Variant calling with GATK

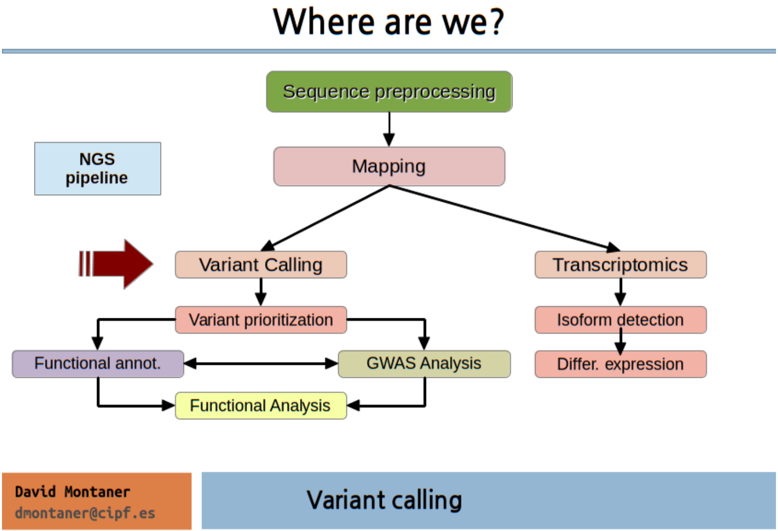
Estudios in silico en Biomedicina

(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2014-11-11

Overview



Where are we?

NGS pipeline

Sequence preprocessing

Mapping

Variant Calling

Transcriptomics

Variant prioritization

Isoform detection

Functional annot.

GWAS Analysis

Differ. expression

Functional Analysis

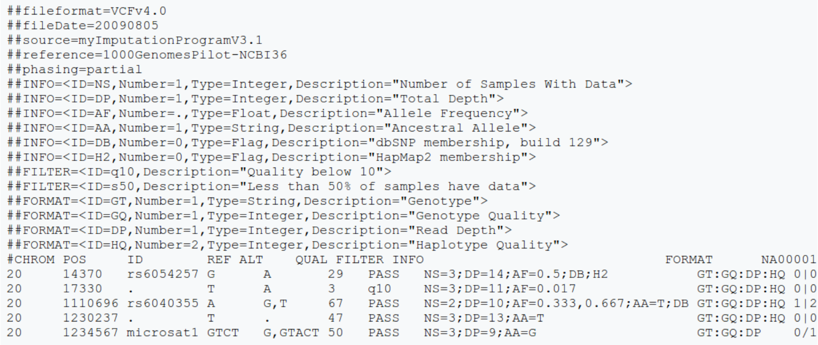
David Montaner
dmontaner@cipf.es

Variant calling

Genomic Variation

- SNPs / single nucleotide variants
- Structural Variants:
  - CNV: Copy number variable regions
    - Deletions
    - Duplications
  - Insertions
  - Inversions
  - Translocations
  - Inversions

File Format



```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##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=.,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">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0
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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1
```

VCF file format

- CHROM: chromosome
- POS: position
- ID: name
- REF: reference base(s)
- ALT: non-reference alleles
- QUAL: quality score of the calls (phred scale)
- FILTER: PASS / filtering_tag
- INFO: additional information
- FORMAT: describes further extra columns

VCF file format: INFO

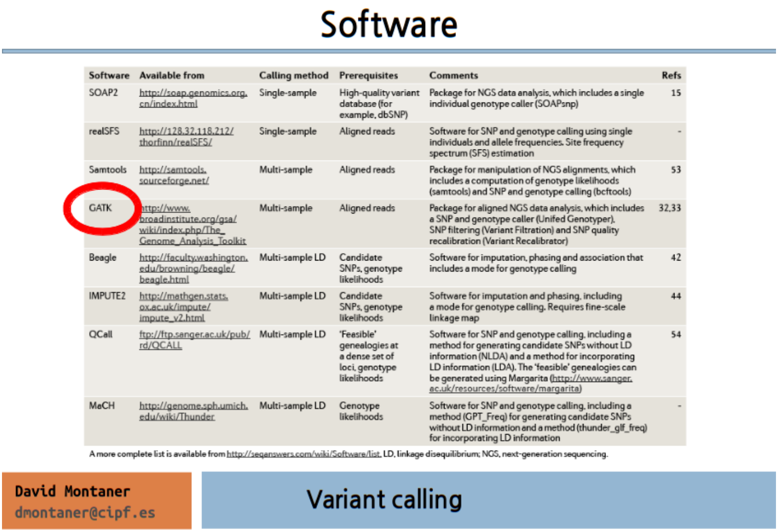
INFO column: semicolon separated fields

`<key>=<data>[,data]`

Some reserved (but optional) keys:

- AA ancestral allele
- AC allele count in genotypes, for each ALT allele, in the same order as listed
- AF allele frequency
- CIGAR cigar string describing how to align an alternate allele to the reference allele
- DB dbSNP membership
- 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

