Figures HW4

Yushi Tang

1a. Insert list of genomic regions (BED format): (help)

1b. Or BED file upload: Choose File no file selected

2. Select:

Organism: Homo sapiens

Assembly: hg38

Background: Mixed

Descriptors: Jaspar 2018 NR

Additional descriptors: Choose File no file selected

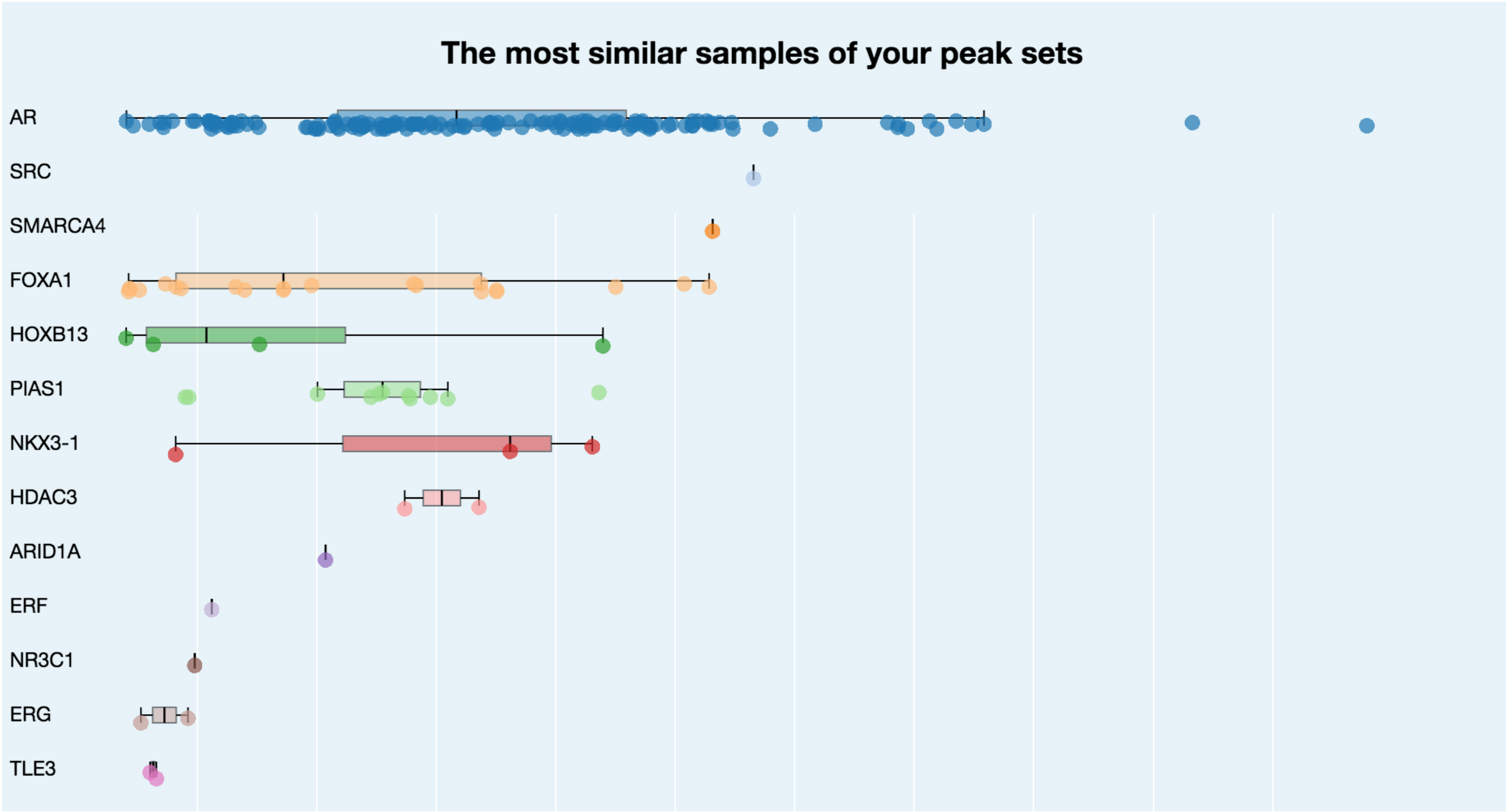
Compute global pvalue for additional descriptors

