Introduction to Germline Variant Discovery

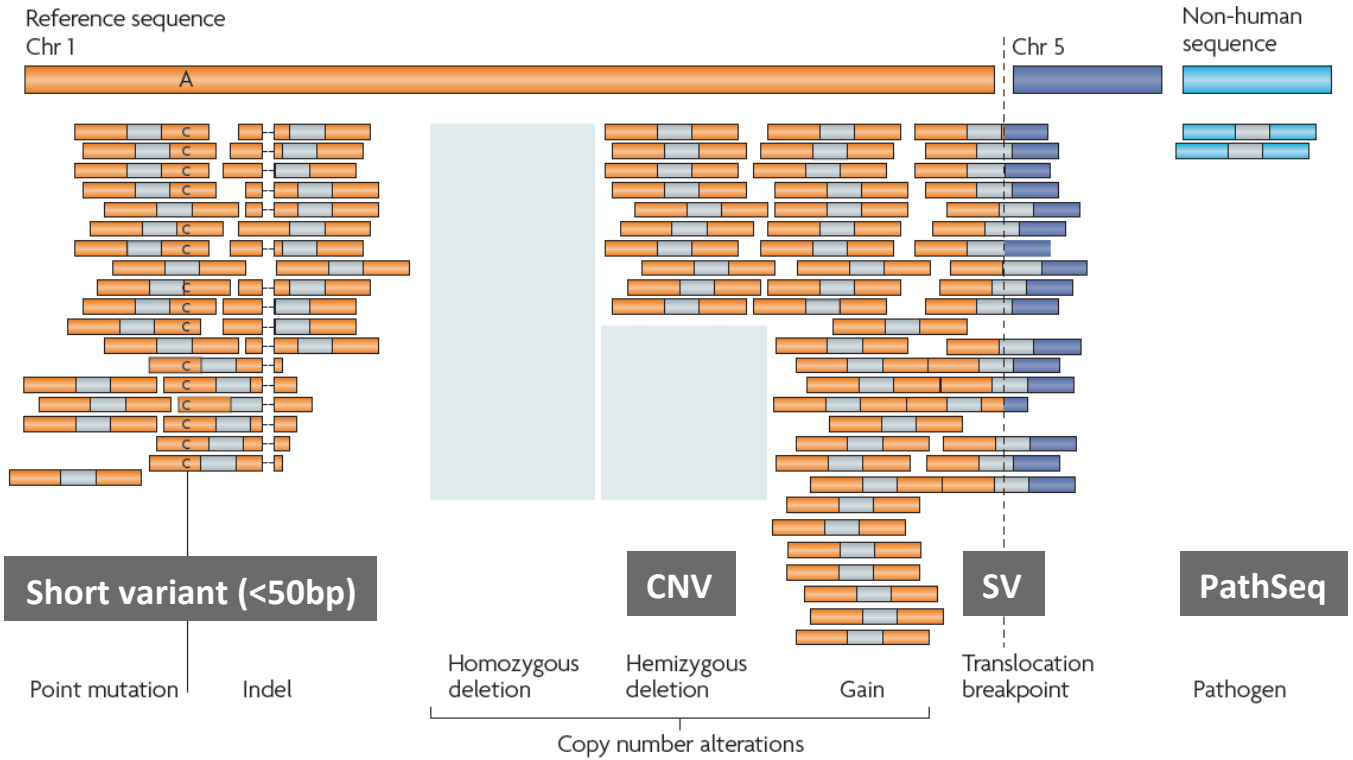
Key considerations and workflow logic

BROAD INSTITUTE

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

gatk

GATK detects different types of genomic variants for two contexts



Germline
Various ploidies

Somatic
Mixed purity, heterogeneous (pooled) or mosaic

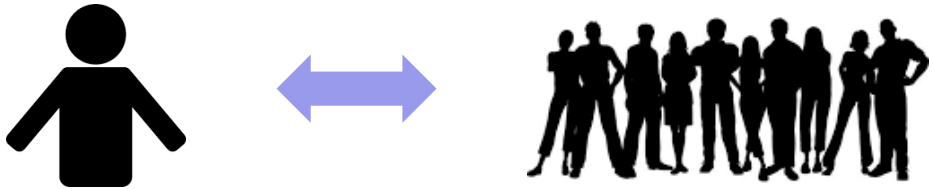
Calling variation against a *reference* enables comparative analyses

AGTGGAGCTGGGAAAGCAGCTGGCTAA
GAAAATAGAGCCTGAGCTTGATGGCAGT
GCTCAAGTGACCTCTCACGACGCTTCT

3 billion sites in the human genome

Humans share 99.5% DNA with any other human

We share commonly variant sites and most of these are *biallelic*

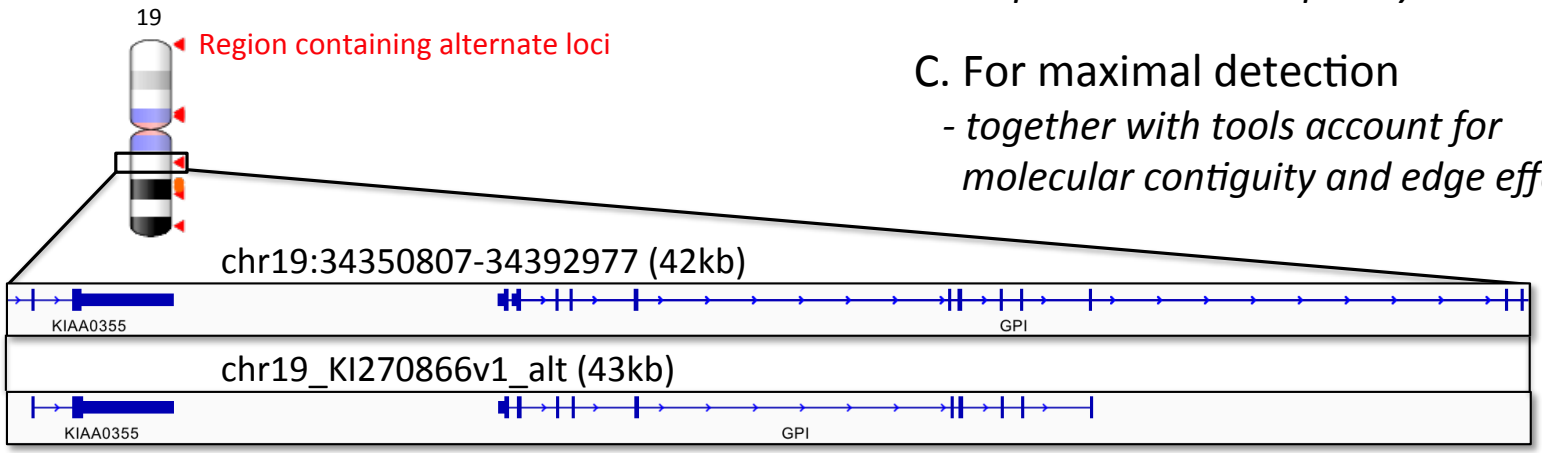


AGCTGGCTAAGAAAATAGAGCCTGAGCT

A good reference enables better variant detection

NCBI Build 34/hg16 (2003)

