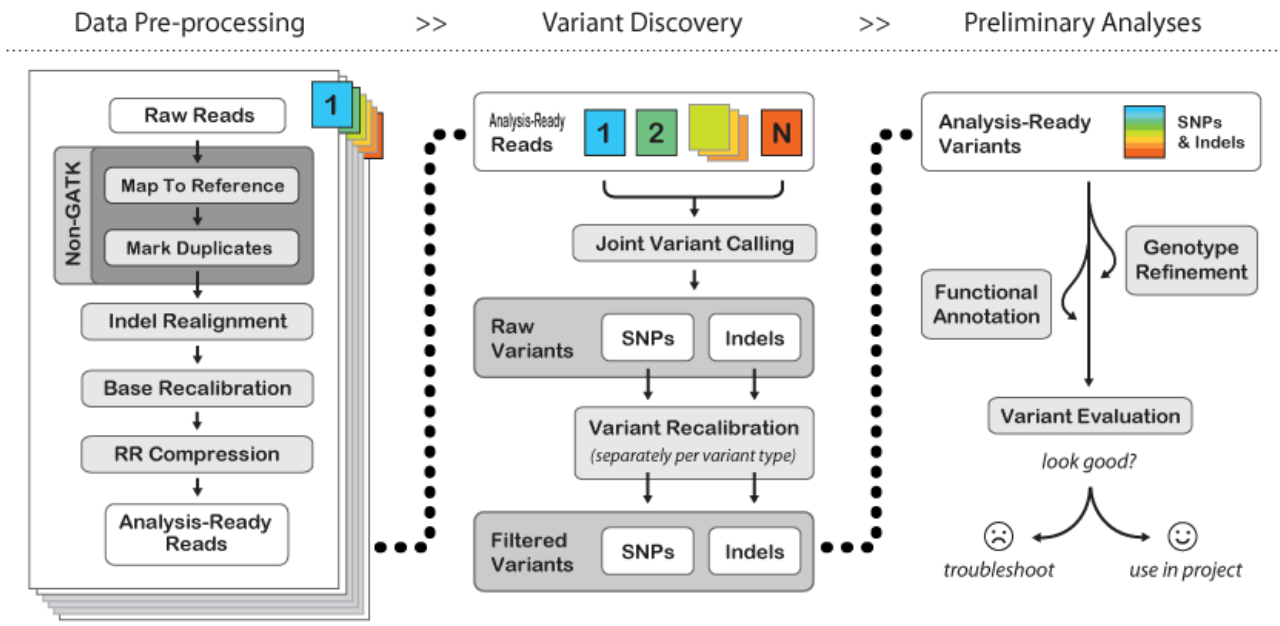
Working with Called Variants in *Bioconductor*

Martin Morgan (mtmorgan@fhcrc.org)

4 February 2014

Before *Bioconductor*: calling variants!

The following illustrates a variant calling work flow best practice; GATK[1]. Alternatives exist, e.g., *R* / *Bioconductor* *VariantTools*.



1 Variant Call Format (VCF) files

The Variant Call Format (VCF) is a complex file format to store information about variants. It is described in a specification[2].

```
##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">
```

[1] http://www.broadinstitute.org/gatk/guide/best-practices
[2] https://github.com/samtools/hts-specs

```
##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">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

According to the VCF specifcation, the REF and ALT columns from the example shows (in order): (a) a good simple SNP; (b) a possible SNP that has been filtered out because its quality is below 10; (c) a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error); (d) a site that is called monomorphic reference (i.e. with no alternate alleles); and (e) a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Evidence on which the variants have been called are in the INFO column.

Genotype data are given for three samples. Two of the samples are phased ('|') and the third unphased (''), with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

Alternatives formats exist, including MAF[3] (mutation annotation format).

2 Working with variants

The vignette Introduction to *VariantAnnotation* provides a great introduction to working with variants in *R*.

Exercise 1 *Explore section 2.1 of the vignette, to become comfortable with basic input and exploration of the data.*

Exercise 2 *Explore section 2.2 of the vignette to become familiar with selective input of the file.*

Exercise 3 *Explore the `readGeno` and `readInfo` functions for rapidly getting simple representations of specific VCF information.*

3 Looking for SNPs in regulatory regions

This section walks through a brief exercise illustrating how variant calls from one public data resource (the Human Genome Project) can be combined with regulatory information from a second public resource (the ENCODE) project to arrive at novel insights. It is based on section 5.1 of the `filterVcf`: Extract Variants of Interest from a Large VCF File vignette in the *VariantAnnotation* package.

Exercise 4 *This exercise does some preparation of a subset of a VCF file.*

- *Read the chr7 subset of data available in the VariantAnnotation package.*
- *Filter the data to identify simple SNPs – REF alleles of a single character; ALT alleles with only one allele, and with the allele a single character.*
- *What other filters might be appropriate? See section 4 of the `filterVcf` vignette for additional ideas.*

[3] https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format

Solution: Start with data from a a subset of chromosome 7

```
library(VariantAnnotation)
fl <- system.file("extdata", "chr7-sub.vcf.gz",
                  package="VariantAnnotation")
vcf <- readVcf(fl, "hg19")
```

Select just those variants that appear to be simple SNPs, with a single REF and ALT allele and a all ALT alleles a single character.

```
isSimpleRef <- nchar(ref(vcf)) == 1
isSimpleAlt <-
    ## 'sapply': apply a function to each ALT record
    ## 'length(elt) == 1: the record has only one allele...
    ## 'nchar(elt) == 1: ... and the allele is a single character
    sapply(alt(vcf), function(elt) (length(elt) == 1) && nchar(elt) == 1)
keep <- isSimpleRef & isSimpleAlt
vcf <- vcf[keep]
nrow(vcf)
```

```
## [1] 988
```

Exercise 5 *Follow section 5.1 of the* `filterVcf` *vignette to load CTCF transcription factor binding regions identified in MCF-7 Breast Cancer Cell Lines as GRanges objects.*

Exercise 6 *Follow section 5.2 of the* `filterVcf` *vignette to find SNPs in the CTCF binding region.*