Next generation DNA sequencing (NGS) instruments produce gigabases per run, but the short read lengths and small size of sequenced fragments result in gaps, misassembled contigs, collapsed repeats and missing sequences, leaving these regions to be finished manually, if at all. A technology that provides long range sequence linkage from short reads is needed for accurate, economical de novo assembly of genomes. We developed a >90% efficiency clone-free mate pair library construction technology that incorporates Chimera Code™ to distinguish true mate pairs from false junctions where other technologies lack the ability, and a 40 kb Fragmented Mate Pair (pNGS™-FOS) for constructing long mate pair libraries propagated in E. coli. NixSeq® Long Mate Pair NGS libraries were constructed using a reference E. coli strain, Thermus aquaticus, and pNGS™-FOS libraries were constructed for 2.5 Gb Miscanthus sinensis and a 1 Gb fish genome. Without mate pair libraries the genome assemblies contained numerous unordered contigts. The addition of >90% efficient NixSeq® data allowed accurate de novo assembly and closing of the microbial small genomes and a remarkable reduction in the number of scaffold contigs for Miscanthus sinensis and the fish genome.

**METHODS AND RESULTS**

**Why Is Mate Pair Data Required?**

- **Repetitive Genomes**
- **Every contig begins and ends with a repeat**
- **De novo Assembly**
- **Even small genomes and BACs incompletely sequenced**
- **Structural Variant Detection**
- **Indels/rearrangements are subtle**
- **Gap Closure & Genome Finishing**

**Genomic context scrambled**

**Importance of mate pair information for genome assembly.**

All genomes contain complex genetic elements that make assembly almost impossible with standard NGS strategies. Mate Pair sequence data allows software to assemble short reads into scaffolds with the correct contig position and orientation.

**Mate Pair Distance Histograms**

*Figure 4. Long mate pair libraries from two samples.*

An 8 kb NixSeq® mate pair library was constructed using gel-free methods, and a 10-20 kb mate pair library using gel isolation. Resulting true mate pairs were mapped against the respective reference genome to determine the resulting mate pair distances.

**Single Scaffold Assembly of E. coli K12**

*Figure 5. De novo assembly of e. coli k12 genome.*

2.5M fragment reads were assembled de novo into 163 contigs over 1 kb by SPAdes 3.1. Scaffolding was performed with commercial software using 3.2M 8 kb mate pair maps. The single scaffold was compared to a reference genome with Mauve 2.3.1.

**De novo Assembly and Closing the Thermus aquaticus Genome with NixSeq® Library Data**

*Figure 6. De novo assembly of thermus aquaticus genome.*

T. aquaticus random sheared DNA was used to generate a fragment library (Mlleib), and 5 kb and 8 kb NixSeq® mate pair libraries. 1M fragment reads were assembled using SPAdes 3.1 and the resulting contigs were scaffolded sequentially using 1M 5 kb and 1M 8 kb NixSeq® true mate reads.

**Effect of pNGS™ FOS mate pair data on assembly of large genomes**

**Table 2. Comparison of pNGS™ FOS mate pair data on assembly of large genomes**

**Effect of pNGS™ FOS mate pair data on assembly of large genomes**

*Figure 7. pNGS™ FOS libraries from large plant and animal genomes.*

pNGS™ FOS library mate pair data improves assemblies of large genomes by spanning gaps left by the assembly of fragment libraries only (A), thereby markedly increasing scaffold sizes (B & C). Lambda packaging ensures a predictable span distribution that lends greater confidence to the arrangement and orientation of contigs.

**Comparison of Long Read Technologies**

*Figure 8. Comparison of size distributions from commercially available NGS sequencing technologies.*

Mate pair size distributions were plotted for various NixSeq® and pNGS™ FOS libraries, as compared to Pac Bio long reads (blue) and Illumina TrueSeq synthetic long reads (green). The NixSeq® and pNGS™ FOS Mate Pair Technology enables user-definable mate pair libraries.

**CONCLUSIONS**

A new paradigm for constructing >90% efficient mate pair libraries has been developed. Chimera Code™ helps prevent false junctions, while incorporation of a Junction Code™ identifies mate pair junctions. NixSeq® and pNGS™-FOS Mate Pair Technology enables accurate, economical assembly of BACs and genomes and is compatible with both Ion Torrent and Illumina platforms.

**ACKNOWLEDGEMENTS**

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

**REFERENCES**

Lucigen is registered to ISO 13485 quality management standard for medical devices through BS9, Inc. This material only for non-human diagnostic use.