H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to NGS Variant Calling

Bioinformatics analysis and annotation of variants in NGS data workshop
Cape Town, 4th to 6th April 2016

Sumir Panji, Amel Ghouila, Gerrit Botha

Learning Outcomes

- Types of variants
- Rationale for calling variants in NGS data
- Overview of types of variant callers
- Different strategies used in variant calling
- Input files used and output files generated by variant callers

Types of Variants

- Single nucleotide polymorphisms (SNPs) – difference in a single base pair from a reference

Reference: ATGCCGTATTCCGTATTCGGACCTTA

Sample 1: ATGCCGTATTCCATATTCGGACCCTA

Sample 2: ATGCCGTATTCCGTATTCGGACCCTA

Sample 3: ATGCCGTATTCCGTATTCGGACCCTA

Sample 4: ATGCCTTATTCCGTATTCGGACCCTA

Types of Variants

- Also known as a single nucleotide variations (SNVs)

Reference: ATGCCGTATTCCGTATTCGGACCTTA

Sample 1: ATGCCGTATTCCATATTCGGACCCTA

Sample 2: ATGCCGTATTCCGTATTCGGACCCTA

Sample 3: ATGCCGTATTCCGTATTCGGACCCTA

Sample 4: ATGCCTTATTCCGTATTCGGACCCTA

- Constitute ~ 90% of all genetic variations between humans

Types of Variants

- Insertions and deletions (INDELS) are small insertions or deletions in a genome in comparison with a reference

Reference: ATGCCGTATTCCGTA- - - TTCGGACCTTA

Sample 1: ATG - - - TATTCCATA- - - TTCGGACCCTA

Sample 2: ATGCCGTATTCCGTAGGTTTCGGACCCTA

Sample 3: ATGC – GTATTCCGTA- - -TTCGGACCCTA

Sample 4: ATGCCTTATTCCGTAGGTTTCGGA - - - TA

Types of Variants

- INDELS < then 50 basepairs referred to as microindels
- INDELS differ from SNPs/SNVs, the latter results in a bp replacement keeping the number of bases the same
- INDELs change the overall number of bps
- INDELS that are not multiples of 3 bps cause frameshift mutations

Reference:	ATG CCG TAT TCC GTA TTC GGA CCT TAA
Sample 1:	ATG - - - - TA TTC CAT ATT CGG ACC CTA A
Sample 2:	ATG GCG TAT TCC GTA TTC GGA CCC TAA
Sample 3:	ATG CCT TAT TCC GTA TTC GGA - - - TAA

Types of Variants

- Structural variations:
  - Copy number variants (CNVs) – deletions or duplications of the same region in a genome (usually 1Kbp to 3Mbp in size)
  - Inversion – the region of the genomes has flipped (usually on the same chromosome or region)
  - Translocation – exchange of parts between two non-homologous chromosomes

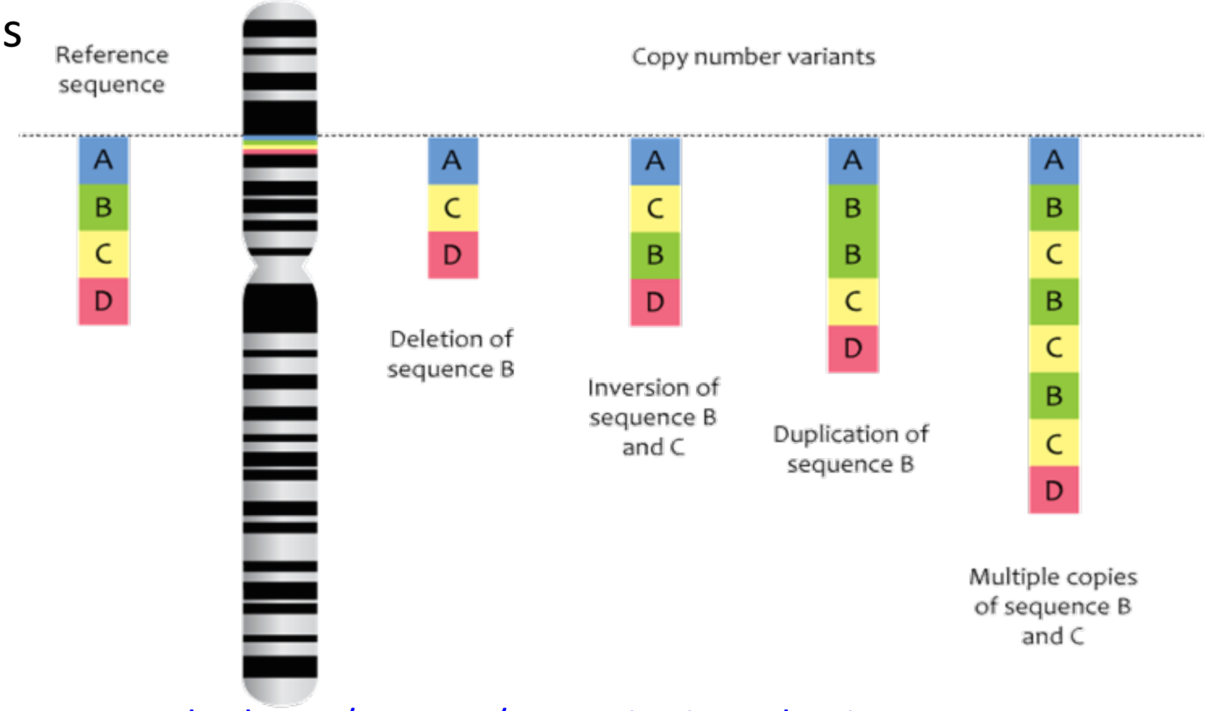

Reference sequence
Copy number variants
A
B
C
D
A
C
D
Deletion of sequence B
A
C
B
D
Inversion of sequence B and C
A
B
B
C
D
Duplication of sequence B
A
B
C
B
C
B
C
D
Multiple copies of sequence B and C

Image from: http://www.emedmd.com/content/genomics-introduction

Types of Variants

- There are two classes of SNVs that occur in the literature
- Constitutional / germline mutations – these are inherited from the parents and present in every cell
- Somatic mutations are mutations that occur during the lifetime of an individual
- Usually, when looking for rare disease contributing SNVs one is using germline variations e.g diabetes
- For non-heritable diseases e.g some cancers, the contribution of somatic mutations in relation to disease state is studied by comparing tumor vs normal samples

Why call variants in NGS?

