NextGENe™ by SoftGenetics

April 2009

File Formats Used in NextGENe Software for Short Sequence Reads

Megan Manion, Kevin LeVan, Jacie Wu, Ni Shouyong, CS Jonathan Liu

Introduction

Next generation DNA sequencing technologies have drastically increased the speed and quantity at which genetic sequence information can be produced. However the unique qualities of this data, including short read lengths, high error rates compared to Sanger sequencing and large data volume, require specialized software applications for analysis. NextGENe software is designed to analyze the short sequence reads from massively parallel sequencing systems such as the Illumina® Genome Analyzer (Solexa sequencing technology) using the sequencing by synthesis technique, the Applied Biosystems SOLiD™ System using sequencing by ligation and the Roche Genome Sequencer System (454 Sequencing) utilizing pyrosequencing technology. Each instrument supplier uses a different format for organizing the reads and assigning quality scores. NextGENe is able to process various instrument outputs from each of the above systems.

NextGENe converts all sequencing output files into a standard fasta format prior to analysis. For SOLiD System reads, the csfasta format can also be used directly since immediate conversion to fasta format can result in error propagation. Quality scores can be used to trim reads or remove entire low-quality reads. Instrument read files are converted into fasta format using the Format Conversion Tool.

Fasta format

Lines starting with ">" are comment lines; this symbol also marks the beginning of each sequence record. The sequence lines follow the comment line.

Figure 1: An example of fasta format is shown. The comment line for each is the name assigned to the given read.

File: s_5.fasta

```
>s_5_0001_5_1_84_598
GTTATTTAACATAAGGTTATAGAACTCTCTACACTT
>s_5_0001_5_1_482_766
GTATAGAGTTCTATAACCTTATGTTAAATAACCTCA
>s_5_0001_5_1_742_905
GCTGCTAATTAATGAAGGTTATAGAACTTAATTGGT
```

Figure 1

Figure 2: NextGENe's Format Conversion Tool accepts instrument file formats and converts them to standard fasta format. Quality scores can be used to remove low quality reads or trim low quality ends of reads.

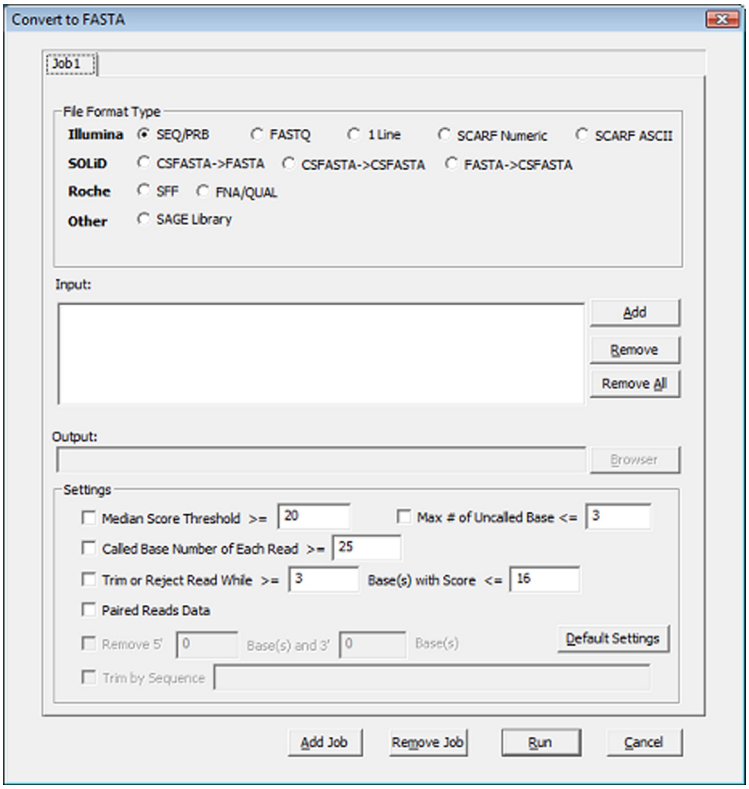

Figure 2

Format Conversion Tool

NextGENe's Format Conversion Tool is used to convert instrument output files into the standard fasta format. Users can select the file format of their data, upload sample file(s), choose quality settings (if desired) and click Run to begin format conversion. Quality scores can be used to remove entire reads with a low median quality score or containing a set number of consecutive uncalled bases or to trim from the 3' end of reads when consecutive bases have a quality score below a selected threshold. When quality metrics are used, reads that are removed or bases that were trimmed are saved in the "removed.fasta" file while all converted reads that met quality thresholds are in the "converted.fasta" file. A log file is also created to provide a summary of the process including the number of reads converted, the number that failed conversion and a breakdown of the number of reads or bases that failed due to each quality standard.

