BMEG-3105 Data analytics for personalized genomics and precision medicine

Lecture 17: Genomics data analysis

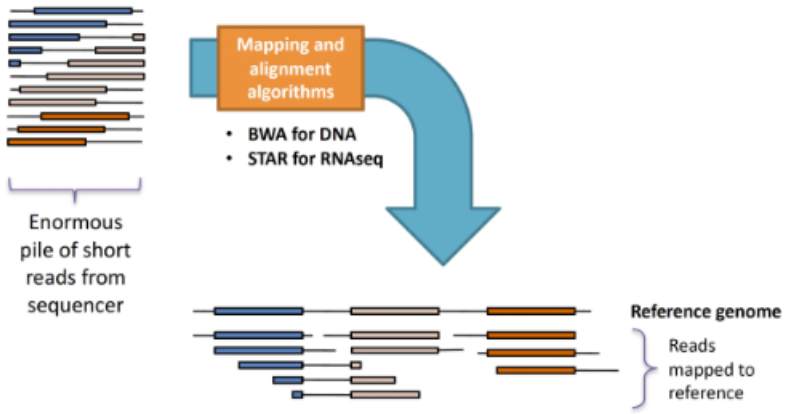
Name: Lee KinHei SID: 1155158901

1. Genome

We must distinguish between actual variation (real change) and errors (artifacts) for example, if we only have single read character difference, we may not sure if it is a real change or an error, but, if there are multiple reads that indicate the same change, we may think that it is an actual change instead of error.

Data processing pipeline

1.1 Map the reads produced by the sequencer to the gene reference



Input format: FASTQ

Each character in the sequence will have a quality score and it is represented by ASCII characters from “!” to “K”

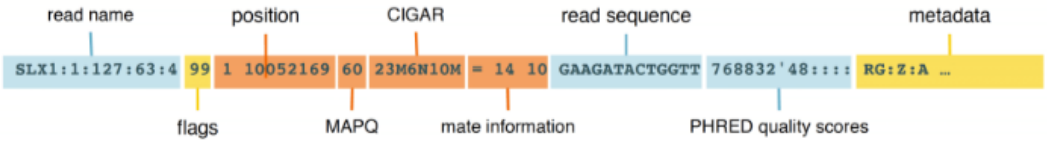
Output: SAM/BAM (SAM stored as text, BAM stored as binary format)

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C       BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
        AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G       DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFCCC
        AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C       DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
        AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
        0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

HEADER lines starting with @ symbol describing various metadata for *all* reads

```
@HD VN:1.6 SO:coordinate ——— BAM header line
@SQ SN:seq1 LN:394893    ——— Reference sequence dictionary entries
@SQ SN:seq2 LN:92783
@RG ID:A SM:SAMPLE_A     ——— Read group(s)
```

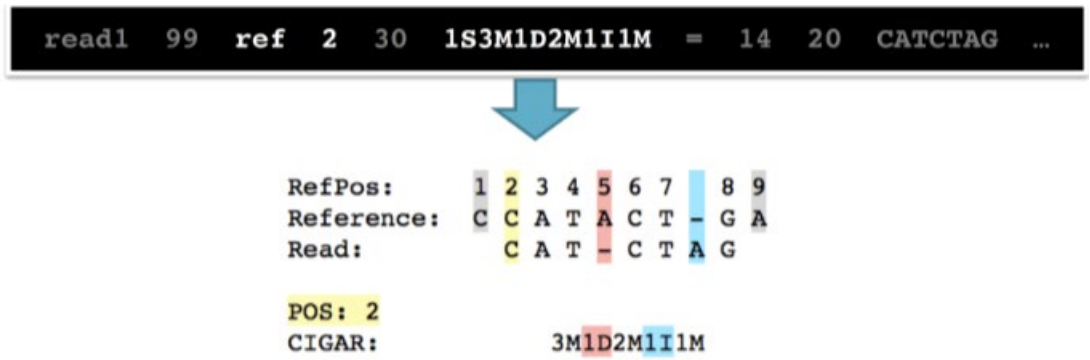
RECORDS containing structured read information (1 line per read/record)



- Added mapping info summarizes **position**, **quality**, and **structure** for each **read**
- Mate information points to the read from the other end of the molecule (other in a pair)

