Taq: neat TaqStart) and finally dH_2O to a total of 10 μl per reaction. Meanwhile approximately 1 – 10 ng of DNA were pipetted into a 96 well microtitre plate. After vortexing, the mastermix was aliquoted into microtitre plate strip lids.

Following a centrifugation step, the samples were cycled in a Perkin-Elmer 9700 PCR machine (PE Biosystems, Foster City, CA). Thermal profiles were 38 cycles of 30 seconds each at 95°, 55° and 72° with a 4 minute 95° initial denaturation step and a 10 minute 72° final extension, in some cases a touchdown procedure was used, decreasing the annealing temperature by 0.5°/cycle for the first 8 cycles. The Prism panel markers were cycled per manufacturer's instructions. PCR products were diluted and pooled, prior to loading on a 96-lane, ABI 377 sequencer (PE Biosystems, Foster City, CA). In this way 3-18 markers were loaded in a lane allowing between 288 and 1,728 genotypes per gel to be called. The average number of loci per lane was 10. Size calling was performed using Genescan software. A control individual was run at least twice on each gel to standardise for gel to gel shifts in migration. 455H14 is a novel locus in the PAC of the same name. Note DXS1047 was not typed in the Ashkenazim and so was not used in the analysis.

Table 1. Details of the admixture linkage disequilibrium panel. The column 'PCR kit' shows the multiplex grouping. 'Conc' is the final concentration of the primer in micromoles per litre. 'cM' is the position along the X chromosome in centimorgans according to the Genethon map. 'Jim' shows the control genotype on an ABI 377 for gel to gel variation in mobility.

No.	Kit	Locus/Primer	Conc	Primer sequence	Label	cM	Jim
1	**XLD1**	DXS7103u	0.30	CACACACCCCTACCTGGA	FAM	16	135
2	XLD1	DXS7103r	0.30	CCCTAGAAGTTTTGCCCC	--	--	--
3	XLD1	DXS8036u	0.60	ATACAAACTGCCCCACTTCC	HEX	28	125
4	XLD1	DXS8036r	0.60	CTCTGGNTCCTGTCCTGG	--	--	--
5	XLD1	DXS8061u	0.12	GCTTGAAGTGTCCATGAGGTATC	TET	205	145
6	XLD1	DXS8061r	0.12	AGAAGCTGATGTGCTCCCTG	--	--	--
7	XLD1	DXS1232u	0.40	ACCAACAGCCTAATAATGC	FAM	171	190
8	XLD1	DXS1232r	0.40	AGAGATGGGAGCAGCA	--	--	--
9	XLD1	DXS984u	0.35	TTTCTGTCTGCCAAGTGTTT	HEX	171	172
10	XLD1	DXS984r	0.35	CCTACTCCATTCCACACT	--	--	--
11	XLD1	DXS8085u	0.20	TCAAAGAGGTTTTGCCAC	TET	75	160
12	XLD1	DXS8085r	0.20	AGATAAAGACATCCTGCCTAGTTC	--	--	--
13	**XLD2**	DXS8027u	0.50	GTGAGACGCTGTCTTGG	FAM	44	240
14	XLD2	DXS8027r	0.50	AGCTGCTGTACTAATAACATAGG	--	--	--
15	XLD2	DXS1067u	0.08	TATGTCCTCAGACTATTCAGATGCC	HEX	53	223
16	XLD2	DXS1067r	0.08	CCTCCAGTAACAGATTTGGGTG	--	--	--
17	XLD2	DXS1204u	0.20	ATGAACCCTTAACTCATTTAGCAGG	TET	92	244
18	XLD2	DXS1204r	0.20	AGCNTGCACCAACATGCC	--	--	--
19	XLD2	DXS8087u	0.25	GGAGTCCCTGAGGCAG	FAM	210	285
20	XLD2	DXS8087r	0.25	AAGGCCAGCAGCATCA	--	--	--
21	XLD2	DXS8009u	0.50	AGCCTTCGTCCTCTATATATTTC	HEX	158	250
22	XLD2	DXS8009r	0.50	GCAATTCAAAAGATTCTGATTAATT	--	--	--
23	XLD2	DXS8092u	0.30	CACCCTATGGCCTAGC	TET	100	271
24	XLD2	DXS8092r	0.30	ACCCAAACTTGCTCAGG	--	--	--
25	**XLD3**	DXS8099u	0.20	AGCTGGTTTTGTGATTCTGC	FAM	44	108
26	XLD3	DXS8099r	0.20	GGCCTTGAGATGTAGCCA	--	--	--
27	XLD3	DXS996u	0.45	AAATTCTTGCTTAGGCCACTCTAGG	FAM	11	155

28	XLD3	DXS996r	0.45	ACGTTGTTCTGGATCGTATGCTAGG	--	--	--
29	XLD3	DXS1036u	0.16	TGCAGTTTATTATGTTTCCACG	HEX	53	148
30	XLD3	DXS1036r	0.16	GCCATTGATAAGTGCCAGAT	--	--	--
31	XLD3	DXS1205u	0.40	CCTACGCATGTGGCTC	HEX	175	186
32	XLD3	DXS1205r	0.40	ATTAATGGCTTAGAGTACTTTTTCA	--	--	--
33	XLD3	DXS1223u	0.20	TGCCAATTTTTGCTTTTGTTATG	TET	14	168
34	XLD3	DXS1223r	0.20	AGGATTTTGCCACTCACTTCA	--	--	--
35	**XLD4**	DXS8078u	0.30	TGCATCCCCATAGTAATTGGT	FAM	159	199
36	XLD4	DXS8078r	0.30	CAAATGGCAGGATTTCCC	--	--	--
37	XLD4	DXS1212u	0.20	TGGAAGCATGAGAAATCACATCCT	FAM	153	234
38	XLD4	DXS1212r	0.20	TGGCATTACAAGCCCTCAAGTC	--	--	--
39	XLD4	DXS8081u	0.60	GATCCTCTGACCCTCATTTC	HEX	135	229
40	XLD4	DXS8081r	0.60	TGATAGGACTTTGCAGTGG	--	--	--
41	XLD4	DXS1053u	0.30	TTAAGGAAGTATGAGGCTCCA	TET	27	196
42	XLD4	DXS1053r	0.30	TTGGTGCAAAAGTAATCACG	--	--	--
43	XLD4	DXS8013u	0.50	CCAACCCAACTGTCTATCAA	TET	172	222
44	XLD4	DXS8013r	0.50	GTTTGGTTTTCCATTCCTGA	--	--	--
45	XLD4	DXS8098u	0.25	CAGAGGCATTTACCAAGC	TET	153	252
46	XLD4	DXS8098r	0.25	CCCCTGGAGCAAAGAC	--	--	--
47	**XLD5**	DXS1220u	0.22	AGCGAGAGTCTGACCCAC	FAM	135	213
48	XLD5	DXS1220r	0.22	GGGGCCTATAAAATGGAG	--	--	--
49	XLD5	DXS1206u	0.40	CACAATCTGCTGTCTGCAAT	HEX	157	167
50	XLD5	DXS1206r	0.40	AGCATGGGACTTCTCAACC	--	--	--
51	XLD5	DXS8038u	0.60	GTGGACTGTCTCCGTAACC	HEX	157	137
52	XLD5	DXS8038r	0.60	CCAAGATGTGAGCATTTTTC	--	--	--
53	XLD5	DXS1211u	0.35	CCCTCCAATCTGGCAGAA	TET	167	159
54	XLD5	DXS1211r	0.35	AAGACCTGGGTTTGGCCT	--	--	--
55	XLD5	DXS1192u	0.25	GTTGCCAACTGCTGGAACG	TET	167	128
56	XLD5	DXS1192r	0.25	TGTGGTGCAGGGAAGCC	--	--	--
57	**XLD6**	DXS8086u	0.37	GACACATAATTGTGTATTAGCCAG	FAM	199	235

58	XLD6	DXS8086r	0.37	TGACTCACCCTGCAGATC	--	--	--
59	XLD6	DXS8014u	0.50	GGCAAAGTTGTCAGAGGC	FAM	63	272
60	XLD6	DXS8014r	0.50	CAAATGGCTTGTTCCAGTT	--	--	--
61	XLD6	DXS8073u	0.30	GAAAATGTCTGGTGTGCTAC	HEX	186	216
62	XLD6	DXS8073r	0.30	TTAATATCTCAGGGCTAGAGTCC	--	--	--
63	XLD6	DXS1062u	0.14	GAGATGTGTGACCTTGAGCACT	HEX	165	236
64	XLD6	DXS1062r	0.14	GTTGCCTGTTAAGCACTTTGAATC	--	--	--
65	XLD6	DXS1203u	0.18	CCTGAATTTCCCCAGC	TET	110	213
66	XLD6	DXS1203r	0.18	TCCCACTTGCCAACTC	--	--	--
67	XLD6	DXS8068u	0.25	ACTTAGCATAATGTCCTCCAG	TET	159	239
68	XLD6	DXS8068r	0.25	CCTCTGTAAAGAACCTGCAC	--	--	--
69	**XLD7**	DXS8025u	0.30	AGCTGGGCAGCAGAGC	TET	63	197
70	XLD7	DXS8025r	0.30	TGGGTAATTCTTTGACATCACTC	--	--	--
71	XLD7	DXS1193u	0.15	AATTCTGACTCTGGGGC	HEX	199	137
72	XLD7	DXS1193r	0.15	TTATTTTAAGGTGAGTATGGTGTGT	--	--	--
73	**XLD8**	DXS8057u	0.50	GGGGTAATATGCAGCCTC	FAM	154	221
74	XLD8	DXS8057r	0.50	AGCCACCATGCCTAGC	--	--	--
75	XLD8	DXS8032u	0.45	CATTTTATTTTGCTTTGTATTTGGC	HEX	92	193
76	XLD8	DXS8032r	0.45	CTCCTAGAACAGTACCTGACACG	--	--	--
77	XLD8	DXS8072u	0.27	AGGAAGTAAAAATTTACGGTTGT	HEX	161	219
78	XLD8	DXS8072r	0.27	TCTCCCTATCCAACTCATGC	--	--	--
79	XLD8	DXS8094u	0.30	GCCATTGTAAAATAAAATTCAG	TET	164	228
80	XLD8	DXS8094r	0.30	ATGGTCTTGAGTCACTGTCT	--	--	--
81	**XLD9**	DXS8067u	0.20	CAGGAGTCCAAGGCTGCT	FAM	146	225
82	XLD9	DXS8067r	0.20	CACAGAGTGATACCCTGTCTCTAAA	--	--	--
83	XLD9	DXS8012u	0.20	AGCTTAGCAAGCCCAAGTAA	HEX	73	183
84	XLD9	DXS8012r	0.20	GGACATAAACATTCAGACCATAAC	--	--	--
85	XLD9	DXS994u	0.40	CTGTCCTACCCTGTACTGTCAC	HEX	162	212
86	XLD9	DXS994r	0.40	TATTGTCCTACTGGGCATAGAG	--	--	--
87	XLD9	455H14u	0.30	GATTCAAGAACCCTCTCTTG	TET	110	130

88	XLD9	455H14r	0.30	TGGGGTAGATTTAGAAATGATA	--	--	--
89	XLD9	DXS6801u	0.20	AGTCATTTCCTCTAACAAGTCTCC	FAM	110	133
90	XLD9	DXS6801r	0.20	TCCAGAGAGTCAGAATCAGTAGG	--	--	--
91	XLD9	DXS8059u	0.20	TTCCAGGTGCCACCAAG	TET	152	211
92	XLD9	DXS8059r	0.20	CAGTGCATTCCAGCCAGATAC	--	--	--
93	**XLD10**	DXS8105u	0.50	TTTAGTTTCCTGCCACG	FAM	11	152
94	XLD10	DXS8105r	0.50	TATTCCATGTTTTTCATATTGAG	--	--	--
95	XLD10	DXS8028u	0.40	TGATGACACTCGGACTGC	FAM	189	238
96	XLD10	DXS8028r	0.40	GAAATAATAATACTTGCCTTGCCT	--	--	--
97	XLD10	DXS8082u	0.25	TGACACCTCCAAAGGCTC	HEX	100	218
98	XLD10	DXS8082r	0.25	ACCTCCTTGGTTAAATGTATTCC	--	--	--
99	**PrismS**	DXS1227u	0.30	Proprietary information	FAM	176	92
100	PrismS	DXS1227r	0.30	Proprietary information	--	--	--
101	PrismS	DXS990u	0.30	Proprietary information	FAM	110	130
102	PrismS	DXS990r	0.30	Proprietary information	--	--	--
103	PrismS	DXS8091u	0.30	Proprietary information	HEX	198	86
104	PrismS	DXS8091r	0.30	Proprietary information	--	--	--
105	PrismS	DXS8051u	0.30	Proprietary information	NED	16	132
106	PrismS	DXS8051r	0.30	Proprietary information	--	--	--
107	PrismS	DXS1106u	0.30	Proprietary information	HEX	121	133
108	PrismS	DXS1106r	0.30	Proprietary information	--	--	--
109	PrismS	DXS1047u	0.60	Proprietary information	HEX	160	164
110	PrismS	DXS1047r	0.60	Proprietary information	--	--	--
111	**PrismM**	DXS986u	0.30	Proprietary information	FAM	100	155
112	PrismM	DXS986r	0.30	Proprietary information	--	--	--
113	PrismM	DXS987u	0.30	Proprietary information	FAM	27	211
114	PrismM	DXS987r	0.30	Proprietary information	--	--	--
115	PrismM	DXS8043u	0.30	Proprietary information	NED	188	162
116	PrismM	DXS8043r	0.30	Proprietary information	--	--	--
117	PrismM	DXS1068u	0.30	Proprietary information	HEX	63	261

118	PrismM	DXS1068r	0.30	Proprietary information	--	--	--
119	PrismM	DXS1060u	0.30	Proprietary information	NED	10	252
120	PrismM	DXS1060r	0.30	Proprietary information	--	--	--
121	PrismM	DXS993u	0.30	Proprietary information	FAM	73	288
122	PrismM	DXS993r	0.30	Proprietary information	--	--	--
123	**PrismL**	DXS1214u	0.30	Proprietary information	HEX	53	293
124	PrismL	DXS1214r	0.30	Proprietary information	--	--	--
125	PrismL	DXS1226u	0.30	Proprietary information	NED	37	296
126	PrismL	DXS1226r	0.30	Proprietary information	--	--	--
127	PrismL	DXS8055u	0.30	Proprietary information	HEX	136	317
128	PrismL	DXS8055r	0.30	Proprietary information	--	--	--
129	PrismL	DXS991u	0.30	Proprietary information	NED	92	328
130	PrismL	DXS991r	0.30	Proprietary information	--	--	--
131	PrismL	DXS1001u	0.30	Proprietary information	HEX	149	200
132	PrismL	DXS1001r	0.30	Proprietary information	--	--	--
133	PrismL	DXS1073u	0.30	Proprietary information	FAM	208	311
134	PrismL	DXS1073r	0.30	Proprietary information	--	--	--

Analysis. LD was measured by a test of independence between alleles at pairs of loci (Slatkin 1994), which is only sensitive to marginal frequencies. This extension of Fisher's exact test for R x C contingency tables was implemented using ARLEQUIN (Schneider et al. 1997). The probability of finding a table with the same marginal totals which has a probability equal to or less than the observed table was obtained using a Markov chain to efficiently explore the space of all possible randomly-shuffled tables (Guo and Thompson 1992). These LDp values were calculated by first dememorising the chain for 10,000 steps and running the chain in batches until the error on the P value was less than 0.001. Contingency tables testing for interactions between genetic distance or delta product and LDp were evaluated using χ^2. Corrections for multiple comparisons involving the windows covering different genetic distances were made using the Dunn-Sidak sequential correction (Sokal and Rohlf 1995), $P' = 1-(1-\alpha)^{1/k}$, where k = the rank of the P value from highest to lowest, and using $\alpha = 0.05$. The correction is conservative in our case, as the windows are overlapping and thus non-independent. Composite delta values for each locus were calculated as the sum of all allele-by-allele deltas of like sign. The log-linear test for a three-way interaction between LDp, cM and $\delta_1\delta_2$ was implemented in STATISTICA.

