Data analytics for personalized genomics and precision medicine

Lecturer: Yu LI(CSE)

Email: liyu@cse.cuhk.edu.hk

Lecture 15: Genomics analysis

Friday, October 24, 2025

Why do we care about Variants?

* Under 3.2 billion sites in the human genome, any 2 humans share 99.5% DNA

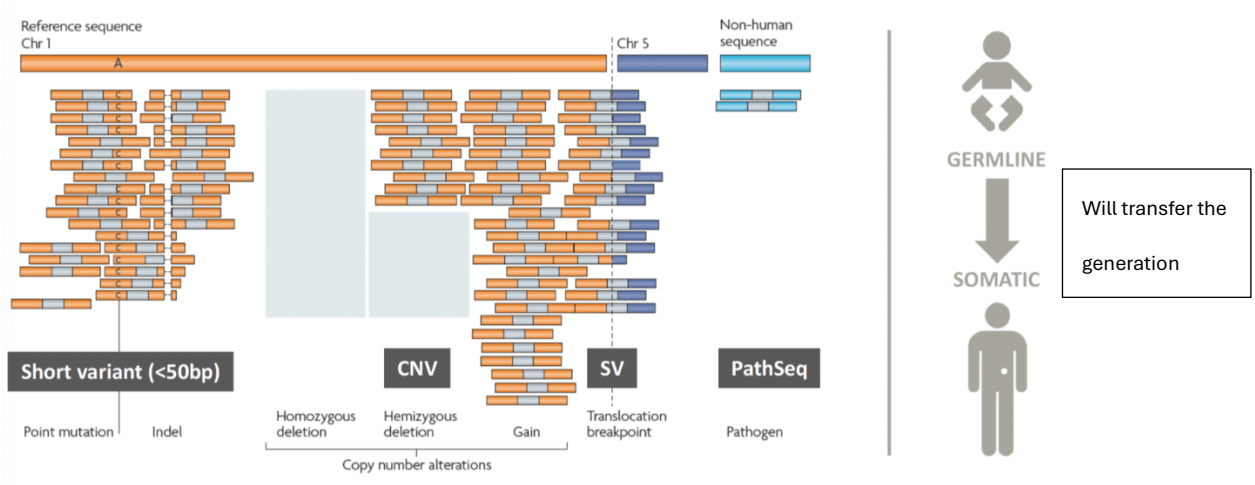
-> are the same, we need to do selection

* Genetic differences among people lead to differences in disease risk and response to treatment

* Genetic variation is used to find genes and variants that contribute to disease

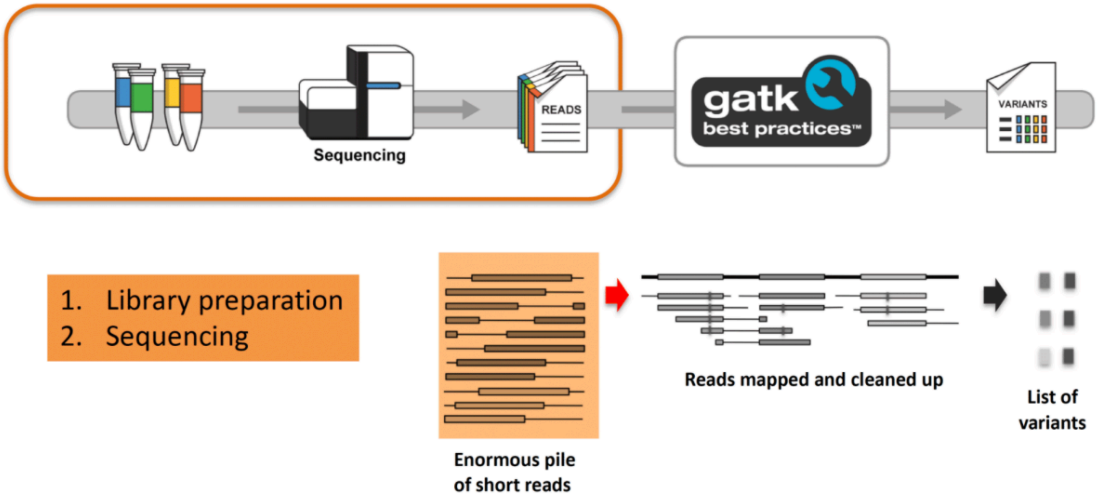
* Cancer is a genetic variant at multiple levels

