Analysis of neo-antigens to identify T-cell neo-epitopes in human Head & Neck cancer

Project –XX1001

Customer Detail

Table of Contents

1. Bioinformatics analysis pipeline

The following bioinformatics steps were performed for analysis of the data

1.1. Read quality check - We check the following parameters from fastq file

- Base quality score distribution
- Sequence quality score distribution
- Average base content per read
- GC distribution in the reads
- PCR amplification issue
- Check for over-represented sequences
- Adapter trimming

Based on quality report of fastq files we trim sequence read where necessary to only retain high quality sequence for further analysis. In addition, the low-quality sequence reads are excluded from the analysis. The adapter trimming was performed using fastq-mcf program (version - 1.04.676).

1.2. Read alignment – The paired-end reads are aligned to the reference human genome primary assembly hg19 downloaded from UCSC database. Alignment was performed using BWA program (version = bwa-0.7.8). While running BWA trimming was performed using –q = 20. Reads after alignment with mapping quality >= 20 with insert size >= 60bp and <=1000bp were taken further for analysis. Any paired reads aligning to two different chromosomes were totally discarded from the analysis.

1.3. Duplicate read removal – The aligned reads are first sorted using Picard tool (picard-tools-1.115) Sort Sam command and then the read duplicates were removed using Picard MarkDuplicates command.

1.4. Read realignment – After removing duplicate reads the reads were realigned around the known indels provided by GATK group.

1.5. Base recalibration – After performing realignment the base re-calibration step is performed. After recalibration, the quality score of each base is more accurate. For details please look into the following website: http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr. Known variant position is taken into account to recalibrate the quality score.

1.6. Variant calling – After performing realignment we used Strelka program to call the somatic variants. Default parameter settings were used to call the variants.

1.7. Variant annotation and comparison – The identified variants are annotated using our in-house program (VariMAT). The gene model used for annotation was downloaded from Ensembl database Release 75. The variant class prediction is done using VeP Release 75 and internal annotation pipeline (VariMAT).

