European Y chromosome diversity: population movement, geography and language

Jobling, Mark A.
University of Leicester
Department of Genetics
University Road
Leicester LE1. 7RH, UK

Y giza kromosomak gizonezkoen sexua determinatzen du, eta beraz, haploidea da, eta ez gurutzagarria. Nahiz eta Y kromosoma locus bakarrekoa den, eta beraz, deriba eta hautespena gerta badaitezke ere, ADNaren segmentu horrek informazio ugari ematen du giza populazioen historia eta egiturez. Azterlan honek horren gaitasuna frogatu du sexu baterako edo besterako makurtzeko gehigarriak aztertzerakoan. Y kromosoma lanabes analitiko egokiena izan daiteke emakumeen migrazio handia gertatu den kasuetan.

Giltza-Hitzak: Y kromosoma. Populazioaren historia. Gehigarriak. Hautespen naturala. Deriba genetikoa.

El cromosoma Y humano determina el sexo masculino, y es por lo tanto haploide y no - recombinable. Aunque el cromosoma Y es de locus único, y por lo tanto susceptible de deriva y selección, este segmento de ADN resulta altamente informativo en el análisis de historiales y estructuras de poblaciones humanas. Este estudio ha demostrado su poder en cuanto a la disección de sucesos de aditivos de inclinación hacia uno u otro sexo. El cromosoma Y puede ser la mejor herramienta analítica en casos en los que se ha producido una considerable migración femenina.

Palabras Clave: Cromosoma Y. Historial de población. Aditivos. Selección natural. Deriva genética.

Le chromosome humain Y détermine le sexe masculin, et il est donc haploïde et non recombinable. Bien que le chromosome Y soit d'un seul lucus, et donc susceptible de dérive et de sélection, ce segment d'ADN est grandement informatif sur l'analyse des parcours et des structures des populations humaines. Cette étude a démontré son pouvoir en ce qui concerne la dissection d'événements d'additifs d'inclinaison vers l'un ou l'autre sexe. Le chromosome Y peut être le meilleur outil analytique dans les cas où s'est produite une considérable migration féminine.

Mots Clés: Chromosome Y. Parcours de population. Additifs. Sélection naturelle. Dérive génétique.

XV Congreso de Estudios Vascos: Euskal zientzia eta kultura, eta sare telematikoak = Ciencia y cultura vasca, y redes telemáticas = Science et culture basque, et réseaux télématiques = Basque science and culture, and telematic networks (15. 2001. Donostia). – Donostia : Eusko Ikaskuntza, 2002. - P 171-176. - ISBN: 84-8419-949-5.

1. WHY Y?

In humans, as in all mammals, the Y chromosome is dominantly male sex-determining, through the action of a single gene, *SRY* (Sinclair *et al.*, 1990). Because of its role in sex-determination, the Y is constitutively haploid, and therefore escapes recombination over most of its length. Recombination does occur with the X in the two pseudoautosomal regions at the tips of the short and long arms, but the majority of the chromosome, lying between these regions, is exempt from the reshuffling effects of recombination. For this reason, differences between contemporary Y chromosomes arise only as a result of mutation events in their ancestors, and not as a result of recombinational reshuffling. Y chromosomes therefore contain a relatively simple record of their past (Jobling and Tyler-Smith, 1995), and are analogous to the maternally inherited, non-recombining and functionally haploid mtDNA. All modern Y chromosomes can be traced back ('coalesce') to a single common ancestor at some point in the past.

The reason for coalescence to one ancestor is the variance in the number of sons that men have. Some have many, while some have none at all. In fact, this variance is larger for males than it is for females, and the Y is therefore particularly susceptible to drift - stochastic changes in the frequencies of Y types. If drift is too strong, it could erase patterns which might otherwise reflect the histories and relationships of populations.