Results

Populations. The admixture LD panel was genotyped in the Lemba (n = 96), a Bantu-Semitic hybrid population, and in three other populations with less evidence of admixture. The Lemba are a Southern African group who speak a variety of Bantu languages and claim Jewish ancestry. Y chromosome analysis (Spurdle and Jenkins

1996; Thomas et al. 2000) indicates that approximately 68% of the Lemba Y chromosomes are Semitic and 32% are Bantu in origin. Although it was not possible to distinguish an Arabic contribution from a Jewish one, the high frequency of the Cohen Modal Haplotype, a putative Hebrew signature, may indicate a specifically Judaic component in the Lemba gene pool. Mitochondrial DNA studies (Soodyall et al. 1996) found no evidence of female-mediated admixture. As putative parental populations of the Lemba, as well as examples of populations with less evidence of admixture, we chose South African Bantu (n = 86) and Ashkenazi Jews (n = 80). The latter are also of interest because of their frequent use in genetic epidemiology (Wright et al. 1999) and may have undergone some degree of admixture with Europeans. As an alternative second parental population we used a sample of Ethiopians (n = 77) (described in methods). The Y chromosomes in this sample were more similar to the non-Bantu Lemba chromosomes than those of the Ashkenazim (data not shown). Only the first 34 markers listed in the Methods were genotyped in the Ethiopian sample.

Extent and consistency of linkage disequilibrium. LD was measured using an extension of Fisher's exact test giving a P value we refer to as LDp. In total, twice as many marker pairs are in significant LD (LDp<0.05) in the Lemba (296 of the 2145 pairs, 13.8%) compared to the Ashkenazi Jews (7.0%), Bantu (7.7%) and Ethiopians (6.4%). To investigate the tendency of marker pairs to be in LD as a function of genetic distance, we compared the LD in a given distance interval to that between unlinked marker pairs using contingency tables. The tables compare the counts of significant and non-significant LDp values in a sliding window of 6 genetic distances (0 through 5 cM, 1-6 cM etc) with the numbers of significant and non-significant

unlinked pairs (distance ≥50 cM) (Fig. 2). The comparison to unlinked pairs is to control for any inherent tendencies of the loci to be non-randomly associated. We developed this approach rather than using the more common Mantel test because we are not interested in demonstrating a trend in the relationship of LD with distance. Rather, we want to identify the point at which marker pairs show LD equivalent to that seen for unlinked pairs. As each locus is involved in many pairwise comparisons, they are obviously not independent and so our use of χ^2 is not intended as a rigorous statistical treatment, rather as an exploratory analysis.

The Lemba show elevated LD at very large genetic distances compared with unlinked markers: there is a significant excess of pairs in LD out to the 19-24 cM interval (Fig. 2). This contrasts sharply with the Bantu and Ashkenazim in which excess LD extends only to the 1-6 cM window and the Ethiopians who show only weak LD in the 0-5 cM window. Using LDp<0.01 as a significance threshold does not alter the results. After correction for multiple comparisons, P values for 17 of the 20 intervals out to 19-24 cM are significant in the Lemba.

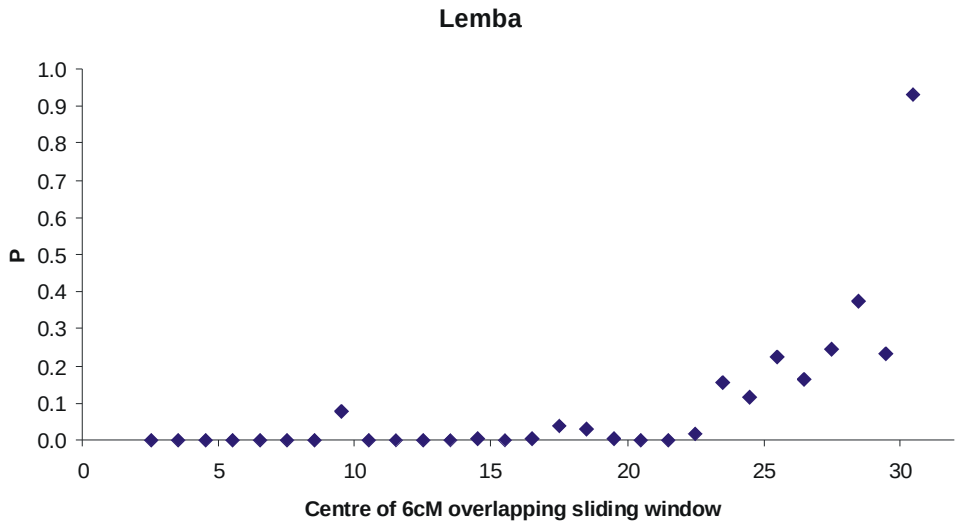
To better localise the point at which excess LD disappears in the Bantu and Ashkenazi Jews, we constructed contingency tables with 1 cM windows: only the 0, 1 and 2 cM tables are significant in each population (not shown). In the Ethiopians, none of the intervals contain a significant excess of LD. In the Lemba, excess LD extends to the 19-24 cM window, but it is not possible to use 1 cM windows due to the lower number of pairs in the 19-24 cM interval. However, with 2 cM windows, no tables beyond 21-22 cM are significant. LD is thus maintained across genetic distances an order of magnitude greater in the Bantu-Semitic hybrid population than in the other populations. It should be noted, however, that the comparison to unlinked

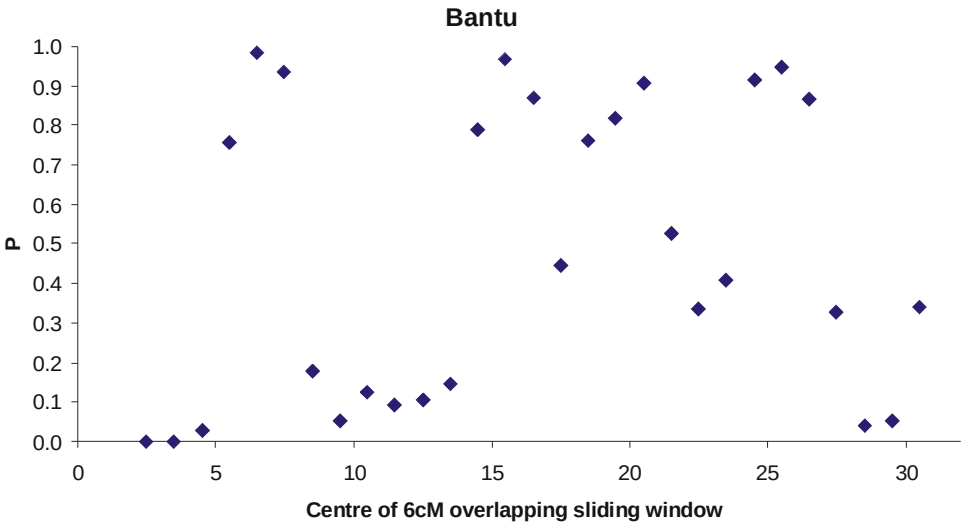
pairs demonstrates that the LD is not independent of genetic distance as would be observed in a substructured population sample.

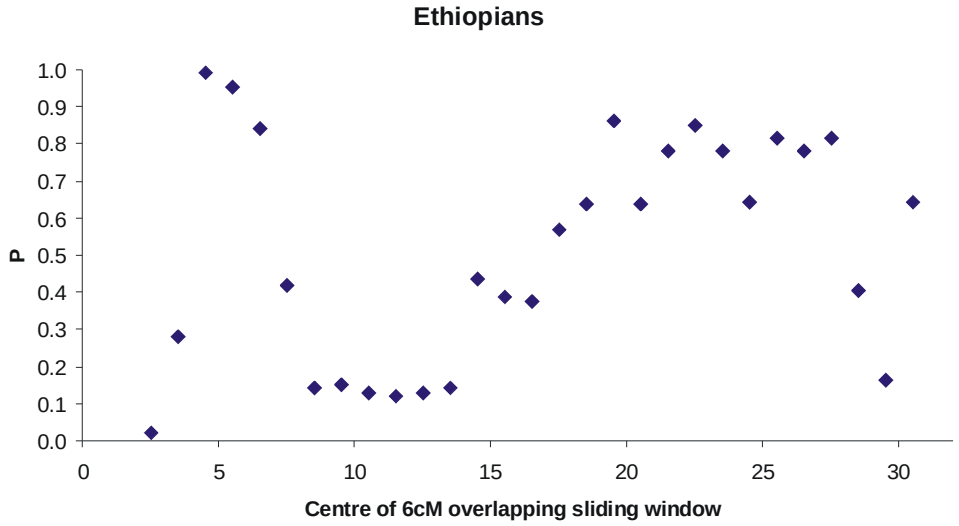
Beyond the extent of LD, the consistency of disequilibrium in the Lemba is also remarkable. In the 0-2 cM interval, in the Bantu and Ashkenazim, 46% and 38%, respectively of pairs are in LD, but in the Lemba one and a half times as many pairs – 63% – show significant LD in this interval (Fig. 3). The proportion of pairs in LD rapidly decreases with increasing genetic distance in all populations except the Lemba. For markers between 3-23 cM, 20% of pairs show significant LD in the Lemba, in contrast to only 4-7% in the other three populations. Thus, the Lemba have increased LD at most genetic distances, but LD is increased only in the 0-2 cM interval in the Bantu and Ashkenazi Jews. 38% and 25%, respectively of marker pairs separated by exactly 2 cM show LD in these populations contrasting with 0% in the Ethiopians. Fewer marker pairs were genotyped in the Ethiopians, however only 18% of pairs in the 0-2 cM interval show LD. Very little allelic association is observed at any genetic distance in the Ethiopians.

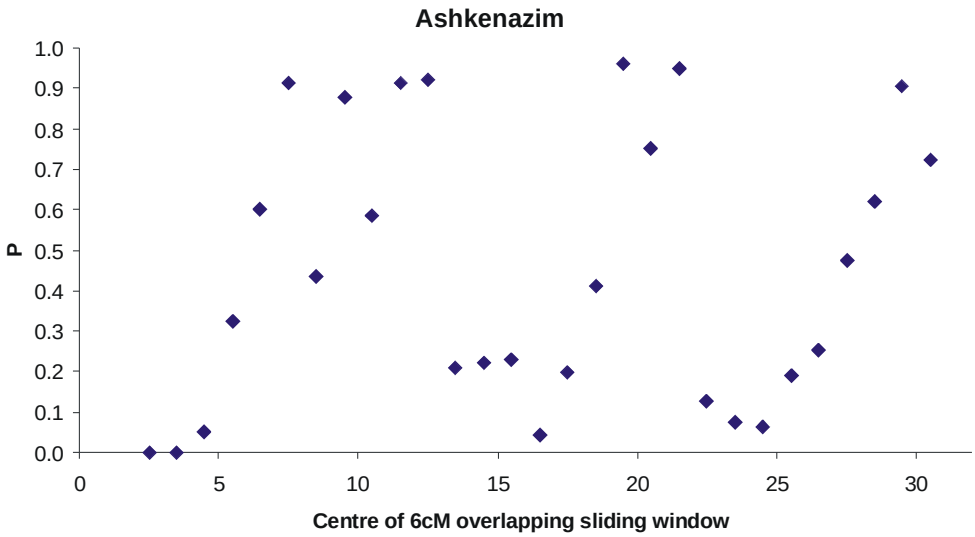
Even for unlinked markers, the Lemba show an increased level of LD: almost twice as many pairs are in disequilibrium (10%) compared to the other three populations (6%), which could reflect differences due either to recent admixture, drift or substructure. As the extent of LD was compared with that between unlinked markers, it is therefore all the more remarkable that an excess extends out to around 21 cM. Furthermore, as the admixture is thought to be principally male-mediated, X-linked markers will show reduced LD compared to autosomal markers, although this is offset as the X chromosome has less opportunity to recombine and a lower effective population size.

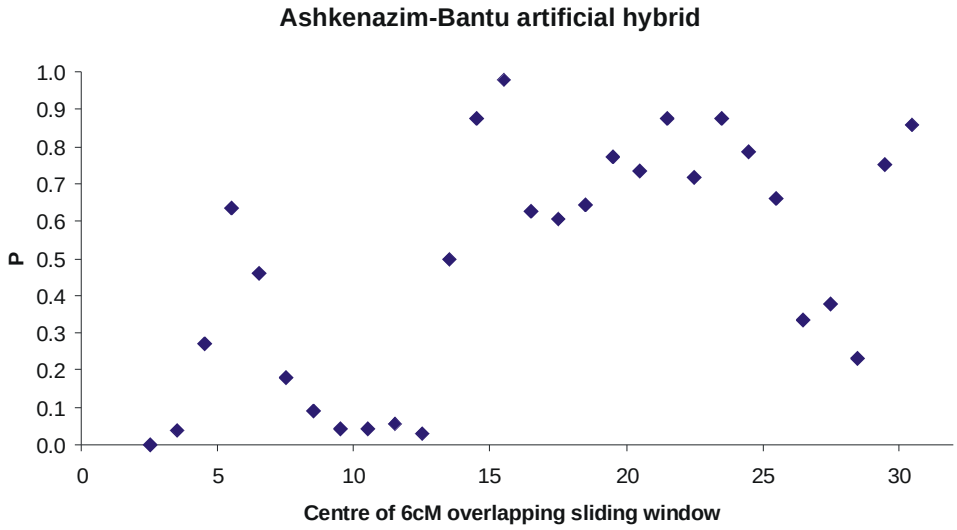
Figure 2 Extent of LD. Excess of significant LD is tested for by comparing pairs of markers within a given range of genetic distances to unlinked pairs. For each point in the graphs, a contingency table is evaluated in which the first column contains the numbers of significant and non-significant marker pairs with genetic distances in a window delimited by r-2.5 and r+2.5, where r is plotted on the X-axis. In all cases the second columns contain the number of significant and non-significant LDp values for all unlinked marker pairs.


Lemba
P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
Centre of 6cM overlapping sliding window


Bantu
P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
Centre of 6cM overlapping sliding window


Ethiopians
P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
Centre of 6cM overlapping sliding window


Ashkenazim
P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
Centre of 6cM overlapping sliding window


Ashkenazim-Bantu artificial hybrid
P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
Centre of 6cM overlapping sliding window

