Santa Fe Institute 6-11 Feb 2006 Meeting – notes

Richard Villems
Tartu

Dear all,
The following is not what I plan to present in my talk – I wrote the text in order to save precious meeting time - a sort of general introduction, a background. Very hectic as such, but makes my life easier in a meaning that more time will be left to concentrate on problems and ideas that are directly relevant to the problems I promised to talk about – initial steps of the colonization of Eurasia and Australia by anatomically modern humans.

Opening remarks

As geneticists, we cannot contribute much into the question about the "beginning" of *Homo sapiens.* And it is not clear whether we ever can, because this "beginning", from genetics point of view, is something we cannot define so far. And in any case, classification is an agreement between scholars, not an empirical finding. For example, one may classify Neanderthals among *Homo sapiens*, like some schools of palaeoanthropologists suggest. In this case we have two branches of *Homo sapiens* in the *Hominine* clade – *H. sapiens sapiens* (us) and, secondly, *H. s. neanderthalensis* – the latter being a sister group that became extinct relatively recently. Most fossil specialists assume their split from a common ancestor about a half a million of years ago, in Africa. One of such schemes, drawn very recently by Rob Foley, puts this split in between 0.3 – 0.7 MYA *(1-2 slides)*. This very wide window in such a recent time depth tells us that much more fieldwork is still needed.

On the other hand, most palaeoanthropologists seem to agree that modernization of "an early" *Homo* in the line of our (i.e. *H.s.s*.) direct predecessors can be seen starting about 300,000 YBP. Classification is of course something on what most of researchers must agree, otherwise it would be just someone's opinion. It is interesting to notice here that, during the last decade or so, new fossil findings have forced to shift "the beginning" of *Homo sapiens sapiens* from ~120,000 YBP first to about 150,000 YBP and, recently, to about 190,000 YBP. This is very remarkable – not the particular time depth, but the fact that one of the basic milestones turned out to be so flexible. Again – genetics has so far little to say here – it is a domain of palaeoanthropologists. However, it is not totally excluded that, in one good day, genetics will find something in our genes, comparing e.g. us and chimpanzees, suggesting that certain changes (mutations) either altered expression patterns of some genes or have changed some characteristics of proteins coded by some genes, or else – that had profound impact on some mental processes (linguistic abilities ?) - and suggest "geneticists" idea about the emergence of *H.s.s.* There are more than one group working in this direction already for several years.

Evolution of population genetics

The fundamentals of human population genetics trace back to pioneering observations from the beginning of the previous century that not just people but populations from different geographic areas tend to differ in blood groups. This, obviously very important from practical medicine discovery was first systematically studied during

the WW I. Later, and in particular starting from the post-WW II time, thanks to progress in analytical biochemistry (electrophoretic separation of proteins), impressive amount of work has been carried out, using the so-called "classical markers" – a new term coined to mark not the beginning but the end of the "classical era" and beginning of the DNA era. Though the first publications of this new era date back for almost a generation, it became truly wide-spread starting from the second half of the 90s.

When "classical" and "the DNA era" is compared, people often stress that while the best studies of the first period operated with 100 or so markers, the DNA era allows to increase this number very significantly (orders of magnitude). That is true, but perhaps not the most important difference. One of the intriguing differences was that the DNA era made it possible (by far more easy) to assess variation in two haploid genomes – in the Y chromosome and mtDNA. Haploid genomes in mammals are perfect for phylogenetic analysis because of the lack of recombination and because they allow to follow separately paternally and maternally inherited phylogenies. The second, and some authors suggest that the most important aspect is that it became now possible to rely on molecular clock and to get, imperfect as they might be, time estimates for the molecular events (mutations) that one infers from the topology of phylogenetic trees.

It is often stressed that molecular evolution shaping variation in mtDNA and Y-chomosomal sequences is neutral – is not caused by natural selection. However, one must add here that new information about global variation of mtDNA, gathered during a few recent years (and still not really digested in depth), forces researchers to take selection more seriously than it has been done so far. Furthermore, care should be taken speaking about natural selection. Natural selection in a sense of purifying selection is for sure always active, in particular in mtDNA, tightly packed with genes. Therefore, stressing "neutrality" in mtDNA phylogenetics has never been, professionally speaking, understood as a process where mutations occur and accumulate randomly all over mtDNA genome. The aspect that is hotly debated at present, addresses entirely different question – should we expect that some of mutations, shaping human mtDNA phylogenetic tree at present or has done it recently, is or has been under positive selection and, perhaps more importantly, don't we need to take into account that the process of the accumulation of variation in the mtDNA pool involves, in recent limbs and twigs of the tree, accumulation of (lots of) mildly deleterious mutations, that are afterwards under slight negative selection and eliminated from a population? I am touching these problems here not for further in-depth discussions, but in order to indicate that we have unsolved problems that, at least in theory, may significantly influence how to use molecular clock properly: how to calibrate it etc.

