Data file format

VCF text files with .vcf extension

VCF files have a complex structure with the start consisting of a number of lines starting with '##' that describe the data format in the rest of the file. This is followed by a single line starting with '#' that describes the fields/columns in the rest of the file as outlined in the table below. The contents of the **FORMAT** and **Sample** columns are variable with the requirements of `AgileMultiIdeogram` outlined in the 2nd table.

Header	Description	Possible values	Note	Required
#CHROM	Chromosome	The name of the reference sequence in the reference genome. For autosomal chromosomes it is a number (1-22). X, Y and MT typically refer to the mitochondrial and sex chromosomes. Other sequences may be present that contain unlocated sequences, common viruses or common alternative haplotypes these tend to have complex names. These names may have the 'chr' prefix.	Only variants with values of 1 to 22 or chr1 to chr22 are retained	Yes
POS	Position	The variant's location in the reference sequence (The first base of the reference may be referenced as 0 and not 1).	Any whole number	Yes
ID	Variant name	This can be any text value, but is typically either a RS id or a '.'	`AgileMultiIdeogram` may filter variants based on the presence of 'RS' at the start of the name	Optional
REF	Variant's wild type allele	Any possible sequence of A, C, G and T	Only alleles of one nucleotide are retained	Yes
ALT	Variant's alternative allele(s)	Any possible sequence of A, C, G and T. If a variant is heterozygous for two none reference alleles each allele will be separated by a comma.	Only variants with a single, 1 base Alt allele are retained	Yes

Header	Description	Possible values	Note	Required
QUAL	Quality score	Any number or '.'	An aligner/variant caller specific score of the variant's quality	No
FILTER	Variant filtering describing any filtering performed on the data	Any value	Not used	No
INFO	Variant specific data created by variant caller	List of key value pairs		No
FORMAT	Structure and format of variant specific data, written as a series of keys (described in the header section) separated by colon (**:**)	The contents are variable depending on variant caller and variant type	See below	Yes
<JT706> (Sample/patient name)	Alphanumeric text	Variant specific data as described in the **FORMAT** column	For multiple sample VCF files, this column will be repeated for each samples See below	Yes

The contents of the **Format** and **Sample** columns are variable, being variant caller and variant type specific. For example a VCF file created by GATK's 'haplotypecaller' may have a format field like this ' GT:AD:DP:GQ:PL', table below describes its structure.

Key	Value	Description	Required

Key	Value	Description	Required
GT	Variant's genotype	Each allele is numbered with the reference allele = 0 and the alternative allele's value given by its position in the Alt column list of alleles. The programs ignores variants with multiple alternative alleles so this value is always 1. Homozygous variants are shown as 1/1, while heterozygous positions are either 0/1 or 1/0. Homozygous reference genotypes (0/0) are typically not present in single sample files and are ignored in multi-sample VCF files	Yes if -V option used
AD	Two numbers separated by a ','	Allelic depths for the ref and alt alleles in the order listed	Yes if -V option not used
DP	Number	Approximate read depth (poor quality reads may be ignored)	Yes if -V option not used
GQ	Number	Genotype Quality as calculated by the variant caller	No
PL	Three numbers separated by ','	Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification	No

As can be seen from the two tables the programs require the chromosome, position and allele sequences that are always present in a vcf file. However, they also require the allele read depth and total read depth data (as detailed in columns 9 and 10+ of the file) if the -V option is not used; this data may be omitted by some variant callers. Consequently it is suggested that GATK's haplotypecaller is used to make the files.

Affymetrix tab-delimited text files with .xls extension

Header	Chromosome	Chromosomal Position	dbSNP RS ID	Call
Data type	Text with or without 'chr' prefix	Number	Text	Text (AA, AB, BB or NoCall)

Affymetrix birdseed files files with .txt extension

Header	Chromosome	Chromosomal Position	dbSNP RS ID	Call
Data type	Text with or without 'chr' prefix	Number	Text	Text (AA, AB, BB or NoCall)