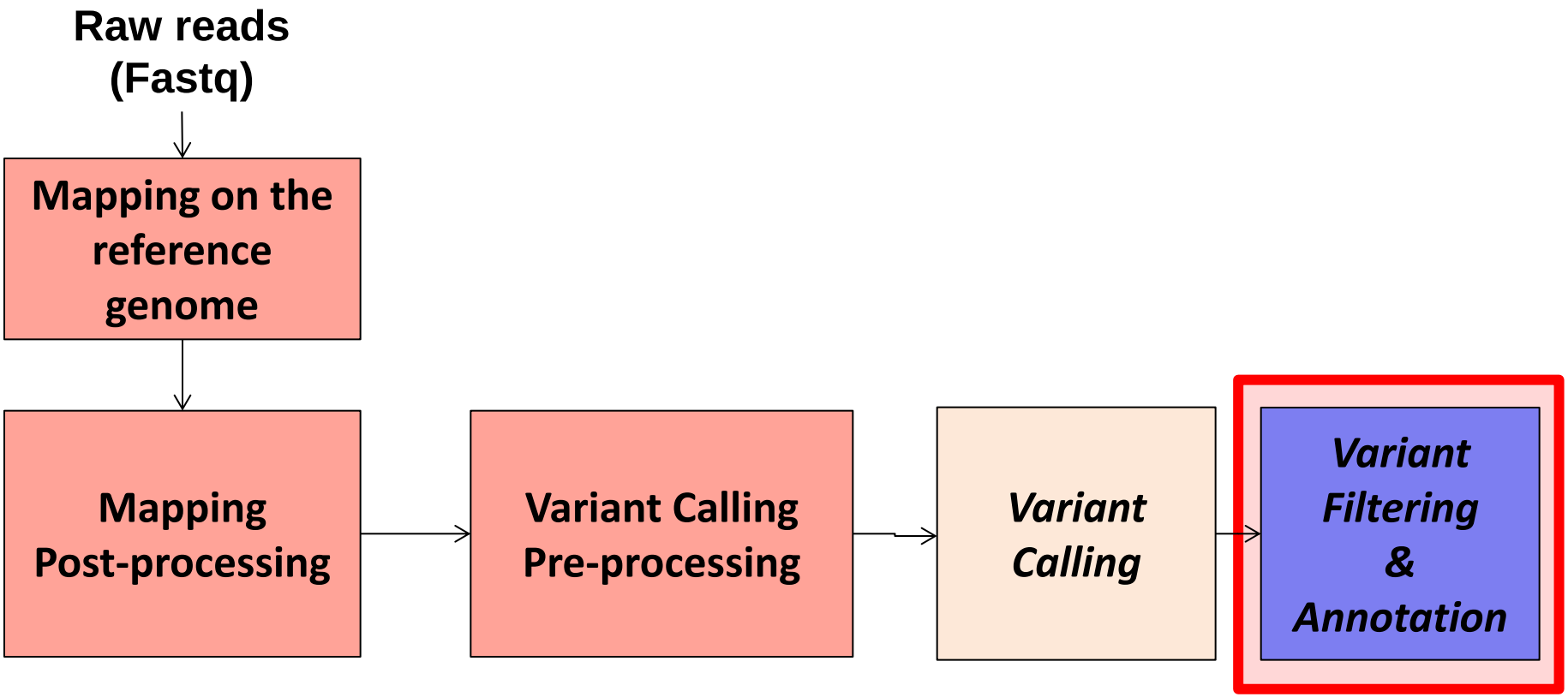
Variant Filtering

Ecole de Bioinformatique – Roscoff

octobre 2015

Workflow


Raw reads
(Fastq)
Mapping on the reference genome
Mapping Post-processing
Variant Calling Pre-processing
Variant Calling
Variant Filtering & Annotation

Why filter ?

Some use cases :

- Extract a subset of variants (localization, type)
- Combine variants from several analyses
- Compare obtained variants from several data types (RNA-Seq, Exome-Seq, Whole Genome)
- Identify new variants compare to a reference list
- Apply specific filter for Chip design
- …

HOWTO ?