GRCh38 (2013)



A. For reads to map correctly
- *encompasses population diversity*
- *enables even complicated variants*

B. For even alignment coverage
- *reflects copy number & pseudo autosomal regions*
- *sequesters low complexity & extraneous sequences*

C. For maximal detection
- *together with tools account for molecular contiguity and edge effects*

** Image from GRC <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/> modified by S.H. Lee 2016*

So you can focus on the interesting analyses

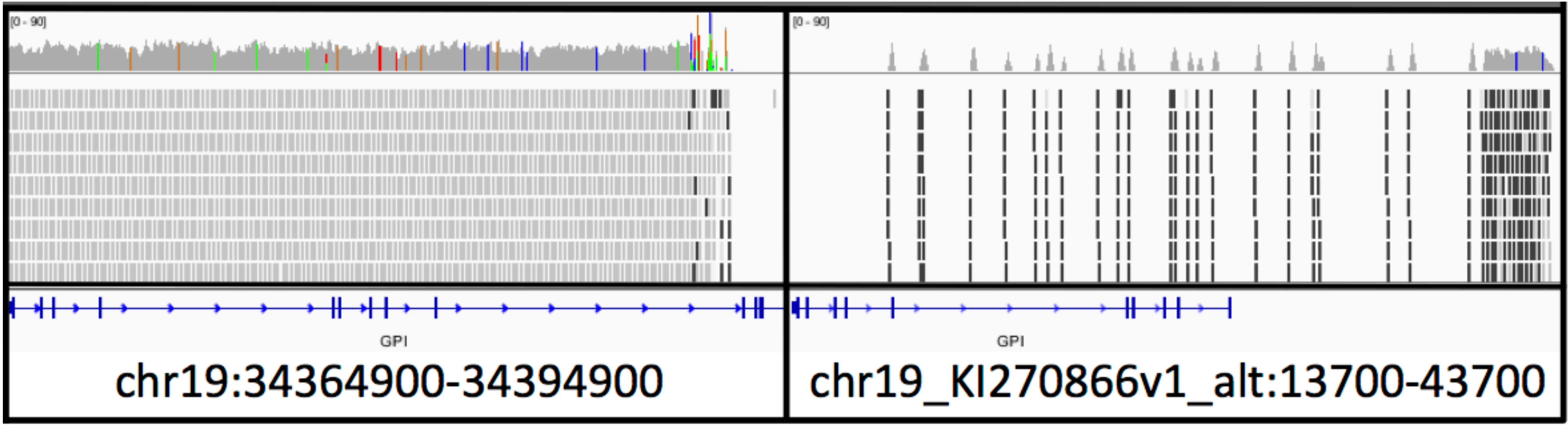
Gtex portal is at https://www.gtexportal.org/home/; figure from https://gatkforums.broadinstitute.org/gatk/discussion/8180

Our brains process visuals 60,000x faster than text

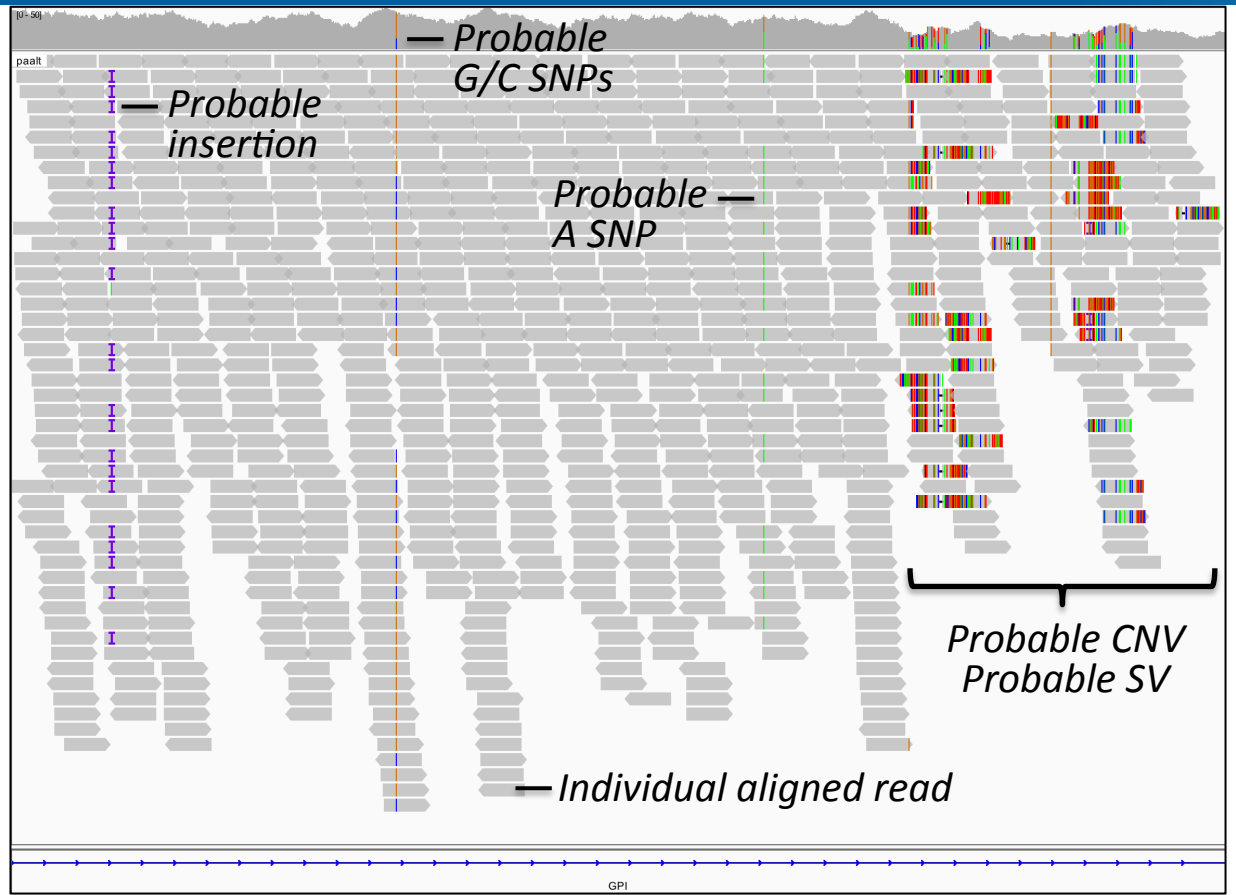
Two SAM format records https://samtools.github.io/hts-specs/

```
chr19_KI270866v1_alt_4hetvars_26518_27047_0:0:0_0:0:0_931	83	chr19	34377685	60	151M	=	34377337	-499	CTTTTCG
TGGGCCCTGAATTCTTATTCTCTGATGCTATGTCTCCCCGCAGGGGCCCCTCATGGTGACTGAAGCCCTTAAGCCATACTCTTCAGGAGGTCCCCGCGTCTGGTATGTCTCCAACATTGATGGAACTCACATTGCCAAAACCCT	I
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIN
M:i:1	MD:Z:52A98	AS:i:146	XS:i:0	SA:Z:chr19_KI270866v1_alt,26897,-,151M,8,0;	pa:f:0.967
chr19_KI270866v1_alt_4hetvars_26518_27047_0:0:0_0:0:0_931	2131	chr19_KI270866v1_alt	26897	8	151M	chr19	34377337	0	C
TTTTCGTGGGCCCTGAATTCTTATTCTCTGATGCTATGTCTCCCCGCAGGGGCCCCTCATGGTGACTGAAGCCCTTAAGCCATACTCTTCAGGAGGTCCCCGCGTCTGGTATGTCTCCAACATTGATGGAACTCACATTGCCAAAACCCTI
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIN
M:i:0	MD:Z:151	AS:i:151	XS:i:146	SA:Z:chr19,34377685,-,151M,60,1;	XA:Z:chr19,-34377685,151M,1;
```