Run! Reset!

```
hw4q3_tumor_over_normal_peak_74
55      9.54144

........

************

9988 regions acquired...
Running Pscan_Chip...
Please wait... done
```

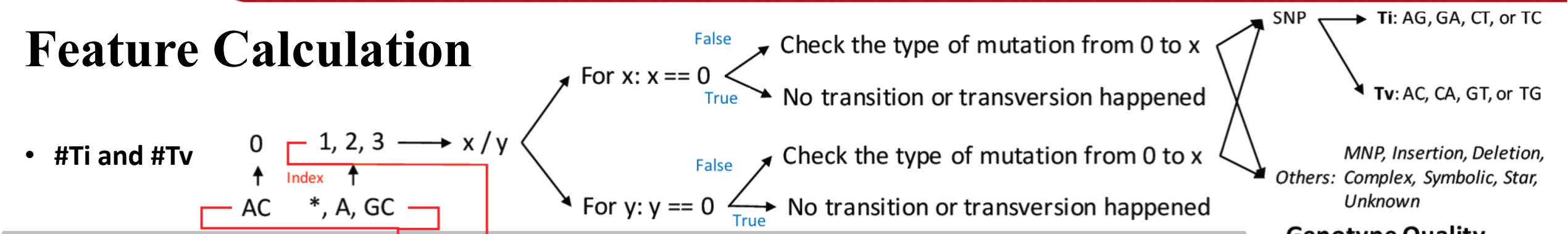
Download txt file

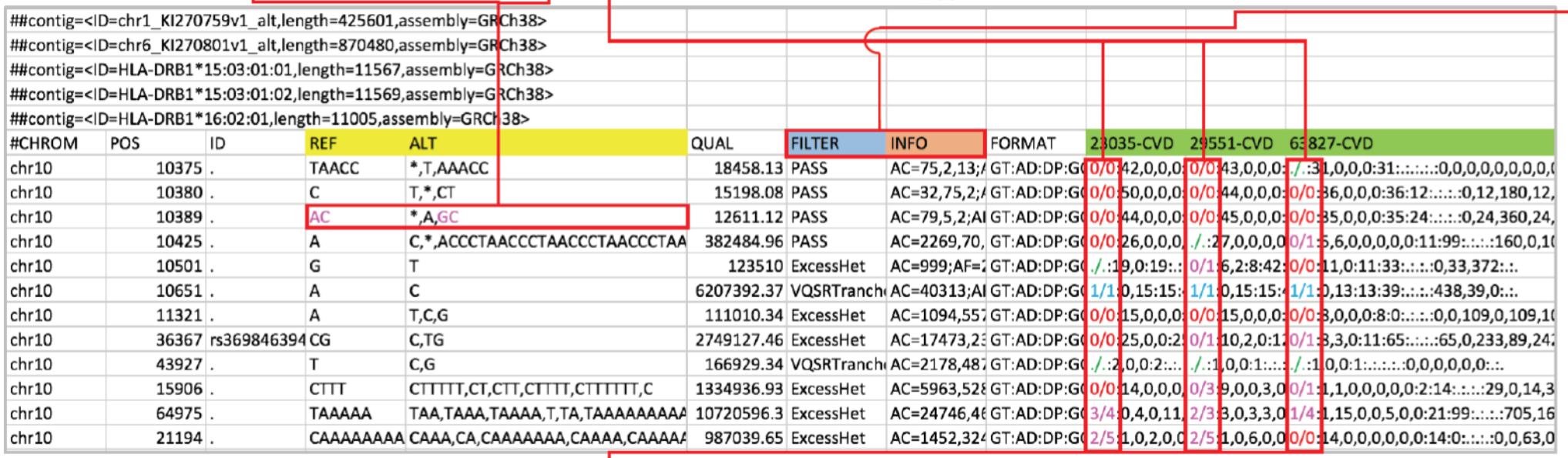
Name	ID	L.PV	L.O/U	G.PV	G.O/U	SP.COR	P.POS	P.POS.PV
Ar	MA0007.3	0		0		0.0191	[-2,8]	5.6E-13
NR3C2	MA0727.1	0		0		0.0119	[-2,8]	3.8E-10
FOXP1	MA0481.2	0		0		0.0126	[1,11]	3.7E-10
FOXF2	MA0030.1	0		0		0.0115	[12,22]	0.0535
Foxo1	MA0480.1	9.3E-274		0		0.002	[-5,5]	4.1E-5
FOXD1	MA0031.1	0		0		-0.0387	[-1,9]	2.6E-6
Foxa2	MA0047.2	0		0		0.0097	[-14,-4]	8.8E-9
FOXA1	MA0148.3	0		0		0.0077	[2,12]	4.5E-9
FOXP2	MA0593.1	0		0		0.0018	[-15,-5]	1.1E-7
FOXG1	MA0613.1	0		0		-0.076	[-16,-6]	3.3E-8
Foxj2	MA0614.1	0		0		-0.024	[-1,9]	2.1E-10
FOXI1	MA0042.2	0		0		-0.0742	[-12,-2]	2.2E-6
FOXL1	MA0033.2	0		0		-0.0739	[3,13]	1.6E-6
FOXO3	MA0157.2	0		0		-0.0142	[-16,-6]	4.6E-6
NR3C1	MA0113.3	0		0		0.0107	[-2,8]	5.2E-11
FOXB1	MA0845.1	0		0		-0.0031	[-15,-5]	0.0263
FOXC1	MA0032.2	0		0		0.0019	[-15,-5]	0.0348
FOXC2	MA0846.1	0		0		0.0046	[-14,-4]	0.0003
FOXD2	MA0847.1	0		0		-0.0746	[-2,8]	1.3E-8
FOXO4	MA0848.1	0		0		-0.0726	[-2,8]	2.0E-8
FOXO6	MA0849.1	0		0		-0.0692	[-2,8]	7.9E-9
FOXP3	MA0850.1	0		0		-0.0723	[-2,8]	1.6E-7
Foxj3	MA0851.1	0		0		0.0027	[3,13]	0.0021


