Kanza Batool Haider
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 8/8/2018

WEEK 2: STUDY ON VCF FILE SPECIFICATIONS

The first step for working on any project is to understand and explore the content of its data. In order to perform annotation on the variants to predict cancerous mutations, the data that I shall be required to work upon is typically a **VCF file,** a file in which only the variants are reported instead of the whole sequence which could be as long as 3 billion bp incase of a human genome making **700 MB** file size or if we also align the sequence w.r.t to a reference genome (FASTQ file) then the size goes up to **200GB.** But since, only about 0.1% of the genome is different among individuals, which equates to about 3 million variants (aka mutations) in the average human genome. This means it is much smarter way to only store for information for just the places where any given individual differs from the normal "reference" genome. This is exactly what a vcf file is and the size of a vcf file is normally onto the range of only **125MB.**

VCF stands for Variant Call Format. It is a standardized text file format stores information for:

- SNP
- Indel
- Structural variation calls.

SNP: Single Nucleotide Polymorphism; are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called the nucleotide. SNPs occur normally throughout a person's DNA. They occur once in every 300 nucleotides on average, which means that for an average DNA with 3 billion nucleotides there are 10 million SNPS in the human genome. (3 billion / 300 = 10 million)

For e.g.: a SNP may replace the nucleotide Cytosine (C) with the nucleotide Thymine (T) in a certain stretch of DNA.

Reference Sequence: ATCTTCAGCCAC AAAAGATGAAGTT

SNP: ATCTTCAGCCAT AAAAGATGAAGTT

Indel: is a molecular biology term for an insertion or deletion of bases in the genome of an organism.

Reference Sequence: ATCTTCAGCCATAAAAGATGAAGTT

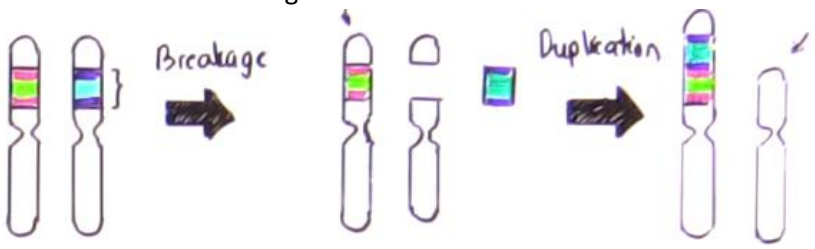
Deletion (3bp): ATCTTCAGCCAAAGATGAAGTT

Insertion (4bp): ATCTTCAGCCATATGTGAAAGATGAAG

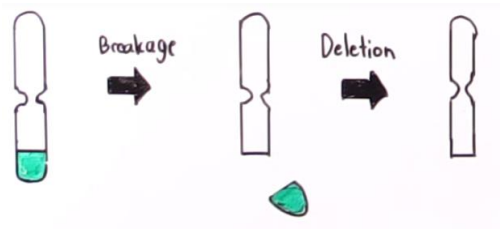
Structural Variations: under certain circumstances, a section of a chromosome can break off. The detached variant can lead to some of the example structural variants shown below:

Kanza Batool Haider
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 8/8/2018

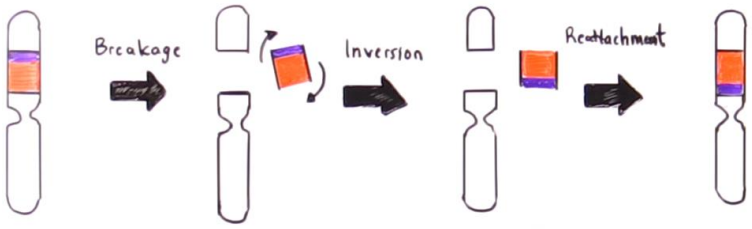
Inversion: The deleted segment reverses its orientation and reattaches to the original chromosome.



