NextGENe™ by SoftGenetics

March 2009

Sequence Analysis Using Barcode/Index Tags of Pooled Samples with NextGENe Software

Megan Manion, Kevin LeVan, Ni Shouyong and CS Jonathan Liu

Introduction

The development of next generation sequencing technologies such as the Genome Analyzer FLX by Roche Applied Science (454 Sequencing), the SOLiD™ system from Applied Biosystems and the Illumina Genome Analyzer have drastically lowered sequencing costs while increasing speed and the quantity of information gathered. These technologies give reliable sequence read-outs of 25-75 bps for Illumina and SOLiD reads and up to 500 bp for Roche/454 reads with approximately 1-100 million reads per sequencing run. For certain applications aimed at sequencing small genomes or select regions of DNA this amounts to an excess of sequencing data producing greater coverage than necessary. For instance, the sequencing of a chloroplast genome by high throughput sequencing can produce roughly 12,500x coverage (1). The use of multiplex sample tags (or barcodes) allows for optimization of 2nd generation sequencing technologies by pooling samples and sequencing multiple samples in parallel (2). Tags are used to distinguish samples from one another. In addition to the sequencing of small genomes, other applications where barcoding is useful include sequencing a specific gene region in multiple patients or sequencing the same region between different species to study phylogenetic relationships. When barcode tags are used for multiplexing, software is needed to accurately parse sample files according to these tags prior to analysis.

NextGENe software is an easy-to-use, Windows based program that is able to analyze all 2nd Generation sequencing data with barcode sample tags. NextGENe can be used to parse sample files according to barcode tags (MIDs). Since different instruments output barcode data in different formats, NextGENe offers a flexible tool to accommodate these. The Barcode Sorting function can be used for files where barcodes are included in the sequence of reads or in the read name. For multiplex data that includes barcode tags within the sequences, NextGENe is able to trim the tag sequences from reads and parse reads according to tags. After parsing files, each sample can be analyzed independently by NextGENe for a variety of different applications. Since samples are pooled in this type of analysis, input DNA for each sample is reduced. Thus, when a large genomic region or a whole genome is of interest, care must be taken to maintain adequate coverage. We have found that SNPs can be detected when a minimum of 10-20 reads are aligned to a location.

Different platforms use different methods for producing reads with barcode tags. Illumina multiplexed sequencing uses the Genome Analyzer with the Paired-End Module. There are three steps in processing sample DNA to produce barcoded reads; adding the sequence index to the library, PCR amplification on solid surface, and the sequencing read-out of the sample DNA with the index. In the first step, Adapters are ligated to DNA fragments during sample preparation, with the adapter at 5' end containing the sequencing primer site and the adapter at the 3' end containing the index sequencing primer site. The index sequence, together with attachment primer for contacting the solid surface, is linked to the index sequencing primer. The index sequence for Illumina data is generally six bp in length. The resultant library of the sample with index sequence contains attachment site at 5' end, sequencing primer (5' end), DNA insert of interest, index sequence primer, index sequence and the attachment site at 3' end.

In the second step, Bridge amplification is done on the solid surface using attachment primers. The third step is to read barcode sample sequences. Illumina uses two sequencing reads, or three reads for paired end analysis with barcode tags. First, Read 1 Sequencing Primer generates the sequence of the sample DNA. Second, after Read 1 product is removed, the Index Sequencing Primer is annealed to the index adapter for sequencing the barcode tag, or index. Third, for paired-end samples, an optional third read is used to sequence the strand complementary to the original template strand. We have determined that the barcode sequence error in the Illumina system is about 3%.

