Genome Biology

REVIEW

A world in a grain of sand: human history from genetic data

Vincenza Colonna[1,2], Luca Pagani[1,3], Yali Xue[1] and Chris Tyler-Smith[1]*

Abstract

Genome-wide genotypes and sequences are enriching our understanding of the past 50,000 years of human history and providing insights into earlier periods largely inaccessible to mitochondrial DNA and Y-chromosomal studies.

To see a world in a grain of sand …

William Blake,
Auguries of Innocence

The genome of each individual is a temporary assemblage of DNA segments brought together for a single generation by a combination of chance, ancestry, recombination and natural selection. These segments have different histories because of recombination and can thus provide independent information about ancestry, the focus of this review. However, the ancestry of different segments is not entirely independent. Humans are not a single randomly mating population: we are subdivided, and these subdivisions into bands, tribes, clans, ethnic groups, nations and so on are of great interest to both scientists and non-scientists. Thus the thousands of different genomic segments in any individual do not trace back to ancestors randomly spread around the globe; segment ancestry is constrained by population history. Two non-recombining segments of the genome, mitochondrial DNA (mtDNA) and the Y chromosome, have been used for decades to study genetic histories [1,2]. Sometimes mtDNA and the Y chromosome share the same history, but often they do not, and such differences alert us to some of the complexities of the human past [3]. But mtDNA and the Y chromosome provide only two perspectives. Recent advances in technology provide access to most of the genome, and increasingly to the genomes of companion species. Here, we consider how this wider perspective is beginning to inform our view of human history. We will see that it is possible to probe much further back into the past, into a period in which the uniparental markers are uninformative yet key evolutionary events took place, and even to speculate about when humans might have begun to wear clothes or to start reconstructing the genomics of former populations before their contact with modern expansions.

Genome-wide data can be obtained by either genotyping samples or re-sequencing them. Genotyping provides information about the allelic state of positions in the genome (currently up to five million, mostly single nucleotide polymorphisms, SNPs) that have prior evidence of variability [4]; such studies are relatively low-cost and routine, and have been performed on massive numbers of samples. Whole-genome sequencing, by contrast, provides information about new as well as known variants; the technologies are still developing rapidly [5,6] and have provided our first glimpses of population-scale samples of hundreds of individuals [7]. Here, we discuss how samples have been chosen for studies of human history and some of the resulting sample sets, together with a number of developments in analytical approaches. We also focus on a few case studies, chosen to illustrate how whole-genome analyses compare with conclusions from mtDNA and the Y chromosome analysis, how studying other species informs our understanding of humans, and how genetic analyses themselves compare with conclusions from other sources of insights into history: archaeology, language, oral traditions or written records. We begin with examples based entirely on the analysis of modern human populations, and then move to studies that have used more diverse sources of genetic information: ancient DNA and other species associated with humans.

Sampling

If we wanted to sample an animal or plant species for genetic study, we would probably choose samples at random throughout its range. Human population

*Correspondence: cts@sanger.ac.uk
[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK
Full list of author information is available at the end of the article

sampling is never done like this. 'Related' individuals (meaning 'related in the last two to three generations') are almost always excluded; known recent migrants are also often excluded, or sometimes used as representatives of their source population; populations of particular interest to the samplers are often included. These considerations are illustrated by some of the most widely used sample sets available to the field (Figure 1a-c). We see, for example, the total lack of samples from Australia in these sets, the poor representation of India, and the predominantly migrant and admixed samples collected from the Americas. A representation of the variation in these populations (Figure 2a) reveals additional features of the samples: the high level of variation combined with sparse representation of African populations, for example. The samples chosen for the HapMap [8] and 1000 Genomes [7] projects (Figure 1b,c) reflected the medical, rather than evolutionary, emphases of these projects, and all the sampling choices were limited by political constraints. Nevertheless, these are the key samples for many inferences about human history. Individual studies may use particular additional samples, as we will see in some of the examples given, but will usually be interpreted in the context of these shared resources.

Advances in whole-genome analyses

Population genetics aims to understand the observed distribution of genetic variability and to infer histories of populations from genetic data. For this purpose, what really matters are the differences between sequences (variants). Developments in the past few years now enable millions of variants in thousands of individuals to be analyzed.

Differences can be summarized using statistics (such as allele frequencies or genetic distances) and used to quantify relationships among individuals or populations using clustering techniques, such as principal components analysis (PCA) [9]. With PCA, the information from n polymorphisms is summarized at the individual or population level with n artificial variables (components). Usually the first two or three (principal) capture much of the information and provide a picture of the relationships between the samples (Figure 2a). Model-based clustering techniques (STRUCTURE-like methods; Figure 2b) [10-12] go a little further, estimating the probability of an individual belonging to a certain genetic cluster, the ancestry coefficient. A model with k possible genetic clusters is assumed, and for each individual, k ancestry coefficients (summing to one) are calculated from the genotypes. Patterns of admixture can be visualized at the individual or population level when clusters do not coincide with populations, suggesting past demographic events such as migrations. We see, for example, the admixture in the 1000 Genomes American samples where three components are seen in most individuals (Figure 2b); k1 (light orange) represents a likely African origin, k2 (light blue) a European or West/South Asian contribution, and k5 (yellow) a Native American origin.

One way to estimate relationships that make use of individual whole-genome sequences is through the D statistic [13]. This statistic compares the sharing of derived alleles in two individuals with a third, and thus measures whether the two are equally distant from the third, or whether one is closer. This last situation is taken as evidence for greater genetic exchanges between them - for example, through admixture. More sophisticated inferences about past demographic events can be attained by fitting statistics derived from simulation under specific demographic models (for example, [14-16]) to empirical data. There are fast algorithms that can accommodate the growing mass of empirical data [16-18]. The underlying rationale is that because the parameters (such as population sizes or split times) in the simulation are known, a good fit to the empirical data indicates that the observed pattern of genetic variability might have been produced by that model. This approach, however, remains computationally intensive and it is unclear how fully the simple models used capture key elements of human history.

