

# NGS - quality control, alignment, visualisation

File types

# File types

fasta	sequences
fastq	reads
sam/bam	alignments
bed	regions
gff	annotations
vcf	variants

# fasta

- Plain sequence: \*.fasta or \*.fa
- Nucleotides or amino acids (proteins)
- Useful command:

```
grep -c "^>" sequence.fasta
```

```
sequence.fasta
```

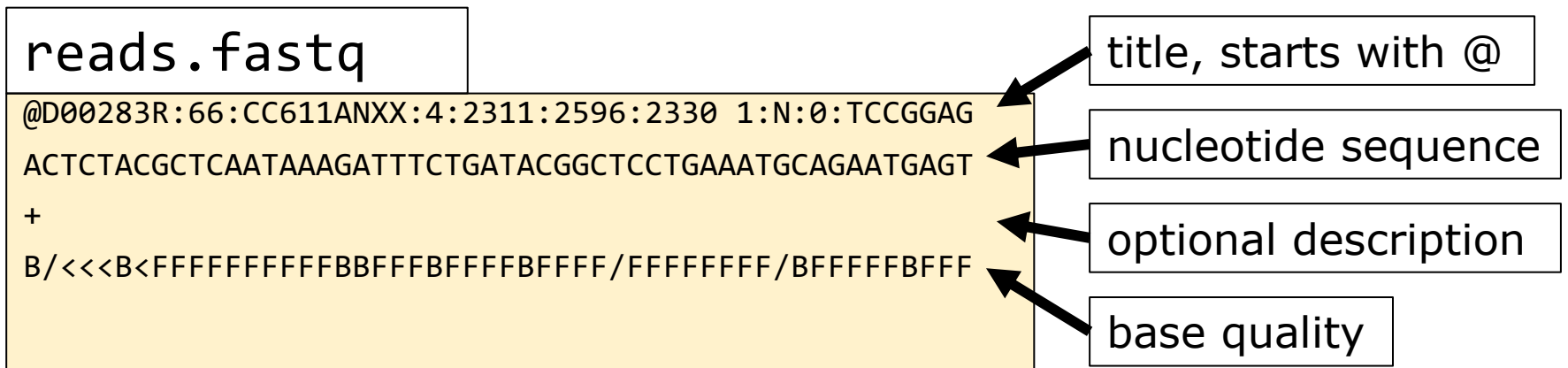
```
>sequence title1
```

```
ATCGTATCT
```

```
>sequence title2
```

