INTERNET MEETS THE HUMAN GENOME

Lovro Rabuzin[a], Réka Gyetvai[b], Julia Hamblin-Trué[c], Lukas Lorenz[d], Inga Patarčić[e]

Affiliations

[a]V. gymnasium, Klaićeva 1, 10000 Zagreb,[b]Bela III Secondary School, 6500 Baja, Szent Imre tér 5., Hungary,[c] IB Programme at the Berlin International School, Lentzeallee 8/14, 14195 Berlin, Germany,[d] G11, Geringergasse 2, 1110 Vienna, Austria,[e] Berlin Institute for Molecular Systems Biology at Max-Delbrück-Centre for Molecular Medicine

Abstract

In this project we studied how genetic background of lactose intolerance, handedness, speech and language comprehension capabilities and cognitive function changed over time and as humans evolved. We compared single nucleotide polymorphisms (SNPs) in genomes of people who lived during 6000 BC to today. Besides humans, we also analyzed genotypes of chimpanzees, orangutans, Neanderthals and Denisovans. First, we identified SNP previously linked to handedness (rs11855415) and eight polymorphisms linked to cognitive functions. Then, for those positions we determined the values of the DNA bases in genomes of human ancestors and related species. From our results, we can assume that the number of left-handed individuals has been increasing with time, but the number of risk alleles for cognitive functions has stayed about the same. Finally, we determined that, based on genomic data from human ancestors, it should be possible to infer their characteristics, which don't manifest on human remains and artefacts.

Introduction:

The technologies used to determine exact values of DNA bases (A – adenine, C - cytosine, G – guanine and T - thymine) in the human genome have gone through a rapid growth and development in the last decade. Genotyping determines the exact values of DNA bases for predefined locations in the genome, today most commonly around 600,000 positions in the genome, while sequencing tries to cover the whole human genome (approximately 3 billion base pairs).

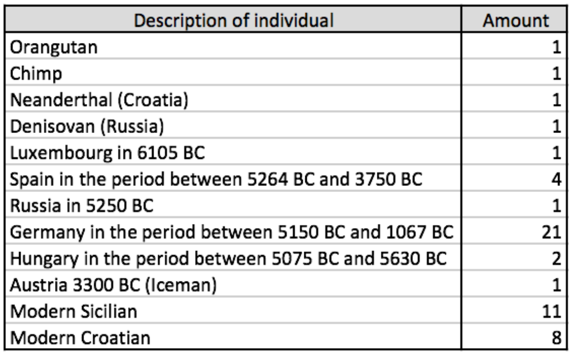
Today, very little genetic material is needed to determine the genotype of the individual, which is why scientists have been able to get the genotypes from the remainings of people who have lived in periods between 6000 BC and 1000 BC from different parts of Europe. Alongside the genotypes of ancient individuals, they also managed to genotype a Neanderthal and a Denisovan, which lived around 30 000 i 41 000 years BC. All of that data is publicly available (Haak et al. 2015).

In this research, we have used the capabilities of the internet, namely the fact that a large number of human genomes is publicly available and we have decided to download the data and analyze the genomes of ancient and modern individuals. More accurately, by comparing the values of DNA bases (A, C, G and T) for positions in the genomes associated with certain characteristics, we can assume the status of these characteristics in ancient humans. That possibility was very fascinating to us because, until now, scientists have been able to assume about the characteristics of human ancestors only based on their remains (bones and artifacts). Since many characteristics of ancient individuals do not manifest on the remains, such as lactose intolerance, handedness, speech and language comprehension capabilities and cognitive functions, we have tried to find out more about these characteristics based on the data from the genomes available to us.

