SOLiD™ System Mate-Paired Libraries Detect and Define Large Genetic Rearrangements

Introduction

Structural variations are found throughout the genome[1] and have been implicated in several diseases[2,3,4]. Techniques such as array-based comparative genomic hybridization (CGH) allow rapid analysis of genome-wide structural changes. However, the effectiveness of array technologies is limited because they lack the resolution required to identify small copy number variants (CNV), and they cannot detect balanced translocations.

The advent of massively parallel sequencing platforms, such as the SOLiD™ System, allow rapid and precise mapping of structural variations, including translocations, across the entire genome of complex organisms. The data described below demonstrates how large mate-paired libraries, and the simultaneous sequencing of tens of millions of sequence tags, enable the identification and precise positioning of chromosomal rearrangements.

The Power of Mate-Paired Analysis

The power of mate-paired libraries in assembling complex genomes and identifying genetic variants derives from the way in which the libraries are constructed. Randomly sheared fragments are size-selected, and the ends of the fragments are brought into close proximity by re-circularization upon a synthetic oligonucleotide adaptor. Further manipulations produce a library of fragments that allow the sequence of both ends of each fragment to be determined sequentially.

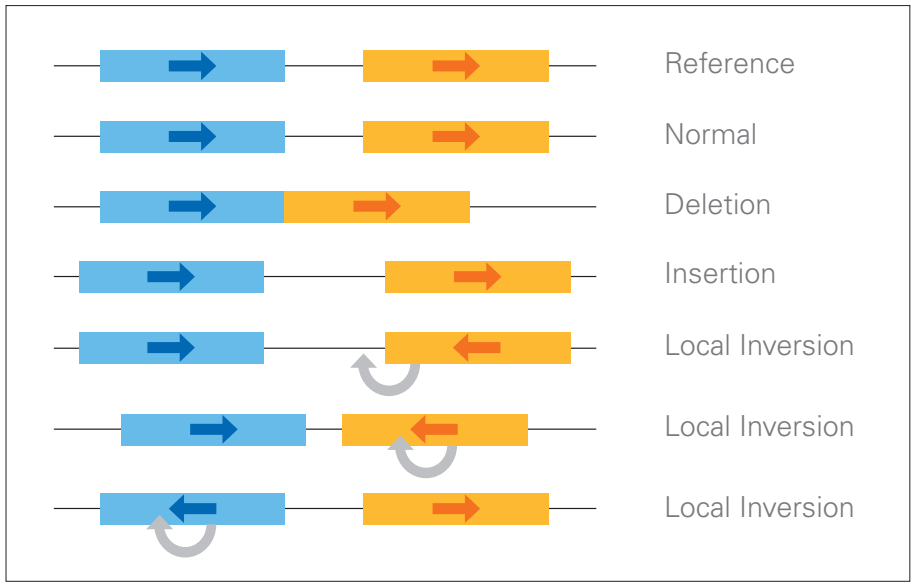

Figure 1. Characterization of structural variation using mate-paired analysis. Blue and orange bars represent 25bp mate-paired sequence tags. The arrows represent the direction of the sequence of the tags against the reference. If the tags map in opposite directions relative to the reference, it may indicate the presence of a rearrangement. The apparent distance between the tags, relative to the reference genome, indicates an insertion or deletion if it differs significantly from the library insert size.

When the sequences of the two mate-pairs are aligned to the reference genome, their orientation and the apparent distance between them is determined. Because both sequences originated from a single DNA fragment of known size, the direction and distance is known and can be compared to the result after mapping to the reference genome. Mate-pairs that map in the same orientation on the same DNA strand are classified as "normal". These normal mate-pairs can be used to assemble a scaffold of the genome that has a simple DNA alignment.

Additionally, mate-pairs with orientation, strand location, or distances that differ from expected values are also important since they may be used to map a variety of structural rearrangements. A distance between two tags that is greater or less than expected may indicate the presence of insertions or deletions. Alternatively, if the first mate-paired sequence maps to one strand in the forward orientation, while the second sequence maps to the other strand in the reverse orientation (referred to as "broken mate-pairs") it would be indicative of a rearrangement occurring between the first and second mate-pair (Figure 1). The following analysis will focus on rearrangements detected by mate-pairs.

Methods

Mate-paired libraries, with inserts of 0.6, 1, 2.5, or 4 kb, were generated from 30 µg of randomly sheared genomic DNA (Yoruban NA18507, Coriell Institute), according to the Applied Biosystems mate-paired library construction protocol (Figure 2). The libraries were clonally amplified on paramagnetic beads, deposited onto a glass slide, and sequenced according to standard Applied Biosystems protocols for the SOLiD System.