Anyhow, a big leap ahead – employing molecular clock concept - has all the sudden changed the way how researchers could interpret their results and to reconstruct the past. A new term "archaeogenetics" (I am not here to propagate this term, though) has been coined and is increasingly wider used for exploring great many essentially interdisciplinary problems. Of course, one should not forget that the two haploid genetic systems are, strictly speaking, both single loci and, therefore, care must be taken in the interpretation of the obtained results in terms of populations: their migrations, intra- and inter-population diversities and distances etc. However, except

some, largely non-constructive criticism (often dilettante), the value of employing the two "marker systems" turned out to be truly significant for the progress in the whole area. And it is worth to stress here that careful mtDNA and Y-chromosome geneticists try to avoid the term "population" in a "serious talk" – we speak about phylogeny and phylogeography of our "maternal and paternal lineages". Yet notice that if you find a certain variant of mtDNA or Y chromosome in a given area, it inevitably means that a real person – a woman or a man, must have carried it there either recently or generations back.

Subjectively speaking – it seems to me that the main source for misunderstandings (among researchers in the field of population genetics) is that there is still a considerable fraction of people who disparately try to ignore phylogenetics, treating individual mutations as "classical markers", forgetting that mutations occur in increasingly more precisely understood "phylogenetic environment" – in trunks, limbs and twigs of the reconstructed from the extant genetic variation tree

It needs to be added that the last few years witness an enormous burst of new research in something what may be taken (not entirely correctly) as reincarnation of the "classical marker" era – HapMap has produced several millions of single nucleotide polymorphisms in our "usual" autosomal chromosomes – to be explored in order to study, among other goals, shaping of the existing at the present global genetic variation. For sure at least a few per cent out of these millions turn out to be informative in the context of population genetics and interesting times are waiting ahead. But it is not the topic of these notes. See Rosenberg et al. (2005) in PloS Genetics Vol. 6, e70 and refs therein for further reading.

Coalescence ages of mtDNA and Y-chromosomal trees and the emergence of *Homo sapiens sapiens* – a place for quite common mis-interpretation

Coming back to the emergence of *Homo sapiens sapiens*, its dating, it has been now and then suggested that the time window for this "event" – 150,000 – 200,000 years - *finds excellent support* from most of the mtDNA studies as well as from some Y-chromosomal work. Just a month ago I found such an enthusiastic passage in a manuscript of an otherwise very useful review paper by an outstanding archaeologists. It is, however, unclear whether such a "support" has much (or any) logics behind it. MtDNA estimates for time depth since the "African mitochondrial Eve" , coming from the combining of the molecular clock concept with the coalescence tree concept offers, strictly speaking, an estimate for the beginning of the expansion of *Homo sapiens sapines* mtDNA lineages – such lineages that have survived till today and can be sampled in extant populations both in Africa and/or elsewhere.

It is perhaps useful to remind here that coalescence ages for the majority of our autosomal genes, calculated, broadly speaking, in a similar way, vary between 500,000 and one million years. And there is no contradiction involved between these numbers and those calculated from mtDNA variation. On the contrary, the two estimates (for haploid and autosomal genes) are in a reasonably good agreement with a prediction – trees for haploid genes with their four times lower effective population size ought to coalesce roughly four times faster than trees for autosomal genes. Therefore, discussed above coincidence between "the age" of the human mtDNA (and perhaps the Y) tree and the suggested "age" of *H. s. s*. as a species, is not a pair of logically related parameters. Yet this coincidence is in a way convenient, allowing to

look to the branching pattern of these two haploid trees and ascribed to them coalescence ages in parallel with various other events – climatic changes, advancements in technologies, archaeological findings etc. etc. With data sets large enough, one can propose "educated guesses" about human migrations. Many of such inferences address events that are deep in pre-historic time, often lacking satisfactory (or any) primary archaeological evidence. Imperfect as our genetics-based interpretations certainly are, they are often the only source for hypotheses, addressing important and intriguing questions about the (pre)history of our species.

