

NGS output

NGS output

- Millions of sequences -> size of file up to GBytes
- Output format:
 - Fasta
 - FASTQ \approx Fasta + Quality (goodness of base call)

FASTQ files

- Name: FASTQ \approx Fasta + Quality (goodness of base call)
- Structure:
 - 1) '@' sequence identifier and description
 - 2) Raw sequence (in letters)
 - 3) '+' sequence identifier (again) – opt.
 - 4) Quality score per nucleotide, char encoded

```
@M70273:8:000000000-AJLMP:1:1101:14452:1861 1:N:0:1
TAACTACTTTGGGAATGTTAGCCTGGACAAACAATTTGATGAATGTCGTGTTTCTTTCTGAATT
+
5,,5</5<@A--+,+6-AC/.88A,+6-, -7,+7+8AC..9...9..9-.88CAEFFFECE---5A
@M70273:8:000000000-AJLMP:1:1101:14458:1948 1:N:0:1
CAGTGAAACGATATACTCCAGCCCGCATTGCCCTGGGCTGCCAGGGTGCCAAACCAAGGAACCTCTT
+
====-99/@@@@@AAE8C;-8C>CC7EE-9.977+++7++A--++555@A-55>A+,+,-,AFFFE
@M70273:8:000000000-AJLMP:1:1101:14505:2082 1:N:0:1
GTGCTGTTTCATCACTGTGCCATTGCAGGTTTATTTGAAATACAACAATGTCCAAGAGGAAAGCACTGC
+
????B??B?BBBBBBBFBFFHHHHFFHHHHFH009EFFHDFEFEG@FHHFGFD?D-CEFFHDFE
@M70273:8:000000000-AJLMP:1:1101:14399:2091 1:N:0:1
TGCCTCCCTTTCCAATGGACTATTTAGAGAAGAAATGGAGCTGTCACCCACATCAAGATTGAGAACACTG
+
????ABA?DDDDDDDFGGGFFIIIIHHIIFHHII@FHHIIIIIGFF>EHHFFGHHIFHFGHAFGH
@M70273:8:000000000-AJLMP:1:1101:16927:2095 1:N:0:1
CCTATCATATATGCCTTAGTTTTGATGAAANATATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
??AA?BBBEDDEEEEEGGGGGIIIIIII#7AFHII#####5#####
@M70273:8:000000000-AJLMP:1:1101:18171:2095 1:N:0:1
TTGTGATCCACATTCTCTCCATTGTAGNGCAAATNNNNNNNNNNNNNNNNNNNNNTNNTCNTTNNNTNN
+
????BBBDDDDDDGGGGGIIIIHHI#7AEFHI#####7###5#5###5##
@M70273:8:000000000-AJLMP:1:1101:19337:2095 1:N:0:1
GCCGCCATGCGCGGCATGATGAACTCCGCTGCTGNNNNNNNNNNNNNNNNNNNNNTNNTTNTTNNNCAN
+
????ABAADDDDDDFFFFHIIIIHHHHHHI#####5###5#5###44#
@M70273:8:000000000-AJLMP:1:1101:14484:2097 1:N:0:1
CTGGACTGATATGTGATTTATCTTTCAACAGCCACGGCCAGATCCAGTGAAAAACAAGCTCTCATGTC
+
????A?BB?DDDDDDDBGGGGGIIIIIIIIIIIIHHHHHFFGHHHHHFHHHHHHHIHHIHHIHHI
@M70273:8:000000000-AJLMP:1:1101:16321:2100 1:N:0:1
TAGATGCTTTTAAACTAAGTTACCTGACTTNCCTTATNNNNNNNNNNNNNNNNNNNTNNGCNGCNCNN
+
?????BBBDDDDDDDFGGGIIIFHHI#7AFHFG#####7###5#5###5##
```

Phred Quality Score

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	99.9%

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

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

SAM Format

Sequence Alignment/Map (SAM) Format

SAM Format

- TAB-delimited text
- header section (optional): lines start with '@'
- alignment section with 11 mandatory fields

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

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

```
@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 AAAAGATAAAGGATA *
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 *
```

SAM Format

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2 ¹⁶ -1]	bitwise FLAG
3	RNAME	String	* [!-()+-<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 ²⁹ -1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2 ⁸ -1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	* = [!-()+-<>-~] [!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	[0,2 ²⁹ -1]	Position of the mate/next segment
9	TLEN	Int	[-2 ²⁹ +1,2 ²⁹ -1]	observed Template LENgth
10	SEQ	String	* [A-Za-z=.]+	segment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

VCF Format

The Variant Call Format

VCF files

- Lines starting with `##`: arbitrary number of meta-information lines
- Line starting with `#`: column definition (8 mandatory):
 - CHROM = chromosome
 - POS = start position of the variant
 - ID = unique identifier of the variant (e.g. Number for SNPs)
 - REF = reference allele
 - ALT = comma separated list of alternate alleles
 - QUAL = phred-scaled quality score
 - FILTER = site filtering information
 - INFO = user extensible annotation (e.g. snpEff, Annovar)
 - • FORMAT = an (optional) extensible list of fields for describing the SAMPLE column
 - • SAMPLE COLUMN = free

VCF Format

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Mandatory header lines (indicated by a red arrow pointing to the first line)

Optional header lines (meta-data about the annotations in the VCF body) (indicated by a black arrow pointing to the remaining header lines)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Reference alleles (GT=0) (indicated by a blue arrow pointing to the first '0' in the first row)

Alternate alleles (GT>0 is an index to the ALT column) (indicated by a blue arrow pointing to the '1' in the first row)

Deletion (indicated by a blue arrow pointing to the in the last row)

SNP (indicated by a blue arrow pointing to the 'A,AT' in the first row)

Large SV (indicated by a blue arrow pointing to the 'T,CT' in the second row)

Insertion (indicated by a blue arrow pointing to the 'G' in the third row)

Other event (indicated by a blue arrow pointing to the 'CT' in the second row)

Phased data (G and C above are on the same chromosome) (indicated by a blue arrow pointing to the '|1:100' in the second row)