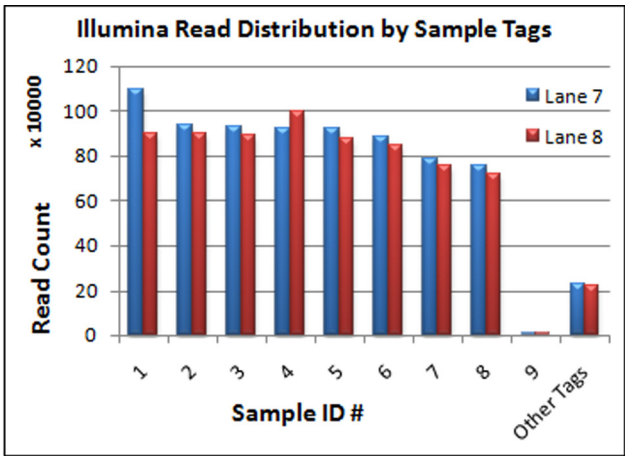

Figure 1: Read distribution for two lanes of Illumina multiplex data with sample tags in read names is shown. NextGENe's Barcode Sorting Tool can automatically detect sample tag count by evaluating the number of reads with each tag sequence. For this data, eight sample tags were used to create new files. All other tags, which made up roughly 3% of the total reads, were included in an "Other Tags" file. These low frequency tags are considered the result of sequencing errors. The column for Sample ID 9 represents the most common tag from the "Other Tags" file.

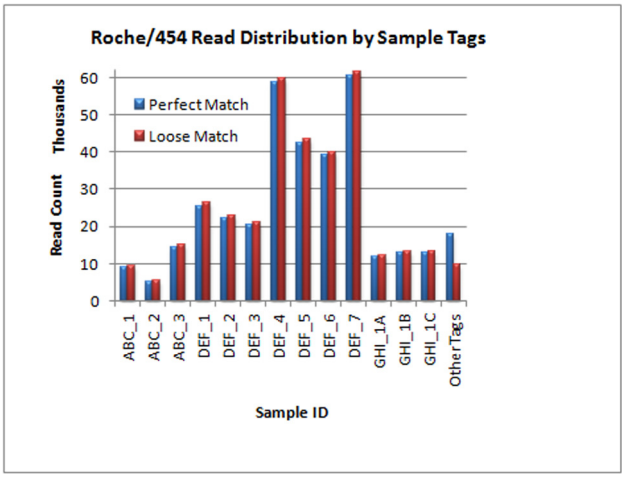

Figure 2: The read distribution for one Roche multiplex file is shown when a Barcode/Primer file is used with a Perfect Match or a Loose Match. The "Other Tags" column includes all reads with tags that were not in the Barcode/Primer file, likely the result of sequencing errors When perfect match is used, the error rate is 5.1%, while the error rate when Loose Match is used is 2.7%.

Roche FLX uses a different approach to accomplish the same three general steps. First, reverse and forward primers, which consist of an adapter sequence, the barcode sequence, and a cloning linker, are attached to sample DNA during sample preparation. Sequence tags are generally ten to twenty bp in length. Primers are attached using blunt-end ligation of the cloning-linker to the sample DNA fragment. Once samples are prepared, multiple samples can be pooled and processed simultaneously using the standard protocols. Second, Emulsion PCR is done to amplify samples. Third, samples are sequenced using the pyrosequencing technique, measuring the light given off when dNTPs are added to sample DNA templates during polymerase reaction. Roche uses a single read to sequence the sample tag, or multiplex identifier (MID), in conjunction with the sample DNA fragment. We have determined the error rate for multiplexed reads from the Roche FLX system to be about 5% when perfect matching is used and 2.7% when loose matching is used, allowing one error. The Barcode Sorting Tool in NextGENe software parses reads with barcode/index tags into multiple files. A user may specify the number of the tags or software may determine them automatically. Multiple tags could be associated with one sample.

Procedure

1. Open Tools: Barcode Sorting in NextGENe's main toolbar.
2. Specify if Barcode Tags are in Sequence or in Read Name.
3. Add sample file.
4. If Barcodes are in Sequence:
    a. Choose to Import Barcode/Primer File or allow software to automatically determine tags.
    b. When Importing a Barcode/Primer File, select "Perfect Match" or "Loose Match" for matching tag sequences.
    c. When allowing software to Determine Automatically, input the barcode tag length and, if known, the number of tags.
5. If Barcodes are in Read Name:
    a. Check "Total Number of Tags" and input expected number of tags, if known.
    b. Alternatively, leave "Total Number of Tags" unchecked to allow software to automatically determine the number of tags.
6. Specify Output location.
7. Click "Run" to begin processing.
8. The Barcode Sorting Tool closes when processing is completed.

Users can create a Barcode/Primer File by listing information about Sample IDs, forward and reverse primers/barcode tags. NextGENe also supports more than two tags for the same sample. An example of a Barcode/Primer File is shown. Each line includes Patient ID, followed by the forward barcode tag and finally the reverse barcode tag, each separated by the tab key.

