ChIP-PIPE

1. WORKFLOW

This pipeline takes FastQ file as input.

1.1 Mapping

The first step of this pipeline is to map reads to the reference genome.

While mapping reads to reference genome, there might be mismatches, deletions or insertions which cause low mapping quality. The mapping quality corresponding to the identity between the reads and the reference genome spot. Low quality means low reliability. BWA retain only one alignment for each read, while Bowtie provides a custom option to decide how many alignments will be retained in alignment file. A chimeric alignment is represented as a set of linear alignments that do not have large overlaps. Typically, one of the linear alignments in a chimeric alignment is considered the "representative" alignment, and the others are called "supplementary" and are distinguished by the supplementary alignment flag. Supplementary alignments also are recorded in alignment file. If a query have multiple alignments(e.g. repetitive sequence), bwa will choose one randomly, bowtie will retain alignments according to demand.

Here we test these two alignment tools using the following command:

BWA using the default set:

```
bwa aln ${ref_dir} ${fastq_dir} > ${name}.sai
bwa samse ${ref_dir} ${name}.sai ${fastq_dir} > ${name}.sam
```

BOWTIE:

```
bowtie -v 3 -k 1 --best -S -x dm3 ${ref_dir}/dm3 \
${fastq_dir} > ${name}.sam
```

`-v 3` allow at most 3 mismatches while the query is aligned, and `-k 1 --best` as mentioned above, retains the best alignment only.

1.2 Qualification

The second step is to filter out the low quality alignment. The threshold is set by user using `-q` option. Low threshold will retain more alignments therefore more enrichment signal, but this may lead to false positive enrichment. Also this step will retain one alignment if there are secondary alignment or supplement alignment.

Command used in this step is:

```
samtools view -h -q ${map_quality} ${name}.sam -F 07404 -o \
${name}.q30.uni.sam
```

1.3 Deduplication

While preparing the sample for sequencing, multiple copies of one fragment may produced by PCR. They do not represent any of the enrichment of target protein but are presented in the FastQ file, they share be moved. Also, while sequencing a single amplification cluster, incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument. They are called duplicates, the latter one is called optical duplicate. After alignment, alignments that start and end at exactly position will be regarded as duplicates, PICARD will recognize these alignments and mark them with a flag. After being marked, they will be removed.

Command applied at this step is:

```
java -jar $Picard MarkDuplicates \
I=${name}.q30.uni.sorted.bam \
O=${name}.q30.uni.marked_dup.bam M=marked_dup_metrics.txt
samtools view -F 07404 ${name}.q30.uni.marked_dup.bam -o
${name}.q30.uni.dedup.bam
```

1.4 Peak calling

The peak-calling step identifies significantly enriched loci (peaks) in the genome.

There are three types of enriched regions: sharp, broad and mixed. Sharp peaks are generally found for protein-DNA binding or histone modifications at regulatory elements, whereas broad regions are often associated with histone modifications that mark domains and there are no clear peak summits and sequence specificity, for example, transcribed or repressed regions. In this pipeline, peak type can be specify by `-p option`.

As MACS is the recommended tool used for peak calling without control sample, here we test two different version of MACS: MACS2.1.0 and MACS3.0.0a6.

Command for MACS2.1.0:

```
macs2 callpeak -t ${name}.q30.uni.dedup.bam -g dm -n ${name}
--${peak_type}
```

Command for MACS3.0.0a6:

```
macs3 callpeak -t ${name}.q30.uni.dedup.bam -g dm -n ${name}
--${peak_type}
```

1.5 Peak visualization

MACS produce .bed format file for visualization. These bed file can be viewed in UCSC genome browser, IGV, or WashU Epigenome browser.

2. Test data

Follow data sets are processed(table 1):

Table 1. Data tested in this pipeline.

factor	GEO	fastq	peak type
H3K27ac	GSM1017494	SRR585051.fastq	broad
H3K27me3	GSM480157	SRR038285.fastq	broad
RNAPII	GSM1017403	SRR585050.fastq	narrow

factor	GEO	fastq	peak type
SuHw	GSM685610	SRR121533-SRR121539.fastq	narrow

2.1 Mapping

All the mapping results are shown in table 2. As mentioned above, we set BOWTIE to allow at most one alignment in the sam file. While doing mapping and q30, BOWTIE retains more alignment than BWA as BWA alignment is 15 ~30% less than BOWTIE alignment. But after deduplication, this difference can decreases to less than 15%. It seems like that BOWTIE have more duplicates in its alignment file.

