

File Formats Used in NextGENe™ Software for Short Sequence Reads

Kevin LeVan, Jacie Wu, Changsheng Jonathan Liu

Introduction

The new "Next Generation" DNA Sequence technologies generate millions to hundreds of millions of short sequence reads. Each system employs a unique technology:

- ◆ Illumina® Genome Analyzer: employs the Solexa sequencing technology of PCR on a surface
- ◆ ABI's SOLiD™ System utilizes emulsion PCR and sequencing by ligation
- ◆ Roche Genome Sequencer FLX System uses the pyrosequencing technology developed by 454 Life Sciences™

SoftGenetics' NextGENe software analyzes the short sequence reads from all of the above sequencing systems. The software is an excellent choice for SNP/Indel detection, Digital Gene Expression, Transcriptome Analysis, ChIPSeq, as well as *de novo* assembly.

NextGENe interprets sequence reads and quality scores that are in fasta format by default. The quality scores are used in a unique fashion to make processing possible on current desktop computers. Tools are present to remove the low quality reads. Downstream analysis tools can be used to condense short sequence reads into reads of about twice the length while filtering out reads containing base calling errors. The software assumes that the quality of the first 25 bases is higher than 99% accurate (equivalent to 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. High coverage of sequence reads will generate more reliable condensed sequences.

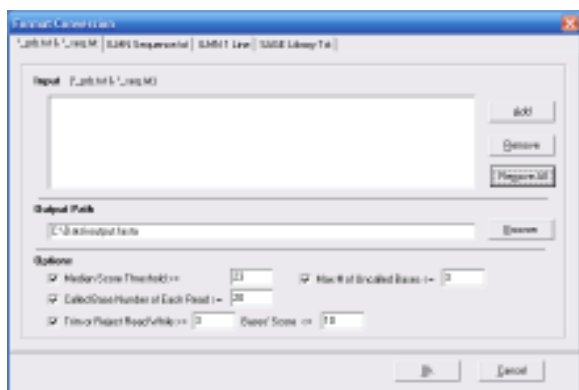


Figure 1: This tool will use the Illumina PRB and SEQ files and convert them to a fasta file. Reads can be trimmed or filtered from output based on base call quality scores.

Data Formats

All of the listed formats can be converted into information recognizable by the NextGENe software package.

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.

```
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
GCTGCTAATTAATGAAGTTATAGAACTTAATTGGT
```

In the above example, three sequence reads are shown, each containing 36 nucleotides. The comment line for each is the name assigned to the given read.

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

```
File: s_1_0001_seq.txt
1      1      137      689
AACATAATGTTGTTCACTGAGAACACATTGCACTCAA
1      1      87      649
TATTGCAACTTGTTTAATTTTTTCATGCCATTATCA
1      1      121     642
TACATGATTTCGCACTTTGGTAAATAGCTACTTTTAT
```

The sequence file records the channel number, tile number, x position and y position of each sequence read. One flow cell contains 8 channels and 300 tiles.

```
File: s_1_0001_prb.txt
```

```
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
```

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%.

NextGENe converts the seq.txt and prb.txt into a fasta format file using a Format Conversion tool shown in Figure 1. Several options are available for filtering out low quality information. Entire reads with a low median quality score can be removed. A filter is available to remove reads containing a set number of consecutive uncalled bases. Also, low quality base calls can be trimmed from the 3' end of reads when quality is consistently below a selected threshold.

Illumina Sequence Format

Illumina has an additional format starting with @ that combines the sequence reads and quality score into one file. The quality score is represented as an ascii character. The ascii character is associated with the quality score. The ascii 0 means 0 in quality score. Others are shifted.

File: s_1_sequence.txt

```
@ILMN-GA001_3_208HWAAAX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAAX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAAX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAAX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```

Roche/454 data format

Roche uses the fasta (*.fna) file and quality (*.qual) file.

File: MR3.2008_02_01.1.fna

```
>000007_1940_1402 length=172 uaccno=E4UQSRD01E0MP4
TAACAATCGAGCGAAGTCCCGTGAAGCTGTTACTTCATGATCACACAGGCGCIG
GCTCTCAGGCAAAACAGGTACGICTACGATAGGTTCCATGAAAAGTCCAGTTTGCCGA
GCTCTGGCTCCTTTTGACGCACAGTGGAACTTCTTGTTCACGGAAATTGCA
```

File: MR3.2008_02_01.1.qual

```
>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
```

Note

If you have any problems loading data into NextGENe software, we will convert them for you free of charge. Please contact us for our demo software and format conversion tools are free of charge.

Additional Application Notes for the Analysis of “Next Generation Data” with NextGENe Software:

- **Analysis of Digital Gene Expression Using Short Sequence Reads with NextGENe Software.**
- **SNP and Micro Indel Detection with NextGENe Software.**
- *de novo* **Assembly of Short Sequence Reads with NextGENe Software.**
- **Transcriptome Analysis Using NextGENe Software.**
- **NextGENe Software Tools for Analysis of Protein-DNA Interactions by ChIPSeq.**

Request a copy today: info@biogene.com