Characterization of Oncogenes in Pretreatment and Post-treatment Sequence Data

Samuel Chatmon

National Cancer Institute

Background

- Obtained Two SRA fastq files from the SRA database.
- Files were labeled as WES Nextgen Sequence data which were obtained from a melanoma patient before and after treatement
- Wanted to examine the sequences for the variant levels in Ocogenes

Project Overview

- Fastq files were obtained from NCBI's SRA database and transfer using the Fastq-dump tool
- The complete hg38 human genomic sequence was downloaded from the UCSC Genomic Browser Site
- The top twenty genes that contains somatic mutations for melanoma was obtained from the Cosmic Cancer Browser site and their chromosome number and position was obtained from the UCSC site
- Both sets of fastq sequences were aligned against the human genome with the BWA aligner

Project Overview (continued)

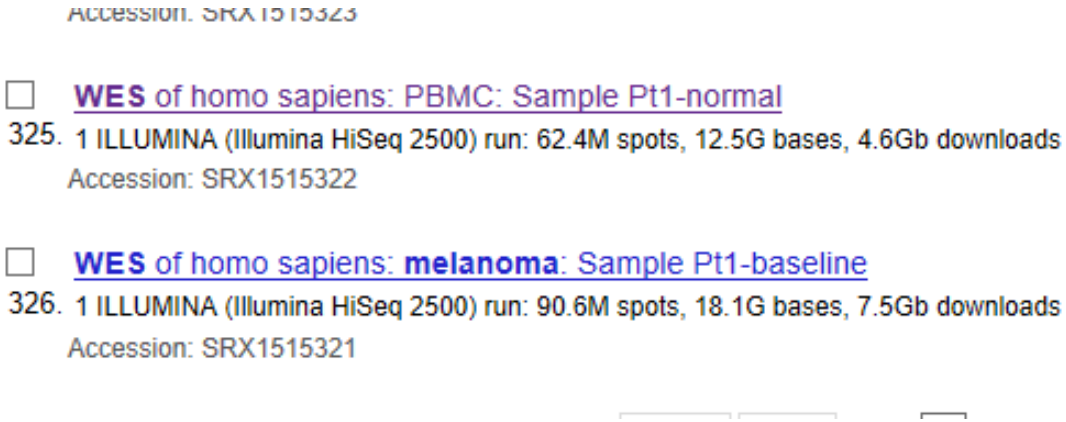
- Samtools and BCFtools were employed to create VCF files that the contained base substitutions, deletions and insertions in the data.
- Python programming and bash scripting were utilized to score the number of times variants were found in the genes

First Issue: Forgot about HIPAA Compliance

```
manager@bl8vbox[Desktop] fastq-dump SRR6431487                    [ 5:08PM]
2018-05-09T17:09:06 fastq-dump.2.9.0 err: query unauthorized while resolving que
ry within virtual file system module - failed to resolve accession 'SRR6431487'
- Access denied - please request permission to access phs001469/HMB-PUB-NPU-MDS
in dbGaP ( 403 )
2018-05-09T17:09:06 fastq-dump.2.9.0 err: item not found while constructing with
in virtual database module - the path 'SRR6431487' cannot be opened as database
or table
manager@bl8vbox[Desktop]                                          [ 5:09PM]
```

Fortunately, A Few Files were downloadable

Fortunately, a Few Files were downloadable.



Study: WES of patient derived pre-treatment metastatic melanoma on anti-PD-1 antibody treatment. We will call treatment srr8 and pre-treatment srr7 for short.

Second Issue Issue: Did not consider memory or CPU requirements

Bowtie Alignment ~12 hours in Oracles VirtualMachine ~12gb memory and 2 cpu's

Step 2: Choose an Instance Type

	Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance	IPv6 Support
	GPU graphics	g3.8xlarge	32	244	EBS only	Yes	10 Gigabit	Yes
	GPU instances	g2.8xlarge	32	60	2 x 120 (SSD)	-	10 Gigabit	-
	GPU compute	p2.8xlarge	32	488	EBS only	Yes	10 Gigabit	Yes
	GPU compute	p3.8xlarge	32	244	EBS only	Yes	10 Gigabit	Yes
	Memory optimized	r4.8xlarge	32	244	EBS only	Yes	10 Gigabit	Yes
	Memory optimized	x1e.8xlarge	32	976	1 x 960 (SSD)	Yes	Up to 10 Gigabit	Yes
	Storage optimized	i2.8xlarge	32	244	8 x 800 (SSD)	-	10 Gigabit	Yes
	Storage optimized	h1.8xlarge	32	128	4 x 2000	Yes	10 Gigabit	Yes