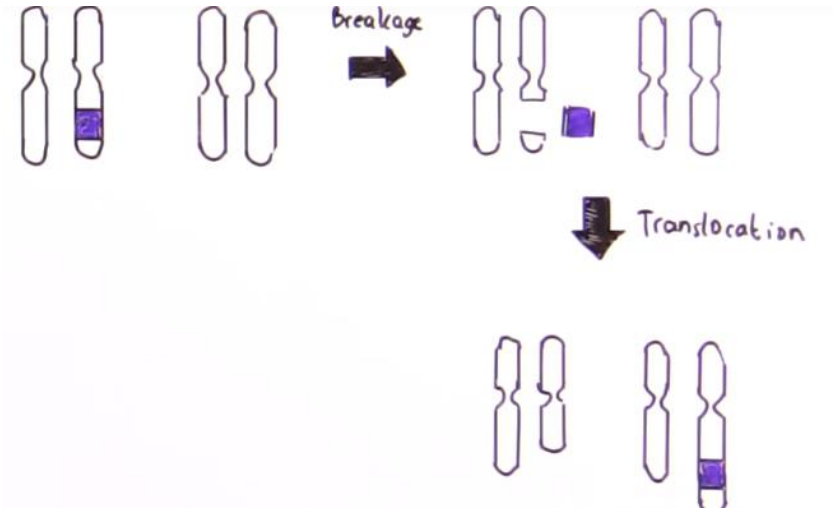
Deletion: If the detached fragment of DNA does not reattach to the original chromosome



Duplication: If a segment of DNA breaks off and attaches into the homologous chromosome



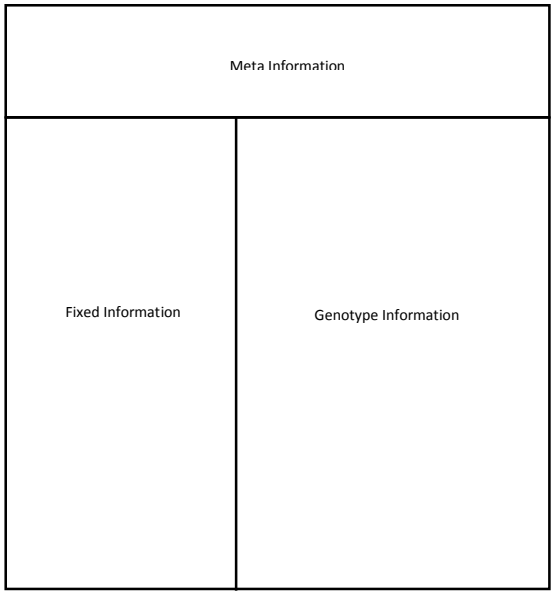
Translocation: If a segment of DNA breaks off and attaches onto a non-homologous chromosome



Kanza Batool Haider
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 8/8/2018

VCF File Structure

A vcf file can be thought of as having three sections – a vcf header (Meta data info), a fix region and a gt region.



Meta Information vcf header:

- Each vcf Meta line begins with a ##.
- The information in the Meta region defines the abbreviations used elsewhere in the file.
- It may document the software used to create the vcf file and version of the vcf file.

Example:

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Fixed Region:

Kanza Batool Haider
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 8/8/2018

The first eight columns of this table contain information about each variant. This data is fixed over all samples.

CHROM: The name of the sequence (typically a chromosome) on which the variation is being called. This sequence is usually known as 'the reference sequence', i.e. the sequence against which the given sample varies.

POS: The 1-based position of the variation on the given sequence.

ID: The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".". Multiple identifiers should be separated by semi-colons without white-space.

REF: The reference base (or bases in the case of an indel) at the given position on the given reference sequence.

ALT: The list of alternative alleles at this position.

QUAL: A quality score associated with the inference of the given alleles.

FILTER: A flag indicating which of a given set of filters the variation has passed.

INFO: An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: "<key>=[, data]".