Ancient DNA and archaeogenetics – clarification as far as popular terminology is concerned

"Ancient DNA " is a term used for DNA, extracted from archaeological remains – bones, hair, soft tissue if available and even from soil in caves. Contrary to early expectations (DNA from dinosaur eggs, DNA from 17 million years old magnolia leafs, DNA from insects (or bacteria in the guts of insects), fixed inside 30 million years old Dominican amber, DNA from microorganisms, extracted from perfect salt crystals of even deeper times, etc.etc., today specialists in this field seem to put the lower time limit for this technology to about 100,000 years before the present – a far cry from dinosaurs and, unfortunately, from our 4-5 million years old bipedal australopithecine ancestors as well. It is in particular pity that there seems to be no hope whatsoever to compare directly *Hominine* genes before and after a boundary at about 2 megayears ago, when, all the sudden, brains of our direct ancestors started to expand much faster than our cousins. Yet this 100,000 years boundary is very promising time depth if we keep in mind peopling of Eurasia and Australia.

Irrespective of possible verbal associations, archaeogenetics operates predominantly by interpreting past from genetic variation of extant human populations. Progress in ancient DNA (aDNA) has been very slow and to obtain reliable results turned out to be notoriously complicated. Some success stories are well known (as well as some pitfalls).

I am not going to review here the existing literature about ancient (mt)DNA, but it seems worth to comment some of them.

First of all – the Neanderthal DNA. More precisely, ancient mtDNA, extracted from Neanderthal fossils, with first publication from 1997. By now, with already several different Neanderthal amtDNAs (still small fragments) studied, one can be quite sure that our first cousins form a separate from us clade and that the detected distance between the two clades suggests their departure time, fitting well into a time window that is accepted by most of palaeoanthropologists – i.e. around 500,000 YBP. The finding that so far studied fragments of amtDNA from different Neanderthals form a clade (cluster of closely related lineages), different from the corresponding cluster for anatomically modern humans, is much more important than a question, whether all nucleotide positions in all "restored" sequences are correct. Furthermore, it is worth to stress here that in apioneering and influential 1996 year paper by M. Richards and colleagues (AJHG 1996, 59:185-203) about the western Eurasian variation of mtDNA, the authors, *inter alia*, predicted that any remaining Neanderthal mtDNA in the pool of the present-day human mtDNA (i.e. assuming admixture between the two speciae) should be readily visible because of by far different sequence. Furthermore, Richards et al even predicted this difference quantitatively (assuming certain speed of molecular clock and split 500,00 years ago) and, as a matter of fact, got it right – the

real Neanderthal mtDNA did indeed differ from an "average" modern human in close to 30 positions within the hypervariable segment I.

Not so many years ago, there was a provocative ancient DNA study of Australian human fossils, older than even most of the Neanderthals (bones used for DNA extraction). This paper was published in PNAS, with typical for that time problems – no parallel confirmation of results, no molecular cloning of amplified products. Furthermore, at that time very little was known about mtDNA variation among extant Australian Aborigines. In short, authors suggested that sequences of mtDNA fragments they obtained are absolutely unique and prove that (at least a fraction of) Aboriginal genetics cannot be explained by recent out-of-Africa scenario. Alas – by now this study is not mentioned any more, because everything was wrong in this paper – sequences, their analysis and, inevitably, conclusions. This is a good example, because original conclusions have been "dictated ideologically" – there were (and still are) scholars, who do not accept recent out-of-Africa spread of anatomically modern humans. Irrespective of the fact that by now, thanks to hard work of many laboratories, reconstructed phylogenetic trees of mtDNA and the Y chromosome are unequivocally identifying Africa as the source of the present-day maternal and paternal lineages worldwide. It does not necessarily mean that interbreeding between people who moved out of Africa some 50,000 – 70,000 years ago and possible earlier migrants, is formally excluded. However, so far all such claims, based on genetic evidence, have been far from convincing.

As a second example, I briefly mention a more recent period paper by Haak et al.(Nature 2005, 310:1016-1918) where the authors from Mainz and their colleagues have been just lucky to find out that the LBK fossils in Europe, excavated from a number of sites, encompassed, at high frequencies, a variant of mtDNA (haplogroup N1a) that is very rare among the present-day Europeans – suggesting, as the most likely interpretation, that the contribution of the first carriers of Neolithic farming culture in central continental Europe into the present-day European mtDNA pool is probably negligible. Whether these first farmers (their mtDNA) came directly from the Near East, is yet another question: to be able to provide indirect, yet convincing evidence for that, one needs to draw, once more, a winning lottery ticket – to find, somewhere in Anatolia or Mesopotamia, early Neolithic skeletons, encompassing N1a-type mtDNA genomes. Unfortunately, one may add here that going southwards, the chance to extract mtDNA from fossil bones becomes increasingly less probable because spontaneous damage of DNA correlates positively with ambient temperature. This is why there seems to be little hope