Materials:

Haak *et al.* 2015. contains genotypes of around 2000 persons with 350 thousand positions, but we used the genotypes of only 53 individuals. The chosen genotypes have been chosen either based on the time period in which the individuals lived (spaced out evenly through different periods) or so that the chosen individuals have lived in spatially different locations. Of the 53 chosen genotypes, 19 belong to modern humans (from which 8 of Croatian and 11 of Sardinian descent) as well as one orangutan and one chimpanzee (Table 1.). The reference genome hg19 has been used in all analyses. The GWAS Catalog (https://www.ebi.ac.uk/gwas/) was used to identify locations of single nucleotide polymorphisms in the genome associated with features of interest (SNP corresponds to location in the genome with length of one base and which differs among unrelated individuals).

Table 1. Categories of individuals and corresponding number of genotypes used in the analysis



Description of individual	Amount
Orangutan	1
Chimp	1
Neanderthal (Croatia)	1
Denisovan (Russia)	1
Luxembourg in 6105 BC	1
Spain in the period between 5264 BC and 3750 BC	4
Russia in 5250 BC	1
Germany in the period between 5150 BC and 1067 BC	21
Hungary in the period between 5075 BC and 5630 BC	2
Austria 3300 BC (Iceman)	1
Modern Sicilian	11
Modern Croatian	8

Methods

We used two research approaches to infer characteristics of human ancestors based on their genomic data.

In the first approach, we chose the following phenotypes: 1) lactose intolerance, 2) handedness, which is whether the individuals were right- or left-handed, and 3) capability of speech and comprehension of language. When we chose all three of our desired phenotypes, the next step was to find the polymorphisms which affect how the phenotypes manifest. Polymorphisms associated with these traits were identified in the GWAS Catalog. However, many of the identified polymorphisms weren't used in further analysis because the risk factors for identified alleles were not reported in original publications.

We converted the locations from the reference genome hg38 reported in GWAS Catalog to the reference genome hg19, which was used in Haak *et al.* 2015., using VarSome (https://varsome.com). For further data analysis; determining the exact values of the DNA bases in the genotypes of the individuals, we used the programming language R (R. Core Team 2015) and the integrated development environment (IDE) Rstudio. We also used the function package GenomicRanges to make the process of finding overlaps between our SNPs and the genotypes of the individuals far simpler (Lawrence *et al.* 2013). The analysis was repeated for each of the 53 analysed genomes individually and the results were analysed using risk factors for certain alleles.

The second approach represented the reverse process of analysis. First, we downloaded full GWAS Catalog and then chose only polymorphisms which affect the list of characteristics of interest. In this case, the list of characteristics was a bit longer than in previously since we chose SNPs which affect the cognitive functions of the individuals defined as: psychomotor speed, and/or learning and remembering, and/or intelligence, attention and executive functioning. For identified SNPs we determined the base pairs values in the genomes of previously analyzed individuals and we analyzed the results using risk factors for alleles.

Results

The polymorphisms which we identified and that affect the handedness of an individual are: rs11855415 (Scerri et al. 2010), rs883565 and rs296859 (Armour et al. 2014). For lactose intolerance we found the SNPs rs182549 and rs4988235 (Enattah et al 2002.). For speech and language comprehension capabilities, we found SNPs in the area of the FOXP2 gene (Lai et al. 2001): rs17137124 and rs1456031(Premi *et al.* 2012). Unfortunately, using the first approach (a simple overlap of phenotype associated SNP locations with genotyped positions of the analyzed individuals), we could not identify the DNA bases for most of the polymorphisms since there was no direct overlap. Only a single SNP - rs11855415 - of all the previously mentioned ones, overlapped in location with the full set of 350 thousand analysed polymorphisms from the Haak *et al.* 2015.. Based on this one polymorphisms, results indicate that the number of left-handed people has been increasing with time and growth of the human population.