Thousands of aligned records on IGV



What variants look like in a genome browser



Depth of coverage

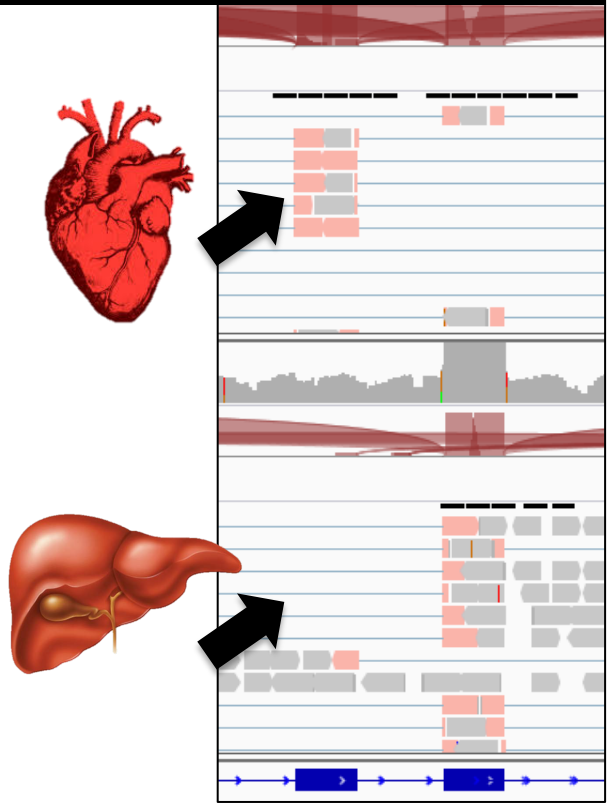
Non-reference bases are colored; reference bases are grey

A C G T

Reference genome

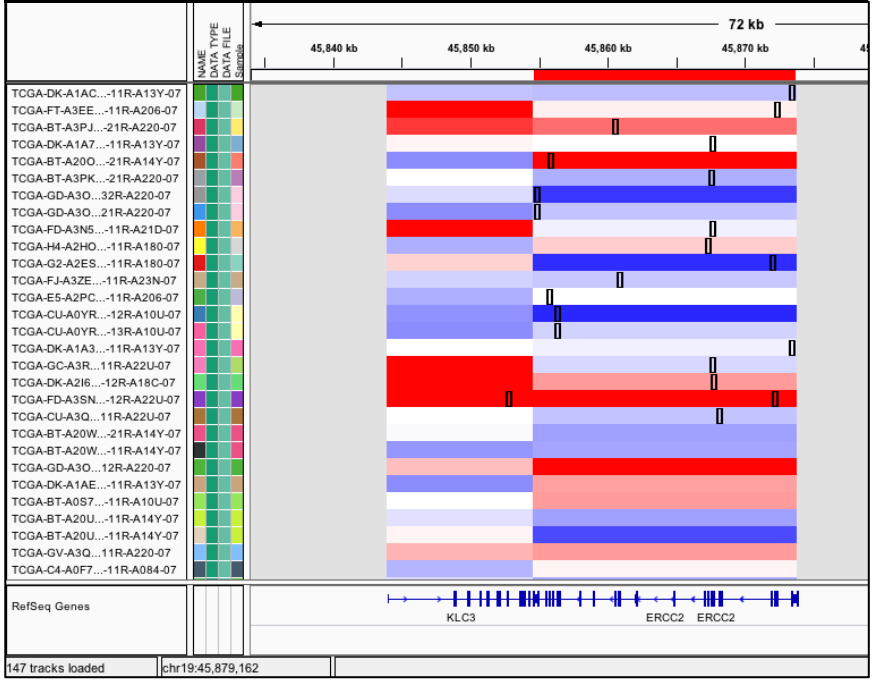
IGV is the *Integrative Genomics Viewer*

