RoAM: computational reconstruction of ancient methylomes and identification of differentially methylated regions

Yoav Mathov[1,2], Naomi Rosen[1], Chen Leibson[1], Eran Meshorer[1,2], Benjamin Yakir[3], Liran Carmel[1]

[1] Department of Genetics, The Alexander Silberman Institute of Life Sciences, Faculty of Science, the Hebrew University of Jerusalem, Jerusalem 9190401, Israel.

[2] Edmond and Lily Safra Center for Brain Sciences (ELSC), the Hebrew University of Jerusalem, Jerusalem 9190401, Israel.

[3] Department of Statistics and Data Science, the Hebrew University of Jerusalem, Jerusalem 9190500, Israel.

Abstract

Identifying evolutionary changes in DNA methylation bears a huge potential for unraveling adaptations that have occurred in modern humans. Over the past decade, computational methods to reconstruct DNA methylation patterns from ancient DNA sequences have been developed, allowing for the exploration of DNA methylation changes during the past hundreds of thousands of years of human evolution. Here, we introduce a new version of RoAM (Reconstruction of Ancient Methylation), a flexible tool that allows for the reconstruction of ancient methylomes, as well as the identification of differentially methylated regions between ancient populations. RoAM incorporates a series of filtering and quality control steps, resulting in highly reliable DNA methylation maps that exhibit similar characteristics to modern maps. To showcase RoAM's capabilities, we used it to compare ancient methylation patterns between pre- and post-Neolithic revolution samples from the Balkans. Differentially methylated regions separating these populations are shown to be associated with genes related to regulation of sugar metabolism. Notably, we provide evidence for overexpression of the gene PTPRN2 in post-Neolithic revolution samples. PTPRN2 is a key regulator of insulin secretion, and our finding is compatible with hypoinsulinism in pre-Neolithic revolution hunter-gatherers. Additionally, we observe methylation changes in the genes EIF2AK4 and SLC2A5, which provide further evidence to metabolic adaptations to a changing diet during the Neolithic transition. RoAM offers powerful algorithms that position it as a key asset for researchers seeking to identify evolutionary regulatory changes through the lens of paleoepigenetics.

Introduction

Identifying changes in gene expression levels is a powerful tool to study evolutionary shifts and adaptations. Specifically, many efforts have been directed to identifying changes in gene regulation along the human lineage (1–3). The rise of ancient DNA (aDNA) offers a potential to study regulatory changes that shaped modern humans and might have affected our recent evolution. Given that changes in gene regulation are difficult to read directly from the DNA sequence, and that RNA rarely survives in ancient remains, DNA methylation stands out as the best proxy of ancient gene expression levels (2, 4). DNA methylation, which in mammals affects cytosines in the context of CpG positions, is a key epigenetic mark that is tightly associated with gene expression levels (5). Incidentally, unlike other epigenetic marks such as histone modifications, DNA methylation is highly stable and remains on aDNA for extended periods (6).

More than a decade ago, we and others developed computational methods for reconstructing premortem DNA methylation patterns from aDNA (7, 8). These techniques used the fact that deamination – the main chemical degradation of aDNA – turns methylated cytosines into thymines and unmethylated cytosines into uracils (6). In preparing the aDNA for sequencing, a treatment with uracil-specific excision reagent (USER) is often used (6), generating an asymmetry between methylated and unmethylated cytosines, which can be used to distinguish between them by carrying out statistical analysis of the number of $C \rightarrow T$ transitions in CpG positions. A high rate of $C \rightarrow T$ indicates premortem hypermethylation, while the converse indicates hypomethylation. These pioneering methods have not only opened avenues for exploring epigenetic regulations in ancient samples, but also enabled the reconstruction of genome-wide methylation maps and the identification of differentially methylated regions (DMRs) between ancient samples. These works founded the field of paleoepigenetics, which provided significant insights into various aspects of human evolution (2, 4, 9)

Several of these aDNA reconstruction algorithms were published as tools, including RoAM (8), epiPALEOMIX (10), and its successor DAMMET (11). However, RoAM was distributed as Matlab code, and its use was limited. DAMMET, which employs a maximum likelihood estimation to calculate methylation levels based on the $C \rightarrow T$ transition counts, is the most recent tool, but has several limitations. First, it is assumed that all four nucleotides have the same frequency of 0.25, ignoring GC content biases. Specifically, the GC content of the human genome is known to be 40.9% (12). Second, it is assumed that there is an equal probability of 1/7 for each dinucleotide that can be read as a CpG due to a mutation, which does not align with established mutation rates in humans (13, 14). Notably, mutations are not uniformly distributed, with $C \rightarrow T$ mutations being particularly prevalent (15). Third, the estimation procedure includes cytosines outside of a CpG context, which are assumed to be unmethylated. However, some small levels of non-CpG methylation are known to exist (16), especially in embryonic cells and specific mature cell types such as neurons. Not much is known about the prevalence and significance of non-CpG methylation in tissues that are present in the fossil record, particularly bones and teeth. This, combined with the assumption that the deamination rate of cytosines in non-CpG context is identical to that of cytosines within CpG context, can introduce bias into the reconstructed map. Indeed, previous works showed that methylation maps provided by DAMMET show more hypomethylation than expected (17, 18).

Moreover, DAMMET generates methylation maps, but does not compare them to identify DMRs. RoAM, on the other hand, does include a method to detect DMRs, but could originally do so only between pairs of samples. As the number of published aDNA samples continues to grow, the detection of DMRs between large groups of samples has become desirable, as they have the

potential to unveil DNA methylation differences between populations, within the same population across different time points, and between closely related species.

Here, we present a new python version of RoAM (Reconstruction of Ancient Methylation), which removes many of the aforementioned limitations. It is feature-rich, flexible, easy to use, and its code is freely available. It allows for the generation of premortem genome-wide aDNA methylation maps, as well as the detection of DMRs between groups of samples. The current version contains numerous improvements over the original software, including novel methods for filtering out true $C \rightarrow T$ mutations. RoAM does not use assumptions about nucleotide frequencies, and does not rely on cytosines outside of CpG context.

To demonstrate the capabilities of RoAM, we reconstructed the methylomes of 14 samples from the Balkans and detected DMRs between pre-Neolithic revolution specimens and post-Neolithic revolution ones. As expected from the short time span separating these populations, we detected only four DMRs with modest levels of methylation change. However, the genes that are associated with these DMRs might hold clues to nutritional changes that occurred during the Neolithic transition. Notably, PTPRN2, which is involved in insulin response to glucose stimulus, is predicted to be overexpressed in post-Neolithic revolution individuals, as expected for high carbohydrate diet. Methylation changes were also found in EIF2AK4, a sensor for amino acid deprivation that also regulates insulin, and SLC2A5, the main fructose transporter. In total, these findings may provide clues to regulatory changes that might have accompanied the major changes in diet and lifestyle that ensued following the Neolithic revolution.

Methods

RoAM performs two main tasks. First, it reconstructs the methylation map of ancient samples. Then, it compares groups of samples to detect DMRs between them (Figure 1). In the following, we thoroughly describe these two parts.

Part I. Reconstructing premortem ancient DNA methylation

The main input for this part is a BAM file of an ancient individual. Each BAM file is analyzed through five consecutive steps (Figure 1): (1) basic processing of the BAM file; (2) diagnosis step to automatically determine filtering parameters; (3) carrying out the filtering to remove non-informative CpG positions; (4) estimating the deamination rate; and (5) reconstructing the premortem DNA methylation.

In addition to the BAM file, RoAM requires several input parameters that instruct it how to process the specific sample. Although not mandatory, it is highly recommended to provide a reference present-day DNA methylation map generated from the same tissue from which the aDNA was extracted. The reason for this is that such a reference allows for more accurate estimation of the deamination rate and methylation reconstruction. We provide such reference for present-day human bone aligned to hg19, see https://carmelab.huji.ac.il/data.html. A full description of all input parameters can be found in the README file in the GitHub page of RoAM, https://github.com/swidler/roam-python.

RoAM provides the user with two outputs that contain the reconstructed methylation map. One is a simple BED file, and the other is a python object that also contains all parameters that were used by the algorithm. This python object is later required for the DMR detection part.

Step I: BAM file processing

RoAM reads all relevant information from a BAM file into a python object. This object stores some descriptive characteristics of the sample, such as the sample name, species and library preparation method. In addition, it holds relevant summary statistics of the sequencing data, specifically nucleotide counts in each CpG position. At later steps, more information is stored in this object, such as the filtering parameters, the estimated deamination rate, and the reconstructed methylation values. BAM files are processed, one chromosome at a time, using Python's *pysam* module (https://github.com/pysam-developers/pysam). During this step, RoAM filters out low quality reads, and computes base counts for each position in a way that depends on the library preparation method. For single-strand libraries, each strand is treated independently, and only $C \to T$ transitions are relevant. For double-strand libraries, complementary $G \to A$ events along the opposite strand are counted as well.

This step is the most time-consuming part of the pipeline. However, it needs to be executed only once per sample, after which the saved object can be re-used.

Step II: Diagnosis

Filtering in RoAM is aimed at removing CpG positions whose cytosine and thymine counts may be spurious or uninformative to the reconstruction process. We identify two groups of such CpG positions. First, some positions show suspiciously high coverage that suggest they are a result of PCR duplications. Second, some positions show $C \to T$ counts suggestive of premortem mutations rather than deamination.