Remainder of page left intestinally blank

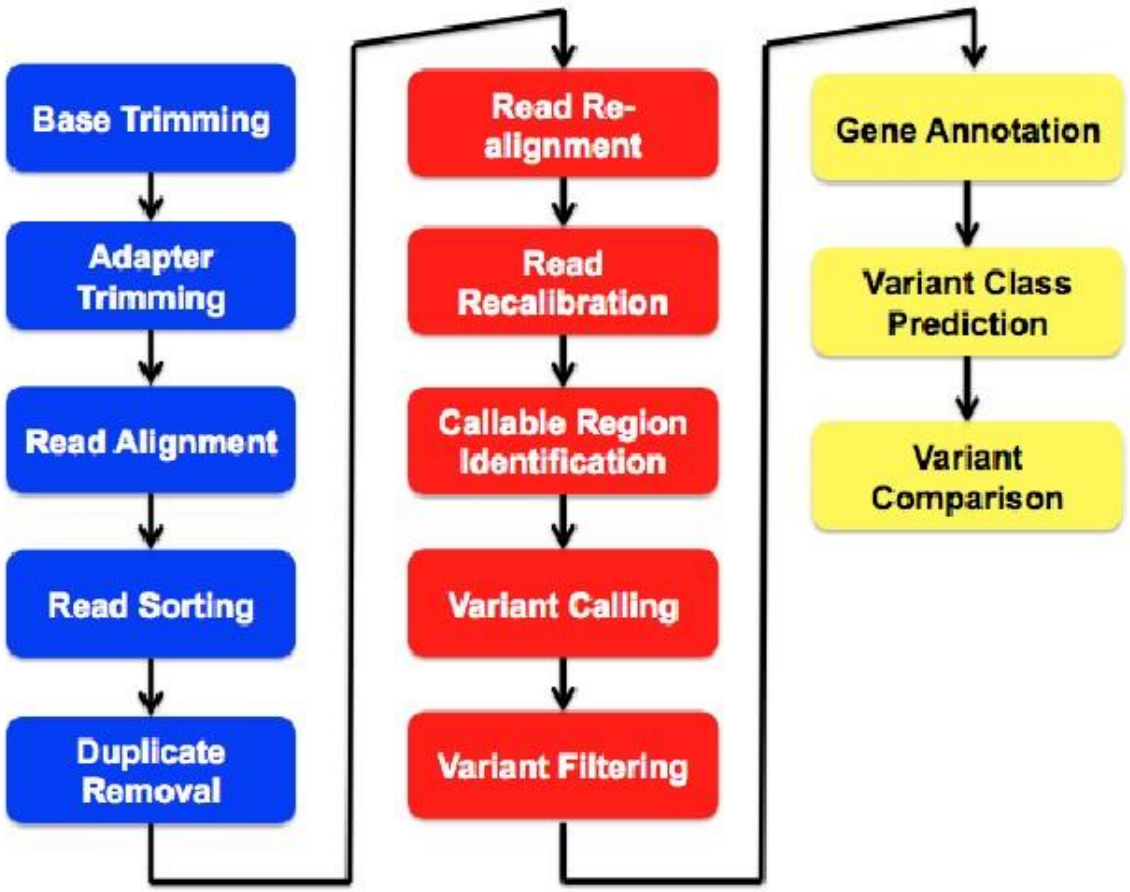

Fig. 1. Bioinformatics analysis pipeline

Remainder of page left intestinally blank

2. Analysis results

2.1. Data summary

The primary data summary of the samples is provided in Table1. Further raw fastq data quality report can be seen in supplementary file (Supp_Results_S1).

Table 1. Primary data summary

	Sample1-Tumor	Sample1-Blood	Sample2-Tumor	Sample2-Blood
Total paired-end	59,500,561	61,023,124	61,888,012	65,154,012
Total Reads	119,001,122	122,046,248	123,776,024	130,308,024
Read length (bp)	100x2	100x2	100x2	100x2
Total Data (Gb)	11.9	12.2	12.3	13.0
Total data (>Q30)	90.01	92.1	90.2	90.1
Average Sequencing Depth	200	200	200	200

2.2. Alignment & Filter

The overall alignment summary is provided in Table 2. More than 98% of the total reads aligned to the reference genome for all the samples. The aligned reads are filtered for mapping quality, insert size estimates and cross mapping read. More than 85% of the reads passes the alignment filtering parameters. More than 85% of the reads have mapping quality >= Q30, which reflects very good alignment quality (Fig. 2). Insert size distribution for all aligned reads is shown in Fig. 3. All samples have average insert size around 200bp. Overall ~ 4-10% of aligned reads were found to be duplicate in all samples. The duplicate reads were removed from further processing.

Table 2. Read alignment summary

	Sample1-Tumor	Sample1-Blood	Sample2-Tumor	Sample2-Blood
Total pre-processed reads	118,990,110	122,022,404	123,752,145	130,300,020
Alignment %	98.36	99.42	99.17	99.43
Passed Alignment %	87.12	89.23	85.87	89.59
Passed aligned reads duplicate %	10.21	4.32	11.29	8.57