Use specific tools to rewrite / annotate VCF File.

Reminder (VCF Format) :

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##FILTER=<ID=LowQual,Description="Low quality">
…
.
##contig=<ID=chr12,length=133851895>
##reference=file:///tmp/13905.1.galaxy.q/tmp-gatk-MPGS7G/gatk_input.fasta

#CHROM  POS     ID          REF  ALT  QUAL    FILTER  INFO                                                                                  FORMAT          Pickrell
Chr12   406292  rs2229351   G    A    994.77  .       AC=1;AF=0.500;AN=2;DB;DP=69;Dels=0.00;FS=4.853;MQ=37.00;MQ0=0;                        GT:AD:DP:GQ:PL  0/1:33,36:66:99:1023,0,994
Chr12   416046  rs35042439  C    CT   391.73  .       AC=1;AF=0.500;AN=2;DB;DP=46;FS=0.000;MQ=37.49;MQ0=0;QD=8.52;                          GT:AD:DP:GQ:PL  0/1:22,17:46:99:429,0,521
```

Which Tools ?

VcfTools (http://vcftools.sourceforge.net/)

- C++ /Perl
- Tool / Tool suite
  - **Filter**
  - **Convert**
  - **Compare**
  - **Stats (LD estimate, …)**
  - **Merge**
  - **Annotate**

- Available via Toolshed (limited implementation)

Which Tools ?

BcfTools (http://www.htslib.org/man/bcftools/) well suited to Samtools

- C++
- Tool suite:
  - **Annotate (edit, add annotations)**
  - **Concat (concatenate data from same sample)**
  - **Filter**
  - **Merge**
  - **Stats**
  - **View (subset, filter, convert)**
  - **Call (variant calling)**

- Available via Toolshed (limited implementation)

Which Tools ?

vcflib (https://github.com/ekg/vcflib) , well suited to FreeBayes

- C++ library
- Tool suite :
  - **Comparison (union, intersection, combine vcf files)**
  - **Format conversion (export to tsv, SQLite, Bed file)**
  - **Filtering (filter with expression, subsample, variant types)**
  - **Annotation (from other VCF File, Bed file, nearest variant)**
  - **Variant representation (complex variant, multiallellic)**
  - **Genotype manipulation (remove aberrant genotype, provide GL _> genolikelihood)**
  - **Classification of variants (heterozygosity, by annotation, pcr primers)**
- Available via Toolshed (well maintained)

Which Tools ?

GATK (https://www.broadinstitute.org/gatk/)

- Java
- Tool suite :
  - **Filtering (filter with expression, …)**
  - **Selection (criteria, reference file)**
  - **Annotation (reference file, VCF specific tag)**
  - **Comparison (union, intersection, combine vcf files)**
  - **Evaluation (report, stratification)**
  - **…**

- Available via Toolshed (well maintained)

HOWTO ?

- Remove variant entry or add Filter info (Hard Filtering):

Before Filtering

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Pickrell
Chr12	406292	rs2229351	G	A	994.77	.	AC=1;AF=0.500;AN=2;DB;DP=69;Dels=0.00;FS=4.853;MQ=37.00;MQ0=0;	GT:AD:DP:GQ:PL	0/1:33,36:66:99:1023,0,994
Chr12	416046	rs35042439	C	CT	391.73	.	AC=1;AF=0.500;AN=2;DB;DP=46;FS=0.000;MQ=37.49;MQ0=0;QD=8.52;	GT:AD:DP:GQ:PL	0/1:22,17:46:99:429,0,521

Filter : MQ < 30.0

After Filtering

##FILTER=<ID=LowQual,Description="Low quality, mapping < 30.0">

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Pickrell
Chr12	406292	rs2229351	G	A	994.77	**PASS**	AC=1;AF=0.500;AN=2;DB;DP=69;Dels=0.00;FS=4.853;MQ=37.00	GT:AD:DP:GQ:PL	0/1:33,36:66:99:1023,0,994
Chr12	416046	rs35042439	C	CT	391.73	**PASS**	AC=1;AF=0.500;AN=2;DB;DP=46;FS=0.000;MQ=37.49	GT:AD:DP:GQ:PL	0/1:22,17:46:99:429,0,521

Filter : **PASS** (Not Filtered), **.** (no data, filtering not performed), **LowQual** (Filtered)

HOWTO ?

