

Swati Ranade, Heather Peckham, Joel Malek, Chris Clouser, Jason Warner, Jeffery Ichikawa, Clarence Lee, Brittney Coleman, Michael Laptewicz, Alena Antipova, Alan Blanchard, Gina Costa and Kevin McKernan
Advanced Genetic Analysis, Applied Biosystems, 500 Cummings Center-Suite 2400, Beverly MA 01915, USA

Summary

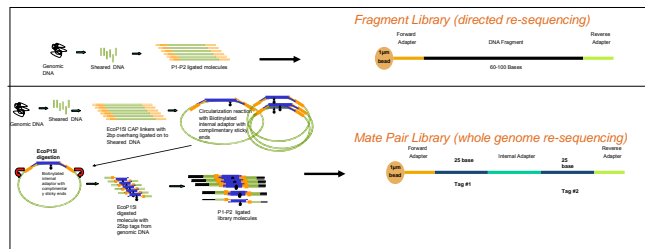
High-throughput sequencing technologies have greatly increased the power of human genome variability studies. We have prepared multiple paired end libraries with insert sizes ranging from 500bp to 10kb and fragment libraries with an average insert size of 60-100bp using the NA18507 DNA belonging to a Yoruba individual. These libraries are being extensively sequenced using SOLiD (ABI's next generation sequencing platform). The deep sequence coverage obtained from these diverse libraries, not only helps identify SNPs in this genome but also submicroscopic structural variations like insertions/deletions. In this high density genome re-sequencing effort we address the challenges of constructing multiple sized mate pair libraries of complex genomes.

SOLiD™ for Structural Variation Studies

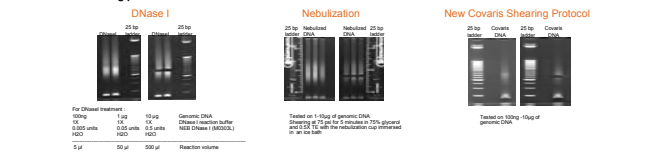
The SOLiD™ sequencing system uses stepwise cyclic ligation and has been developed for high throughput DNA sequencing. The fragment and mate-paired library construction methods employed afford genome sequencing of short fragment (1 x <50 bases) and mate-paired (2 x 25 bases) DNA libraries. Recent improvements have demonstrated performance of >6 Gb per single tag (fragment library) and >8 Gb per dual tag (mate-pair library) for a single instrument run. This high throughput sequencing capability is only expected to improve making SOLiD a very amicable technology for Structural Genomic studies. Such studies however, warrant a tedious sample preparation process especially for paired end libraries which are extremely important for improved mapping and structural analysis of genomes.

We have developed HPLC mediated library construction methods for high throughput mate pair library construction that facilitate automated sizing of sheared DNA molecules with a range of different sizes. Similarly new and improved fragment library protocols have been developed to facilitate library construction from as low as 100ng of DNA. Whole genome sequencing applications for structural variation studies require very high physical and sequencing coverage the present study is intended to be a pilot for understanding the advantages of sequencing different insert sized libraries and establishing a sample preparation protocol for such studies.

SOLiD™ Library Construction



Fragment Library Construction
Flexible shearing protocols



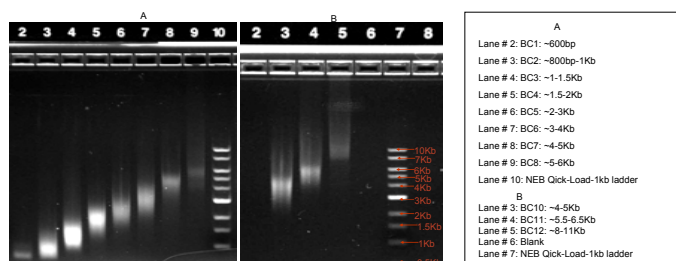
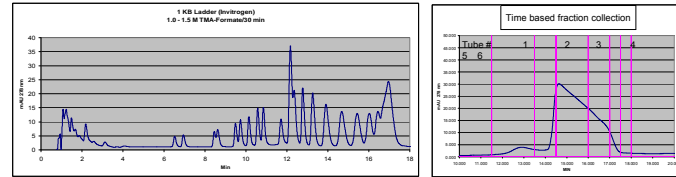
Fragment Library Performance

