Variant Calling

CHRIS FIELDS

MAYO-ILLINOIS COMPUTATIONAL GENOMICS WORKSHOP,
JUNE 8, 2022

Overview

Variant calling and use cases

Errors vs. actual variants

Experimental design (GATK focus)

Small variant (SNV/Small Indel) analysis

- GATK Pipeline
- Formats encountered within

Structural Variation Analysis (SV)

Association analysis (briefly)

Variant Calling

As the name implies, we’re looking for differences (variations)

- **Reference** – reference genome (hg38, GRCh38)
- **Sample(s)** – one or more comparative samples

Start with raw sequence data

End with a human (or other organism) ‘diff’ file, recording the variants

Additional information added downstream:

- Filters (quality of the calls)
- Functional annotation

Variations

Difference between 2 individuals : 1 every 1000 bp

- ~ 2.7 million differences

Small (<50 bp)

- SNV – single nucleotide (`**SNPs**`)
- Small insertions or deletions (‘**Indels**’)

Large (structural variations)

- Indels > 50 bp
- Copy Number Variations
- Inversions
- Translocations
- Chromosomal fusions

Variations

Mainly focus on diploid organisms

- Human:
  - 22 pairs of autosomal chromosomes
    - One from mother, one from father
  - 2 sex chromosomes (female XX, male XY)
    - One from mother, one from father (where does Y come from for male offspring)
  - Mitochondrial genome (generally maternally inherited)
    - 100-10,000 copies per cell

Variation can be in

- One chromosome (heterozygous, or ‘het’)
- Both chromosomes (homozygous, or ‘hom’)

Use cases

Medicine

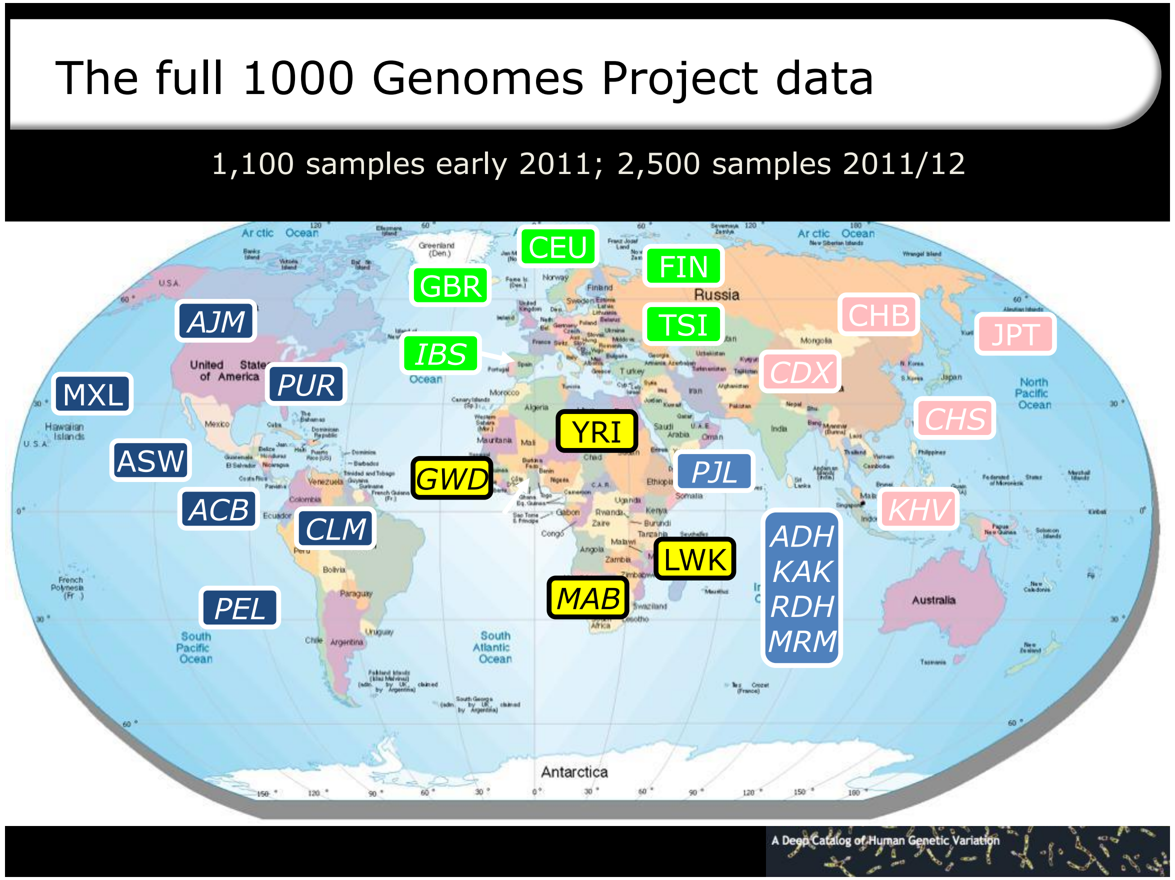
- Hereditary or genetic diseases, genetic predisposition to disease
- Normal vs. tumor analyses
- Heteroplasmy

Population genetics

- GWAS

Population genetics

The 1000 Genomes Project

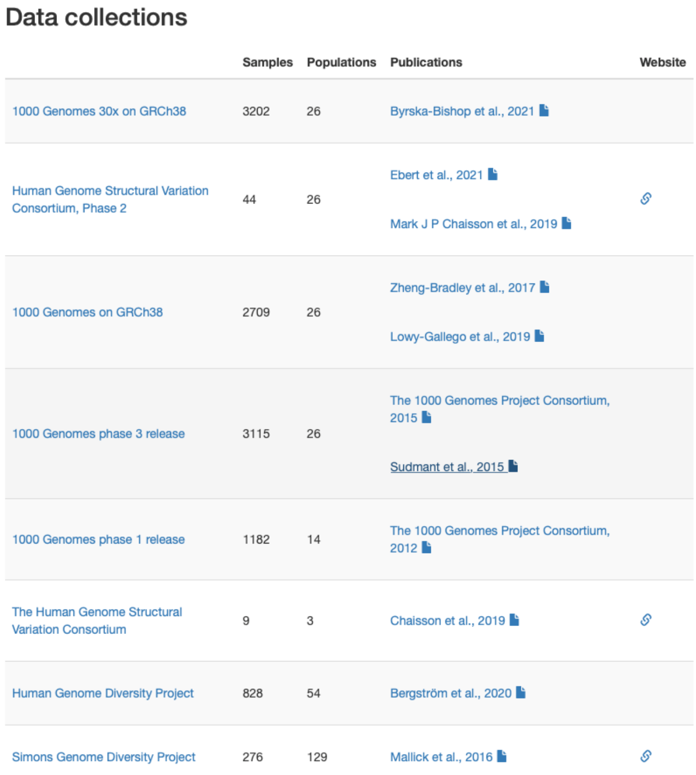


The full 1000 Genomes Project data

1,100 samples early 2011; 2,500 samples 2011/12

A Deep Catalog of Human Genetic Variation

[IGSR: The International Genome Sample Resource]

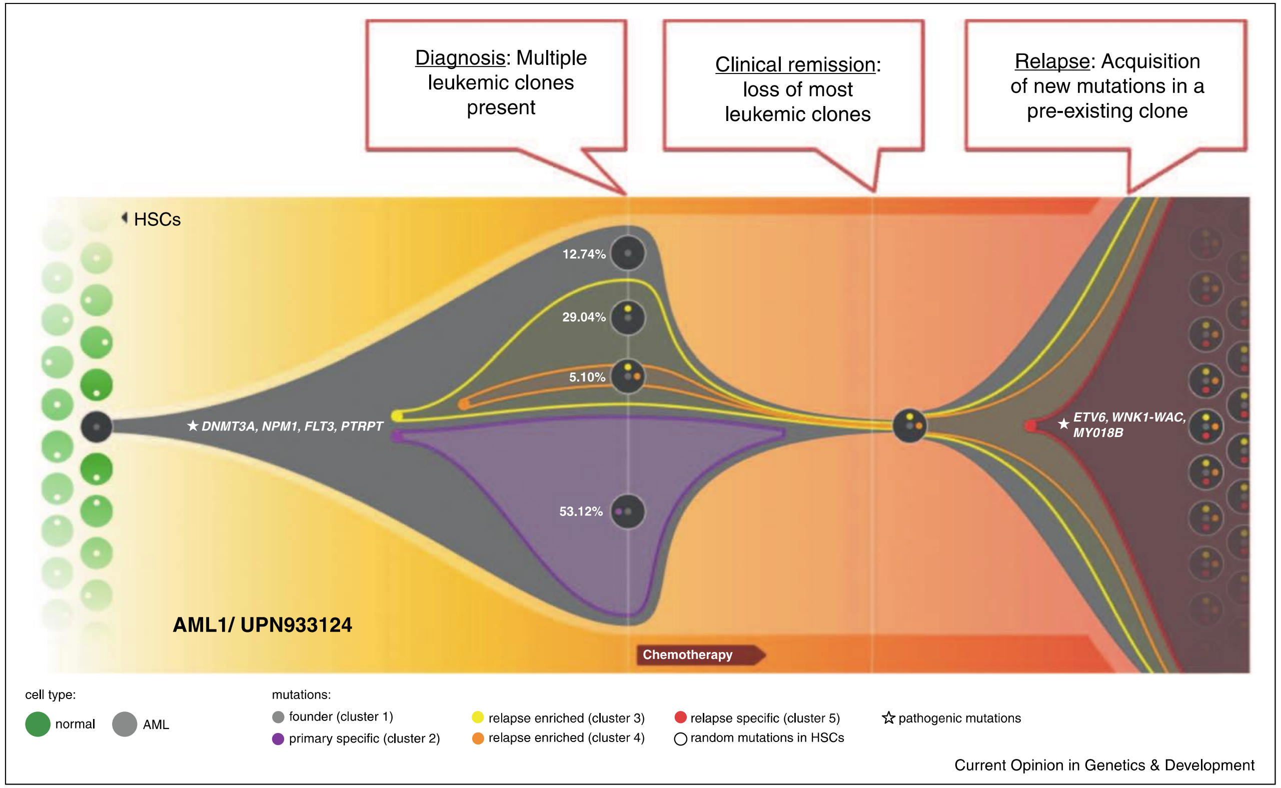


Data collections

	Samples	Populations	Publications	Website
1000 Genomes 30x on GRCh38	3202	26	Byrska-Bishop et al., 2021	
Human Genome Structural Variation Consortium, Phase 2	44	26	Ebert et al., 2021 Mark J P Chaisson et al., 2019	🔗
1000 Genomes on GRCh38	2709	26	Zheng-Bradley et al., 2017 Lowy-Gallego et al., 2019	
1000 Genomes phase 3 release	3115	26	The 1000 Genomes Project Consortium, 2015 Sudmant et al., 2015	
1000 Genomes phase 1 release	1182	14	The 1000 Genomes Project Consortium, 2012	
The Human Genome Structural Variation Consortium	9	3	Chaisson et al., 2019	🔗
Human Genome Diversity Project	828	54	Bergström et al., 2020	
Simons Genome Diversity Project	276	129	Mallick et al., 2016	🔗

Cancer

Figure 2



Diagnosis: Multiple leukemic clones present

Clinical remission: loss of most leukemic clones

Relapse: Acquisition of new mutations in a pre-existing clone

Current Opinion in Genetics & Development

Model of the clonal progression process that occurs between the initial (*de novo*) and relapse presentation in AML patients. At diagnosis, this patient has an oligoclonal disease characterized by four different subclones, each present at a specific proportion in the tumor cell population and with a specific mutational profile. Chemotherapy used to induce the patient into remission decreases clonal heterogeneity but a single subclone persists, acquires new mutations, and again proliferates in the bone marrow as a relapse-specific subclone.

E. Mardis, Current Opinion in Genetics & Development 2012, 22:245–250

Variants vs. Errors

Must distinguish between actual **variation** (real change) and **errors** (artifacts) introduced into the analysis

Errors can creep in on various levels:

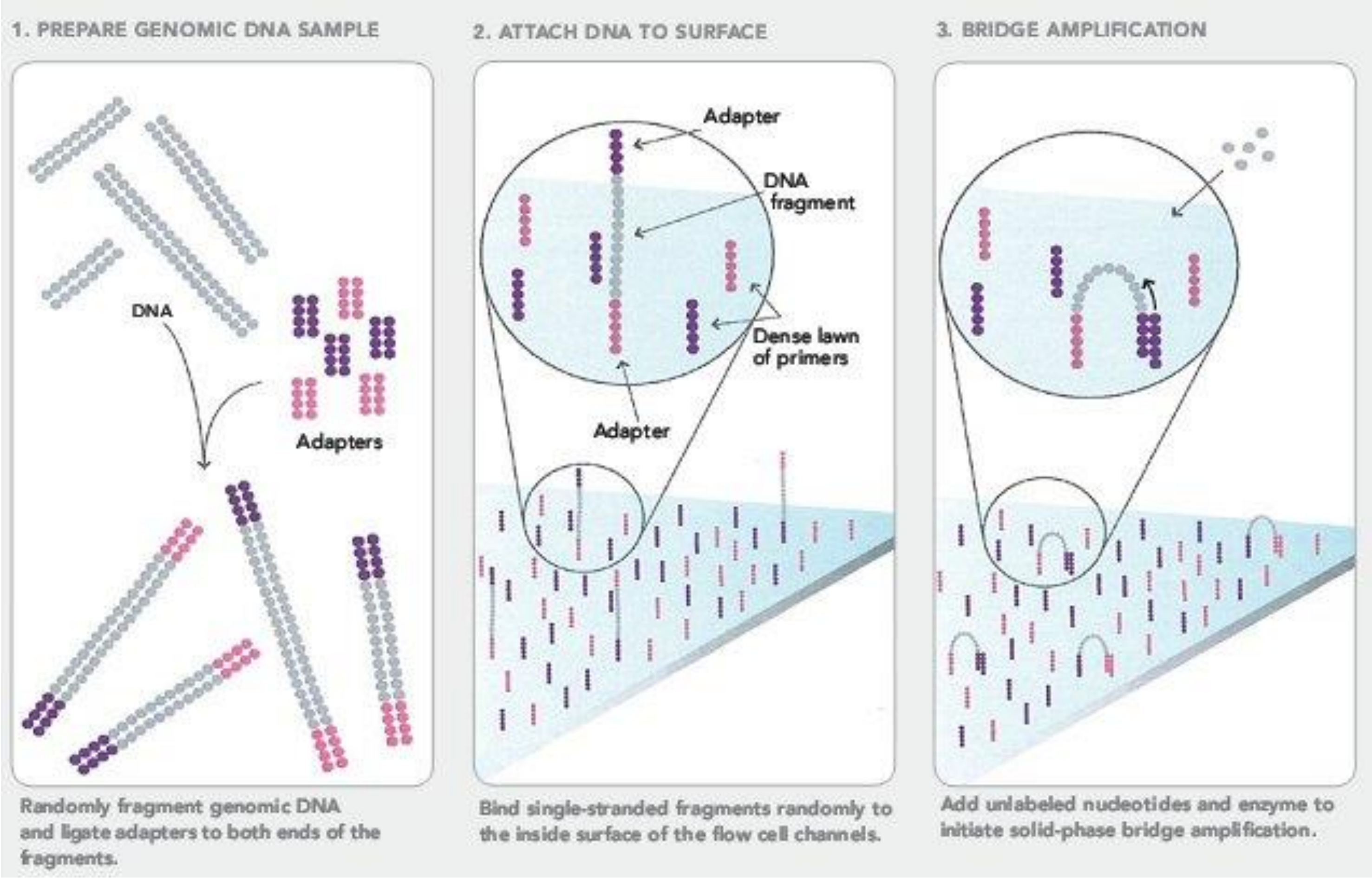
- **PCR artifacts** (amplification of errors)
- **Sequencing** (errors in base calling)
- **Alignment** (misalignment, mis-gapped alignments)
- **Variant calling** (low depth of coverage, few samples)
- **Genotyping** (poor annotation)

Try to control for these when possible to **reduce false positives** w/o incurring (worse) false negatives

How do sequencing errors occur?

Illumina Sequencing

Video!

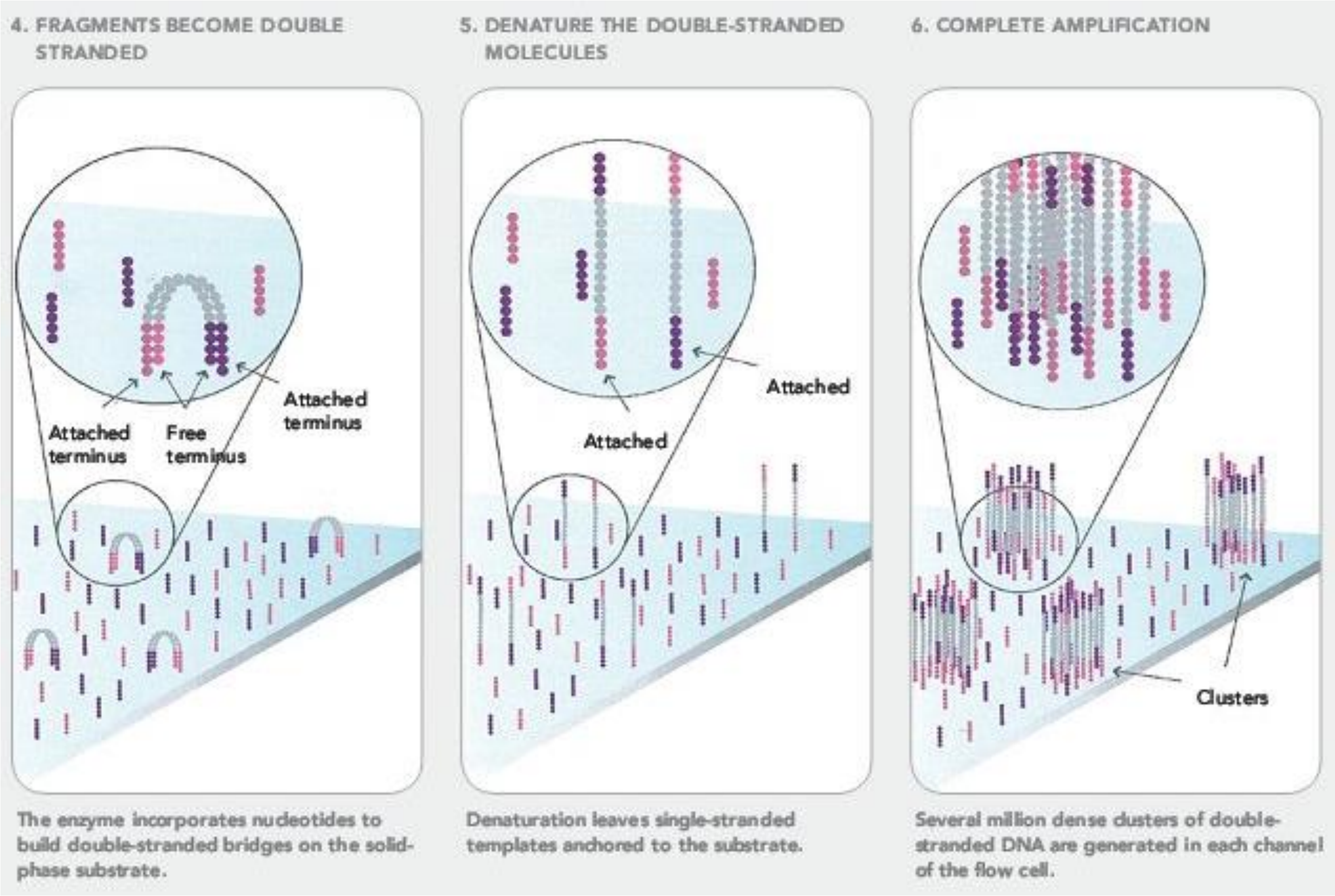


Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Illumina Sequencing



4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

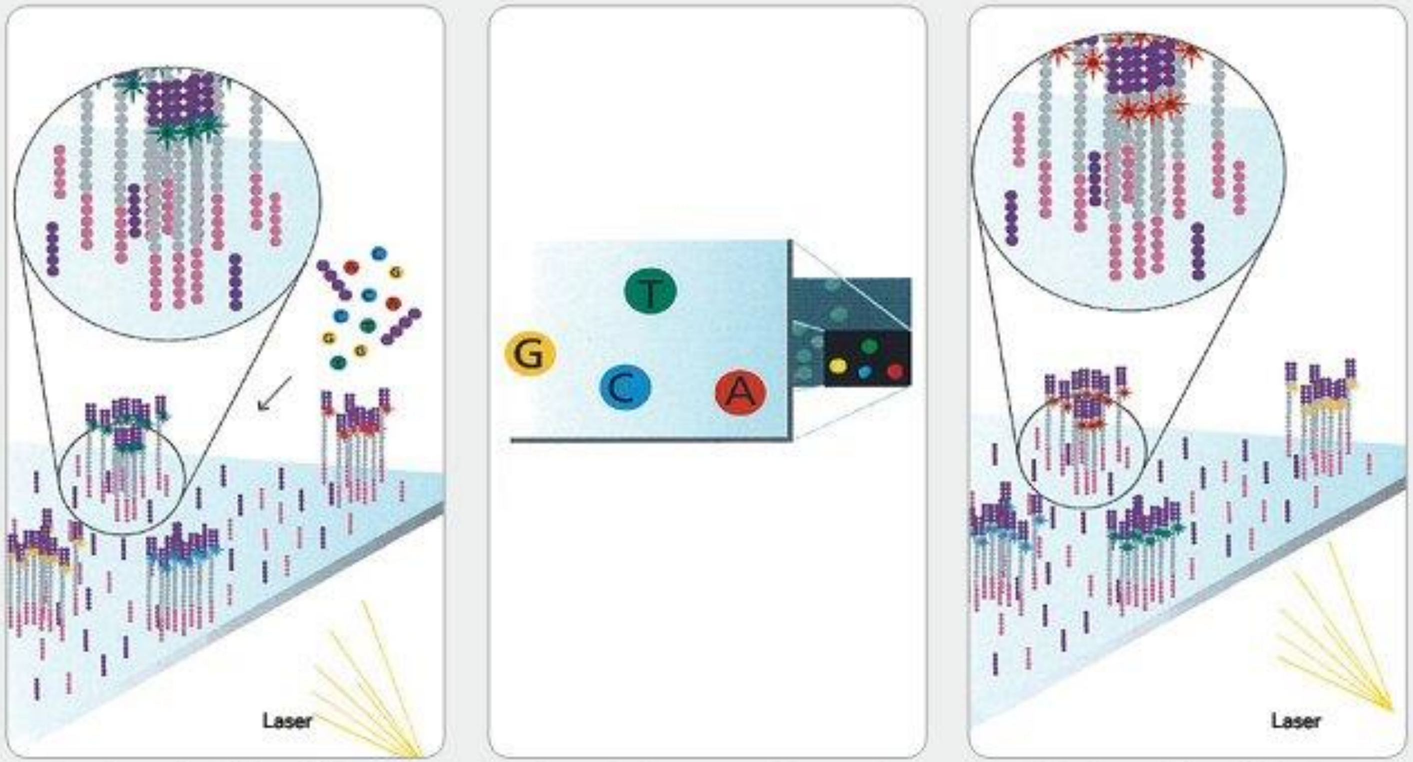
5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Illumina Sequencing



7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

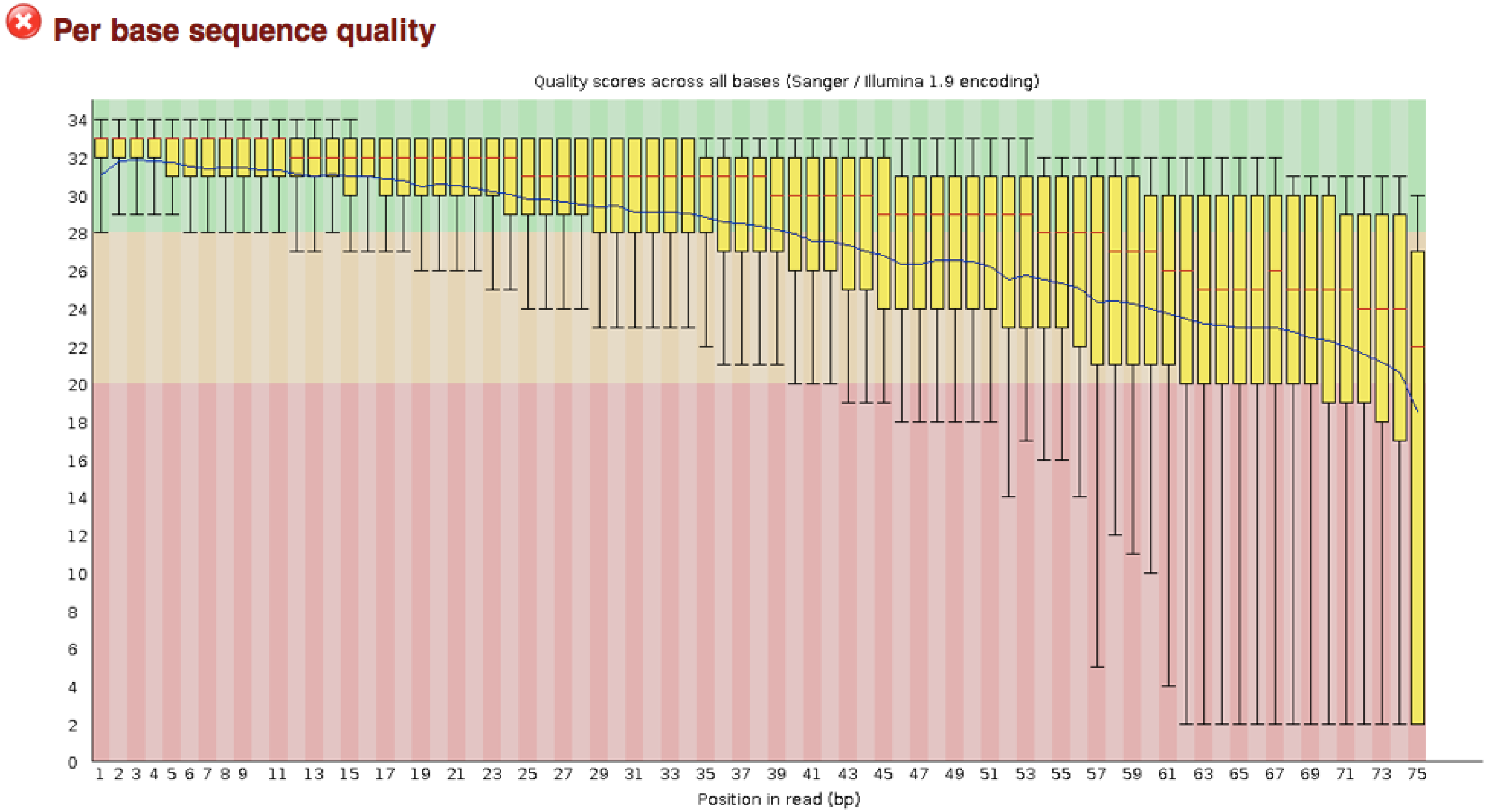
8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

Check sequence data!



