GATK Best Practices for Variant Discovery

Introduction to Germline Variant Discovery

Key considerations and workflow logic

BROAD INSTITUTE

http://software.broadinstitute.org/gatk/

gatk

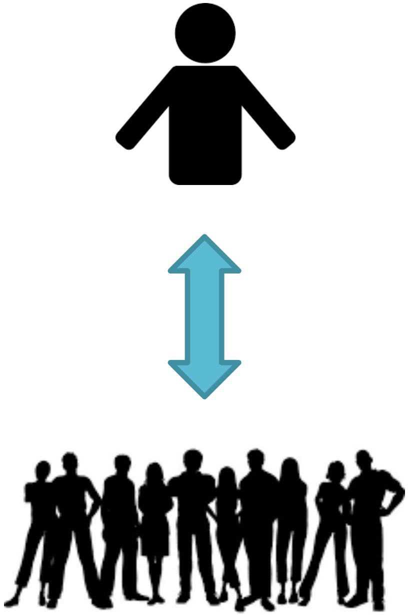
Discover **variants** relative to a reference genome

- Genetic changes in individuals **relative to a reference genome**
  - Germline (inherited)
  - Somatic (cancer)

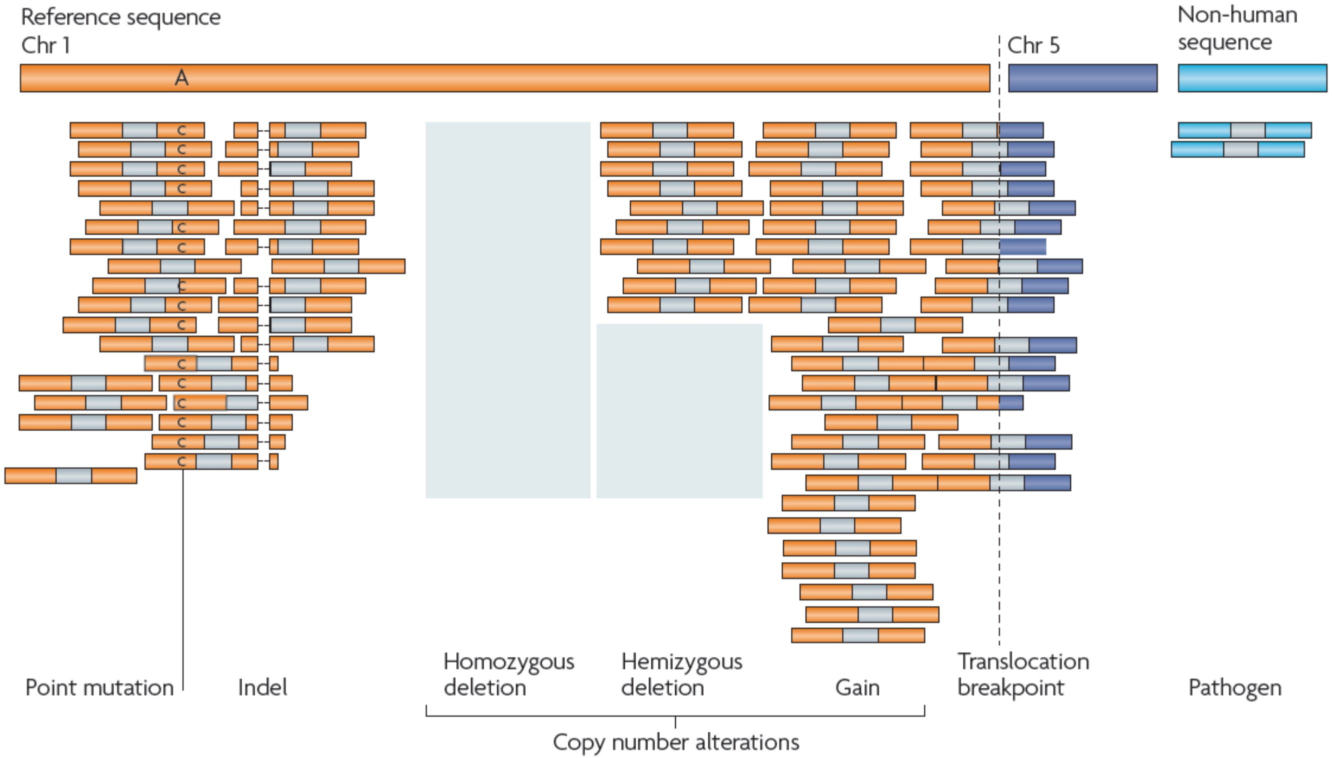
- **Reference genome** = a standardized genomic sequence

- Human genome reference sequence
  - Previous standard: hg19 / b37
  - New standard: hg38

- Other organisms
  - Many have a fully assembled reference available
  - Many still do not -> must make one



