Introduction to Bioinformatics

Paper Discussion - Saliva cell type DNA methylation reference panel for epidemiological studies in children

Opening

- Review a paper send by Pak Htet
- What is the paper about
- Understanding the new terms that are in the paper
- Understanding the bioinformatics pipeline that is in the paper

What is the paper about

- Creation of saliva cell-type reference panel
  - Primary Saliva DNA Methylation reference panel is not available for children
  - Helps other researchers to take into account cell-type proportions when performing research on association of disease or exposure
  - Result is implemented in R library called ewastools
- Proving that saliva immune and epithelial cells have distinct DNA methylation profiles which can drive whole-saliva DNA methylation measures

Important keywords that you may have not heard yet

- Reference panels
- Cofounding variables
- Heterogeneity
- ENCODE
- ewastools
- PCA
- Cell proportion estimation

How was the research in the paper conducted?

Collection of Saliva samples from Children

- 18 large cell fractions
- 20 CD45+ cell fractions
- 18 whole saliva
- 4 Oragene kit

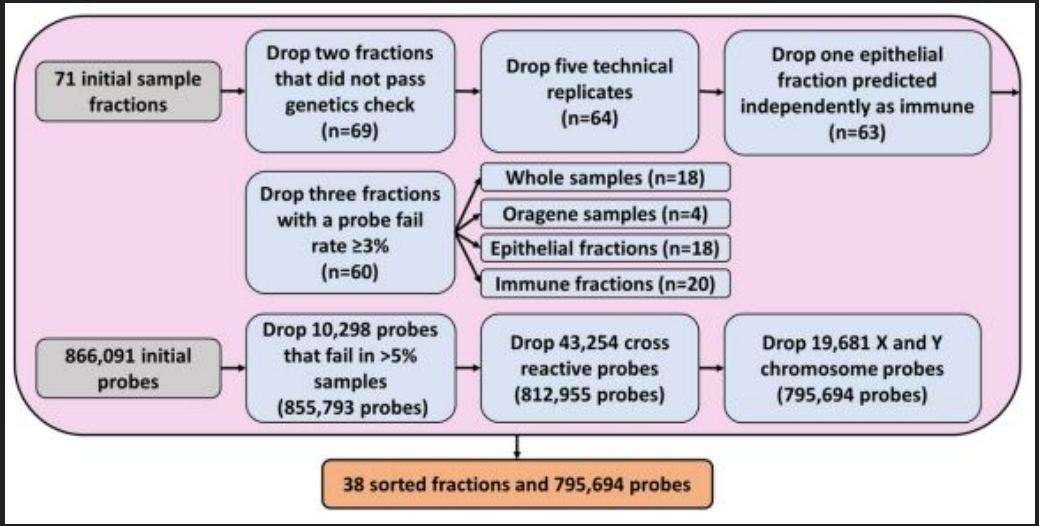
DNA methylation analysis

- Normal comparison between epithelial and immune cells

Cell Proportion Estimation

Study Samples and Saliva Collection

- Children between the ages of 7 and 17 years
- Sample of 22 children from schools in Ann Arbor, Michigan.
- Prior to saliva collection, participants did not eat or drink for 30 minutes



