RAPID COMMUNICATION

Genomics & Informatics Vol. 6(2) 87-90, June 2008

An Optimized Strategy for Genome Assembly of Sanger/pyrosequencing Hybrid Data using Available Software

Haeyoung Jeong[1] and Jihyun F. Kim[1,2]*

[1]Laboratory of Microbial Genomics, Systems Microbiology Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB) and [2]Field of Functional Genomics, School of Science, Korea University of Science and Technology (UST), Daejeon 305-806, Korea

Abstract

During the last four years, the pyrosequencing-based 454 platform has rapidly displaced the traditional Sanger sequencing method due to its high throughput and cost effectiveness. Meanwhile, the Sanger sequencing methodology still provides the longest reads, and paired-end sequencing that is based on that chemistry offers an opportunity to ensure accurate assembly results. In this report, we describe an optimized approach for hybrid *de novo* genome assembly using pyrosequencing data and varying amounts of Sanger-type reads. 454 platform-derived contigs can be used as single non-breakable virtual reads or converted to simpler contigs that consist of editable, overlapping pseudoreads. These modified contigs maintain their integrity at the first jumpstarting assembly stage and are edited by fragmenting and rejoining. Pre-existing assembly software then can be applied for mixed assembly with 454-derived data and Sanger reads. An effective method for identifying genomic differences between reference and sample sequences in whole-genome resequencing procedures also is suggested.

Abbreviations: CelAsm (Celera Assembler)

Keywords: hybrid assembly, pyrosequencing, resequencing

The 454 sequencing platform (Roche Applied Science GS 20 or GS FLX), which is based on massively parallel sequence determination by pyrosequencing on clonally amplified genome fragments that are captured on microscopic beads, is becoming more and more popular in genome sequencing applications (Margulies *et al.,* 2005). Its characteristics, which are superior to the traditional Sanger method - such as high production rate with an affordable cost, absence of cloning bias, and ability to go beyond strong secondary structure - enlarge its field of application in genome technology. Although there are several commercial next-generation sequencing technologies that have become available in recent years (Shendure *et al.,* 2004), 454 pyrosequencing is the only one that can be used for *de novo* genome sequencing among the high-throughput, short-read sequencing technologies due to its long read length (∼250 bp in GS FLX; announced to be extended to 400 bp by the end of 2008).

Many sequencing centers, however, may want to mix a limited amount of traditional Sanger-type sequences, usually generated from fosmid libraries, for scaffolding purposes. Also, a few may want to mix a considerable amount of Sanger read data to 454 pyrosequencing data to produce more accurate results. Among the SFF tools that Roche Applied Science provides for the handling of raw data files, SFFINFO can generate FASTA and quality score files from an SFF file. Although the converted files can be assembled using PHRAP (http://www.phrap.org/), it does not ensure correct assembly because the quality scores that are generated from 454 data are not compatible with those from Sanger reads. Further, PHRAP has problems with handling massive reads (usually hundreds of thousands from an SFF file). A recent report has demonstrated that GS assembler programs (gsAssembler for *de novo* assembly and gsMapping for reference-guided assembly; http://www.454.com/enabling-technology/the-software.asp) that are supplied by Roche Applied Science are ideal for correct assembly of 454 data that are short and inherently error-rich (Chaisson and Pevzner, 2008).

Recent versions (1.1.02.15 and later) of GS assembler programs support mixed assembly with Sanger-type reads, but their performance is not well known at present. Moreover, because pre-existing assembly software such as PHRAP and CelAsm (Huson *et al.,* 2001) do not directly support data that are produced by 454 machines, 454-derived contigs (GS contigs) should be used as if they were individual reads or be shredded to generate many overlapping 'pseudoreads' (Goldberg *et al.,* 2006). Pseudoreads, made from GS contigs to emulate the read size of standard Sanger data (ca. 600 bp), are virtual reads whose stepping between consecutive

*Corresponding author: E-mail jfk@kribb.re.kr
Tel +82-42-860-4412, Fax +82-42-879-8595
Accepted 14 June 2008

shreds are controlled such that the underlying GS read depth can be represented. For example, an 890-bp contig that has an average depth of 21.7 can be converted into 32 600-bp pseudoreads. The stepping between consecutive pseudoreads can be given by (contig_length - pseuoread_length)/(num_pseudoreads - 1), or 9.35 bp. It not only reduces the effective data size but also minimizes misassemblies by co-incorporating Sanger data at the read level.

In this communication, we introduce a general strategy for genome assembly, either *de novo* sequencing or resequencing, that uses both 454 pyrosequence data and Sanger reads. The key consideration is that GS contigs are processed as if they were normal 'reads'.

