华大基因研究院
BGI research

FASTA

Definition: In bioinformatics, FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes.

Example:

```
>gi|163051|gb|M63452.1|BOVFGG Bovine gamma globin gene and globin(PSI-2) pseudogene,complete cds
GGAGAGTAGAGTTTCTGAGTTTAGACACACTGAATCAGCCAATCACAGATGAAGAGCACTGAGCAACAAG
AGTTCATCTTACATTCCCCCAAACCAATGAACTTGTATTATGCCCTGGGCTAATCTGCTCTCAGAAGCAG
GGAGGGCAGGAGGCTGGGTGGGGCTCACAAGGAAGACCAGGGCCCCTACTGCTTACACATGCTTTTGACA
```

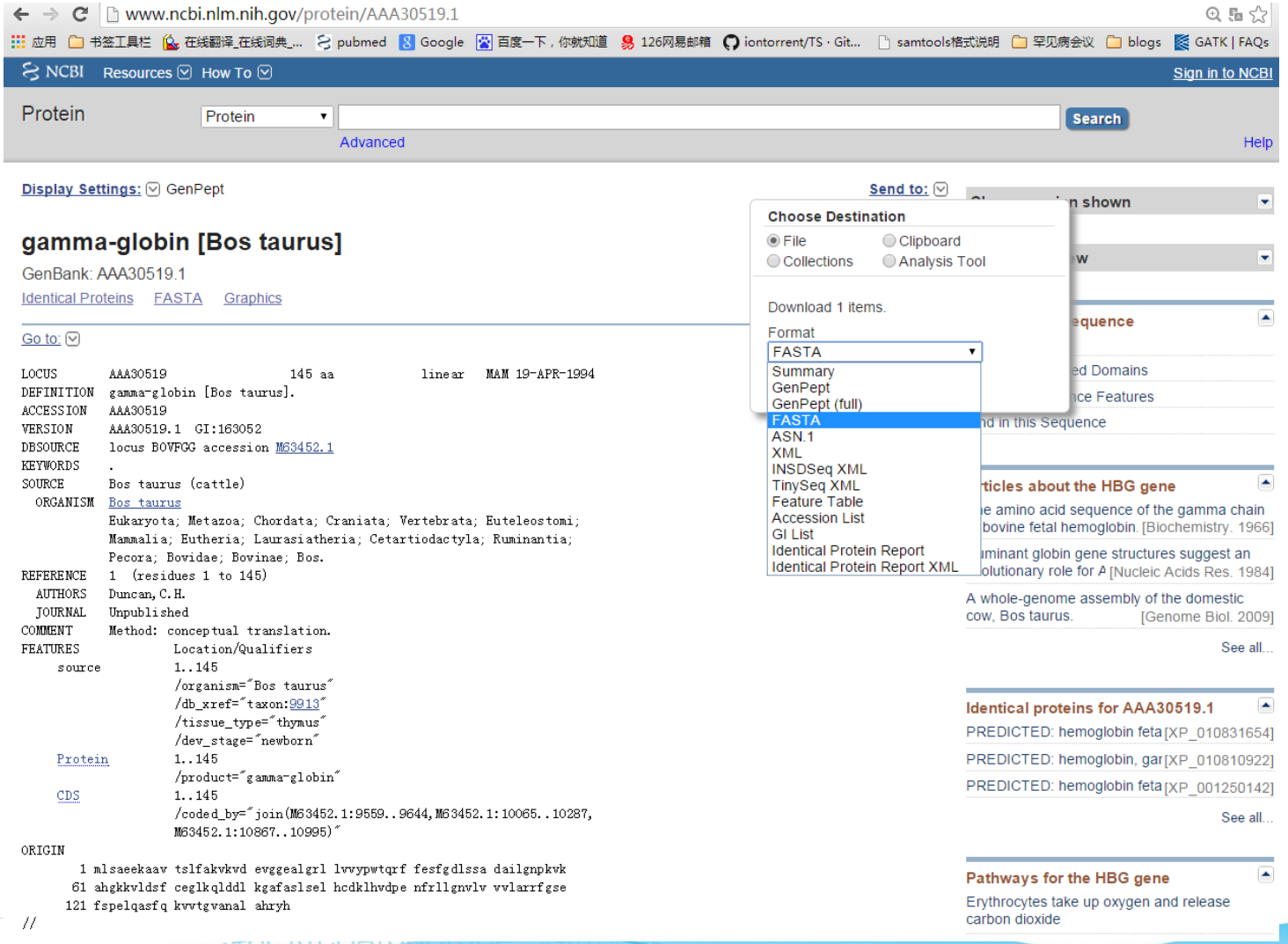
```
>gi|163052|gb|AAA30519.1| gamma-globin [Bos taurus]
MLSAEEKAAVTSLFAKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSSADAILGNPKVKAHGKKVLDSF
CEGLKQLDDLKGAFASLSELHCDKLHVDPENFRLLGNVLVVVLARRFGSEFSPELQASFQKVVTGVANAL
AHRYH
```

