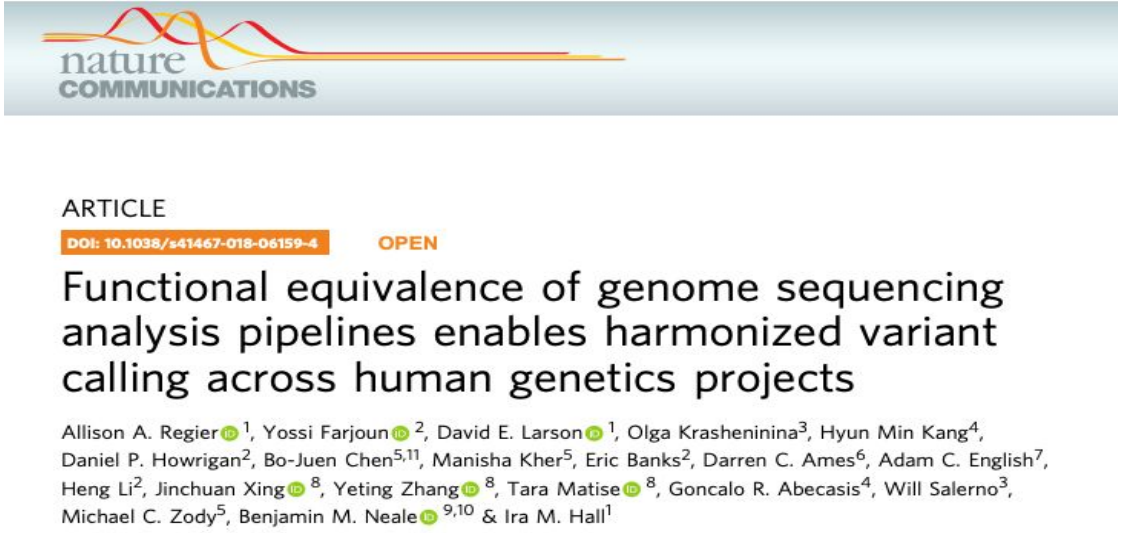
Genomic Variations



nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-06159-4 OPEN

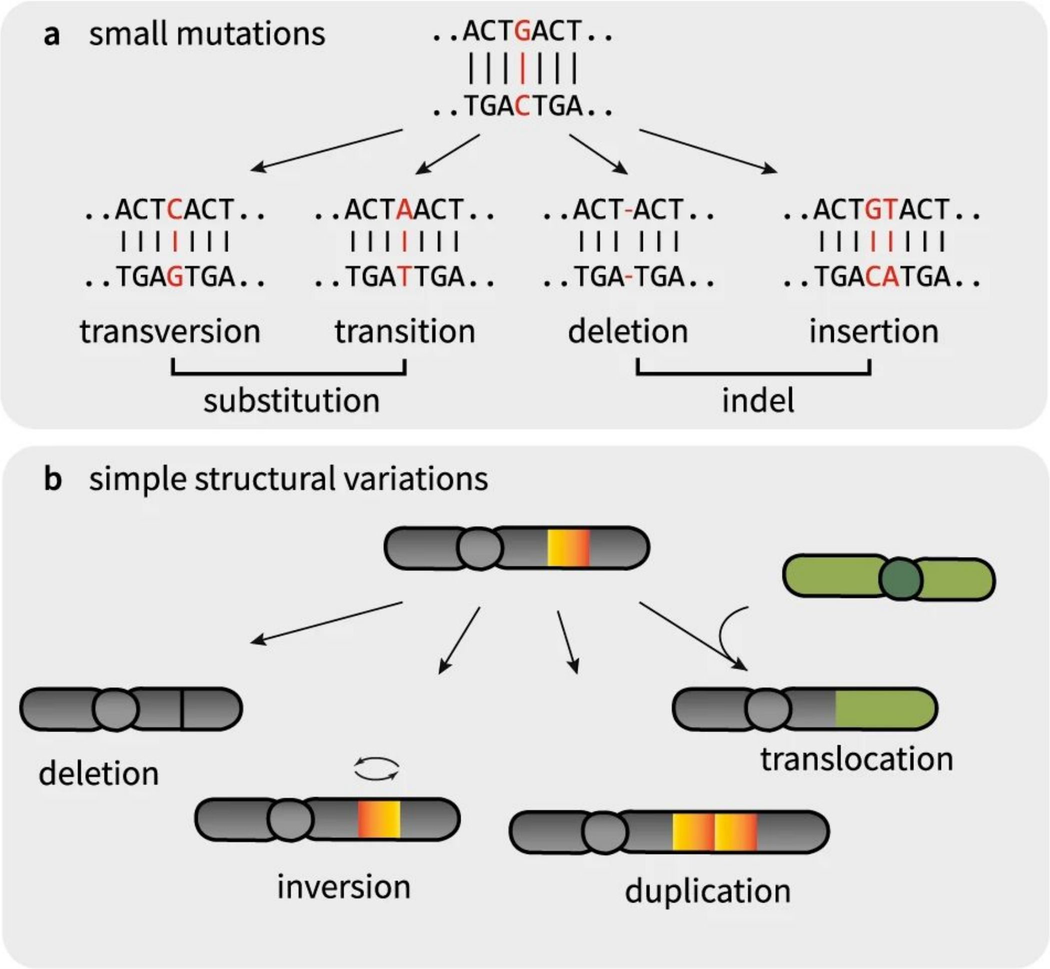
Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects

Allison A. Regier[1], Yossi Farjoun[2], David E. Larson[1], Olga Krasheninina[3], Hyun Min Kang[4], Daniel P. Howrigan[2], Bo-Juen Chen[5,11], Manisha Kher[5], Eric Banks[2], Darren C. Ames[6], Adam C. English[7], Heng Li[2], Jinchuan Xing[8], Yeting Zhang[8], Tara Matise[8], Goncalo R. Abecasis[4], Will Salerno[3], Michael C. Zody[5], Benjamin M. Neale[9,10] & Ira M. Hall[1]

Sherine Awad
Bio2Core Facility

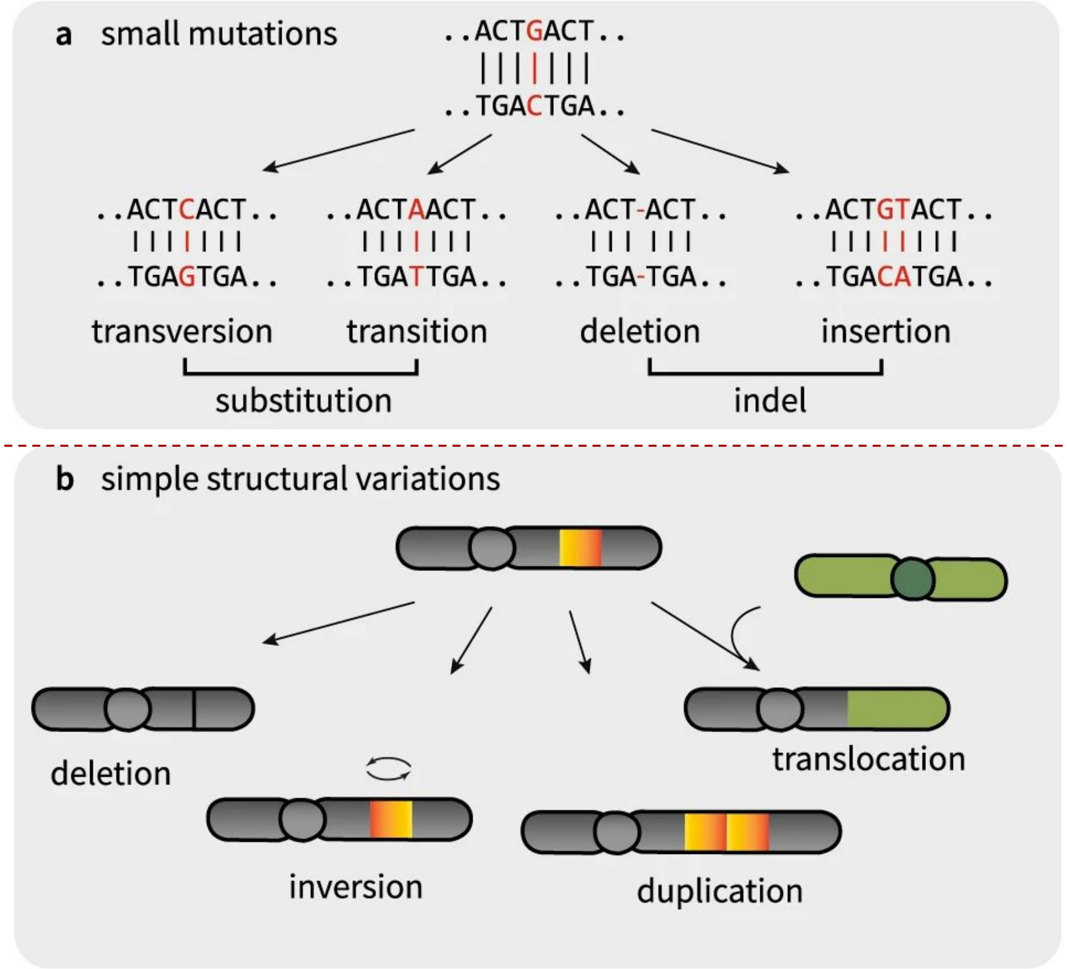
Variants

Structure Variants



Yi, K., & Ju, Y. S. (2018). Patterns and mechanisms of structural variations in human cancer. *Experimental & molecular medicine*, *50*(8), 1-11.

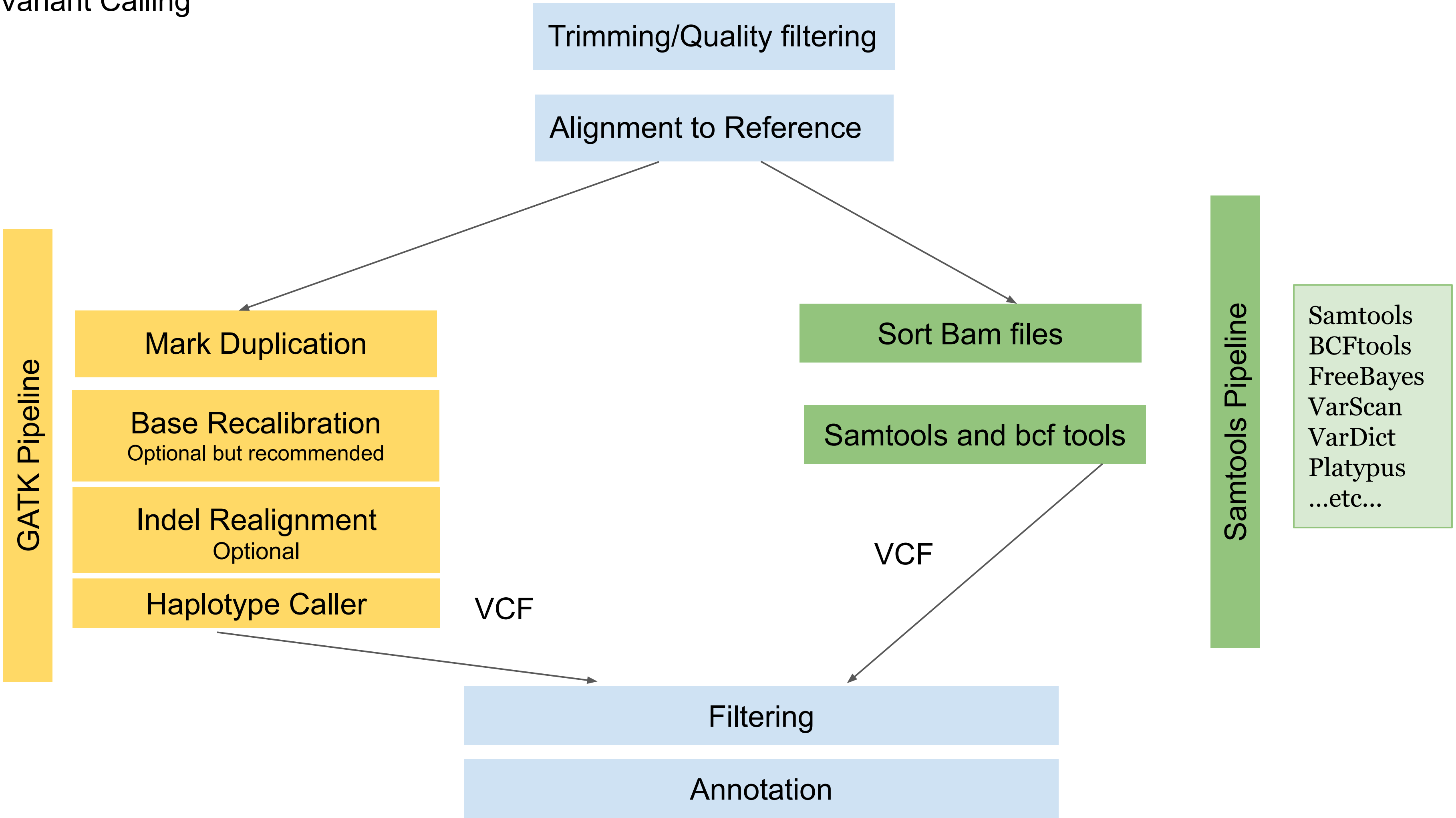
Variants



Structure Variants

Yi, K., & Ju, Y. S. (2018). Patterns and mechanisms of structural variations in human cancer. *Experimental & molecular medicine*, *50*(8), 1-11.