2. AIMS OF Y CHROMOSOME RESEARCH IN HUMAN EVOLUTIONARY STUDIES

In the Y, we have a large (~60Mb) segment of non-recombining, paternally inherited DNA - what would we like to do with it? First, we need markers that will allow us to distinguish between different kinds of Y chromosomes (polymorphisms). We now have many of these, of different varieties, and with different properties. Second, we want to build a phylogenetic tree which relates haplotypes defined by these polymorphisms, and shows the true relationships between Y chromosomes. This tree should have a defined root, and dates on its branchpoints. The tree turns out to be trivial to build for binary polymorphisms, and trivial to root. Dating is much more difficult.

Having defined our Y types and their relationships, we want to look in human populations at the distributions and frequencies of these types, and ask what these can tell us about the histories of populations - in particular, about past demographies: mating practices, migrations, colonisations and range expansions. We also want to examine the relationships of the distribution of Y chromosomes with linguistic and geographic barriers, and compare Y-chromosomal geographic patterns with those of maternally and biparentally inherited systems. This could tell us about the different behaviours of men and women in the past. Finally, we must bear in mind that there are other forces which can affect the distribution of genetic diversity among populations - is selection acting upon human Y chromosomes?

3. TYPES OF Y MARKER

The Y chromosome carries a wide range of different kinds of polymorphic marker, which differ widely in their mutational rates and processes. Markers such as minisatellites and microsatellites have high mutation rates, around 10^{-3} to 10^{-2} per locus per generation (Jobling *et al.*, 1999; Kayser *et al.*, 2000). They are not the markers of choice for tree-building, but have uses in dating lineages and in forensics (Jobling *et al.*, 1997) and genealogy (Jobling, 2001). They will not be discussed in this paper. In contrast, base substitutions (SNPs) on the Y have mutation rates of around 10^{-8} per base per generation (Thomson *et al.*, 2000), show identity by descent, and can therefore be regarded as unique events in human evolution.

Tree-building, using binary markers and the principle of parsimony, is trivial. A single tree emerges from any set of Y markers. The one referred to throughout this paper (Figure 1) employs 33 binary markers to define 35 kinds of Y chromosomes ('haplogroups'), given arbitrary numbers from 1 to 35 (Jobling and Tyler-Smith, 2000; Kalaydjieva *et al.*, 2001). The tree roots in Africa, which is consistent with an African origin for modern Y chromosomes.

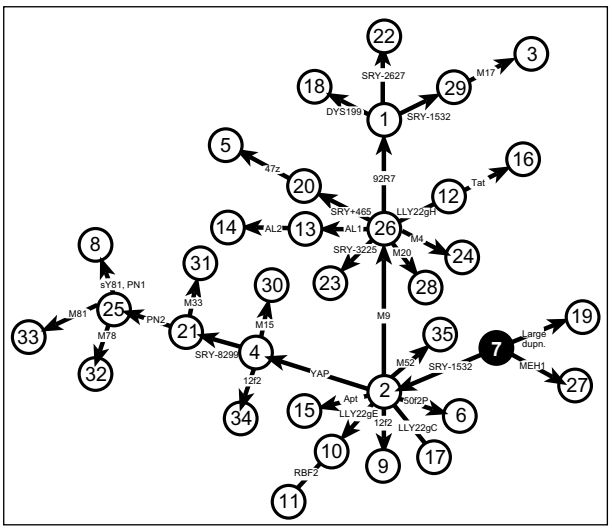

Figure 1: Maximum parsimony tree of Y-chromosomal haplogroups.

Each circle represents a compound binary marker haplotype ('haplogroup'), and contains the number assigned to it. Lines between circles indicate binary mutations, whose names are given on the lines. Where the lines have arrowheads these point to the derived state of the marker, as determined from great ape-human comparisons. Lines lacking arrowheads do so because it is not possible to determine their ancestral state, and for this reason the tree is not rooted. Maximum parsimony rooting would place the root in hg7 (filled circle), which is sub-Saharan African specific. Based on previously published information (Jobling and Tyler-Smith, 2000; Kalaydjieva *et al.*, 2001).