The first step in the analysis of mate-paired sequences is to align Tag 1 and then Tag 2 to the reference sequence. Unique sequences were aligned against the human genome (NCBI, b36, hg18) and classified according to the orientation of the first and second tags, relative to each other and to the distance they map from each other on the reference genome (Table 1). Next, broken mate-pairs (orange rows in Table 1) were binned together and scanned for grouping in genomic locations that would be consistent with rearranged sequences by database queries and visual inspections. Multiple sets of "broken" mate-pairs which all map to the same chromosomal region provide evidence for a potential structural variation at that location.

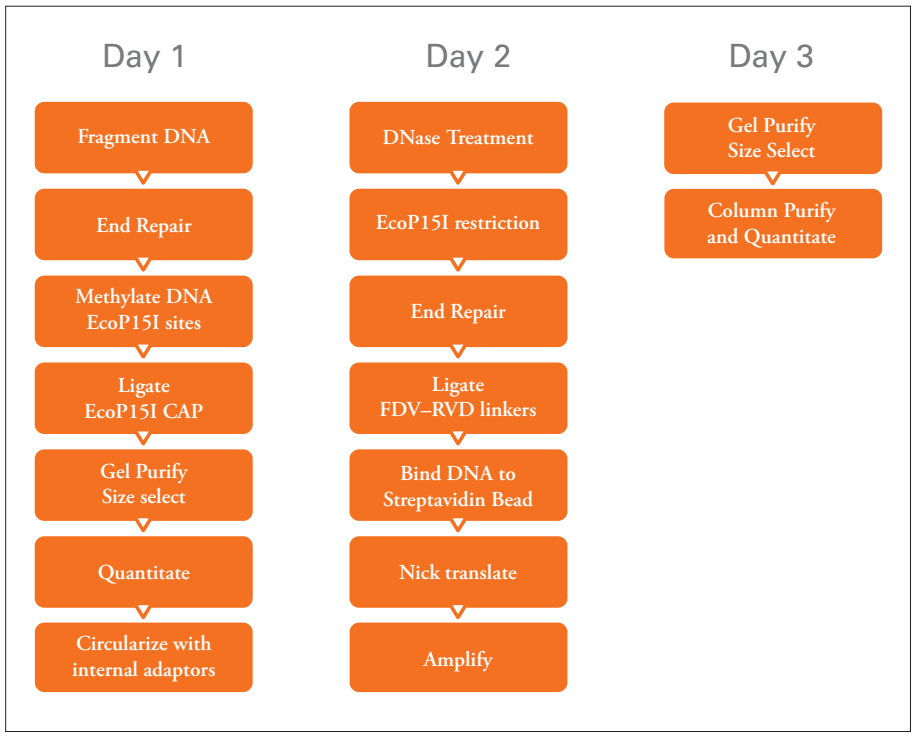

Figure 2. Mate-Paired Analysis Protocol

Results

Multiple sequencing runs were performed on the SOLiD Analyzer, and each run generated at least 1.65 gb of mappable mate-paired data per slide. It should be noted that the SOLiD System can analyze two slides in one run, each run can therefore generate sufficient data to establish a high-resolution structural variation map for two samples. The majority of mate-paired tags from the 0.6, 1, 2.5, and 4 kb libraries aligned across the genome as expected, and provided exceptional depth of coverage. Alignment of "broken" mate-pairs indicated the presence of several potential rearrangements including one on Chromosome 10 (Figure 3). The large number of unique mate-paired tags generated in this experiment made it possible to define two breakpoints to within 117 and 135 nucleotides respectively, suggesting that the inversion is between 36,310 and 36,560 bases in size (Figure 3). PCR primers can now be designed using this information to amplify across both breakpoints, thereby establishing by capillary electrophoresis the exact nucleotide where these breakpoints occur.

The normal and broken mate-paired tags that map to this region of Chromosome 10 were visualized in the SOLiD Alignment Browser (Figure 4). Gaps (indicated by the red arrows) could, in principle, be the result of

TABLE 1. Binning of Mate-Paired Tags.

Bin Code **Letters 1 and 2 = Strand and orientation** **Letter 3 = Distance**	**A = Reference**	**B = Insertion**	**C = Deletion**
AA = Same strand and orientation	AAA	AAB	AAC
AB = Same strand, reverse orientation	ABA	ABB	ABC
BA = Opposite strands, oriented away from each other	BAA	BAB	BAC
BB = Opposite strands, oriented toward each other	BBA	BBB	BBC

* The third letter refers to the distance between the two tags on the reference genome in relation to the expected insert size.
A = identical with the expected insert size; B = too small; C = too large.

incomplete coverage. However, when the coordinates of the broken mate-pairs from Figure 3 are superimposed on this alignment, it is apparent that the ends of the broken mate-pairs map to the positions of the two gaps in the alignment (the black diagonal lines in Figure 4).