Condition	Sheared DNA size	µg of second size selected DNA	ng of circles recovered post plasmid safe, precipitation and column purification	% circularization efficiency	% of PCR cycles for library amplification	Unique normal mates	Mean Library Insert size / SD	Physical coverage	Sequence Coverage
600bp_BC1	1µg	228ng	23%	13cycles (over amplified)	>53 Million (Library not saturated)	600 / 58	10.75 X	0.9X	
800bp-1kb_BC2	12.3µg	567ng	5%	13cycles (over amplified)	>63 Million (Library not saturated)	804/ 184	16.95X	1.05X	
1-1.5kb_BC3	14.16µg	603ng	4.3%	13cycles (over amplified)	>83 Million (Library not saturated)	1179/ 221	32.63X	1.38X	
1.5-2kb_BC4	10.7µg	471ng	4.5%	13cycles (over amplified)	>44 Million (Library not saturated)	171/ 316	25.1X	0.73X	
2-3kb_BC5	10.8 µg	525ng	5%	15cycles	>25 Million (Library not saturated)	284/1611	24.47X	0.43X	
3-4kb_BC6	13.7µg	483ng	3.5%	15cycles					
4-5kb-BC7	3.33µg	201ng	6%	19cycles					
4-5kb-BC10	7.5µg	390ng	5.2%	17 cycles					
5-6kb_BC8	2.27 µg	204ng	10%	20cycles (Fair amplification visible in 19cycles)					
6-7_BC11	4µg	300ng	7.5%	22cycles (Fair amplification visible in 19cycles)					
10-12_BC12	2µg	210ng	10.5%	22cycles					

Mate Pair Library Construction

CORIELL repository, Yoruban 18507 DNA used in the International HAPMAP project (Catalog ID NA18507)
HPLC Conditions

Column: TOSOH TSKGEL DNA-NPR 2.5u 4.6X75 Column, no guard column, no sample filter.
Column not thermostated, Room temp = 22-24 C
Solvents: (TMA = Tetramethylammonium); "A": 1.0 M TMA-Formate, 20 mM TRIS base, HCl to pH 9.0 / "B": 1.5 M TMA-Formate, 20 mM TRIS base, HCl to pH 9.0
Run conditions: 0% to 100% "B" linear over 30 min., 0.5mL/min.



Mate Pair Library Performance Matrix

Sheared DNA size	µg of second size selected DNA	ng of circles recovered post plasmid safe, precipitation and column purification	% circularization efficiency	% of PCR cycles for library amplification	Unique normal mates	Mean Library Insert size / SD	Physical coverage	Sequence Coverage
600bp_BC1	1µg	228ng	23%	13cycles (over amplified)	>53 Million (Library not saturated)	600 / 58	10.75 X	0.9X
800bp-1kb_BC2	12.3µg	567ng	5%	13cycles (over amplified)	>63 Million (Library not saturated)	804/ 184	16.95X	1.05X
1-1.5kb_BC3	14.16µg	603ng	4.3%	13cycles (over amplified)	>83 Million (Library not saturated)	1179/ 221	32.63X	1.38X
1.5-2kb_BC4	10.7µg	471ng	4.5%	13cycles (over amplified)	>44 Million (Library not saturated)	171/ 316	25.1X	0.73X
2-3kb_BC5	10.8 µg	525ng	5%	15cycles	>25 Million (Library not saturated)	284/1611	24.47X	0.43X
3-4kb_BC6	13.7µg	483ng	3.5%	15cycles				
4-5kb-BC7	3.33µg	201ng	6%	19cycles				
4-5kb-BC10	7.5µg	390ng	5.2%	17 cycles				
5-6kb_BC8	2.27 µg	204ng	10%	20cycles (Fair amplification visible in 19cycles)				
6-7_BC11	4µg	300ng	7.5%	22cycles (Fair amplification visible in 19cycles)				
10-12_BC12	2µg	210ng	10.5%	22cycles				

