Visualization of distinct DNA regions of the modern human relatively to a Neanderthal genome

Diogo Pratas, Morteza Hosseini, Raquel M. Silva,
Armando J. Pinho & Paulo J. S. G. Ferreira

IEETA/DETI/iBiMED, University of Aveiro, Portugal
{pratas,seyedmorteza,raquelsilva,ap,pjf}@ua.pt

Abstract. Species-specific DNA regions are segments that are unique or share high dissimilarity relatively to close species. Their discovery is important, because they allow the localization of evolutionary traits that are often related to novel functionalities and, sometimes, diseases.
We have detected distinct DNA regions specific in the modern human, when compared to a Neanderthal high-quality genome sequence obtained from a bone of a Siberian woman. The bone is around 50,000 years old and the DNA raw data totalizes more than 418 GB. Since the data size required for localizing efficiently such events is very high, it is not practical to store the model on a table or hash table. Thus, we propose a probabilistic method to map and visualize those regions. The time complexity of the method is linear. The computational tool is available at `http://pratas.github.io/chester`.
The results, computed in approximately two days using a single CPU core, show several regions with documented neanderthal absent regions, namely genes associated with the brain (neurotransmiters and synapses), hearing, blood, fertility and the immune system. However, it also shows several undocumented regions, that may express new functions linked with the evolution of the modern human.

Keywords: DNA patterns, Bloom filters, Ancient DNA, paleogenomics

1 Introduction

Given a set of books and a text, the question "Is this text unique, i.e., absent from each and every book?" poses no fundamental difficulties. The problem becomes much more difficult when we are given a very large collection of text *fragments*, in random order, and containing a large number of changes (substitutions, editions and deletions). The question now is: "Identify and locate in these books all the unique segments, not present in any of these *fragments*". We have shown a way to proceed when the fragments can be assembled in reasonably large texts [1]. Others have even proposed a distance [2]. However, if the volume of the data is very large, with fragments very small, typically from 30 to 300 symbols, and in random order, new challenges arise. This paper proposes solutions to these

challenges, in the context of the recent field of computational paleogenomics, using ancient DNA.

Until a few years ago, it was only possible to obtain DNA sequences from present-day species. Mostly due to the works of Pääbo's group on methods and techniques for retrieving DNA sequences from archaeological and paleontological remains, the first *time travels* to ancient DNA hominins became possible [3–5].

One of the closest and most interesting extinct hominid groups, relative to modern humans, is the Neanderthal. Neanderthals populated Eurasia from 350,000 ± 50,000 to 35,000 ± 5,000 years ago. The availablility of Neanderthal sequences emerged as pieces [3–5], complete mitochondrial [6] and genome draft sequences [7].

The first public *complete* Neanderthal genome was released on 2010 [8], although sequenced with a low coverage. It was acquired from a woman toe phalanx bone with approximately 50,000 years, found in the Denisova Cave, in the Altai mountains of Siberia. This finding made it possibe to analyze the *complete* Neanderthal whole genome at a computational level. The high coverage version (∼30-fold) of the Neanderthal genome was released in 2014 [9].

We use the high coverage whole Neanderthal genome (raw data) to localize and visualize distinct regions of the modern human DNA, using a modern human reference assembly, GRC37 (`https://www.ncbi.nlm.nih.gov/grc`). Recent reports have suggested that modern humans interbred with Neanderthals when they arrived to Europe [10]. This and the similarity between the Neanderthal and human genomes means that any specific, unique regions in the DNA of modern humans may be of very limited extent.

The detection of these regions, using ancient DNA, is a very complex challenge, for the following main four reasons:

1. The large volume of data involved in the analysis ($>$ 418 GB);
2. The need to deal with raw data: ancient DNA is not assembled (random order);
3. Contamination of ancient DNA samples [11];
4. High degree of substitutions in the ancient DNA data, mostly caused by PCR amplifications [12] and postmortem degradation [13].

Aware of these problems, we propose in the next section an unsupervised, probabilistic method that is is alignment-free with respect to the ancient DNA. The method uses a model that is able to trade off precision and space/time resources, maintaining resonable precision values. After describing the method, we show its results: an exhaustive identification of DNA regions that are found in the DNA of modern humans but not on Neanderthal DNA. Finally, we present some conclusions.