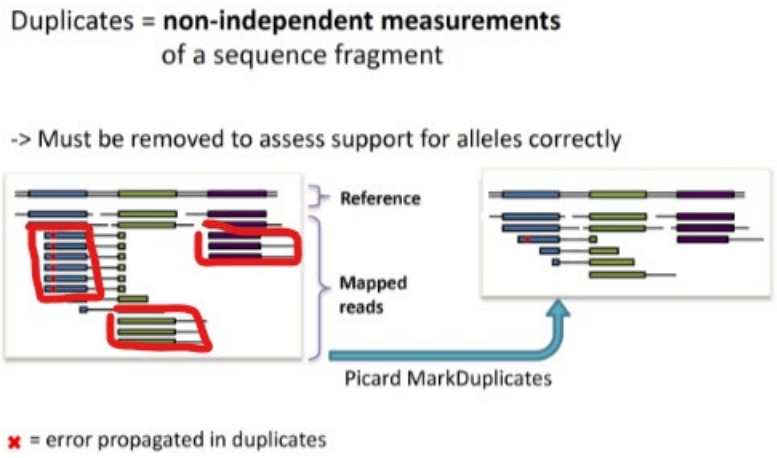
Here we want to specifically mention CIGAR (alignment)

CIGAR → alignment report



We can see starting from POS:2 we have 3 matches from index 2,3,4 so 3M will be there. Then, 1 deletion from the reference so 1D for index 5. 2 matches for index 6 and 7. 1 insertion to the reference after index 7 to match Read A. 1 more match afterward for index 8. All of the above created a sequence mapping with the best score.

1.2 Mark duplicates



The duplicates may mislead us to let us believe the random errors as actual variants because when it increases the number of reads and increases the confidence score in variant calling.

Where these duplications come from?

1. Library duplicates caused by PCR
2. Optical duplicates

1.3 Variant Calling

Variant Call Format (VCF)

Simple comparison where the reference and the read is different

HEADER

```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
```

RECORDS

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	DP=14;AF=0.5	GT:GQ:DP	0/0:48:1	1/0:48:8	1/1:43:5
20	1230237	.	T	.	47	PASS	DP=13	GT:GQ:DP	0/0:54:7	0/0:48:4	0/0:61:2
20	1234567	.	GT	G	50	PASS	DP=9	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

POS: The start coordinate of the variant

REF: Found in reference genome

ALT: Found in sample you are studying

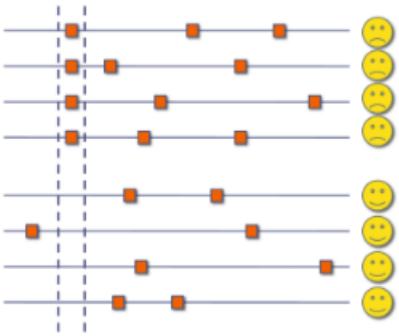
INFO: additional information eg DP=combined depth across samples, AF=allele frequency for each ALT allele in the same order as listed

We should have joint variant calling to increase the credibility

Downstream analysis after data preprocessing pipeline

Genome-wide association studies(GWAS)

Try to determine whether specific variant(s) in many individuals can be associated with a trait like a disease.



Spot the variant that is common amongst all affected but absent in all unaffected

The ideal case (for some rare Mendelian diseases)

But in reality, we have much more cases to study and thus, we need to adjust the statistical significance value to increase the credibility.

Bonferroni Correction

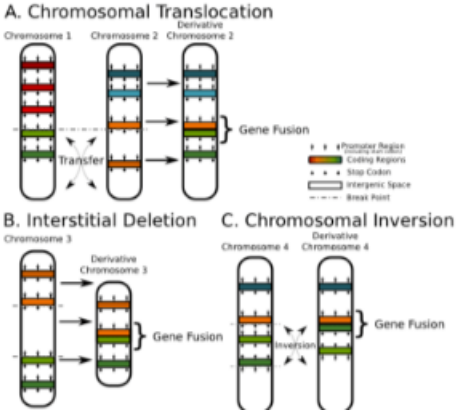
Originally, p-value of 0.05 should be enough for a single test but how about 1 million tests.

We need to adjust the p-value to p-value/number of tests = 5 * 10^-8 to make the whole test more convincible.

