

# Supplemental Material for: “Privacy Preserving Epigenetic PaceMaker Stronger Privacy and Improved Efficiency”

Meir Goldenberg, Loay Mualem, Amit Shahar, Sagi Snir, and Adi Akavia<sup>[0000–0003–0853–3576]</sup>

University of Haifa, Israel

meirgold@hotmail.com, loaymua@gmail.com, ashaha16@campus.haifa.ac.il,  
ssagi@research.haifa.ac.il, adi.akavia@gmail.com

## Table of Contents:

1. Supplemental Figure S1:  $X$ ,  $\beta$ , and  $y$
2. Supplemental Figure S2: EPM Algorithm on cleartext data
3. Supplemental Figure S3: Notations summary
4. Supplemental Figure S4: Algorithm for finding a lower bound of the plaintext modulus
5. Supplemental Figure S5: Our Secure EPM Protocol
6. Supplemental Figure S6: Simplified EPM Functionality
7. Supplemental Figure S7: Synthetic data generation and simulation
8. Supplemental Figure S8: RSS vs number of individuals in the dataset.

$$X = \left[ \begin{array}{cccc|cccc}
 t_1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\
 t_2 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\
 \vdots & & & & \vdots & & & \\
 t_m & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\
 0 & t_1 & \dots & 0 & 0 & 1 & \dots & 0 \\
 0 & t_2 & \dots & 0 & 0 & 1 & \dots & 0 \\
 \vdots & & & & \vdots & & & \\
 0 & t_m & \dots & 0 & 0 & 1 & \dots & 0 \\
 \vdots & & & & \vdots & & & \\
 \vdots & & & & \vdots & & & \\
 0 & \dots & 0 & t_1 & 0 & \dots & 0 & 1 \\
 0 & \dots & 0 & t_2 & 0 & \dots & 0 & 1 \\
 \vdots & & & \vdots & & & & \\
 0 & \dots & 0 & t_m & 0 & \dots & 0 & 1
 \end{array} \right], \beta = \begin{bmatrix} r_1 \\ \vdots \\ r_n \\ - \\ s_1^0 \\ \vdots \\ s_n^0 \end{bmatrix}, Y = \begin{bmatrix} \hat{s}_{1,1} \\ \hat{s}_{1,2} \\ \vdots \\ \hat{s}_{n,m} \end{bmatrix}$$

Supplemental Figure S1:  $X$ ,  $\beta$ , and  $y$

**Parameters:** Number individuals  $m$ , sites  $n$ , and iterations  $\text{iter}$ .

**Input:** A matrix  $\hat{S} = \hat{s}_{i,j}$  holding observed methylation levels on all sites  $i \in [n]$  and for each individual  $j \in [m]$ .

**Output:**  $t_1, \dots, t_m$

**Steps:**

1. **Initialization:** Initialize  $t_1, \dots, t_m$  to random age values and  $\text{isFail} = \text{FALSE}$ . Let  $X$  be the  $mn \times 2n$  matrix associated with  $t_j$ s as specified in Figure S1. Let  $y$  be a  $mn$ -dimension vector holding the entries of  $\hat{S}$  from top down, left to right (i.e.  $y_{im+j} \leftarrow \hat{s}_{i,j}$ )
2. For  $\text{iter}$  iterations do:
  - (a) **Site step:** Solve the linear regression problem on input  $(X, y)$ , denote the solution by the length  $2n$  vector  $\beta = (r_1, \dots, r_n, s_1^0, \dots, s_n^0)$  (cf. Figure S1)
  - (b) **Time step:** For each  $j \in [m]$ , calculate

$$t_j \leftarrow \frac{\sum_i r_i (\hat{s}_{i,j} - s_i^0)}{\sum_i r_i^2} \quad (1)$$

unless the denominator is zero in which case we set  $t_j$  to an arbitrary value and set  $\text{isFail} = \text{TRUE}$ .

