

ABSTRACT

Next generation sequencing (NGS) technologies can rapidly and economically produce a draft genome of an organism de novo. However, the quality of the draft data is seldom more than 80% complete with >10e5 contigs for large genomes, which is insufficient for many applications. Most contigs begin and end with a repeat with existing library construction technologies. Sequence data that is closer to 95% finished with the unambiguous order and placement of genes would have the greatest utility for scientific and commercial research. New molecular tools that bridge the gaps between massively parallel short read sequencing technologies (35-1,500 bases) and the need for large scaffolds (>100,000 bases) to accurately assemble complex repeat rich genomes are needed. We have successfully developed 40 kb mate-pair NGS libraries by designing and constructing a novel pNGS fosmid system. Our results show that ~70% pNGS fosmid paired-end sequences can be obtained by either Illumina or 454 sequencing, which is significantly better than existing long-span mate-pair systems. We have also developed a clone free long-span mate-pair NGS library construction technology for 10-300kb inserts.

NGS Transcription-free BAC/FOS Vector

We have developed a new BAC/fosmid cloning system that includes a transcription-free pNGS BAC/Fosmid vector, a simple system for copy induction of BAC/fosmid DNA, and primer binding sites for next-generation sequencing on Illumina or 454 platforms immediately flanking the insert cloning site (Fig. 1 & 2). The lack of active transcription or translation due to the absence of a *lacZ* promoter and gene, plus the presence of transcriptional stop signals, results in higher stability of cloned inserts compared to conventional vectors.