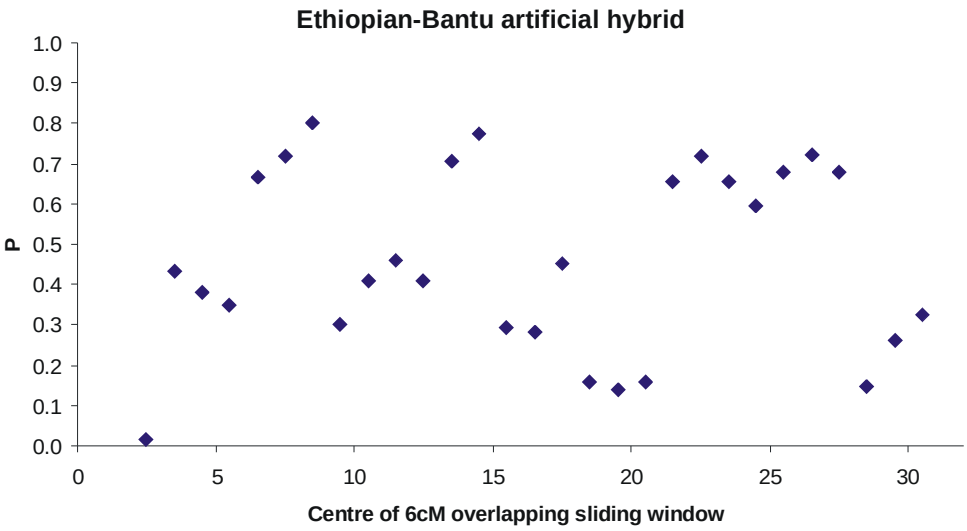
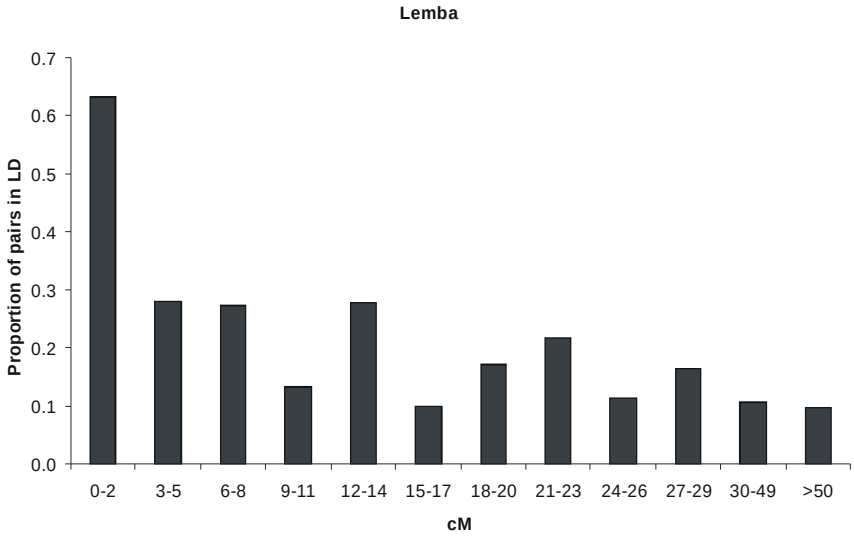
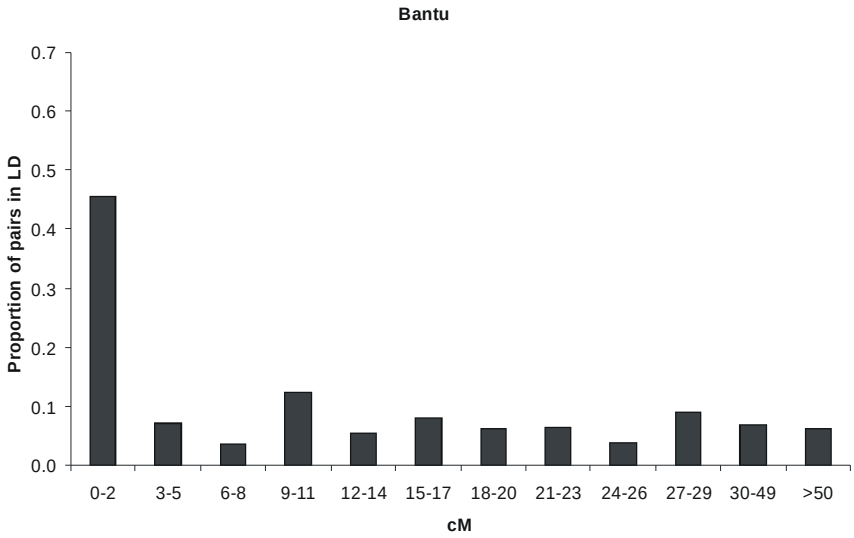
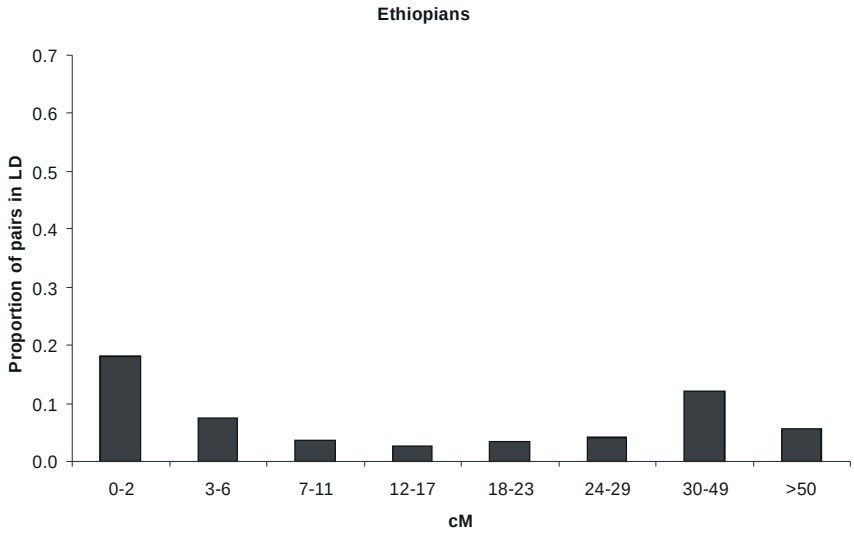
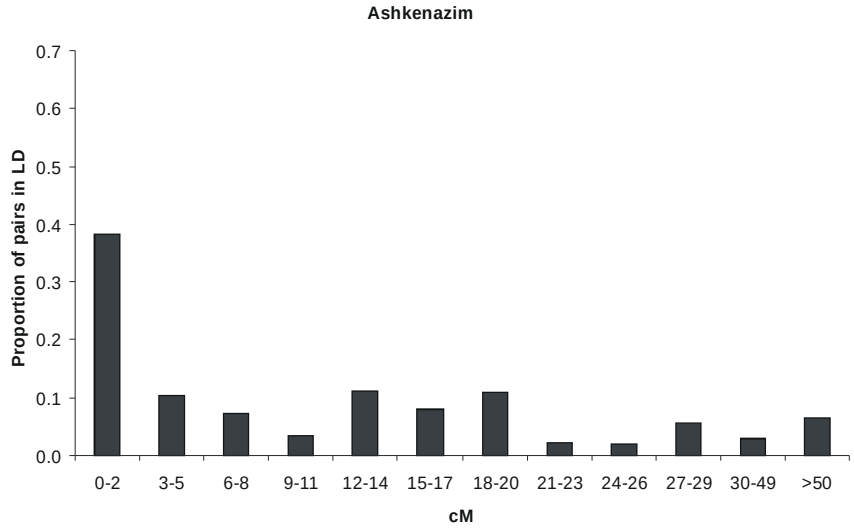

Ethiopian-Bantu artificial hybrid
P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
Centre of 6cM overlapping sliding window

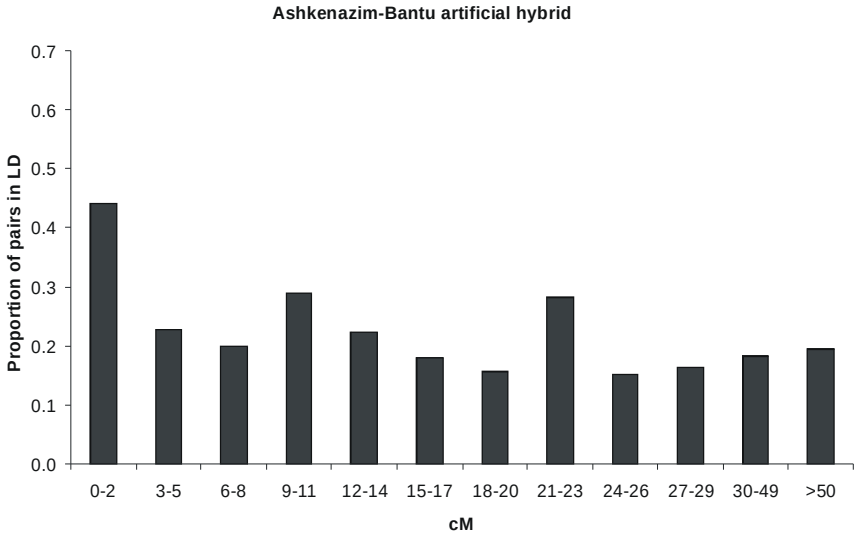
Figure 3 Consistency of LD at different genetic distances. Histograms of the proportion of locus pairs with LDp<0.05 in different distance classes. Note the distance classes are wider in the Ethiopians as fewer markers were genotyped.

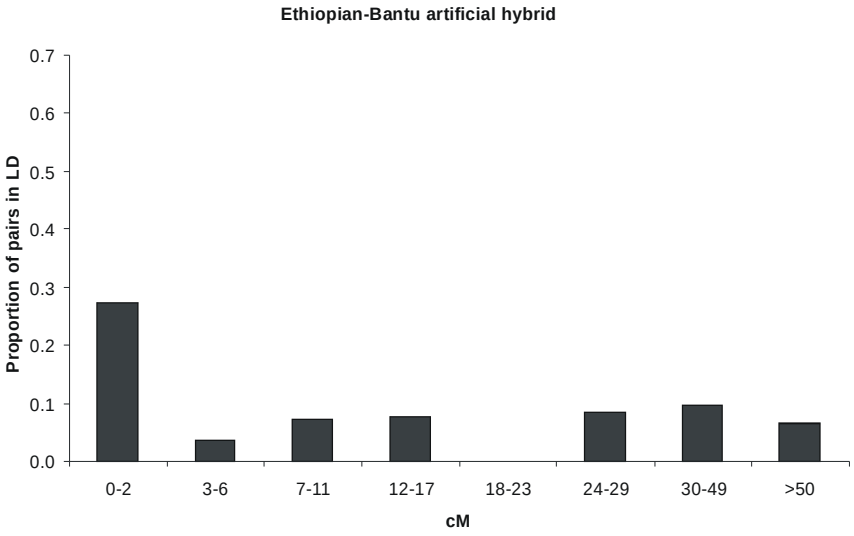

Lemba
Proportion of pairs in LD
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0-2
3-5
6-8
9-11
12-14
15-17
18-20
21-23
24-26
27-29
30-49
>50
cM


Bantu
Proportion of pairs in LD
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0-2
3-5
6-8
9-11
12-14
15-17
18-20
21-23
24-26
27-29
30-49
>50
cM


Ethiopians
Proportion of pairs in LD
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0-2
3-6
7-11
12-17
18-23
24-29
30-49
>50
cM


Ashkenazim
Proportion of pairs in LD
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0-2
3-5
6-8
9-11
12-14
15-17
18-20
21-23
24-26
27-29
30-49
>50
cM


Ashkenazim-Bantu artificial hybrid
Proportion of pairs in LD
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0-2
3-5
6-8
9-11
12-14
15-17
18-20
21-23
24-26
27-29
30-49
>50
cM

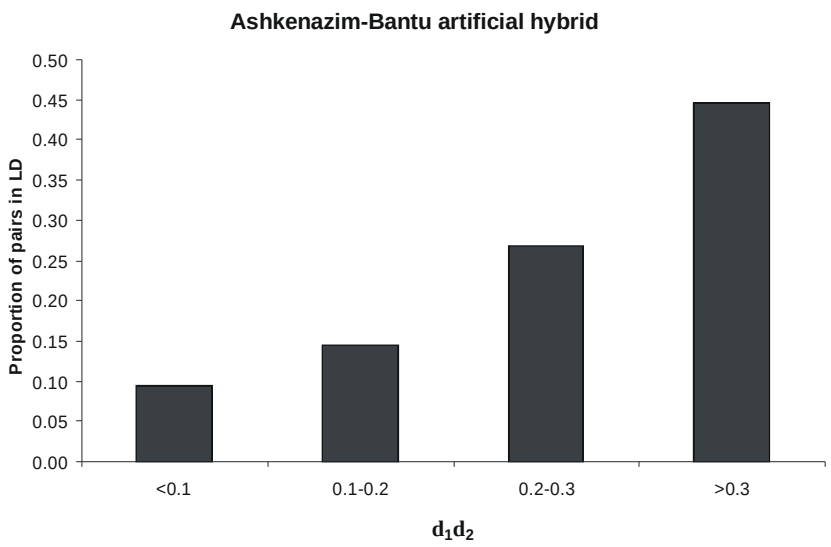

Ethiopian-Bantu artificial hybrid
Proportion of pairs in LD
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0-2
3-6
7-11
12-17
18-23
24-29
30-49
>50
cM

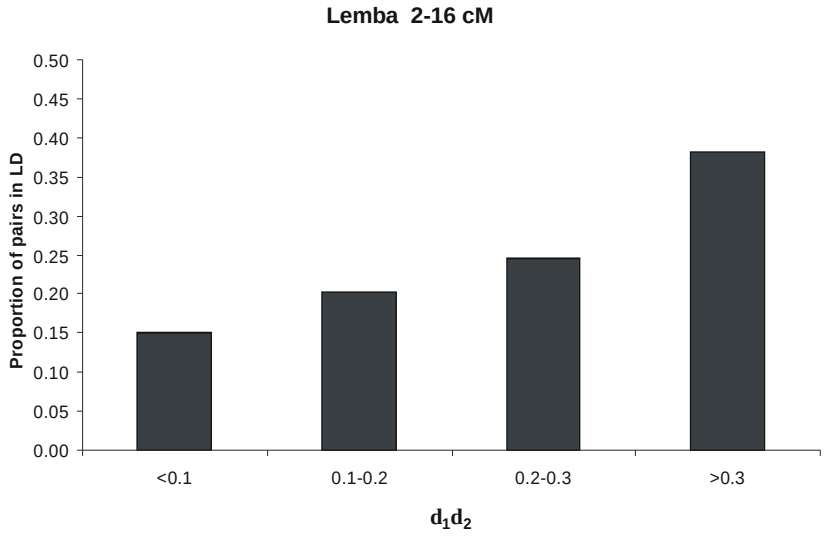
Is the Lemba LD generated by admixture? To the extent that the Lemba LD has resulted from admixture between Semitic and Bantu peoples, the differentials in allele frequencies (δ) between the parental populations should be predictive of the LD observed (Stephens et al. 1994). In particular, in the case of two alleles at each locus, at a given genetic distance, the product of the differentials at two loci ($\delta_1\delta_2$) should be linearly related to the LD (Briscoe et al. 1994; Chakraborty and Weiss 1988). We used the composite delta for multi-allelic systems suggested by Shriver *et al.* (1997) (see Methods). The predicted relationship between delta and LD provides a framework both to investigate whether LD is admixture-generated and for the identification of populations that have given rise to hybrid groups. We used two combinations of putative parental populations for the Lemba: Ashkenazi-Bantu and Ethiopian-Bantu.

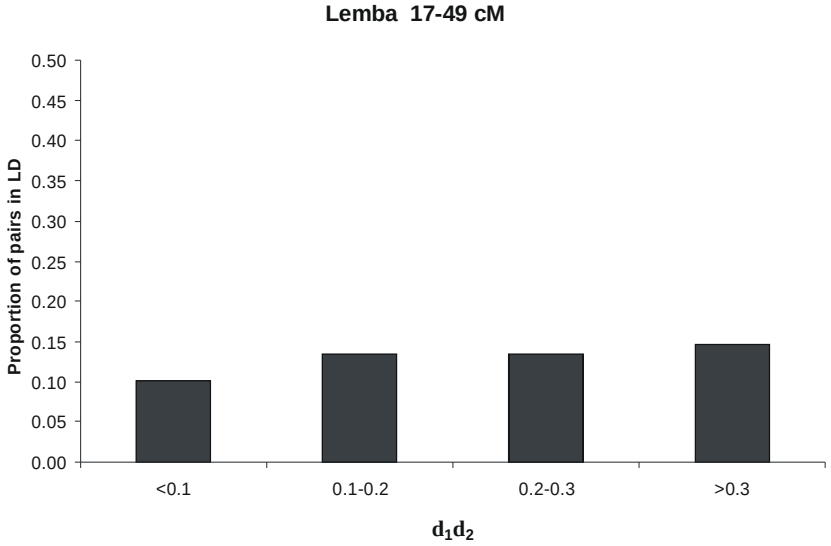
In relating δ values to LD, we exclude marker pairs in the 0-1 cM interval as these show considerable LD in the parental populations, some of which will have been transmitted to the Lemba, obscuring the signal of admixture-generated LD. In fact, 66% of significant Lemba pairs between 0-1 cM are also significant in the Bantu or Ashkenazim compared to only 21% between 2-16 cM. Using the Ashkenazi-Bantu δ values, the proportion of significant Lemba pairs between 2-16 cM increases in higher delta product classes (Fig. 4). 38% of pairs with a delta product above 0.3 are in LD but only 15% of pairs with $\delta_1\delta_2$ below 0.1. This effect is not observed at distances over 17 cM where the proportion of significant pairs in different $\delta_1\delta_2$ classes varies from 10-15%. To assess the significance of this difference, we built a contingency table by counting the number of significant ($LDp<0.05$) and non-significant marker pairs with delta products above and below the mean (0.17). Between 2-16 cM, significantly more pairs in the high $\delta_1\delta_2$ category show LD (P =

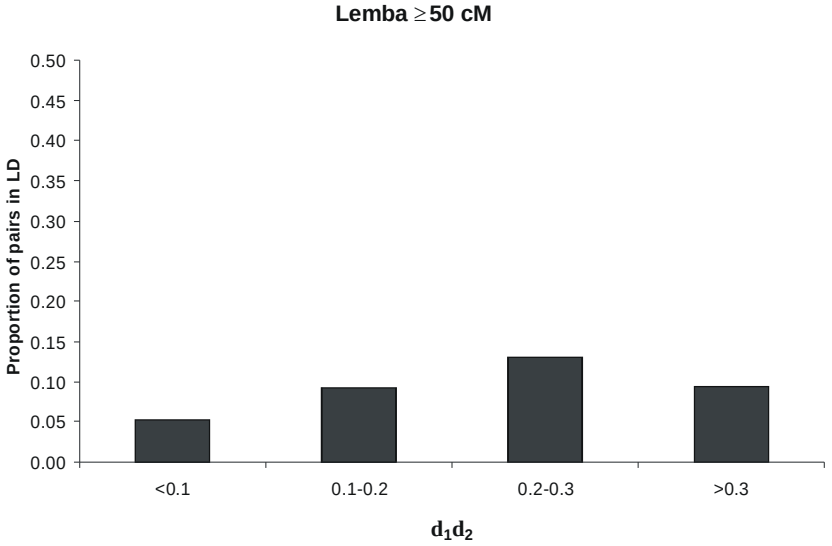
0.04). For unlinked pairs, this contingency table is not significant (P = 0.2): there is little effect of $\delta_1\delta_2$ on the probability of showing significant LD, even though both the delta products and LD are influenced by allele frequencies. When the 2-16 cM contingency table is evaluated using the Ethiopians as the second parental population there is no significant interaction between $\delta_1\delta_2$ and LDp using either the Ashkenazi-Bantu threshold of 0.17, or the Bantu-Ethiopian mean $\delta_1\delta_2$ value. We also tested whether the degree of association between LDp and Ashkenazi-Bantu $\delta_1\delta_2$ is significantly different for linked and unlinked marker pairs using multi-way tables, however the three-way interaction is not significant (P = 0.3).