3. If  $\text{isFail} = \text{FALSE}$  return  $(t_1, \dots, t_m)$  (else return  $\perp$ ).

Supplemental Figure S2: EPM Algorithm on cleartext data

- $m$  and  $n$  are the number of individuals and methylation sites respectively.
- $\hat{s}_{i,j}$  is the observed methylation at individual  $j$  and site  $i$ .
- $\vec{t} = (t_1, \dots, t_m)$  are the e-ages as updated throughout the computation, where each  $t_j$  is represented as a pair  $(t_{\text{num},j}, t_{\text{denom}})$  of its numerator and denominator (initialize to  $(t_{\text{num},j}, t_{\text{denom}}) = (t_j, 1)$ ).
- $\psi_k$  and  $\phi_k$  (for  $k \in [m]$ ) are defined by:

$$\psi(t_{\text{num},k}, t_{\text{denom}}) = (-m \cdot t_{\text{num},k} + \sum_{j \leq m} t_{\text{num},j}) \cdot t_{\text{denom}} \quad (2)$$

$$\phi(t_{\text{num},k}) = t_{\text{num},k} \sum_{j \leq m} t_{\text{num},j} - \sum_{j \leq m} t_{\text{num},j}^2 \quad (3)$$

- The  $n \times nm$ -dimensional matrices  $U, L$  and  $nm$ -dimensional vector  $Y$  are defined by:

$$U(\vec{t}_{\text{num}}, t_{\text{denom}}) = \begin{bmatrix} \psi_1 \cdots \psi_m & 0 & \cdots & 0 \\ 0 \cdots 0 & \psi_1 \cdots \psi_m & 0 & \cdots \cdots 0 \\ \vdots & & & \vdots \\ 0 & \cdots & 0 & \psi_1 \cdots \psi_m \end{bmatrix} \quad (4)$$

$$L(\vec{t}_{\text{num}}) = \begin{bmatrix} \phi(t_1) \cdots \phi(t_m) & 0 & \cdots & 0 \\ 0 \cdots 0 & \phi(t_1) \cdots \phi(t_m) & 0 & \cdots \cdots 0 \\ \vdots & & & \vdots \\ 0 & \cdots & 0 & \phi(t_1) \cdots \phi(t_m) \end{bmatrix} \quad (5)$$

$$Y = (\hat{s}_{1,1}, \dots, \hat{s}_{1,m}, \hat{s}_{2,1}, \dots, \hat{s}_{2,m}, \dots, \hat{s}_{n,1}, \dots, \hat{s}_{n,m})^T \quad (6)$$

- **MatVecMult** is the function that given a matrix and a vector (of matching dimensions) returns their product.
- $\vec{r} = (r_1, \dots, r_n)$  and  $\vec{s}^0 = (s_1^0, \dots, s_n^0)$  are sites' rates and methylation at birth values, respectively, as computed throughout the EPM algorithm; and  $\vec{r}', \vec{s}'^0$  their scaling by the factor  $\Lambda^{-1}$  defined by:

$$\Lambda^{-1}(\vec{t}_{\text{num}}) = \left( \sum_{j \leq m} t_{\text{num},j} \right)^2 - m \sum_{j \leq m} t_{\text{num},j}^2 \quad (7)$$

- The e-age are updated (in each time step) according to the following formulae:

$$t_{\text{num}}(\Lambda^{-1}, \vec{r}', \vec{s}'^0, (\hat{s}_{i,j})_{i \in [n]}) = \sum_{i \leq n} r'_i (\Lambda^{-1} \hat{s}_{i,j} - s'^0_i) \quad (8)$$

$$t_{\text{denom}}(\vec{r}') = \sum_{i \leq n} r_i'^2 \quad (9)$$

Supplemental Figure S3: Notations summary: inputs, variables, and functions to be homomorphically evaluated in our protocol.

**Input:** Positive integers  $m, n, \text{iter}, \ell$  (the parameters of the EPM functionality).

**Output:** A positive integer  $N_0$  (a lower bound on the plaintext modulus  $N$  for the privacy preserving EPM protocol)

**Steps:**

1. Initialize:  $\alpha \leftarrow 1, \sigma \leftarrow 10^\ell, \tau \leftarrow 100 \cdot 10^\ell$
2. For  $\text{iter}$  iterations do:
  - (a)  $N_0 \leftarrow n \cdot 6m^4 \cdot \tau^3 \cdot \sigma^2 \cdot \alpha$
  - (b)  $\tau \leftarrow N_0$
  - (c)  $\alpha \leftarrow n \cdot 4m^4 \cdot \tau^2 \cdot \sigma^2 \cdot \alpha^2$
3.  $N_0 \leftarrow N_0 \cdot 2$
4. Return  $N_0$

Supplemental Figure S4: Algorithm for finding a lower bound of the plaintext modulus. For simplicity we assume that the upper age bound is 100 and methylation values are in range of (0, 1).