De novo Sequencing

A general decision tree for *de novo* genome sequencing is shown in Fig. 1. 454 pyrosequencing at 20X sequence coverage is usually enough to produce a high-quality draft. For a conventional microbial genome project that employs paired-end Sanger sequencing on genomic libraries, end sequences from a fosmid library that has 10X clone coverage is sufficient for generating scaffolds. This also would be an appropriate choice when both 454 pyrosequencing and fosmid end sequencing with Sanger chemistry are utilized.

For completion of a genome sequencing project, CONSED with primer design, contig/read editing, scaffold viewing, and other plentiful features (http://www.phrap.org/) is the most preferred software. Though an ACE file that has a complete folder structure readable by CONSED can be produced by GS assembler programs, it is not fully compatible. Specifically, one read can appear in multiple contigs if it spans the boundary between two contigs, of which only one corresponds to a repeat region. After assembly, the names of many reads often are converted into non-standard ones to reflect the position of an aligned region as well, which may hamper the proper understanding of read information by CONSED, such as for the assembly view feature. We therefore highly recommend using PHRAP or other assembler software that generates 100% CONSED-compatible ACE files if users are going to finish the genome project.

If finishing is not scheduled (right side of the first decision step in Fig. 1), mixed assembly using gsAssembler would be the most convenient solution. Because there are upper limits for the number of Sanger reads that can be incorporated depending on system memory and the amount of data, GS contigs (with quality scores) can be converted to virtual reads and then assembled with Sanger reads using PHRAP. If there are reads that are larger than 65,536 bp, PHRAP.LONGREADS should be used. MKTRACE in the CONSED package is a convenient tool for producing fake traces and PHD files with nucleotide FASTA files and quality score files as input.

In case of scaffolding with just a small amount of Sanger reads, reads can be incorporated directly in CONSED. For this application, individual ACE files are produced from each GS contig after conversion to vir-

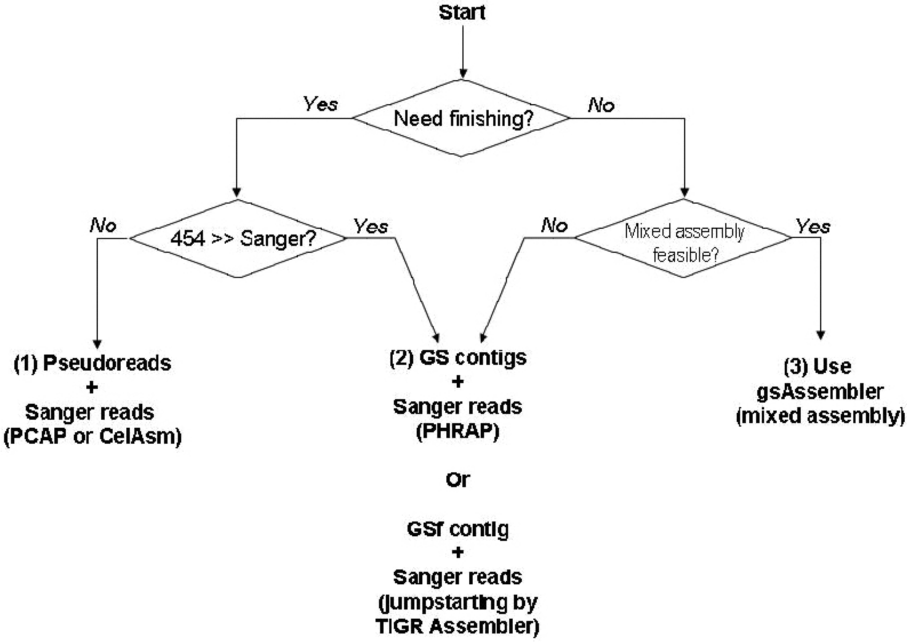
Start
Need finishing?
Yes
No
454 >> Sanger?
No
Yes
Mixed assembly feasible?
No
Yes
(1) Pseudoreads + Sanger reads (PCAP or CelAsm)
(2) GS contigs + Sanger reads (PHRAP)
Or
GSf contig + Sanger reads (Jumpstarting by TIGR Assembler)
(3) Use gsAssembler (mixed assembly)

Fig. 1. Decision tree for a genome project that employs both 454 pyrosequencing and Sanger sequencing. The first choice depends on whether finishing will be carried out or not, because CONSED-compatible ACE output from the assembler software is crucial for finishing. The amount of Sanger data then determines whether pseudoread generation and mixed assembly will be used or not (1 or 2). If finishing is not scheduled, mixed assembly using gsAssembler is the most convenient way, whereby a small amount of Sanger data is just enough for scaffolding purposes (3).

tual reads (via PHD files) and are merged into a single ACE file. Sanger reads can then be incorporated into the contigs by 'Add New Reads' in CONSED. Because this menu does not automatically extend or join contigs (it only compares newly added reads with existing contigs), users must validate the results of the additive assembly and process them manually.