Different types of genomic variants



- Point mutation: can change the signal
- SV-translocation breakpoint just like structure

Process of discover the genetic variants



For the sequence mapping recap

T	A	A	T	G	C	C	A	T	G	G	A	T	G
					C	C	A						
	2	3	3	3	2	0	2	3	3	2	3	3	

The no. means the difference

For this example, compare GCCA with CCA, they are the same

⇨ The no. equal 0

Example 2:

T	A	A	T	G	C	G	A	T	G	G	A	T	G
					C	C	A						
	2	3	3	3	2	1	3	3	3	2	3	3	

Compare GCGA with CCA, they have one difference

⇨ The no. equal 1

⇨ Its doesn't mean that having the point mutation (Because only one sample)

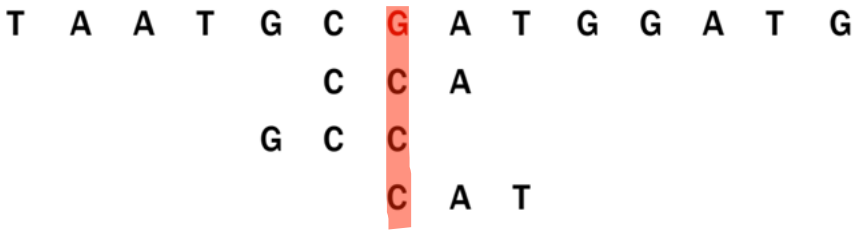
Variants VS Errors

* Distinguish between actual variation (real change) and errors (artifacts)

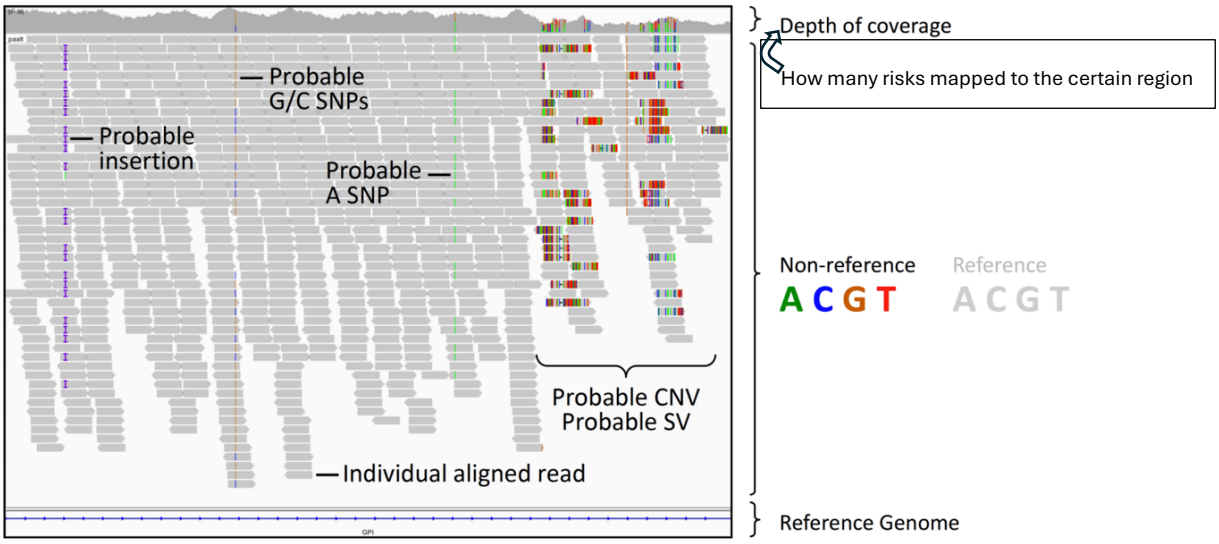
* Errors can creep in on different levels:

- PCR artifacts (amplification of errors)
- Sequencing (errors in base calling) 1% error
- Alignment (misalignment, mis-gapped alignments)
- Variant calling (low depth of coverage, few samples)
- Genotyping (poor annotation)

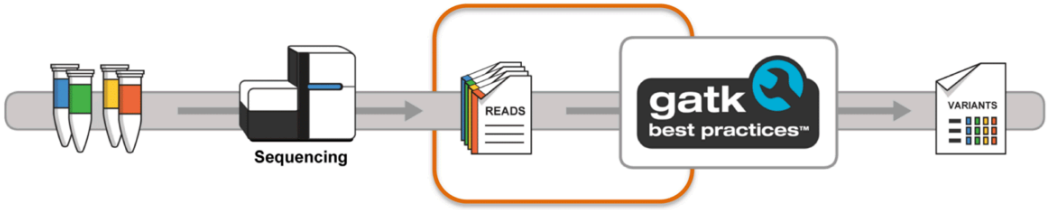
* This situation is more reliable



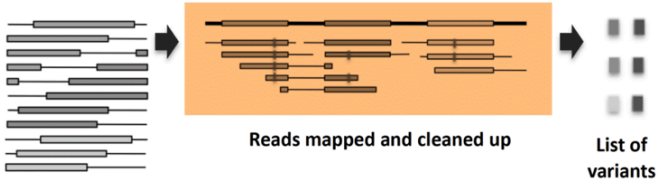
Genome Browser



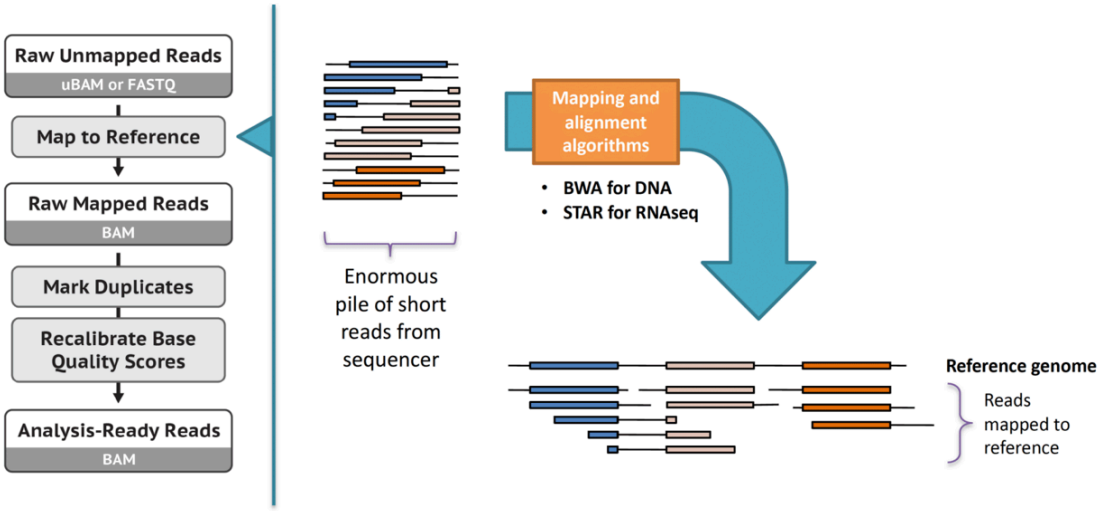
Data pre-processing step



1. Mapping
2. Marking duplicates
3. Base recalibration



SETP 1 : Map the reads produced by the sequence to the reference

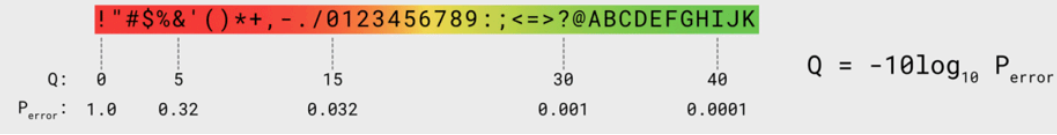


* Input format: FASTQ

FASTQ file sample:

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAACGTCCAGACCCGCGAACGGGTGATGCGGGCCCTGGGCAAACGGTTGCACCCGGATCTGCCCGATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFBFFFFFFFFFF7FFFF<FF
```

Quality scores as ASCII characters:



⇨ P error value => smaller means lower error

* Output format: Sequence/Binary Alignment Map(SAM/BAM)

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

coding

⇨ Binary format is to reduce file size

⇨ Coding applies from 001.fastq

HEADER lines starting with @ symbol describing various metadata for *all* reads

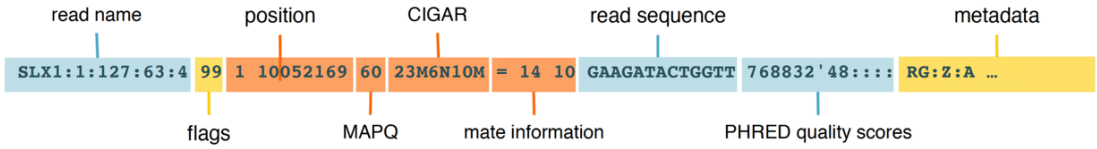
```
@HD VN:1.6 SO:coordinate
@SQ SN:seq1 LN:394893
@SQ SN:seq2 LN:92783
@RG ID:A SM:SAMPLE_A
```

BAM header line

Reference sequence dictionary entries

Read group(s)

RECORDS containing structured read information (1 line per read/record)