- Variation in DNA sequences function as markers to study Mendelian and non-monogenic complex diseases

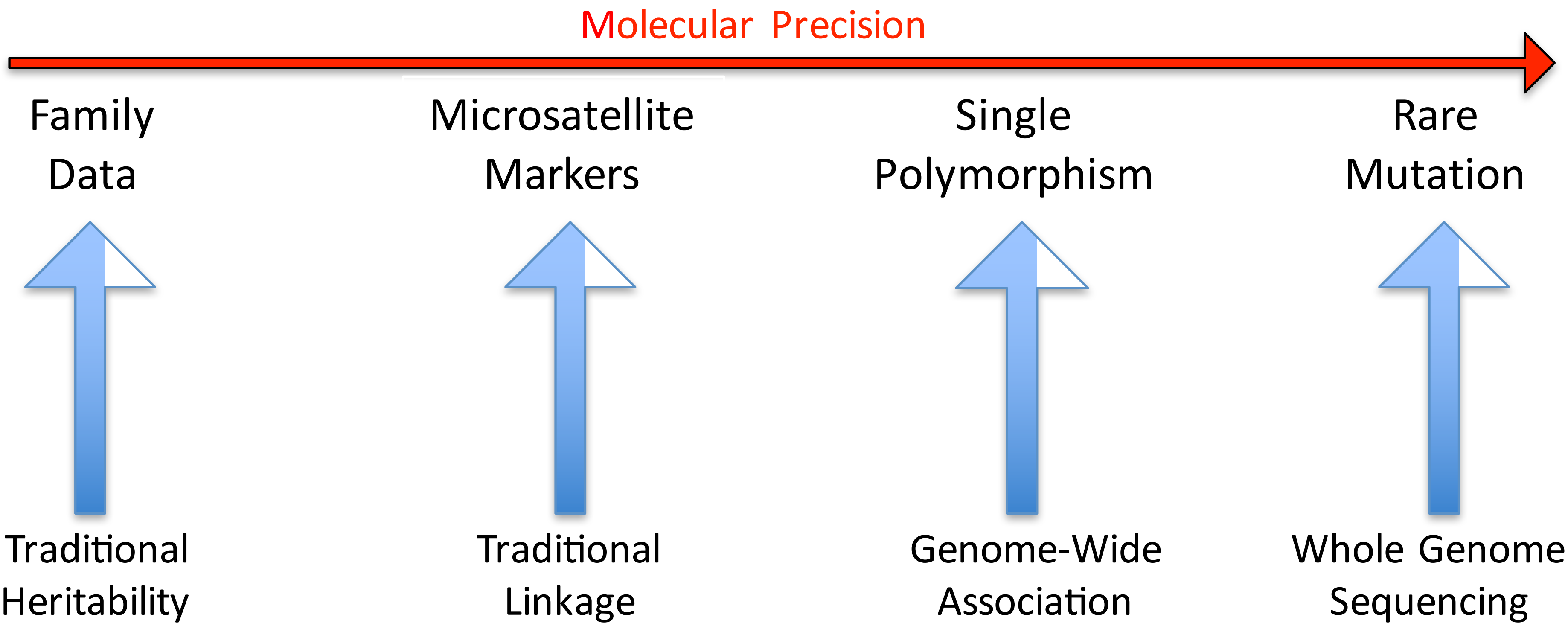
- Pharmocogenomics - understand responses in drug treatments e.g polymorphisms in CYP2C9 linked with elevated risks in anticoagulation and of bleeding events amongst warfarin patients (PMID: 11926893)

- Increasing data from diverse human populations are being generated – leads to higher confidence / understanding of genetic variation e.g 1,000 (1K) genomes project, 100,000 (100K) genomes project, maybe a 1,000K project soon?

- ↑ sequencing output + ↓ cost = more data = greater resolution / precision in the study of genetic variation disease association for rare variants

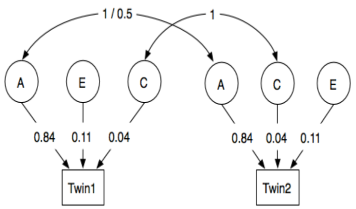
Why call variants in NGS?

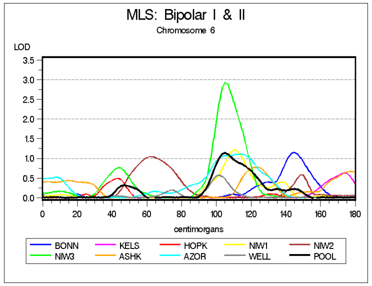
Slide courtesy of Prof. Matt McQueen – University of Boulder Colorado

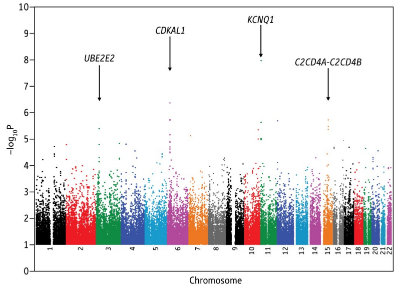


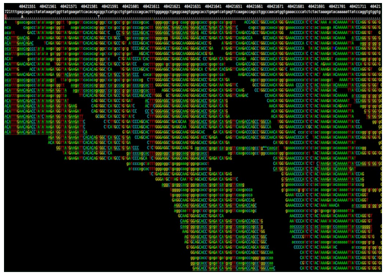
Molecular Precision

Family Data	Microsatellite Markers	Single Polymorphism	Rare Mutation
Traditional Heritability	Traditional Linkage	Genome-Wide Association	Whole Genome Sequencing









Why call variants in NGS?

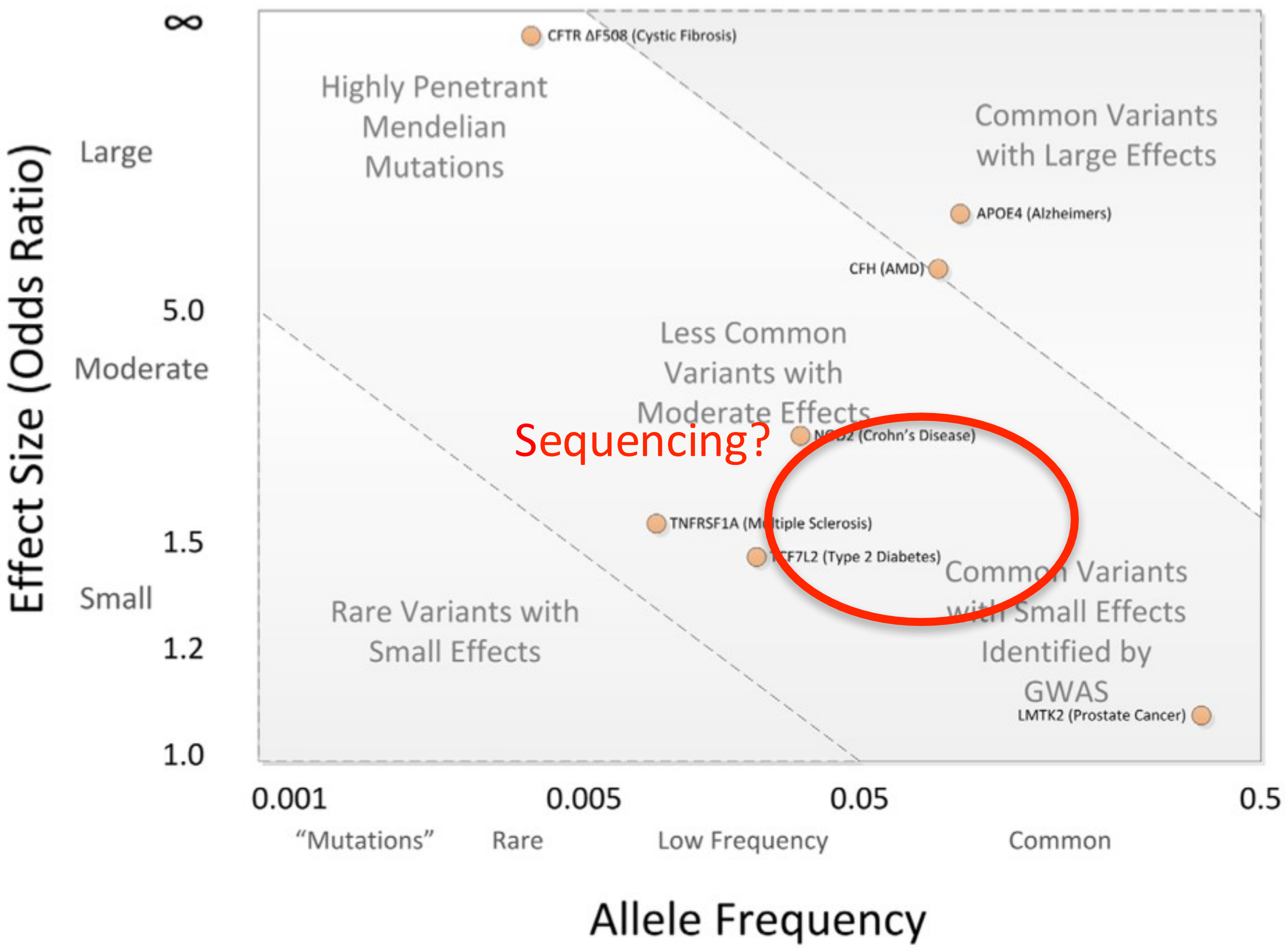
- Enables studies of complex disease associations between genotype and phenotype and the effect of variant on phenotype

