

OPINION

Genetic ancestry and the search for personalized genetic histories

Mark D. Shriver and Rick A. Kittles

Public demand and the development of large public and private databases of genetic information across human populations has encouraged the development of the new and rapidly growing field of genetic ancestry testing. Both the promise of the science that underlies this field and a lack of a full understanding of its limitations have fuelled the increased public interest in genetic ancestry testing.

The popularity of ancestry and genealogical research has grown rapidly over the past 10 years. Recently called “America’s latest obsession”¹, genealogical research has become the fastest growing hobby in many communities in the United States. Increasingly, DNA technology is being used to supplement historical documents for researching genealogy. Genetic data in the form of DNA polymorphisms sampled from different human populations are powerful tools for inferring human population history, exploring genealogy and estimating individual ancestry.

The estimation of personalized genetic histories (PGHs) is not just a pastime. The diverse genetic origins of human populations (see FIG. 1) have created challenges for the medical and social sciences that are also driving interest in understanding and defining PGHs. For example, estimates of PGHs can be used to adjust for POPULATION and ADMIXTURE STRATIFICATION^{2,3} to deconvolute environmental and genetic effects on complex diseases. Moreover, PGH can have a role in medical

risk calculation, ADMIXTURE MAPPING⁴⁻⁷, forensic investigations (see below), PERSONAL INTEREST GENOMICS and the assessment of ancestry for socio-political purposes (for example, adoption record access, affirmative action qualifications, Native American tribal affiliation). For whatever purpose, it is clear that an increasing number of members of New World populations are seeking more information on their Old World ancestries. Perhaps the most prominent example is the desire of many African–Americans to identify their ancestral communities⁸⁻¹⁰ (see BOX 1). However, many others are also eager to learn more about

their Jewish, European, Asian, African and/or Native American ancestors¹¹.

Genetic ancestry testing is seen by some to be controversial¹²⁻¹⁵. For the most part, these controversies have more to do with the complex history of ‘race’, discrimination and injustice than the science behind PGH. Regardless of these perceived controversies, the wide public interest in PGH and the uses of genetic ancestry tracing highlight the need for an overview of this area. Moreover, the burgeoning number of PGH companies that now offer fee-for-service tests for genetic ancestry (see TABLE 1) indicates the timeliness of an overview of the issues. Here, we aim to highlight why there is such strong public interest in PGH, and to discuss some of the misconceptions about this rapidly expanding field, as well as its limitations. First, we summarize the two principal analytical approaches being applied for the estimation of PGH. We then discuss the constraints and limitations placed on PGH estimation and the issues that surround its integration into society before concluding with an assessment of the future of the field.

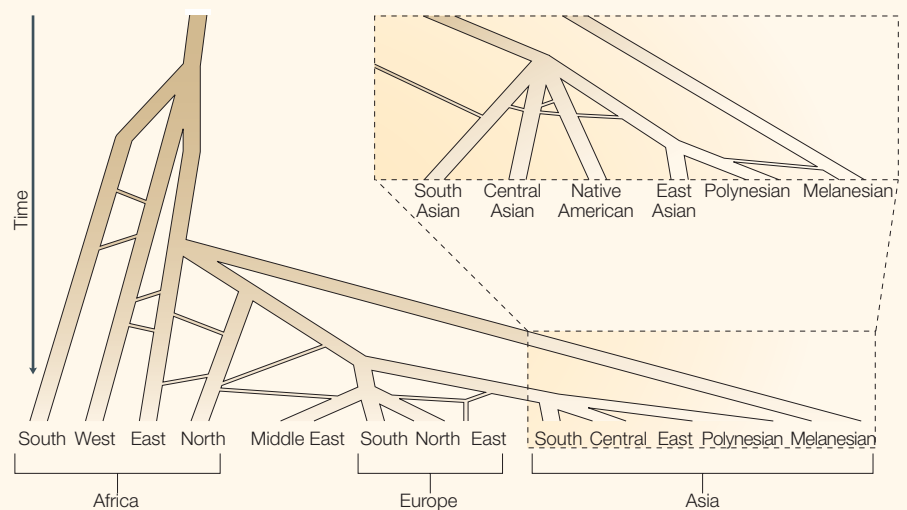


Figure 1 | **A schematic tree of evolutionary relationships among the principal human population groups.** Primary branch points in human evolution are shown, as well as admixture and gene-flow events (indicated as horizontal or slightly sloping connections between adjacent branches).

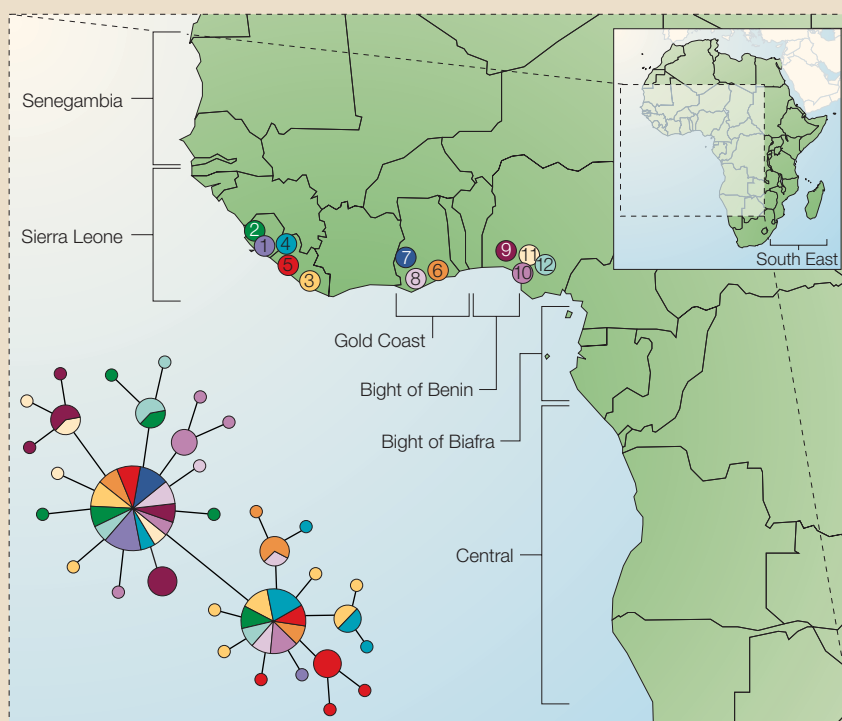
Box 1 | Inferring the genetic ancestry of African-Americans