**Parties:** Data Owners  $DO_1, \dots, DO_m$ , Machine Learning Server MLE and Crypto Service Provider CSP.

**Common Parameters:** the number of data owners/individuals  $m$  and methylation sites  $n$ , the precision  $\ell$ , the number of iterations  $iter$ , a fully homomorphic encryption scheme  $\mathcal{E} = (\text{KeyGen}, \text{Enc}, \text{Dec}, \text{Eval})$ , and a security parameter  $\lambda$ . Let  $N$  be an integer s.t.  $N \geq N_0$ , where computations are in the ring  $\mathbb{Z}_N$  of integers modulo the  $N$  (unless explicitly stated otherwise).

**Input:** Each  $DO_j$  holds methylation values  $(\hat{s}_{i,j})_{i \in [n]} \in (0, 1)^n$  (MLE and CSP have no input).

**Output:** Predicted epigenetic ages  $t_1, \dots, t_m$

**I. Pre-processing phase (key generation and data upload):**

1. **Setup:** The CSP executes  $(pk, sk) \leftarrow \text{KeyGen}(1^\lambda, N)$  and publishes the public key  $pk$ .
2. **Data upload:** Each  $DO_j$  rounds her input to  $\ell$  decimal digits, encrypts the rounded values with the public key  $pk$ , and sends the ciphertexts to the MLE.

**II. Secure computing Phase:** MLE, upon receiving ciphertexts  $\hat{s}_{i,j}$  for the methylation values  $\hat{s}_{i,j}$  ( $i \in [n], j \in [m]$ ), initializes the e-ages  $t_j$  to uniformly random in  $[0, \tau]$  where  $\tau$  (for  $\tau$  as set in Figure S4) represented as a numerator and denominator pair  $(t_{\text{num},j}, t_{\text{denom}}) = (t_j, 1)$ , and does the following:

1. For  $iter$  iterations do:

- (a) **Site step:** Homomorphically solve the linear system of the EPM site step (Figure S2, Step 2a) as follows:
  - i. Implicitly construct  $\mathbf{U}, \mathbf{L}$  and  $\mathbf{Y}$  by homomorphically computing:

$$\begin{aligned}\psi_k &\leftarrow \text{Eval}(pk, \psi; t_{\text{num},k}, t_{\text{denom}}) \\ \phi_k &\leftarrow \text{Eval}(pk, \phi; t_{\text{num},k})\end{aligned}$$

and using  $(\hat{s}_{i,j})_{i \in [n], j \in [m]}$ , respectively. We remark that  $\overrightarrow{t_{\text{num}}}$  is in cleartext form at the 1st iteration, and in encrypted form in subsequent ones.

- ii. Homomorphically compute ciphertexts encrypting the numerator of the linear system's solution:

$$\begin{aligned}\overrightarrow{\mathbf{r}'} &\leftarrow \text{Eval}(pk, \text{MatVecMult}; \mathbf{U}, \mathbf{Y}) \\ \overrightarrow{\mathbf{s}^{\prime 0}} &\leftarrow \text{Eval}(pk, \text{MatVecMult}; \mathbf{L}, \mathbf{Y})\end{aligned}$$

and the corresponding denominator:  $\Lambda^{-1} \leftarrow \text{Eval}(pk, \Lambda^{-1}; \overrightarrow{t_{\text{num}}})$ .

- (b) **Time step:** Homomorphically update the e-ages  $t_j$  (cf. Figure S2, Step 2b) by computing their updated numerators and denominator:

$$\begin{aligned}t_{\text{num},j} &\leftarrow \text{Eval}(pk, t_{\text{num}}; \Lambda^{-1}, \overrightarrow{\mathbf{r}'}, \overrightarrow{\mathbf{s}^{\prime 0}}, (\hat{s}_{i,j})_{i \in [n]}) \quad \forall j \in [m] \\ t_{\text{denom}} &\leftarrow \text{Eval}(pk, t_{\text{denom}}; \overrightarrow{\mathbf{r}'})\end{aligned}$$

2. Publish the ciphertexts for the updated e-ages:  $t_{\text{num},1}, \dots, t_{\text{num},m}$  and  $t_{\text{denom}}$ .