Per base sequence quality

Sequence quality

Different technologies have different errors, error rates

- **Illumina** – substitution errors (0.1%)
- **PacBio and Oxford Nanopore** – (10-15%) Indels, primarily around homopolymer track errors

Represented as a quality score (Phred)

- Q = -10log10(e)

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.90%
40	1 in 10000	99.99%
50	1 in 100000	~100.00%

Formats: FASTQ – ‘sequence with quality’

```
@HWI-ST1155:109:D0L23ACXX:5:1101:2247:1985 1:N:0:GCCAAT
NTTCCTTTGACAAATATTAAAATTAAGAATCAAATATGGTAGTGTATGCCAAGACCTAGTCTGAGTCAGTAGGAT
+
#1=DDFFFHHHHHJJJJJJIJJJIJJIJIJJJJJJIJI?FHFHEIJEIIIEGFFHHGIGHIJEIFGIJHGDIII
```

Three ‘variants’ – Sanger, Illumina, Solexa (Sanger is most common)

May be ‘raw’ data (straight from seq pipeline) or processed (trimmed for various reasons)

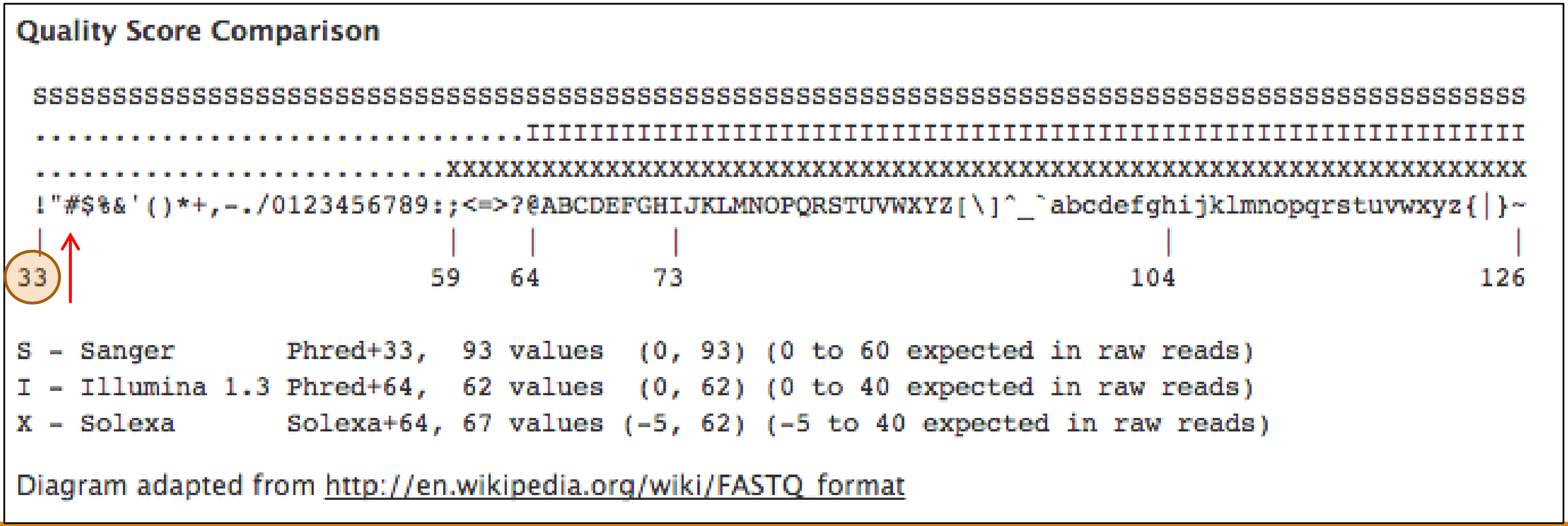
Can hold 100’s of millions of records **per sample**

Files can be very large (100’s of GB) apiece

Formats: FASTQ – ‘sequence with quality’

```
@HWI-ST1155:109:D0L23ACXX:5:1101:2247:1985 1:N:0:GCCAAT
NTTCCTTTGACAAATATTAAAATTAAGAATCAAATATGGTAGTGTATGCCAAGACCTAGTCTGAGTCAGTAGGAT
+
#1=DDFFFHHHHHJJJJJIJJJIJJIJIJJJJJJJIJI?FHFHEIJEIIIEGFFHHGIGHIJEIFGIJHGDIII
```

Very low Phred score, less than 10 (35 - 33 = 2)



Quality Score Comparison

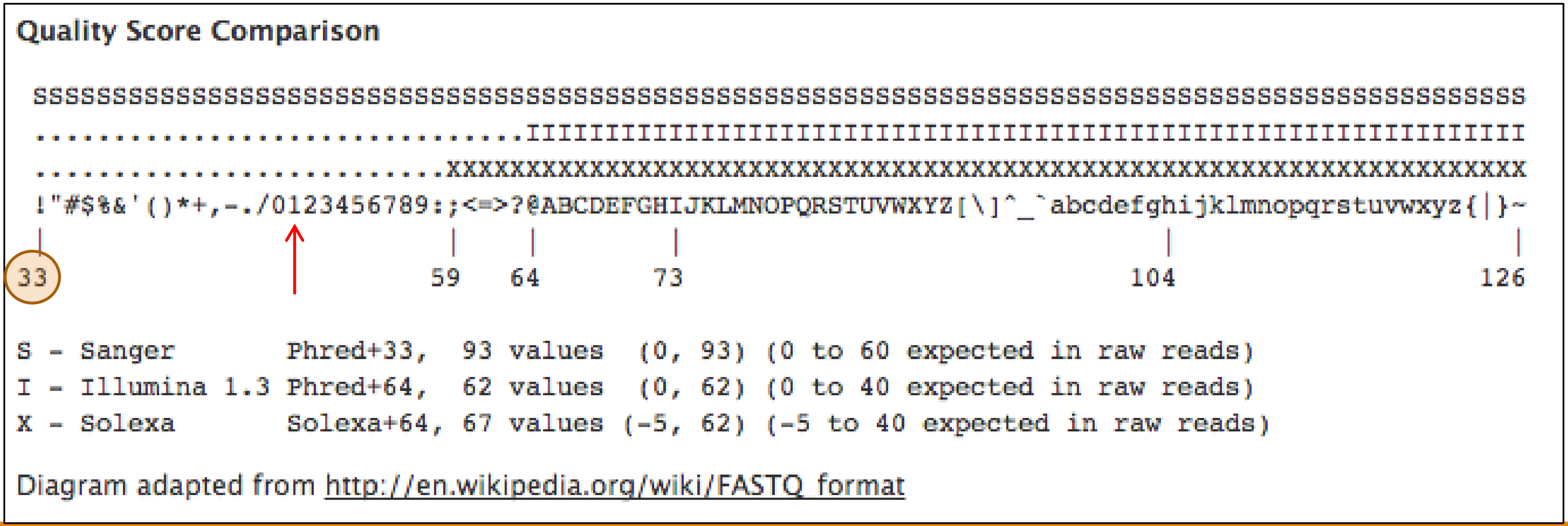
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                       59   64       73                            104                   126

S - Sanger        Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3  Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa        Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

Diagram adapted from http://en.wikipedia.org/wiki/FASTQ_format

Formats: FASTQ – ‘sequence with quality’

```
@HWI-ST1155:109:D0L23ACXX:5:1101:2247:1985 1:N:0:GCCAAT
NTTCCTTTGACAAATATTAAAATTAAGAATCAAATATGGTAGTGTATGCCAAGACCTAGTCTGAGTCAGTAGGAT
+
#1=DDFFFHHHHHJJJJJIJJJIJJIJIJJJJJJJIJI?FHFHEIJEIIIEGFFHHGIGHIJEIFGIJHGDIII
```

Low Phred score, < 20 (49– 33 = 16)



Quality Score Comparison

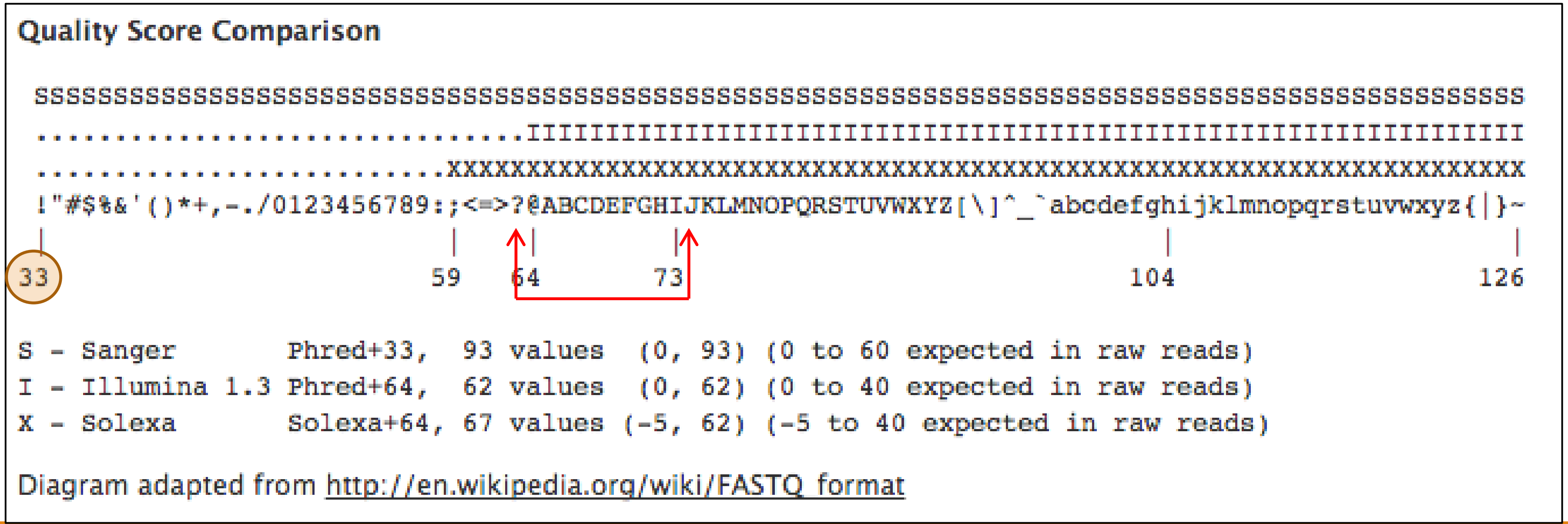
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                             104                  126

S - Sanger        Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3  Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa        Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

Diagram adapted from http://en.wikipedia.org/wiki/FASTQ_format

Formats: FASTQ – ‘sequence with quality’

```
@HWI-ST1155:109:D0L23ACXX:5:1101:2247:1985 1:N:0:GCCAAT
NTTCCTTTGACAAATATTAAAATTAAGAATCAAATATGGTAGTGTATGCCAAGACCTAGTCTGAGTCAGTAGGAT
+
#1=DDFFFHHHHHJJJJJIJJJIJJIJIJJJJJJJIJI?FHFHEIJEIIIEGFFHHGIGHIJEIFGIJHGDIII
```

High quality reads, Phred score >= 30 (63 – 33 = 30)



Quality Score Comparison

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                       59    64      73                            104                   126

S - Sanger        Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3  Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa        Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

Diagram adapted from http://en.wikipedia.org/wiki/FASTQ_format

Basic Experimental Design

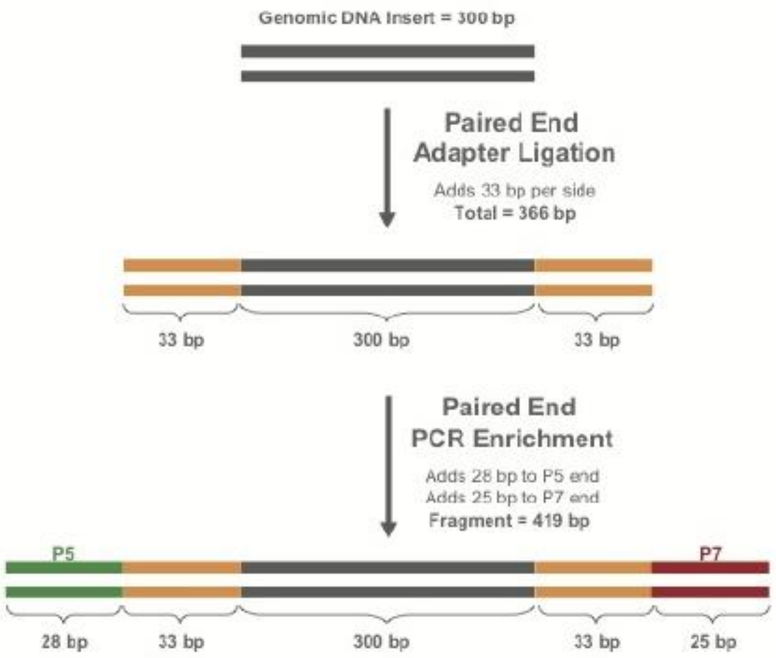
Terminology

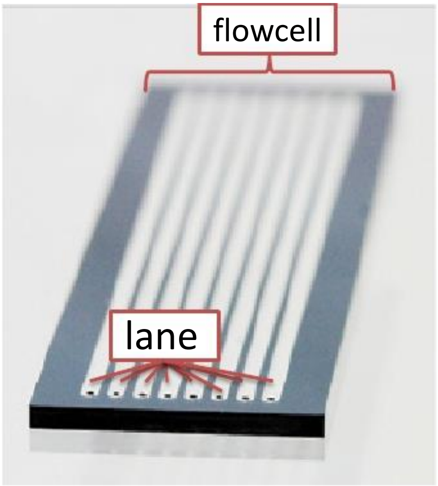
Lane – Physical sequencing lane

Library – Unit of DNA prep pooled together

Sample – Single individual

Cohort – Collection of samples analyzed together



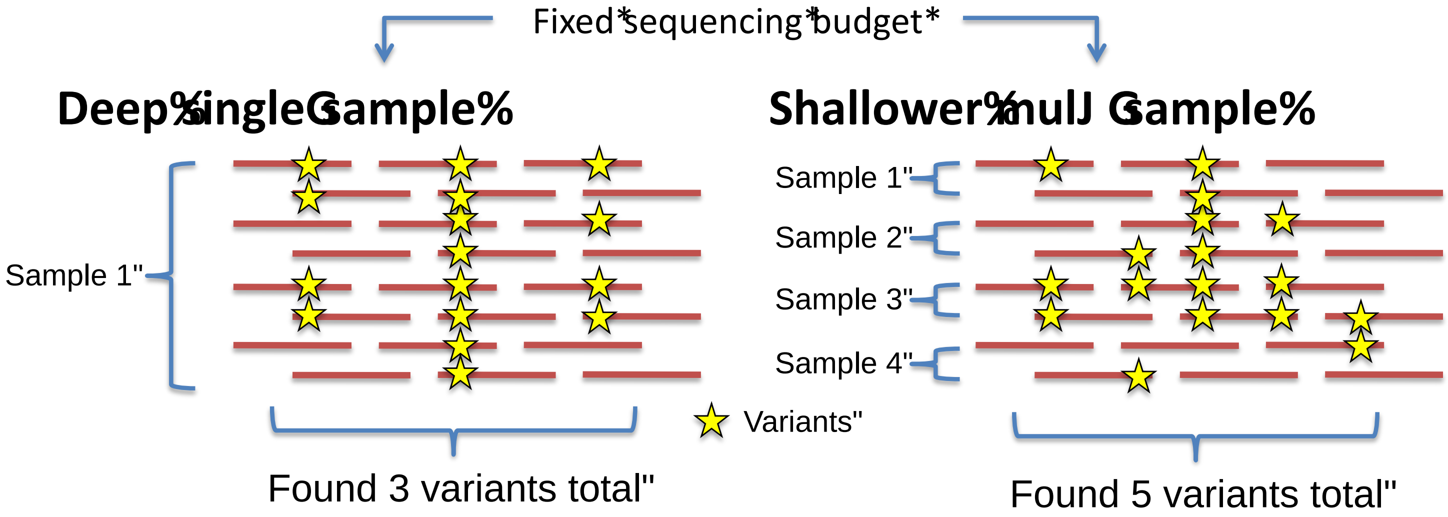


Terminology

WGS vs Exome Capture

- **Whole genome sequencing** – everything
  - High cost if per sample is deep sequence (>25-30x)
  - Can run multisample low coverage samples
- **Exome capture** – targeted sequencing (1-5% of genome)
  - Deeper coverage of transcribed regions
  - Miss other important non-coding regions (promoters, introns, enhancers, small RNA, etc)

Single*vs.*mul) 6sample*analysis*



Fixed*sequencing*budget*

Deep%singleGsample%

Shallower%mulJ Gsample%

Sample 1"

Sample 1"

Sample 2"

Sample 3"

Sample 4"

☆ Variants"

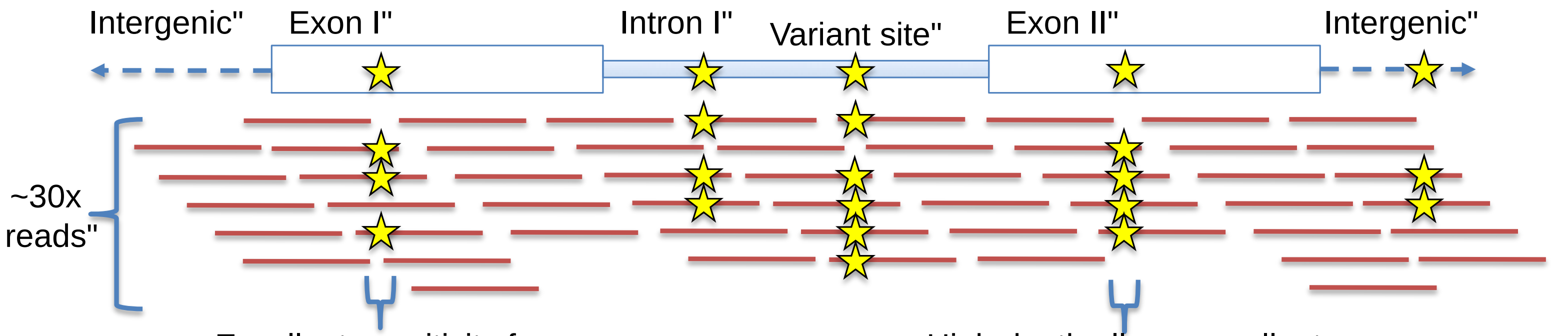
Found 3 variants total"

Found 5 variants total"

- Higher*sensi) vity*for*variants* in*the*sample*
- More*accurate*genotyping*per* sample*
- Cost:*no*informa) on*about* other*samples*

- Sensi) vity*dependent*on* frequency*of*varia) on*
- Worse*genotyping*
- More*total*variants* discovered*

High6pass*sequencing*design*



Excellent sensitivity for hetero- and homozygotes"

High depth allows excellent genotype calling"

Data requirements per sample

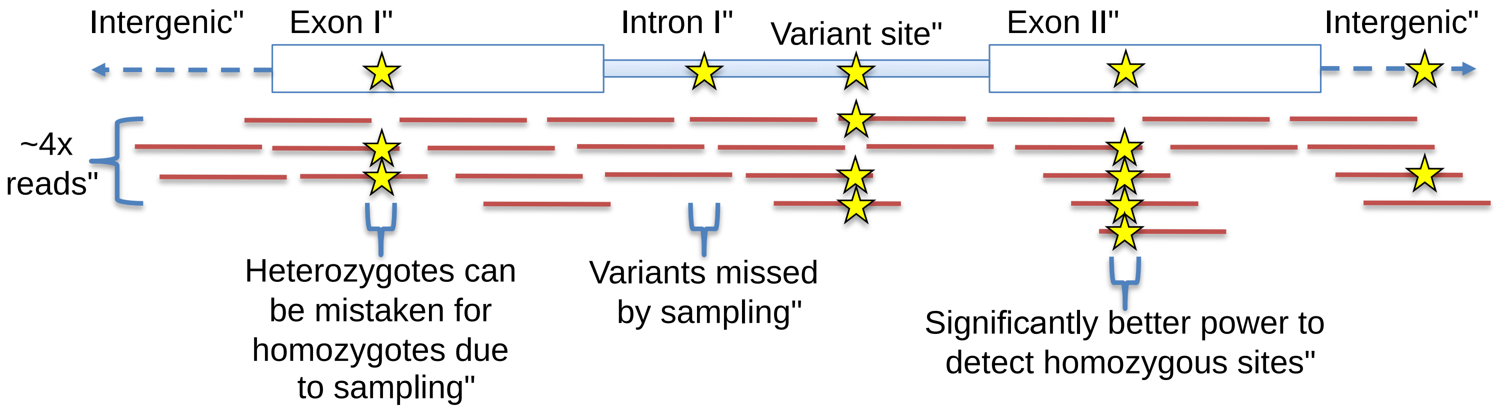
Target bases	3 Gb
Coverage	Avg. 30x
# sequenced bases	100 Gb
# per lane (HiSeq 4000)	~1
# per lane (NovaSeq, S4)	~8-9

Variant detection among multiple samples

Variants found per sample	~3-5M
Percent of variation in genome	>99%
Pr{singleton discovery}	>99%
Pr{common allele discovery}	>99%

NovaSeq 6000 = 750-850 Gb/lane (2x150nt, S4 lane)

Low6pass*sequencing*design*



Intergenic"

Exon I"

Intron I"

Variant site"

Exon II"

Intergenic"

~4x reads"

Heterozygotes can be mistaken for homozygotes due to sampling"

Variants missed by sampling"

Significantly better power to detect homozygous sites"

Data requirements per sample

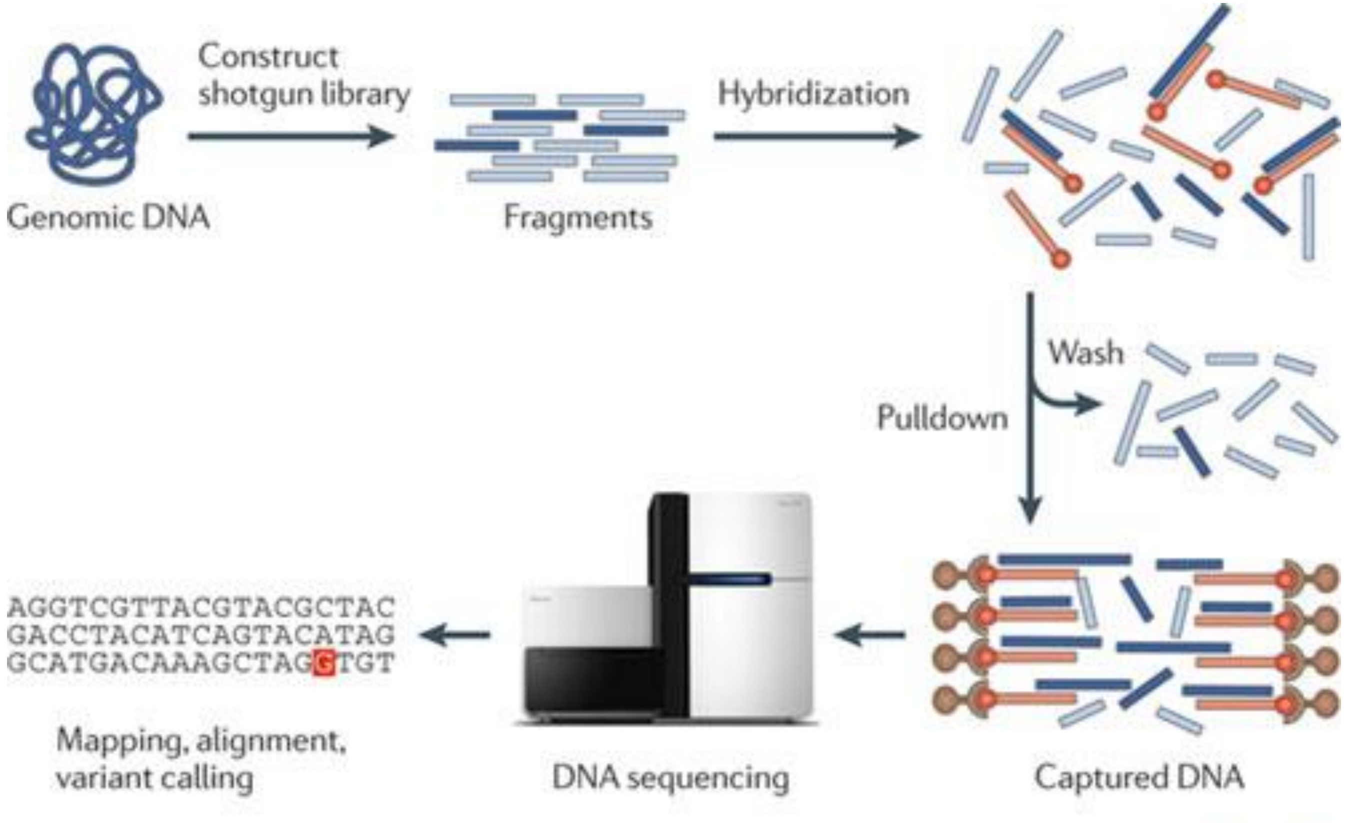
Target bases	3 Gb
Coverage	Avg. 5x
# sequenced bases	15 Gb
# per lane (HiSeq 4000)	~6
# per lane (NovaSeq, S4)	~50

Variant detection among multiple samples

Variants found per sample	~3M
Percent of variation in genome	~90%
Pr{singleton discovery}	< 50%
Pr{common allele discovery}	~99%