- Understand VCF Format File
- Identify specific tags
- Fix Thresholds
- Find external ressources (dbSNP) to exclude / keep known Variants (other VCF File)
- Limit analysis to specific genomic locations (BED File)

Which Criteria ?

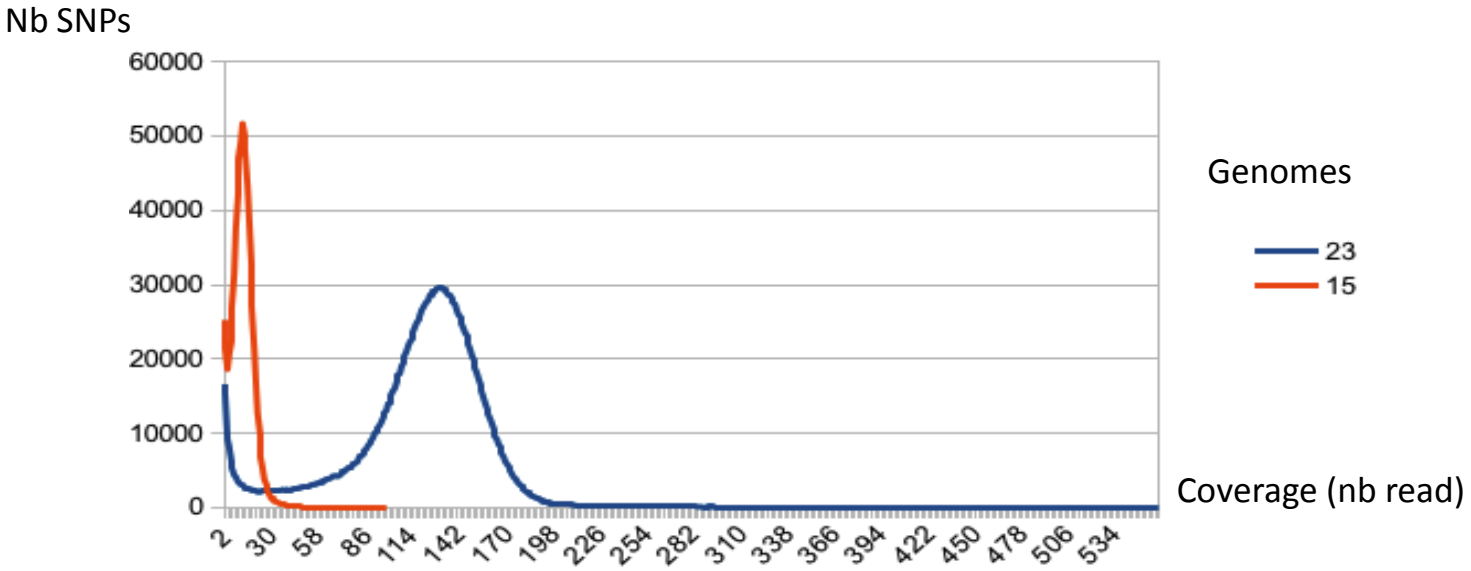
It depends of:

- Variant caller : methods, available info, VCF specific tags
- Data Type : DNA-Seq, Exome-Seq, RNA-Seq,
- Sequencing Technology : (depth, protocole)
- Reference genome: reliability of the reference sequence
- Studied species: Genome features (Transposable Elements, Tandem Repeats)
- Available ressources: reference variant sets

Which Criteria ?

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*

- Depth (DP): Min / Max => d ± 4√d, d = average Read Depth



- Low DP : mapping errors, sequencing errors
- High DP : CNVs or Repeat Regions, mapping errors
- Reliable with High coverage > 40X
- DNA-Seq OK, Exome-Seq NOK

Which Criteria ?

