Project Title

Tumor micro-environment analysis from human cancer

Customer Detail

XXXXXXX

Report

Date: XXXXX

Table of Contents

1. Overview

1.1 Sample information

Following sample was sequenced and analyzed

Sample – SAM1, SAM2, SAM3, SAM4

1.2. Sample QC

Sample Name	Qubit (ng/µl)	RIN Value	QC Pass/Fail
SAM1	49	8	PASS
SAM2	53	8.2	PASS
SAM3	60	9	PASS
SAM4	58	8.7	PASS

RNA sequencing: TruSeq RNA Access Library Prep Kit

2. Analysis results

2.1. Data summary

The Human samples sequencing was done using Illumina HiSeq-2500 platform. TruSeq RNA Access Library Prep Kit was used to perform RNA sequencing. The steps in our bioinformatic analysis is described in **Appendix** A. For sample id or description refer to **Appendix B.**

Table 1: Data summary for RNA sequencing

RNA-seq samples	SAM1	SAM2	SAM3	SAM4
Total Reads	130,259,172	99,456,156	137,150,984	135,530,574
Total data (Gb)	13.03	9.95	13.72	13.55
Average read length (bp)	100	100	100	100
Avg. base quality (Phred)	37.88	37.96	37.69	37.92
Total data >= Q30 (%)	95.33	95.62	94.78	95.46
Total data >= Q20 (%)	97	97.2	96.61	97.08
GC (%)	50.84	49.56	50.29	49.65

2.2. Alignment & Filter

The read alignment summary for transcriptome data is provided in Table 2. Alignment using STAR (2.4.1) aligner is provided in the table. Reads mapping to ribosomal and mitochondrial genome were removed before performing alignment. Overall 96% of the total pre-processed reads mapped to the reference gene model/genome. The raw read counts were estimated using HTSeq-0.6.1. Read count data were normalized using DESeq2.

Table 2: Read alignment summary for RNA sequencing

RNA-seq data	SAM1	SAM2	SAM3	SAM4
Total Reads	130,259,172	99,456,156	137,150,984	135,530,574
# of Reads after adapter trimming	128,769,470	98,441,440	135,069,854	133,774,656

# of Reads after contamination removal	125,481,094	97,032,836	131,796,562	131,129,200
# of aligned reads	122,483,778	93,976,696	127,838,802	128,464,814
% of aligned reads	97.61	96.85	97	97.97

2.3 Hierarchical clustering and PCA analysis

Hierarchical clustering and PCA analysis is performed for normalized counts of all the protein coding genes. Euclidean distance and the complete linkage clustering method is used for hierarchical clustering. Analysis is performed using the R language and additional packages: ggplot2, reshape2 and ggrepel. Clustering and PCA analysis provided in figure 1a and 1b respectively.