Variant Calling



Trimming/Quality filtering

Alignment to Reference

GATK Pipeline

Mark Duplication

Base Recalibration
Optional but recommended

Indel Realignment
Optional

Haplotype Caller

VCF

Sort Bam files

Samtools and bcf tools

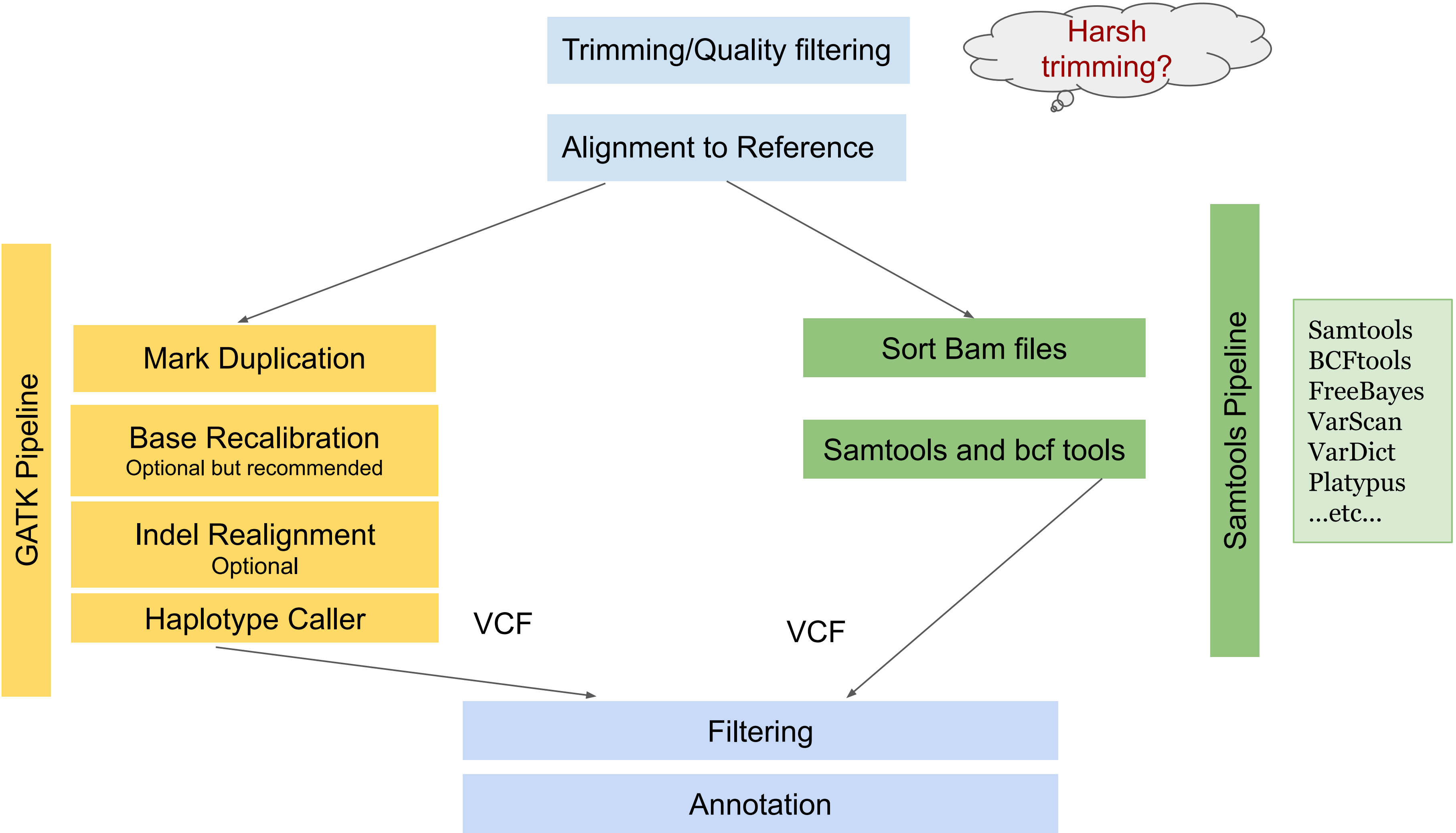
VCF

Samtools Pipeline

Samtools
BCFtools
FreeBayes
VarScan
VarDict
Platypus
...etc...

Filtering

Annotation



Trimming/Quality filtering

Harsh trimming?

Alignment to Reference

GATK Pipeline

Mark Duplication

Base Recalibration
Optional but recommended

Indel Realignment
Optional

Haplotype Caller

VCF

Samtools Pipeline

Sort Bam files

Samtools and bcf tools

VCF

Samtools
BCFtools
FreeBayes
VarScan
VarDict
Platypus
...etc...

Filtering

Annotation

Mark Duplicates

MarkDuplicates is important in removing PCR duplicates -- which can introduce bias in your variant calling. If you did not mark duplicates, you would risk having over-representation in your sequence of areas preferentially amplified during PCR.

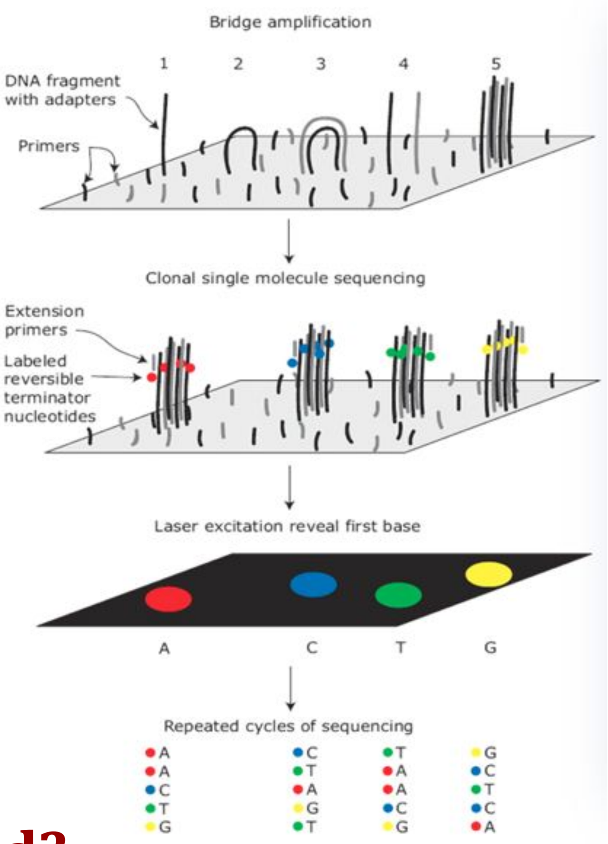
Almost all statistical models for variant calling assume some sort of independence between measurements. The duplicates (if one assumes that they arise from PCR artifact) are not independent. This lack of independence will usually lead to a breakdown of the statistical model and measures of statistical significance that are incorrect.

Ref	A
Alt	G
	G
	G
	G
	A

Base Recalibration

1. Sequencers are more often **over-confident** than **under-confident**, but we do occasionally see runs from sequencers that seemed to suffer from low self-esteem.

2. Recalibrate base quality scores in order to correct sequencing errors and other experimental artifacts.
3. G shoud be A **NO**
   We are 60% confident in this call instead of 85%

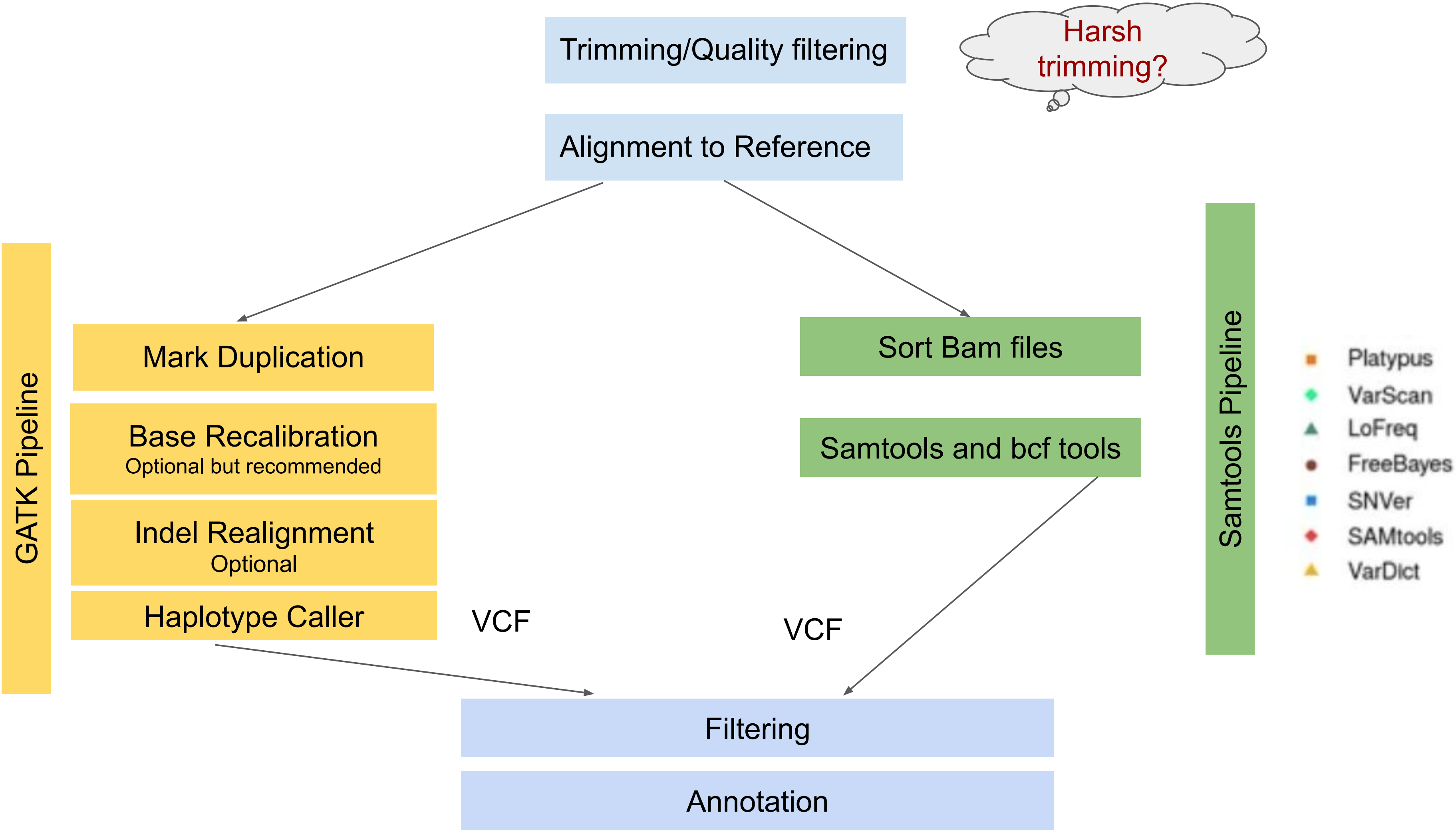

Bridge amplification
1
2
3
4
5
DNA fragment with adapters
Primers
Clonal single molecule sequencing
Extension primers
Labeled reversible terminator nucleotides
Laser excitation reveal first base
A
C
T
G
Repeated cycles of sequencing
A A C T G
C T A G T
T A A C G
G C T C A

Are these base qualities over estimated?

Indel Realignment

- Reads near detected indels are realigned to remove alignment artifacts.

- Genome aligners can only consider each read independently. Depending on the variant event and its relative location within a read, the aligner may favor alignments with mismatches or soft-clips instead of opening a gap in either the read or the reference sequence.
- Local realignment around indels allows correcting mapping errors made by genome aligners and make read alignments more consistent in regions that contain indels.



Trimming/Quality filtering

Harsh trimming?

Alignment to Reference

GATK Pipeline

Mark Duplication

Base Recalibration
Optional but recommended

Indel Realignment
Optional

Haplotype Caller

VCF

Samtools Pipeline

Sort Bam files

Samtools and bcf tools

VCF

■ Platypus
◆ VarScan
▲ LoFreq
● FreeBayes
■ SNVer
◆ SAMtools
▲ VarDict

Filtering

Annotation

Filtering = Throwing data

Every pipeline give guidelines for filtering. GATK offers V**ariant Quality Score Recalibration (VQSR)** or **hard filtering** when the set is too small for VQSR :

QualByDepth (QD) <2

This is the variant confidence (from the QUAL field) divided by the unfiltered depth of non-reference samples)

FisherStrand (FS) > 60.0

This is the Phred-scaled probability that there is strand bias at the site. When there little to no strand bias at the site, the FS value will be close to 0.

MappingQualityRankSuMest (MQRankSum) < -12.5

This is the u-based z-approximation from the Rank Sum Test for mapping qualities. It compares the mapping qualities of the reads supporting the reference allele and the alternate allele.

Etc ...

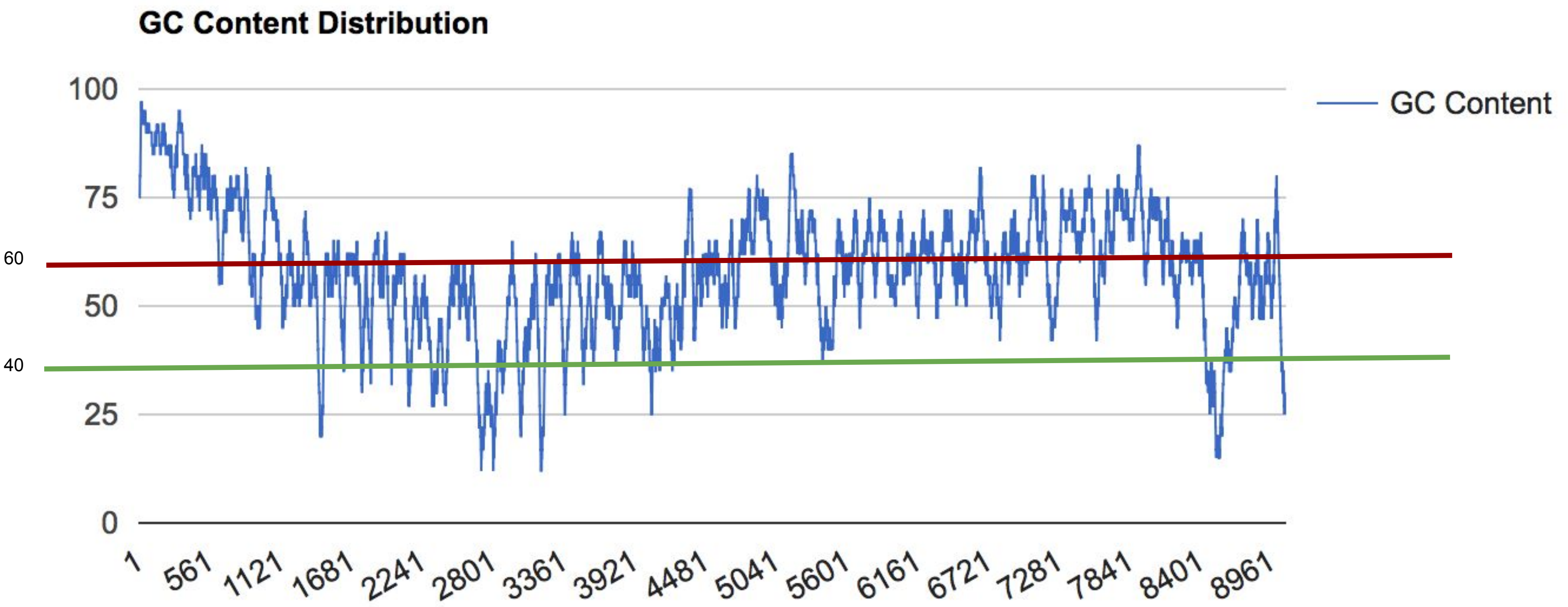
➔ Failing to take GC content into account may lead to false findings/Missings

➔ We expect low coverage in high/low GC content regions

- ❑ **Effect of DNA melting, etc. But not necessarily affect the alignment**

- Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PloS one*, *8*(4), e62856.
- Rieber, N., Zapatka, M., Lasitschka, B., Jones, D., Northcott, P., Hutter, B., ... & Pfister, S. (2013). Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PloS one*, *8*(6), e66621.

PPP1R26



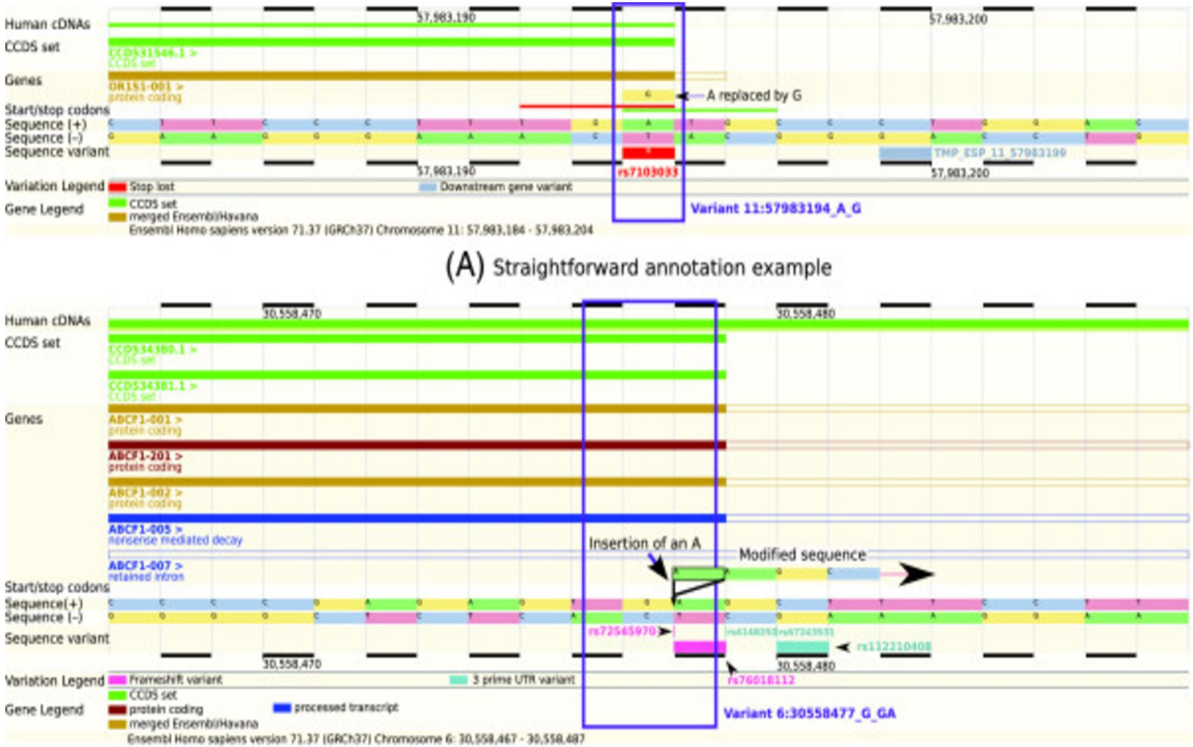
Summary: Full Length(9105bp) | A(21% 1920) | T(22% 1874) | G(30% 2773) | C(27% 2538)

Jia, P., Li, F., Xia, J., Chen, H., Ji, H., Pao, W., & Zhao, Z. (2012). Consensus rules in variant detection from next-generation sequencing data. *PloS one*, *7*(6), e38470.
Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PloS one*, *8*(4), e62856.

Chicago

Annotation

Choice of software has a large effect on variant annotation



(A) Straightforward annotation example

(B) More complex annotation example

Stop codon TGA

Example A: A is replaced by G so TGA become TGG.
Unambiguously: Stop loss variant

Example B: Penultimate base of the exon. A is inserted. TGA becomes TGA

=It is a single-base insertion, so could be annotated as a **frameshift variant.**

=It is an insertion in a stop codon, so could be a **stop-loss variant.**

=The final codon, TGA (stop codon), remains TGA with this variant (insertion of a single base A), so it is actually **a synonymous variant.**

Annovar annotates it as frameshift insertion and VEP as stop-loss, when using ensembl transcripts

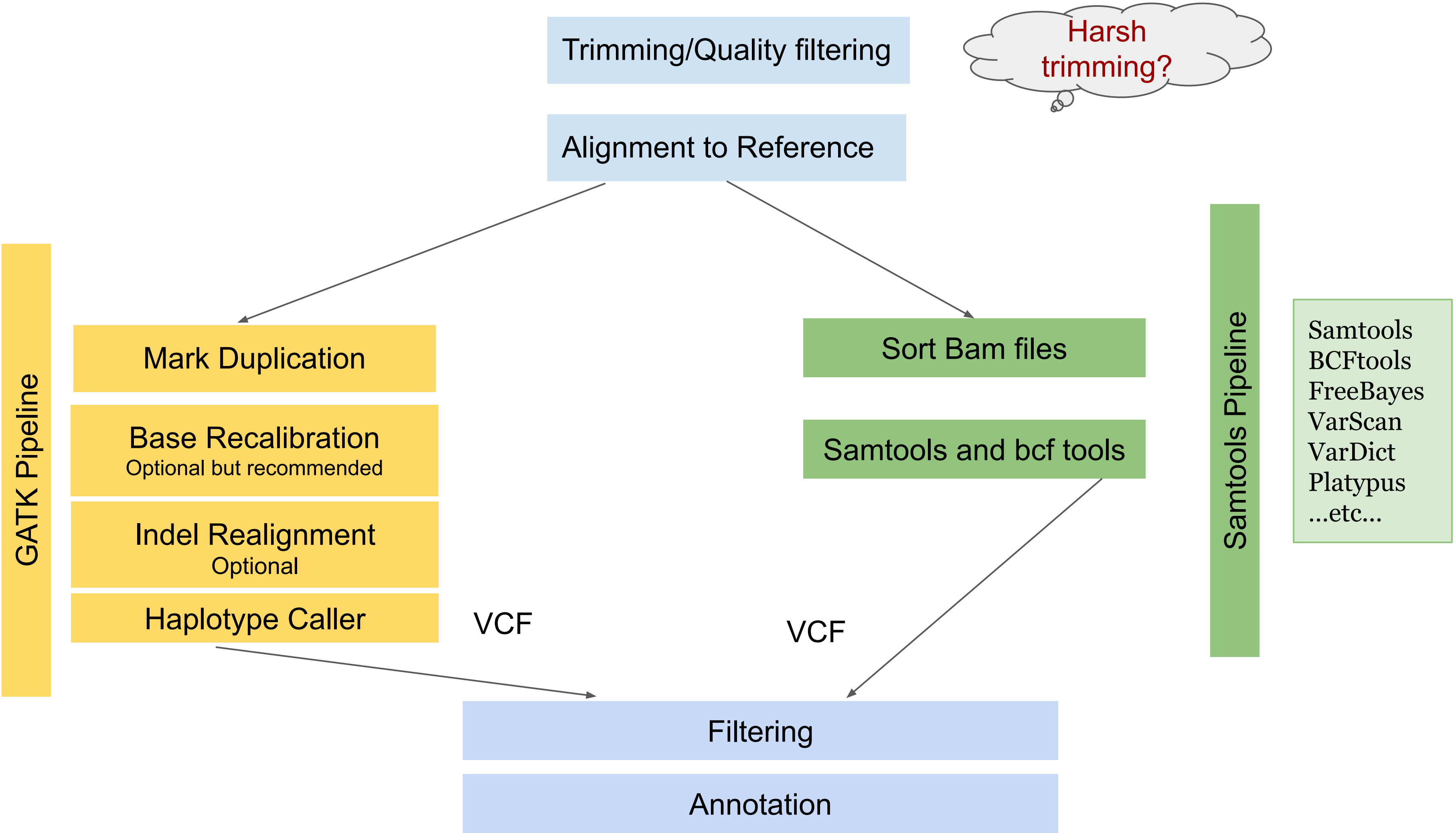
McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., Cazier, J.-B., & Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, *6*(3), 26.

GnomAD - ExAC Get a MAF from populations: look for **rare** variants and filter polymorphism **(How rare is rare?)**

Mutation Taster: DNA sequence variants for their disease-causing potential.

GERP: if a variant is in highly conserved region then it is likely pathogenic

Etc



Trimming/Quality filtering

Harsh trimming?

Alignment to Reference

GATK Pipeline

Mark Duplication

Base Recalibration
Optional but recommended

Indel Realignment
Optional

Haplotype Caller

VCF

Sort Bam files

Samtools and bcf tools

VCF

Samtools Pipeline

Samtools
BCFtools
FreeBayes
VarScan
VarDict
Platypus
...etc...

Filtering

Annotation

PPP1R26: A curly hair gene?

F171

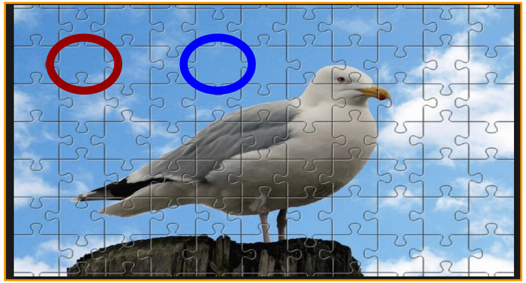
F74

c.C181T:p.R61W

Highly conserved region
Amino acid is conserved in Human, Gorilla, Chimpanzee, Cynomolgus Monkey, Green Monkey
Kinship test !

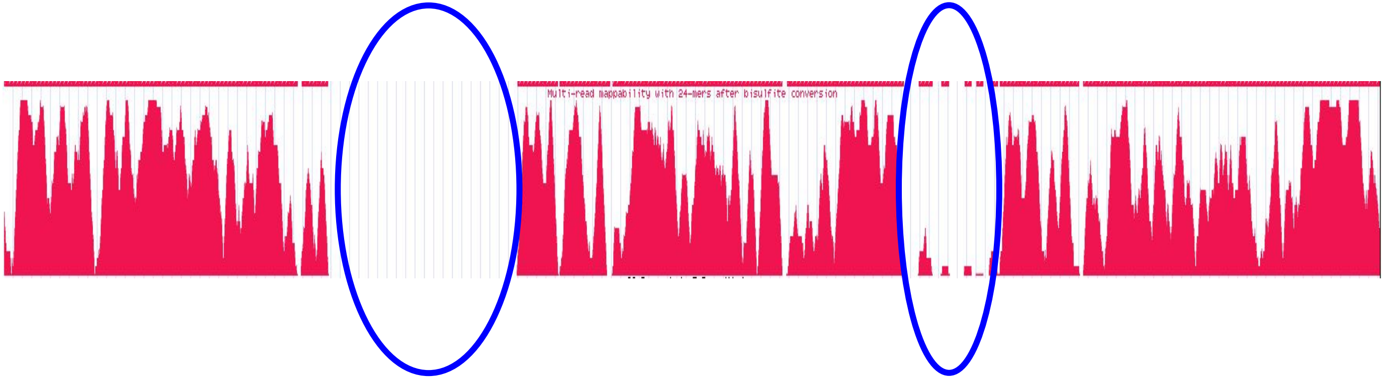
König, I. R. (2011). Validation in genetic association studies. *Briefings in bioinformatics*, *12*(3), 253-258.

Human Genome exhibits lots of similarity !



Region Mappability: Duke uniqueness (35 bp)

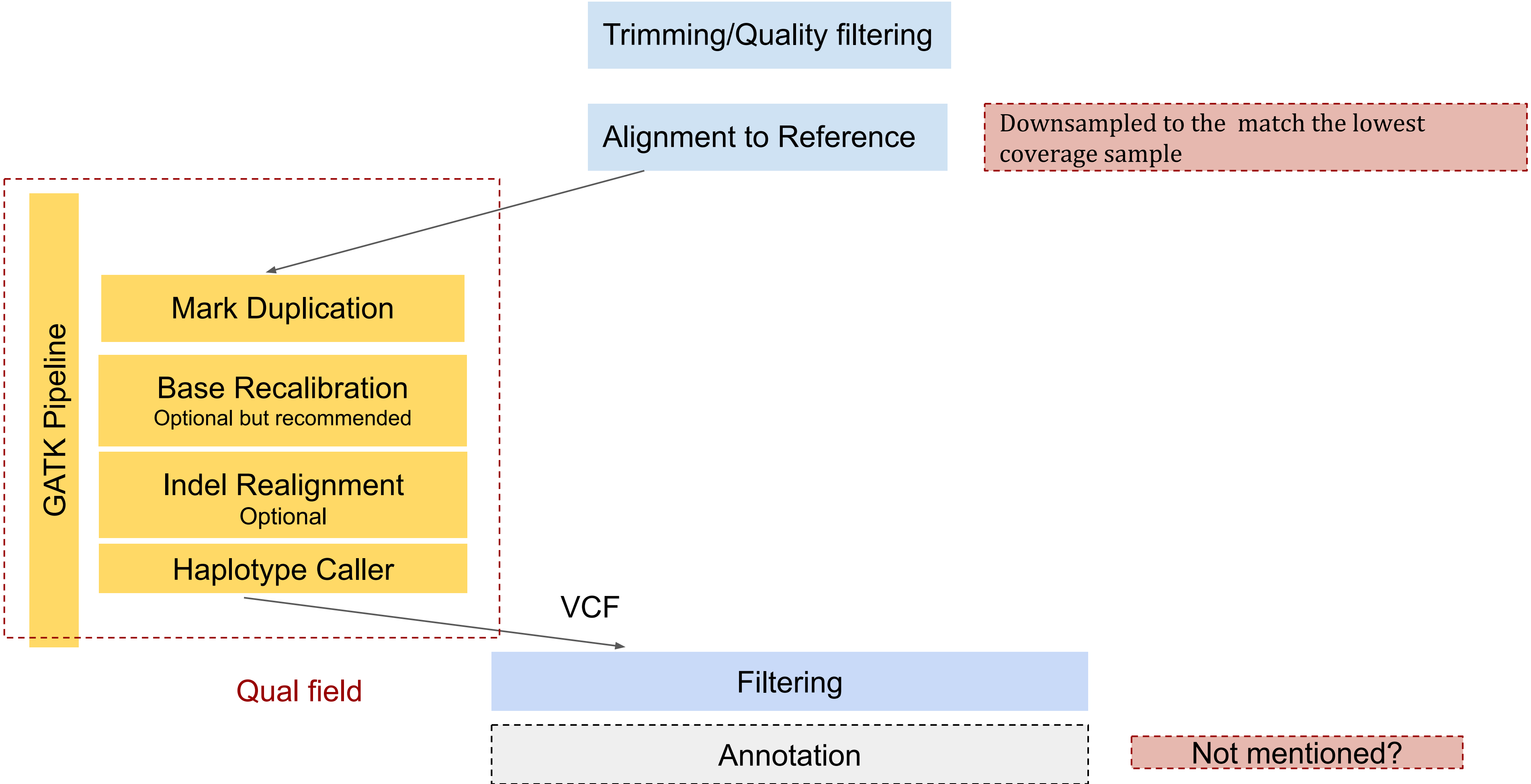
Region mappality gives greater noise reductions.



The Duke uniqueness tracks display how unique is each sequence on the positive strand starting at a particular base and of a particular length. Thus, the 20 bp track reflects the uniqueness of all 20 base sequences with the score being assigned to the first base of the sequence. Scores are normalized to between 0 and 1 with 1 representing a completely unique sequence and 0 representing the sequence occurs >4 times in the genome (excluding chrN_random and alternative haplotypes). A score of 0.5 indicates the sequence occurs exactly twice, likewise 0.33 for three times and 0.25 for four times.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., ... & Li, H. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43.

Pipeline as of the paper



Trimming/Quality filtering

Alignment to Reference

Downsampled to the match the lowest coverage sample

GATK Pipeline

Mark Duplication

Base Recalibration
Optional but recommended

Indel Realignment
Optional

Haplotype Caller

VCF

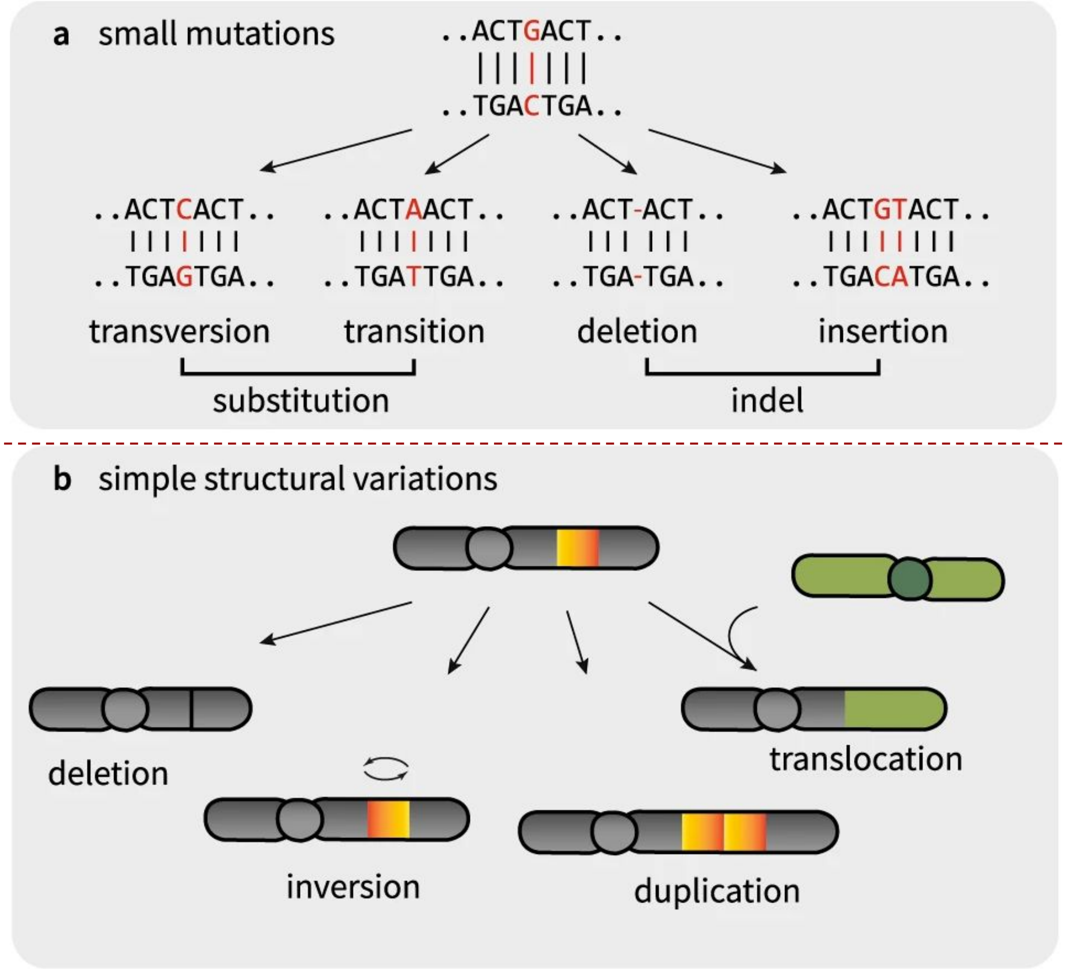
Qual field

Filtering

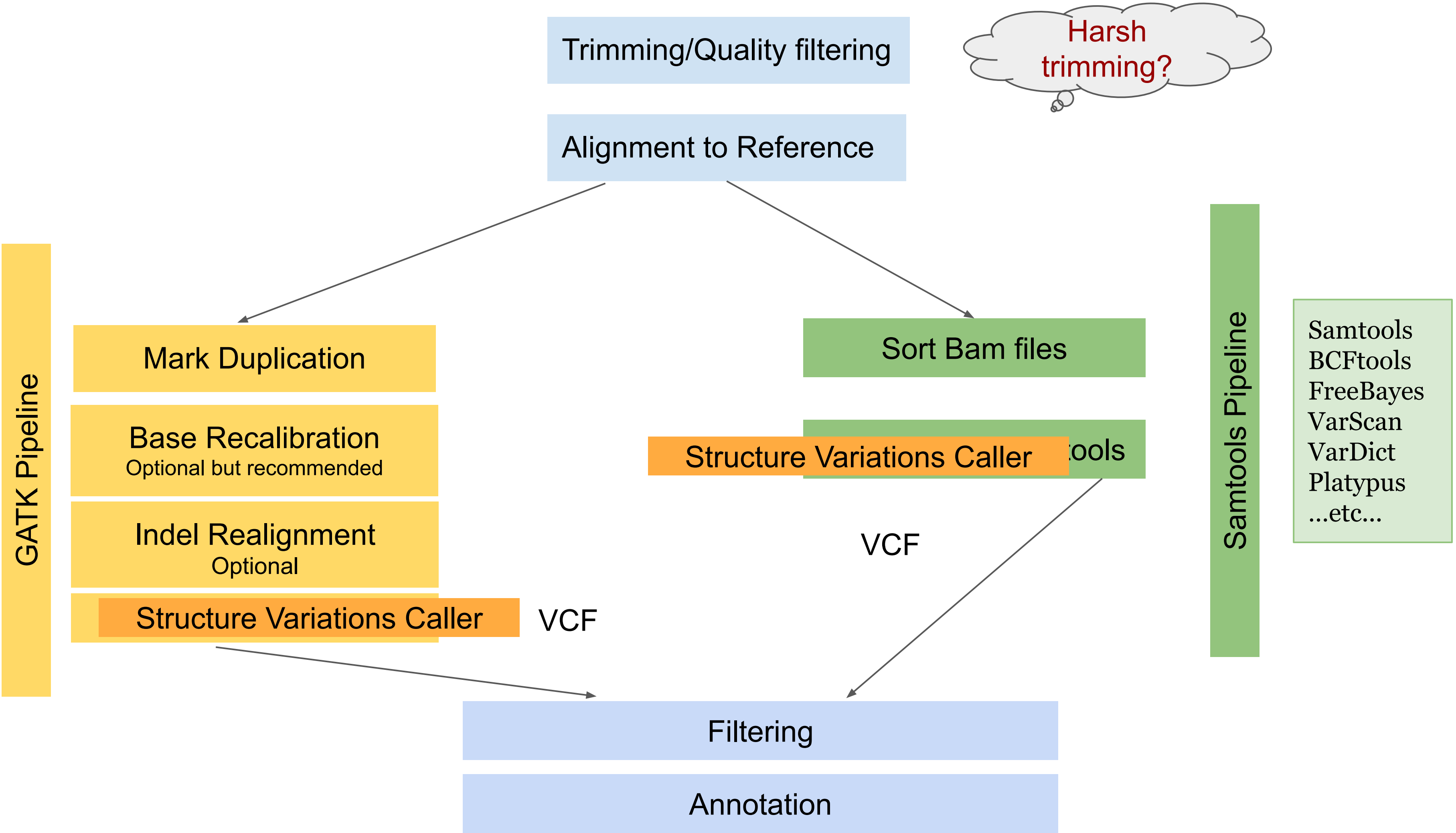
Annotation

Not mentioned?

Structure Variants



Yi, K., & Ju, Y. S. (2018). Patterns and mechanisms of structural variations in human cancer. *Experimental & molecular medicine*, *50*(8), 1-11.



Trimming/Quality filtering

Harsh trimming?

Alignment to Reference

GATK Pipeline

Mark Duplication

Base Recalibration
Optional but recommended

Indel Realignment
Optional

Structure Variations Caller

VCF

Sort Bam files

ools

Structure Variations Caller

VCF

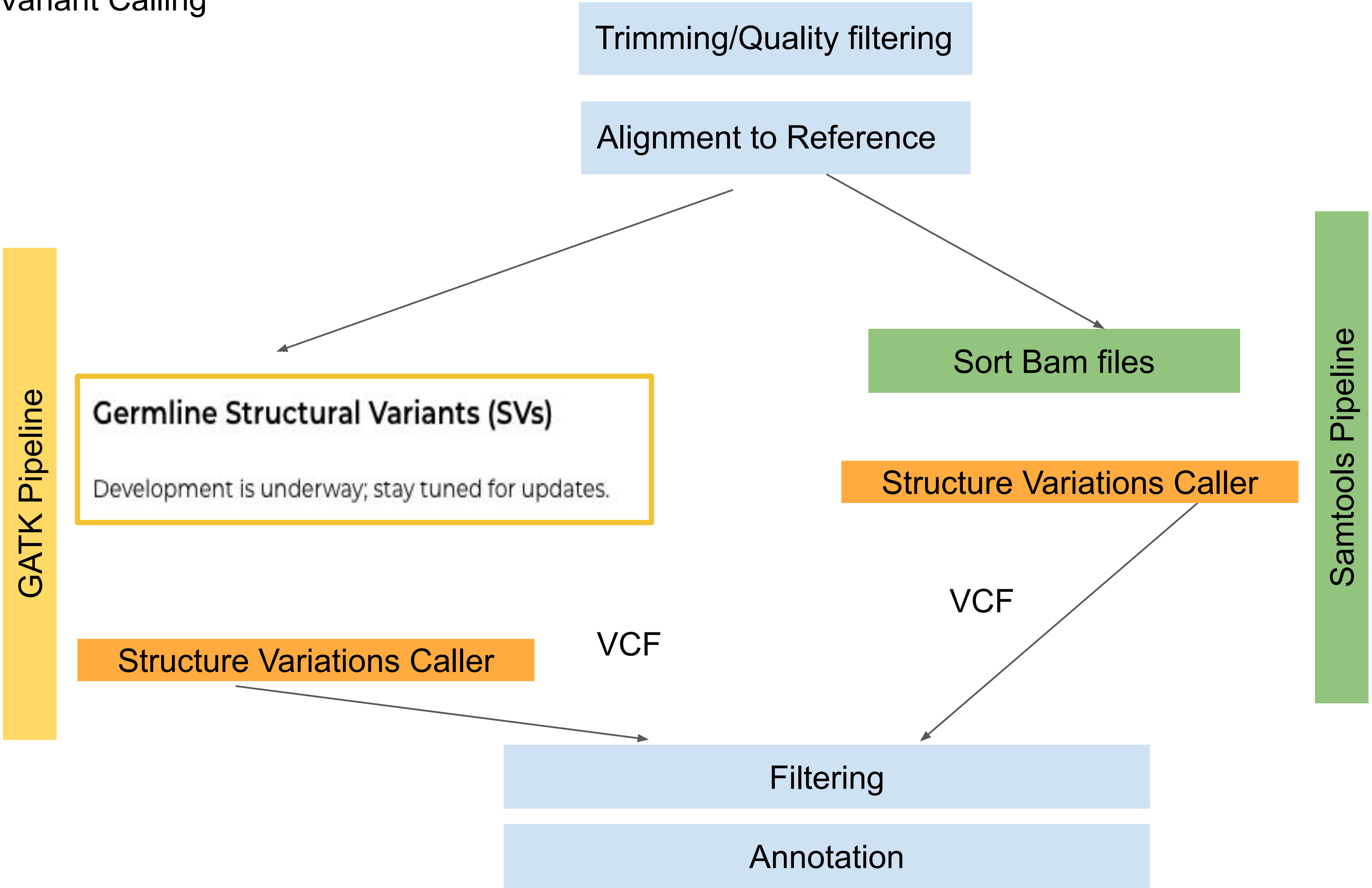
Samtools Pipeline

Samtools
BCFtools
FreeBayes
VarScan
VarDict
Platypus
...etc...

Filtering

Annotation

Variant Calling



Trimming/Quality filtering

Alignment to Reference

GATK Pipeline

Germline Structural Variants (SVs)

Development is underway; stay tuned for updates.

Structure Variations Caller

VCF

Samtools Pipeline

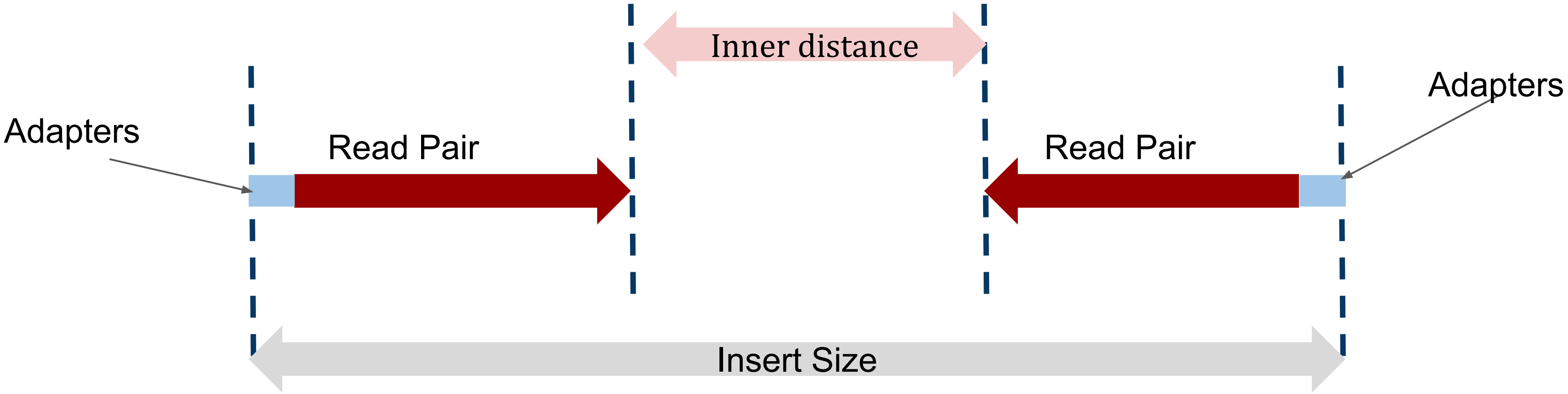
Sort Bam files

Structure Variations Caller

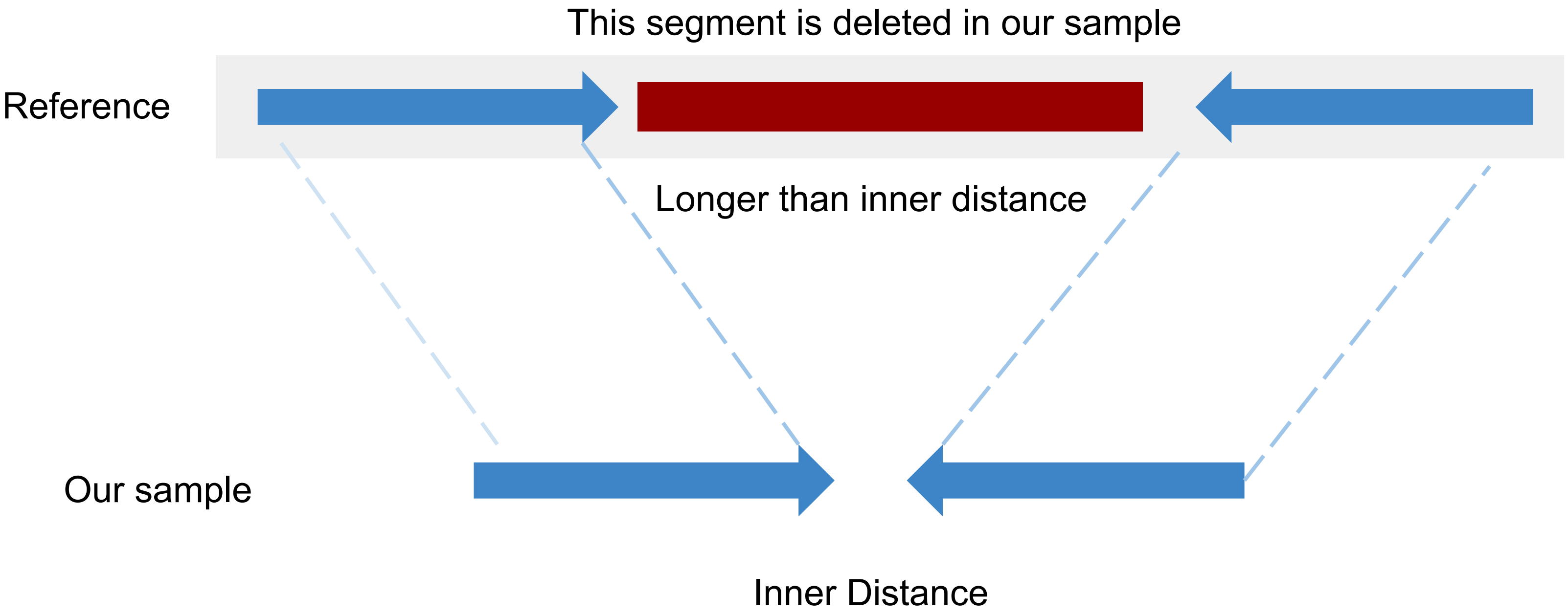
VCF

Filtering

Annotation

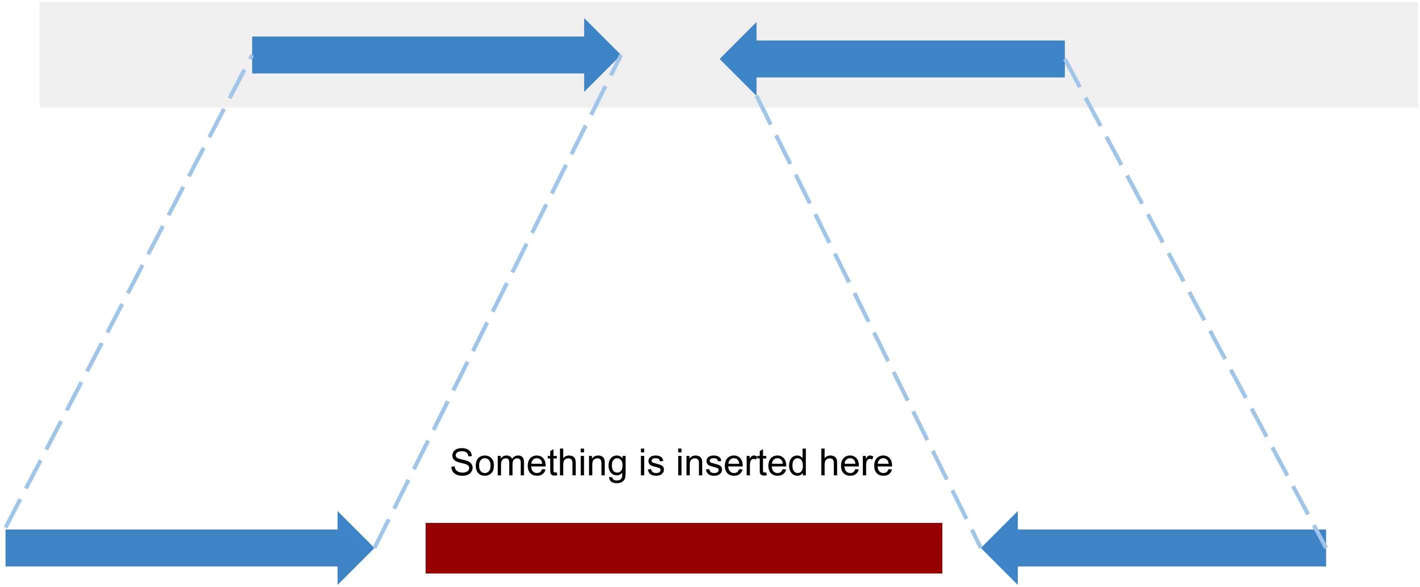


- We know the read length
- We know the insert size
- We need the inner distance, just a subtraction !


This segment is deleted in our sample
Reference
Longer than inner distance
Our sample
Inner Distance

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, *3*, 92.

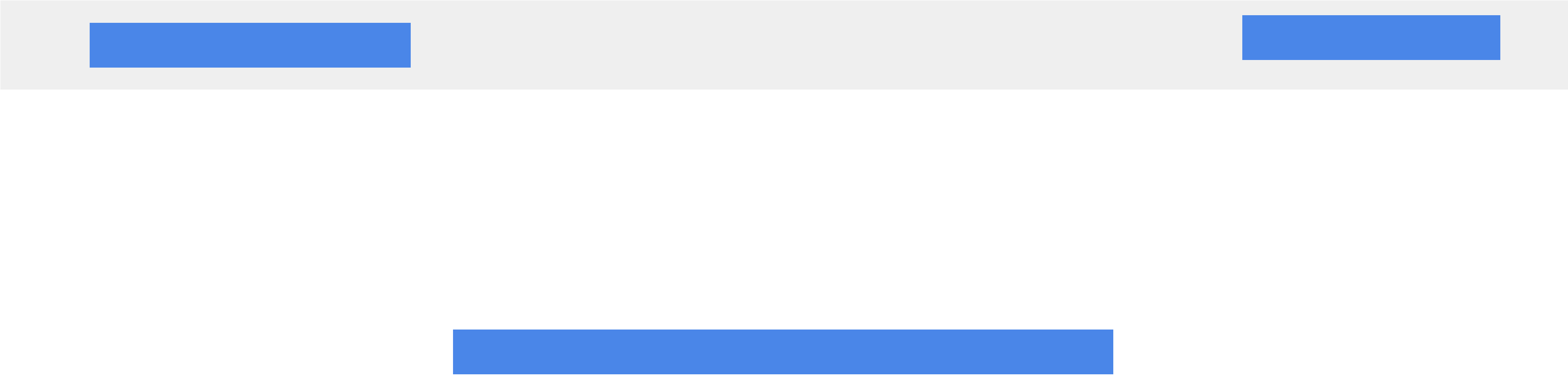
Shorter distance than our inner distance



Something is inserted here

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, *3*, 92.

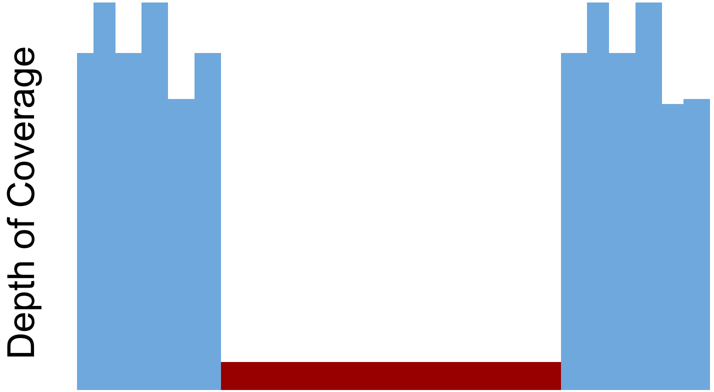
Split read



Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, *3*, 92.

Depth of Coverage

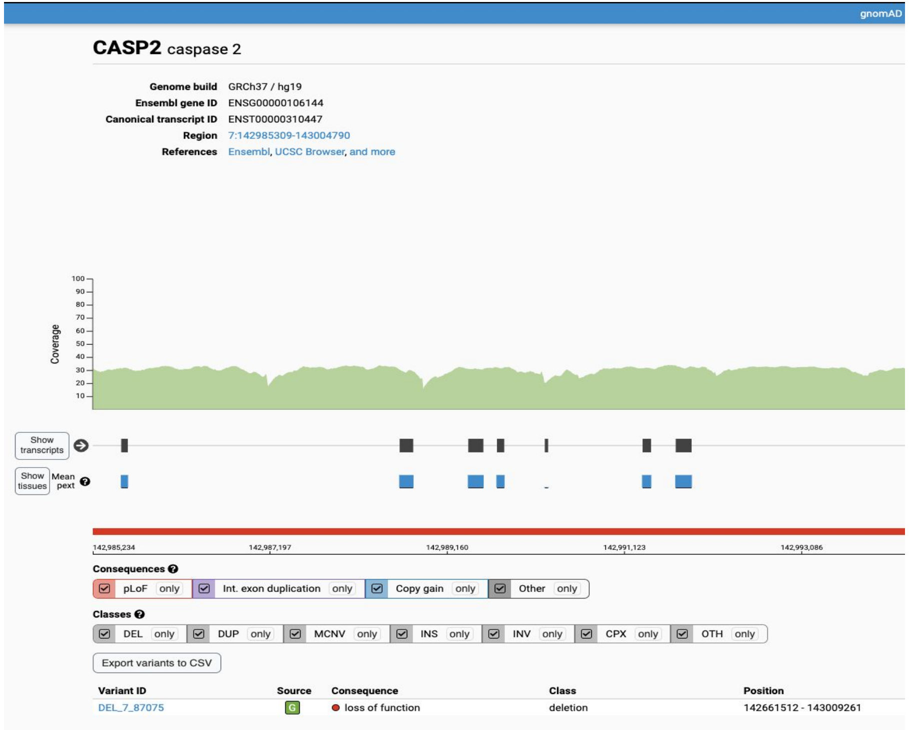
Unifrom Coverage



- Tiddit rely on depth so it forces user to mark duplicates.
- Delly uses read direction so it doesn’t force marking duplicates.

Tattini, L., D’Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, *3*, 92.
Ameur, A., Dahlberg, J., Olason, P., Vezzi, F., Karlsson, R., Martin, M., ... & Thutkawkorapin, J. (2017). SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *European Journal of Human Genetics*, *25*(11), 1253-1260.
Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333-i339.

GnomAD



gnomAD

CASP2 caspase 2

Genome build	GRCh37 / hg19
Ensembl gene ID	ENSG00000106144
Canonical transcript ID	ENST00000310447
Region	7:142985309-143004790
References	Ensembl, UCSC Browser, and more

Consequences

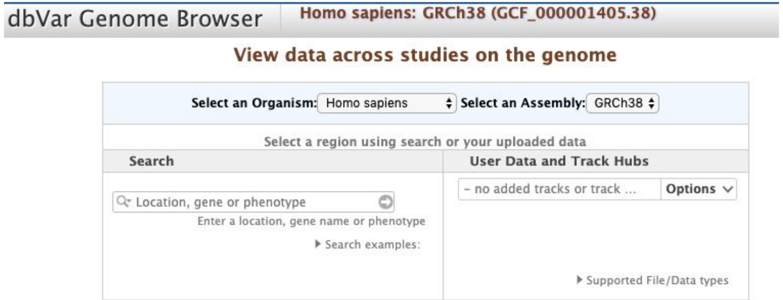
☑ pLoF only ☑ Int. exon duplication only ☑ Copy gain only ☑ Other only

Classes

☑ DEL only ☑ DUP only ☑ MCNV only ☑ INS only ☑ INV only ☑ CPX only ☑ OTH only

Export variants to CSV

Variant ID	Source	Consequence	Class	Position
DEL_7_87075	G	● loss of function	deletion	142661512 - 143009261



dbVar Genome Browser Homo sapiens: GRCh38 (GCF_000001405.38)

View data across studies on the genome

Select an Organism: Homo sapiens Select an Assembly: GRCh38

Select a region using search or your uploaded data

Search	User Data and Track Hubs
Location, gene or phenotype Enter a location, gene name or phenotype Search examples:	– no added tracks or track ... Options Supported File/Data types

Database of Genomic Variants

A curated catalogue of human genomic structural variation

DECIPHER GRCh37

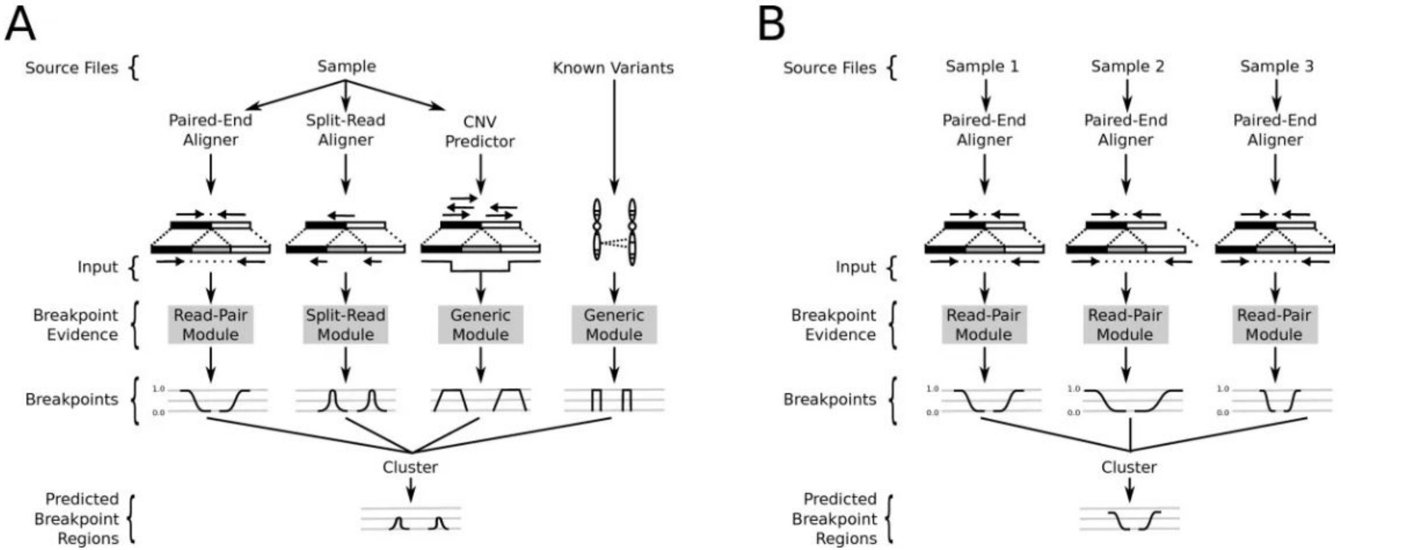
SURVIVOR (Merging, comparing, ..etc): Essential tool, traditional tools or eye merging won't work as simple as point SNP

Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast.
Jeffares, Daniel C; Jolly, Clemency; Hoti, Mimoza; Speed, Doug; Shaw, Liam; Rallis, Charalampos; Balloux, Francois; Dessimoz, Christophe; Bähler, Jürg; Sedlazeck, Fritz J.
Nature communications, Vol. 8, 14061, 24.01.2017, p. 1-11. DOI:10.1038/NCOMMS14061

AnnotSV (Annotation): Using traditional tools doesn't work, as it is not one point SNP

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., & Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, *34*(20), 3572-3574.

Paper used LUMPY (MSQ field for quality)



The LUMPY framework for integrating multiple structural variation signals. (A) A scenario in which LUMPY integrates three different sequence alignment signals (read-pair, split-read and read-depth) from a genome single sample. Additionally, sites of known variants are provided to LUMPY as prior knowledge in order to improve sensitivity. **(B)** A single signal type (in this case, read-pair) that is integrated from three different genome samples. We present these as example scenarios and emphasize that multi-signal and multi-sample workflows are not mutually exclusive. CNV, copy number variation.

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome biology*, *15*(6), R84.

Functional Equivalence (FE)

They define FE to be a shared property of two pipelines that can be run independently on the same raw WGS data to produce two output files that, upon analysis by the same variant caller(s), produce virtually indistinguishable genome variation maps.

Datasets

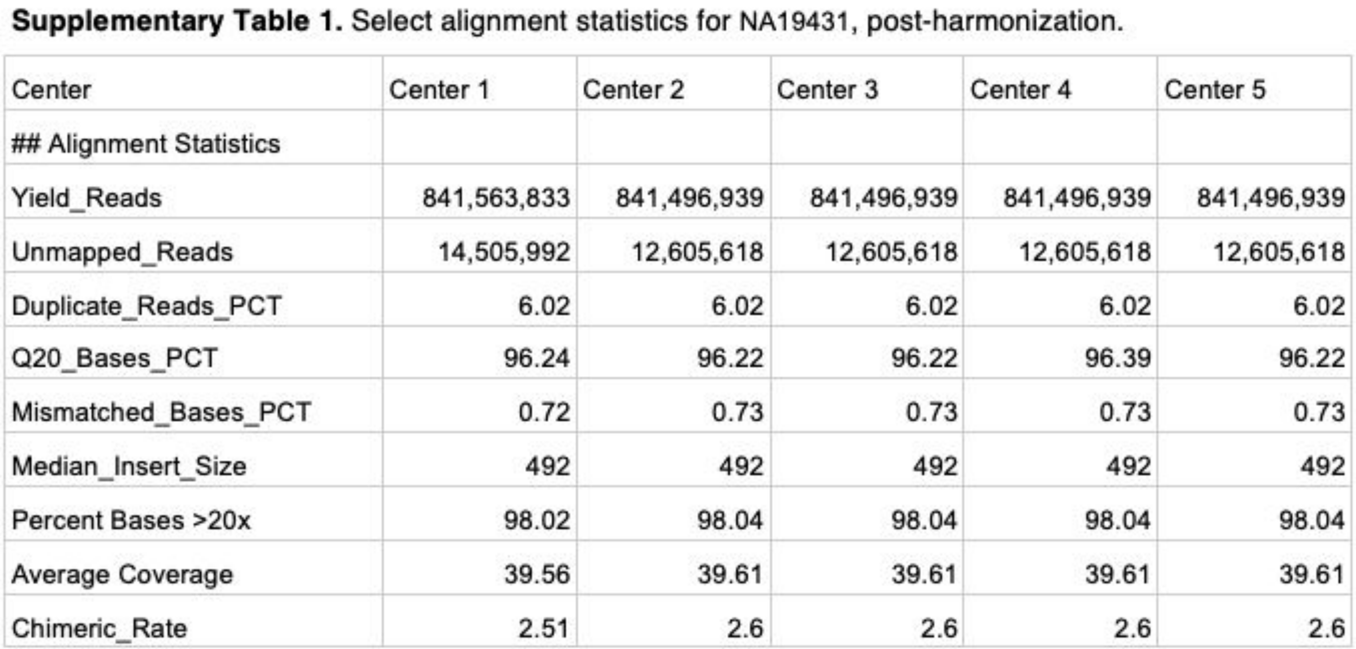
14 dataset, different ancestory of well studied samples from 1000 Genomes, includes:

4 independently sequenced NA12878 (CEPH)

2 replicates of NA19238 (Yoruban)

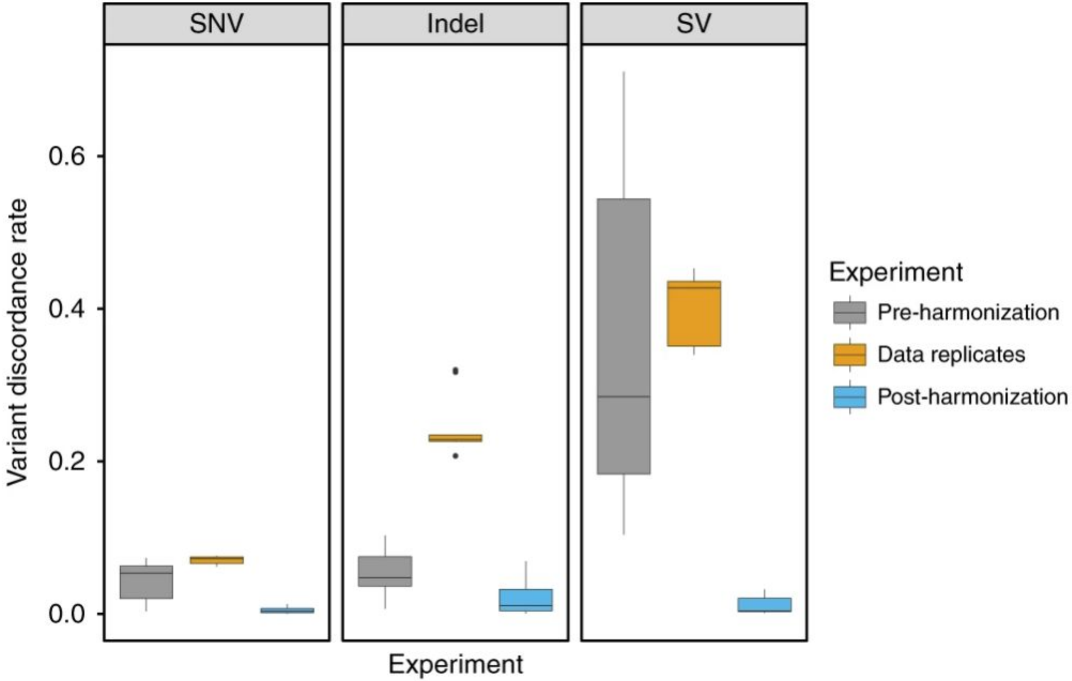
Deep coverage (20x) illumina HiSeq

Five centers of five pipelines?



Supplementary Table 1. Select alignment statistics for NA19431, post-harmonization.