Visualize RNA splice isoforms



Visualize diverse genomic data

Overlay mutations on heatmap data



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

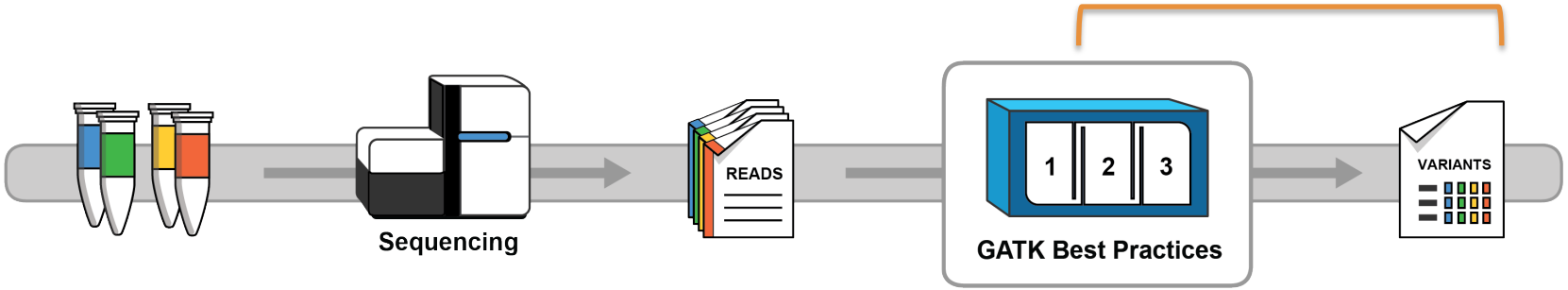
From format specification in https://samtools.github.io/hts-specs/VCFv4.2.pdf

VCF format supports CNVs and SVs

```
...
##INFO=<ID=BKPTID,Number=.,Type=String,Description="ID of the assembled alternate allele in the assembly fil
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
###INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">
#CHROM POS ID    REF   ALT   QUAL       FILTER     INFO       FORMAT     NA00001
1 2827694  rs2376870 CGTGGATGCGGGGAC  C     . PASS SVTYPE=DEL;END=2827708;HOMLEN=1;HOMSEQ=G;SVLEN=-14 GT:GQ
2  321682        .     T               <DEL> 6 PASS SVTYPE=DEL;END=321887;SVLEN=-205;CIPOS=-56,20;CIEND=-10,6
3 12665100       .     A               <DUP> 14 PASS SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500,500;CIEND
```

Symbolic allele in angle-bracketed ID

From format specification at https://samtools.github.io/hts-specs/VCFv4.2.pdf



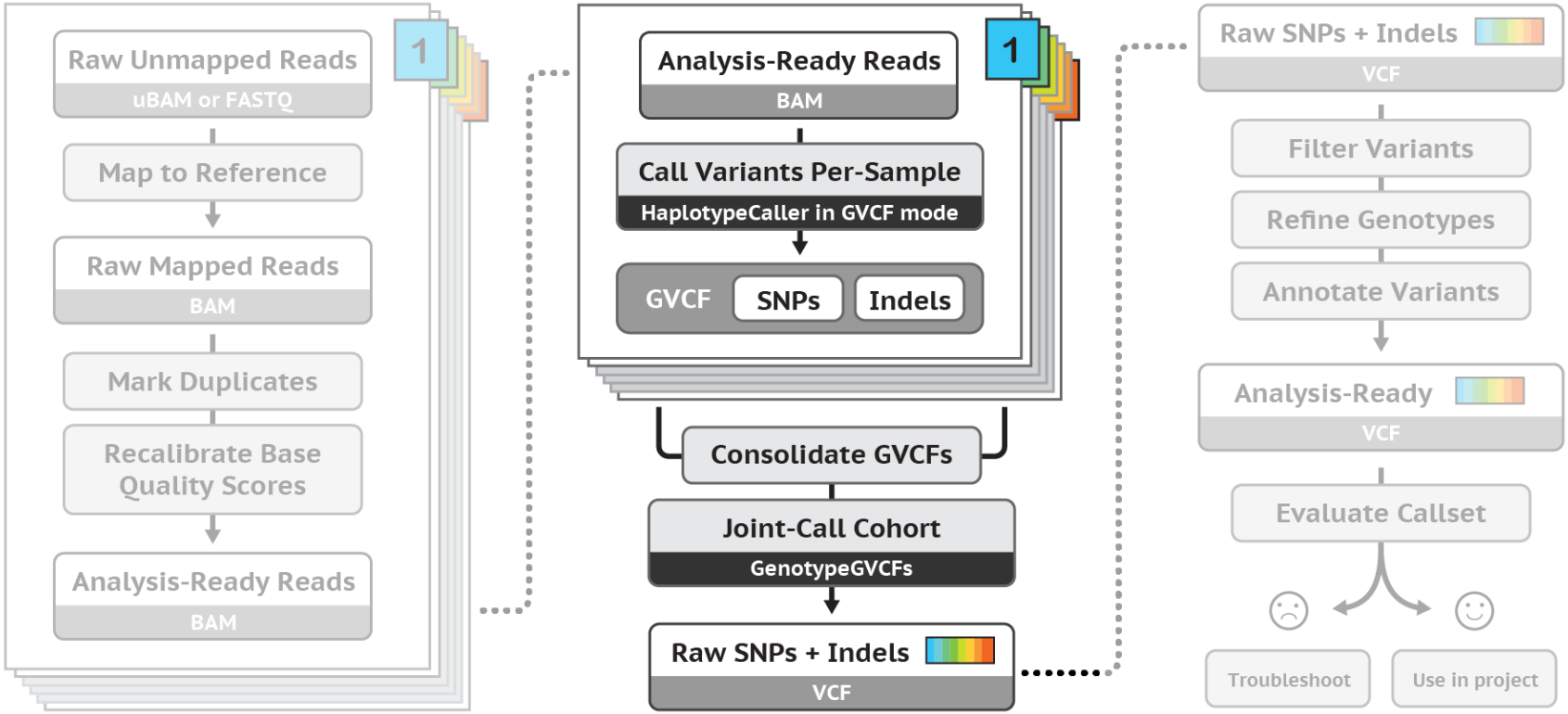
THE WORKFLOW

CNV central concept: Panel of Normals (PoN)

