Classification of individuals based on self-reported smoking status is prone to errors due to under-reporting and poor recall of long-term smoking history. Traditional biomarkers, such as cotinine, only measure short-term exposure. Cigarette smoking strongly influences DNA methylation, with differential methylation patterns detected among current, former and never smokers. We developed, **EpiSmoker** a robust prediction tool to infer the smoking status based on the whole-blood DNA methylation profiles.

**METHODOLOGY**
- DNA methylation data from DILGOM (N=517 and Age range: 25-74) a sub-cohort of FINRISK was used as the training dataset.
- After stringent QC, quantile normalization was performed. Quantiles were later used to fit the distributions of validation datasets.
- Multinomial LASSO regression was performed to select a parsimonious set of CpG sites predictive of smoking status (Figure 1).
- Tuning parameter, optimal lambda was chosen through a tenfold internal cross-validation from a sequence of 100 lambda values.
- 121 methylation sites and the corresponding training coefficients selected from this regression were used to predict the smoking statuses in the three test datasets.
- Three independent and external test datasets were used to assess the performance.

### RESULTS
**EpiSmoker** showed high accuracy for predicting current and never smokers in all the 3 test datasets.
- Figures 2 - 4 illustrate results from the Finnish Twin Cohort.
- Precise classification of self-reported former smokers from current and never smokers depends on the extent and duration of smoking prior to smoking cessation. Longer cessation time leads to reversal of methylation profiles in former smokers making them indistinguishable from never smokers (Figures 3 & 4).

### ADVANTAGES of **EpiSmoker**
- Considers three smoking statuses
- Implementation of quantile normalization on test datasets using quantiles from discovery dataset has allowed transferring prediction model from one study to another. Thereby allowing and improving cross-study performance.
- Serves as a robust predictor of biological smoking status based on methylation data.
- Globally applicable to all populations.
- Available on GitHub: https://github.com/sailalithabollepalli/EpiSmoker
- Output provided as Smoking status labels, in HTML and CSV file formats.