These analyses extract far more information from genome-wide data than from single loci such as mtDNA or the Y chromosome. Although demographic inferences have been made using single loci - for example, suggesting that 20 to 45 thousand years ago (KYA) most humans lived in South Asia [19] - such conclusions are highly dependent on the sampling strategy [20]. In any case, extant lineages from both mtDNA and the Y chromosome coalesce <200 KYA [21,22], which prevents inferences about earlier demography. By contrast, inference about effective population size back to several million years ago (MYA) has been made using whole-genome sequences [23]. This approach estimated the coalescence time of the maternal and paternal copies of each genomic segment and examined the distribution of effective population size at different time intervals inferred from these. It identified a significant reduction in population size in the past 100 KY, more marked in European and East Asian populations than in Africans, and a shared demography before that. The population size was larger between 100 KYA and 200 KYA, which might reflect population substructure at that time. Interestingly, considerable exchange between sub-Saharan Africans and Europeans/Asians was inferred until 20 to 40 KYA, consistent with some more standard models that estimated an unexpectedly low split time between Asians and Europeans of 23 KYA [16].

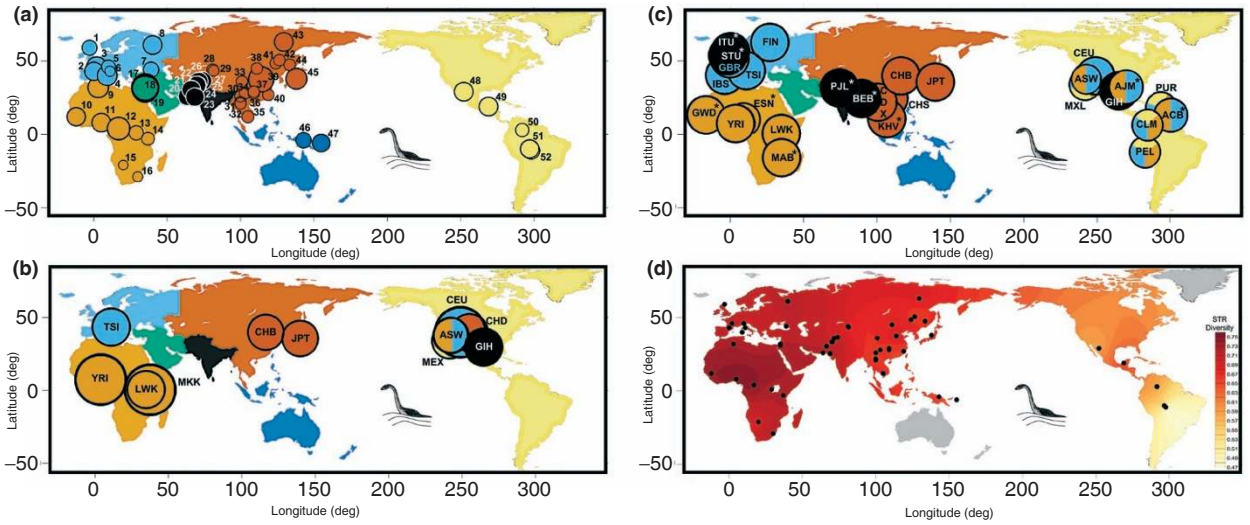

Figure 1. Widely used population samples and worldwide genetic diversity. (a) Human Genome Diversity Project (HGDP) [30], **(b)** HapMap [60], **(c)** 1000 Genomes Project [7], **(d)** Genetic diversity in the HGDP samples [26]. In (a-c), samples are represented by circles with an area proportional to the sample size, placed according to their sampling location, but colored according to the predominant continental ancestry. Numbers and acronyms identify individual samples and details are given in Additional file 1; asterisks in (c) indicate populations that are still being collected. In (d), diversity measurements [26] were interpolated between sampling sites (dots); Australia and Greenland are grayed out because of a lack of data. See Additional file 1 for further details.

Another great advantage of sequence data is that they provide an unbiased estimate of genomic variability. African individuals carry more SNPs than Europeans: 3.3 million each compared with 2.7 million according to one study [7]. However, somewhat counter-intuitively, the number of variants in a population sample of more than a few thousand is actually larger in Europeans [24,25]. This is because the European (or Asian) populations carry vast numbers of extremely rare variants that are seldom shared between individuals, a consequence of their recent explosive demographic growth.

Worldwide diversity patterns: the serial founder model and early expansion patterns

The broad pattern of global human genetic variation is well established: autosomes, the X chromosome, mtDNA and the Y chromosome all generally show higher genetic diversity in African populations than in non-Africans. In addition, non-African populations carry only a fraction of the common (and hence old) African genetic variants. Furthermore, phylogenetic trees from mtDNA, the Y chromosome and autosomal regions most commonly root in Africa, with non-African populations having subsets of the lineages. These conclusions, derived before the days of large-scale sequencing, were unsurprisingly reinforced by whole-genome sequences [7]. Such observations are readily explained by a recent and predominantly African origin for modern humans and expansion of a subgroup into the rest of the world, but detailed genetic analyses have allowed more sophisticated models to be developed.