BWI Alignment ~3 hours in Amazon AWS. But Alas, ,most data on machine. Decided to continue virtual machine and use BWA instead.

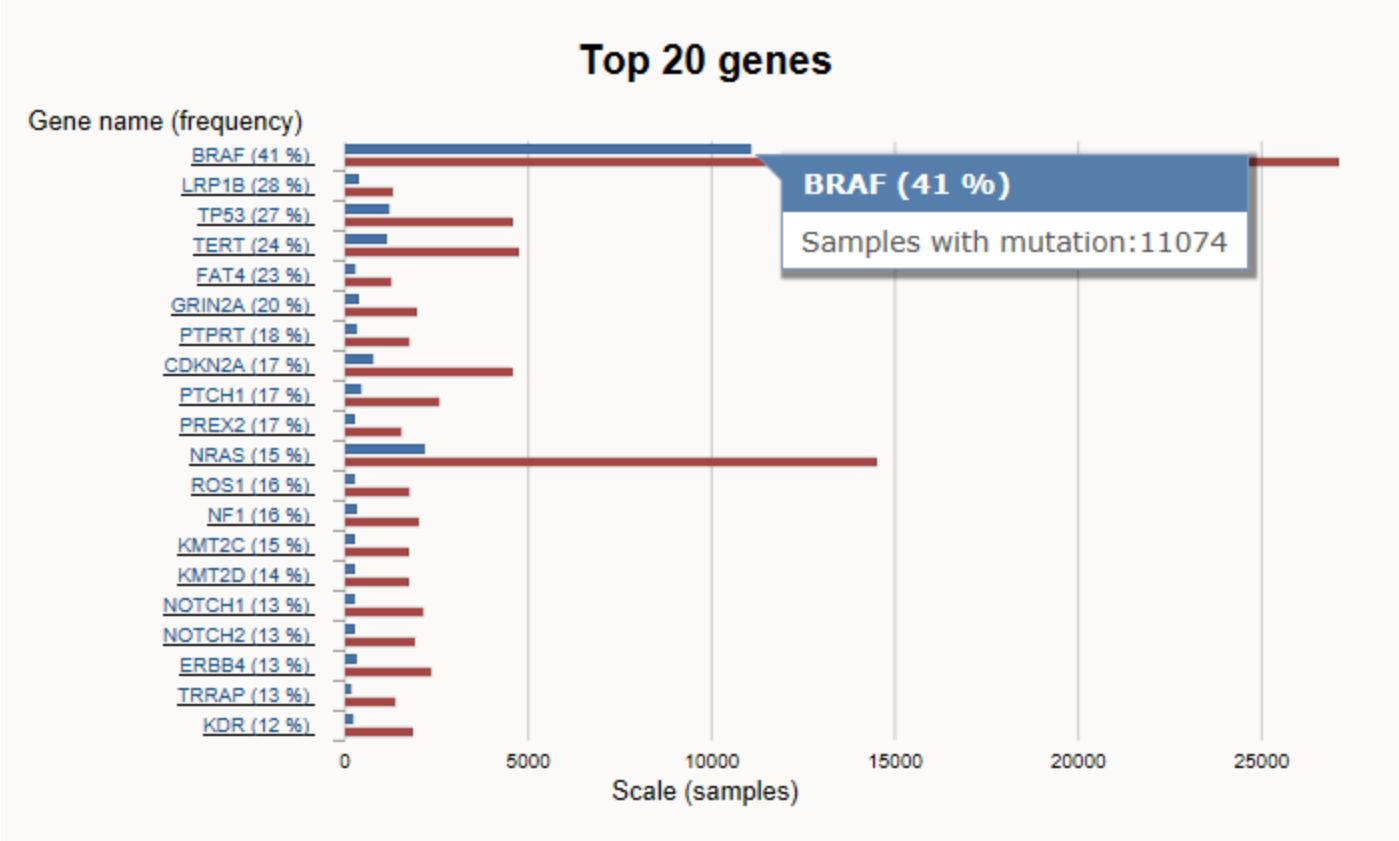
Finally can use Samtools to create bam file of alignments.

```
BF<FF<BB000BFB00BFB<<0<<<       NM:i:1  MD:Z:96A3       AS:i:96 XS:i:20
SRR3083838.3    83      chr16   52446097        60      100M    =       52446013        -184    TCTCTGAAAAACAGGGCATATGCTGACACTGGCTTCTGTGGCT
CATTGGGATCTTTCTTTTTCTTTTTCTTTGGAGTCTTGGGCTTCTTGCCAGAGTCTG       <FFBFBFBBBBFFFBBBBFFFFFFFFBBBFFFFFFFFFIIIIIIIFIIIIIIFFIFIBIFIIIIIIIIFFFFIIII
IFIFFIIFFBFFFFFFFFFFFFBBB       NM:i:0  MD:Z:100        AS:i:100        XS:i:23
SRR3083838.3    163     chr16   52446013        60      100M    =       52446097        184     AGGCTGTCCCACATGGATGCTACAATTTTTGAGACCTCTCCAA
