# **Demultiplexing** Illumina<sup>®</sup> MiSeq<sup>™</sup> Data with NextGENe<sup>®</sup> 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.

01		
File Format Type	•	
Illumina	FASTQ	
C SOLID	CSFASTA->FASTA	
C Roche	FNA/QUAL -	
C Ion PGM	FASTQ	
C Other	SAGE Library	
iput:		
C: \Data \Custome	er Data\Illumina\MiSeg\s_G1_L001_R1_001.fastq	Add
2: \Data \Custome	er Data ynumna Wiiseg ys_g1_001_R1_002.tastg er Data Villumina Wiiseg ys_g1_001_R2_001.fastg er Data Villumina Wiiseg ys_g1_001_R2_002_fastg	<u>R</u> emove
para (custome	er Data (diunnina (hised (\$_01_001_K2_002.hasid	Remove All
utput:		
utput: C:\Data\Custome	er Data VIIumina WiSeq \Fasta	Set
utput: C:\Data\Custome Settings V Median Sci	er Data\Illumina\MISeq\Fasta ore Threshold >= 20	<u>S</u> et
C:\Data\Custome Settings V Median Sci Called Bas	er Data Ullumina WISeq (Fasta ore Threshold >= 20 IV Max # of Uncalled Bases <= se Number of Each Read >= 25	Set
Itput: C:\Data\Custome Settings ✓ Median Sci ✓ Called Bas ✓ Trim or Rej	er Data'lllumina'MSeq\Fasta ore Threshold >= 20	<u>Set</u>
Settings Vertings C:\Data\Custome Settings Vertings Called Bas Verting Trim or Re Verting Paired Rea	er Data'Illumina'MSeq\Fasta ore Threshold >= 20 iv Max # of Uncalled Bases <= ive Number of Each Read >= 25 iject Read when >= 3 Base(s) with Score <= 16 ads Data Save As FASTA File(*, fasta) v	Set
Itput: C:\Data\Custome Settings ✓ Median Sci ✓ Called Bas ✓ Trim or Re ✓ Paired Rez ✓ Remove 5	er Data Vilumina WiSeq VFasta ore Threshold >= 20 V Max = of Uncalled Bases <= se Number of Each Read >= 25 gect Read when >= 3 Base(s) with Score <= 16 ads Data Save As FASTA File(*, fasta) v 7 0 Base(s) and 3' 0 Base(s) <u>Defau</u>	Set
Itput: C:\pata\Custome Settings ✓ Median Sci ✓ Called Bas ✓ Trim or Re ✓ Paired Rez ← Remove 5 ← Trim by Se	er Data Vilumina WiSeq VFasta ore Threshold >= 20 V Max # of Uncalled Bases <= se Number of Each Read >= 25 gect Read when >= 3 Base(s) with Score <= 16 ads Data Save As FASTA File(*.fasta) v tr 0 Base(s) and 3' 0 Base(s) <u>Defau</u> equence	Set
itput: C: \pata \Custome Settings C Called Bas C Called Bas C Trim or Re C Paired Res Remove 5' Trim by Se Custom Lin	er Data VIIlumina WISeq\Fasta ore Threshold >= 20	3 it Settings





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



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.

File Format Typ	2	
Illumina	FASTQ	
C SOLID	CSFASTA->FASTA	
C Roche	FNA/QUAL	
C Ion PGM	FASTQ	
C Other	SAGE Library	
nput:		
C: Data Custom	er Data\Illumina\MiSeq\s_G1_L001_I1_001.fastq	Add
C: \Data (Custom) C: \Data (Custom) C: \Data (Custom)	r Data yuumma ymseq ys_o1_L00 _11_002.18\$10 rr Data\Jllumina\MiSeq ys_o1_L001_12_001.fastq rr Data\Illumina\MiSeq ys_o1_C01_12_002.fasta	Remove
er para leastern	n one frammet med h of foot 1 n foot upped	Remove All
C: \Data\Custom	er Data Ullumina (MiSeq (Fasta	Set
Median So	ore Threshold >= 20	3
Called Ba	e Number of Each Read >= 25	
Trim or Re	ject Read when >= 3 Base(s) with Score <= 16	
Paired Re	ads Data Save As FASTA File(*.fasta)	
Remove 5	0 Base(s) and 3' 0 Base(s) Defaul	t Settings
Trim by Se	auence	×
	obou on	
Custom Li	nker	

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.