4. GLOBAL GEOGRAPHICAL DIFFERENTIATION OF Y CHROMOSOME HAPLOTYPES

Having defined Y haplogroups, we can ask how they are distributed among worldwide populations. Distribution is highly non-random, with many haplogroups showing strong population-specificity (Jobling and Tyler-Smith, 2000). How does this compare with the distribution of other markers, elsewhere in the human genome? In studies comparing genetic differentiation in geographical space of Y chromosomes, autosomes, and mtDNAs (Seielstad *et al.*, 1998), the Y shows the steepest genetic differentiation of all segments of the genome. There are a number of possible reasons for this, but the favoured one is patrilocality: when a man and woman who have different birthplaces marry, the woman generally moves to the man's home rather than vice versa. The effect of this is a continuous, generally short-range migration of women, and a stasis of men. The maternal genetic landscape becomes homogenised, while the paternal landscape becomes more and more differentiated.

5. EUROPEAN Y CHROMOSOME DIVERSITY

We undertook a study of European Y diversity (Rosser *et al.*, 2000) to address questions of the origins of agriculture in Europe, which, if spread by demic diffusion (Menozzi *et al.*, 1978), might be expected to lead to continent-wide clines with foci in the near East, and to examine the relationship of patterns of Y diversity with those of linguistic and geographical differentiation.

We determined the haplogroups of 3689 Y chromosomes from 48 European and circum-European populations. Ten of a possible 12 haplogroups were found, of which six made up the vast majority of the chromosomes. From simple visual inspection, the distribution of chromsomes in different haplogroups is highly non-random (Figure 2). Spatial autocorrelation analysis shows that overall variation is clinally distributed, as is variation for five of the six major haplogroups. Two haplogroups, hg1 and hg9, representing over 40% of the chromosomes, show complementary continent-wide clines with north-west/ south-east axes, and are compatible with an origin in demic diffusion originating in the Neolithic. Other haplogroups show more regional clines, and reflect other population movements. In the absence of convincing evidence, we are reluctant to assign these clines to specific historical or prehistorical events.

Principal components analysis and Mantel testing shows that affiliations between populations are based primarily on geography, rather than on language.

Genetic barrier analysis of these data reveals the major barriers to lie in Central Europe, running north-south. These correlate strikingly with the 'suture zones' identified from the analysis of patterns of genetic diversity of several plant and animal species (Taberlet *et al.*, 1998), and explained as a result of the re-expansion of populations from glacial refugia in the southern peninsulas of Europe after the last glacial maximum.

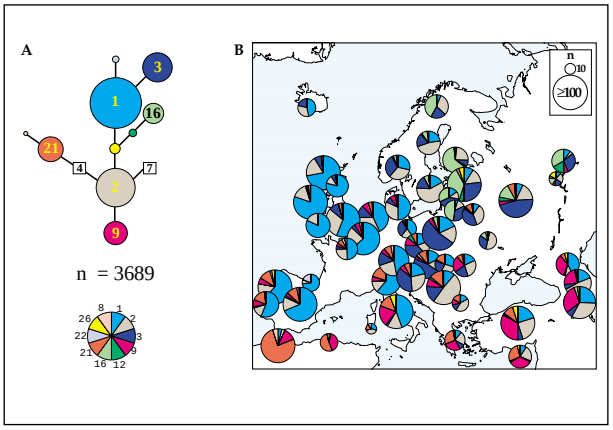

Figure 2: Distribution of Y-chromosomal haplogroups in Europe

A) A simplified version of the tree shown in Figure 1, illustrating the relative frequencies of the 10 observed haplogroups in a sample of 3689 European and circum-European Y chromosomes. Areas of circles are proportional to haplogroup frequency. Haplogroups 4 and 7 were not observed.

B) Map showing the distribution of haplogroups among 48 populations. Each pie chart represents a population, with area proportional to sample size, up to a number of 100 or greater. The area of each coloured sector is proportional to the frequency of the corresponding haplogroup - for key to colours, see A.

Figure drawn from published data (Rosser *et al.*, 2000).