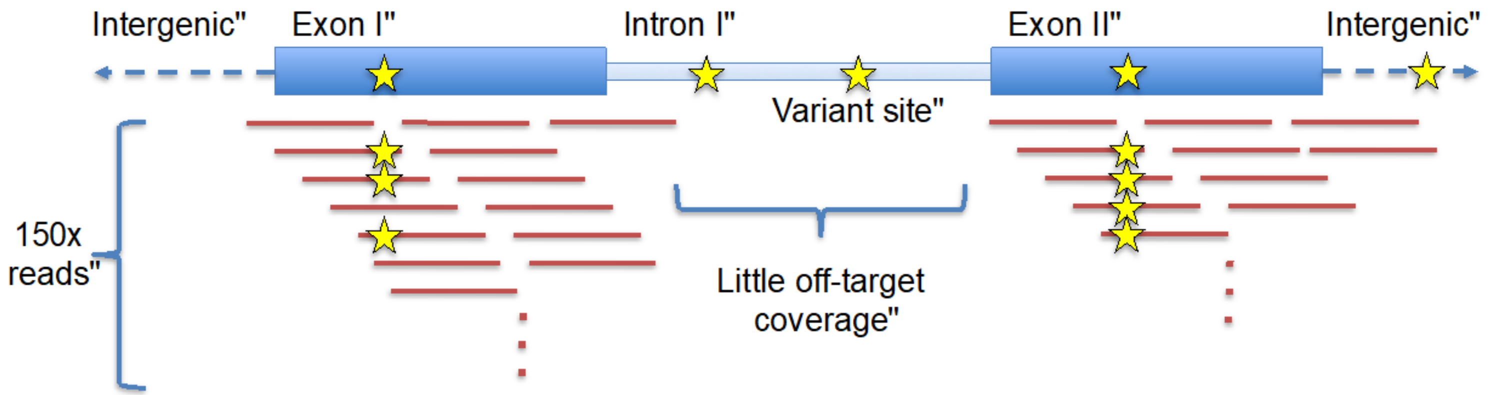
NovaSeq 6000 = 750-850 Gb/lane (2x150nt, S4 lane)

Exome Capture



Nature Reviews | Genetics

Exome*capture*sequencing*design*



Based on Illumina 'DNA Prep with Enrichment' panel

Data requirements per sample

Target bases	45-60 Mb
Coverage	>90% 20x*
# sequenced bases	4 Gb
# per lane (HiSeq 4000)	20
# per lane (NovaSeq, S4)	~384*

Variant detection among multiple samples

Variants found per sample	~25-45k
Percent of variation in genome	0.005
Pr{singleton discovery}	~95%
Pr{common allele discovery}	~95%

NovaSeq 6000 = 750-850 Gb/lane (2x150nt, S4 lane)

General variant calling pipelines

Common pattern:

- Align reads
- Optimize alignment
- Call variants
- Filter called variants
- Annotate

Tool/Workflow examples

Examples (standard variant calling)

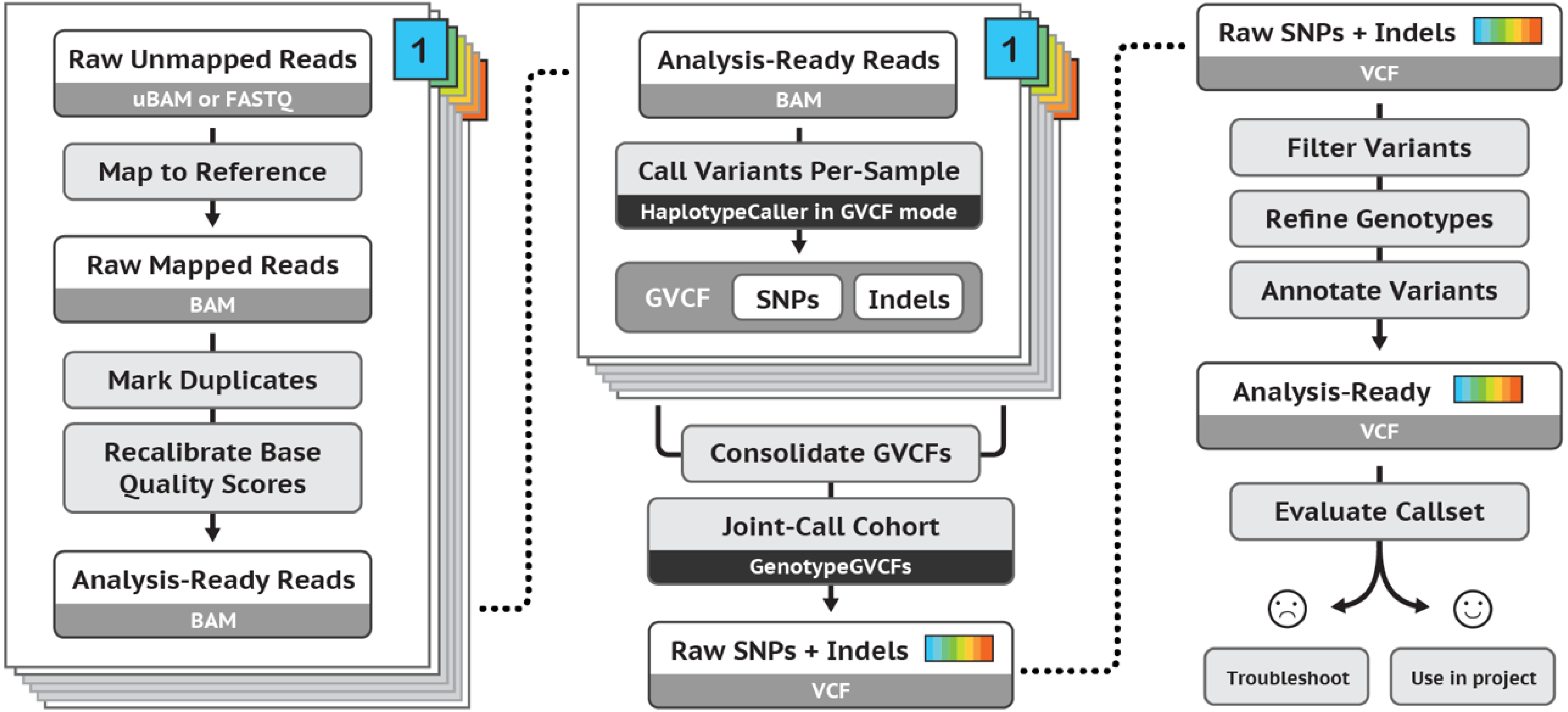
- ***Genome Analysis Toolkit (GATK)***
- samtools mpileup
- VarScan2
- freeBayes
- Commercial
  - Illumina DRAGEN – GATK using FPGA
  - Sentieon – accelerated CPU
  - Parabricks – GPU-based

Tool/Workflow examples

Specialized purposes

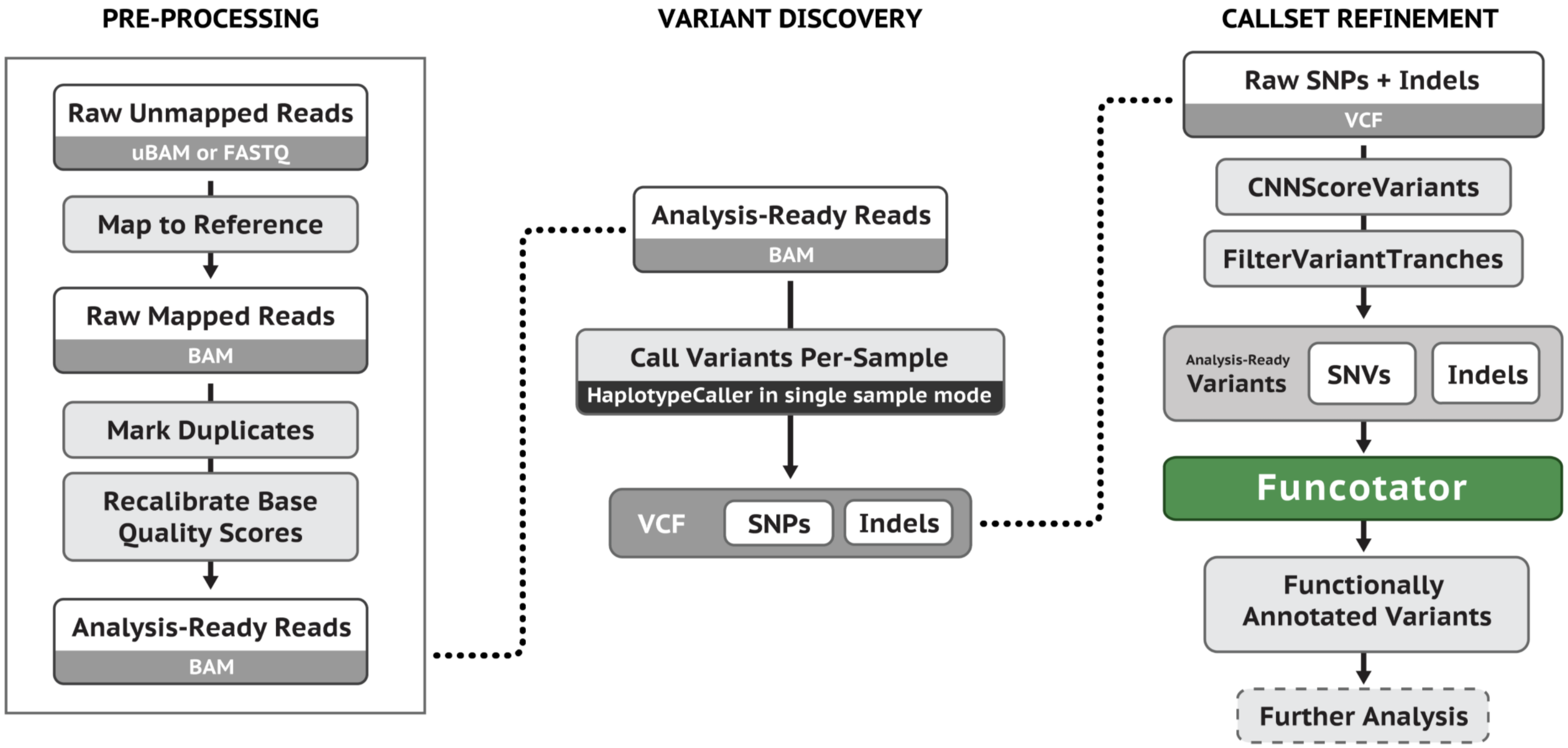
- Copy number variation and structural variation
- Cancer (tumor sample analyses)
- RNA-Seq

GATK – Multi-sample Germline Calls



GATK Germline Calling

GATK – Single Sample Germline Calls



PRE-PROCESSING

Raw Unmapped Reads
uBAM or FASTQ

Map to Reference

Raw Mapped Reads
BAM

Mark Duplicates

Recalibrate Base Quality Scores

Analysis-Ready Reads
BAM

VARIANT DISCOVERY

Analysis-Ready Reads
BAM

Call Variants Per-Sample
HaplotypeCaller in single sample mode

VCF SNPs Indels

CALLSET REFINEMENT

Raw SNPs + Indels
VCF

CNNScoreVariants

FilterVariantTranches

Analysis-Ready Variants SNVs Indels

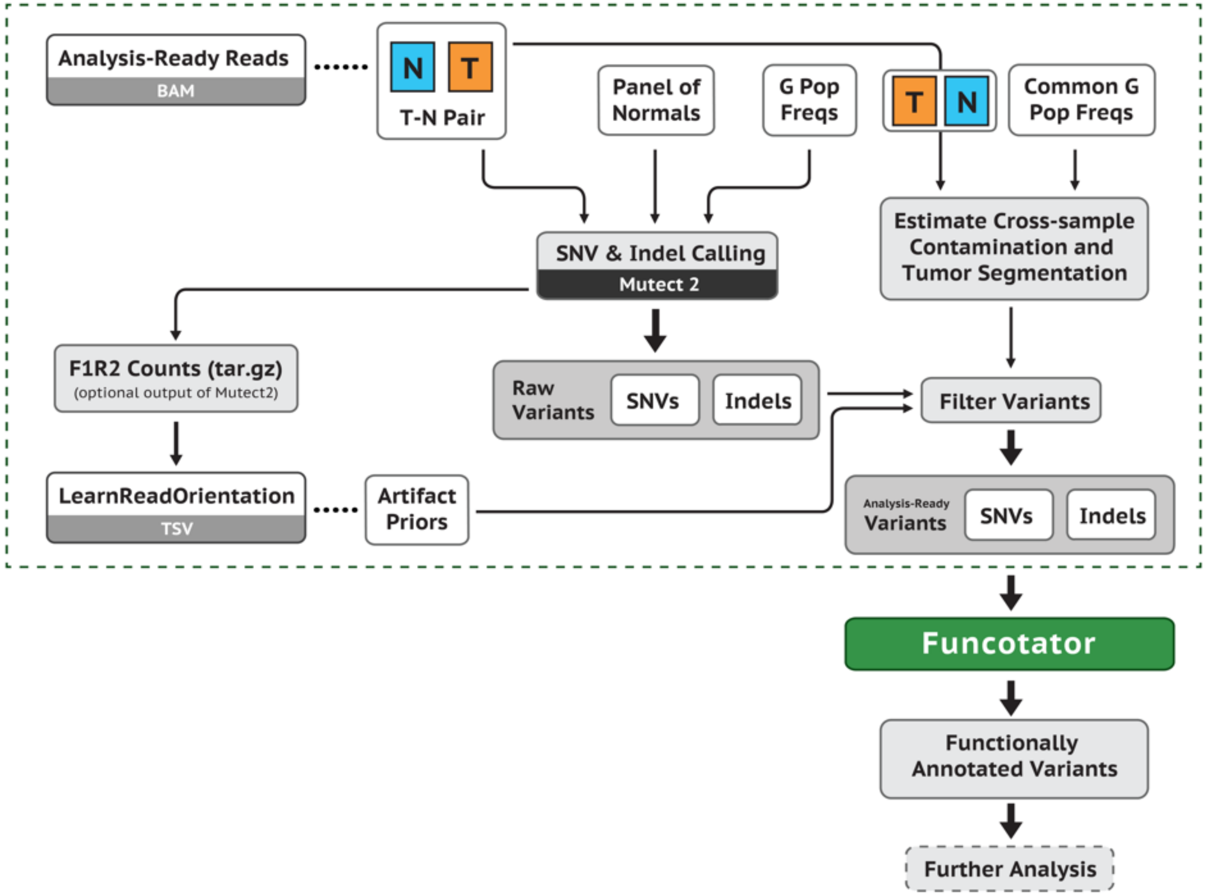
Funcotator

Functionally Annotated Variants

Further Analysis

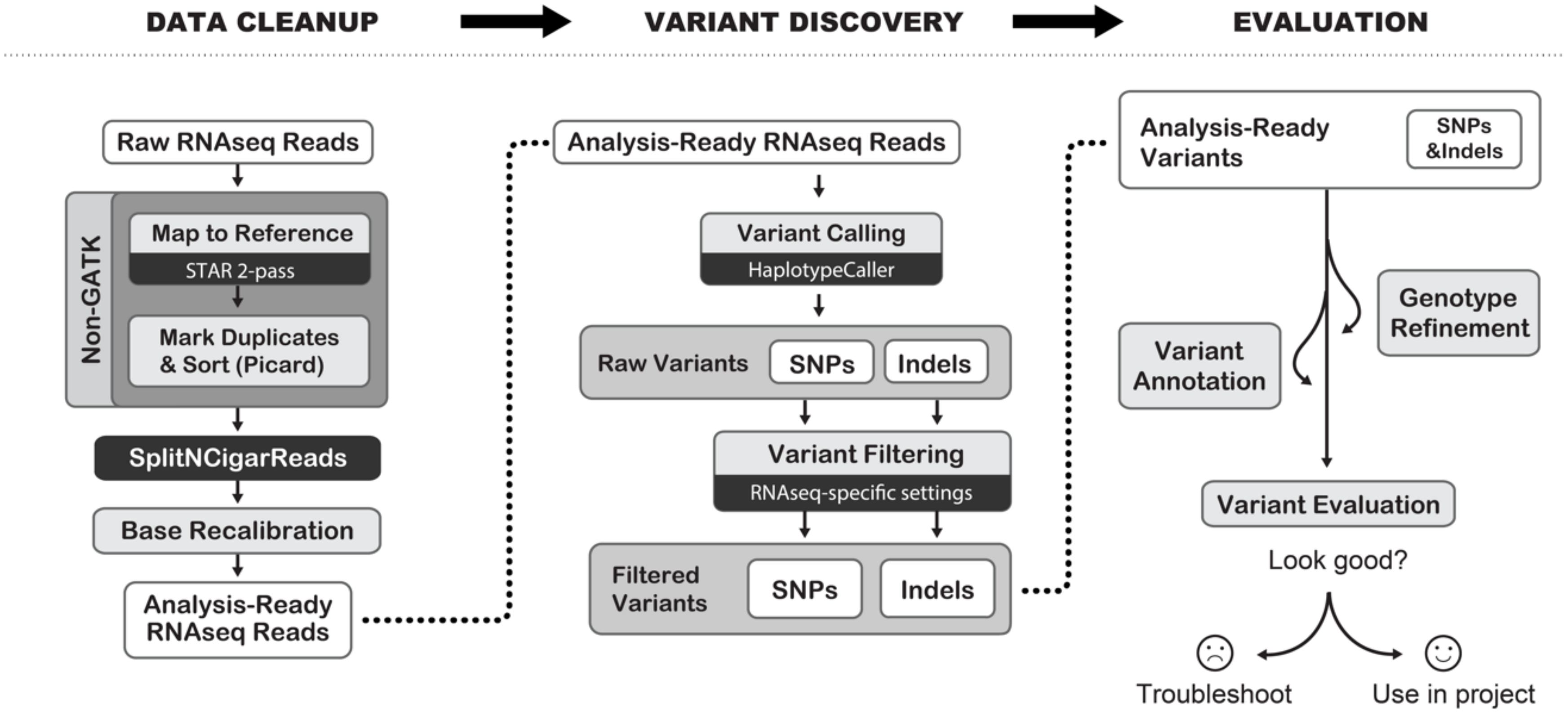
[GATK Germline Calling]

GATK – Somatic Calls (Tumor)



GATK Somatic Callling

GATK – RNA-Seq

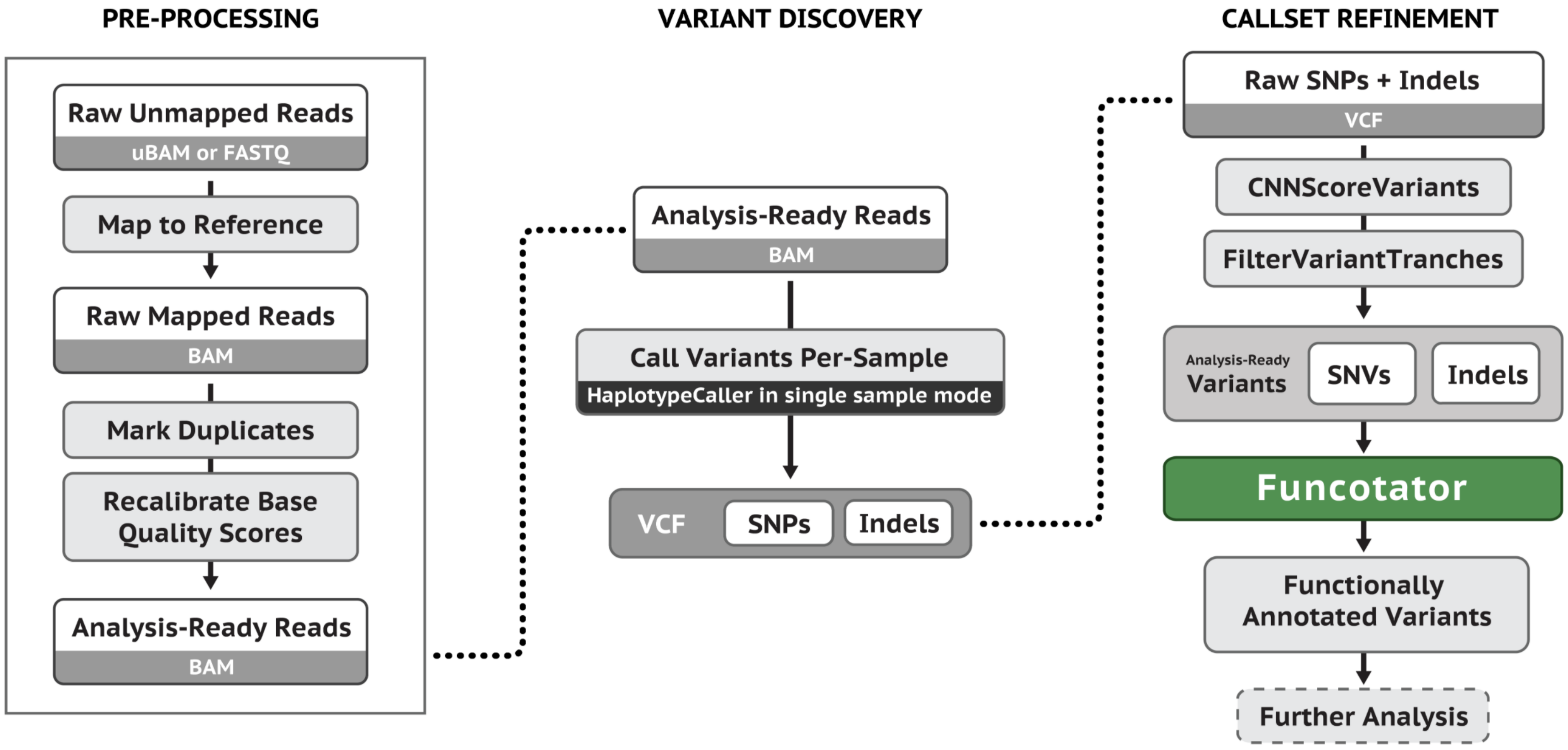


GATK RNA-Seq

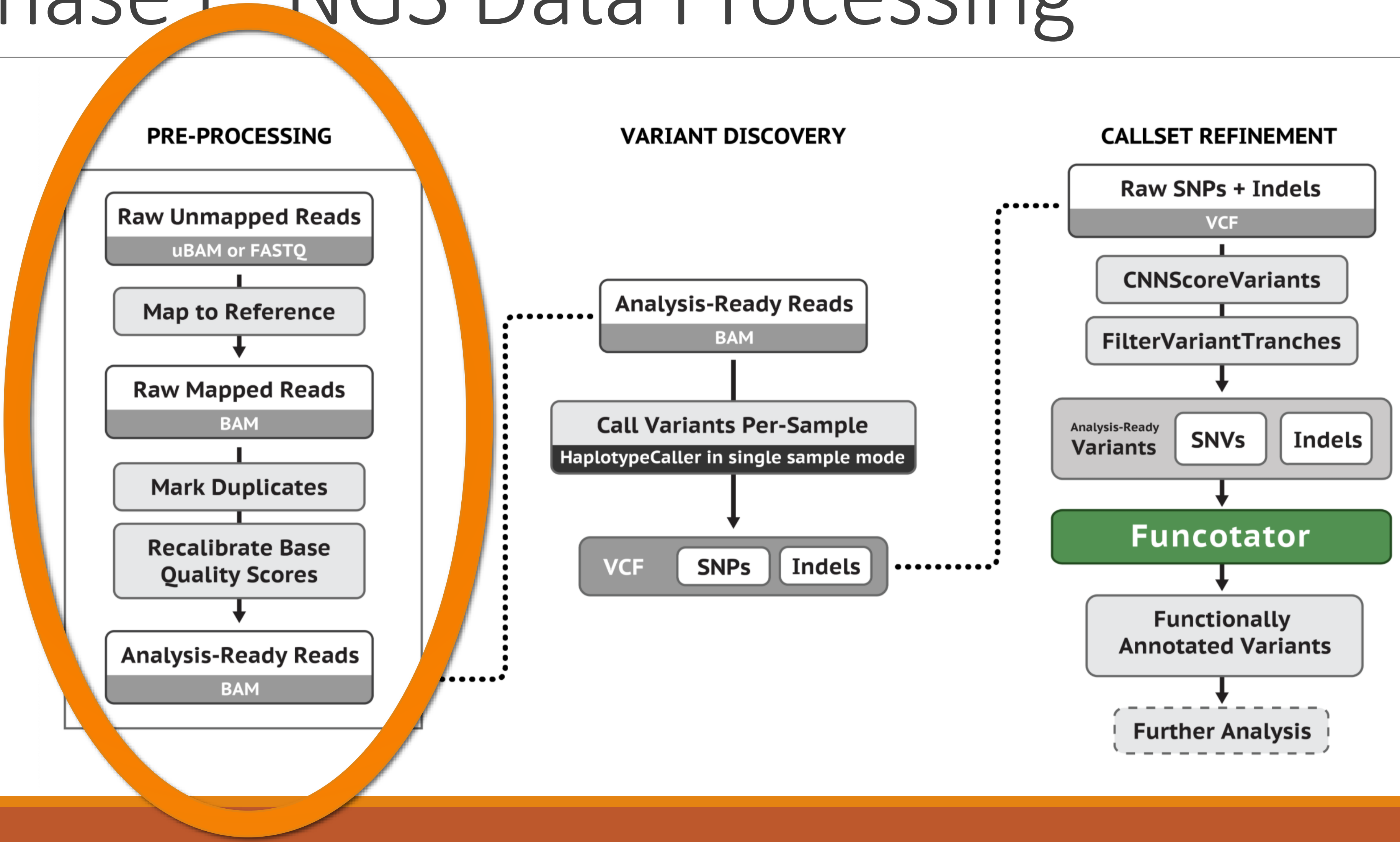
Standard GATK Pipeline

aka ‘Best Practices’