Merge Files	C Split Files	C Sequence T	irim
C Arrange Paired Reads	C Remove Duplicate Read	ds C Reverse/Co	mplement Seq
nput:			
: \Data\Customer Data\Illumina\	MiSeg\Fasta\s G1 L001 R1 00	1 converted.fasta	Add
C:\Data\Customer Data\Illumina\	MiSeq\Fasta\s_G1_L001_R1_00	2_converted.fasta	
			Remove
			Remove All
•			
•			
<			
<ul> <li>Iutput:</li> <li>C:\Data\Customer Data\Illumina\</li> </ul>	MiSeqWerged\\$_G1_L001_R1.f	asta	Set
<ul> <li>Lutput:</li> <li>C:\Data\Customer Data\Illumina\</li> </ul>	™ MiSeqWerged\s_G1_L001_R1.f	asta	<u>Set</u>
<ul> <li>Lutput:</li> <li>C: \Data \Customer Data \IIlumina \</li> <li>Settings</li></ul>	MSeqWerged\s_G1_L001_R1.f	asta	<u>Set</u>
<ul> <li>Settings</li> <li>Max Size of Each New File:</li> </ul>	/////////////////////////////////////	asta Tax # of Uncalled Bases <≈	<u>Set</u>
Ling     Ling     Ling     Ling     Max Size of Each New File:     Called Base Number of Eac	m MiSeqWerged g_G1_L001_R1.f 2 (MB) F M h Read >= 50	asta 1ax # of Uncalled Bases <<	<b>Set</b>
Liput:     Cityput:     Citypata/Customer Data/Jilumina'     Settings     Max Size of Each New File:     ✓ Called Base Number of Eac	MiSeqWerged/ø_G1_L001_R1.f       2     (MB)       bh Read >=     50	asta Iax # of Uncalled Bases <<	<b>Set</b>
C:Data/Customer Data/Illumina' C:Data/Customer Data/Illumina' Settings Max Size of Each New File:	MSeqWergedk_G1_001_R1.f 2 (MB) F M h Read >= 50 >= 3 Base(s) with S(	asta lax # of Uncalled Bases << core <= 15	<b>Set</b>
	MiSeqWerged%_G1_001_R1.f           2         (MB)           bh Read >=         50           >=         3         Base(s) with S4           ase(s) and 3'         2         Base	asta tax # of Uncalled Bases <- core <= 15 c(a)	<u>Set</u>
	MiSeqWerged/s_G1_001_R1.fl       2     (MB)       h Read >=     50       >=     3     Base(s) with 54       ase(s) and 3'     2     Base	asta lax # of Uncalled Bases <- core <= 15 e(a)	<u>Set</u>
C: [Data/Customer Data'][Ilumina' Settings Max Size of Each New File: Called Base Number of Eac Trim or Reject Read when Remove 5' 2 B Save the Trimmed Reads/t	//MSeq!Werged!/a_G1_L001_R1.f           2         (MB)         [7" /// /// /// /// /// /// /// /// /// /	iasta lax # of Uncalled Bases <= core <= 15 c(s)	
	Miseq/Merged/s_G1_L001_R1.f           2         (MB)         [7] M           h Read >=         50         50           >=         3         Base(s) with Sc           ase(s) and 3'         2         Base           2ual in One Line         Vabal/Sustamer Data/SOLD/H*         10	iasta lax # of Uncalled Bases << core <= 15 e(s) 12424\Primers_20s.txt	<u>Şet</u>

Figure 3: Merging Files

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



#### **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,

	$\mathcal{O}$	1		
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

ocation: C Barcode in Se	equence C Barcode in Read Na	ame 👎 Barcode i	n Separate Fi
nput:			
C: \Data \Customer Data \I C: \Data \Customer Data \I C: \Data \Customer Data \I C: \Data \Customer Data \I	llumina WiSeq Werged \s_G1_L00 llumina WiSeq Werged \s_G1_L00 llumina WiSeq Werged \s_G1_L00	1_I1.fasta 1_I2.fasta 1_R1.fasta	Add
C: pata (Customer Data µ	liumina (MiSeq (Merged \s_G 1_LUU	1_R2.fasta	Remove
			Remove All
•	III	Þ	
Sample 1 ATCACGAC Sample 2 ACAGTGGT	TGAACCTT		
Sample 3 CAGATCCA Sample 4 ACAAACGG Sample 5 ACCCAGCA Sample 6 AACCCCCC Match Type:	TGAACCTT TGAACCTT TGAACCTT TGAACCTT TGAACCTT TGAACCTT TGAACCTT TGAACCTT	-	
Sample 3 CAGATCCA Sample 4 ACAACGG Sample 5 ACCCAGCA Sample 6 AACCCACCTC Match Type:	TGAACCTT TGAACCTT	₹ fTags: 16	
Sample 3 CAGATCCA Sample 3 CACAACGG Sample 5 ACCCAGCA Sample 6 AACCCCTC Match Type:	TGAACCTT TGAACCTT TGAACCTT tGAACCTT tGAACCTT tGAACCTT TGAACCTT	FTags: 16 Advar	nced Settings

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.

Trademarks are property of their respective owners



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

