## Week 2 Homework

## 2.COVID-19 UCSC genome browser and bedtools:

- a. Get BED files from NCBI genes annotations, microdeletions, variants of concern
- b. Find which genes have the most microdeletions and variants of concern (intersectBed)

```
(base) anamkidwai@anamkidsmacbook Desktop % cut -f 4 microdeletions output.bed | sort | uniq -c | sort -nr | awk
'{print "chr1", $2, $2+1, "gene_"$2, $1}' > top_genes_microdeletions.bed
(base) anamkidwai@anamkidsmacbook Desktop % cut -f 4 variants of concern output.bed | sort | uniq -c | sort -nr |
awk '{print "chr1", $2, $2+1, "gene_"$2, $1}' > top_genes_variants_of_concern.bed
(base) anamkidwai@anamkidsmacbook Desktop % cat top genes microdeletions.bed
command line uses the intersectBed tool to find the overlapping regions between the "covid19.bed" file containing
NCBI genes annotations and the "covid.part2" file containing microdeletions.
chr1 ORF1ab 1 gene ORF1ab 84
chr1 ORF1a 1 gene ORF1a 64
chr1 S 1 gene S 15
chr1 ORF3a 1 gene ORF3a 11
chr1 ORF7b 1 gene ORF7b 6
chr1 ORF7a 1 gene ORF7a 6
chr1 E 1 gene E 6
chr1 ORF8 1 gene ORF8 5
chr1 N 1 gene N 3
(base) anamkidwai@anamkidsmacbook Desktop % cat top genes variants of concern.bed
command line uses the intersectBed tool to find the overlapping regions between the "covid19.bed" file containing
NCBI genes annotations and the "covid.part3" file containing variants of concern.
chr1 S 1 gene S 9
chr1 N 1 gene N 6
chr1 ORF1ab 1 gene ORF1ab 5
chr1 ORF1a 1 gene ORF1a 4
chr1 ORF8 1 gene_ORF8 3
```

```
(base) anamkidwai@anamkidsmacbook Desktop % cut -f 4 microdeletions_output.bed | sort | uniq -c | sort -nr | awk '(print "chr1", $2, $2+1, "gene_"$2, $1)' > top_genes_microdeletions.bed

(base) anamkidwai@anamkidsmacbook Desktop % cut -f 4 variants_of_concern_output.bed | sort | uniq -c | sort -nr | awk '(print "chr1", $2, $2+1, "gene_"$2, $1)' > top_genes_variants_of_concern.bed

chr1 ORFia 1 gene_ORFia 84

chr1 SR 1 gene_ORFia 64

chr1 SR 1
```

## c. Find the closest microdeletion to each gene (closestBed).

```
(base) anamkidwai@anamkidsmacbook Desktop % closestBed -a ~/Desktop/covid19.bed -b ~/Desktop/covid.part2 -io > closest_microdeletion_to_genes_no_overlap.bed (base) anamkidwai@anamkidsmacbook Desktop % cat closest_microdeletion_to_genes_no_overlap.bed NC_045512v2 265 13483 ORF1a 0 + 265 13483 0 1 13218, 0, NC 045512v2 262 263 1b 300 +
```

NC_045512v2	265	21555	ORF1ab	0	+	265	21555	0	2	13203,80	88,
0,13202, NC_0455	512v2	262	263	1b	300	+					
NC_045512v2	21562	25384	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	21561	21562	1b	600	+						
NC_045512v2	25392	26220	ORF3a	0	+	25392	26220	0	1	828,	0,
NC_045512v2	26293	26305	12b	300	+						
NC_045512v2	26244	26472	E	0	+	26244	26472	0	1	228,	0,
NC_045512v2	26486	26487	1b	300	+						
NC_045512v2	26522	27191	M	0	+	26522	27191	0	1	669,	0,
NC_045512v2	26501	26502	1b	300	+						
NC_045512v2	27201	27387	ORF6	0	+	27201	27387	0	1	186,	0,
NC_045512v2	27403	27484	81b	300	+						
NC_045512v2	27393	27759	ORF7a	0	+	27393	27759	0	1	366,	0,
NC_045512v2	27784	27785	1b	300	+						
NC_045512v2	27755	27887	ORF7b	0	+	27755	27887	0	1	132,	0,
NC_045512v2	27701	27741	40b	300	+						
NC_045512v2	27893	28259	ORF8	0	+	27893	28259	0	1	366,	0,
NC_045512v2	27836	27837	1b	300	+						
NC_045512v2	28273	29533	N	0	+	28273	29533	0	1	1260,	0,
NC_045512v2	28257	28258	1b	300	+						
NC_045512v2	29557	29674	ORF10	0	+	29557	29674	0	1	117,	0,
NC_045512v2	28894	28903	9b	300	+						