GATK Pipeline – Germline Calls



Phase I : NGS Data Processing



Phase I

NGS Data Processing

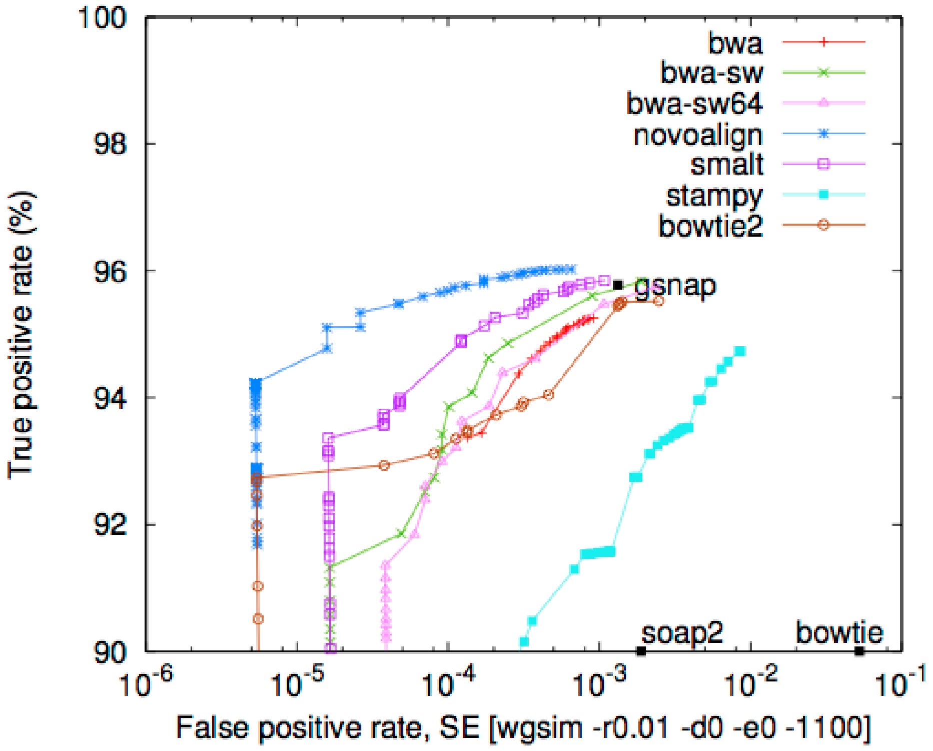
- Alignment of raw reads
- Duplicate marking
- Base quality recalibration
- ***Local realignment no longer required***

Phase I : Alignment of raw reads

Accuracy

- **Sensitivity** – maps reads accurately allowing for errors or variation

- **Specificity** – maps to the correct region



Heng Li's aligner assessment

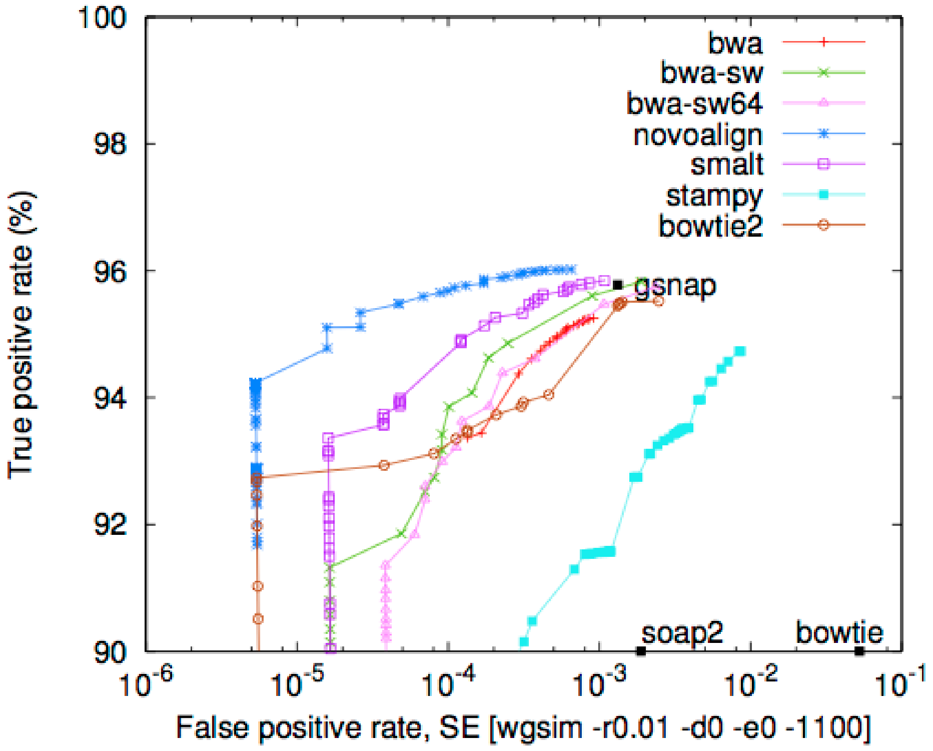
Phase I : Alignment of raw reads

Accuracy assessed using simulated data

BWA MEM is currently recommended

Unique vs. multi-mapped reads

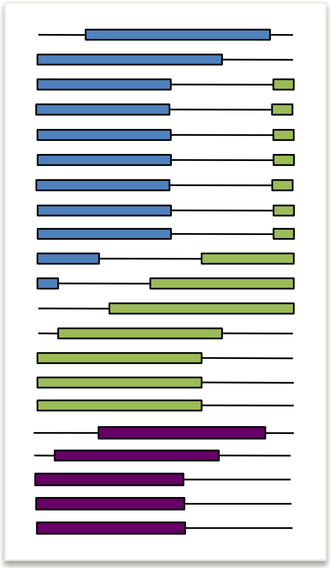
- Should we retain reads mapping to repetitive regions?
- May depend on the application

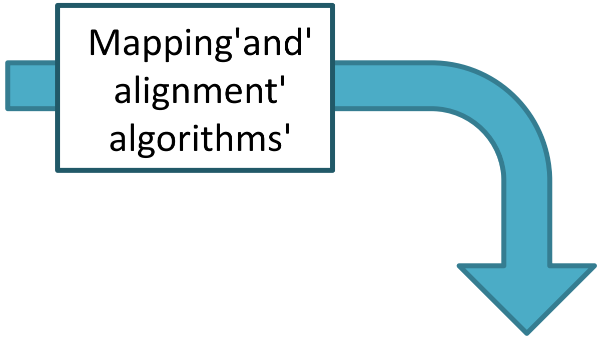


Heng Li’s aligner assessment

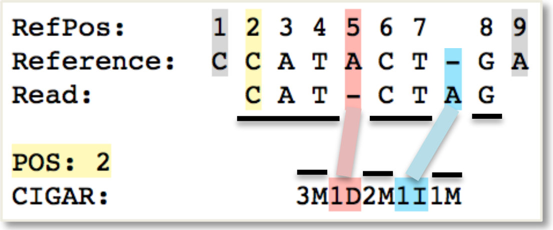
Mapping'short'reads'to'a'reference'is'simple'in'principle'

Enormous)pile)of)short) reads)from)NGS)

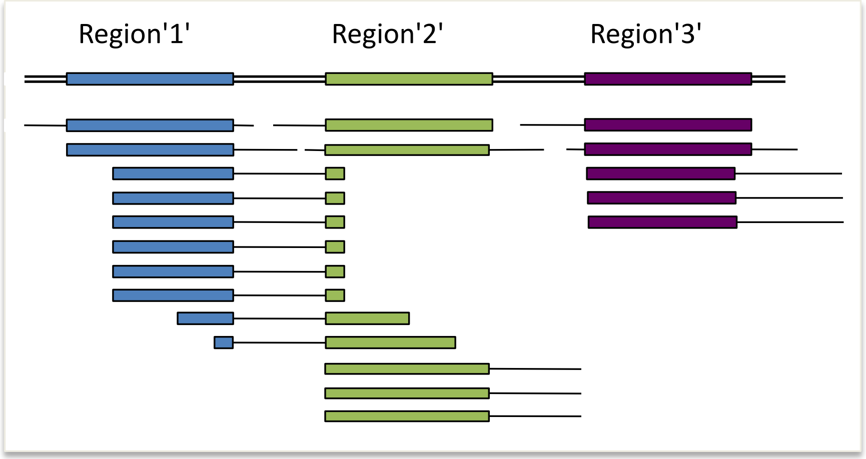




Iden; fy'where'the'read'matches' the'reference'sequence'and'record' match'details'as'CIGAR'string'



```
RefPos:     1 2 3 4 5 6 7   8 9
Reference:  C C A T A C T - G A
Read:         C A T - C T A G

POS: 2
CIGAR:          3M1D2M1I1M
```

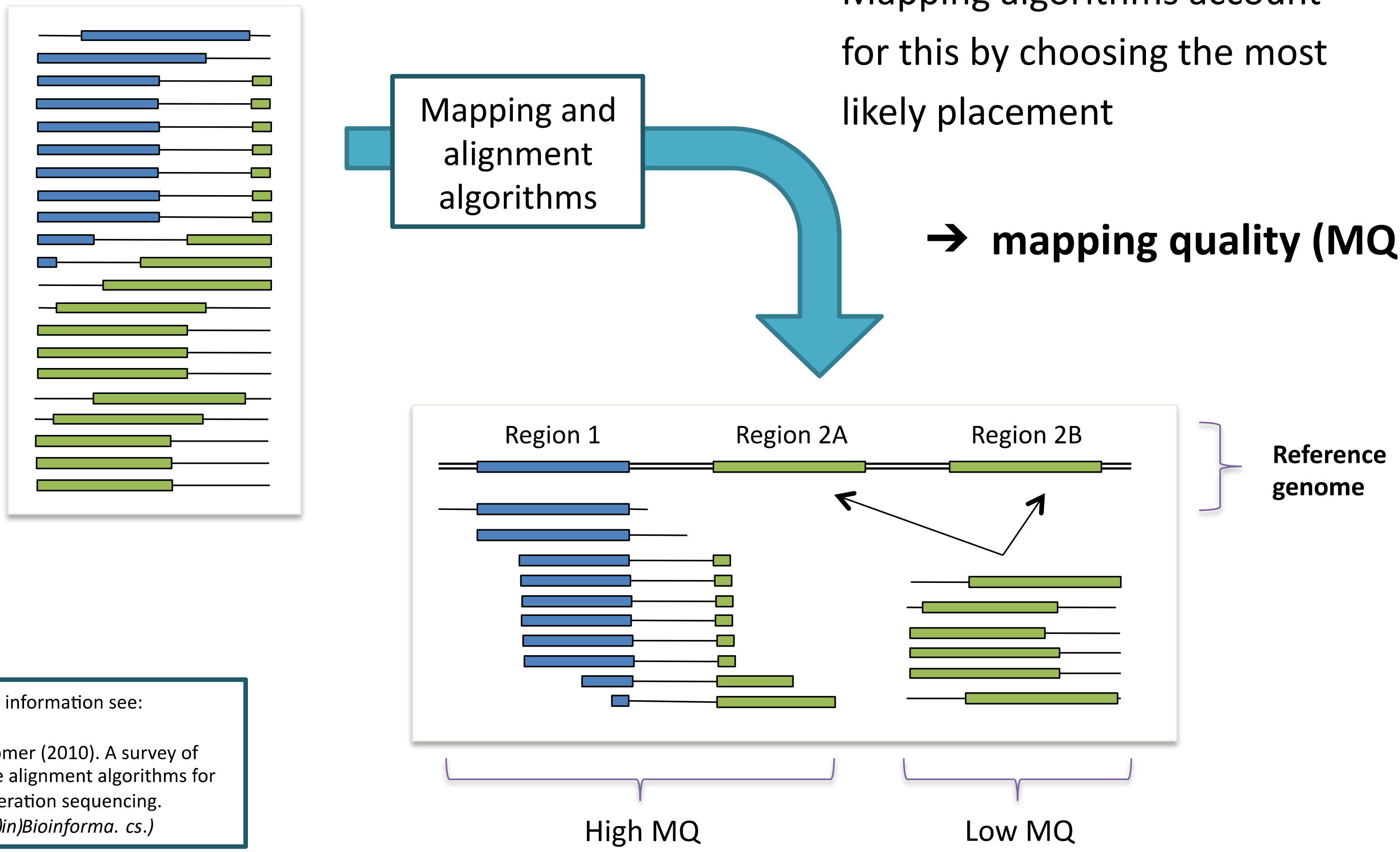


Reference) genome)

Reads' mapped'to' reference'

But mapping is actually very hard because of mismatches (true mutations or sequencing errors), duplicated regions etc.!

Enormous pile of short reads from NGS



Mapping algorithms account for this by choosing the most likely placement

➔ **mapping quality (MQ)**

For more information see:

Li and Homer (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings)in)Bioinforma. cs.)*

Alignment output : SAM/BAM

SAM – Sequence Alignment/Map format

- SAM file format stores alignment information
- Normally converted into BAM (text format is mostly useless for analysis)

Specification: http://samtools.sourceforge.net/SAM1.pdf

Contains FASTQ reads, quality information, meta data, alignment information, etc.

Files are typically very large: Many 100’s of GB or more

Alignment output : SAM/BAM

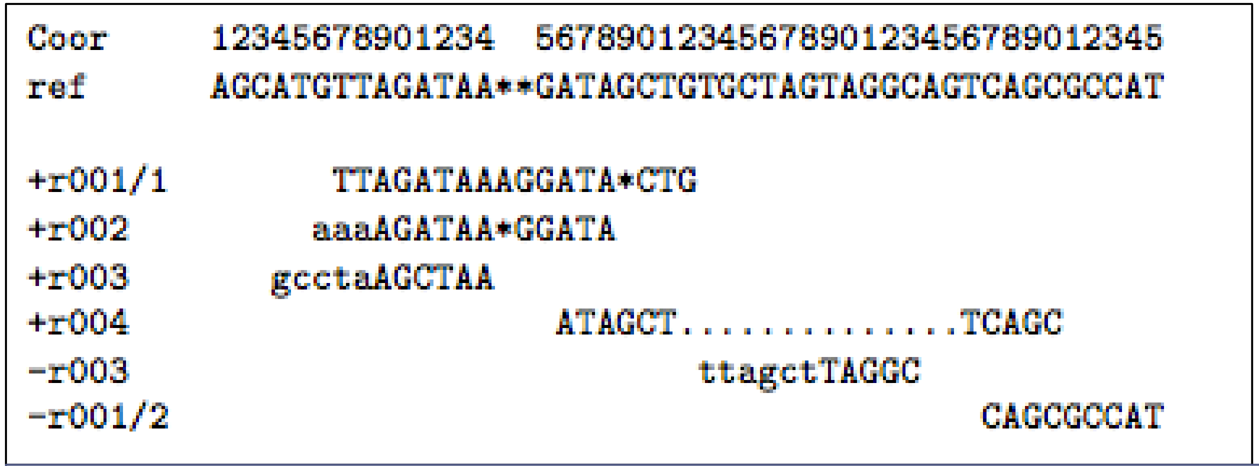
BAM – BGZF compressed SAM format

- May be unsorted, or sorted by sequence name or genome coordinates
- May be accompanied by an index file (.bai) (only if coord-sorted)
- Makes the alignment information easily accessible to downstream applications (large genome file not necessary)
- Relatively simple format makes it easy to extract specific features, e.g. genomic locations
- BAM is the compressed/binary version of SAM and is not human readable. Uses a specialize compression algorithm optimized for indexing and record retrieval (bgzip)

Files are typically very large: 1/5 of SAM, but still very large

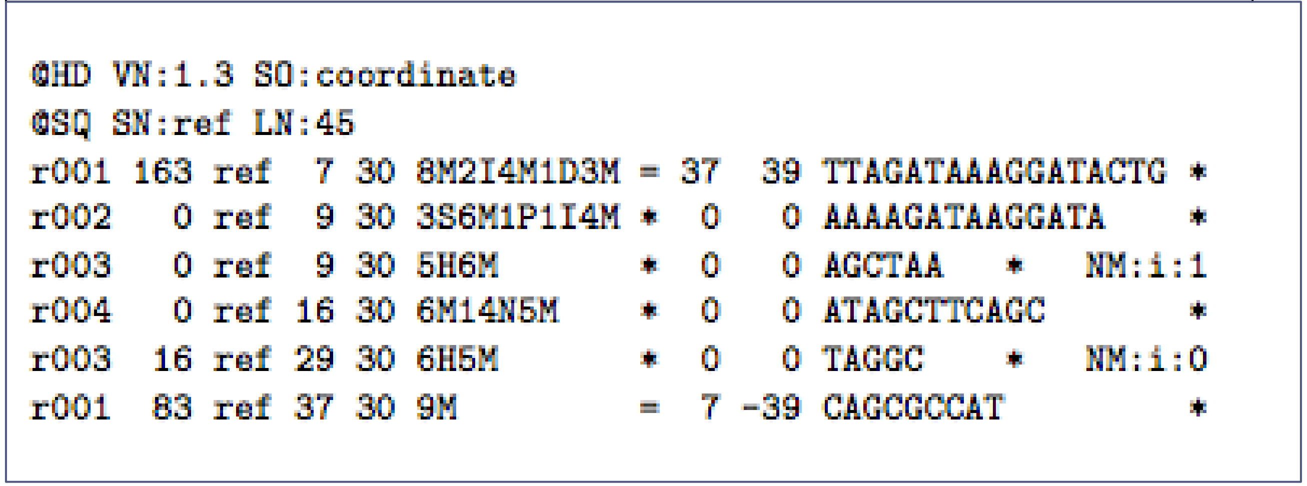
Alignment output : SAM/BAM

Alignment



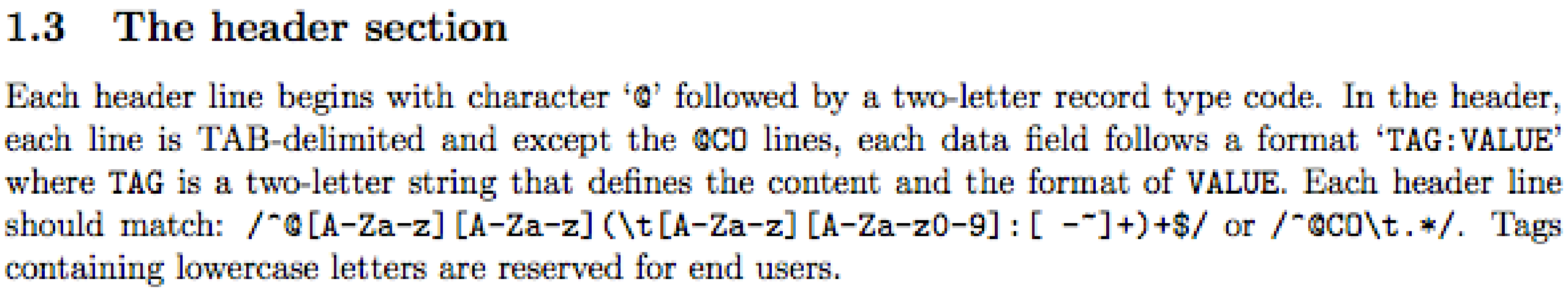
```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                        CAGCGCCAT
```

SAM format



```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

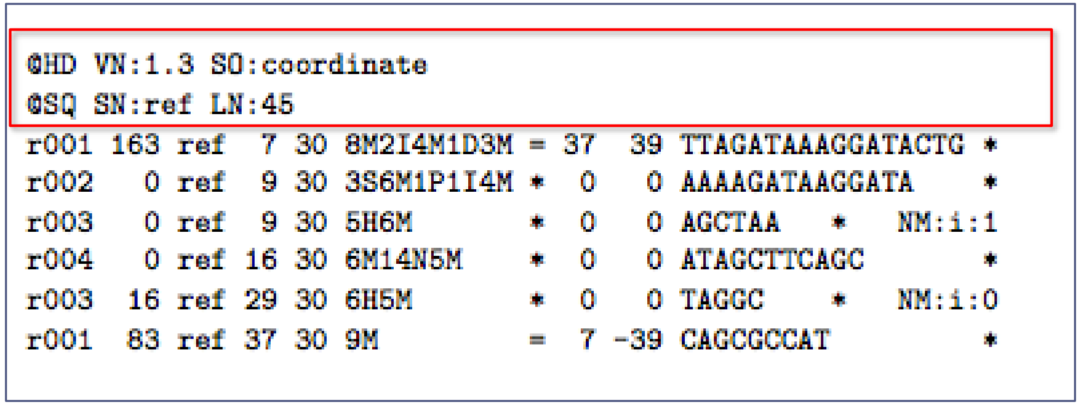
Alignment output : SAM/BAM



1.3 The header section

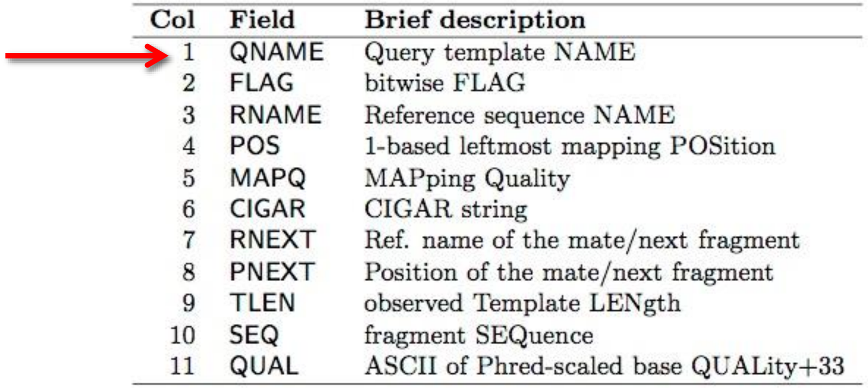
Each header line begins with character '@' followed by a two-letter record type code. In the header, each line is TAB-delimited and except the @CO lines, each data field follows a format 'TAG:VALUE' where TAG is a two-letter string that defines the content and the format of VALUE. Each header line should match: /^@[A-Za-z][A-Za-z](\t[A-Za-z][A-Za-z0-9]:[-~]+)+$/ or /^@CO\t.*/. Tags containing lowercase letters are reserved for end users.

SAM format



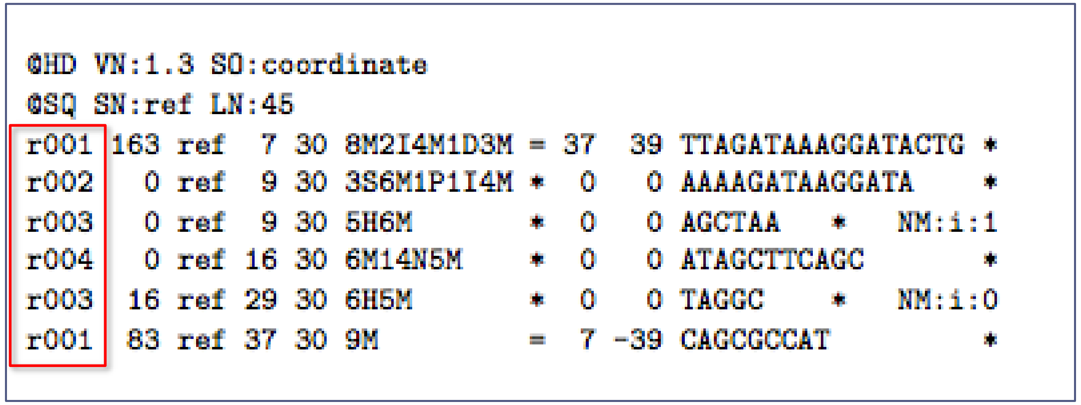
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA  *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC   *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

1.4 The alignment section: mandatory fields



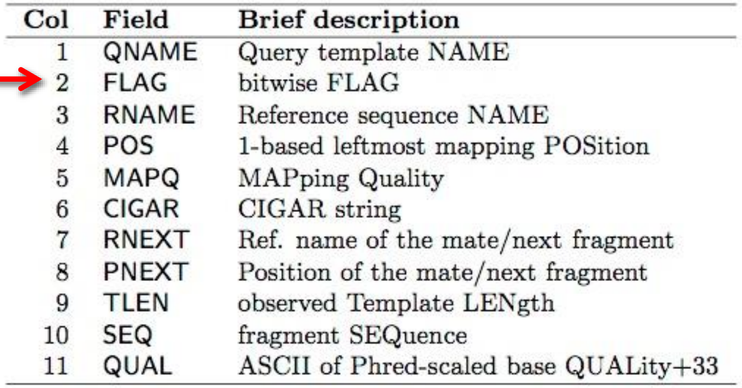
Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next fragment
8	PNEXT	Position of the mate/next fragment
9	TLEN	observed Template LENgth
10	SEQ	fragment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

SAM format



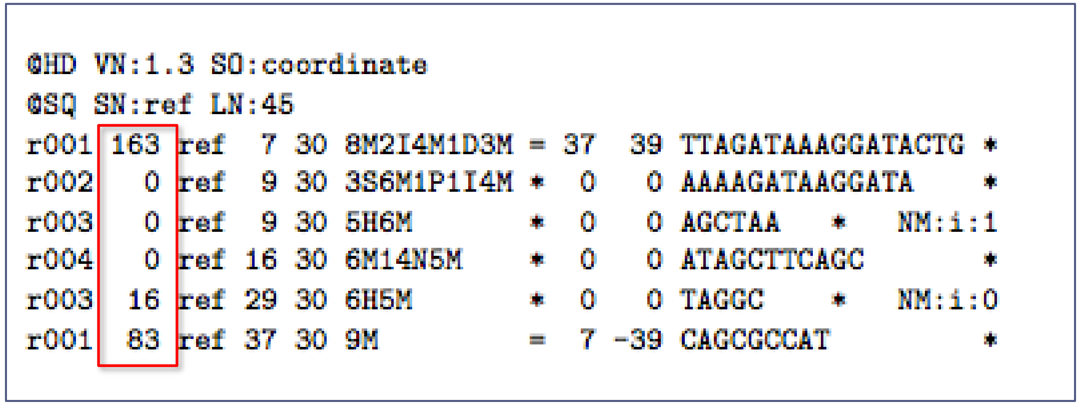
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M * 0    0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       * 0    0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    * 0    0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       * 0    0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         = 7  -39 CAGCGCCAT         *
```

1.4 The alignment section: mandatory fields



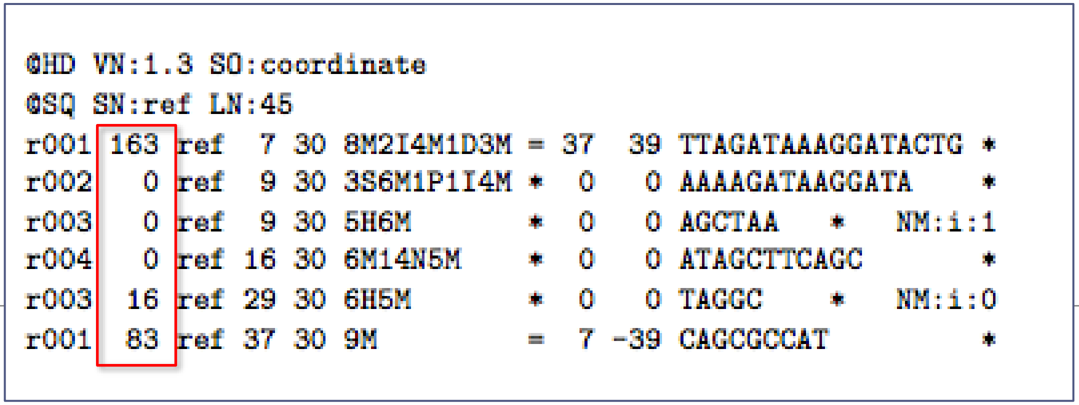
Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next fragment
8	PNEXT	Position of the mate/next fragment
9	TLEN	observed Template LENgth
10	SEQ	fragment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