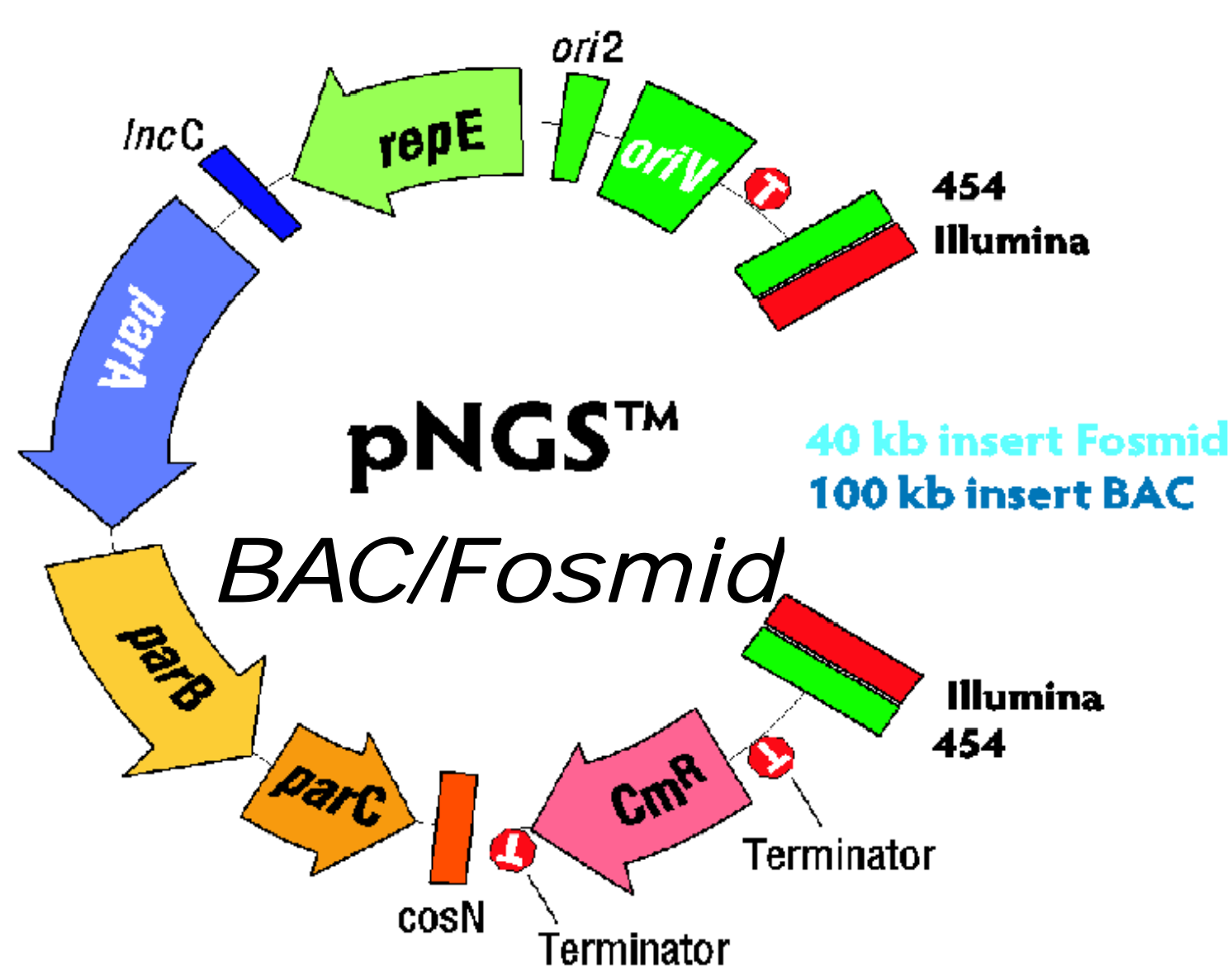


Figure 1. pNGS™ BAC/fosmid vector. *ori2*, *repE*, *IncC* - origin of replication (single copy); *oriV* - inducible origin of replication; *parA,B,C*- partition genes; *CmR*- chloramphenicol resistance gene; *cosN* - lambda packaging signal; T - transcription terminators. Position of Illumina and 454 NGS primers for amplification and sequencing are indicated.