Dase) allalikti	dwai@anam	kidsmacbo	ook Deskt	ор % с	at close	st_microd	eletion_	to_gene	es_no_ove	erlap.bed									
C 045512v2	265	13483	ORF1a	а		265	13483	А		13218,		NC 045512v2	262	263	1b	300			
C_045512V2 C 045512V2	265	21555	ORF1ab		-	265	21555	a	2	13218,		0.13202.	262 NC 045		262	263	1b	300	
C_045512v2	21562	25384	OKLIAD	9	-	21562	25384	a	1	3822,		NC_045512v2	21561	21562	1b	600		300	
			5						1		0,						+		
C_045512v2	25392	26220	ORF3a	0		25392	26220	0	1	828,	0,	NC_045512v2	26293	26305	12b	300	+		
C_045512v2	26244	26472	E	0		26244	26472	0	1	228,	0,	NC_045512v2	26486	26487	1b	300			
C_045512v2	26522	27191	М	0		26522	27191	0	1	669,	0,	NC_045512v2	26501	26502	1b	300			
C 045512v2	27201	27387	ORF6	0		27201	27387	0	1	186,	0,	NC 045512v2	27403	27484	81b	300			
C 045512v2	27393	27759	ORF7a	0		27393	27759	0	1	366,	ø,	NC 045512v2	27784	27785	1b	300			
C 045512v2	27755	27887	ORF7b	0		27755	27887	0	1	132,	0,	NC 045512v2	27701	27741	40b	300			
C 045512v2	27893	28259	ORF8	0		27893	28259	0	1	366.	0,	NC 045512v2	27836	27837	1b	300			
C 045512v2	28273	29533	N	0	+	28273	29533	0	1	1260,	0,	NC 045512v2	28257	28258	1b	300	+		
C 045512v2	29557	29674	ORF10	a	+	29557	29674	0	1	117.	ø,	NC 045512v2	28894	28903	9b	300	+		

## d. Get the parts of the genome that do NOT have a variant of concern.

 $(base)\ anamkidwai@anamkidsmacbook\ Desktop\ \%\ subtractBed\ -a\ \sim/Desktop/covid19.bed\ -b\ \sim/Desktop/covid.part3 > genome\_without\_variants\_of\_concern.bed$ 

		_	_	_						
(base) anamkid	lwai@ar	amkidsı	nacbook I	Desktop	% cat g	genome_	_without	_variant	s_of_co	ncern.bed
NC_045512v2	265	3265	ORF1a (	0	+	265	13483	0	1	13218, 0,
NC_045512v2	3268	5386	ORF1a (	0	+	265	13483	0	1	13218, 0,
NC_045512v2	5389	6952	ORF1a (	0	+	265	13483	0	1	13218, 0,
NC_045512v2	6955	11287	ORF1a (	0	+	265	13483	0	1	13218, 0,
NC_045512v2	11296	13483	ORF1a (	0	+	265	13483	0	1	13218, 0,
NC_045512v2	265	3265	ORF1ab (	0	+	265	21555	0	2	13203,8088,
0,13202,										
NC_045512v2	3268	5386	ORF1ab (	0	+	265	21555	0	2	13203,8088,
0,13202,										
NC_045512v2	5389	6952	ORF1ab (	0	+	265	21555	0	2	13203,8088,
0,13202,										
NC_045512v2	6955	11287	ORF1ab (	0	+	265	21555	0	2	13203,8088,
0,13202,										