The panel of normals (PoN) forms
the baseline for what coverage is normal

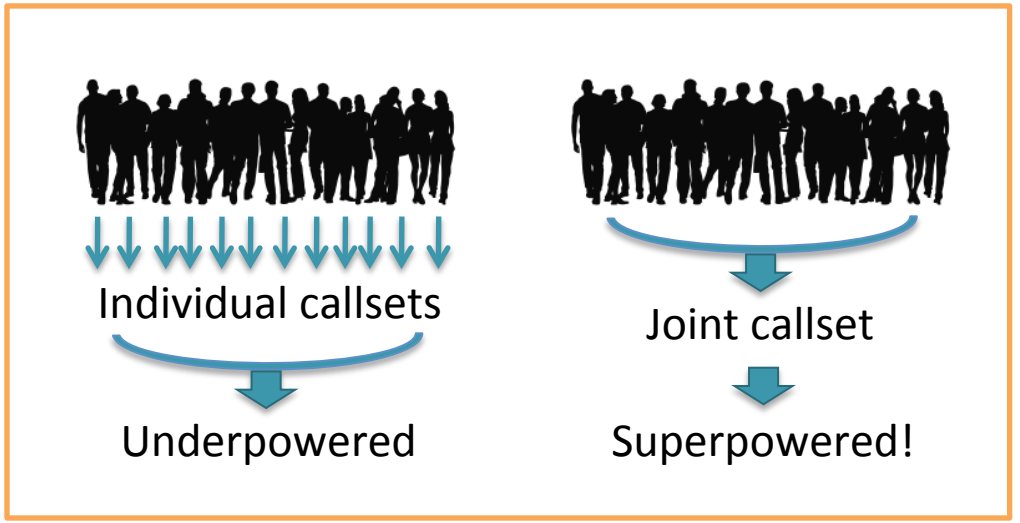
choose constituents with care

Short variants central concept: joint calling

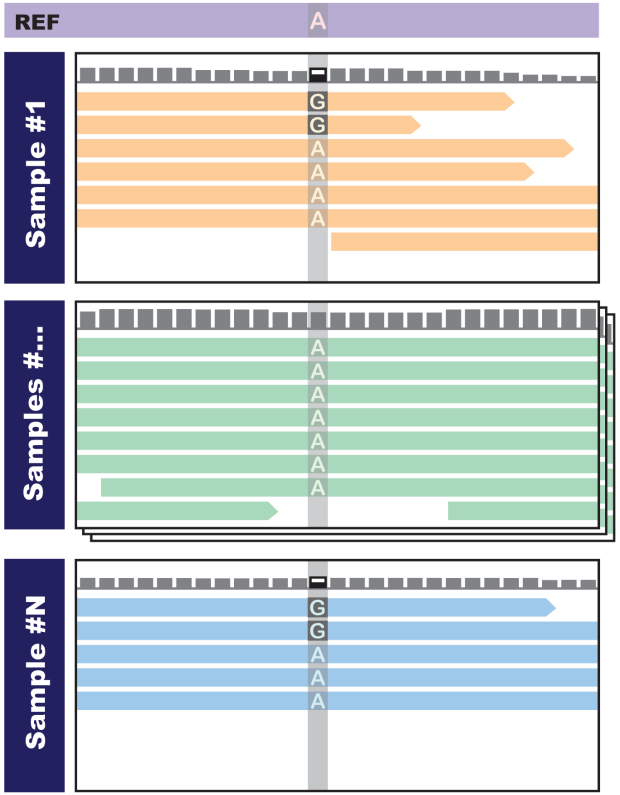


Joint analysis empowers discovery

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



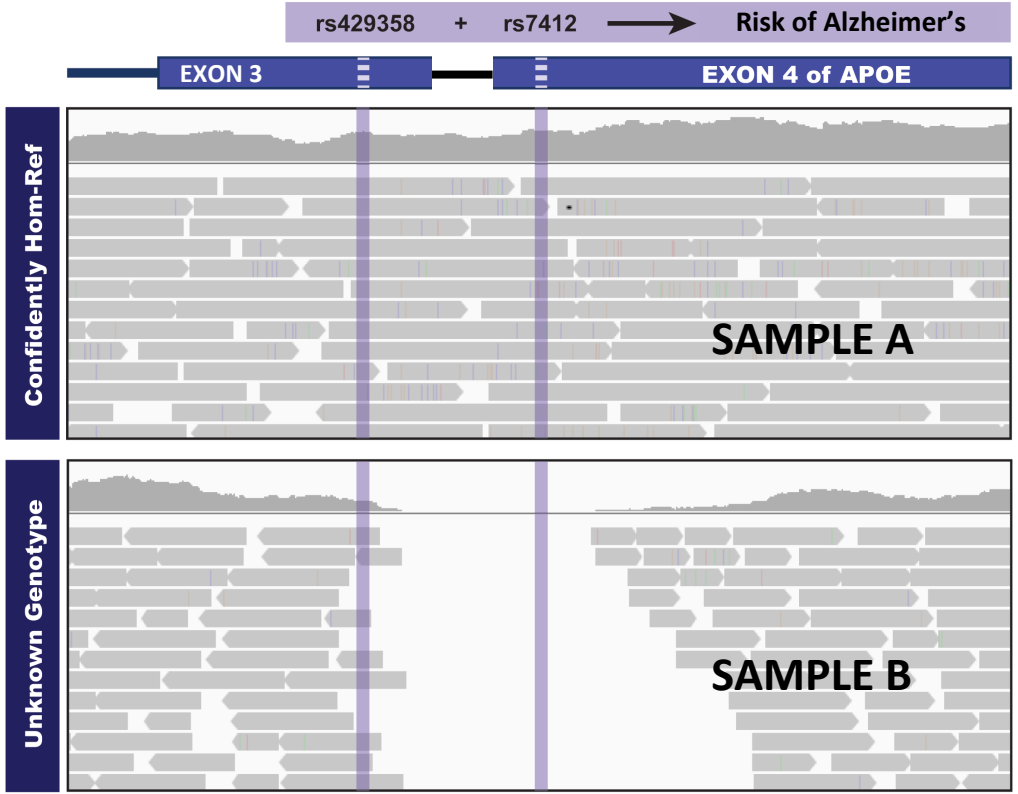
Discovery is empowered at difficult sites



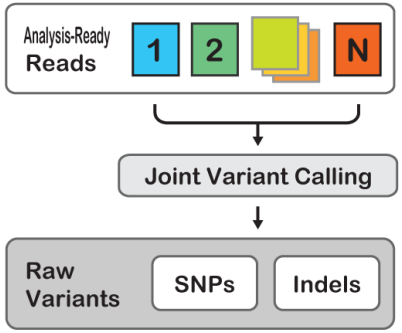
- Sample #1 or Sample #N alone:
  - weak evidence for variant
  - may miss calling the variant
- Both samples seen together:
  - unlikely to be artifact
  - call the variant more confidently