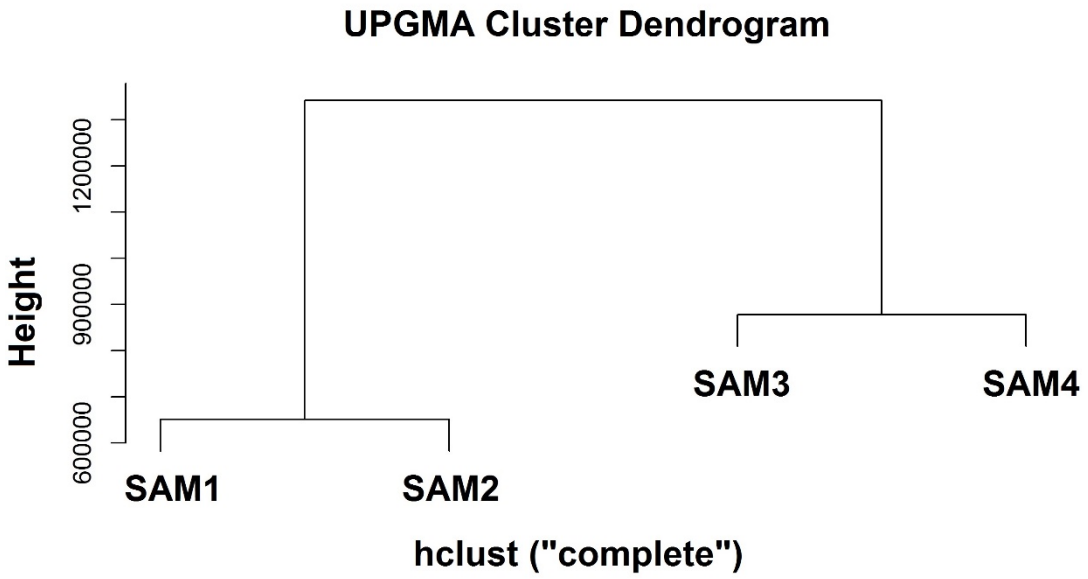

Figure 1a. Hierarchical clustering analysis

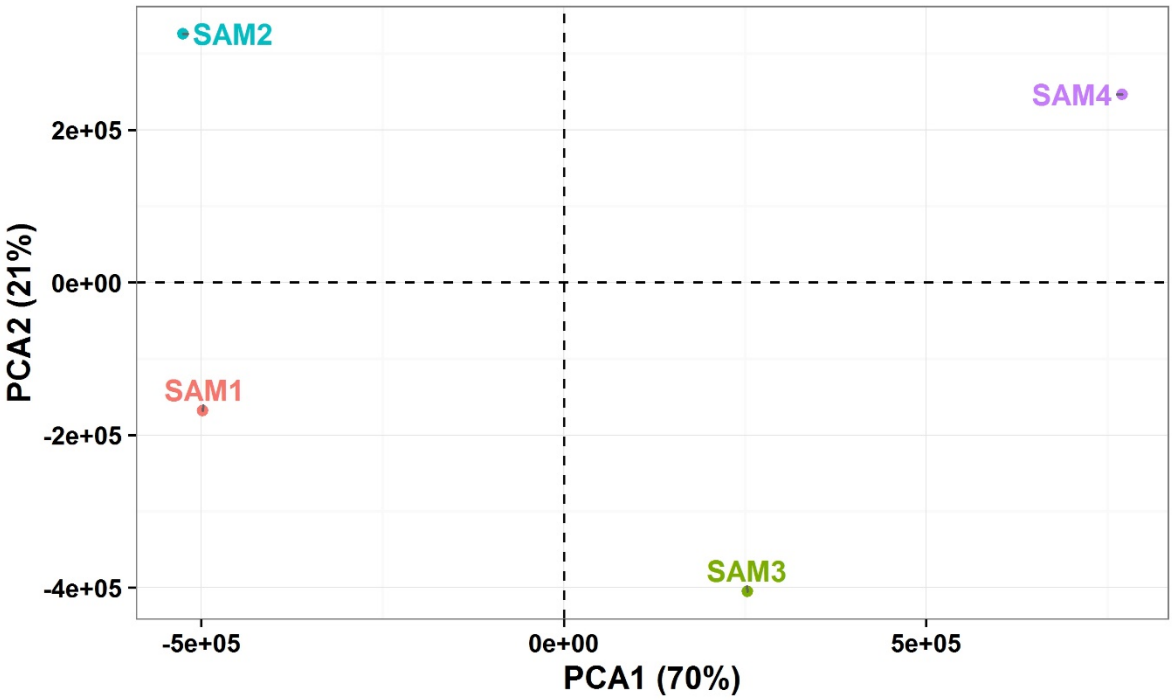

Figure 1b. Principal component analysis

2.4. Tumor microenvironment analysis

Tumor microenvironment analysis is performed using ssGSEA for all the protein coding genes at multiple levels i.e. Tumor content, innate immune cells and adaptive immune cells. Scores calculated for each level is provided in figure 2, 3 and table 3.