6. EUROPEAN Y CHROMOSOME DIVERSITY - INDEPENDENT STUDIES

In a an independent study (Semino *et al.*, 2000), a different set of binary Y markers was used to analyse a different set of European population samples. Encouragingly, the overall patterns are very similar in both studies. However, this study is keen to assign Neolithic or Palaeolithic origins to particular haplogroups, to estimate their ages based on scanty evidence, and then to equate the age estimate for a lineage with that of a population or population movement. This has to be done with caution, and statements to the effect that 80% of the population of Europe is Palaeolithic, must be taken with a pinch of salt. Unfortunately, we do not know the haplogroup distribution of the Neolithic farmers who began their westward migration 10,000 years ago.

7. EUROPEANS ABROAD

Identifying the pattern of lineages in European lineages allows the identification of European admixture in non-European populations. An example is the ~30% European Y-chromosomal admixture in the Cook Islands, Eastern Polynesia (Hurles *et al.*, 1998), which contrasts with zero European mtDNA admixture (Sykes *et al.*, 1995). This is likely to reflect the composition of the first European

ships of contact, as well as differences in male and female sexual behaviour. An even more striking picture emerges from our unpublished work on Greenlandic Inuit, who possess only Native American mtDNAs (Saillard *et al.*, 2000), but 58% European Y chromosomes. These Y chromosomes could date to mediaeval times, from the vanished Viking colonies of south Greenland, or from the more recent Dano-Norwegian colonisation after 1720. The latter seems rather more probable given the extreme sex bias of the admixture in this case.

8. WHAT CAN WE KNOW ABOUT INTER- AND INTRACONTINENTAL POPULATION MOVEMENT?

The general conclusion based on a number of admixture studies is that intercontinental gene flow is relatively easy to see and interpret: source populations are easily differentiated, and the events responsible for admixture are often historically documented. In contrast, intracontinental gene flow is more difficult, often because the populations involved share relatively recent common origins: for example, using genetics to trace the incursion of the Vikings and Saxons into the British Isles, as is popular today, seems unlikely to be easy. There is also a general lack of knowledge about the haplogroup compositions of migrating populations, and gene flow cannot always be assumed to be unidirectional. One approach to these issues is to identify 'signature' lineages of one population, and to seek them in another. If this can be done, it at least unambiguously demonstrates relationship or contact between the populations.

Studies such as our European diversity survey demonstrate clearly that drift has not erased coherent patterns of Y-chromosomal variation which have their roots in the distant past. Indeed, despite the potential problem of drift, the Y chromosome may be the best marker in some cases, where there has been high female migration. The negative effect of drift on the Y is that it makes quantitative relationships, such as migration rates and population-separation times, imprecise. Dating of Y-chromosomal lineages is inherently unreliable because of problems of microsatellite mutation rate estimation, sampling, generation time and demographic history, and in general it seems a poor idea to equate the age of a lineage with the age of a population or event. However, lineage ages can sometimes suggest upper limit for population or migration ages.

9. HOW FAR IS NATURAL SELECTION INFLUENCING THE DISTRIBUTION OF Y DIVERSITY?

Because the Y chromosome is a single locus, and carries genes involved in male fertility, it is a potential target for selection. This needs investigation if we are to be confident in our interpretations of geographical patterns of Y diversity in terms of population history. We have taken the approach of trying to identify Y haplotypes which are predisposed to or protected against particular deleterious phenotypes. These studies concentrate on male infertility, which can be caused by deletion of one of three non-overlapping regions of the Y chromosome long arm known as *AZF* (for *azoospermia factor*) *a*, *b*, and *c* (Vogt *et al.*, 1996). We have evidence to suggest that *AZFa* infertility arises preferentially from a subset of haplotypes. The mechanism of mutation in this case is aberrant recombination between 10kb direct repeats (HERV elements) flanking a region containing two genes, which causes their deletion (Blanco *et al.*, 2000). The efficiency of aberrant recombination of this kind is influenced by the degree of sequence identity between the repeats; in some Y lineages, one repeat has 'converted' the other, so that the extent of identity is greater than in other lineages. These lineages are therefore predisposed to mutation. The frequency of *AZFa* infertility is low (about 10^{-5}), and so the selective effect here is likely to be very weak, and overwhelmed by stochastic effects.