RoAM runs a diagnostic procedure that automatically suggests parameter values that should be used during filtering. The user can manually override these suggestions.

Identifying PCR duplications

Let t_i and c_i be the counts of thymines and cytosines, respectively, in CpG position i, and let $n_i = t_i + c_i$ be the total count in this position. As a first step, we use a crude outlier removal process to remove positions with extreme values of n_i. Then, we use a more refined method to remove additional outlying positions. For the first step, we compute the 25th and 75th percentiles of all n_i values, denoted p_{25} and p_{75}, respectively. We compute the interquartile range as $r = p_{75} - p_{25}$, and remove all positions where $n_i > p_{75} + s \cdot r$. Here, s is a parameter called *span*, set by default to 5.

For the second step, let $N(c)$ be the histogram of the remaining n_i values, counting the number of CpG positions with coverage c. $N(c)$ resembles normal distribution truncated at 1, and with a heavier right-tail. To account for the truncation, we use the following method to estimate the parameters of the distribution. Let c_m be the coverage level that maximizes $N(c)$. Based on the three coverage levels $c_m - 1, c_m$, and $c_m + 1$, we estimate the parameters a, b, and c of the best-fitting binomial $N(x) = ax^2 + bx + c$. We estimate μ, the mean of the normal distribution, as the point where this function is maximized, $\hat{\mu} = -b/2a$. We call the maximum value of the function $N_0 = N(\hat{\mu}) = -b^2/4a + c$. To estimate the standard deviation, σ, we find the first bin in the histogram, b, such that b is greater than μ and for which $N(b) \leq t \cdot N_0$, where t is a parameter that is set by default to 0.1. The smaller t is, the more the approximation accounts for the heavy tail. Let $f = N(b)/N_0$ be the ratio of these two bins. Then,

$$N(b) = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(b-\hat{\mu})^2}{2\sigma^2}} = f \cdot N_0 = \frac{fN}{\sigma\sqrt{2\pi}},$$

where N is the total number of counts, $N = \sum_c N(c)$. Therefore, σ can be estimated using this ratio by

$$\hat{\sigma} = \frac{b - \hat{\mu}}{\sqrt{2\ln\left({}^{1}/_{f}\right)}}.$$

Given the estimates $\hat{\mu}$ and $\hat{\sigma}$, we further remove CpG positions that might be a result of PCR duplicates by filtering out all positions whose coverage exceeds a threshold c_T, determined as the coverage that would yield an expected value of counts less than one. We find c_T by solving for

$$\frac{N}{\hat{\sigma}\sqrt{2\pi}} e^{-\frac{(c_T-\hat{\mu})^2}{2\hat{\sigma}^2}} < 1,$$

giving

$$c_T = \left\lfloor \hat{\mu} + \hat{\sigma}\sqrt{2\ln\left(\frac{N}{\hat{\sigma}\sqrt{2\pi}}\right)} \right\rfloor.$$

Removal of PCR duplicates is done by removing all positions with

$$n_i > c_T. \tag{1}$$

Detecting $C \to T$ mutations

$C \to T$ mutations are detected in two ways. As we detailed in past papers (8, 19), in aDNA that was sequenced using single-strand libraries, mutations can be distinguished from deamination by examining the opposite strand for $G \to A$ transitions.

Here, we developed an alternative technique, that is also applicable for aDNA sequenced using double-strand libraries. It is based on the analysis of the histogram H of t_i and removing CpG positions with $C \to T$ rates that are too high and likely represent a $C \to T$ mutation. We examine each coverage level c, independently, and therefore can now assume that we are only looking at the $N(c)$ CpG positions whose coverage is c. Denoting by p the probability of a $C \to T$ transition, we assume that H represents a mixture of three binomial distributions: (1) homozygous mutations, where $p_1 \approx 1$. (2) heterozygous mutations, where $p_2 = 0.5$. (3) non-mutated sites that went through deamination, where $p_3 \approx 0.01$; We use expectation maximization (EM) to estimate the p's and w's in the binomial mixture model (BMM):

$$\Pr(t_i = t) = \sum_{k=1}^{3} w_k B(t|c, p_k),$$

where w_k is the weight of the k'th distribution. According to the EM formulation, in each iteration we update the following three magnitudes:

$$r_{ik} = \frac{w_k P_{ki}}{\sum_{k'} w_{k'} P_{k'i}},$$

$$w_k = \frac{1}{n}\sum_{i=1}^{n} r_{ik},$$

$$p_k = \frac{\sum_{i=1}^{n} t_i r_{ik}}{N \sum_{i=1}^{n} r_{ik}}.$$

Here, $P_{ki} = B(t_i|c, p_k)$. To make the computations more efficient, we avoided summing over $i = 1, \dots, n$ and rather summed over the values of the histogram, $g = 0, \dots, N$. Hence, we modified the above iterations to read:

$$r_{gk} = \frac{w_k P_{kg}}{\sum_{k'} w_{k'} P_{k'g}},$$

$$w_k = \frac{1}{n} \sum_{g=1}^{N} H_g r_{gk},$$

$$p_k = \frac{\sum_{g=1}^{N} g \cdot H_g r_{gk}}{N \sum_{g=1}^{N} r_{gk}},$$

where P_{kg} is the value of the k'th distribution for the value of the gth bin. Once we have estimated all these parameters, we decide on the threshold k_c, defined as

$$w_3 B(k_c|c, p_3) = w_2 B(k_c|c, p_2).$$

Writing the binomial distribution explicitly, we get

$$w_3 \binom{c}{k_c} p_3^{k_c} (1-p_3)^{c-k_c} = w_2 \binom{c}{k_c} p_2^{k_c} (1-p_2)^{c-k_c}.$$

Taking log from both sides,

$$\ln w_3 + k_c \ln p_3 + (c - k_c) \ln(1-p_3) = \ln w_2 + k_c \ln p_2 + (c - k_c) \ln(1-p_2).$$

This gives

$$k_c \ln \frac{p_3}{p_2} + c \ln \frac{1-p_3}{1-p_2} - k_c \ln \frac{1-p_3}{1-p_2} = \ln \frac{w_2}{w_3},$$

or

$$k_c = \frac{\ln \frac{w_2}{w_3} - c \ln \frac{1-p_3}{1-p_2}}{\ln \frac{p_3(1-p_2)}{p_2(1-p_3)}}.$$

In our case, we force $p_2 = 0.5$, hence

$$k_c = \left\lceil \frac{\ln \frac{w_2}{w_3} - c \ln 2(1-p_3)}{\ln \frac{p_3}{1-p_3}} \right\rceil,$$

where the $\lceil \cdot \rceil$ operator means the closest integer from above. We then remove all CpG positions whose coverage is c and where $t_i \geq k_c$.

We can evaluate the number of true deaminated positions that are missed by this filtering,

$$w_3 N(c) \sum_{k=k_c}^{N} \binom{c}{k} p_3^k (1-p_3)^{c-k}.$$

Similarly, we can evaluate the number of positions with false deamination, which is

$$w_2 N(c) \sum_{k=0}^{k_c-1} \binom{c}{k} p_2^k (1-p_2)^{c-k} + w_1 N(c) \sum_{k=0}^{k_c-1} \binom{c}{k} p_1^k (1-p_1)^{c-k},$$

which, after substituting $p_2 = 0.5$, gives

$$w_2 N(c) \left(\frac{1}{2}\right)^c \cdot \sum_{k=0}^{k_c-1} \binom{c}{k} + w_1 N(c) \sum_{k=0}^{k_c-1} \binom{c}{k} p_1^k (1-p_1)^{c-k}.$$

These computations of false negative and false positive rates are reported during the diagnosis step.

Next, from within all positions with coverage $1 \leq c \leq c_T$, we remove those for which

$$t_i \geq k_c. \tag{2}$$

Step III: Filtering

The filters outlined in Step II are used to determine which sites to remove. First, we clean PCR duplicates. This is done by setting a threshold c_T and removing every site where $n_i > c_T$ (see Eq. 1). Then, let i denote all the sites with a given coverage level c, where $c = 1, \ldots, c_T$. We remove all sites for which $t_i \geq k_C$ (see Eq. 2), as they are suspected as true mutations.

For single-strand libraries, we add a filter on the $G \to A$ transitions in the opposite strand. To this end, we set a maximum number of allowed A's, a_L, as well as a minimum $G \to A$ ratio, T_m (default: 0.25). Then, we remove all sites where

$$a_i > a_L \text{ and } \frac{a_i}{a_i + g_i} \geq T_m. \tag{3}$$

After filtering, a merging procedure is applied to combine information from both Cs of the same CpG position (on opposite strands). The methylation state of these two Cs should be identical (20), thus merging the counts of Cs and Ts from both strands increases the amount of information obtained from each CpG position.