Different types of variants

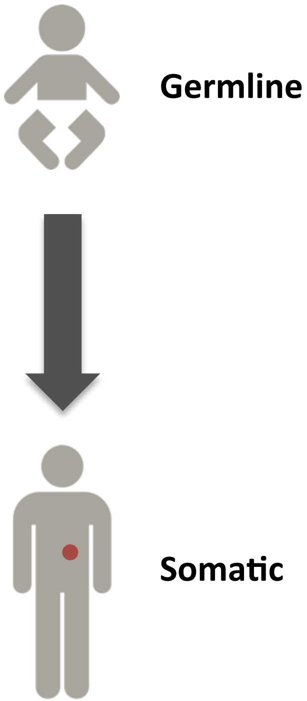


SNP/SNV

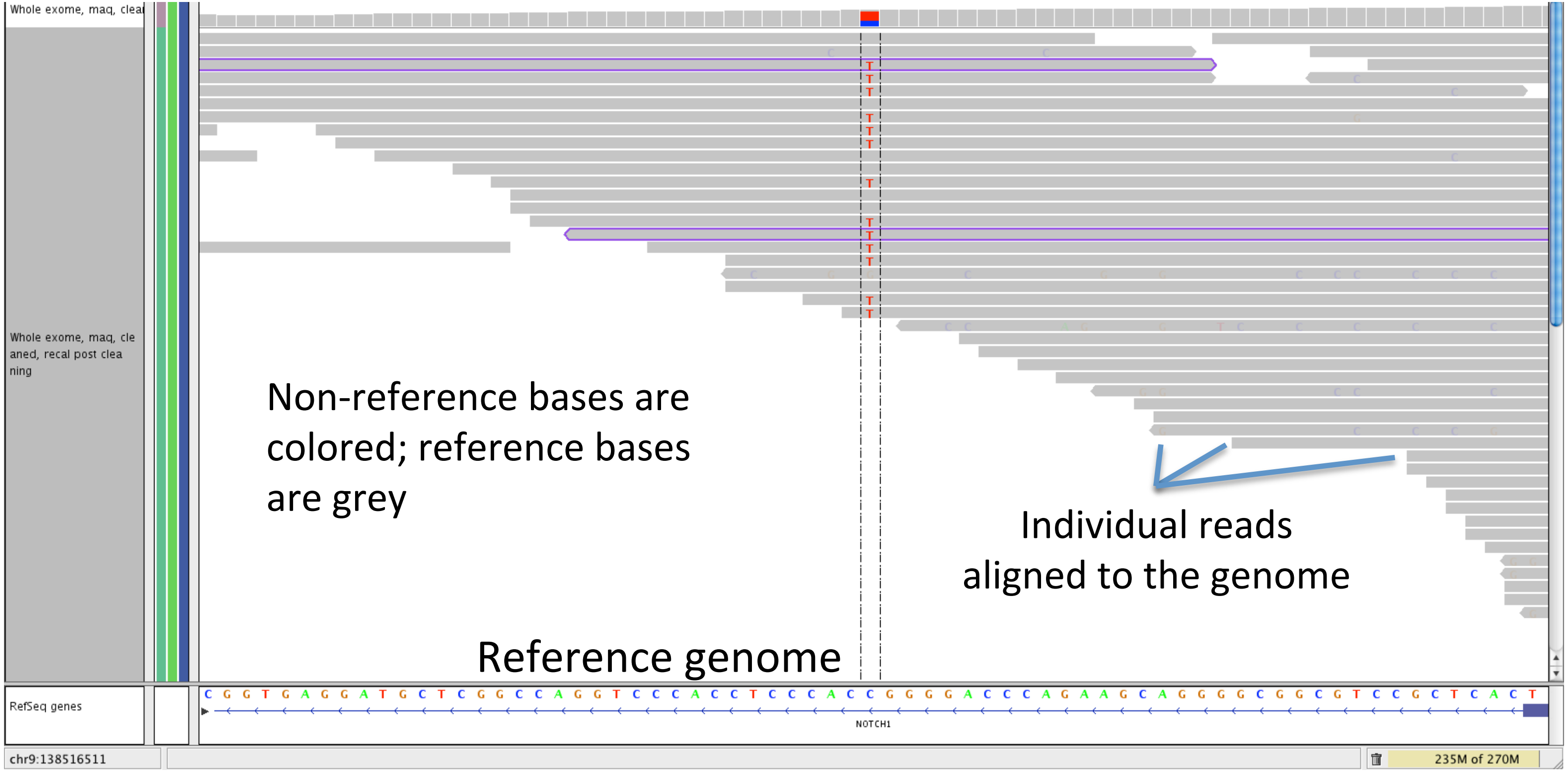
Indel

CNV/CNA

SV



This is what a good SNP looks like in a genome browser


Whole exome, maq, clea
T
Whole exome, maq, cleaned, recal post cleaning
RefSeq genes
C G G T G A G G A T G C T C G G C C A G G T C C C A C C T C C C A C C G G G G A C C C A G A A G C A G G G G C G G C G T C C G C T C A C T
NOTCH1
chr9:138516511
235M of 270M

Depth of coverage

Probable C/T SNP

Non-reference bases are colored; reference bases are grey

Individual reads aligned to the genome

Reference genome

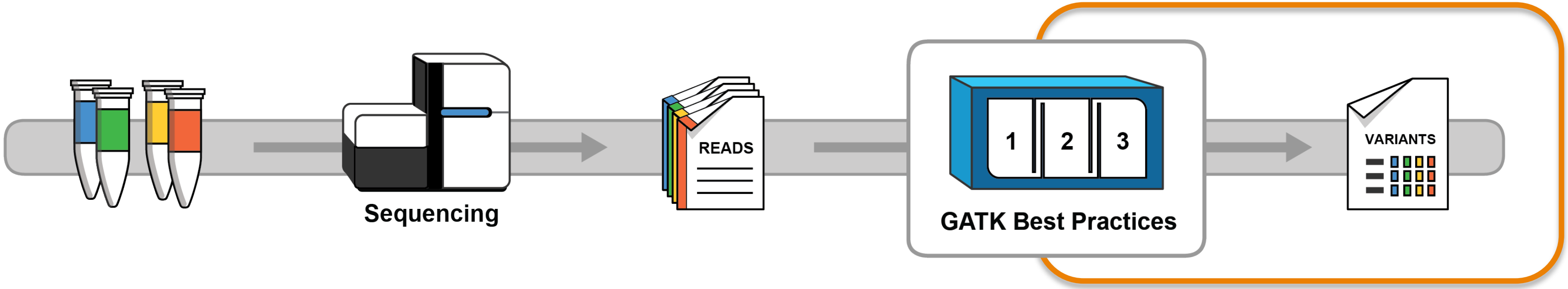
Short variants are reported in VCF: Variant Call Format

```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">
##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">
#CHROM POS    ID        REF ALT  QUAL  FILTER INFO          FORMAT    NA00001   NA00002   NA00003
20     14370   rs6054257 G   A    29    PASS   DP=14;AF=0.5  GT:GQ:DP  0/0:48:1  1/0:48:8  1/1:43:5
20     1230237 .         T   .    47    PASS   DP=13         GT:GQ:DP  0/0:54:7  0/0:48:4  0/0:61:2
20     1234567 .         GT  G    50    PASS   DP=9          GT:GQ:DP  0/1:35:4  0/2:17:2  1/1:40:3
```