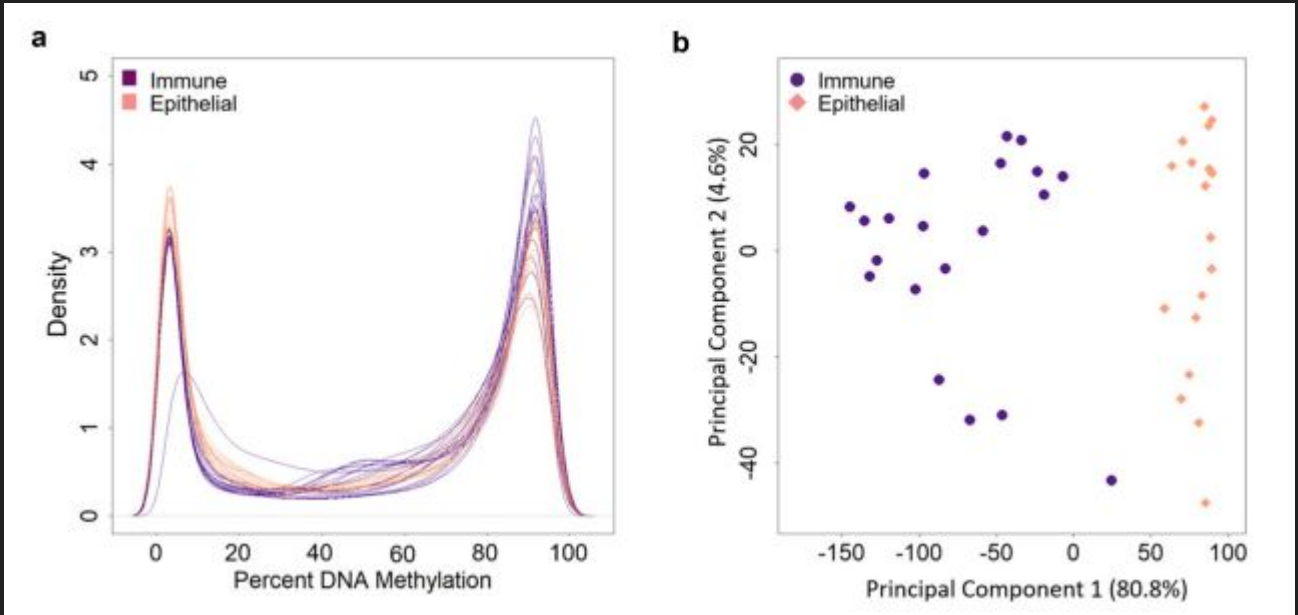
DNA methylation data preprocessing

- EPIC BeadArray IDAT image files were processed and control metrics were assessed using the **ewastools** package
- Background correction was performed using noob in the **minfi** package
- Detection p-values were calculated to identify failed probes. Samples with >3% probes exceeding the detection p-value = 0.01 were dropped.
- Cell-type proportions were estimated for each sample using a reference panel generated from ENCODE and adult white blood cell data, implemented in ewastools
- DNA methylation was measured at 866,091 sites filtered to 795,694

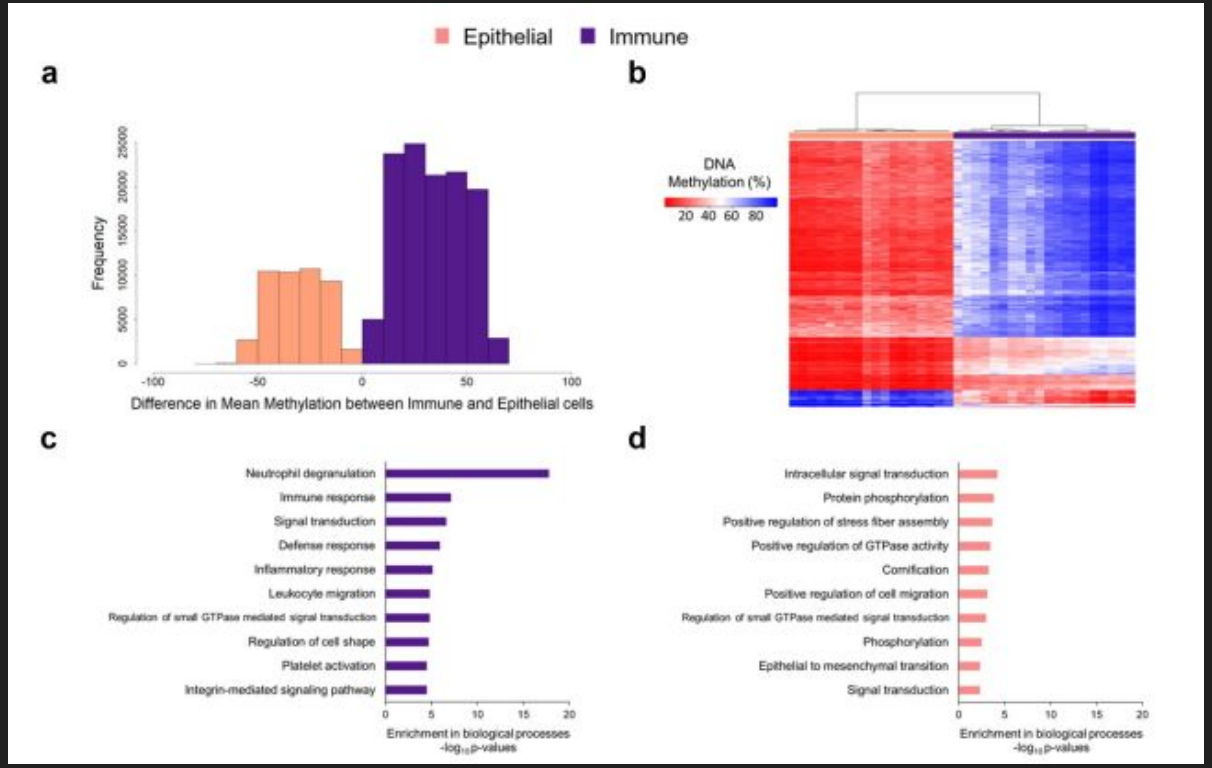
Statistical and Bioinformatics Analysis

- Sample descriptive statistics on demographic and laboratory measures. For continuous variables (age, cell count, cell viability, sample volume) and for categorical variables (sex, race, illness status).
- Density plots were constructed, to visualize DNA methylation distributions by cell type.
- Principal component analysis was conducted to summarize variations in the DNA methylation data
- To test for differences in DNA methylation between all 18 epithelial and 20 immune cell samples, unpaired t-tests were used at each DNA methylation site.
- GO were constructed for hypermethylated and hypomethylated probes

Resulted Analysis



Resulted Analysis



So how about the Cell Type Estimation?

- Cell type estimation is being done using the Houseman algorithm
- Houseman algorithm (created in 2012) is the algorithm that is being used to estimate the cell proportion from the DMR data that was generated

Whole Cell Sample group

Sample groups where the cell subgroups are distinguished