Step IV: Estimation of deamination rate

The deamination rate is estimated using the same technique we have previously detailed (19), and is based on the C and T counts in CpG positions whose methylation in the modern reference is above a certain threshold, m_h. This parameter should be close to one and is exactly one by default. As the maximum-likelihood estimator of the methylation in a site is

$$m_i = \frac{t_i}{\pi n_i},$$

limiting ourselves to positions where $m_i = 1$ lets us estimate the degradation rate by

$$\hat{\pi} = \frac{\sum t_i}{\sum n_i}.$$

In a case where a reference is not available, one can estimate the degradation rate by assuming knowledge of the global mean methylation in the sample, m_g. Then, we estimate the deamination rate by

$$\hat{\pi} = \frac{\sum t_i}{m_g \sum n_i},$$

where the sum is over all positions in the genomes.

This function also computes the local methylation rate for each chromosome, which allows testing for homogeneity of the deamination rate across chromosomes.

Step V: Methylation reconstruction

By default, RoAM uses histogram matching to reconstruct the methylation maps by finding the non-linear transformation $m_i = f(t_i/n_i)$ that makes the histogram of m_i as close as possible to that of a reference methylation map in a modern bone (21, 22), where m_i is the estimated methylation at position i.

By design, we obtain a histogram of methylation values which resembles the equivalent histogram from modern-day sample. It stands in contrast to DAMMET, which tends to show shifts towards low methylation (Figure 2). We used the χ^2 statistic to compare the histograms of maps reconstructed using RoAM and DAMMET to a bone map that was not used as a reference. Whereas $\chi^2_{RoAM} = 7 * 10^5$, the same statistic for DAMMET was one order of magnitude larger, $\chi^2_{Dammet} = 2.53 * 10^6$, indicating larger distance between the histograms.

Histogram matching serves as the default method for methylation reconstruction, but users may choose two other methods. One is the truncated linear transformation,

$$m_i = \max\left(0, \min\left(1, \frac{t_i}{\pi n_i}\right)\right),$$

as described and used previously (19). To achieve a smooth truncation, RoAM also offers a third method, called the logistic transformation, where

$$m_i = \tanh\left(\frac{t_i}{\pi n_i}\right).$$

Given the typically small thymine counts in each CpG position, t_i, RoAM reconstructs methylation in windows of W consecutive CpG positions (W is always set as an odd number, and the reconstructed methylation in the window is assigned to the middle position). The user may determine the window size they wish to use, but RoAM includes two methods to automatically determine the window size.

Probability-based method. We require that the probability of observing no thymines in a window for a minimum methylation level m_0 be less than p_0. This translates into

$$\Pr(t = 0) = (1 - \pi m_0)^n < p_0,$$

where t is the total thymine count in the window, and n is the total count of thymines and cytosines. Taking log of both sides we get

$$n \cdot \ln(1 - \pi m_0) < \ln p_0,$$

meaning that we have to have

$$n > \frac{\ln p_0}{\ln(1 - \pi m_0)}$$

in the window. If the window is covered by the average effective coverage, C , then $n = WC$. This translates into the following window size:

$$W = \left\lceil \frac{1}{C} \cdot \frac{\ln p_0}{\ln(1 - \pi m_0)} \right\rceil.$$

Relative-error-based method. We require that the relative error in estimating the methylation, when the true methylation is m_0, be lower than $1/k$. Let the estimator for the methylation be

$$m = \frac{t}{\pi n}.$$

Then, its mean is

$$E(m) = \frac{1}{\pi} E\left(\frac{t}{n}\right) = \frac{1}{\pi} \pi m_0 = m_0,$$

and its variance is

$$s^2(m) = \frac{1}{\pi^2} V\left(\frac{t}{n}\right) = \frac{\pi m_0 (1 - \pi m_0)}{\pi^2 n}.$$

We require that $s(m)/E(m) < 1/k$, hence

$$\frac{\pi m_0 (1 - \pi m_0)}{\pi^2 m_0^2 n} < \frac{1}{k^2},$$

or

$$n > \frac{k^2 (1 - \pi m_0)}{\pi m_0}.$$

Again, using the average effective coverage, C, this translates into

$$W = \left\lceil \frac{k^2 (1 - \pi m_0)}{\pi m_0 C} \right\rceil.$$

The default that RoAM uses is the probability-based method.

Part II. DMR detection

Once reconstruction of methylation has been achieved for multiple samples, a second part of RoAM is designed to detect and statistically validate DMRs between two groups of samples. This process comprises the following steps (Figure 1): (1) DMR detection between the two groups, (2) the use of simulations to adjust the parameters of the DMR-calling algorithm to reach a desired level of false discovery rate (FDR), and (3) annotation of the final list of DMRs. The algorithm provides a table with a list of all the DMRs, their location, annotation, and the methylation level in each of the samples, as well as the combined estimated methylation in each group.

Step I: DMR detection

Let us first examine a group of S samples. We assume that the methylation across members of the group is homogeneous, and denote the common methylation value in window j as m_j.

Let us look at sample i. We assume that the observed number of T bases in window j is binomially distributed, $t_{ij} \sim B(n_{ij}, m_j \pi_i)$, where π_i is the deamination rate of the sample and n_{ij} is the sum of the Cs and Ts in each CpG in window *j* in sample *i.* The likelihood of sample i is

$$L_{ij} = \binom{n_{ij}}{t_{ij}} (m_j \pi_i)^{t_{ij}} (1 - m_j \pi_i)^{n_{ij} - t_{ij}},$$

and the log-likelihood

$$\ell_{ij} = t_{ij}\ln\left(m_j\pi_i\right) + \left(n_{ij} - t_{ij}\right)\ln(1 - m_j\pi_i) + B_i,$$

where B_i is a term that is independent of m_j. The total log-likelihood of all S samples in the group

$$\ell_j = \sum_{i=1}^{S} t_{ij}\ln\left(m_j\pi_i\right) + \sum_{i=1}^{S}\left(n_{ij} - t_{ij}\right)\ln\left(1 - m_j\pi_i\right) + B,$$

where $B = \sum_i B_i$ is a term independent of m_j. The score function with respect to m_j is:

$$\frac{d\ell_j}{dm_j} = \sum_{i=1}^{S}\frac{t_{ij}}{m_j} - \sum_{i=1}^{S}\frac{\left(n_{ij} - t_{ij}\right)\pi_i}{1 - m_j\pi_i} = \frac{T_j}{m_j} - \sum_{i=1}^{S}\frac{\left(n_{ij} - t_{ij}\right)\pi_i}{1 - m_j\pi_i},$$

where

$$T_j = \sum_{i=1}^{S} t_{ij}.$$

To look for the maximum likelihood estimator, we should equate this to zero. This can be done numerically, using, e.g., the Newton-Raphson method. For this,

$$\frac{d^2\ell_j}{dm_j^2} = -\frac{T_j}{m_j^2} - \sum_{i=1}^{S}\frac{\left(n_{ij} - t_{ij}\right)\pi_i^2}{\left(1 - m_j\pi_i\right)^2}.$$

Given m_j^t is the approximate solution at iteration t, the solution at iteration $t + 1$ is given by

$$m_j^{t+1} = m_j^t - \frac{d\ell/dm_j(m_j^t)}{d^2\ell/dm_j^2(m_j^t)}.$$

In order to get an initial guess, we may obtain an approximated solution using

$$\frac{1}{1 - m_j\pi_i} \approx 1 + m_j\pi_i.$$

Hence,

$$\frac{d\ell_j}{dm_j} \approx \frac{T_j}{m_j} - \sum_{i=1}^{S}\left(1 + m_j\pi_i\right)\left(n_{ij} - t_{ij}\right)\pi_i.$$

This simplifies into

$$\frac{d\ell_j}{dm_j} \approx \frac{T_j}{m_j} - \sum_{i=1}^{S}\left(n_{ij} - t_{ij}\right)\pi_i - m_j\sum_{(i=1)}^{S}\left(n_{ij} - t_{ij}\right)\pi_i^2.$$

Further approximating by neglecting terms of the order of π_i^2, we get

$$\frac{d\ell_j}{dm_j} \approx \frac{T_j}{m_j} - \sum_{i=}^{S}\left(n_{ij} - t_{ij}\right)\pi_i.$$

This can be written as

$$\frac{d\ell_j}{dm_j} \approx \frac{T_j}{m_j} - N_j^{\pi} + T_j^{\pi},$$

where

$$N_j^{\pi} = \sum_{i=1}^{S} \pi_i n_{ij}\,, \qquad T_j^{\pi} = \sum_{i=1}^{S} \pi_i t_{ij}.$$

The approximate solution is therefore

$$m_j^0 \approx \frac{T_j}{N_j^{\pi} - T_j^{\pi}}.$$

The Fisher information for estimating m_j is equal to the expectation of the negative second derivative of the log-likelihood function,

$$I(m_j) = -E\left(\frac{d^2\ell_j}{dm_j^2}\right) = E\left(\frac{T_j}{m_j^2}\right) + \sum_{i=1}^{S} E\left(\frac{(n_{ij} - t_{ij})\pi_i^2}{(1 - m_j\pi_i)^2}\right)$$

The empirical Fisher information is the evaluation of this negative second derivative at the estimated value of the parameter, and may serve as an approximation of the Fisher information,

$$\hat{I}(\hat{m}_j) = \left(\frac{T_j}{\hat{m}_j^2}\right) + \sum_{i=1}^{S} \frac{(n_{ij} - t_{ij})\pi_i^2}{(1 - \hat{m}_j\pi_i)^2},$$

where $\hat{m}_j$ is the estimator obtained from the iterations of the Newton-Raphson algorithm. The empirical Fisher information is computed as part of the implementation of the algorithm. Finally, we may approximate the variance of the estimator via the inverse of the empirical Fisher information:

$$V(\hat{m}_j) \approx 1/\hat{I}(\hat{m}_j).$$

After computing $m_{1,j}$ and $m_{2,j}$, the methylation levels in every genomic window for the two groups, we can compare the two using a similar approach to the one used in (19). To this end, we define the two statistics

$$\ell_j^{+} = \frac{m_{1,j} - m_{2,j} - \Delta}{\sqrt{V(m_{1,j}) + V(m_{2,j})}},$$

$$\ell_j^{-} = \frac{m_{2,j} - m_{1,j} - \Delta}{\sqrt{V(m_{1,j}) + V(m_{2,j})}}.$$

Here, Δ is a parameter of the algorithm, associated with the desired minimal methylation difference we wish to detect between the two groups. Next, we use a cumulative sum for ℓ_j^{+} and ℓ_j^{-} to identify DMRs, as described in (19). In brief, we define the vectors Q^{+} and Q^{-} of the same length as ℓ_j^{+} and ℓ_j^{-}, as

$$Q_0^{+} = 0, Q_j^{+} = \max\left(Q_{j-1}^{+} + \ell_j^{+}, 0\right).$$

Q^{-} is defined in an analogous way.

