ABSTRACT

Massively parallel next-generation DNA sequencing (NGS) instruments can now produce 10^9 bases per run. However, the utility of the data from this new technology for de novo genome assembly is compromised by three technical limitations; the length of reads is far too short, the size of the sequenced segments is also short, and most of the methods utilize random sequencing and assembly strategies. Current technologies cannot resolve repeats or segmentally duplicated sequences, which results in large missing segments of eukaryotic genomes and failure to close small prokaryotic genomes. These limitations make it very difficult and expensive to assemble large or small genomes de novo. We have developed a new high efficiency long-span mate-pair library construction technology and software for de novo sequence assembly. Our strategy involves the use of multiple methods to prevent chimera formation, including the incorporation of Chimera Codes™ that allow one to distinguish true mate pairs from false junctions for de novo sequence assembly. Ion Torrent 400 bp sequence data from B Ac and 284 kb E. coli mate-pair libraries confirmed our expected results of >90% assembled reads (114,869 of 126,930) showing consistent alignment across the entire genome. Long-span, mate-pair sequences with long connectivity and high physical coverage of a genome can be used for de novo assembly and to detect genomic variation more efficiently. Here we present early data demonstrating feasibility and efficiency of these methods.

METHODS AND RESULTS

**Why Is Mate Pair Data Required?**

- Repetitive Genomes
- De novo Assembly
- Structural Variant Detection
  - Indels/rearrangements are subtle
- Gap Closure & Genome Finishing
  - Genomic context scrambled

**Complex Genomes Contain Millions of Repeats**

- Assembler Can Not Distinguish the Repeats
- Mate Pair Information Spans Repeats

**Figure 1. Importance of mate pair information for genome assembly.** All genomic sequences contain repetitive elements, indels and rearrangements that make assembly so difficult with standard Next Generation Sequencing strategies. Mate Pair sequence data allows software to assemble short reads into scaffolds with the Correct contig position and orientation.

**Figure 2. Schematic of Lucigen’s strategy for mate pair library construction.** Genomic DNA is sheared to the desired size, end-repaired, A-tailed and ligated to barcode adaptors prior to size selection. The insert is ligated to a unique coupler and samples are then treated with exonuclease to remove unwanted DNA and then digested with a selection of endonucleases to produce the correct sized Di-tags. Biotin capture allows for the removal of unwanted DNA fragments prior to the addition of a Junction Code adaptor and re-circularization.

**Figure 3. Detection of chimeras with Chimera Codes.** Engineered into the Lucigen NxMate couplers are Chimera Codes that allow software to distinguish between true mate pairs and chimeras.

**Figure 4. Example of chimera detection with Chimera Codes.** A) Ion Torrent sequencing shows the correct Chimera Codes on genomic di-tags 1 and 2. B) Example of incorrect Chimera Codes on the two di-tags.

**Figure 5. Insert Size Selection and resulting NxMate data.** Sheared DNA that was adapted and size selected (gel images), was used to create Mate Pair libraries. Analysis of sequence data shows that the mate pair insert sizes are similar to the original sheared DNA.

**Figure 6. Detection of chimeric sequences with Chimera Codes.** Library sequenced data was sorted by Chimera Codes and the correct code pairs were plotted against each other (diagonal green). Incorrect pairs and their percentage of total reads are shown in black. Prior to removal of chimeras from data, 84% of mate pairs are correct.

**Figure 7. Analysis of mate pair libraries.** Mate pair libraries were analyzed to show the percentage of good mate pairs in the libraries (All Mate Pairs) and the percentage of mate pairs after chimeras and indeterminate data have been removed (True Mate Pairs).

**Figure 8. DNAStar Software analysis of NxMate Data.** Depth of coverage and Inconsistent mates.

**Figure 9. Sequence assembly for two repeat rich mouse BACs.** The sequences for two mouse BAC clones were assembled with DNAStar software using Ion Torrent 400 bp fragmental and 5 Kb Mate pair sequence data. Despite having over 50% repeat sequence, the two BACs were each assembled into single scaffolds of 171 Kb and 143 Kb.

**CONCLUSIONS**

NxMate NGS Mate Pair Technology enables more accurate assembly of BACs and genomes. A new paradigm for constructing high efficiency mate pair libraries has been developed. Chimera Code™ helps prevent false junctions. Junction Code™ identifies mate pair junctions.

**ACKNOWLEDGEMENTS**

This work was partly supported by NHGRI grants to ML and DM.