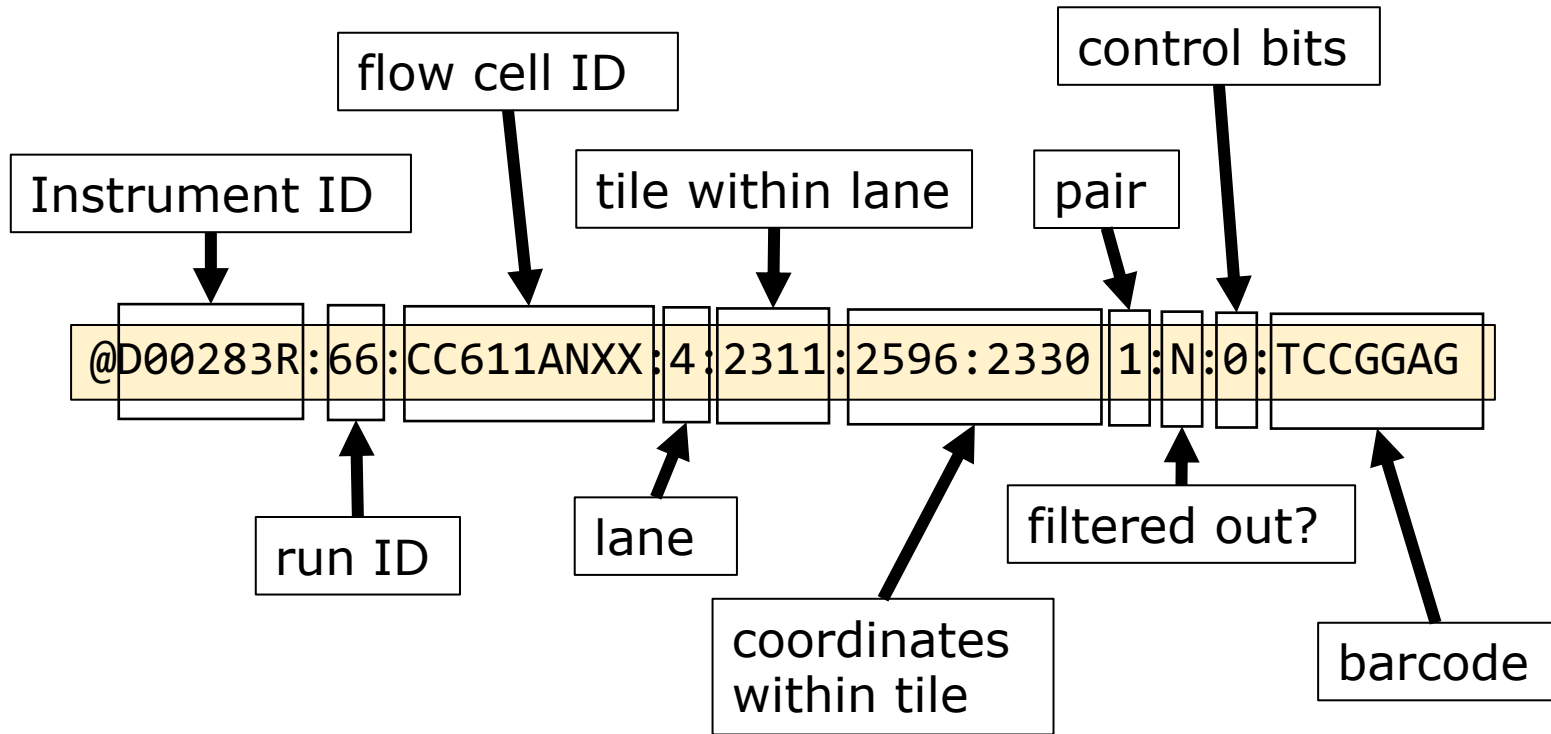
```
ATGATGACGT
```

# fastq files



!"#\$%&'()\*+,-./0123456789:;<=>?@ABCDEFGHI  
| | | | |  
0.2.....26...31.....41

# fastq header



# Quiz Question 9

# Sam

sequence alignment format

Aim: alignments

# sam header

```
@HD      VN:1.0  SO:coordinate
@SQ      SN:U00096.3      LN:4641652
@PG      ID:bowtie2      PN:bowtie2      VN:2.4.1
CL: "/opt/miniconda3/envs/ngs/bin/bowtie2-align-s \
--wrapper basic-0 \
-x /home/ubuntu/ecoli/ref_genome//ecoli-strK12-MG1655.fasta \
-1 /home/ubuntu/ecoli/trimmed_data/paired_trimmed_SRR519926_1.fastq \
-2 / home/ubuntu/ecoli/trimmed_data/paired_trimmed_SRR519926_2.fastq"
```

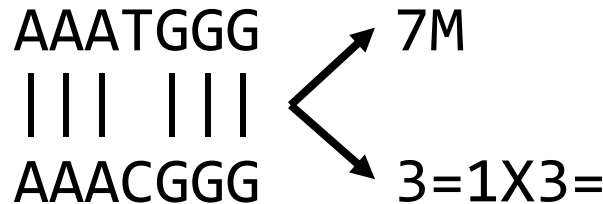


<b>SAM column</b>	<b>example</b>
read name	@D00283R:66:CC611..
flag	89
reference	U00096.3
start position	12513
mapping quality	42
CIGAR string	5M2D7M3S
reference name mate is mapped	=
start position mate	12928
fragment length	430
sequence	CATCACCATTCCCAC
base quality	@>4:4C@89+&9CC@
optional	AS:i:-2
optional	XN:i:0