Header

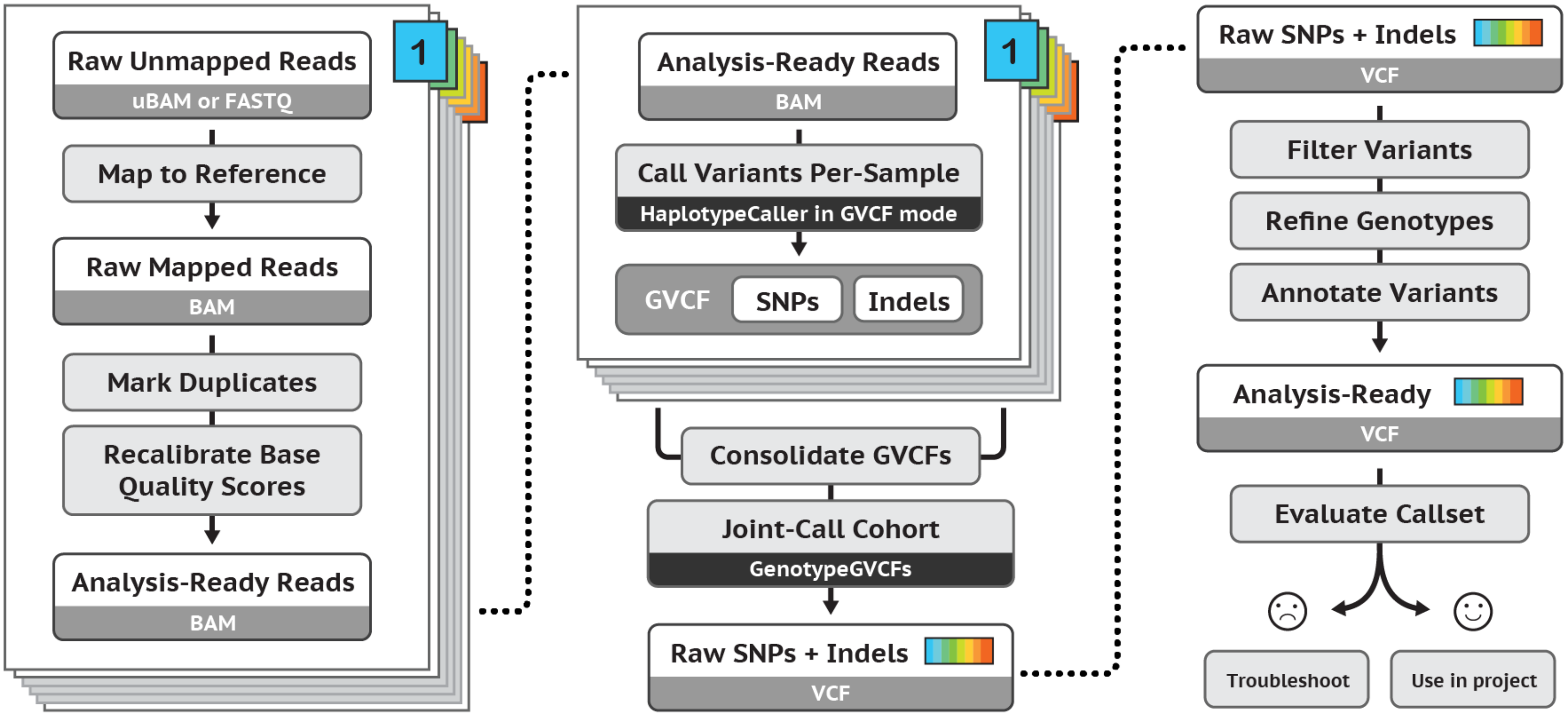
Records

Format specification in
https://samtools.github.io/hts-specs/VCFv4.2.pdf

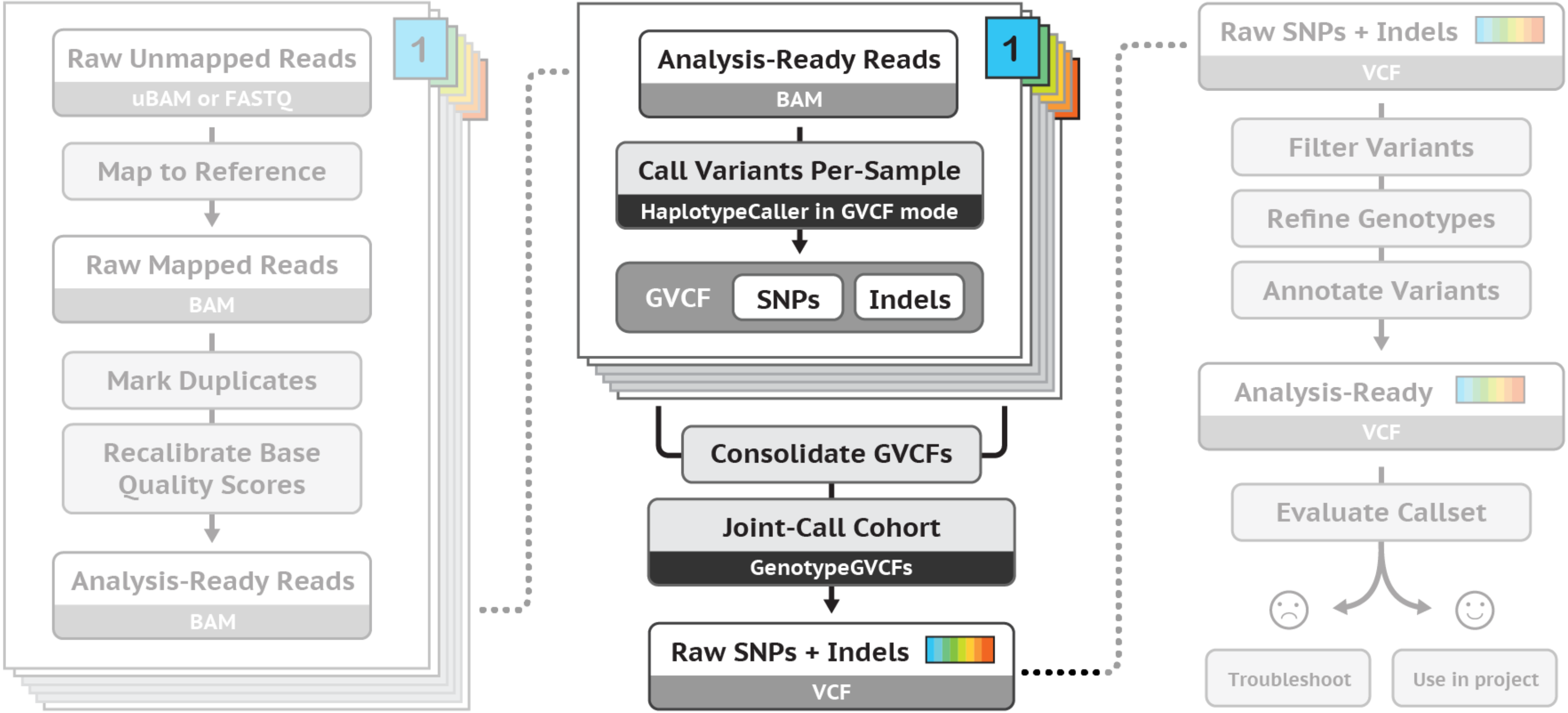


THE WORKFLOW

Best Practices for Germline SNP & INDEL Discovery

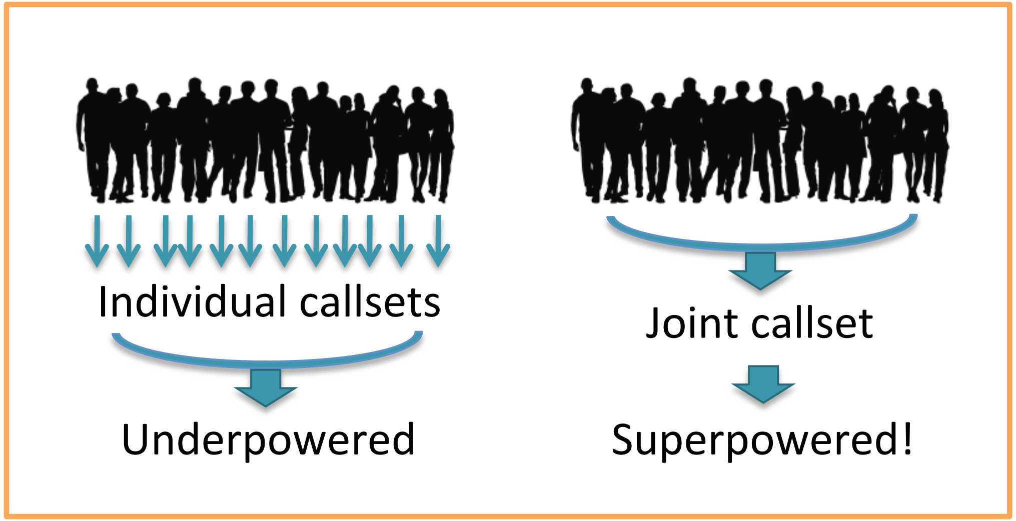


Central concept: **joint calling**