Sample_ID	Forward Tag	Reverse Tag
ABC_1	GTGAGGCTTGTCTCAAAGATTAAGCC	GTGAGGCCTGCTGCCTTCCTTGGA
ABC_2	TACGCGCTTGTCTCAAAGATTAAGCC	TACGCGCCTGCTGCCTTCCTTGGA
ABC_3	GTCACGCTTGTCTCAAAGATTAAGCC	GTCACGCCTGCTGCCTTCCTTGGA

Table 1: An example Barcode/Primer File is shown in tab-delimited text format.

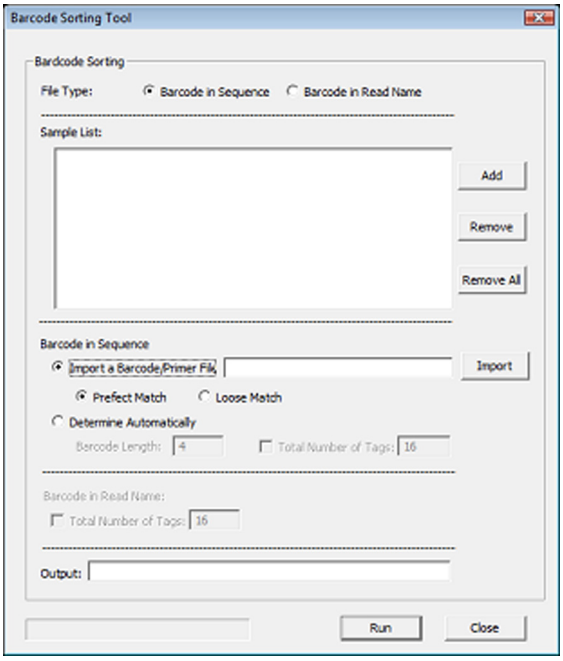

Figure3: NextGENe's Barcode Sorting Tool is flexible to handle barcoded data in different formats.

Methodology

Tags in Read Name used by Illumina

Illumina Genome Analyzer reads the DNA and index sequence in two steps. The barcode tags/index sequences are included in the read name and are not part of the sequence read. Users can input the number of tags expected, if known, so that NextGENe parses the sample file according to the most common tags up to the set value. If the number of barcode tags is not known the software can automatically determine the number of tags present at a significant frequency.

```
@R0174436_30DL5AAXX_164:7:1:243:1870#CAGATC
GTGTGGTGGCCTTGGTATGTGCTTTCTCGTGTTAC
+R0174436_30DL5AAXX_164:7:1:243:1870#CAGATC
ZZZUZZZZZZZZZZLZZZZZZZZZZZZTZTZZMV
@R0174436_30DL5AAXX_164:7:1:286:1769#TGACCA
GATTAGTCGGTTGTTGATGAGATATTTGGAGGTGG
+R0174436_30DL5AAXX_164:7:1:286:1769#TGACCA
ZZZZZZZZZZZZYZZYZZYYZZZVZZZYLVULZVC
```

Figure 4: Illumina sequencing reads output in FASTQ format with barcodes included in read name. The end of the read's name line shows "#" symbol followed by the read's barcode sequence.

When determining the sequence tag count automatically, only the most common tag sequences are used to parse files. This is to avoid parsing reads according to tag sequences that are the results of sequencing errors. For automatic detection of tag count two criteria are used to determine the minimum frequency for an observed tag to be used for file parsing. The Barcode Sorting Tool evaluates the number of reads with each observed tag in order of descending frequency. When the count of reads that contain a sample tag is less than 10% of the count for the previous tag, that tag is considered to be an error, is not used, and barcode sorting is complete. Also, once 95% of the sample reads have been parsed by barcode, one additional tag is used for sorting and then sorting is completed. Reads containing all other tags that are found are grouped into a separate file.

Tags in Sequence used in Roche systems.

When multiplex sample tags are included in the sequence, NextGENe can detect tags automatically or users can create and input a Barcode/Primer File in tab-delimited text format that contains information about the tags used.

