FALL 2022 Data analytics for personalized genomics and precision medicine Course introduction

Lecturer: Yu LI (李煜) from CSE Liyu95.com, liyu@cse.cuhk.edu.hk

2 November

Course agenda:

Genome

Variant calling (very complicated)

GWAS

RNA-seq

Gene-fusion—structural variant



Peak calling

Genome (genetic variant level to study cancer)

Why we care about variant?

- Human share 99.5% of genome, but the 0.5% difference could lead to lots of diseases and difference response to treatment.
- One of the examples is cancer which is caused by genome variant in multiple levels.

Types of genome variants:

- 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



Ways to discover genetic variants:

- Library preparation (from gene bank sometimes)
- Sequencing (way similar to mapping)

Variant VS error:

- Variant: real change
- Error: man-made and creep in on various experimental manipulations (PCR, sequencing, etc.)

Procedure for data pre-processing:



• Step1: map the reads produced by the sequence to the reference, with an input format

named FASTQ, a text-based format for storing both a biological sequence and its corresponding quality scores. And output a sequence or binary alignment map with header and structured read information. The following pics is FASTQ and BAM format respectively:

ТАТТО По кондина совоемоте 11 ведин-100 стопитичностьющиходожительносторожительстворского пользовательства и пользовательство и пользовательство совоемотель 11 ведин-100 намалителятелятелятелятелятелятелятелятелятеля		HEADER lines starting with @ symbol describing various metadata for all reads <pre>@HD VM:1.6 S0:coordinate — BAM header line</pre>						
05RR6407486.1 1 length=100 CCTCGTCTACAGCGACAAC GATTTGACCTACGTCGAAGTG +SRR6407486.1 1 length=100	Sequence zame DNX requence Quality line break	eso sn:seq1 LN:394893 — Reference sequence dictionary entries eso sn:seq2 LN:92783 — Read group(s) erc ID:A sn:sample_A — Read group(s)						
DDBDDFFFFFFFFFFFF FBFFFFFFFFFFFFFFFF	Quality scores	RECORDS containin	g structured r	read informa CIGAR	tion (1 line per read, read sequence	/record) metadata		
Orability serves as ASCII: characters: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	= -10log. _g P _{arror}	SLX1:1:127:63:4 9 flag	9 1 10052169 (gs M/	60 23M6N10M =	14 10 GAAGATACTGGT	T 768832 '48:::: RG:Z:A PHRED quality scores		

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

Variant calling:

After analysis of data from the above operations, the variant calling details be like:(this is a procedure finding variants from reads)



However, after analysis of genomes, we should joint some data together to conclude a result because a single genome data is unpowered, a joint call set will be more useful for data analysis.

Further downstream analysis for cancer diseases:

GWAS: a study about the SNPs which is of huge amount, trying to determine whether

specific gene variant is related to a disease. For example: pot the variant that is common amongst all affected but absent in all unaffected.

To find the correlation between SNPs to diseases:

Bonferroni correction, is just to adjust P-value, suppose we have N SNPs, then the adjusted P-value is 0.05/N.

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

A basic procedure:



Mapping procedure which is helpful for identifying gene fusion

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

Advantage for RNA-Seq:

Detecting fusion in RNA-seq requires much less sequencing



Principle for RNA-Seq detecting DNA fusion:

Epigenome (gene expression level to study cancer)

Structure of chromosome and co-factors give chance to regulate gene expression:



The overall data analytics pipeline for epigenetics

(A) Sample preparation and sequencing





Peak shape is apparent in random background

The output of peak calling is BED file

Based on the chromosome, start of gene, end of gene and the label of gene

track	name="ItemRGBD	emo" descript	ion="Ite	em RGB	demo	onstration"	visibility=2	itemRgb="0n"
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