

BMEG3105 Lecture 17

WU Sio Fong 1155173201

1. Cancer

❖ Definition of cancer:

Disease in which some of the body's cells grow uncontrollably and spread to other parts of the body

❖ How to study cancer?

- Genetic variants
 - Genome
 - Gene fusion (RNA-seq)
- Abnormal gene expression
 - Genome (genetic information)
 - Epigenome (environment)
 - Transcriptome (direct measurement)

2. Overview of today lecture

❖ Genome

- Variant calling
- GWAS

❖ RNA-seq

- Gene fusion---structural variant

❖ Epigenome

- Peak calling

3. Genome

❖ Variant calling

Reason:

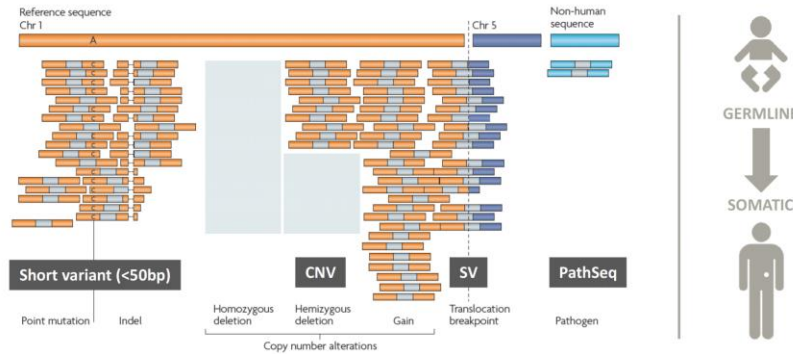
- 3.2 billion sites in the human genome
 - Any 2 humans share 99.5% DNA
 - Can efficiently describe a genome with relation to a reference
- Genetic differences can lead to differences in disease risk and response to treatment
- Genetic variation can be used to find genes and variants that contribute to disease
- Cancer: genetic variants at multiple levels

Types of variant calling:

- Short variant: point mutation, indel(<50bp)
- CNV: homozygous deletion, hemizygous deletion, gain

- SV: translocation breakpoint (gene shift from other location)
- PathSeq: pathogen (non-human)

Different types of genomic variants



Genomics analysis

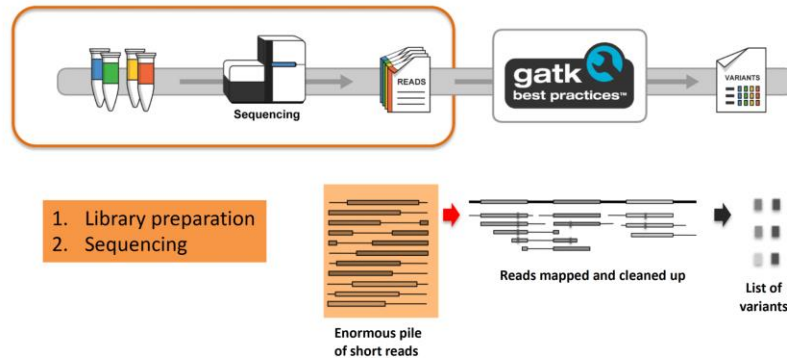
Yu Li

Lecture 17-10

Ways to discover genetic variants:

Library preparation and sequencing

How to discover the genetic variants?



Genomics analysis

Yu Li

Lecture 17-11

Variants VS errors

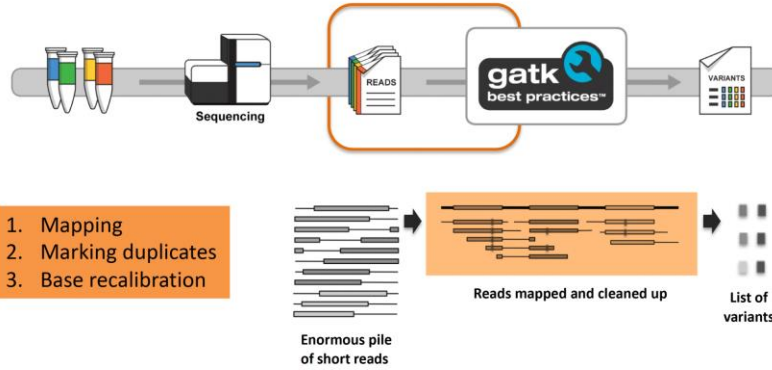
actual variation (real change)

errors (artifacts) -- Errors creep in on various levels

- PCR artifacts (amplification of errors)
- Sequencing (errors in base calling)
- Alignment (misalignment, mis-gapped alignments)
- Variant calling (low depth of coverage, few samples)
- Genotyping (poor annotation)