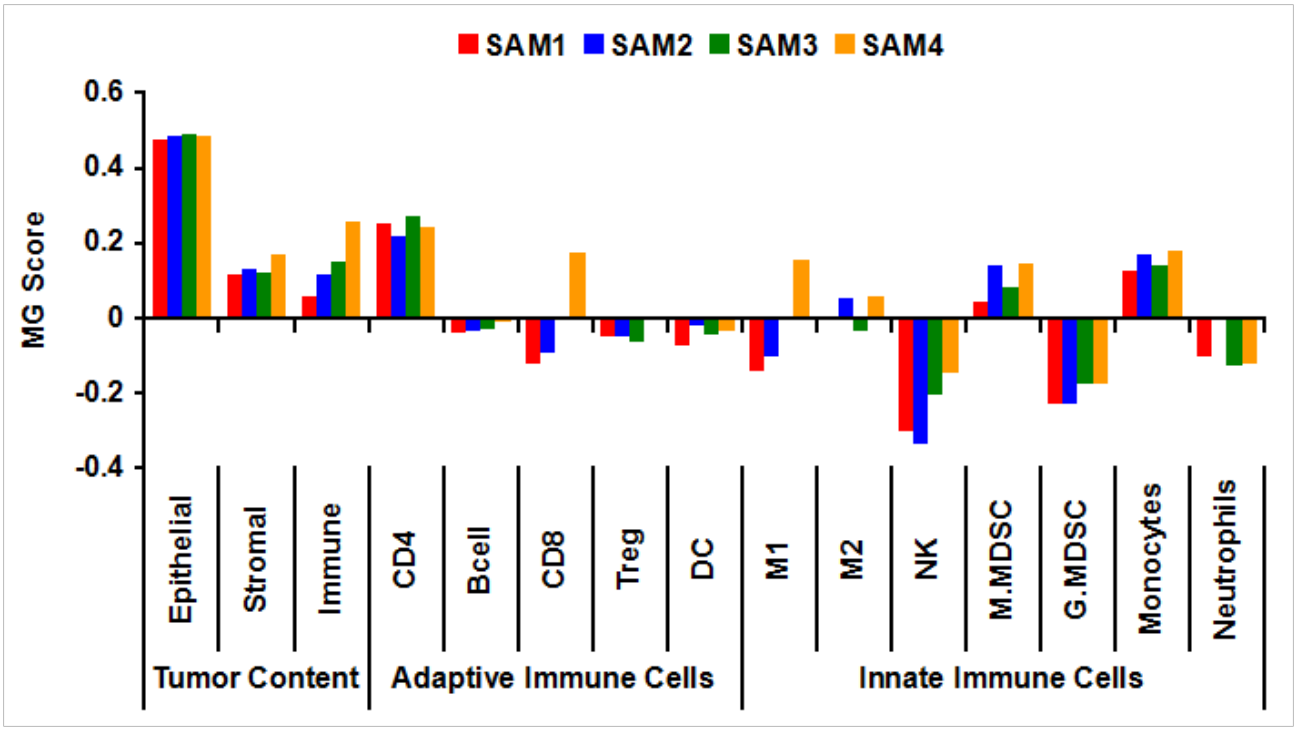

Figure 2. Infiltrated immune cells in the tumor microenvironment

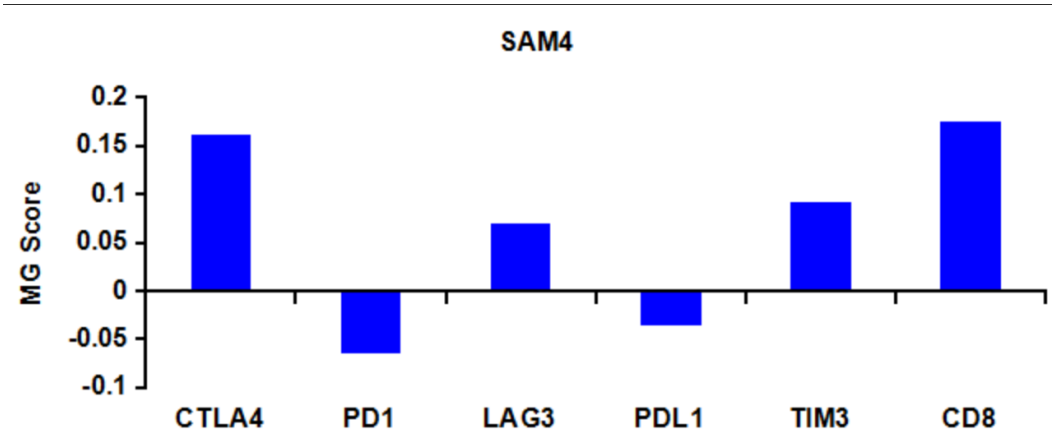

Figure 3. SAM4 has upregulated markers of anergic and exhausted CD8 T cells

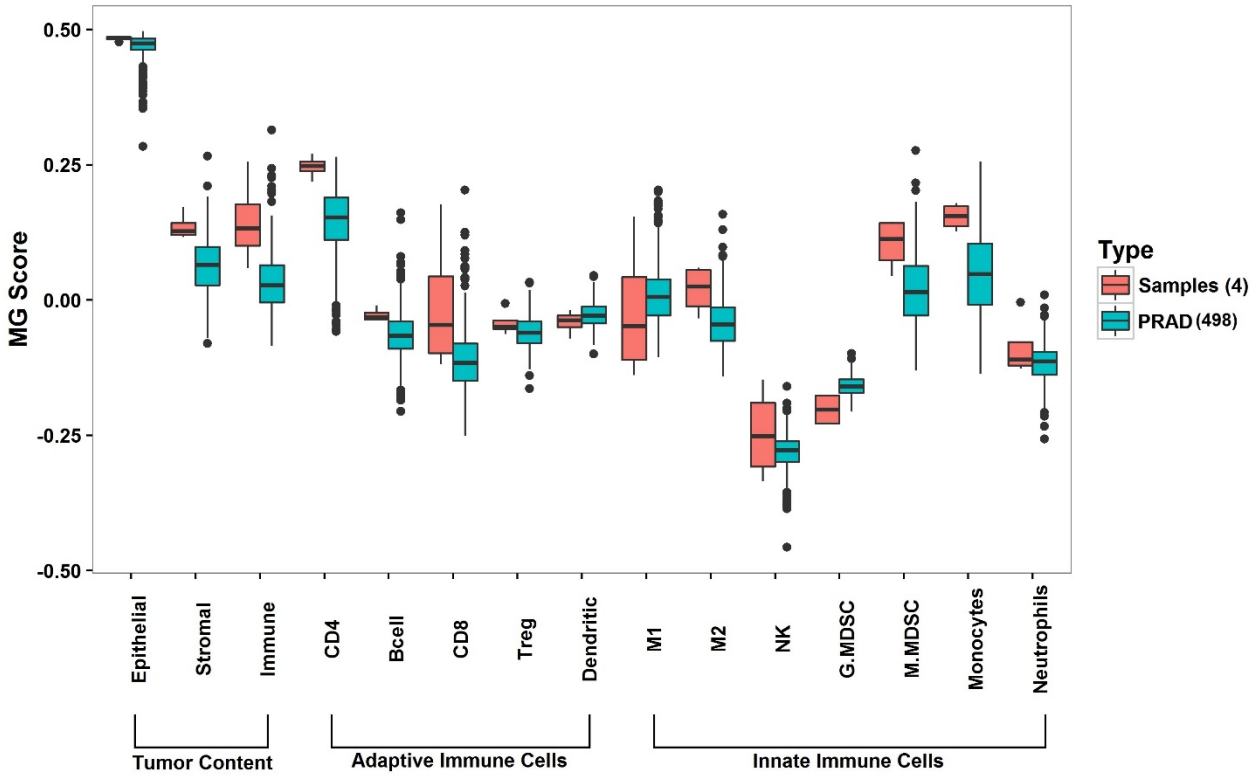

Figure 4. Tumor microenvironment analysis and comparison with TCGA Prostate adenocarcinoma.

Table 3. Score Distribution for different cell types

Score Distribution	Cell Type	SAM1	SAM2	SAM3	SAM4
Tumor Content	Epithelial	0.4768	0.4842	0.4888	0.4848
	Stromal	0.1162	0.1329	0.1214	0.1712
	Immune	0.0591	0.1141	0.1508	0.2555

Adaptive Immune Cells	CD4	0.2509	0.219	0.2706	0.2445
	B cell	-0.0371	-0.0353	-0.0286	-0.0098
	CD8	-0.1194	-0.0911	-0.0008	0.1759
	Treg	-0.0509	-0.0481	-0.0629	-0.0066
	DC	-0.0715	-0.0183	-0.0431	-0.0322
Innate Immune Cells	M1	-0.139	-0.1012	0.0046	0.1544
	M2	-0.004	0.054	-0.0342	0.0598
	NK	-0.2986	-0.3351	-0.2042	-0.1474
	M.MDSC	0.0442	0.1417	0.0832	0.1443
	G.MDSC	-0.2285	-0.228	-0.1767	-0.1749
	Monocytes	0.1267	0.1714	0.1394	0.179
	Neutrophils	-0.102	-0.0048	-0.1271	-0.1189

Conclusions

The tumor microenvironment analysis of the four samples indicates infiltration of both innate and adaptive immune cells (Figure 1). The epithelial and stromal content are comparable between the four samples. Tumor-4 has higher immune content than other tumors. This is reflected by infiltration of both adaptive and innate immune cells. Among the adaptive infiltrate, CD4 T cells are detected in all the tumors. Only Tumor-4 has high infiltration of CD8 T cells. We analyzed the functional state of CD8 T cells by correlating expression of CD8 signature with markers of anergy and exhaustion and show that CD8 T cells are functionally active. Further, high expression of CTLA-4 was detected in SAM4 samples supporting the anergic and exhausted state of CD8 T cells. The lower level of Treg cells in this tumor suggests that CTLA-4 expression may not

be associated with Treg compartment. The innate immune compartment is dominated by infiltration of MDSCs and monocytes in all tumors. Among the MDSCs, M-MDSC is more abundant in all the four tumors. Tumor-4 shows higher infiltration of M1 macrophages compared to other tumors. Tumor-4 is unlikely to benefit from the high infiltration of CD8 T cells because of their anergic and exhausted state.

