FAST and Sequence Data Manipulation

Dana L Carper and Travis J Lawrence
Quantitative and Systems Biology
University of California, Merced

Sequence File Formats

- FASTA
  - Sequence (Nucleotide or Amino acids) represented by one letter codes
- FASTQ
  - Sequence (Nucleotides) represented by one letter codes
  - Quality scores for each nucleotide
- Both are text based storage formats
  - If they are text why cant we process them like text?
  - Lets take a look at each of these formats closer

FASTA Format

Example of 2 FASTA records

>AD01000001|NC_003070|Plant|Arabidopsis thaliana|306384|306456|Val|TAC|0|0||||T|

agaccaataaacttcttctgctctctctactaatggaatcagtttttgttttagaataacagtaacgttatattacgtatctctatagataatgccacaaGGTGCTGTGGTGTAGTGGTTATCACGTTTGCCTTACACGCAAAAGGTCTCCAGTTCGATCCTGGGCAGCACCAattgtgttttgcaatttttaataagaaaaatgcaaacttccttttttctttttatatacagaccaaaaaattggtatgatttactcaagaatgattgttct

>AD01000002|NC_003070|Plant|Arabidopsis thaliana|515494|515566|Phe|GAA|0|0||T||

gaactcaaattgctgagtgttactgatttgtccgtataaaattagaagataatcatgaatttagggtttgaataaagtgattaatgaaaccaaagcaaaaGCGGGGATAGCTCAGTTGGGAGAGCGTCAGACTGAAGATCTGAAGGTCGCGTGTTCGATCCACGCTCACCGCAttttttaatattgtttatgttttattcaaagcccattggatctttattctcttttaaatatgtgtccatttagtgtgttctaccgagcgcgtttggcccgta

FASTA Format

Each record has two lines

>AD01000001|NC_003070|Plant|Arabidopsis thaliana|306384|306456|Val|TAC|0|0||||T|

agaccaataaacttcttctgctctctctactaatggaatcagtttttgttttagaataacagtaacgttatattacgtatctctatagataatgccacaaGGTGCTGTGGTGTAGTGGTTATCACGTTTGCCTTACACGCAAAAGGTCTCCAGTTCGATCCTGGGCAGCACCAattgtgttttgcaatttttaataagaaaaatgcaaacttccttttttctttttatatacagaccaaaaaattggtatgatttactcaagaatgattgttct

>AD01000002|NC_003070|Plant|Arabidopsis thaliana|515494|515566|Phe|GAA|0|0||T||

gaactcaaattgctgagtgttactgatttgtccgtataaaattagaagataatcatgaatttagggtttgaataaagtgattaatgaaaccaaagcaaaaGCGGGGATAGCTCAGTTGGGAGAGCGTCAGACTGAAGATCTGAAGGTCGCGTGTTCGATCCACGCTCACCGCAtttttttaatattgtttatgttttattcaaagcccattggatctttattctcttttaaatatgtgtccatttagtgtgttctaccgagcgcgtttggcccgta

FASTA Format

Each sequence begins with “>”

```
>AD01000001|NC_003070|Plant|Arabidopsis thaliana|306384|306456|Val|TAC|0|0||T||||
agaccaataaacttcttctgctctctctactaatggaatcagtttttgttttagaataacagtaacgttatattacgtatctctatagataatgccacaaGGTGCTGTGGTGTAGTGGTTATCACGTTTGCCTTACACGCAAAAGGTCTCCAGTTCGATCCTGGGCAGCACCAattgtgtttgcaatttttttaataagaaaaatgcaaacttccttttttctttttttatatacagaccaaaaaattggtatgatttactcaagaatgattgttct
>AD01000002|NC_003070|Plant|Arabidopsis thaliana|515494|515566|Phe|GAA|0|0|T|||||
gaactcaaattgctgagtgttactgatttgtccgtataaaattagaagataatcatgaatttagggtttgaataaagtgattaatgaaaccaaagcaaaaGCGGGGATAGCTCAGTTGGGAGAGCGTCAGACTGAAGATCTGAAGGTCGCGTGTTCGATCCACGCTCACCGCAttttttaatattgtttatgttttattcaaagcccattggatctttattctcttttaaatatgtgtccatttagtgtgttctaccgagcgcgtttggcccgta
```