Mutation at *AZFc*, also causing infertility, is also caused by recombination between direct repeats, this time 229kb in length (Kawaguchi *et al.*, 2001). This event occurs more frequently that *AZFa* (about 10^{-4}), and deletes a 3.5Mb region containing many genes. Males carrying a 50f2/C deletion (Jobling *et al.*, 1996), including all of haplogroup 16, common east of the Baltic (Figure 2), are likely to be protected against these *AZFc* deletions, as they lack one of the participating direct repeats.

In another example where an effect exists but a potential mechanism is lacking, a study of Danish men with low sperm count reveals that they include a significantly higher proportion of individuals from one lineage, haplogroup 26, than do controls (Krausz *et al.*, 2001).

It is clear that some deleterious phenotypes arise from subsets of Y haplotypes, and these represent targets for negative selection. Apart from *AZFa* and *AZFc* infertility another proven example is XY translocation leading to sex-reversal (Jobling *et al.*, 1998). However, most of these are low frequency phenotypes, and their effects are therefore likely to be overwhelmed by stochastic phenomena. Another approach to identifying selective influence is to seek evidence in patterns of diversity: if diversity is lower than expected, this can indicated recent 'selective sweeps', which would reflect positive selection. Thus far, such studies have failed to reveal evidence for sweeps; however, their reliability is in doubt, as it is very difficult to distinguish the signal of positive selection from that of population expansion (Jobling and Tyler-Smith, 2000).

On balance, most of the evidence we have to date suggests that positive or negative selective effects have not been strong, and that most of the

patterning we see results from population history and drift, not selection.

10. CONCLUSIONS

Despite the potential problems of drift and selection inherent in a single locus such as the Y chromosome, this segment of DNA is proving highly informative in analysis of the histories and structures of human populations. Y analysis has shown its power in the dissection of sex-biased admixture events, and it may be the best tool for the job in cases where there has been substantial female migration. Recent developments in SNP and microsatellite discovery have enormously improved the available haplotype resolution, and this will have an impact on genealogical and forensic studies, as well as the evolutionary and haplotype association studies described above.

ACKNOWLEDGEMENTS

I am a Wellcome Trust Senior Fellow in Basic Biomedical Science (grant no. 057559). I thank my colleagues Zoë Rosser, Matt Hurles, Elena Bosch, Turi King and Andy Lee for contributions to many of these studies. I also thank the many collaborators in the European project, Søren Nørby and Francesc Calafell for collaboration in the Inuit project, Ken McElreavey, Csilla Krausz and others for the Danish project, and Nabeel Affara, Patricia Blanco and Carole Sargent for the *AZFa* project.

REFERENCES

BLANCO P, SHLUMUKOVA M, SARGENT CA, JOBLING MA, AFFARA N, HURLES ME (2000) Divergent outcomes of intra-chromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. J Med Genet 37:752-758

HURLES ME, IRVEN C, NICHOLSON J, TAYLOR PG, SANTOS FR, LOUGHLIN J, JOBLING MA, SYKES BC (1998) European Y-chromosomal lineages in Polynesia: a contrast to the population structure revealed by mitochondrial DNA. Am J Hum Genet 63:1793-1806

JOBLING MA, TYLER-SMITH C (1995) Fathers and sons: the Y chromosome and human evolution. Trends Genet 11:449-456

JOBLING MA, SAMARA V, PANDYA A, FRETWELL N, BERNASCONI B, MITCHELL RJ, GERELSAIKHAN T, DASHNYAM B, SAJANTILA A, SALO PJ, NAKAHORI Y, DISTECHE CM, THANGARAJ K, SINGH L, CRAWFORD MH, TYLER-SMITH C (1996) Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. Hum Mol Genet 5:1767-1775