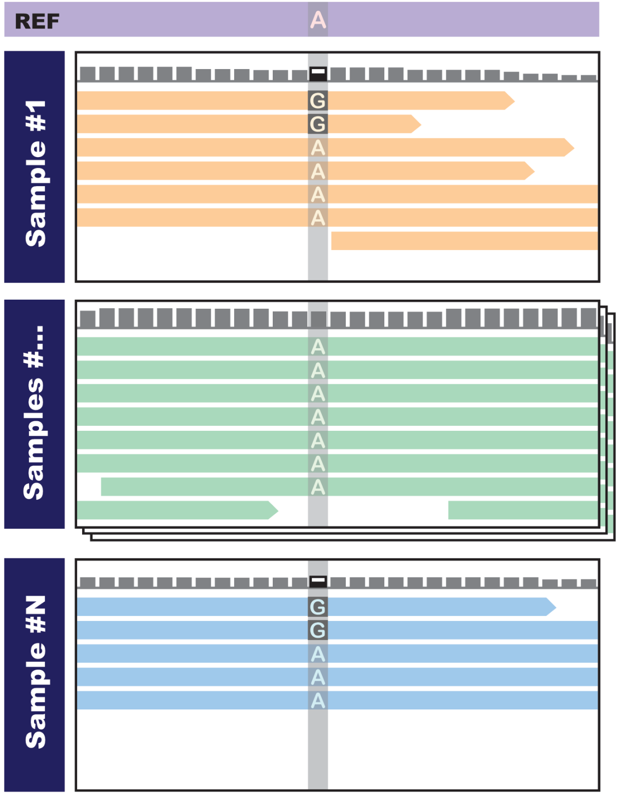


Joint analysis empowers discovery

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



Discovery is empowered at difficult sites



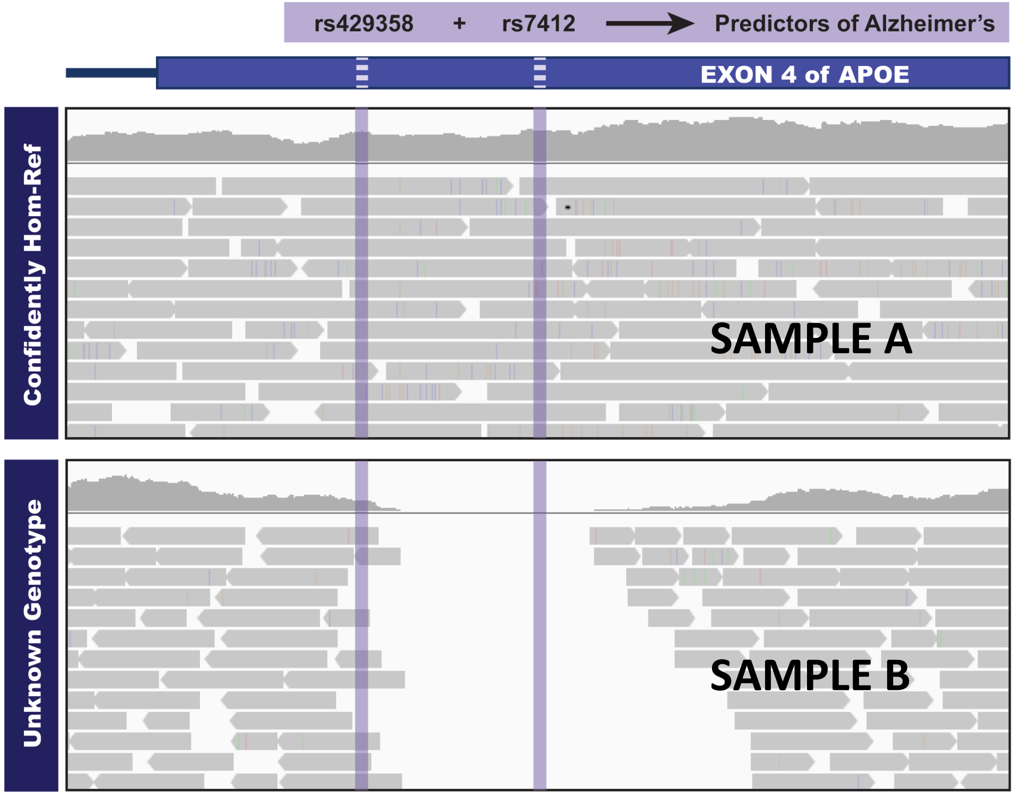
- Sample #1 or Sample #N alone:
  - **weak evidence for variant**
  - **may miss calling the variant**