- “Common disease – common variant” hypothesis – multiple common variants provide a cumulative contribution to an observed phenotype (usually GWAS)

- “Common disease – rare variant” hypothesis - multiple rare variants with a large effect size cause the observed phenotypes

- What about common diseases with multiple rare variants that occur at low frequency with moderate or small effect sizes?

Why call variants in NGS?



Bush & Moore (2012), *PLOS Comp Biol.*

Slide courtesy of Prof. Matt McQueen – University of Boulder Colorado

Variant calling in NGS data

- Currently most variant calling is done on Whole Exome Sequence data (WES) and not whole genome sequence (WGA) data

- WES identifies variants in ~1% of the human genome that codes for proteins

- On average 12,000 variants are found in coding regions (although this number might be biased to well studied European populations and will increase if looking at less well studied African genomes)

- WGS not commonly employed due to current cost implications and also computational data storage and analysis

- WGS is good for finding non-coding, regulatory and intronic variants

- On average ~5 million variants can be obtained compared to a reference (although might be a larger number if looking at African genomes)

Variant calling in NGS data

- Difference between calling a SNP/SNV identification and variant calling

- A SNP/SNV is a basepair difference from the reference sequence e.g A ←→ T

- In cases of low coverage sequencing (x5), there is a high chance that only one chromosome of a diploid organism has been captured

- To mitigate this bias, higher sequence coverage is used, especially with decreasing costs e.g for clinical genomics x50 or x100 coverage is used

Variant calling in NGS data

- Genotyping is determining what sets of alleles are present / inherited at a given location, and at what frequency these occur

- SNP/SNV calling provides information on which location the polymorphism differs from the reference sequence

- When only one WES/WGS is used, genotyping and SNP calling are similar, with multiple WES/WGS the rate of false positives increases with sample size if just looking at SNV/SNP positions

- Genotype likelihoods are calculated for each individual at the position the SNP/SNV has been found to determine what allele the SNP/SNV might originate from

Types of variant callers

- Variant calling tools can be divided into 4 classes based on the types of variants they are designed to identify:

1. Germline callers – used in finding predisposing variants for monogenic, rare and complex diseases – uses a single input file
2. Somatic callers – used for cancer studies comparing normal vs tumor – uses two input files (case/control)
3. CNV callers – callers that identify CNVs
4. Structural variants (SV) callers – callers that identify SVs that are larger then CNVs

- This talk and practical session will focus on germline callers

- Somatic callers require different thresholds compared to germline callers due to the low signal to noise ratio as somatic variations occur at low frequency