RNA-seq data analysis

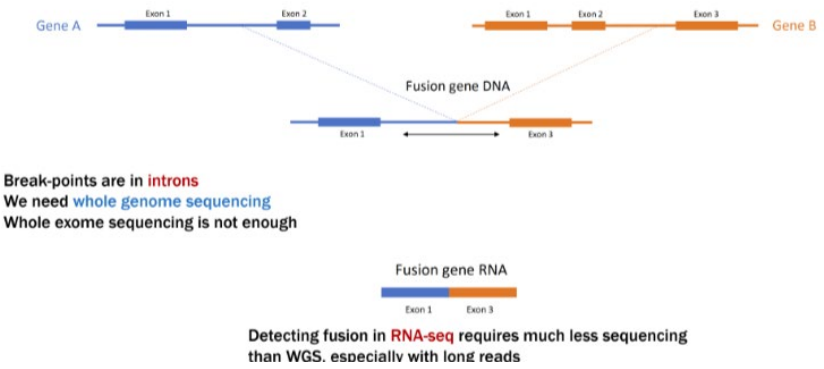
2.1 gene fusion

First fusion gene was described in cancer cells in early 1980s and it means novel gene formed by fusion of two distinct wild type genes.



Gene fusion is a specific kind of structural variant related to cancer

2.2 RNA-seq for gene fusion detection

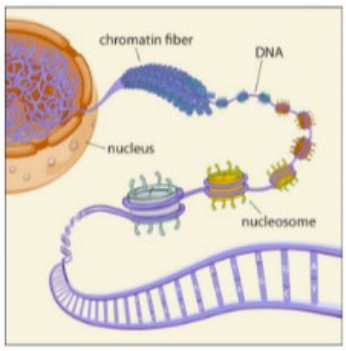


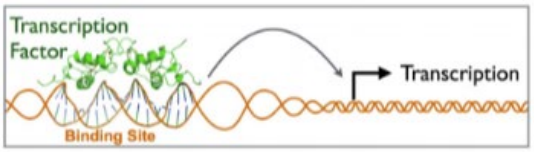
In gene fusion, after the fusion gene DNA translated to protein, we will have only Exon1 and Exon3 which is a mature mRNA. The break point will happen in the introns, so it is impossible to detect if we only check the mature mRNA. Thus, we need to do the whole genome sequencing.

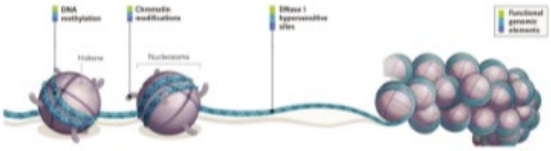
We will find a very long gap between reads which maybe a gene fusion.

Epigenome

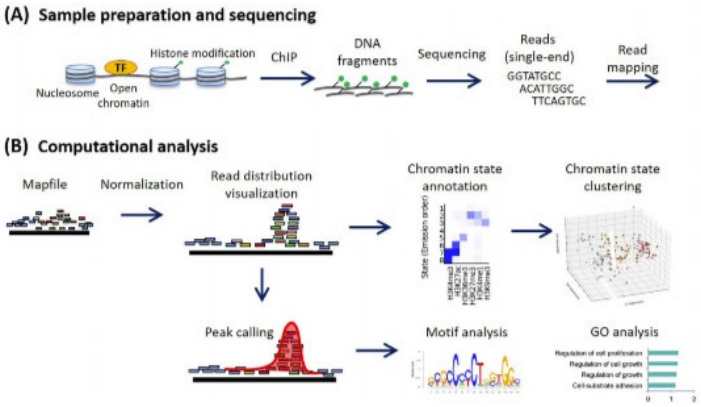
Try to identify the location of the modified DNA as well as the modified Histon.







3.1 Overall data analytics pipeline for epigenetics



Peak calling

It is a statistical testing that finds the peak by contrasting the peak shape and random background.

Peak calling output: Browser Extensible Data

```
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196    127472363    Pos1    0    +    127471196    127472363    255,0,0
chr7    127472363    127473530    Pos2    0    +    127472363    127473530    255,0,0
chr7    127473530    127474697    Pos3    0    +    127473530    127474697    255,0,0
chr7    127474697    127475864    Pos4    0    +    127474697    127475864    255,0,0
chr7    127475864    127477031    Neg1    0    -    127475864    127477031    0,0,255
chr7    127477031    127478198    Neg2    0    -    127477031    127478198    0,0,255
chr7    127478198    127479365    Neg3    0    -    127478198    127479365    0,0,255
chr7    127479365    127480532    Pos5    0    +    127479365    127480532    255,0,0
chr7    127480532    127481699    Neg4    0    -    127480532    127481699    0,0,255
```

3.2 Histone marks and chromatin accessibility