Figure 3: A log file is created to describe the format conversion results.

```
[Filter Results]
[Total Reads in the Input File]: 10651414
        [Reads Converted Successfully]: 10179982
        [Reads Failed to Convert]: 471432
                [Reads Filtered by "Median Score"]: 352853
                [Reads Filtered by "Uncalled Bases"]: 62188
                [Reads Filtered by "Called Base Number in Each Read"]: 111
                [Reads Rejected by "Base's Score"]: 56280
        [Reads Trimmed by "Base's Score"]: 1177117
        [Trimmed Bases by "Base's Score"]: 17764943
```

Figure 3

Instrument Output Formats

All of the listed formats can be converted into fasta format, the standard format used by NextGENe, using the Format Conversion Tool.

Illumina Formats

Illumina Sequence and Probability Files (*_seq.txt, *_prb.txt)

The SEQ file includes the nucleotide sequence for each read and the PRB file contains the quality score of each possible base (ACGT) for the given cycle number. Four numbers, -40 40 -40 -40, each separated by a space, are the quality scores associated for each possible nucleotide, ACGT respectively. The tab is used to separate the bases of each cycle. A negative number is indicative of an impossible base call. A positive number shows the possibility of base call, 20 meaning high confidence and error of 1%. Several options are available for filtering out low quality information using these quality scores.

```
File: s_1_0001_seq.txt
1       1       137     689     AACATAATGTGTTCACTGAGAACACATTGCACTCAA
1       1       87      649     TATTGCAACTTGTTTAATTTTTTCATGCCATTATCA
1       1       121     642     TACATGATTTGCATTTGGTAAATAGCTACTTTTTAT
1       1       6       591     C...T.......T.......................
```

Figure 4

Figure 4: An example seq.txt file is shown. The file records the channel number, tile number, x position and y position of each sequence read. One flow cell contains 8 channels and approximately 300 tiles.

```
40 -40 -40 -40      40 -40 -40 -40      -40 40 -40 -40      40 -40 -40 -40
-40 -40 -40 40      40 -40 -40 -40      40 -40 -40 -40      -40 -40 -40 40
-40 -40 40 -40      -40 -40 -40 40      -40 -40 40 -40      -40 -40 -40 40
-40 -40 -40 40      -40 40 -40 -40      40 -40 -40 -40      -40 40 -40 -40
-40 -40 -40 40      -40 -40 40 -40      40 -40 -40 -40      -40 -40 40 -40
40 -40 -40 -40      40 -40 -40 -40      -40 40 -40 -40      40 -40 -40 -40
-40 40 -40 -40      40 -40 -40 -40      -40 -40 -40 40      -40 -40 -40 40
-40 -40 40 -40      -40 40 -40 -40      40 -40 -40 -40      -40 40 -40 -40
-40 -40 -40 40      -40 40 -40 -40      40 -40 -40 -40      37 -37 -40 -40
```

Figure 5

Figure 5: An example prb.txt file is shown. Quality scores are shown for each possible nucleotide at each base position.

FASTQ Format

Illumina has an additional format, FASTQ, in which each read starts with "@", which combines the sequence reads and quality score into one file. The quality score is represented as an ASCII character. Illumina outputs FASTQ files in ASCII-64 encoding. In this format, subtracting 64 from the ASCII code results in the corresponding quality value. Under the Solexa numbering scheme quality values can theoretically be as low as -5, which has an ASCII encoding of 59.

Figure 6: An example of Illumina reads in FASTQ format is shown.

```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```

Figure 6

Qseq Format

Illumina also offers a format which is similar to FASTQ format but differs in how the quality scores are calculated. Qseq format uses ASCII-64 encoding of Phred quality scores. Subtraction of 64 from the provided quality score results in the Phred score for the base.

```
HWUSI-EAS521    2       1       26      0       76      0       1
        .GGCAGCGGGCAGGGCGAGCCAATGCGTGTGGGGGGGGGGGGGCTCGCAGTGGGGGGGGAACGGCGAGTGCGGGGGG
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB      0
HWUSI-EAS521    2       1       26      0       562     0       1
        .GGGAAATAGCTTTCACGCCTTTAGATAATTTCATAAAAATCATAGCGCCAAATGGGGAGCAAACTACCATACACC
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB      0
HWUSI-EAS521    2       1       26      0       1113    0       1
        .ATCTTTAACAGACCAAGACTGGGCCACAAGCCTCCAGACTGTAACCACTGCTTCAAAGAGGCTTAGGCAGGCAGG
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB      0
HWUSI-EAS521    2       1       26      0       244     0       1
        .GGCTGGGTATGAGTCAGGGGGCTCCAGAGAGACAGAACCAGTCGGACATCGACAGATAGATGCGGGGGGAGTTAT
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB      0
```

Figure 7

Figure 7: qseq file format uses ASCII-64 to encode for Phred quality scores.

SoftGenetics LLC 100 Oakwood Ave, Suite 350 State College, PA 16803 USA
Phone: 814/237/9340 Fax 814/237/9343
www.softgenetics.com email: info@softgenetics.com

Illumina Single Line Format
Illumina can produce a format that merges the base calls and the quality scores for each base into one line.