Feature: Begins with a single-line description whose first column is a greater-than (">") symbol, followed by lines of sequence data.

Extension	Meaning	Notes
fasta (.fas)	generic fasta	Any generic fasta file.
ffn	FASTA nucleotide coding regions	Contains coding regions for a genome.
fna	fasta nucleic acid	Used to generically specify nucleic acids.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA
faa	fasta amino acid	Contains amino acids. A multiple protein fasta file can have the more specific extension mpfa.

华大基因研究院
BGI research

FASTA



华大基因研究院
BGI research

FASTA

Reference: http://en.wikipedia.org/wiki/FASTA_format

Excise: 请下载人的gamma-globin核酸序列。

华大基因研究院
BGI research

FASTQ

Definition: FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are encoded with a single ASCII character for brevity.

Example:

```
@A60W0:1:1101:13989:1481#NTCTGCCT/1
AGGAACACATCATTTAAACAACTAGCTCTCAAGAATACTTTGAAATTGGTCTTTTTTAAATATTATTNTTNNNTTTTATTATTNTACT
+
BCBCBFFFCFFFGGGGGGGGGGHHHHHHHHHHHHHHHGHHHHHHHHHHHHHHGHHHHHGHIHHHHHHHH#BB###BBFFGHHGHH#BFFG
```

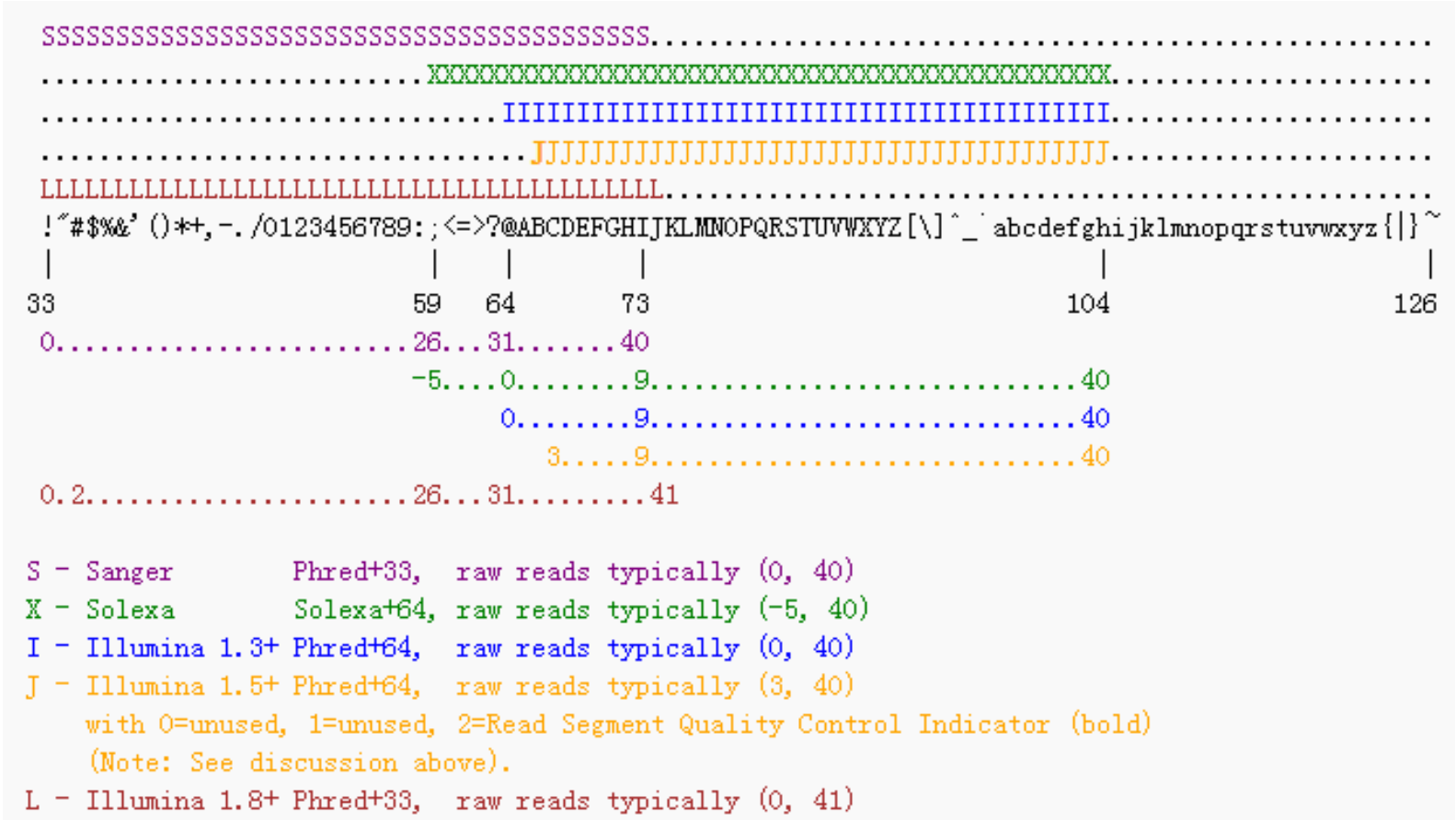
Feature:

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description.
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier.
- Line 4 encodes the **quality values** for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

  - **Phred quality scores** were originally developed by the program Phred to help in the automation of DNA sequencing in the Human Genome Project.
  - Phred quality scores are assigned to each nucleotide base call in automated sequencer traces.
  - If a quality score of 30 is assigned to a base, the chances that this base is called incorrectly are 1 in 1000.
  - The most commonly used method is to count the bases with a quality score of 20 and above.

华大基因研究院
BGI research

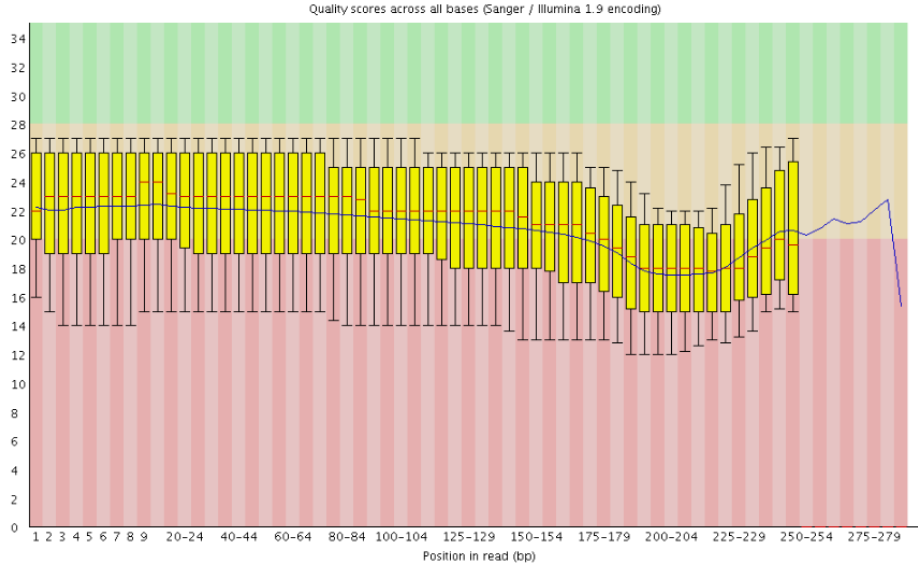
FASTQ



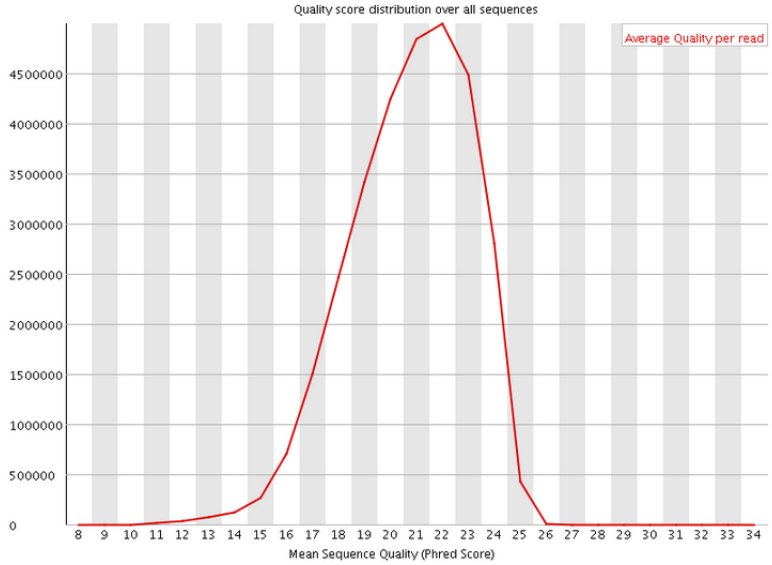
华大基因研究院
BGI research

FASTQ

Per base sequence quality



Per sequence quality scores



华大基因研究院
BGI research

FASTQ

Bioinformatics Advance Access published April 1, 2014

Genome Analysis

Trimmomatic: A flexible trimmer for Illumina Sequence Data

Anthony M. Bolger[1,2], Marc Lohse[1] and Bjoern Usadel[2,3,*]

[1]Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany.

[2]Institut für Biologie I, RWTH Aachen, Worringer Weg 3, 52074 Aachen, Germany.

[3]Institut of Bio- and Geosciences: Plant Sciences, Forschungszentrum Jülich, Leo-Brandt-Straße, 52425 Jülich, Germany

java -classpath <path to trimmomatic jar> org.usadellab.trimmomatic.TrimmomaticSE [-threads <threads>] [-phred33 | -phred64] [-trimlog <logFile>] <input> <output> <step 1> ...

-phred33 or -phred64 specifies the base quality encoding. If no quality encoding is specified, it will be determined automatically (since version 0.32). The prior default was -phred64.

华大基因研究院
BGI research

FASTQ

Reference: http://en.wikipedia.org/wiki/FASTA_format

http://en.wikipedia.org/wiki/Phred_quality_score

Excise: 请举例说明什么步骤可能会用到碱基的quality score。

华大基因研究院
BGI research

SAM / BAM

Definition: SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with `@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

华大基因研究院
BGI research

SAM / BAM

Example: (alignment section)

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

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

The corresponding SAM format is:

```
@HD VN:1.5 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 5S6M       *  0    0 GCCTAAGCTAA        * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref  16 30 6M14N5M    *  0    0 ATAGCTTCAGC        *
r003  2064 ref  29 17 6H5M       *  0    0 TAGGC              * SA:Z:ref,9,+,5S6M,30,1;
r001    83 ref  37 30 9M         =  7  -39 CAGCGGCAT          * NM:i:1
```

header

optional fields

alignment section

11 mandatory fields

华大基因研究院
BGI research

SAM / BAM

Feature: (header)