Li, H. (2014). *Bioinformatics*

- Low Complexity regions (LC) : exclude variants located in LC regions
- Variant Quality (QU) : exclude variants with low quality
- Double Strand filter (DS) : exclude variants with number of reads (ALT allele) below a defined Threshold on reverse or forward strand
- Fisher Strand filter (FS) : reference / no-reference reads highly correlated with strand.
- Allele Balance (AB) : HET > 30%

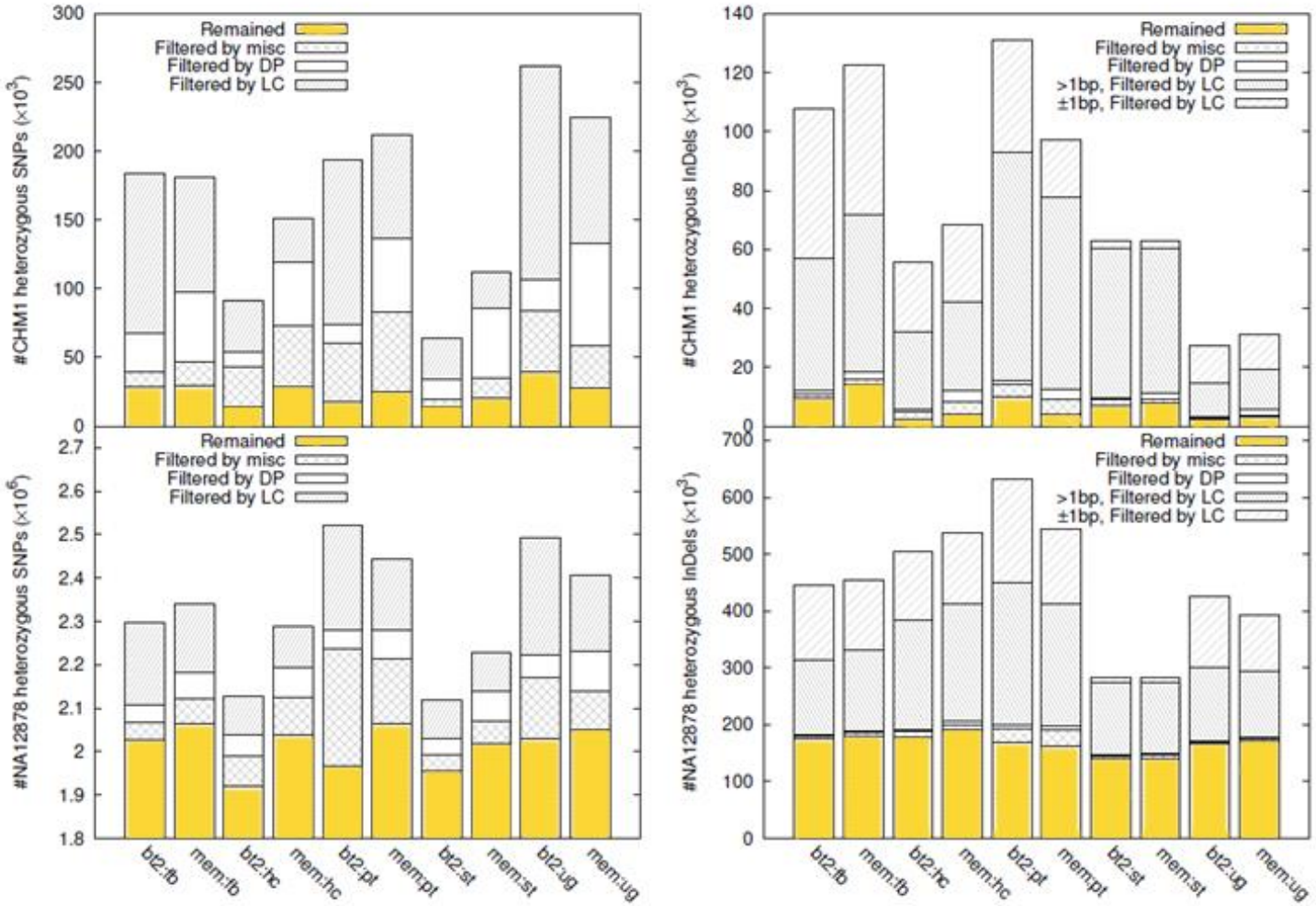
Meynert et al (2014). *BMC Bioinformatics*

- Coverage uniformity vs coverage depth (whole genome vs exome-seq) -> critical for heterozygous sites

Which Criteria ?

