

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

reads.fastq

```
@D00283R:66:CC611ANXX:4:2311:2596:2330 1:N:0:TCCGGAG
ACTCTACGCTCAATAAAGATTTCTGATACGGCTCCTGAAATGCAGAATGAGT
+
B/<<<B<FFFFFFFFFBBFFFBFFFBFFFF/FFFFFFFF/BFFFBFFF
```

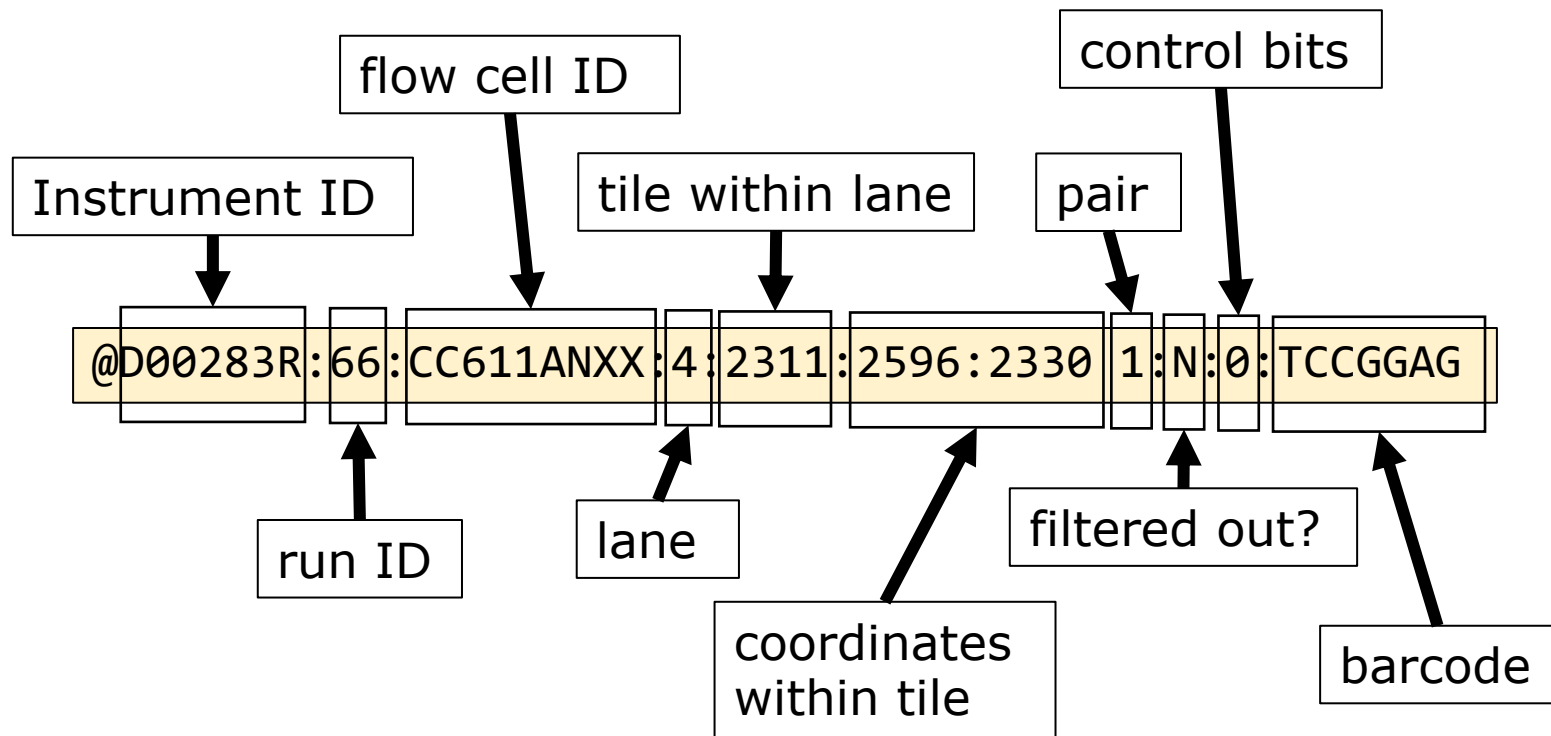
title, starts with @

nucleotide sequence

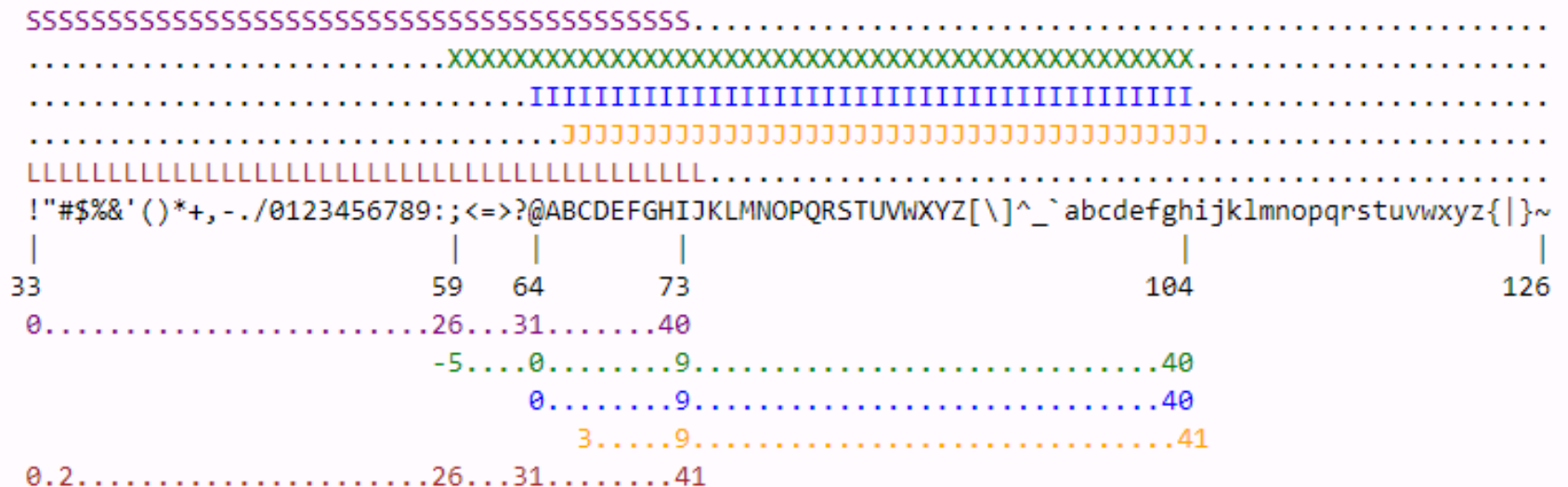
optional description

base quality

fastq header



Base quality (phred)



S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 41)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

Question

With `grep -c "^>"` it is easy to count the number of sequences in a fasta file. Why doesn't `grep -c "^@"` work for fastq sequences?

- A. @ is a special character for regular expressions
