A Simple GUI-based Sequencing Format Conversion Tool for the Three NGS Platforms

Arang Rhie, Sanduk Yang, Kyung-Eun Lee, Chin Ting Thong and Hyun-Seok Park*

Department of Computer Science, Ewha Womans University, Seoul 120-750, Korea

Abstract

To allow for a quick conversion of the proprietary sequence data from various sequencing platforms, sequence format conversion toolkits are required that can be easily integrated into workflow systems. In this respect, a format conversion tool, as well as quality conversion tool would be the minimum requirements to integrate reads from different platforms. We have developed the Pyrus NGS Sequencing Format Converter, a simple software toolkit which allows to convert three kinds of Next Generation Sequencing reads, into commonly used fasta or fastq formats. The converter modules are all implemented, uniformly, in Java GUI modules that can be integrated in software applications for displaying the data content in the same format.

Availability: You may download the conversion module from Sourceforge.net (https://sourceforge.net/projects/ngssequencealig/files/PyrusNGSSequencingFormatConverter.zip/download).

Keywords: sequence format conversion, next generation sequencing

Introduction

Among the NGS (Next Generation Sequencing) technologies, three distinct platforms have attained wide diffusion (Horner *et al*., 2010; Shendure and Ji, 2008): the Roche Genome Sequencer System (Droege and Hill, 2008; Rothberg and Leamon, 2008), the Illumina® Genome Analyzer (Bennett, 2004), and the Applied Biosystems SOLiD™ System (Pandey *et al*., 2008; Porreca *et al*., 2006). However, these instrument suppliers use different formats for organizing the reads and assigning quality scores (Cock *et al*., 2010). Thus, the management and analysis of next-generation sequencing data requires the development of format conversion tools to integrate huge quantities of sequence reads. The three instrument suppliers all provide their own analysis tools: Illumina/Solexa Genome Analyzer (http://www.illumina.com/software.ilmn), Roche 454 (http://454.com/products-solutions/analysis-tools/index.asp), and the Solid (http://solidsoftwaretools.com/gf/) software development community. Some of them even support the development of open-source bioinformatics tools.

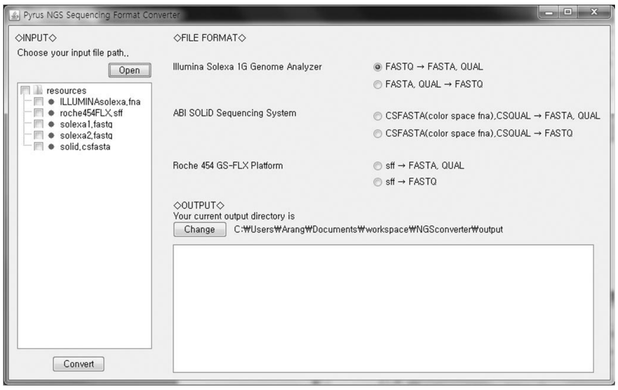
However, there are few open source systems to integrate the different reads from the same samples. On the other hand, multiple technical platforms or different versions of the same platform were used for a large-scale sequencing project, in many studies (MacLean *et al*., 2009; Miller *et al*., 2010). Thus, it became necessary to pool information across these multiple sources to derive a consensus molecular profile for each sample.

Thus, we developed the Pyrus Sequence Format Converter, written in Java. The format conversion tool allows the software to accept sequence read files from these three different platforms.

The Pyrus Sequence Format Convertsion Tool

The Pyrus NGS Sequence Format Converter is designed to read sequence data from the Illumina® Genome Analyzer, the Applied Biosystems SOLiD™ System, and the Roche Genome Sequencer System. The system converts all sequencing reads into a standard fasta or fastq format. Fig. 1 shows the Graphical User Interface of the system. Users can select the file format of their data, upload sample files, select input files to convert, and click Convert to begin format conversion. The directory file tree shows the files with extension of fna, fasta, csfasta, fastq, and sff. The input directory tree does not show .qual files, which are assumed to be named the same as the corresponding FASTA formatted file so that user does not have to choose the coherent .qual file. By pressing the Convert button, the Format Converter automatically checks with input file extensions and proceeds, once the file and formats are correctly mapped. User can change the output directory by pressing the Change button.