Types of variant callers

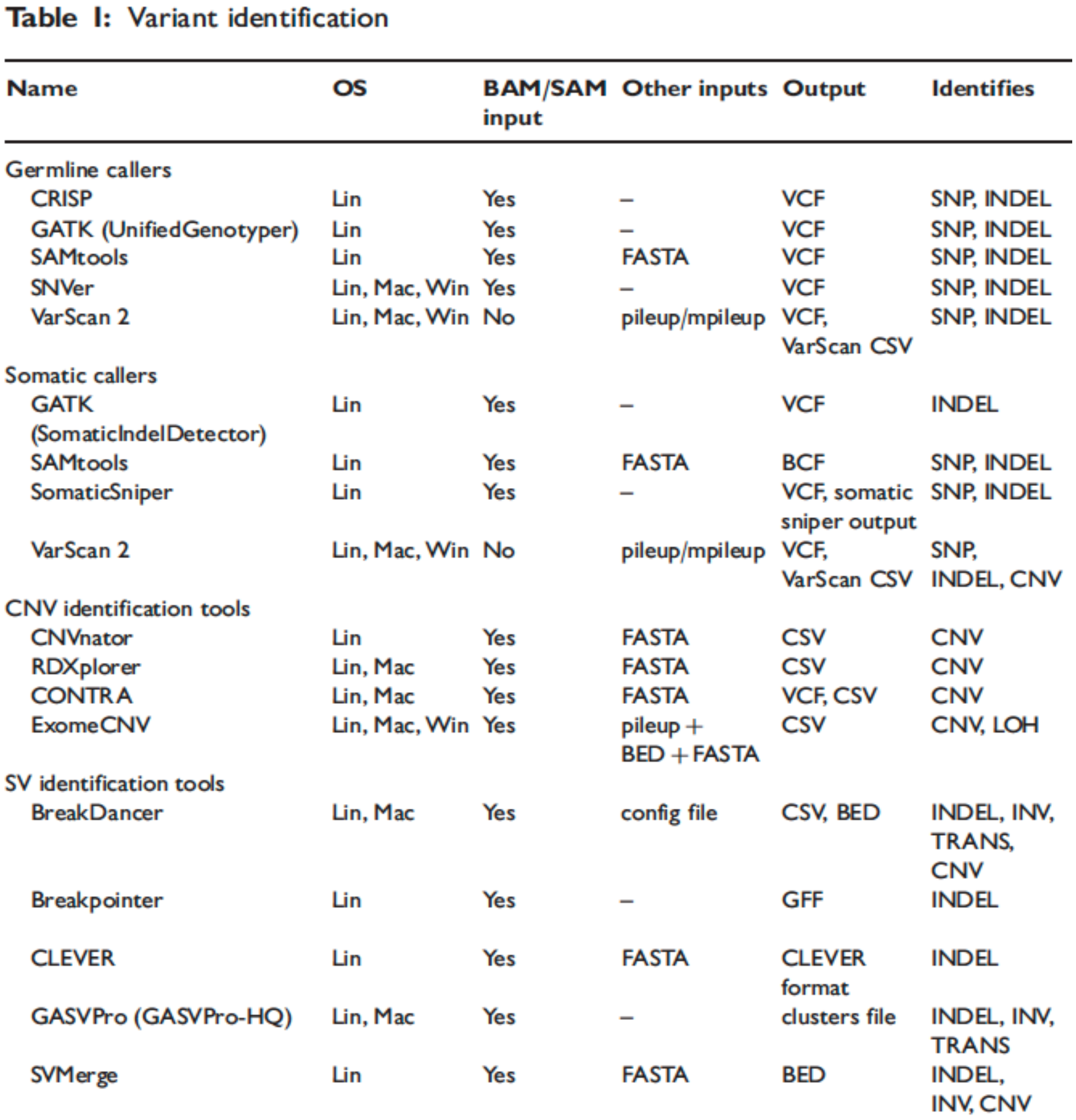


Table I: Variant identification

Name	OS	BAM/SAM input	Other inputs	Output	Identifies
Germline callers					
CRISP	Lin	Yes	–	VCF	SNP, INDEL
GATK (UnifiedGenotyper)	Lin	Yes	–	VCF	SNP, INDEL
SAMtools	Lin	Yes	FASTA	VCF	SNP, INDEL
SNVer	Lin, Mac, Win	Yes	–	VCF	SNP, INDEL
VarScan 2	Lin, Mac, Win	No	pileup/mpileup	VCF, VarScan CSV	SNP, INDEL
Somatic callers					
GATK (SomaticIndelDetector)	Lin	Yes	–	VCF	INDEL
SAMtools	Lin	Yes	FASTA	BCF	SNP, INDEL
SomaticSniper	Lin	Yes	–	VCF, somatic sniper output	SNP, INDEL
VarScan 2	Lin, Mac, Win	No	pileup/mpileup	VCF, VarScan CSV	SNP, INDEL, CNV
CNV identification tools					
CNVnator	Lin	Yes	FASTA	CSV	CNV
RDXplorer	Lin, Mac	Yes	FASTA	CSV	CNV
CONTRA	Lin, Mac	Yes	FASTA	VCF, CSV	CNV
ExomeCNV	Lin, Mac, Win	Yes	pileup + BED + FASTA	CSV	CNV, LOH
SV identification tools					
BreakDancer	Lin, Mac	Yes	config file	CSV, BED	INDEL, INV, TRANS, CNV
Breakpointer	Lin	Yes	–	GFF	INDEL
CLEVER	Lin	Yes	FASTA	CLEVER format	INDEL
GASVPro (GASVPro-HQ)	Lin, Mac	Yes	–	clusters file	INDEL, INV, TRANS
SVMerge	Lin	Yes	FASTA	BED	INDEL, INV, CNV

Table adapted from Pabinger S *et al* Brief Bioinform. 2014 Mar;15(2):256-78; PMID: 23341494

Types of variant callers

- Variant calling methods can be divided into 2 categories:

1. Heuristic methods:
   - Use several sources of information linked with the data
   - VarScan2 is partly heuristic and determines a genotype based on minimum coverage of 33, minimum base quality of 20 and a predefined allele frequency
   - They have a high computational overhead so are much less commonly used compared to probabilistic models

Types of variant callers

- Variant calling methods can be divided into 2 categories:

2. Probabilistic methods:

   - Use a “genotype likelihood” framework that is based on Bayesian probability approach

   - Prior information such as patterns of linkage disequilibrium are joined with other information such as errors in base calling, alignment score to provide a statistical measure of uncertainty

   - Posterior probabilities use data such as the Phred quality score to help calculate each genotype within this framework

Types of variant callers

- “Bayes’ formula: A mathematical expression showing that a posterior probability can be found as the prior probability multiplied by the likelihood divided by constant” *

- “Prior probability: In the context of this Review, the probability of a genotype calculated without incorporating information from the next-generation sequencing data. Prior probabilities can be obtained from a set of reference data types of callers” *

- * Reference: Nielsen, Rasmus et al. “Genotype and SNP Calling from next-Generation Sequencing Data.” *Nature reviews. Genetics* 12.6 (2011): 443–451. *PMC*. Web. 3 Apr. 2016.

$$P(A \mid B) \propto P(A) \cdot P(B \mid A)$$

- P(genotype|data) ∝ P(data|genotype)P(genotype)
- P(genotype) : prior probability for variant
- P(data|genotype): likelihood for observed(called) allele type

- https://en.wikipedia.org/wiki/Bayes'_theorem
- https://en.wikipedia.org/wiki/Bayesian_inference

Specific variant callers

- Popular variant callers include GATK, SAMTools and FreeBayes
- Genome Analysis Toolkit (GATK) is a package of genome tools created by the Broad Institute for the 1000 genomes project
- Two main variant calling programs: UnifiedGenotyper and HaplotypeCaller
- UnifiedGenotyper is used to callSNVs and INDELS separately, deprecated for HaplotypeCaller
- HaplotypeCaller detects SNVs, INDELS with better accuracy due to realignment steps incorporated

- https://www.broadinstitute.org/gatk/about/
- http://gatkforums.broadinstitute.org/gatk/discussion/3151/should-i-use-unifiedgenotyper-or-haplotypecaller-to-call-variants-on-my-data
- https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php

Specific variant callers

- SAMtools is also a software suite for working with NGS data
  - Samtools – manipulate SAM/BAM/CRAM file formats
  - BCFtools – manipulate BCF2/VCF/gVCF and calling SNVs and INDELS
  - HTSlib – a C library for reading and writing NGS data
- MPileup from SAMtools calls the SNVs by scanning every position in the genome/exome, calculates every possible genotype and then assigns likelihoods that the genotype is present in the sample
- BCFtools uses these computed, assigned genotype likelihoods to call the SNVs and INDELs
- Differs to GATK in the models used to estimate the genotypes likelihoods and also uses predefined filters (GATK obtains filter parameters from the data)
- http://www.htslib.org/

Specific variant callers

- FreeBayes works on the concept of haplotype alignments and is designed to find small SNVs and INDELS
- Uses these haplotypes blocks to call variants based on the literal sequences of reads that fall into that haplotype block as opposed to calling from the actual alignment
- The authors claim that this avoids the problems of alignment based variant detection where identical sequences may have multiple possible alignments
- Similar to GATK and SAMtools, but appears to have a much more robust Bayesian framework that can incorporate polyploidy analysis (useful for plant genomes)
- https://github.com/ekg/freebayes#readme

Specific variant callers

- GATK, SAMTools and FreeBayes can take BAM files as input
- GATK, SAMTools and FreeBayes need a reference sequence for variant calling
- GATK, SAMTools and FreeBayes generate a VCF file that is used for further (tertiary analysis)
- GATK, SAMTools and FreeBayes can run on an HPC Unix/Linux environment
- SAMTools and FreeBayes are available on Galaxy, GATK is deprecated on Galaxy (also does not have HaplotypeCaller)

What variant caller to use

- There is no "right" answer as to which variant caller to use
- Variant callers aim to be as sensitive as possible, which leads them to call as many variants possible within the statistical framework that they incorporate
- Rationale behind this is it is better to call some false positives rather then sacrifice any potential true positives as the latter scenario is much worse for biomedical research
- The user is then left to use other sources of data to determine if this variant called is of any biological significance (tertiary analysis)

What variant caller to use

- Common approach is to use 2 to 3 variant callers
- Problem is there is little concordance on the variants identified
- This has led to a proliferation of Venn diagrams in the literature