JOBLING MA, PANDYA A, TYLER-SMITH C (1997) The Y chromosome in forensic analysis and paternity testing. Int J Legal Med 110:118-124

JOBLING MA, WILLIAMS G, SCHIEBEL K, PANDYA A, MCELREAVEY K, SALAS L, RAPPOLD GA, AFFARA NA, TYLER-SMITH C (1998) A selective difference between human Y-chromosomal DNA haplotypes. Curr Biol 8:1391-1394

JOBLING MA, HEYER E, DIELTJES P, DE KNIJFF P (1999) Y-chromosome-specific microsatellite mutation rates re-examined using a minisatellite, MSY1. Hum Mol Genet 8:2117-2120

JOBLING MA, TYLER-SMITH C (2000) New uses for new haplotypes: the human Y chromosome, disease, and selection. Trends Genet 16:356-362

JOBLING MA (2001) In the name of the father: surnames and genetics. Trends Genet 17:353-357

KALAYDJIEVA L, CALAFELL F, JOBLING MA, ANGELICHEVA D, DE KNIJFF P, ROSSER ZH, HURLES ME, UNDERHILL P, TOURNEV I, MARUSHIAKOVA E, POPOV V (2001) Patterns of inter- and intra-group genetic diversity in the Vlax Roma as revealed by Y chromosome and mitochondrial DNA lineages. Eur J Hum Genet 9:97-104

KAWAGUCHI TK, SKALETSKY H, BROWN LG, MINX PJ, CORDUM HS, WATERSTON RH, WILSON RK, SILBER S, OATES R, ROZEN S, PAGE DC (2001) The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet 29:279-286

KAYSER M, ROEWER L, HEDMAN M, HENKE L, HENKE J, BRAUER S, KRÜGER C, KRAWCZAK M, NAGY M, DOBOSZ T, SZIBOR R, DE KNIJFF P, STONEKING M, SAJANTILA A (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66:1580-1588

KRAUSZ C, QUINTANA-MURCI L, RAJPERT-DE MEYTS E, JORGENSEN N, JOBLING MA, ROSSER ZH, SKAKKEBAEK NE, MCELREAVEY K (2001) Identification of a Y chromosome haplogroup associated with reduced sperm counts. Hum Mol Genet 10:1873-1877

MENOZZI P, PIAZZA A, CAVALLI-SFORZA LL (1978) Synthetic maps of human gene frequencies in Europeans. Science 201:786-792

ROSSER ZH, ZERJAL T, HURLES ME, ADOJAAN M, ALAVANTIC D, AMORIM A, AMOS W, et al (2000) Y-chromosomal diversity within Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 67:1526-1543

SAILLARD J, FORSTER P, LYNNERUP N, BANDELT H-J, NØRBY S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718-726

SEIELSTAD MT, MINCH E, CAVALLI-SFORZA LL (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20:278-280

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 290:1155-1159

SINCLAIR AH, BERTA P, PALMER MS, HAWKINS JR, GRIFFITHS B, SMITH MJ, FOSTER JW, FRISCHAUF A-M, LOVELL-BADGE R, GOODFELLOW PN (1990) A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Nature 346:240-244

SYKES B, LEIBOFF A, LOW-BEER J, TETZNER S, RICHARDS M (1995) The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am J Hum Genet 57:1463-1475

TABERLET P, FUMAGALLI L, WUST-SAUCY AG, COSSON JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. Mol Ecol 7:453-464

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 USA 97:7360-7365

VOGT PH, EDELMANN A, KIRSCH S, HENEGARIU O, HIRSCHMANN P, KIESEWETTER F, KÖHN FM, SCHILL WB, FARAH S, RAMOS C, HARTMANN M, HARTSCHUH W, MESCHEDE D, BEHRE HM, CASTEL A, NIESCHLAG E, WEIDNER W, GRÖNE H-J, JUNG A, ENGEL W, HAIDL G (1996) Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. Hum Mol Genet 5:933-943