AA ancestral allele

AC allele count in genotypes, for each ALT allele, in the same order as listed

AF allele frequency for each ALT allele in the same order as listed: use this when estimated from primary data, not called genotypes

AN total number of alleles in called genotypes

BQ RMS base quality at this position

CIGAR cigar string describing how to align an alternate allele to the reference allele

DB dbSNP membership

DP combined depth across samples, e.g. DP=154

END end position of the variant described in this record (for use with symbolic alleles)

H2 membership in hapmap2

H3 membership in hapmap3

MQ RMS mapping quality, e.g. MQ=52

MQ0 Number of MAPQ == 0 reads covering this record

NS Number of samples with data

SB strand bias at this position

SOMATIC indicates that the record is a somatic mutation, for cancer genomics

VALIDATED validated by follow-up experiment

1000G membership in 1000 Genomes

Example:

```
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G
```

Genotype Information:

The organization of each cell containing a genotype and associated information is specified in column nine, The FORMAT column.

```
FORMAT      NA00001         NA00002         NA00003
GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
GT:GQ:DP    0/1:35:4        0/2:17:2        1/1:40:3
```

GT: genotype, encoded as allele values separated by either of / or |.

- /: genotype unphased
- |: genotype phased

DP: read depth at this position for this sample (Integer)

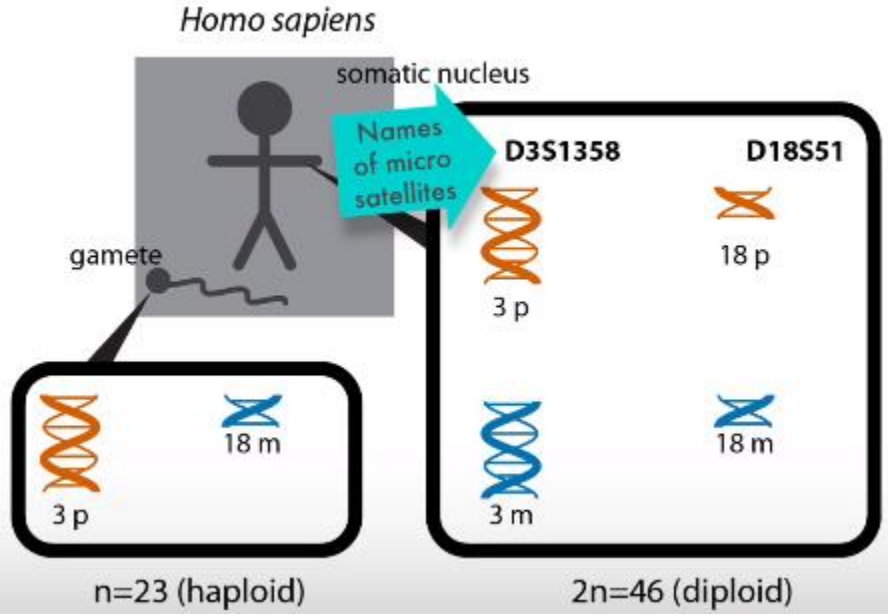
GQ: conditional genotype quality, encoded as a phred quality −10log10 p (genotype call is wrong, conditioned on the site's being variant) (Integer)

HQ: haplotype qualities, two comma separated phred qualities (Integers)

Understanding a Genotype and Haplotype representation

Here is a male human and representation of his gamete. Looking in a somatic cell, we'll see that this individual has 23 pairs of chromosomes. One set came from his mother and the other from his father.

Kanza Batool Haider
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 8/8/2018



For example: one chromosome is 3p (paternal) and the other is 3 m (maternal)

Other chromosome is 18p (paternal) and the other is 18m (maternal)

When we look inside the haplotype gamete, we see that it has just one copy of each chromosome. In this above example case, the particular gamete inherited 3p the paternal version of chromosome and the maternal version of chromosome 18. (And which parent's allele goes into the gamete is random)

We can look at the genotypes of two of these markers D3S1358 on chromosome 3 and D18S51 on chromosome 18

Writing a Genotype:

For a cell like somatic cell that contains multiple versions of one chromosome, we can write a genotype, which is a list if all alleles for the same (or homologous) chromosome.