Converting a GS contig into a single virtual read greatly reduces data size, which facilitates ACE file manipulation by CONSED. This simplification process also means loss of assembly information. Misjoining of two repeat sequences, frequently reported from 454 pyrosequencing-driven genome assembly, would be minimized if mixed assembly is carried out entirely by gsAssembler. A GS contig that is processed as a read in our suggested strategy is literally a minimal unit and cannot be torn apart and rejoined; editing only at the nucleotide level is allowed. Therefore, a trick is required to break or rejoin the virtual reads. From each GS contig, a simplified contig (GSf contig; 'f' stands for 'fragmentable') that consists of overlapping pseudoreads with an appropriate offset can be produced (Fig. 2). These pseudoreads will be used only in the context of assembly in TIGR Assembler's CONTIG file format. TIGR Assembler is the only software that makes jumpstart assembly possible, in which pre-existing contigs are retained and compared with new reads (Pop and Kosack, 2004). It is different from Add New Reads in CONSED in that incoming reads also are compared pairwise. After that, users can break or rejoin regions that are derived from GS contigs on the basis of pseudoread overlap.

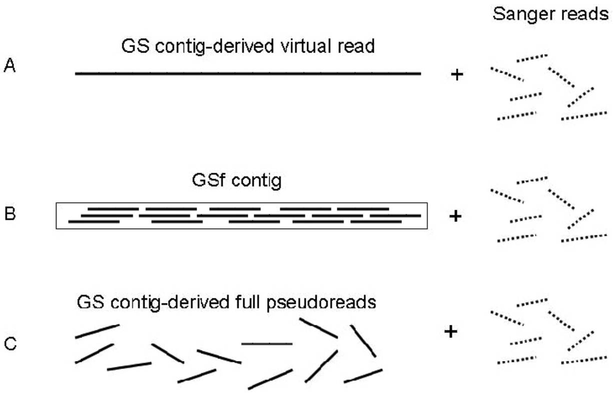
Sanger reads
A
GS contig-derived virtual read
+
B
GSf contig
+
C
GS contig-derived full pseudoreads
+

Fig. 2. Schematic drawing of GS contig processing and assembly. A, the simplest approach, in which each GS contig is converted into a virtual read that can be edited only at the nucleotide level. B, GS contigs are converted into simpler contigs that consist of several overlapping pseudoreads. By means of jumpstarting the assembly, they can maintain their integrity as contigs during the first hybrid assembly with Sanger reads. Contig-level editing, such as break and rejoin, can then be performed on the basis of mate pair information that is derived from the Sanger reads that are aligned on them. We propose to use a constant offset between pseudoreads for simplicity. C, pseudoreads can be compared with each other as well as incoming Sanger reads as originally suggested by Goldberg *et al.*

If the user is willing to finish the genome sequencing (left part of Fig. 1), the relative abundance of Sanger reads over 454 pyrosequence data should be considered. When the amount of available Sanger reads is low as compared with that of 454 data, virtual reads from GS contigs and Sanger reads can be assembled by PHRAP, as mentioned above. If there is more than 2～3X greater coverage of Sanger data available, often it is better to produce overlapping pseudoreads with varying offset between adjacent reads to reflect raw read depth along the assembly. In this case, running PCAP (Huang *et al.,* 2003) or CelAsm on a multiprocessor machine is a good choice for mixed assembly, because hundreds of thousands of pseudoreads are produced from GS assembly with ～20X sequencing coverage. We have written two simple Perl scripts that (1) break contigs at low read coverage (<2) and (2) produce overlapping pseudoreads that take into account read depth from the 454Contigs.ace file.

When pseudoreads are generated either with constant offset or with varying ones that reflect the raw read depth, assignment of an appropriate quality score can be an issue. Because GS contigs usually have a uniform quality score over the entire length, a constant value such as 10～20 should be adequate for most applications. Setting a lower quality value - for example, by dividing all 454 phred quality score values by 4 - might be assigned, as there are concerns about exaggerated quality scores (http://www.genome.ou.edu/454proto/454-LongreadsontheGS20v3.html).

Small contigs that are produced by gsAssembler, which often are generated from repetitive regions from the genome, sometimes contain useful information that is relevant to repeat-induced, over-collapsed misassembly. Because they occupy a small amount of the assembly result, they can be maintained as they are with GS raw read information, such that re-assembly and read-level manipulation are possible in CONSED. A hybrid ACE file that consists of virtual reads and real 454 assemblies can be made easily by simple text editing.