In DMRs where group 1 is hypermethylated compared to group 2, we will obtain a sequence of positive values for ℓ_j^+ resulting in an elevation in the values of Q^+. We define the DMR as the region $[a, b]$ between the last zero $Q_a = 0$, and the highest value $Q_b = Q_{max}$ up to the next zero (19).

Each DMR is characterized by several properties, such as its genomic length, the number of CpG positions it harbors, and its Q_{max}. This allows for further filtering of DMRs, to achieve a desired false discovery rate (FDR), as explained in the next section.

Step II: Simulations and FDR

We employ simulations to filter out DMRs in such a way that we achieve a desired FDR level. The detailed procedure can be found in our previous paper (19). In short, we imitate the deamination process in each sample by generating Ts using a binomial process, where the coverage in each position and the deamination rate are kept constant for each sample. The methylation value in each position is determined in advance, and is kept constant across the samples from both groups, to model the null hypothesis of no methylation differences between the groups.

Subsequently, we apply the same DMR detection procedure to the simulated data and count the number of detected DMRs. This number represents the number of DMRs detected under the null hypothesis. Repeating this many times (typically 100 times), we may compute the expected fraction of false DMRs within our original list of DMRs. By default, we set an FDR threshold of 0.05, but this parameter can be adjusted by the user. Given that the simulated DMRs originate from the null hypothesis, they tend to be shorter and have smaller Q_{max}. Consequently, the algorithm applies a range of thresholds for the minimum number of CpG sites and Q_{max}, looking for a combination that would achieve the desired FDR level. If multiple sets of parameters achieve this FDR level, we select the one that filters out the fewest of the original DMRs.

Step III: Annotation

The final step of this part creates annotations of the final DMR list. Two types of annotations are currently implemented: associating DMRs with gene bodies and promoters, and with CpG islands. Users are required to provide the location data (gene list and CpG island list). These files are provided in https://carmelab.huji.ac.il/data.html for hg19. By default, RoAM defines the promoter region of each gene as 5,000 bp upstream of the transcription start site (TSS) to 1,000 bp downstream, but this can be set by the user. In addition, annotation can be done against any list of genomic segments, inserted as a BED file.

Results

Epigenetics in general, and DNA methylation in particular, may respond to changes in internal or external conditions (23, 24). Research has unveiled connections between numerous environmental factors and alterations in DNA methylation (25–28). Consequently, even short bouts of environmental or lifestyle transitions may make epigenomic imprints that can be read.

Motivated by this, we decided to investigate potential epigenetic imprints of the Neolithic transition using RoAM. The Neolithic revolution is a pivotal milestone in human history, representing a significant shift in lifestyle, from primarily that of hunting and gathering to a more sedentary one based on agriculture and animal husbandry. Pre-Neolithic revolution humans typically lived nomadic life, relying on wild plants and animals for sustenance. The adoption of agriculture and domestication practices led to significant changes in diet, disease load, levels of physical activity and many other aspects of life. These changes might have been accompanied by biological and physiological changes (e.g. (29)). Emerging evidence suggests that DNA methylation in some

genomic loci is sensitive to such lifestyle factors (30, 31). Hence, we decided to compare the ancient epigenomes of pre-Neolithic revolution to those of post-Neolithic revolution.

Ancient individuals sequenced to high coverage are still not abundant, and tend to represent very different populations. Comparing pre- to post-Neolithic revolution individuals across many different populations potentially adds confounding factors. To address this, we limited our study to 14 high-coverage individuals that come from the same region, the Balkans. Nine are pre-Neolithic revolution individuals, and five are post-Neolithic revolution ones (Table 1). Two samples, I1116 and I5725 are dated to a later period compared to the other post-Neolithic revolution samples. Whereas this chronological difference might introduce biases to the analysis, we nevertheless decided to include these samples in order to enlarge the number of post-Neolithic revolution individuals, which was anyway already small compared to the pre-Neolithic revolution group. Indeed, as will be shown below, these two samples sometimes show methylation patterns that are somewhat different than those of the other post-Neolithic revolution individuals. Genomic data of these ancient individuals were downloaded from the Allen Ancient Genome Diversity Project (32). Notably, petrous bone was the source for DNA extraction in all 14 samples, further reducing potential effects of confounding factors.

We applied RoAM to reconstruct methylation for each sample and then detected DMRs between these two groups. Given the short time span separating the two groups, we expected to find only small methylation changes between them. We have therefore set the minimum methylation difference between groups (the Δ parameter, see Methods) to the very low value of 0.1.

Methylation patterns can exhibit significant variations between two distinct tissues within the same individual (33). As a result, much of the research in the field of paleoepigenetics has concentrated on the evolutionary aspects of the skeletal system (19, 34). The relevance of skeletal DMRs to changes in lifestyle following the Neolithic transition is debated. However, we have shown that there are loci in the genome where differential methylation in one tissue may reflect differential methylation in another tissue, as long as the methylation change occurs early during embryogenesis (4). To help in focusing on differential methylation that arose during such early developmental times, we crossed our results with published methylation data derived from blood samples of modern hunter-gatherers and farmers in Africa (35), where differentially methylated sites separating these population have been identified. We only considered genes that featured methylation changes in both bone and blood.

Our conservative analysis yielded only four DMRs between these pre- and post-Neolithic revolution populations (Table 2). Three of the DMRs are located inside gene bodies and, interestingly, all four overlap CpG islands, suggesting a possible regulatory role of these methylation changes. The DMR with the highest Q_{max} (406.4) was found inside the gene body of the PTPRN2 gene (also known as IA-2β, Figure 3A), which also harbors the third-highest number of differentially methylated sites in blood, separating modern African hunter-gatherers from farmers, with a total of 62 such sites. PTPRN2 is a transmembrane protein present in dense-core vesicles and represents a major auto antigen of type 1 diabetes (36). Previous works found that PTPRN2 has a key role in insulin secretion in response to glucose stimulus, and suppression or knocking down of this gene can impair this process (37–40). Epigenetic regulation on PTPRN2 has been examined previously, and a DNA methylation change in a CpG site within this gene, which does not overlap with the detected DMR, has been associated with childhood obesity (41).

To find the predicted effects of the DNA methylation changes on the expression level of this gene, we looked at the correlation between the methylation in this DMR and the expression of PTPRN2

across 22 tissues of present-day humans taken from the Roadmap dataset (42). We found a significant positive correlation ($R = 0.65$, $P = 8.8 \cdot 10^{-4}$), suggesting that PTPRN2 was expressed in higher levels in post-Neolithic revolution individuals (Figure 3B). This implies lower insulin response to glucose stimulus in the pre-Neolithic revolution individuals.

Another DMR was detected inside EIF2AK4 (also known as GCN2, Figure 4A), a sensor for amino acid deprivation and a regulator of lipid metabolism and gluconeogenesis (43, 44). This kinase plays a crucial role in maintaining homeostasis during amino acid deprivation. When under a leucine-deprived diet, EIF2AK4 reduces insulin levels and increases insulin sensitivity (45, 46). However, in mice consuming a high fat diet, the opposite effect is shown, where EIF2AK increases blood insulin levels and decreases insulin sensitivity (47). Furthermore, EIF2AK4 is also implicated in diabetes, as its knockout in diabetic mice results in a decrease in serum fasting glucose and improved cardiac symptoms (48).

The third DMR was detected in the gene MAST1 (Figure 4B), a kinase that plays a role in the central nervous system (49–51). No relation of this gene to metabolism or any function that might be related to the Neolithic revolution is currently known. The fourth DMR is not located on a promoter or a gene body (Figure 4C).