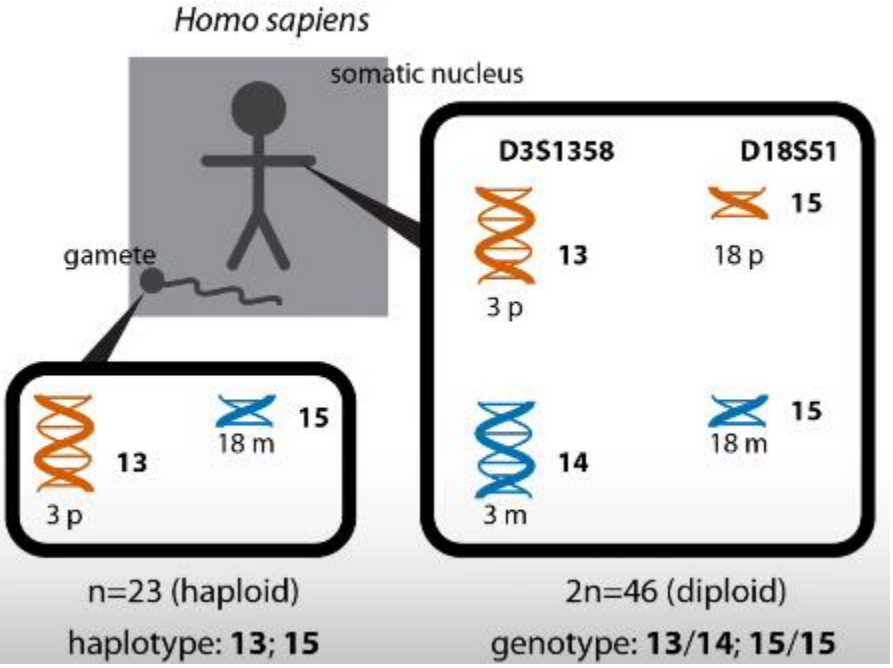
In this case, chromosome 3's genotype is **13/14** and chromosome 18's genotype is **15/15**

Writing a haplotype:

The haplotype gamete, because it doesn't have second copy of either of the chromosomes, contains no slashes- simply a semi colon to indicate the difference between one chromosome and the next: **13; 15**

In sum, a genotype is a list of all of the alleles for one or more loci inside of a somatic cell. In a haplotype cell like a gamete, there is no genotype. Rather there is a haplo- or half-type, listing the single allele on each chromosome present inside that cell.

Kanza Batool Haider
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 8/8/2018



Representing variation in VCF records

Creating VCF entries for SNPs and small indels

Example 1

For example, suppose we are looking at a locus in the genome:

Example Sequence Alteration

Example	Sequence	Alteration
Ref	a t C g a	C is the reference base
1	a t G g a	C base is a G in some individuals
2	a t - g a	C base is deleted w.r.t. the reference sequence
3	a t CAg a	A base is inserted w.r.t. the reference sequence

Representing these as VCF records would be done as follows:

1. A SNP polymorphism of C/G → {C, G} → C is the reference allele

2. A single base deletion of C → {tC, t} → tC is the reference allele

3. A single base insertion of A → {tC, tCA} → tC is the reference allele

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	3	.	C	G	.	PASS	DP=100
20	2	.	TC	T	.	PASS	DP=100
20	2	.	TC	TCA	.	PASS	DP=100

Example 2

Suppose I see a following in a population of individuals and want to represent these three segregating alleles:

Example	Sequence	Alteration
Ref	a t C g a	C is the reference base
1	a t G g a	C base is a G in some individuals
2	a t - g a	C base is deleted w.r.t. the reference sequence

In this case there are three segregating alleles: {tC, tG, t} with a corresponding VCF record:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	2	.	TC	TG,T	.	PASS	DP=100

Example 3

Now suppose I have this more complex example:

Example	Sequence	Alteration
Ref	a t C g a	C is the reference base
1	a t - g a	C base is is deleted w.r.t. the reference sequence
2	a t -- a	C and G bases are deleted w.r.t. the reference sequence
3	a t CAg a	A base is inserted w.r.t. the reference sequence

There are actually four segregating alleles: {tCg, tg, t, tCAg} over bases 2-4. This complex set of allele is represented in VCF as:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	2	.	TCG	TG,T,TCAG	.	PASS	DP=100

For completeness, VCF records are dynamically typed, so whether a VCF record is a SNP, Indel, Mixed, or Reference site depends on the properties of the alleles in the record.

