Data pre-processing before running CEllFi

Fei-Man Hsu

January 12, 2023

CEllFi is a python package that models DNA methylation calls from references with non-negative least squares and estimate the fraction of cell types in a DNA mixture.

$$argmin_x ||Ax - b||_2^2$$

with the constraint $x \geq 0$, $\sum x = 1$, A denotes an $m \times n$ matrix.

Dependencies

- python2, [Recommend] Building the environment with virtualenv
  - pandas
  - numpy
  - scipy
- CELLFi package
  - Click for reference *.bed* files and scripts, or contact Dr. Dennis Montoya for Github access

Differentially methylated regions (DMRs)

Here we demonstrate the CEllFi deconvolution with blood cells' whole genome bisulfite sequencing (WGBS) dataset **GSE186458**. Thirty six blood cell WGBS libraries are included for references:

- GSM5652277_Blood-T-CD3-Z000000TV
- GSM5652278_Blood-T-CD3-Z000000UP
- GSM5652279_Blood-T-CD4-Z000000TT
- GSM5652280_Blood-T-CD4-Z000000U7
- GSM5652281_Blood-T-CD4-Z000000UM

- GSM5652282_Blood-T-CD8-Z000000TR
- GSM5652283_Blood-T-CD8-Z000000U5
- GSM5652284_Blood-T-CD8-Z000000UK
- GSM5652285_Blood-T-CenMem-CD4-Z00000417
- GSM5652286_Blood-T-CenMem-CD4-Z0000041D
- GSM5652287_Blood-T-CenMem-CD4-Z0000041N
- GSM5652288_Blood-T-Eff-CD8-Z00000419
- GSM5652289_Blood-T-Eff-CD8-Z0000041F
- GSM5652290_Blood-T-Eff-CD8-Z0000041Q
- GSM5652291_Blood-T-EffMem-CD4-Z00000416
- GSM5652292_Blood-T-EffMem-CD4-Z0000041C
- GSM5652293_Blood-T-EffMem-CD4-Z0000041M
- GSM5652294_Blood-T-EffMem-CD8-Z0000041A
- GSM5652295_Blood-T-EffMem-CD8-Z0000041G
- GSM5652296_Blood-T-Naive-CD4-Z0000041E
- GSM5652297_Blood-T-Naive-CD8-Z0000041B
- GSM5652298_Blood-T-Naive-CD8-Z0000041H
- GSM5652299_Blood-NK-Z000000TM
- GSM5652300_Blood-NK-Z000000U1
- GSM5652301_Blood-NK-Z000000UF
- GSM5652302_Blood-Monocytes-Z000000TP
- GSM5652303_Blood-Monocytes-Z000000U3
- GSM5652304_Blood-Monocytes-Z000000UH
- GSM5652313_Blood-Granulocytes-Z000000TZ
- GSM5652314_Blood-Granulocytes-Z000000UD
- GSM5652315_Blood-Granulocytes-Z000000UT
- GSM5652316_Blood-B-Z000000TX

- GSM5652317_Blood-B-Z000000UB
- GSM5652318_Blood-B-Z000000UR
- GSM5652319_Blood-B-Mem-Z0000041J
- GSM5652320_Blood-B-Mem-Z0000041K

We aim to build references for B cell, monocyte, granulocyte, nature killer cell (NK cell), and T cell. To achieve this, we first identify cell-type specific hypo-DMRs with the *metilene* software:

```
metilene -M 500 -m 10 -d 0.3 -t 2 -v 0.7 --mode 1 --mtc 2
--groupA 'Bcell_' --groupB 'nonBcell_' Bcell_metilene_input.tsv
> Bcell_DMR.tsv
```

The above example demonstrates that we demand DMRs with at least 500bp in size with minimum 10 `CpG` sites that show differential methylation level >= 30% between B cell and the other cell types. Further filtering could be applied to acquire hypo-DMRs based on the delta methylation level in column 4 of the output Bcell_DMR.tsv file. We further characterize naïve T cell's hypo-DMRs from other T cell.