What variant caller to use

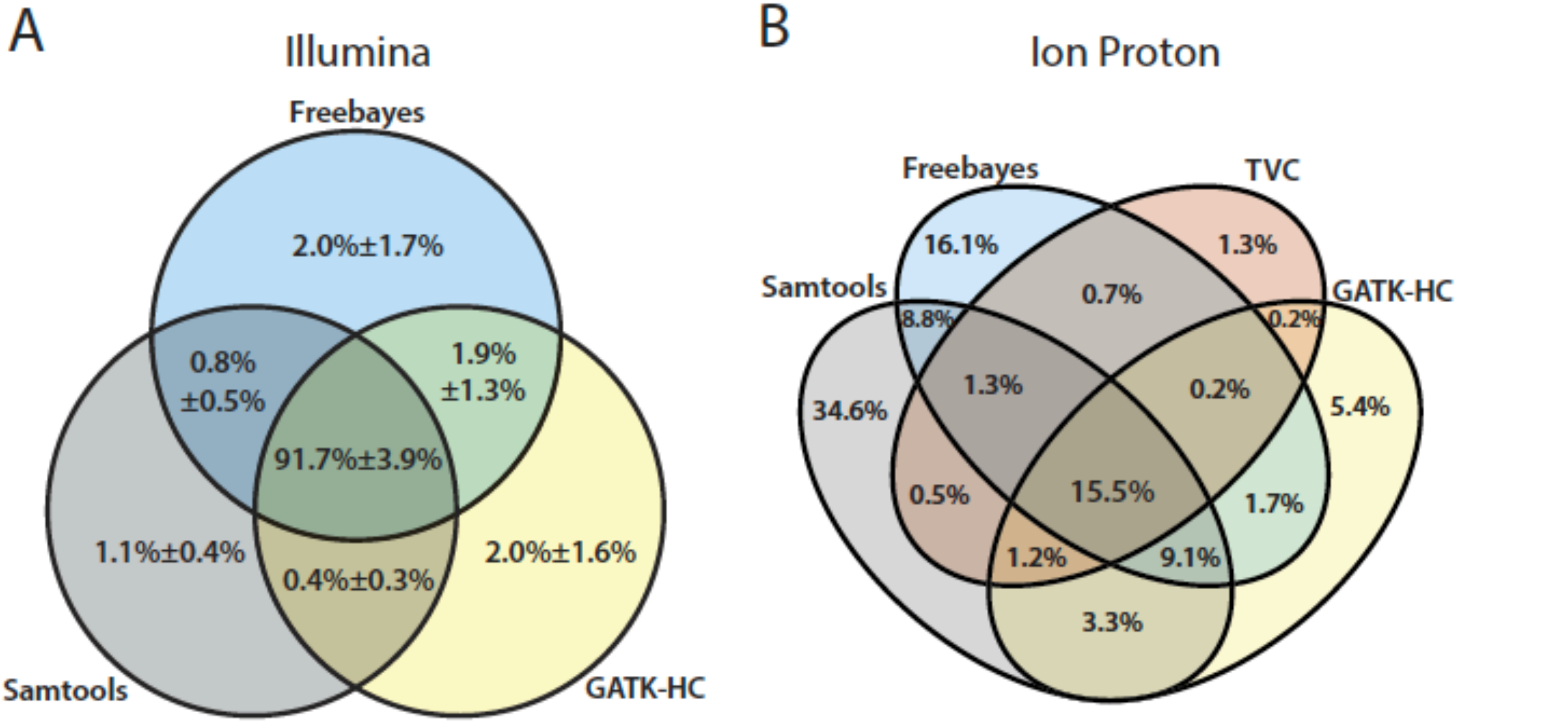


Figure 3. Venn diagrams summarizing called variants by different callers. The mean percentage with standard deviation of confidence variant calls with equal to or higher than the quality score threshold of 20 are represented for (A) Illumina data sets and (B) Ion Proton data set.

Hwang S. Sci Rep. 2015 Dec 7;5:17875. doi: 10.1038/srep17875; PMID: 26639839

What variant caller to use

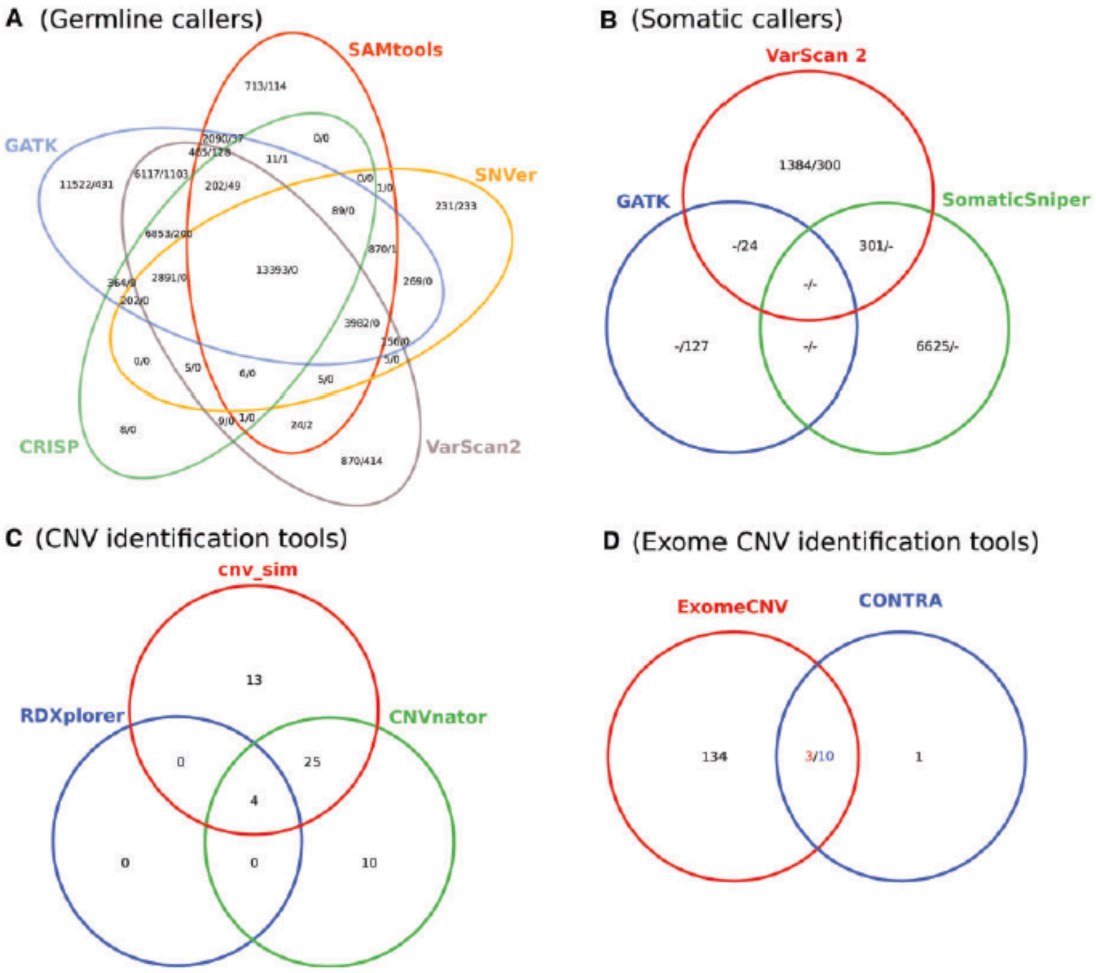


Figure 2: Venn diagrams showing the number of identified variants for tested germline (A), somatic (B), CNV (C) and exome CNV (D) tools. The depicted numbers in (A) and (B) report identified SNPs and INDELs. Venn diagram (C) shows the overlap between known (cnv.sim) and predicted CNVs. Figure (D) illustrates the overlap between CONTRA and ExomeCNV. The intersection numbers were adjusted to reflect that 10 CNVs detected by CONTRA are located within 3 CNVs reported by ExomeCNV.