2 Method

The straighforward approach to solve this problem is the following: lookup each word of size k (k-mer) found in modern human DNA sequence in the Neanderthal

whole data. Clearly, this approach is totally unfeasible, because, for each word, we would have to perform a search on the entire ancient data (> 418 GB). For this reason, we need to invert the problem [14].

The fundamental idea of our method is the following: imagine that there is a model able to capture all the possible existing words found in the Neanderthal data. This would enable us to determine if words from the modern human DNA sequences are unique or not: we would only have to find out whether they exist in the model or not. Furthermore, we would be able to localize the unique words with precision, because — unlike the ancient DNA data — the modern DNA is assembled.

The crucial task here is of course the design of the model. That is the problem to which we now turn.

2.1 Choosing the model

Given the large volume of the non-assembled data, a good model is crucial for an efficient computation. If one uses a binary vector to store all the possible entries indicating if a certain k-mer exists or not in the sequence, we would need 4^k bits. For $k = 30$, it would require 128 petabytes of memory, which is impracticable on current computers. Basing the model on a data structure such as a hash table would certainly be more reasonable, but the memory usage becomes dependent on the number of inserted elements. In this case, we could need over 284 billion entries. This is still unfeasible in computers with 256 GB of memory. Note that although there is redundancy with a coverage of ∼30, the high degree of substitutions in the ancient DNA data creates a new range of additional words that need to be stored in the hash table.

A third option is a probabilistic data structure, namely a Bloom filter [15], which trades space resources by precision. We have determined that a suitably large Bloom filter, with the optimum number of hash functions, can provide precisions very close to the deterministic approach at a fraction of the memory cost.

The number of elements that need to be presented to the Bloom filter, g, in this case $g = 284,388,216,658$, will be handled by a Bloom filter based on a vector of dimension m. For proper working, the condition $m \geq g$ must be respected. The number of hash functions, h, that minimizes the probability of false positives, is approximately given by

$$h = \frac{m}{g} \ln 2. \tag{1}$$

Asymptotically, for a given false positive probability p, the length m of a Bloom filter is proportional to the number of elements being filtered, g. For finite values, we have

$$p \leq \left(1 - e^{-h(g+0.5)/(m-1)}\right)^h. \tag{2}$$

For the case of multiple hash functions, the method uses universal hashing, given by $f_i(x) = (ax+b) \mod q$, where a and b are two large pseudo-random numbers, different for each i, and q is a prime number such that $q \geq m$.

2.2 Algorithm

After setting the parameters of the size of the Bloom filter (m), the threshold (t) and the k-mer size (k), the method works as follows:

- For each ancient DNA sequence (sN_i):
  - Given m, calculate the number of optimal hash functions (h), using Eq. 1;
  - The probability p is calculated using Eq. 2 and the rounded value of h;
  - Using a slidding window, the model updates each possible k-mer in the Bloom filter;
  - Freeze the model (stop updating).
  - For each human reference chromosomal sequence (sH_i):
    - On the frozen model, search for each k-mer and also for the inverted complemented k-mer, storing the result as a boolean in a file (bH_i);
  - Do an exclusive OR on each index element of each boolean file (bH_i) and store the result in a global boolean file (bH);
  - Filter the global boolean file array (bH), given a certain window, and store it in a file containing reals for each element (fH).
  - Segment the filtered results (fH), according to a threshold t, and store them in a file (pH), along with their relative positions;
- Read the relative positions of the regions from each file stored (pH_i) and paint these in an image.

The disk writes are required mainly because of the volume of data. Thus, computers with disks having a fast access time will provide better performance.

Note that the time complexity of the method is linear in the number of ancient DNA sequences (sN) times n. Therefore, if we concatenate all the sN_i sequences in a single one, the time complexity looses the constant and becomes n. However, for a very large volume of data, setting a Bloom filter without increasing its size increases the probability of false positives. As such, this is a trade-off game.

2.3 Implementation

We have implemented the method in a fully automatic command line tool (CHESTER), written in the C language, so that it can be portable across multiple operating systems. The tool is divided into three programs: CHESTER-map (for mapping the regions), CHESTER-filter (for filtering and segmenting the regions) and CHESTER-visual (for visualizing the regions). The method can be applied to any genomic sequence, in FASTA, FASTQ or SEQ (ACGTN) format. Filtering and visual parameter setting is also available in the program. It can be accessed, under GPLv3 license (free for research studies), at `http://pratas.github.io/chester`.