AGGTTGCATTGGGGTTTTGACCTTTAATTGCAGCCTGTGTGTCTCTGAAAAACAGGG       BBBFFFFFFFBBBF00<BFFIIIIIIIIIIIFFIIIIIIIIIIIII7FFFBFFIIIIBFFIFFFFFFFFFFFFF
FFBFBBBBBBBFFFBBBFBB<BBFB       NM:i:1  MD:Z:14A85      AS:i:95 XS:i:0
SRR3083838.4    83      chr11   62617164        40      100M    =       62617033        -231    CTCCCGAAGCCGCTGGGCTTTGGGCGTGAGGACCACCAGGTGT
GTGAAGAGGAGGCCAGAGGCAGCCAGCAGCCCTGAGACCAGCGGGGTGGGACCCTCA       FFFBBFFBFFFBBFFFFFFFFFFFFFFFFFFFBFBBFFFFFFFFFFFFFIIIIFFFFFIIIIIIIIIIIFIFFFII
IIIFFIIIIIIIFFFFFFFFFFBBB       NM:i:0  MD:Z:100        AS:i:100        XS:i:100        XA:Z:chr3,+160452739,100M,0;
SRR3083838.4    163     chr11   62617033        60      100M    =       62617164        231     TAGACGACTCCTTGGGTCCCTGGTGGTGGTGGGTCCTTCTCCC
CACCCACAGCACCCCCTCTGCCCCGGAGCCAGTCCAGGGCCTTGTTCCGCTGCTCGA       BBBFFFFFFFFFFFI0<BFFIFIIFFFFFIBFFFFFFIIIFFIIIIIIIIIFBFFFFFFFBFBBFFFFBBBBFFFF
```