Only four DMRs out of an original set of 155,693 had properties that meet the thresholds set by the simulations to get an FDR below 0.05 (see Methods). Although not statistically significant, we noticed an interesting DMR that was very close to crossing the required thresholds. This DMR overlap CpG island, and is located within the SLC2A5 gene (also known as GLUT5, Figure 4D), a major fructose transporter in the gut. Many works showed that the presence of fructose stimulation can enhance, even within hours, SLC2A5 expression in the small intestine of adult animals. The same is true of neonatal and weaning pups that do not normally consume fructose and have low levels of SLC2A5 expression in their intestines (52–57). SLC2A5 is also associated with diabetes and obesity, as the gene is differentially expressed in insulin-sensitive tissues of patients with type 2 diabetes and in mouse models for diabetes and obesity, such as muscle (58) and fat tissues (59, 60).

Discussion

We introduce here RoAM, a user-friendly program designed to provide a complete analysis pipeline for computational reconstruction of ancient methylomes and the identification of DMRs that distinguish ancient populations from each other. As the significance of evolutionary epigenetics is in the rise, RoAM proves to be a valuable tool for researchers seeking to integrate paleoepigenetic insights into their studies.

An advantage of RoAM is that new features are constantly added, gradually providing it with even more power. For example, a primary limitation of the reconstruction algorithm is that it cannot work on low-coverage samples, as the counts of Cs and Ts may be too low to allow for reasonable standard error of the estimator. To overcome this, we have introduced the concept of pooling, where counts from many low-coverage samples from the same group are amalgamated to provide a methylation map that represents the entire population (22). Pooling has already been integrated into RoAM, making it a viable tool for analyzing methylation maps in populations with low-coverage samples.

There are several limitations of the current software. First, the current implementation does not allow comparisons across more than two groups. Second, the code is limited to detecting DMRs between two groups that are exclusively composed of ancient samples. Ideally, we would like to

integrate modern samples in the analyses, such that each group can potentially consist of a mixture of modern and ancient samples. Third, the algorithm exclusively performs methylation reconstruction on samples subjected to USER treatment (6). However, this treatment is not universally performed in aDNA library preparation. Finally, it does not account for difference in deamination rates along the read (6). We are currently actively working on developing solutions for all these limitations. RoAM will continue to be maintained and updated, with each solution promptly implemented in the code, providing RoAM with the capability to handle a growing number of samples of different types.

To showcase the algorithm, we conducted methylation reconstruction on 14 pre- and post-Neolithic revolution samples from the Balkans, and identified four DMRs that distinguish between them. The genes associated with these DMRs provide insights into understanding dietary changes that were induced by the Neolithic revolution. Two classic hypotheses claim that hypoinsulinism in pre-Neolithic revolution hunter gatherers provided an adaptive advantage. The carnivore connection hypothesis (61, 62) suggests that hunter-gatherers diet was high in protein and low in carbohydrates, and that in such conditions, insulin resistance would confer an evolutionary advantage, as it allows redirection of glucose to specific requirements such as embryonic development and brain functions. The thrifty genotype hypothesis (63) suggests that hypoinsulinism was a preferred strategy for storing food in times of food scarcity due to the instability in food sources. Glucose is specifically important for fetal development and to the function of the brain, and therefore hypoinsulinism can be a good adaptation to accommodate and supply the body needs when experiencing low or unstable glucose availability. In line with these claims, previous studies reported that hunter-gatherers from the north-western Kalahari display lower levels of blood insulin, while genetically similar communities that have adopted a sedentary lifestyle for 15 years show an increase in insulin levels during this period (64, 65). Additional work showed that short term consumption of a paleolithic diet can decrease insulin secretion (66). These works indicate hypoinsulinism in hunter gatherers, and specifically at lower insulin secretion. Our most pronounced DMR resides in PTPRN2, suggesting overexpression of this gene in post-Neolithic revolution individuals compared to pre-Neolithic revolution ones. Given its role in insulin secretion in response to glucose, this finding lends further credence to the claim that hunter-gatherers experienced hypoinsulinism. Further evidence for methylation changes in PTPRN2 that correlate with hunting and gathering lifestyle can be found in an independent study that compared methylation levels in modern hunter gatherers and genetically related farmers in Africa (35). In this study, PTPRN2 stand out as the gene with the third-highest number of differentially methylated sites, amounting to a total of 62 sites.

Another DMR lies with the EIF2AK4 gene. EIF2AK4 regulates insulin level and sensitivity in response to varied dietary components, and specifically during amino-acid deprivation. Changes in the expression level of this gene may be linked to the dietary shift during the Neolithic revolution. A plausible explanation for this change could be attributed to food scarcity, potentially resulting in the deprivation of certain amino acids for hunter-gatherers.

We also found a DMR within the SLC2A5 gene. This DMR is filtered out because of our very strict criteria, but it was just below the threshold, so we decided to discuss it here, as it might be potentially related to the Neolithic dietary transition. SLC2A5 is a fructose transporter, whose expression levels change when fructose consumption is increased. Our data do not allow us to determine the sign of the correlation between the methylation in this DMR and the gene expression, hence we cannot conclusively determine whether the methylation changes are associated with up or

down regulation of this gene in post-Neolithic revolution times. However, we do observe a notable methylation change in this gene, that warrants further experimental examination.

It should be recognized that the DMRs we detect here likely represent just a small minority of the methylated changes that accompanied the Neolithic transition. First, we have used very strict filtering criteria, increasing precision on the expense of sensitivity. Second, the use of DNA methylation maps from bones means that we are able to identify only those methylation changes that occurred very early during embryogenesis, and simultaneously affect multiple tissues (4). We will not be able to observe tissue-specific methylation changes, where the methylation change is not shared with bone. We anticipate the existence of such tissue-specific methylation changes, particularly in genes associated with immune functions and metabolism.

Acknowledgements

This publication was made possible through the support of a grant from the John Templeton Foundation (Grant ID# 61739 to L.C and E.M). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. This study was also funded by the Israel Science Foundation [ISF grant 2436/22 to L.C. and B.Y.]. We wish to thank Shani Vaknine Treidel for help with the graphics.

L.C. is the Snyder Granadar chair in Genetics. E.M. is the Arthur Gutterman Family chair for Stem Cell Research.

References

1. Reilly,S.K. and Noonan,J.P. (2016) Evolution of Gene Regulation in Humans. *Annu. Rev. Genomics Hum. Genet.*, **17**, 45–67.

2. Mathov,Y., Batyrev,D., Meshorer,E. and Carmel,L. (2020) Harnessing epigenetics to study human evolution. *Curr. Opin. Genet. Dev.*, **62**, 23–29.

3. Gokhman,D., Agoglia,R.M., Kinnebrew,M., Gordon,W., Sun,D., Bajpai,V.K., Naqvi,S., Chen,C., Chan,A., Chen,C., *et al.* (2021) Human-chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. *Nat. Genet.*, **53**, 467–476.

4. Gokhman,D., Meshorer,E. and Carmel,L. (2016) Epigenetics: It's Getting Old. Past Meets Future in Paleoepigenetics. *Trends Ecol. Evol.*, **31**, 290–300.

5. Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.

6. Briggs,A.W., Stenzel,U., Meyer,M., Krause,J., Kircher,M. and Pääbo,S. (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.*, **38**, e87.

7. Pedersen,J.S., Valen,E., Velazquez,A.M.V., Parker,B.J., Rasmussen,M., Lindgreen,S., Lilje,B., Tobin,D.J., Kelly,T.K., Vang,S., *et al.* (2014) Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.*, **24**, 454–466.

8. Gokhman,D., Lavi,E., Prüfer,K., Fraga,M.F., Riancho,J.A., Kelso,J., Pääbo,S., Meshorer,E. and Carmel,L. (2014) Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science*, **344**, 523–527.

9. Llamas,B., Willerslev,E. and Orlando,L. (2017) Human evolution: a tale from ancient genomes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **372**, 20150484.

10. Hanghøj,K., Seguin-Orlando,A., Schubert,M., Madsen,T., Pedersen,J.S., Willerslev,E. and Orlando,L. (2016) Fast, Accurate and Automatic Ancient Nucleosome and Methylation Maps with epiPALEOMIX. *Mol. Biol. Evol.*, **33**, 3284–3298.

11. Hanghøj,K., Renaud,G., Albrechtsen,A. and Orlando,L. (2019) DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage. *GigaScience*, **8**, giz025.

12. Piovesan,A., Pelleri,M.C., Antonaros,F., Strippoli,P., Caracausi,M. and Vitale,L. (2019) On the length, weight and GC content of the human genome. *BMC Res. Notes*, **12**, 106.

13. Nachman,M.W. and Crowell,S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.

14. Lynch,M. (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci.*, **107**, 961–968.

15. Beletskii,A. and Bhagwat,A.S. (1996) Transcription-induced mutations: Increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.*, **93**, 13919–13924.

16. Jang,H.S., Shin,W.J., Lee,J.E. and Do,J.T. (2017) CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. *Genes*, **8**, 148.

17. Sawyer,S., Gelabert,P., Yakir,B., Lizcano,A.L., Sperduti,A., Bondioli,L., Cheronet,O., Neugebauer-Maresch,C., Teschler-Nicola,M., Novak,M., *et al.* (2023) Improved detection of methylation in ancient DNA. 10.1101/2023.10.31.564722.

18. Borodko,D.D., Zhenilo,S.V. and Sharko,F.S. (2023) Search for differentially methylated regions in ancient and modern genomes. *Vavilov J. Genet. Breed.*, **27**, 820–828.

