Demultiplexing Illumina® MiSeq™ Data with NextGENe® Software

February 2012

Megan Manion, Kevin LeVan, John McGuigan, Shouyong Ni, CS Jonathan Liu

Introduction

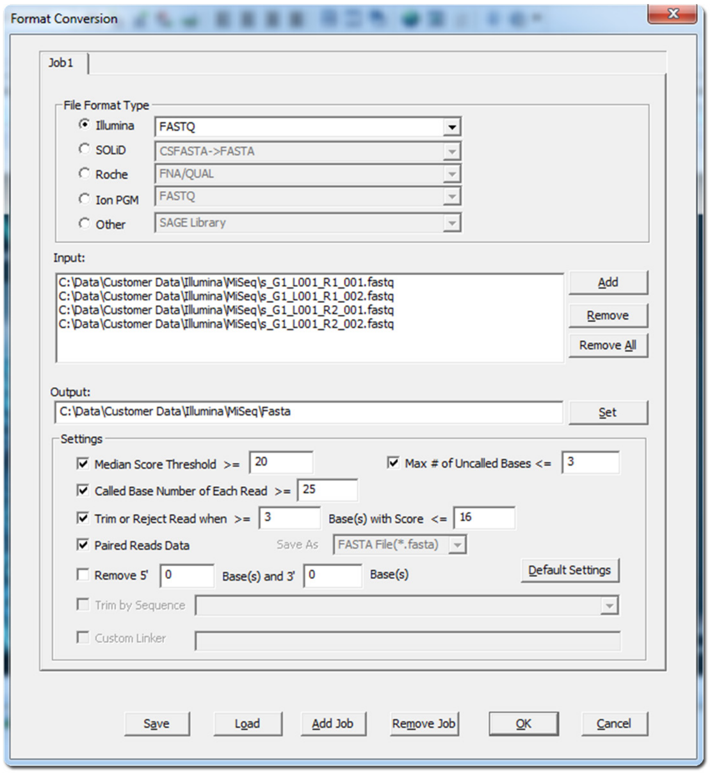
Illumina's MiSeq desktop sequencing system quickly generates more than 1.5 Gb of sequencing data per run. In many cases, if a single sample is sequenced per lane, this can provide a greater coverage depth than needed. In these cases, samples can be indexed so that multiple samples can be sequenced together (as part of the same lane). Bioinformatics solutions are then necessary to demultiplex samples during analysis.

NextGENe Software includes the Barcode Sorting Tool to demultiplex indexed sequencing data from all Next Gen systems, including the Illumina MiSeq system. NextGENe prepares the data for sorting by converting the file format and merging corresponding files, steps that can be processed in less than 2 minutes on a Windows 64-bit desktop PC. The sorting itself takes about 35 minutes for a dataset containing 96 samples.

Procedure

Convert Sample Files to Fasta Format

1. Open the Format Conversion Tool by going to Tools > Format Conversion.
2. Click Add to load all read files for a lane
   (i.e. s_G1_L001_R1_001.fastq, s_G1_L001_R1_002.fastq, s_G1_L001_R2_001.fastq and s_G1_L001_R2_002.fastq).
3. Under File Format Type Illumina and Fastq is automatically selected.
4. Choose the output folder name and location.
5. Choose Settings to quality filter data. For most datasets, the default quality filter settings are suitable.
6. Be sure to also select Paired Reads Data.
7. Click Ok.



File Format Type

- ● Illumina: FASTQ
- ○ SOLiD: CSFASTA->FASTA
- ○ Roche: FNA/QUAL
- ○ Ion PGM: FASTQ
- ○ Other: SAGE Library

Input:

```
C:\Data\Customer Data\Illumina\MiSeq\s_G1_L001_R1_001.fastq
C:\Data\Customer Data\Illumina\MiSeq\s_G1_L001_R1_002.fastq
C:\Data\Customer Data\Illumina\MiSeq\s_G1_L001_R2_001.fastq
C:\Data\Customer Data\Illumina\MiSeq\s_G1_L001_R2_002.fastq
```

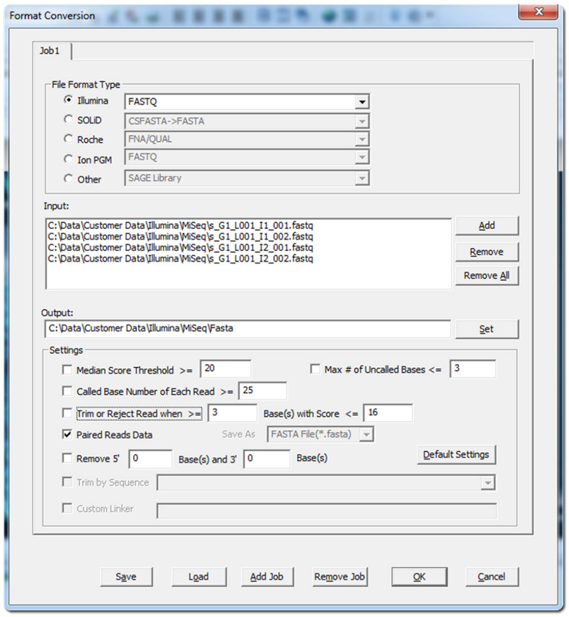
Output: C:\Data\Customer Data\Illumina\MiSeq\Fasta