Haploid cells

Diploid cells



Li, H. (2014). *Bioinformatics,* misc = AB, DS, FS

GATK: VCF Format

FORMAT Tags

GT	Genotype , 0/0, 0/1, 1/1
GQ	Genotype Quality (Highest value = 99)
AD / DP	Depth per Allele / Depth = global coverage
PL	Genotype Likelihoods , max 0 (Phred Score)

INFO Tags

AC,AF,AN	(AC) Alleles Count, and (AF) Allele Frequency for each ALT allele, (AN)Total number of allele
DB	If present, then the variant is in dbSNP.
DP	Coverage (reads that passed quality metrics)
DS	Were any of the samples downsampled because of too much coverage?
MQ and MQ0	Root Mean Square Mapping Quality and Mapping Quality Zero total count
BaseQualityRankSum	Test : quality of Reference reads vs ALT reads
MappingQualityRankSum	Test : Mapping quality of Reference reads vs ALT reads
ReadPosRankSum	Test: Distance of ALT reads from the end of the reads
HaplotypeScore	Consistency of the site with at most two segregating haplotype
QD	Variant Quality / depth of non-ref samples
FS	Test (Fisher) : Phred score p-value for strand bias
InbreedingCoeff	Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation

http://gatkforums.broadinstitute.org/discussion/1268/how-should-i-interpret-vcf-files-produced-by-the-gatk

GATK: recommended filters

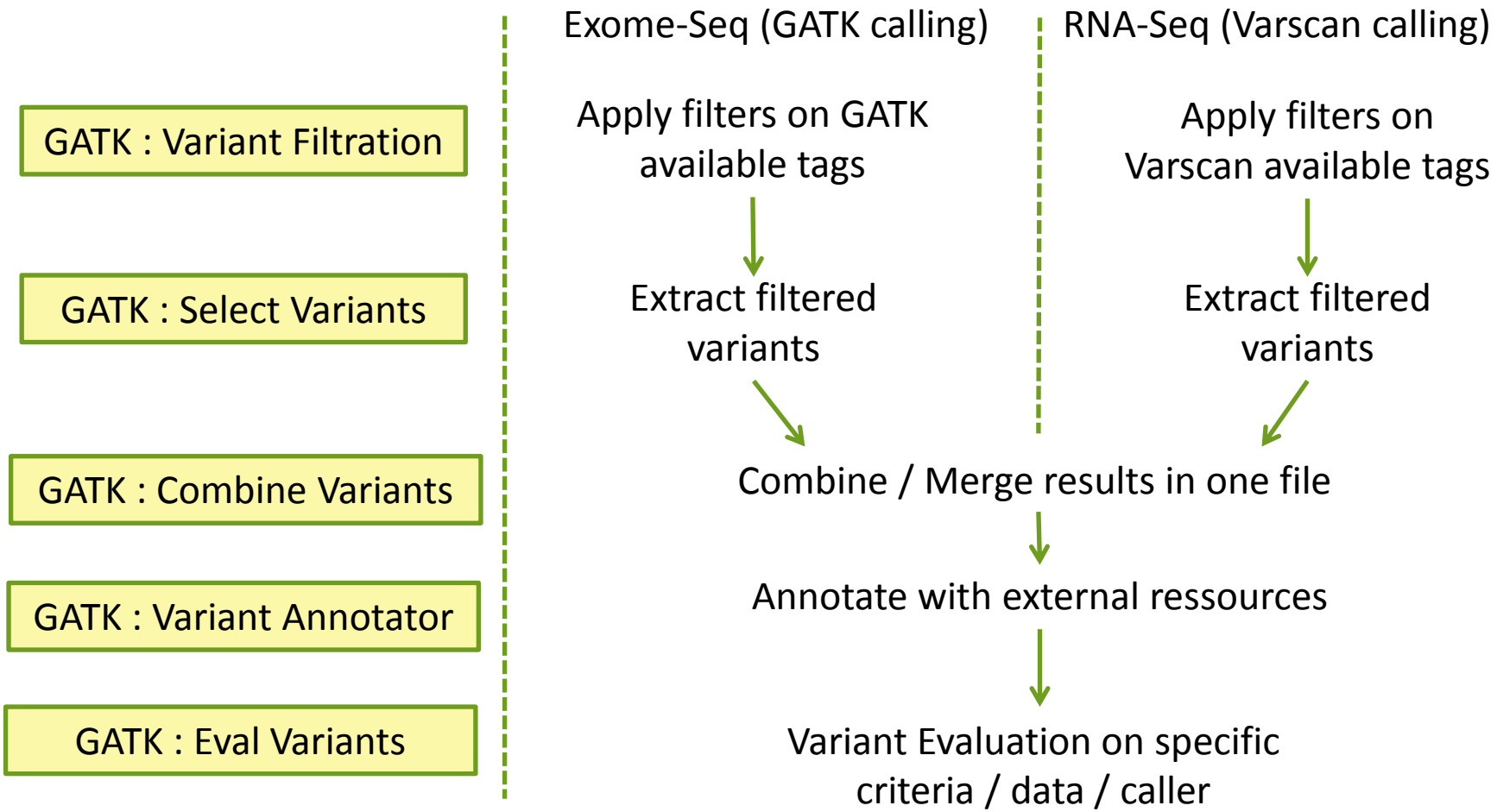
Use case : non-reference variant db, GATK recommended filters for recalibration

- SNPs:
  - **QD < 2.0** (Variant Quality / depth of non-ref samples)
  - **MQ < 40.0** (Mapping Quality)
  - **FS > 60.0** (Phred score Fisher's test p-value for strand bias)
  - **HaplotypeScore > 13.0** (Consistency of the site with at most two segregating haplotype)
  - **MQRankSum < -12.5** (Mapping quality of Reference reads vs ALT reads)
  - **ReadPosRankSum < -8.0** (Distance of ALT reads from the end of the reads)

- INDELs:
  - **QD < 2.0** (Variant Quality / depth of non-ref samples)
  - **ReadPosRankSum < -20.0** (Distance of ALT reads from the end of the reads)
  - **InbreedingCoeff < 0.8**
  - **FS > 200.0** (Phred score Fisher's test p-value for strand bias)

http://gatkforums.broadinstitute.org/discussion/3225/how-can-i-filter-my-callset-if-i-cannot-use-vqsr-recalibrate-variants

Tutorial


GATK : Variant Filtration
GATK : Select Variants
GATK : Combine Variants
GATK : Variant Annotator
GATK : Eval Variants
Exome-Seq (GATK calling)
RNA-Seq (Varscan calling)
Apply filters on GATK available tags
