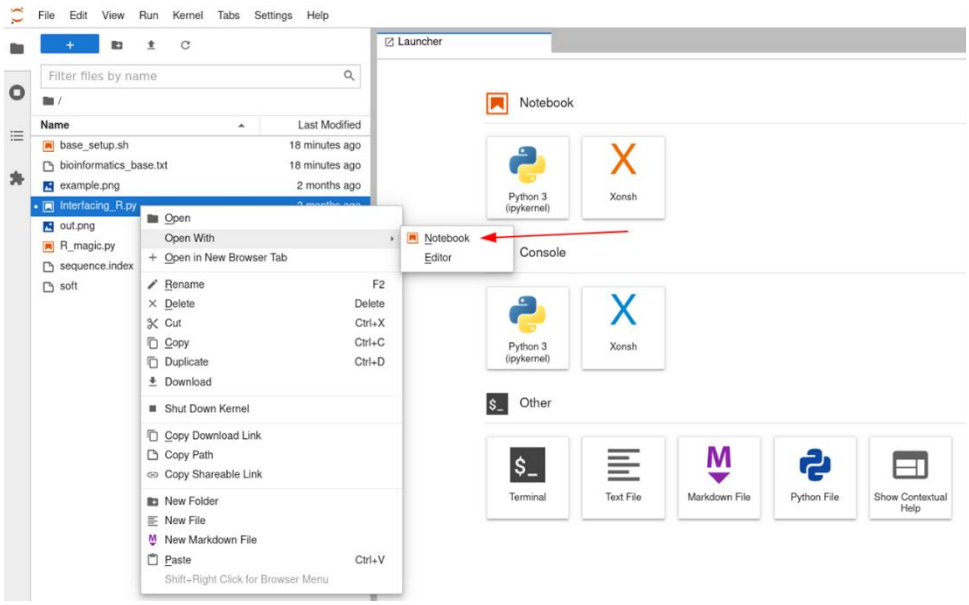
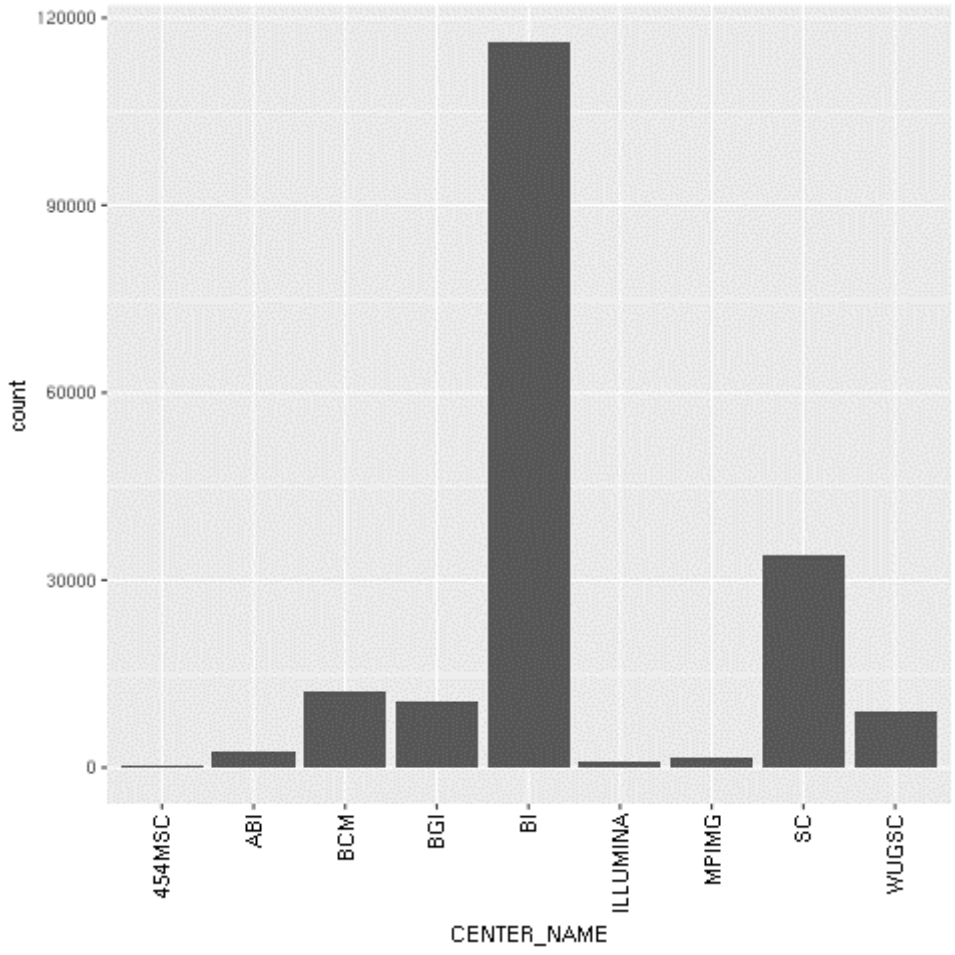
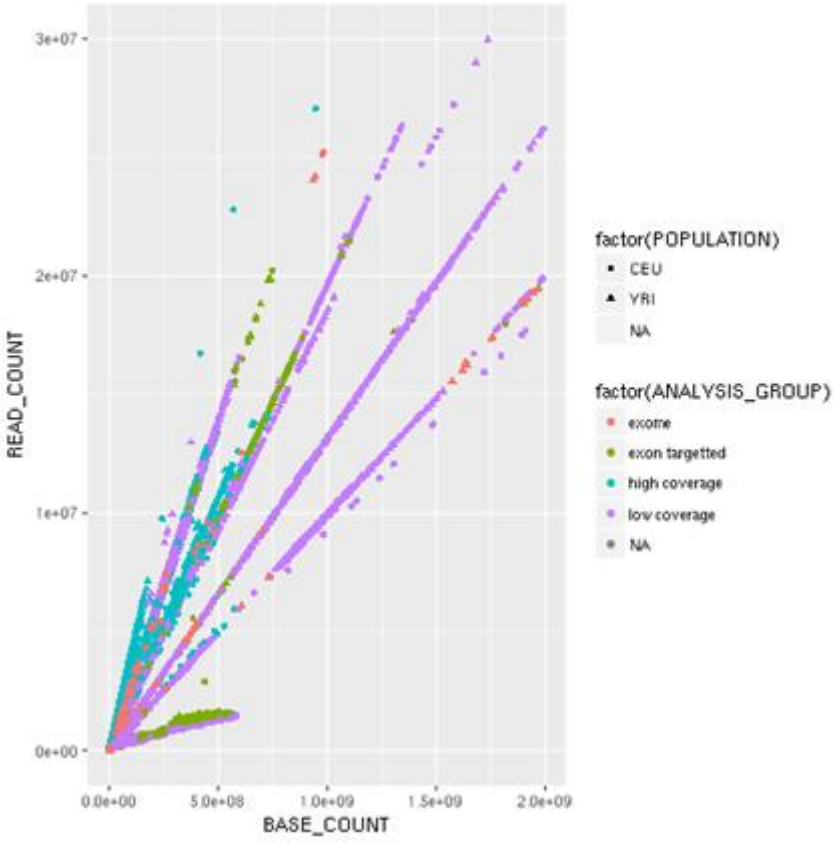
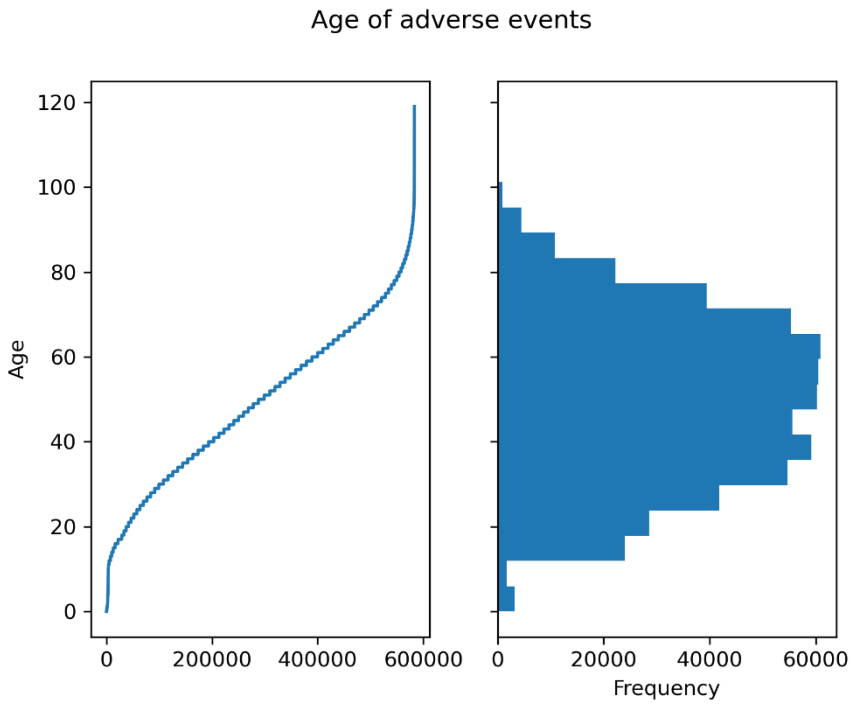
Chapter 1: Python and the Surrounding Software Ecology






3e+07
2e+07
1e+07
0e+00
READ_COUNT
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
BASE_COUNT
factor(POPULATION)
CEU
YRI
NA
factor(ANALYSIS_GROUP)
exome
exon targetted
high coverage
low coverage
NA