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

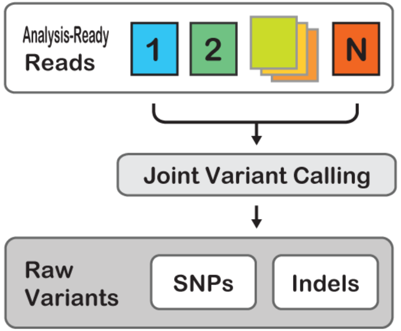
And we get full information at all sites of interest

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

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

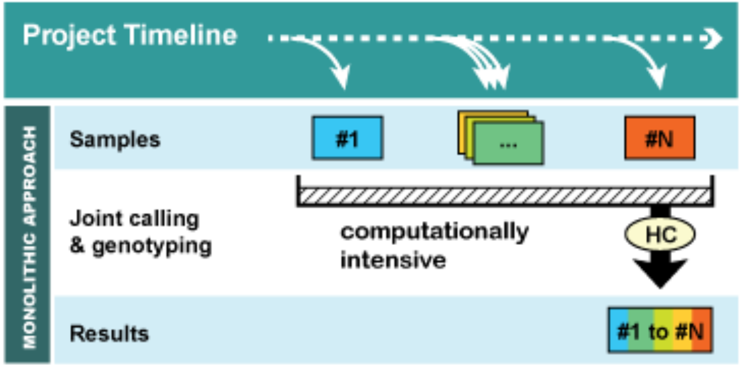


Traditional **multi-sample calling** approach : very inefficient



Compute requirements scale very badly with number of samples!!!

It gives us the right answers, but…

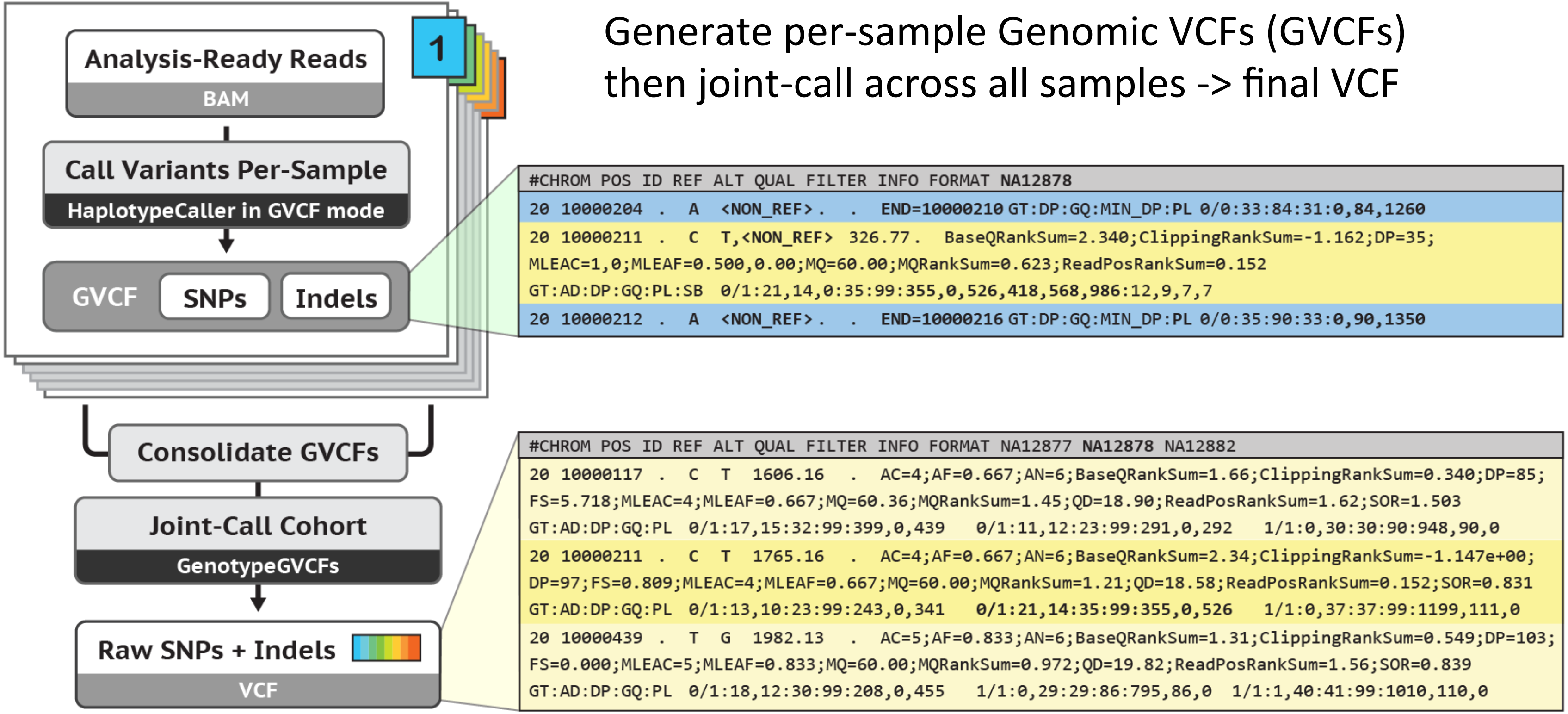


Want to add new samples?

Got to re-run pipeline from scratch! The N+1 problem!

Solution: the GVCF-based joint calling workflow

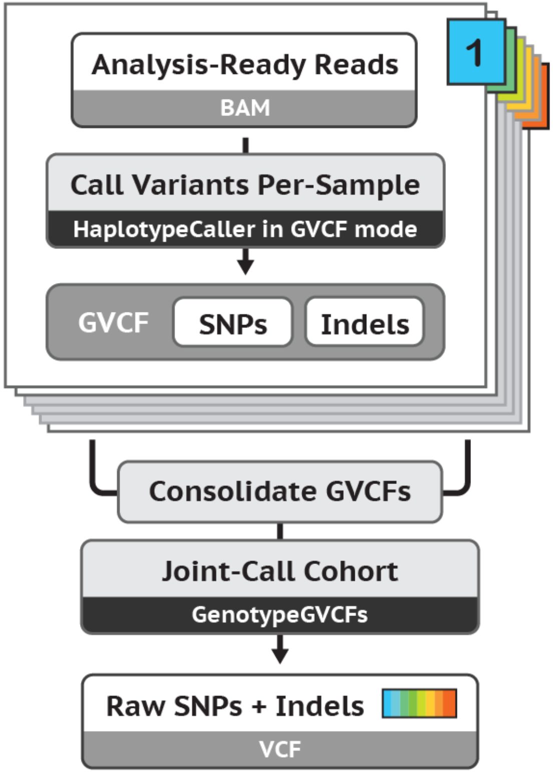
Generate per-sample Genomic VCFs (GVCFs)
then joint-call across all samples -> final VCF