FASTA Format

Sequence name and description

```
>AD01000001|NC_003070|Plant|Arabidopsis thaliana|306384|306436|Val|TAC|0|0|||||
agaccaataaacttcttctgctctctctactaatggaatcagtttttgttttagaataacagtaacgtt
atattacgtatctctatagataatgccacaaGGTGCTGTGGTGTAGTGGTTATCACGTT
TGCCTTACACGCAAAAGGTCTCCAGTTCGATCCTGGGCAGCACCAattgtgtttt
gcaattttttaataagaaaaatgcaaacttccttttttctttttatatacagaccaaaaaattggtatg
atttactcaagaatgattgttct
>AD01000002|NC_003070|Plant|Arabidopsis thaliana|515494|515566|Phe|GAA|0|0|||||
gaactcaaattgctgagtgttactgatttgtccgtataaaattagaagataatcatgaatttagggttt
gaataaagtgattaatgaaaccaaagcaaaaGCGGGGATAGCTCAGTTGGGAGAGC
GTCAGACTGAAGATCTGAAGGTCGCGTGTTCGATCCACGCTCACCGCAtttttt
taatattgtttatgttttattcaaagcccattggatctttattctcttttaaatatgtgtccatttagtgtgt
tctaccgagcgcgtttggcccgta
```

FASTA Format

Sequence name and description

Followed by a new line character

```
>AD01000001|NC_003070|Plant|Arabidopsis thaliana|306384|306436|Val|TAC|0|0||||| \n
agaccaataaacttcttctgctctctctactaatggaatcagtttttgttttagaataacagtaacgtt
atattacgtatctctatagataatgccacaaGGTGCTGTGGTGTAGTGGTTATCACGTT
TGCCTTACACGCAAAAGGTCTCCAGTTCGATCCTGGGCAGCACCAattgtgtttt
gcaatttttaataagaaaaatgcaaacttccttttttctttttatatacagaccaaaaaattggtatg
atttactcaagaatgattgttct
>AD01000002|NC_003070|Plant|Arabidopsis thaliana|515494|515566|Phe|GAA|0|0||||| \n
gaactcaaattgctgagtgttactgatttgtccgtataaaattagaagataatcatgaatttagggttt
gaataaagtgattaatgaaaccaaagcaaaaGCGGGGATAGCTCAGTTGGGAGAGC
GTCAGACTGAAGATCTGAAGGTCGCGTGTTCGATCCACGCTCACCGCAtttttt
taatattgtttatgttttattcaaagcccattggatctttattctcttttaaatatgtgtccatttagtgtgt
tctaccgagcgcgtttggcccgta
```

FASTA Format

Sequence can stretch across multiple lines

```
>AD01000001|NC_003070|Plant|Arabidopsis thaliana|306384|306456|Val|TAC|0|0||T||||
agaccaataaacttcttctgctctctctactaatggaatcagtttttgttttagaataacagtaacgtt
atattacgtatctctatagataatgccacaaGGTGCTGTGGTGTAGTGGTTATCACGTT
TGCCTTACACGCAAAAGGTCTCCAGTTCGATCCTGGGCAGCACCAattgtgtttt
gcaatttttaataagaaaaatgcaaacttccttttttctttttatatacagaccaaaaaattggtatg
atttactcaagaatgattgttct
>AD01000002|NC_003070|Plant|Arabidopsis thaliana|515494|515566|Phe|GAA|0|0|T|||||
gaactcaaattgctgagtgttactgatttgtccgtataaaattagaagataatcatgaatttagggttt
gaataaagtgattaatgaaaccaaagcaaaaGCGGGGATAGCTCAGTTGGGAGAGC
GTCAGACTGAAGATCTGAAGGTCGCGTGTTCGATCCACGCTCACCGCAtttttt
taatattgtttatgttttattcaaagcccattggatctttattctcttttaaatatgtgtccatttagtgtgt
tctaccgagcgcgtttggcccgta
```

FASTQ Format

Example of 1 FASTQ record

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGTGGCT
TCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAA
AACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAAT
TCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATCCACAGAACTTTCCAGA
GATGGATTGGTGCCTTCGGGAACTGTGAGACAGGTGCTGCATGGCTGTCGTCA
GCTCGTGTTGTGAAATGTTGGGTTAAGTCACG
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/B9//BF
AB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??FFC;???FC;@
9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>B9B;;B
F@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:B11A1B0AA
//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

FASTQ Format

Each record has 4 lines

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGTGGCT
TCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAA
AACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAAT
TCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATCCACAGAACTTTCCAGA
GATGGATTGGTGCCTTCGGGAACTGTGAGACAGGTGCTGCATGGCTGTCGTCA
GCTCGTGTTGTGAAATGTTGGGTTAAGTCACG
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/B9//BF
AB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??FFC;???FC;@
9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>B9B;;B
F@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:B11A1B0AA
//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

FASTQ Format

Each sequence begins with “@”

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGT
GGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGC
AAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAG
CATGTGGTTTAATTCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATC
CACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAG
GTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCA
CG
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/
B9//BFAB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??
FFC;???FC;@9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>
B9B;;BF@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:
B11A1B0AA//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