SAM format

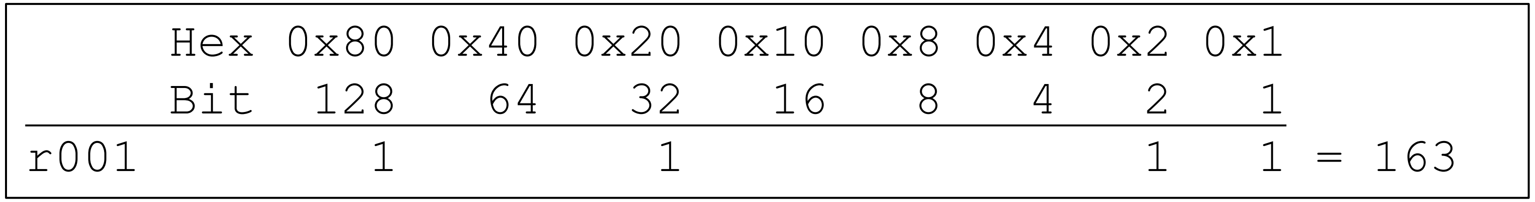


```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA  *  NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC   *  NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

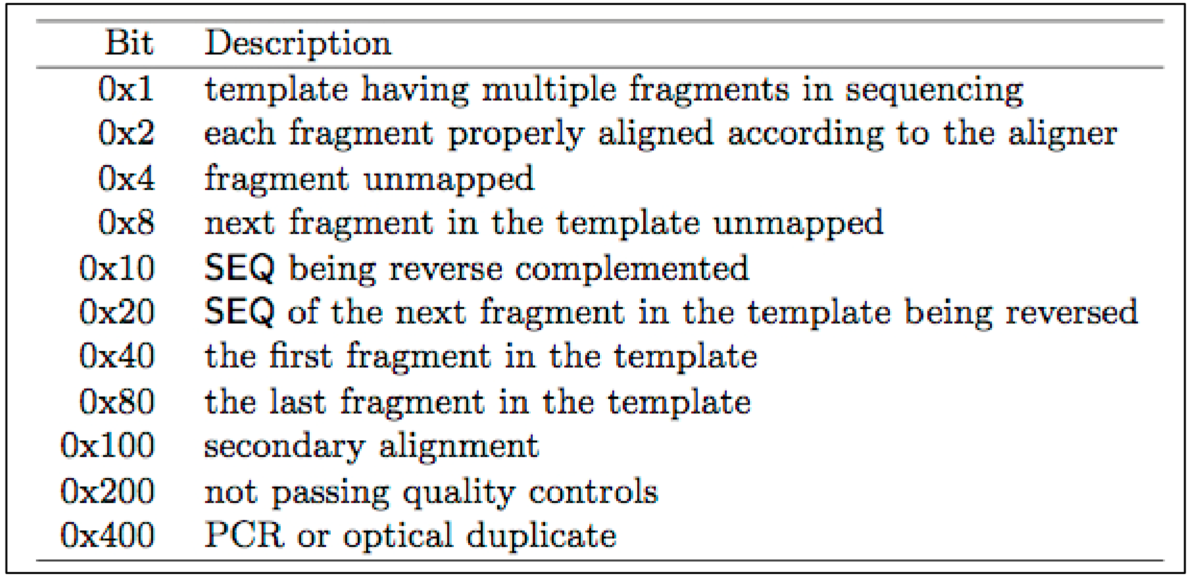
Bit Flags



```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA  *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC   *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

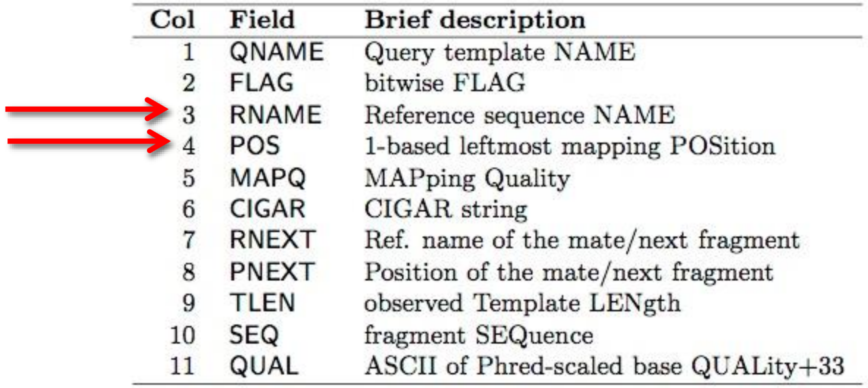


	Hex	0x80	0x40	0x20	0x10	0x8	0x4	0x2	0x1	
	Bit	128	64	32	16	8	4	2	1	
r001		1		1				1	1	= 163



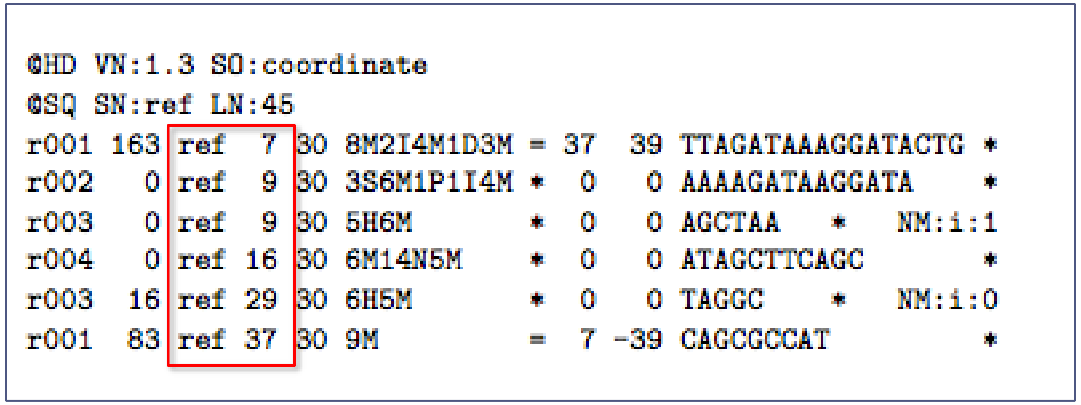
Bit	Description
0x1	template having multiple fragments in sequencing
0x2	each fragment properly aligned according to the aligner
0x4	fragment unmapped
0x8	next fragment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next fragment in the template being reversed
0x40	the first fragment in the template
0x80	the last fragment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

1.4 The alignment section: mandatory fields

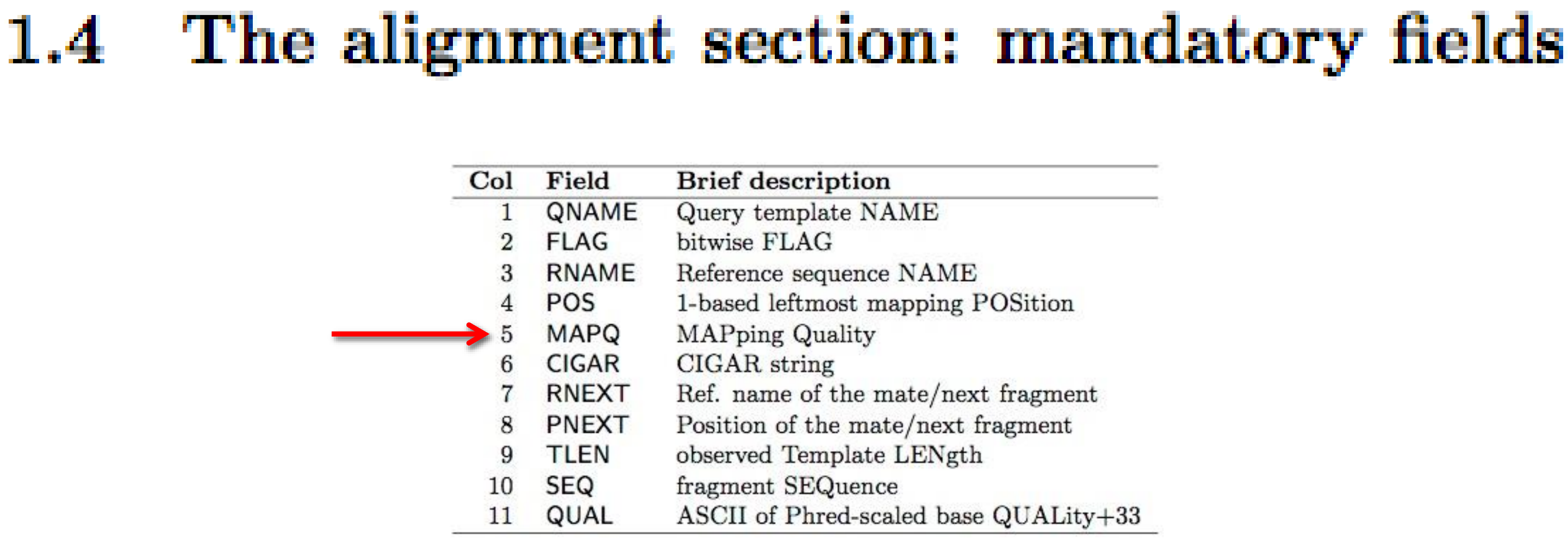


Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next fragment
8	PNEXT	Position of the mate/next fragment
9	TLEN	observed Template LENgth
10	SEQ	fragment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

SAM format



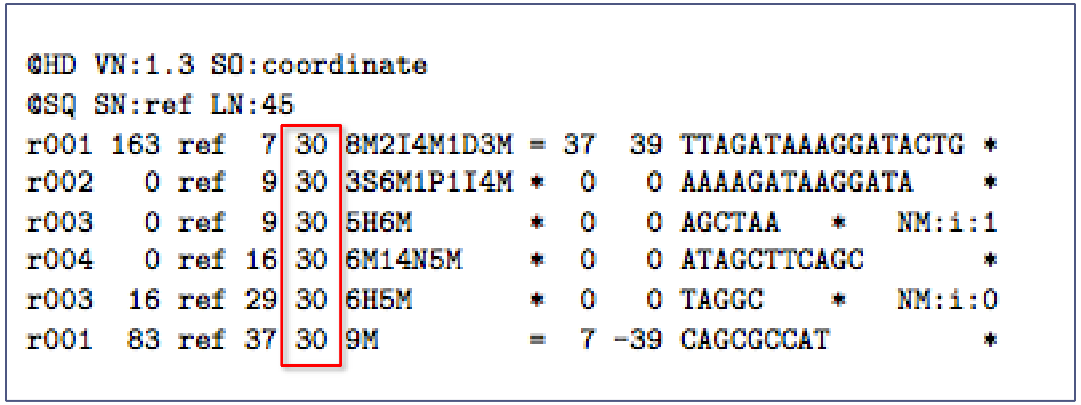
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA  *    NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC   *    NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```



1.4 The alignment section: mandatory fields

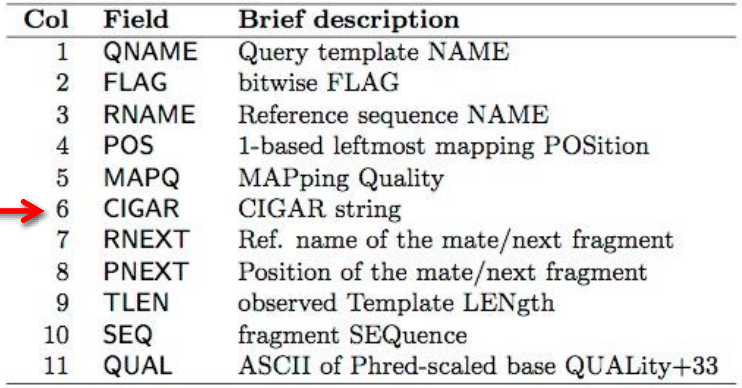
Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next fragment
8	PNEXT	Position of the mate/next fragment
9	TLEN	observed Template LENgth
10	SEQ	fragment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

SAM format



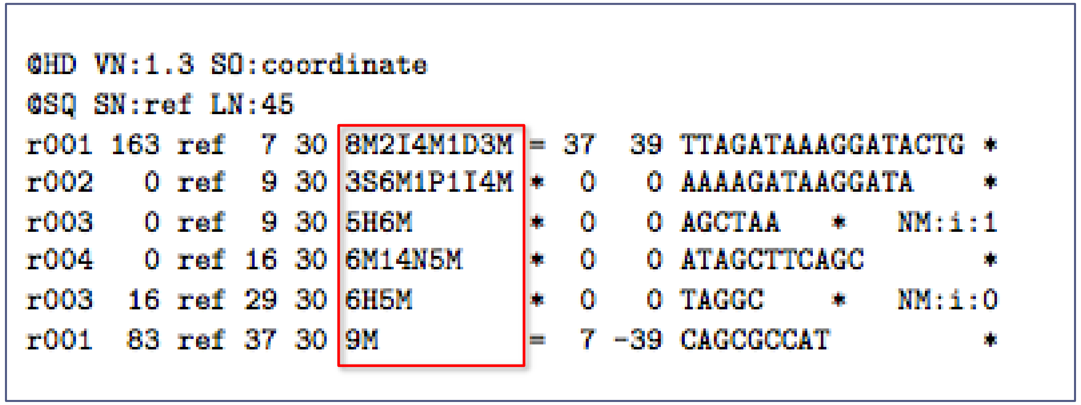
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M * 0    0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       * 0    0 AGCTAA  *   NM:i:1
r004   0 ref 16 30 6M14N5M    * 0    0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       * 0    0 TAGGC   *   NM:i:0
r001  83 ref 37 30 9M         = 7  -39 CAGCGCCAT         *
```

1.4 The alignment section: mandatory fields



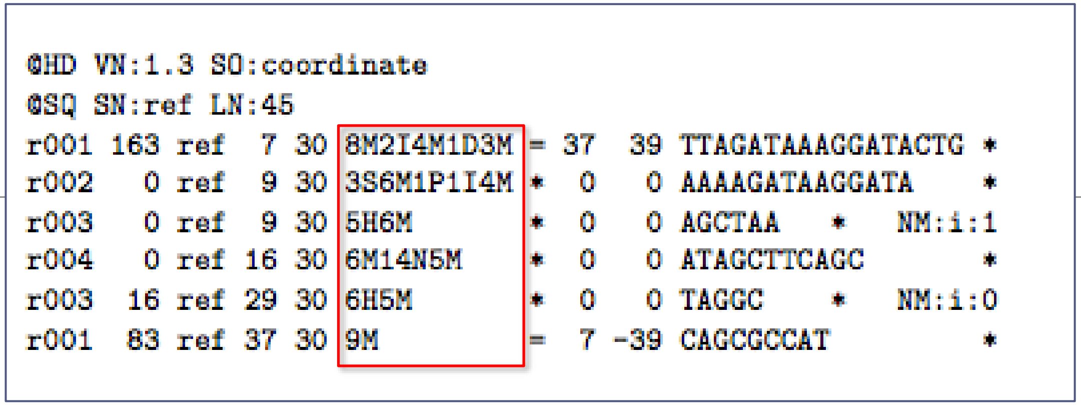
Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next fragment
8	PNEXT	Position of the mate/next fragment
9	TLEN	observed Template LENgth
10	SEQ	fragment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

SAM format

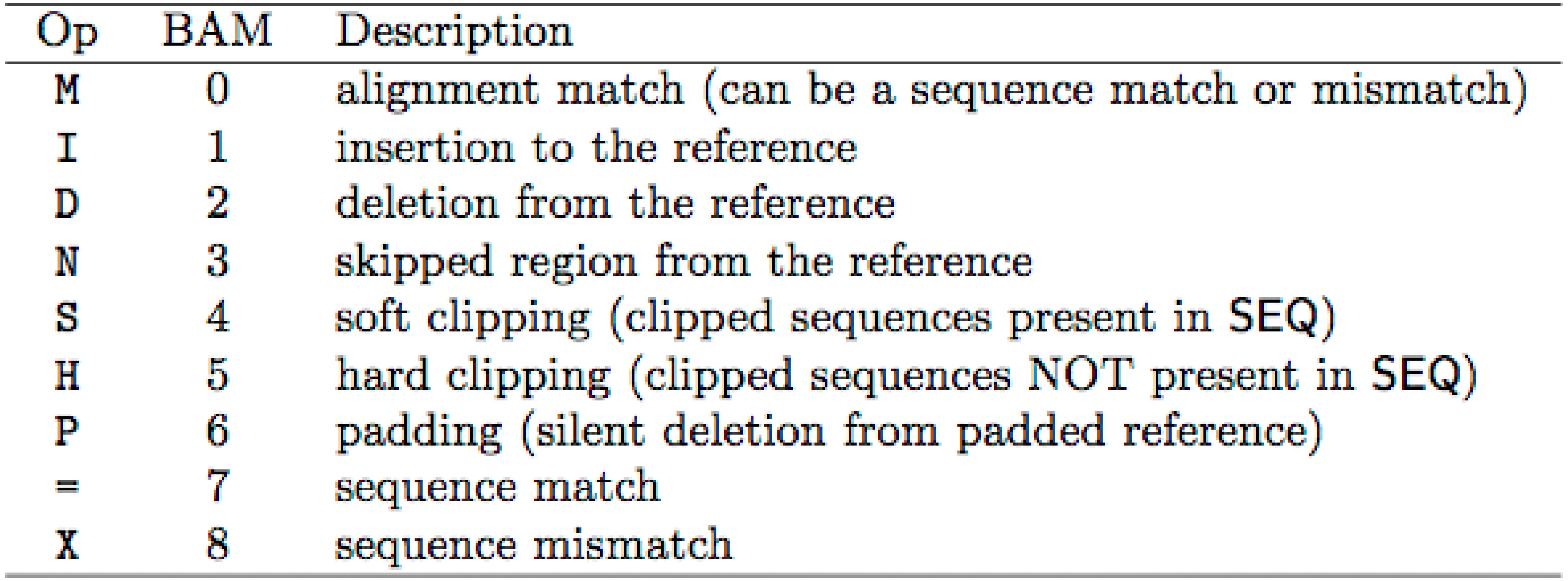


```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA     *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC      *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

CIGAR

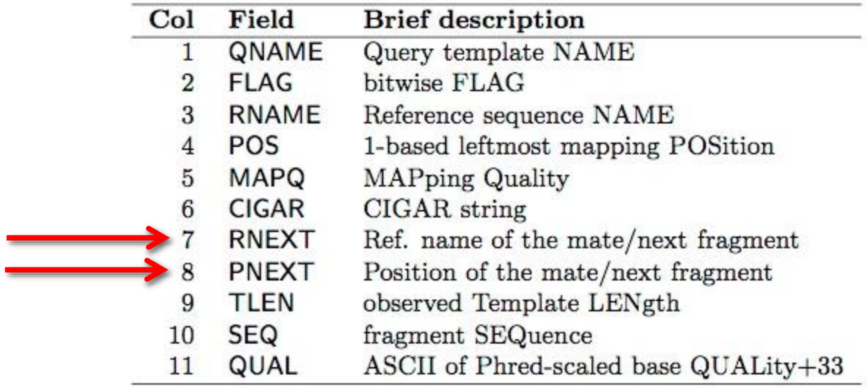


```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA  *    NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC   *    NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```



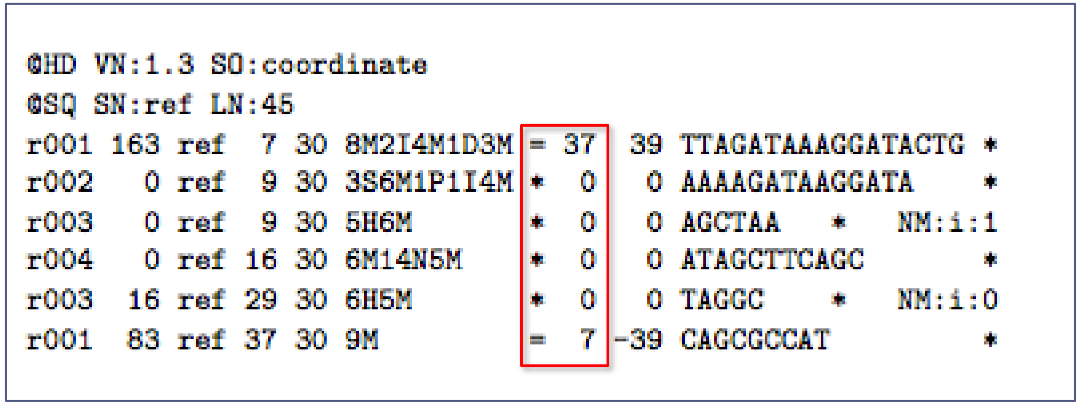
Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

1.4 The alignment section: mandatory fields



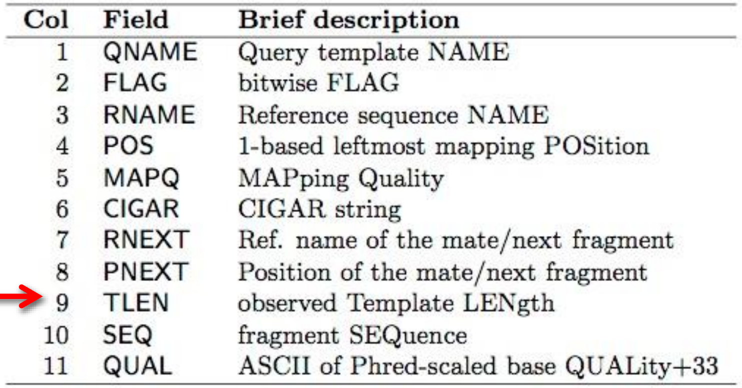
Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next fragment
8	PNEXT	Position of the mate/next fragment
9	TLEN	observed Template LENgth
10	SEQ	fragment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

SAM format



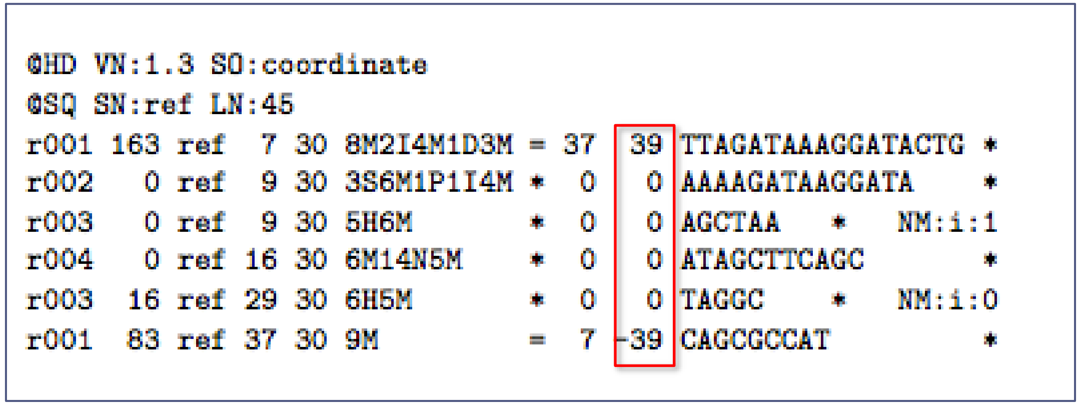
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA     *    NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC      *    NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

1.4 The alignment section: mandatory fields



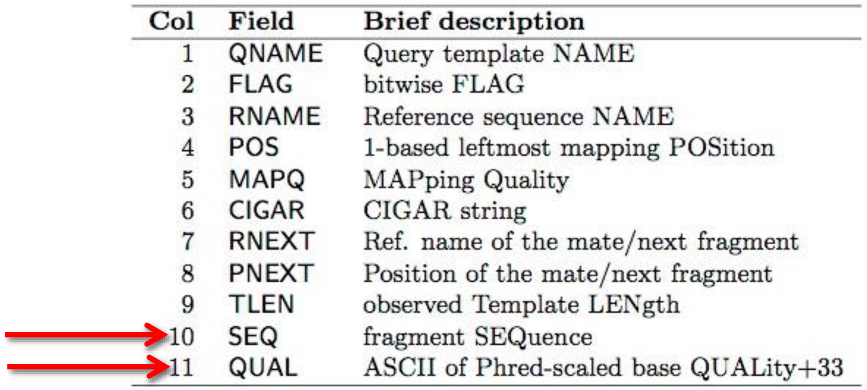
Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next fragment
8	PNEXT	Position of the mate/next fragment
9	TLEN	observed Template LENgth
10	SEQ	fragment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

SAM format



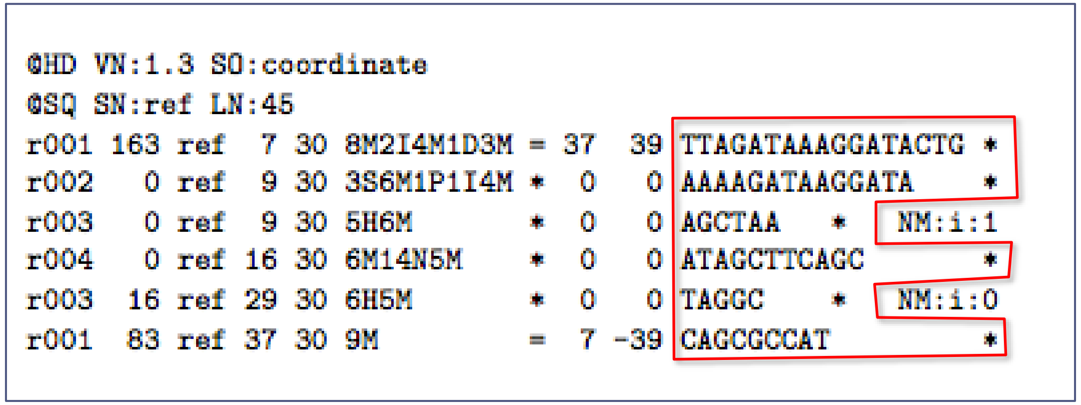
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA  *    NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC   *    NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

1.4 The alignment section: mandatory fields



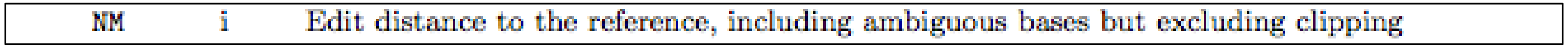
Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next fragment
8	PNEXT	Position of the mate/next fragment
9	TLEN	observed Template LENgth
10	SEQ	fragment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

SAM format



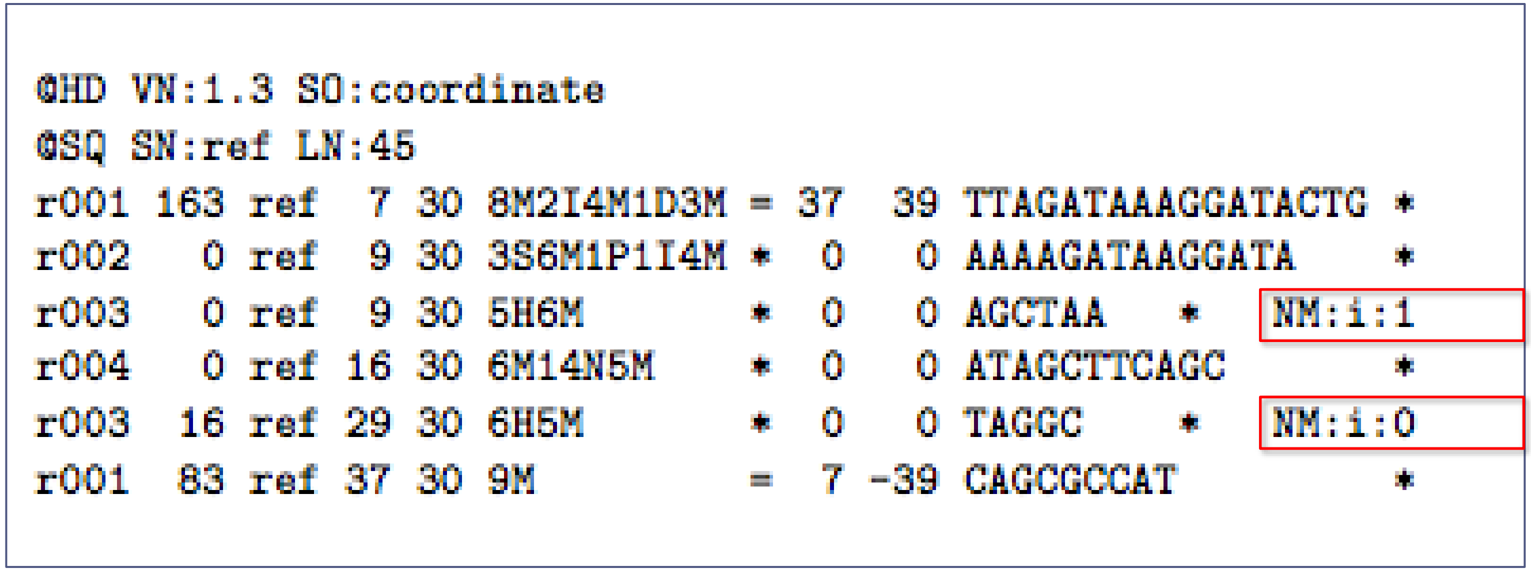
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA  *  NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC   *  NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

1.5 The alignment section: optional fields



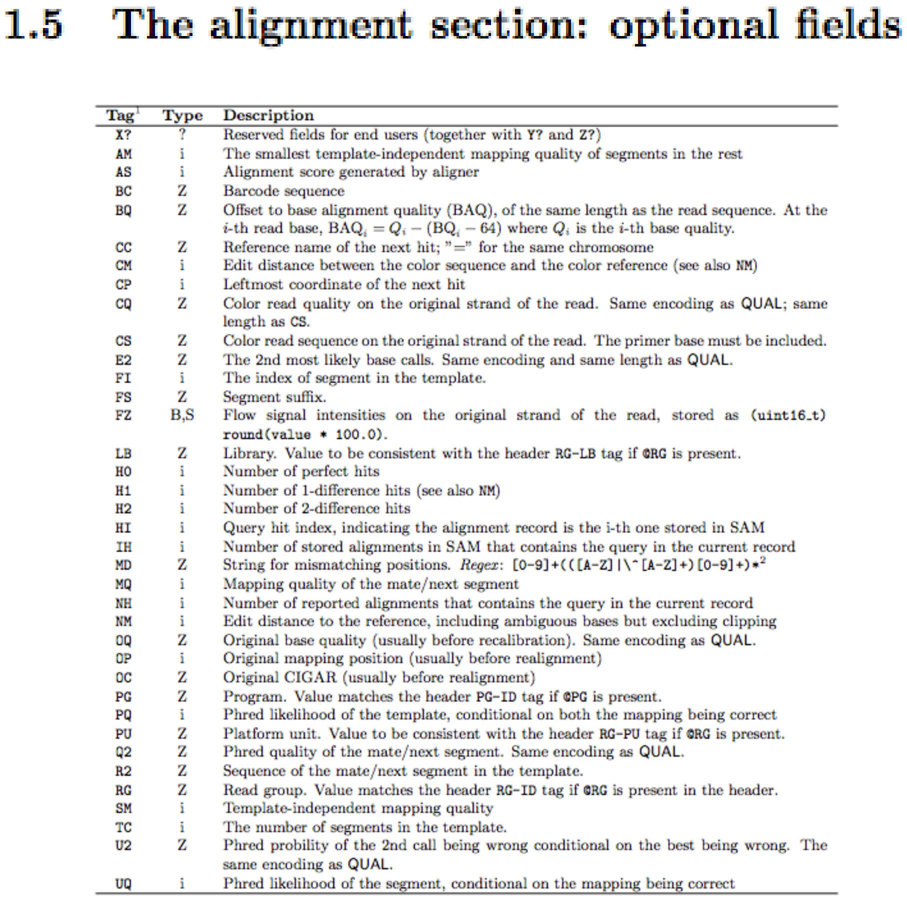
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping

SAM format



