Parco Tecnologico Padano

Francesco Strozzi

Head of the Bioinformatics Core Facility

Practical Sessions
CIHEAM 2015 – Variation Calling

What is Variation Calling ?

The process to determine if the genomics sequence of a particular individual has differences compared to a reference genome.

These differences can be:

- Point mutations, involving only a single nucleotide change (SNPs)
- Structural variations, differences that span multiple bases (InDels or large rearrangements)

Variation Calling in short

Variations are called using two main approaches:

- Genotyping chips: Quick and cheap, but you can call only variations already annotated on the chip and it's designed just for SNPs
- NGS: More expensive but you can discover also new variations and analyse structural changes as well.

Variation Calling: the NGS way

So you have sequenced your samples….

- First step is to QC your samples (e.g. using FastQC)
- Then, clean your reads! (e.g. using Trimmomatic or Sickle)
- After this, map the reads on your reference genome (use BWA)

Variation Calling: the NGS way

So you now have a BAM file….

- Sort it !! (Using Samtools)
- Looks for duplicated reads and be sure to mark them (using Samtools or Picard)
- Realign the reads around the InDels (Using GATK)
- Recalibrate the quality scores (Using GATK)

Variation Calling: the NGS way

So you now have a BAM file properly prepared…

Which caller do I need to use ? Here is where science and myth will cross their roads.

There are many callers, all equally good, tested and widely used. The main ones are:

- Samtools (http://www.htslib.org/doc/samtools-1.2.html)
- GATK (https://www.broadinstitute.org/gatk/)
- FreeBayes (https://github.com/ekg/freebayes)

Other tools exists out there, the better approach is to download them and try for yourself!

Variation Calling: the NGS way

And now a bit of theory…

Bayesian model

$$\Pr\{G|D\} = \frac{\Pr\{G\}\Pr\{D|G\}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

Prior of the genotype

Likelihood of the genotype

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2} \right) \text{ where } G = H_1 H_2$$