Pabinger S *et al* Brief Bioinform. 2014 Mar;15(2):256-78; PMID: 23341494

What variant caller to use

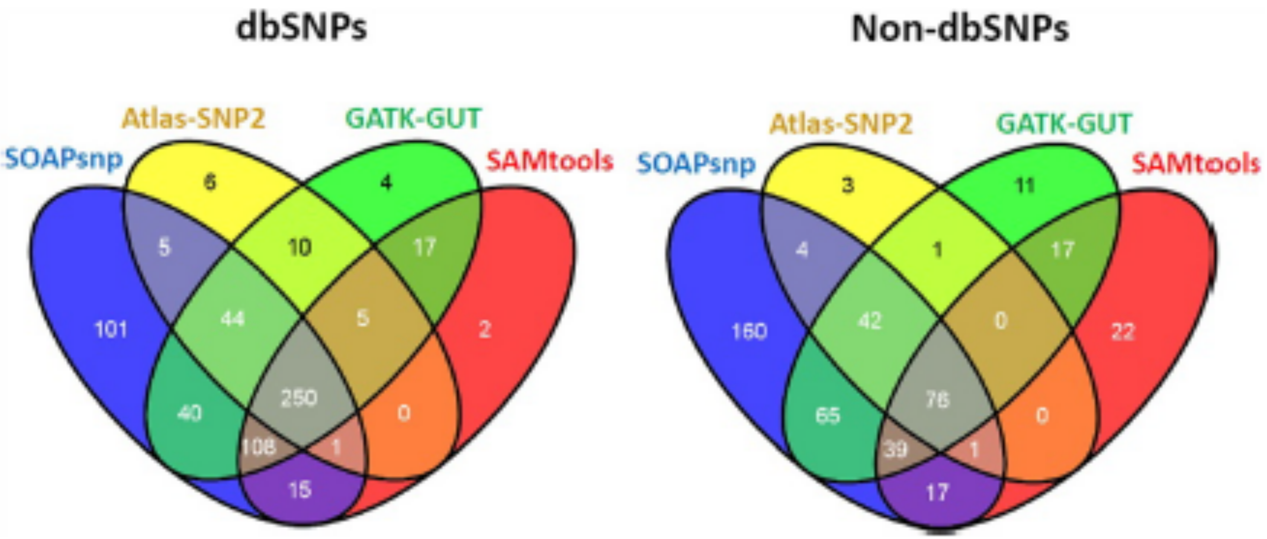


Figure 3 The comparison results of trimmed data without any post-output filters. All SNVs require ≥ 3X coverage.

- In this publication Yu and Sun provide a good set of metrics, and recommend to use multiple callers
- But if you could only choose one, go for GATK

Yu and Sun; BMC Bioinformatics201314:274; PMID:24044377

What variant caller to use

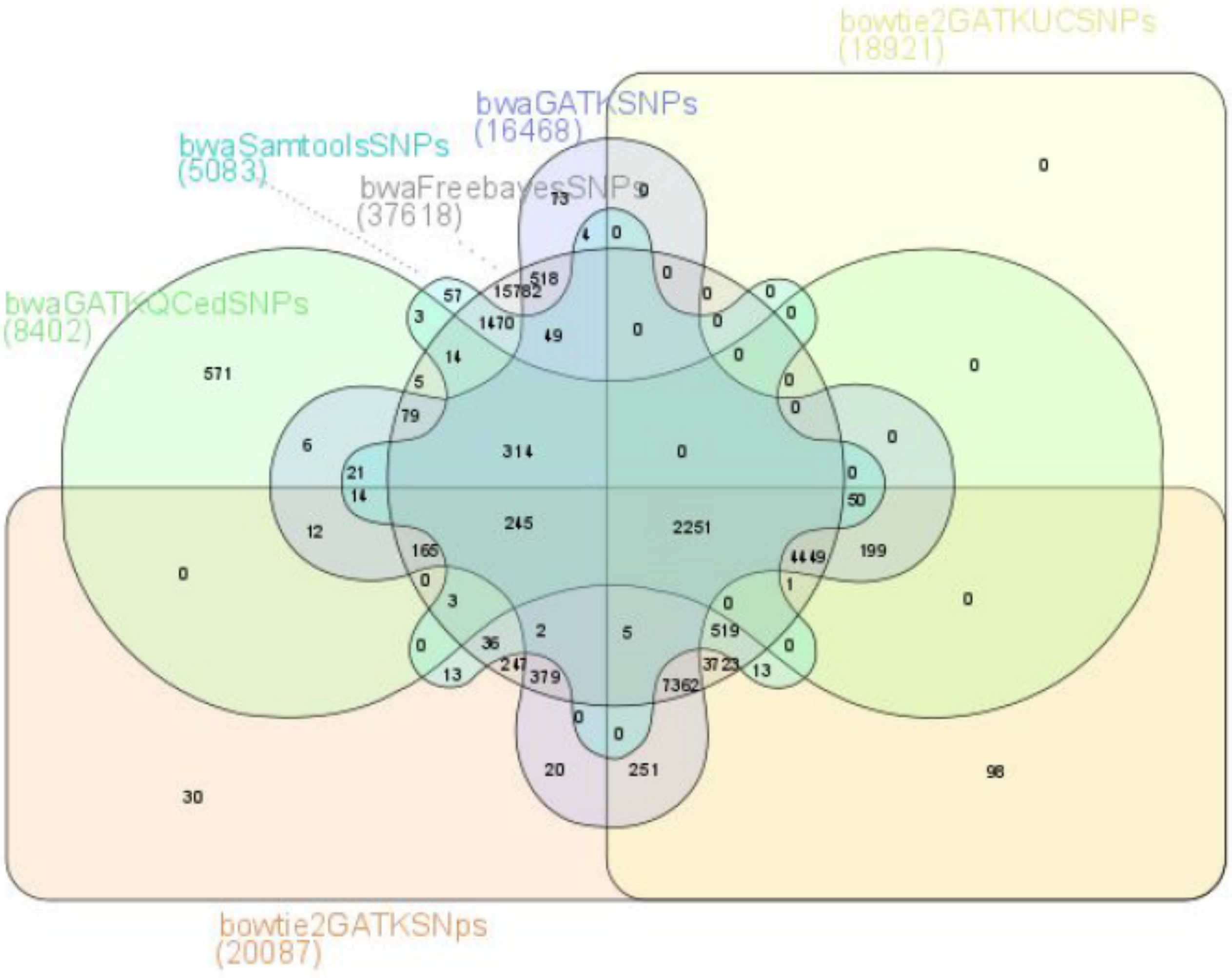


Figure 9. SNP overlap between pipelines (bwaGATK, bwaGATKQCed, bowtie2GATK, bowtie2GATKUG, bwaSamtools, bwaFreebayes).

H3ABioNet CBIO NGS Node accreditation variant calling assessment

What variant caller to use

- Ideally use 2 to 3 well documented variant callers and compare the results
- Not always feasible, see what similar publications are doing (supplementary information is very useful place to get the methods used)
- Useful to follow the same protocols (this also includes software versions) if want to compare your results directly with a previously published study
- Keep up to date with published papers comparing various NGS tools / pipelines and GATK forums
- Also find a good program to draw Venn diagrams

Types of calling

- There are different ways of doing variant calling based on the study design

- Joint calling – calling a group of samples at the same time which do not require access to all the BAM files

- Less computationally intensive and possible with GATK's incremental joint calling where genomic VCFs (gVCFs) are used with new batches of BAM files

- Useful when doing large population based genomic studies and the sequence data arrives in batches