Apply filters on Varscan available tags
Extract filtered variants
Extract filtered variants
Combine / Merge results in one file
Annotate with external ressources
Variant Evaluation on specific criteria / data / caller

GATK: Variant Filtration

- Modify FILTER column (Hard Filtering)
- Criteria on INFO Tags
- Criteria on FORMAT Tags
- Handle missing Values

https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_filters_VariantFiltration.php

GATK: Select Variants

- Direct selection (exclude filtered variants)
- Criteria on INFO Tags
- Criteria on FILTER Tags
- No Criteria on FORMAT Tags
- Intersection / Union with other VCF Files
- Exclude / Include samples
- Selected genomic regions (BED File)

https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_variantutils_SelectVariants.php

Criteria: JEXL expression

JEXL = Java Expression Language

- Key, value
- Case-sensitive (Uppercase, Lowercase, MQ ≠ mq)
- Type-sensitive : `##FORMAT=<ID=AD,Number=.,Type=`**`Integer`**`,Description="..."`
  - **Integer = 2**
  - **Float = 2.0**
  - **String = "two"**
- Operators:
  - **Relational: ==, !=, <,>,<=, >=**
  - **Logical: && (AND) , || (OR)**

http://gatkforums.broadinstitute.org/discussion/1255/what-are-jexl-expressions-and-how-can-i-use-them-with-the-gatk

GATK: Combine Variants

- Combine Variant from VCF files:
  - **Several samples**
  - **Different methods**

- Add "set" tag in INFO column
  - **set = file1, unique to file1**
  - **set = Intersection, found in all files**

https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_variantutils_CombineVariants.php

GATK: Eval Variants

Calculate quality control metrics

- 1 / multiple vcf
- Stratification modules (Novelty, Sample,…)
- Metrics modules:
  - **CompOverlap: overlap between eval and comp sites**
  - **CountVariants: counts of variant classes in the sample**
  - **TiTvVariantEvaluator: Transition/Transversion Variant Evaluator**

https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_varianteval_VariantEval.php

GATK: Eval Variants

Example: Human expected Ti/Tv :

- **Ti (A<->G, C<->T) twice as frequently as Tv (A<->C, A<->T, G<->C,G<->T)**
- **Ti/Tv > 3 in coding regions (exome)**

http://genome.sph.umich.edu/wiki/SNP_Call_Set_Properties

Eval Report:

#:GATKTable:14:12:%s:%s:%s:%s:%s:%d:%d:%.2f:%d:%d:%.2f:%d:%d:%.2f:

TiTvVariantEvaluator	CompRod	EvalRod	JexlExpression	Novelty	nTi	nTv	tiTvRatio	nTiInComp	nTvInComp	TiTvRatioStandard