Chapter 2: Using Data Processing Libraries: numpy, pandas, arrow, and zarr

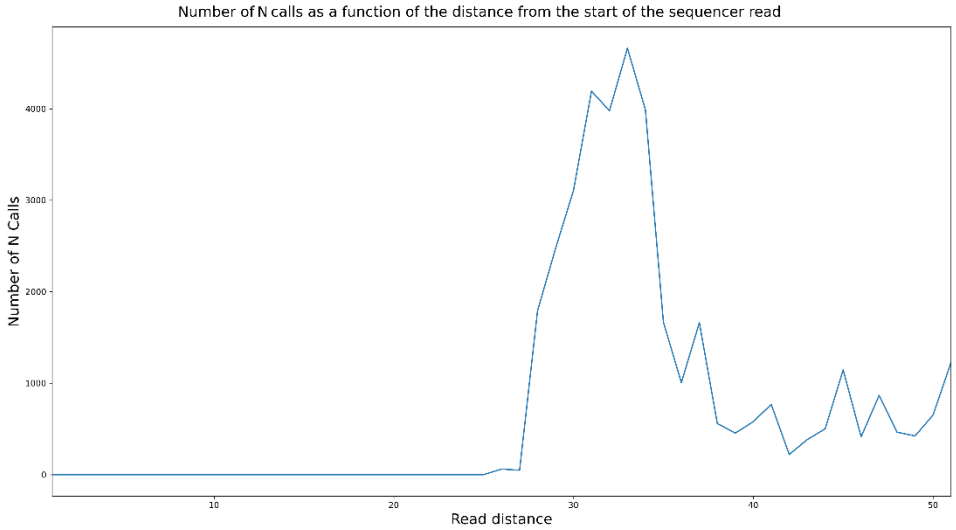


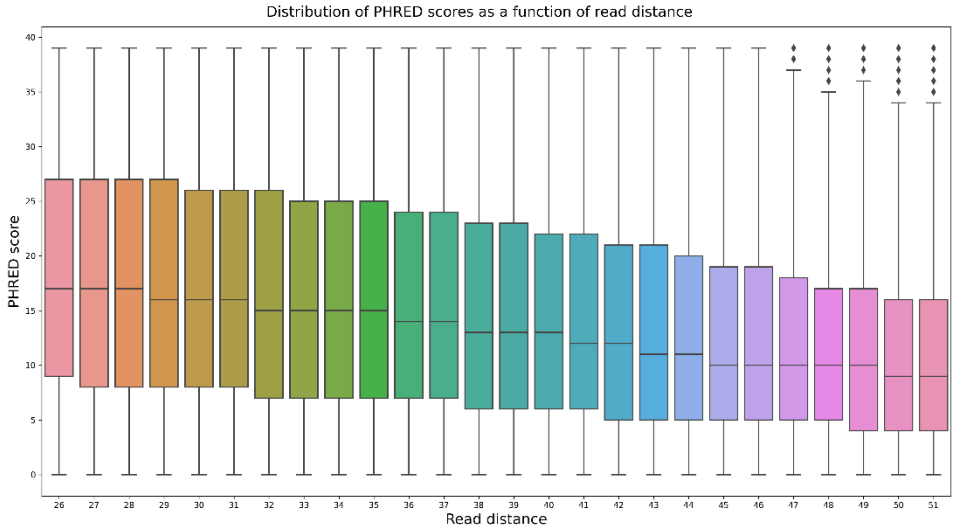
[21]:

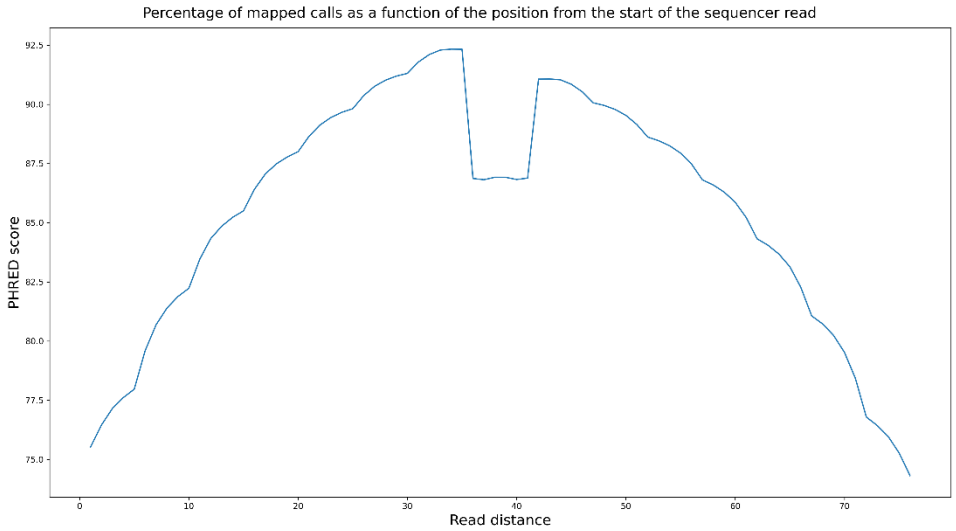
STATE	size	rank
CA	62821	0
FL	38209	1
TX	36512	2
NY	34921	3
PA	23646	4

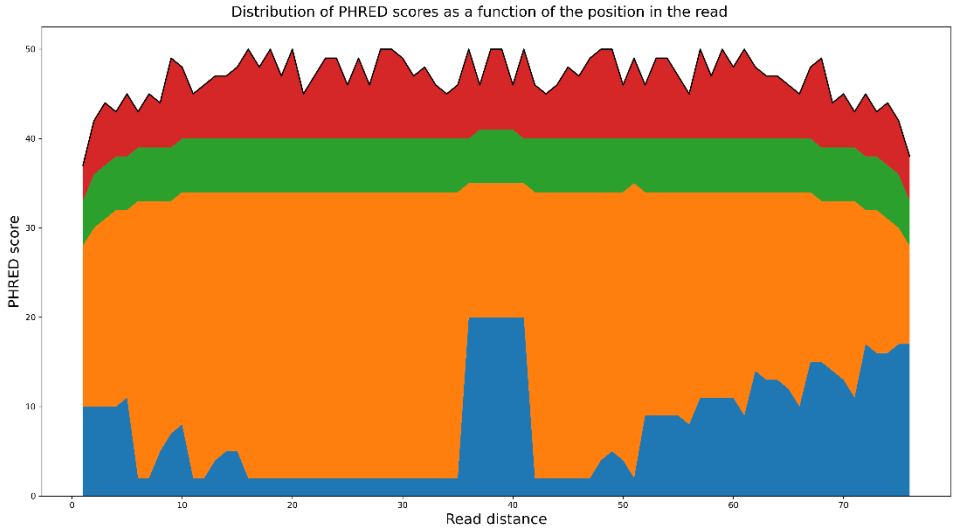
```
array([[ 7, 28, 34, 26,  4,  0],
       [ 3, 19, 30, 39,  7,  0],
       [ 7, 28, 35, 24,  3,  0],
       [ 6, 28, 33, 27,  4,  0],
       [ 5, 26, 34, 29,  4,  0]], dtype=uint8)
```