Figure 4 Proportion of pairs in LD in different Ashkenazi-Bantu delta product classes. Lemba pairs are divided into three genetic distance classes, however all pairs are shown for the Ashkenazi-Bantu artificial hybrid population.


Ashkenazim-Bantu artificial hybrid
Proportion of pairs in LD
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
<0.1
0.1-0.2
0.2-0.3
>0.3
d_1d_2


Lemba 2-16 cM
Proportion of pairs in LD
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
<0.1
0.1-0.2
0.2-0.3
>0.3
d_1d_2


Lemba 17-49 cM
Proportion of pairs in LD
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
<0.1
0.1-0.2
0.2-0.3
>0.3
d_1d_2


Lemba ≥50 cM
Proportion of pairs in LD
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
<0.1
0.1-0.2
0.2-0.3
>0.3
d_1d_2

Artificial hybrid populations. To better evaluate the predicted effect of Bantu-Semitic admixture on LD, we constructed an artificial hybrid population by randomly sampling 45 X chromosomes from each of two putative parental populations of the Lemba: the Ashkenazi Jews and the Bantu. In the resulting artificial hybrid population, 20.5% of all marker pairs are in LD – one and a half times as many as in the Lemba. As expected, the substructuring results in disequilibrium largely independent of genetic distance, with 19.6% of unlinked markers in LD (Fig. 3). Also as expected, a higher proportion of pairs are in LD in higher delta product classes regardless of the genetic distance: 45% of pairs with $\delta_1\delta_2>0.3$ are in LD but only 9% of pairs with $\delta_1\delta_2<0.1$ (Fig. 4). Significantly more pairs with above-average delta products show LD ($P<10^{-18}$). As it only takes three generations to reduce by half the LD at 20 cM (and a single generation for unlinked markers), comparison of the LD levels in the artificial hybrid and the Lemba populations suggests that at least some admixture must have occurred extremely recently in the latter, or that there is some substructure. The significant difference between partially linked and unlinked loci, however, rules out substructure as the sole source of the LD in the Lemba. In contrast, when an artificial Bantu-Ethiopian hybrid is created in the same manner, very little LD is generated (Fig. 3). Only 7.5% of pairs are in LD, comparable to that in the parental populations, indicating that Ethiopian-Bantu differentiation is not sufficient to produce the disequilibria observed in the Lemba. Moreover, there is little relationship between delta products and LD in this artificial hybrid (not shown). Combining the fact that Bantu-Ethiopian delta values do not predict Lemba LD with the general lack of differentiation between these putative parentals, we conclude that the Ethiopians are not a good representation of the non-Bantu parental group of the Lemba.

Ancestral LD remains detectable in a stratified population. Because substructure creates LD at all genetic distances, no overall tendency to disequilibrium above that between unlinked pairs is observed in the artificial hybrid populations (Fig. 2). Interestingly, however, excess LD can still be detected at the smallest genetic distances. This excess is only significant in the Ashkenazi-Bantu artificial population, in which 44% of pairs are in LD in the 0-2 cM interval in comparison with 20% for unlinked pairs (Figs. 2 and 3). This LD must have two sources: LD inherited from the parental populations (the LD useful in association studies) and that created *de novo* by admixture/sample stratification (spurious LD in mapping studies). However, using multiple pairs of markers in the 0-2 cM interval, this ancestral LD remains detectable over and above the stratification LD observed with Bantu-Semitic levels of sub-population differentiation. Furthermore, the proportion of pairs showing extreme LD ($LDp<10^{-4}$) is ten-fold higher for pairs within 0-1 cM (17%) compared to unlinked pairs (1.7%). Case-control association studies may thus overcome spurious signals of association with designs using multiple marker pairs to assess the magnitude of LD generated by stratification and correcting association statistics accordingly (Pritchard et al. 2000; Reich and Goldstein 2001).

Variance and LD. To investigate the relationship between the variability of a locus and LD, we evaluated contingency tables comparing significant and non-significant pairs with above and below the mean repeat count variance (6.2). In the Lemba between 0-16 cM, there is a significant positive interaction between the mean variance and LDp ($P = 0.005$). This may be because high variance loci have greater power to show significant LD or could reflect the positive relationship between the

mean repeat count variance in the parental populations and the delta values ($P < 10^{-10}$). No such relationship between variance and LD is seen in the Bantu, Ethiopians and Ashkenazi Jews in this interval. In fact in the Jews and Ethiopians, for tightly linked markers (0-3 cM), the opposite trend is observed: low variance locus pairs have a tendency to show more significant LD. In this case the LD is probably older and mutation has contributed to the reduction of LD at the high variance loci.

Discussion

Earlier work has claimed to have identified admixture generated LD, most recently at the Duffy locus in African Americans (Lautenberger et al. 2000; Parra et al. 1998). The observation of a significant excess of haplotypes carrying together the alleles commonly found in European populations suggests admixture has indeed contributed to the observed LD. However, another recent study of the Duffy locus in the parental African population revealed strong evidence of selection (Hamblin and Di Rienzo 2000) which has probably generated LD around the locus. An unknown amount of the LD observed in African Americans may be such ancestral LD inherited from the African parental population. In order to provide conclusive evidence of the effect of demography on LD, it is critical to analyse multiple genomic regions (Freimer et al. 1997). Our analysis demonstrates the presence of a significant excess of LD between markers up to 21 cM apart compared to that between unlinked markers in the admixed Lemba. Moreover, analysis of allele frequency differentials between the Bantu and Ashkenazi Jews shows a significant predictive effect on the LD observed in the Lemba, demonstrating that the LD was generated by Bantu-Semitic admixture. The elevated LD at unlinked markers suggests some of the Lemba LD may be due to very

recent admixture, but as there is no relationship with Ashkenazi-Bantu allele frequency differentials, this admixture possibly involved a different parental population. In any case the difference between linked and unlinked intervals (compared with that in our artificial hybrid populations) rules out structure as the only source of excess LD.

Significant excess LD in comparison with unlinked markers is observed in the Ashkenazi Jews out to 2 cM, as may be expected from admixture with European populations and possible founder effects (although improbably tight bottlenecks may be required to generate this level of LD (Kruglyak 1999)). The presence of excess LD to 2 cM in the Southern Bantu contrasts with data indicating that African populations show less LD than non-Africans at the *CD4*, *PAH* and *DM* loci (Kidd et al. 2000; Tishkoff et al. 1996; Tishkoff et al. 1998). The observed LD may be due to a population-specific event such as an older admixture event as there is both genetic and linguistic evidence of admixture in Southern Africa between the indigenous Khoisan peoples and the incoming Bantu (Cavalli-Sforza et al. 1994). Moreover, the admixture likely occurred in the last 20 generations, which would be recent enough for considerable LD to persist today below 2 cM. In contrast to these populations, Ethiopians have a much lower amount of LD, emphasising the powerful influence of unknown aspects of a population's demographic history on the pattern of LD.

Under a set of simplifying assumptions concerning demographic history, it was predicted that for a case-control study of typical sample size, useful LD would be unlikely to extend beyond 3 kb (Kruglyak 1999). Contrary to this expectation, we have observed significant LD in 25-38% of marker pairs separated by 2 cM (~2000 kb on average) in the Bantu and Ashkenazi Jews, two populations with little evidence of extreme demographic histories. It should be noted that our sample sizes (between 77

and 96) correspond to only moderately sized case control studies unless the number of individuals required to detect association between a causal variant and the trait it influences is large (cf. Kruglyak 1999). The discrepancy between expectation and observation suggests that very few, if any real populations will match the assumptions and therefore predictions of such a simplified demographic history. In the case of the Lemba, a population with evidence of a more extreme demographic history, we observe LD across genetic distances more than three orders of magnitude greater than that predicted from simple demographic assumptions. Despite the considerable LD generated by admixture, however, analysis of an artificial hybrid shows that the ancestral LD between tightly linked markers is detectable over and above the spurious associations created by stratification. The profound effect of demographic history on background LD we document here indicates that it will not generally be possible to predict patterns of LD *a priori*, but will rather be necessary to empirically evaluate the patterns in all populations of interest. The observed patterns of LD do, however, present an important (Tishkoff et al. 1996; Tishkoff et al. 1998) and largely untapped source of information regarding human evolutionary history.

Chapter 4

Population genetic structure of variable drug response

Geographic patterns of genetic variation, including variation at drug metabolising enzyme (DME) loci and drug targets, suggest that geographic structuring of inter-individual variation in drug response may be common. This raises the questions of how to represent human population genetic structure in the evaluation of drug safety and efficacy, and of how to relate this structure to drug response. Here we address these questions by (a) inferring the genetic structure present in a heterogeneous sample, and (b) comparing the distribution of DME variants across the inferred genetic clusters of individuals. We find that commonly used ethnic labels are both insufficient and inaccurate representations of the inferred genetic clusters, and that drug metabolising profiles, defined by the distribution of DME variants, differ significantly between the clusters. We note, however, that the complexity of human demographic history means that there is no obvious natural clustering scheme, nor obvious appropriate degree of resolution. Our comparison of drug metabolising profiles across the inferred clusters, however, establishes a framework for assessing the appropriate level of resolution in relating genetic structure to drug response.

Introduction

Many drugs that show therapeutic potential never reach the market because of adverse reactions in some individuals, while other drugs in common use are only effective in a fraction of the population in which they are prescribed. This variation in drug response depends on many factors such as sex, age, and the environment as well as genetic determinants. Since the 1950's pharmacogenetic studies have systematically identified allelic variants at genes that influence drug response, including both drug metabolising enzymes (DMEs) (Weber 1997) and drug targets (Evans and Relling 1999), for example the cytochrome P450 monooxygenase *CYP2D6* (Gough et al. 1990; Kagimoto et al. 1990) and in the N-acetyl transferase *NAT2* (Blum et al. 1991) genes. Detailed functional analysis of variants at genes such as these has clearly demonstrated the importance of genetic variation in drug responses. For example analysis of mutant NAT2 alleles revealed reduction in enzyme activity half life in one case and defective translation leading to reduced enzyme protein in another (Blum et al. 1991), while common *CYP2D6* variants include a frameshift leading to a truncated non-functional protein and a splice site mutation resulting in the absence of the protein (Gough et al. 1990; Kagimoto et al. 1990). These and other examples suggest the possibility of genetic tests to predict an individual's response to specific drugs, ultimately allowing medicines to be tailored to specific genetic make ups. Because of the potential commercial and clinical significance of such personalised medicines, understanding the genetic role of variable drug response is one of the primary challenges of biomedical research.

In addition to concerns surrounding individual variation in drug response, however, the geographic structuring of certain variants has focussed attention on the

possible importance of average differences in drug response across populations. Genetic polymorphisms in DMEs, which are likely to be responsible for much of the phenotypic variation in drug response, all vary in frequency among populations (Evans and Relling 1999), some by as much as twelve-fold (Weber 1997). For example, the well-known poor metaboliser phenotype of debrisoquine oxidation is due to variant alleles of *CYP2D6*. Between 5% and 10% of Europeans (but only ~1% of Japanese) have loss of function variants at this locus which affect the metabolism of more than 40 drugs, including commonly used agents such as β-blockers, codeine and tricyclic antidepressants. *CYP2D6* ultra-rapid metaboliser alleles (*i.e.* duplications) also vary in frequency, even within Europe, from ~10% in Northern Spain to 1-2% in Sweden (Bernal et al. 1999). These polymorphisms can lead to acute toxic responses, unwanted drug-drug interactions and also therapeutic failure (e.g. in the case of *CYP2D6* duplications) (Meyer and Zanger 1997; Weber 1997).

These observations make clear that for some drugs the tradeoffs between efficacy and adverse drug reaction will differ not only between individuals but will show average differences in different populations (ICH 1998). Thus genetically structured populations may be composed of two or more sub-populations with distinct drug reaction profiles, and thus may be better considered separately in some contexts. This raises the question of the appropriate way to infer human population genetic structure in the context of the evaluation of drug safety and efficacy, and of how to relate this inferred genetic structure to drug response. In order to address this problem we have used presumably neutral microsatellite markers to infer genetic clusters for a heterogeneous population, as for example may be used in large-scale drug trials of sufficient size to allow detection of both genetic and environmental effects (e.g. Phase III trials). We then compared the frequencies of functionally significant alleles at

DME loci across the inferred clusters as an easily defined surrogate for drug response. Using this approach we (a) demonstrate that there is considerable scope for population-genetic structuring in drug response in diverse metropolitan populations, due to variation among such clusters in DME allele frequencies; (b) establish a framework for determining the appropriate level of resolution (i.e. the number of inferred clusters that should be used) in relating this population-genetic structuring to drug response; and (c) demonstrate that commonly used ethnic labels (e.g. Black, Caucasian and Asian/other) are insufficient and inaccurate descriptions of human genetic structure.

Materials and Methods

All subjects were unrelated males. The following X-linked microsatellites were genotyped (Ch. 2): *DXS984*, *996*, *1036*, *1053*, *1062*, *1203*, *1204*, *1205*, *1206*, *1211*, *1212*, *1220*, *1223*, *7103*, *8014*, *8061*, *8068*, *8073*, *8085*, *8086*, *8087* and *8099*. The following chromosome one microsatellites were genotyped: *D1S196*, *206*, *213*, *249*, *255*, *450*, *484*, *2667*, *2726*, *2785*, *2797*, *2800*, *2836*, *2842*, *2878* and *2890*. All form part of the ABI Prism linkage mapping panel 1 and were amplified according to manufacturers instructions. Individuals were assigned into clusters by the program STRUCTURE (Pritchard et al. 2000) using the admixture model, with no correlation in allele frequencies among populations and a burn in time of at least 1 million steps, followed by another million steps of the Markov Chain for data collection. Multiple runs were carried out for each set of conditions to be sure that the chain had converged and in total more than 500 runs were performed. The intronic C734A transversion in *CYP1A2* was genotyped by sequencing as were two SNPs in *NAT2*: C481T defining allele *5 (in complete allelic association with Ile113Thr) and G590A

(giving Arg197Gln) defining allele *6. All other alleles were classed as *4 while both mutant allele frequencies were combined for binary analyses. The deletion allele in glutathione-S-transferase mu 1 (*GSTM1*) was genotyped using GSTM4 amplification as an internal control (Krajinovic et al. 1999). The C191T transition (giving Pro187Ser) in *DIA4* (Gaedigk et al. 1998) and the G117A (leading to a truncated protein) transition in *CYP2C19* (Goldstein and Blaisdell 1996) were genotyped by PCR-RFLP (Table 1). *GSTM1* and RFLP amplicons were fluorescently labelled and sizes determined on an ABI 3100 automated sequencer (Applied Biosystems). *CYP2D6* SNPs were typed by allele-specific PCR followed by nested reamplification-RFLP detection of the following 'key' mutations (Gaedigk et al. 1999): C100T (Pro34Ser) (alleles *10 and *4), G1846A (splicing defect) (allele *4), A2549del (frameshift) (allele *3), 2613-2615AGAdel (Lys281del) (allele *9) and C2850T (Arg296Cys) (allele *2). All other chromosomes were denoted *1 although this will include some non-wild type alleles. *CYP2D6**1 was considered to have normal activity, and all other alleles were treated as reduced activity for the binary analyses. *CYP2D6* amplicons were labelled using fluorescent primers, then pooled and sized on an ABI 377 automated sequencer (Applied Biosystems) (genotyping details were personally communicated by Ben Fletcher). In the case of *GSTM1* the assay does not allow the differentiation between homozygous and heterozygous presence of the non-deletion allele. In this case calculations were performed on genotype frequencies, homozygous deletion versus homozygous or heterozygous for the non-deletion allele. DME allele frequencies in the clusters were calculated by distributing an individual's genotype among the clusters according to the proportion of ancestry STRUCTURE determined the individual had in each cluster. When individuals were forced into the cluster in which they had the most ancestry, the results changed very little (not

shown). In order to meet the assumption of a multinomial distribution, χ^2 tables were evaluated after placing individuals into the clusters in which they had most ancestry. We estimated the accuracy of our genotyping by retesting a number of samples from each population. Error rates varied from 0%-7% for the DME polymorphisms and from 0% to 5% for the microsatellites.