Procedure for data pre-processing:

Data pre-processing step



1. Mapping
2. Marking duplicates
3. Base recalibration

Genomics analysis

Yu Li

Lecture 17-16

Step1: map the reads produced by the sequence to the reference
 Input: FASTQ, a text-based format for storing both a biological sequence and its corresponding quality scores.

Output: SAM/BAM.

SAM: Sequence Alignment Map (store in text)

BAM: Binary Alignment Map (store in binary only i.e., 1& 0)

The following pictures are FASTQ and BAM format respectively.

Input format: FASTQ

```

FASTQ file example:
@SRR6487486.1.1.length=188
CCCTCGCTACAGCGACAAC...GATTGACCTACGCGAAGTG
+
SRR6487486.1.1.length=188
BBBBBFFFFFFFFFFFFFF...FFFFFFFFFFF
    
```

Labels: Sequence name, DNA sequence, Quality line break, Quality scores

Row: 1
Seq: 188

Quality scores as ASCII characters:
 ! 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J K
 P_{error}: 1.8 0.32 0.022 0.001 0.0001

Q = -10 log₁₀ P_{error}

Genomics analysis

Yu Li

Output format: Sequence/Binary Alignment Map (SAM/BAM)

```

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
@RG ID:A SM:SNPEX_A Read group(s)

RECORDS containing structured read information (1 line per read/record)
read name position CIGAR read sequence metadata
SLX1.1.1237.62.4 99 | 10052169 60 23068106 - 14 10 GAAGATACTGGTT 768832 48... RG:2:A ...
    
```

Labels: flags, MAPQ, mate information, PHRED quality scores

- 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)

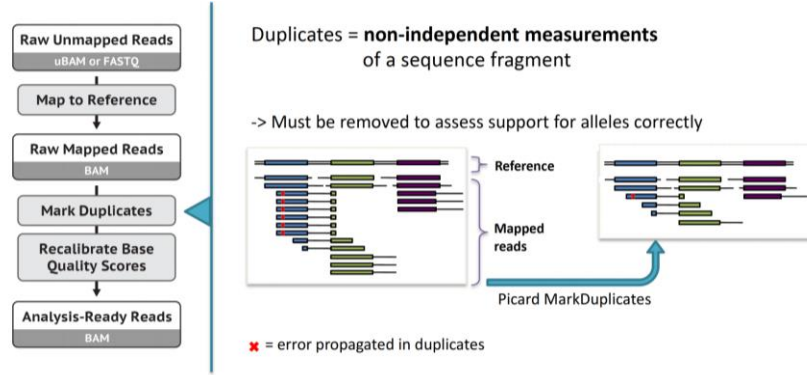
<http://samtools.github.io/hts-specs/SAMv1.pdf>

Lecture 17-19 Genomics analysis

Yu Li

Lecture 17-21

Step2: mark duplication to reduce duplications which come from some experimental manipulations.

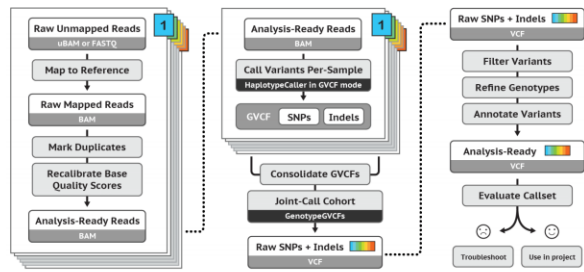


Error may be due to noise/ error/ duplicate (we must do this to avoid we mistake error as real variance)