FASTQ Format

Followed by sequence information

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGT
GGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGC
AAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAG
CATGTGGTTTAATTCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATC
CACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAG
GTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCA
CG
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/
B9//BFAB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??
FFC;???FC;@9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>
B9B;;BF@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:
B11A1B0AA//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

FASTQ Format

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0 \n
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGT
GGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGC
AAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAG
CATGTGGTTTAATTCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATC
CACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAG
GTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCA
CG
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/
B9//BFAB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??
FFC;???FC;@9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>
B9B;;BF@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:
B11A1B0AA//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

Followed by sequence information

Followed by a new line character

FASTQ Format

Actual nucleotide sequence can stretch across multiple lines

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGT
GGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGC
AAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAG
CATGTGGTTTAATTCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATC
CACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAG
GTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCA
CG
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/
B9//BFAB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??
FFC;???FC;@9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>
B9B;;BF@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:
B11A1B0AA//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

FASTQ Format

Actual nucleotide sequence can stretch across multiple lines

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGT
GGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGC
AAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAG
CATGTGGTTTAATTCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATC
CACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAG
GTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCA
CG \n
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/
B9//BFAB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??
FFC;???FC;@9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>
B9B;;BF@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:
B11A1B0AA//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

FASTQ Format

@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0

TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGT
GGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGC
AAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAG
CATGTGGTTTAATTCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATC
CACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAG
GTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCA
CG

+

1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/
B9//BFAB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??
FFC;???FC;@9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>
B9B;;BF@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:
B11A1B0AA//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111

Third line begins with "+" then can contain other sequence information

FASTQ Format

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGT
GGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGC
AAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAG
CATGTGGTTTAATTCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATC
CACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAG
GTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCA
CG
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/
B9//BFAB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??
FFC;???FC;@9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>
B9B;;BF@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:
B11A1B0AA//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

The last line contains quality scores

FASTQ Format

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGT
GGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGC
AAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAG
CATGTGGTTTAATTCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATC
CACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAG
GTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCA
CG
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/
B9//BFAB/1F;/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??
FFC;???FC;@9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>
B9B;;BF@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:
B11A1B0AA//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

Quality scores is a measure of how accurate the base is

FASTQ Format

```
@M02450:19:000000000-AMR66:1:1101:14559:1829 1:N:0:0
TTAGTCCATGCCGTAAACGATGTCGTCTTGTAGTTTGTTCCCTTGAGTCGT
GGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGC
AAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAG
CATGTGGTTTAATTCGATGCAACGCGAAGAGCCTTACCTGGTCTTGACATC
CACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAG
GTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCA
CG
+
1AAAAFFFFBCFGGGGGGFGGGHE00ABF11122DD1B0BAAFG0111A/
B9//BFAB/1F:/D@@B9A9>EEA9BF9;>>/>>E9/BFE9?;B;/<</<FC90??
FFC;???FC;@9C;@FAC?@:::-<<:1@-
:EC<:<?;?<F19?;<;<<CDFHF;CBF@;</B>:<9EGFBB0:9:FB99FBB;9BF>
B9B;;BF@EEBB1;B119HFGBBBB0B;A?>;/9;FFFBEEFB;1B;DGBD9GAA:
B11A1B0AA//A00B11BA000A0A13A3BGBB1A111BDB1FD>11111
```

Notice that “@” are also quality scores

Why is this format an issue with the command line tools?

- The sequence record is stretched across multiple lines

- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

grep 'AAAAA' Sequences.fa
AAAAATTTTCGATC
```

Want to find sequences that contain AAAAA

Output is the line containing this not the entire sequence record

FAST: Analysis of Sequences Toolbox

- Freely available toolbox
- Designed to work on either FASTA or FASTQ files
- Based on the UNIX philosophy
- Based on the command line tools but designed for FASTA/Q files

fasgrep

- Acts like command line grep command
- Can act on the description line or the sequence
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fasgrep -s ‘AAAAA’ Sequences.fa
```

fasgrep

- Acts like command line grep command
- Can act on the description line or the sequence
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fasgrep -s 'AAAAA' Sequences.fa
```

command

fasgrep

- Acts like command line grep command
- Can act on the description line or the sequence
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fasgrep -s 'AAAAA' Sequences.fa
```

Flag to search the sequence not the description

fasgrep

- Acts like command line grep command
- Can act on the description line or the sequence
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fasgrep -s 'AAAAA' Sequences.fa
>sequence1
AAAAATTTTCGATC
```

Output is the sequence file that matches

fascut

- Cut out particular parts of the sequence (such as the coding sequences)
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC
```

```
fascut 1-3,5,8-13 Sequences.fa
```

fascut

- Cut out particular parts of the sequence (such as the coding sequences)
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fascut 1-3,5,8-13 Sequences.fa
```

Command

fascut

- Cut out particular parts of the sequence (such as the coding sequences)
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fascut 1-3,5,8-13 Sequences.fa
```