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 a Barcode/Primer File is used, reads can be matched to the specified tags using a "Perfect Match" or "Loose Match." When "Perfect Match" is selected the tag for a read must match perfectly with the tag in the Barcode/Primer file to be allocated with that tag. When "Loose Match" is selected, the beginning of the read (tag) is compared to three segments of equal size within the tag sequence; the first half, the second half and the middle region of the tag. For a "Loose Match" only one of these three segments must match exactly to the tag in the Barcode/Primer file. The "Loose Match" method is useful for longer tag sequences, such as the 20 bp tags available with Roche reads, where the likelihood of sequencing errors within the tag region is high.

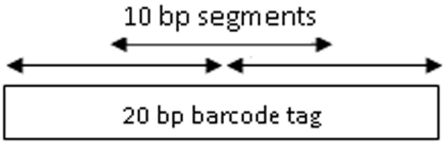

Figure 5: NextGENe's Barcode Sorting Tool is able to allow for a small amount of error in matching tags by comparing the barcode to 3 segments within the read tags and requiring only one segment to match perfectly with the expected tag. Users can also choose to require a perfect match between the expected tag and the read tag. Some portions with same sequence in different barcodes will be ignored in the detemination of the sample.

When detecting tags without the use of a Barcode/Primer file, the Barcode Sorting Tool can detect tag sequences automatically when the tag length is provided.

The total number of tags used can be set by the user or the software can automatically detect the number of tags that are found at a significant frequency. NextGENe's Barcode Sorting Tool parses files according the most frequently observed tags using the method described in the previous section to evaluate whether an observed tag is found at a significantly high frequency.

Results

Following Barcode Sorting, multiple fasta files are created containing reads with tags that were found at high frequency or those specified by the user. A log file is also created once parsing is completed to provide statistics about the process by listing each output file name and the number of reads assigned to each new file.

NextGENe's Barcode Sorting Tool is able to accurately determine the number of tags used in sample preparation. Tags that are found at low frequency, often the result of sequencing errors, are not used for parsing the sample file.

As shown in the ouput results in Figure 6, the Barcode Sorting Tool created new sample files for this data for sample IDs 1-8 as well as an Other Tags file that includes all reads with tags found at frequencies lower than sample ID8. Comparing the number of reads with each tag (Figure 1) allows for clear differentiation between true tags and tags that are the result of sequencing errors. The clear drop-off of read count between Sample ID 8 (the least common tag used for parsing) and ID 9 (the next most common tag) illustrates the distinct boundary between true tags and tags caused by error. The total number of reads in the "Other Tags" file is significantly less than for each of the true tag files. The reads in this file represent 3.1% of the total read count, indicating the barcode sequence error of roughly 3%.

As shown in Figure 7, when sample files are parsed by the Barcode Sorting Tool using a Barcode/Primer file, the new files are named according to the Sample IDs provided for easy identification of samples for further analysis. When the "Loose Match" method is used to allow for one sequencing error within the tag sequence, the error rate can be reduced from roughly 5% when "Perfect Match" is used to roughly 3%.

Figure 6: A list of output sample files produced by the Barcode Sorting Tool using automatic detection of tag sequences and tag count is shown. The convert.log file contains information about how the file was parsed including number of reads in each new file. The OtherTags.fasta file includes reads with tags found at low frequency. For this sample file, eight tags were found at high enough frequency to be used for parsing reads

Figure 7: A list of output sample files produced by the Barcode Sorting Tool using a Barcode/Primer File is shown. The convert.log file contains information about how the file was parsed including number of reads in each new file. The OtherTags.fasta file includes reads with tags other than those specified by the user.

Discussion

The utilization of multiplex sequencing techniques allows multiple samples to be sequenced together, ideal for the sequencing of small genomes or genomic regions while taking full advantage of the high output of massively parallel sequencing systems. NextGENe software provides a Windows-based user friendly application for parsing sample files according to tags. Following sorting, NextGENe can be used for further analysis in a variety of different applications.

NextGENe includes a unique, patent-pending Condensation Tool™ that can be used to improve read quality by lengthening reads and statistically removing instrument errors. NextGENe also includes software modules for de novo assembly, SNP and Indel Detection, ChIP-Seq, Transcriptome, small RNA discovery and quantification and SAGE analysis. Both de novo assembly and SNP/Indel Detection can be analyzed with or without the use of paired reads.

References

1. R Crocc et al. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research. 36(19): e122.
2. P Parameswaran et al. 2007. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Research. 35(19): e130.

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