Analyses of global genome-wide datasets - initially short tandem repeats (STRs) in the Human Genome Diversity Project (HGDP) panel (Figure 1a) - revealed a strong negative correlation between genetic diversity and migration distance from East Africa [26,27] (Figure 1d). The authors proposed a serial founder model to account for this relationship: a subgroup set out from a source population to colonize a neighboring region, expanded to form a secondary source population and the process was repeated in successive steps away from the origin. The correlation was further confirmed by whole-genome SNP data in the same panel [28] and by extensive re-sequencing data (thus largely free from ascertainment bias) in an independent sample set [29]. Although the model was established using autosomal variants, Y-STR diversity in the HGDP panel was consistent with it and showed corresponding decreases in Y-chromosomal coalescence time and effective population size, which further supported the model [30].

An extension of the model would be to identify the region within Africa that provides the best candidate for the origin; the above analyses simply assumed an origin in East Africa. Two such studies have pointed to south-western Africa origin [31,32]: for example, the ≠Khomani Bushmen from South Africa have the lowest linkage disequilibrium among the populations examined (a characteristic of an ancient population) [31]. This conclusion

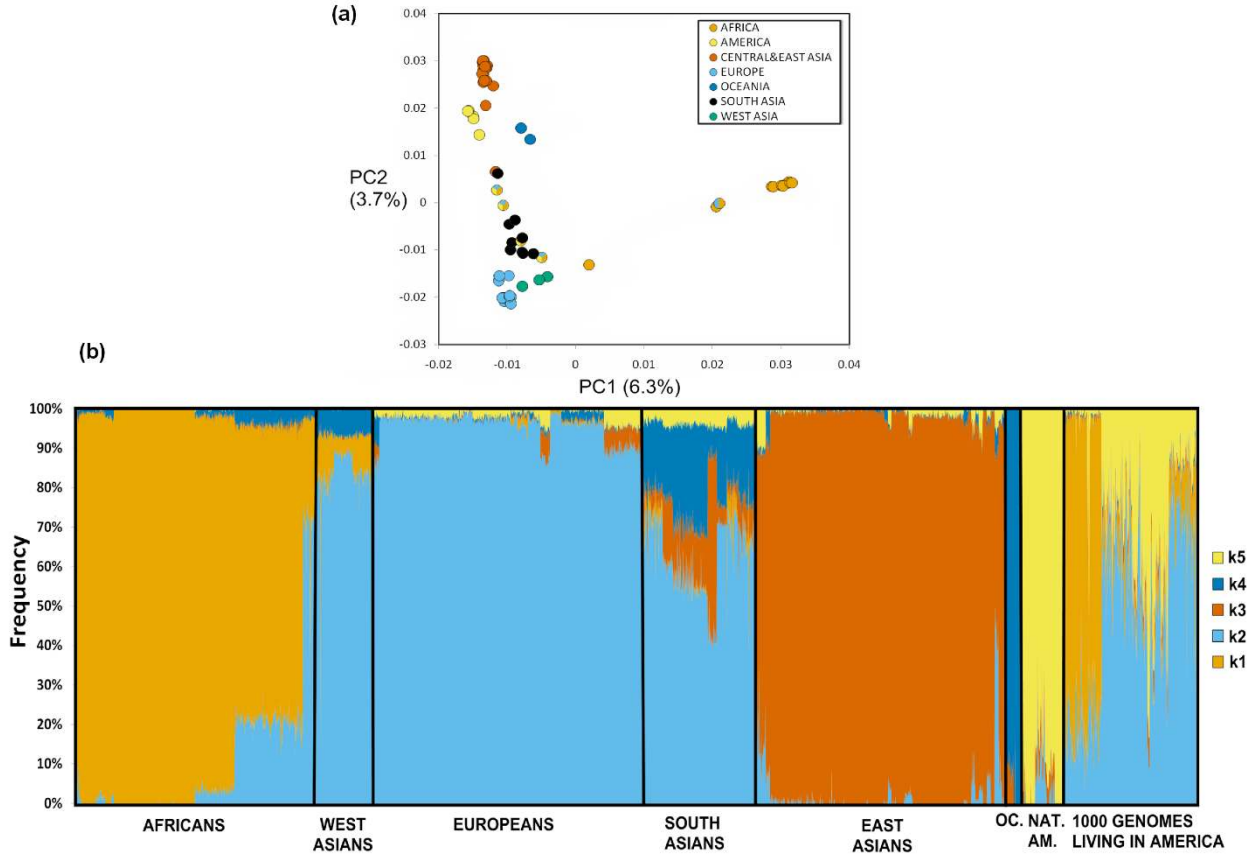

Figure 2. Representation of individual and population relationships in the combined HGDP and 1000 Genomes samples. (a) Principal components plot, where each dot represents a population, colored according to ancestry as in Figure 1. **(b)** STRUCTURE-like plot, where each thin vertical line represents an individual with inferred ancestry from five clusters, k1 to k5. Individuals are grouped by population, and the regions of origin are indicated. Oc, Oceania; Nat. Am., Native American. See Additional file 1 for further details.

must, however, be interpreted with caution because of the limited representation of East African populations in these comparisons, and the admixture and migration that have occurred in Africa since the origin, influencing the genetic properties of the modern populations analyzed.

This genetic model has stimulated new analyses of non-genetic datasets. An examination of 37 morphometric characteristics in 4,666 male skulls drawn from 105 populations worldwide revealed a decrease in phenotypic variability mirroring the loss in genetic diversity and suggested central/southern Africa as the origin [33]. Similarly, the number of phonemes used in a global sample of 504 languages was also well explained by a serial founder model of expansion from an inferred origin in southern/western Africa [34]. These strong patterns of human genetic, morphological and linguistic variation support a single African origin for most of human diversity, but do not preclude low levels of admixture with archaic humans [13,35] - a 'leaky replacement' model where most but not all archaic diversity was replaced by an expansion from Africa 50 to 70 KYA.

