File Formats

John McCallum
Marcus Davy
Samantha Baldwin

NGS common File formats

Input and outputs should stick to standard common formats

- *Fasta – Raw nucleotide/peptide format*
  (From Sanger sequencing, assembly etc)

- *Fastq – Raw sequence information*
  (From NGS illumina Roche 454 etc)

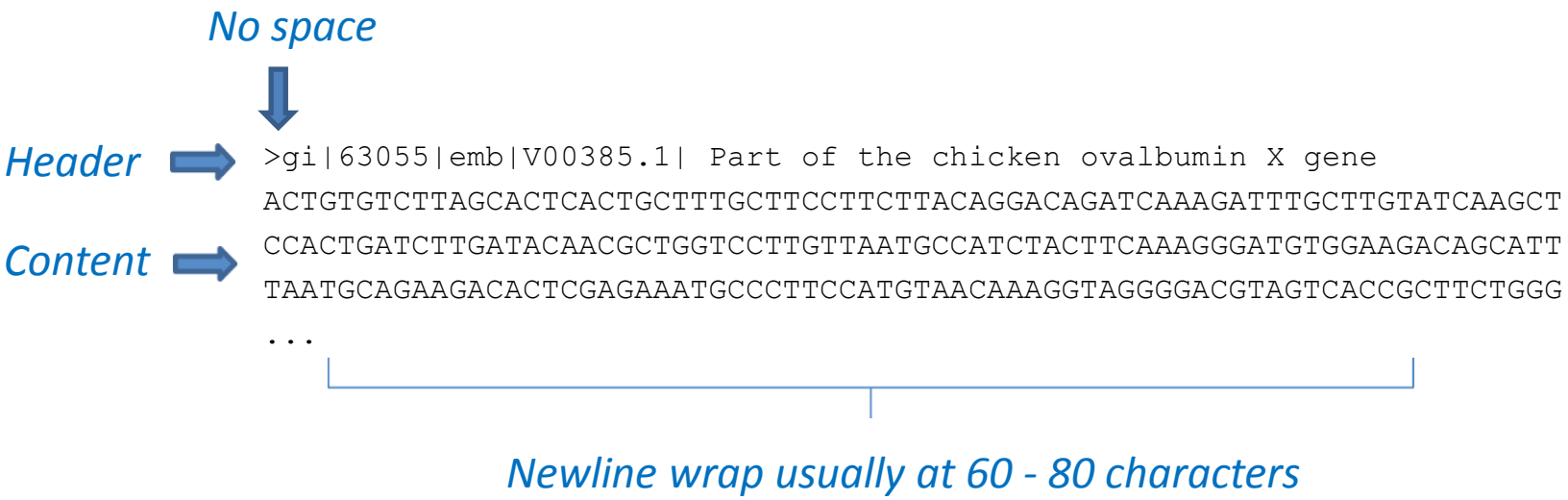
- *sam/bam format – Sequence Alignment/Map format*
  (Standard alignment mapper output format)

- *GFF – General feature format*
  (Describes genes and other features of DNA, RNA and protein sequences)

  - *Some standards are mature, others are evolving*

Fasta format

- Text-based format for storing nucleotide/peptide sequence(s)
- Restricted to *IUPAC alphabet* letters



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

Fastq format

- Four lines per sequence record
- Variable width format *(header x=..., y=... not padded)*
- *Makes parsing the format more difficult*

Line 1: @Header

Line2: Sequence string

Line3: +Header *(or just +)*

Line4: Quality string

...

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

Fastq format header

- A single record

Header ➡ `@HWI-EAS209_0025_FC427:6:1:1041:14884#ACAGTG/2`

Sequence ➡ `AATTTGTTTGTTTGTTGTTTATTTTTTTGTTAGTTTCGTTTGTTGTTTGTTTGGATTTCCTCTGTGTTGAGTATTT`

`+HWI-EAS209_0025_FC427:6:1:1041:14884#ACAGTG/2`

Quality ➡ `_________QZNUSUISNQW__U^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB`

- Illumina header contains several fields

HWI-EAS209	Unique machine identifier
0025	Run number
FC427	Unique flowcell identifier
6	Lane number
1	Tile number
1041	X coordinate within tile
14884	Y coordinate within tile
#ACAGTG	illumina barcode multiplexing index tag
/2	Pair number (1 or 2)

Fastq format quality

- *Quality scores encoded as ASCII characters*
- *Format has been evolving*

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

 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, 40)
    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)
```

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/

Sequence alignment map (sam)

De facto standard for storing alignment data

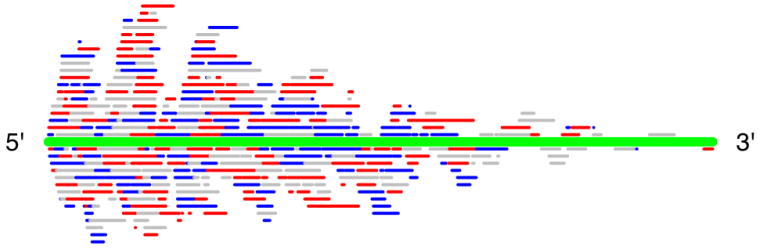
SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is **flexible** enough to store all the alignment information generated by various alignment programs;
- Is **simple** enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is **compact** in file size;
- Allows most of operations on the alignment to **work on a stream** without loading the whole alignment into memory;
- Allows the file to be **indexed by genomic position** to efficiently retrieve all reads aligning to a locus.

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

SAM format structure

- SAM file –TAB delimited text format
- BAM file – binary version



Header ➡

Alignment ➡

```
@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         *
...
```

SAMtools is a ***sourceforge*** project which provides utilities to manipulate SAM/BAM files

SAM format specification: ***http://samtools.sourceforge.net/SAM1.pdf***

General Feature Format (GFF)

- TAB delimited text format for describing genes and other features associated with DNA, RNA and Protein sequences.
- GFF files contain features and coordinates but no sequence

```
##gff-version 3
```

seqid	*source*	*type*	*start*	*end*	*score*	*strand*	*phase*	*attributes*
Ctg123	.	Exon	1300	1500	.	+	.	ID=exon00001
Ctg123	.	Exon	1050	1500	.	+	.	ID=exon00002
Ctg123	.	Exon	3000	3902	.	+	.	ID=exon00003
Ctg123	.	Exon	5000	5500	.	+	.	ID=exon00004
Ctg123	.	Exon	7000	9000	.	+	.	ID=exon00005

```
...
```

http://www.sequenceontology.org/gff3.shtml
http://gmod.org/wiki/GFF