From approximately 1619 to 1850, tens of millions of indigenous West and Central Africans from 7 coastal regions (see figure) were kidnapped and transported to the Americas⁵⁴. This process led to significant aspects of the history, identity and culture of the enslaved Africans being wiped away for succeeding generations. Records of births and deaths during the period of slavery are substandard at best and non-existent in most cases⁴³. Genetic information has proved to be a useful way of supplementing these inadequate historical documents.

European shipping and naval records provide detailed information on the companies and organizations that were involved in the transatlantic slave trade. Based on these records, personalized genetic history (PGH) companies have been able to target regional populations from which Africans were enslaved and to gather information on maternal and paternal lineages within these populations. Mitochondrial DNA (mtDNA, which defines maternal lineages) and polymorphic Y-chromosome DNA markers (which define paternal lineages) have both been extremely powerful tools for forensic and genealogical studies (see BOX 2), and so were the logical source of genetic markers for this task. However, until

recently, little was known about mtDNA and Y-chromosome variation in West and Central African populations^{27,49,55–62}. Moreover, migration within Africa over the past 400 years means that mtDNA and Y-chromosome lineages found in these populations now do not necessarily reflect those present at the time of enslavement. For this reason, the best that PGH analyses can do is to link marker lineages that are found in African-Americans to present-day African populations.

Even this task is not straightforward given the number of marker lineages that are shared among different African populations. For example, the figure shows the phylogeny of several West African Y-chromosome HAPLOTYPES (belonging to HAPLOGROUP E^{63,64}), many of which are shared among different ethnic groups. Each haplotype is shown as a circle that is proportional in size to its frequency in the region and coloured to reflect the ethnic groups in which it is found (see map). Several common haplotypes (~3–5% in frequency) are found in all the populations in the region. However, the rarer haplotypes are restricted to particular ethnic groups. Some groups, such as the Mende (1) and Temne (2) in Sierra Leone, are similar and share several paternal lineages. Populations are numbered as follows: 1, Mende; 2, Temne; 3, Kru; 4, Kissi; 5, Basso; 6, Ga; 7, Akan; 8, Fante; 9, Yoruba; 10, Itsekiri; 11, Urhobo; 12, Edo (M.B. Doura and R.A.K., unpublished observations).



Lineage-based analyses

PGH tests can be divided into two types: lineage-based tests, which target mitochondrial DNA (mtDNA) and the non-recombining Y chromosome (NRY), and autosomal marker-based tests, which use ANCESTRY INFORMATIVE MARKERS (AIMs) to estimate BIOGEOGRAPHICAL ANCESTRY (BGA). At present, most PGH companies use lineage-based approaches (see TABLE 1).

Maternally (mtDNA) and paternally (NRY) inherited DNA have been useful for studying human evolution and genealogical inference^{16–18}. These sources of markers, which are not subject to recombination, define haplotypes that can easily be used to reconstruct maternal and paternal lineages. They also have a lower EFFECTIVE POPULATION SIZE than autosomal loci and so are more sensitive to the effects of genetic drift, which makes them better markers of population structure. mtDNA has the

added advantage that it evolves more rapidly than nuclear DNA because it has a higher nucleotide substitution rate.

At present, more than 400 well-characterized polymorphic loci on the NRY can be analysed using PCR¹⁹. These include SNPs, *Alu* insertion/deletion polymorphisms²⁰ and microsatellites^{21,22}. Although the informative stretches of mtDNA and NRY represent less than 1% of the entire human genome, they have proved to be powerful tools for identifying and defining maternal and paternal lineages that have improved our understanding of human migrations and solidified our appreciation that ultimately all humans share a small group of recent common ancestors.

Maternal and paternal lineages. Lineage-based ancestry tests are popular because mtDNA and NRY haplotypes can provide information that is regionally specific^{19,23–28}

(see BOX 2). For this reason, companies can now offer tests to determine whether an individual has paternal or maternal lineages that originate from Native American, European, African or Asian populations (see TABLE 1). The process is straightforward: DNA is extracted from buccal cells that are collected on swabs, and informative genetic markers are sequenced or genotyped (BOX 2). The genetic markers are then compared with a reference database of haplotypes that have been identified in specific populations to search for a match (BOX 3).