A genome sequence derived from a 100-year-old Australian Aboriginal hair sample has provided new insights into the early migration patterns [36]. Using a variant of the *D* statistic (designated D_{4P}), which assesses allele sharing between four individual genomes, the authors [36] investigated the sharing patterns between African, European, East Asian (Han Chinese) and Australian Aboriginal genomes. They observed more sharing between East Asian and European than between East Asian and Aboriginal genomes, and so proposed an initial split after the exit from Africa between the Aboriginal ancestors on the one hand, and the ancestors of modern Europeans and Chinese on the other, a conclusion supported by an independent study based on genotyping a much larger number of individuals [37].

The Jewish Diaspora

Population movements have, of course, continued since the initial expansion out of Africa, and one, initiated by the destruction of the First Temple in Jerusalem by Nebuchadnezzar II of Babylon (c. 586 BCE) and known

as the Jewish Diaspora, has been studied intensively using both historical and genetic data. The Diaspora led to the dispersal of Jewish people from the Levant to many parts of the world, and traditionally three main groups were recognized: Ashkenazim (Eastern European Jewish populations), Sephardim (Southern European Jewish populations) and Mizrahim (Middle Easterners) [38]. Other Jewish groups outside Israel include Ethiopian and Indian Jews. 'Jewishness' is inherited from the mother, so a shared Middle Eastern origin for mtDNA contrasted with a more diverse origin for Y chromosomes might be expected, at least in exogamous communities, and has indeed been reported [39,40].

Two independent studies have examined the relationships between Jewish and surrounding non-Jewish groups ('hosts') using genome-wide SNP data [38,41]. Both came to the conclusion that although the whole-genome data show detectable introgression from the local populations into the genetic pool of the Diaspora people, both autosomal and uniparental markers point towards a clear Middle Eastern origin for most of the individuals studied. In particular, the mtDNA and genomic markers show strong evidence for a Middle Eastern origin, matching the expectations from the matrilineal pattern of Jewishness (Figure 3). Furthermore, the genomic information provided for the first time a signature of the demographic expansion documented as the 'demographic miracle' of the Ashkenazi Jews during the past 500 years [38], whereas little evidence for inter-Diaspora gene flow was observed. The authors [38] write: 'Admixture with surrounding populations had an early role in shaping world Jewry, but, during the past 2,000 years, may have been limited by religious law as Judaism evolved from a proselytizing to an inward-looking religion.'

However, a major exception to this scenario was provided by the Ethiopian and Indian Jews. These two groups show a much greater extent of host genetic introgression, effectively clustering together with the local non-Jewish populations in both PCA and STRUCTURE-like analyses (Figure 3c). The genomic clustering explains observations from the mtDNA data [39,40] in which a high proportion of local maternal lineages were shared with the Ethiopian and Indian populations (Figure 3a), potentially as a result of more relaxed criteria of self-identification in these Jewish communities. This pictures the migration of these people as a star-shaped pattern centered on the Levant, with founder effects and genetic drift [42] acting differently on the individual branches.

To further emphasize the importance of social rules in shaping the genetic landscape of a population, we can make a parallel with the more recent (about 1,000 YA) Roma (Gypsy) expansion from a North Indian source population towards South-Western Europe. Although the available data still rely on a limited set of markers [43-45],

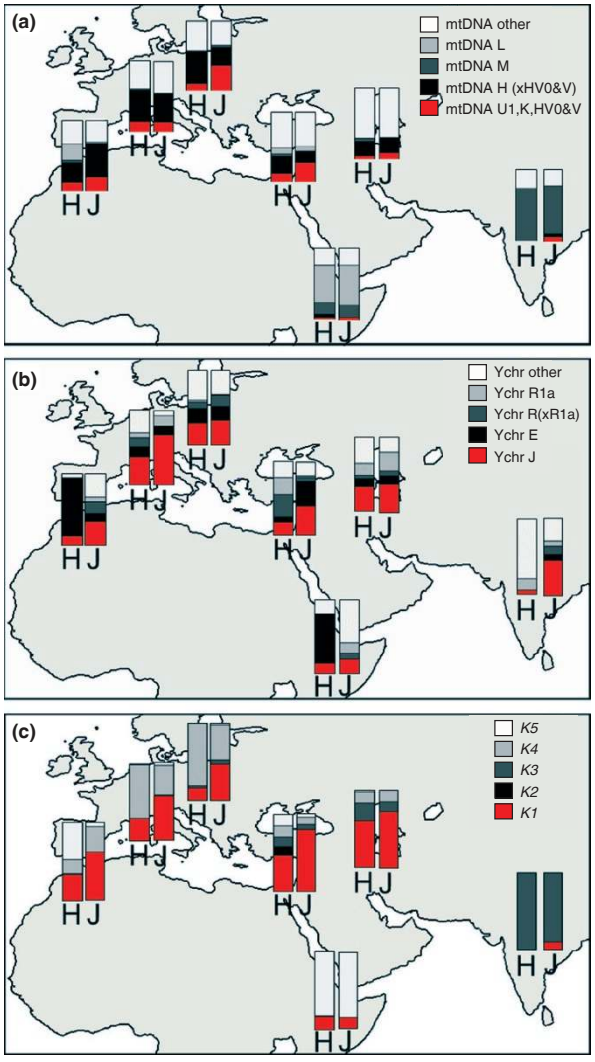

Figure 3. Genetic relationships between Jewish Diaspora (J) and nearby non-Jewish (host, H) populations. (a) mtDNA analysis; **(b)** Y-chromosomal analysis; **(c)** autosomal SNP genotype analysis. Lineages (a,b) or ancestry components (c) were divided into five classes, and in each panel one class (red) is present at higher frequency in most Jewish populations (J) than in the corresponding host population (H), illustrating a genetic heritage component shared by most Jewish populations. See Additional file 1 for further details.

the more flexible inheritance of Roma status is reflected in a serial dilution of the source (Indian) gene pool, characterized by stepwise admixture with local 'host' populations for both uniparental and autosomal markers. It is thus striking that the different migration pattern and relaxation in self-identification criteria allowed a greater extent of mixture of Roma with hosts in half of the time of the Jewish Diaspora.