Types of calling

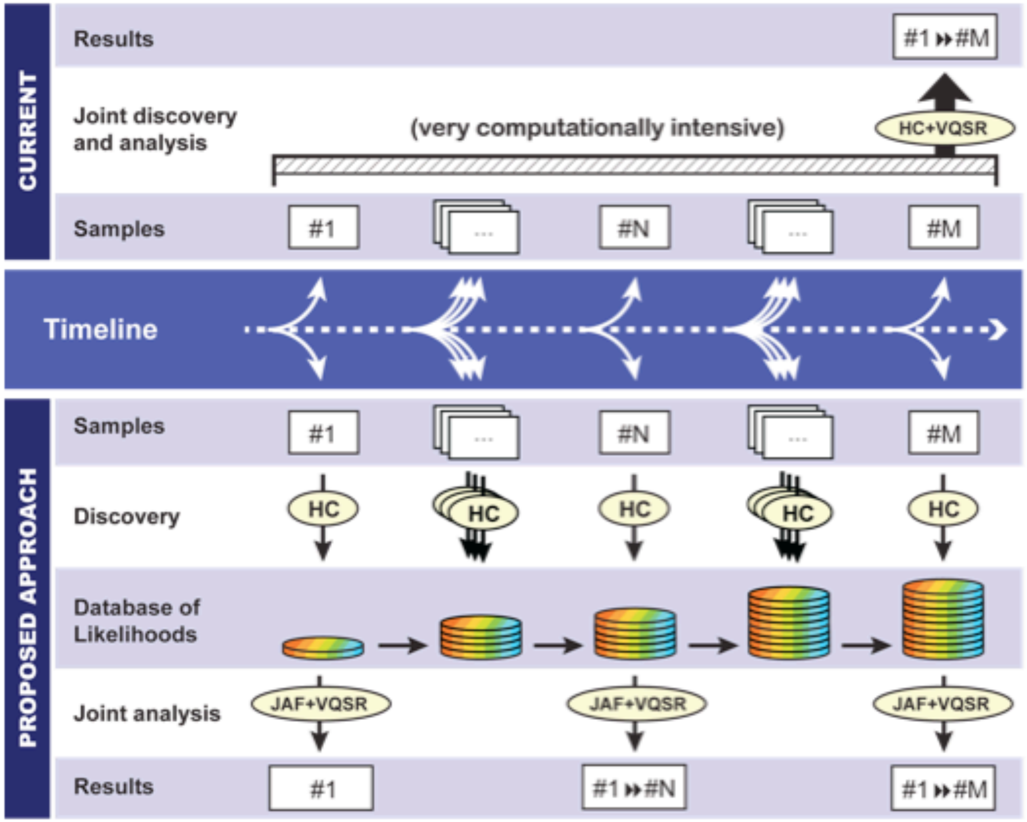


Figure 3: schematic illustrating the current and proposed methods for variant calling. Currently, all samples must be processed together throughout the entire pipeline, making it extremely computationally intensive and limiting it to pre-planned freezes. In contrast, the proposed approach enables modularized processing that is much less intensive and more flexible, as variant calling can be performed cumulatively as new data arrives.

Image obtained from: http://gatkforums.broadinstitute.org/gatk/discussion/3686/why-do-joint-calling-rather-than-single-sample-calling-retired

Types of calling

- Pooled or batch calling – traditional approach where all the BAMs from a sample are used to call variants

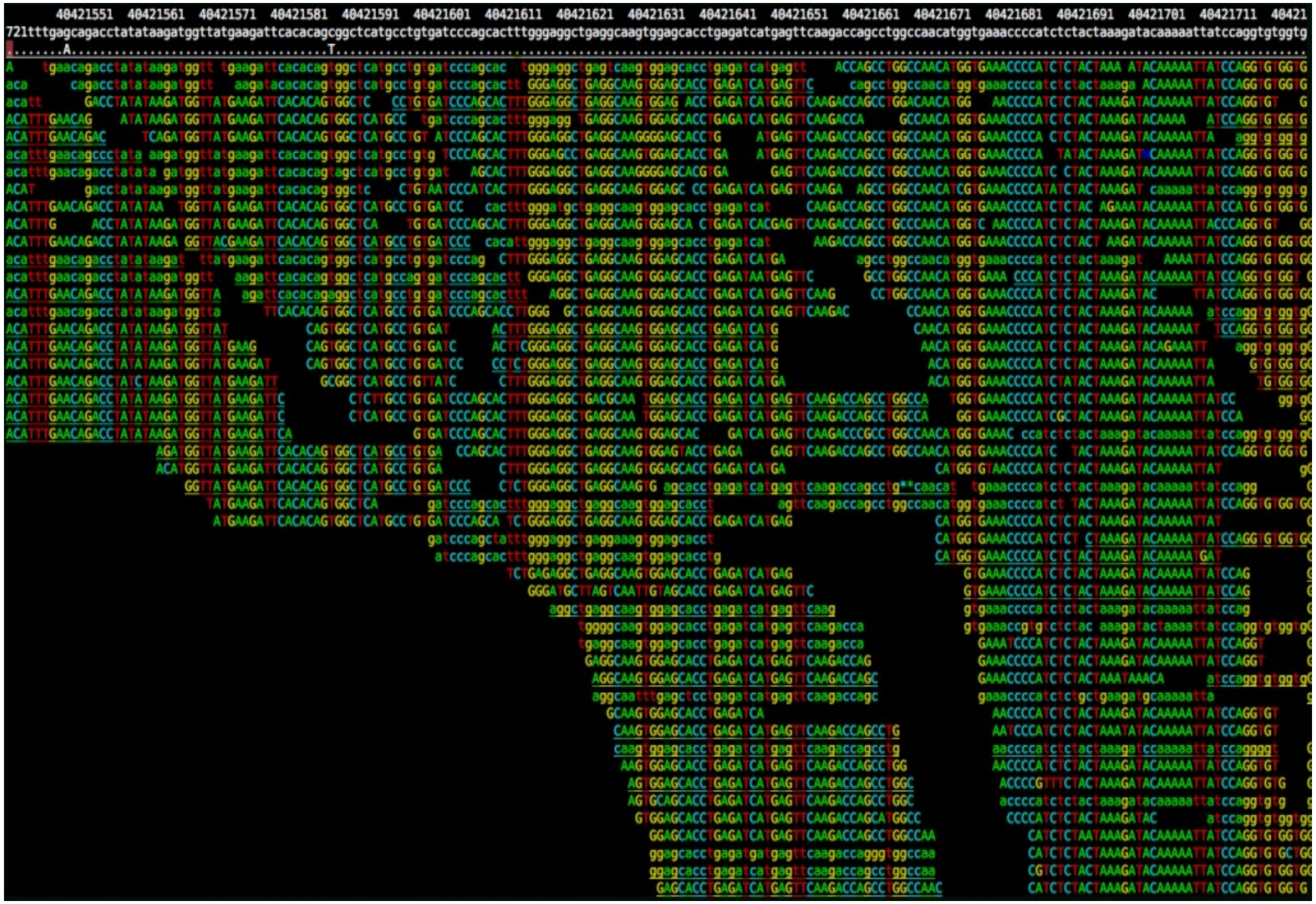
- Scales quite poorly in terms of computational costs as more samples are added

- Single sample calling – using a single sample to identify variants can be used for variant calling in cancer

- As the methods are statistical, the greater the sample size, the greater the power of the study and hence the higher confidence in results (especially for low frequency, rare variants)

Potential pitfalls of variant calling

- When a variant is identified, how can you be certain it is real?
- One way is to look at the read depth coverage for the position of the variant – the more reads the more confident one is



```
       40421551  40421561  40421571  40421581  40421591  40421601  40421611  40421621  40421631  40421641  40421651  40421661  40421671  40421681  40421691  40421701  40421711  40421
721tttgagcagacctatataagatggttatgaagattcacacagcggctcatgcctgtgatcccagcactttgggaggctgaggcaagtggagcacctgagatcatgagttcaagaccagcctggccaacatggtgaaaccccatctctactaaagatacaaaaattatccaggtgtggtg
........A...................................T...................................................................................................................................
```