The accuracy of lineage matching depends on the size and sampling of the database that is used to match mtDNA or Y-chromosome lineages to particular populations or geographical regions. The level of geographical resolution depends on both the sampled haplotype and the populations included in the database (BOX 2; see also REF. 29). Many databases that

Table 1 | Information on personalized genetic history companies and groups

Company/organization	Web site	Products	Types of marker	Number of markers	Database access
African Ancestry	www.africanancestry.com	Paternal lineage (PatriClan) Maternal lineage (MatriClan)	YSTRs mtDNA HVSI	10 N/A	Private Private
DNA Heritage	www.dnaheritage.com	Paternal (surname) lineage	YSTRs	37	Public
DNA Print	http://www.ancestrybydna.com	Biogeographical ancestry	Autosomal SNPs	175	Private
Family Tree	www.familytreedna.com	Paternal (surname) lineage Maternal lineage Native American maternal lineage Native American paternal lineage Y search database	YSTRs mtDNA HVSI and HVSII mtDNA HVSI and HVSII YSTRs YSTRs	37 N/A N/A 37 37	Private Private Private Private Public
Gene Tree*	www.genetree.com	Paternity testing Sibling testing Native American maternal lineage Native American paternal lineage	YSTRs mtDNA HVSI and HVSII mtDNA HVSI and HVSII YSTRs	37 N/A N/A 37	N/A N/A Private Private
International Forensic Y-User Group	www.ystr.org	N/A	YSTRs	11	Public
Oxford Ancestors	www.oxfordancestors.com	Paternal lineage (Y-Line) Maternal lineage (MatriLine)	YSTRs mtDNA HVSI	10 N/A	Private Private
Relative Genetics*	www.relativegenetics.com	Paternal (surname) lineage Maternal lineage Native American maternal lineage Native American paternal lineage	YSTRs mtDNA HVSI and HVSII mtDNA HVSI and HVSII YSTRs	37 N/A N/A 37	Private Private Private Private
Roots for Real	www.rootsforreal.com	Maternal lineage	mtDNA HVSI and HVSII	N/A	Private
Sorenson Molecular Genealogy Foundation	www.smgf.org	N/A	YSTRs	24	Controlled public
Trace Genetics	www.tracegenetics.com	Native American maternal lineage	mtDNA HVSI and HVSII	N/A	Private

*Subsidiaries of Sorenson Genomics. HVSI, HVSII, hypervariable segment I, II; LLC, limited liability company; mtDNA, mitochondrial DNA; N/A, not applicable; YSTRs, Y-chromosome short tandem repeats.

are derived from published research are too small and lack samples in certain geographical regions. Nonetheless, the high regional specificity of mtDNA and Y-chromosome haplotypes means that some less common lineages can be traced to particular ethnic groups or locales^{29,30}. However, tracing the more common haplotypes to a particular location is problematic (see BOX 1 and BOX 3).

Surname matching. NRY lineages are correlated with surnames^{31,32}, so these lineages can be compared among individuals with the same or similar surnames to identify previously unconnected relatives. The potential to do such surname matching has led to an increased interest in establishing PGH databases for this purpose (see TABLE 1). For many of these databases, identifiers linked to each surname are used to allow the person who submitted the surname and genetic information to be contacted anonymously. For many individuals, these databases can provide information about specific ancestors to help them get past 'brick walls' in their genealogies.

The standard surname-matching procedure for PGH companies is to compare markers on the Y chromosome of two or more men to determine relatedness and, if possible, an estimate of when their most recent common ancestor (MRCA) lived^{16,33}

(see BOX 3). However, there is no consensus among these companies about whether people in surname databases who share the same lineage can be contacted. The few publicly accessible databases (TABLE 1) maintain strict confidentiality of participants' information when linking individuals together in genetic genealogies, whereas some private surname-matching databases are not so well monitored.

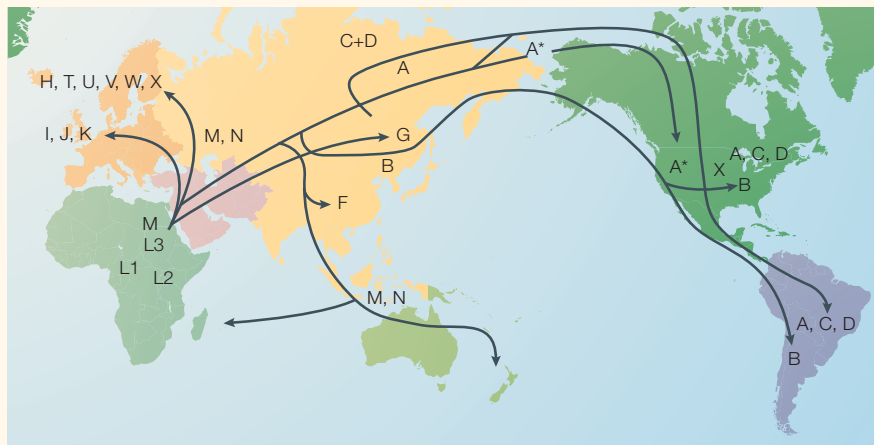
Biogeographical ancestry analyses

One major disadvantage of PGH estimates that use lineage-based analyses is that they focus on a single maternal or paternal lineage and therefore neglect the contribution of the vast majority of an individual's ancestors to their genome (see below). One solution to this problem is to use many autosomal genetic markers that distinguish among ancestral groups. Ancestry informative markers (AIMs; also known as population-specific alleles (PSAs)^{34,35}, ethnic difference markers (EDMs)³⁶ and mapping by admixture linkage disequilibrium (MALD) markers^{37,38}) are autosomal genetic markers that show substantial differences in allele frequency across population groups. These groups can range from relatively local clusters (for example, Southern European/Northern European) to larger continental distinctions (for example, African/non-African). Allele-frequency

databases that catalogue AIMs have lagged far behind other genomic databases. Nonetheless, there has been a tenacious set of researchers that has searched for and developed AIMs^{36–39}. The main impetus for the development of these AIM panels has been the promise of admixture mapping^{40,41}.