```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *  NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *  NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Alignment output : SAM/BAM



1.5 The alignment section: optional fields

Tag[1]	Type	Description
X?	?	Reserved fields for end users (together with Y? and Z?)
AM	i	The smallest template-independent mapping quality of segments in the rest
AS	i	Alignment score generated by aligner
BC	Z	Barcode sequence
BQ	Z	Offset to base alignment quality (BAQ), of the same length as the read sequence. At the i-th read base, $\text{BAQ}_i = Q_i - (\text{BQ}_i - 64)$ where Q_i is the i-th base quality.
CC	Z	Reference name of the next hit; "=" for the same chromosome
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CP	i	Leftmost coordinate of the next hit
CQ	Z	Color read quality on the original strand of the read. Same encoding as QUAL; same length as CS.
CS	Z	Color read sequence on the original strand of the read. The primer base must be included.
E2	Z	The 2nd most likely base calls. Same encoding and same length as QUAL.
FI	i	The index of segment in the template.
FS	Z	Segment suffix.
FZ	B,S	Flow signal intensities on the original strand of the read, stored as (uint16_t) round(value * 100.0).
LB	Z	Library. Value to be consistent with the header RG-LB tag if @RG is present.
H0	i	Number of perfect hits
H1	i	Number of 1-difference hits (see also NM)
H2	i	Number of 2-difference hits
HI	i	Query hit index, indicating the alignment record is the i-th one stored in SAM
IH	i	Number of stored alignments in SAM that contains the query in the current record
MD	Z	String for mismatching positions. *Regex*: [0-9]+(([A-Z]|\^[A-Z]+)[0-9]+)*[2]
MQ	i	Mapping quality of the mate/next segment
NH	i	Number of reported alignments that contains the query in the current record
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping
OQ	Z	Original base quality (usually before recalibration). Same encoding as QUAL.
OP	i	Original mapping position (usually before realignment)
OC	Z	Original CIGAR (usually before realignment)
PG	Z	Program. Value matches the header PG-ID tag if @PG is present.
PQ	i	Phred likelihood of the template, conditional on both the mapping being correct
PU	Z	Platform unit. Value to be consistent with the header RG-PU tag if @RG is present.
Q2	Z	Phred quality of the mate/next segment. Same encoding as QUAL.
R2	Z	Sequence of the mate/next segment in the template.
RG	Z	Read group. Value matches the header RG-ID tag if @RG is present in the header.
SM	i	Template-independent mapping quality
TC	i	The number of segments in the template.
U2	Z	Phred probility of the 2nd call being wrong conditional on the best being wrong. The same encoding as QUAL.
UQ	i	Phred likelihood of the segment, conditional on the mapping being correct

Too many to go over!!!

Alignment output : SAM/BAM

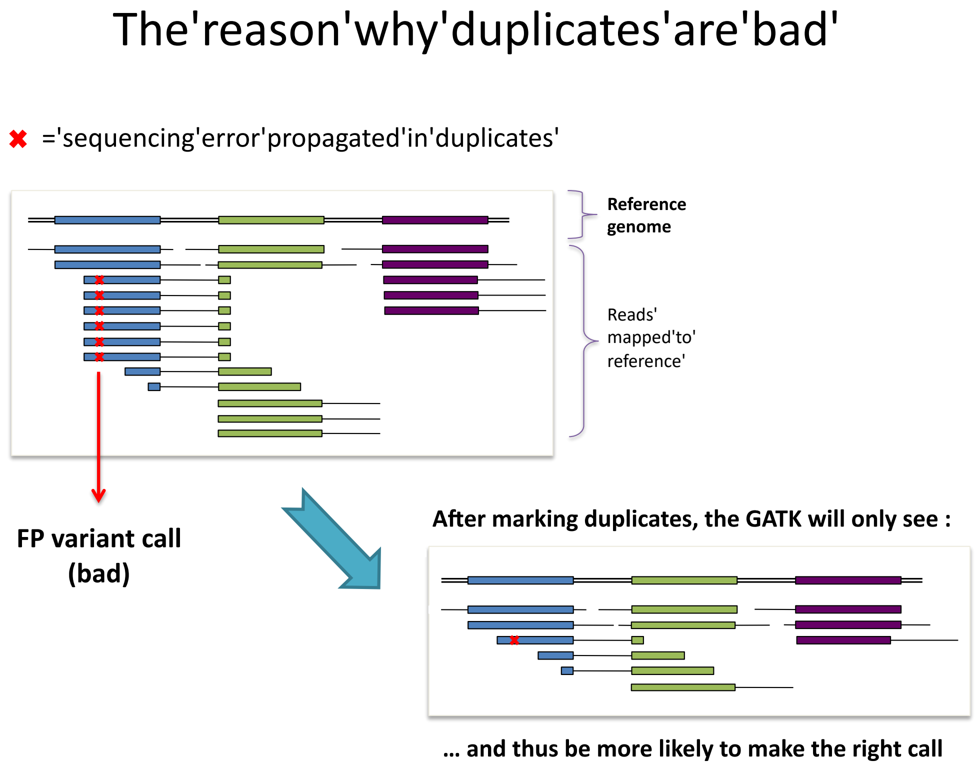
Tools

- **samtools**
- **Picard**

Mining information from a properly formatted BAM file:

- Reads in a region (good for RNA-Seq, ChIP-Seq)
- Quality of alignments
- Coverage
- …and of course, differences (variants)

Phase I : Duplicate Marking



The'reason'why'duplicates'are'bad'

✖ ='sequencing'error'propagated'in'duplicates'

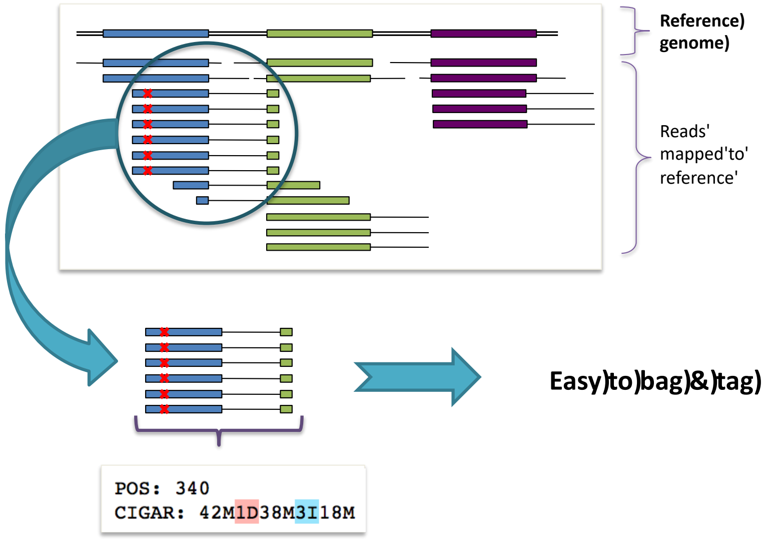
FP variant call (bad)

After marking duplicates, the GATK will only see :

… and thus be more likely to make the right call

Phase I : Duplicate Marking

Duplicates'have'the'same'star; ng'posi; on''
and'the'same'CIGAR'string'



Reference)
genome)

Reads'
mapped'to'
reference'

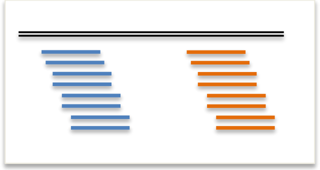
Easy)to)bag)&)tag)

POS: 340
CIGAR: 42M1D38M3I18M

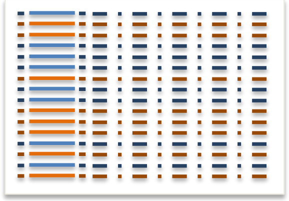
Phase I : Sorting, Read Groups

A'quick'diversion'about'sor; ng'and'read'groups'

The information for this:

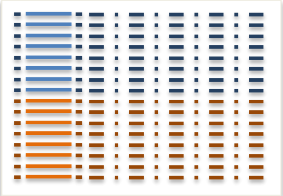


… is actually stored as a text file with one line per read which from far away looks like this:



The reads are in no particular order…

… but the GATK wants reads to be sorted by starting position like this:



So we need to explicitly sort the SAM file…

And'while'we're'at'it,'let's'add'**read group** informa; on'if'it'isn't'already' there,'so'**the GATK will know what read belongs to what sample**'(that's' kind'of'important).'

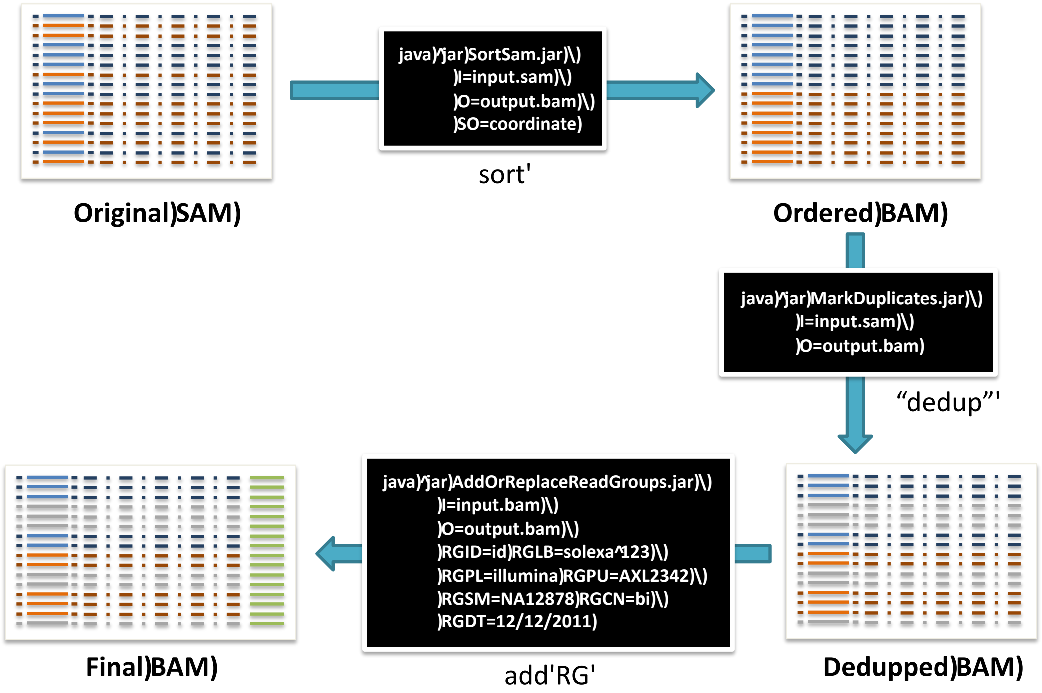
Terminology

Read groups – information about the samples and how they were run

- ID – Simple unique identifier
- Library
- Sample name
- Platform – sequencing platform
- Platform unit – barcode or identifier
- Sequencing center (optional)
- Description (optional)
- Run date (optional)

Phase I : Sorting, Read Groups

Typical'workflow'using'Picard'tools'to'mark'duplicates'*et)al.)*



```
java)^jar)SortSam.jar)\)
        )I=input.sam)\)
        )O=output.bam)\)
        )SO=coordinate)
```

```
java)^jar)MarkDuplicates.jar)\)
        )I=input.sam)\)
        )O=output.bam)
```

```
java)^jar)AddOrReplaceReadGroups.jar)\)
        )I=input.bam)\)
        )O=output.bam)\)
        )RGID=id)RGLB=solexa^123)\)
        )RGPL=illumina)RGPU=AXL2342)\)
        )RGSM=NA12878)RGCN=bi)\)
        )RGDT=12/12/2011)
```

Phase I : Base Quality Score Recalibration

Quality scores from sequencers are biased and somewhat inaccurate

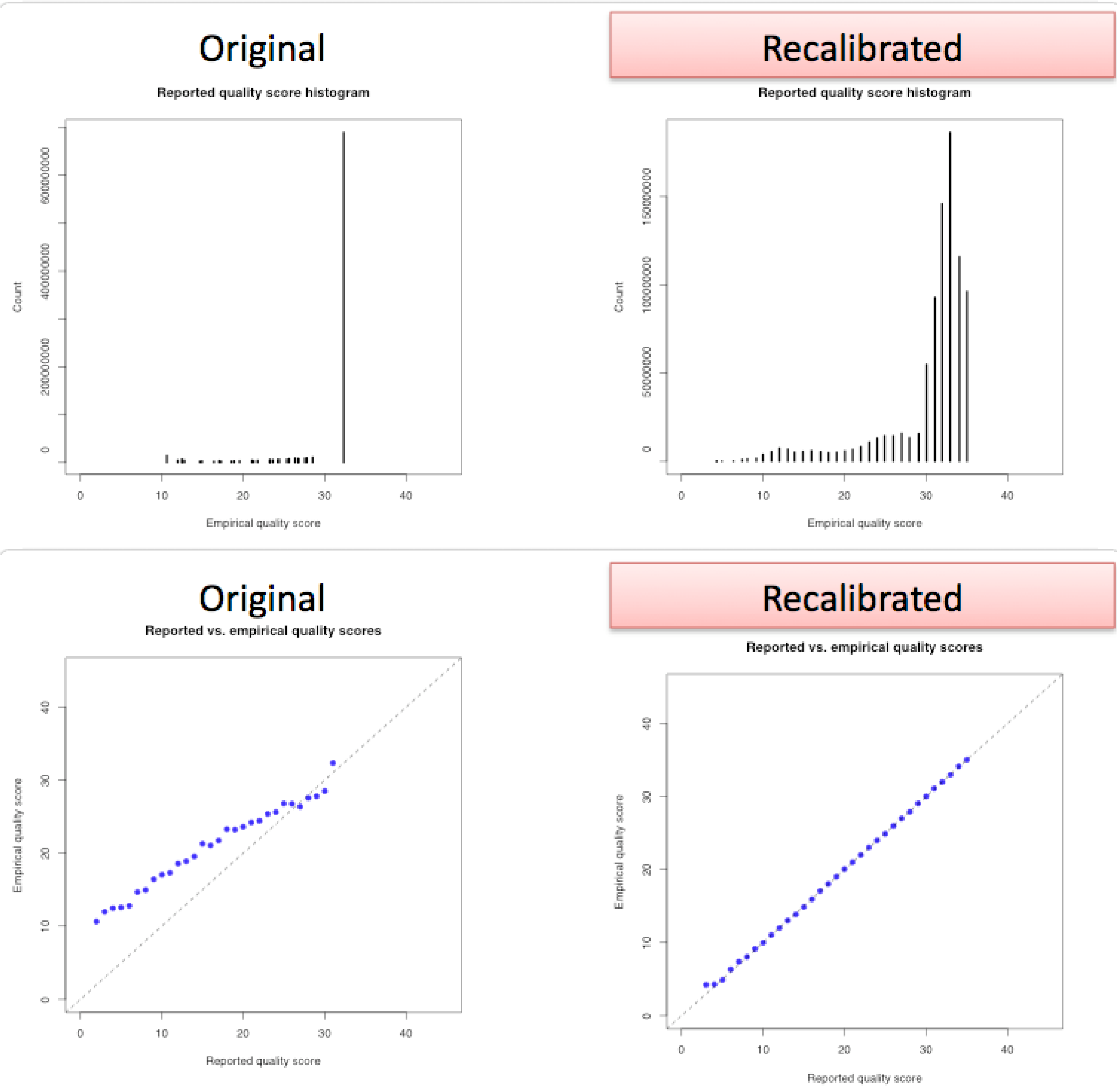
Quality scores are critical for all downstream analysis

Biases are a major contributor to bad variant calls

Caveat:

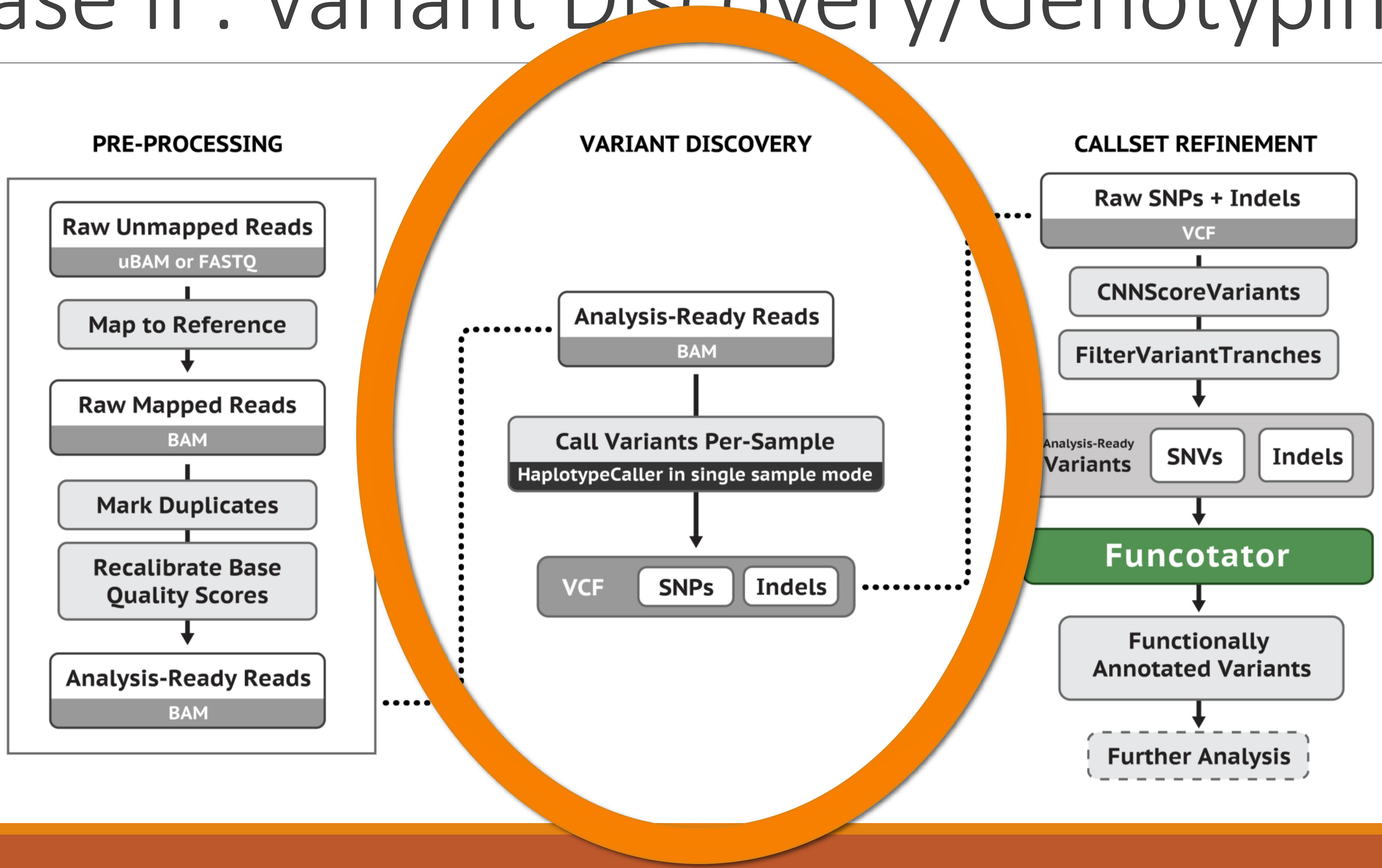
- **In practice, requires having a known set of variants (dbSNP)**

BQSR (GATK Link)



Also works for binned quality scores (NovaSeq)

Phase II : Variant Discovery/Genotyping



Phase II : Variant Calling

This is where we actually call the variants

Prior steps leading up to this help remove potential causes of variant calling errors

I’ll be covering their current recommended caller, the **HaplotypeCaller,** and a little on the use of the legacy **UnifiedGenotyper** (not included in GATK v4)