The most similar samples of your peak sets
AR
SRC
SMARCA4
FOXA1
HOXB13
PIAS1
NKX3-1
HDAC3
ARID1A
ERF
NR3C1
ERG
TLE3

Feature Calculation

- **#Ti and #Tv**





```
##contig=<ID=chr1_KI270759v1_alt,length=425601,assembly=GRCh38>
##contig=<ID=chr6_KI270801v1_alt,length=870480,assembly=GRCh38>
##contig=<ID=HLA-DRB1*15:03:01:01,length=11567,assembly=GRCh38>
##contig=<ID=HLA-DRB1*15:03:01:02,length=11569,assembly=GRCh38>
##contig=<ID=HLA-DRB1*16:02:01,length=11005,assembly=GRCh38>
```

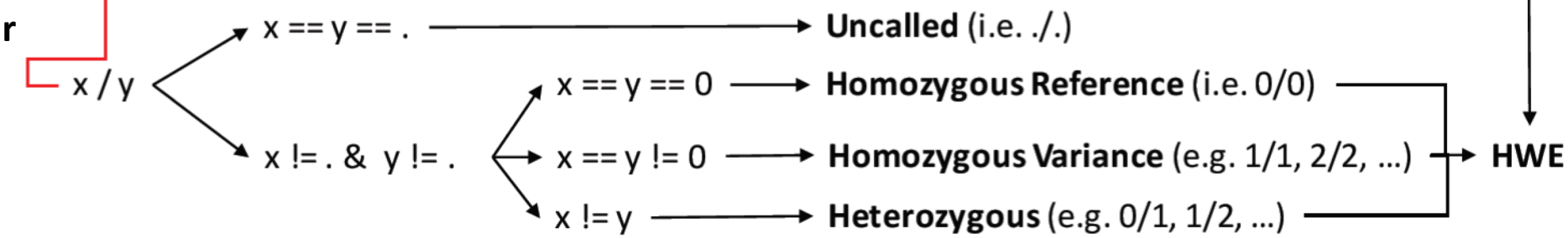
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	23035-CVD	29551-CVD	63827-CVD
chr10	10375	.	TAACC	*,T,AAACC	18458.13	PASS	AC=75,2,13;A	GT:AD:DP:GQ	0/0:42,0,0,0	0/0:43,0,0,0	./.:31,0,0,0:31:.:.:.:.:0,0,0,0,0,0,0,0,0,0
chr10	10380	.	C	T,*,CT	15198.08	PASS	AC=32,75,2;A	GT:AD:DP:GQ	0/0:50,0,0,0	0/0:44,0,0,0	0/0:36,0,0,0:36:12:.:.:.:0,12,180,12,
chr10	10389	.	AC	*,A,GC	12611.12	PASS	AC=79,5,2;AI	GT:AD:DP:GQ	0/0:44,0,0,0	0/0:45,0,0,0	0/0:35,0,0,0:35:24:.:.:.:0,24,360,24,
chr10	10425	.	A	C,*,ACCCTAACCCTAACCCTAACCCTAA	382484.96	PASS	AC=2269,70,	GT:AD:DP:GQ	0/0:26,0,0,0	./.:27,0,0,0	0/1:5,6,0,0,0,0,0:11:99:.:.:.:160,0,1
chr10	10501	.	G	T	123510	ExcessHet	AC=999;AF=2	GT:AD:DP:GQ	./.:19,0:19:.:.	0/1:6,2:8:42:	0/0:11,0:11:33:.:.:.:0,33,372:.:.
chr10	10651	.	A	C	6207392.37	VQSRTranch	AC=40313;AI	GT:AD:DP:GQ	1/1:0,15:15:.	1/1:0,15:15:4	1/1:0,13:13:39:.:.:.:438,39,0:.:.
chr10	11321	.	A	T,C,G	111010.34	ExcessHet	AC=1094,557	GT:AD:DP:GQ	0/0:15,0,0,0	0/0:15,0,0,0	0/0:8,0,0,0:8:0:.:.:.:0,0,109,0,109,1
chr10	36367	rs369846394	CG	C,TG	2749127.46	ExcessHet	AC=17473,23	GT:AD:DP:GQ	0/0:25,0,0:2	0/1:10,2,0:1	0/1:3,3,0:11:65:.:.:.:65,0,233,89,24