# CIGAR strings

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

Almost never used



# Quiz Question 10

# sam flags

Bit	Description	
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

	<b>read paired?</b>	<b>properly aligned?</b>	<b>unmapped?</b>	<b>mate unmapped?</b>	<b>flag</b>
<b>read1</b>	1	1	0	0	3
<b>read2</b>	1	0	0	1	9
<b>read3</b>	1	0	1	1	13

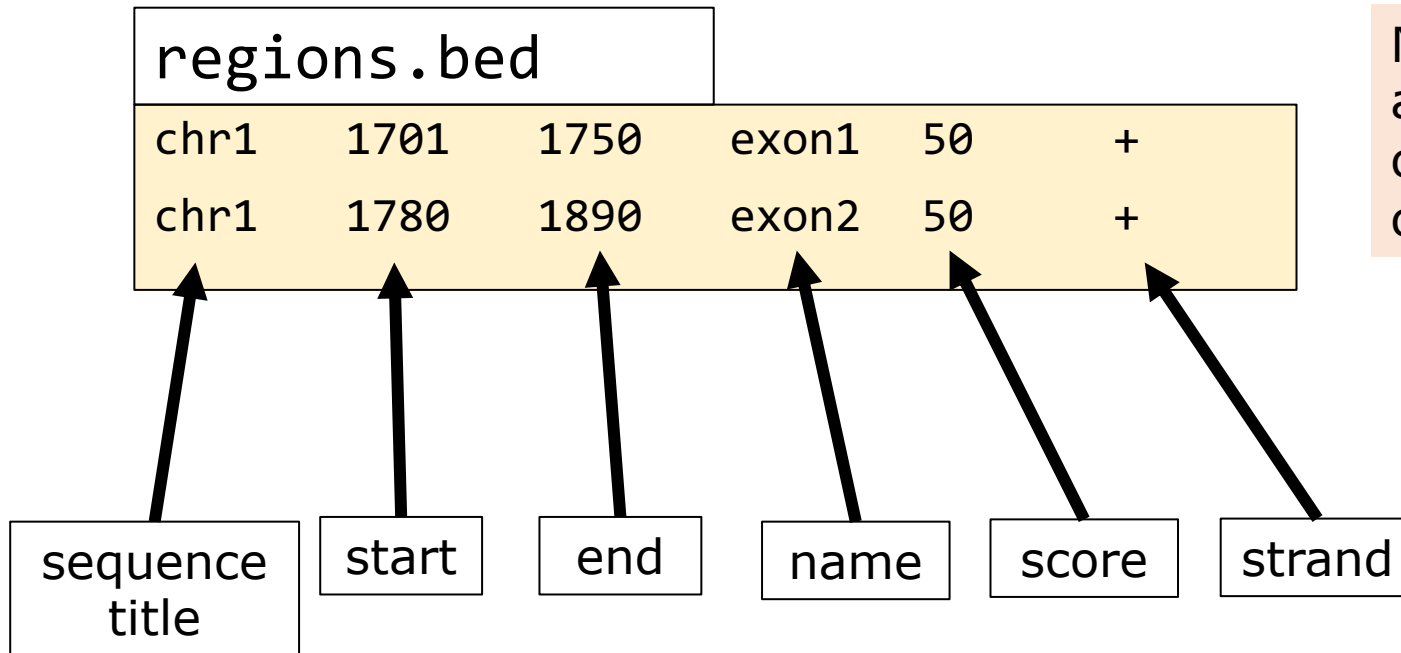
1	2	4	8
---	---	---	---

# Quiz Question 11

# bed

## Browser Extensible Data

Aim: specify regions



Numbering starts at 0!  
chr1:1702-1750  
chr1:1781-1890

# gff

## General Feature Format

Aim: annotation

seq name	source	feature	start	end	score	strand	frame	attributes
1	ensembl	mRNA	339070	346959	.	-	.	ID=...;
1	ensembl	exon	339070	339312	.	-	.	Parent=; ...
1	ensembl	CDS	339070	339312	.	-	0	ID=;...

