How do I get set up?

- Download all files to your local directory. Run the ***InfiniumPurify.py*** and ***InfiniumDMC.py*** scripts for purity estimation and DM calling.
- Dependencies: numpy, scipy, rpy2

InfiniumPurify: tumor purity estimation

In cancer genomics research, one important problem is that the solid tissue sample obtained from clinical settings is always a mixture of cancer and normal cells. The sample mixture brings complication in data analysis and results in biased findings if not correctly accounted for. We develop a simple but effective method to estimate purities from the DNA methylation 450k array data.

Usage: python InfiniumPurify.py <-f filename> <-t cancer_type> [...]

Get tumor purity from 450K array data. -f and -c options are needed!

Options: --version Show program's version number and exit

-h, --help Show this help message and exit.

-f FILENAME, --filename=FILENAME The file name of 450K array, with directory

-c CANCERTYPE, --CancerType=CANCERTYPE Cancer type, in TCGA abbreviation format, can be chosen from 'BRAC','LUAD','COAD',... (See abbr.txt for detail)

Input tumor methlylation file has four columns, with each corresponding to probe name, beta value, related gene, chromosome and genomic coordinate, respectively. Where beta value can be obtained from minfi or missMethyl packages. See an example in '450k_example.txt'.

Output is estimated purity.

Example

```
python InfiniumPurify.py -f 450k_example.txt -c LUAD
```

Now works for all 32 TCGA cancer types:

- ACC: Adrenocortical carcinoma
- BLCA: Bladder Urothelial Carcinoma
- BRCA: Breast invasive carcinoma
- CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma
- CHOL: Cholangiocarcinoma
- COAD: Colon adenocarcinoma
- DLBC: Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
- ESCA: Esophageal carcinoma
- GBM: Glioblastoma multiforme
- HNSC: Head and Neck squamous cell carcinoma
- KICH: Kidney Chromophobe
- KIRC: Kidney renal clear cell carcinoma

- KIRP: Kidney renal papillary cell carcinoma
- LAML: Acute Myeloid Leukemia
- LGG: Brain Lower Grade Glioma
- LIHC: Liver hepatocellular carcinoma
- LUAD: Lung adenocarcinoma
- LUSC: Lung squamous cell carcinoma
- MESO: Mesothelioma
- OV: Ovarian serous cystadenocarcinoma
- PAAD: Pancreatic adenocarcinoma
- PCPG: Pheochromocytoma and Paraganglioma
- PRAD: Prostate adenocarcinoma
- SARC: Sarcoma
- SKCM: Skin Cutaneous Melanoma
- STAD: Stomach adenocarcinoma
- TGCT: Testicular Germ Cell Tumors
- THCA: Thyroid carcinoma
- THYM: Thymoma
- UCEC: Uterine Corpus Endometrial Carcinoma
- UCS: Uterine Carcinosarcoma
- UVM: Uveal Melanoma

InfiniumDMC: differential methylation analysis

A generalized linear model to incorporate the purity information for differential methylation (DM) analysis in cancer versus normal comparison. When normal controls are not available, we propose a “control-free” DM calling method by testing the correlation between methylation levels and tumor sample purities of a CpG site. The analyses of TCGA data demonstrate that the proposed differential methylation methods provide more sensitive and biologically plausible results compared with existing method.

Usage: python InfiniumDMC.py <-t tumorFile> <-p purityFile> [-n normalFile] [...]

Differential Methylation analysis accounting for tumor purities. If normal samples are provided, a generalized least square model is executed for DM analysis. If there is no normal control, a control-free DM model is used for testing the regression coeffcient. -t and -p options are needed!

Options: --version show program's version number and exit

-h, --help Show this help message and exit.

-t TUMORFILE, --tumorFile=TUMORFILE tumor file, rows are CpG sites, columns are tumor samples.

-p PURITYFILE, --purityFile=PURITYFILE tumor purities estimated by InfiniumPurify, should have the same number with tumor files.

-n NORMALFILE, --normalFile=NORMALFILE normal file, rows are CpG sites, columns are normal samples.

Example 1: DM calling with normal controls

```
python InfiniumDMC.py -t tumor.txt -n normal.txt -p purity.txt -n normal.txt
```

tumor.txt and normal.txt are Infinium 450K array files of tumor and normal samples, purity.txt is tumor purity file estimated from InfiniumPurify, or user specific purities.

- Output file

  CpgSite Statistic P-value

  cg13332474 6.77357464971 3.56754439712e-11

  cg00651829 6.12894376895 1.80533429812e-09

  cg17027195 1.08498219974 0.27845660427

  cg09868354 2.31455408983 0.0210448694546

  cg03050183 -4.20565716333 3.08937733964e-05

  cg06819656 -2.91160511207 0.00375866922763

  cg04244851 -14.1940307326 1.27809047711e-38

  cg19669385 21.8567784859 1.13670562756e-74

  ...

Example 2: Control-free DM calling

```
python InfiniumDMC.py -t tumor.txt -n normal.txt -p purity.txt
```

tumor.txt is Infinium 450K array file of tumor samples, purity.txt is tumor purity file estimated from InfiniumPurify, or user specific purities.

- Output file

  CpgSite P-value

  cg13332474 0.45703201837

  cg00651829 0.113555062473

  cg17027195 0.0158541421073

  cg09868354 0.0

  cg03050183 9.94559773726e-06

  cg01989731 NA

  cg06819656 0.184717611127

  cg04244851 0.999999996308

  cg19669385 0.999121475655

  cg04244855 0.987516661682

  ...

Who do I talk to?

- Naiqian Zhang (naiqian@wfu.edu.cn) & Xiaoqi Zheng (xqzheng@shnu.edu.cn)
- Shanghai Normal University & Amory University

Reference

- Naiqian Zhang, Hua-Jun Wu, Weiwei Zhang, Jun Wang, Hao Wu, Xiaoqi Zheng, Predicting tumor purity from methylation microarray data. Bioinformatics 2015, 31:3401-3405.
- Xiaoqi Zheng, Naiqian Zhang, Hua-Jun Wu, Hao Wu, Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. Genome Biology, in revision.