Settings

- ☑ Median Score Threshold >= 20 ☑ Max # of Uncalled Bases <= 3
- ☑ Called Base Number of Each Read >= 25
- ☑ Trim or Reject Read when >= 3 Base(s) with Score <= 16
- ☑ Paired Reads Data Save As FASTA File(*.fasta)
- □ Remove 5' 0 Base(s) and 3' 0 Base(s)
- □ Trim by Sequence
- □ Custom Linker

Figure 1: Format Conversion for Read Files

When format conversion is completed a message will appear indicating the process is completed.
For each read file loaded, a *_converted.fasta and *_removed.fasta file are created in the output folder.
8. Re-open the Format Conversion Tool to convert your index files.
9. Click Add to load all index files
   (i.e. s_G1_L001_I1_001.fastq, s_G1_L001_I1_002.fastq, s_G1_L001_I2_001.fastq and s_G1_L001_I2_002.fastq).
10. Under File Format Type Illumina and Fastq is automatically selected.
11. Choose the output folder location.
12. Deselect ALL settings except Paired Reads Data.
13. Click Ok.



Format Conversion

Job1

File Format Type
● Illumina FASTQ
○ SOLiD CSFASTA->FASTA
○ Roche FNA/QUAL
○ Ion PGM FASTQ
○ Other SAGE Library

Input:
```
C:\Data\Customer Data\Illumina\MiSeq\s_G1_L001_I1_001.fastq
C:\Data\Customer Data\Illumina\MiSeq\s_G1_L001_I1_002.fastq
C:\Data\Customer Data\Illumina\MiSeq\s_G1_L001_I2_001.fastq
C:\Data\Customer Data\Illumina\MiSeq\s_G1_L001_I2_002.fastq
```

Output: C:\Data\Customer Data\Illumina\MiSeq\Fasta

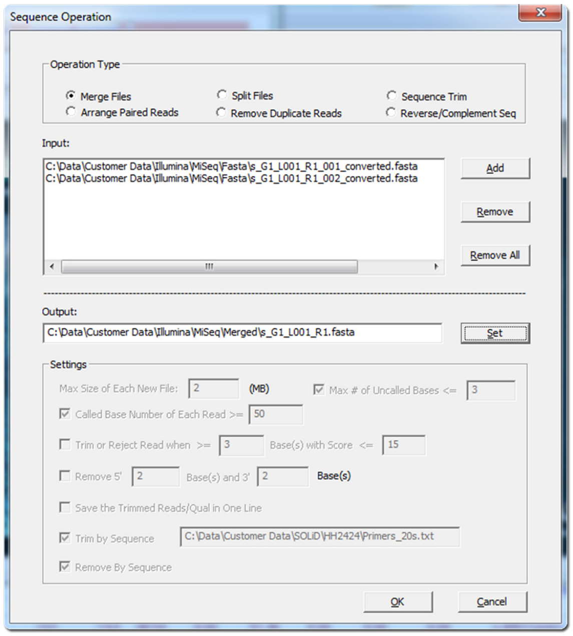
Settings
☐ Median Score Threshold >= 20 ☐ Max # of Uncalled Bases <= 3
☐ Called Base Number of Each Read >= 25
☐ Trim or Reject Read when >= 3 Base(s) with Score <= 16
☑ Paired Reads Data Save As FASTA File(*.fasta)
☐ Remove 5' 0 Base(s) and 3' 0 Base(s)
☐ Trim by Sequence
☐ Custom Linker

Figure 2: Format Conversion for Index Files

When format conversion is completed a message will again appear indicating the process is completed. For each index file loaded, a *_converted.fasta file and *_removed.fasta file are created in the output folder (although in this case the *_removed.fasta file will be empty since no quality filter settings were applied).

Merge Files

Prior to Barcode Sorting the corresponding *_001 and *_002 files need to be merged.
1. Open the Sequence Operation Tool by going to Tools > Sequence Operation.
2. Under Operation Type choose Merge Files.
3. Load corresponding *_001 and *_002 files to be merged .
   (i.e. s_G1_L001_I1_001_converted.fasta and s_G1_L001_I1_002_converted.fasta)
4. Choose output folder location and input name for the merged files.
   It will be useful to input a meaningful name, like s_G1_L001_I1_merged.fasta
5. Click OK.
Repeat this for each set of files.