**III. Post processing phase:** The CSP decrypts to obtain  $t_{\text{num},j} \leftarrow \text{Dec}(sk, t_{\text{num},j})$ , for all  $j \in [m]$ , and  $t_{\text{denom}} \leftarrow \text{Dec}(sk, t_{\text{denom}})$ , and publish the cleartext e-ages' numerators and denominator  $(\overrightarrow{t_{\text{num}}}, t_{\text{denom}})$ . These values represent the e-ages:  $t_1 \leftarrow \frac{t_{\text{num},1}}{t_{\text{denom}}}, \dots, t_m \leftarrow \frac{t_{\text{num},m}}{t_{\text{denom}}}$  where division is over the reals (output  $\perp$ , if  $t_{\text{denom}} = 0$ ).

Supplemental Figure S5: Our Secure EPM Protocol. See Supplemental Figure S3 for the definition of the functions  $\Lambda^{-1}, \psi, \phi, \text{MatVecMult}, t_{\text{num}}, t_{\text{denom}}$ , matrices  $\mathbf{U}, \mathbf{L}$ , vector  $\mathbf{Y}$ , and the bounds  $\tau$  and  $N_0$ . Ciphertexts are denoted in **Boldface**, i.e.,  $\overrightarrow{t_{\text{num}}}, \Lambda^{-1}, t_{\text{denom}}, \mathbf{U}, \mathbf{L}, \mathbf{Y}, \overrightarrow{\mathbf{r}'}$  and  $\overrightarrow{\mathbf{s}^{\prime 0}}$  decrypt to  $\overrightarrow{t_{\text{num}}}, \Lambda^{-1}, t_{\text{denom}}, \mathbf{U}, \mathbf{L}, \mathbf{Y}, \mathbf{r}'$  and  $\mathbf{s}^{\prime 0}$  respectively. We note that  $\Lambda^{-1}$  denotes both a function and its output  $\Lambda^{-1}(\overrightarrow{t_{\text{num}}})$ .

**Parties:** Data-Owners  $DO_1, \dots, DO_m$ , a Machine Learning Engine (MLE), and a Crypto Service Provider (CSP).

**Common Parameters:** The number of individuals  $m$ ; the sites  $s_1, \dots, s_n$ ; the precision  $\ell$ , where all values in  $\mathbb{R}$  are scaled to  $[-\delta, \delta]$  with a precision of  $\ell$  digit; number of iterations  $iter$ . Denote by  $N$  the smallest sufficiently larger integer as to guarantee that no overflow occurs during the computation, i.e., we require that  $N > N_0$  for  $N_0$  as specified in Figure S4

**Input:** Each data owner  $j$  holds observed methylation levels  $\hat{s}_{1,j}, \dots, \hat{s}_{n,j}$ .

**Output:** All parties receive as output the epigenetic age estimation  $t_1, \dots, t_m$  that are the output of the EPM algorithm (Figure S2) when executed on input  $(\hat{s}_{i,j})_{i \in [n], j \in [m]}$  for  $iter$  iterations.

**Leakage Profile:** The common parameters and the outputted epigenetic ages  $t_j$  provided in the following representation. Each  $t_j$  is represented as a pair of integers whose ratio equals  $t_j$ , concretely, by the numerator and denominator computed in the Time Step at the final iteration of the EPM algorithm of Figure S2.