- Added mapping info summarizes **position**, **quality**, and **structure** for each **read**
- Mate information points to the read from the other end of the molecule (other in a pair)

Read name = ID

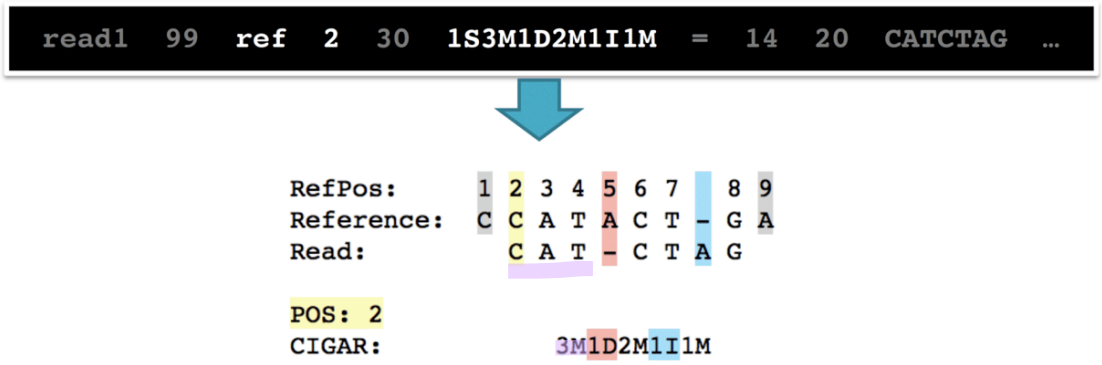
MAPQ = mapping quality

Position = the coordinate with mapping

CIGAR = mapping result

CIGAR summarizes alignment structure

CIGAR = Concise Idiosyncratic Gapped Alignment Report



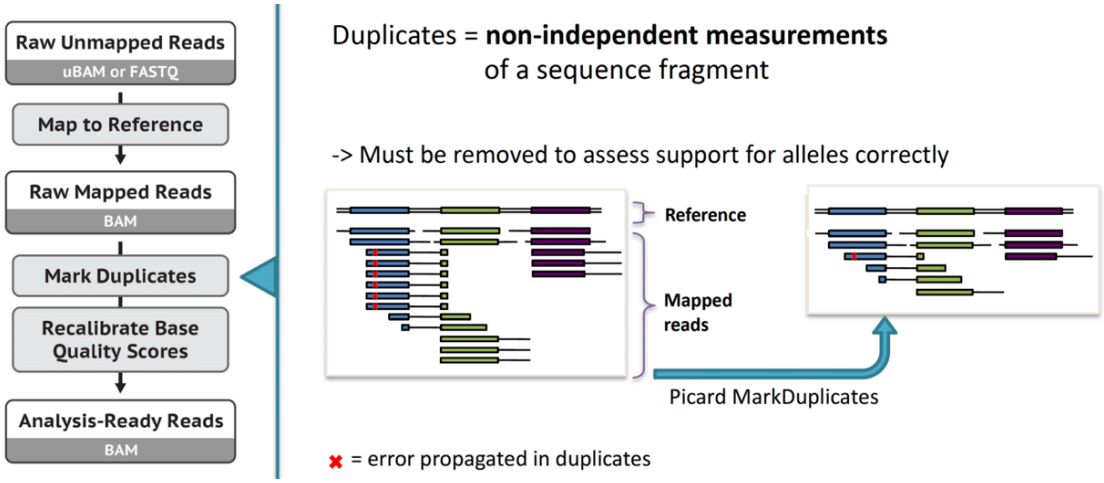
⇨ “Pos :2” means the starting point for the mapping

⇨ “3M” means 3 matched

⇨ “1D” means 1 deletion

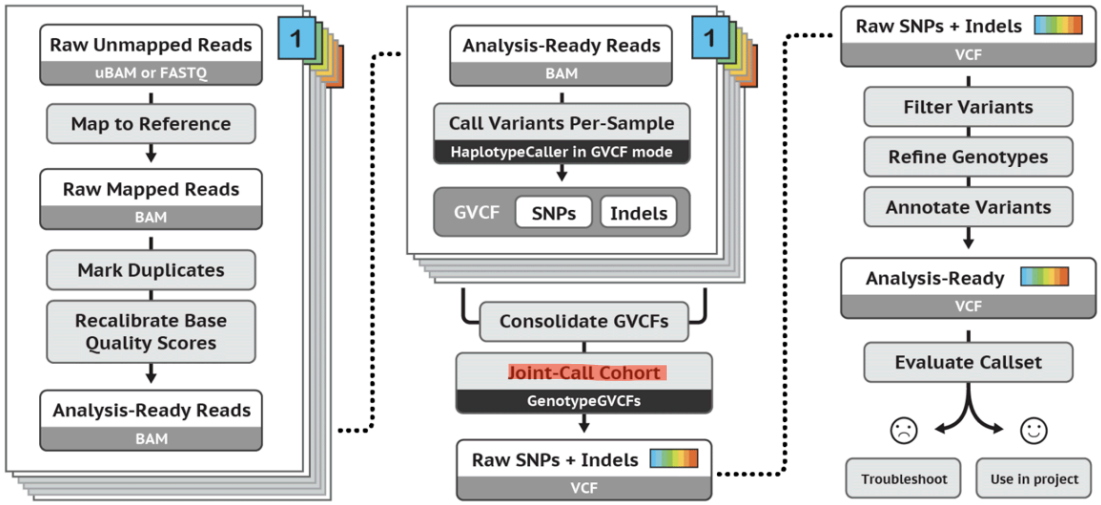
⇨ “1I” means 1 insertion

STEP 2 : Mark duplicates to mitigate duplication artifacts



For having the X, we want to reduce it

STEP 3 : Variant calling in more detail



For the Joint-call Cohort , it is supper powerful

*Variant Call Format (VCF)

HEADER