*Corresponding author: E-mail neo@ewha.ac.kr
Tel +82-2-3277-2831, Fax +82-2-3277-2306
Accepted 17 June 2010



Pyrus NGS Sequencing Format Converter

◇INPUT◇

Choose your input file path,,

Open

- □ resources
  - □ ● ILLUMINAsolexa.fna
  - □ ● roche454FLX.sff
  - □ ● solexa1.fastq
  - □ ● solexa2.fastq
  - □ ● solid.csfasta

Convert

◇FILE FORMAT◇

Illumina Solexa 1G Genome Analyzer
- ● FASTQ → FASTA, QUAL
- ○ FASTA, QUAL → FASTQ

ABI SOLiD Sequencing System
- ○ CSFASTA(color space fna),CSQUAL → FASTA, QUAL
- ○ CSFASTA(color space fna),CSQUAL → FASTQ

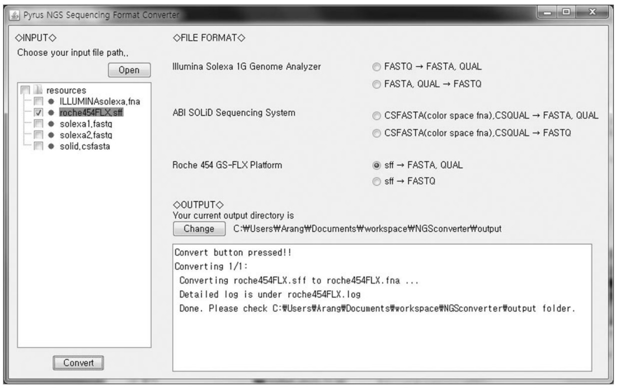
Roche 454 GS-FLX Platform
- ○ sff → FASTA, QUAL
- ○ sff → FASTQ

◇OUTPUT◇

Your current output directory is

Change C:₩Users₩Arang₩Documents₩workspace₩NGSconverter₩output

Fig. 1. User Interface of the Pyrus NGS Sequence Format Converter: users select the file format of their data, upload sample files, and click Change to begin format conversion. File format "FASTQ to FASTA, QUAL" has been chosen by default. For SOLiD^{TM} System reads, the csfasta format is used. A log file is created to provide a summary of the process.



Pyrus NGS Sequencing Format Converter

◇INPUT◇

Choose your input file path,,

Open

- □ resources
  - □ ● ILLUMINAsolexa.fna
  - ☑ ● roche454FLX.sff
  - □ ● solexa1.fastq
  - □ ● solexa2.fastq
  - □ ● solid.csfasta

Convert

◇FILE FORMAT◇

Illumina Solexa 1G Genome Analyzer
- ○ FASTQ → FASTA, QUAL
- ○ FASTA, QUAL → FASTQ

ABI SOLiD Sequencing System
- ○ CSFASTA(color space fna),CSQUAL → FASTA, QUAL
- ○ CSFASTA(color space fna),CSQUAL → FASTQ

Roche 454 GS-FLX Platform
- ● sff → FASTA, QUAL
- ○ sff → FASTQ

◇OUTPUT◇

Your current output directory is

Change C:₩Users₩Arang₩Documents₩workspace₩NGSconverter₩output

```
Convert button pressed!!
Converting 1/1:
 Converting roche454FLX.sff to roche454FLX.fna ...
 Detailed log is under roche454FLX.log
 Done. Please check C:₩Users₩Arang₩Documents₩workspace₩NGSconverter₩output folder.
```

Fig. 2. A snapshot after converting 454 Roche GSFLX .sff file into a regular FASTA and Phred quality valued Qual file.