```
>1-1-137-689     AACATAATGTTCACTGAGAACACATTGCACTCAA   U0
>1-1-87-649      TATTGCAACTTGTTTAATTTTTCATGCCATTATCA   U1
>1-1-121-642     TACATGATTTGCATTTGGTAAATAGCTACTTTTAT   U0
```

Figure 8

Figure 8: Sequence reads and quality scores are merged into one line in the Illumina Single Line Format.

SCARF (Solexa Compact ASCII Read Format)
One of Illumina's single line formats is referred to as SCARF format. This format contains all information for one read in a single line. From left to right each line contains the read name, nucleotide sequence, quality scores for each position, and more information. Illumina's pipeline can produce SCARF files with quality scores in ASCII or numeric format. When using the Format Conversion Tool, be sure to choose the appropriate quality score format (SCARF Numeric versus SCARF ASCII).

Figure 9: Sequence reads in Illumina's SCARF file format with quality scores in numeric format are shown.

```
HWI-EAS102_3 : 6 : 1 : 897 : 791 : AATGTCAATCTGAGTT . . . TTT : 40  40  40  40  40 . .
HWI-EAS102_3 : 6 : 1 : 930 : 291 : AATGTACTTTTTCTAA . . . CTA : 40  29  14  17  16 . .
HWI-EAS102_3 : 6 : 1 : 944 : 665 : AATCGATCCCCTTCCC . . . TTC : 40  34  33  40  40 . .
```

Figure 9

SOLiD Formats
The SOLiD System produces color-space sequence reads in a fasta format labeled as CSFASTA. These reads can be retained and analyzed in color-space by NextGENe, or they can be converted to base-space prior to analysis. The Format Conversion Tool offers options for cleaning of the CSFASTA files. By choosing CSFASTA→ FASTA, NextGENe converts the reads from color-space to base-space. By choosing CSFASTA→ CSFASTA the output file remains in color-space. This option can be used to quality trim reads while maintaining color-space.

Note: Errors in color-space can lead to the propagation of errors downstream within the read when converted to base-space with the Format Conversion tool. NextGENe's Condensation Tool™ helps correct many of these errors, so it is recommended to leave reads in color space until after Condensation.

Roche/454 data formats
SFF Format
The Roche/454 platform produces sequence reads in SFF format. This file contains the flow grams and quality scores for each read in addition to other information.

```
>000007_1940_1402 length=172 uaccno=E4UQSRD01E0MP4
TAACAATCGAGGCGAAGTCCCGTGAGAAGCTGTTTACTTCTCATGATCACACAGGCGCTG
GCTCCTCAGGCAAACAGGTACGTCTACGATAGGTTCCATGAAAAGTCCAAGTTTGGCCGA
GCTCTGGCTCCTTTTGACGCACAGTGGAACTTCCTTGTTCACGGAAATTGCA
```

Figure 10: An example .fna file is shown

FNA/QUAL Format
454 Sequencing's Genome Sequencer also produces read outputs in a fasta (*.fna) file and quality (*.qual) file.

```
>000007_1940_1402 length=172 uaccno=E4UQSRD01E0MP4
28 35 28 27 34 27 26 25 25 28 31 24 26 27 32 25 27 27 32 28 6  28 27 27 27 27 33 26
27 26 27 27 34 30 10 27 25 34 27 28 22 28 27 26 26 27 27 26 27 25 22 23 28 27 18 20
23 27 27 29 21 25 25 34 26 27 24 25 32 24 22 33 28 7  25 20 30 22 28 27 24 25 28 28
28 27 28 26 27 25 23 33 25 35 28 34 27 27 25 28 38 34 21 8  25 27 34 27 31 23 22 36
32 17 29 21 32 24 24 27 28 19 27 28 26 34 28 23 25 35 28 38 34 21 8  26 26 27 25 27
21 28 28 27 27 34 27 34 27 25 30 21 34 26 33 25 26 35 28 20 28 25 34 27 37 33 15 33
25 23 28 25
```

Figure 11: An example .qual file is shown containing the quality scores for the .fna file in Figure 8.

SAGE Library Format
SAGE libraries can be formatted into the same fasta format to be used as a reference for sequence alignment. For libraries that are organized in the format shown in Figure 11, the "SAGE Library" option in the Format Conversion Tool can be used to convert to fasta.

```
GGACCTGCGCCGCCCAG, Hs. 529989, RNATSET2, Ribonuclease T2 . . .
ATGTCAACTTCCTTGTA,  Hs339, P2RY2, Purinergic receptor . . .
CCTGTAATATCAGCTAC, Hs. 187388, TMC7, Transmembrane channel . . .
```

Figure 12: Example SAGE library format that can be converted to fasta format.

Discussion

NextGENe's Format Conversion Tool allows the software to accept sequence read output files in a variety of formats from different 2nd generation instruments. Once reads have been converted, NextGENe's Condensation Tool™ can be used to statistically polish short sequence reads by correcting low frequency instrument errors. The software assumes that the quality of the first 25 bases is higher than Phred 20 and the quality of the remainder of the read is of lower quality, the software relies on the 5' end of the sequences when condensing and lengthening the reads. The Condensation Tool can also be used to increase read length and reduce read count. Further, NextGENe can be used to analyze reads for many different applications including parsing files according to sample tags (or barcodes), SNP/Indel detection, *de novo* assembly, transcriptome analysis, ChIP-Seq, SAGE and small RNA discovery and quantification.