Biogeographical ancestry (BGA) is an expression coined to describe the aspects of PGH that can be computed using the ancestry information contained in AIMs³. Factors such as isolation by distance, range expansions, land bridges, maritime technologies, ice ages, and cultural and linguistic barriers have all affected human migration and mating patterns in the past and have therefore shaped the present worldwide distribution of genetic variation⁴². BGA in the broad sense is the quantitative representation of the effects of these factors. Specifically, BGA aims to estimate a person's ancestry in terms of the proportional representation of AIMs from a selection of ancestral populations (see, for example, BOX 4). PGH analyses are based on statistical calculations whereby confidence intervals can and should be estimated to represent their precision or lack thereof (see BOX 3). Clear calculation and presentation of the statistics and confidences of PGH measures is an important challenge that we feel PGH companies need to address.

Box 2 | Sex-linked markers and the stories that they tell



Phylogenetic methods are routinely used to identify human mitochondrial DNA (mtDNA) and Y-chromosome haplogroups and the informative genetic markers that define them. Most personalized genetic history (PGH) tests for maternal ancestry involve sequencing approximately 400 bp of hypervariable segment I of the mitochondrial DNA CONTROL REGION and confirming the haplogroup through diagnostic SNP typing. mtDNA haplogroups are continent-specific, with little mixing of mtDNA haplogroups from different continents³³.

mtDNA analyses provided our first genetic window into the past, detailing the history of maternal lineages across human populations. The oldest mtDNA haplogroups are found in Africa (L1, L2 and L3; see figure). The pan-African haplogroup L3 radiated to form MACROHAPLOGROUPS M and N. Macrohaplogroups M and N arose in North Eastern Africa, and individuals that had M and N mtDNA subsequently left Africa to colonize Europe and Asia ~60,000–80,000 years ago^{61,62,65}. Haplogroups H, I, J, N1b, T, U, V, W and X are mainly derived from macrohaplogroup N and make up almost all of the mtDNA types found in Europe⁶⁶. In Asia, macrohaplogroups N and M radiated to generate mtDNA lineages A, B, C, D, F and G. Native Americans are known to have Asian ancestry because only five haplogroups (A, B, C, D and X) encompass all of the mtDNA variation in the New World, four of which came from Asia⁶⁷.

Increasingly, Y-specific polymorphisms, such as SNPs^{27,64}, microsatellites^{22,26,68–72} and combinations of these, are being used to trace male lineages in the same way that mtDNA has been used to trace female lineages^{48,49,58,60}. Compound Y-chromosome haplotypes take advantage of the high resolution of microsatellites in defining individual lineages in addition to well-characterized binary SNPs, which anchor short tandem repeat (STR)-defined lineages within a larger set of Y-chromosome haplogroups. There are 18 main Y haplogroups defined by the Y-Chromosome Consortium (YCC) (see REFS 19,64 for reviews). Similar to mtDNA, the oldest haplogroups are found in Africa (haplogroups A and B). Haplogroups C, D and H are found in Asia and in low frequency in the Americas, whereas haplogroup E is common in sub-Saharan Africa and some parts of the Mediterranean^{49,58}. Haplogroups G, H and I are mainly European. Haplogroups J and F are found in Africa, Asia and Europe. Haplogroups L, M, N, O, P and Q have been observed in non-African populations from Asia, Europe and the Americas. Haplogroup R is observed in Europe, Asia, Oceania, the Americas and West Africa. Figure courtesy of D.T. Schulthess, Gene Tree.

One notable practical application of BGA methods is in forensic investigations in which ancestry estimates might provide police with vital clues and help to direct investigations. A recent example from the United States highlights the usefulness of these methods. By May 2003, the murders of five women in Louisiana had been linked through the analysis of forensic short tandem repeat (STR) marker panels; however, there were no hits when compared against the national Combined DNA Index System (CODIS) database of convicted felons. For months, on

the basis of eyewitness accounts and psychological profiles, the police had restricted their investigations to white men, screening more than 600 men in various DNA-based dragnets. Late in the investigation, the state police sought the assistance of a company, the services of which include commercially available PGH tests. The results of these tests revealed that the perpetrator was mainly of West African descent. This information caused the police to reconsider their focus: they broadened their investigation to include non-white suspects. Previously

overlooked leads that indicated that an African–American man might be involved were re-examined. This change in direction ultimately led to new evidence and the arrest of a suspect whose STR profiles matched those of the perpetrator.

Constraints and limitations

Three main factors determine the accuracy and precision of PGH estimates: the number of markers used, the quality of the reference databases (that is, the geographical spread and number of samples included) and the levels of genetic differentiation among populations in the region(s) being considered.

Markers and references. In general, the more markers used, the more reliable are the PGH estimates. The decreasing expense of genotyping and sequencing technologies has allowed many PGH companies to increase the number of markers that they use. However, the addition of new markers also means that the database of reference samples should be typed for these new markers. Ideally, a standard set of genetic markers would be used to allow comparisons with data from the scientific literature and other public and private databases. To some extent, this is already happening: there is certainly significant overlap in markers that different companies use (TABLE 1).

Lineage-based PGH companies use reference databases that range in size from a few thousand to tens of thousands of haplotypes. Most companies do not share database information. However, there have been several partnerships built around reference databases for specific populations, such as West Africans and Native Americans. Nonetheless, these partnerships are few and there is definitely room for increased sharing of reference-database information among companies and between the companies and their customers. In particular, most PGH companies do not provide details about the number and geographical spread of samples included. So, in general, it is difficult to assess exactly what limitations the quality of these databases put on PGH estimation.

Lineage-based analyses. Lineage-based products (mtDNA and NRY) provide a high level of ancestry information due to the lack of recombination, high mutation rates and more restricted geographical distributions compared with autosomal markers (see BOX 2). However, as previously mentioned, such tests provide information about only one line of descent (that is, one ancestor per generation) out of many that contributed to the present genetic make-up of an individual.