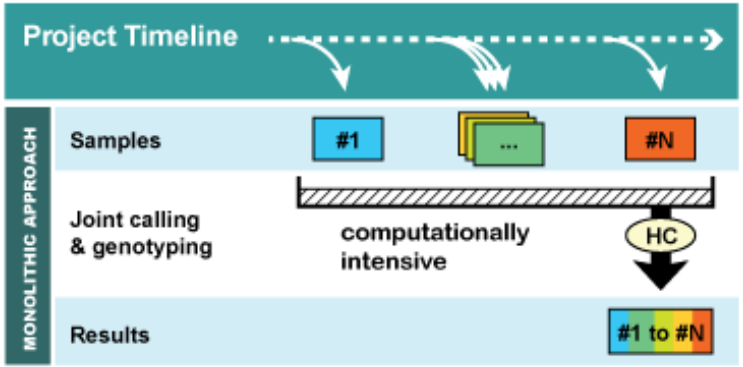
And we get full information at all sites of interest

- **Analyzed individually:**
  - No call for either sample
  - Very different reasons!
- **In joint analysis with other samples:**
  - Hom-ref call and no-call genotypes emitted



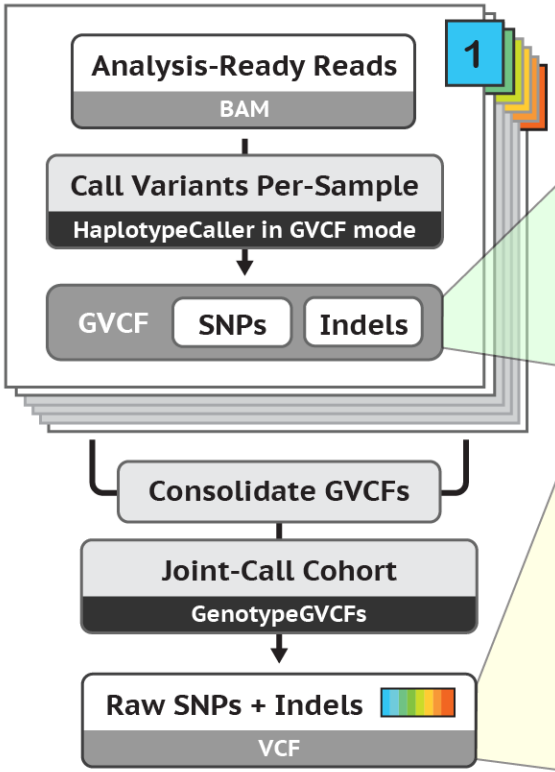
Traditional multi-sample calling approaches scale poorly





- Increasing number of samples increases compute requirement
- It gives us the right answers, BUT
- Want to add new samples? Having to re-run pipeline from scratch gives rise to the N+1 problem.