Software



Software

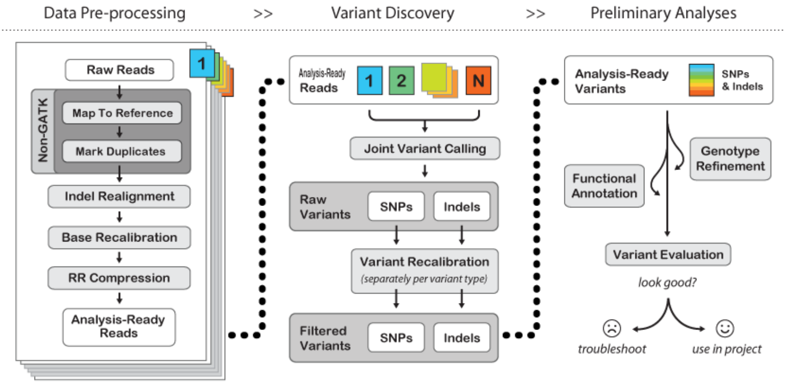
Software	Available from	Calling method	Prerequisites	Comments	Refs
SOAP2	http://soap.genomics.org.cn/index.html	Single-sample	High-quality variant database (for example, dbSNP)	Package for NGS data analysis, which includes a single individual genotype caller (SOAPsnp)	15
realSFS	http://128.32.118.212/thorfinn/realSFS/	Single-sample	Aligned reads	Software for SNP and genotype calling using single individuals and allele frequencies. Site frequency spectrum (SFS) estimation	-
Samtools	http://samtools.sourceforge.net/	Multi-sample	Aligned reads	Package for manipulation of NGS alignments, which includes a computation of genotype likelihoods (samtools) and SNP and genotype calling (bcftools)	53
GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit	Multi-sample	Aligned reads	Package for aligned NGS data analysis, which includes a SNP and genotype caller (Unifed Genotyper), SNP filtering (Variant Filtration) and SNP quality recalibration (Variant Recalibrator)	32,33
Beagle	http://faculty.washington.edu/browning/beagle/beagle.html	Multi-sample LD	Candidate SNPs, genotype likelihoods	Software for imputation, phasing and association that includes a mode for genotype calling	42
IMPUTE2	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Multi-sample LD	Candidate SNPs, genotype likelihoods	Software for imputation and phasing, including a mode for genotype calling. Requires fine-scale linkage map	44
QCall	ftp://ftp.sanger.ac.uk/pub/rd/QCALL	Multi-sample LD	'Feasible' genealogies at a dense set of loci, genotype likelihoods	Software for SNP and genotype calling, including a method for generating candidate SNPs without LD information (NLDA) and a method for incorporating LD information (LDA). The 'feasible' genealogies can be generated using Margarita (http://www.sanger.ac.uk/resources/software/margarita)	54
MaCH	http://genome.sph.umich.edu/wiki/Thunder	Multi-sample LD	Genotype likelihoods	Software for SNP and genotype calling, including a method (GPT_Freq) for generating candidate SNPs without LD information and a method (thunder_glf_freq) for incorporating LD information	-

A more complete list is available from http://seqanswers.com/wiki/Software/list. LD, linkage disequilibrium; NGS, next-generation sequencing.

David Montaner
dmontaner@cipf.es

Variant calling

GATK Best Practices work flow



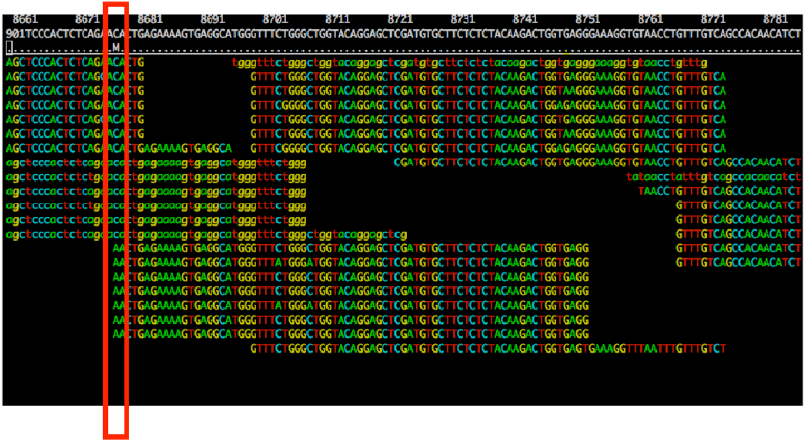
Mark duplicates

- All NGS sequencing platforms are NOT single molecule sequencing
- PCR -> duplicate DNA fragments in the final library.
- If there is a base variation it will have high depth support
- Can result in false SNP calls