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
.............................
.............................
.............................
@SQ     SN:chr20        LN:63025520
@SQ     SN:chr21        LN:48129895
@SQ     SN:chr22        LN:51304566
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@SQ     SN:chrM LN:16569
@RG     ID:0    PL:ILLUMINA     PU:0    LB:bar  SM:Amplicon
@PG     ID:bwa  PN:bwa  VN:0.7.10-r789  CL:/ifs2/BC_MD/DEV/WorkF
low/Sofware/bwa-0.7.10/bwa sampe -n 1 hg19.fa output1P.fq.gz.sai
 output2P.fq.gz.sai output1P.fq.gz output2P.fq.gz
```

@HD: The first line if present.
"VN"– Format version; *
"SO"– Sorting order of alignment;

@SQ: Reference sequence dictionary.
"SN" – Reference sequence name; *
"LN" – Reference sequence length; *

@RG: Read group.
"ID" – Read group identifier; *
"PL" – Platform used to produce the reads; exclusive to several values;
"LB" – library; important for de-dup;
"SM" – Sample;

@PG – Program;
"ID" – Program record identifier; *
"PN" – Program name;
"CL" – Command line;

Tags with "*" are required when the record type is present.

华大基因研究院
BGI research

SAM / BAM

Feature: (Alignment section)

