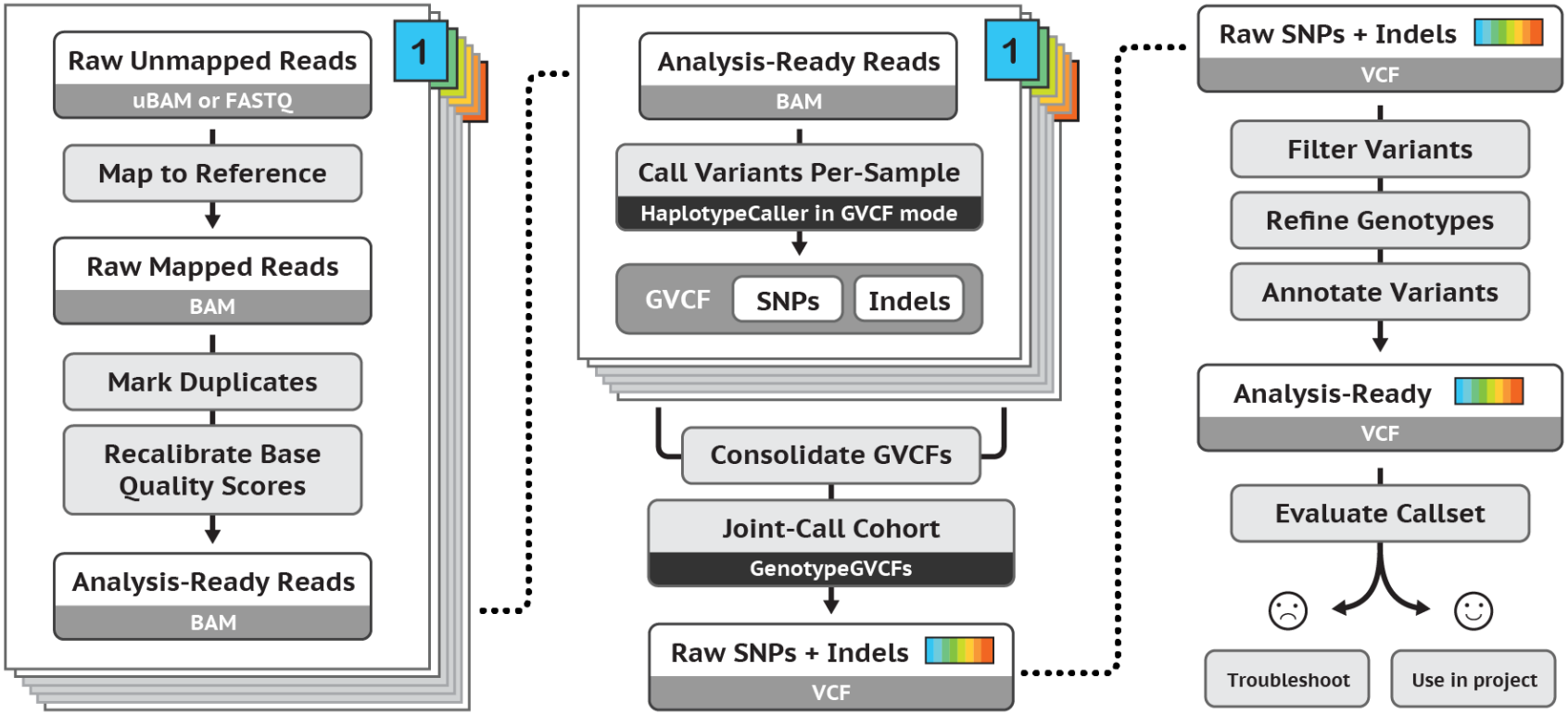
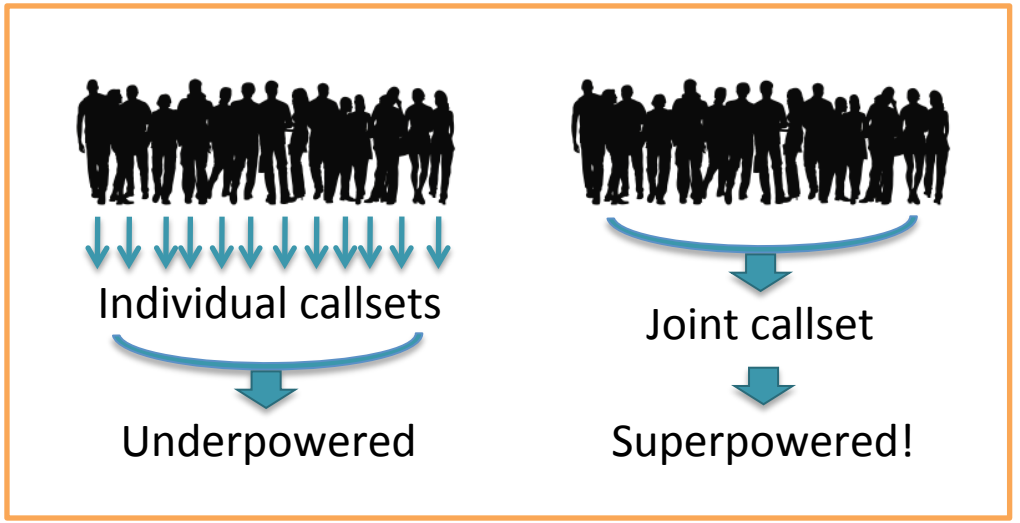
Germline Variant Discovery Highlights