Whole-genome Resequencing

runMapping is the most convenient way to identify SNP-type genomic differences by aligning 454 pyrosequencing reads to the reference sequence. In addition to as-

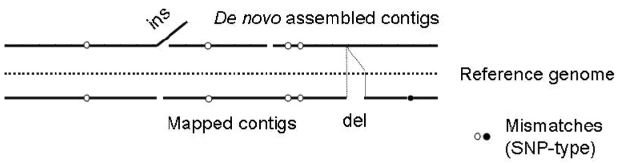

Fig. 3. Identifying differences between the sample and the reference by comparing the results of the 454 pyrosequencing data that are directly mapped to the reference sequence and the alignment of the *de novo* assembled contigs to the reference. The closed circle denotes a sequence mismatch that is represented only by one contig set.

sembly results, high-confidence differences also are written to a text file (454HCDiffs.txt) in which there are at least three reads, one read for each direction. This application usually produces larger contigs in smaller numbers than gsAssembler does, if the reference genome sequence is sufficiently close to the sample.

In most cases, however, variations occur, rather than base-to-base alterations (SNPs and small indels) such as large-scale insertions, deletions, inversions, and translocations. Identical copies of IS elements are a main cause of genome rearrangements. Genomic segments that are horizontally transferred from other bacteria or regions that are genetically manipulated on purpose cannot be identified by a standard runMapping procedure. For example, sample-specific deletion poses no 454 raw reads on the region that corresponds to the reference sequence. If *de novo*-assembled contigs are aligned to the reference sequence, however, contigs that span the deleted segment will easily be identified due to its partial alignment and discrepancy with the reference.

We therefore suggest using contig sets that are derived from *de novo* assembly (gsAssembler) and mapped assembly (gsMapper) to uncover the maximum number of possible differences (Fig. 3). The actual procedure is similar to the one that is shown in the center panel of Fig. 1. Two sets of GS-derived contigs are converted to virtual chromatograms and accompanying PHD files. In this case, the reference sequence also is converted to a single sequence read. It is then converted to a contig via a PHD file, and two contig sets that are derived from the GS data (as virtual reads) will be added in CONSED using Add New Reads. Because the number of reads that is to be added will be around 100 or more, we highly recommend inspecting the individual alignment result in CONSED. Each insertion/deletion/inversion candidate is then subject to confirmation by PCR amplification of the suspicious area and end sequencing of the product. It is better to confirm SNP-type variations that might appear in only one set of contigs, too (filled circle in Fig. 3).

In 454 pyrosequencing-based genome projects, we have successfully applied this hybrid approach either for *de novo* sequencing or for resequencing. They include *Escherichia coli* BL21(DE3) and its derivatives (resequencing), *Donghaeana dokdonensis* DSW-6 (*de novo* sequencing), *Hansenula polymorpha* DL-1 (*de novo* sequencing), and a few others. Resolving highly repetitive reads that are usually piled up as short, high-depth contigs is still a challenging task. Maintaining raw reads and their assembly structures that are generated from repetitive regions and combining them separately with other types of Sanger data (transposon-mediated sequencing or mini-scale shotgun for fosmid clones that harbor repeats) often is required for the completion of a 454-based genome project. Development of third-party software that is optimized for short-read fragment assembly also should accelerate advancements in the new era of genome technology (Sundquist *et al.,* 2007).

Acknowledgements

This work was supported by the 21st Century Frontier Microbial Genomics and Applications Center Program, Ministry of Education, Science and Technology, Korea.

References

Chaisson, M.J., and Pevzner, P.A. (2008). Short read fragment assembly of bacterial genomes. *Genome Res.* 18, 324-330.

Goldberg, S.M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., and Friedman, R., *et al.* (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. USA* 103, 11240-11245.

Huang, X., Wang, J., Aluru, S., Yang, S.P., and Hillier, L. (2003). PCAP: a whole-genome assembly program. *Genome Res.* 13, 2164-2170.

Huson, D.H., Reinert, K., Kravitz, S.A., Remington, K.A., Delcher, A.L., Dew, I.M., *et al.* (2001). Design of a compartmentalized shotgun assembler for the human genome. *Bioinformatics* 17 Suppl 1, S132-139.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376-380.

Pop, M., and Kosack, D. (2004). Using the TIGR assembler in shotgun sequencing projects. *Methods Mol. Biol.* 255, 279-294.

Shendure, J., Mitra, R.D., Varma, C., and Church, G.M. (2004). Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* 5, 335-344.

Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P., and Batzoglou, S. (2007). Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE* 2, e484.