NC_045512v2	11296	14406	ORF1ab	0 0	+	265	21555	0	2	13203,8	3088,
0,13202,	1.4.400	01555	ODE1 1	0		265	21555	0	2	12202 (	2000
NC_045512v2	14409	21555	ORF1at	0	+	265	21555	0	2	13203,8	3088,
0,13202,			_								
NC_045512v2	21562	21764	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	21770	21990	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	21993	23062	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	23065	23269	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	23272	23401	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	23404	23602	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	23605	23707	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	23710	24505	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	24508	24913	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	24916	25384	S	0	+	21562	25384	0	1	3822,	0,
NC_045512v2	25392	26220	ORF3a	0	+	25392	26220	0	1	828,	0,
NC_045512v2	26244	26472	E	0	+	26244	26472	0	1	228,	0,
NC_045512v2	26522	27191	M	0	+	26522	27191	0	1	669,	0,
NC_045512v2	27201	27387	ORF6	0	+	27201	27387	0	1	186,	0,
NC 045512v2	27393	27759	ORF7a	0	+	27393	27759	0	1	366,	0,
NC_045512v2	27755	27887	ORF7b	0	+	27755	27887	0	1	132,	0,
NC 045512v2	27893	27971	ORF8	0	+	27893	28259	0	1	366,	0,
NC_045512v2	27974	28046	ORF8	0	+	27893	28259	0	1	366,	0,
NC_045512v2	28049	28109	ORF8	0	+	27893	28259	0	1	366,	0,
NC_045512v2	28112	28259	ORF8	0	+	27893	28259	0	1	366,	0,
NC_045512v2	28273	28279	N	0	+	28273	29533	0	1	1260,	0,
NC_045512v2	28282	28879	N	0	+	28273	29533	0	1	1260,	0,
NC_045512v2	28885	28975	N	0	+	28273	29533	0	1	1260,	0,
NC 045512v2	28978	29533	N	0	+	28273	29533	0	1	1260,	0,
NC 045512v2	29557	29674	ORF10		+	29557	29674	0	1	117,	0,
	'					'		-		- ,	,

pase) anamki	dwai@anam	kidsmacb	ook Deskt	ор % с	at genom	e_without	_variants	_of_c	oncern.be	ed		
C 045512v2	265	3265	ORF1a	0		265	13483	0	1	13218,	0,	
C_045512v2	3268	5386	ORF1a	õ		265	13483	ě			ø,	
C 045512v2	5389	6952	ORF1a	0		265	13483	ø	1	13218,	0,	
C 045512v2	6955	11287	ORF1a	0		265	13483	0	1	13218,	0,	
C 045512v2	11296	13483	ORF1a	0		265	13483	0	1	13218,	0,	
C_045512v2	265	3265	ORF1ab	0		265	21555	0	2	13203,80	988,	0,13202,
C 045512v2	3268	5386	ORF1ab	0		265	21555	ø	2	13203,80		0,13202,
C 045512v2	5389	6952	ORF1ab	0		265	21555	0	2	13203,80		0,13202,
C 045512v2	6955	11287	ORF1ab	0		265	21555	0		13203,80	988,	0,13202,
C_045512v2	11296	14406	ORF1ab	0		265	21555	0	2	13203,80	988,	0,13202,
C 045512v2	14409	21555	ORF1ab	0		265	21555	0	2	13203,80	988.	0,13202,
IC 045512v2	21562	21764	s	0		21562	25384	0		3822.	0,	
IC 045512v2	21770	21990		0		21562	25384	0		3822,	0,	
C_045512v2	21993	23062	s	0		21562	25384	0	1	3822,	0,	
C_045512v2	23065	23269	S	0		21562	25384	0	1	3822,	0,	
IC 045512v2	23272	23401	s	0		21562	25384	0		3822,	0,	
IC 045512v2	23404	23602		0		21562	25384	0		3822,	0,	
IC_045512v2	23605	23707	s	0		21562	25384	0	1	3822,	0,	
IC_045512v2	23710	24505	S	0		21562	25384	0	1	3822,	0,	
IC_045512v2	24508	24913	S	0		21562	25384	0		3822,	0,	
IC 045512v2	24916	25384		0		21562	25384	0		3822,	0,	
IC_045512v2	25392	26220	ORF3a	0		25392	26220	0		828,	0,	
IC_045512v2	26244	26472	E	0		26244	26472	0	1	228,	0,	
IC 045512v2	26522	27191		0		26522	27191	0		669,	0,	
IC_045512v2	27201	27387	ORF6	0		27201	27387	0		186,	0,	
IC_045512v2	27393	27759	ORF7a	0		27393	27759	0		366,	0,	
IC_045512v2	27755	27887	ORF7b			27755	27887			132,	0,	
IC_045512v2	27893	27971	ORF8	0		27893	28259	0		366,	0,	
IC_045512v2	27974	28046	ORF8			27893	28259			366,	0,	
IC_045512v2	28049	28109	ORF8	0		27893	28259	0		366,	0,	
IC_045512v2	28112	28259	ORF8			27893	28259			366,	0,	
C_045512v2	28273	28279				28273	29533			1260,	0,	
C_045512v2	28282	28879		0		28273	29533	0		1260,	0,	
C_045512v2	28885	28975		ø		28273	29533	ō		1260,	ø,	
C_045512v2	28978	29533	N	0		28273	29533	0	1	1260,	0,	
IC 045512v2	29557	29674	ORF10	0		29557	29674	ē	1	117.	0,	