3 Results

For the results reported in this paper, we have included one script, available at `https://raw.githubusercontent.com/pratas/chester/master/ancient/runNeanderthalGRC37.sh`, that allows replicating the entire results under a Linux OS.

For the analysis in this paper, we have used a Bloom filter with size of 64 GB ($m = 549,755,813,888$). The maximum probability of a false positive, p, was 0.008205. We have split the Neanderthal whole genome into 5 parts, where $sN = 5$ and, hence, using a complexity time of $5n$. In our server (Intel Xeon CPU E7320 at 2.13 GHz), it took approximately 51 hours to run the full experiment (without parallelization).

The full map is shown in Figure 1. As it can be seen, there are multiple regions of singularity/divergence. Comparison with the supplementary information 19b of [9] shows that we have found more regions. This results from the deeper filter used here, namely with 0.95, and while Prufer *et al.* have focused on the regions that had 20-fold excess over randomized rank assignments, we used the whole regions. Moreover, times are much higher for their analysis.

We have found, for thresholds of 0.85, 0.90 and 0.95, respectively, 125, 170 and 2169 regions. From the 170 regions (threshold=0.90), we have searched for genes present in at least half of the region. From those, we have detected: spermatogenesis associated 45 (149643, C1, protein coding), metastasis associated 1 family member 3 (57504, C2, protein coding), major facilitator superfamily domain containing 6 (54842, C2, protein coding), cell adhesion molecule 2 (253559, C3, protein coding), calsyntenin 2 (64084, C3, protein coding), LOC100287290 (100287290, C3, protein coding), sperm tail PG-rich repeat containing 2 (285555, C4, protein coding), TBC1 domain containing kinase (93627, C4, protein coding), pyroglutamylated RFamide peptide receptor (84109, C4, protein coding), KIAA0825 (285600, C5, protein coding), complement C4B (Chido blood group) (721, C6, protein coding), contactin associated protein-like 2 (26047, C7, protein coding), CUB and Sushi multiple domains 1 (64478, C8, protein coding), POTE ankyrin domain family member A (340441, C8, protein coding), leucine rich repeat and Ig domain containing 2 (158038, C9, protein coding), MAM and LDL receptor class A domain containing 1 (340895, C10, protein coding), myosin IIIA (53904, C10, protein coding), chromosome 10 open reading frame 11 (83938, C10, protein coding), MRGPRG antisense RNA 1 (283303, C11, ncRNA), phosphodiesterase 3B (5140, C11, protein coding), ELKS/RAB6-interacting/CAST family member 1 (23085, C12, protein coding), OVOS (408186, C12, protein coding), single-pass membrane protein with coiled-coil domains 2 (341346, C12, protein coding), synaptotagmin 1 (6857, C12, protein coding), HEAT repeat containing 4 (399671, C14, protein coding), neurexin 3 (9369, C14, protein coding), immunoglobulin heavy locus (3492, C14, protein coding), methyltransferase like 22 (79091, C16, protein coding), sorting nexin 29 (92017, C16, protein coding), envoplakin like (645027, C17, protein coding), BCAS3, microtubule associated cell migration factor (54828, C17, protein coding), BCL2, apoptosis regulator (596, C18, protein coding), coiled-coil domain containing 102B (79839, C18, protein