```
##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">
```

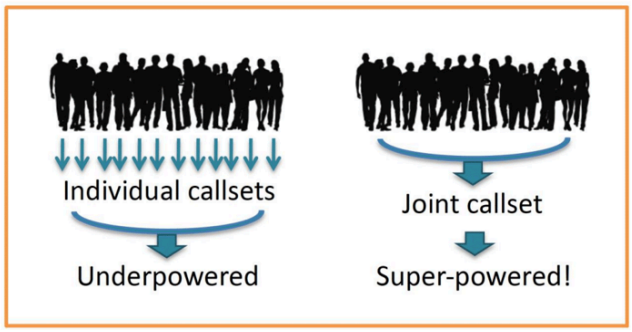
RECORDS

#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

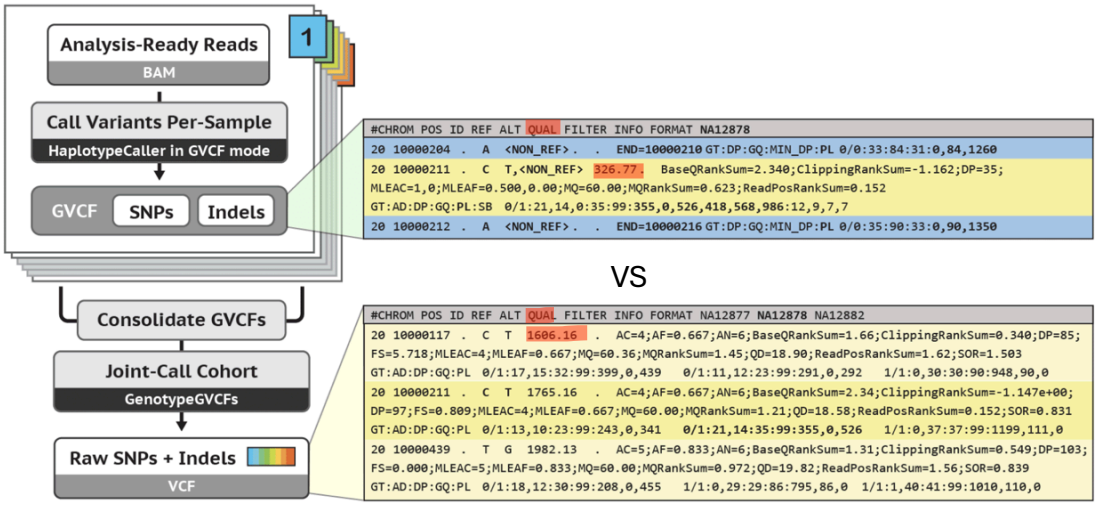
- ⇨ For the “header”, all the record within the file
- ⇨ “Record” refers to each variant
- ⇨ “#CHROM” means chromosome no.
- ⇨ “POS” means position
- ⇨ “REF” means reference

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



From pre-sample GVCFs to final multi-sample VCF



⇨ They have different quality

⇨ Higher quality means more sample which is more credible

Conclusion and significate part

- ❖The pipeline
  - ➢A concrete tool you can use in the future
  - ➢You know what you are expecting from each step. And which file you are looking for
- ❖The file format
  - ➢We talked about reads a lot of time. What are they in the real analysis?
  - ➢It's for practice. We want to avoid the case that you learn a lot but you still cannot resolve real-life problems
  - ➢You know what to input to a specific step. If you get an error, you know what to change
- ❖Trouble-shooting
  - ➢For example, in real-life, you have a nice BAM/SAM file, but your VCF file is empty. Is it because of programming bugs, file formats, or no variants?
  - ➢Hopefully, our introduction to the pipeline will be useful
  - ➢Usefulness is more important than exams
- ❖The reasons that we need to do the steps
  - ➢For example, why we would like to remove the duplicates
- ❖The ability to read the records in those files
  - ➢Given an alignment, you should be able to convert it into a CIGAR string
  - ➢Given a VCF record, you should know what has been changed
- ❖How different factors affect the quality of the mapping and the variant calling
  - ➢Errors VS variants
  - ➢Duplicates
  - ➢Depth/coverage
  - ➢Sequence quality

⇨ Should know the reason why duplicate