The first Greenlanders

Greenland was among the last places on the earth to be reached by humans and is currently inhabited by the

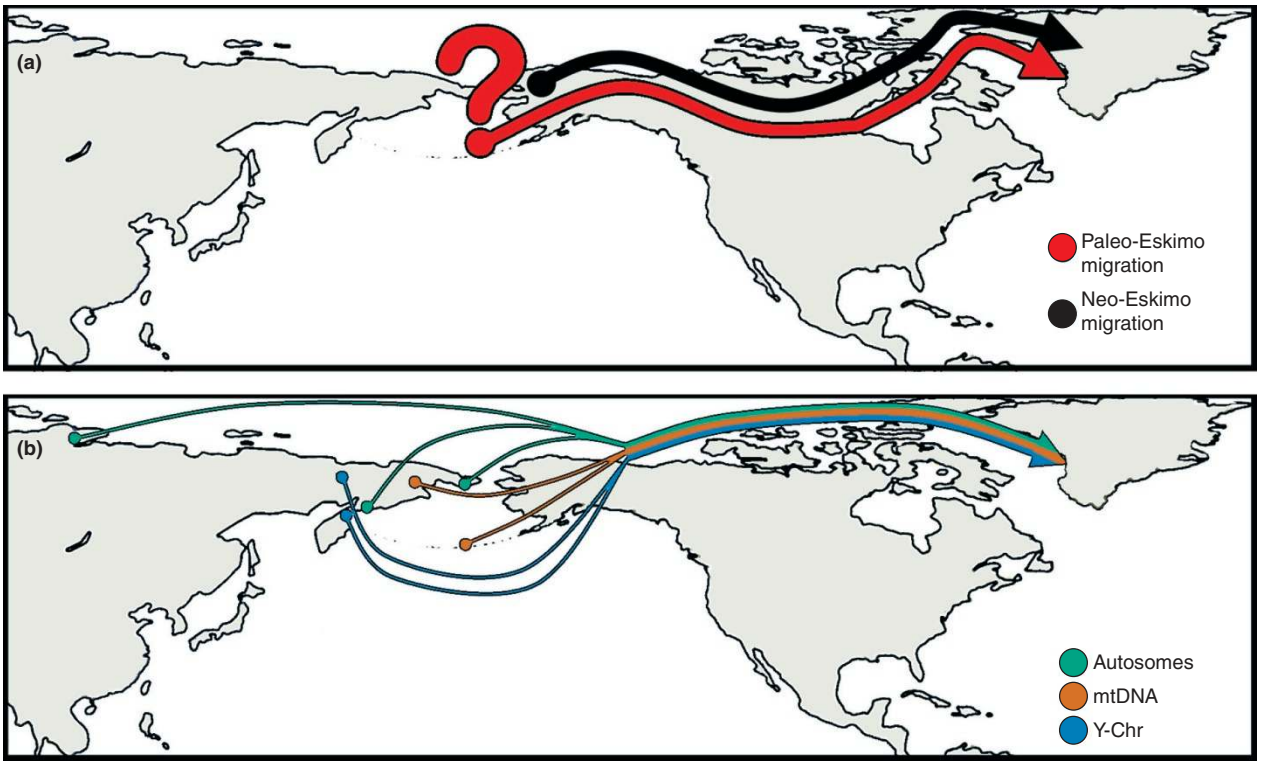

Figure 4. Archaeological and genetic evidence for the peopling of Greenland. (a) Archaeological evidence suggests two waves of migration to Greenland: Paleo-Eskimo (Saqqaq) about 4,500 YA (red) and Neo-Eskimo (Inuit) about 1,000 YA (black). The place of origin of the Saqqaq migration was unknown. **(b)** Genetic evidence suggests a Saqqaq origin in Siberia. The modern populations most similar to the Saqqaq mtDNA, Y chromosome and autosomes are all located in Siberia or the Aleut Islands. These maps represent initial and final locations, but not the route taken.

Inuit, who have their closest relatives in Canadian Inuit. Archaeological evidence suggests two waves of migration into Greenland from Siberia and Alaska (Old and New World Beringia, respectively; Figure 4a). The first migration started about 4.5 KYA and involved the Paleo-Eskimo populations, which included at least three different groups identified by distinct archaeological cultures (Saqqaq, Pre-Dorset and Dorset). The Inuit arrived in Greenland as a second wave (Neo-Eskimo) that took place about 1,000 YA [46]. Although both genetics and archaeology support a recent (1,000 YA) Beringian origin for the Neo-Eskimo populations [47], the origin of the Paleo-Eskimo groups was unknown, apart from the geographical proximity of Native American and Beringian populations to Greenland. The availability of an exceptionally well-preserved hair sample from an ancient Paleo-Eskimo individual from the Saqqaq culture, originating from the west coast of Greenland, allowed this issue to be investigated using genetics.