```
60W0:1:1101:28181:16441#GTCTGCCT    QNAME
113         FLAG
chr1        RNAME
436459      POS
0           MAPQ
150M        CIGAR
chr8        RNEXT
46543       PNEXT
0           TLEN
TCAGCCTCCGGAGTAGCTGGGACTACAGGCATGCCCCACCACATTCGGCTAATTTTTAAAATTTTTTTGTAGAGACAGGGTCTCACTATA
TTGTCCAGGCTGCTCTCAAAATCCTGGCCTAAAGTGATCCTCCTGCCTCAGCCTCCTAAG    SEQ
HHHGGFDGGHHHHHHHHHHHHHHHHHHHHHHGGGGEEGHHFHGGGFHHHHHHHHHHGHHHHGGHHHHHHHHHHHHHFHFFHHHGFHHHHH
HHHHHHHHHHHHHHHGGHHHHGDHHHHGHHHHHHHHHHGGGGGGGGGGGGFCCCFFFCCBCC    QUAL
XT:A:R      X? reserved fields for end users
NM:i:1      NM: Edit distance to the reference, including ambiguous bases but excluding clipping
SM:i:0      SM: Template-independent mapping quality
AM:i:0      AM: The smallest template-independent mapping quality of segments in the rest
X0:i:4
X1:i:1
XM:i:1
XO:i:0
XG:i:0
MD:Z:147G2  MD: String for mismatching positions. It aims to achieve SNP/indel calling without looking at the reference
XA:Z:chr8,+46464,150M,1;chr5,-180863013,150M,1;chr1,+547999,150M,1;chr6,-171020695,150M,2;
```

华大基因研究院
BGI research

SAM / BAM

Feature: (Alignment section, mandatory fields)

QNAME: Query template NAME; Reads/segments having identical QNAME are regarded to come from the same template. Template is a DNA/RNA sequence part of which is sequenced on a sequencing machine or assembled from raw sequences.

Flag: Bitwise FLAG. Each bit has different meaning.

113	FLAG
chr1	RNAME
436459	POS
0	MAPQ

Flag: 113 Explain

Explanation:

- [x] read paired
- [ ] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [x] read reverse strand
- [x] mate reverse strand
- [x] first in pair
- [ ] second in pair
- [ ] not primary alignment
- [ ] read fails platform/vendor quality checks
- [ ] read is PCR or optical duplicate
- [ ] read paired
- [ ] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [ ] read reverse strand

RNAME: Reference sequence NAME of the alignment.
An unmapped segment without coordinate has a "*" at this field.