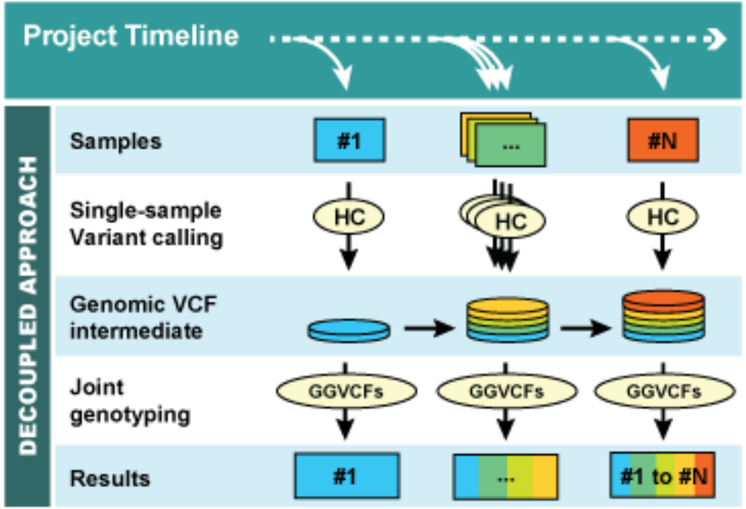


```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204  .   A  <NON_REF> .    .   END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211  .   C  T,<NON_REF> 326.77 .  BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
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
```

Same results as old approach - but scalable and incremental!