Tools

- Samtools: samtools rmdup or samtools rmdupse
- Picard/GATK: MarkDuplicates

Duplicated induce biased SNP calls

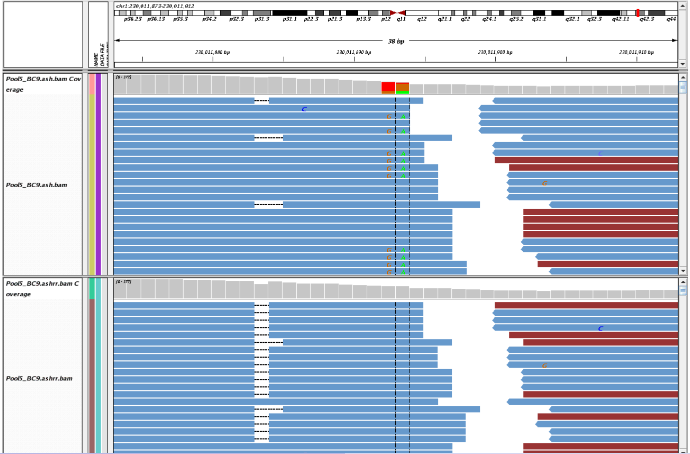


```
 8661      8671      8681      8691      8701      8711      8721      8731      8741      8751      8761      8771      8781
901TCCCACTCTCAGAACACTGAGAAAAGTGAGGCATGGGTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGGGAAAGGTGTAACCTGTTTGTCAGCCACAACATCT
.................M.............................................................................................................
AGCTCCCACTCTCAGAACACTG              tgggtttctgggctggtacaggagctcgatgtgcttctctctacaagactggtgagggaaaggtgtaacctgtttg
AGCTCCCACTCTCAGAACACTG                 GTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGGGAAAGGTGTAACCTGTTTGTCA
AGCTCCCACTCTCAGAACACTG                 GTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTAAGGGAAAGGTGTAACCTGTTTGTCA
AGCTCCCACTCTCAGAACACTG                 GTTTCGGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGAGAGGGAAAGGTGTAACCTGTTTGTCA
AGCTCCCACTCTCAGAACACTG                 GTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGGGAAAGGTGTAACCTGTTTGTCA
AGCTCCCACTCTCAGAACACTG                 GTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTAAGGGAAAGGTGTAACCTGTTTGTCA
AGCTCCCACTCTCAGAACACTGAGAAAAGTGAGGCA   GTTTCGGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGAGAGGGAAAGGTGTAACCTGTTTGTCA
agctcccactctcagaacactgagaaaagtgaggcatgggtttctggg              CGATGTGCTTCTCTCTACAAGACTGGTGAGGGAAAGGTGTAACCTGTTTGTCAGCCACAACATCT
agctcccactctctgaacactgagaaaagtgaggcatgggtttctggg                                                   tataacctatttgtcagccacaacatct
agctcccactctcagaacactgagaaaagtgaggcatgggtttctggg                                                     TAACCTGTTTGTCAGCCACAACATCT
agctcccactctctgaacactgagaaaagtgaggcatgggtttctggg                                                           GTTTGTCAGCCACAACATCT
agctcccactctcagaacactgagaaaagtgaggcatgggtttctggg                                                           GTTTGTCAGCCACAACATCT
agctcccactctcagaacactgagaaaagtgaggcatgggtttctgggctggtacaggagctcg                                           GTTTGTCAGCCACAACATCT
                 AACTGAGAAAAGTGAGGCATGGGTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGG              GTTTGTCAGCCACAACATCT
                 AACTGAGAAAAGTGAGGCATGGGTTTATGGGATGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGG              GTTTGTCAGCCACAACATCT
                 AACTGAGAAAAGTGAGGCATGGGTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGG
                 AACTGAGAAAAGTGAGGCATGGGTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGG
                 AACTGAGAAAAGTGAGGCATGGGTTTATGGGATGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGG
                 AACTGAGAAAAGTGAGGCATGGGTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGG
                 AACTGAGAAAAGTGAGGCATGGGTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGG
                                       GTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGTGAAAGGTTTAATTTGTTTGTCT
```

INDEL Realignment

Local realignment of all reads at a specific location simultaneously to minimize mismatches to the reference genome.

Reduces erroneous SNPs refines location of INDELS.



Base quality re-calibration

Re-calibrate base quality scores in order to correct sequencing errors and other experimental artifacts:

- Analyze patterns of covariation in the sequence data: creates a report that will be used later.
- Generate before/after plots: check the effect before you apply it to your sequence data.
- Apply the re-calibration to your sequence data: transform your BAM files.
- Requires a reference genome and a catalog of known variable sites.
- The known sites are used to build the covariation model and estimate empirical base qualities.

Calling: GATK

- Probabilistic method: Bayesian estimation of the most likely genotype.
- Calculates many parameters for each position of the genome.
- SNP and indel calling.
- Used in many NGS projects, including the 1000 Genomes Project, The Cancer
- Genome Atlas, etc.
- Base quality re-calibration.
- Indel realignment
- Uses standard input and output files.
- Many tools for manage VCF files.
- Multi-sample calling
- http://www.broadinstitute.org/gatk/