TiTvVariantEvaluator	dbsnp	input_0	Gatk	all	640	241	2.66	634	241	2.63
TiTvVariantEvaluator	dbsnp	input_0	Gatk	known	636	239	2.66	633	241	2.63
TiTvVariantEvaluator	dbsnp	input_0	Gatk	novel	4	2	2.00	1	0	1.00
TiTvVariantEvaluator	dbsnp	input_0	Intersection	all	95	30	3.17	88	29	3.03
TiTvVariantEvaluator	dbsnp	input_0	Intersection	known	95	30	3.17	88	29	3.03
TiTvVariantEvaluator	dbsnp	input_0	Intersection	novel	0	0	0.00	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	Varscan	all	192	53	3.62	162	45	3.60
TiTvVariantEvaluator	dbsnp	input_0	Varscan	known	172	44	3.91	162	44	3.68
TiTvVariantEvaluator	dbsnp	input_0	Varscan	novel	20	9	2.22	0	1	0.00
TiTvVariantEvaluator	dbsnp	input_0	none	all	927	324	2.86	1398900	692871	2.02
TiTvVariantEvaluator	dbsnp	input_0	none	known	903	313	2.88	883	314	2.81
TiTvVariantEvaluator	dbsnp	input_0	none	novel	24	11	2.18	1398017	692557	2.02

PRACTICAL !

Appendix: Phred Score

$$Q = -10 \log_{10} P \quad \text{and} \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Appendix: filters

- Note that the InbreedingCoeff statistic is a population-level calculation that is only available with 10 or more samples. If you have fewer samples you will need to omit that particular filter statement.
- **For shallow-coverage (<10x): you cannot use filtering to reliably separate true positives from false positives. You must use the protocol involving variant quality score recalibration.**
- The maximum DP (depth) filter only applies to whole genome data, where the probability of a site having exactly N reads given an average coverage of M is a well-behaved function. First principles suggest this should be a binomial sampling but in practice it is more a Gaussian distribution. Regardless, the DP threshold should be set a 5 or 6 sigma from the mean coverage across all samples, so that the DP > X threshold eliminates sites with excessive coverage caused by alignment artifacts. Note that **for exomes, a straight DP filter shouldn't be used** because the relationship between misalignments and depth isn't clear for capture data.
- That said, all of the caveats about determining the right parameters, etc, are annoying and are largely eliminated by variant quality score recalibration.
- https://www.broadinstitute.org/gatk/guide/article?id=3225
- http://gatkforums.broadinstitute.org/discussion/2806/howto-apply-hard-filters-to-a-call-set