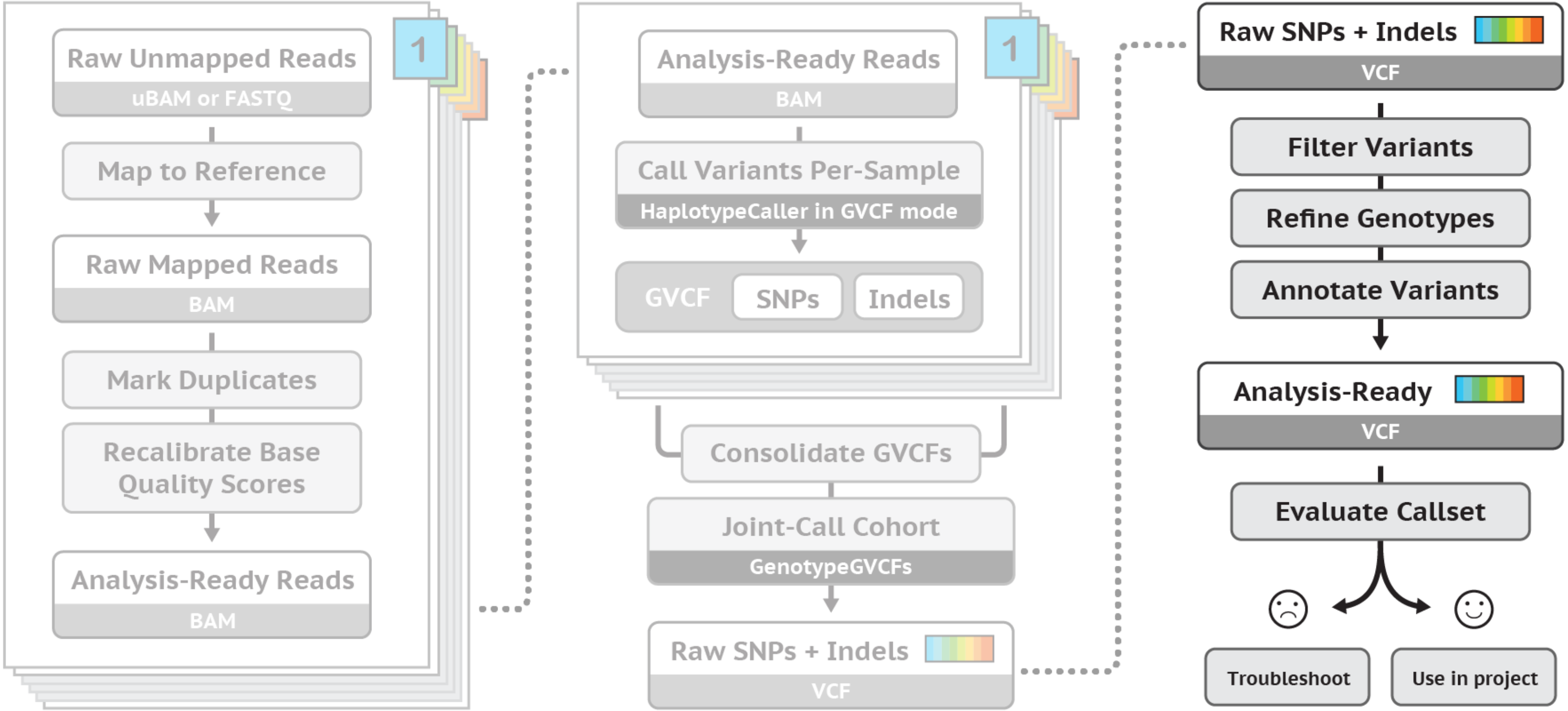




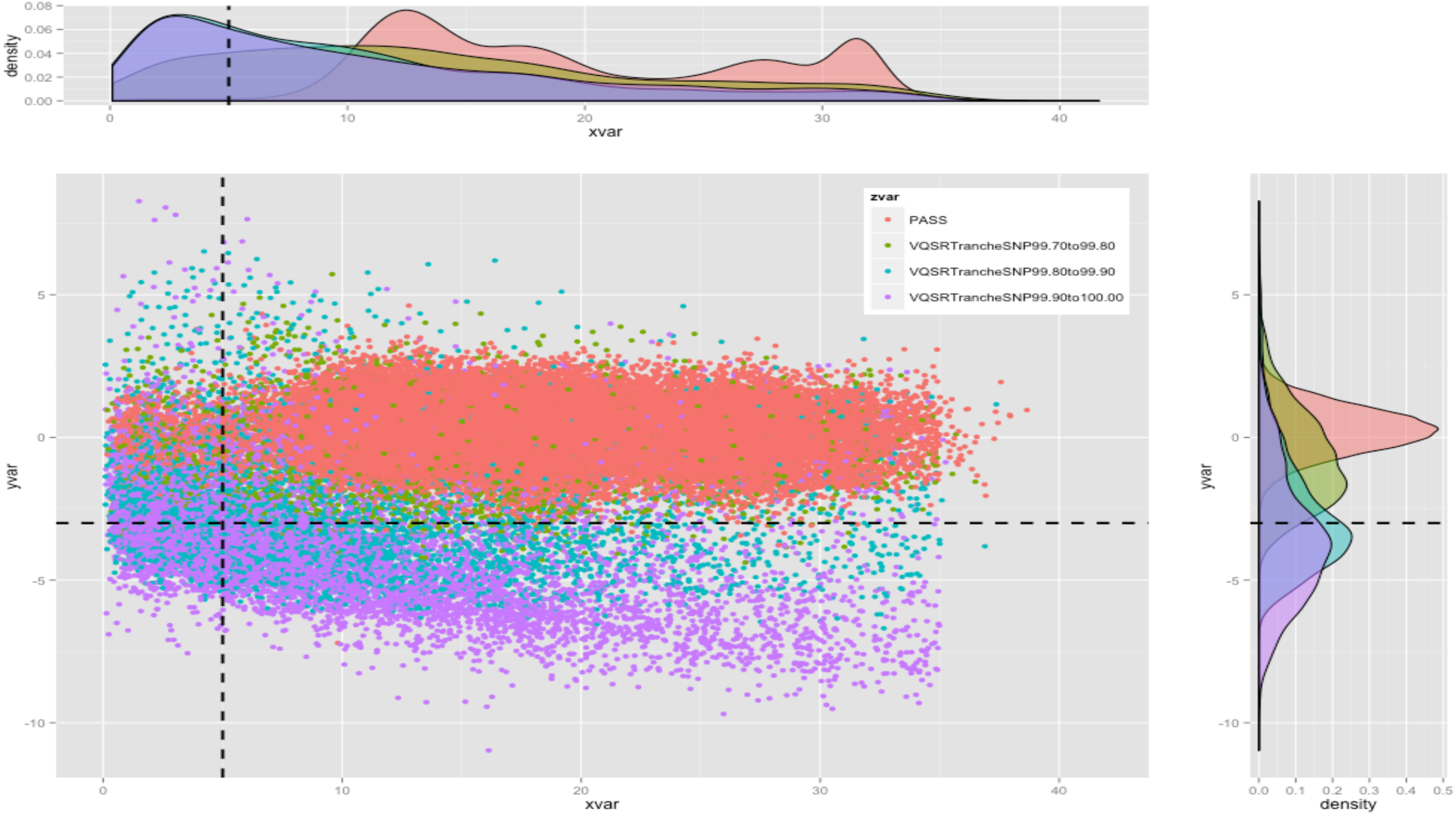
Scales linearly with number of samples!

Want to add a new sample? Make a GVCF for that sample then re-call the cohort at will!