chr10	43927	.	T	C,G	166929.34	VQSRTranch	AC=2178,487	GT:AD:DP:GQ	./.:2,0,0:2:.:.	./.:1,0,0:1:.:.:.	./.:1,0,0:1:.:.:.:.:0,0,0,0,0,0:.:.
chr10	15906	.	CTTT	CTTTTT,CT,CTT,CTTTT,CTTTTTT,C	1334936.93	ExcessHet	AC=5963,528	GT:AD:DP:GQ	0/0:14,0,0,0	0/3:9,0,0,3,0	0/1:1,1,0,0,0,0,0:2:14:.:.:.:29,0,14,3
chr10	64975	.	TAAAAA	TAA,TAAA,TAAAA,T,TA,TAAAAAAAAA	10720596.3	ExcessHet	AC=24746,46	GT:AD:DP:GQ	3/4:0,4,0,11	2/3:3,0,3,3,0	1/4:1,15,0,0,5,0,0:21:99:.:.:.:705,16
chr10	21194	.	CAAAAAAAA	CAAA,CA,CAAAAAAA,CAAAA,CAAAAA	987039.65	ExcessHet	AC=1452,324	GT:AD:DP:GQ	2/5:1,0,2,0,0	2/5:1,0,6,0,0	0/0:14,0,0,0,0,0,0:14:0:.:.:.:0,0,63,0

Genotype Quality

PASS — Excess or VQSR

- AC: allele count in genotypes
- QD: variant quality by depth
- DP: combined depth
- AN — Coverage = 2DP / AN
- AF — min(AF, 1-AF) — MAF

- **#Het, #HomRef, and #HomVar**



Genotype Quality (GQ & PL)

- Phred-scale genotype likelihoods rounded to the closest integer (PL)

$$PL = -10 \times \log P(Genotype|Data)$$

  - P(Genotype|Data) is the conditional probability of the Genotype given the sequence Data that we have observed
  - Normalization: subtract the value of the lowest PL from all the values, and normalize the values across all genotypes so that the PL value of the most likely genotype is 0
  - E.g. Alleles with A (Ref) and T (Alt) with conditional probabilities as:

P(AA|Data) = 0.000001, P(AT|Data) = 0.000100, P(TT|Data) = 0.010000

Genotype	A/A	A/T	T/T
Raw PL	-10*log(0.000001) = 60	-10*log(0.000100) = 40	-10*log(0.010000) = 20
Normalized PL	PL(0/0) = 60 – 20 = 40	PL(0/1) = 40 – 20 = 20	PL(1/1) = 20 – 20 = 0

- Genotype quality (GQ)
  - The difference between the second lowest PL and the lowest PL

Genotype Quality (GQ & PL)

- Where are GQ and PL located in original VCF file?