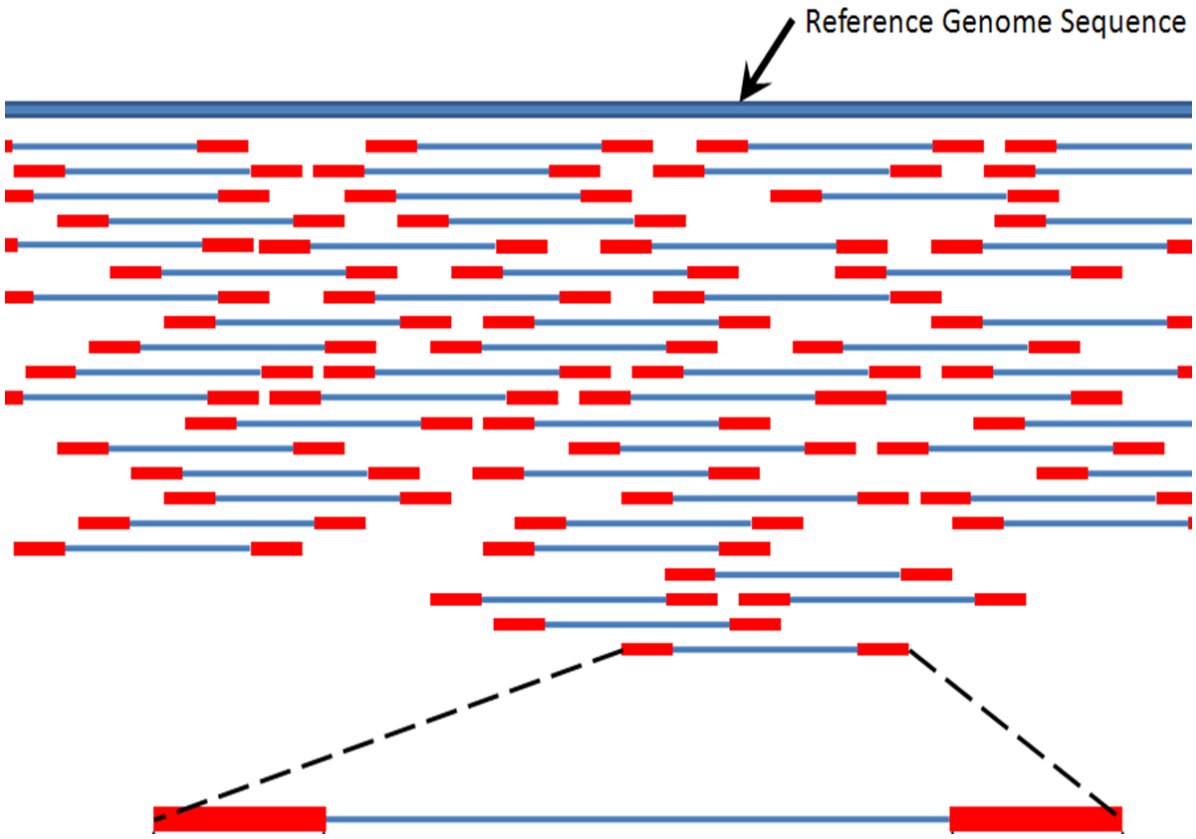
Diploid assumption

$$\Pr\{D|H\} \text{ is the haploid likelihood function}$$

Inference: what is the genotype G of each sample given read data D for each sample?