and BCFtools to create vcf file which contains all variants.

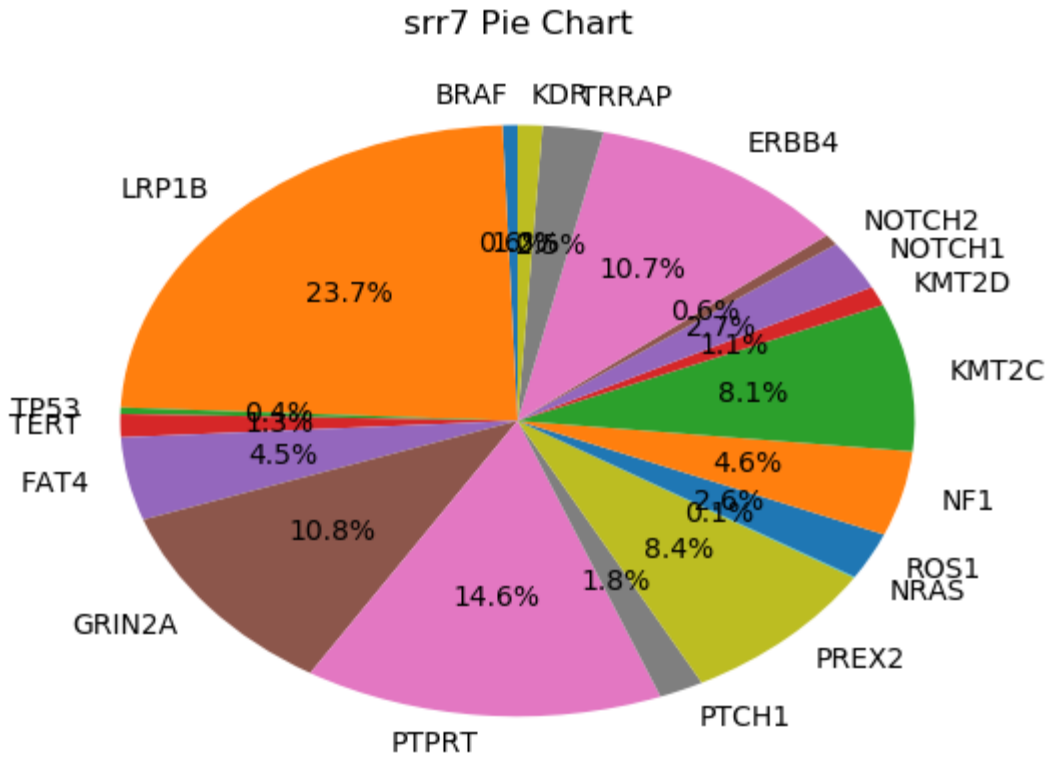
```
GT:PL   0/1:131,0,181
chr1    121104  .       C       G       4.29312 .       DP=10;VDB=0.14;SGB=-0.453602;RPB=0.1875;MQB=1;MQSB=1;B
B=0.3125;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,7,0,2;MQ=60       GT:PL   0/1:37,0,180
chr1    121191  .       C       A       5.71907 .       DP=3;SGB=-0.379885;RPB=1;MQB=1;MQSB=1;BQB=1;MQ0F=0.666
67;AC=2;AN=2;DP4=2,0,0,1;MQ=13  GT:PL   1/1:31,0,0
chr1    126113  .       C       A       3.22451 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=30
GT:PL   1/1:30,3,0
chr1    129285  .       G       A       80      .       DP=427;VDB=0.0145003;SGB=-0.693147;RPB=0.000820385;MQB
4.60393e-06;MQSB=0.00284594;BQB=0.207822;MQ0F=0.299766;ICB=1;HOB=0.5;AC=1;AN=2;DP4=99,130,109,44;MQ=14  GT:PL
0/1:113,0,159
chr1    133129  .       G       A       20.0638 .       DP=9;VDB=0.176409;SGB=-0.556411;RPB=0.406082;MQB=0.406
82;MQSB=0.714286;BQB=0.649731;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=3,2,4,0;MQ=34  GT:PL   0/1:53,0,124
chr1    134876  .       C       T       13.2351 .       DP=26;VDB=0.132453;SGB=-0.616816;RPB=0.557095;MQB=0.40
543;MQSB=0.569783;BQB=0.710273;MQ0F=0.5;ICB=1;HOB=0.5;AC=1;AN=2;DP4=9,10,4,2;MQ=13      GT:PL   0/1:46,0,96
chr1    135040  .       T       C       76      .       DP=71;VDB=0.337694;SGB=-0.691153;RPB=0.782415;MQB=0.98
025;MQSB=0.38374;BQB=0.321351;MQ0F=0.0704225;ICB=1;HOB=0.5;AC=1;AN=2;DP4=51,2,16,2;MQ=27        GT:PL   0/1:11
,0,156
chr1    136635  .       T       G       5.68565 .       DP=10;VDB=0.0389723;SGB=-0.511536;RPB=0.28717;MQB=0.57
341;MQSB=0.25;BQB=0.574341;MQ0F=0.5;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,2,3,0;MQ=17   GT:PL   0/1:37,0,20
chr1    136817  .       T       C       36.981  .       DP=9;VDB=0.348933;SGB=-0.556411;RPB=0.810265;MQB=0.810
65;MQSB=0.4;BQB=0.405132;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=3,0,2,2;MQ=25       GT:PL   0/1:70,0,29
chr1    138565  .       G       A       58      .       DP=57;VDB=0.726211;SGB=-0.683931;RPB=0.841869;MQB=0.72
74;MQSB=0.0637448;BQB=0.986986;MQ0F=0.0877193;ICB=1;HOB=0.5;AC=1;AN=2;DP4=33,8,6,7;MQ=24        GT:PL   0/1:92
--More--
```