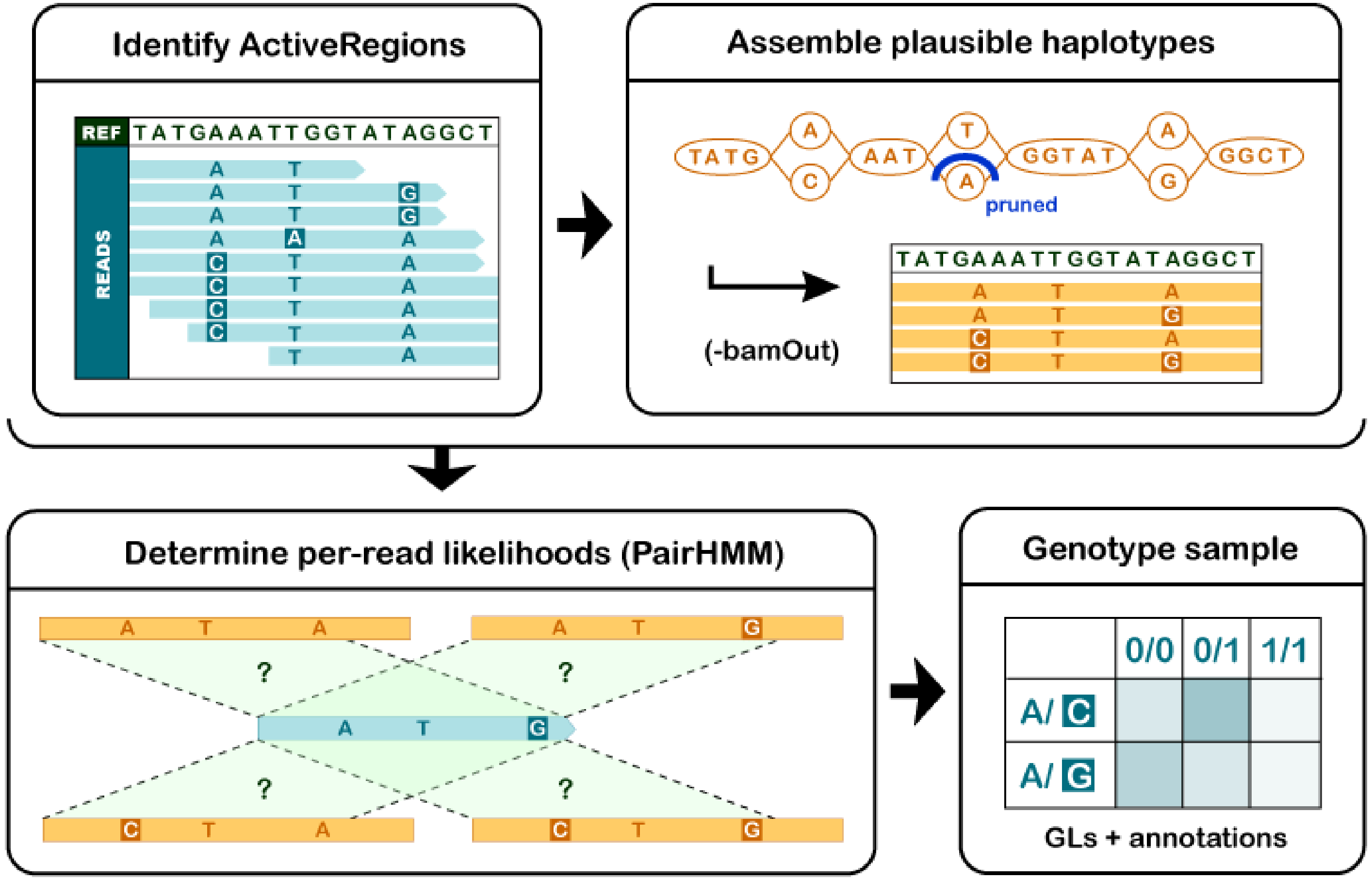
Phase II : Variant Calling *HaplotypeCaller*

Assembly-based approach

- Define active regions (evidence for variation)
- Re-assemble active region, align against reference
  - Get a list of *possible* haplotypes
  - No need for local realignment
- Determine likelihoods based on reads compared to haplotypes
- Find most likely genotype at each site, emit as a call

HaplotypeCaller in a nutshell (Broad Institute)

Phase II : Variant Calling
HaplotypeCaller



HaplotypeCaller in a nutshell (Broad Institute)

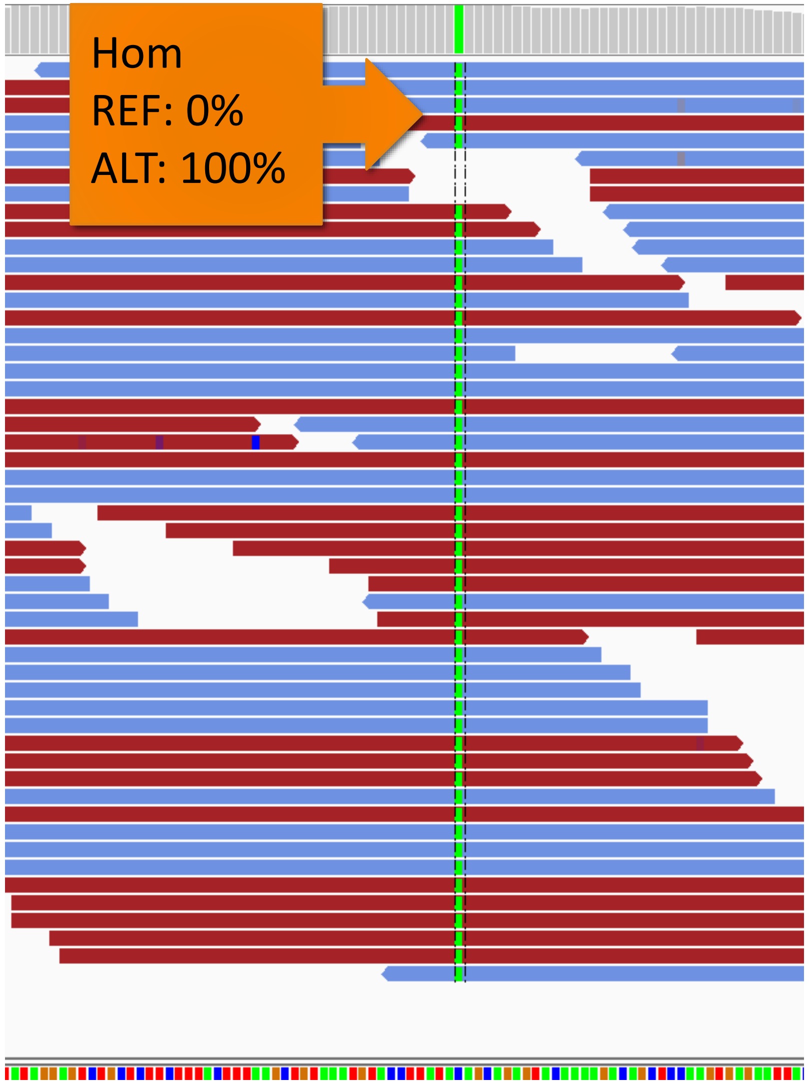
Phase II : Variant Calling *UnifiedGenotyper*

In general, uses a probabilistic method, e.g. Bayesian model

- Determine the possible SNP and indel alleles
- Only “good bases” are included:
  - Those satisfying minimum base quality, mapping read quality, pair mapping quality, etc.
- Compute, for each sample, for each genotype, likelihoods of data given genotypes
- Compute the allele frequency distribution to determine most likely allele count; emit a variant call if determined
- If we are going to emit a variant, assign a genotype to each sample

Note this assumes alignment is correct, **so should perform local realignment**

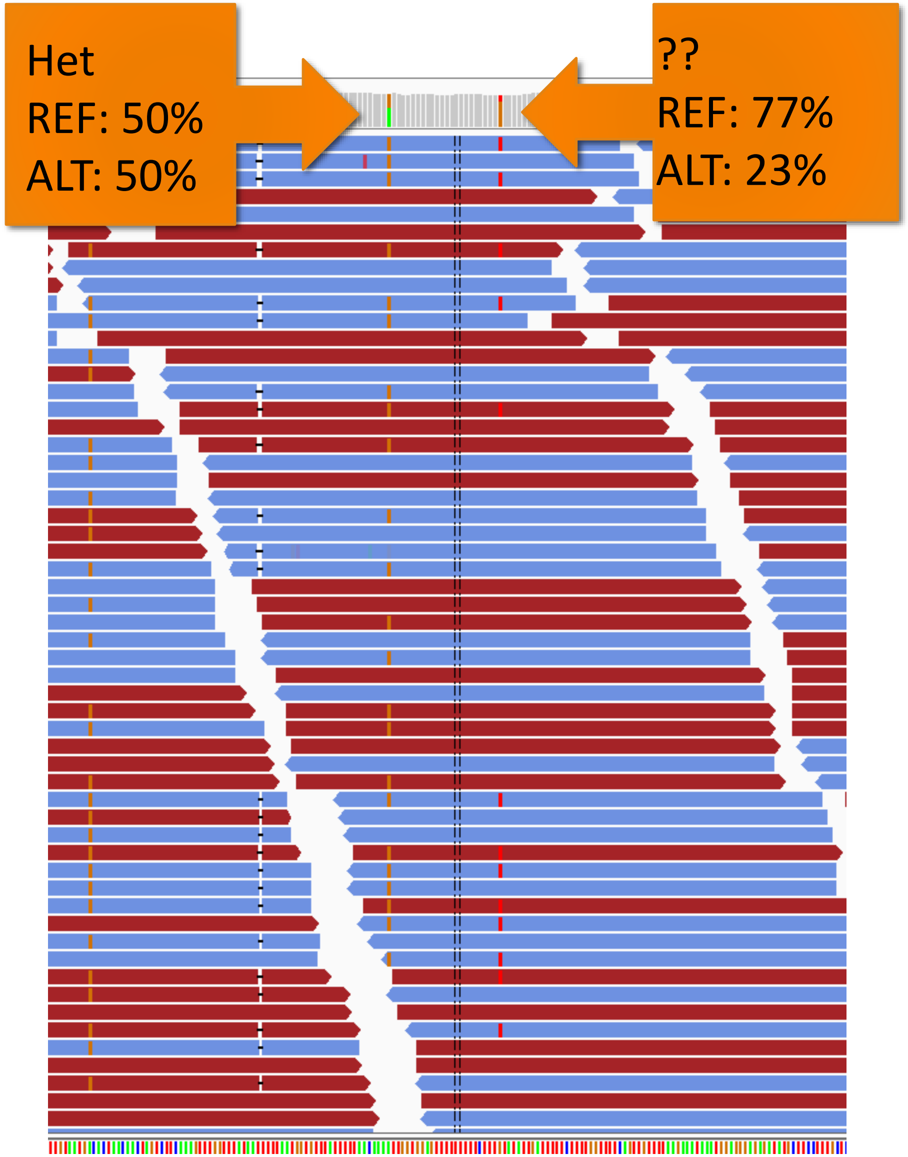
No longer recommended nor supported (not in GATK v4!)



Hom

REF: 0%

ALT: 100%



Het

REF: 50%

ALT: 50%

??

REF: 77%

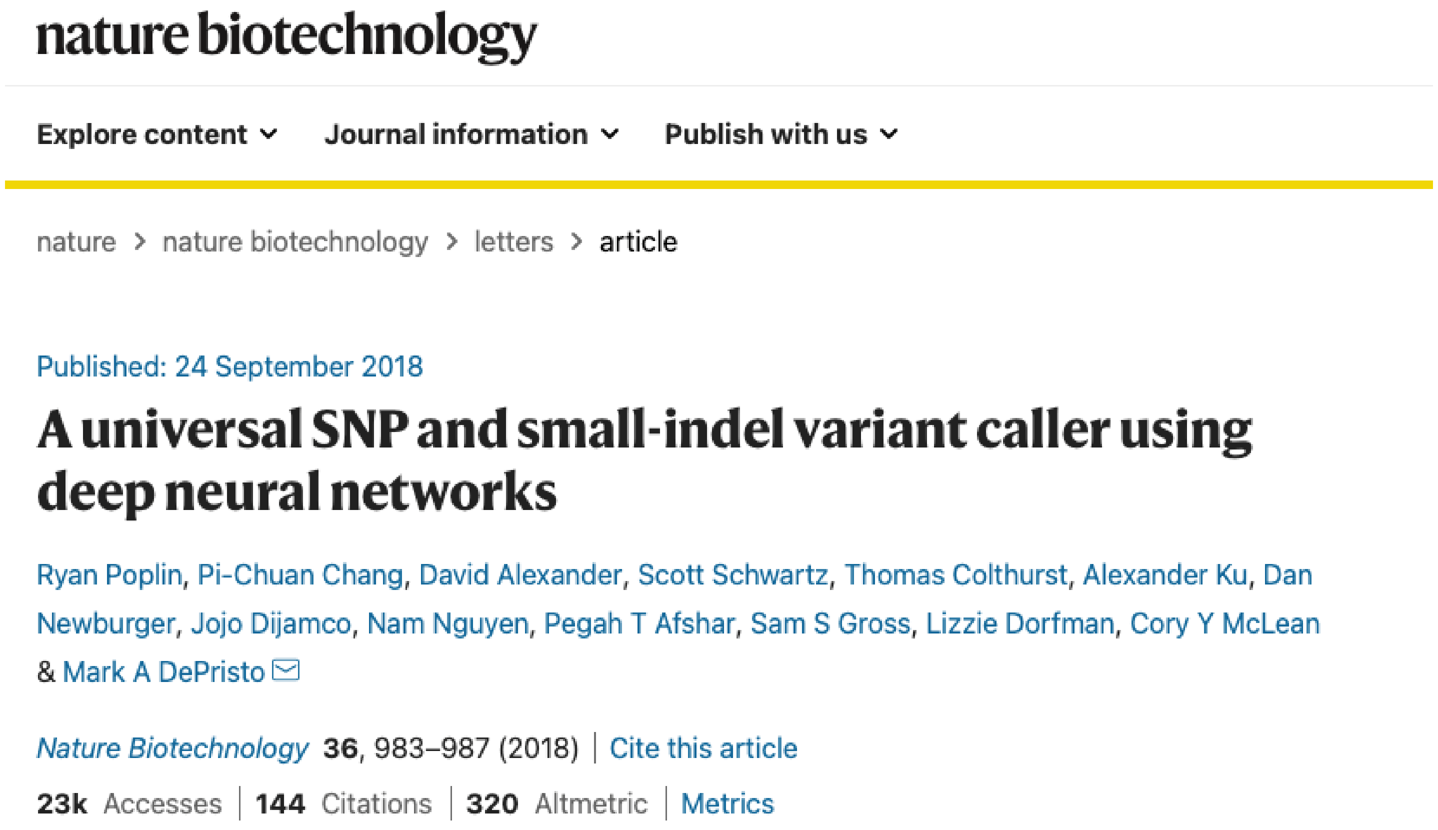
ALT: 23%

Side note: DeepVariant

‘Deep learning’ based variant calling tool

NOT PART OF GATK

Considered more accurate than HC



nature biotechnology

nature › nature biotechnology › letters › article

Published: 24 September 2018

A universal SNP and small-indel variant caller using deep neural networks

Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean & Mark A DePristo

Nature Biotechnology **36**, 983–987 (2018) | Cite this article

23k Accesses | **144** Citations | **320** Altmetric | Metrics

https://github.com/google/deepvariant

Output?

Variant calling output: VCF

VCF (Variant Call Format)

Like SAM/BAM, also has a versioned specification

- **From the 1000 Genomes Project**
- **http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41**

Structure (2 parts):

- **Header (metadata)**
- **Variant calls (one or more samples)**

Variant calls have multiple fields (right)

COL	FIELD	DESCRIPTION
1	CHROM	Chromosome name
2	POS	1-based position. For an indel, this is the position preceding the indel.
3	ID	Variant identifier. Usually the dbSNP rsID.
4	REF	Reference sequence at POS involved in the variant. For a SNP, it is a single base.
5	ALT	Comma delimited list of alternative seuqence(s).
6	QUAL	Phred-scaled probability of all samples being homozygous reference.
7	FILTER	Semicolon delimited list of filters that the variant fails to pass.
8	INFO	Semicolon delimited list of variant information.
9	FORMAT	Colon delimited list of the format of individual genotypes in the following fields.
10+	Sample(s)	Individual genotype information defined by FORMAT.

Formats: VCF

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Formats: VCF - Header

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Formats: VCF – Variant calls

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Formats: VCF – Chromosome and position

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Formats: VCF - ID

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC   G,GTCT 50   PASS   NS=3;DP=9;AA=G                GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Formats: VCF – Reference and alternate alleles

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Formats: VCF – Variant quality

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC  G,GTCT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

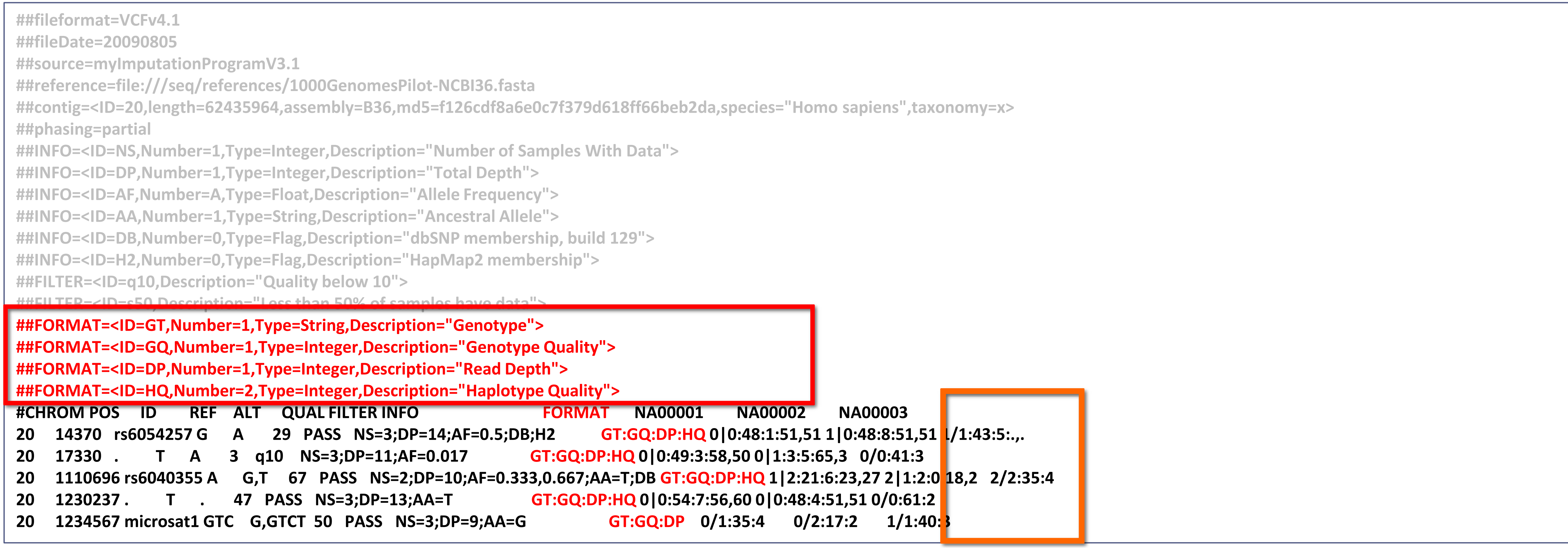
Formats: VCF – Filter

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC  G,GTCT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Formats: VCF – Variant information (across samples)

