Colorspace to Illumina FastQ

Colorspace formats and Transformations

```
==> 5500xl_DA2_2013_08_12_1_IGF_FC1_01_Library17_F3.csfasta <==
>1_11_259_F3
T303202210001123002120020110233120030023222011120021101131030321030003311220
>1_11_974_F3
T202210101121000203122100202202003231311023113211110103300103301012101212211
>1_11_1721_F3
T010033310012033020000220002020010033002231220013123000000003011000000002300
>1_11_1786_F3
T012131122210002131003100211230000012102131220302103233013103330031230132220
>1_11_1859_F3
T213201230300030020203003211202210213102113111303330201030313112312302220311
```

```
==> 5500xl_DA2_2013_08_12_1_IGF_FC1_01_Library17_F3.QV.qual <==
>1_11_259_F3
31 31 30 31 31 28 31 31 30 31 31 31 31 31 29 30 28 31 31 23 28 28 31 30 31
31 31 31 31 30 31 31 31 31 30 31 31 31 31 31 31 31 31 31 30 31 32 30 31 19
30 28 31 28 28 21 28 13 31 31 31 17 31 30 17 31 26 31 26 17 30 21 27 21 29
>1_11_974_F3
31 17 31 28 31 28 31 31 31 28 27 23 23 27 31 31 23 27 31 27 17 13 17 31 14
23 13 27 21 23 21 21 9 31 12 18 14 12 14 21 13 14 9 15 9 14 14 18 17 21 14
23 23 12 14 13 28 15 21 21 21 15 24 12 17 18 14 14 9 28 13 14 9 17 9
>1_11_1721_F3
31 30 21 14 17 23 14 27 23 29 17 31 31 14 29 20 31 29 31 28 31 31 31 31 31
31 31 23 14 31 31 27 13 14 30 28 21 17 31 30 31 31 14 31 31 26 31 28 29 31
26 27 23 12 27 12 17 30 19 17 17 12 14 21 27 23 14 27 21 14 31 14 21 14 31
>1_11_1786_F3
31 31 29 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31
14 31 31 28 31 31 31 31 30 31 31 31 14 31 28 23 31 31 31 28 25 29 30 27 31
31 31 29 31 29 17 14 31 17 31 14 21 23 14 31 14 30 17 30 26 13 21 29 14 14
>1_11_1859_F3
31 31 31 31 31 31 31 31 31 31 31 29 31 31 31 31 31 31 31 31 31 31 31 31 31
31 31 31 31 31 31 31 17 31 26 31 31 21 31 31 31 26 31 31 31 31 17 31 27 26
31 31 17 28 30 23 30 31 28 27 31 27 21 27 31 31 17 27 28 31 31 14 31 30 31
```

This link on biostars shows different types of SOLiD to fastq conversions. What we really need is the last format that is described in the link. It is the format where the rest of the sequence will be N's after encountering the first dot. And the quals should be represented in the phred 33 scale. Several scripts have been available online but since none of them worked like we wanted, I have written a C++ code to directly produce the fastq formated file with our required features by taking csfasta and qual flies as an input.

Transformation Programs and Code

Output from my own C++ code

It has the following features:

- It has a straight forward help function that can be displayed like this.

```
$ ./cq2ip33fq -h
Program Usage
    Examples:
    cq2ip33fq --csfasta input.csfasta --qual input.qual > output.fastq
    cq2ip33fq --csfasta input.csfasta --qual input.qual | gzip >
output.fastq.gz
:
  -h [ --help ]         displays this help message
  -c [ --csfasta ] arg  csfasta file
  -q [ --qual ] arg     qual file
```