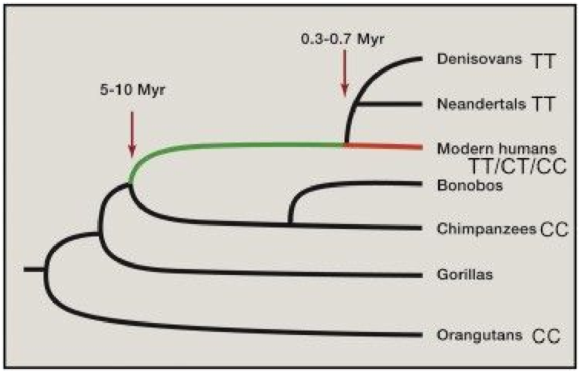
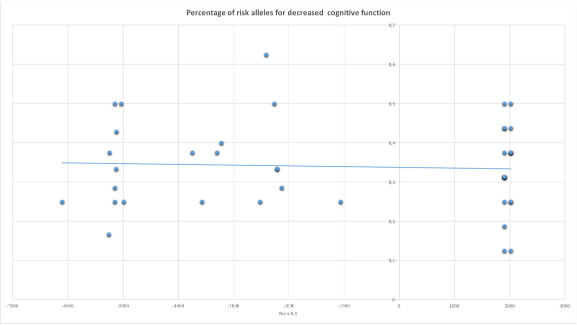


Figure 1. Risk allele variants reported for handedness in hominids and primates overlapping a schematic representation of the evolution tree

We also determined that the analysed individuals of the Neanderthals and the Denisovans have TT on this location in the genome, while orangutans and chimpanzees have CC. In the genomes of the modern humans, we found all three possible variants: CC,CT,TT.

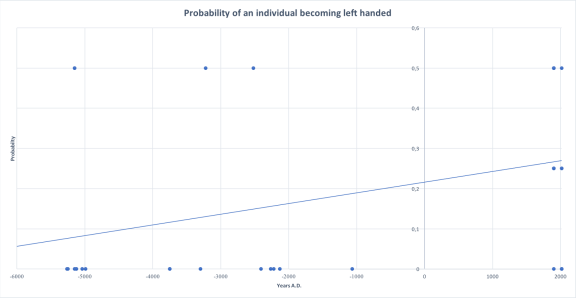
Using the second approach, we identified eleven different polymorphisms which have risk alleles for cognitive functions. However, after the literature search for each one of those SNPs we omitted three of the polymorphisms, as the risk factors for certain alleles had not been mentioned. Eight final SNPs which we included in the further analysis are: rs2300290, rs719714, rs2116081, rs11096990, rs7800418, rs16953622, rs1003247, rs10954361. We determined the base pair values for each of these polymorphisms and analyzed the results using the risk factors.



Graph 1. Percentage of risk alleles for decreased cognitive function. Each dot represents one individual. X-axis corresponds to the time scale (6000BC - today).

Discussion

rs11855415 is the only polymorphism associated with the human characteristic of handedness for which we could determine the values of the DNA bases in the genomes of human ancestors and related species. Individuals can have bases TT, CT or CC for rs11855415, where a larger number of T bases means a greater likelihood of an individual being left-handed. If an individual has the TT allele, they have a 50% chance of becoming left-handed, if they have CT, they have 25% chance of becoming left-handed, and if they have the CC allele, 100% of them are right-handed. The fact that we found CC alleles in the chimps' genomes, confirms previous research which suggested that chimpanzees are predominantly right-handed (Hopkins *et al.* 2004), however 90% of modern humans are right-handed as well (Hardyck *et al.* 1977). Our results indicate that the number of left-handed people has probably been increasing with time and growth of the human population. This was hypothesized in the previous research by Toth et al. 1985, however, before now, it has not been tested on genomes of people from the paleolithic and neolithic.



Graph 2. Probabilities of individuals becoming left handed. Each dot represents one individual. X-axis corresponds to the time scale (6000 BC - today).

Our results therefore confirm the previous scientific researches, but the sample size is not nearly comprehensive enough to infer about whole populations. Likewise, for more detailed analysis, a larger number of SNPs which affect the observed characteristic should be found, as well as using a larger number of genotyped positions in the analysed genomes. Therefore, this paper could be extended by using more recent publications, which contain about a million determined base pairs for the same set of analysed individuals (Mathieson et al. 2015). By also using a larger sample size, more accurate conclusions could be made on a population-wide level. These modifications were not implemented since we had time constraints and both modifications would require a lot of time, taking into account that we were using average computers which are quite slow at processing that kind of data.