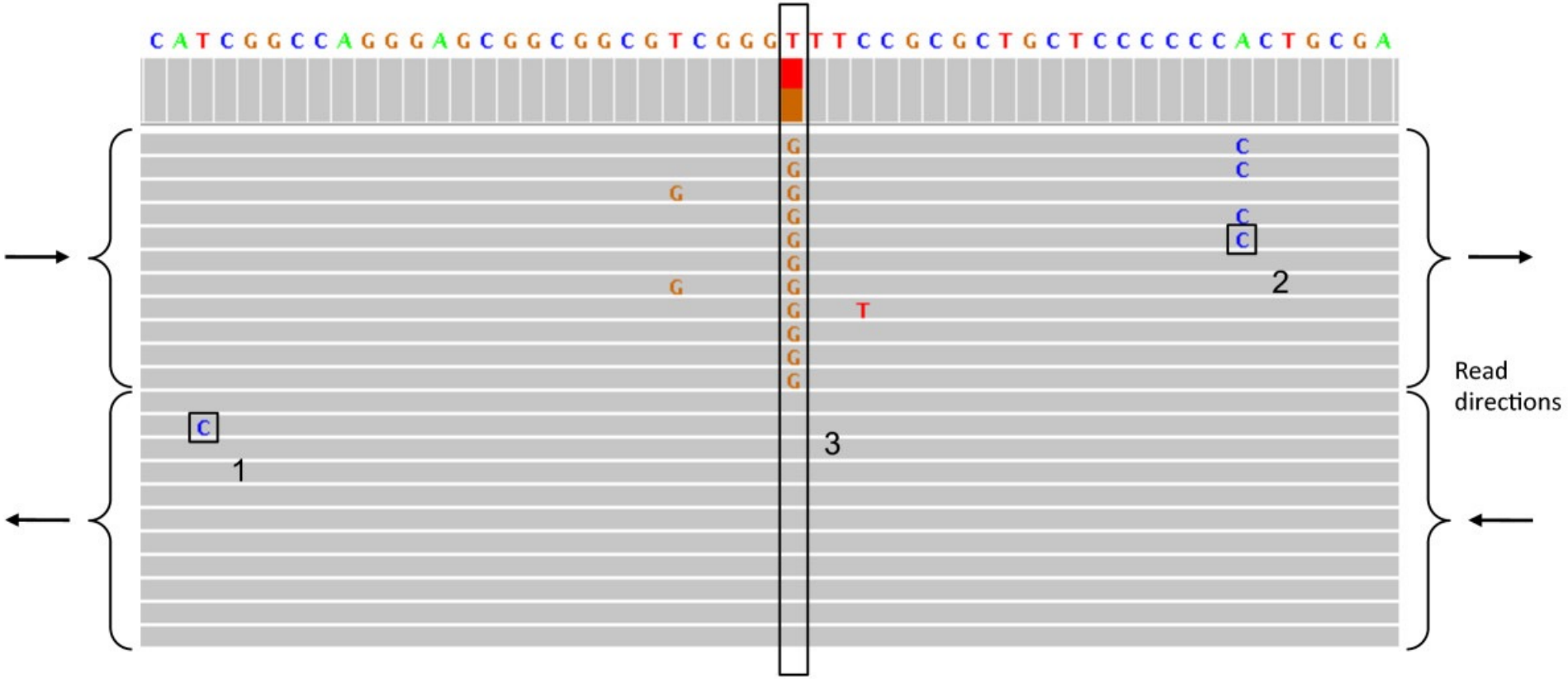
Parco Tecnologico Padano

Variation Calling: the NGS way



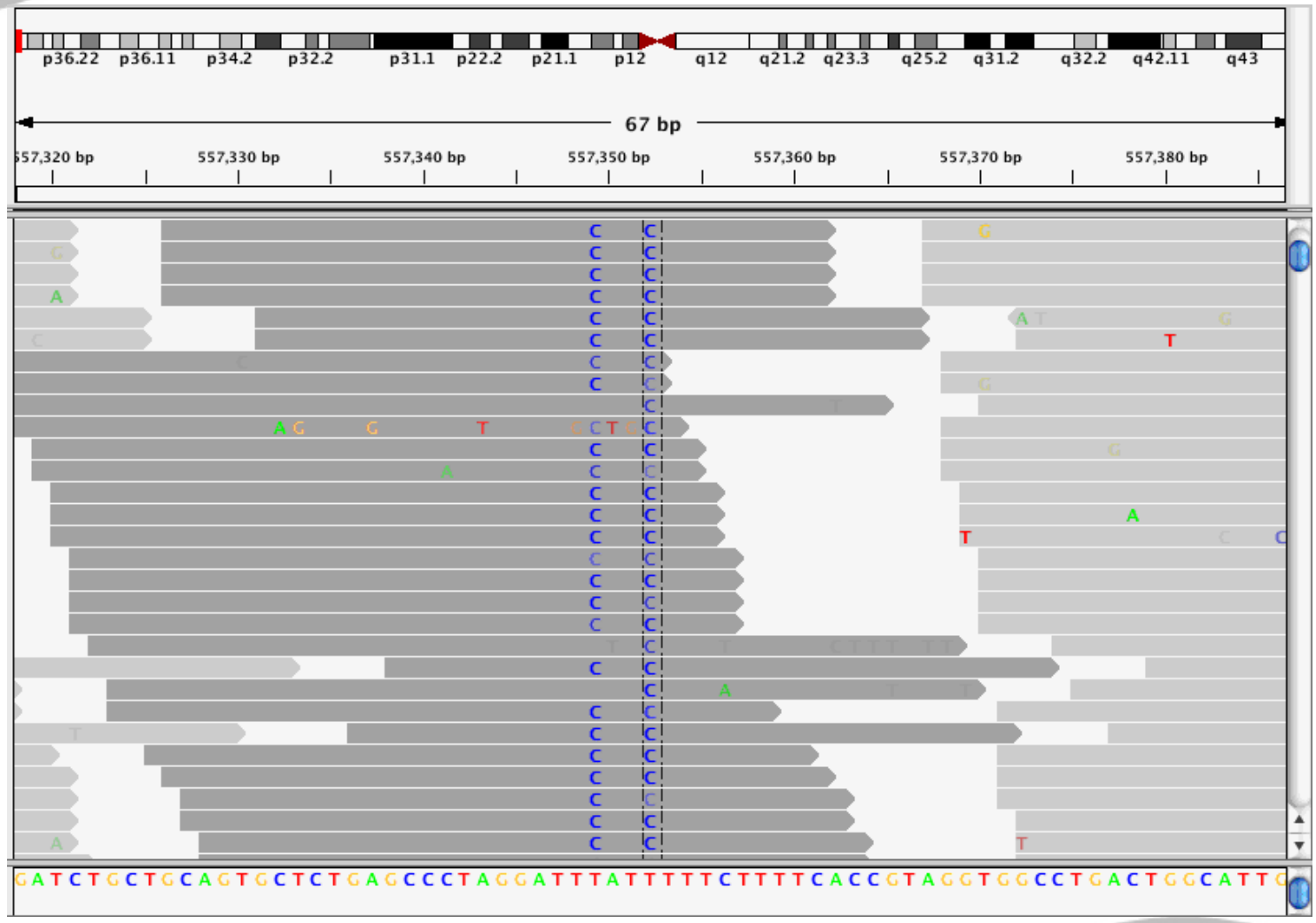
Parco Tecnologico Padano

Variation Calling: the NGS way



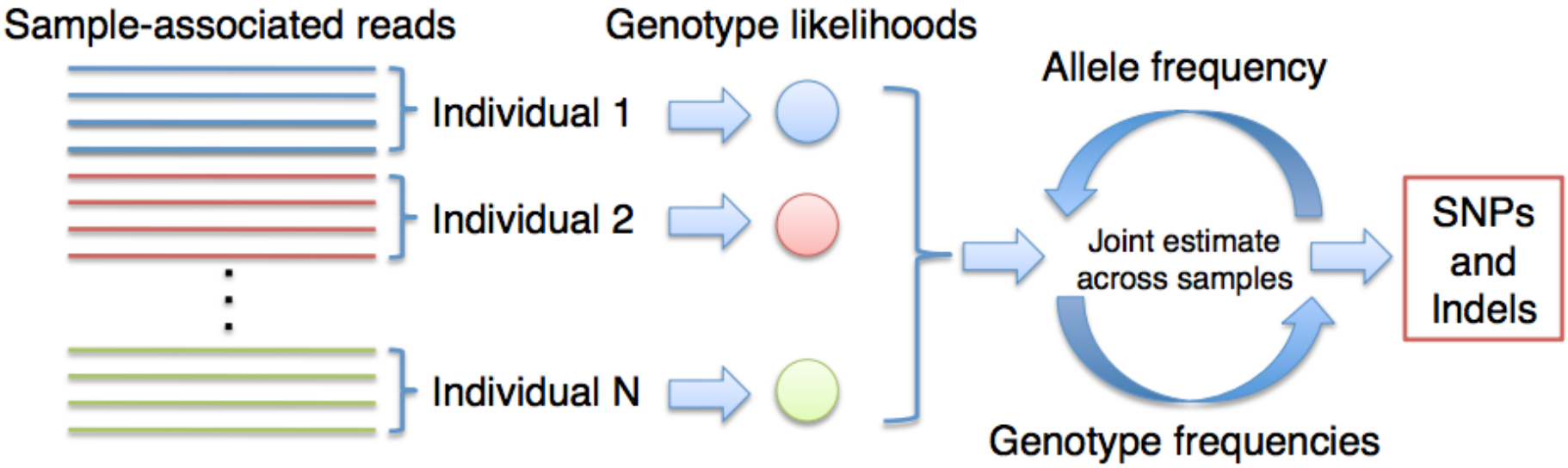
Parco Tecnologico Padano

Variation Calling: the NGS way

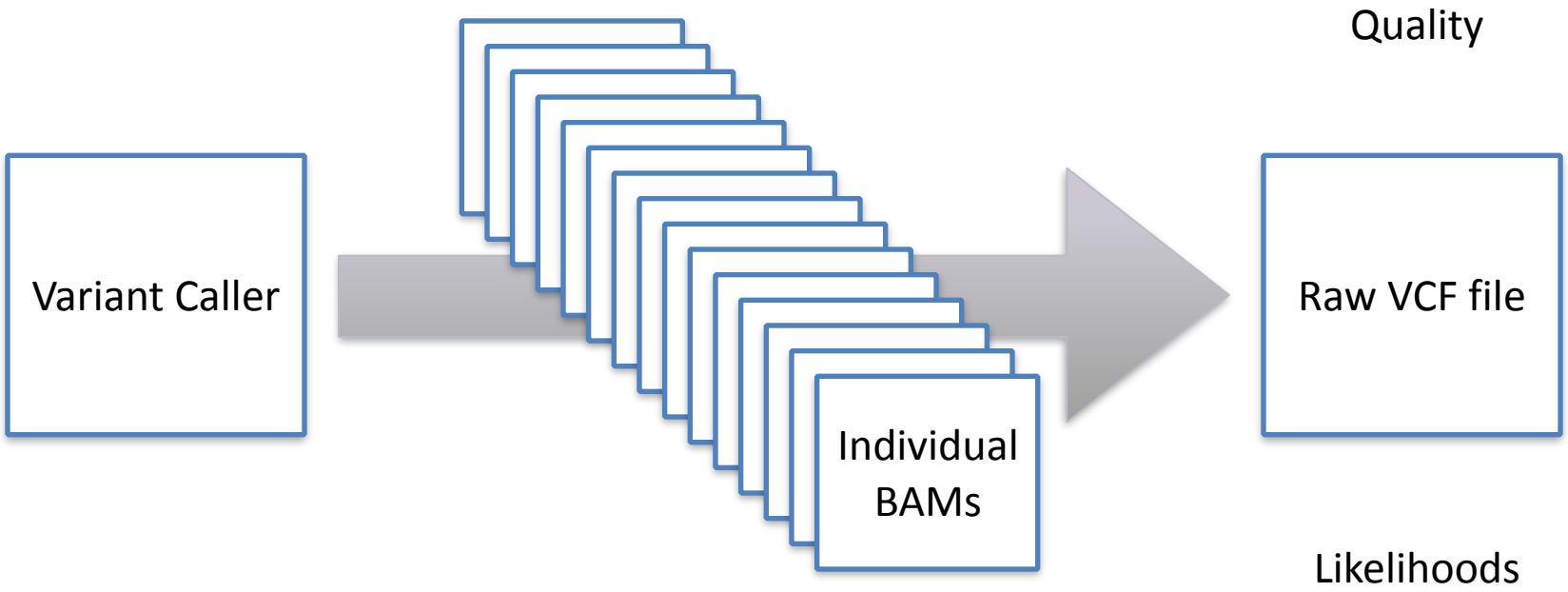


Parco Tecnologico Padano

Variation Calling: the NGS way



Variation Calling: the NGS way



A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Li H. Bioinformatics 2011*

A framework for variation discovery and genotyping using next-generation DNA sequencing data. *DePristo M. et al. Nature Genetics 2011*

Parco Tecnologico Padano

Variation Calling: the VCF

```
##fileformat=VCFv4.1
##fileDate=20140903
##source=freeBayes v9.9.2-29-g9ed353c
##reference=/storage/genomes/bt_umd31/Bos_taurus.UMD3.1.68.dna.toplevel.fa
##phasing=none
##commandline="/storage/software/freebayes-0.9.10/bin/freebayes […]

##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">


##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality, the Phred-scaled marginal (or unconditional) probability of the
called genotype">
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype Likelihood, log10-scaled likelihoods of the data given the called genotype for
each possible genotype generated from the reference and alternate alleles given the sample ploidy">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">
##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of quality of the reference observations">
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">
##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of quality of the alternate observations">


#CHROM  POS    ID     REF    ALT    QUAL    FILTER  INFO    FORMAT  Sample_XYZ
1       340    .      G      A      342.963 .       AB=0.764706;ABP=13.3567;AC=1;AF=0.5;AN=2;AO=13;CIGAR=1X;DP=17 […]   GT:DP:RO:QR:AO:QA:GL
0/1:17:4:142:13:497:-10,0,-7.83057
```