19. Gokhman,D., Nissim-Rafinia,M., Agranat-Tamir,L., Housman,G., García-Pérez,R., Lizano,E., Cheronet,O., Mallick,S., Nieves-Colón,M.A., Li,H., *et al.* (2020) Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nat. Commun.*, **11**, 1189.

20. Breiling,A. and Lyko,F. (2015) Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics Chromatin*, **8**, 24.

21. Gonzalez, Refael C. and Woods, Richard E. (1992) Digital Image Processing 3rd ed. Pearson Education.

22. Barouch,A., Mathov,Y., Meshorer,E., Yakir,B. and Carmel,L. (2024) Reconstructing DNA methylation maps of ancient populations. *Nucleic Acids Res.*, 10.1093/nar/gkad1232.

23. Yona,A.H., Frumkin,I. and Pilpel,Y. (2015) A Relay Race on the Evolutionary Adaptation Spectrum. *Cell*, **163**, 549–559.

24. Gokhman,D., Malul,A. and Carmel,L. (2017) Inferring Past Environments from Ancient Epigenomes. *Mol. Biol. Evol.*, **34**, 2429–2438.

25. Heijmans,B.T., Tobi,E.W., Stein,A.D., Putter,H., Blauw,G.J., Susser,E.S., Slagboom,P.E. and Lumey,L.H. (2008) Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 17046–17049.

26. Galanter,J.M., Gignoux,C.R., Oh,S.S., Torgerson,D., Pino-Yanes,M., Thakur,N., Eng,C., Hu,D., Huntsman,S., Farber,H.J., *et al.* (2017) Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *eLife*, **6**, e20532.

27. Waterland,R.A., Kellermayer,R., Laritsky,E., Rayco-Solon,P., Harris,R.A., Travisano,M., Zhang,W., Torskaya,M.S., Zhang,J., Shen,L., *et al.* (2010) Season of Conception in Rural Gambia Affects DNA Methylation at Putative Human Metastable Epialleles. *PLOS Genet.*, **6**, e1001252.

28. Portales-Casamar,E., Lussier,A.A., Jones,M.J., MacIsaac,J.L., Edgar,R.D., Mah,S.M., Barhdadi,A., Provost,S., Lemieux-Perreault,L.-P., Cynader,M.S., *et al.* (2016) DNA methylation signature of human fetal alcohol spectrum disorder. *Epigenetics Chromatin*, **9**, 25.

29. Latham,K. (2013) Human Health and the Neolithic Revolution: an Overview of Impacts of the Agricultural Transition on Oral Health, Epidemiology, and the Human Body. *Neb. Anthropol.*

30. McKay,J.A. and Mathers,J.C. (2011) Diet induced epigenetic changes and their implications for health. *Acta Physiol.*, **202**, 103–118.

31. Zhang,Y. and Kutateladze,T.G. (2018) Diet and the epigenome. *Nat. Commun.*, **9**, 3375.

32. Mallick,S., Micco,A., Mah,M., Ringbauer,H., Lazaridis,I., Olalde,I., Patterson,N. and Reich,D. (2023) The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes. *bioRxiv*, 10.1101/2023.04.06.535797.

33. Pai,A.A., Bell,J.T., Marioni,J.C., Pritchard,J.K. and Gilad,Y. (2011) A Genome-Wide Study of DNA Methylation Patterns and Gene Expression Levels in Multiple Human and Chimpanzee Tissues. *PLOS Genet.*, **7**, e1001316.

34. Gokhman,D., Mishol,N., de Manuel,M., de Juan,D., Shuqrun,J., Meshorer,E., Marques-Bonet,T., Rak,Y. and Carmel,L. (2019) Reconstructing Denisovan Anatomy Using DNA Methylation Maps. *Cell*, **179**, 180-192.e10.

35. Fagny,M., Patin,E., MacIsaac,J.L., Rotival,M., Flutre,T., Jones,M.J., Siddle,K.J., Quach,H., Harmant,C., McEwen,L.M., *et al.* (2015) The epigenomic landscape of African rainforest hunter-gatherers and farmers. *Nat. Commun.*, **6**, 10047.

36. Lu,J., Li,Q., Xie,H., Chen,Z.J., Borovitskaya,A.E., Maclaren,N.K., Notkins,A.L. and Lan,M.S. (1996) Identification of a second transmembrane protein tyrosine phosphatase, IA-2beta, as an autoantigen in insulin-dependent diabetes mellitus: precursor of the 37-kDa tryptic fragment. *Proc. Natl. Acad. Sci. U. S. A.*, **93**, 2307–2311.

37. Kubosaki,A., Gross,S., Miura,J., Saeki,K., Zhu,M., Nakamura,S., Hendriks,W. and Notkins,A.L. (2004) Targeted Disruption of the IA-2β Gene Causes Glucose Intolerance and Impairs Insulin Secretion but Does Not Prevent the Development of Diabetes in NOD Mice. *Diabetes*, **53**, 1684–1691.

38. Kubosaki,A., Nakamura,S. and Notkins,A.L. (2005) Dense Core Vesicle Proteins IA-2 and IA-2β: Metabolic Alterations in Double Knockout Mice. *Diabetes*, **54**, S46–S51.

39. Cai,T., Hirai,H., Zhang,G., Zhang,M., Takahashi,N., Kasai,H., Satin,L.S., Leapman,R.D. and Notkins,A.L. (2011) Deletion of Ia-2 and/or Ia-2β in mice decreases insulin secretion by reducing the number of dense core vesicles. *Diabetologia*, **54**, 2347–2357.

40. Xu,H., Abuhatzira,L., Carmona,G.N., Vadrevu,S., Satin,L.S. and Notkins,A.L. (2015) The Ia-2β intronic miRNA, miR-153, is a negative regulator of insulin and dopamine secretion through its effect on the Cacna1c gene in mice. *Diabetologia*, **58**, 2298–2306.

41. Lee,S. (2019) The association of genetically controlled CpG methylation (cg158269415) of protein tyrosine phosphatase, receptor type N2 (PTPRN2) with childhood obesity. *Sci. Rep.*, **9**, 4855.

42. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

43. Hinnebusch,A.G. (1994) The eIF-2α kinases: regulators of protein synthesis in starvation and stress. *Semin. Cell Biol.*, **5**, 417–426.

44. Xu,X., Hu,J., McGrath,B.C. and Cavener,D.R. (2013) GCN2 regulates the CCAAT enhancer binding protein beta and hepatic gluconeogenesis. *Am. J. Physiol.-Endocrinol. Metab.*, **305**, E1007–E1017.

45. Xiao,F., Huang,Z., Li,H., Yu,J., Wang,C., Chen,S., Meng,Q., Cheng,Y., Gao,X., Li,J., *et al.* (2011) Leucine Deprivation Increases Hepatic Insulin Sensitivity via GCN2/mTOR/S6K1 and AMPK Pathways. *Diabetes*, **60**, 746–756.

46. Yin,H., Yuan,F., Jiao,F., Niu,Y., Jiang,X., Deng,J., Guo,Y., Chen,S., Zhai,Q., Hu,C., *et al.* (2021) Intermittent Leucine Deprivation Produces Long-lasting Improvement in Insulin Sensitivity by Increasing Hepatic Gcn2 Expression. *Diabetes*, **71**, 206–218.

47. Liu,S., Yuan,J., Yue,W., Bi,Y., Shen,X., Gao,J., Xu,X. and Lu,Z. (2018) GCN2 deficiency protects against high fat diet induced hepatic steatosis and insulin resistance in mice. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.*, **1864**, 3257–3267.

48. Feng,W., Lei,T., Wang,Y., Feng,R., Yuan,J., Shen,X., Wu,Y., Gao,J., Ding,W. and Lu,Z. (2019) GCN2 deficiency ameliorates cardiac dysfunction in diabetic mice by reducing lipotoxicity and oxidative stress. *Free Radic. Biol. Med.*, **130**, 128–139.

49. Tripathy,R., Leca,I., van Dijk,T., Weiss,J., van Bon,B.W., Sergaki,M.C., Gstrein,T., Breuss,M., Tian,G., Bahi-Buisson,N., *et al.* (2018) Mutations in MAST1 Cause Mega-Corpus-Callosum Syndrome with Cerebellar Hypoplasia and Cortical Malformations. *Neuron*, **100**, 1354-1368.e5.

50. Jing,T., Ma,J., Zhao,H., Zhang,J., Jiang,N. and Ma,D. (2020) MAST1 modulates neuronal differentiation and cell cycle exit via P27 in neuroblastoma cells. *FEBS Open Bio*, **10**, 1104–1114.

51. Garland,P., Quraishe,S., French,P. and O'Connor,V. (2008) Expression of the MAST family of serine/threonine kinases. *Brain Res.*, **1195**, 12–19.

52. David,E.S., Cingari,D.S. and Ferraris,R.P. (1995) Dietary Induction of Intestinal Fructose Absorption in Weaning Rats. *Pediatr. Res.*, **37**, 777–782.

53. Shu,R., David,E.S. and Ferraris,R.P. (1997) Dietary fructose enhances intestinal fructose transport and GLUT5 expression in weaning rats. *Am. J. Physiol.-Gastrointest. Liver Physiol.*, **272**, G446–G453.