Fosmid Di-Tag Sequencing

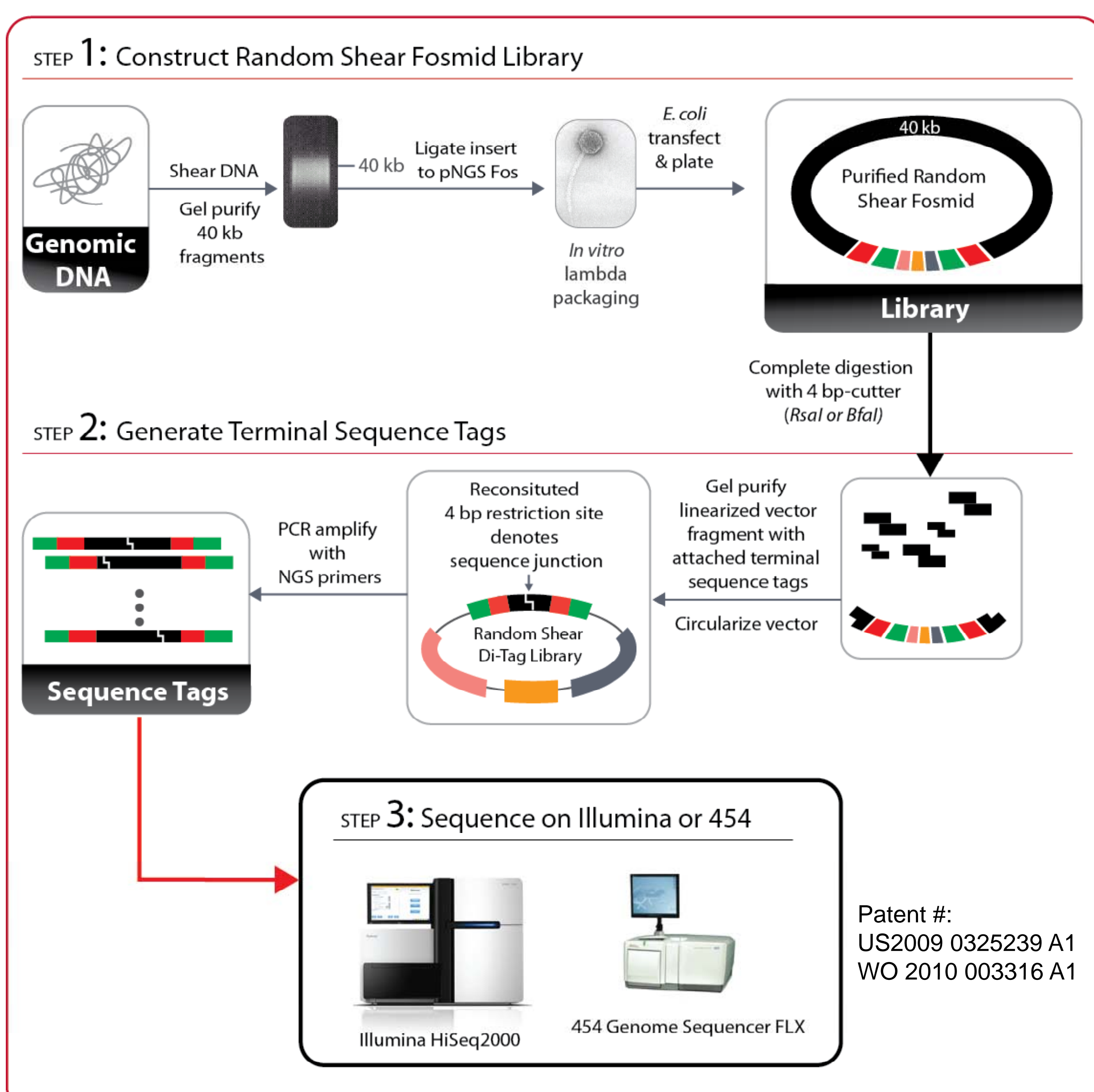


Figure 2. The scheme for generating 40kb fosmid paired end reads by Illumina and 454 sequencing platforms.

NGS Fosmid Library Construction

Genomic DNA was prepared from *Mus musculus* strain C57Bl6J, randomly sheared, size-selected, and ligated into the pNGS fosmid vector. DNA was packaged using lambda extracts and transfected into *E. coli* cells. The cells were plated at high density on large trays and two pools of DNA were prepared, one consisting of about 8,000 and the other 35,000 clones. The DNA was purified as a mixed pool and completely digested with the frequent endonuclease restriction enzymes BfaI or CviQI (Fig. 3 left panel). The fosmid vector with paired-ends were gel-purified (Fig. 3 right panel) and recirculized by ligation. The recirculized fosmid pair-ends were used as templates for PCR amplification with either 454 or Illumina primers (Fig. 4).

