cell-free DNA methylation biomarker discovery in ALS

Mybaits probe capture design

Summer 2018

Project Goal

- Given three tissues of interest- brain, muscle, leukocytes- and CpG sites found using sparse PCA1, design probes to capture cell free DNA that can illustrate tissue of origin between ALS and control cell-free DNA.
- List of CpG of interest contain ~100-1,000 CpGs determined to be differentially methylated given reference tissue profiling from ENCODE2, Roadmap3, and other annotation projects.

Code Design

- Code designed to find all reads covering a CpG of interest in WGBS data (4 ALS and 4 CTRL merged)
- A **SIZE** parameter was used to search for the amount of base pairs around a CpG of interest to design probes for.
- Each CpG of interest was required to have at least 2 other CpGs to form coupled blocks of CpGs4

```
README.md
data
   muscle_cpgs_125.hg38.txt
output
   muscle_probes_125.txt
   muscle_probes_125_patterns_all.txt
src
    __init__.py
    design
       __init__.py
       MethylRead.py
       cfDNA.py
    run.py
    run.sh
    utils
        __init__.py
```

```
io.py
```

Output

Probes file

- Has extension `_probes_SIZE.txt` where **SIZE** is the number of nucleotides around a CpG of interest assessed.
  - ex: `muscle_probes_125.txt`
- Contains 3 flavors of fasta entries per CpG of interest, 1) for reference sequence 2) for fully methylated sequence 3) read observed in the WGBS data
- Reference header:
  - parameters: search range around CpG | number of CpGs in region | tissue | reference
  - ex: `>chr1:1093248-1093499|number_of_cpgs=5|tissue=muscle|reference`
- Methylated header:
  - parameters: search range around CpG | number of CpGs in region | tissue | fully converted methylated
  - **Fully converted** means all C's in the sequence not part of a CpG dinucleotide are converted to a T
  - ex: `>chr1:1093248-1093499|number_of_cpgs=5|tissue=muscle|fully_converted_methylated`
- Read header:
  - parameters: Range read covers | number read in sequence
  - ex: `>chr1:1093282-1093374|read_number=1`
  - Alignment depicted to the reference strand. "." indicates a converted/non-methylated base (C-T mismatch for forward strand, G-A mistmatch for reverse strand)

Sample Entry:

```
>chr1:1093248-1093499|number_of_cpgs=5|tissue=muscle|reference
TGCTCCCTCTCTGGTTAAAGGGCATCCTGAGGGCCACATTAAGTCACAAAACATCATTTTGATTCAGGAACCAGAAGTCCAAGATTTCAAT
>chr1:1093248-1093499|number_of_cpgs=5|tissue=muscle|fully_converted_methylated
TGTTTTTTTTTTGGTTAAAGGGTATTTTGAGGGTTATATTAAGTTATAAAATATTATTTTGATTTAGGAATTAGAAGTTTAAGATTTTAAT
>chr1:1093282-1093374|read_number=1 |reverse_strand
TGCTCCCTCTCTGGTTAAAGGGCATCCTGAGGGCCACATTAAGTCACAAAACATCATTTTGATTCAGGAACCAGAAGTCCAAGATTTCAAT
                                  .|||||||.|||||||||||||||||.|||||..|||||.||.|||||.||||||||
----------------------------------GACATTAAATCACAAAACATCATTTTAATTCAAAAACCAAAAATCCAAAATTTCAAT
>chr1:1093282-1093374|read_number=2|forward_strand
TGCTCCCTCTCTGGTTAAAGGGCATCCTGAGGGCCACATTAAGTCACAAAACATCATTTTGATTCAGGAACCAGAAGTCCAAGATTTCAAT
                                  .|.|||||||.|.||||.||.|||||||||.|||||..||||||..|||||||.|||
----------------------------------TATATTAAGTTATAAAATATTATTTTGATTTAGGAATTAGAAGTTTAAGATTTTAAT
```

Patterns file

- Extension `_patterns_all.txt`
- Binary representation of the CpG patterning in a given region around a CpG of interest for all the reads covering that CpG
- This file illustrates the variability of the cfDNA reads covering a region
- 0 = Unmethylated observation for read
- 1 = Methylated observation for read
- "-" = not covered by read
- "." = read covered, but it is an incorrect base
- Header gives the genomic range covered, and % methylated for each CpG in locus. NA indicates no reads covered that CpG. " * " indicates CpG of interest
- Ex: `> chr1:7224041-7224292(cpg*1*: 0.781, cpg2: 0.962, cpg3: NA)`

Sample entry:

```
> chr1:110347251-110347502(cpg1: 1.0, cpg*2*: 0.891, cpg3: 0.985, cpg4: 0.969, cpg5: 0.99, c
111----
111----
111----
111----
111----
111----
111----
111----
1111---
1010---
1111---
1011---
1011---
1111---
1111---
1111---
--10111
--11111
--.1111
--.1111
--11111
--11111
```

References

1 Rahmani et al. Nat Methods 2017 "Sparse PCA Corrects for Cell-Type Heterogeneity in Epigenome-Wide Association Studies ".

2 ENCODE Consortium, 2012, ENCODE encyclopedia, Version 4: Genomic Annotations.

3 NIH Roadmap Epigenomics Mapping Consortium, 2015, NIH Roadmap Epigenomics Project Data Listings.

4 Guo et al, Nat Gen, 2017, Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA.