GATK Best Practices for Variant Discovery

Joint variant calling

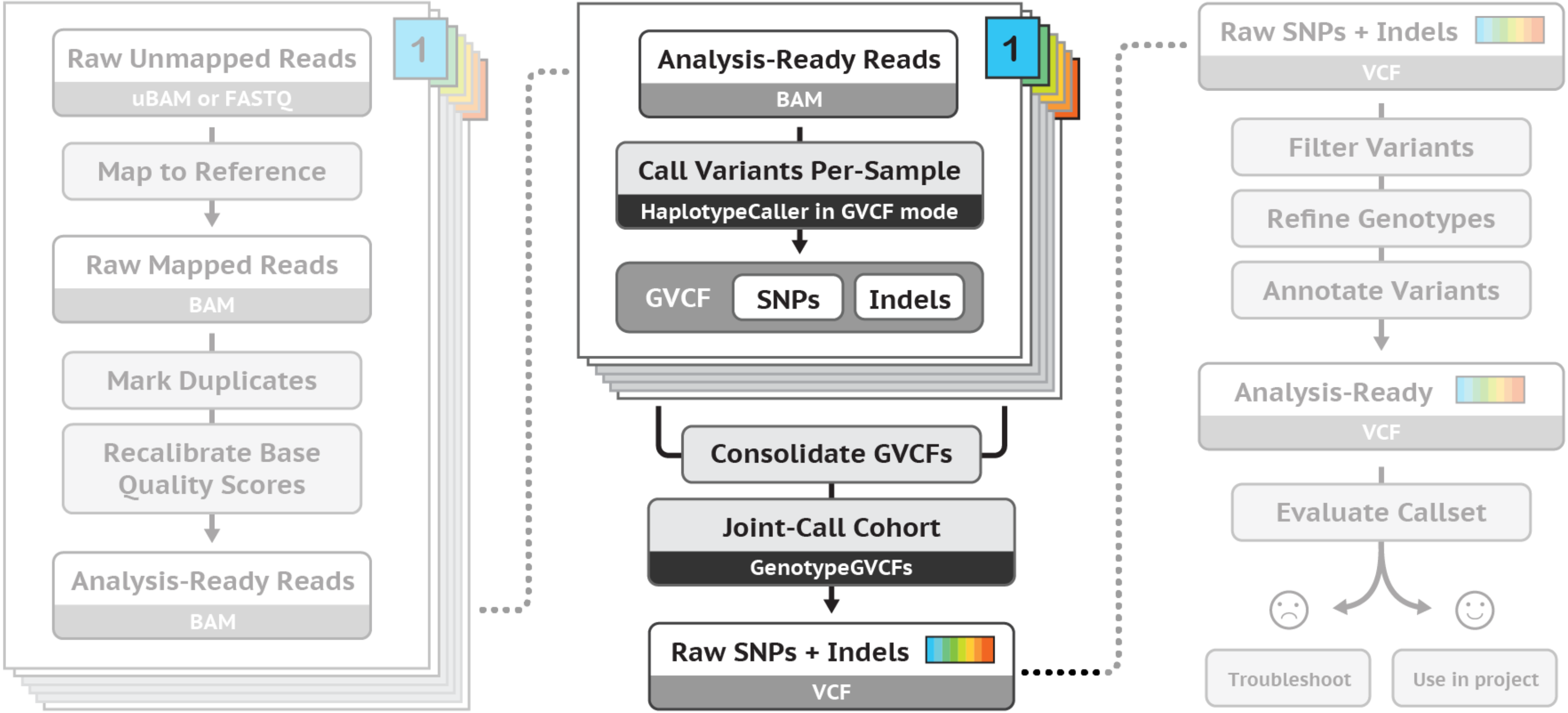
GVCF-based workflow using GenomicsDB and GenotypeGVCFs

BROAD INSTITUTE

https://software.broadinstitute.org/gatk/

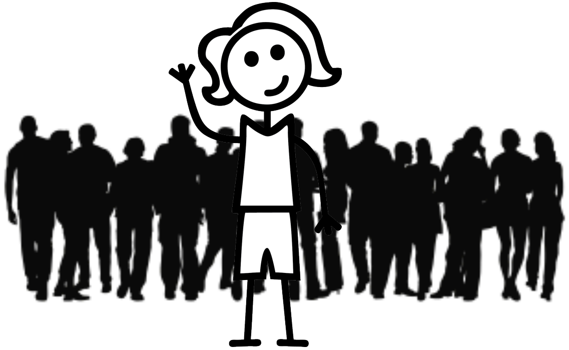
gatk

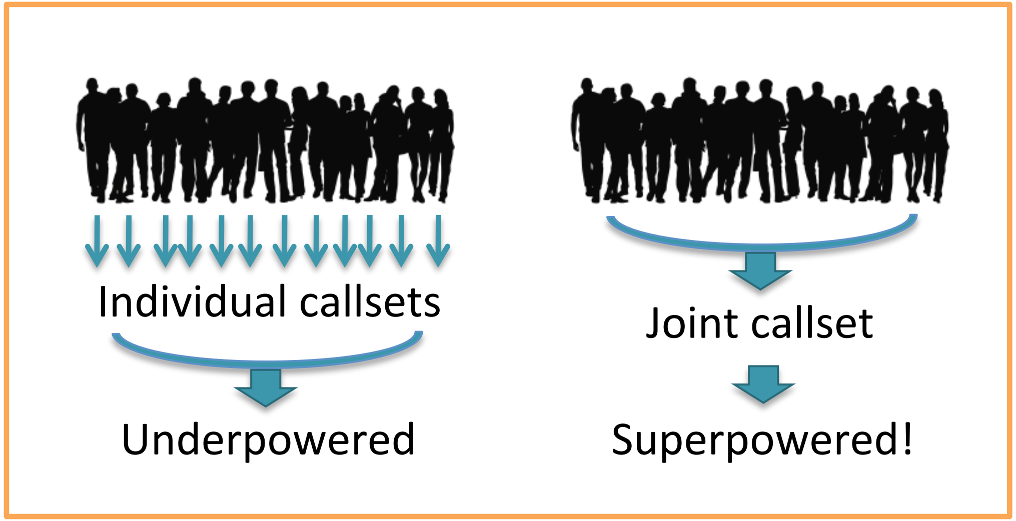
Best Practices for Germline SNP & INDEL Discovery



Joint analysis empowers discovery

- Single genome in isolation: almost never useful
- Family or population data add valuable information
  - rarity of variants
  - *de novo* mutations
  - ethnic background





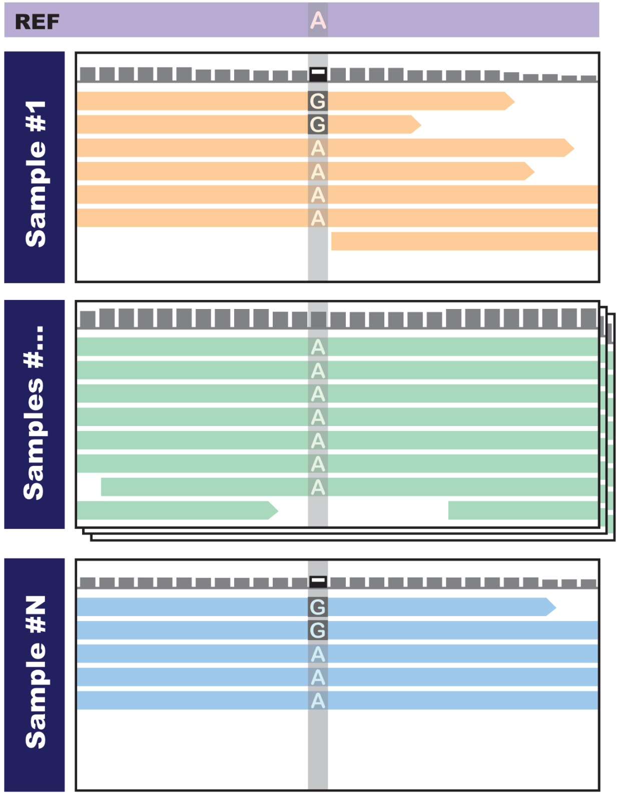
Individual callsets

Underpowered

Joint callset

Superpowered!

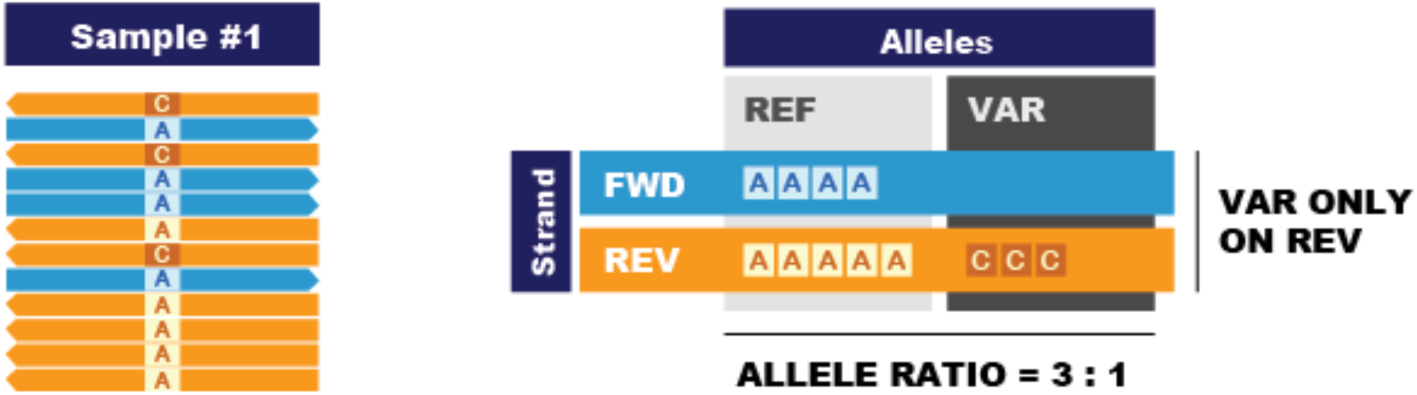
Discovery is empowered at difficult sites



- Sample #1 or Sample #N alone:
  - **weak evidence for variant**
  - **may miss calling the variant**

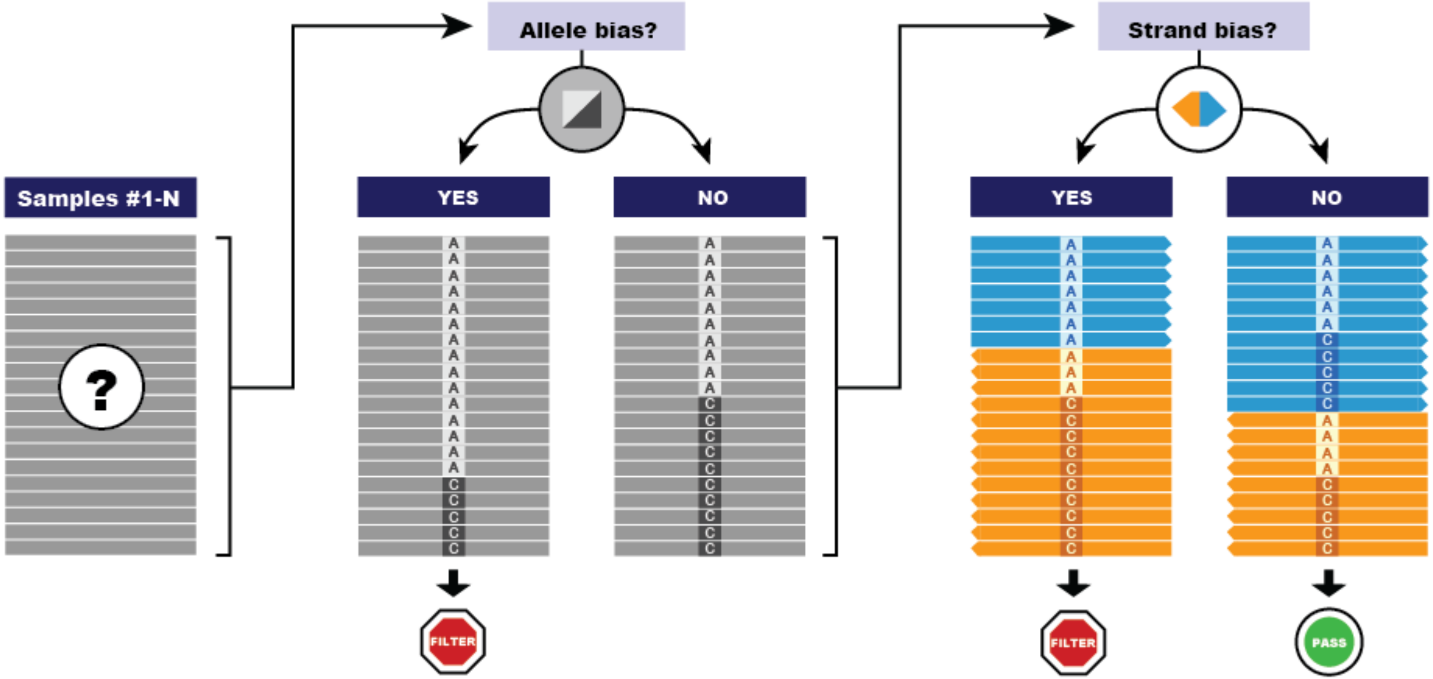
- Both samples seen together:
  - **unlikely to be artifact**
  - **call the variant more confidently**

Joint analysis helps resolve bias issues



Single sample showing strand and allelic biases – would you call it?

Joint analysis helps resolve bias issues

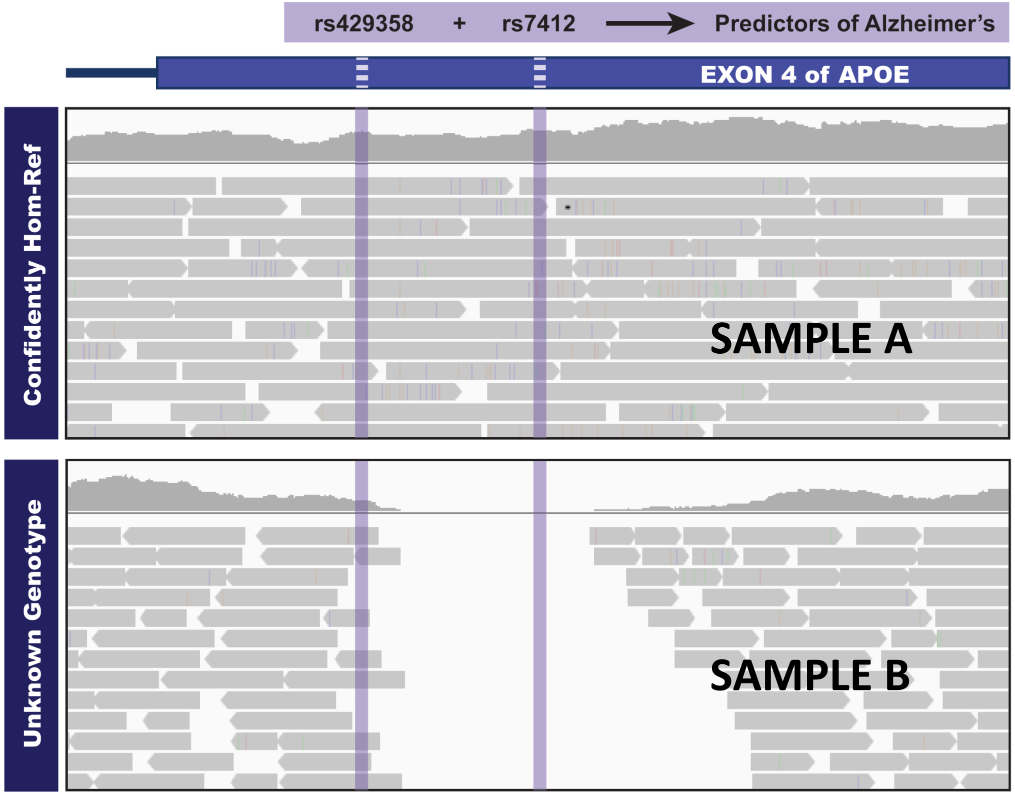


Decision process using evidence from multiple samples to filter out sites showing systematic biases

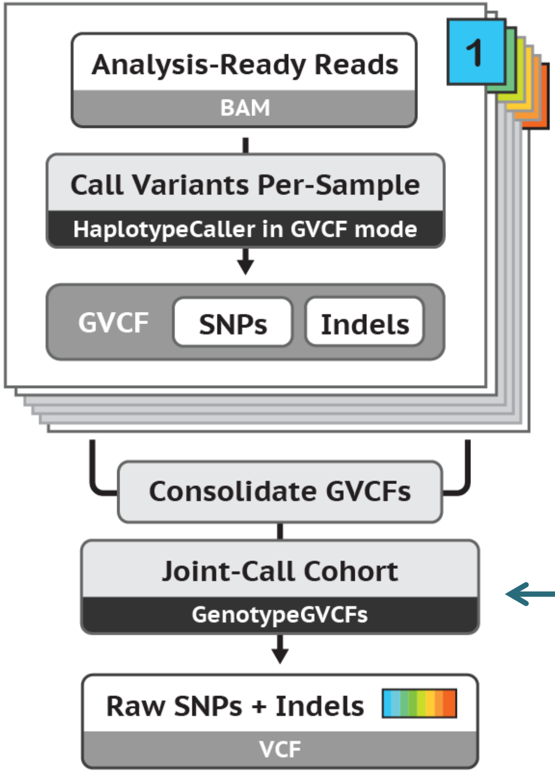
And we get full information at all sites of interest

- **Analyzed individually:**
  - No call for either sample
  - Very different reasons!

- **In joint analysis with other samples:**
  - Hom-ref call and no-call genotypes emitted



Joint calling implemented as a two-step process for scalability



1

Analysis-Ready Reads

BAM

Call Variants Per-Sample

HaplotypeCaller in GVCF mode

GVCF SNPs Indels

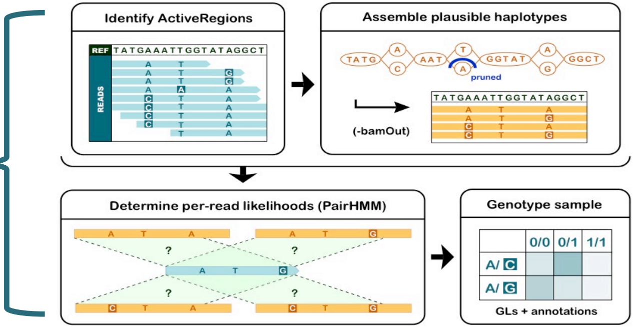
Consolidate GVCFs

Joint-Call Cohort

GenotypeGVCFs

Raw SNPs + Indels

VCF



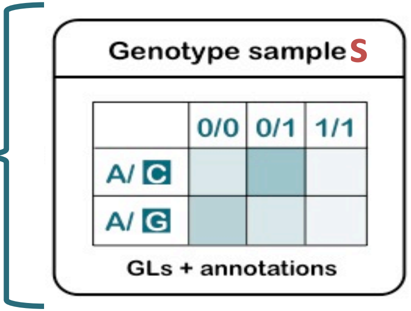
Identify ActiveRegions

Assemble plausible haplotypes

Determine per-read likelihoods (PairHMM)

Genotype sample

- Run HC in **GVCF mode** to emit GVCF



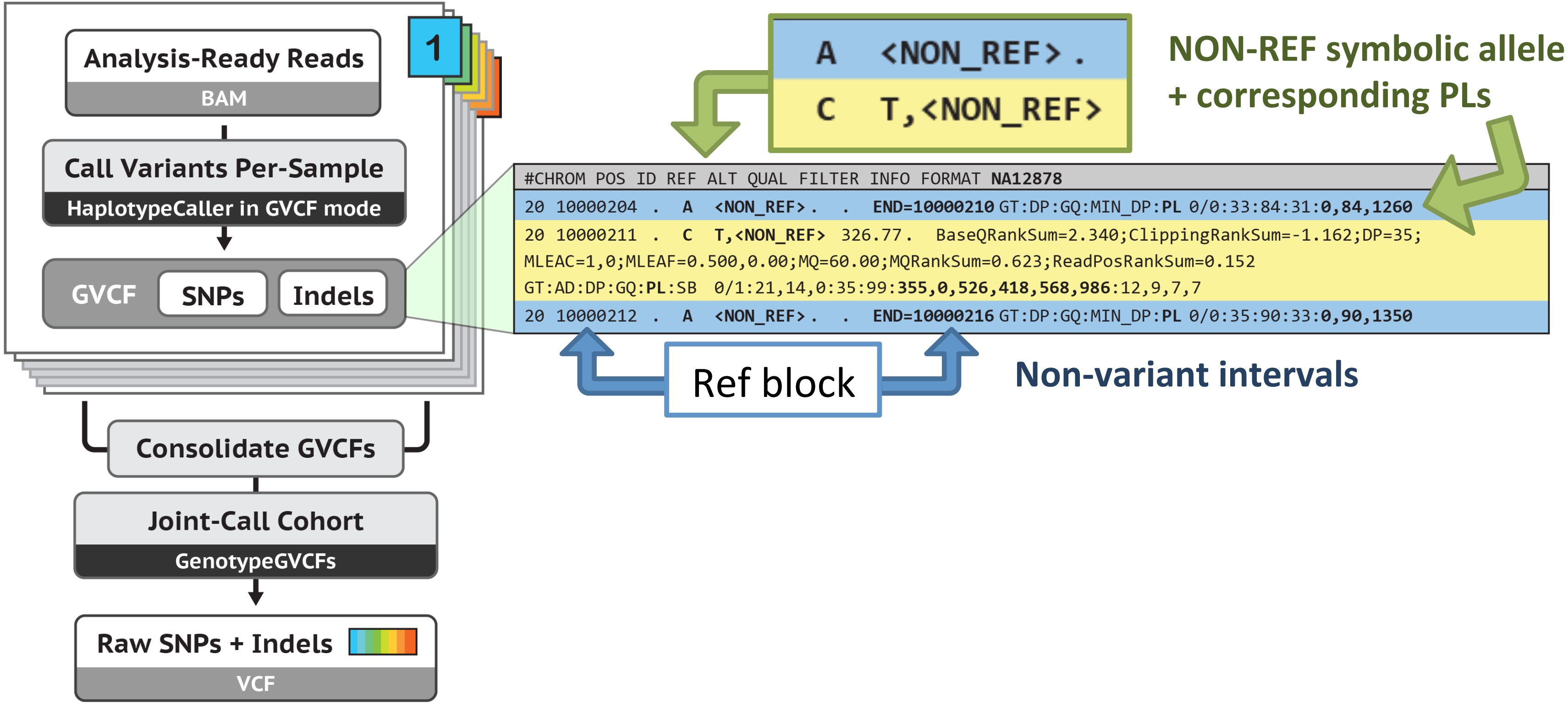
Genotype sampleS

	0/0	0/1	1/1
A/ C			
A/ G			

GLs + annotations

- Run GenotypeGVCFs to re-genotype samples with **multi-sample model**

GVCF intermediate contains reference confidence estimate

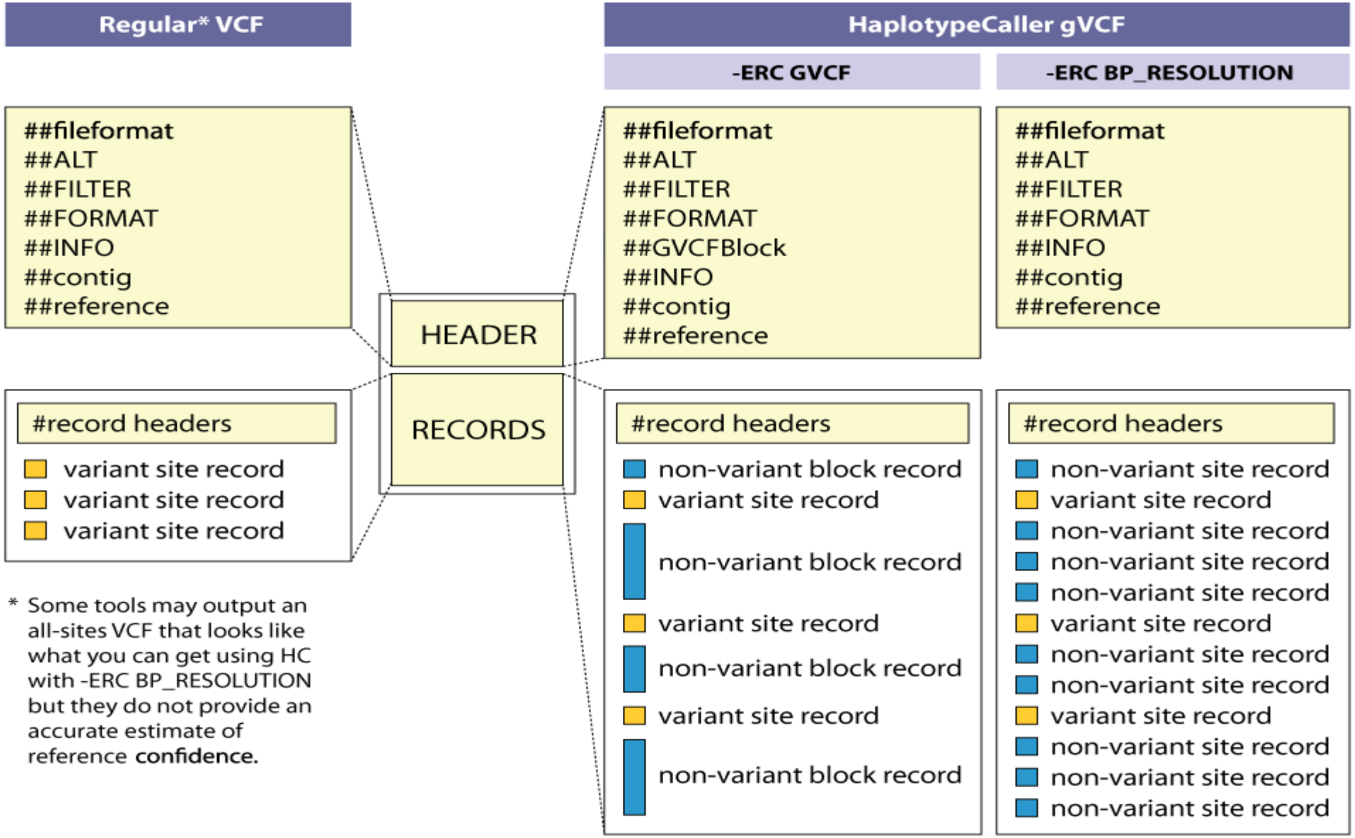


```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204 .  A  <NON_REF> .  .  END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211 .  C  T,<NON_REF> 326.77 .  BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
GT:AD:DP:GQ:PL:SB  0/1:21,14,0:35:99:355,0,526,418,568,986:12,9,7,7
20 10000212 .  A  <NON_REF> .  .  END=10000216 GT:DP:GQ:MIN_DP:PL 0/0:35:90:33:0,90,1350
```

NON-REF symbolic allele + corresponding PLs

Non-variant intervals

GVCFs are valid VCFs with extra information



Regular* VCF

##fileformat
##ALT
##FILTER
##FORMAT
##INFO
##contig
##reference

#record headers
🟨 variant site record
🟨 variant site record
🟨 variant site record

HEADER

RECORDS

HaplotypeCaller gVCF

-ERC GVCF

##fileformat
##ALT
##FILTER
##FORMAT
##GVCFBlock
##INFO
##contig
##reference

#record headers
🟦 non-variant block record
🟨 variant site record
🟦 non-variant block record
🟨 variant site record
🟦 non-variant block record
🟨 variant site record
🟦 non-variant block record

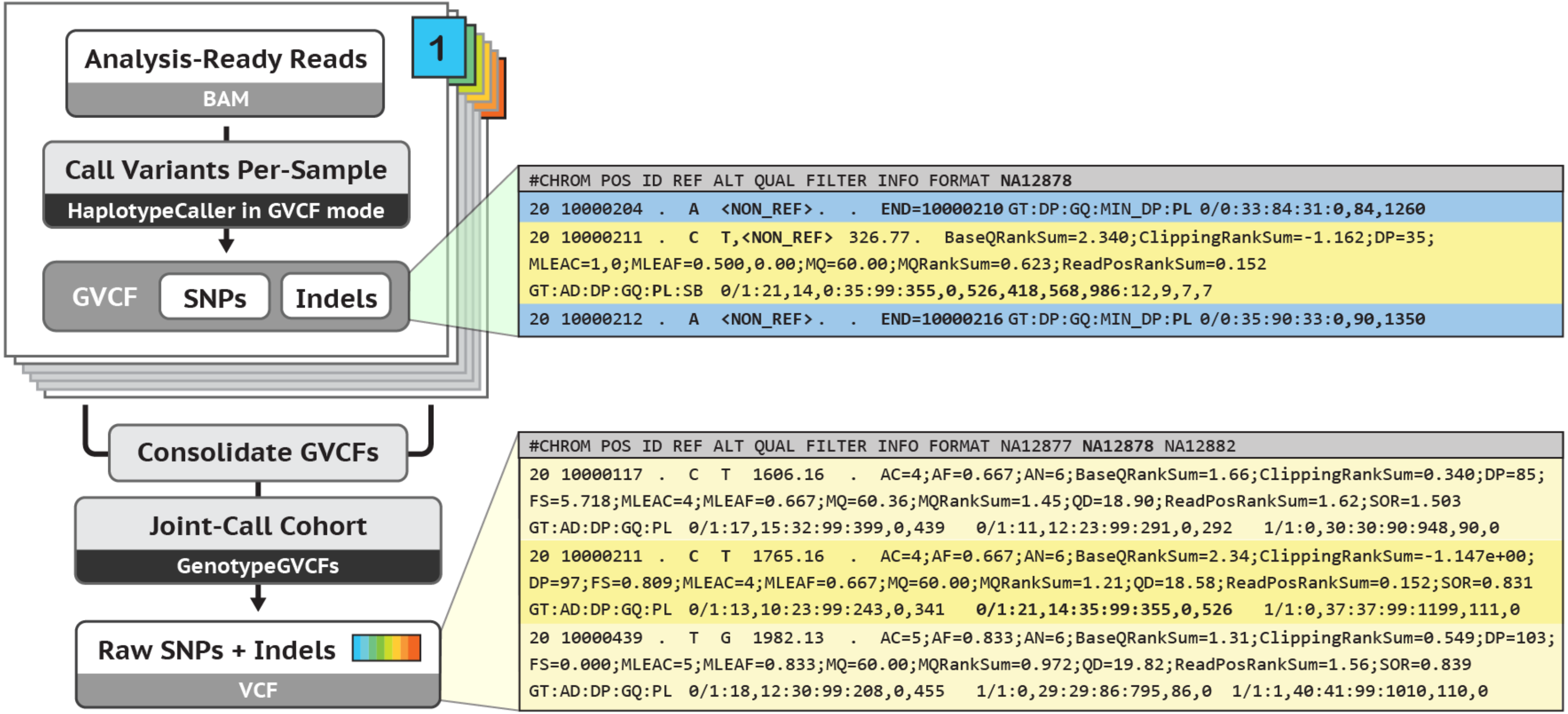
-ERC BP_RESOLUTION

##fileformat
##ALT
##FILTER
##FORMAT
##INFO
##contig
##reference

#record headers
🟦 non-variant site record
🟨 variant site record
🟦 non-variant site record
🟦 non-variant site record
🟦 non-variant site record
🟨 variant site record
🟦 non-variant site record
🟦 non-variant site record
🟨 variant site record
🟦 non-variant site record
🟦 non-variant site record
🟦 non-variant site record

* Some tools may output an all-sites VCF that looks like what you can get using HC with -ERC BP_RESOLUTION but they do not provide an accurate estimate of reference confidence.

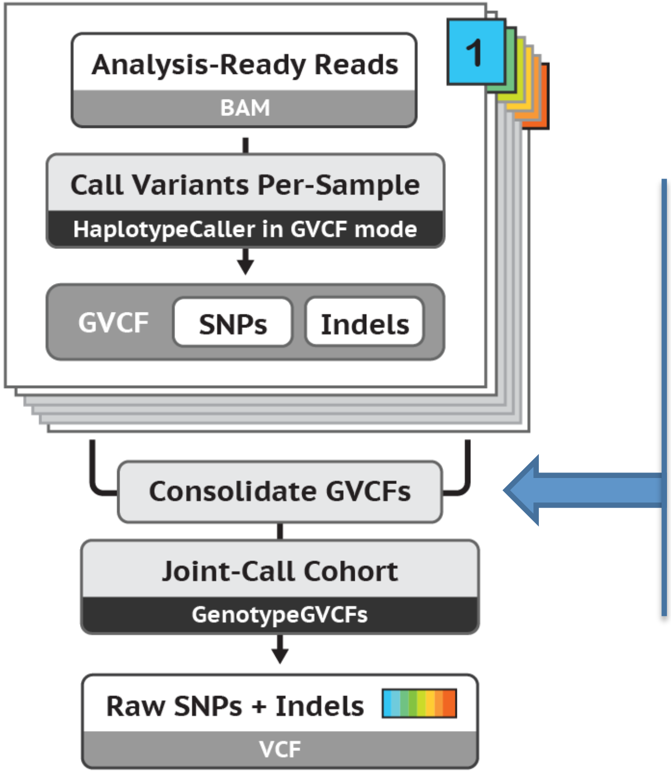
Joint calling produces final multi-sample VCF



```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204 .  A  <NON_REF> .  .  END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211 .  C  T,<NON_REF> 326.77 .  BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
GT:AD:DP:GQ:PL:SB  0/1:21,14,0:35:99:355,0,526,418,568,986:12,9,7,7
20 10000212 .  A  <NON_REF> .  .  END=10000216 GT:DP:GQ:MIN_DP:PL 0/0:35:90:33:0,90,1350
```

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12877 NA12878 NA12882
20 10000117 .  C  T  1606.16  .  AC=4;AF=0.667;AN=6;BaseQRankSum=1.66;ClippingRankSum=0.340;DP=85;
FS=5.718;MLEAC=4;MLEAF=0.667;MQ=60.36;MQRankSum=1.45;QD=18.90;ReadPosRankSum=1.62;SOR=1.503
GT:AD:DP:GQ:PL  0/1:17,15:32:99:399,0,439   0/1:11,12:23:99:291,0,292   1/1:0,30:30:90:948,90,0
20 10000211 .  C  T  1765.16  .  AC=4;AF=0.667;AN=6;BaseQRankSum=2.34;ClippingRankSum=-1.147e+00;
DP=97;FS=0.809;MLEAC=4;MLEAF=0.667;MQ=60.00;MQRankSum=1.21;QD=18.58;ReadPosRankSum=0.152;SOR=0.831
GT:AD:DP:GQ:PL  0/1:13,10:23:99:243,0,341   0/1:21,14:35:99:355,0,526   1/1:0,37:37:99:1199,111,0
20 10000439 .  T  G  1982.13  .  AC=5;AF=0.833;AN=6;BaseQRankSum=1.31;ClippingRankSum=0.549;DP=103;
FS=0.000;MLEAC=5;MLEAF=0.833;MQ=60.00;MQRankSum=0.972;QD=19.82;ReadPosRankSum=1.56;SOR=0.839
GT:AD:DP:GQ:PL  0/1:18,12:30:99:208,0,455   1/1:0,29:29:86:795,86,0  1/1:1,40:41:99:1010,110,0
```

Need to consolidate GVCFs before joint calling!



Necessary for efficient scaling

- **In GATK 3.x : CombineGVCFs**
  Hierarchical merge on batches of 200 samples max; outputs GVCF
- **In GATK 4.x : GenomicsDBImport**
  All samples processed in a single command; outputs datastore

Consolidating GVCFs

With CombineGVCFs:

```
gatk CombineGVCFs \
   -R reference.fasta \
   -V sample1.g.vcf \
   -V sample2.g.vcf \
   -O combined.g.vcf
```

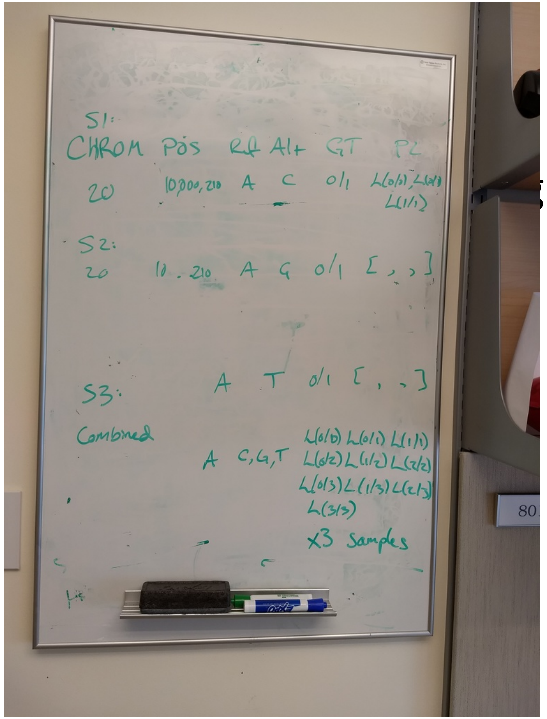
CombineGVCFs does not scale well

With GenomicsDBImport:

```
gatk GenomicsDBImport \
   -R reference.fasta \
   -V sample1.g.vcf \
   -V sample2.g.vcf \
   -L chr20 \
   --genomicsdb-workspace-path gvcfs_db
```

GenomicsDBImport scales well but must be run on a single interval at a time

GenomicsDB is a datastore / better representation than CombineGVCFs



S1:

CHROM	Pos	Ref	Alt	GT	PL

20	10,000,210	A	C	0/1	L(0/0), L(0/1)
					L(1/1)

S2:

20	10 . 210	A	G	0/1	[, ,]

S3:

		A	T	0/1	[, ,]

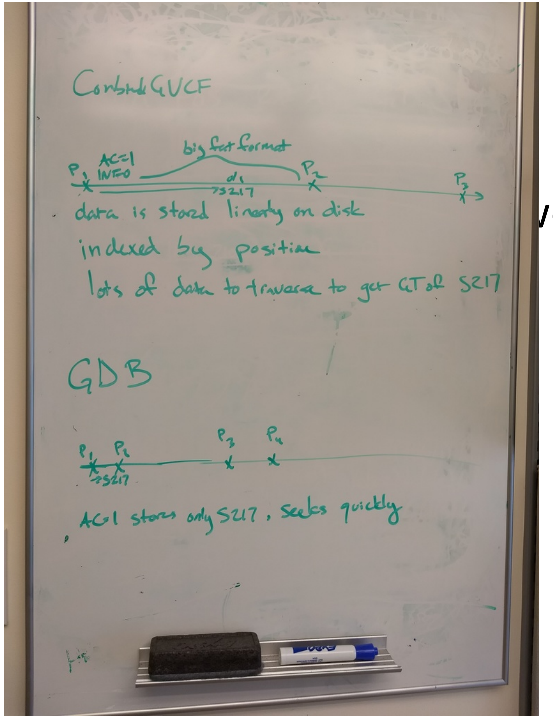
Combined

		A	C,G,T		L(0/0) L(0/1) L(1/1)
					L(0/2) L(1/2) L(2/2)
					L(0/3) L(1/3) L(2/3)
					L(3/3)

x3 samples

Combined GVCFs have big PL arrays at multiallelic sites

goal: add some deta … vorks



CombinedGVCF

data is stored linearly on disk

indexed by position

lots of data to traverse to get GT of S217

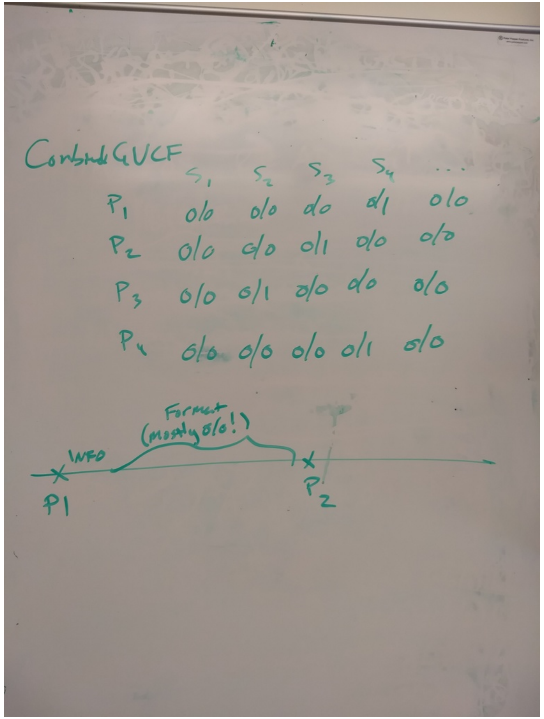
GDB

AC=1 stores only S217, seeks quickly

Query by sample in GDB is fast

Storage size comparisons

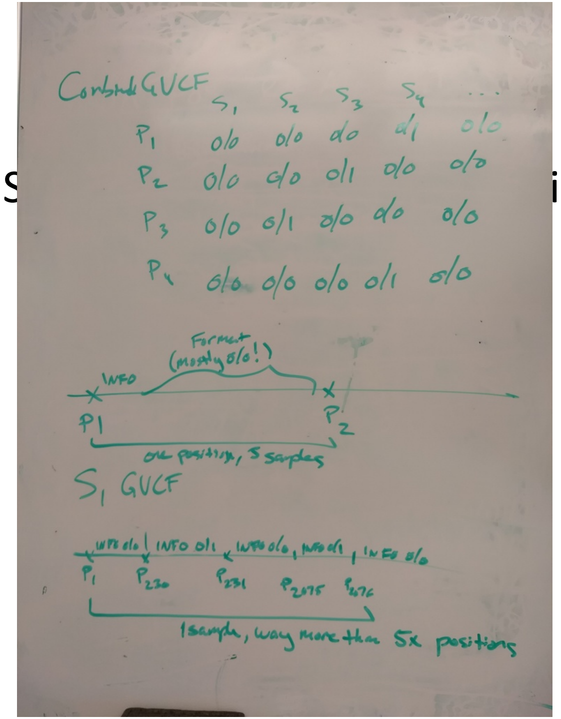
S il ab ks



Combined GVCF

	S_1	S_2	S_3	S_4	...
P_1	0/0	0/0	0/0	0/1	0/0
P_2	0/0	0/0	0/1	0/0	0/0
P_3	0/0	0/1	0/0	0/0	0/0
P_4	0/0	0/0	0/0	0/1	0/0

Data is sparse (mostly repeated values from ref block) but gets bigger when we combine

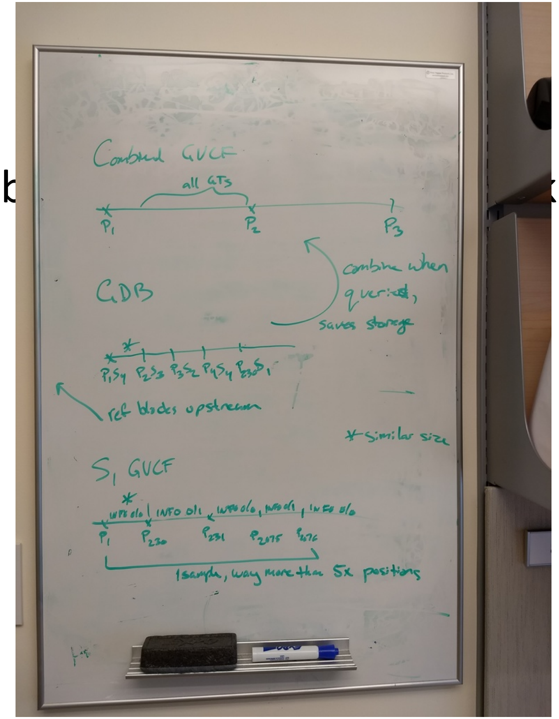


Combined GVCF

	S_1	S_2	S_3	S_4	...
P_1	0/0	0/0	0/0	0/1	0/0
P_2	0/0	0/0	0/1	0/0	0/0
P_3	0/0	0/1	0/0	0/0	0/0
P_4	0/0	0/0	0/0	0/1	0/0

INFO
Format (mostly 0/0!)
P1
P2
one position, 5 samples
S1 GVCF
INFO 0/0 | INFO 0/1 , INFO 0/0 , INFO 0/1 , INFO 0/0
P1
P230
P231
P2075
P2076
1 sample, way more than 5x positions

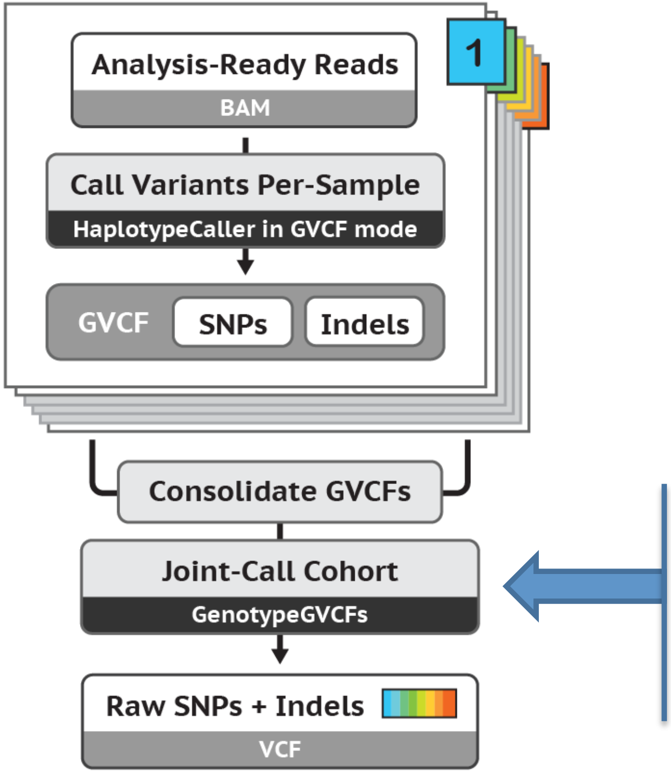
Comparison of combined storage vs single GVCF



Combined GVCF
all GTs
P1
P2
P3
combine when queried, saves storage
GDB
*
P1S4 P2S3 P3S2 P4S4 P230S1
ref blocks upstream
*similar size
S1 GVCF
*
INFO 0/0 | INFO 0/1 , INFO 0/0 , INFO 0/1 , INFO 0/0
P1
P230
P231
P2075
P2076
1 sample, way more than 5x positions

Theoretical size comparison

Joint calling with GenotypeGVCFs



- **GenotypeGVCFs** can take either a **single GVCF file** (can be a merged multi-sample GVCF from CombineGVCFs) or a **GenomicsDB datastore**
- No more multiple inputs! (unlike GATK3)

Running GenotypeGVCFs

On a single- or multi-sample GVCF:

```
gatk GenotypeGVCFs \
   -R reference.fasta \
   -V variants.g.vcf \
   -O final_variants.vcf
```

On a GenomicsDB workspace:

```
gatk GenotypeGVCFs \
   -R reference.fasta \
   -V gendb://gvcfs_db \
   -O final_variants.vcf
```

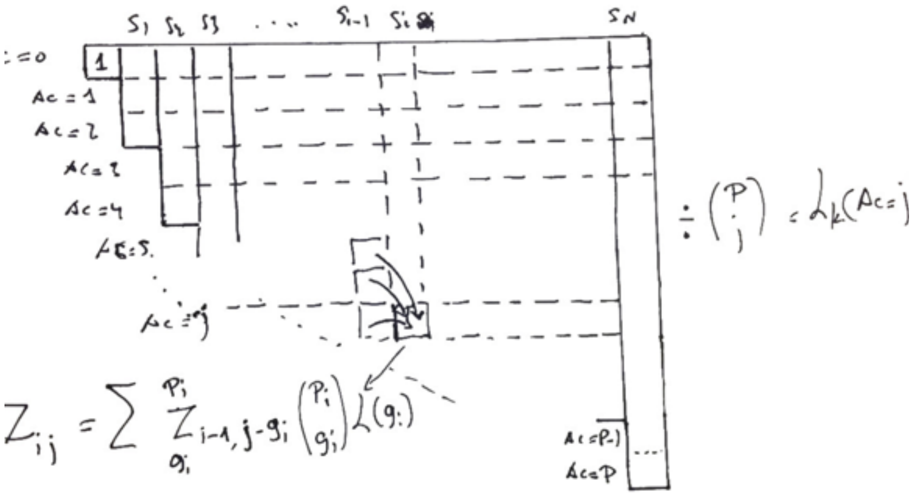
GenotypeGVCFs cannot take multiple inputs (unlike the GATK3 version)

Multi-sample QUAL calculation

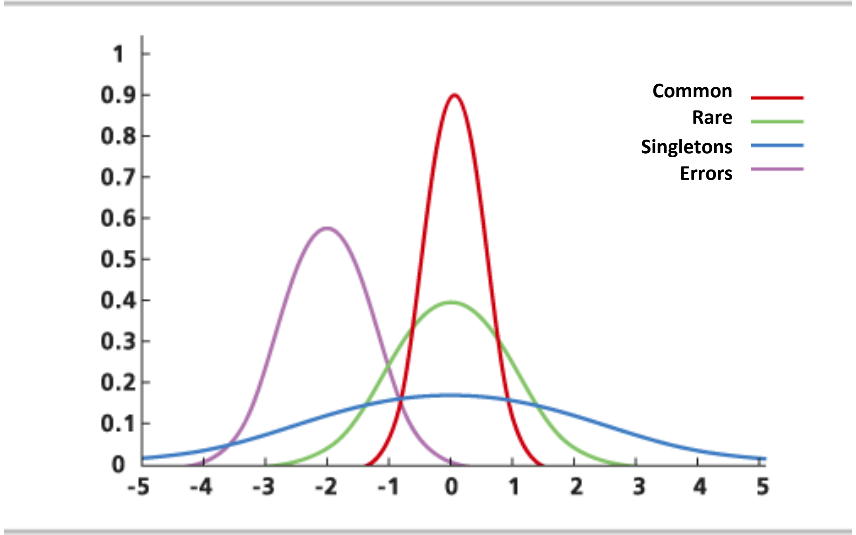
- Uses human SNP heterozygosity 1/1000 bases = Phred scale Q30
  (can be modified)

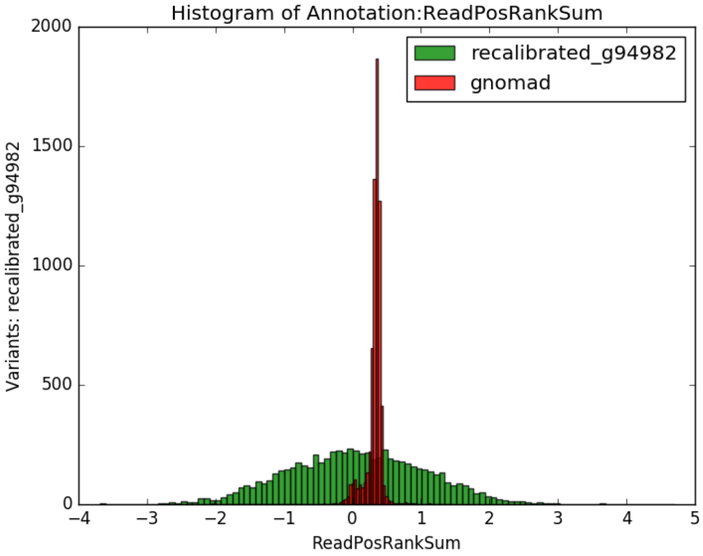
- QUAL > 30 means a variant is more likely than this base level

- **Heuristic for the QUAL score:** *sum(PL[0])-30* across samples


S_1 S_2 S_3 . . . . S_{i-1} S_i S_N
i=0
1
AC=1
AC=2
AC=3
AC=4
AC=5
AC=j
AC=P-1
AC=P
$\div \binom{P}{j} = L_k(AC=j)$
$Z_{ij} = \sum_{g_i}^{P_i} Z_{i-1,j-g_i} \binom{P_i}{g_i} L(g_i)$

Combination of annotations stabilizes distributions





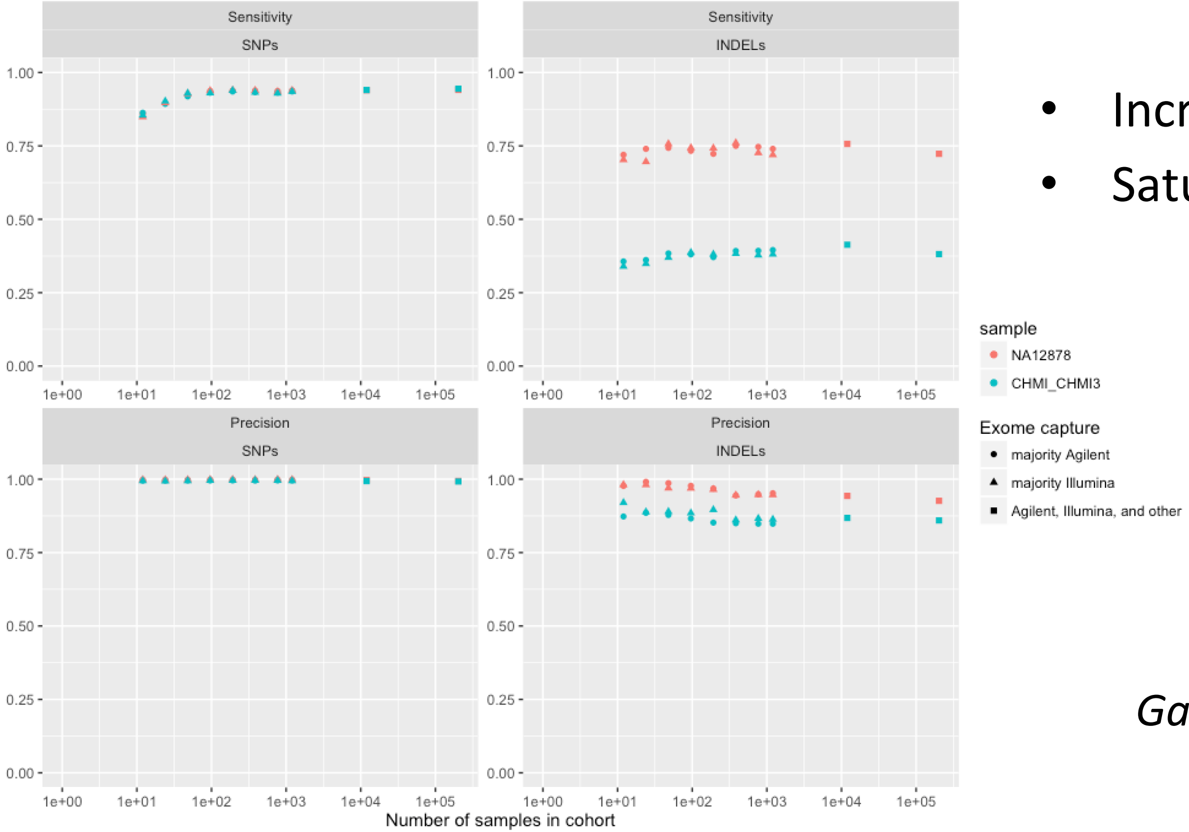
Theoretical distribution of annotation values

- Distinct for TP vs FP
- Tighter for common variants

Annotation values for same variants called in :

- Single-sample run (recalibrated_g94982, green)
- Multi-sample run (gnomad, red)

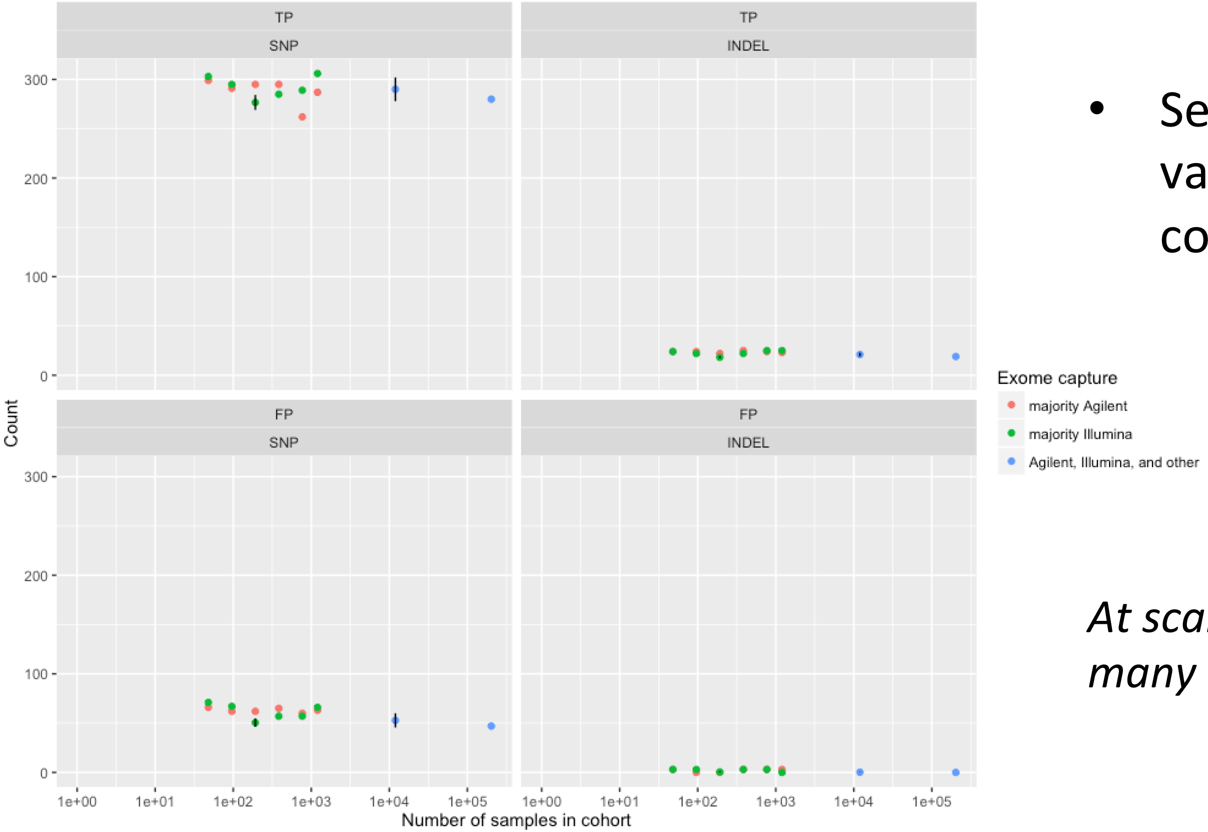
Use of a larger cohort increases sensitivity



- Increased sensitivity
- Saturation ~600 samples

Gauthier et al., 2016 (ASHG)

No loss of accuracy on singletons



- Set of singleton truth variants compared across cohort sizes

At scale of largest cohort, many are no longer singletons

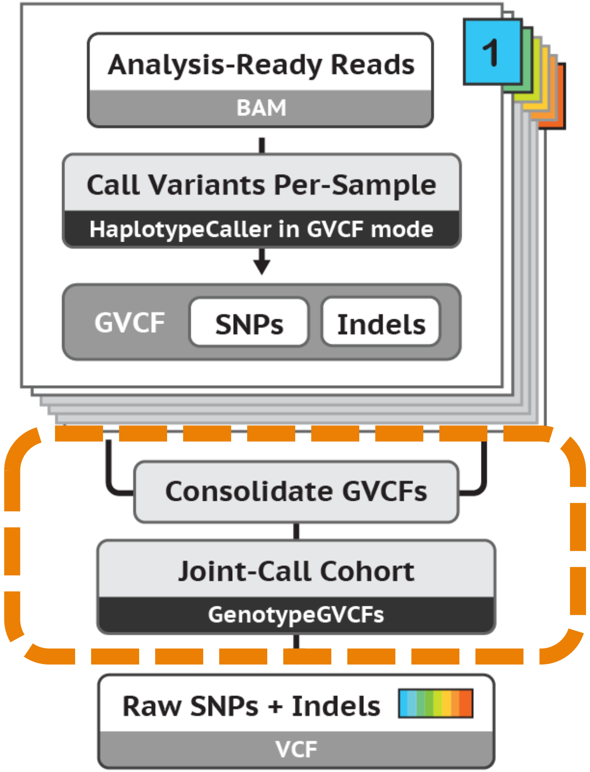
Joint calling scales massively better + faster with GenomicsDB

GATK 3

With CombineGVCFs

+ GenotypeGVCFs

- **Subset of samples**
- **3,000 WGS**
- **6 weeks**



GATK 4

With GenomicsDB

+ GenotypeGVCFs

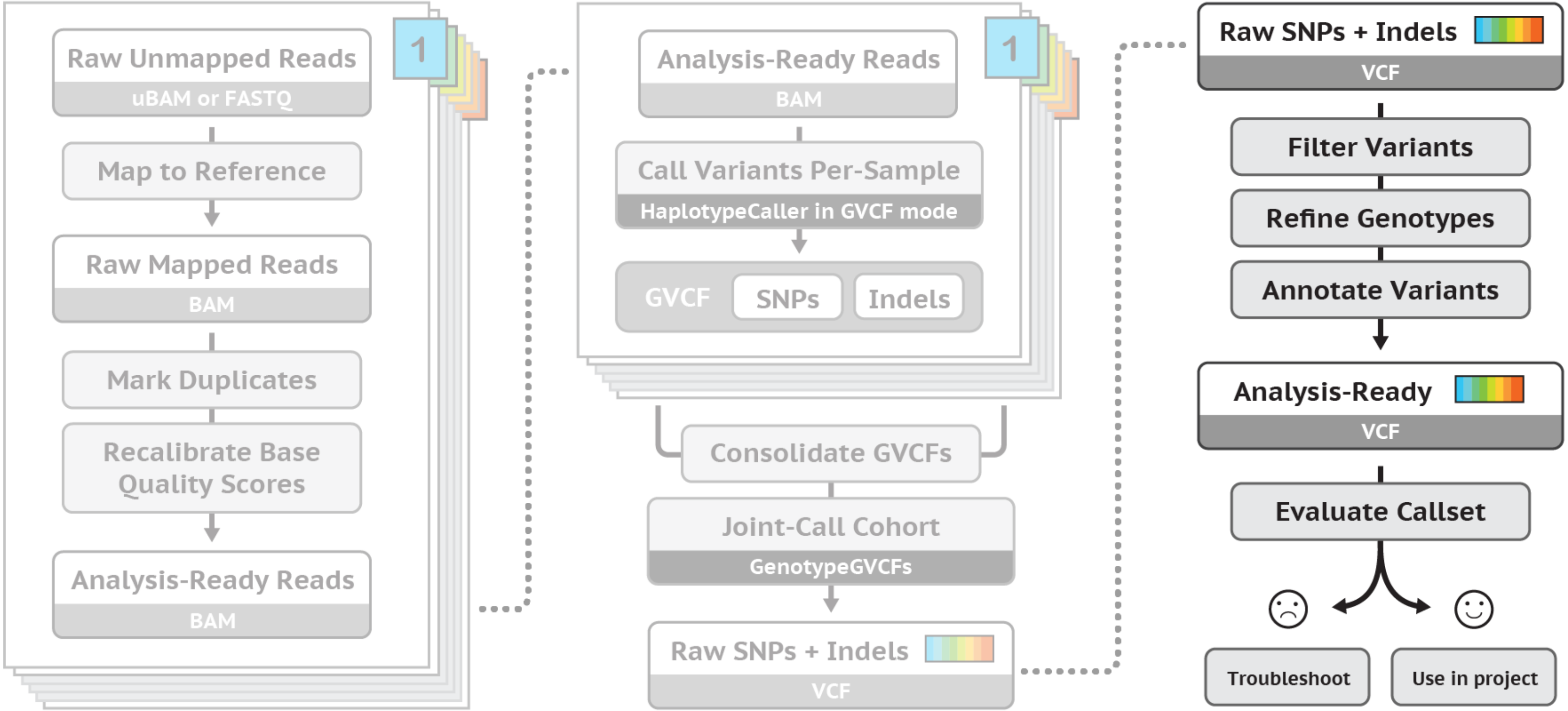
- **Full gnomAD v1**
- **15,000 WGS**
- **2 weeks**

→ 5x increase in scale

→ processed 3x faster

NOW DOING 76,000 WGS!

Next steps: filtering and other callset refinements



Raw Unmapped Reads
uBAM or FASTQ
1
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
Analysis-Ready Reads
BAM
1
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project