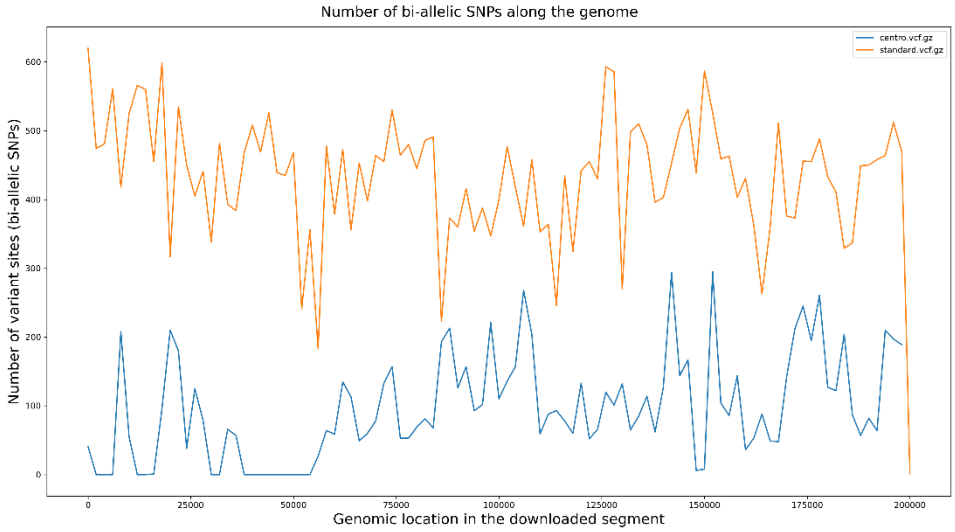
Chapter 3: Next Generation Sequencing

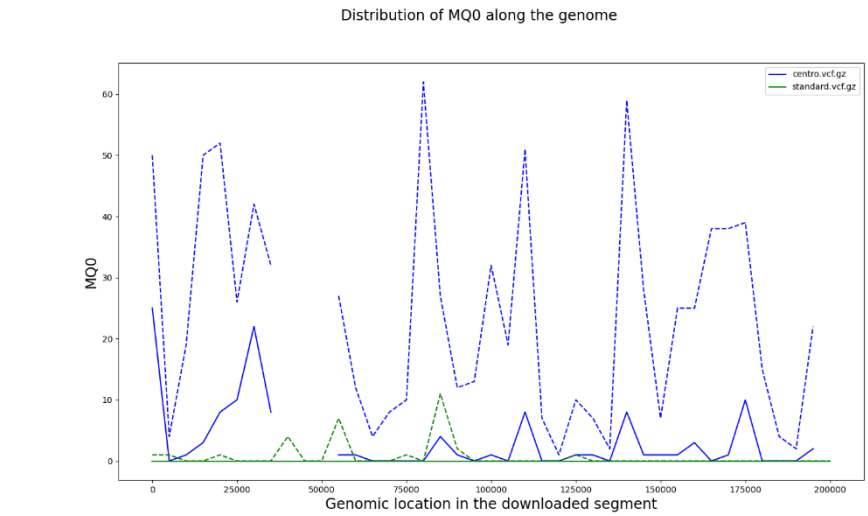


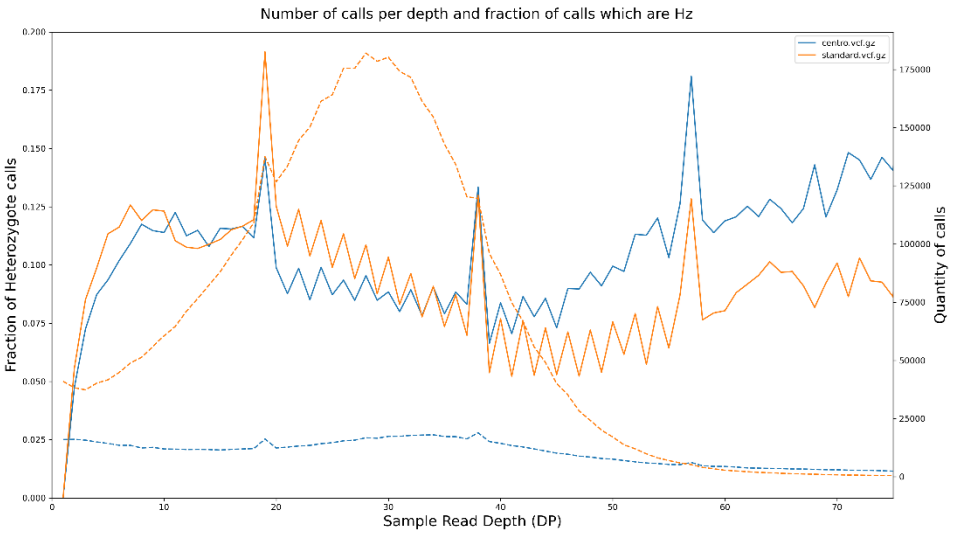


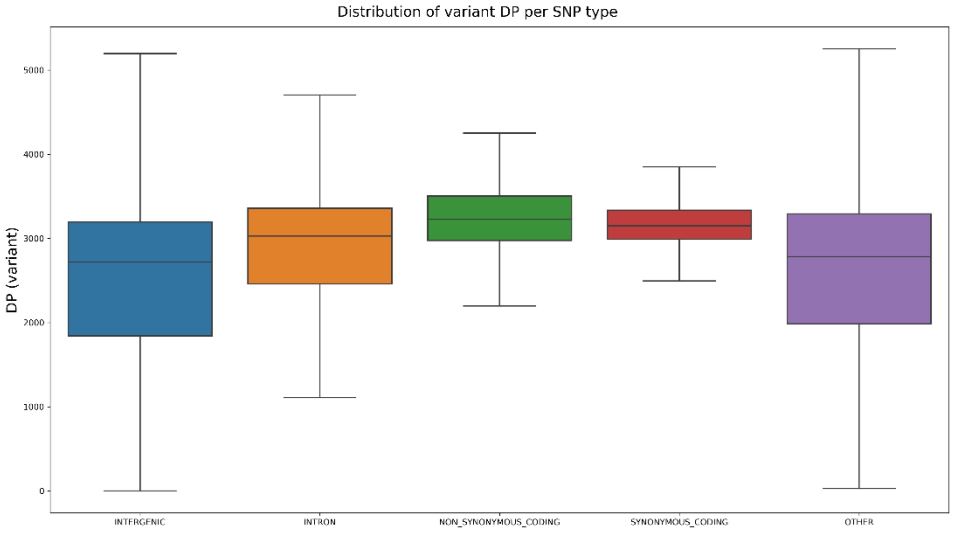



Distribution of PHRED scores as a function of the position in the read
PHRED score
Read distance


Number of bi-allelic SNPs along the genome
centro.vcf.gz
standard.vcf.gz
Number of variant sites (bi-allelic SNPs)
Genomic location in the downloaded segment

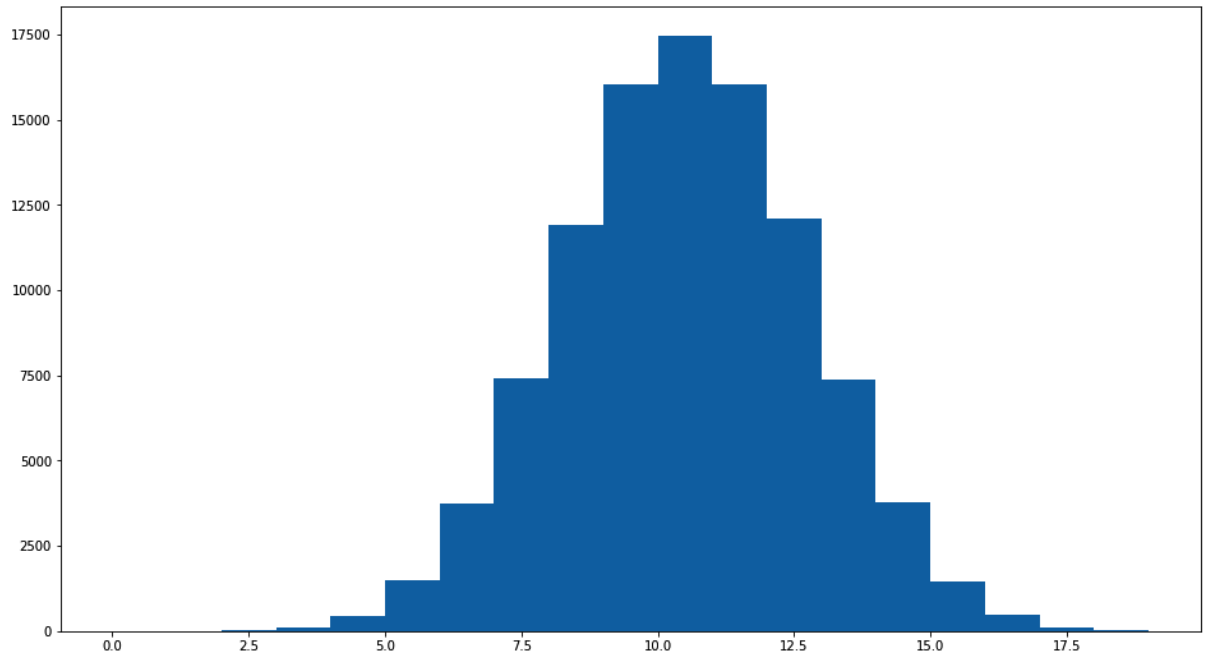

Distribution of MQ0 along the genome
centro.vcf.gz
standard.vcf.gz
MQ0
Genomic location in the downloaded segment

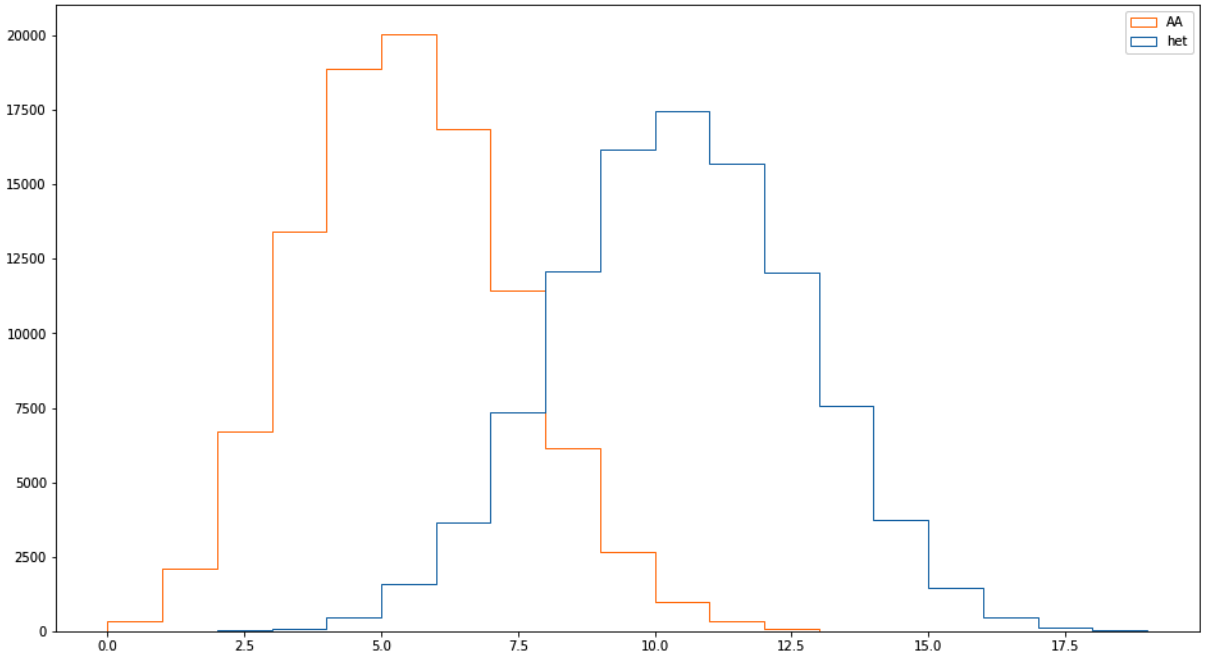

Number of calls per depth and fraction of calls which are Hz
centro.vcf.gz
standard.vcf.gz
Fraction of Heterozygote calls
Quantity of calls
Sample Read Depth (DP)

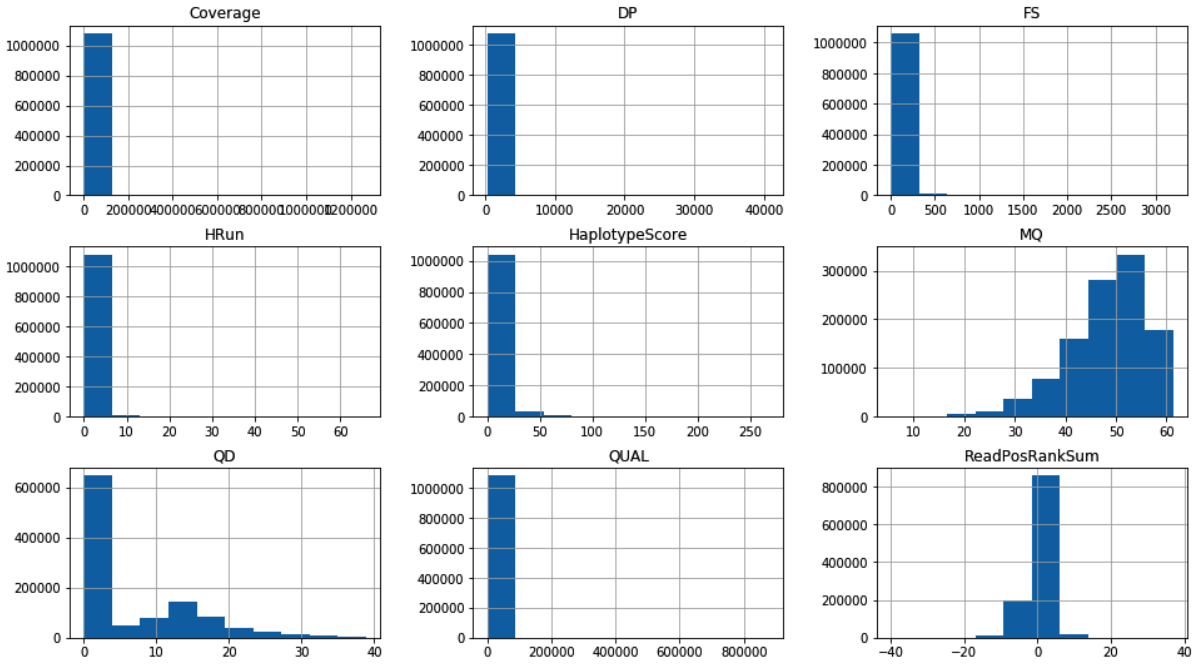

Distribution of variant DP per SNP type
DP (variant)
INTERGENIC
INTRON
NON_SYNONYMOUS_CODING
SYNONYMOUS_CODING
OTHER