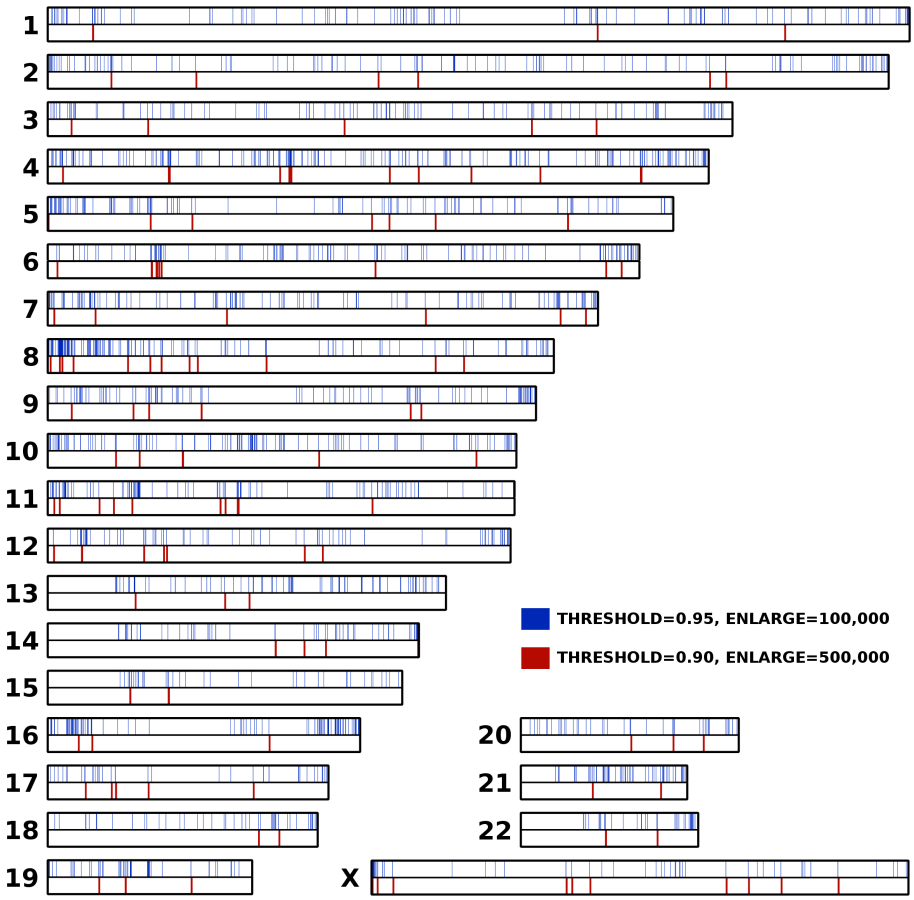

Fig. 1. Modern human chromosomal singular regions relatively to a Neanderthal. CHESTER-map ran with $k = 30$, while CHESTER-filter with "-u 100 -w 20000". Different colors represent specific running parameters, described in the figure. The "ENLARGE" represents a region that is increased with a certain number of bases only for visualization purposes.

coding), trans-2,3-enoyl-CoA reductase (9524, C19, protein coding), breast carcinoma amplified sequence 1 (8537, C20, protein coding), ASMTL antisense RNA 1 (80161, CX, ncRNA), neuroligin 4, X-linked (57502, CX, protein coding), acyl-CoA synthetase long-chain family member 4 (2182, CX, protein coding), sarcoma antigen 2 and pseudogene (644717, CX, pseudo).

The majority are protein coding regions, while a lower part is distributed on pseudo-genes and ncRNA. From the coding regions, we highlight the following: spermatogenesis associated 45 (149643), myosin IIIA (53904), ELKS/RAB6-interacting/CAST family member 1 (23085) and synaptotagmin 1 (6857). The gene spermatogenesis associated 45 (149643) has been described recently as hav-

ing an important role in reproductive efficacy and success [16]. The gene myosin IIIA (53904) encodes a protein that plays an important role in hearing in humans. Three different recessive, loss of function mutations in the encoded protein have been shown to cause nonsyndromic progressive hearing loss [17]. The protein that is encoded by ELKS/RAB6-interacting/CAST family member 1 (23085) is a member of a family of RIM-binding proteins. RIMs are active zone proteins that regulate neurotransmitter release. Changes in the gene have been associated with autism [18]. The gene synaptotagmin 1 (6857) encodes a protein that participates in triggering neurotransmitter release at the synapses [19].

Although these results look very interesting, we need to be aware that the amplification of the sequencing process might be creating several substitutional mutations, namely C→T and G→A [20]. Therefore, future work on these results are needed to study if the differences between these regions may or not be given by these characteristics, and, if yes, to assess its impact.

4 Conclusions

We have proposed a method (and tool) to detect distinct regions of the modern human DNA, when compared to a Neanderthal high-quality genome with more than 418 GB of raw data. The method uses a fully automatic probabilistic approach to map and visualize these regions. The time complexity of the method is linear, being able to compute the results using only two days in a single core CPU.