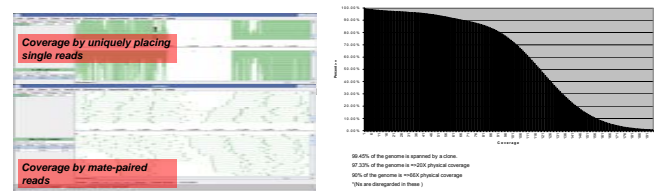
Libraries in sequencing pipeline

The % circularization efficiency value is also affected by the multiple precipitation and column purification steps especially due to the large volumes of circularization reactions. As is seen in the above table whenever the amounts of DNA was larger, the reaction volumes often went up to 5 to 8ml and the recovery of circles was affected.

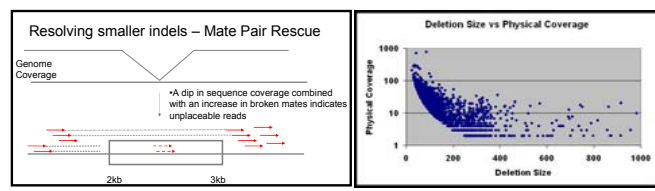
SOLiD Sequence Data

SOLiD™ sequencing of both fragment (60-100 bp) and mate pair (600bp-10K insert size) libraries though not yet complete, has already yielded us at least 100 X physical and about 4.49X sequence coverage of the Yoruba- NA18507 DNA

Multiple Mate Pair Libraries are advantageous as the varying insert lengths can span the unmappable repetitive regions



High physical coverage allows detection of small insertions and deletions based on deviation of average insert sizes spanning a genomic region



Structural variations detected

Category	Count
Deletions	45,252
Homozygous	44,310
Heterozygous	942
Insertions	22,387
Homozygous	22,358
Heterozygous	29
Double Deletions	93

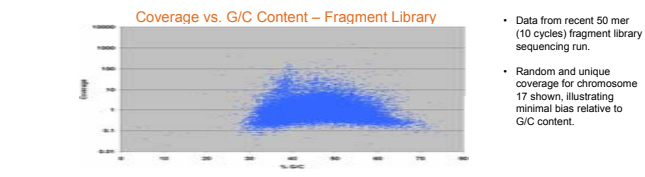
SNP detection

Chr 7: At 4X Sequence Coverage: 75% in dbSNP
10 ENCODE Regions: 91% Found in dbSNP

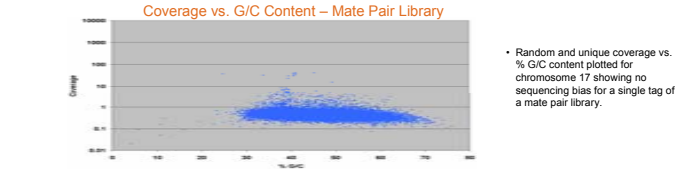
Conclusion:

We have prepared multiple mate paired libraries with a range of insert sizes for SOLiD mediated comprehensive re-sequencing of the Yoruba (NA18507) genome. SOLiD sequencing produced short sequencing reads with almost 100X physical and about 4.9X sequence coverage. Paired end libraries consisting of two short tags that were originally separated by a known distance in the target genome, allowed assembly where the target genome has deletions, insertions, duplications, inversions and rearrangements. The use of paired ends also overcame the problem with placement of short reads on repetitive genomes. As more data is being generated we are able to establish a high-resolution map of the diploid structural variations present in this Yoruban individual compared to the human genome reference sequence. Once all the libraries are sequenced, this study will help us establish a sample preparation pipeline for re-sequencing of complex genomes

Relevant talk and Poster at AGBT:
High-resolution Structural Variation Detected with Ultra High-throughput Sequencing of Paired End Libraries: Talk by Heather Peckham; February 7th 2008
Ligation-Based High-Throughput Sequencing and 2-Base Encoding for Large Scale Human SNP Detection: Poster by Stephen F. McLaughlin



- Data from recent 50 mer (10 cycles) fragment library sequencing run.
- Random and unique coverage for chromosome 17 shown, illustrating minimal bias relative to G/C content.



- Random and unique coverage vs. % G/C content plotted for chromosome 17 showing no sequencing bias for a single tag of a mate pair library.