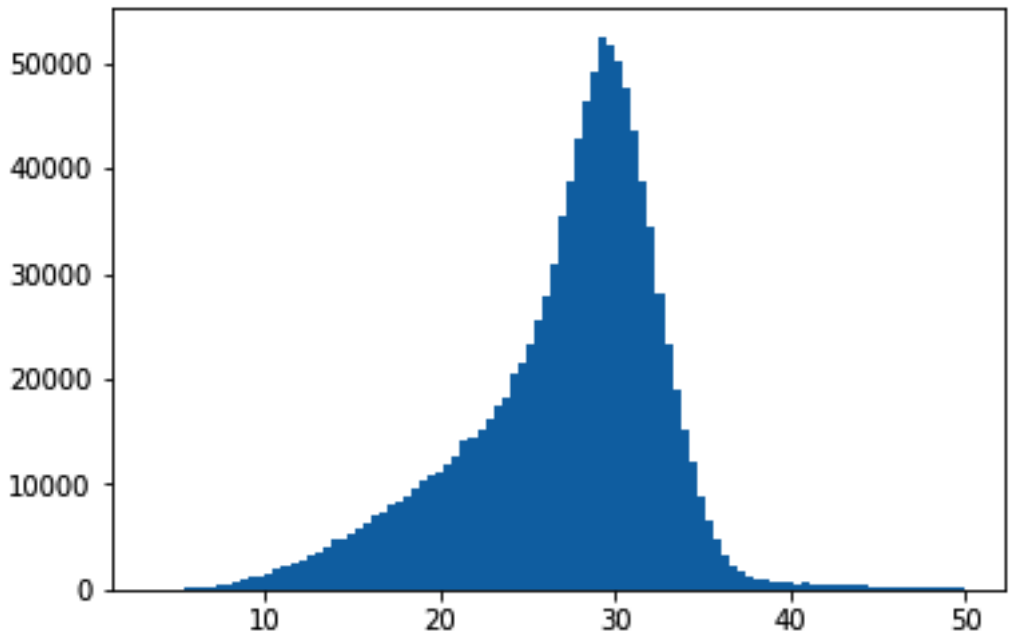
$$10^{-\frac{Q}{10}}$$

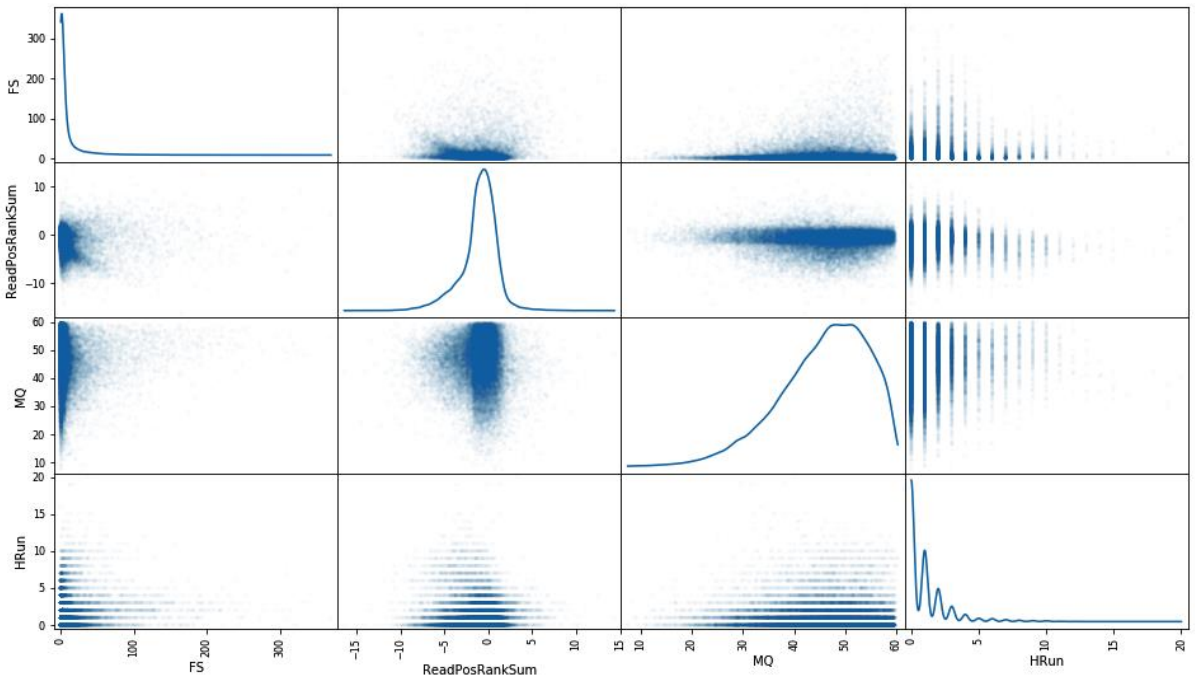
Chapter 4: Advanced NGS Data Processing

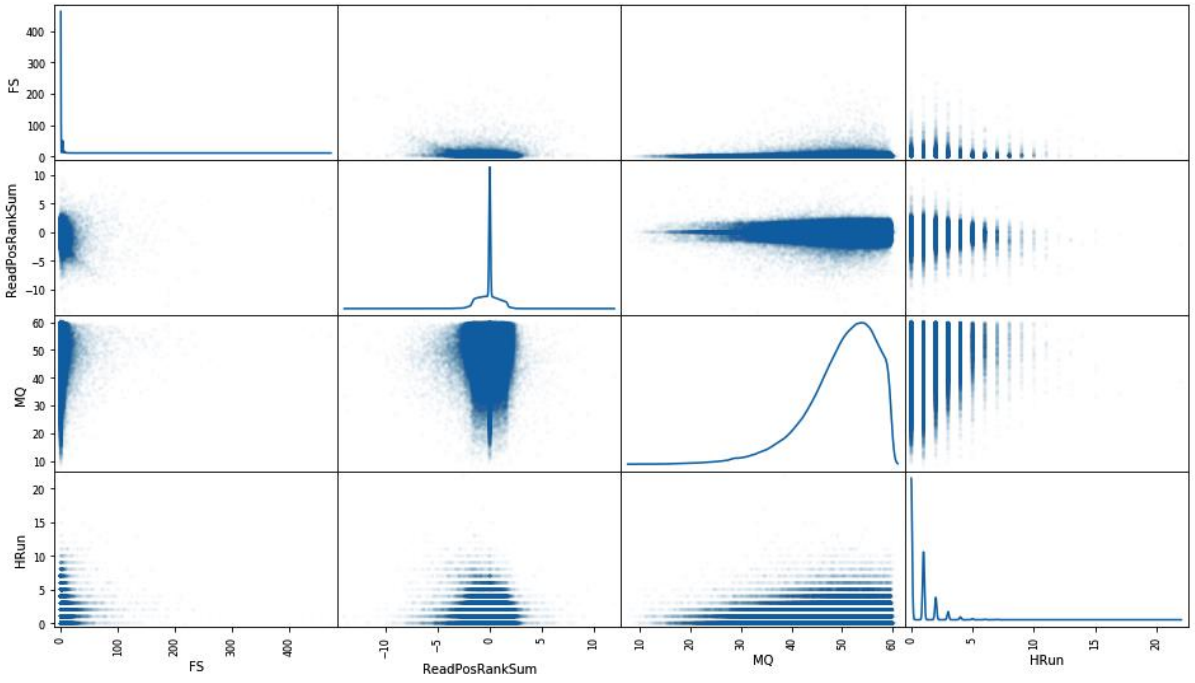


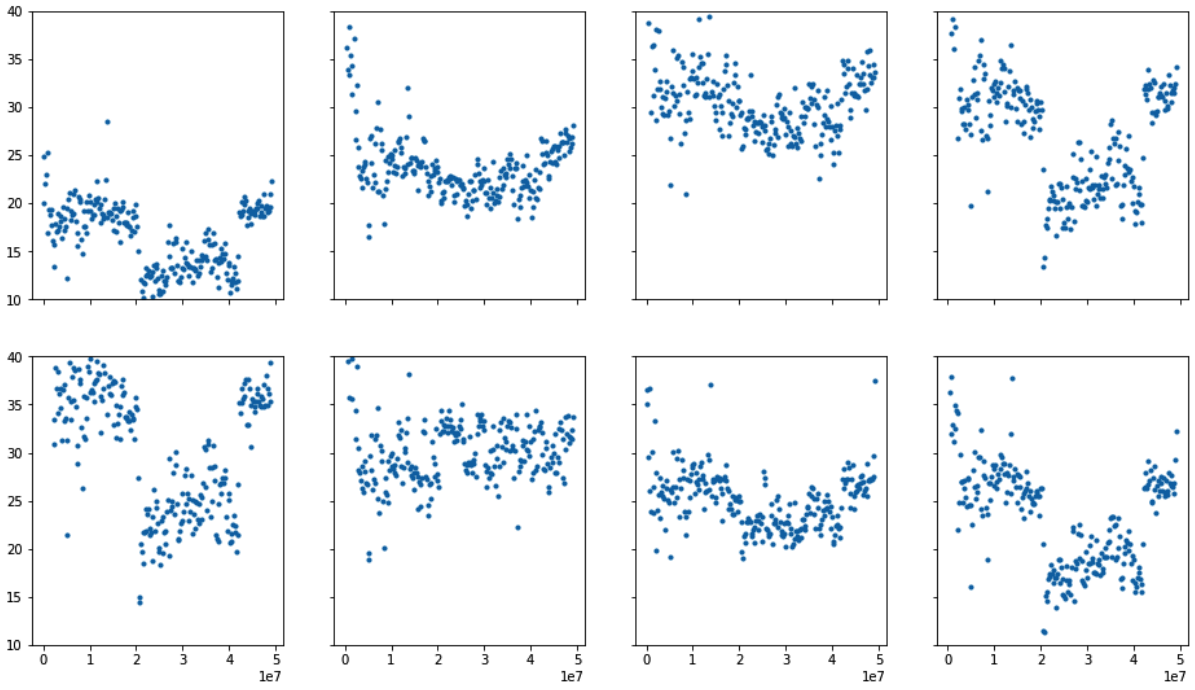



Coverage
DP
FS
HRun
HaplotypeScore
MQ
QD
QUAL
ReadPosRankSum

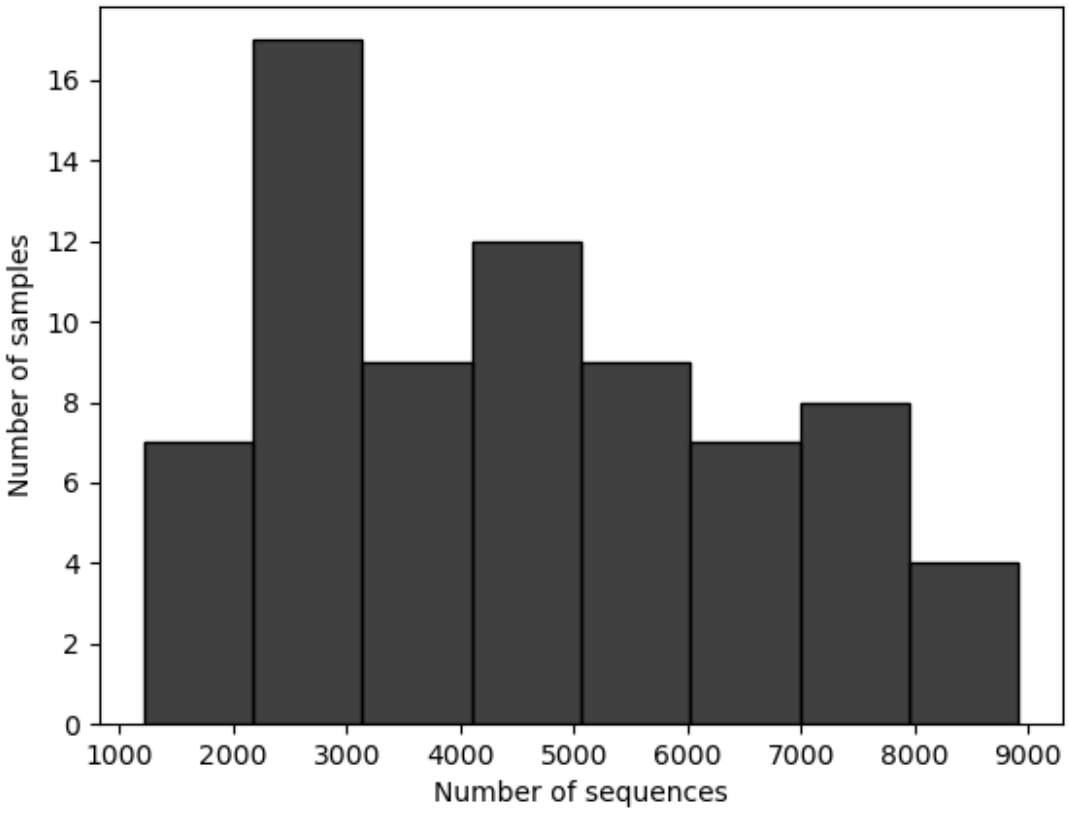



FS
ReadPosRankSum
MQ
HRun
FS
ReadPosRankSum
MQ
HRun

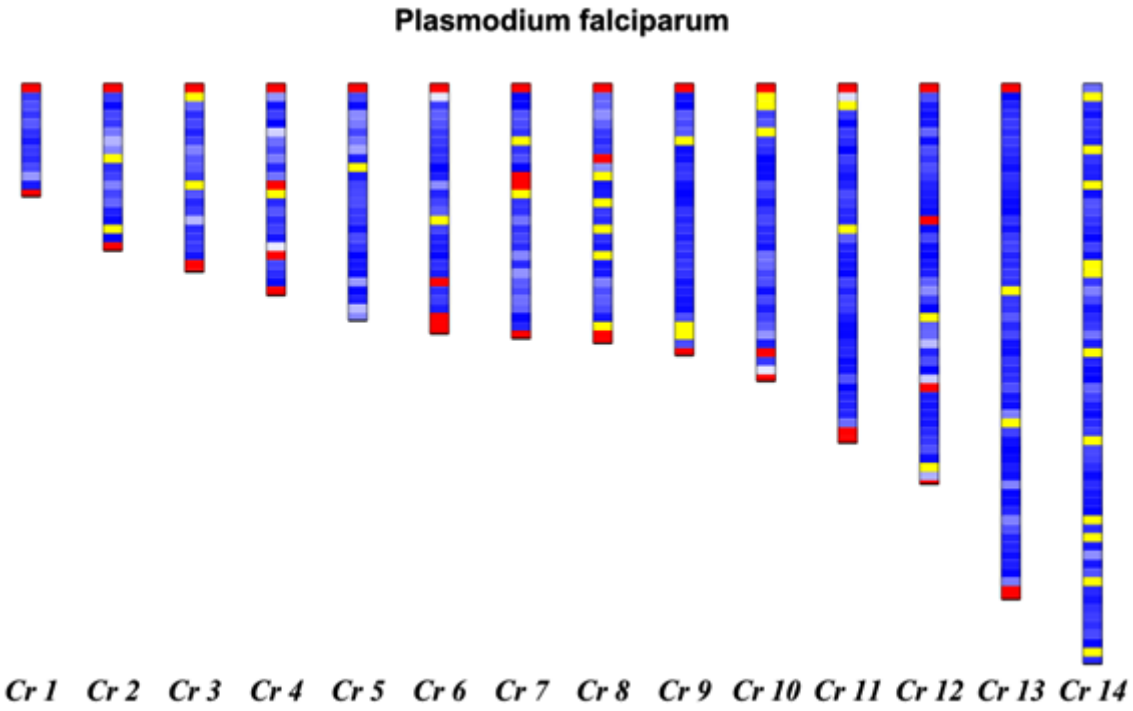

FS
ReadPosRankSum
MQ
HRun
FS
ReadPosRankSum
MQ
HRun



Forward Reads Frequency Histogram

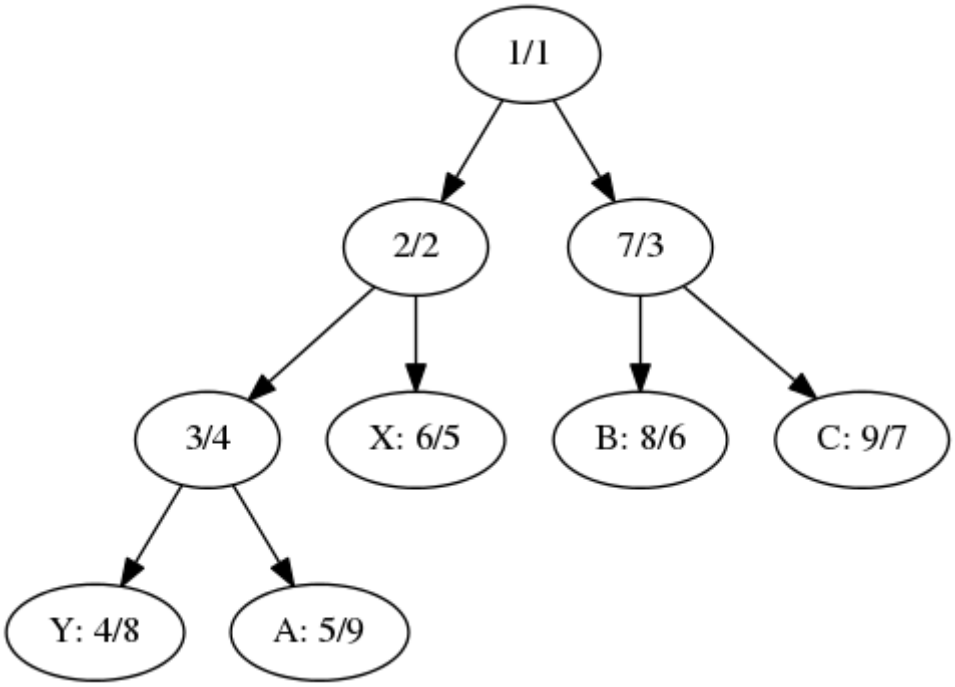


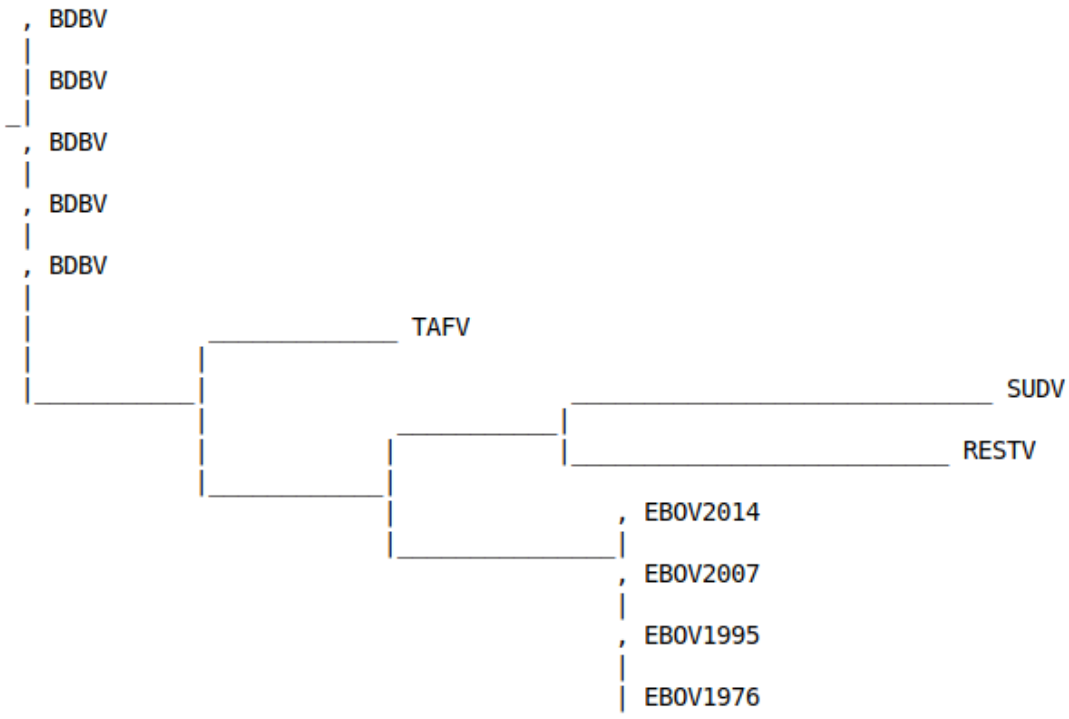
Chapter 5: Working with Genomes



Chapter 7: Phylogenetics

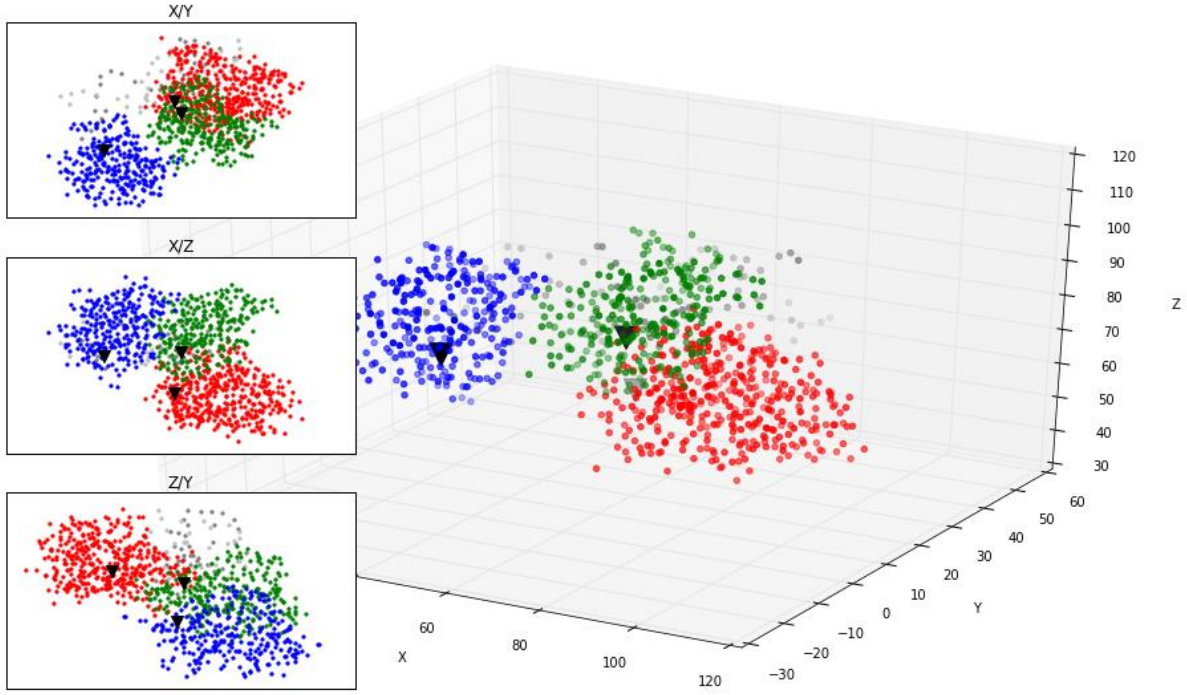
	seg_sites (RESTV)	nuc_div (RESTV)	taj_d (RESTV)	wat_theta (RESTV)	seg_sites (SUDV)	nuc_div (SUDV)	taj_d (SUDV)	wat_theta (SUDV)