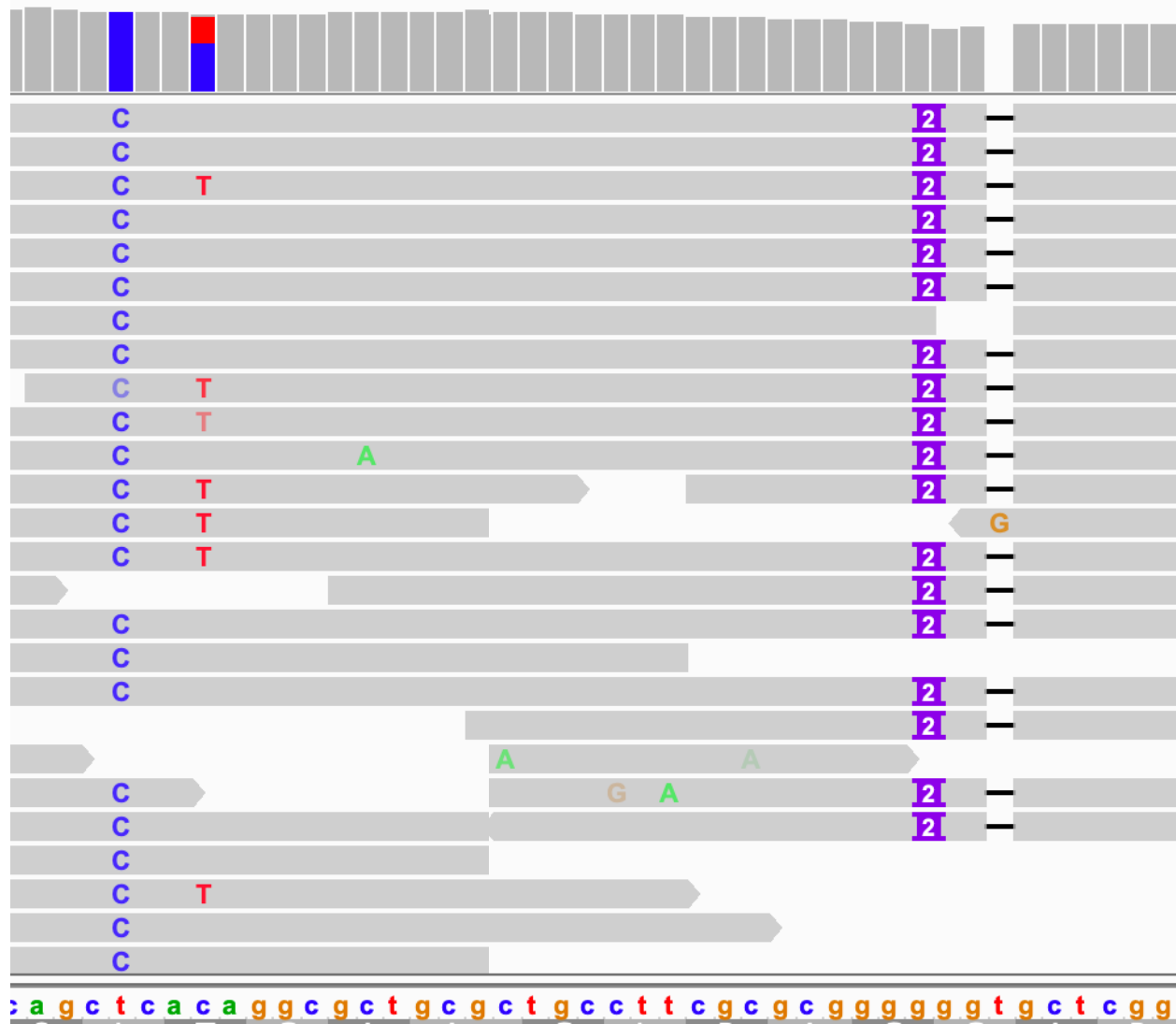


# vcf

## Variant Call Format

### Aim: variants

```
##fileformat=VCFv4.2
##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
```



# vcf

```
##fileformat=VCFv4.3
##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
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
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
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
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
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

samples