Obtained 20 more frequent Mutated Genes from Cosmic Site



Utilized Python and BASH scripting to score variants in each data set : List of Genes and Pie Chart for srr7

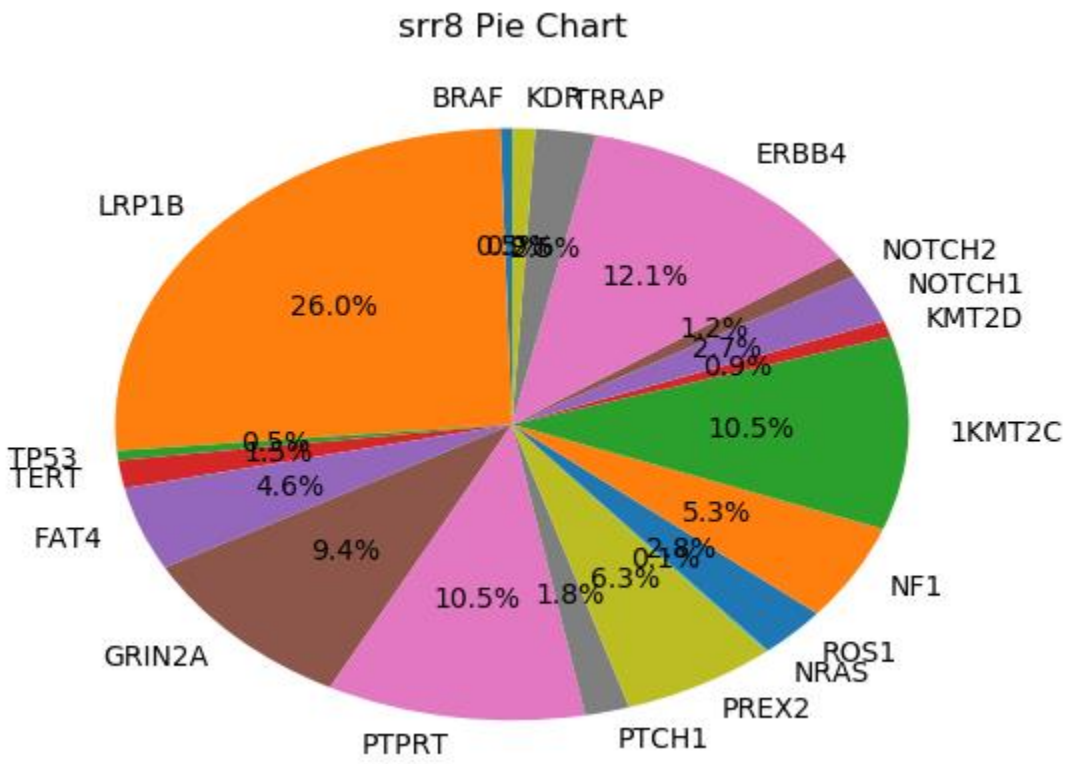
srr7	
Number	**Gene Name**
1	BRAF
2	ERBB4
3	FAT4
4	GRIN2A
5	KDR
6	KMT2C
7	KMT2D
8	LRP1B
9	NF1
10	NOTCH1
11	NOTCH2
12	NRAS
13	PREX2
14	PTCH1
15	PTPRT
16	ROS1
17	TERT
18	TP53
19	TRRAP