Hardships involved in working with aDNA are numerous and there is rich literature on it. Even mtDNA extracted from the well-studied Tyrolean Iceman, a mere ~5300 years old body found at the border of Italy and Austria, high in mountains, melted out from glacier, is still re-investigated. The original publication of its mtDNA sequence (fragments of hypervariable region, published about decade ago) documented, in the extracted DNA, the presence of a mixture of at least 5 different mtDNAs, and there is still a controversy about the exact nucleotide sequence of the likely "authentic Iceman" mtDNA variant within this mixture, although it seems to be certain that this man carried Hg K variant of mtDNA. However, this fact does not provide much novelty – very extensive knowledge about the phylogeny and phylogeography of the

extant Europeans suggests that the coalescence age of mtDNA hg K, widespread in Europe, is much deeper than 5000 YBP.

While there was "a hunt" for Neanderthals (their mtDNAs) and several independent studies (different geographically and covering rather wide geography) are by now published , we know very little about pre-glacial and glacial era (early-mid Upper Palaeolithic), as well as early post-glacial, pre-Holocene, aDNA. A few papers published have not, for various reasons, convinced so far "the community". Furthermore – one needs to repeat - singular findings are usually not of great help – archaeogenetics needs "populations" to be studied and singular finding(s), over tens of thousands of years, offer little. Unless something extraordinary would be found. Like South-East Asian mtDNA in remains of children in the urnfield close to Carthage. But that is a different story.

The finding of Haak et al., however, teaches us yet another lesson: signals for expansions can be almost lost, being picked up only by luck. Namely, frequency of N1a in the present-day European mtDNA pool is in average 100-200 times lower than among the first LBK farmers. In theory, there is nothing to wonder, but observing near loss of otherwise potentially clear genetic signal for a revolutionary event in the history of *H. s. s*. in Europe – for the first archaeologically unequivocal, rapid and wide expansion of Neolithic industry in the center of the continent starting some 7500 years ago – makes one wonder how many such signals have been lost forever (in the geneology of our maternally inherited mtDNA). Unless we analyze aDNA and are lucky enough to detect something odd – so odd that accomplished with such work usual worries about systemic contamination, damage-related artifacts etc., can be (almost certainly) ruled out. Perhaps it shows quite clearly how limited can heuristic value of genetic evidence be in drawing conclusions, based on the "lack of evidence". In more general terms – a coalescence tree drawn from data based on experimentally detected genetic variation of a gene is only an approximation of "the true" phylogenetic tree of this gene, because much of the earlier existed variation can be considered as lost forever. It is not a question about the tips of the tree – even extensive sampling does not offer possibilities to catalogue all variation. It is a question of a loss of a supposedly strong signal of migration by a mechanism that cannot be realistically ascribed to random genetic drift within the pioneer community, but to the "dilution" of original carriers of radically new technology that changed the history, among the pre-existing hunters-gatherers. "Neolithic genes" (here: N1a mtDNA) got almost lost likely because this new lifestyle, leading to rapid expansion of population, was very soon picked up by still much more numerous community of hunters-gatherers, who started to expand in numbers equally fast. This is one of the explanations. There may have been also other mechanisms involved, including natural selection. The "aboriginal" HG people were likely better adapted to the local environment and provided they became agriculturalists, this long-time better adaptation would have preserved as selective advantage.
Unfortunately, because of about 1000-fold lower molar concentration of the Y-chromosome, compared that for mtDNA, makes it orders of magnitude more complicated to analyze the corresponding ancient DNA (NRY aDNA). While literature about amtDNA contains tens of papers published during the last decade (forget their reliability for a moment), there is, to the best of my knowledge, no phylogenetically informative NRY aDNA study so far published. Taken together, one should probably not account for rapid progress in this direction and the ideal:

comparing simultaneously, in quantitatively meaningful numbers, amtDNA and NRY aDNA (of the same or close locations) with their present-day counterparts, does not seem to be a target easily achievable.

Sometimes lousy use of words disorientates readers, in particular in papers for interdisciplinary audience

As I already mentioned, in case of mtDNA and Y chromosome, we deal with single loci out of tens of thousands. And their contrasting way of inheritance is an additional aspect to be carefully considered before speaking about "populations".