Table 1. Details of drug metabolising enzyme polymorphism assays. T is the primer annealing temperature and enz, the enzyme used in the digest.

Gene	Polymorphism	Assay Reference	Type	Primer 1	Primer 2	Sizes	T	Enz
CYP1A2	C734A	Wilson et al. 2001	Sequencing	ctactccagccccagaagtg	gaagggaacagactgggaca	318	57	-
NAT2	C481T	Wilson et al. 2001	Sequencing	cagatgtggcagcctctagaat	ggaacaaaatgatgtggttataaatg	285	60	-
NAT2	G590A	ditto	ditto	ditto	ditto	ditto	ditto	ditto
GSTM1	gross deletion	Krajinovic et al. 1999	PCR	cgccatcttgtgctacattgcccg	atcttctcctcttctgtctc	1=230,157; 0=157	58	-
				ttctggattgtagcagatca	amplifies GSTM4 control fragment			
CYP2C19	G117A	Goldstein et al. 1996	PCR-RFLP	cagagcttggcatattgtatc	gtaaacacacaactagtcaatg	G=109; A=321	53	*Sma*I
DIA4	C191T	Gaedigk et al. 1998	PCR-RFLP	cctgaggcctccttatcag	caaagaggctgcttggagc	C=283; T=151	60	*Hin*fI

Results and Discussion

We genotyped 16 chromosome one microsatellites from the ABI prism panel 1 (an average of 17 centimorgans (cM) apart) and 23 X-linked microsatellites (≥2 cM apart) (Ch. 2) in each of eight populations: South African Bantu-speakers (46), Amharic and Oromo-speaking Ethiopians from Shewa and Wollo provinces collected in Addis Ababa (48), Ashkenazi Jews (48), Armenians (48), Norwegian-speakers from Oslo (47), Chinese from Sichuan in SW China (39), Papua New Guineans from Madang (48) and Afro-Caribbeans collected in London (30).

Genetic structure. A model-based clustering method implemented by the program STRUCTURE (Pritchard et al. 2000) was used to assign individuals to sub-clusters on the basis of these genetic data, ignoring their actual population affiliations. This mimics a scenario in which there is cryptic population structure, or no information as to the ethnic origin of the individuals. Briefly, the model implemented in STRUCTURE assumes K clusters, each characterised by a set of allele frequencies at each locus, the admixture model then estimates the proportion of each individual's genome having ancestry in each cluster. We estimated $\Pr(X|K)$, where X represents the data, using a model allowing admixture, for K between 1 and 6. From this and a uniform prior on K between 1 and 6, we estimated, $\Pr(K|X)$ using Bayes' theorem (Pritchard et al. 2000) (Table 2). Virtually all of the posterior probability density is on $K = 4$.

Table 2. Inferring the number of clusters.

K	ln Pr(X|K)	Pr(K|X)
1	-33680.97	~0
2	-32650.80	~0
3	-32046.80	~0
4	-31943.23	1.000
5	-31972.33	~0
6	-31987.10	~0

Table 3. Proportion of membership of each sampled population in STRUCTURE-defined sub-clusters. PNG is Papua New Guinea.

Population	A	B	C	D
Bantu	0.04	0.02	0.93	0.02
Ashkenazim	0.96	0.01	0.01	0.02
Ethiopia	0.62	0.08	0.24	0.06
Norway	0.96	0.02	0.01	0.01
Armenia	0.90	0.04	0.02	0.05
China	0.09	0.05	0.01	0.84
PNG	0.02	0.95	0.01	0.02
Afro-Caribbean	0.21	0.03	0.73	0.03

The apportionment of individuals from each of the eight populations into the four STRUCTURE-defined clusters (Table 3) broadly corresponds to four geographical areas: Western Eurasia, Sub-Saharan Africa, China and New Guinea. Importantly 62% of the Ethiopians fall in the first cluster, which encompasses the majority of the Jews, Norwegians, and Armenians, demonstrating that placement of these individuals in a 'black' cluster would be an inaccurate reflection of the genetic structure. Only 24% of the Ethiopians are placed in the cluster with the Bantu and most of the Afro-Caribbeans, however 21% of the Afro-Caribbeans are placed in cluster A with the West Eurasians, no doubt reflecting genetic exchange with Europeans. Finally China and New Guinea are placed almost entirely in separate clusters, indicating that the ethnic label Asian is also an inaccurate description of population structure.

Consideration of only the X-linked microsatellites for the purposes of clustering supported $K = 3$, with a very similar clustering to that for the entire data set except that the Chinese and New Guinean clusters were combined into one. When only the chromosome one microsatellites were used the clustering is essentially the same as for the whole data set. This discrepancy may be explained by one of two factors: (i) a lack of resolution in the X chromosome microsatellites, or (ii) a biological factor such as the different number of X chromosomes and autosomes carried by males and females. In order to differentiate these hypotheses we carried out structure runs on the chromosome one data using an equal amount of information to that available from the X chromosome (22 alleles). The chromosome one microsatellites continued to support $K = 4$, indicating that a lack of resolution in the X chromosome microsatellites may not have been the explanation. Perhaps the facts that the X chromosome spends one sixth of the time longer in the female germline than chromosome one does and that females are known to have a higher migration rate

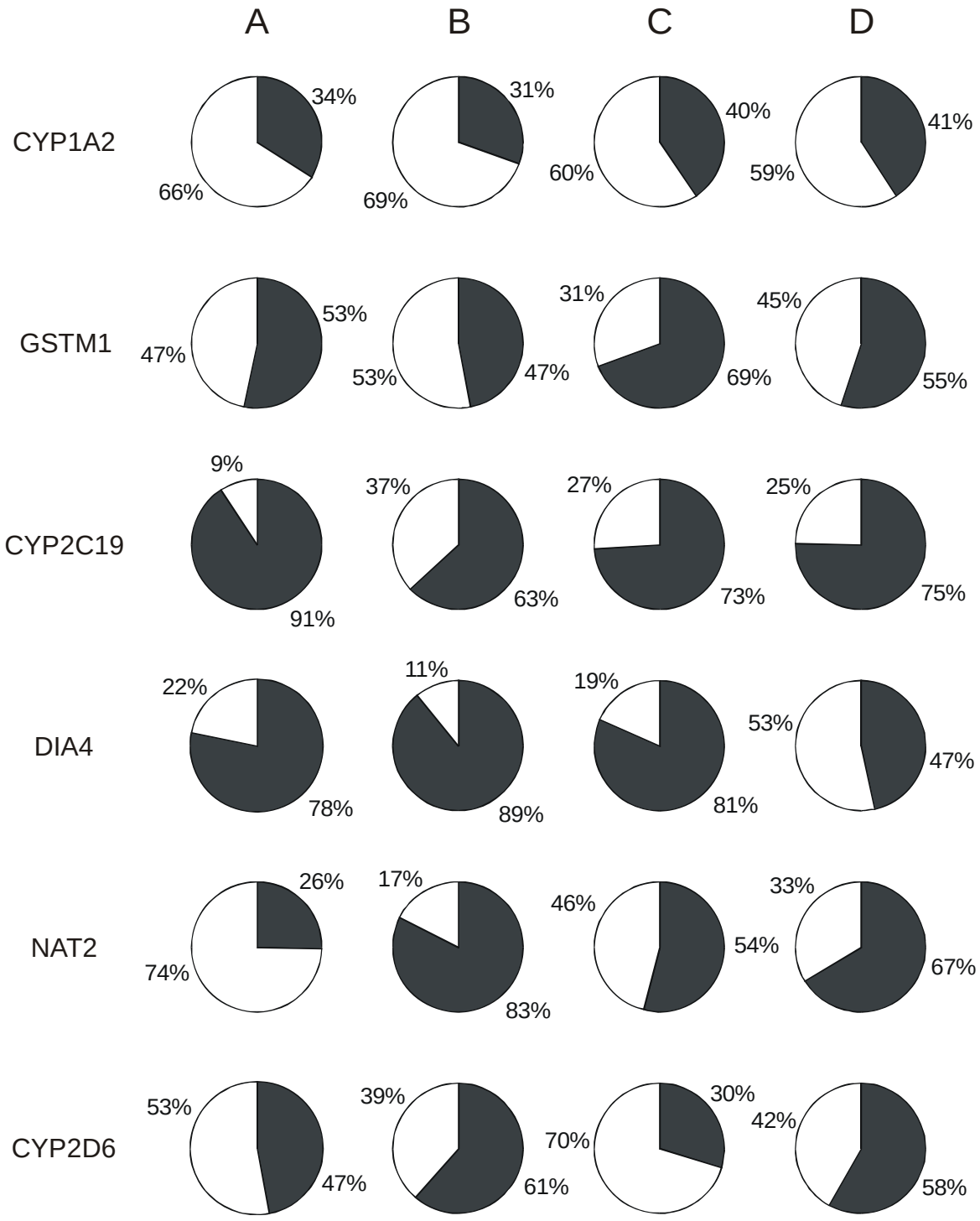
than males (Seielstad et al. 1998) have served to decrease the power of the X-linked loci to detect genetic structure. Smaller random subsets of the loci support a variety of values for *K* and do not agree on the clustering scheme (not shown). This is likely because there are no natural clusters as there has not been a history of bifurcation in human populations. Our results indicate that a reasonably high number of loci should be used in order that consistency in clustering is achieved. For example, one approach would be to use one marker from each chromosome arm. All analyses we present use the full data set, resulting in four clusters (Table 3).

Drug metabolising enzymes. Our selection of DMEs includes representatives of both phase I (oxidation or reduction) and phase II (conjugation) drug metabolism. Three enzymes of the phase I cytochrome P450 family were included: *CYP1A2*, *CYP2C19* and *CYP2D6*. Three conjugating or phase II metabolism enzymes were also included: *NAT2*, NAD(P):quinone oxidoreductase (*DIA4*) and glutathione-S-transferase mu 1 (*GSTM1*). We determined allele frequencies at eleven variants in these six DMEs, all of which are known to be functionally significant (Fig. 1).

Fig. 1

Allele frequencies at each of the DME variants in the STRUCTURE-defined clusters. In all but the last two, black is wild type and white is mutant, for *CYP2D6* all mutant alleles are pooled as white, and for *NAT2* both tested mutant alleles (*5 and *6) are pooled as white. Several drugs and carcinogens are metabolised by cytochrome P450 1A2 (*CYP1A2*), including the analgesic acetaminophen (Tylenol®). This enzyme is also thought to be involved in the metabolism of antipsychotic drugs (Basile et al. 2000). Polymorphism in *CYP2C19* is responsible for the classical mephenytoin poor metaboliser phenotype but diazepam, barbiturates and antidepressants are also

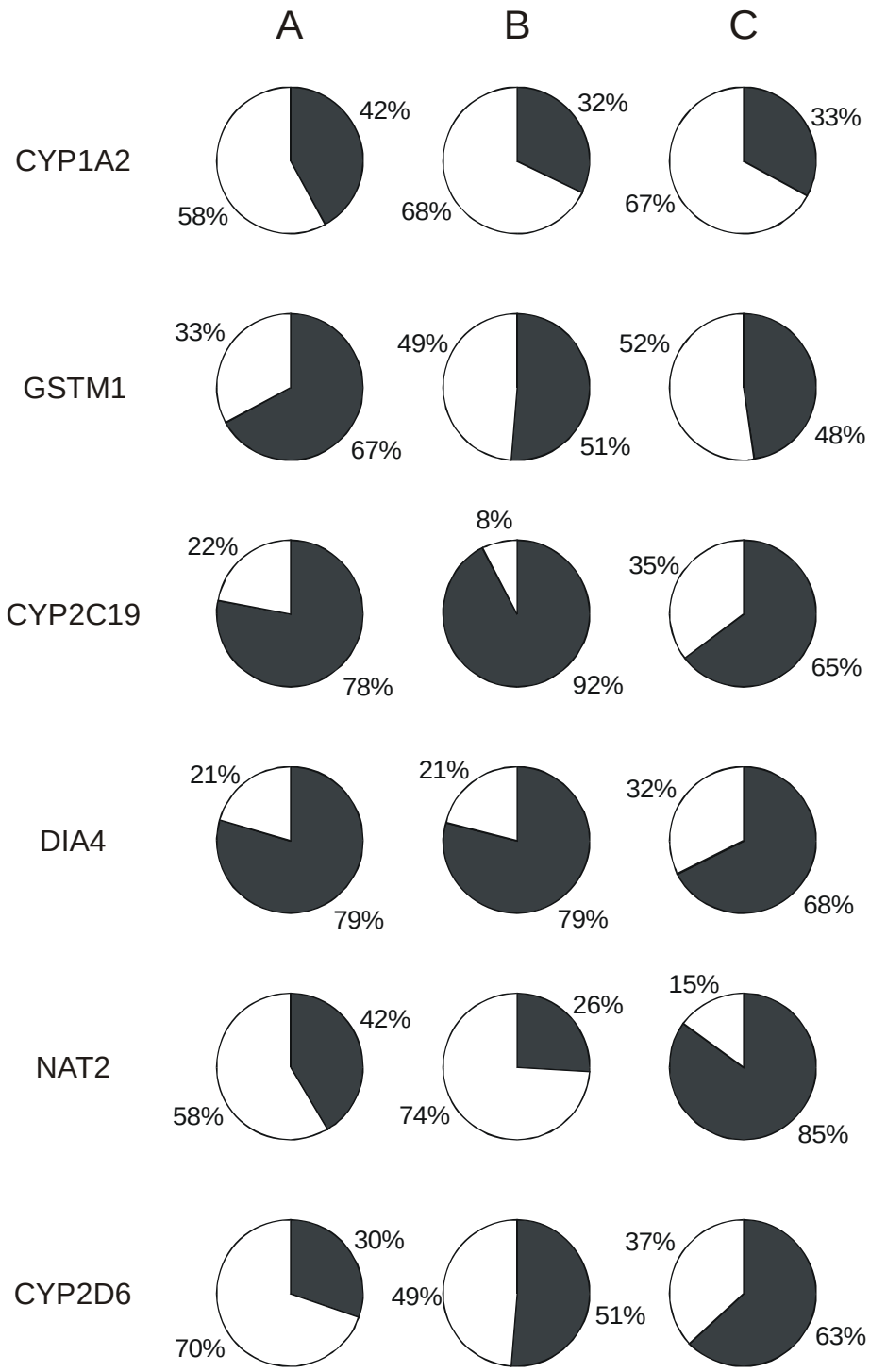
metabolised by this enzyme (Ferguson et al. 1998). The classical debrisoquine poor metaboliser phenotype is due to polymorphism in *CYP2D6* (Meyer and Zanger 1997). *NAT2* is responsible for the classical isoniazid polymorphism (Blum et al. 1991). Quinones are converted to stable hydroquinones by NAD(P):quinone oxidoreductase (*DIA4*) which also bioactivates antitumour quinones and nitrobenzenes (Gaedigk et al. 1998). Glutathione-S-transferase mu 1 (*GSTM1*) conjugates various electrophilic compounds including potent environmental carcinogens such as aflatoxin B_1 epoxides (Weber 1997). The two *NAT2* polymorphisms we genotyped both result in slow acetylator alleles which lead to increased risks of drug toxicity and of certain cancers (Blum et al. 1991; Weber 1997). Of the *CYP2D6* alleles we assayed: *CYP2D6**1 is wild type, *3 and *4 have no activity (which can lead to an acute toxic response to some drugs), and *2, *9 and *10 have reduced activity (Daly et al. 1996; Gaedigk et al. 1999). The *CYP1A2* variant genotyped leads to increased enzyme inducibility in smokers (Sachse et al. 1999). The major polymorphism in *CYP2C19* responsible for the mephenytoin poor metaboliser trait was genotyped. After the administration of various drugs this variant can lead to bone marrow toxicity, fatal blood disorders, and other adverse responses (Weber 1997). Increased susceptibilities to various cancers are associated with the deletion polymorphism in *GSTM1* genotyped here, dramatically so for smokers (Krajinovic et al. 1999; Weber 1997). The mutation in *DIA4* leads to a complete absence of the protein and thus loss of the protection against the toxic and carcinogenic effects of quinones (Gaedigk et al. 1998).