54. Shu,R., David,E.S. and Ferraris,R.P. (1998) Luminal fructose modulates fructose transport and GLUT-5 expression in small intestine of weaning rats. *Am. J. Physiol.-Gastrointest. Liver Physiol.*, **274**, G232–G239.

55. Jiang,L., David,E.S., Espina,N. and Ferraris,R.P. (2001) GLUT-5 expression in neonatal rats: crypt-villus location and age-dependent regulation. *Am. J. Physiol.-Gastrointest. Liver Physiol.*, **281**, G666–G674.

56. Cui,X.-L., Ananian,C., Perez,E., Strenger,A., Beuve,A.V. and Ferraris,R.P. (2004) Cyclic AMP Stimulates Fructose Transport in Neonatal Rat Small Intestine. *J. Nutr.*, **134**, 1697–1703.

57. Douard,V., Cui,X.-L., Soteropoulos,P. and Ferraris,R.P. (2008) Dexamethasone Sensitizes the Neonatal Intestine to Fructose Induction of Intestinal Fructose Transporter (Slc2A5) Function. *Endocrinology*, **149**, 409–423.

58. Stuart,C.A., Howell,M.E.A. and Yin,D. (2007) Overexpression of GLUT5 in Diabetic Muscle Is Reversed by Pioglitazone. *Diabetes Care*, **30**, 925–931.

59. Hajduch,E., Darakhshan,F. and Hundal,H.S. (1998) Fructose uptake in rat adipocytes: GLUT5 expression and the effects of streptozotocin-induced diabetes. *Diabetologia*, **41**, 821–828.

60. Litherland,G.J., Hajduch,E., Gould,G.W. and Hundal,H.S. (2004) Fructose transport and metabolism in adipose tissue of Zucker rats: Diminished GLUT5 activity during obesity and insulin resistance. *Mol. Cell. Biochem.*, **261**, 23–33.

61. Brand Miller,J.C. and Colagiuri,S. (1994) The carnivore connection: dietary carbohydrate in the evolution of NIDDM. *Diabetologia*, **37**, 1280–1286.

62. Brand-Miller,J.C., Griffin,H.J. and Colagiuri,S. (2011) The Carnivore Connection Hypothesis: Revisited. *J. Obes.*, **2012**, e258624.

63. Neel,J.V. (1962) Diabetes Mellitus: A “Thrifty” Genotype Rendered Detrimental by “Progress”? *Am. J. Hum. Genet.*, **14**, 353–362.

64. Joffe,B.I., Jackson,W.P.U., Thomas,M.E., Toyer,M.G., Keller,P., Pimstone,B.L. and Zamit,R. (1971) Metabolic Responses to Oral Glucose in the Kalahari Bushmen. *Br. Med. J.*, **4**, 206–208.

65. *Jenkins,T.,**Joffe, B.I., *Panz, V.R., ***Ramsay, M. and ***Seftel H.C. (1987) Transition from a hunter-gatherer to a settled life-style among the !Kung San (bushmen): effect on glucose tolerance and insulin secretion. *South Afr. J. Sci.*, **83**, 410.

66. Frassetto,L.A., Schloetter,M., Mietus-Synder,M., Morris,R.C. and Sebastian,A. (2009) Metabolic and physiologic improvements from consuming a paleolithic, hunter-gatherer type diet. *Eur. J. Clin. Nutr.*, **63**, 947–955.

Tables

Sample name	Group	Sample Age (K years)	Location	Coverage
I1507	Pre-Neolithic revolution	7.7	Hungary (Tiszaszolos-Domaháza)	22.42
I4873	Pre-Neolithic revolution	7.9	Serbia (Vlasak)	25.76
I4875	Pre-Neolithic revolution	8.5	Serbia (Vlasak)	21.48
I4877	Pre-Neolithic revolution	8.5	Serbia (Vlasak)	27.44
I4878	Pre-Neolithic revolution	7.8	Serbia (Vlasak)	25.3
I4914	Pre-Neolithic revolution	8.1	Serbia (Vlasak)	24.85
I5233	Pre-Neolithic revolution	8	Serbia (Padina)	23.55
I5235	Pre-Neolithic revolution	10.8	Serbia (Padina)	25.61
I5236	Pre-Neolithic revolution	10	Serbia (Padina)	26.91
I1116	Post-Neolithic revolution	1	Serbia (Gomolova)	26.97
I1496	Post-Neolithic revolution	7	Hungary (Apc-Berekalya I)	29.96
I2520	Post-Neolithic revolution	5.1	Bulgaria (Dzhulyunitsa)	24.47
I5077	Post-Neolithic revolution	7	Croatia (Sopot)	27.79
I5725	Post-Neolithic revolution	2.5	Croatia (Sv Kriz)	27.49

Table 1. List of samples used in this work. Sample names are taken from the Allen Ancient Genome Diversity Project.

Chrom.	DMR start (hg19)	DMR end (hg19)	Qmax	# CpGs	Gene	Pre-Neolithic revolution methylation	Post-Neolithic revolution methylation	Methylation difference
7	157405128	157407252	406.4	145	PTPRN2	0.36	0.65	0.29
2	131008391	131011881	311.9	119		0.35	0.64	0.29
19	12983432	12985003	287.4	118	MAST1	0.50	0.77	0.27
15	40266418	40269319	255.8	79	EIF2AK4	0.40	0.74	0.34

Table 2. DMRs separating pre- and post-Neolithic revolution samples from the Balkan (ordered by $Q_{\max}$, from largest to smallest).

Figures

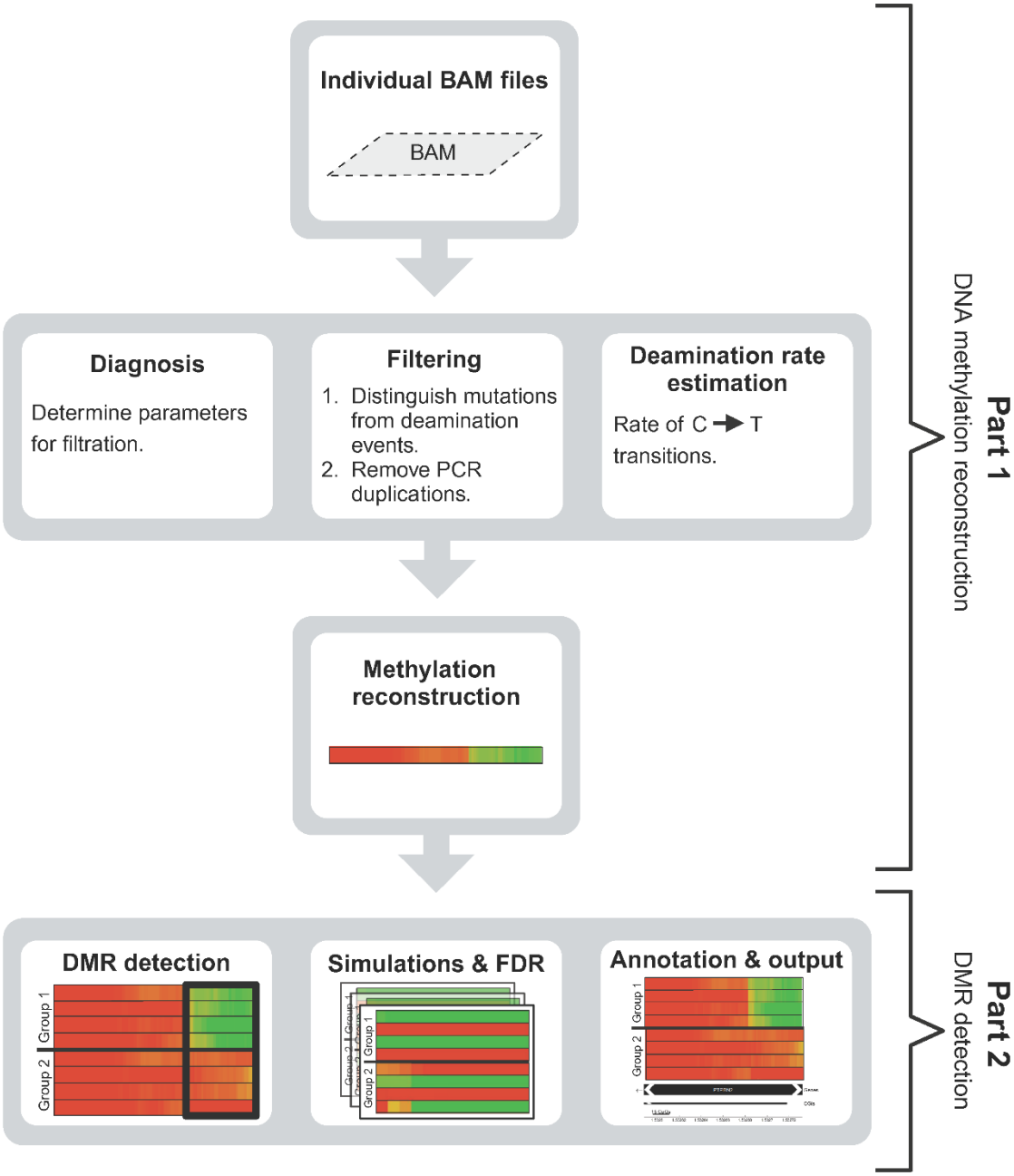

Figure 1. The RoAM pipeline is split into two parts. In Part I, RoAM starts with BAM files of ancient genomes, and reconstructs the individual methylation maps. In Part II, RoAM detects differentially methylated regions (DMRs) distinguishing two groups of ancient samples. (created with Biorender)