NP	113.0	0.020659	-0.482275	49.489051	118.0	0.029630	1.203522	56.64
L	288.0	0.018143	-0.295386	126.131387	282.0	0.024193	1.412350	135.36
VP35	43.0	0.017427	-0.553739	18.832117	50.0	0.027761	1.069061	24.00
VP40	61.0	0.026155	-0.188135	26.715328	41.0	0.023517	1.269160	19.68





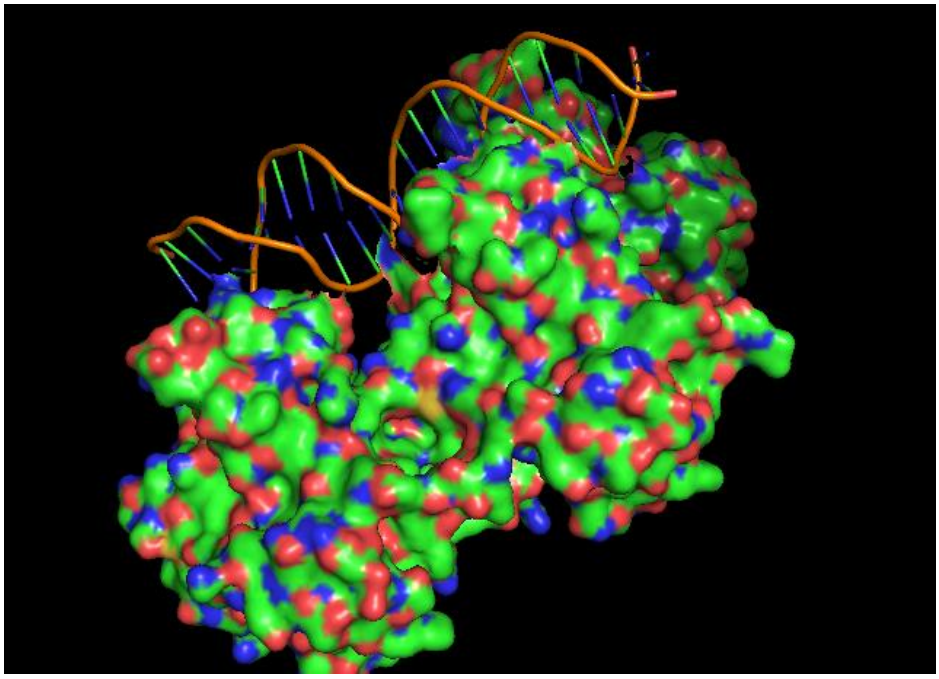
Chapter 8: Using the Protein Data Bank

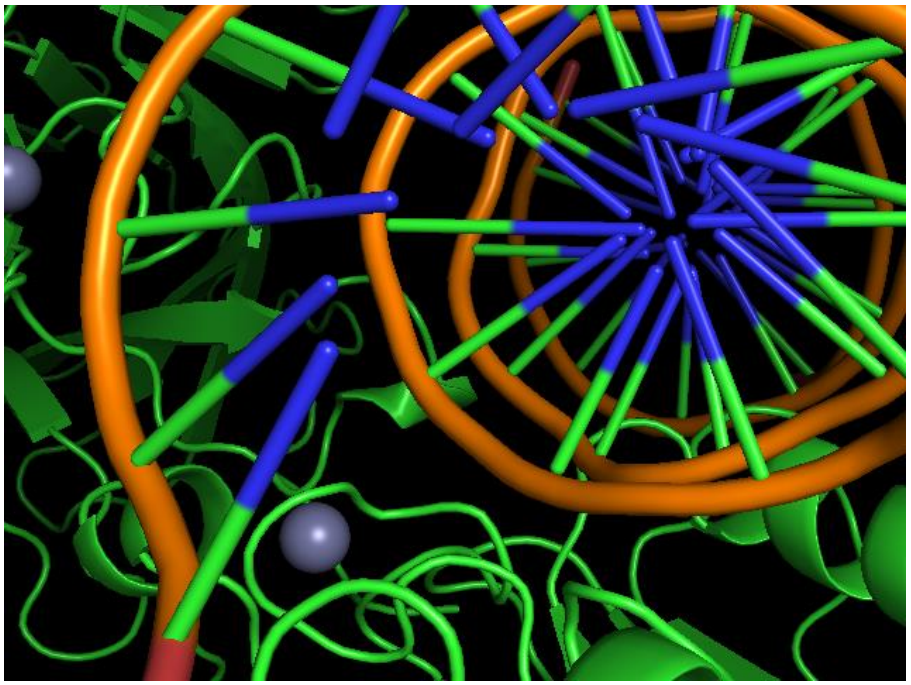
	Entry	Entry name	Length	Organism	ID	Cross-reference (PDB)	Cross-reference (HGNC)
0	Q42578	PER53_ARATH	335	Arabidopsis thaliana (Mouse-ear cress)	3702	1PA2;1QO4;	NaN
1	P79820	P53_ORYLA	352	Oryzias latipes (Japanese rice fish) (Japanese...	8090	NaN	NaN
2	Q7Z419	R144B_HUMAN	303	Homo sapiens (Human)	9606	NaN	21578;
3	Q9TUB2	P53_PIG	386	Sus scrofa (Pig)	9823	NaN	NaN
4	A7TJT7	SUB22_VANPO	442	Vanderwaltozyma polyspora (strain ATCC 22028 /...	436907	NaN	NaN