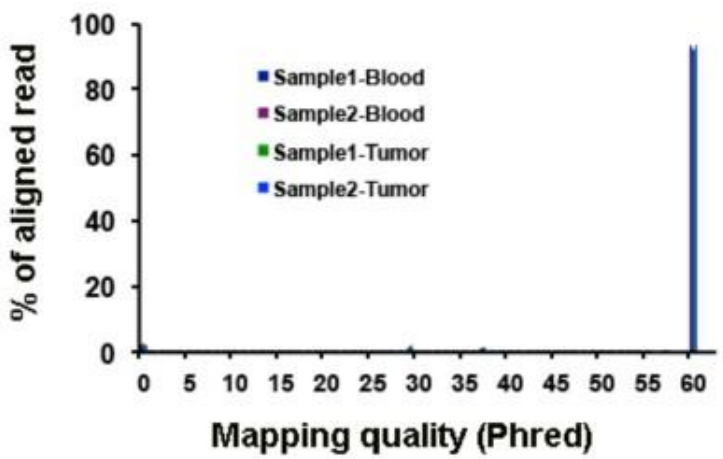

Fig. 2. Mapping quality distribution for aligned reads

Remainder of page left intestinally blank

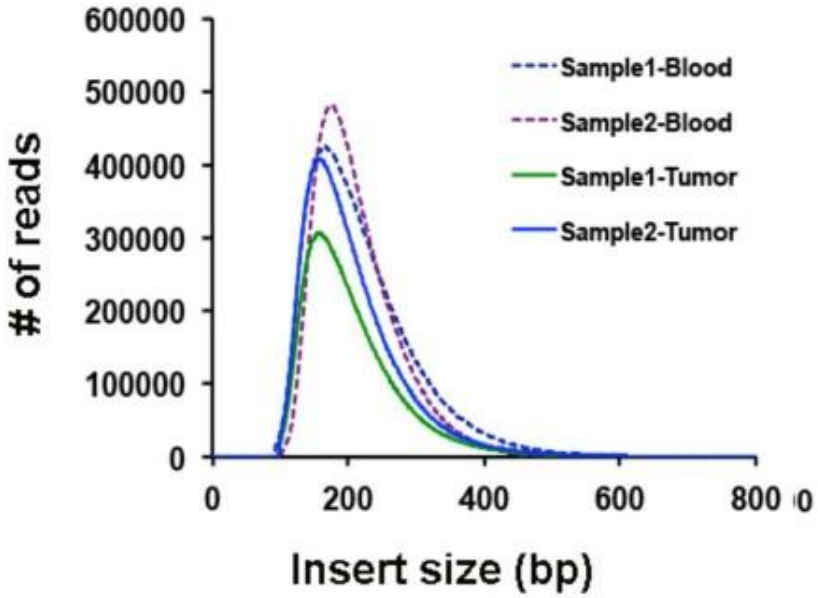

Fig. 3. Insert-size distribution of the aligned reads

Remainder of page left intestinally blank

2.3. Coverage Summary

The coverage statistics is calculated using the filtered aligned reads compared against the capture exon regions. The hg19 bed coordinates provided by the vendor was used to calculate the coverage summary. The overall coverage summary is provided in Table 3.

Table 3. Coverage summary for the sample on the target site

	Sample1-Tumor	Sample1-Blood	Sample2-Tumor	Sample2-Blood
Panel coverage (%)	98.25	98.52	97.58	97.61
Average read depth (X)	110.91	115.12	107.21	105.23
Panel region 1X-10X (%)	4.38	4.09	4.24	5.02
Panel region 11X-30X (%)	6.34	7.26	6.57	6.16
Panel region 31X-50X (%)	13.32	11.47	14.86	11.06
Panel region 51X-100X (%)	33.98	31.23	32.23	36.12
Panel region > 100X (%)	40.23	44.47	39.68	39.25
Aligned reads on-target (%)	75.32	78.23	72.82	75.97

2.4. Variant calling and annotation

Before performing the variant calling the reads were realigned and base-recalibrated using GATK-Lite program. For realignment and base-recalibration the standard files provided by GATK. The variant calling was performed using Strelka program (v2.0.15) with default parameters. The complete annotation file of the variants is provided as supplementary file **(Supplementary Results).** The variant class summary of passed variants is provided in Table 4. The mutations found in OncoMD are provided in Table 5.

Table 4. Variant classification of on-target and passed variants

Tier1	Tier2	Tier3	Tier4	Sample1-Tumor	Sample2-Tumor
Inside Gene	Exonic	Non-Synonymous	Missense	1821	10
			Missense_SS_PRX		
			Nonsense	2	1
			Nonsense_SS_PRX	0	0
			StartCodon	0	0
			StopLoss	0	0
			Frameshift	3	7
			Frameshift_SS_PRX	1	0
			InFrame	0	1
			InFrame_SS_PRX	0	0
		Synonymous	Silent	45	43
			Silent_SS_PRX	3	1
			3UTR	25	21
			3UTR_SS_PRX	0	0
			5UTR	0	0
			5UTR_SS_PRX	0	0
			Exon_Noncoding	17	14
			Exon_Noncoding_SS_PRX	0	0
			Intron_CDS_Conflict	0	0
	Intronic	Splice-Site	SS_ACR	2	0
			SS_DNR	0	1
		Proximal Splice-site	SS_ACR_PRX	17	13
			SS_DNR_PRX	12	8
		Intron	Intron	93	101
	Other		Mature_miRNA	0	0
			None	0	0
Intergenic			Intergenic	2	1

2.5. HLA prediction and binding analysis

The HLA type was performed using Seq2HLA (v2.2) program on the RNA-seq data. The raw data summary of RNA-seq blood samples is provided in Table 5. The default settings of the program were used for HLA class I, II subtype prediction. The HLA type and p-value for prediction is provided in Table 6. The binding analysis was performed using NetMHCcons program for Class I.

Table 5. Primary data summary of RNA-seq samples

RNA-seq	Sample1-Blood	Sample2-Blood
Total reads	53,586,242	64,249,498
# of Reads after adapter trimming	53,524,562	63,041,510
# of Reads after contamination removal	49,919,569	58,327,700
# of aligned reads	43,789,373	55,615,242
% of aligned reads	87.72	95.35

Table 6. HLA Class I subtypes result from Seq2HLA program from RNA-seq samples

Sample	HLA Class-I				
	Locus	Allele 1	Confidence	Allele 2	Confidence
Sample1	A	A*01:01	0.003778207	A*11:01	0.005693306
	B	B*07:05'	0.02564003	B*40:06	0.000690862
	C	C*15:02	8.69E-12	C*07:02	0.0123953
Sample2	A	A*26:01	0.0172551	A*02:11'	0.002925319
	B	B*07:02	0.004728552	B*15:18'	0.00029772
	C	C*07:04'	4.04E-14	C*07:04	0.024963

2.6. Analysis of neo-antigen expression from RNA-seq data

Normalized transcript levels expressed as FPKM (Fragment Per Kilobase per Million) were computed for all transcripts. A cutoff value of 0.5 was used to select expressed genes. Neo-antigens derived from expressed genes were selected further for neo-epitope prediction. Analysis was done to determine a correlation between mutant allele frequency at the DNA and the transcript level as shown in Fig. 4.