Nucleotides I want to cut

fascut

- Cut out particular parts of the sequence (such as the coding sequences)
- For example: Sequences.fa

>sequence1
AAAA**A**TT**TTCGAT**C
>sequence 2
AATT**T**CC**GGAATC**C

fascut 1-3,5,8-13 Sequences.fa
>sequence1
AAAATTCGAT
>sequence 2
AATTGGAATC

Output is the sequence file description and the cut sequence

fastr

- Transform nucleotide, get rid of gaps, squish multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC
```

```
fastr –s “T” “U” Sequences.fa
```

fastr

- Transform nucleotide, get rid of gaps, squish multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –s “T” “U” Sequences.fa
```

Command

fastr

- Transform nucleotide, get rid of gaps, squish multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –s “T” “U” Sequences.fa
```

Act on the sequences

fastr

- Transform nucleotide, get rid of gaps, squish multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –s “T” “U” Sequences.fa
```

What I want to be replaced

fastr

- Transform nucleotide, get rid of gaps, squish multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –s “T” “U” Sequences.fa
```

What to replace it with (DNA to RNA)

fastr

- Transform nucleotide, get rid of gaps, squish multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –s “T” “U” Sequences.fa
>sequence1
AAAAAUUUUCGAUC
>sequence2
AAUUUCCGGAAUCC
```

Output all T’s replaced with U’s

fastr

- Transform nucleotide, get rid of gaps, squish multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –snS  “A” Sequences.fa
```

fastr

- Transform nucleotide, get rid of gaps, squish multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –snS  “A” Sequences.fa
```

Act on sequences

fastr

- Transform nucleotide, get rid of gaps, squish multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –snS  “A” Sequences.fa
```

No replace

fastr

- Transform nucleotide, get rid of gaps, squash multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –snS  “A” Sequences.fa
```

Squash multiple characters

fastr

- Transform nucleotide, get rid of gaps, squash multiple characters
- For example: Sequences.fa

```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –snS  “A” Sequences.fa
```

Character I want it to act on

fastr

- Transform nucleotide, get rid of gaps, squash multiple characters
- For example: Sequences.fa

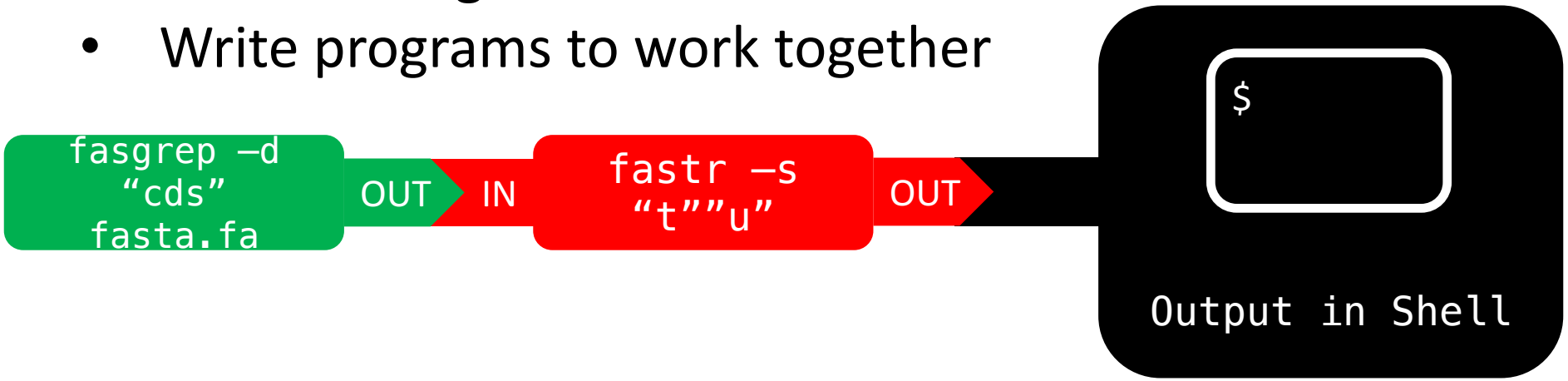
```
>sequence1
AAAAATTTTCGATC
>sequence 2
AATTTCCGGAATCC

fastr –snS  “A” Sequences.fa
>sequence1
ATTTTCGATC
>sequence2
ATTTCCGGATCC
```

Places with repeating A’s are squashed to just one A

Combining Commands

- Unix Philosophy:
  - Do one thing and do it well
  - Write programs to work together



- `$ fasgrep –d “cds” fasta.fa | fastr –s “t” “u”`