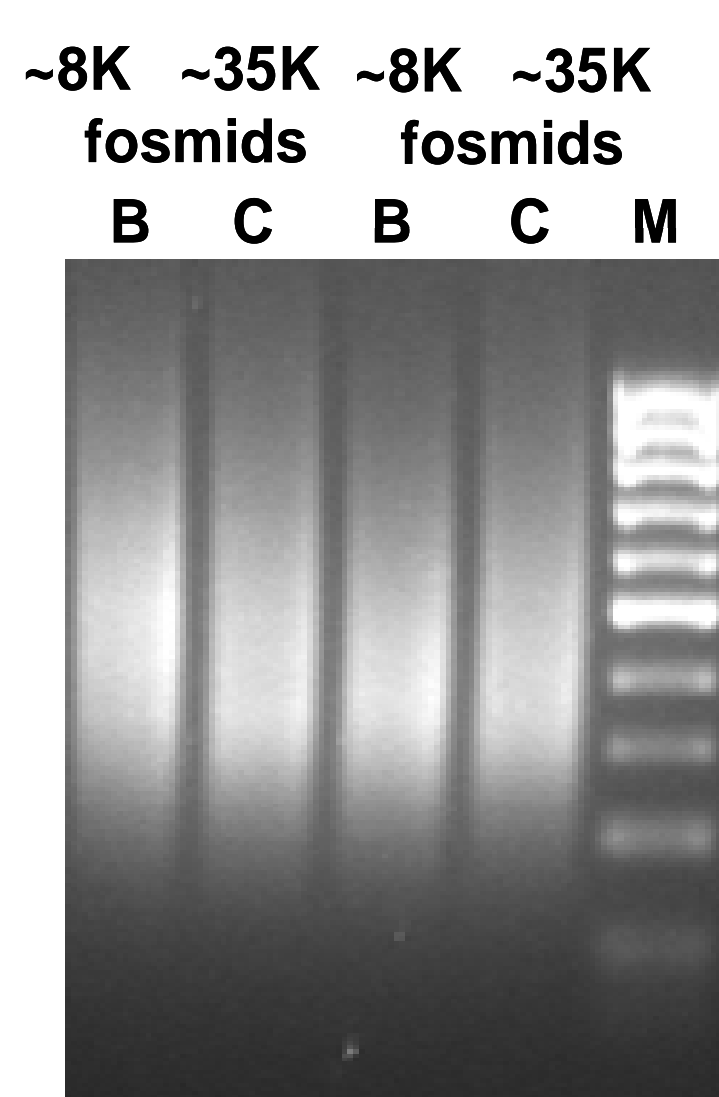


Figure 4. PCR amplification of re-circularized NGS fosmid paired-ends + vector backbone with the primer set for Illumina sequencing. Similar results were seen with 454 primers (data not shown). The size ranges from 200-1,000 bp.

