Mapping Efficiency

Problems and Goals: It is interesting to study the percentage of trophoblast's DNA in pregnant women's DNA. However, there is a problem that the mapping efficiency of the DNA is too low, typically they are under 30%, and this may make the quality of the results low and affect the final interpretations. The goal of the project is to improve the mapping efficient by optimizing parameters in bowtie1 of bismark.

Background:

1. How Bowtie works: Sequence reads are first transformed into fully bisulfite-converted forward (C->T) and reverse read (G->A conversion of the forward strand) versions, before they are aligned to similarly converted versions of the genome (also C->T and G->A converted). Sequence reads that produce a unique best alignment from the four alignment processes against the bisulfite genomes (which are running in parallel) are then compared to the normal genomic sequence and the methylation state of all cytosine positions in the read is inferred. A read is considered to align uniquely if one alignment exists that has with fewer mismatches to the genome than any other alignment (or if there Is no other alignment).

2. How to measure mapping efficiency:

Mapping efficiency=(Number of sequences with a unique best alignment)/(Number of sequences analyzed)

3. Parameters involved in affecting mapping efficiency

-n: The maximum number of mismatches permitted in the "seed", default 1

-l: The "seed length"; i.e., the number of bases of the high quality end of the read to which the -n ceiling applies, default 28

-e: Maximum permitted total of quality values at all mismatched read positions throughout the entire alignment, not just in the "seed", default is 70

Methods:

We improve the mapping efficiency by changing the –n to 3, therefore roughly 10% of error allowed on the seed. We also change the –e to 300. There are total 150 bases in each read before preprocessing, and assume quality values of 20 (which is 1 error in 100) is a criterion we can trust as mismatch, this methods will also allow roughly 10% of mismatches.

Results:

The first group of tests is conducted on the projects in450, which are two compliment sequences of methyl-maxiSeq single-end human sapiens data hg19. Using the default settings, the results are summarized as below:

Group	Mapping efficiency	Unique CpGs	Avg CpG Coverage	Bisulfite Conversion Rate
0621-12-r1	43%	83419	21X	>99%
0621-12-r2	43%	82384	22X	>99%

After changing the –n=3, -e=300, the mapping efficiency can be improved;

Group	Mapping efficiency	Unique CpGs	Avg CpG Coverage	Bisulfite Conversion Rate
0621-12-r1	48%	92541	22X	>99%
0621-12-r2	49%	90425	23X	99%

The mapping efficiency improves from 43% to 48%-49% after adjusting the parameters

The next step is to implement on the data we are interested, the data of trophoblast's DNA and pregnant women's DNA. We used a subset of 30,000,000 reads for each test. Since the data is a subset, we don't show the Unique CpGs and Avg CpG Coverage here since it is not compatible to the whole data. With default settings, the mapping efficiency of the whole data is:

Group	Mapping efficiency	Bisulfite Conversion Rate
Girl-trophoblast	24%	98%
Boy-trophoblast	26%	99%
Plasma-Pregnant	21%	99%

And after we change the settings with –n=3, -e=300. The mapping efficiency is shown below:

Group	Mapping efficiency	Bisulfite Conversion Rate
Girl-trophoblast	38%	97%
Boy-trophoblast	36%	98%
Plasma-Pregnant	31%	98%

The result shows the new setting can improve the mapping efficiency for more than 10%, which is promising.

Summary and Future Directions

One concern of this setting is that now it allows roughly 10% of the error using the optimized parameters, which may influence the final accuracy of the

result. Another concern is that the mapping efficiency is still slightly below 40%, therefore the quality of maxiseq processing may still not high enough.

To improve the result, one way maybe to increase –e, and of course the larger –e, the more error it will introduce. We may need to have some test to verify the robustness to make sure it won't lead a wrong result. Another way maybe try some other aligner, for example bowtie2, BWA, STAR et al.