A
B
C
D
CYP1A2
34%
66%
31%
69%
40%
60%
41%
59%
GSTM1
53%
47%
47%
53%
31%
69%
45%
55%
CYP2C19
9%
91%
37%
63%
27%
73%
25%
75%
DIA4
22%
78%
11%
89%
19%
81%
53%
47%
NAT2
26%
74%
17%
83%
46%
54%
33%
67%
CYP2D6
53%
47%
39%
61%
30%
70%
42%
58%

There are notable DME allele frequency differences between the genetically identified clusters (Fig. 1) for five of six reported loci. To assess differentiation across clusters, we counted allele frequencies in each of the clusters and calculated χ^2; we also tested for differences in allele frequencies using logistic regression. Using both methods, the allele frequency distributions are highly significantly different for four of the six loci (significant for *NAT2, CYP2C19, DIA4* and *CYP2D6*). The pattern is particularly striking at *CYP2C19* where the frequency of the mutant allele (the mephenytoin polymorphism) in cluster B is more than four times that in cluster A (significant at $P < 0.0001$). Extreme differentiation is also evident between clusters B and D for *DIA4* where the frequency of the mutant allele (which provides no protection against the toxic effects of quinones) differs by almost five-fold ($P < 0.0001$). This is a noteworthy difference since clusters B and D would be combined as Asian in current drug evaluation using ethnic labels. *NAT2* also shows significant differentiation between these two clusters as well as among the others. Strong to modest differences in allele frequencies are observed for the other DMEs between at least two pairs of the clusters in each case. To further explore cluster differentiation we counted the number of loci for which there are significant allele frequency differences (using χ^2) for each of the pairs of clusters. Without correcting for multiple comparisons, this number varied from 2 out of 6 for B vs. D to 5 out of 6 for B vs. C. Given the important differences in drug response determined by these variants, the scope for genetic structuring in drug response is manifestly high. The trade off between therapeutic response and adverse drug reactions will differ between the different clusters identified here thus making it critical to perform this kind of genetic analysis to check for such effects in any phase III clinical trial.

Fig. 2

Allele frequencies at each of the DME variants in the ethnically labelled groups. See Fig. 1 legend for details. A presents Bantu, Ethiopian and Afro-Caribbean frequencies; B those for Norwegians, Ashkenazi Jews and Armenians; and C those for Chinese and New Guineans.


A
B
C
CYP1A2
42%
58%
32%
68%
33%
67%
GSTM1
33%
67%
49%
51%
52%
48%
CYP2C19
22%
78%
8%
92%
35%
65%
DIA4
21%
79%
21%
79%
32%
68%
NAT2
42%
58%
26%
74%
15%
85%
CYP2D6
30%
70%
49%
51%
37%
63%

We compared how informative the genetic clusters are versus commonly used ethnic labels by counting the DME allele frequencies in the grouping resulting from commonly used labels: Caucasian (Norwegian, Ashkenazi Jew, Armenian), Black (Bantu, Ethiopian, Afro-Caribbean) and Asian (Chinese, New Guinean) (Fig. 2). The case of *DIA4* is noteworthy. The large frequency difference between clusters B and D, driven by the differentiation between China and New Guinea, is averaged when both populations are lumped and so the mutant allele frequency is only one and a half times as high as that in the other two groups. Indeed, the overall differentiation for the ethnic groups is not significant after correction for multiple comparisons. Note that in no case did we observe the reverse in our data. That is, the ethnic labels never show sharp differentiation that is not observed in the clusters. Furthermore, only in the case of *CYP2D6* are the allele frequency differentials as high as they are for genetically defined clusters. Thus, although there is some DME allele frequency differentiation between the ethnically labelled groups, in most cases it is less than that seen for the genetic clusters. To demonstrate this formally, we fitted logistic regression models to the allele data using membership in the genetic clusters as the explanatory variables and tested for the increase in goodness of fit obtained by adding the ethnic labels as explanatory variables. We then compared this to the increase in goodness of fit obtained by adding the genetic cluster information to the ethnic group information. Of those DME loci that showed significant differentiation in either the clusters or the ethnic groups (i.e. *NAT2, CYP2C19, DIA4* and *CYP2D6*), in three out of four cases genetic cluster information to ethnic labels was more significant than adding ethnic labels to genetic clusters. Only for *CYP2D6* was the outcome the reverse.

Multilocus interactions: Undesirable drug reactions or interactions may also be due to the possession by an individual of variants at two (or more) loci, as appears to be the case in the increased susceptibility to colorectal cancer in individuals with a rapid/rapid metaboliser phenotype at *CYP1A2* and *NAT2*, greatly so for those who prefer well cooked meat (Kohlmeier et al. 1997). It is therefore important to consider not only allele frequency differences between the inferred clusters, but also frequency differences for multilocus genotypes. There are large frequency differentials between the clusters we have identified for multilocus genotypes which may give rise to phenotypic combinations like this, in fact the frequency of the combination *CYP1A2**A/A, *NAT2**4/- observed in cluster B (47%) is more than twice that seen in clusters A (19%) or C (22%; $P<0.01$ for overall differentiation). When such interactions are important they may be apparent using the genetic analysis described here from the distribution of drug response across inferred clusters.

By carrying out the clustering analysis with the number of clusters set to different values, we can compare the extent of differentiation among the clusters in order to assess the appropriate level of resolution. In the context of a Phase III trial the appropriate benchmark would be in terms of the amount of the total variation in drug response explained by the genetic clusters. A surrogate would be to carry out exact tests of differentiation (Raymond and Rousset 1995) on relevant functional polymorphisms, stopping when an increase in the number of clusters does not appreciably increase the degree of differentiation. We note, however, that the clustering properties of STRUCTURE can be unstable across different values of K and so implementation of such an analysis using STRUCTURE would not be straightforward.

It is well known that there are inter-ethnic differences in DME allele frequencies and thus in drug response. Our focus here, however, has been to assess the scope for average difference in drug response across genetically inferred clusters. Not only can these be derived in the absence of knowledge about ethnicity (or geographic origin), we also show the inferred clusters are more informative than commonly used ethnic labels. Because of the potential for average difference in drug response to have clinical significance, we conclude that it is not only feasible but a clinical priority to assess genetic structure as a routine part of drug evaluation.

When the most important genes influencing response to a particular drug or group of drugs has been defined, it should be possible to personalise medicine on the basis of an individual's genotype, assuming that routine individual genotyping is commercially and technically feasible. Short of such detailed knowledge, however, it is important to assess whether drugs work similarly in different genetic subgroups. The appropriate level of clustering may be evaluated empirically by assessing the amount of variation in response explained by the inferred clusters. Finally, we have shown that the common ethnic labels currently available to regulatory authorities show a poor correspondence with genetically inferred clusters.

Analysis of Population Structure in Biomedical Research. Our implementation of STRUCTURE is meant principally to demonstrate that the familiar ethnic labels are not accurate guides to genetic structure and not as a definitive description of the structure in human populations. In fact the results of STRUCTURE can be quite difficult to interpret. Notably, statistical difficulties may arise when assessing convergence and the assessment of the appropriate value of K is currently non-rigorous (Pritchard et al. 2000). These and other issues can lead to anomalous outcomes, for example, when an implausible value of K is supported in which one of

the clusters is more or less empty. Second, results may vary for biological reasons such as when markers are affected differentially by forces acting on the genome, such as gene flow. Detailed analysis of STRUCTURE output and other clustering schemes should thus be explored using a standard battery of markers in a global sample of human populations in order to arrive at a canonical clustering scheme for use in biomedical research. Such an evaluation would need to be geographically exhaustive, and would need to include a sufficient number of markers throughout the genome to ensure that resulting clustering scheme is robust in the sense that similar results would be obtained with different marker and sample sets.

Chapter 5

Discussion

My thesis aimed to illustrate the academic and applied uses of human genetic variation. Each of the chapters focused on a different aspect of human variation, but all three reflect how the variation we observe has been shaped by population history into non-random patterns of association or structuring. The last two chapters demonstrate that the consequences of these patterns are crucial to biomedical applications from predicting variable drug reaction to mapping the genetic determinants of common diseases such as hypertension and asthma.

The second chapter used high-resolution genealogical systems as well as multilocus systems to unravel parts of the history of the British Isles. As well as the intrinsic historical interest, such fine-scale population structuring may well prove important in the distribution of functional variation. The third chapter used a panel of markers to show that demographic history, and admixture in particular dramatically affects the pattern of linkage disequilibrium (LD) in the genome. Preliminary studies of LD will thus be required in all populations in which mapping is to be carried out. Chapter four provides an example of how genetic structuring of human populations may be very important clinically. Average differences in drug response (inferred from functional variation in drug metabolising enzymes) were predicted among individuals clustered on the basis of variation at neutral markers, indicating that the trade-off between efficacy and safety would be different in each cluster.

In Chapter two I used different genetic systems to investigate the demographic effect of cultural change using the British Isles as a model system. When a culture spreads into a new region, the process by which this occurs may involve any grading between, on the one extreme, a mass movement of people through to the other extreme of a movement of ideas only. I sought to assess the demographic context of

cultural change in the British Isles, concentrating on Scandinavian influence in Orkney and also on the Neolithic and Iron Age transitions.

Strong population genetic structuring of Y chromosome variation was revealed between the samples representing the possible Orcadian source populations: Norway, on the one hand, and Ireland and Wales, on the other. The distribution of haplotypes in Orkney was consistent with a large Scandinavian component in the paternal heritage of the islands. When chromosomes were filtered by including only those carried by bearers of surnames that are endemic to the archipelago, the Scandinavian component was increased, exactly as would be predicted if the Norse legacy in Orkney was both cultural and genetic.

The similarity of the Y chromosome distributions in the Welsh and Irish indicated that they are likely to represent the pre-Anglo-Saxon populations of the British Isles and Ireland. Two extreme models as to the origin of these populations may again be contrasted, one in which the arrival of the Neolithic and afterwards Celtic material cultures involved folk migrations and one which postulates only cultural change during these transitions. I thus compared the Celtic samples with Basques, considered to be descended from the Palaeolithic inhabitants of Europe, and also to Near Easterners as representatives of Neolithic farmers. The Celtic and Basque Y chromosome distributions were extremely similar, but were very different to those in the Near East, suggesting that farming spread to the British Isles without large scale immigration. To assess whether other genomic regions gave the same signals as the Y chromosome, I compared patterns of variation in mitochondrial DNA (mtDNA) and at X-linked microsatellites. Much less genetic structuring was apparent for these marker systems, however principal components analysis of each genetic system revealed the same major pattern as in the Y chromosome data: the Basques and Near Easterners

occupy opposite poles of the first axis. However, for both systems influenced by female movement, the Celtic populations are close to the centre of the plot, in contrast to their position in the Y chromosome plot: next to the Basques. This indicates that they have undergone more female-mediated gene flow from other European populations than the Basques have. The Neolithic, Iron Age or other cultural transitions must have involved the immigration of women.

There is much further work that could be done. The resolution afforded by the Y chromosome, for instance, is marker-limited. The size of the euchromatic portion of the chromosome (~30 megabases) (Hurles and Jobling 2001), together with the mutation rate (1.24×10^{-9}/site/year) (Thomson et al. 2000) predicts an average of 0.9 mutations/chromosome/generation (assuming a 25 year generation time). An almost perfectly resolved genealogy is thus theoretically possible; markers that are useful in a particular branch need only be discovered. Therefore, given sample sizes at least an order of magnitude greater than common at present and enough fine-scale population structure in Norway, megabase-scale Y chromosome polymorphism discovery may have the power to identify which fjord a particular Viking ancestor set sail from!

While this may be a distant prospect, new Y chromosome single nucleotide polymorphisms (SNPs) have improved the resolution of the tree, reducing the need to use paraphyletic groups such as haplogroups 1 and 2. Thus almost all the European haplogroup two (hg2) chromosomes are derived at M170 (Underhill et al. 2000) and almost all the hg1 chromosomes at $DYS194_{469}$ (defining hg1L, a sister group of hg3) (Hammer et al. 2000). There are, however, no SNPs available as yet that split Western European hg1L chromosomes into sizeable subclades. Such markers would allow investigation of the hypothesis that a northward movement from a glacial refugium in Iberia was responsible for much of the ancestry of Northern Europe (Torroni et al.

1998; Torroni et al. 2001). Any structuring between Norwegian and Atlantic hg1L Y chromosomes revealed by new markers would allow more accurate estimates of Norse contributions in Orkney and elsewhere in the British Isles to be made.

The required markers may be among the thousands of new Y-linked SNPs in the National Institute of Health's dbSNP database (http//:www.ncbi.nlm.nih.gov/SNP/), otherwise SNP discovery may be performed using denaturing high performance liquid chromatography (Underhill et al. 1997; Underhill et al. 2000). For questions requiring more resolution within hg1, dbSNP should provide a wealth of markers as many of the SNPs will lie on the branches leading to common European haplotypes within hg1. This is because the SNP consortium markers that constitute much of dbSNP were ascertained on a panel of Europeans, Asians, Africans and Amerindians (Collins et al. 1998) and from the world-wide distribution of haplogroups (Karafet et al. 1999; Underhill et al. 2000) it can be expected that at least half of the Europeans and possibly all of the native Americans will be members of hg1 or its derivatives. Thus a large proportion of the SNPs in the database will lie on branches (a) between hg1 and the hg8/hg21 clade likely represented by most of the African part of the polymorphism discovery panel (Seielstad et al. 1994) and (b) between the different hg1 chromosomes in the panel. A diverse set of chromosomes within the clade of interest should be screened – for instance by selecting on the basis of their microsatellite genotypes (as a substitute for knowing their true genealogy) – with each new dbSNP until sufficient internal genealogical structure were revealed. Sequencing a stretch around each SNP would allow discovery of novel SNPs simultaneously with dbSNP screening. Informative markers could then be genotyped in the populations of interest.

It would also be possible to increase the resolution of both the X chromosome and mtDNA systems used in the chapter. The entire 16.5 kilobase (kb) mtDNA molecule can be sequenced which allows construction of a strongly supported tree analogous to the Y genealogy (Richards and Macaulay 2001). Choosing mtDNAs for complete sequencing on the basis of their hypervariable segment I sequences has revealed, for instance, new markers that subdivide the most frequent European clade, hgH (Finnila et al. 2001). Genotyping with these or other new markers may reveal hidden phylogeographic structure in Europe. Increased resolution can be achieved on the X chromosome simply by increasing the number of microsatellites. While no evidence of genetic structuring was found using 23 X-linked microsatellites, genotyping more microsatellites, possibly from throughout the genome, would reveal whether STRUCTURE or other clustering programs support the pattern exhibited in the principal components analysis.

Model-based inference of admixture coefficients, migration rates or population separation times might also be performed using the Y chromosome data but as the Y genealogy is only one realisation of the evolutionary process estimates have a very high variance. Using multiple unlinked regions of the genome would drastically reduce this source of error by incorporating independent genealogical histories. For instance, sequencing of low recombination regions or regions contained within long blocks of LD (Daly et al. 2001) would increase the chances of recovering sufficient mutations to give genealogical resolving power within Europe (always presuming there is structuring between the different source populations). Alternatively inferences could be made from unlinked microsatellite data. Combining these two approaches by developing SNP haplotype systems around a microsatellite or minisatellite (e.g. (Alonso and Armour 2001)) would have the advantages of

increased resolution at the tips of the tree while reducing the allele size homoplasy which results from the use of tandem repeat markers alone.

In Chapter three I investigated the effect of demographic history on the patterns of LD in different populations. The optimal strategy for testing association of SNPs and complex disease susceptibility requires an understanding of the extent and consistency of LD in the genome. I thus developed a systematic framework of 66 microsatellites for evaluating background LD in multiple regions across the entire X chromosome. I then applied this panel to populations with differing demographic histories including the Lemba, a southern African population with genetic evidence of Semitic admixture. Ethiopians, Ashkenazi Jews and South African Bantu were also included, as putative parental populations of the Lemba with less evidence of admixture. It is known that population stratification can generate spurious associations, however, even in the absence of genetic structuring, historical admixture between genetically differentiated populations can result in increased LD, which is predicted to persist for many generations.

The Lemba showed increased LD out to ~21 centimorgans (cM), where the number of marker pairs in LD was equivalent to that seen for unlinked markers. In the other populations the extent of LD was an order of magnitude less than in the Lemba: this cut off point was below 3 cM in each case. The Lemba also had a higher consistency of LD: one and a half times as many tightly-linked marker pairs were in LD compared to the other populations. If the Lemba LD resulted from admixture, the differentials in allele frequencies (delta) between the parental populations should be predictive of the LD observed. Excluding tightly-linked marker pairs (where ancestral LD may obscure that generated by admixture) and marker pairs at distances over 17

cM (where many fewer pairs are in LD), using the product of the Ashkenazi-Bantu delta values for each marker pair, the proportion of Lemba marker pairs in LD was increased for pairs in higher classes of delta product. Significantly more pairs with an above average delta product were in LD, indicating that the disequilibria were due to Bantu-Semitic admixture. The fact that the delta values are only predictive of Lemba LD for partially linked markers and not for unlinked pairs rules out genetic structuring as the only source of the LD.

In contrast, construction of artificial hybrid populations made up of an equal mixture of Ashkenazi and Bantu X chromosomes generated LD in a distance-independent manner. As expected, significantly more pairs with above-average delta products were in LD, again regardless of genetic distance. Ethiopian-Bantu artificial hybrids, on the other hand, showed little LD due to the lack of differentiation between the parental populations. Together with the fact that the Bantu-Ethiopian delta values do not predict LD in the Lemba, it would thus appear that the Ethiopians are not a good candidate for the second parental population of the Lemba. The profound effect of population history on LD documented in chapter three indicates that background LD will have to be evaluated in all populations of interest prior to mapping. LD will also provide an important source of information regarding human evolutionary history.

There is again an enormous amount of further work that could be carried out. With regard to genetic history, the non-Bantu parental population of the Lemba appears to be Semitic, but it is not known which modern population is most closely related to these enigmatic ancestors. Taking advantage of its unparalleled geographic structuring, Y chromosome polymorphism discovery in the Lemba to increase the genealogical resolving power still further, in this case along with dense sampling of

Jewish and other Middle Eastern candidate populations, would once again be the recommended course of action. Multiple autosomal haplotype systems or microsatellites could then be used to estimate admixture proportions (Bertorelle and Excoffier 1998). The deep (African versus non-African) differentiation between the parental populations would facilitate accurate estimation, however the possibility that there was only paternal Semitic input (Soodyall et al. 1996; Thomas et al. 2000) should be taken into account.

The date of this input is also unknown. Taking the Ashkenazim and Bantu as parental populations, the decay of disequilibrium with time could be investigated from the start point of the artificial hybrid populations (with a lower proportion of Jewish X chromosomes to account for the sex-biased admixture). The major difference in the hybrid populations' pattern of LD compared to that in the Lemba is that there was no fall off with distance. The time since admixture could thus be investigated by evolving artificial hybrid populations through Wright-Fisher simulations and resampling chromosomes each generation until the fall off with distance is similar to that seen in the Lemba. In this way distributions of the number of marker pairs in LD could be generated for each genetic distance class at each generation since admixture. The number of generations that have passed since admixture (assuming the Wright-Fisher and sudden admixture models) and our confidence in this number could then be estimated.

Populations without LD-generating events in their recent history are predicted only to show increased LD over scales more than an order of magnitude shorter than that in the Lemba. Therefore, if we are to evaluate the patterns of LD in all populations in which mapping will be carried out, much better coverage of distances below 1 cM is required than the markers used in chapter three provide. Linked SNPs

are the system of choice at kilobase scales as there is not a sufficient density of polymorphic microsatellites for such fine-scale work. Furthermore, as LD which is detectable only over tens of kilobases must have been generated much longer ago than the long range LD observed in the Lemba, the much slower mutation rate of SNPs compared to microsatellites means that less of the signal will be lost by mutation.

Recent work has shown that the fine-scale structure of LD in the human genome may be block-like due to the presence of recombination hotspots separated by low-recombination regions, in which most gametes bear one of a very small number of haplotypes (Daly et al. 2001). If proven to be a genome-wide phenomenon, it will be critical that this spatial heterogeneity in recombination rate is taken into account in the design of future association studies. Empirical descriptions of this block structure must be undertaken to guide their design and interpretation. The focus of previous studies on the average extent of LD (Reich et al. 2001) is much less important than the differences in within- and between-block LD among populations. Thus multiple SNP systems, each encompassing a number of adjacent blocks should be developed for initial assessment of this structure in different populations.

A geographically exhaustive population set should be screened to investigate whether ancient demographic events have shaped the block structure of LD differently in different parts of the world. For example, it has been suggested that, during human history, both the out of Africa and the 'into-America' putative bottlenecks generated disequilibria which are still observed across entire continents (Tishkoff et al. 1996). Multiple examples of populations thought to exemplify different demographic histories such as supposedly constant-sized populations (e.g. Saami, !Kung), 'isolates' (e.g. Sardinians, Finns, Ashkenazi Jews), 'sub-isolates' (e.g. single villages within Finland or Sardinia), admixed populations (e.g. African

Americans, Latin Americans) and bottlenecked populations (e.g. Pacific Islanders, Old Order Amish) should also be included to contrast with the expanded, out-bred populations.

After an extreme bottleneck (or other LD-generating event), markers will be in LD across blocks as well as within blocks (the pre-block phase) until sufficient recombination occurs at the hotspots to break down across-block associations. The population will then enter the block phase, during which time phenotypes may be mapped to blocks but not within them. After a further period of time (depending on the ratio of the hotspot to non-hotspot recombination rates), recombination will occur within blocks (the post-block phase), allowing mapping to sites within blocks.

Over kilobase distances, the patterns observed may not reflect those seen over the more often investigated cM-scale distances. This is because it is not yet clear how the recombination rate within blocks, the intensity and the density of hotspots combine to determine the overall rate of recombination in large genomic segments and, therefore, the large-scale pattern of LD. Furthermore, in the case of admixture, marker pairs which in the parental populations show either (a) insufficient differentiation or (b) association of the opposite alleles will not be in LD in the hybrid population. Admixture generates excess LD across large distances because there is usually very little LD at these scales, however at distances where there is commonly LD, it is just as likely to remove as to generate LD. Factors such as gene conversion may also influence short range LD (Ardlie et al. 2001).

Genotyping with multiple SNP systems will reveal populations' block structure phase as well as the effect of different demographic events on the dynamics of progression through the phases. Not all regions of the genome will evolve through the phases at the same time, as not all hotspots will have the same intensity, another

reason to include multiple regions in descriptions of LD. Mapping may be simulated using the data from different populations by denoting one of the SNPs as the causal variant and evaluating how well variant localisation proceeds using the observed associations with the remaining markers. This will provide a framework for investigating the utility for mapping of populations in different block structure phases.

Both shared and population-specific signatures of LD will also, of course, illuminate human genetic history by pointing to historical founder and other LD-generating events. In addition, the haplotype data produced would provide multiple genealogical systems with which to dissect admixture and, with the addition of microsatellites in each block, perhaps date events in human history using multiple independent realisations of the evolutionary process.

In chapter four I investigated the population structure of variable drug response. Inter-individual variation in drug efficacy and toxicity depends on many factors, including allelic variants in drug metabolising enzymes (DMEs) and drug targets. Attention has focussed on the population structuring of this variation because of the possibility that genetically structured populations may be composed of two or more distinct subgroups with different drug reaction profiles. However, during trials of drug safety and efficacy, if individuals are clustered, it is according to racial labels such as Black, Caucasian and Asian. I used a model-based method implemented in the program STRUCTURE to infer genetic clusters on the basis of microsatellite genotypes. A model with four clusters was supported using a heterogeneous population composed of Norwegians, Ashkenazi Jews, Armenians, Chinese, New Guineans, South African Bantu, Ethiopians and Afro-Caribbeans.

The clusters corresponded broadly to the geographical areas Western Eurasia, Sub-Saharan Africa, China and New Guinea. However, 62% of Ethiopians fell in the Western Eurasian cluster along with almost all the Jews, Norwegians and Armenians, thus demonstrating the inaccuracy of the commonly used ethnic labels Black and Caucasian as surrogates for knowledge of the actual genetic structure. Around one fifth of the Afro-Caribbeans' ancestry is also inferred to be in the Western Eurasian cluster, presumably reflecting European introgression. China and New Guinea are placed almost entirely in separate clusters, indicating that the ethnic label Asian is insufficient to describe population structure accurately.

The frequencies of functionally significant alleles at DME loci were then compared across the inferred clusters as a proxy for drug response. Significant frequency differences were observed among the inferred clusters for four of the six DMEs included. Allele frequencies differed by up to fourfold, even for the Chinese and New Guinean clusters that would have been combined as Asian in an analysis using ethnic labels. I compared how informative genetic clusters were compared to the ethnic labels Caucasian, Black and Asian by calculating DME allele frequencies in these groups. The genetic clusters were more informative than those identified by skin colour, in fact ethnically labelled groups never showed sharp differentiation that was not observed in the genetic clusters. Average difference in drug reaction among groups may thus be identified without knowledge of ethnicity or race. Given the potential of such difference to have clinical significance, until such time as the most important determinants of drug efficacy and toxicity may be genotyped cheaply, assessment of genetic structure should become a routine part of drug evaluation.

A canonical description of human population structuring is thus required for use in biomedical research. To this end, the analyses of chapter four may be extended

to include a geographically exhaustive set of populations, in particular more populations from Africa where the increased time depth may have given rise to strong differentiation, as well as Amerindians and Aboriginal Australians, who were omitted. Increasing the number of neutral markers will also be required for consistency of clustering, using one from each chromosome arm is a simple way of maximising their independence. The higher resolution afforded by increasing the number of markers may increase the number of clusters supported, however the optimum number will be determined by the resolution which explains the most variation in drug response. As pointed out in chapter four, STRUCTURE was used to demonstrate that genetic clustering was superior to phenotypic clustering, however detailed analysis should be performed using other clustering algorithms in order to obtain a standardised scheme.

A valuable test of the application of genetic clustering schemes in drug trials would be to use actual drug response data from a heterogeneous sample instead of the DME genotypes employed as proxies in chapter four. Non-genetic factors influencing drug response, such as health, age and the environment could then be controlled for.

In the future era of personalised medicine when all the variants influencing drug response are known and genotyping them is technically and commercially feasible, genetic clustering will no longer be required. The discovery of these functional variants will be expedited by an understanding of the block structure of haplotypes in DMEs, drug transporters and drug targets. A sufficient panel of haplotype-tagging SNPs could then be assembled for association studies of drug response. The population distribution of haplotypes and locations of block boundaries will also be important in predicting variable drug reaction.

The different applications illustrated here underline the importance of human genetic variation to medical science. Population genetics already plays several key roles in this field but as technologies improve it will become ever more central to our health and wellbeing. The post-genomic surge in polymorphism discovery will be followed by an unprecedented quantity of empirical population work as described here. We can hope that the consequent improvements in the understanding of the diseases that fill hospitals in the developed world will lead to novel therapies. The benefits for genetic history will, of course, be enormous: the genealogies of hundreds of loci will be resolved to give a truly genomic picture of the origin of man and his history.

Acknowledgements

So many people have helped in so many different ways to make this thesis possible. First I must thank the Natural Environment Research Council for my studentship. I thank Rachel ‘philosophical good of chimpanzees’ Whiteley, Stuart ‘I rock’ Macdonald, Isabelle ‘I’m not French!’ Colson, David ‘Eiffel tower’ Reich and Julia ‘not my cup of tea’ Gockel for help and company in Oxford and London. Cristian ‘no but I mean’ Capelli and Helen ‘everything they said about Kabul airport was true!’ Roberts have provided the same in London. I want to thank Stuart in particular for being a top office mate – there is no way I would have done all the work if he had not been doing the same. I would also like to thank Cristian for the endless genetic history brainstorms. Thanks are due to Fiona ‘Essex is a nice place’ Gratrix, Alice ‘some of us have been at work for four hours’ Smith and Richard ‘culturally I’m in another universe’ Marguerie de Rotrou for technical assistance, to Keeley ‘I had to eat a baked potato to calm down’ Ribas for bringing a sparkle to the department, to Mark ‘joy of plex’ Thomas, Neil ‘have I ever told you about?’ Bradman and Mike ‘regression’ Weale for interesting discussions, samples or expertise. Martin ‘Palaeolithic’ Richards gave invaluable advice during the mtDNA work and Chris ‘I am the type specimen’ Tyler-Smith, in particular, did the same for Y chromosomes. Vincent ‘Marvel at the results and start telling stories about them. But take them with heaps of salt.’ Macaulay helped with statistics. Numerous people have shared data before publication. I would also like to thank Mark ‘This is no proposal for an early origin of tourism’ Jobling and Peter ‘seminar’ Donnelly for being my examiners. Outside of the lab, many people in Orkney, Edinburgh, Oxford and London have

unwittingly helped this thesis come to fruition by providing respite in nights in the Turf, the Exmouth and beyond as well as fear and loathing in Finsbury Park: Nicola ‘soft on the outside, hard on the inside’ Crossland, Dave ‘dramas=0’ Clancy, Melanie ‘do not start!’ West, Claire ‘that’s like never having left Lambeth, man!’ Grant, Michelle ‘wonder-woman’ Keaney, Mike ‘intensity and frequency’ Cant and Lucian ‘tequila’ Sweet. Angelika ‘do you want to come and stroke the dog’s tail?’ Kritz brought me more joy than I can write about here – thanks for everything schatzi. I should also like to thank all those people who collected the human DNA samples without which none of this work would be possible and, of course, the sample donors, especially the people of Orkney. David ‘how many gels is that?’ Goldstein provided inspiration and excellent supervision throughout this thesis. Finally, I should like to thank my sister Karen ‘Wilson’ Nielsen and my parents for encouragement and support.

References

Alonso S, Armour JA (2001) A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outstide Africa. Proc Natl Acad Sci U S A 98:864-9.

Ammerman A, Cavalli-Sforza L (1984) The Neolithic Transition and the Genetics of Populations in Europe. Princeton University Press, Princeton

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457-65

Angius A, Melis PM, Morelli L, Petretto E, Casu G, Maestrale GB, Fraumene C, Bebbere D, Forabosco P, Pirastu M (2001) Archival, demographic and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits. Human Genetics 109:198-209

Ardlie K, Liu-Cordero SN, Eberle MA, Daly M, Barrett J, Winchester E, Lander ES, Kruglyak L (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am J Hum Genet 69:582-9.

Aynacioglu AS, Sachse C, Bozkurt A, Kortunay S, Nacak M, Schroder T, Kayaalp SO, Roots I, Brockmoller J (1999) Low frequency of defective alleles of cytochrome P450 enzymes 2C19 and 2D6 in the Turkish population. Clin Pharmacol Ther 66:185-92.

Barbujani G, Pilastro A, De Domenico S, Renfrew C (1994) Genetic variation in North Africa and Eurasia: neolithic demic diffusion vs. Paleolithic colonisation. Am J Phys Anthropol 95:137-54

Barnes MP (1998) The Norn Language of Orkney and Shetland. Shetland Times, Lerwick

Basile VS, Ozdemir V, Masellis M, Walker ML, Meltzer HY, Lieberman JA, Potkin SG, Alva G, Kalow W, Macciardi FM, Kennedy JL (2000) A functional polymorphism of the cytochrome P450 1A2 (CYP1A2) gene: association with tardive dyskinesia in schizophrenia. Mol Psychiatry 5:410-7

Bathum L, Johansson I, Ingelman-Sundberg M, Horder M, Brosen K (1998) Ultrarapid metabolism of sparteine: frequency of alleles with duplicated CYP2D6 genes in a Danish population as determined by restriction fragment length polymorphism and long polymerase chain reaction. Pharmacogenetics 8:119-23.