- B. The @ can also occur elsewhere in a fastq file
- C. There is no specific title for fastq files

sam

sequence alignment format

Aim: alignments

SAM column	example
read name	SRR519926.5
flag	89
reference	U00096.3
start position	222
mapping quality	42
CIGAR string	187M
reference name mate is mapped	=
start position mate	222
fragment length	339
sequence	CATCACCATTCCCAC
base quality	@>4:4C@89+&9CC@
optional	AS:i:-2
optional	XN:i:0

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

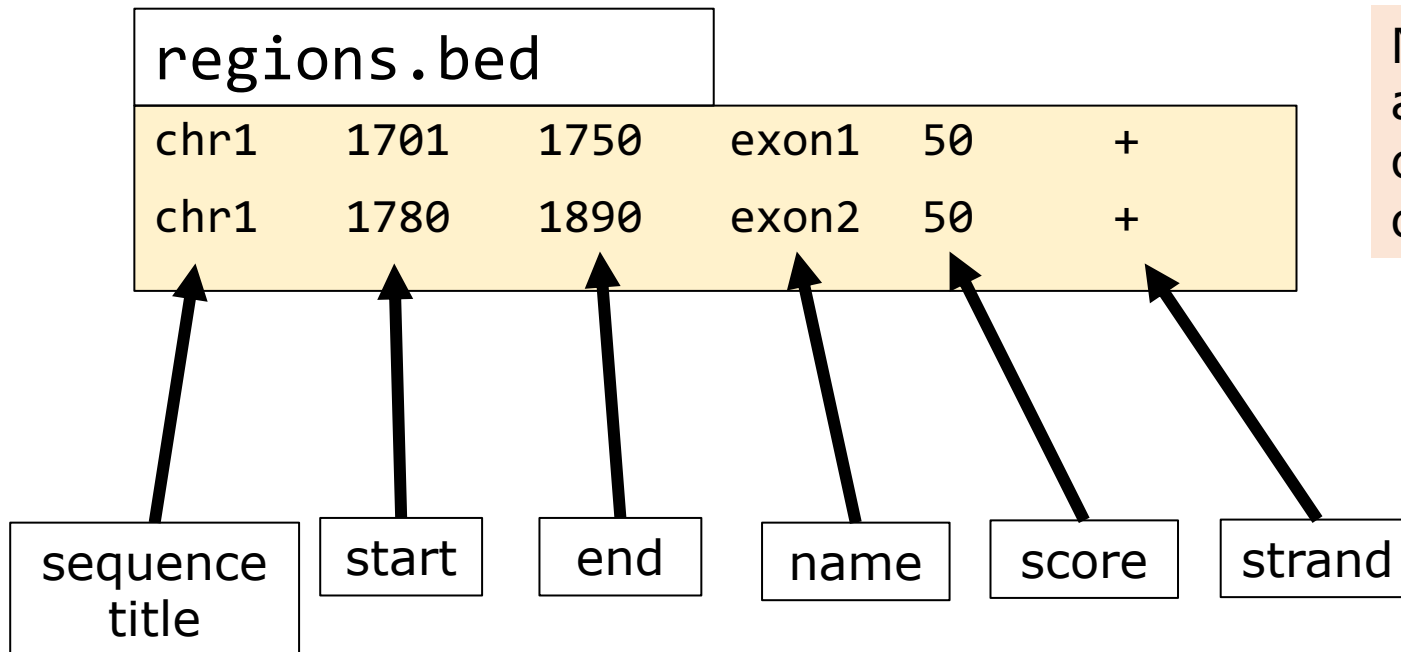
	multiple segments?	properly aligned?	unmapped?	next unmapped?	flag
read1	0	1	0	0	2
read2	0	0	1	1	12
read3	0	1	0	1	10

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

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
```