- An example output for the reads with matching readnames.

```
$ ./cq2ip33fq --csfasta F3_6reads.csfasta --qual F3_6reads.qual
@1_11_259_F3
AATCCTCAAAACAGCCCTGAAAGGTGGATACTTTAAAGCTCTTGTGAAAGTGGTGCAATTAGTTAAAATACAGAA
+
@@?@@=@@?@@@@@>?=@@8==@?@@@@@?@@@@?@@@@@@@@@@?@A?@4?=@==6=.@@@2@?2@;@;2?6<6>
@1_11_974_F3
CCTCAACCACTGGGGAATGAGTTTCCTCCTTTAGCATGTTCGTGCTGTGTTGGCGGGTTATTGGTCAACTGAGTG
+
@2@=@=@@@=<88<@@8<@<2.2@/8.<6866*@-3/-/6./*0*//326/88-/.=066609-23//*=./*2*
@1_11_1721_F3
TGGGCGCAAACTTANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
@?6/28/<8>2@@/>5@>@=@@@@@@@8/@@<./?=62@?@@/@@;@=>@;<8-<-2?422-/6<8/<6/@/6/@
```

- In the above example, you will also see that after encountering a missed call i.e a 'dot' in the colorspace, the rest of the sequence will be Ns. This is because it is unreliable to decode after encountering the a missed call; as the error will cascade through the rest of the sequence.

```
# Representation of the colorspace and its corresponding basespace sequence

T01003331001203.020000220002020010033002231220013123000000003011000000002300
 TGGGCGCAAACTTANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

- The program throws an error when the readnames doesn't match between csfasta and qual files.

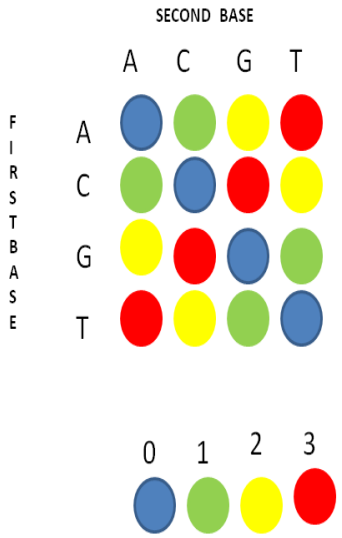
```
$ ./cq2ip33fq --csfasta F3_6reads_1wrongname.csfasta --qual F3_6reads.qual
@1_11_259_F3
AATCCTCAAAACAGCCCTGAAAGGTGGATACTTTAAAGCTCTTGTGAAAGTGGTGCAATTAGTTAAAATACAGAA
+
@@?@@=@@?@@@@@>?=@@8==@?@@@@@?@@@@?@@@@@@@@@@?@A?@4?=@==6=.@@@2@?2@;@;2?6<6>

ERROR: csfasta readname >1_11_946_F3 is not equal to the qual readname
>1_11_974_F3.
Please check if both the csfasta file and qual file has equal number of
reads with matching readnames.
```

- The time of execution for 3 million 75bp reads. Marcin's python code took 14m54.733s for F3 and 11m14.612s for F5-RNA reads.

```
# With Compression (3 million 75bp F3 Reads)
$ time ./cq2ip33fq \
--csfasta 5500xl_DA2_2013_08_12_1_IGF_FC1_01_Library17_F3.csfasta \
--qual 5500xl_DA2_2013_08_12_1_IGF_FC1_01_Library17_F3.QV.qual | \
gzip > F3.fastq.gz

real    11m38.551s
user    13m41.925s
sys 0m32.688s

# With Compression (3 million 35bp F5-RNA Reads)
$ time ./cq2ip33fq \
--csfasta 5500xl_DA2_2013_08_12_1_IGF_FC1_01_Library17_F5-RNA.csfasta \
--qual 5500xl_DA2_2013_08_12_1_IGF_FC1_01_Library17_F5-RNA.QV.qual | \
gzip > F5-RNA.fastq.gz

real    5m41.050s
user    6m29.568s
sys 0m21.469s

# Without Compression
$ time ./cq2ip33fq \
--csfasta 5500xl_DA2_2013_08_12_1_IGF_FC1_01_Library17_F3.csfasta \
```

```
--qual 5500xl_DA2_2013_08_12_1_IGF_FC1_01_Library17_F3.QV.qual \
> test_output.fastq

real    10m40.395s
user    9m58.467s
sys 0m42.002s
```

The code is based on the following 2-base colorspace encoding scheme.



	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

AA	0
CA	1
GA	2
TA	3
AC	1
CC	0
GC	3
TC	2
AG	2
CG	3
GG	0
TG	1
AT	3
CT	2
GT	1
TT	0