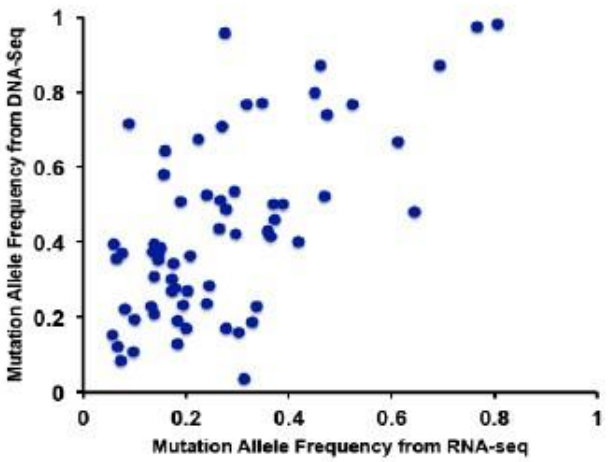

Fig. 4. Correlation of mutant allele frequency between exome sequencing and RNA sequencing datasets

2.7. Proteasomal and Immunoproteasomal processing of peptides

During Class-I presentation, 9-mer peptides are generated from proteins via multiple processing steps regulated by proteasomal and immunoproteasomal protease complex. Peptides having high affinity binding to HLA, and predicted to express in cells based on transcript profile of genes were further analyzed for the presence of protease cleavage sites using proprietary algorithms.
A list of potential T-cell neo-epitopes from a Head and Neck cancer sample is shown in Table 7.

Remainder of page left intestinally blank

Table 7. List of potential T-cell neo-epitopes from two Head and Neck cancer samples

Chrom	*Pos*	*Ref*	*Alt*	*Gene*	*Amino Acid Change*	*ClassI HLA Allele*	*Sample #*
chr12	7362289	T	A	PEX5	p.L545Q	HLA-A02:01	1
chr17	7577547	C	T	TP53	p.G245D	HLA-A11:01	1
chr15	25605560	C	A	UBE3A	p.V578F	HLA-B15:17	2
chr2	217297533 G		T	SMARCAL1	p.R476L	HLA-C03:02	2

Remainder of page left intestinally blank

APPENDIX

Variant Class and annotation terms	Description
3UTR	3UTR of the transcript
3UTR_SS_PRX	3UTR of the transcript and within 2bases from splice site
5UTR	5UTR of the transcript
5UTR_SS_PRX	5UTR of the transcript and within 2bases from splice site
Exon_Noncoding	Exonic region of Noncoding gene/transcript
Exon_Noncoding_SS_PRX	Exonic region of Noncoding gene/transcript and within 2bases from splice site
Frameshift	Frameshift in CDS region
Frameshift_SS_PRX	Frameshift in CDS region and within 2bases from splice site
InFrame	Inframe in CDS region
InFrame_SS_PRX	Inframe in CDS region and within 2bases from splice site
Intergenic	Intergenic region / Non gene region
Intron	Intronic region of the transcript
Intron_CDS_Conflict	Something wrong in the gene model. Not able to decide
Mature_miRNA	Present in mature miRNA region as per VeP program
Missense	Missense in CDS region
Missense_SS_PRX	Misense in CDS region and within 2bases from splice site
None	None of the variant class. Mostly is some non-coding region of the gene
Nonsense	Nonsense in CDS region
Nonsense_SS_PRX	Nonsense in CDS region and within 2bases from splice site
Silent	Silent in CDS region
Silent_SS_PRX	Silent in CDS region and within 2bases from splice site
SS_ACR	Splice site acceptor site of the transcript
SS_ACR_PRX	Between 3-10 bases from the acceptor site of the transcript
SS_DNR	Splice site donor site of the transcript
SS_DNR_PRX	Between 3-10 bases from the donor site of the transcript

StartCodon	Start codon of the transcript
StopLoss	Stop codon of the transcript
PCT	Protein Coding Transcript
NCT	Non-coding Transcript
NMDT	Nonsense mediated decay transcript