Center	Center 1	Center 2	Center 3	Center 4	Center 5
## Alignment Statistics					
Yield_Reads	841,563,833	841,496,939	841,496,939	841,496,939	841,496,939
Unmapped_Reads	14,505,992	12,605,618	12,605,618	12,605,618	12,605,618
Duplicate_Reads_PCT	6.02	6.02	6.02	6.02	6.02
Q20_Bases_PCT	96.24	96.22	96.22	96.39	96.22
Mismatched_Bases_PCT	0.72	0.73	0.73	0.73	0.73
Median_Insert_Size	492	492	492	492	492
Percent Bases >20x	98.02	98.04	98.04	98.04	98.04
Average Coverage	39.56	39.61	39.61	39.61	39.61
Chimeric_Rate	2.51	2.6	2.6	2.6	2.6

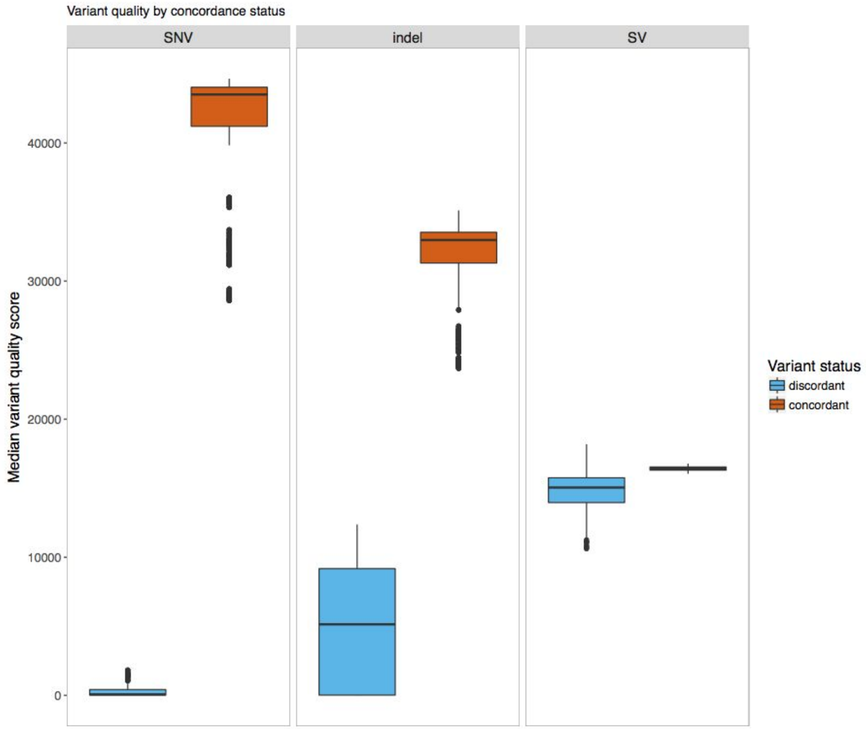


Pairwise variant discordance rates were calculated between pipelines from each of five centers (pre-harmonization and post-harmonization) as well as between independent sequencing replicates of the same individuals processed by the same pipeline (data replicates). From left, single nucleotide (SNV) and small insertion/deletion (indel) variants were detected with GATK, and structural variants (SV) with LUMPY. The pre-harmonization and post-harmonization comparisons include 14 independently sequenced samples. The data replicate comparisons include four replicates of NA12878 and two replicates of NA19238. Note that the extremely high levels of discordance for SVs pre-harmonization are largely due to variable use of decoy sequences in the reference genomes used by the different centers. The center line is the median, the upper and lower hinges are the first and third quartiles, and the whiskers extend to the largest/smallest values no further than 1.5 * inter-quartile range from the hinge

- Applied final pipeline version to an independent set of 100 genomes comprising 8 trios from 1000 genomes project and 19 quads from Simons Simplex Collection

- Diverse ancestory: vitnamese, S. Han Chineese, peurto rican, Gambian, Caucasian.

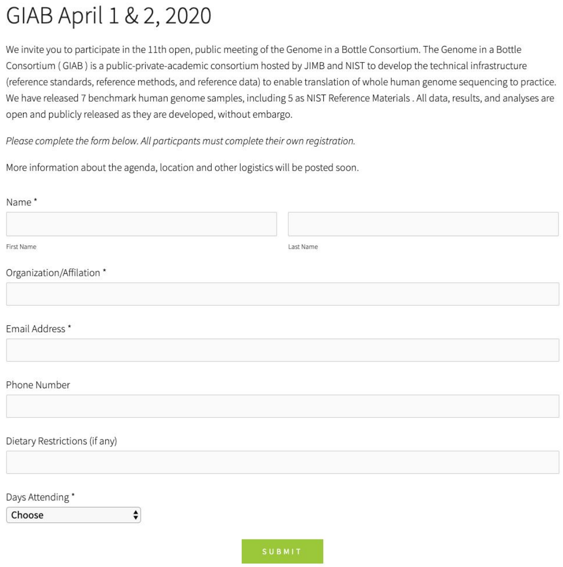
- The most majority of GATK variants (97.2%) are identified in data from all five pipelines with only 1.74% unique to a single pipeline.
- !!!!



Supplementary Figure 3. Variant quality score by concordance status. The median variant quality score (QUAL field from the GATK VCF; MSQ INFO field from the LUMPY SV VCF) was calculated for each sample, with variants partitioned by their status in each pairwise pipeline comparison. The center line is the median, the upper and lower hinges are the first and third quartiles, and the whiskers extend to the largest/smallest values no further than 1.5 * inter-quartile range from the hinge.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$



GIAB April 1 & 2, 2020

We invite you to participate in the 11th open, public meeting of the Genome in a Bottle Consortium. The Genome in a Bottle Consortium (GIAB) is a public-private-academic consortium hosted by JIMB and NIST to develop the technical infrastructure (reference standards, reference methods, and reference data) to enable translation of whole human genome sequencing to practice. We have released 7 benchmark human genome samples, including 5 as NIST Reference Materials . All data, results, and analyses are open and publicly released as they are developed, without embargo.

Please complete the form below. All particpants must complete their own registration.

More information about the agenda, location and other logistics will be posted soon.

Name *

First Name

Last Name

Organization/Affilation *

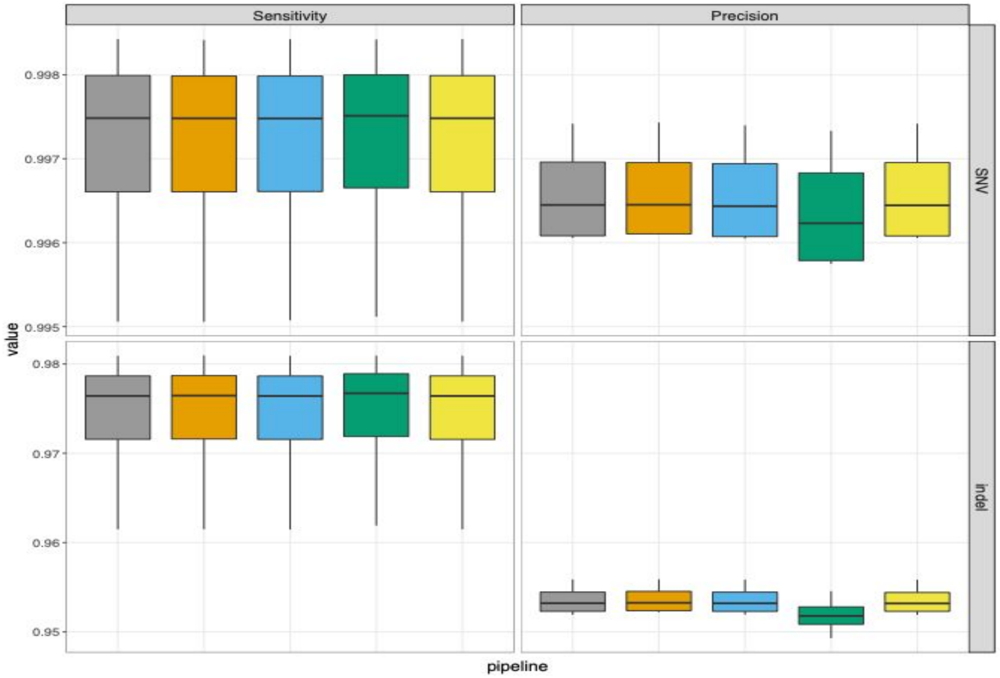
Email Address *

Phone Number

Dietary Restrictions (if any)

Days Attending *

Choose



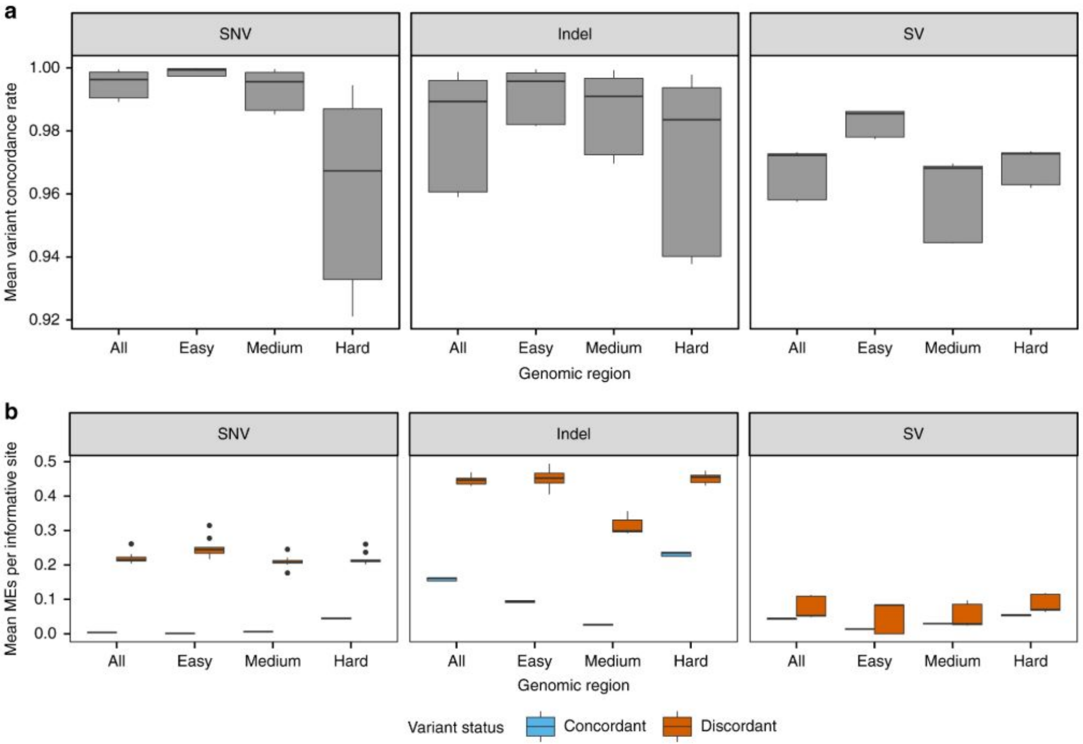
Supplementary Figure 2. Comparison to gold standard variants. Sensitivity and precision to the GiaB gold standard variants were very similar across pipelines for all four NA12878 replicates, although one center with different pipeline components is slightly more sensitive and less precise. The center line is the median, the upper and lower hinges are the first and third quartiles, and the whiskers extend to the largest/smallest values no further than 1.5 * inter-quartile range from the hinge.

https://www.nist.gov/programs-projects/genome-bottle

Mendelian Error rate calculation

- SNPs and indel classified into **shared between pipelines or unique to one pipeline**
- Exclude **missing genotypes or uniformly homozygous genotypes**
- Used a python script to classify a variant as **uninformative, informative with no Mendelian error, or informative with Mendelian error**

GC content?



Variant concordance and Mendelian error (ME) rates were calculated for different variant classes and genomic regions using 100 samples, including 8 trios from the 1000 Genomes Project and 19 quads from the Simons Simplex Collection. **a** Variant concordance rates were calculated from pairwise comparisons across five pipelines for 100 samples. **b** Mendelian error rates were calculated using informative sites in 44 parent-offspring trios, for variants classified as concordant and discordant in pairwise comparisons between five pipelines. The center line is the median, the upper and lower hinges are the first and third quartiles, and the whiskers extend to the largest/smallest values no further than 1.5 * inter-quartile range from the hinge

Summary

- Filtering on qual ?
- 5 pipelines?
- Harmonized pipeline doesn’t necessarily work on a different dataset. Tweeking parameters could be dataset dependent !
- Concordance and uniquness/validation, Skipping filtering
  - These 97% could be low quality and need to be thrown away, relying on qual is not enough
  - These uniqueness could be true variant when validated, so other pipelines are missing it
- Truthset? Sanger sequence validated?
- Annotation

Thank You