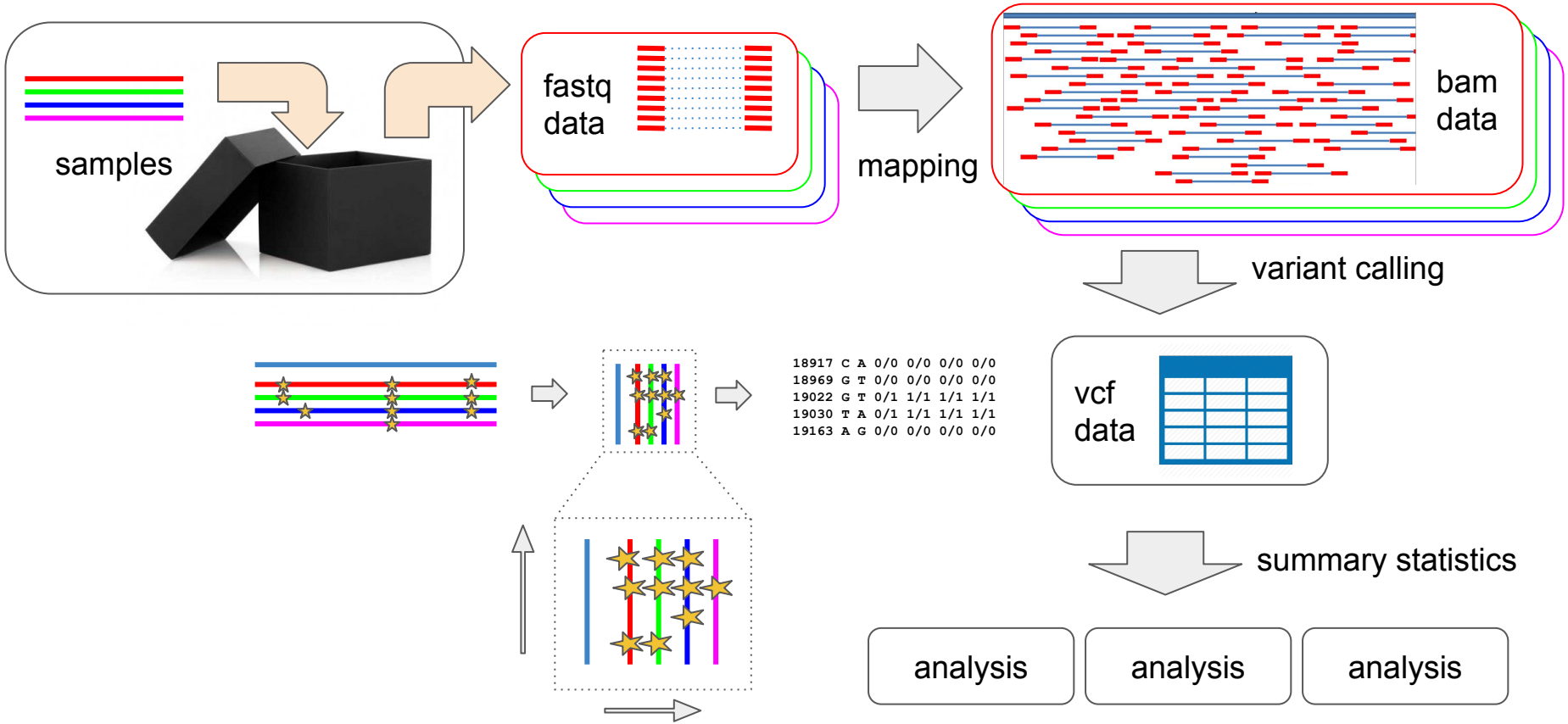
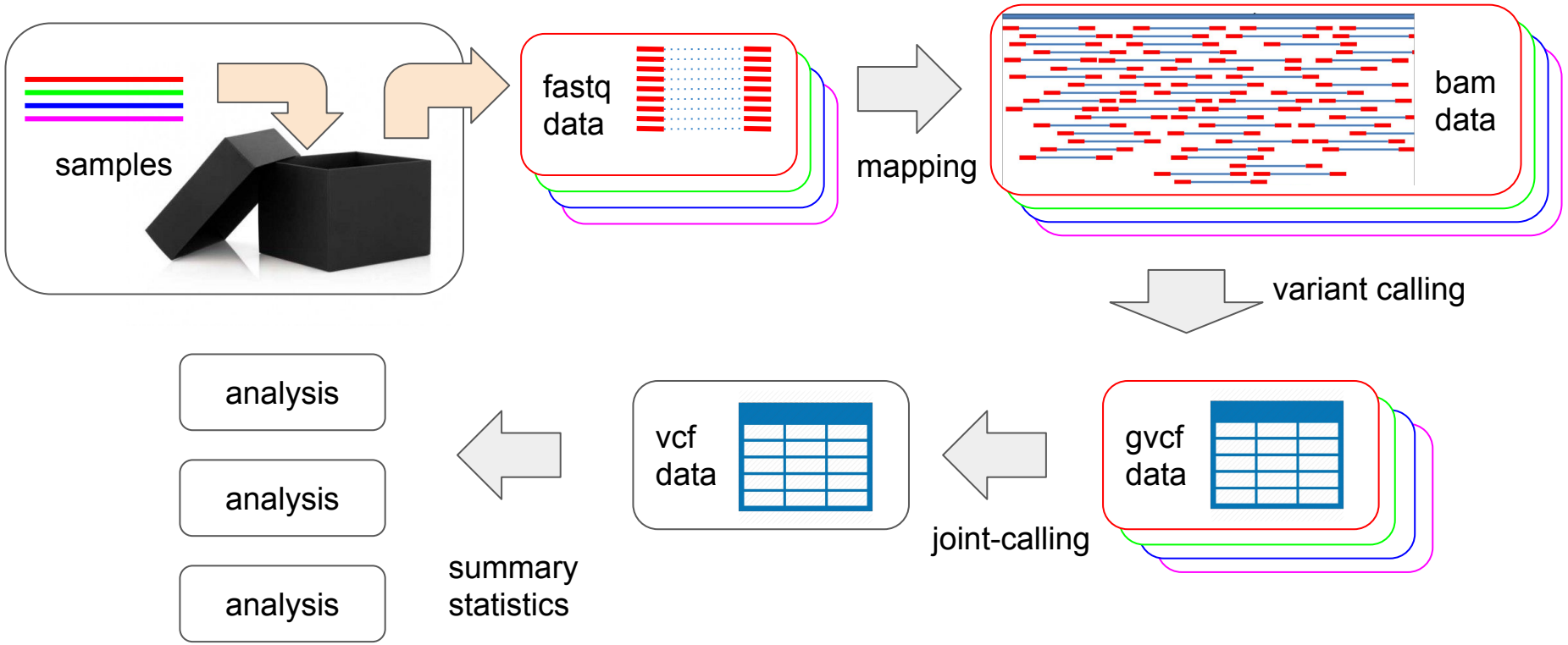
17 March, 2016:

From fastq to vcf

Overview of resequencing data analysis



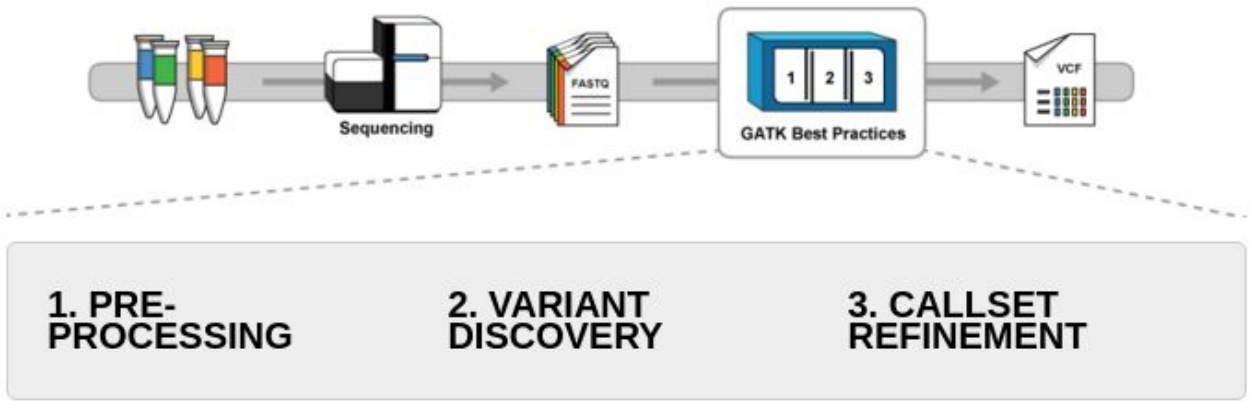
Overview of resequencing data analysis



Resequencing data analysis

Many different software for mapping and variant calling

- we use bwa, samtools and GATK, loosely following *GATK Best Practices*
- https://www.broadinstitute.org/gatk/guide/best-practices.php
- bwa and samtools are free; GATK is *gratis* but not *libre*
- different approach: ANGSD (http://popgen.dk/wiki/index.php/ANGSD)



sam format: alignment to reference

fastq:

```
@K00137:79:H3KNNBBXX:3:1101:1030:21850
TTTGACCATTATTGTTACTTTTAAAGCTTTTGCAGAAGGAGTAAACAGCTAAAACAGAATGTATGAAGAAGGCTTTAGCTTACACTGTTTCTATCAAGCAAATGACCTTTTTTAATCATTA
+
AAFFFKKKKKKKKKKKKKKKKKKKKKKAKKKKKKAKKK7FKKKKKKKKKKKKKKKKKKKKFKKKKKKKFKFAFFKKKKKAFKKKKFKKKKKK<FKFKKKKKKKAAKK<AKKK,,,,,,,A,
```

sam:

QNAME FLAG RNAME POS MAPQ CIGAR RNEXT PNEXT TLEN SEQ QUAL <optional fields>

```
K00137:79:H3KNNBBXX:3:1101:1030:21850   163     ctg7180000008138        10224   0       150M    =       10362   288
TTTGACCATTATTGTTACTTTTAAAGCTTTTGCAGAAGGAGTAAACAGCTAAAACAGAATGTATGAAGAAGGCTTTAGCTTACACTGTTTCTATCAAGCAAATGACCTTTTTTAATCATTA
AAFFFKKKKKKKKKKKKKKKKKKKKKKAKKKKKKAKKK7FKKKKKKKKKKKKKKKKKKKKFKKKKKKKFKFAFFKKKKKAFKKKKFKKKKKK<FKFKKKKKKKAAKK<AKKK,,,,,,,A,
NM:i:5  MD:Z:113C7T0C16T4T5   AS:i:125        XS:i:125        RG:Z:sample-1
```

sam format: alignment to reference

https://samtools.github.io/hts-specs/SAMv1.pdf

http://genome.sph.umich.edu/wiki/SAM → CIGAR format

Phred scores (https://en.wikipedia.org/wiki/Phred_quality_score)

P = probability of being wrong

Q = Phred-scaled quality score

$Q = -10 \log_{10} P$

$P = 10^{-Q/10}$

$Q = 30 \rightarrow P = 10^{-30/10} \rightarrow P = 1/1000$

$P = 1/1000 \rightarrow Q = -10 \log_{10} (1/1000) \rightarrow Q = 30$

In fastq/sam format, ASCII(Q + 33) http://www.asciitable.com

sam format: alignment to reference

flags:

- using binary number: 1=yes, 0=no

Bit	Description
1	template having multiple segments in sequencing
2	each segment properly aligned according to the aligner
4	segment unmapped
8	next segment in the template unmapped
16	SEQ being reverse complemented
32	SEQ of the next segment in the template being reverse complemented
64	the first segment in the template
128	the last segment in the template
256	secondary alignment
512	not passing filters, such as platform/vendor quality controls
1024	PCR or optical duplicate
2048	supplementary alignment

vcf format: variant calls

http://samtools.github.io/hts-specs/VCFv4.3.pdf

http://www.1000genomes.org/wiki/Analysis/vcf4.0

METAINFORMATION
- data origin
- analysis history
- included fields

HEADER
- for the table below

DATA
- in a table format

vcf format: variant calls

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T   .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

vcf format: variant c

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele F
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestr
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP mem
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 m
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genot
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Geno
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T   .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                   GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Five variant positions

1. good simple SNP,
2. possible SNP filtered out due to low quality
3. site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error)
4. site that is called monomorphic reference (i.e. with no alternate alleles)
5. microsatellite with two alternative alleles, one a deletion of 3 bases (TCT), and the other an insertion of one base (A).

Genotype data for three samples, two phased, one unphased

1. genotypes
2. per sample genotype quality
3. read depth
4. haplotype qualities (for phased samples)

vcf and gvcf

vcf only includes variable positions

1. variable positions (with probability of call being correct)
   - if variants called for one sample, only variants of **that** sample
   - if called for a set of samples, only variants of **that** set of samples

Variant calling performed from bam files, computationally hard

- with a new/changed set of samples, calling has to be done again
- calling hundreds or thousands of samples gets really heavy

vcf and gvcf

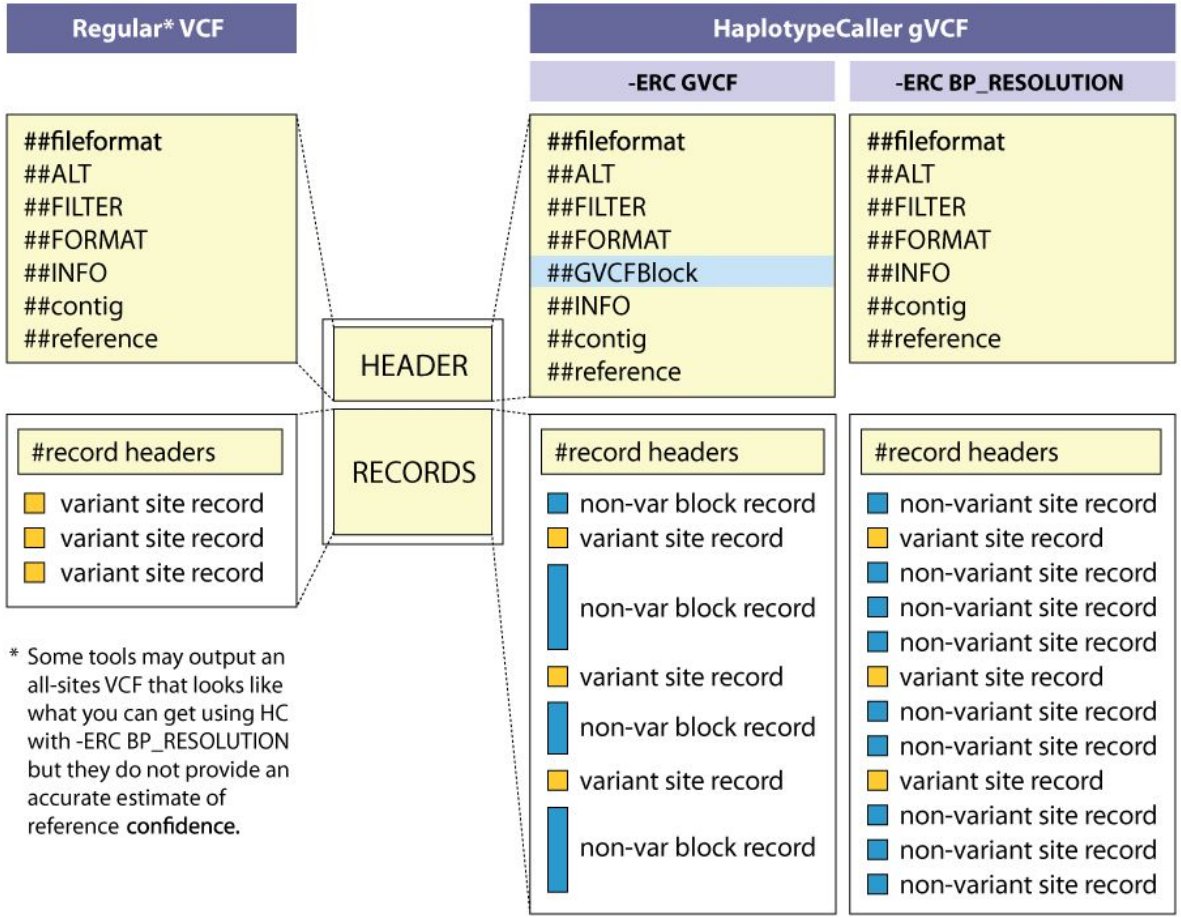
gvcf includes

1. variable positions (with probability of call being correct)
2. probability that intermediate positions are not variants
   - this could be done for every single position but the file gets big
   - typically non-variant positions are combined to intervals
     - approximate probabilities for sites within an interval (or band)

calling vcf from multiple gvcf files is computationally light

- adding samples is easy, large sample sets can be handled

https://www.broadinstitute.org/gatk/guide/article?id=4017


Regular* VCF
HaplotypeCaller gVCF
-ERC GVCF
-ERC BP_RESOLUTION
##fileformat
##ALT
##FILTER
##FORMAT
##INFO
##contig
##reference
HEADER
##fileformat
##ALT
##FILTER
##FORMAT
##GVCFBlock
##INFO
##contig
##reference
##fileformat
##ALT
##FILTER
##FORMAT
##INFO
##contig
##reference
RECORDS
#record headers
variant site record
variant site record
variant site record
#record headers
non-var block record
variant site record
non-var block record
variant site record
non-var block record
variant site record
non-var block record
#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.

gvcf

```
##GVCFBlock=minGQ=0(inclusive),maxGQ=5(exclusive)
##GVCFBlock=minGQ=20(inclusive),maxGQ=60(exclusive)
##GVCFBlock=minGQ=5(inclusive),maxGQ=20(exclusive)

#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  NA12878
20  10000000    .   T   <NON_REF>   .   .   END=10000116    GT:DP:GQ:MIN_DP:PL  0/0:44:99:38:0,89,1385
20  10000117    .   C   T,<NON_REF> 612.77  .   BaseQRankSum=0.000;ClippingRankSum=-0.411;DP=38;MLEAC=1,0;MLEAF=0.500,0.00;
  MQ=221.39;MQ0=0;MQRankSum=-2.172;ReadPosRankSum=-0.235   GT:AD:DP:GQ:PL:SB   0/1:17,21,0:38:99:641,0,456,691,519,1210:6,11,11,10
20  10000118    .   T   <NON_REF>   .   .   END=10000210    GT:DP:GQ:MIN_DP:PL  0/0:42:99:38:0,80,1314
20  10000211    .   C   T,<NON_REF> 638.77  .   BaseQRankSum=0.894;ClippingRankSum=-1.927;DP=42;MLEAC=1,0;MLEAF=0.500,0.00;
  MQ=221.89;MQ0=0;MQRankSum=-1.750;ReadPosRankSum=1.549    GT:AD:DP:GQ:PL:SB   0/1:20,22,0:42:99:667,0,566,728,632,1360:9,11,12,10
20  10000212    .   A   <NON_REF>   .   .   END=10000438    GT:DP:GQ:MIN_DP:PL  0/0:52:99:42:0,99,1403
20  10000439    .   T   G,<NON_REF> 1737.77 .   DP=57;MLEAC=2,0;MLEAF=1.00,0.00;
  MQ=221.41;MQ0=0 GT:AD:DP:GQ:PL:SB   1/1:0,56,0:56:99:1771,168,0,1771,168,1771:0,0,0,0
20  10000440    .   T   <NON_REF>   .   .   END=10000597    GT:DP:GQ:MIN_DP:PL  0/0:56:99:49:0,120,1800
20  10000598    .   T   A,<NON_REF> 1754.77 .   DP=54;MLEAC=2,0;MLEAF=1.00,0.00;
  MQ=185.55;MQ0=0 GT:AD:DP:GQ:PL:SB   1/1:0,53,0:53:99:1788,158,0,1788,158,1788:0,0,0,0
20  10000599    .   T   <NON_REF>   .   .   END=10000693    GT:DP:GQ:MIN_DP:PL  0/0:51:99:47:0,120,1800
20  10000694    .   G   A,<NON_REF> 961.77  .   BaseQRankSum=0.736;ClippingRankSum=-0.009;DP=54;MLEAC=1,0;M
```

Analysis pipeline

1. preparations: indexing, dictionary — for reference

2. alignment (or mapping)
3. sorting and duplicate removal
4. re-alignment around indels
5. variant calling: genome vcf

— for each sample

6. joint-calling: vcf
7. vcf data filtering

— combined data