Variation Calling: the VCF

Things you can do with a VCF…

- A VCF file holds all the variations calling information for one or multiple samples
- It can be compressed using tools such as *bgzip*, specifically designed to handle large datasets
- It can be indexed, to improve access to the information
- VCF files are a standard format used by public databases and international initiatives (e.g. 1000 genomes)
- Tools performing variations filtering and comparisons works directly with VCF files

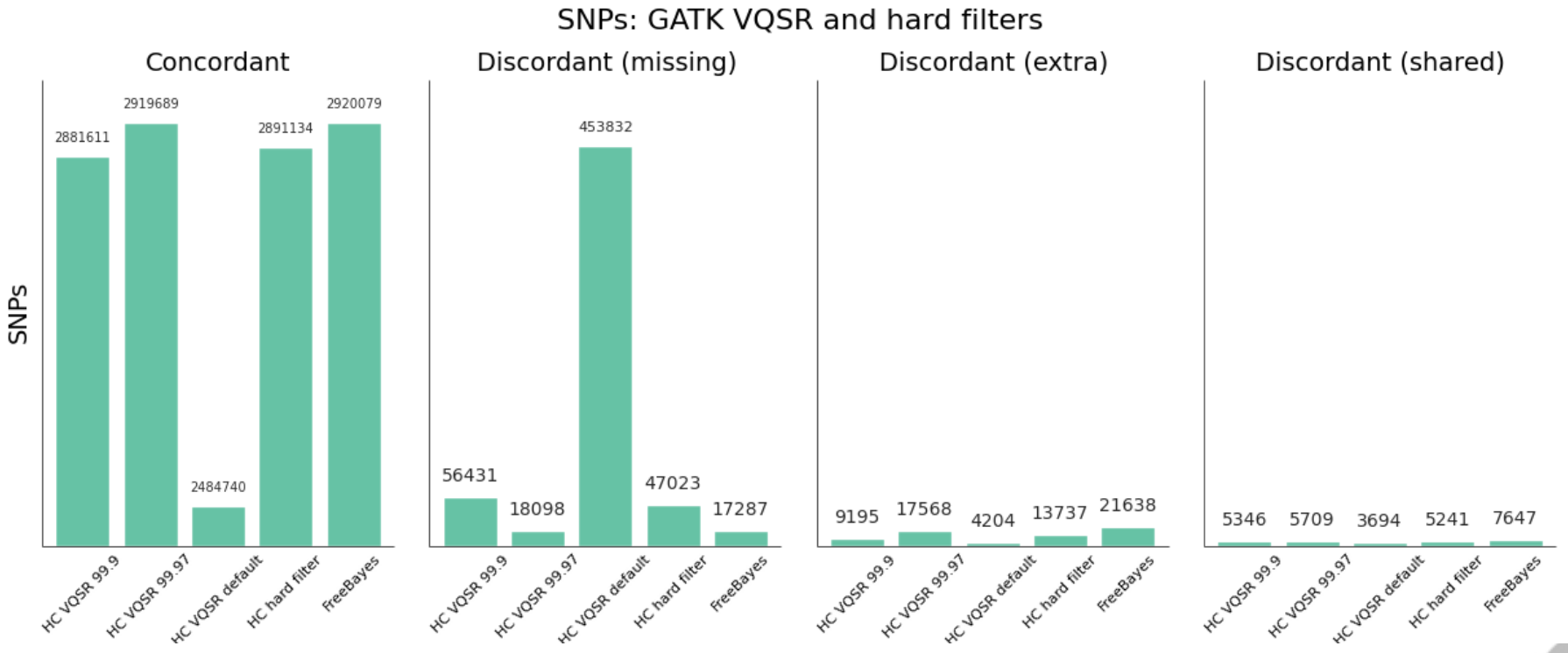
Variation Calling: the callers war

GATK	FreeBayes	SamTools
Developed by Broad Institute	Developed by Marth Lab Boston College	Community based development

Benefits for variation calling analysis:

- Some “competition” helped improving the algorithms and performances
- Few years ago there was a high divergence among variation callers
- Results concordance has now dramatically improved

