Data analytics for personalized genomics and precision medicine

Lecture 17: Genomics data analysis

Lecturer: Yu LI (李煜) from CSE

Liyu95.com, liyu@cse.cuhk.edu.hk

Wednesday, 2 November 2022

***** Outline of the Lecture:

- About previous lecture
- Overview of this lecture
- ➢ Genome
- ≻ RNA-Seq

1.About previous lecture

- More about cancer
 - Definition of cancer:
 - Disease in which some of the body's cells grow uncontrollably and spread to other parts of the body
 - Study cancer at multiple level:



- Genetic variants
 - Genome
 - Gene fusion (RNA-seq)
- Abnormal gene expression

- Genome (genetic information)
 - Can be differed even when the DNA is not different
- Epigenome (environment)
- Transcriptome (direct measurement)

2. Overview about this lecture

- Data analytics for cancer genomics
 - Genome: variant calling, genome association study
 - > Epigenome: what is it, peak calling, differential peak calling
 - RNA-seq: DEG, gene fusion

3.Genome

- ✤ Variant calling
 - Reason that variants are important
 - 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 used to find genes and variants that contribute to disease
 - Cancer: genetic variants at multiple levels
 - Different types of genomic variants



➢ How to discover genetic variants (1st part)



- Library preparation and Sequencing
- Variants VS errors
 - Important to distinguish between actual variation (real change) and 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)
- Data pre-processing step (2nd part)



- Mapping, Marking duplicates, Base recalibration
- Step 1: Map the reads produced by the sequencer to the reference



Input format (FASTQ)

• 4 rows \rightarrow 1 read

a = 100100 + 10 + 17 + 1001 + 1001 + 1000 + 1 + 10 + 0 + 0 +	
GAUDISUSIUSHTANSUMAXIISIIUSIISUSIISUUU ISUSACTI+GICGIICG TGCAGCAGCTAATGAGGAACCACTTCCTCCCCCCAGCCGCTCTAAATACCTCAGAACAATAGGATCAT +	CATAATAATCCCCTAGTCTGAACT
8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF	FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
TGCAGAAGAAGAAAGCACAAGTATTTACGCCTATCCTTCATATTTTCCGCAAGGTAACTATCTCGGTT +	TCATATCGAGATTTATATAGAATC
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF	FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
TGCAGGAAGTTATGCAGGGGCATCCTGTATTATTAAATAGAGCACCTACTCTTCATAGATTAGGTATA + 	CAGGCGTTCCAACCTATTTTAGTG
FF-88F8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF	FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
TGCAGAGTACATCAACAAAAGAAACCTAACTGCCCTACCGGCAAACCGGTAGAGTACCCTTCCCCAAA +	AGTATTACTCCCAGTCAATATAAG
85888555 555555555555555555555555555555	EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
8F888FFF-FFFFFFFFFFFFFFFFFFFFFFFFFFFFF	FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
<pre>@SRR6407486.1 1 length=100</pre>	Sequence name
<pre>@SRR6407486.1 1 length=100 CCTCGTCTACAGCGACAAC GATTTGACCTACGTCGAAGTG</pre>	Sequence name DNA sequence
<pre>@SRR6407486.1 1 length=100 CCTCGTCTACAGCGACAAC GATTTGACCTACGTCGAAGTG +SRR6407486.1 1 length=100</pre>	Sequence name DNA sequence Quality line break
<pre>@SRR6407486.1 1 length=100 CCTCGTCTACAGCGACAAC GATTTGACCTACGTCGAAGTG +SRR6407486.1 1 length=100 BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF</pre>	Sequence name DNA sequence Quality line break Quality scores
<pre>@SRR6407486.1 1 length=100 CCTCGTCTACAGCGACAAC GATTTGACCTACGTCGAAGTG +SRR6407486.1 1 length=100 BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF</pre>	Sequence name DNA sequence Quality line break Quality scores

• Quality scores (ASCII characters)

Quality	score	s as ASCII	characters:				
	! " #	#\$%&'()	*+,/01234567	7 <mark>89:;<=></mark> ?@ABCDEF	GHIJK	<	
						$0 = -10 \log_{10}$	Р
Q:	0	5	15	30	40	510	error
P _{error} :	1.0	0.32	0.032	0.001	0.0001		

• Output format (SAM, BAM)