Best Practices for Germline SNP & INDEL Discovery

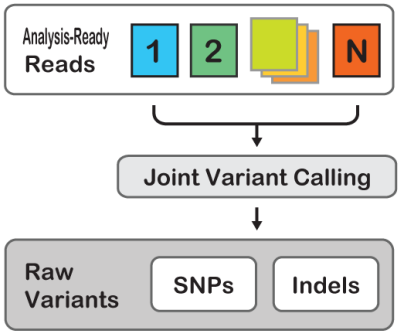


Joint analysis empowers discovery

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

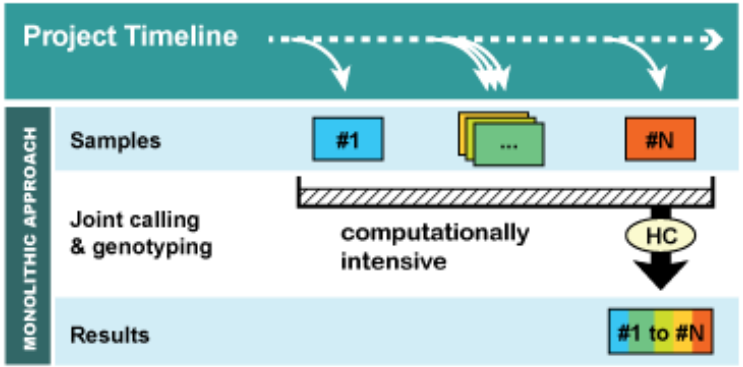


Traditional **multi-sample calling** approach : very inefficient



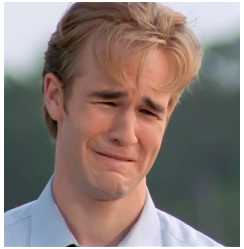
Compute requirements scale very badly with number of samples!!!

It gives us the right answers, but…

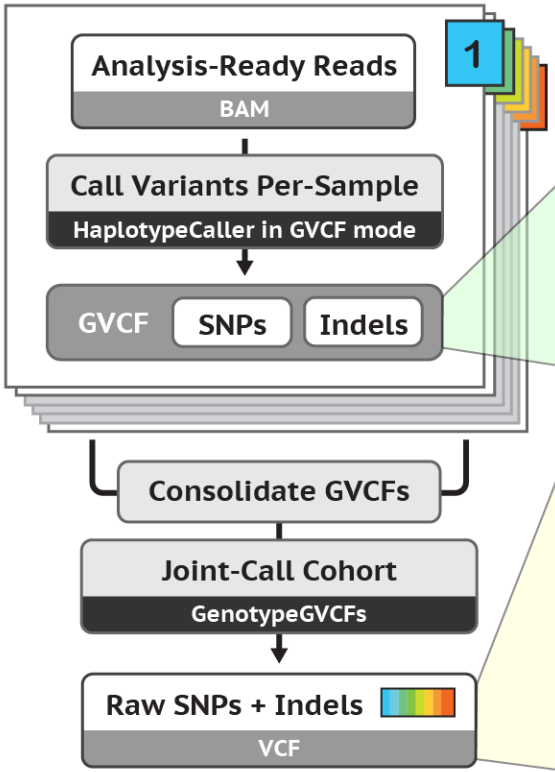


Want to add new samples?

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



Solution: the GVCF-based joint calling workflow

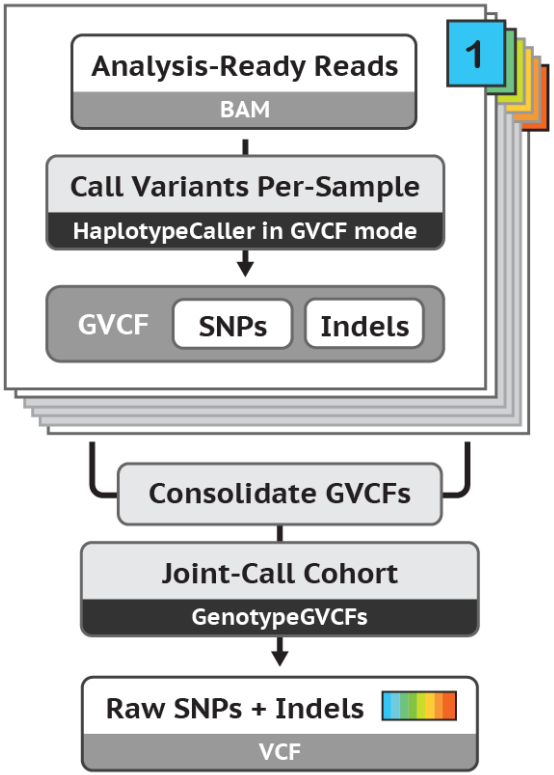


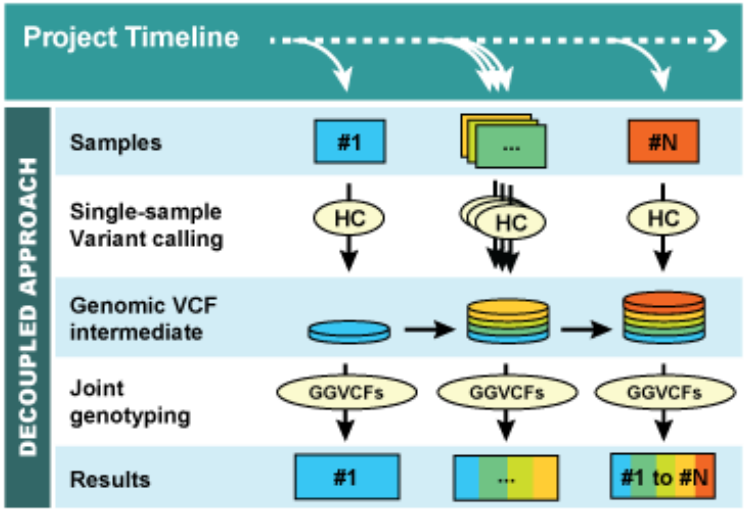
Generate per-sample Genomic VCFs (GVCFs)
then joint-call across all samples -> final VCF

```
#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