Box 3 | Statistical considerations

A detailed discussion of the statistical details that underlie personalized genetic history (PGH) analyses is beyond the scope of this article. Nonetheless, PGH is basically an application of population genetics, a highly statistical field of science. As such, we believe that most PGH companies are yet to properly address the uncertainties that surround PGH estimates and how these should be presented to their current and potential customers. For example, even a perfect match between a customer and a sample in the reference database is not a straightforward result. A tempting conclusion to reach from such a match would be that the customer shares an ancestor with the reference sample donor in the past few generations. However, such a conclusion is much too strong: many more markers than are typically used (or even available) are needed to estimate the maximum number of past generations in which two matching mtDNA or non-recombining Y chromosome (NRY) haplotypes could share an ancestor³³.

Another statistical problem is posed when haplotypes are found in multiple geographical locations (for example, many of the haplotypes in BOX 1). In such cases, statistics such as likelihood ratio tests can be used to compare the probability that the source is population 1 versus population 2. The test statistic is often calculated as a ratio of the haplotype frequency (genotype frequency for diploid markers) in population 1 divided by the genotype frequency in population 2. The larger this ratio, the greater the support is for population 1 being the original source of the haplotype. Clearly, such comparisons will not often produce high levels of confidence when only one marker is used. However, when multiple unlinked ancestry informative markers (AIMs) are available, such an approach can be useful³⁵. An extension of this method is the application of maximum likelihood to biogeographical ancestry (BGA) estimation, whereby instead of comparing two populations, a whole series of populations that differ in admixture levels are compared simultaneously. The admixed population in which the multilocus AIM genotype is most frequent is identified as the maximum-likelihood estimate. Support intervals can then be calculated as selections of admixed populations that have likelihood estimates within a specified interval of the maximum (see BOX 4).

Such examples illustrate that PGH estimation is far from being an exact science. For this reason, companies that offer PGH products should always present their data with confidence intervals on the estimates and provide statistical tutorials on how the estimates are calculated and what they mean.

For example, if you go back 14 generations (approximately 350 years), each person has a maximum of 16,384 direct ancestors (the actual number of ancestors might be much smaller as this calculation assumes infinite population size and non-consanguineous mating).

For many customers of lineage-based tests, there is a lack of understanding that

their maternal and paternal lineages do not necessarily represent their entire genetic make-up. For example, an individual might have more than 85% Western European 'genomic' ancestry but still have a West African mtDNA or NRY lineage. In addition, lineage-based tests performed within a single family might reveal multiple lineages from diverse ethnic groups. Several lineage-based

PGH companies now use professional genealogists to help to identify living relatives with different maternal/paternal lineages that can provide information about ancestors who have contributed autosomal DNA (but not mtDNA/NRY) to the present genetic make-up of a customer (see REFS 44,45 for a discussion of traditional methods of tracing family history). However, we suggest that these companies need to put more effort into disclosing the types of result that a potential customer could expect to receive from such approaches and the statistical precision and power of any conclusions.

Biogeographical ancestry analyses. Similar to lineage-based analyses, BGA-based PGH estimates need to be considered in the context of the different potential sources of the genetic variants assayed. In particular, the less restricted geographical distributions of autosomal markers, relative to the sex-linked markers that are used for lineage-based analyses, mean that many potential ancestral patterns will be consistent with any given result. For example, a person might show 75% West African and 25% Western European ancestry in a BGA estimate because three grandparents are from West Africa and one is from Western Europe or because all four grandparents are of East African ancestry. Similarly, both recent and ancient admixture events can lead to intermediate BGA estimates in such analyses. For this reason, such results should be carefully and clearly presented and interpreted for customers, as should the parental population groups that provide the reference points for such analyses (see BOX 4). In particular, PGH providers should not misrepresent how informative such tests can be.

Glossary

ADMIXTURE MAPPING

(Mapping by admixture linkage disequilibrium). Mapping of genes for traits or diseases that have different genetic risks in two or more populations that have admixed recently to form a third hybrid population.

ADMIXTURE STRATIFICATION

Heterogeneity in a population that is composed of recent admixture of different ethnic groups that differ in marker allele frequencies.

ANCESTRY INFORMATIVE MARKERS

Genetic markers that show substantial differences in allele frequency across population groups.

BIOGEOGRAPHICAL ANCESTRY

The aspects of personalized genetic histories that can be computed using the ancestry information that is contained in ancestry informative markers.

CONTROL REGION

The section of mitochondrial DNA that does not code for any proteins. It consists of two segments (I and II) in which mutations are especially frequent.

EFFECTIVE POPULATION SIZE

The size of the ideal population in which the effects of random drift would be the same as observed in the actual population.

GENETIC DETERMINISM

The idea that genes determine and shape everything about a person.

HAPLOGROUP

(Also known as macrolineage). A group of haplotypes that share common ancestry defined by shared sequence.

HAPLOTYPE

A specific segment of DNA sequence that is inherited as a unit.

MACROHAPLOGROUPS

A group of haplogroups that are closely related and share a recent common ancestor.

ONE-DROP RULE

(Also known as the rule of hypodescent). A socially constructed classification system that was initiated during slavery in America that proclaimed that any person with any known African 'black' ancestry would have the same legal status as a 'pure' African.

PERSONAL INTEREST GENOMICS

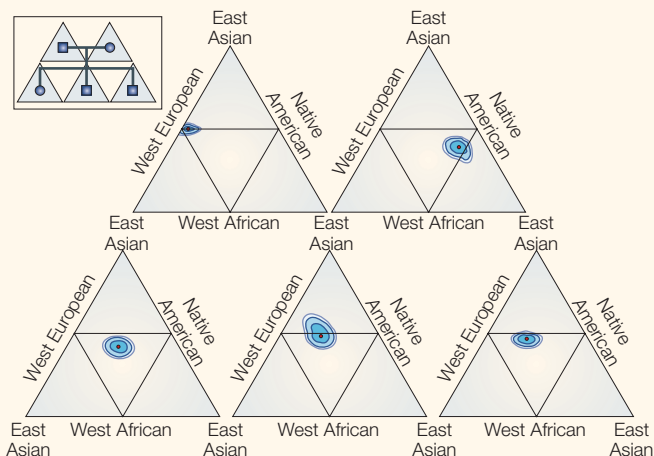
The personal or recreational use of genetic ancestry information.