Total number of variants = 1985

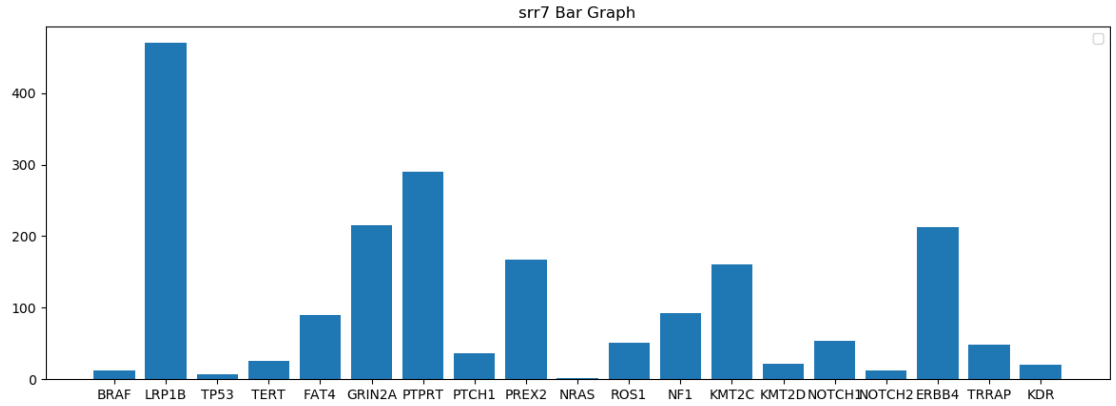
Utilized Python and BASH scripting to score variants in each data set : List of Genes and Pie Chart for srr8

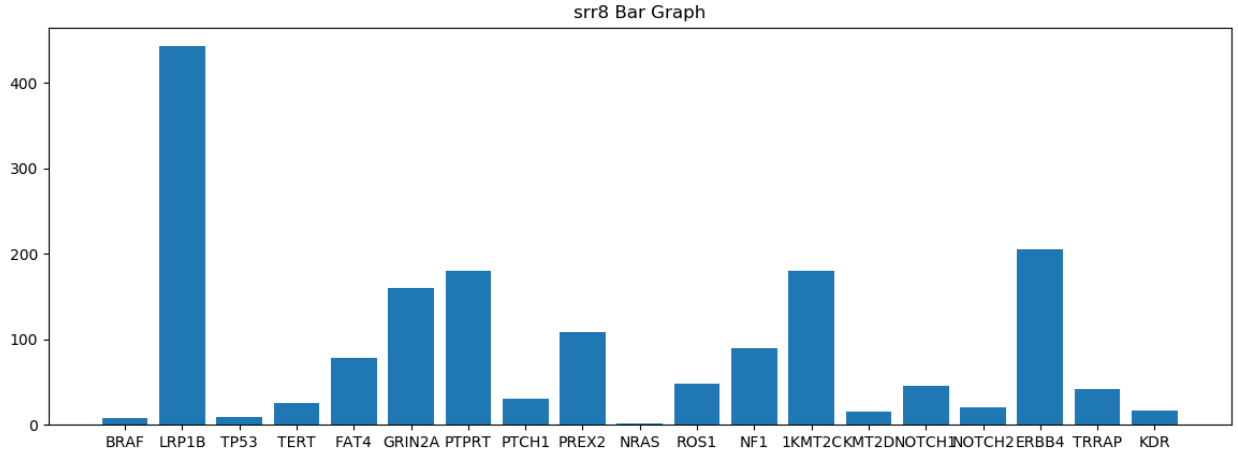
srr8	
Number	**Gene Name**
1	BRAF
2	ERBB4
3	FAT4
4	GRIN2A
5	KDR
6	KMT2C
7	KMT2D
8	LRP1B
9	NF1
10	NOTCH1
11	NOTCH2
12	NRAS
13	PREX2
14	PTCH1
15	PTPRT
16	ROS1
17	TERT
18	TP53
19	TRRAP



Total number of variants = 1707

Utilized Python and BASH scripting to score variants in each data set : Bar Graphs for srr7 and srr8





Conclusion

- Can’t really hypothesize anything by just using simple visualization tools.
- Most likely would need to look at much more data and using Transcriptomics tools and star alignment tool