Variant calling:

After analysis of data from the above operations, we want to find variants from reads, so we do:

Variant calling in more detail



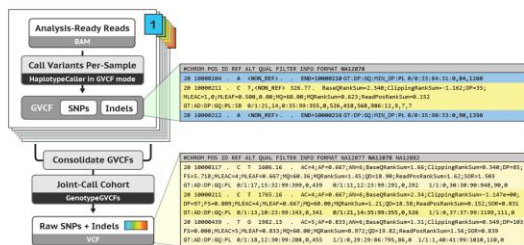
Genomics analysis

Yu Li

Lecture 17-27

VCF is some information in below form:

From per-sample GVCFs to final multi-sample VCF



Genomics analysis

Yu Li

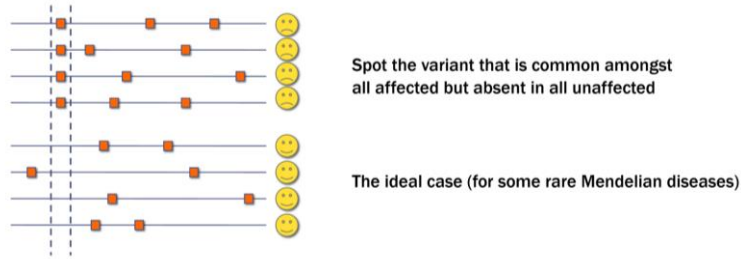
Lecture 17-30

However, after analysis of genomes, we should joint some data together to conclude a result because a single genome data is always unpowered, a joint call set could provide more valuable information.

4. Further downstream analysis for cancer diseases:

GWAS:

Trying to determine whether specific gene variant is related to a disease



To find the correlation between SNPs to diseases:

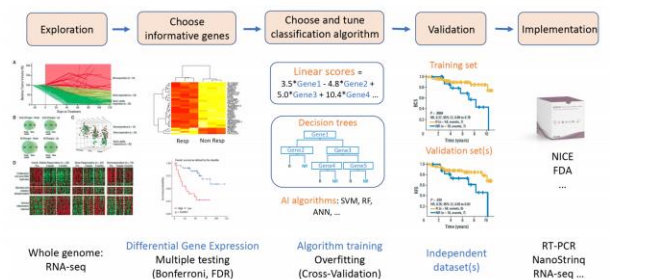
Bonferroni correction --adjust P-value

Adjusted p-value= p-value/ number of tests

RNA-seq data analysis (genetic variant level to study cancer)

A basic procedure of RNA-seq data analysis

RNA-seq data analysis

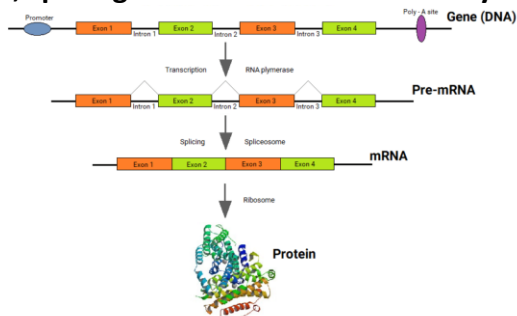


Genomics analysis

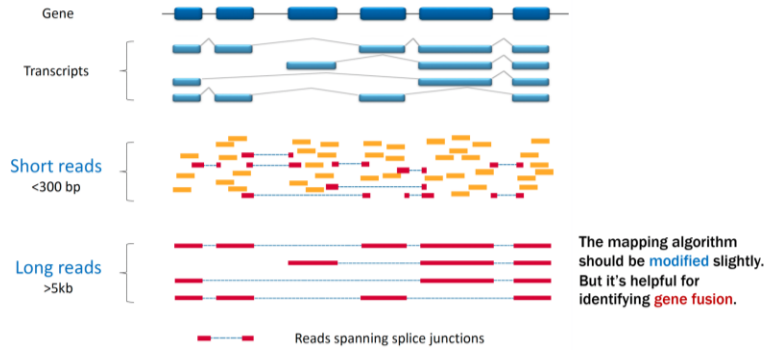
Yu Li

Lecture 17-41

Transcription, splicing and translation of a eukaryotic gene



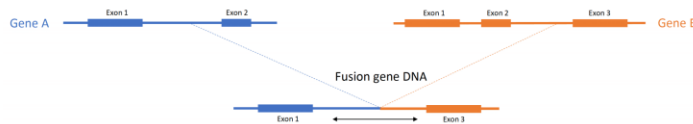
Mapping spanning splice junctions



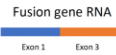
Gene fusion:

- Novel gene formed by fusion of two distinct wild type genes
- Is a specific kind of structural variant related to cancer
- In cancer: produced by somatic genome rearrangements

RNA-seq for gene fusion detection



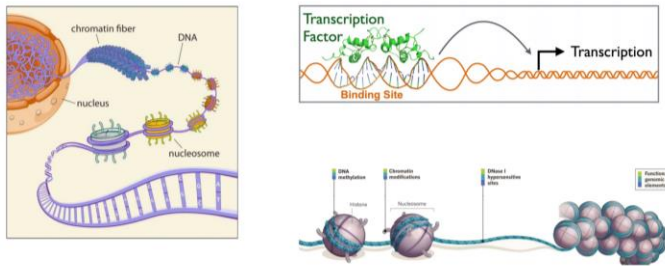
Break-points are in introns
 We need whole genome sequencing
 Whole exome sequencing is not enough



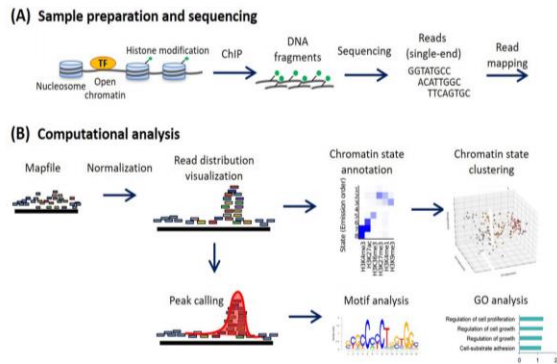
Detecting fusion in RNA-seq requires much less sequencing than WGS, especially with long reads

5. Epigenome

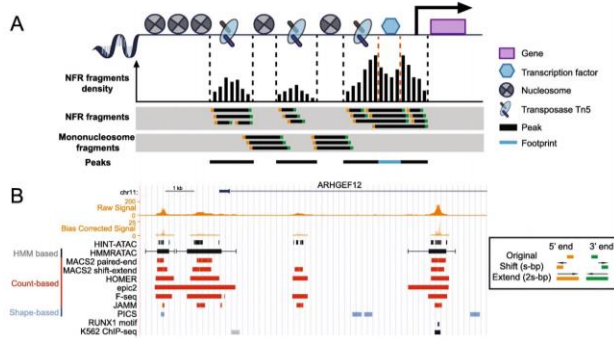
Structure of epigenome



The overall data analytics pipeline for epigenetics



Peak calling

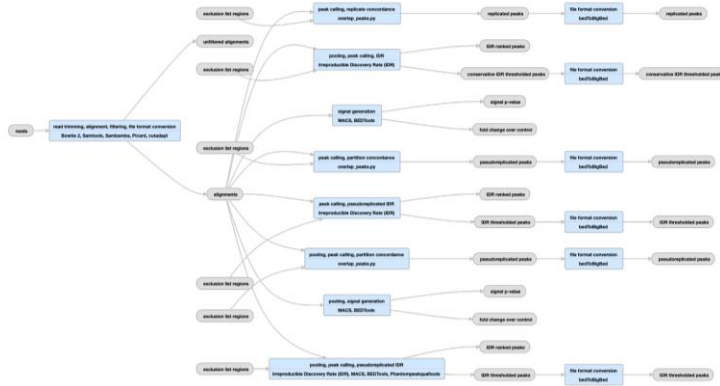


The output of peak calling (Browser Extensible Data (BED) format)

```

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
    
```

The Entire Detailed Pipeline (ATAC-seq as an example)



Histone marks and chromatin accessibility