We also compared the four tumors with prostate cancer from TCGA data. All the four prostate tumor samples in this analysis showed higher infiltration of innate and adaptive immune cells compared to median infiltration of immune cells in prostate tumors.

APPENDIX A

Bioinformatics analysis

The following bioinformatics steps were performed for analysis of the data (Figure 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) and cutadapt (version - 1.8dev).

Contamination removal - For the RNA-Seq analysis we begin by removing the unwanted sequences, especially nonpolyA tailed RNAs from the sample (assuming that poly-A tailed RNAs are sequenced). The unwanted sequences include – mitochon-drial genome sequences, ribosomal RNAs, transfer RNAs, adapter sequences and others. Contamination removal was performed using Bowtie2 (version - 2.2.4).

Read alignment – The paired-end reads are aligned to the reference human genome Feb. 2009 release downloaded from UCSC database (GRCh37/hg19).The chromosome fasta file was downloaded from the following website

(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz). GTF file was downloaded from the following website (ftp://ftp.ensembl.org/pub/release75/gtf/homo_sapiens/Homo_sapiens.GRCh37.75.gtf.gz). Alignment was performed using STAR (2.4.1).

Expression estimation – The aligned reads are used for estimating expression of the genes. The raw read counts were estimated using HTSeq-0.6.1. Read count data were normalized using DESeq2.

Hierarchical clustering and PCA analysis - Hierarchical clustering and PCA analysis is performed for normalized counts. Euclidean distance and the complete linkage clustering method is used for hierarchical clustering. Analysis is performed using the R language and additional packages: ggplot2, reshape2 and ggrepel.

Analysis of tumor microenvironment by OncoPept*TUME*™ -

The tumor microenvironment for the sample is analyzed at multiple levels. First tumor content analysis is done using expression of gene signatures associated with Epithelial, Stromal and Immune cells. Next the immune cell compartment is further stratified into following two different immune cell types Adaptive immune cells which are - CD8 T-cells, CD4 T-cells, T-regulatory cells, B-cells, Dendritic cells and Innate immune cells which are - M1 and M2 Macrophages, NK cells, Myeloid derived suppressor cells, Granulocytic myeloid-derived suppressor cells, Monocytes and Neutrophils.

The signature-based scoring was calculated using the ssGSEA (single cell Gene Set Enrichment Analysis) method. The score is a composite value calculated from the expression level of all genes in the signature. Scores are high if all the genes are coordinately regulated. The co-expression level of genes belonging to a signature defines the relative abundance of a specific cell type.

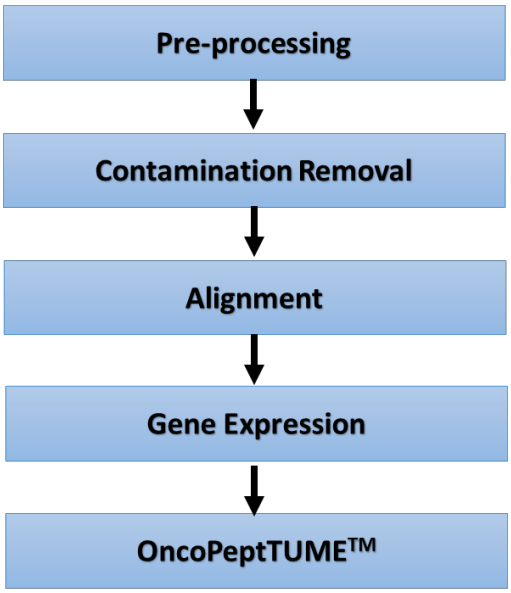

Figure 5: Bioinformatics analysis pipeline

APPENDIX B

SampleID	Sequencing type	Sample information
SAM1	RNA-Seq	RNA from SAM1
SAM2	RNA-Seq	RNA from SAM2
SAM3	RNA-Seq	RNA from SAM3
SAM4	RNA-Seq	RNA from SAM4

List of additional files

Table1. Raw_Normalized_Counts.xlsx – Raw and Normalized counts for each sample

QC_Report.pdf – Data quality control report