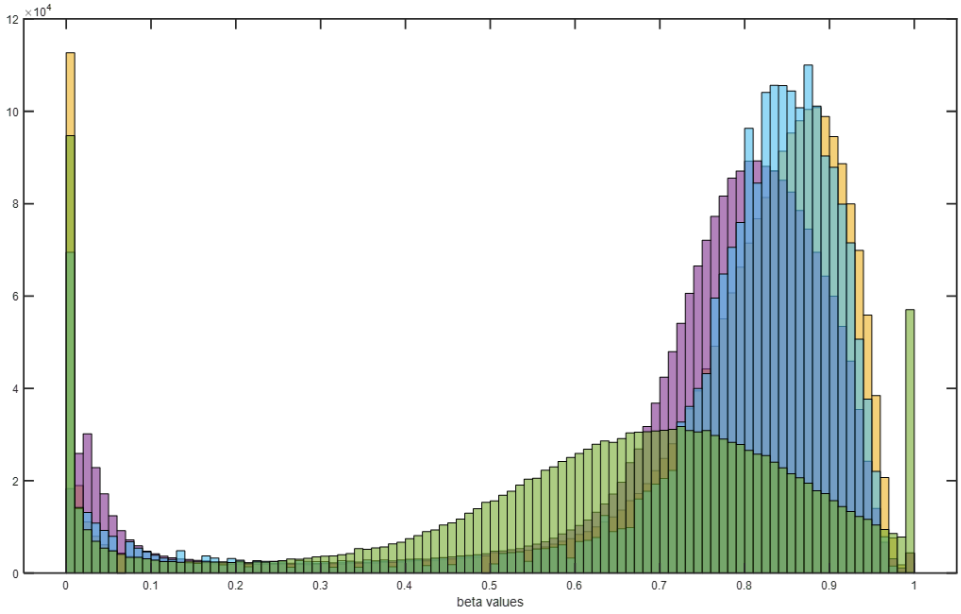

Figure 2. Genome-wide histograms of methylation levels (chromosome 1). Ancient DNA methylation in sample I1116 was reconstructed by RoAM (blue) and by DAMMET (green). Two modern bone samples are shown. One (yellow) was used as a reference in RoAM, and another (purple) that was not used in RoAM.

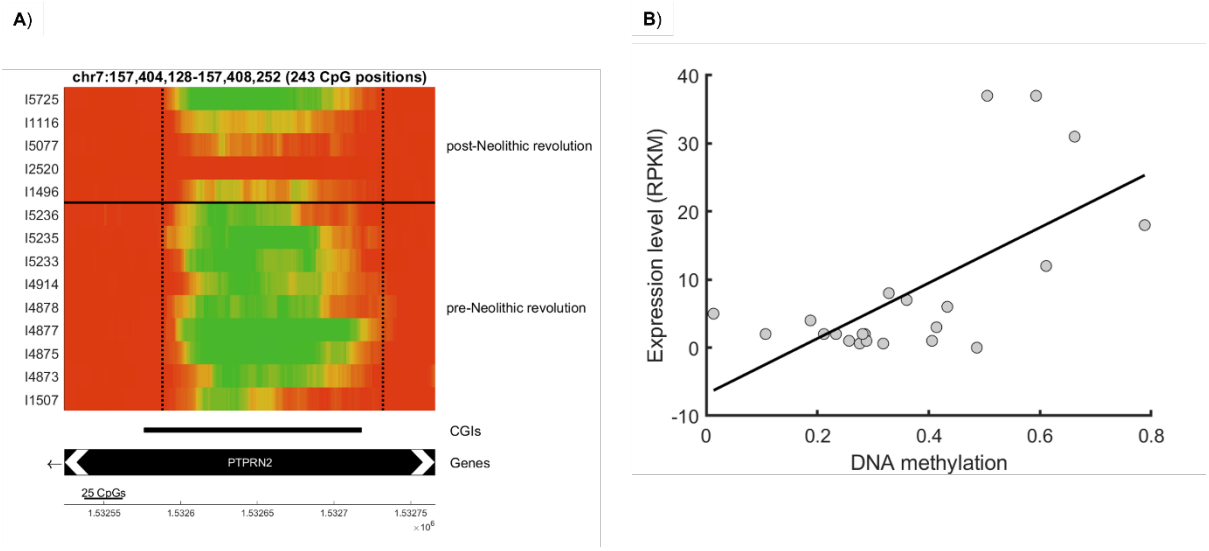

Figure 3. The DMR within the PTPRN2 gene. **A)** Reconstructed DNA methylation of the 14 Balkan samples. The DMR (dashed vertical lines) distinguishes between post-Neolithic revolution samples (upper lanes) and pre-Neolithic revolution ones (lower lanes). Methylation is color coded, from low methylation in green to high methylation in red. Lower lanes describe the genomic locations of CpG islands (CGIs) and genes. This DMR intersects a CpG island. **B)** Expression level of the PTPRN2 gene as a function of the mean methylation within the DMR, in 22 modern human tissues.

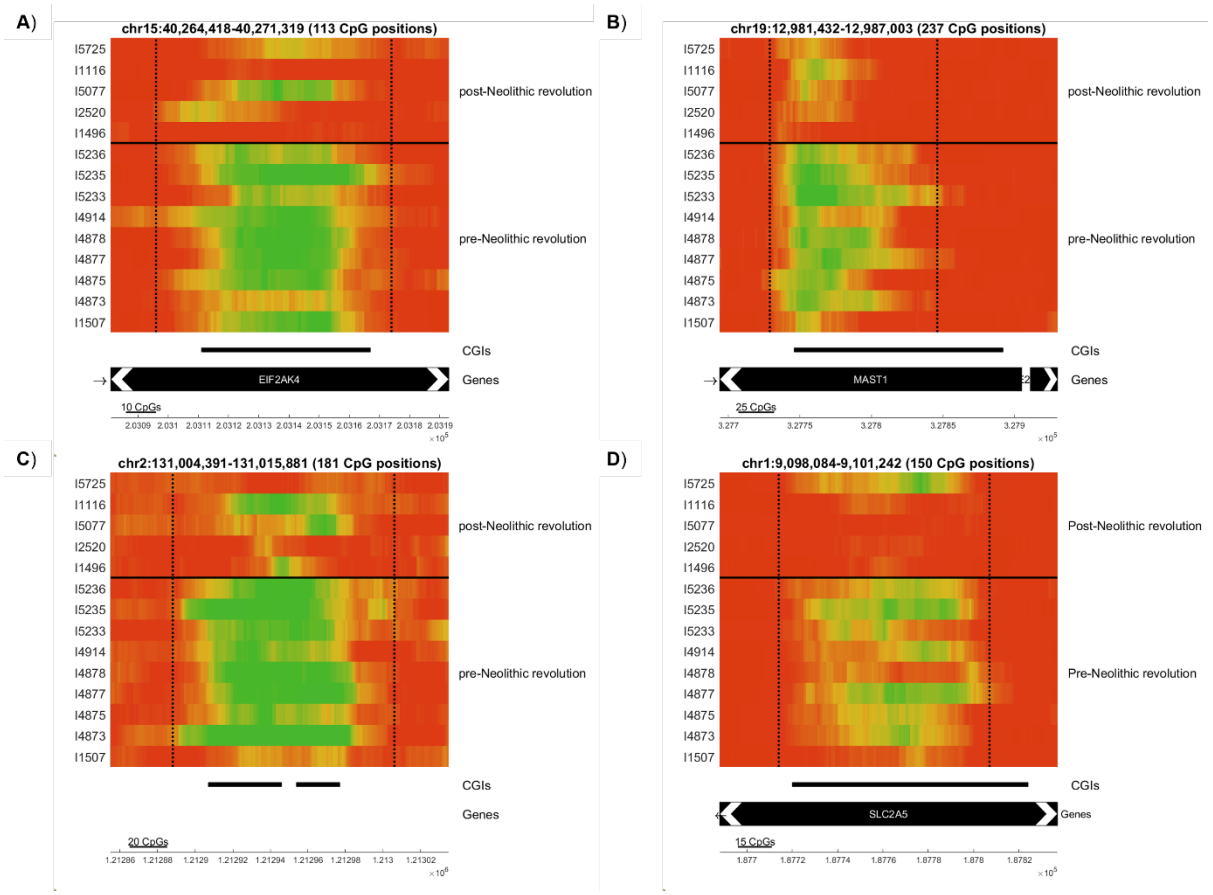

Figure 4. Additional DMRs (bounded by dashed vertical lines) that distinguish between post-Neolithic revolution samples (upper lanes) and pre-Neolithic revolution ones (lower lanes). Methylation is color coded, from low methylation in green to high methylation in red. Lower lanes describe the genomic locations of CpG islands (CGIs) and genes. **A-C)** The additional three DMRs detected in the analysis. **D)** The DMR with the highest Q_{max} that did not pass significance threshold. All DMRs intersect CpG islands.