Table 2. Reads statistics across the processed.

	FASTQ reads	Mapping tool	Quality mapped reads		Deduplicated reads	
H3K27ac	35816233	BWA	28338917	79.12%	17925121	50.04%
		BOWTIE	34168481	95.40%	21027188	58.71%
H3K27me3	8828433	BWA	1891069	21.42%	1683513	19.07%
		BOWTIE	4548945	51.53%	3756489	42.55%
RNAP II	22599246	BWA	15127865	66.94%	9319733	41.24%
		BOWTIE	19598701	86.72%	12213570	54.04%
SuHw	203088537	BWA	126392273	62.23%	42513651	20.93%
		BOWTIE	-	-	-	-

2.2 Peak calling

Different ways of peak calling are presented in figure 1. Here shows that

Figure 1. Peaks visualized with WashU Epigenome browser. NCBI peak file using bowtie v0.12.7 do the map, peak calling use MACS v1.4.1. 14 in the name of peak file means that this peak file is called using MACS v2.1.0 and b in the name of peak file means that this peak file using BOWTIE to do the alignment. No 14 or b in the name means the peak file is produced using MACS3 and BWA respectively.

MACS v2.1.0 and MACS v1.4.1 call broader peaks than MACS3. While looking into the region covered by peaks, we can see that MACS3 called peaks have less superpositions than the other two older versions, which can mean higher resolution. Also, we can observe that most of the peaks are overlapped between different procedure's peak file. But still there are several intervals do not overlap, so I did some statistics in table 3.

Table 3. Peak statistics

	Total peak number	NCBI -		- NCBI	
NCBI peak	5695	-	-	-	-
BWA_MACS2	4449	1487	26.1%	463	10.4%
BOWTIE_MACS2	6117	1331	23.3%	1894	31.0%
BWA_MACS3	12072	197	3.4%	1726	14.3%

BOWTIE_MACS3	15070	158	2.8%	4413	29.3%

As can see, the older version of MACS calls less peaks than MACS3, it agree with the figure 1 that MACS call small intervals and peak regions are highly overlapped. Also bedtools is used to find out the unique peaks.

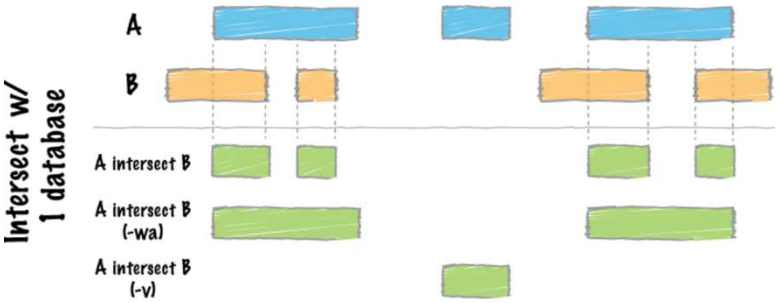

Figure 2. Scheme of how the unique peaks are selected. If a peak is presented in A, and B have no peak intersect with, then this peak will write to file.

Statistics are presented in table 3. NCBI – means peaks that present in NCBI peak file and not in the corresponding file. MACS2 will loss about 20% peaks than MACS v1.4.1 which may because of the improvement of algorithm that abandons some false peaks. MACS3 loss less peaks than MACS2, this may be because MACS3 called small peak have a few intersect will MACS 1.4.1 will cause the peak non-unique. Also, there are some peaks present in MACS2 called peak file. BOWTIE have 15~20% more unique peaks than BWA which can be inferred from the alignment file in table 2. BOWTIE retains more alignments than BWA which lead to more peaks.

3. Conclusion

Based on the result in part2, to make the peaks subtler and more credible, I chose BWA and MACS3 for the final pipeline.