POS: 1-based leftmost mapping POSition of the first matching base. (1-based VS. 0-based)

MAPQ: MAPing Quality. Mapping quality scores are computed differently by each aligner.

CIGAR: CIGAR string.

150M	CIGAR
chr8	RNEXT
46543	PNEXT

RNEXT: Reference sequence name of the primary alignment of the NEXT read in the template.

PNEXT: Position of the primary alignment of the NEXT read in the template.

TLEN: Signed observed Template LENth. If all segments are mapped to the same reference, the unsigned observed template length equals the number of bases from the leftmost mapped base to the rightmost mapped base. The leftmost segment has a plus sign and the rightmost has a minus sign. The sign of segments in the middle is undefined. It is set as 0 for single-segment template or when the information is unavailable. Segment is a contiguous sequence or subsequence.

0	TLEN

SEQ: segment SEQuence.

QUAL: ASCII of base QUALity plus 33 (same as the quality string in the Sanger FASTQ format).

华大基因研究院
BGI research

SAM / BAM

Feature: (Alignment section, mandatory fields)

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

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

- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

华大基因研究院
BGI research

SAM / BAM

Feature: (Alignment section, optional fields)

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

MD: The MD field aims to achieve SNP/indel calling without looking at the reference. For example, a string `10A5^AC6‘ means from the leftmost reference base in the alignment, there are 10 matches followed by an A on the reference which is different from the aligned read base; the next 5 reference bases are matches followed by a 2bp deletion from the reference; the deleted sequence is AC; the last 6 bases are matches. The MD field ought to match the CIGAR string.

华大基因研究院
BGI research

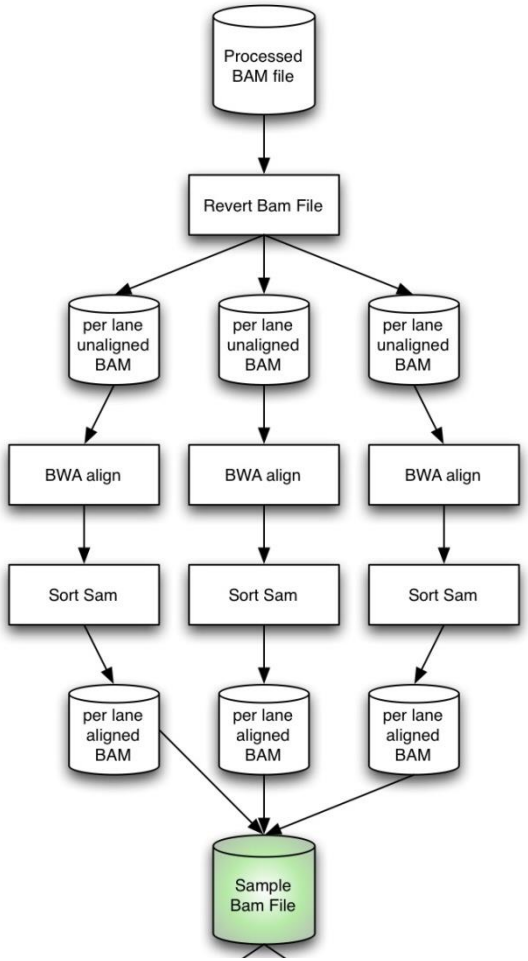
SAM / BAM

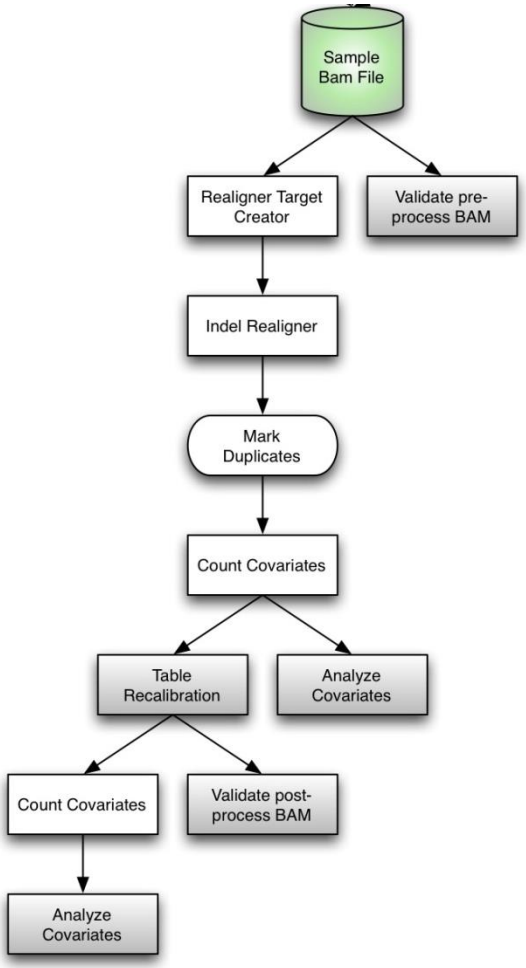
Samtools: (Tools for alignments in the SAM format)