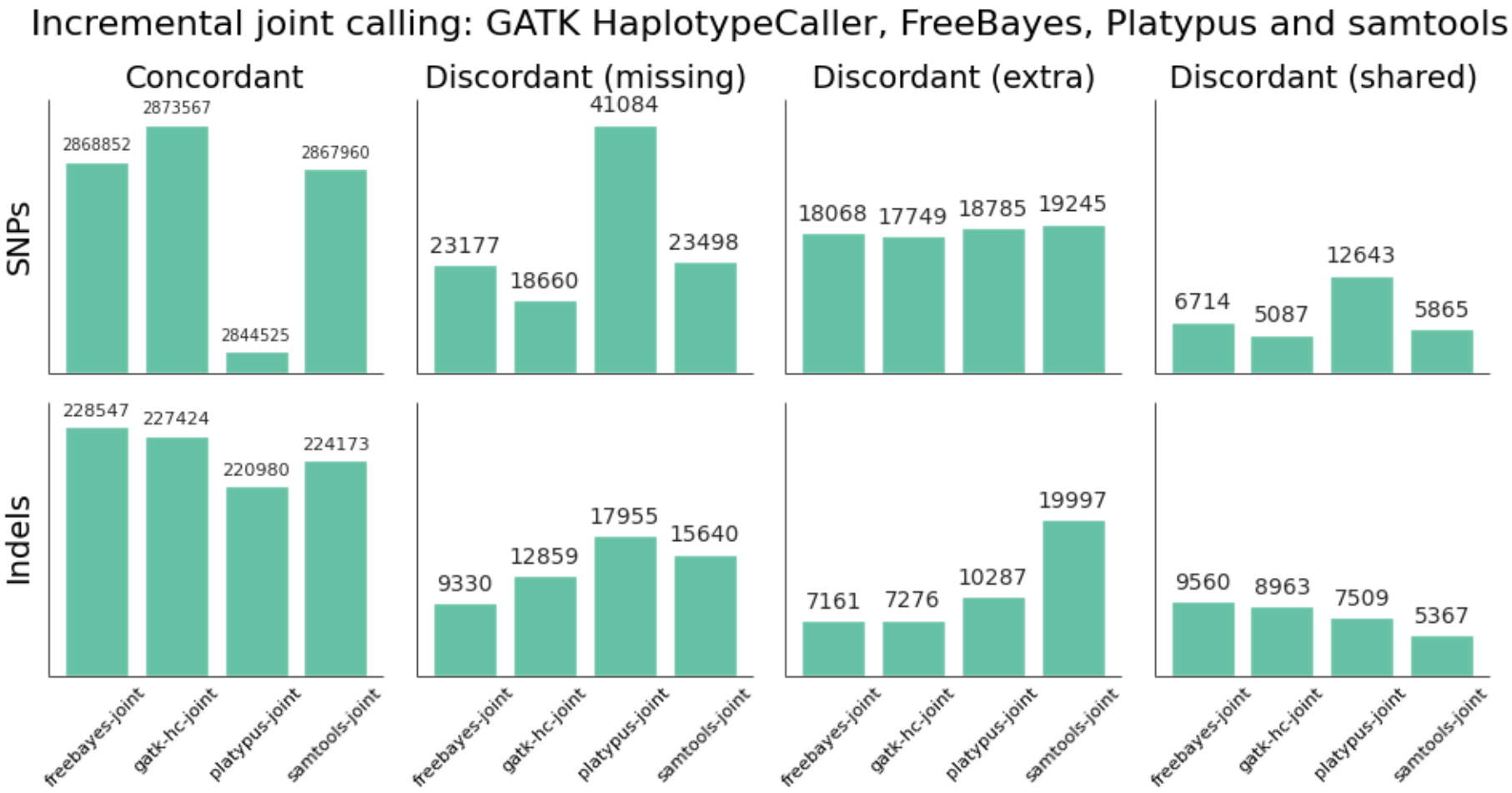
Variation Calling: the callers war



http://bcb.io/2014/05/12/wgs-trio-variant-evaluation/

Parco Tecnologico Padano

Variation Calling: the callers war



http://bcb.io/2014/10/07/joint-calling/

Variation Calling: callers quick facts

GATK

- UnifiedGenotyper (UG)
- HaplotypeCaller (HC)
- UG requires BAM processing
- HC can be used directly (slow)