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

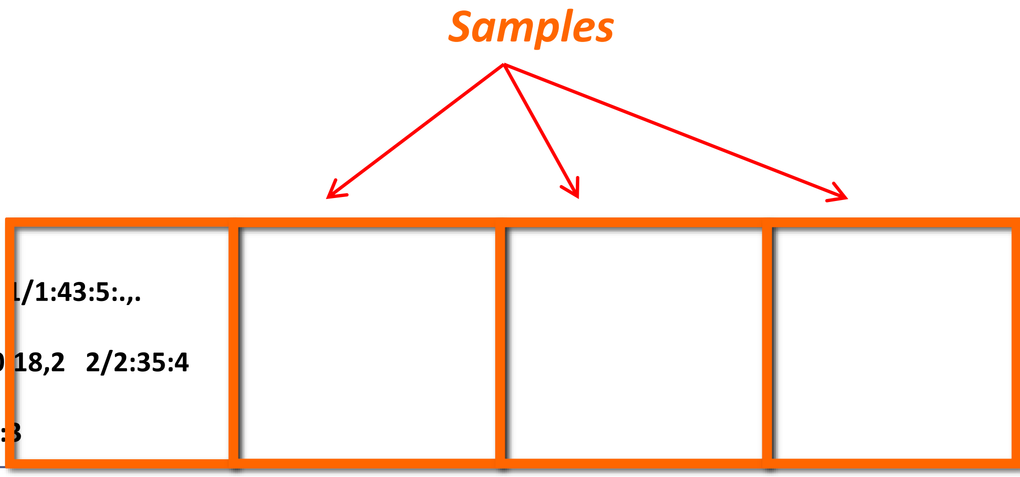
Formats: VCF - Per-sample format information



```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC  G,GTCT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Formats: VCF – Formats - Variant per-sample information

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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 GTC  G,GTCT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```



Annotations

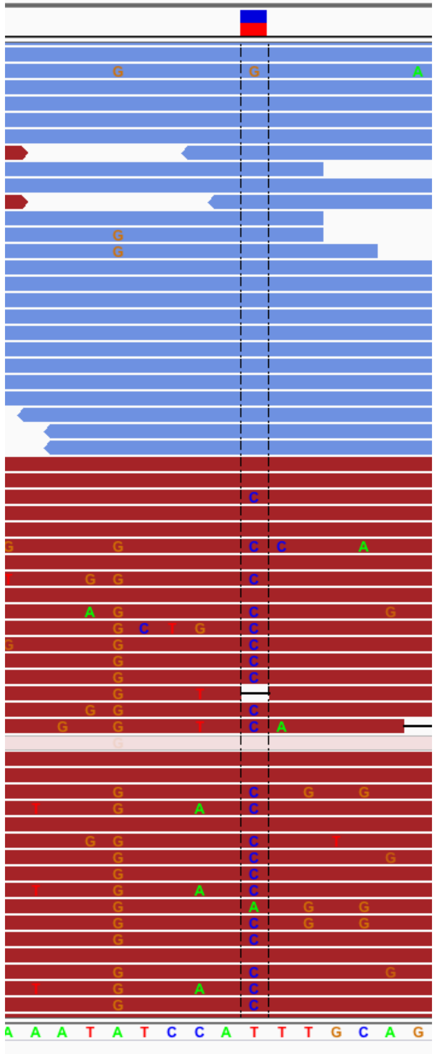
Additional data included for variants that help assess quality of the call

Complete list link (GATK)

Can include measures/scores for:

- Quality of variant call (QD)
- Total or allele-specific read depth
- Read base quality metrics
- Strand bias (FS) – at right
- Base quality (QD)
- Consanguinity (InbreedingCoefficient)
- Tumor/normal somatic calling information (TLOD, NLOD)

Strand Bias

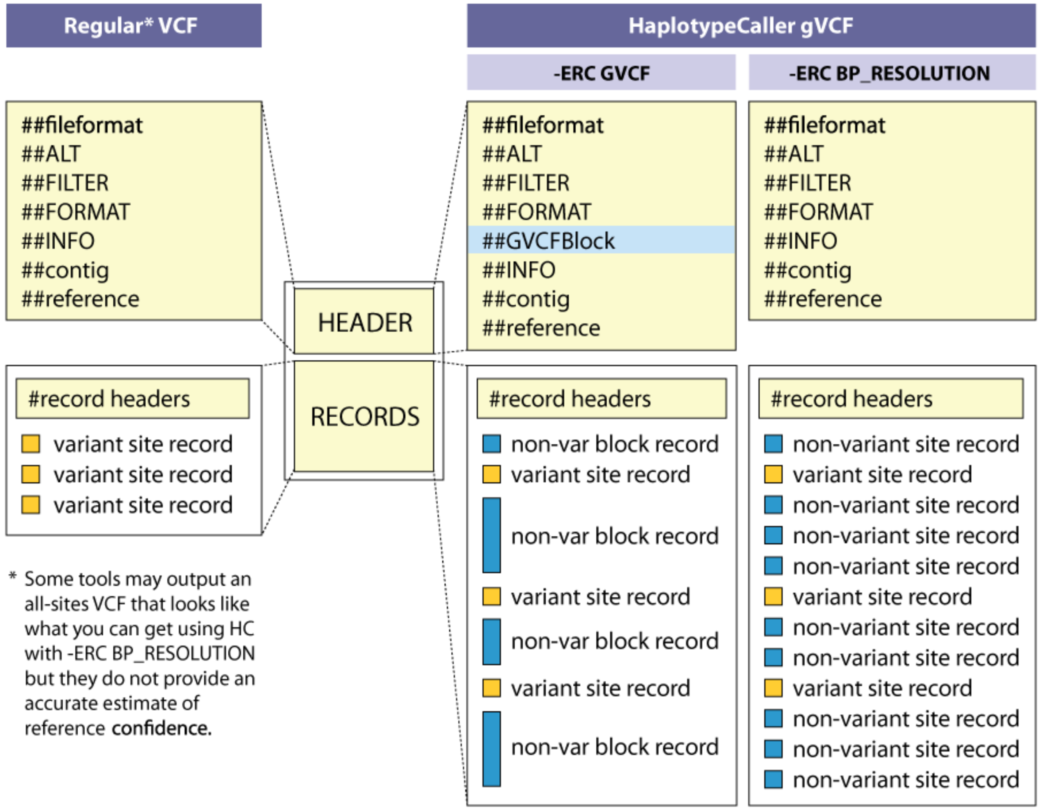


GVCF

Genomic VCF

A VCF file that contains a record for every site (regardless if there is a variant or not)

Highly recommended for multi-sample calling



Regular* VCF

##fileformat
##ALT
##FILTER
##FORMAT
##INFO
##contig
##reference

#record headers

🟨 variant site record
🟨 variant site record
🟨 variant site record

HEADER

RECORDS

HaplotypeCaller gVCF

-ERC GVCF

##fileformat
##ALT
##FILTER
##FORMAT
##GVCFBlock
##INFO
##contig
##reference

#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

-ERC BP_RESOLUTION

##fileformat
##ALT
##FILTER
##FORMAT
##INFO
##contig
##reference

#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.

Phase II : Filtering

Two basic methods:

- Hard filtering
- Variant quality score recalibration (VQSR)

Phase II : Hard Filtering

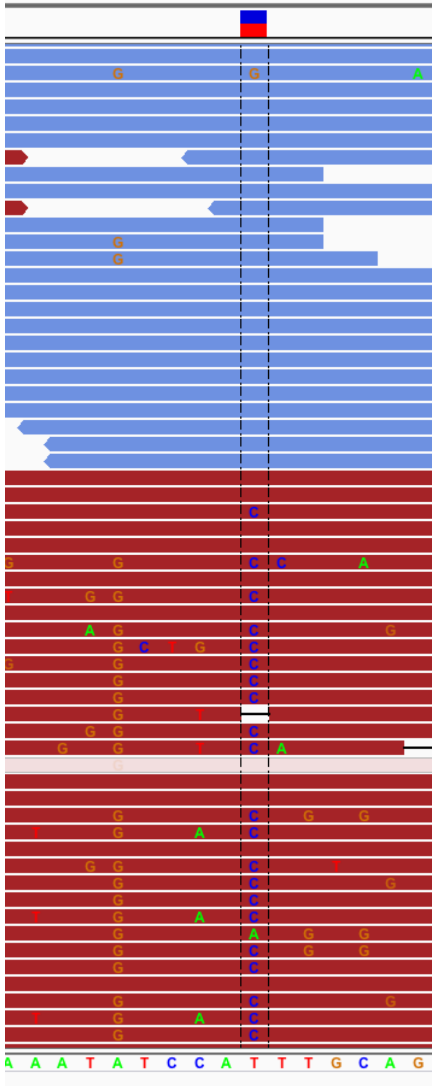
Reducing false positives by e.g. requiring

- Sufficient Depth
- Variant to be in >30% reads
- High quality
- **Strand balance**
- Etc etc etc

Very high dimensional search space

- … so, very subjective!

Strand Bias



Phase II : Hard Filtering

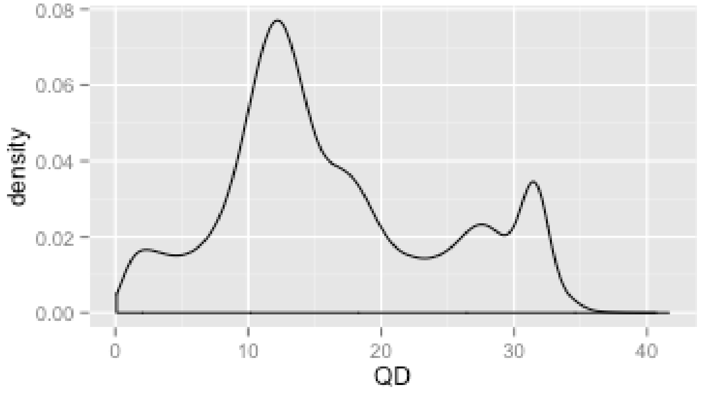
Great overview here

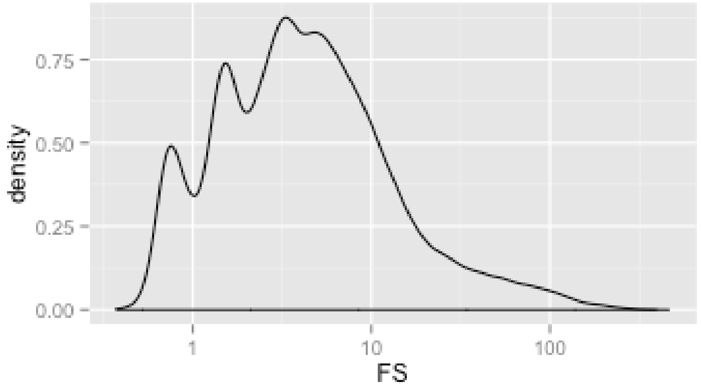
Some starting points here

Visualize distribution of annotation value, pick cutoff

Most informative annotations:

- QD – normalized quality
- FS – strand bias
- SOR - strand bias
- MQ – mapping qual of reads
- MQRankSum - mapping qual of reads
- ReadPosRankSum - position of alleles in read





Phase II : Variant Quality Score Recalibration (VQSR)

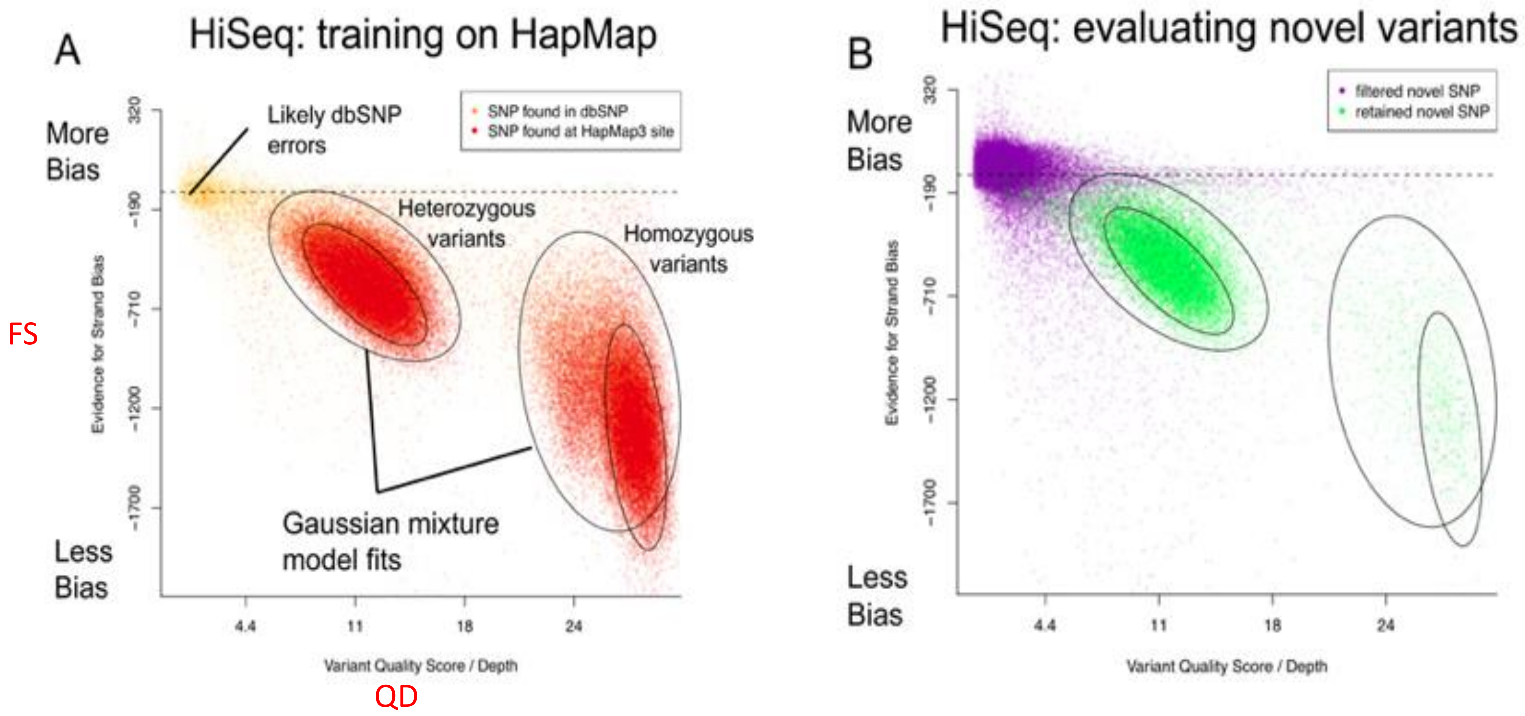
Considered GATK ‘best practice’

Train on trusted variants (e.g. HapMap)

Require the new variants to live in the same hyperspace

Potential problems:

- Over-fitting
- Biasing to features of known SNPs



Phase II : Variant Quality Score Recalibration (VQSR)

Considered GATK ‘best practice’

Train on trusted variants (e.g. HapMap)

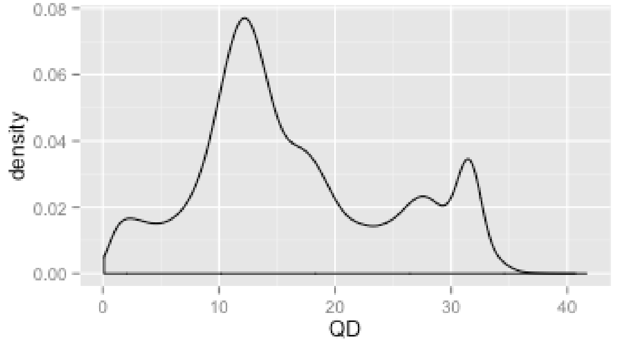
Require the new variants to live in the same hyperspace

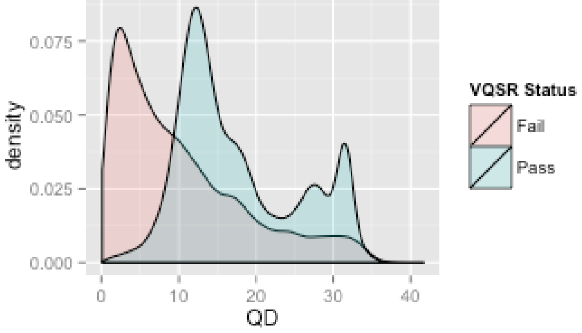
Potential problems:

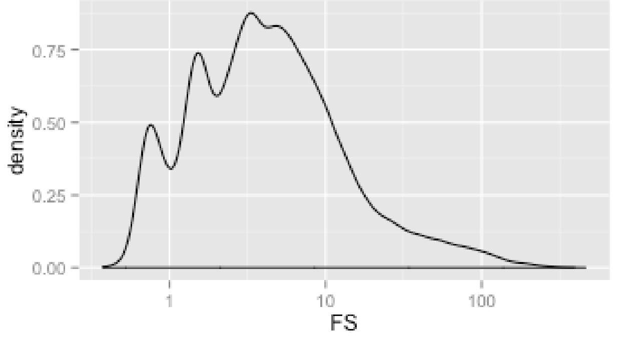
- Over-fitting
- Biasing to features of known SNPs

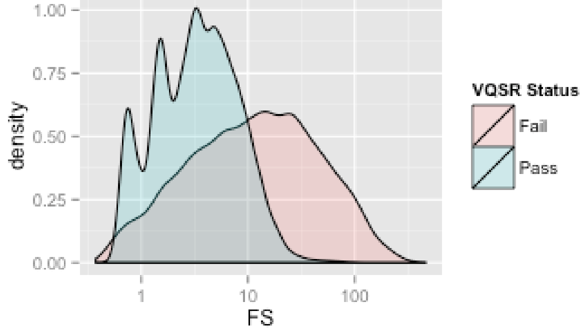
Hard Filtering

VQSR

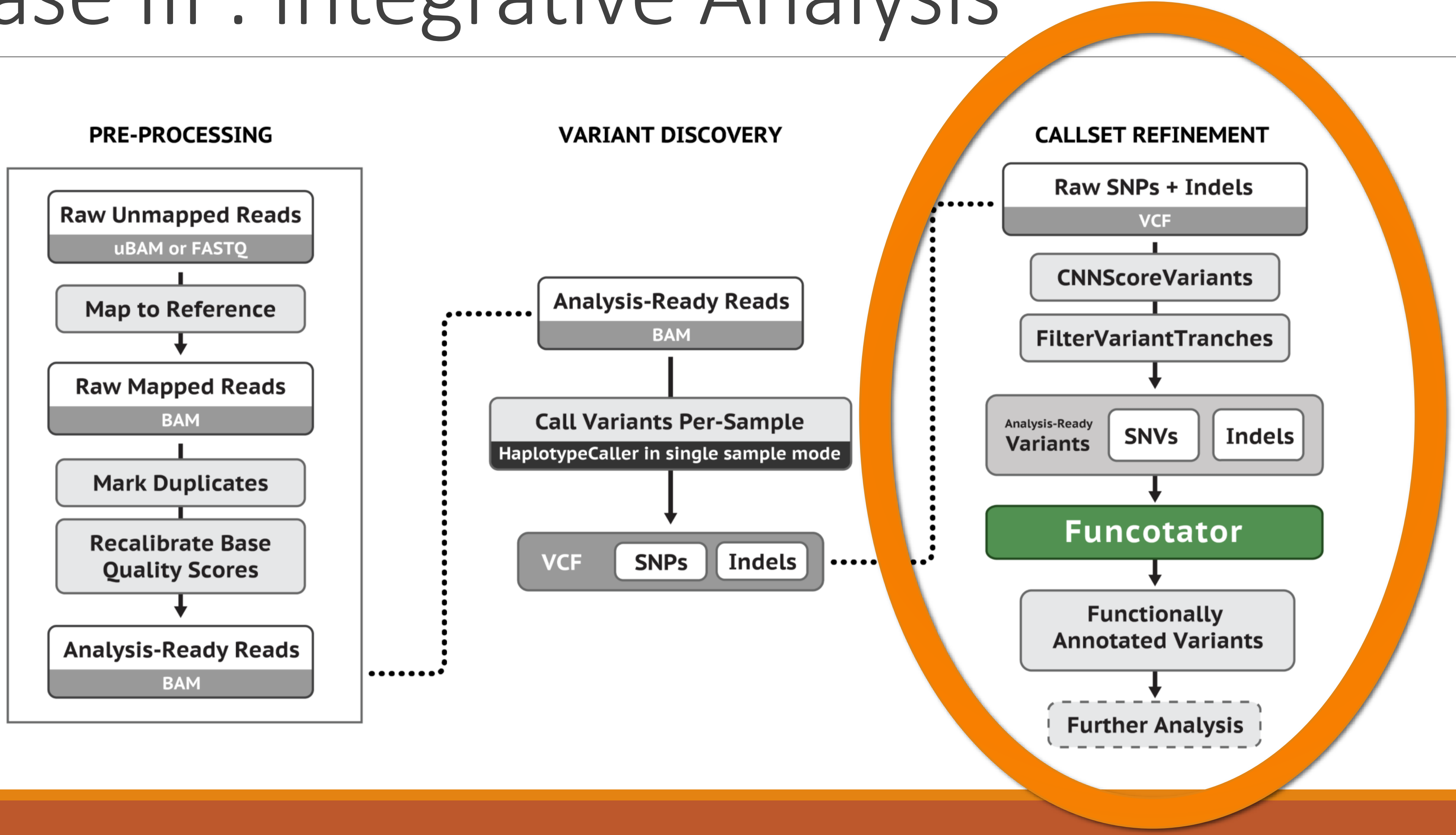








Phase III : Integrative Analysis



PRE-PROCESSING

Raw Unmapped Reads
uBAM or FASTQ

Map to Reference

Raw Mapped Reads
BAM

Mark Duplicates

Recalibrate Base Quality Scores

Analysis-Ready Reads
BAM

VARIANT DISCOVERY

Analysis-Ready Reads
BAM

Call Variants Per-Sample
HaplotypeCaller in single sample mode

VCF SNPs Indels

CALLSET REFINEMENT

Raw SNPs + Indels
VCF

CNNScoreVariants

FilterVariantTranches

Analysis-Ready Variants SNVs Indels

Funcotator

Functionally Annotated Variants

Further Analysis

Phase III : Functional Annotation

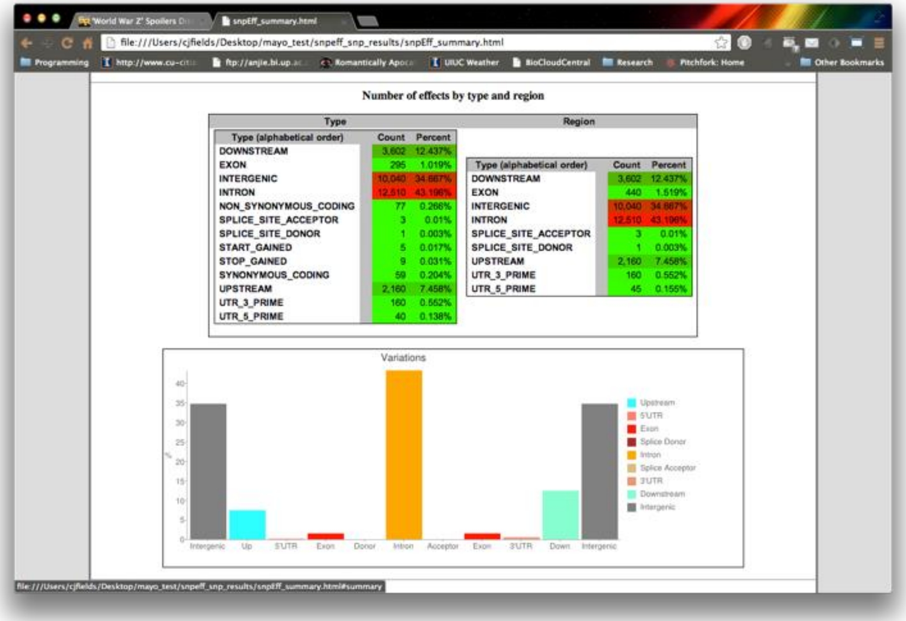
Are these mutations in important regions?

- Genes? UTR?
- Are they changing the coding sequence?

Would these changes have an affect?

Tools:

- SnpEff/SnpSift
- Annovar



Number of effects by type and region

Type

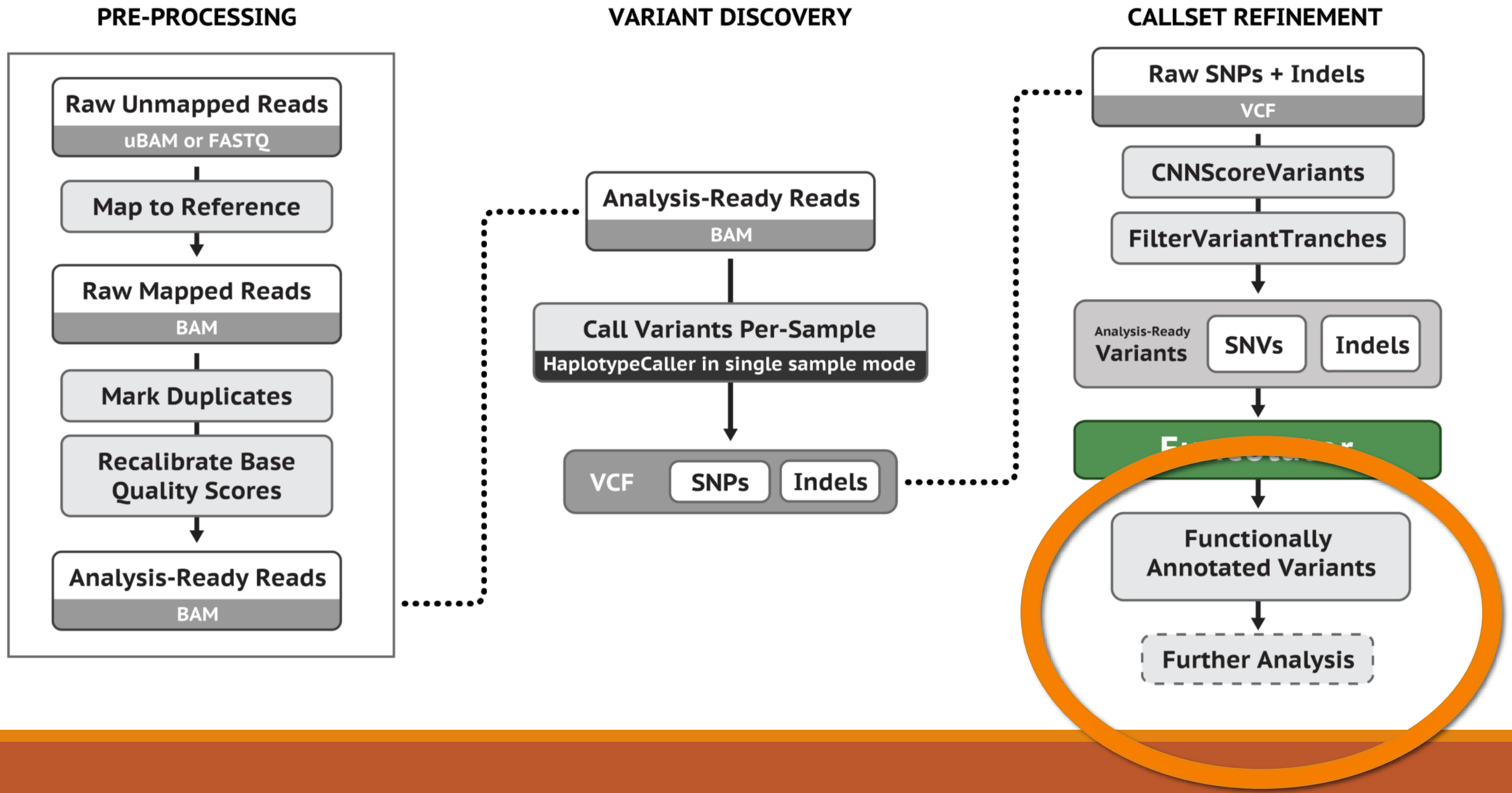
Type (alphabetical order)	Count	Percent
DOWNSTREAM	3,602	12.437%
EXON	295	1.019%
INTERGENIC	10,040	34.667%
INTRON	12,510	43.196%
NON_SYNONYMOUS_CODING	77	0.266%
SPLICE_SITE_ACCEPTOR	3	0.01%
SPLICE_SITE_DONOR	1	0.003%
START_GAINED	5	0.017%
STOP_GAINED	9	0.031%
SYNONYMOUS_CODING	59	0.204%
UPSTREAM	2,160	7.458%
UTR_3_PRIME	160	0.552%
UTR_5_PRIME	40	0.138%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	3,602	12.437%
EXON	440	1.519%
INTERGENIC	10,040	34.667%
INTRON	12,510	43.196%
SPLICE_SITE_ACCEPTOR	3	0.01%
SPLICE_SITE_DONOR	1	0.003%
UPSTREAM	2,160	7.458%
UTR_3_PRIME	160	0.552%
UTR_5_PRIME	45	0.155%

file:///Users/cjfields/Desktop/mayo_test/snpeff_snp_results/snpEff_summary.html#summary

The end of the (pipe)line



PRE-PROCESSING

Raw Unmapped Reads
uBAM or FASTQ

Map to Reference

Raw Mapped Reads
BAM

Mark Duplicates

Recalibrate Base Quality Scores

Analysis-Ready Reads
BAM

VARIANT DISCOVERY

Analysis-Ready Reads
BAM

Call Variants Per-Sample
HaplotypeCaller in single sample mode

VCF SNPs Indels

CALLSET REFINEMENT

Raw SNPs + Indels
VCF

CNNScoreVariants

FilterVariantTranches

Analysis-Ready Variants SNVs Indels

Funcotator

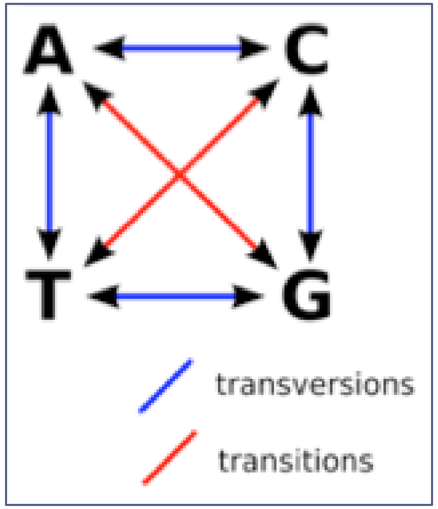
Functionally Annotated Variants

Further Analysis

Follow-up Quality Control

Transi**t**ion/**T**rans**v**ersion ratio (T_i/T_v)

Condition	Expected T_i/T_v
random	0.5
whole genome	2.1
exome	3.0-3.3



- ***bcftools*** can help here

Concordance with known variants: dbSNP, HapMap, 1000genomes

Lower than expected – possibly includes more false positives

Higher that expected – indicates potential bias

Calling variants on cohorts of samples

When running HaplotypeCaller, can use a specific output type called **GVCF (Genotype VCF)**

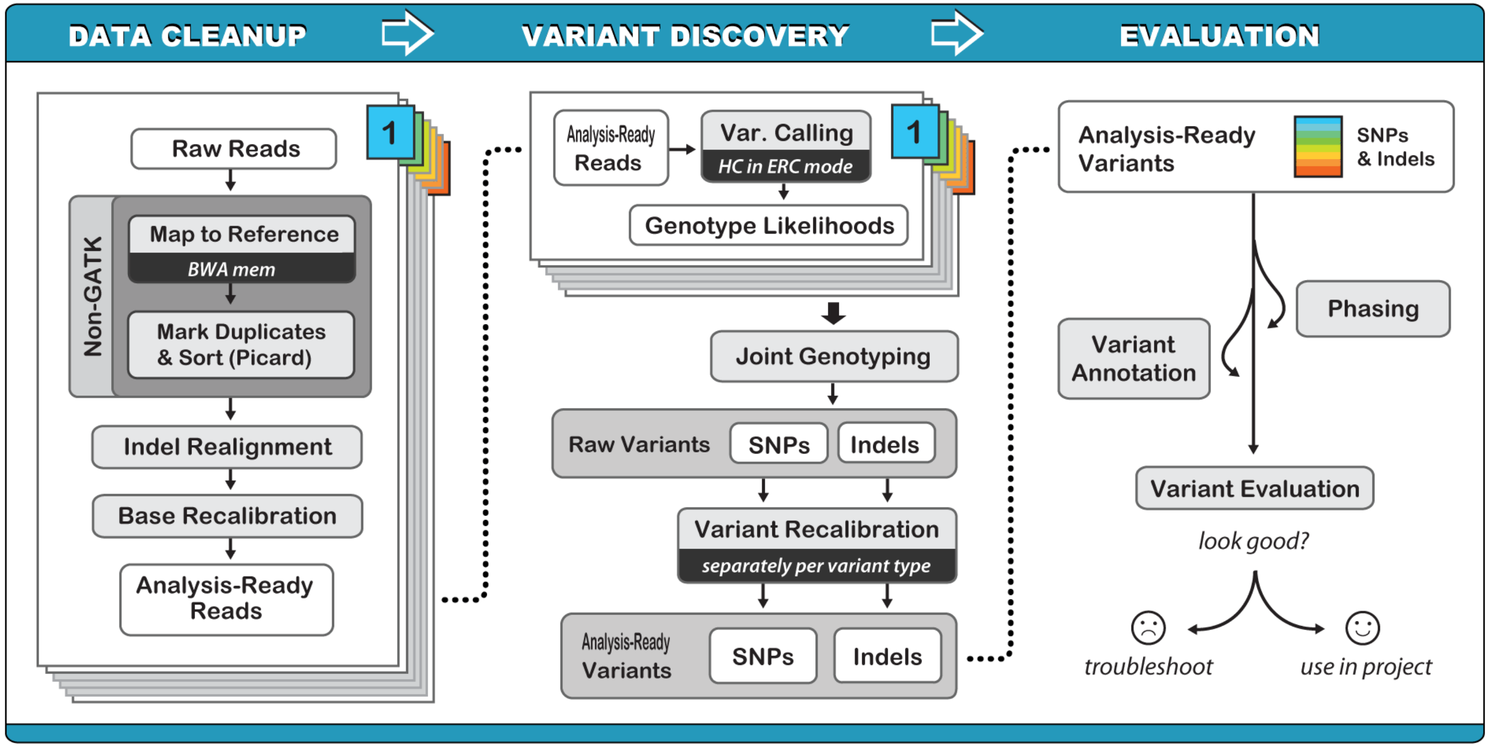
- Contains genotype likelihood and annotation for each site in genome

Perform joint genotyping calls on cohort

Can rerun as needed if more samples added to cohort

Used for ExAC cohort (92K exomes)

Link!



Bonus materials

Structural variants

Acknowledgments

Many figures/slides come from:

- GATK Workshop slides: http://www.broadinstitute.org/gatk/guide/events?id=2038
- IGV Workshop slides: http://lanyrd.com/2013/vizbi/scdttf/
- Denis Bauer (CSIRO): http://www.allpower.de/
- Many varied publications