### Supplemental Figure S6: EPM Functionality, simplified to reveal the output

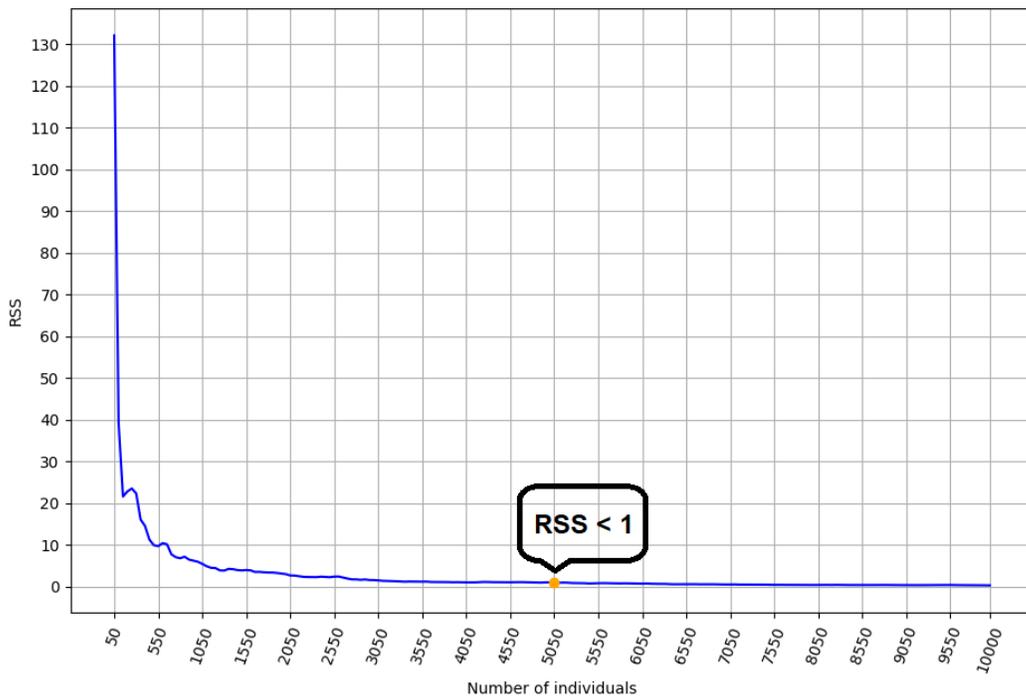
#### I. Data preparation for $N$ individuals:

- Rates ( $r$ ) and methylation at birth ( $s^0$ ):** We define  $r_{min}, r_{max}, s_{min}^0, s_{max}^0$  to be the minimum and maximum rate and  $s^0$  values derived by our cleartext algorithm as part of the empirical evaluation results on real data. For each methylation site  $i$  uniformly sample  $r_i \in [r_{min}, r_{max}]$  and  $s_i^0 \in [s_{min}^0, s_{max}^0]$
- Ages:** We sample  $T$  as a uniformly generated list of chronological age values where for each  $t_j \in T$  it holds that  $0 \leq t_j \leq 99$
- Noise value:** We sample  $\mathcal{E}_{i,j} \sim N(0, \sigma^2)$  as a normally distributed noise value where  $\sigma = 0.15$  as used in [Snir et al., 2016]
- Methylation values:** We compute  $\hat{s}_{ij}$  as the methylation value for individual  $j$  at site  $i$  derived using the formula  $\hat{s}_{ij} = s_i^0 + r_i t_i + \mathcal{E}_{i,j}$  as detailed in [Snir, 2020] using the age, rate, methylation at birth and noise values described above.
- Ground truth epigenetic age values:** We compute our ground truth epigenetic ages  $T_{gt}$  by running the EPM site step using the above methylation and age values as inputs.
- Dataset construction:** The chronological ages ( $T$ ), methylation values ( $\hat{s}_{ij}$ ) and ground truth epigenetic ages ( $T_{gt}$ ) from above will serve as input to the next stage.

**II. Simulation:** Defining  $N$  as the total number of individuals in our synthetic dataset, run the following steps for the simulation:

- Set  $n \leftarrow 50$
- Repeat while  $n < N$ 
  - Derive a subset of  $n$  individuals and their methylation values from the full dataset, starting from the first individual.
  - Run the cleartext EPM algorithm as defined in S2 using the derived dataset for 3 iterations.
  - Let  $T_{predicted}$  be the predicted ages outputted by the EPM algorithm. Calculate  $RSS = \sum_{j \leq 50} (t_{predicted_j} - t_{gt_j})^2$  where  $t_{predicted_j} \in T_{predicted}$  and  $t_{gt_j} \in T_{gt}$
  - Add the calculated RSS to  $RSS\_List$
  - $n \leftarrow n + 50$
- return  $RSS\_List$

### Supplemental Figure S7: synthetic data generation and simulation



Supplemental Figure S8: RSS vs number of individuals in the dataset. RSS measurement is for the first 50 individuals.

## References

- Sagi Snir. Epigenetic pacemaker: closed form algebraic solutions. *BMC genomics*, 21(2):1–11, 2020.
- Sagi Snir, Bridgett M vonHoldt, and Matteo Pellegrini. A statistical framework to identify deviation from time linearity in epigenetic aging. *PLoS computational biology*, 12(11):e1005183, 2016.