POPULATION STRATIFICATION

A population that contains subpopulations that differ in marker allele frequencies owing to nonrandom mating, finite population size and/or geographical barriers.

Box 4 | An example of biogeographical ancestry estimation

In this example of biogeographical ancestry (BGA) estimates for 5 people from 1 family (see pedigree in inset), a database of 176 ancestry informative markers (AIMs) is used to calculate 4 possible 3-way ancestry models with respect to the 4 parental populations included (West African (WA), Western European (WE), East Asian (EA) and Native American (NA)).



These four models (WA/WE/EA, WA/WE/NA, WA/EA/NA and WE/EA/NA) are calculated separately and then combined into an unfolded tetrahedron projection in which each of the four surfaces (inner triangles) is one of the models. The red dot indicates the most probable position for the multilocus AIM genotype of the person in question. To present information on the precision of the result, concentric rings delineate probability spaces relative to the maximum-likelihood estimate (2X, 5X and 10X). For example, all of the values within the first (smallest) confidence ring are less than two times less likely than the maximum. These two-dimensional representations can be cut out and folded into three-dimensional tetrahedrons (four-sided pyramids).

The analysis illustrated correctly shows that the father is primarily of European descent, whereas the mother has substantial Native American ancestry and smaller amounts of Western European and West African ancestry. As we would expect, the three children are intermediate between the two parents and similar to one another. Also evident is some level of variability among siblings that is expected in recently admixed people as a result of the independent assortment of large non-recombined chromosomal segments showing consistent ancestry. Figure courtesy of T. Frudakis, DNA Print Genomics.

used to supplement rather than supplant traditional genealogical tracing methods. The use of PGH to encourage a better understanding of family history has the potential to inform individuals about important issues such as hereditary diseases as well as to help to educate the public about the extent and meaning of genetic variation and human genetics in general.

Although in many ways, the genetic analyses that PGH companies provide are just a supplement to traditional methods, these companies must be aware of the ethical and psychological implications of the services that they offer. Specifically, if an individual regards the meaning and significance of their genetic ancestry highly, the results of PGH tests could deeply affect them⁴⁵. So, it might be important for PGH providers to provide useful psychological resources, such as access to a clinical psychologist who specializes in identity and psychological well-being counselling. In addition to the obligations that PGH companies might have to the psychological well-being of their customers, PGH testing raises several other ethical issues that have been addressed in detail elsewhere^{8,12,14,45}.

The future of PGH

There are already many high-profile examples of studies of genetic history. These include studies that show that many Southern African Lemba people share ancestry with Jewish priests⁴⁷, that some European populations have multiple origins^{48–50}, that Thomas Jefferson, or one of his male relatives, fathered offspring with Sally Hemmings, his enslaved servant⁵¹, and that some African-Americans are highly admixed with Europeans and differ in population substructure^{34,52}. The appeal of such studies to many people who are interested in knowing more about their ancestry means that, in the future, PGH analyses might increasingly influence how individuals and groups define themselves. However, the inherent limitations in the informativeness of various databases might compromise future efforts to achieve optimal reliability and specificity in these analyses. Specifically, improvements to PGH tests will require: larger databases of individuals for parental frequency estimates and lineage comparisons; larger panels of AIMs; and further development of analytical methods and models such as estimating time since admixture^{5,53}. In addition to these technical improvements, PGH companies must be prepared to meet the highest possible standards in communicating with their customers. The accuracy, meaning and applications of PGH tests and their results must be clearly explained to these clients.

PGH and society

PGH has the potential to affect ethnic, religious, 'racial' and family identities. Many of these types of identity overlap and might potentially conflict when individuals attempt to reconcile their social identities with their genetic histories. In particular, non-paternity, adoption and multiple marriages are factors that can disengage PGH from social identity. However, these issues are not new, nor are they unique to PGH: traditional methods of estimating ancestry and identity also have to deal with the same problems.

One potential negative consequence of PGH testing is that the genetically defined ancestral categories that PGH companies use could be misinterpreted as indications of 'real' racial divisions, even if they are explicitly acknowledged as being continuous and, to some extent, arbitrary groups (see article by Bamshad *et al.*⁴⁶ in this issue for a detailed discussion of these issues). Underlying such misconceptions is a strong undercurrent of GENETIC DETERMINISM that is found in the public in general and is shared by some who work in the field and even some of the detractors

of PGH. The challenge for PGH companies is to perform and present PGH tests in a way that educates the public and dispels such misconceptions.

The importance of, and interest in, PGH analyses varies among groups and individuals according to the values of the group and the individual. For example, the African component of the ancestry of African-Americans has, in general, socially defined an individual in this group, regardless of the number of non-African ancestors that they might have (the ONE-DROP RULE). As a result, some African-Americans want to use PGH analyses to reclaim their non-African ancestry. As this example illustrates, for some groups, PGH can be valued as another form of evidence that can be used to reclaim history, culture and knowledge that is denied to individuals^{10,11,43,44}.