```
Command:  view         SAM<->BAM conversion ***
          sort         sort alignment file ***
          mpileup      multi-way pileup *
          depth        compute the depth
          faidx        index/extract FASTA *
          tview        text alignment viewer *
          index        index alignment ***
          idxstats     BAM index stats (r595 or later)
          fixmate      fix mate information
          flagstat     simple stats
          calmd        recalculate MD/NM tags and '=' bases
          merge        merge sorted alignments **
          rmdup        remove PCR duplicates *
          reheader     replace BAM header
          cat          concatenate BAMs
          targetcut    cut fosmid regions (for fosmid pool only)
          phase        phase heterozygotes
```

华大基因研究院
BGI research

SAM / BAM





华大基因研究院
BGI research

SAM / BAM

Reference:

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

https://www.broadinstitute.org/gatk/guide/tagged?tag=workflow

Li Heng, *et al*. The Sequence Alignment/Map format and SAMtools, *Bioinformatics*. Aug 15, 2009; 25(16): 2078–2079.

Excise:

- How to convert SAM to BAM or the reverse?
- How can you get FASTQ data into BAM format and the reserve?
- How to sort an unorder BAM into a sorted BAM?
- What is the canonical ordering of human reference contigs in a BAM file?
- How can you tell if a BAM file has read group and sample information?
- How can you know if your BAM file is valid for the downstream analysis?
- How to extract the unmapped reads with awk?

华大基因研究院
BGI research

VCF/BCF

Definition: VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position. BCF is a binary, compressed equivalent of VCF that can be indexed with tabix and can be efficiently decoded from disk or streams.

Example:

```
##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
```

华大基因研究院
BGI research

VCF/BCF

Definition: VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position. BCF is a binary, compressed equivalent of VCF that can be indexed with tabix and can be efficiently decoded from disk or streams.

Example:

```
##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
```

Meta-information lines

Header line

Data lines

华大基因研究院
BGI research

VCF/BCF

Feature: Meta-information lines

version

FILTERs applied to the data

genotype-level

position-level

programs

contigs

Reference

```
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=VQSRTrancheSNP99.00to99.90,Description="Truth sensitivity tranche level for SNP mod
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for geno
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele,
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the sam
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have bee
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test o
##INFO=<ID=VQSLOD,Number=1,Type=Float,Description="Log odds ratio of being a true variant versus
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[chr1.sort.fix.brecal.bam] read_bu
##ApplyRecalibration="analysis_type=ApplyRecalibration input_file=[] read_buffer_size=null phone
##contig=<ID=chr1,length=249250621,assembly=hg19>
##contig=<ID=chr2,length=243199373,assembly=hg19>
.................................................
##contig=<ID=chr22,length=51304566,assembly=hg19>
##contig=<ID=chrX,length=155270560,assembly=hg19>
##contig=<ID=chrY,length=59373566,assembly=hg19>
##contig=<ID=chrM,length=16569,assembly=hg19>
##reference=file:///ifs1/ST_RNA/USER/liyaqiao/ZPY/rna/ref/Homo_genome/hg19_chunsheng/hg19.fasta
```

华大基因研究院
BGI research

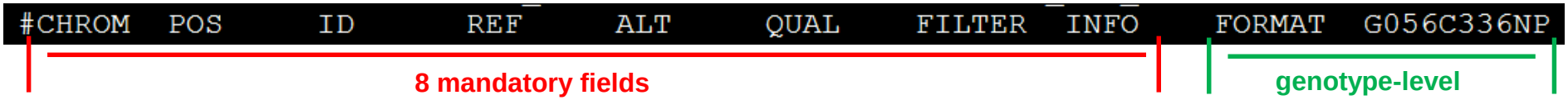
VCF/BCF

Feature: Header line

The header line names the 8 fixed, mandatory columns. These columns are as follows:

1. #CHROM
2. POS
3. ID
4. REF
5. ALT
6. QUAL
7. FILTER
8. INFO

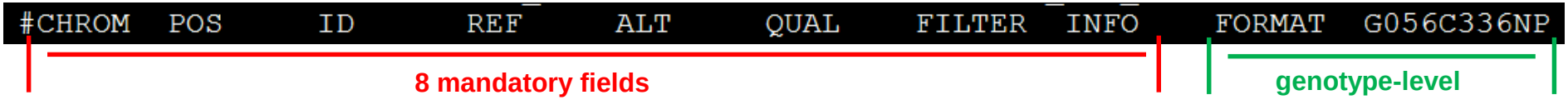
If genotype data is present in the file, these are followed by a FORMAT column header, then an arbitrary number of sample IDs. The header line is tab-delimited.



华大基因研究院
BGI research

VCF/BCF

Feature: Data lines