FreeBayes

- Quite fast
- No need for BAM preprocessing
- The alternative to GATK HC

SamTools

- Recently updated
- No need for BAM preprocessing
- Performs similar to the others

Variation Calling: the practical part

Exercise: Prepare a BAM file only with the reads mapped on chromosome 18

Exercise: Run FreeBayes

Exercise: Run GATK UnifiedGenotyper

Exercise: Run Samtools Variation Calling

BREAK

Short lecture

Exercise: Select only SNPs from a VCF file

Exercise: Compress and index a VCF file

Exercise: Filter the VCF file according to different parameters

Exercise: Run vcf-compare to get statistics on different VCF files

Exercise: Plot the results of vcf-compare using R

Variation Filtering

The process to take raw variants, as they are produced by a caller, and filter them to remove possible false positives and artefacts information.

Variation Filtering

DISCLAIMER: This is an area in active development, and the specifics filtering which works on a particular organism or dataset may not necessary be good for others.

*So there is **no** golden rule….*

Variation Filtering

General principles: All callers outputs a same core set of information useful for filtering:

- QUAL: this is the quality score assigned by the caller to the call itself (Phred score)
- DP: depth of coverage, i.e. how many reads were aligned in the position where the variation was called
- AO / AD: the number of reads supporting each allele observed (FORMAT)

Parco Tecnologico Padano

Variation Filtering

Simple rules of thumb:

- Try different filtering combinations for your dataset
- Filter first on general fields such as QUAL, DP and alternative alleles supporting reads
- The VCF header is your friend:
    - Go into the details of specifics parameters. Every caller has its own algorithms
    - Specifics fields are only emitted by a particular caller and may help in advanced filtering
        - e.g. QD and FS for GATK

Advanced Filtering: GATK

GATK approach uses “trusted” sets of variations:

- These are passed to a Variant Quality Score Recalibration process
- Variants scores are adjusted using information on existing SNPs
- This final set is supposed to be cleaned from false positives and to have an increased accuracy
- Additional filtering done on QD, FS, HaplotypeScore (only HC), InBreedingCoeff (more than 10 samples)

Advanced Filtering: FreeBayes

FreeBayes does many things automatically, simplifying the analysis:

- Realignment around InDels is done automatically
- Base quality recalibration is not needed since it uses an haplotype approach, which looks at variant context (i.e. surrounding sequences) and not just single bases
- Variant Quality Recalibration is not needed as well, since parameters such as reads placement bias and allele balance are built directly into the Bayesian model

As a result, filtering on FreeBayes calls requires less steps

- Filtering on QUAL, DP and AO is normally sufficient

Advanced Filtering: Samtools

Samtools (1.0+) is similar to FreeBayes:

- No need for realignment around InDels or recalibrations
- Filtering on “common” fields such as QUAL and DP is normally sufficient
- Additional filtering can be done using fields such as DV (minimum number of high-quality non-reference reads)

Filtering: general rules

Common rules to follow when filtering VCF data:

- Variation quality is important but it's not everything
- Always look at a good balance between quality and DP
- A DP bigger than twice the average depth may indicate problematic regions where artefacts can be present
- Try to filter also on the number of reads supporting alternative alleles
- Always remember that each organism and dataset require *ad hoc* filters
- Advanced analyses can also be done using genotyping likelihoods (GL) which are emitted by all callers for each sample

Parco Tecnologico Padano

Variation Calling: the practical part 2

Exercise: Select only SNPs from a VCF file

Exercise: Compress and index a VCF file

Exercise: Filter the VCF file according to different parameters

Exercise: Run vcf-compare to get statistics on different VCF files

Exercise: Plot the results of vcf-compare using R