##contig=<ID=chr1_KI270763v1_alt,length=911658,assembly=GRCh38>											
##contig=<ID=chr1_KI270759v1_alt,length=425601,assembly=GRCh38>											
##contig=<ID=chr6_KI270801v1_alt,length=870480,assembly=GRCh38>											
##contig=<ID=HLA-DRB1*15:03:01:01,length=11567,assembly=GRCh38>											
##contig=<ID=HLA-DRB1*15:03:01:02,length=11569,assembly=GRCh38>											
##contig=<ID=HLA-DRB1*16:02:01,length=11005,assembly=GRCh38>											
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT			
chr10	10296	.	T	C	119.87	ExcessHet	AC=22;AF=5.	GT:AD:DP	GQ	MIN_DP:PGT:PID	:PL:RGQ:SB
chr10	10325	.	C	CTAACCCTAA	80.21	ExcessHet	AC=4;AF=8.8	GT:AD:DP	GQ	MIN_DP:PGT:PID	:PL:RGQ:SB
chr10	10375	.	TAACC	*,T,AAACC	18458.13	PASS	AC=75,2,13;A	GT:AD:DP	GQ	MIN_DP:PGT:PID	:PL:RGQ:SB
chr10	10380	.	C	T,*,CT	15198.08	PASS	AC=32,75,2;A	GT:AD:DP	GQ	MIN_DP:PGT:PID	:PL:RGQ:SB
chr10	10389	.	AC	*,A,GC	12611.12	PASS	AC=79,5,2;AI	GT:AD:DP	GQ	MIN_DP:PGT:PID	:PL:RGQ:SB
chr10	10425	.	A	C,*,ACCCTAA	382484.96	PASS	AC=2269,70,	GT:AD:DP	GQ	MIN_DP:PGT:PID	:PL:RGQ:SB
chr10	10501	.	G	T	123510	ExcessHet	AC=999;AF=2	GT:AD:DP	GQ	MIN_DP:PGT:PID	:PL:RGQ:SB
chr10	10651	.	A	C	6207392.37	VQSRTranch	AC=40313;AI	GT:AD:DP	GQ	MIN_DP:PGT:PID	:PL:RGQ:SB
chr10	11321	.	A	T,C,G	111010.34	ExcessHet	AC=1094,557	GT:AD:DP	GQ	MIN_DP:PGT:PID	:PL:RGQ:SB

23035-CVD	29551-CVD	63827-CVD
0/0:22,0:22:0:.:.:.:0,0,377:.:.	0/0:16,0:16:9	0/0:9,0:9:0
0/0:40,0:40:46:.:.:.:0,46,981:.:.	0/0:34,0:34:4	0/0:13,0:13
0/0:42,0,0,0:42:0:.:.:.:0,0,373,0,373,373,0,373,373,373:.:.	0/0:43,0,0,0:	./.:31,0,0,0
0/0:50,0,0,0:50:21:.:.:.:0,21,1218,21,1218,1218,21,1218,1218,1218:.:.	0/0:44,0,0,0:	0/0:36,0,0,
0/0:44,0,0,0:44:48:.:.:.:0,48,720,48,720,720,48,720,720,720:.:.	0/0:45,0,0,0:	0/0:35,0,0,
0/0:26,0,0,0,0,0,0:26:0:.:.:.:0,0,530,0,530,530,0,530,530,530,0,530,530,	./.:27,0,0,0,0	0/1:5,6,0,0
./.:19,0:19:.:.:.:.:0,0,0:.:.	0/1:6,2:8:42:	0/0:11,0:1
1/1:0,15:15:45:.:1|1:10617_G_C:542,45,0:.:.	1/1:0,15:15:4	1/1:0,13:1
0/0:15,0,0,0:15:42:.:.:.:0,42,630,42,630,630,42,630,630,630:.:.	0/0:15,0,0,0:	0/0:8,0,0,0