@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 3read-edit-dist 5read-rea				
lign-edit-dist 2 -i 50 -I 5000max-coverage-intron 5000 -M -o out /data/user446/mapping tophat/index/chr				
20 /data/user446/mapping tophat/L6 18 GTGAAA L007 R1 001.fastg				
HWI-ST1145:74:C101DACXX:7:1102:4284:73714 16 chr20 190930 3 100M * 0 0				
TOTOTOTAATORATORATORATORAATORATORATORATO				
C BBBCCDDCCDDDDCDCCCCDBC2DDDDDDDDDDDDDDD				
AS:1:-15 XM:1:3 X0:1:0 XG:1:0 MD:7:55C20C13A9 NM:1:3 NH:1:2 CC:7:= CP:1:55352714 HT:1:0				
HWT-ST1145-74-C101DACXX-7-1114-2759-41961 16 cbr20 193953 50 100M * 0 0				
AS-i-16 XM-i-3 X0-i-0 XG-i-0 MD-7-60G16T18T3 NM-i-3 NH-i-1				
HUT_STIL45-74-C1010ACXY-7-1204-14760-4030 16 cbr20 278877 50 100M * 0 0				
AS-i-11 XM-i-2 X0-i-0 XG-i-0 MD-7-0085G13 XM-i-2 XH-i-1				
HUT_STITUS_74/C101DACY27-1210-11167-8600 A chr20 271218 50 50M4700D50M * 0				
CANG Company of Alice and Many (stores in toget)				
• SAM: Sequence Alignment Map (store in text)				
• BAM: Binary Alignment Map (story in binary only i.e., 1& 0)				
Difficience of the start of the				

- HEADER lines: start with @ symbol→describe various metadata for all reads

@HD	VN:1.6 SO:coordinate BAM header line
@sg	SN: seq1 LN: 394893 Potoropoo soguopoo dictionary entrico
eso	SN:seq2 LN:92783
@RG	ID:A SM: SAMPLE_A Read group(s)

• RECORDS: contain structured read information \rightarrow 1 line/read

read name		position	CIGAR			read sequence		metadata
SLX1:1:127:63:4	99 1	l 10 <mark>052169</mark> 60	23M6N10M	= 14	10	GAAGATACTGGTT	768832'48::::	RG:Z:A
	ГТ	T						
fla	ags	MAP	'Q ma	te info	rmat	ion P	HRED quality score	es

- 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)

• CIGAR (Concise Idiosyncratic Gapped Alignment Report)

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description	Consumes query	Consumes reference
М	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
s	4	soft clipping (clipped sequences present in SEQ)	yes	no
н	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
Р	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

• "Consumes query" and "consumes reference" indicate whether the CIGAR operation causes the alignment to step along the query sequence and the reference sequence respectively.

- $\bullet\,$ H can only be present as the first and/or last operation.
- $\bullet\,$ S may only have H operations between them and the ends of the CIGAR string.

 $\bullet\,$ For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.

• Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

• Step 2: Mark duplicates to mitigate duplication artifacts

Raw Unmapped Reads	Duplicates = non-independent measurements of a sequence fragment
Map to Reference	-> Must be removed to assess support for alleles correctly
Raw Mapped Reads BAM	Reference
Mark Duplicates Recalibrate Base Quality Scores	Mapped reads
Analysis-Ready Reads	Picard MarkDuplicates
BAM	# = error propagated in duplicates

- Error may be due to noise/ error/ duplicate
- Variant calling

Raw Unmapped Reads uBAM or FASTQ Map to Reference	Analysis-Ready Reads DAM Call Variants Per-Sample HaptotypeCaller in CVCF mode	Raw SNPs + Indels
Raw Mapped Reads BAM Mark Duplicates Recalibrate Base Ouality Scores	GVCF SNPs Indets	Annotate Variants Analysis-Ready
Analysis-Ready Reads BAM	Joint-Call Cohort GenotypeGVCFs Raw SNPs + Indels	Evaluate Callset

- Joint analysis empowers discovery
- Single genome in isolation→almost never useful
- Family or population data \rightarrow add valuable information
 - Rarity of variants

- De novo mutations
- Ethnic background
- GWAS (Genome-wide association studies)
 - Trying to determine whether specific variant(s) in many individuals can be associated with a trait (disease)



▶ In reality: 3.5 million SNPs



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

4.RNA-seq

RNA-seq data analysis



✤ Transcription, splicing and translation of a eukaryotic gene



Mapping spanning splice junctions



➢ Gene fusion



- The first gene was described in cancer cells in early1980s
- Novel gene formed by fusion of two distinct wild type genes
- In cancer: produced by somatic genome rearrangements
- RNA-seq for gene fusion detection
 - Break-points are in introns
 - > Need whole genome sequencing, whole exome sequencing is not enough
 - > Detecting fusion in RNA-seq requires much less sequencing than WGS
 - ➢ Can be detected by RNA-seq