The results show several regions that are associated with the brain, namely neurotransmiters and synapses, hearing, blood, fertility, immune system, among others. These regions may now be studied acording to their expression and meaning in the evolution path. Other regions have also been detected that, although undocumented, may reveal unique functions in the Neanderthal or in the modern human.

Acknowledgments

We thank Martin Kircher, for very helpful comments and explanations, and Cláudio Teixeira, for computational infrastructures. This work was funded by FEDER (Programa Operacional Factores de Competitividade - COMPETE) and by National Funds through the FCT - Foundation for Science and Technology, in the context of the projects `UID/CEC/00127/2013`, `UID/BIM/04501/2013`, `PTCD/EEI-SII/6608/2014` and the grant `SFRH/BPD/111148/2015` to RMS.

References

1. Pratas, D., Silva, R.M., Pinho, A.J., Ferreira, P.J.S.G.: Detection and visualisation of regions of human DNA not present in other primates. In: Proceedings of the 21st RecPad 2015, Faro, Portugal (October 2015)

2. Rahman, M.S., Alatabbi, A., Athar, T., Crochemore, M., et al.: Absent words and the (dis)similarity analysis of DNA sequences: an experimental study. BMC Research Notes **9**(1) (2016) 186
3. Krings, M., Stone, A., Schmitz, R.W., Krainitzki, H., et al.: Neandertal DNA sequences and the origin of modern humans. Cell **90**(1) (1997) 19–30
4. Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., et al.: Analysis of one million base pairs of Neanderthal DNA. Nature **444**(7117) (2006) 330–336
5. Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., et al.: Sequencing and analysis of Neanderthal genomic DNA. Science **314**(5802) (2006) 1113–1118
6. Green, R.E., Malaspinas, A.S., Krause, J., Briggs, A.W., et al.: A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell **134**(3) (2008) 416–426
7. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., et al.: A draft sequence of the Neandertal genome. Science **328**(5979) (2010) 710–722
8. Reich, D., Green, R.E., Kircher, M., Krause, J., et al.: Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature **468**(7327) (2010) 1053–1060
9. Prüfer, K., Racimo, F., Patterson, N., Jay, F., et al.: The complete genome sequence of a Neanderthal from the Altai Mountains. Nature **505**(7481) (2014) 43–49
10. Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., et al.: An early modern human from Romania with a recent Neanderthal ancestor. Nature **524**(7564) (2015) 216–219
11. Skoglund, P., Northoff, B.H., Shunkov, M.V., Derevianko, A.P., et al.: Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. PNAS **111**(6) (2014) 2229–2234
12. Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A., et al.: DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Research **29**(23) (2001) 4793–4799
13. Briggs, A.W., Stenzel, U., Johnson, P.L., Green, R.E., et al.: Patterns of damage in genomic DNA sequences from a Neandertal. PNAS **104**(37) (2007) 14616–14621
14. Silva, R.M., Pratas, D., Castro, L., Pinho, A.J., Ferreira, P.J.S.G.: Three minimal sequences found in Ebola virus genomes and absent from human DNA. Bioinformatics **31**(15) (April 2015) 2421–2425
15. Bloom, B.H.: Space/time trade-offs in hash coding with allowable errors. Communications of the ACM **13**(7) (July 1970) 422–426
16. Lin, Y.L., Pavlidis, P., Karakoc, E., Ajay, J., Gokcumen, O.: The evolution and functional impact of human deletion variants shared with archaic hominin genomes. Molecular biology and evolution (2015) msu405
17. Qu, R., Sang, Q., Xu, Y., Feng, R., et al.: Identification of a novel homozygous mutation in myo3a in a chinese family with dfnb30 non-syndromic hearing impairment. International journal of pediatric otorhinolaryngology **84** (2016) 43–47
18. Silva, I.M., Rosenfeld, J., Antoniuk, S.A., Raskin, S., Sotomaior, V.S.: A 1.5 mb terminal deletion of 12p associated with autism spectrum disorder. Gene **542**(1) (2014) 83–86
19. Baker, K., Gordon, S.L., Grozeva, D., van Kogelenberg, M., et al.: Identification of a human synaptotagmin-1 mutation that perturbs synaptic vesicle cycling. The Journal of clinical investigation **125**(4) (2015) 1670
20. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., et al.: A high-coverage genome sequence from an archaic Denisovan individual. Science **338**(6104) (2012) 222–226