Bengtson JD (1991) Some Macro-Caucasian etymologies. In: Sheveroshkin V (ed) Dene-Sino-Caucasian Languages. Brockmeyer, Bochum

Bernal ML, Sinues B, Johansson I, McLellan RA, Wennerholm A, Dahl ML, Ingelman-Sundberg M, Bertilsson L (1999) Ten percent of North Spanish individuals carry duplicated or triplicated CYP2D6 genes associated with ultrarapid metabolism of debrisoquine. Pharmacogenetics 9:657-60

Bertorelle G, Excoffier L (1998) Inferring admixture proportions from molecular data. Mol Biol Evol 15:1298-311

Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, Comas D (1995) Human mitochondrial DNA variation and the origin of Basques. Ann Hum Genet 59:63-81

Bianchi NO, Catanesi CI, Bailliet G, Martinez-Marignac VL, Bravi CM, Vidal-Rioja LB, Herrera RJ, Lopez-Camelo JS (1998) Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. Am. J. Hum. Genet. 63:1862-71

Blum M, Demierre A, Grant DM, Heim M, Meyer UA (1991) Molecular mechanism of slow acetylation of drugs and carcinogens in humans. Proc Natl Acad Sci U S A 88:5237-41

Bosch E, Calafell F, Santos F, Perez-Lezaun A, Comas D, Benchemsi N, Tyler-Smith C, Bertranpetit J (1999) STR variation is deeply structured by genetic background on the human Y chromosome. Am. J. Hum. Gen. 65:1623-1638

Briscoe D, Stephens JC, O'Brien SJ (1994) Linkage disequilibrium in admixed populations: applications in gene mapping. J Hered 85:59-63

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63:861-9

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31-6

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. Princeton University Press, Princeton, New Jersey

Cavalli-Sforza LL, Minch E (1997) Paleolithic and Neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 61:247-54

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA 85:9119-23

Collins A, Lonjou C, Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci USA 96:15173-7

Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8:1229-31.

Cunliffe B (1997) The Ancient Celts. Oxford University Press, Oxford

Dahl ML, Johansson I, Bertilsson L, Ingelman-Sundberg M, Sjoqvist F (1995) Ultrarapid hydroxylation of debrisoquine in a Swedish population. Analysis of the molecular genetic basis. J Pharmacol Exp Ther 274:516-20.

Daly AK, Brockmoller J, Broly F, Eichelbaum M, Evans WE, Gonzalez FJ, Huang JD, Idle JR, Ingelman-Sundberg M, Ishizaki T, Jacqz-Aigrain E, Meyer UA, Nebert DW, Steen VM, Wolf CR, Zanger UM (1996) Nomenclature for human CYP2D6 alleles. Pharmacogenetics 6:193-201

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229-32.

de Knijff P (2000) Messages through Bottlenecks: On the Combined Use of Slow and Fast Evolving Polymorphic Markers on the Human Y Chromosome. Am J Hum Genet 67:1055-1061

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152-4

Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, Todd JA (2000) The genetically isolated populations of finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nat Genet 25:320-3

Evans WE, Relling MV (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. Science 286:487-91

Ferguson RJ, De Morais SM, Benhamou S, Bouchardy C, Blaisdell J, Ibeanu G, Wilkinson GR, Sarich TC, Wright JM, Dayer P, Goldstein JA (1998) A new genetic defect in human CYP2C19: mutation of the initiation codon is responsible for poor metabolism of S-mephenytoin. J Pharmacol Exp Ther 284:356-61

Finnila S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. Am J Hum Genet 68:1475-84.

Freimer NB, Service SK, Slatkin M (1997) Expanding on population studies. Nature Genet 17:371-3

Gabunia L, Vekua A (1995) A Plio-Pleistocene hominid from Dmanisi, East Georgia, Caucasus. Nature 373:509-12.

Gaedigk A, Gotschall RR, Forbes NS, Simon SD, Kearns GL, Leeder JS (1999) Optimization of cytochrome P4502D6 (CYP2D6) phenotype assignment using a genotyping algorithm based on allele frequency data. Pharmacogenetics 9:669-82

Gaedigk A, Tyndale RF, Jurima-Romet M, Sellers EM, Grant DM, Leeder JS (1998) NAD(P)H:quinone oxidoreductase: polymorphisms and allele frequencies in Caucasian, Chinese and Canadian Native Indian and Inuit populations. Pharmacogenetics 8:305-13

Gamkrelidze T, Ivanov V (1990) The early history of Indo-European languages. Sci. Am. 262:110-116

Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216-34

Goldstein DB, Pollock DD (1997) Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. J Hered 88:335-42

Goldstein DB, Reich DE, Bradman N, Usher S, Seligsohn U, Peretz H (1999) Age Estimates of Two Common Mutations Causing Factor XI Deficiency: Recent Genetic Drift Is Not Necessary for Elevated Disease Incidence among Ashkenazi Jews. Am. J. Hum. Genet. 64:1071-1075

Goldstein DB, Roemer GW, Smith DA, Reich DE, Bergman A, Wayne RK (1999) The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics 151:797-801

Goldstein DB, Zerjal T, Wilson JF, Pandya A, Santos FR, Thomas MG, Bradman N, Tyler-Smith C (in press) Mutation rates at human Y chromosome microsatellites show a linear dependence on repeat count. Genetics

Goldstein JA, Blaisdell J (1996) Genetic tests which identify the principal defects in CYP2C19 responsible for the polymorphism in mephenytoin metabolism. Methods Enzymol 272:210-8

Gough AC, Miles JS, Spurr NK, Moss JE, Gaedigk A, Eichelbaum M, Wolf CR (1990) Identification of the primary gene defect at the cytochrome P450 CYP2D locus. Nature 347:773-6

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361-72

Hamblin MT, Di Rienzo A (2000) Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus. Am J Hum Genet 66:1669-1679

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15:427-41

Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjanazi H, Karafet T, Santachiara-Benerecetti S, Oppenheim A, Jobling MA, Jenkins T, Ostrer H, Bonne-Tamir B (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc Natl Acad Sci U S A

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60:772-89.

Harris EE, Hey J (1999) X chromosome evidence for ancient human histories. Proc Natl Acad Sci U S A 96:3320-4.

Hawkes C (1931) Hill forts. Antiquity 5:60-97

Helgason A, Sigureth ardottir S, Nicholson J, Sykes B, Hill EW, Bradley DG, Bosnes V, Gulcher JR, Ward R, Stefansson K (2000) Estimating scandinavian and gaelic ancestry in the male settlers of iceland. Am J Hum Genet 67:697-717

Hill EW, Jobling MA, Bradley DG (2000) Y-chromosome variation and Irish origins. Nature 404:351-2

Hirszfeld L, Hirszfeld H (1919) Essai d'application des méthodes au problème des races. Anthropologie 29:505-537

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci U S A 92:532-6

Hurles ME, Jobling MA (2001) Haploid chromosomes in molecular ecology: lessons from the human Y. Mol Ecol 10:1599-613.

Huttley GA, Smith MW, Carrington M, O'Brien SJ (1999) A scan for linkage disequilibrium across the human genome. Genetics 152:1711-22

ICH (1998) Ethnic Factors in the Acceptability of Foreign Clinical Data. International Conference on Harmonisation

Jobling MA (1994) A survey of long-range DNA polymorphisms on the human Y chromosome. Hum. Mol. Genet. 3:107-14

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends. Genet. 11:449-56

Kagimoto M, Heim M, Kagimoto K, Zeugin T, Meyer UA (1990) Multiple mutations of the human cytochrome P450IID6 gene (CYP2D6) in poor metabolizers of debrisoquine. Study of the functional significance of individual mutations by expression of chimeric genes. J Biol Chem 265:17209-14

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihara S, de Knijff P, Wiebe V, Griffiths RC, Templeton AR, Hammer MF (1999) Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. Am. J. Hum. Genet. 64:817-31

Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and Linkage Disequilibrium at the Phenylalanine Hydroxylase Locus, PAH, in a Global Representation of Populations. Am J Hum Genet 66:1882-1899

Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. Am J Hum Genet 62:1171-9

Kohlmeier L, DeMarini D, Piegorsch W (1997) Gene-nutrient interactions in nutritional epidemiology. In: Margetts B, Nelson M (eds) Design Concepts in Nutritional Epidemiology. Oxford University Press, Oxford

Krajinovic M, Labuda D, Richer C, Karimi S, Sinnett D (1999) Susceptibility to childhood acute lymphoblastic leukemia: influence of CYP1A1, CYP2D6, GSTM1, and GSTT1 genetic polymorphisms. Blood 93:1496-501

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genet 22:139-44

Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, Schuetz J, Watkins PB, Daly A, Wrighton SA, Hall SD, Maurel P, Relling M, Brimer C, Yasuda K, Venkataramanan R, Strom S, Thummel K, Boguski MS, Schuetz E (2001) Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. Nat Genet 27:383-91.

Laan M, Pääbo S (1997) Demographic history and linkage disequilibrium in human populations. Nature Genet 17:435-8

Lamb G (1981) Orkney Surnames. Paul Harris, Edinburgh

Lamb G (1993) Testimony of the Orkneyingar. Byrgisey, Kirkwall

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lautenberger JA, Stephens JC, O'Brien SJ, Smith MW (2000) Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. Am J Hum Genet 66:969-78

Lonjou C, Collins A, Morton NE (1999) Allelic association between marker loci. Proc Natl Acad Sci U S A 96:1621-6

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B, Torroni A (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control- region sequences and RFLPs. Am J Hum Genet 64:232-49

McKeigue PM (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet 63:241-51

Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. Science 201:786-92.

Meyer UA, Zanger UM (1997) Molecular mechanisms of genetic polymorphisms of drug metabolism. Annu Rev Pharmacol Toxicol 37:269-96

Opdal SH, Rognum TO, Vege A, Stave AK, Dupuy BM, Egeland T (1998) Increased number of substitutions in the D-loop of mitochondrial DNA in the sudden infant death syndrome. Acta Paediatr 87:1039-44

Otte M (1990) The northwestern European plain around 18,000 BP. In: Soffer O, Gamble C (eds) The world at 18,000 BP. Vol. 1. Unwin Hyman, London

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63:1839-51

Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martinez-Arias R, Clarimon J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. Am. J. Hum. Genet. 65:208-19

Perez-Lezaun A, Calafell F, Seielstad M, Mateu E, Comas D, Bosch E, Bertranpetit J (1997) Population genetics of Y-chromosome short tandem repeats in humans. J. Mol. Evol. 45:265-70

Peterson AC, Di Rienzo A, Lehesjoki AE, de la Chapelle A, Slatkin M, Freimer NB (1995) The distribution of linkage disequilibrium over anonymous genome regions. Hum Mol Genet 4:887-94

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220-8

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-59

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170-81

Pult I, Sajantila A, Simanainen J, Georgiev O, Schaffner W, Paabo S (1994) Mitochondrial DNA sequences from Switzerland reveal striking homogeneity of European populations. Biol Chem Hoppe Seyler 375:837-40

Quintana-Murci L, Semino O, Minch E, Passarimo G, Brega A, Santachiara-Benerecetti AS (1999) Further characteristics of proto-European Y chromosomes. Eur. J. Hum. Genet. 7:603-8

Raymond M, Rousset F (1995) An exact test for population differentiation. Evolution 49:1280-1283

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199-204.

Reich DE, Goldstein DB (1998) Genetic evidence for a Paleolithic human population expansion in Africa. Proc Natl Acad Sci USA 95:8119-23

Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4-16.

Renfrew C (1987) Archaeology and Language. Penguin, London

Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt HJ, Sykes B (1996) Paleolithic and neolithic lineages in the European mitochondrial gene pool. Am. J. Hum. Genet. 59:185-203

Richards M, Macaulay V (2001) The mitochondrial gene tree comes of age. Am J Hum Genet 68:1315-20.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, et al (2000) Tracing European Founder Lineages in the Near Eastern mtDNA Pool. Am J Hum Genet 67:1251-1276

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-7

Ritchie A (1993) Viking Scotland. Historic Scotland, London

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-33.

Sachse C, Brockmoller J, Bauer S, Roots I (1997) Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. Am J Hum Genet 60:284-95.

Sachse C, Brockmoller J, Bauer S, Roots I (1999) Functional significance of a C-->A polymorphism in intron 1 of the cytochrome P450 CYP1A2 gene tested with caffeine. Br J Clin Pharmacol 47:445-9

Schneider S, Kueffer J-M, Roessli D, Excoffier L (1997) Arlequin ver. 1.1: A software for population genetic analysis., Genetics and Biometry Laboratory, University of Geneva, Switzerland

Seielstad MT, Hebert JM, Lin AA, Underhill PA, Ibrahim M, Vollrath D, Cavalli-Sforza LL (1994) Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. Hum Mol Genet 3:2159-61

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nat. Genet. 20:278-80

Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti AS (1996) A view of the neolithic demic diffusion in Europe through two Y chromosome-specific markers. Am. J. Hum. Genet. 59:964-8

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA (2000) The genetic legacy of paleolithic homo sapiens sapiens in extant europeans: A Y chromosome perspective. Science 290:1155-9

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60:957-64

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66:262-78

Slatkin M (1994) Linkage disequilibrium in growing and stable populations. Genetics 137:331-6

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457-62

Smith B (2001) The Picts and the Martyrs or how did the Vikings kill the Native Population of Orkney and Shetland? *Northern Studies* 36:7-32.

Sokal RR, Rohlf FJ (1995) Biometry. Freeman, New York

Soodyall H, Vigilant L, Hill AV, Stoneking M, Jenkins T (1996) mtDNA control-region sequence variation suggests multiple independent origins of an Asian-specific 9-bp deletion in sub-Saharan Africans. Am J Hum Genet 58:595-608

Spurdle AB, Jenkins T (1996) The origins of the Lemba "Black Jews" of southern Africa: evidence from p12F2 and other Y-chromosome markers. Am J Hum Genet 59:1126-33

Stephens JC, Briscoe D, O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet 55:809-24

Swisher CC, Curtis GH, Jacob T, Getty AG, Suprijo A, Widiasmoro (1994) Age of the earliest known hominids in Java, Indonesia. Science 263:1118-21.

Taillon-Miller P, Bauer-Sardina I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP, Kwok PY (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat Genet 25:324-8

Templeton AR (1997) Out of Africa? What do genes tell us? Curr Opin Genet Dev 7:841-7.

Thomas MG, Bradman N, Flinn HM (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Hum. Genet. 105:577-581

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N, Goldstein DB (2000) Y-chromosomes travelling South: The Cohen Modal Haplotype and the Origins of the Lemba - The "Black Jews of Southern Africa". Am J Hum Genet 66:674-686

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. Nature 394:138-40

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci U S A 97:7360-5

Thomson WPL (1986) Pict, Norse, Celt and Lowland Scot. In: Berry RJ, Firth HN (eds) The People of Orkney. Orkney Press, Kirkwall

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380-7

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. Am J Hum Genet 62:1389-402

Tishkoff SA, Ruano G, Kidd JR, Kidd KK (1996) Distribution and frequency of a polymorphic Alu insertion at the plasminogen activator locus in humans. Hum Genet 97:759-64

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293:455-62.

Torroni A, Bandelt HJ, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus ML, Bonne-Tamir B, Scozzari R (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62:1137-52

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, et al (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69:844-52.

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML, Wallace DC (1996) Classification of European mtDNAs from an analysis of three European populations. Genetics 144:1835-50

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography [letter] [see comments]. Genome Res 7:996-1005

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-

Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26:358-61

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, et al (2001) The sequence of the human genome. Science 291:1304-51.

Watson E, Forster P, Richards M, Bandelt HJ (1997) Mitochondrial footprints of human expansions in Africa. Am J Hum Genet 61:691-704

Weber WW (1997) Pharmacogenetics. Oxford University Press, Oxford

Wright AF, Carothers AD, Pirastu M (1999) Population choice in mapping genes for complex diseases. Nature Genet 23:397-404

Yu N, Li W (2000) No fixed nucleotide difference between Africans and Non-Africans at the pyruvate dehydrogenase E1 alpha-subunit locus. Genetics 155:1481-3.

Zavattari P, Deidda E, Whalen M, Lampis R, Mulargia A, Loddo M, Eaves I, Mastio G, Todd JA, Cucca F (2000) Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. Hum Mol Genet 9:2947-57.

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhovel W, Fretwell N, Jobling MA, Harihara S, Shimizu K, Semjidmaa D, Sajantila A, Salo P, Crawford MH, Ginter EK, Evgrafov OV, Tyler-Smith C (1997) Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. Am. J. Hum. Genet. 60:1174-83

Zerjal T, Pandya A, Santos FR, Adhikari R, Tarazona E, Kayser M, Evgrafov O, Singh L, Thangaraj K, Destro-Bisol G, Thomas MG, Qamar R, Mehdi SQ, Rosser ZH, Hurles ME, Jobling MA, Tyler-Smith C (1999) The use of Y-

chromosomal DNA variation to investigate population history: recent male spread in Asia and Europe. In: Papiha SS, Deka R (eds) Genomic diversity: applications in human population genetics. Plenum, New York