The hair dated to around 4,000 YA and provided both a high-coverage mtDNA sequence [48] and 20x coverage genomic sequence [49]. Analyses of these data suggested that this individual was not a likely ancestor of the current Greenland Inuit, confirming the theory of different waves of migration into Greenland, the last of which eventually replaced the previous populations. Indeed, the Paleo-Eskimo Saqqaq mtDNA was attributed to the D2a1 haplogroup, which has not been found in the Inuit [48]. The closest lineage to D2a1 is currently found in the Aleut and Sireniky-Yuit populations from the Aleut Islands and extreme East Siberia, respectively, suggesting an Old World Beringian origin for the maternal ancestry of the Saqqaq (Figure 4b). The Y lineage was Q1a*, now known from the Koryaks and Yukaghir of East Siberia [50]. This connection with Siberian populations was confirmed by the autosomal data. Indeed, clustering analysis of polymorphisms from the Saqqaq genome (in comparison with 35 populations from Europe, Asia and America) showed shared ancestry with three Siberian populations (Figure 4b) and excluded detectable admixture or shared origin with Western European and Native American populations. In addition to this ancestry information, analyses of functional variants in genomic DNA provided some insight into phenotypic traits of the Saqqaq individual [49], suggesting blood type A+, brown eyes, dry earwax and dark hair, with, ironically in view of the source material, a predisposition to baldness.

Lessons from associated species

The study of human DNA variation is not the only way to use genetics to decipher our history. Insights can also come from the study of organisms that have established long-term relationships with us (such as parasites and pathogens), providing an independent perspective on their host's evolutionary history. Broadly speaking, samples are collected in different geographical areas and their phylogeographic relationship is investigated to locate their probable origin (usually the place where the parasite's genetic diversity is greatest) and characterize their spread by describing and dating the branching pattern and/or analyzing population-genetic parameters.

As expected, in many cases a strong correlation has been found between parasite origins and spread and both ancient human migrations, such as the out-of-Africa movement [51], and more recent migrations, such as the Austronesian expansion [52] or along the Silk Road trade route [53]. In addition, louse phylogenetic history has also helped to reconstruct aspects of our appearance that do not fossilize, a challenging task using genotype information, and even more so for ancestors without genotype information. Chimpanzees have only head lice and gorillas only pubic, whereas humans have both. Head lice seem to have co-evolved with their respective hosts since the chimpanzee-human split 6 to 7 MYA. By contrast, the divergence time between gorilla and human pubic lice is substantially lower than the gorilla-human split (about 8 MYA), suggesting that pubic lice originated in gorillas and switched to the human lineage around 3.3 MYA [54]. This may indicate that, by that time, the human body was partially free from hair and a new niche was available. Despite the ancient loss of body hair, according to this interpretation, it took a long time for humans to introduce clothes, according to another study of lice. *Pediculus* head and clothing lice split around 70 to 80 KYA [55,56]. If the split occurred as soon as the new clothing niche became available, this date provides us with an estimate of the time of the introduction of clothes.

Conclusions and future directions

What can we learn from these examples about the information contributed by different genetic loci, and the comparison between genetic and non-genetic sources? It is useful to consider this question by time period. In the last about 50 KY, mtDNA, the Y chromosome and whole-genome data are all informative, and the agreements between them are more striking than the disagreements. For example, in both the Diaspora and Saqqaq genome studies, the genome-wide data extended, rather than overturned, the conclusions from the uniparental loci. This does not have to be the case, and is not always found: a family from England, for example, unexpectedly carried a Y chromosome typical of sub-Saharan Africa [57], but the rarity of such examples - the strong correlation between the histories of different genomic regions - provides evidence for marked geographical structure prevailing over this period. Nevertheless, some unresolved issues remain: genome-wide data suggest extensive gene flow between sub-Saharan Africans and non-Africans [23], whereas uniparental data do not [1,2].

Further back into the past, the uniparental loci provide less and less information because fewer lineages have survived to the present: all mtDNA and Y lineages outside Africa trace back to a single ancestor about 50 to 70 KYA, for example. So for understanding the early expansions, the main insights have come from genome-wide data (for example, [36]), and these are our only source of genetic data for the period >200 KYA.

In general, non-genetic data provide far more detailed information for the last few millennia, and the genetic inferences are largely consistent with them (such as the Diaspora). Genetics can add information, such as about the origins of the Paleo-Eskimos, that has not been available from other sources. Again, there are unresolved issues - for example, the genetic contact inferred between populations 20 to 40 KYA contrasts with the distinct regional archaeological records for this period [58]. And genetics can generate hypotheses, such as the adoption of clothing 70 to 80 KYA, that can now be tested by archaeologists.

As whole-genome sequencing becomes cheaper and more routine, it can more readily be used to address questions of evolutionary interest: the complex demography in Africa and the peopling of the Americas and the Pacific, for example. When did the ancestral populations split, how much subsequent gene flow was there, what patterns of migration can be identified? There may be admixed populations in which mtDNA and the Y chromosome from one ancestral source have been entirely lost by drift, but ancestral information can be reconstructed from the autosomal regions: the admixed American populations in the 1000 Genomes Project are allowing this possibility to be tested. Perhaps we can study the genomics of pre-contact Native American or Tasmanian populations, as proposed by the Taíno Genome Project [59] investigating the legacy of a pre-Colombian Puerto Rican population using 1000 Genomes data. Our history is encoded in our genomes, each much smaller than a grain of sand, and we are only now collecting the data and developing the tools to decipher this rich source of information.

Additional files

Additional file 1: Population samples and geographic coordinates used in Figures 1 and 2, together with additional information on the construction of Figures 2 and 3. [61]

Acknowledgements
Our work is supported by The Wellcome Trust (grant number 098051), and additionally an EMBO Short-term Fellowship (ASTF 324-2010) to VC, and the Cambridge European Trust, a Domestic Research Scholarship and Emmanuel College, Cambridge, UK (LP). We thank Jean McEwen for information about the 1000 Genomes samples.

Author details
[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [2]Institute of Genetics and Biophysics 'A. Buzzati-Traverso', National Research Council (CNR), Naples, Italy. [3]The Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Fitzwilliam Street, Cambridge CB2 1QH, UK.

