High Efficiency Long Insert Mate Pair Library Preparation for NGS Platforms

Michael J. Lodes, Svetlana Jasinovica, Erin Ferguson, Amanda Krerowicz, Brendan Keough, Megan Niebauer, David Mead, and Scott Monsma
Lucigen Corporation, 2905 Parmenter Street, Middleton, WI 53562, USA

Lucigen®
Simplifying Genomics
www.lucigen.com

ABSTRACT

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 95% efficiency clone-free mate pair library construction technology that incorporates Chimera Codes™ to distinguish true mate pairs from false mates where other technologies lack this ability, and a 40 kb Fosmid vector (pNGS™-FOS) for constructing long mate pair libraries propagated in *E. coli*. NxSeq® 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 contigs. The addition of 95% efficient NxSeq® mate pair data allowed accurate *de novo* assembly and closing of the microbial small genomes and a remarkable reduction in the number of scaffolds 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 incomplete
- 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.

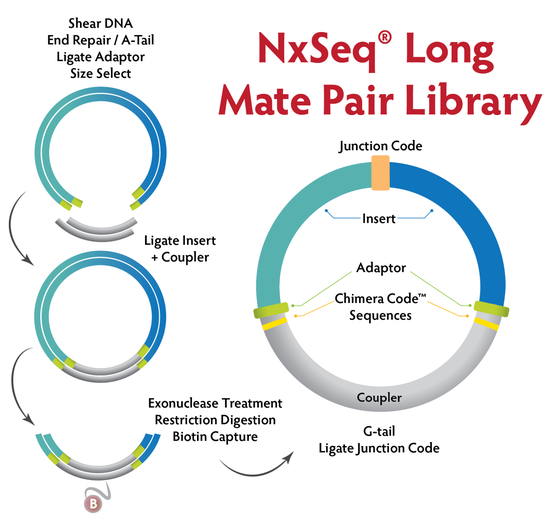


Figure 1. Schematic for user defined 2-20 kb mate pair library construction. Genomic DNA is sheared, end repaired, A-tailed and ligated to barcode adaptors prior to size selection. The insert is ligated to a unique coupler with encrypted Chimera Codes, samples are then treated with exonuclease to remove unwanted DNA, and finally 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. NxSeq® libraries can be made compatible with either IonTorrent or Illumina sequencing platforms.

40 kb NGS Mate Pair Fosmid

US Patent 8329400

Figure 2. Schematic of pNGS FOS library construction workflow. Sheared and end-repaired gDNA is ligated to the pNGS™ FOS vector, packaged into lambda phage, and transfected into *E. coli*. Plasmid DNA is miniprepped, and after restriction of the insert DNA, the construct is rejoined and sequenced using Illumina platforms.

Efficiency of Mate Pair Libraries

	E. coli DH10B 2 kb	*E. coli* DH10B 5 kb	*E. coli* DH10B 8 kb
Raw Reads	6,377,792	5,995,974	6,851,682
Total Mates	2,167,286	2,242,930	3,091,359
True Mate Pairs	2,071,267 (96%)	2,094,413 (93%)	2,938,426 (95%)
Chimera	96,019 (4%)	148,517 (7%)	152,933 (5%)
Avg. Read Length (after split)	170 b	161 b	159 b
Total Mate Pair Bases	352,115,390	337,200,493	467,209,734
Mapped Mate Pair Distance	2,543	5,145	6,191

Figure 3. Analysis of mate pair libraries. Mate pair libraries were analyzed to show the percentage of good mate pairs in the libraries (Total Mates) and the percentage of mate pairs after chimeras and indeterminate data were removed (True Mate Pairs). NxSeq® libraries produce ~95% true mate pairs.

Mate Pair Distance Histograms

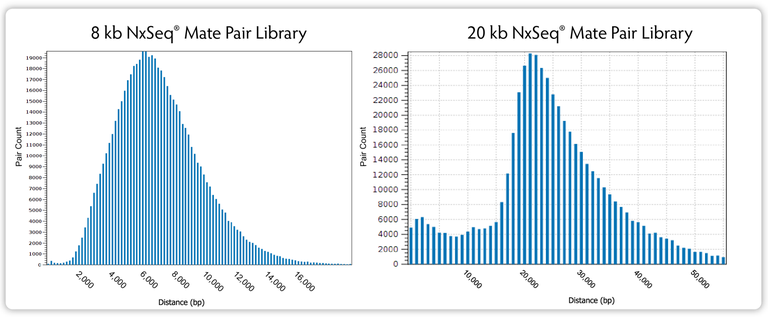


Figure 4. Long mate pair libraries from two samples.
An 8 kb NxSeq® mate pair library was constructed using gel-free methods, and a 20 kb mate pair library using gel isolation. Resulting true mate pairs were mapped against the respective reference genome to determine the resulting mate pair distances.

Single Scaffold Assembly of *E. coli* K12

Fragment Library + 8 kb Mate Pair Library

Fragment library
5,001,861 reads – 4,656,638 mappable
Assembled 2.5 M reads in **SPAdes** with K45

Contigs > 1kb	163
Contigs > 500 bp	184
N50	56,903
Max contig	179,213 bp

8 kb NxSeq library
Megaruptor sheared

Raw reads	6,909,356
True mate pairs	3,288,275 (48% of raw)
Mate pair distance	8,310 bp

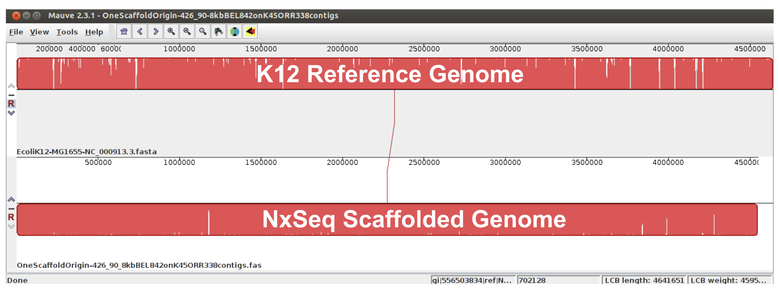


Figure 5. De novo assembly of *E.* coli K12 genome.
2.5M fragment reads were assembled de novo into 163 contigs over 1 kb by SPAdes 3.1. Scaffolding was performed with commercial software using 3.2M 8 kb mate pairs. The single scaffold was compared to a reference genome with Mauve 2.3.1.

