

Vcf file format

The diagram illustrates the VCF file format with a table of variant data. Annotations with arrows point to specific rows and columns, explaining the meaning of the data. The table has columns for chromosome, position, reference and alternative alleles, quality, and format. The first four rows represent different types of variants: a biallelic SNP, a multi-allelic SNP, an indel, and an invariant/monomorphic site. The next three rows show genotypes for three specific variants (CAM002881, CAM002894, and CAM008026), with annotations for homozygous reference (GG), heterozygous (AG), and homozygous alternative (AA) states. The final row shows missing data for the first two variants.

#CHROM	POS	REF	ALT	QUAL	FORMAT	CAM002881	CAM002894	CAM008026
chr1	119401	G	A	96.05	GT:AD:DP:GQ:PL	0/0:11,0:11:15:0,15,225	0/1:4,5:9:27:27,0,274	1/1:15,0:15:45:0,45,491
chr1	131983	C	T,G	1085.57	GT:AD:DP:GQ:PGT:PID:PL	0/0:4,0,0:4:12:....:0,12,99,12,99,99	2/2:0,0,17:17:48:....:720,48,720,48,720,0	0/1:0,5,0:5:15:....:15,0,175,15,175,175
chr1	65534	G	GG	192.44	GT:AD:DP:GQ:PGT:PID:PL	0/0:5,0:5:15:....:0,15,175	./.:2,0:2:6:....:0,6,84	0/0:8,0:8:24:....:0,24,266
chr1	133078	A	.	24.68	GT:AD:DP:RGQ	./.:1,0:1:3	./.:2,0:2:6	0/0:3,0:3:9

Annotations:

- biallelic SNP (points to row 1)
- multi-allelic SNP (points to row 2)
- indel (points to row 3)
- invariant/monomorphic site (points to row 4)
- homozygous for the reference (GG) (points to row 5, column 7)
- heterozygous Genotype (AG) (points to row 5, column 8)
- homozygous for the alternative allele (AA) (points to row 5, column 9)
- missing data (points to row 6, column 7)
- homozygous for the second alternative allele (GG) (points to row 6, column 8)

Vcf filtering

#CHROM	POS	REF	ALT	QUAL	FORMAT	CAM002881	CAM002894	CAM008026
chr1	119401	G	A	96.05	GT:AD:DP:GQ:PL	0/0:11,0:11:15:0,15,225	0/0:9,0:9:27:0,27,274	0/0:15,0:15:45:0,45,491
chr1	131983	C	T,G	1085.57	GT:AD:DP:GQ:PGT:PID:PL	0/0:4,0,0:4:12:...:0,12,99,12,99,99	0/0:17,0,0:17:48:...:0,48,720,48,720,720	0/0:5,0,0:5:15:...:0,15,175,15,175,175
chr1	65534	G	GG	192.44	GT:AD:DP:GQ:PGT:PID:PL	0/0:5,0:5:15:...:0,15,175	./:2,0:2:6:...:0,6,84	0/0:8,0:8:24:...:0,24,266
chr1	133078	A	.	24.68	GT:AD:DP:RGQ	0/0:15,0:15:25	0/0:6,0:6:13	0/0:3,0:3:9

Genotype filtering

minGQ 30 (remove genotypes with quality below 30)
minDP 10 (remove genotypes with less than 10 reads)

#CHROM	POS	REF	ALT	QUAL	FORMAT	CAM002881	CAM002894	CAM008026
chr1	119401	G	A	96.05	GT:AD:DP:GQ:PL	./.	./.	0/0:15,0:15:45:0,45,491
chr1	131983	C	T,G	1085.57	GT:AD:DP:GQ:PGT:PID:PL	./.	0/0:17,0,0:17:48:...:0,48,720,48,720,720	./.
chr1	65534	G	GG	192.44	GT:AD:DP:GQ:PGT:PID:PL	./.	./.	./.
chr1	133078	A	.	24.68	GT:AD:DP:RGQ	0/0:15,0:15:25	./.	./.

Site filtering

max-missing 0.33 (remove sites with more than 33% missing data)
minQ 30 (remove sites with quality below 30)

#CHROM	POS	REF	ALT	QUAL	FORMAT	CAM002881	CAM002894	CAM008026
chr1	119401	G	A	96.05	GT:AD:DP:GQ:PL	0/0:11,0:11:15:0,15,225	./.	0/0:15,0:15:45:0,45,491
chr1	131983	C	T,G	1085.57	GT:AD:DP:GQ:PGT:PID:PL	./.	0/0:17,0,0:17:48:...:0,48,720,48,720,720	./.
chr1	65534	G	GG	192.44	GT:AD:DP:GQ:PGT:PID:PL	./.	./.	./.
chr1	133078	A	.	24.68	GT:AD:DP:RGQ	0/0:15,0:15:25	./.	./.