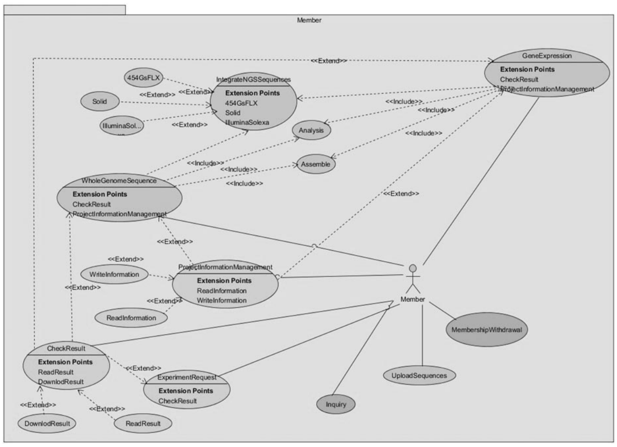


Fig. 3. Use Case Diagram for the SMBA Industrial-Educational Cooperation Project (000358780109): the diagram is created, using the Visual Paradigm UML Suite, community edition (http://www.visual-paradigm.com/).

Integrating Quality Values

Analogously to automated Sanger sequencing, NGS platforms provide quality scores describing the correctness of a base call. NGS platforms have different error profiles and, thus, quality values need to be derived accordingly (Cock *et al.*, 2010; Harismendy *et al.*, 2009). In the Illumina® Genome Analyzer, and the 454 Roche Genome Sequencer System, the meaning of the quality values is relatively close to capillary sequencers, though the range of Sanger Phred quality scores and Illumina or Roche quality scores are different. In the SOLiD^{TM} system, quality scores are assigned to each color and they are calculated using a Phred-like score.

Our system integrates these quality values from different platform, according to the algorithms described in various online resources, though some of the exact relationship between NGS scores and Phred values is not completely known, yet. For an up-to-date discussion on the implications of quality scores for the NGS platforms, there exists an online forum (http://seqanswers.com). Fig. 2 shows a snapshot after converting 454 Roche GSFLX .sff file into a regular FASTA and Phred quality valued Qual file.

Software Release Stage and Conclusion

Recently, a well known bioinformatics integration tool such as Galaxy Tools (Giardine *et al.*, 2005) begins to add new libraries, in regards to NGS data processing. However, the core component and operation libraries are written in C. We developed a Java-based sequence format converter, with graphical user interface. This is part of our ongoing effort to develop a large-scale DNA comparative alignment LIMS system for multiple NGS sequencing platforms, as in Fig. 3. The Java-based project is supported by the SMBA Industrial-Educational Cooperation Project (000358780109) of the Korean government. Our system is currently in pre-alpha release.

References

Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics* 5:433-438.

Cock, P., Fields, C., Goto, N., Heuer, M., and Rice, P. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, *Nucl. Acids Res.* 38, 1767-1771.

Droege, M., and Hill, B. (2008). The Genome Sequencer FLX System-longer reads, more applications, straight forward bioinformatics and more complete data sets. *J. Biotechnol.* 136, 3-10.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* (10), 1451-1455.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., and Frazer, K.A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 10, R32.

Horner, D.S., Pavesi, G., Castrignanò, T., De Meo, P.D., Liuni, S., Sammeth, M., Picardi, E., and Pesole, G. (2010). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief. Bioinfo.* 11, 181-197.

MacLean, D., Jones J.D., and Studholme, D.J. (2009). Application of ‘next-generation’ sequencing technologies to microbial genetics, *Nat. Rev. Microbiol.* 7, 287-296.

Miller, J., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data, *Genomics* 95, 315-327.

Pandey, V., Nutter, R.C., and E, E.P. (2008). Applied Biosystems SOLiD system: ligation-based sequencing. In *Next Generation Genome Sequencing: towards personalized medicine,* Janitz, M, ed. Weinheim, Wiley-VCH, pp. 29-41.

Porreca, G.J., Shendure, J., and Church, G.M. (2006). Polony DNA sequencing. *Curr ProtocMol. Biol.* Chapter 7:Unit:7-8.

Rothberg, J.M., and Leamon, J.H. (2008). The development and impact of 454 sequencing. *Nat. Biotechnol.* 26, 1117-1124.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135-1145.