Sequence Operation

Operation Type
● Merge Files ○ Split Files ○ Sequence Trim
○ Arrange Paired Reads ○ Remove Duplicate Reads ○ Reverse/Complement Seq

Input:
```
C:\Data\Customer Data\Illumina\MiSeq\Fasta\s_G1_L001_R1_001_converted.fasta
C:\Data\Customer Data\Illumina\MiSeq\Fasta\s_G1_L001_R1_002_converted.fasta
```

Output: C:\Data\Customer Data\Illumina\MiSeq\Merged\s_G1_L001_R1.fasta

Settings
Max Size of Each New File: 2 (MB) ☑ Max # of Uncalled Bases <= 3
☑ Called Base Number of Each Read >= 50
☐ Trim or Reject Read when >= 3 Base(s) with Score <= 15
☐ Remove 5' 2 Base(s) and 3' 2 Base(s)
☐ Save the Trimmed Reads/Qual in One Line
☑ Trim by Sequence C:\Data\Customer Data\SOLiD\HH2424\Primers_20s.txt
☑ Remove By Sequence

Figure 3: Merging Files

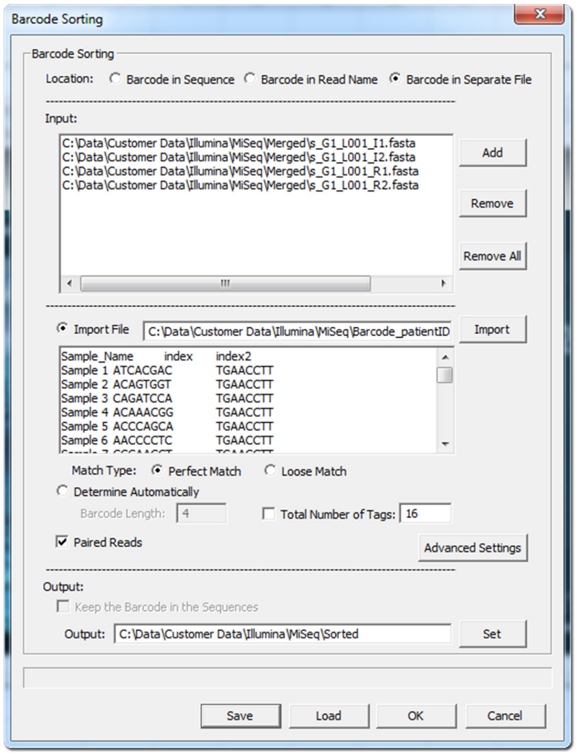
Sort Files

To demultiplex your data you'll need to create a text file which provides sample IDs and tag sequences. This allows NextGENe to output the sorted files with the sample IDs included in the file names. The text file should be in tab-delimited text format, with 3 fields – sample ID, forward tag sequence and reverse tag sequence. Each sample should start a new line. For example,

Sample1	AGCTGATG	ACGGTCGC
Sample2	ACGATCGT	GACGACTG

Your SampleSheet.csv file, provided by Illumina, can be easily modified to satisfy this format.

1. Open the Barcode Sorting Tool by going to Tools > Barcode Sorting.
2. For Location, select Barcode in Separate File.
3. Load all 4 merged files
4. Select Import File and click Import and browse to load the text file you created with the index sequences and sample names.
5. Select the Paired Reads option.
6. Specify the output folder name and location.
7. Click OK



Location: ○ Barcode in Sequence ○ Barcode in Read Name ● Barcode in Separate File

Input:

```
C:\Data\Customer Data\Illumina\MiSeq\Merged\s_G1_L001_I1.fasta
C:\Data\Customer Data\Illumina\MiSeq\Merged\s_G1_L001_I2.fasta
C:\Data\Customer Data\Illumina\MiSeq\Merged\s_G1_L001_R1.fasta
C:\Data\Customer Data\Illumina\MiSeq\Merged\s_G1_L001_R2.fasta
```

● Import File C:\Data\Customer Data\Illumina\MiSeq\Barcode_patientID

Sample_Name	index	index2
Sample 1	ATCACGAC	TGAACCTT
Sample 2	ACAGTGGT	TGAACCTT
Sample 3	CAGATCCA	TGAACCTT
Sample 4	ACAAACGG	TGAACCTT
Sample 5	ACCCAGCA	TGAACCTT
Sample 6	AACCCCTC	TGAACCTT

Match Type: ● Perfect Match ○ Loose Match

○ Determine Automatically

Barcode Length: 4 ☐ Total Number of Tags: 16

☑ Paired Reads

Output:

☐ Keep the Barcode in the Sequences

Output: C:\Data\Customer Data\Illumina\MiSeq\Sorted

Figure 4: Barcode Sorting Tool with MiSeq data files loaded

When sorting is completed, a message will appear indicating the process is completed. In the output folder, you'll find a fasta file for each sample with the sample name shown in the file name, as well as a log file which indicates how many reads were sorted into each sample file.

Discussion

NextGENe's Barcode Sorting Tool can be used to demultiplex indexed sample files from all Next Gen systems, including Illumina MiSeq, GA and HiSeq, Roche/ 454 GS FLX, FLX Titanium and Junior, Applied Biosystems' SOLiD System, and Ion PGM.

Following demultiplexing, NextGENe can be used to analyze sample data for a variety of applications such as SNP/Indel Discovery, Targeted Sequencing, RNA-Seq, de novo assembly and ChIP-Seq. Results are displayed in the interactive NextGENe Viewer, providing a detailed visualization not found in other commercial programs like Lasergene's SeqMan Pro, CLC Bio & DNASTAR's NGEN or in open-source tools like TopHat, Bowtie & BWA.

Acknowledgement

We would like to thank Illumina Corporation for supplying the MiSeq multiplexed data.