To verify the DMRs, we performed supervised hierarchical clustering:

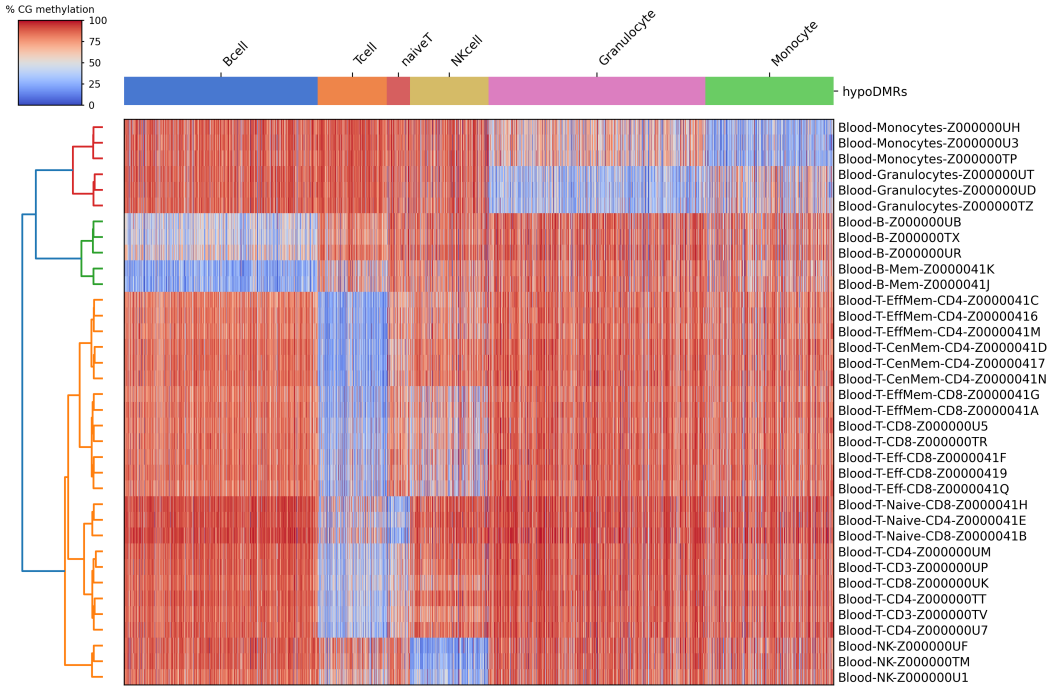

Figure 1: All cell-specific DMRs (n= 110,276)

CpG in these DMRs could be used as references for CEllFi deconvolution. However, it could take hours for CEllFi to construct reference matrix, and, when applied to targeted-bisulfite sequencing data (TBS), the joint matrix of reference and sample would have *nan* in majority, which is not cost-effective. Subset is therefore recommended. The following cluster heatmap shows that the only 60 DMRs, subset from the COVID19 TBS dataset, still are able to distinguish the 6 cell types:

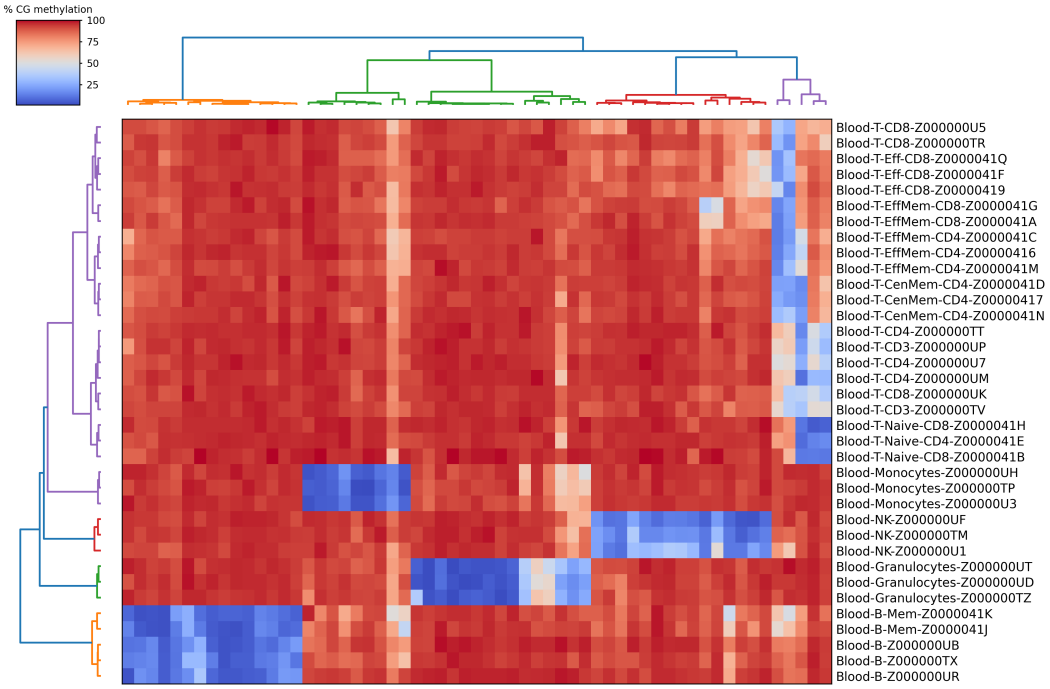

Figure 2: TBS subset cell-specific DMRs (n= 60)

After subsetting the CpG, all reference *.bed* files are moved to ./ref/ directory.

Validation

Three *.conf* files need to be ready before running CEllFi:

- cellfi_reference.conf, binary classification pointing the reference *.bed* to the cell type
- cellfi_sample.conf, sample *.bed* to be deconvoluted
- cellfi_group.conf, cell types to be included in the analysis

Make sure the reference and sample *.bed* files are in the correct directory, and run:

```
FILE    Bcell   Tcell   NKcell  Monocyte        Granulocyte     NaiveT
./ref/GSM5652277_Blood-T-CD3-Z000000TV_all_celltype-specific_hypoDMR_mCG.bed    0       1       0       0       0       0
./ref/GSM5652278_Blood-T-CD3-Z000000UP_all_celltype-specific_hypoDMR_mCG.bed    0       1       0       0       0       0
./ref/GSM5652279_Blood-T-CD4-Z000000TT_all_celltype-specific_hypoDMR_mCG.bed    0       1       0       0       0       0
./ref/GSM5652280_Blood-T-CD4-Z000000U7_all_celltype-specific_hypoDMR_mCG.bed    0       1       0       0       0       0
./ref/GSM5652281_Blood-T-CD4-Z000000UM_all_celltype-specific_hypoDMR_mCG.bed    0       1       0       0       0       0
./ref/GSM5652282_Blood-T-CD8-Z000000TR_all_celltype-specific_hypoDMR_mCG.bed    0       1       0       0       0       0
```

Figure 3: Example of cellfi_reference.conf

```
FILE
./samples/natural_killer_cell_C002CTA1bs_hg38.bed
```

Figure 4: Example of cellfi_sample.conf

```
FILE
./samples/natural_killer_cell_C002CTA1bs_hg38.bed
```

Figure 5: Example of cellfi_group.conf

```
python ./scripts/bs_decon_pipe.py -pipe_o ./ -estimate_refs \\
Bcell,Granulocyte,Monocyte,NKcell,NaiveT,Tcell \\
-detect_delta 0.30 -select_num_hypo 15 -select_ttest_p 0.1 \\
-select_anova_p 0.1 -select_sam_cpg 5 -select_cell_delta 0.3
```

```
Samples Bcell   Granulocyte     Monocyte        NKcell  NaiveT  Residual        Tcell
CD4_CD45_RA_naive_5277_covered  0.0     0.003725486266468322    0.0     0.0521421761117374      0.9398169890433812      0.37926152288397524     0.0
```

Figure 6: Example of cellfi_coeff_meth.txt

Noted that parameters could be adjusted if needed. Several intermediate *.txt* files will be generated, and the *cellfi_coeff_meth.txt* is the final output that shows the fraction of denoted cell types in the query sample.