Solution: the GVCF-based joint calling workflow

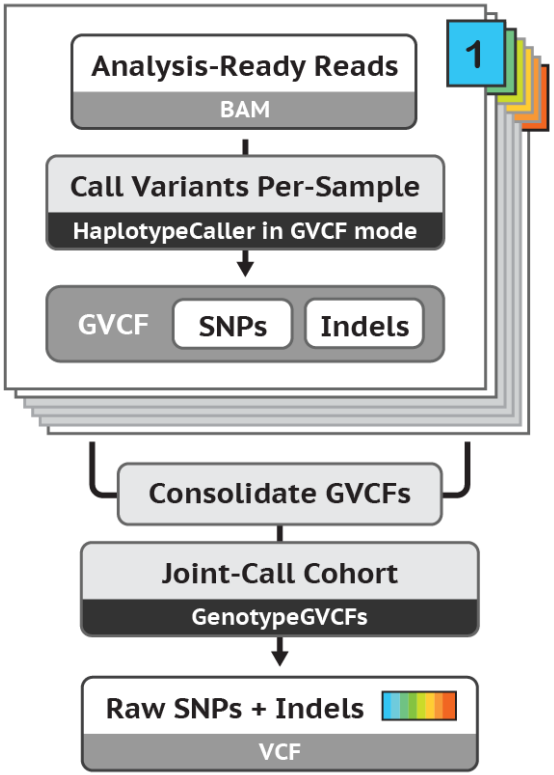


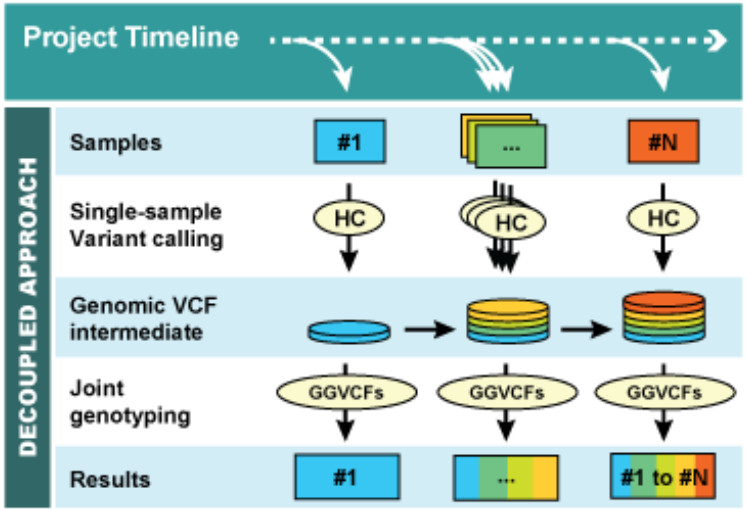
Generate per-sample Genomic VCFs (GVCFs)
then joint-call across all samples for final VCF

```
#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 incrementally scalable



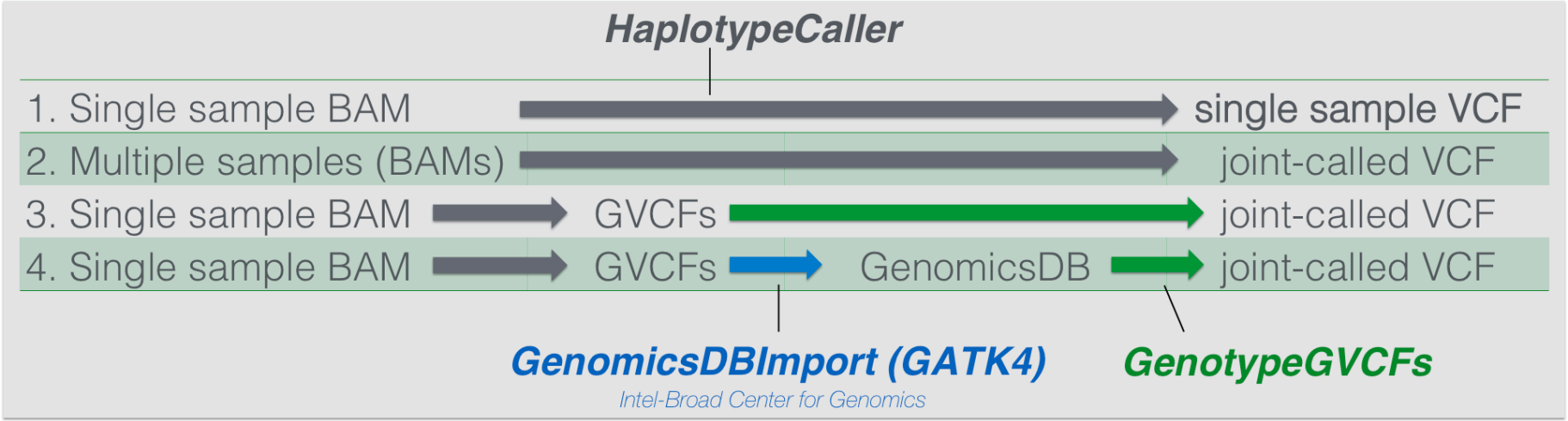


Scales linearly with number of samples

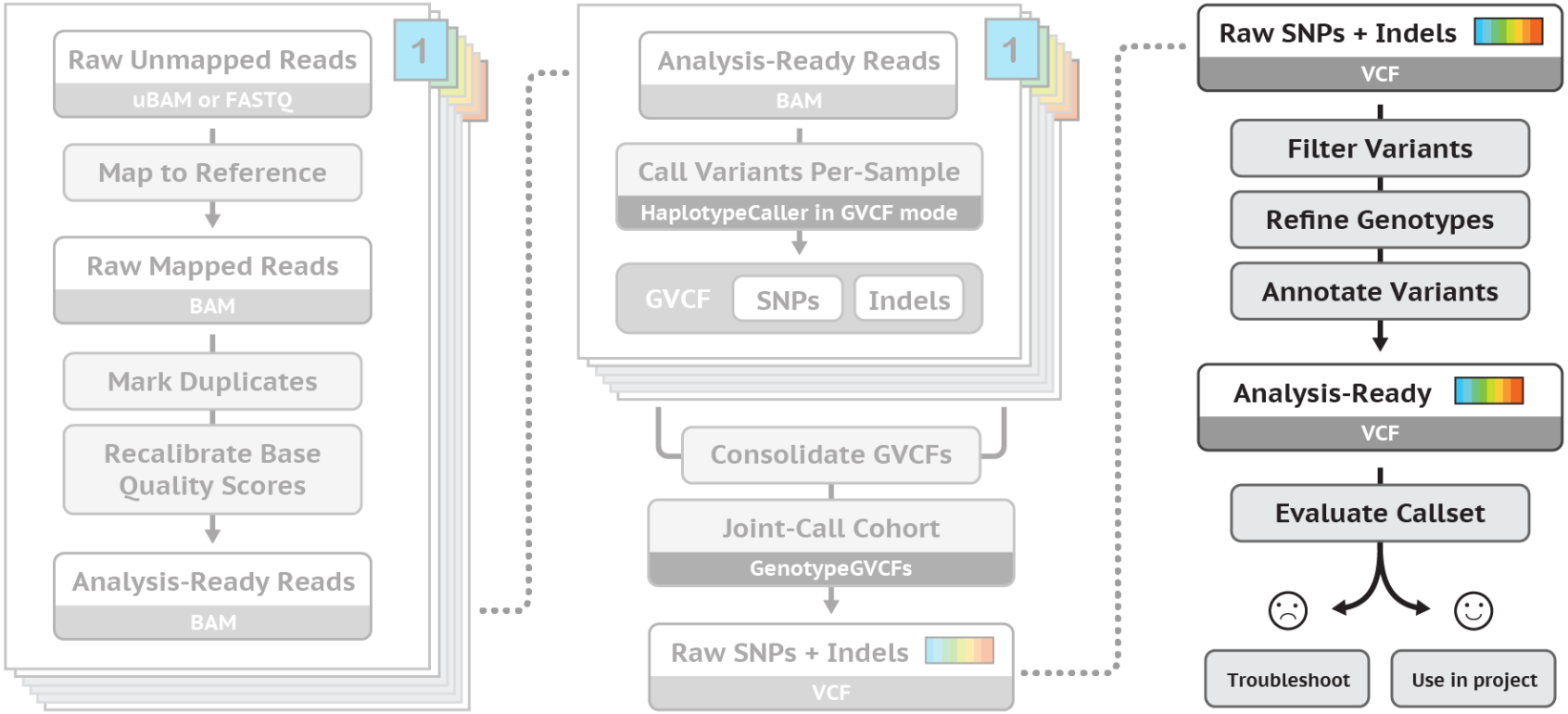
Want to add a new sample?
Make a GVCF for that sample then re-call the cohort

The many hats of Haplotypecaller

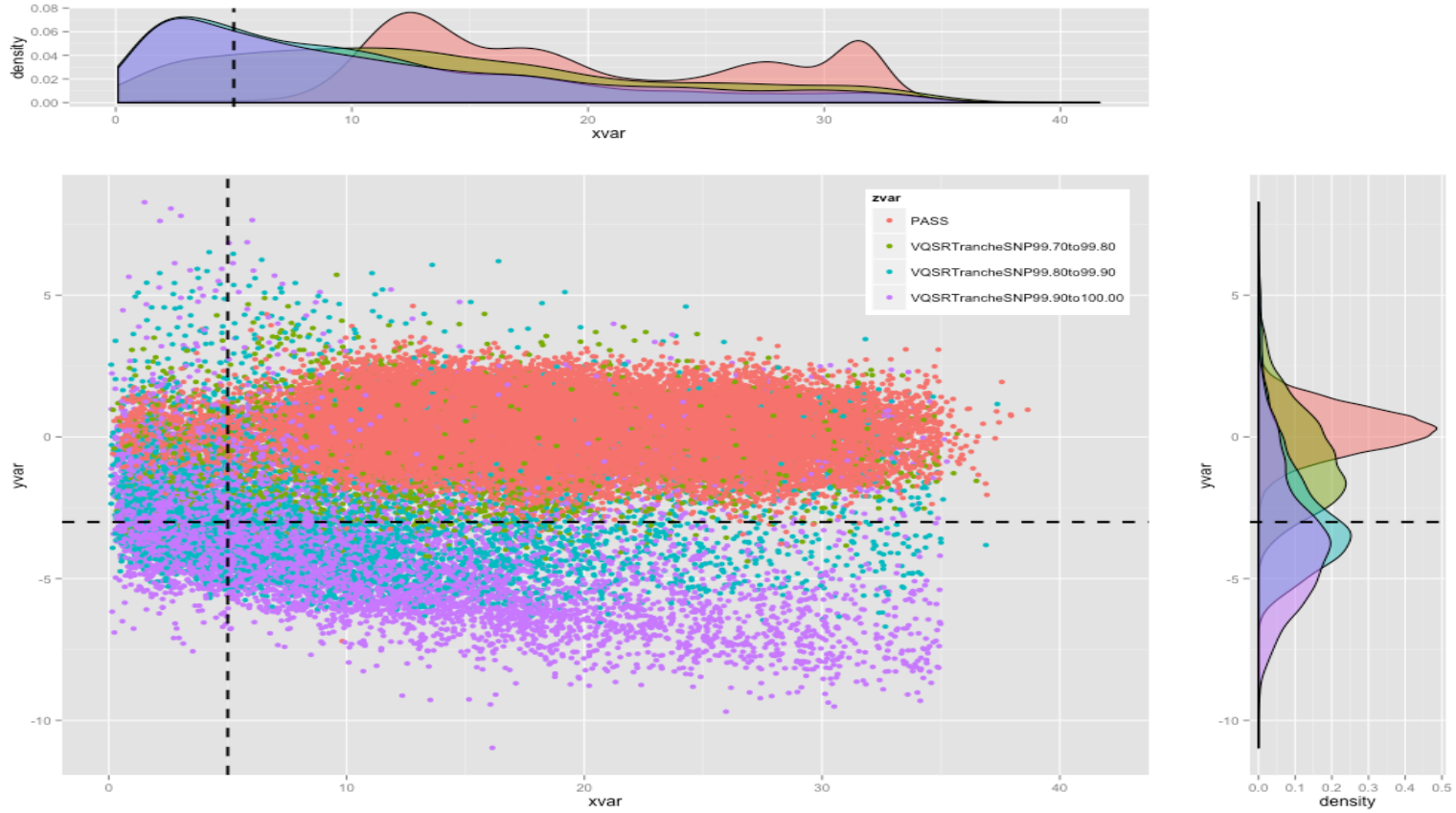
- Joint genotyping samples together (*2–4*) enables calling variants with high confidence and detection of rare variants