5	P56424	P53_MACMU	393	Macaca mulatta (Rhesus macaque)	9544	NaN	NaN
6	Q9W679	P53_TETMU	367	Tetraodon miurus (Congo puffer)	94908	NaN	NaN
7	Q9W678	P53_BARBU	369	Barbus barbus (Barbel) (Cyprinus barbus)	40830	NaN	NaN
8	Q29537	P53_CANLF	381	Canis lupus familiaris (Dog) (Canis familiaris)	9615	NaN	NaN
9	O09185	P53_CRIGR	393	Cricetulus griseus (Chinese hamster) (Cricetul...	10029	NaN	NaN
10	Q8SPZ3	P53_DELLE	387	Delphinapterus leucas (Beluga whale)	9749	NaN	NaN
11	P79892	P53_HORSE	280	Equus caballus (Horse)	9796	NaN	NaN
12	Q9TTA1	P53_TUPBE	393	Tupaia belangeri (Common tree shrew) (Tupaia g...	37347	NaN	NaN
13	P61260	P53_MACFU	393	Macaca fuscata fuscata (Japanese macaque)	9543	NaN	NaN
14	P04637	P53_HUMAN	393	Homo sapiens (Human)	9606	1A1U;1AIE;1C26;1DT7;1GZH;1H26;1HS5;1JSP;1KZY;1...	11998;