As we have already mentioned, by using the dataset that we used, we couldn't find exact values for DNA bases for most polymorphisms associated with lactose intolerance and speech and language comprehension capabilities. By using a more comprehensive dataset, we could probably also identify DNA bases for most of the polymorphisms connected with these human characteristics.

By analysing eight risk alleles associated with cognitive functions, we concluded that the percentage of risk allele has not changed significantly during the time between the paleolithic and today. By using a more comprehensive dataset, we could probably also increase the results of this analysis.

With this research, we have shown that it should be possible to infer about the traits of human ancestors by using their genomic data, especially about the traits that are not manifested on human remains and artifacts.

Conclusions

rs11855415 is the polymorphism associated with the human characteristic of handedness for which we could determine the values of DNA bases in the genomes of human ancestors and related species. We have determined that analysed individuals of the Neanderthals and Denisovans on this location in the genome have TT, whereas orangutans and chimpanzees Have CC. With today's humans we found all three possible variants: CC,CT,TT

By analysing eight risk alleles associated with cognitive functions, we concluded that the percentage of risk alleles has not changed significantly since the paleolithic.

With a more comprehensive dataset, we could probably be able to identify the DNA bases for most of the polymorphisms connected with other human characteristics for which we hadn't been able to find matching genotypes in the analysed dataset.

Literature

Arensburg, Baruch, et al. "A reappraisal of the anatomical basis for speech in Middle Palaeolithic hominids." *American Journal of Physical Anthropology*83.2 (1990): 137-146.

Armour, J. AL, A. Davison, and I. C. McManus. "Genome-wide association study of handedness excludes simple genetic models." *Heredity* 112.3 (2014): 221.

Enattah, Nabil Sabri, et al. "Identification of a variant associated with adult-type hypolactasia." *Nature genetics* 30.2 (2002): 233.

Haak, Wolfgang, et al. "Massive migration from the steppe was a source for Indo-European languages in Europe." *Nature* 522.7555 (2015): 207.

Hardyck, Curtis, and Lewis F. Petrinovich. "Left-handedness." *Psychological bulletin* 84.3 (1977): 385.

Hopkins, William D., et al. "Chimpanzees (Pan troglodytes) are predominantly right-handed: replication in three populations of apes." *Behavioral neuroscience* 118.3 (2004): 659.

Lawrence, Michael, et al. "Software for computing and annotating genomic ranges." *PLoS computational biology* 9.8 (2013): e1003118.

Lai, Cecilia SL, et al. "A forkhead-domain gene is mutated in a severe speech and language disorder." *Nature* 413.6855 (2001): 519.

Lieberman, Philip. "On the Kebara KMH 2 hyoid and Neanderthal speech." (1993): 172-175.

Mathieson, Iain, et al. "Genome-wide patterns of selection in 230 ancient Eurasians." *Nature* 528.7583 (2015): 499-503.

Premi, Enrico, et al. "FOXP2, APOE, and PRNP: new modulators in primary progressive aphasia." *Journal of Alzheimer's Disease* 28.4 (2012): 941-950.

Scerri, Thomas S., et al. "PCSK6 is associated with handedness in individuals with dyslexia." *Human molecular genetics* 20.3 (2010): 608-614.

Uomini, Natalie T. "Handedness in neanderthals." *Neanderthal lifeways, subsistence and technology*. Springer Netherlands, 2011. 139-154.
Team, R. Core. "R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2014." (2015): 99.

Toth, Nicholas. "Archaeological evidence for preferential right-handedness in the lower and middle pleistocene, and its possible implications." *Journal of Human Evolution* 14.6 (1985): 607-614.

https://www.ebi.ac.uk/gwas/ (3.8.2017.)

https://varsome.com (3.8.2017.)