Decoding VCF entries for SNPs and small indels

SNP VCF record

Suppose I receive the following VCF record:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	3	.	C	T	.	PASS	DP=100

This is a SNP since its only single base substitution and there are only two alleles so I have the two following segregating haplotypes:

Kanza Batool Haider
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 8/8/2018

Example	Sequence	Alteration
Ref	a t C g a	C is the reference base
1	a t T g a	C base is a T in some individuals

Insertion VCF record

Suppose I receive the following VCF record:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	3	.	C	CTAG	.	PASS	DP=100

This is an insertion since the reference base C is being replaced by C [the reference base] plus three insertion bases TAG. Again there are only two alleles so I have the two following segregating haplotypes:

Example	Sequence	Alteration
Ref	a t C - - - g a	C is the reference base
1	a t C T A G g a	following the C base is an insertion of 3 bases

Deletion VCF record

Suppose I receive the following VCF record:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	2	.	TCG	T	.	PASS	DP=100

This is a deletion of two reference bases since the reference allele TCG is being replaced by just the T [the reference base]. Again there are only two alleles so I have the two following segregating haplotypes:

Example	Sequence	Alteration
Ref	a T C G a	T is the (first) reference base
1	a T - - a	following the T base is a deletion of 2 bases

Mixed VCF record

For a microsatellite Suppose I receive the following VCF record:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	4	.	GCG	G,GCGCG	.	PASS	DP=100

This is a mixed type record containing a 2 base insertion and a 2 base deletion. There are three segregating alleles so I have the three following haplotypes:

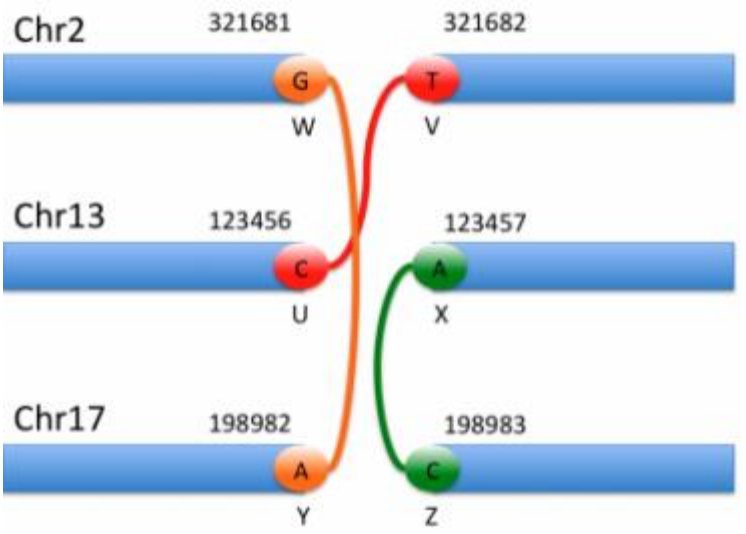
Example	Sequence	Alteration
Ref	a t c G C G - - a	G is the (first) reference base
1	a t c G - - - - a	following the G base is a deletion of 2 bases
2	a t c G C G C G a	following the G base is a insertion of 2 bases

Kanza Batool Haider
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 8/8/2018

Specifying complex rearrangements with breakends

An arbitrary rearrangement event can be summarized as a set of novel adjacencies. Each adjacency ties together 2 breakends. The two breakends at either end of a novel adjacency are called mates. There is one line of VCF (i.e. one record) for each of the two breakends in a novel adjacency. A breakend record is identified with the tag "SYTYPE=BND" in the INFO field.

The example below shows a 3-break operation involving 6 breakends. It exemplifies all possible orientations of breakends in adjacencies. Notice how the ALT field expresses the orientation of the breakends.