Genotype Quality (GQ & PL)

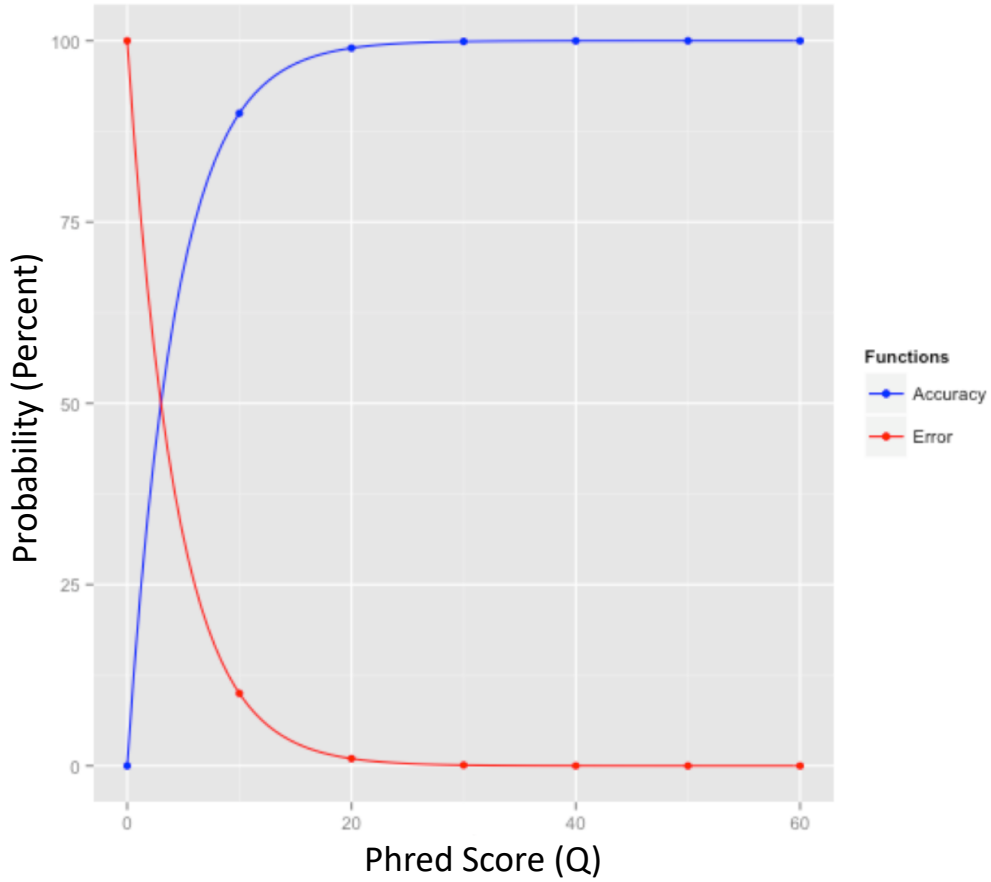
- **Phred-scale in practice:** 20 is recommended by GATK as a threshold
  - The Phred quality score (Q) is logarithmically related to the error probability (E)

$$Q = -10 \times \log E$$

  - As for the estimate of accuracy (A), we can take

$$A = 1 - E = 1 - 10^{-(\frac{Q}{10})}$$

Phred Quality Score	Error	Accuracy(1-Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%



Data Size Comparison Between Formats

Data Sets	VCF	VCF.BGZ	BCF	GDS
1000 Genome	75 GB	14 GB	12.3 GB	2.6 GB
TOPMed Freeze 5 (n=54,000)	120 TB	22 TB	987 GB	143 GB
GSP CCDG Freeze 1 raw (n = 22,000)	436 TB	55 TB	3.6 TB	450 GB
GSP CCDG Freeze 1 after QC (n = 22,000)	21 TB	2 TB	126 GB	21 GB
Compression	1	0.09	0.006	0.001

The files in different formats contain the same data