Published: 21 November 2011

References

1. Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ: **Harvesting the fruit of the human mtDNA tree.** *Trends Genet* 2006, **22:**339-345.
2. Jobling MA, Tyler-Smith C: **The human Y chromosome: an evolutionary marker comes of age.** *Nat Rev Genet* 2003, **4:**598-612.
3. Underhill PA, Kivisild T: **Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations.** *Annu Rev Genet* 2007, **41:**539-564.
4. Ragoussis J: **Genotyping technologies for genetic research.** *Annu Rev Genomics Hum Genet* 2009, **10:**117-133.
5. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data.** *Nat Rev Genet* 2011, **12:**443-451.
6. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11:**31-46.
7. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.
8. The International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426:**789-796.
9. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2:**e190.
10. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155:**945-959.
11. Tang H, Coram M, Wang P, Zhu X, Risch N: **Reconstructing genetic ancestry blocks in admixed individuals.** *Am J Hum Genet* 2006, **79:**1-12.
12. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
13. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, *et al.*: **A draft sequence of the Neandertal genome.** *Science* 2010, **328:**710-722.
14. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: **Calibrating a coalescent simulation of human genome sequence variation.** *Genome Res* 2005, **15:**1576-1583.
15. Beaumont MA, Zhang W, Balding DJ: **Approximate Bayesian computation in population genetics.** *Genetics* 2002, **162:**2025-2035.
16. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD: **Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data.** *PLoS Genet* 2009, **5:**e1000695.
17. Marjoram P, Wall JD: **Fast "coalescent" simulation.** *BMC Genet* 2006, **7:**16.
18. McVean GA, Cardin NJ: **Approximating the coalescent with recombination.** *Philos Trans R Soc Lond B Biol Sci* 2005, **360:**1387-1393.
19. Atkinson QD, Gray RD, Drummond AJ: **mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory.** *Mol Biol Evol* 2008, **25:**468-474.
20. Gunnarsdottir ED, Li M, Bauchet M, Finstermeier K, Stoneking M: **High-throughput sequencing of complete human mtDNA genomes from the Philippines.** *Genome Res* 2011, **21:**1-11.
21. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S: **The dawn of human matrilineal diversity.** *Am J Hum Genet* 2008, **82:**1130-1140.
22. Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D, Scozzari R: **A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa.** *Am J Hum Genet* 2011, **88:**814-818.
23. Li H, Durbin R: **Inference of human population history from individual whole-genome sequences.** *Nature* 2011, **475:**493-496.
24. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR, Wheeler DA, Sabo A, Lusk C, Weiss KG, Akbar H, Cree A, Hawes AC, Newsham I, Varghese RT, Villasana D, Gross S, Joshi V, Santibanez J, Morgan M, Chang K, Iv WH, Templeton AR, Boerwinkle E, Gibbs R, Sing CF: **Deep resequencing reveals excess rare recent variants consistent with explosive population growth.** *Nat Commun* 2010, **1:**131.
25. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD: **Demographic history and rare allele sharing among human populations.** *Proc Natl Acad Sci U S A* 2011, **108:**11983-11988.
26. Prugnolle F, Manica A, Balloux F: **Geography predicts neutral genetic diversity of human populations.** *Curr Biol* 2005, **15:**R159-R160.
27. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL: **Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa.** *Proc Natl Acad Sci U S A* 2005, **102:**15942-15947.
28. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM: **Worldwide human relationships inferred from genome-wide patterns of variation.** *Science* 2008, **319:**1100-1104.
29. Luca F, Hudson RR, Witonsky DB, Di Rienzo A: **A reduced representation approach to population genetic analyses and applications to human evolution.** *Genome Res* 2011, **21:**1087-1098.
30. Shi W, Ayub Q, Vermeulen M, Shao RG, Zuniga S, van der Gaag K, de Knijff P, Kayser M, Xue Y, Tyler-Smith C: **A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations.** *Mol Biol Evol* 2010, **27:**385-393.
31. Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodríguez-Botigué L, Ramachandran S, Hon L, Brisbin A, Lin AA, Underhill PA, Comas D, Kidd KK, Norman PJ, Parham P, Bustamante CD, Mountain JL, Feldman MW: **Hunter-gatherer genomic diversity suggests a southern African origin for modern humans.** *Proc Natl Acad Sci U S A* 2011, **108:**5154-5162.
32. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM: **The genetic structure and history of Africans and African Americans.** *Science* 2009, **324:**1035-1044.
33. Manica A, Amos W, Balloux F, Hanihara T: **The effect of ancient population bottlenecks on human phenotypic variation.** *Nature* 2007, **448:**346-348.
34. Atkinson QD: **Phonemic diversity supports a serial founder effect model of language expansion from Africa.** *Science* 2011, **332:**346-349.
35. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Pääbo S: **Genetic history of an archaic hominin group from Denisova Cave in Siberia.** *Nature* 2010, **468:**1053-1060.
36. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, Kivisild T, Zhai W, Eriksson A, Manica A, Orlando L, De La Vega FM, Tridico S, Metspalu E, Nielsen K, Ávila-Arcos MC, Moreno-Mayar JV, Muller C, Dortch J, Gilbert MT, Lund O, Wesolowska A, Karmin M, Weinert LA, Wang B, Li J, *et al.*: **An Aboriginal Australian genome reveals separate human dispersals into Asia.** *Science* 2011, **334:**94-98.
37. Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME, Saitou N, Wollstein A, Kayser M, Pääbo S, Stoneking M: **Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania.** *Am J Hum Genet* 2011, **89:**516-528.
38. Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, Palamara PF, Morrow B, Friedman E, Oddoux C, Burns E, Ostrer H: **Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry.** *Am J Hum Genet* 2010, **86:**850-859.
39. Behar DM, Metspalu E, Kivisild T, Rosset S, Tzur S, Hadid Y, Yudkovsky G, Rosengarten D, Pereira L, Amorim A, Kutuev I, Gurwitz D, Bonne-Tamir B, Villems R, Skorecki K: **Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora.** *PLoS One* 2008, **3:**e2062.
40. Non AL, Al-Meeri A, Raaum RL, Sanchez LF, Mulligan CJ: **Mitochondrial DNA**