Although PGH is challenging, controversial and has many limitations, it is important. PGH analyses will almost certainly cause some people to change the way that they think about themselves and their identities. However, it is probable that PGH tests will be

How do we ensure that PGH providers meet acceptable standards in practice? We suggest that organizations such as the American Association of Blood Banks for Parentage Testing and the National Forensic Science and Technology Center (International Organization for Standardization (ISO) accreditation) should jointly develop a code of ethical conduct for the PGH industry and should be responsible for accreditation of PGH laboratories. It is probable that this embryonic industry will continue to raise new ethical, social and psychological issues in the future. However, the immediate challenge for PGH providers is to clearly articulate to the public the promise and the limitations of the science.

In summary, the strong public interest in genetic ancestry is grounded in historical and contemporary experiences. The identification of informative DNA markers in human populations indicates that PGH analyses will be able to address some of the many questions that individuals have about their ancestry. However, the most reliable method for inferring ancestry is to combine DNA evidence with other forms of genealogical and historical knowledge (that is, family and court documents). We recommend this mosaic approach to inferring ancestry as it clearly recognizes that identity is not simply that which is ascribed, based on phenotype (skin colour and facial features) or quasi-genetics (the one-drop rule), but is multi-determined.

Mark D. Shriver is at the Department of Anthropology, Penn State University, University Park, Pennsylvania 16802, USA.

Rick A. Kittles is at the Department of Molecular Virology, Immunology and Medical Genetics, Ohio State University, Columbus, Ohio 43210, USA.

Correspondence to M.D.S. and R.A.K. e-mails: mds17@psu.edu; kittles-1@medctr.osu.edu