In order to validate the CEllFi deconvolution, we took the following WGBS data (in the ./samples/ directory):

- CD4_CD45_RA_naïve_5277, pellegrini lab
- CD4_CD45_RO_memory_5277, pellegrini lab
- classical_monocyte_C000S5A1bs, blueprint
- mature_neutrophil_C000S5A2bs, blueprint
- natural_killer_cell_C002CTA1bs, blueprint
- *in silico* synthetic mixture of 30% B cell and 70% Monocyte
- *in silico* synthetic mixture of 70% B cell and 30% Monocyte

Figure 7 bellow shows that all WGBS are with $> 90\%$ composition of the indicated cell types.

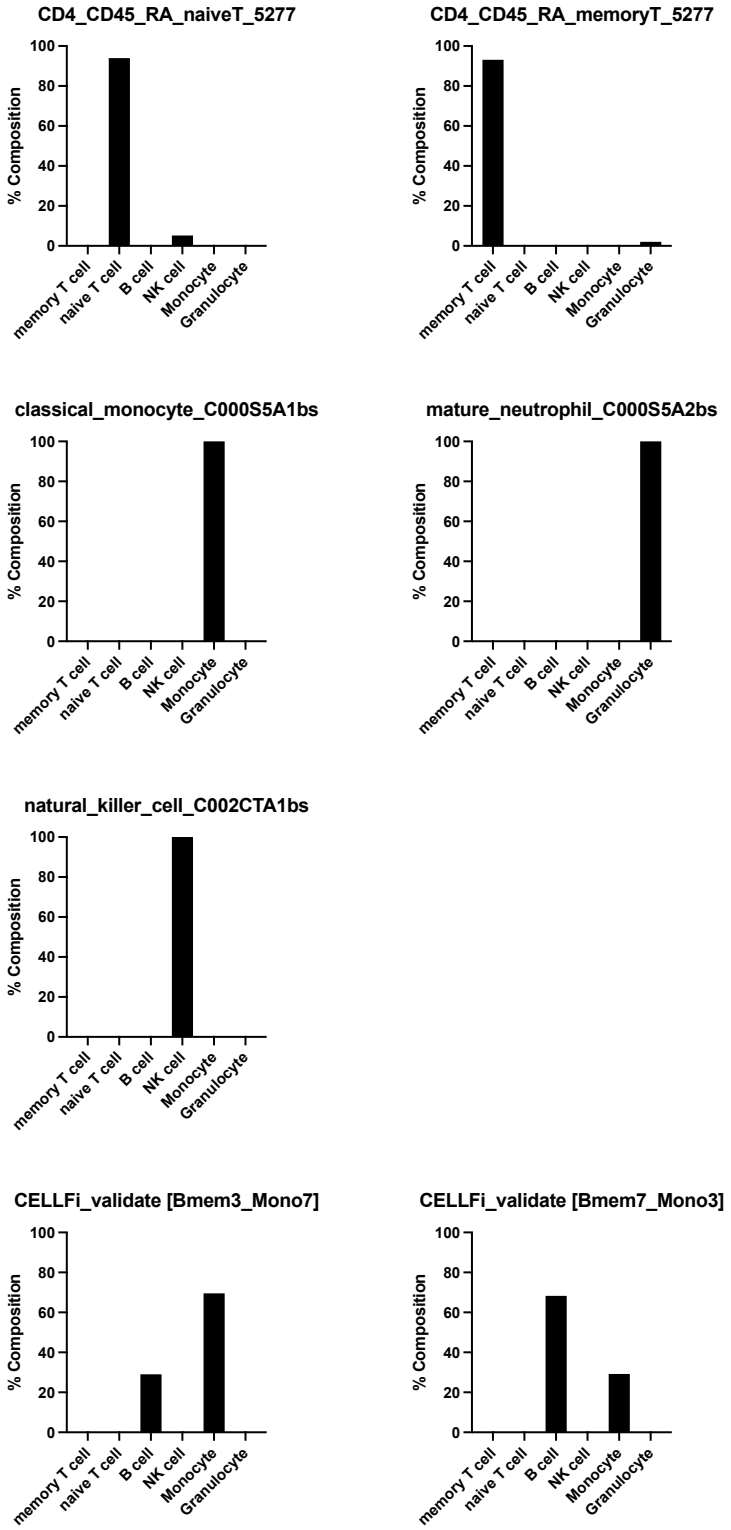

Figure 7: CEllFi validation