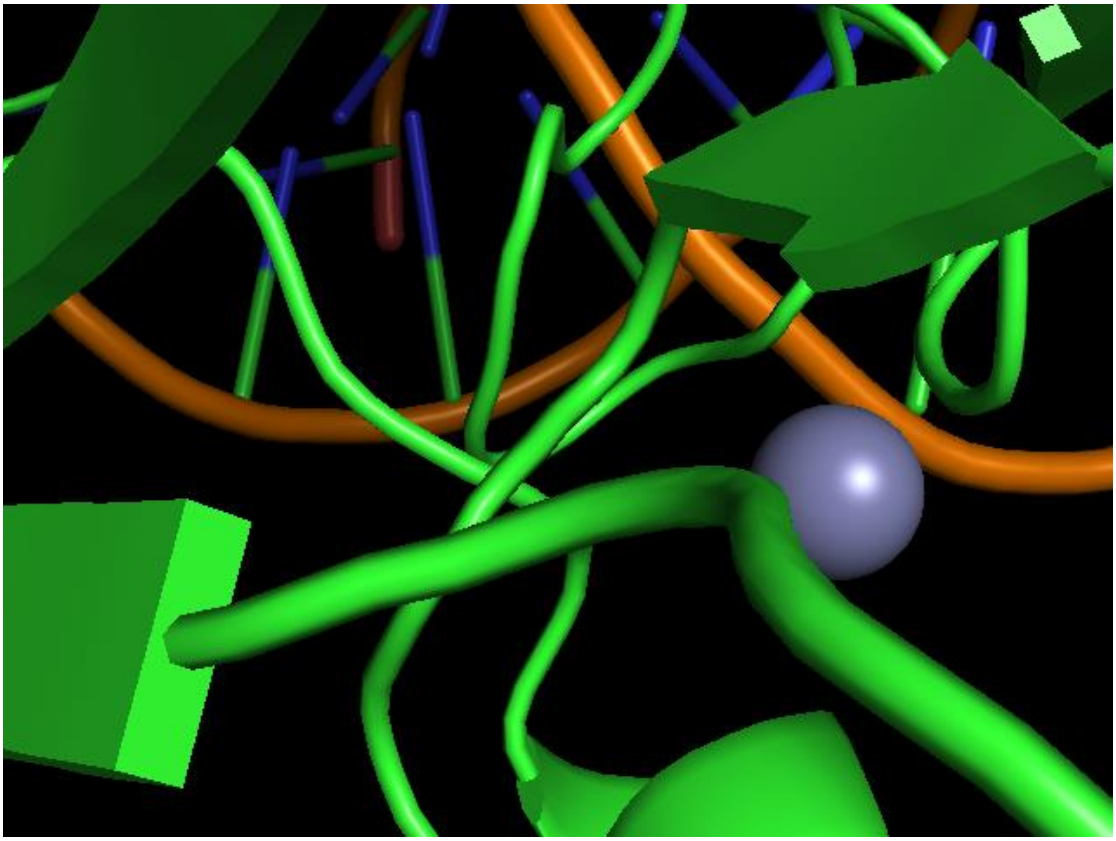


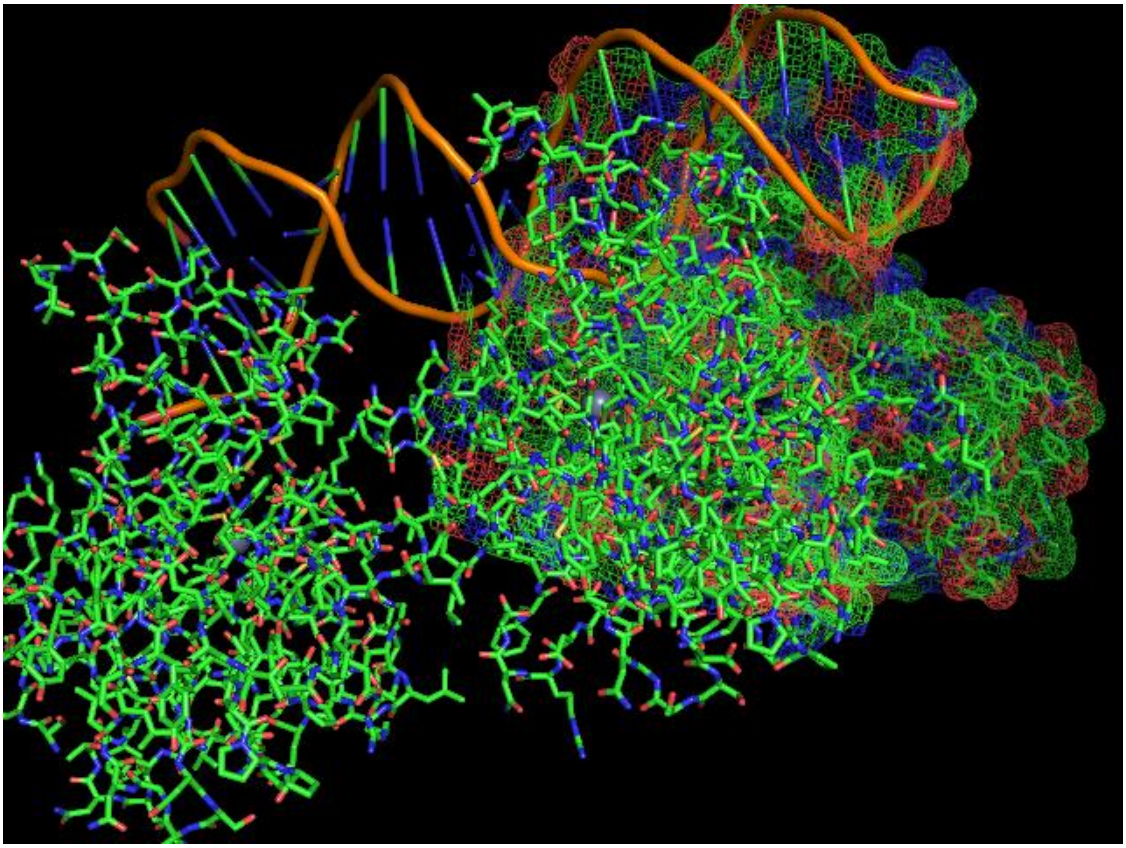
	No Water	Zincs	Water
E	6068.04412	0.00	351.9868
F	6258.20442	0.00	223.9916
A	20548.26300	65.39	3167.8812
B	20368.18840	65.39	1119.9580
C	20466.22540	65.39	1279.9520

	X	Y	Z	X (Mass)	Y (Mass)	Z (Mass)
E	49.727231	32.744879	81.253417	49.708513	32.759725	81.207395
F	51.982368	33.843370	81.578795	52.002223	33.820064	81.624394
A	72.990763	28.825429	56.714012	72.822668	28.810327	56.716117
B	67.810026	12.624435	88.656590	67.729100	12.724130	88.545659
C	38.221565	-5.010494	88.293141	38.169364	-4.915395	88.166711