reveals distinct evolutionary histories for Jewish populations in Yemen and Ethiopia. *Am J Phys Anthropol* 2011, **144:**1-10.
41. Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, Rootsi S, Chaubey G, Kutuev I, Yudkovsky G, Khusnutdinova EK, Balanovsky O, Semino O, Pereira L, Comas D, Gurwitz D, Bonne-Tamir B, Parfitt T, Hammer MF, Skorecki K, Villems R: **The genome-wide structure of the Jewish people.** *Nature* 2010, **466:**238-242.
42. Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST: **Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population.** *Proc Natl Acad Sci U S A* 2010, **107:**16222-16227.
43. Gusmao A, Gusmao L, Gomes V, Alves C, Calafell F, Amorim A, Prata MJ: **A perspective on the history of the Iberian gypsies provided by phylogeographic analysis of Y-chromosome lineages.** *Ann Hum Genet* 2008, **72:**215-227.
44. Gusmao A, Valente C, Gomes V, Alves C, Amorim A, Prata MJ, Gusmao L: **A genetic historical sketch of European Gypsies: The perspective from autosomal markers.** *Am J Phys Anthropol* 2010, **141:**507-514.
45. Mendizabal I, Valente C, Gusmão A, Alves C, Gomes V, Goios A, Parson W, Calafell F, Alvarez L, Amorim A, Gusmão L, Comas D, Prata MJ: **Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective.** *PLoS One* 2011, **6:**e15988.
46. Damas D: *Handbook of North American Indians. Volume 5.* Washington DC: Smithsonian Institution; 1984.
47. Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S: **mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion.** *Am J Hum Genet* 2000, **67:**718-726.
48. Gilbert MT, Kivisild T, Grønnow B, Andersen PK, Metspalu E, Reidla M, Tamm E, Axelsson E, Götherström A, Campos PF, Rasmussen M, Metspalu M, Higham TF, Schwenninger JL, Nathan R, De Hoog CJ, Koch A, Møller LN, Andreasen C, Meldgaard M, Villems R, Bendixen C, Willerslev E: **Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland.** *Science* 2008, **320:**1787-1789.
49. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MT, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, *et al.*: **Ancient human genome sequence of an extinct Palaeo-Eskimo.** *Nature* 2010, **463:**757-762.
50. Malyarchuk B, Derenko M, Denisova G, Maksimov A, Wozniak M, Grzybowski T, Dambueva I, Zakharov I: **Ancient links between Siberians and Native Americans revealed by subtyping the Y chromosome haplogroup Q1a.** *J Hum Genet* 2011, **56:**583-588.
51. Tanabe K, Mita T, Jombart T, Eriksson A, Horibe S, Palacpac N, Ranford-Cartwright L, Sawai H, Sakihama N, Ohmae H, Nakamura M, Ferreira MU, Escalante AA, Prugnolle F, Björkman A, Färnert A, Kaneko A, Horii T, Manica A, Kishino H, Balloux F: ***Plasmodium falciparum*** **accompanied the human expansion out of Africa.** *Curr Biol* 2010, **20:**1283-1289.
52. Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu JY, Maady A, Bernhöft S, Thiberge JM, Phuanukoonnon S, Jobb G, Siba P, Graham DY, Marshall BJ, Achtman M: **The peopling of the Pacific from a bacterial perspective.** *Science* 2009, **323:**527-530.
53. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, Cui Y, Thomson NR, Jombart T, Leblois R, Lichtner P, Rahalison L, Petersen JM, Balloux F, Keim P, Wirth T, Ravel J, Yang R, Carniel E, Achtman M: ***Yersinia pestis*** **genome sequencing identifies patterns of global phylogenetic diversity.** *Nat Genet* 2010, **42:**1140-1143.
54. Reed DL, Light JE, Allen JM, Kirchman JJ: **Pair of lice lost or parasites regained: the evolutionary history of anthropoid primate lice.** *BMC Biol* 2007, **5:**7.
55. Kittler R, Kayser M, Stoneking M: **Molecular evolution of *Pediculus humanus* and the origin of clothing.** *Curr Biol* 2003, **13:**1414-1417.
56. Toups MA, Kitchen A, Light JE, Reed DL: **Origin of clothing lice indicates early clothing use by anatomically modern humans in Africa.** *Mol Biol Evol* 2011, **28:**29-32.
57. King TE, Parkin EJ, Swinfield G, Cruciani F, Scozzari R, Rosa A, Lim SK, Xue Y, Tyler-Smith C, Jobling MA: **Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy.** *Eur J Hum Genet* 2007, **15:**288-293.
58. Jobling MA, Hurles M, Tyler-Smith C: *Human Evolutionary Genetics.* New York and Abingdon: Garland Science; 2004.
59. **The Taino Genome Project** [https://sites.google.com/a/upr.edu/dna-lab/1000genomes/the-taino-genome-project]
60. The International HapMap 3 Consortium: **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, **467:**52-58.
61. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nat Genet* 2006, **38:**904-909.

doi:10.1186/gb-2011-12-11-234
Cite this article as: Colonna V, *et al.*: A world in a grain of sand: human history from genetic data. *Genome Biology* 2011, **12:**234.