```
CHROM   chr1   Chromosome, an identifier from the reference genome or  ID String ("<ID>") pointing to a contig in the assembly file.
POS   1141608  The reference position, with the 1st base having position 1
ID   .    If this is a dbSNP variant it is encouraged to use the rs number(s).
REF   G   Reference base(s)
ALT   A   Alternate base(s)
QUAL   12.05  Phred-scaled quality score for the assertion made in ALT
FILTER   LowQual   Filter status, PASS if this position has passed all filters
INFO   AC=1;AF=0.500;AN=2;BaseQRankSum=-7.360e-01;ClippingRankSum=-7.360e-01;DP=3;FS=0
       .000;GQ_MEAN=29. 00;MLEAC=1;MLEAF=0.500;MQ=60.00;MQ0=0;MQRankSum=-7.360e-01;NCC
       =0;QD=4.02; ReadPosRankSum=0.736  Encoded as a semicolon-separated series of short keys with optional values
FORMAT   GT:AD:DP:GQ:PL
G056C336NP   0/1:1,2:3:29:40,0,29
```

华大基因研究院
BGI research

VCF/BCF

Feature: Genotype representation

```
chr1    873762  .       T   G   [CLIPPED] GT:AD:DP:GQ:PL    0/1:173,141:282:99:255,0,255
chr1    877664  rs3828047   A   G   [CLIPPED] GT:AD:DP:GQ:PL    1/1:0,105:94:99:255,255,0
chr1    899282  rs28548431  C   T   [CLIPPED] GT:AD:DP:GQ:PL    0/1:1,3:4:25.92:103,0,26
```

GT : The genotype of this sample. For a diploid organism, the GT field indicates the two alleles carried by the sample, encoded by a 0 for the REF allele, 1 for the first ALT allele, 2 for the second ALT allele, etc. When there's a single ALT allele (by far the more common case), GT will be either:

0/0 - the sample is homozygous reference
0/1 - the sample is heterozygous, carrying 1 copy of each of the REF and ALT alleles
1/1 - the sample is homozygous alternate In the three examples above, NA12878 is observed with the allele combinations T/G, G/G, and C/T respectively.

GQ: The Genotype Quality, or Phred-scaled confidence that the true genotype is the one provided in GT. The GQ is simply the second most likely PL - the most likely PL. Because the most likely PL is always 0, GQ = second highest PL - 0. If the second most likely PL is greater than 99, we still assign a GQ of 99, so the highest value of GQ is 99.

AD and DP: AD is also known as allele depth. It gives the unfiltered count of reads that support a given allele for an individual sample. The values in the field are ordered to match the order of alleles specified in the REF and ALT fields: REF, ALT1, ALT2 and so on if there are multiple ALT alleles. At the sample level (FORMAT), the DP value is the count of reads that passed the caller's internal quality control metrics. At the site level (INFO), the DP value is the unfiltered depth over all samples.

PL: This field provides the likelihoods of the given genotypes (here, 0/0, 0/1, and 1/1). These are normalized, Phred-scaled likelihoods for each of the 0/0, 0/1, and 1/1, without priors. The most likely genotype (given in the GT field) is scaled so that it's P = 1.0 (0 when Phred-scaled), and the other likelihoods reflect their Phred-scaled likelihoods relative to this most likely genotype.

华大基因研究院
BGI research

VCF / BCF

Reference:

http://samtools.github.io/hts-specs/VCFv4.1.pdf

https://www.broadinstitute.org/gatk/guide/tagged?tag=vcf

Excise:

- Please calculate Ti/Tv for a paritcular VCF file