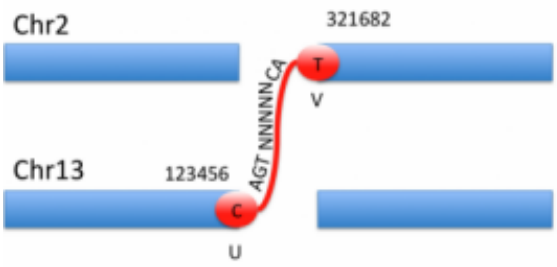


#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
2	321681	bnd_W	G	G]17:198982]	6	PASS	SVTYPE=BND
2	321682	bnd_V	T	]13:123456]T	6	PASS	SVTYPE=BND
13	123456	bnd_U	C	C[2:321682[	6	PASS	SVTYPE=BND
13	123457	bnd_X	A	[17:198983[A	6	PASS	SVTYPE=BND
17	198982	bnd_Y	A	A]2:321681]	6	PASS	SVTYPE=BND
17	198983	bnd_Z	C	[13:123457[C	6	PASS	SVTYPE=BND

Inserted Sequence

Sometimes, as shown in following example, some bases are inserted between the two breakends, this information is also carried in the ALT column:

Kanza Batool Haider
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 8/8/2018



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
2	321682	bnd_V	T	]13 : 123456]AGTNNNNNCAT	6	PASS	SVTYPE=BND;MATEID=bnd_U
13	123456	bnd_U	C	CAGTNNNNNCA[2 : 321682[	6	PASS	SVTYPE=BND;MATEID=bnd_V

For the following variations below, follow example from: http://samtools.github.io/hts-specs/VCFv4.2.pdf

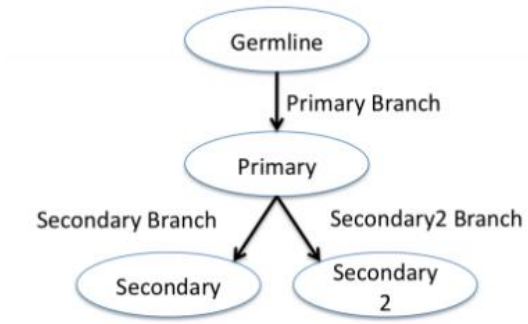
Large Insertions
Multiple mates
Explicit partners
Telomeres
Event modifiers
Inversions
Uncertainty around breakend location
Single breakends
Sample mixtures
Phasing adjacencies in an aneuploid context

Clonal derivation relationships

In cancer, each VCF file represents several genomes from a patient, but one genome is special in that it represents the germline genome of the patient. This genome is contrasted to a second genome, the cancer tumor genome. Hence, a vcf file has the capability to tell the way that a trait for some disease is passed down from one generation to another. In a VCF file, this is represented in the header with PEDIGREE tags.

See for example:

A cancer patient VCF file with 4 genomes: germline, primary tumor, secondary tumor1, and secondary tumor2 as illustrated in fig below. The primary tumor is derived from the germline and the secondary tumors are each derived independently from the primary tumor, in all cases by clonal derivation with mutations.



The PEDIGREE lines would look like:

```
##PEDIGREE=<Derived=PRIMARY-TUMOR-GENOME-ID,Original=GERMLINE-GENOME-ID>
##PEDIGREE=<Derived=SECONDARY1-TUMOR-GENOME-ID,Original=PRIMARY-TUMOR-GENOME-ID>
##PEDIGREE=<Derived=SECONDARY2-TUMOR-GENOME-ID,Original=PRIMARY-TUMOR-GENOME-ID>
```

Final Remarks and next up plan:

A vcf file is a very complicated yet integrated set of variant data. Most of the important chunks of information and understanding needed for a vcf file is defined in this report along with examples. However, the study of vcf demands more time and efforts.

For my next week's target, I plan to:

- Explore python vcf libraries such as the bio-vcf and pyvcf.
- Do the study on the different types of variants for Cancer diagnosis
- Calling variants vs. reference
- Downstream of variant calling
- VariantTools package
- Visualization of variants