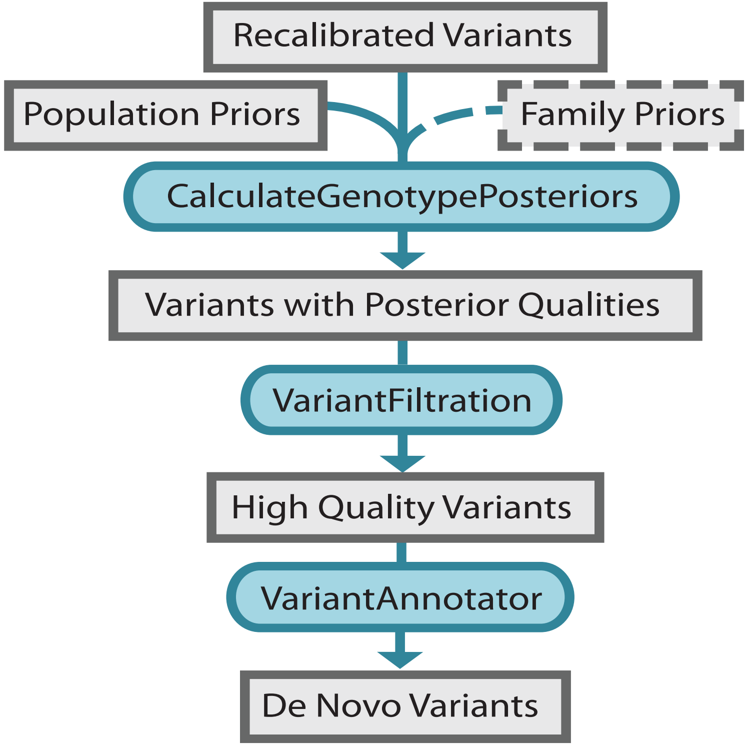
Further refinements: filtering and more

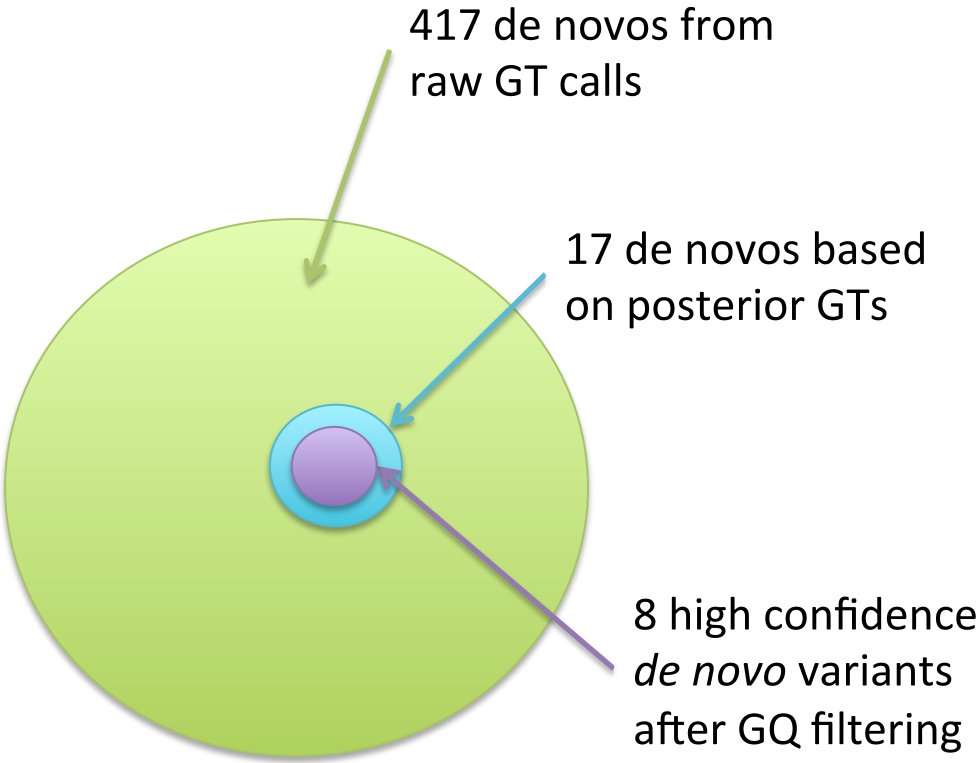


Variant filtering reduces false positives

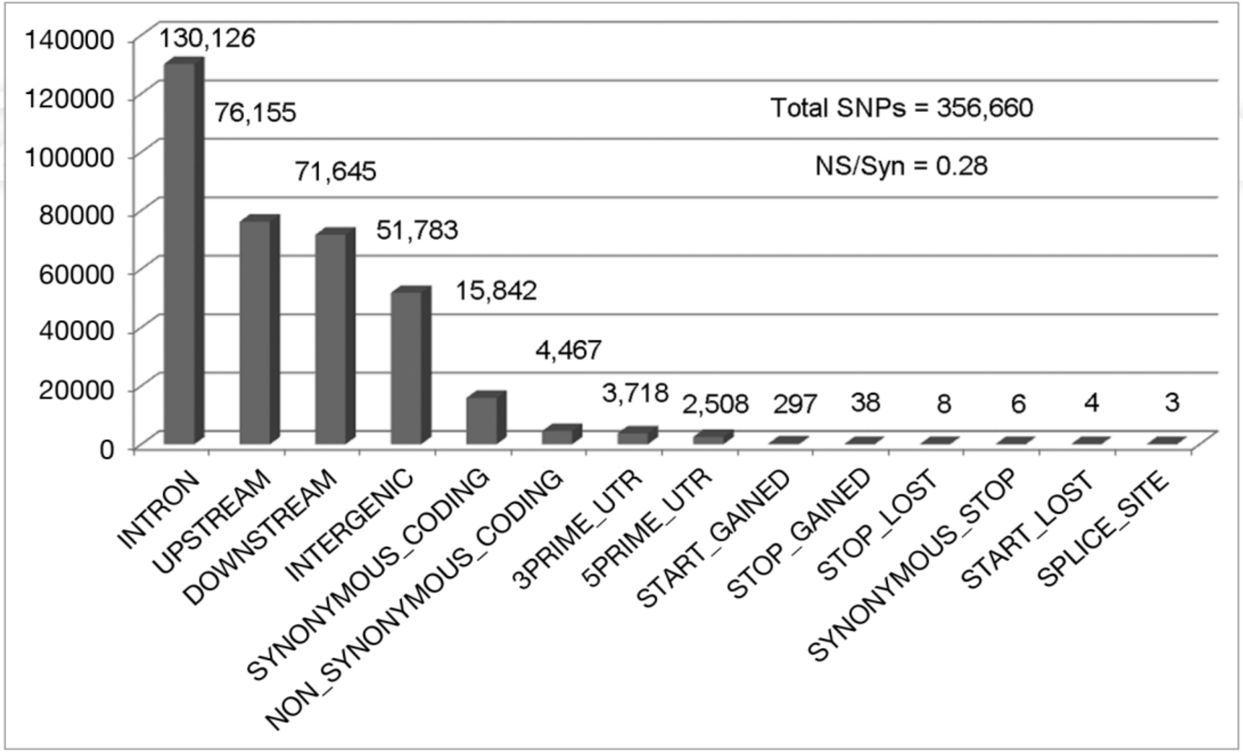


Genotype refinement improves GT quality and *de novo* calls

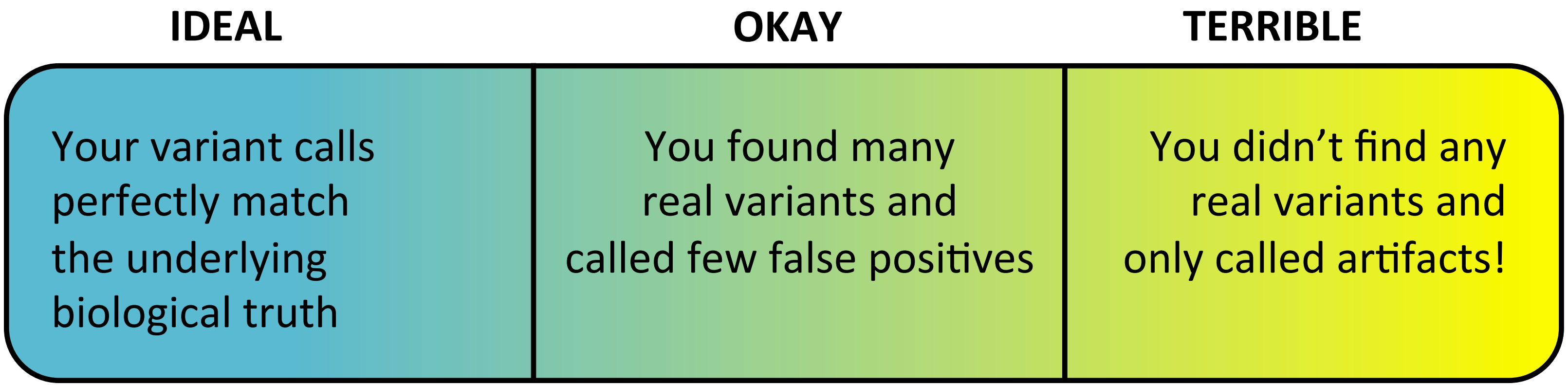



417 de novos from raw GT calls
17 de novos based on posterior GTs
8 high confidence *de novo* variants after GQ filtering

Functional annotation predicts effects of variants



Callset evaluation: where are you on this spectrum?



IDEAL

Your variant calls perfectly match the underlying biological truth

OKAY

You found many real variants and called few false positives

TERRIBLE

You didn’t find any real variants and only called artifacts!

(does not determine veracity of individual variant calls)