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

	JGI Permanent Draft Genome	NxSeq® Mate Pair Closed and Finished Genome and Plasmids
# Contigs > 500 bp	22	NA
Contig N50	180,857	NA
Max contig	343,213	NA
Genome scaffolds	0	1
Max scaffold	343,213	2,158,963
Genome + plasmid size	2,338,193	2,338,240
Plasmid scaffolds	?	4
Plasmid sizes	?	14,047 16,597 78,727 69,906

Figure 6. De novo assembly of *Thermus aquaticus* genome.
The *Thermus aquaticus* genome contains high GC content (68%), making it difficult to assemble with next generation sequencing technologies. The genome was left unfinished in 2008 (NCBI accession number NZ_ABVK00000000.2). The use of two separate 5kb and 8kb NxSeq mate pair libraries enabled the first ever closed and finished genome for this organism, as well as the delineation of four separate plasmids contained within the bacteria.

Effect of pNGS™ fosmid mate pair data on assembly of large genomes

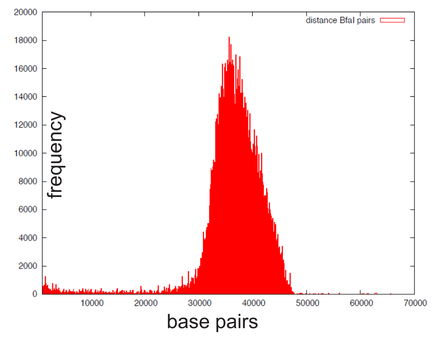
A **2.5 Gb *Miscanthus sinensis* Genome**
pNGS™-FOS Mate Pair Libraries

pNGS Library	Spans	Contigs Spanned
BfaI	252,236	96,205
CviQI	294,410	99,285
EcoRI	30,792	33,073
BamHI	28,510	25,234
HindIII	80,924	47,337

668,886 confirmed spans from all 5 libraries
301,134 gaps spanned

B
N50 = 40 kb before pNGS™ FOS mate pair library
N50 = 102 kb after pNGS™ FOS mate pair library

Span sizes for pNGS BfaI mate pair library



C **1 Gb Fish Genome**

N50 = 80 kb before pNGS™ FOS mate pair library
N50 = 477 kb after pNGS™ FOS mate pair library

Span sizes for pNGS BfaI mate pair library

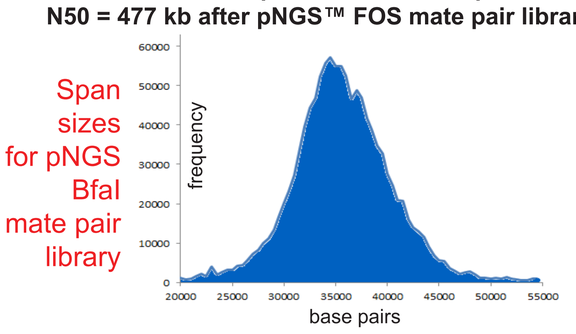


Figure 7. pNGS™ FOS libraries from large plant and animal genomes.
pNGS™ FOS library mate pair data improves assemblies of large genomes by spanning gaps left by the assembly of fragment libraries only (A), thereby markedly increasing scaffold sizes (B & C). Lambda packaging ensures a predictable span distribution that lends greater confidence to the arrangement and orientation of contigs.

Comparison Of Long Read Technologies

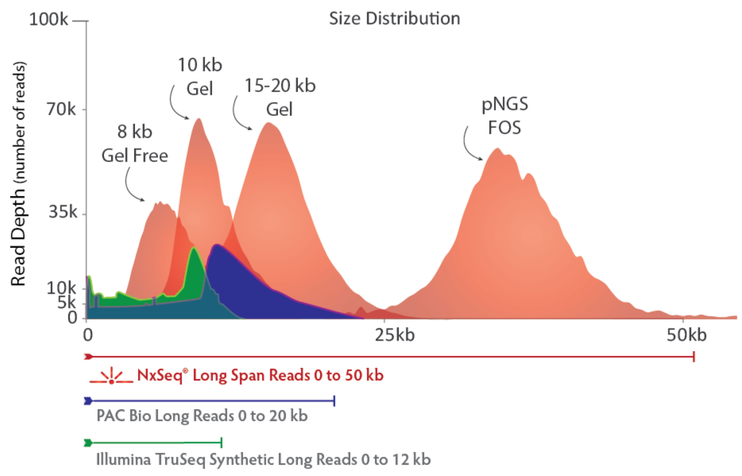


Figure 8. Comparison of size distributions from commercially available NGS sequencing technologies.
Mate pair size distributions are plotted for various NxSeq and pNGS FOS libraries, as compared to Pac Bio long reads (blue) and Illumina TruSeq synthetic long reads (green). The NxSeq and pNGS FOS Mate Pair Technology enables user-definable mate pair libraries.

CONCLUSIONS

A new paradigm for constructing 95% efficient mate pair libraries has been developed. Chimera Code™ helps prevent false mates, while incorporation of a Junction Code™ identifies mate pair junctions. NxSeq® 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.