doi:10.1038/nrg1405

- Hornblower, M. Roots mania. *Time* 153 (19 Apr 1999).
- Hoggart, C. J. *et al.* Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* **72**, 1492–1504 (2003).
- Shriver, M. D. *et al.* Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum. Genet.* **112**, 387–399 (2003).
- Smith, M. W. *et al.* A high-density admixture map for disease gene discovery in African Americans. *Am. J. Hum. Genet.* **74**, 1001–1013 (2004).
- Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979–1000 (2004).
- McKeigue, P. M. 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**, 241–251 (1998).
- McKeigue, P. M. Multipoint admixture mapping. *Genet. Epidemiol.* **19**, 464–467 (2000).
- Lee, S. S., Mountain, J. & Koenig, B. A. The meanings of 'race' in the new genomics: implications for health disparities research. *Yale J. Health Policy Law Ethics* **1**, 33–75 (2001).
- Baylis, F. Blacks as me: narrative identity. *Developing World Bioeth.* **3**, 142–150 (2003).
- Rotimi, C. N. Genetic ancestry tracing and the African identity: a double-edged sword? *Developing World Bioeth.* **3**, 151–158 (2003).
- Brown, K. Tangled roots? Genetics meets genealogy. *Science* **295**, 1634–1635 (2002).
- Elliott, C. & Brodwin, P. Identity and genetic ancestry tracing. *BMJ* **325**, 1469–1471 (2002).
- Johnston, J. & Thomas, M. Summary: the science of genealogy by genetics. *Developing World Bioeth.* **3**, 103–108 (2003).
- Zoloth, L. Yearning for the long lost home: the Lemba and Jewish narrative of genetic return. *Developing World Bioeth.* **3**, 128–132 (2003).
- Johnston, J. Resisting a genetic identity: the black Seminoles and genetic tests of ancestry. *J. Law Med. Ethics* **31**, 262–271 (2003).
- Stumpf, M. P. & Goldstein, D. B. Genealogical and evolutionary inference with the human Y chromosome. *Science* **291**, 1738–1742. (2001).
- Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
- Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
- Jobling, M. A. & Tyler-Smith, C. The human Y chromosome: an evolutionary marker comes of age. *Nature Rev. Genet.* **4**, 598–612 (2003).
- Hammer, M. F. A recent insertion of an *alu* element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol.* **11**, 749–761. (1994).
- Jobling, M. A., Heyer, E., Diehljes, P. & de Knijff, P. Y-chromosome-specific microsatellite mutation rates re-examined using a minisatellite, MSY1. *Hum. Mol. Genet.* **8**, 2117–2120 (1999).
- Kayser, M. *et al.* 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 (2000).
- Budowle, B., Allard, M. W., Wilson, M. R. & Chakraborty, R. Forensics and mitochondrial DNA: applications, debates, and foundations. *Annu. Rev. Genomics Hum. Genet.* **4**, 119–141 (2003).
- Hammer, M. F. *et al.* Hierarchical patterns of global human Y-chromosome diversity. *Mol. Biol. Evol.* **18**, 1189–1203. (2001).
- Hammer, M. F. *et al.* The geographic distribution of human Y chromosome variation. *Genetics* **145**, 787–805. (1997).
- Kayser, M. & Sajantila, A. Mutations at Y-STR loci: implications for paternity testing and forensic analysis. *Forensic Sci. Int.* **118**, 116–121. (2001).
- Underhill, P. A. *et al.* Y chromosome sequence variation and the history of human populations. *Nature Genet.* **26**, 358–361. (2000).
- Wallace, D. C. Mitochondrial DNA sequence variation in human evolution and disease. *Proc. Natl Acad. Sci. USA* **91**, 8739–8746 (1994).
- Stoneking, M. & Soodyall, H. Human evolution and the mitochondrial genome. *Curr. Opin. Genet. Dev.* **6**, 731–736 (1996).
- Soodyall, H., Nebel, A., Morar, B. & Jenkins, T. Genealogy and genes: tracing the founding fathers of Tristan da Cunha. *Eur. J. Hum. Genet.* **11**, 705–709 (2003).
- Sykes, B. & Irven, C. Surnames and the Y chromosome. *Am. J. Hum. Genet.* **66**, 1417–1419 (2000).
- Jobling, M. A. In the name of the father: surnames and genetics. *Trends Genet.* **17**, 353–357 (2001).
- Walsh, B. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics* **158**, 897–912 (2001).
- Parra, E. J. *et al.* Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* **63**, 1839–1851 (1998).
- Shriver, M. D. *et al.* Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Hum. Genet.* **60**, 957–964 (1997).
- Collins-Schramm, H. E. *et al.* Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Am. J. Hum. Genet.* **70**, 737–750 (2002).
- Smith, M. W. *et al.* Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. *Am. J. Hum. Genet.* **69**, 1080–1094 (2001).
- Dean, M. *et al.* Polymorphic admixture typing in human ethnic populations. *Am. J. Hum. Genet.* **55**, 788–808 (1994).
- Pfaff, C. L. *et al.* Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.* **68**, 198–207 (2001).
- Li, W. H. & Nei, M. Stable linkage disequilibrium without epistasis in subdivided populations. *Theor. Popul. Biol.* **6**, 173–183 (1974).
- Chakraborty, R. & Weiss, K. M. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl Acad. Sci. USA* **85**, 9119–9123 (1988).
- Templeton, A. Out of Africa again and again. *Nature* **416**, 45–51 (2002).
- Burroughs, T. *Black Roots: A Beginner's Guide to Tracing the African American Family Tree* (Fireside, New York, 2001).
- Woodtor, D. P. *Finding a Place Called Home: A Guide to African American Genealogy and Historical Identity* (Random House, New York, 1999).
- Winston, C. E. & Kittles, R. A. in *Biological Anthropology and Ethics: From Repatriation to Genetic Identity* (ed. Turner, T.) (SUNY, Albany, in the press).
- Bamshad, M., Wooding, S., Salisbury, B. A. & Stephens, J. C. Deconstructing the relationship between genetics and race. *Nature Rev. Genet.* **5**, 598–609 (2004).
- Thomas, M. G. *et al.* Y chromosomes traveling south: the cohen modal haplotype and the origins of the Lemba—the 'Black Jews of Southern Africa'. *Am. J. Hum. Genet.* **66**, 674–686 (2000).
- Kittles, R. A. *et al.* Dual origins of Finns revealed by Y chromosome haplotype variation. *Am. J. Hum. Genet.* **62**, 1171–1179 (1998).
- Semino, O. *et al.* Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am. J. Hum. Genet.* **74**, 1023–1034 (2004).
- Semino, O. *et al.* The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science* **290**, 1155–1159 (2000).
- Foster, E. A. *et al.* Jefferson fathered slave's last child. *Nature* **396**, 27–28 (1998).
- Parra, E. J. *et al.* Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *Am. J. Phys. Anthropol.* **114**, 18–29 (2001).
- Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G. & McKeigue, P. M. Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.* **74**, 965–978 (2004).
- Curtin, P. D. *The Atlantic Slave Trade: A Census* (University of Wisconsin Press, Wisconsin, 1969).
- Salas, A. *et al.* The making of the African mtDNA landscape. *Am. J. Hum. Genet.* **71**, 1082–1111 (2002).
- Salas, A. *et al.* The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am. J. Hum. Genet.* **74**, 454–465 (2004).
- Cruciani, F. *et al.* Phylogeographic analysis of haplogroup e3b (e-m215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am. J. Hum. Genet.* **74**, 1014–1022 (2004).
- Cruciani, F. *et al.* A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am. J. Hum. Genet.* **70**, 1197–1214 (2002).
- Richards, M. *et al.* Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. *Am. J. Hum. Genet.* **72**, 1058–1064 (2003).
- Scozzari, R. *et al.* Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am. J. Hum. Genet.* **65**, 829–846 (1999).
- Watson, E. *et al.* mtDNA sequence diversity in Africa. *Am. J. Hum. Genet.* **59**, 437–444 (1996).
- Watson, E., Forster, P., Richards, M. & Bandelt, H. J. Mitochondrial footprints of human expansions in Africa. *Am. J. Hum. Genet.* **61**, 691–704 (1997).
- Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12**, 339–348 (2002).
- Hammer, M. F. & Zegura, S. L. The human Y chromosome haplogroup tree: nomenclature and phylogeography of its major divisions. *Annu. Rev. Anthropol.* **31**, 303–321 (2002).
- Quintana-Murci, L. *et al.* Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nature Genet.* **23**, 437–441 (1999).
- Mishmar, D. *et al.* Natural selection shaped regional mtDNA variation in humans. *Proc. Natl Acad. Sci. USA* **100**, 171–176 (2003).
- Malhi, R. S. *et al.* The structure of diversity within New World mitochondrial DNA haplogroups: implications for the prehistory of North America. *Am. J. Hum. Genet.* **70**, 905–919 (2002).

68. Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E. & de Knijff, P. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* **6**, 799–803 (1997).
69. Kayser, M. *et al.* An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *Am. J. Hum. Genet.* **68**, 990–1018 (2001).
70. Roewer, L. *et al.* Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Sci. Int.* **118**, 106–113 (2001).
71. Kayser, M. *et al.* Online Y-chromosomal short tandem repeat haplotype reference database (YHRD) for U.S. populations. *J. Forensic Sci.* **47**, 513–519 (2002).
72. Lessig, R. *et al.* Asian online Y-STR Haplotype Reference Database. *Leg. Med. (Tokyo)* **5** (Suppl 1), S160–S163 (2003).

Acknowledgements

We thank D. T. Schulthess from Sorenson Genomics for providing information on several personalized genetic history companies and T. Fruidakis for access to data from DNAPrint Genomics, and three anonymous reviewers for constructive comments. M.D.S. and R.A.K. are supported by the National Institutes of Health and the Department of Defense.

Competing interests statement

The authors declare competing financial interests: see Web version for details.

Online links

FURTHER INFORMATION

Mark Shriver's web page: <http://www.anthro.psu.edu/biolab>

Access to this links box is available online.