- HaplotypeCaller's ERC GVCF mode breaks down compute per sample towards joint calling (*3 and 4*)
- GenomicsDB *database* stores GVCF data for fast and flexible joint calling (*4*)



Refinement central concepts: filtering and evaluation

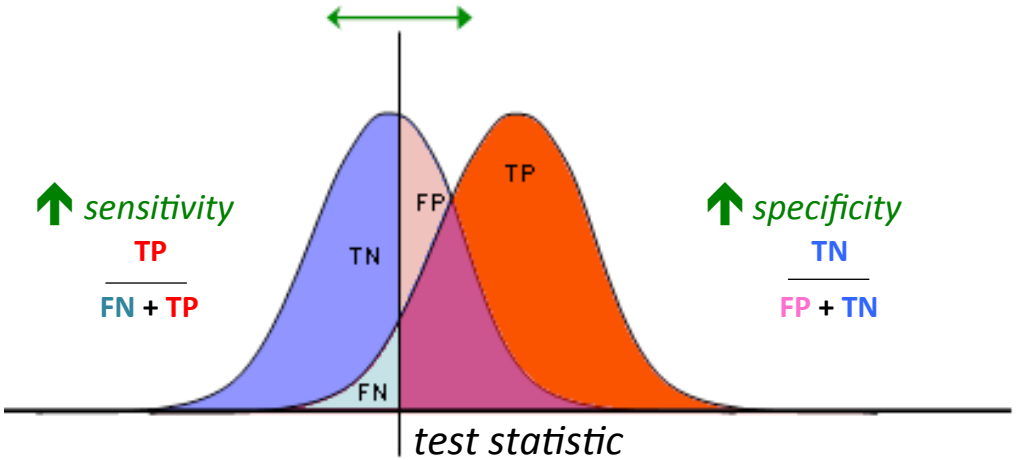


Variant filtering reduces false positives



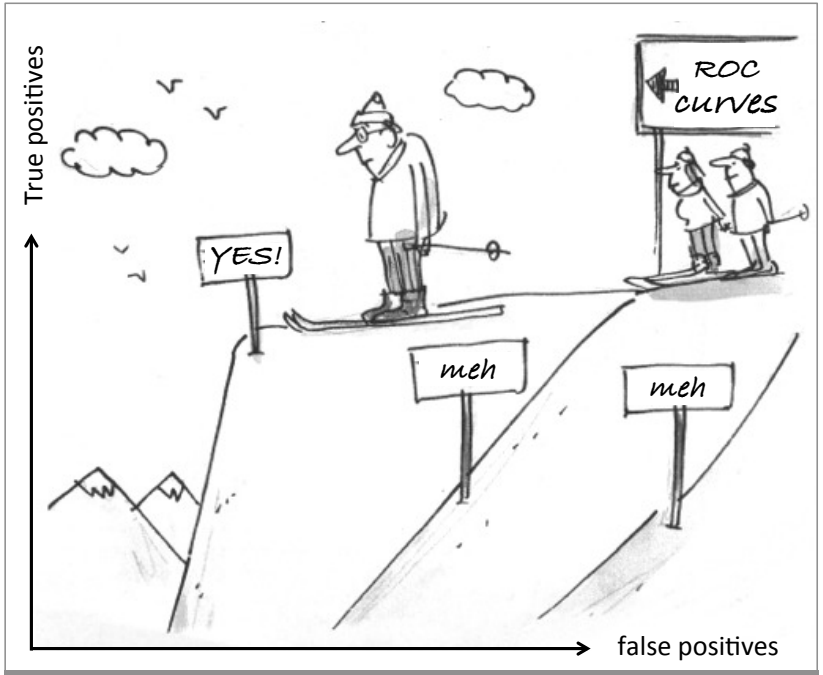
Evaluate callset for downstream analyses

Adapted from Wikipedia https://en.wikipedia.org/wiki/Receiver_operating_characteristic



What is considered true?

- Site level concordance
- Variant level concordance
- Genotype concordance and correct zygosity



ROC curve of TPs vs. FPs
Steeper inclines indicate better test statistic