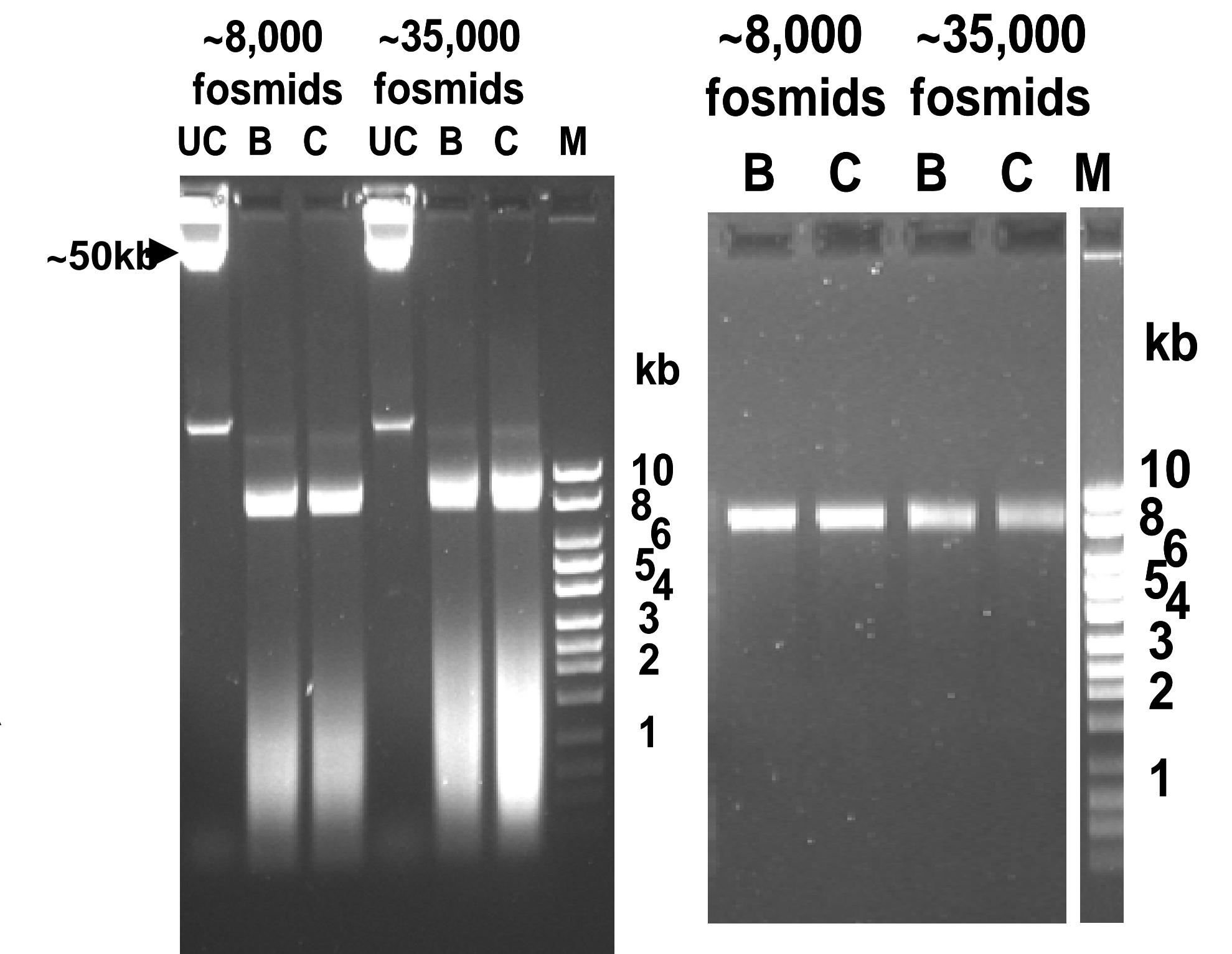


Figure 3. Pooled mouse fosmid DNA was completely digested by two restriction enzymes (Left panel) and the paired-ends with pNGS fosmid backbone purified by gel extraction (Right panel). Lanes: UC, uncut fosmids; B, BfaI/FspBI; C, CviQI/Csp6I; M, 1kb ladder Marker. Large NGS fosmid inserts (~40kb) can be completely removed by restriction digestion with a frequent cutter and the NGS fosmid paired-ends are purified, self-ligated, and amplified for direct NGS PCR.

High Efficiency 40 kb Paired-end Sequencing for Next Generation Platforms

The PCR amplicons containing the 40 kb fosmid paired-ends were cloned and sequenced using conventional Sanger technology in order to determine the efficiency of this system on a small scale. 96 random clones were picked and sequenced from both ends to ensure sequence coverage for larger paired ends. The results are shown in Fig. 5, Tables 1 and 2 and demonstrate a 75% success rate for obtaining ~40kb mate-pair sequences. 10.4% of the amplicons only had one-end sequence, sequencing failures accounted for 5.2% of the total, and 9.4% of the reads were located in repeats that were too complex to precisely determine their chromosomal locations. Some of the sequences could also include potential genomic variations (Table 1 and 2). Large-scale long-span, mate-pair sequencing of a human cell line (GM15510, Coriell) using Illumina technology resulted in 64% of filtered reads accurately mapping to the genome (Table 3). This represents many-fold higher efficiency than existing systems and will allow the accurate assembly of genomes for the first time using next gen sequencing platforms.



Figure 5. An example of ~40kb NGS-fosmid paired-end sequence. The lengths of forward and reverse end sequences are 134 and 58 bp respectively, and the joint sequence of this paired-end is highlighted in blue (CviQI cutting site). The paired-end sequence is from the genomic position between 95110848 and 95069222, which is 41.626 kb apart on chromosome 10 from the complete genome sequence of *Mus musculus* (C57BL/6J).

Table 1 Statistics of successful long-span, mate-pair percentage

	No. of tags	%
Mate-pairs	72	75.0
One-end sequences	10	10.4
Repeats*	9	9.4
Sequencing fail	5	5.2

*potential genomic variation

Table 2. Statistics of sequence length and percentage

Sequence length (bp)	No of sequences*	%
4~11	8	4.2
12~50	21	10.9
51~100	23	12.0
101~200	37	19.3
201~300	25	13.0
301~400	17	8.9
401~500	20	10.4
501~700	12	6.3
701~874	20	10.4
Sequencing fail	9	4.7

* Sequence length <11bp and failed sequences total 8.9%

Table 3 Summary of Illumina long-span, mate-pair sequencing of a human cell line

Stage	No. of reads	%
Pipeline	8,683,854	
Filtered reads	6,506,126	
Reads mapped	5,173,778	79.2
Reads with pair-sequences	4,685,996	90.6
Long-span, mate pairs*	3,018,576	64.4

*Mapped on same chromosome, ~40kb apart, uniquely mapped

Conclusions

- New tools to construct fosmid (~40kb) random shear libraries for next-gen sequencing have been developed. Using sanger sequencing and human genome data we demonstrated that Lucigen's pNGS-fosmid system produces the highest efficiency long-span, mate-pair libraries available.
- 40 kb NGS paired-end libraries can be used to probe fine-scale variation in genomic DNA (i.e. insertions/deletions/inversions/translocations) and to scaffold random short read sequences, thereby improving *de novo* assembly of genomes.
- Other versions of the vector can be used to produce large paired-ends of methylation sensitive sites, or pairs of adjacent restriction enzyme site to maximize the use of the technology.

Acknowledgements

This work was supported by an NHGRI grant to CCW.