Potential pitfalls of variant calling

- Does not always hold true in the case of instrument problems as the same error is propagated repeatedly
- Reads not QC-ed well before the alignment, refinement and variant calling steps will result in false variants (tools are becoming more robust in this regard)
- Possible mapping problems may give rise to a bias in the in the number of reads, base quality scores favoring an alternative allele or the position of the variant in the read

Possible improvements for variant calling

- With GATK one can use variant quality score recalibration (VQSR) which refines the variant qualities and improves precision (true positives)

- Drawbacks of VQSR is ~30 WES datasets required to be effective or WGS for use and the reference variants datasets are limited to few organisms

- Ensure that duplicate sequences have been removed as these will lead to “over scoring” of a variant

- Local realignment of reads around INDELS will help to improve the quality of the variants called

Metrics to help interpret variant quality

- For GATK a Genotype quality score ranges from 0-99 with higher values indicating more confidence in the called variant

- A QUAL parameter in the output VCF for GATK represents a quality probability of an SNV being a homozygous reference, values ≥ 30 is usually used for reliable SNV calling

- The FisherStrand value >60 indicates a strand bias and likely a false positive

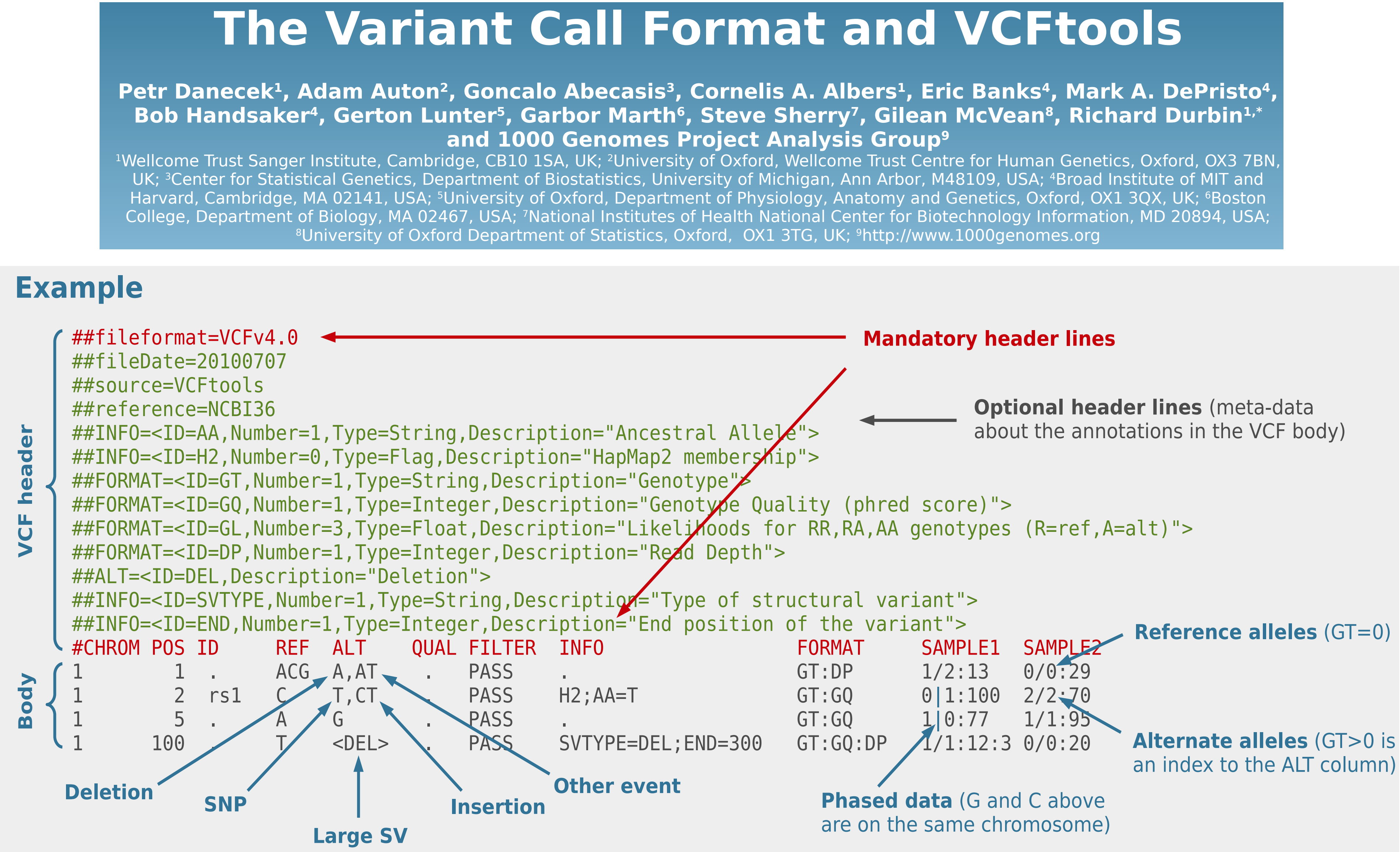
- A 10 base window around a called SNV can be used to check if a SNV is mapped to more then 2 haplotypes using the HaplotypeScore parameter, the lower the score the better (≤ 13)

Output file - VCF

- Variant Call Format file is the output from variant calling tools

- Stanadardized file format in text to represent SNV, INDELS and SV

Output file - VCF



The Variant Call Format and VCFtools

Petr Danecek[1], Adam Auton[2], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Bob Handsaker[4], Gerton Lunter[5], Garbor Marth[6], Steve Sherry[7], Gilean McVean[8], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[9]

[1]Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK; [2]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, OX3 7BN, UK; [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, M48109, USA; [4]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA; [5]University of Oxford, Department of Physiology, Anatomy and Genetics, Oxford, OX1 3QX, UK; [6]Boston College, Department of Biology, MA 02467, USA; [7]National Institutes of Health National Center for Biotechnology Information, MD 20894, USA; [8]University of Oxford Department of Statistics, Oxford, OX1 3TG, UK; [9]http://www.1000genomes.org

Example

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1  SAMPLE2
1        1  .     ACG  A,AT    .   PASS    .                    GT:DP      1/2:13   0/0:29
1        2  rs1   C    T,CT    .   PASS    H2;AA=T              GT:GQ      0|1:100  2/2:70
1        5  .     A    G       .   PASS    .                    GT:GQ      1|0:77   1/1:95
1      100  .     T    <DEL>   .   PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3 0/0:20
```

Body

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Deletion

SNP

Large SV

Insertion

Other event

Source: http://vcftools.sourceforge.net/VCF-poster.pdf

Output file - VCF

- Header section contains information on the dataset:
  - Organism
  - Genome build reference version used
  - Definitions of the annotations used (this usually contains the parameters chosen when running the variant calling experiment)
  - First line indicates VCF version:

  ##fileformat=VCFv4.0

  - The FILTER lines tell you what filters have been applied to the data:

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

Practical

- Use the alignment QC-ed alignment dataset created in the previous practical as input for the variant calling tools in Galaxy (in this case FreeBayes)
- Generate a file that has variant calls (VCF)
- Look at the sections of the VCF to determine what organism was used, the VCF file format version number, number of variants called, the number of variants that pass your threshold