The second troublemaker is perhaps confusing coalescence age of a particular haplogroup, with its "time depth" – exactly because of sloppy use of terms. Split from the Most Recent Common Ancestor and time signaling expansion of a particular lineage can be very different. Seems trivial, but nevertheless causes confusion now and then. Both in case of mtDNA and Y-chromosome phylogenies. In case of Y-chromosome papers, where time is almost exclusively measured by counting diversity of Short Tandem Repeats (STRs) within a particular Single Nucleotide Polymorphism (SNP)–defined haplogroup, the outcome reflects (at best) a signal from the beginning of the expansion of this cluster. Only a fraction of papers (usually with Lev Zhivotovsky as a co-author) provides estimates that reflect time since the split of the MRCA – sometimes several times older than the former estimate. The same can be seen in mtDNA papers. Most of secondary literature cites, e.g. Richards et al (2000) and declares that "haplogroup U5 is about 40,000 – 50,000 years old". Perhaps close to correct, but some 4-5 years ago it was not so easy to explain to the wider audience that in fact, this complex haplogroup consists of some 5 – 6 sub-haplogroups (with nice star-like phylogenetic trees), exhibiting largely very close to each other coalescence ages around the Pleistocene-Holocene boundary some 10,000 years ago.

Many more examples can be given. In one of the previous SFI meetings I briefly discussed and illustrated a case study based on a common all over West Asia (Iran incl.), North Africa and Europe (extending to Central Asia, West Siberia) haplogroup T. It seems that in this year's meeting it is worth to return to this example, because it demonstrates quite convincingly that: (i) the coalescence age summary for a haplogroup, over wide geographic are, results in an estimate of dubious value; (ii) signal for expansion of the same haplogroup can, even within Europe, differ profoundly; (iii) information obtained from such a more detailed study has heuristic value and can provoke interesting interdisciplinary discussions. For example – if, for the northern part of Europe (Scandinavia, maybe even northern UK, southeast Baltic and northern East Europe), we see simultaneously different (sub)haplogroups of mtDNA lineages which display coalescence ages around 12,000 BP and 3000 BP – should we treat it as a strong hint about migration of the carriers of the latter (shorter coalescence age) to areas where such a haplo was not present earlier? Furthermore, migrations pose particularly complex problems to geneticists, because so much depends on the size of moving populations. One can go to the extreme: recent massive arrival of Europeans to North America brought over Atlantic mtDNA and Y-chromosomal diversity not much different from that in Europe. Hence, their coalescence age calculations would not say anything about the time of this migration and it might seem even silly to bring such an example. It is silly because we know real demographic history. However, for most of the 50,000 – 70,000 years since the likely beginning of the colonization of Eurasia and beyond, hardly anything is firmly

known. And therefore, we can easily fall into pitfalls in interpretations – in situ expansion *vs* massive migration, carrying alongside the pre-existing genetic diversity. Adding here population bottlenecks complicates situation further. This is why archaeological background, first of all at least some estimates about historic (i.e. pre-historic) population densities in a given area, is so valuable.
My planned paper will be about the initial phase of "Out-of-Africa", currently estimated at about 50,000 – 70,000 YBP. I try to argue why, during the last 5 years, there is an increasingly stronger evidence that the major (and perhaps even the only important) route was the coastal one, over the Horn of Africa, around coastal Indian Peninsula, to a branching point in Southeast Asia, with one branch taking people to Australia and the other one to South China.
However, as additional examples (if there will be time), I would like to illustrate migrations by explaining some recent papers, covering very different demographic events, like pinpointing the source population for mtDNA variation in Polynesia, establishing "recent" North African – Northeast European connections, "counting Ashkenazi mothers" – i.e. very different projects that nevertheless share common methodological approach.

Here is the list of a few papers, relevant to my talk and to the "notes":

Out-of-Africa

Macaulay et al. (2005) *Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes*. Science, 308:1034-1036

Thangaraj et al. (2005) *Reconstructing the origin of Andaman Islanders.* Science 308:996.

Sun et al. (2005) *The dizzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes*. Mol. Biol. Evol. Epub ahead of print , 16 Dec.

Polynesia:

Thangaraj et al (2005) *Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations*. PloS Biol. 3(8):e247

Saami

Achilli et al. (2005) *Saami and Berbers – an unexpected mitochondrial DNA link*. Am. J. Hum. Genet. 76:883-886

Tambets et al. (2004) *The western and eastern roots of the Saami – the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes*. Am. J. Hum. Genet. 74:661-682

Ancient DNA

Cooper et al. (2004) *Ancient DNA: Would the real Neanderthal please stand up?* Current Biology 14:431-433.

Haak et al. (2005) *Ancient DNA from the first European farmers in 7500*-year-*old Neolithic sites*. Science, 310:1016-1018

Ashkenazim

Behar et al. (2006) *The matrilineal ancestry of Ashkenazi Jewry: Portrait of a recent founder event*. Am. J. Hum, Genet. Epub ahead of print, 10 Jan