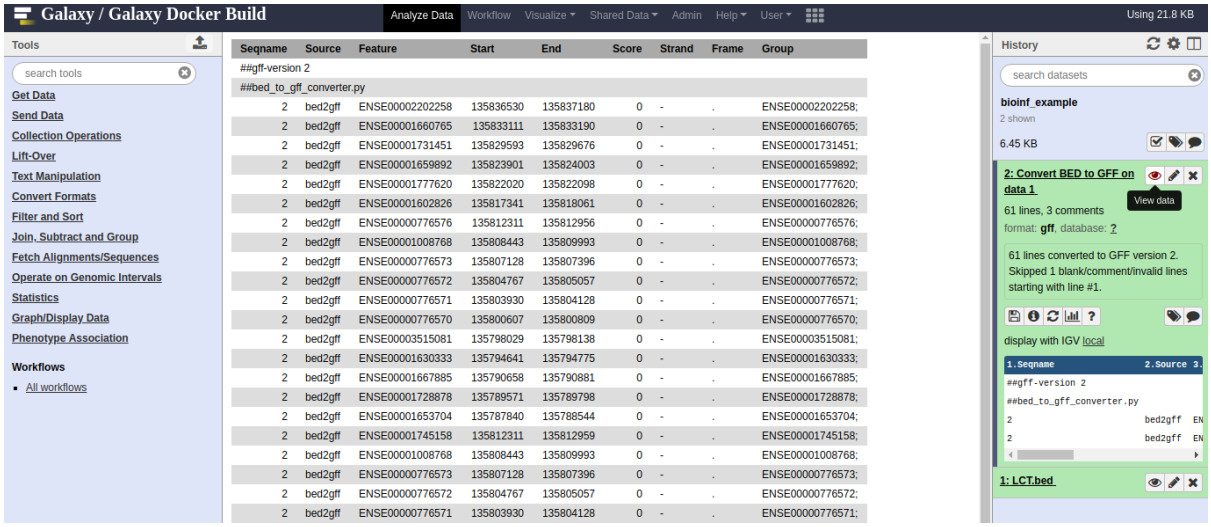


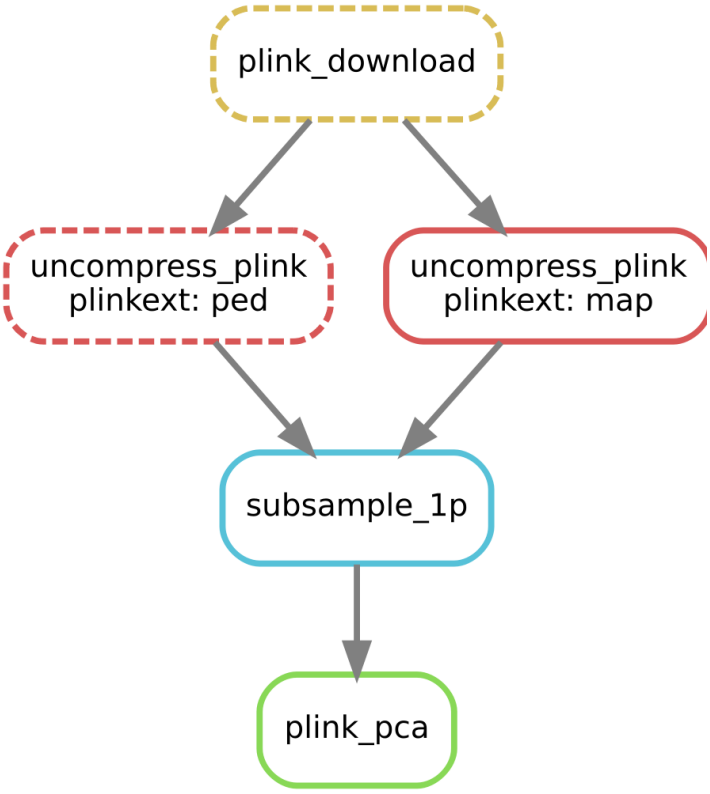






Chapter 9: Bioinformatics Pipelines



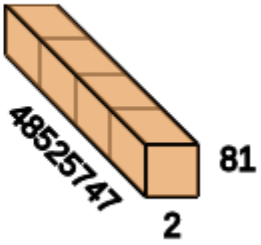


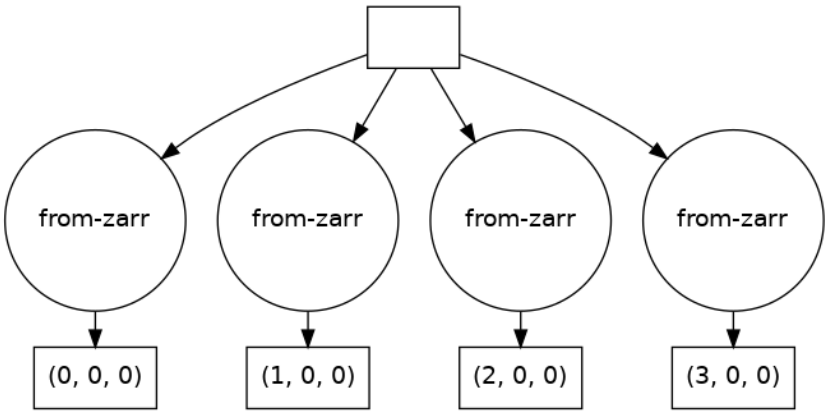
Chapter 11: Parallel Processing with Dask

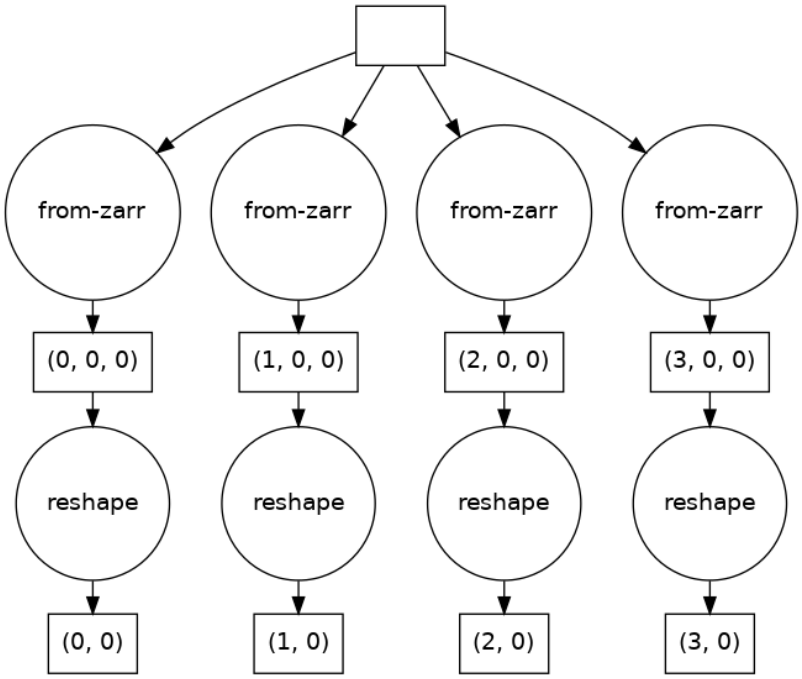
```
[6]: mosquito
```

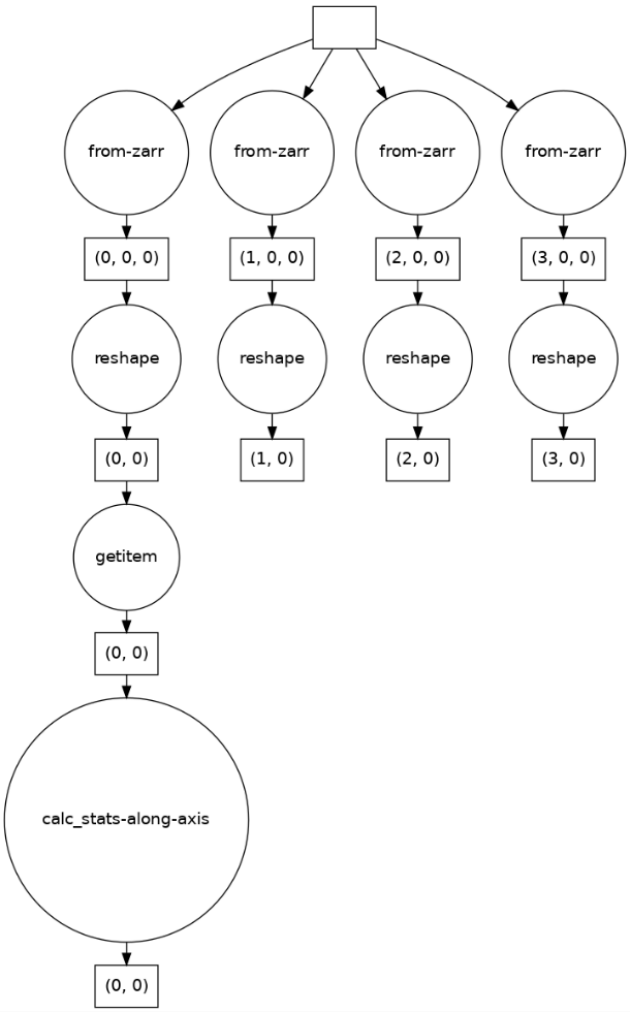
[6]:

	Array	Chunk
Bytes	7.32 GiB	1.83 GiB
Shape	(48525747, 81, 2)	(12131437, 81, 2)
Count	5 Tasks	4 Chunks
Type	int8	numpy.ndarray









Client

Client-d1eb4034-9b1e-11ec-8b81-ac675d70f79c

Connection method: Direct

Dashboard: http://127.0.0.1:34771/status

▾ Scheduler Info

Scheduler

Scheduler-96f917c0-2f31-472d-a272-b30e402cb5c4

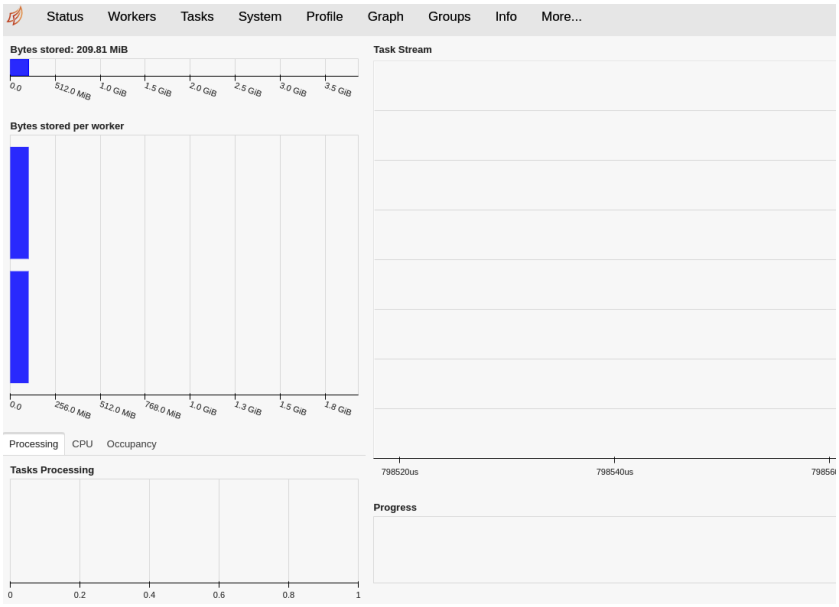
Comm: tcp://192.168.2.23:8786	**Workers:** 2
Dashboard: http://192.168.2.23:34771/status	**Total threads:** 2
Started: 4 minutes ago	**Total memory:** 3.87 GiB

▾ Workers

▾ Worker: tcp://127.0.0.1:36143

Comm: tcp://127.0.0.1:36143	**Total threads:** 1
Dashboard: http://127.0.0.1:37425/status	**Memory:** 1.93 GiB
Nanny: tcp://127.0.0.1:36727	
Local directory: /home/tantao/write/bio3/Bioinformatics-with-Python-Cookbook-third-edition/Chapter11/dask-worker-space/worker-ddfsc3ke	
Tasks executing: 0	**Tasks in memory:** 0
Tasks ready: 0	**Tasks in flight:** 0
CPU usage: 10.0%	**Last seen:** Just now
Memory usage: 98.43 MiB	**Spilled bytes:** 0 B
Read bytes: 43.78 kiB	**Write bytes:** 27.79 kiB

▸ Worker: tcp://127.0.0.1:41149



	Array	Chunk
Bytes	7.32 GiB	28.61 MiB
Shape	(48525747, 81, 2)	(300000, 50, 2)
Count	325 Tasks	324 Chunks
Type	int8	numpy.ndarray