Analysis of five other human DNA samples; HCT116 colon cancer cell line (unpublished SOLiD System data), MCF7 breast cancer cell line (unpublished SOLiD System data), Independent human cell line3, Independent Yoruban line (NA18505)[2] and presumed normal individual (NA15510)[2], show rearrangements that map to this same region of Chromosome 10 (Figure 5). Four of the five examples demonstrate rearrangements of a similar size, mapping to a region separated by only a few thousand bases, providing corroborative evidence that a rearrangement has occurred at this location on Chromosome 10.

Conclusion

High throughput sequencing of mate-paired libraries provides a valuable method for characterizing genomic rearrangements, such as insertions, deletions, duplications, rearrangements, and translocations. A robust protocol has been developed to generate millions of mappable mate-paired sequences from libraries that range from 0.6 to 4 kb using the SOLiD System. The sequences generated from a single run of the SOLiD System provide sufficient coverage to detect most structural changes and precisely define the breakpoints of particular rearrangements.

References

[1]McClintock, B. *Genetics* (1941) **26**:234–282 (Article 49).

[2]Korbel, et al. *Science* (2007) **318**:420–426 (Supplemental Data).

[3]Tusen, et al. *Nature Genetics* (2005) **37**:727–732 (Supplemental Data, Figure S1 (B35)).

[4]Pennisi, *Science* (2007)318:1829-1941.

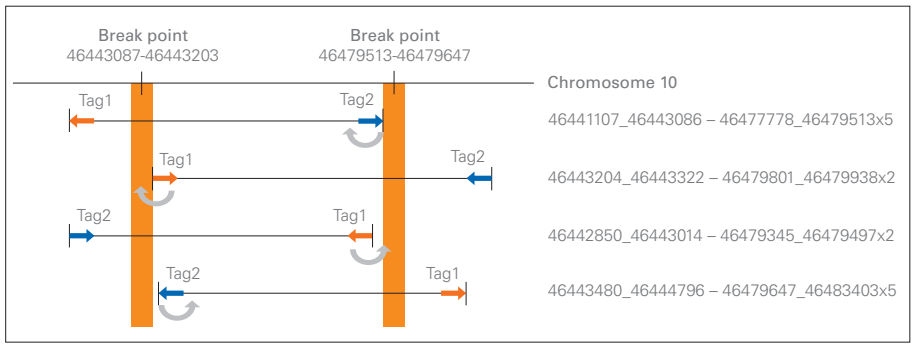

Figure 3. Graphical representation of breakpoints identified on Chromosome 10. The numbers on the right show the genomic coordinates for the broken mate-paired grouped tags mapping to this region. The last number of each group refers to the number of unique tags within that group (i.e., the first mate-pair with coordinates 46441107_46443088_46477778_46479513 is represented 5 times in the data set). However, most tags were unique (1.2X average).

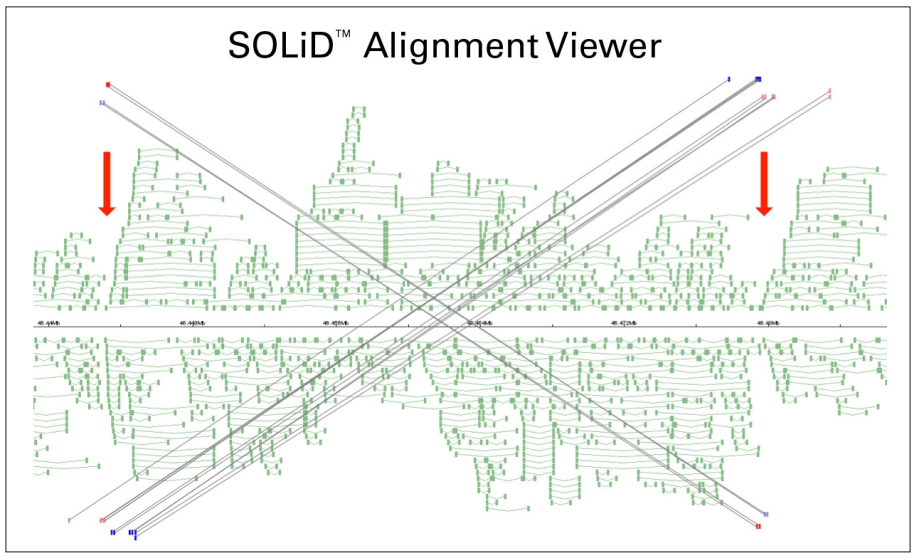

Figure 4. Mate-paired sequence tags mapped to Chromosome 10, visualized in the SOLiD™ Alignment Browser. Individual tags are joined by lines. Green lines represent normal tags. Black lines link tags that map across breakpoints.

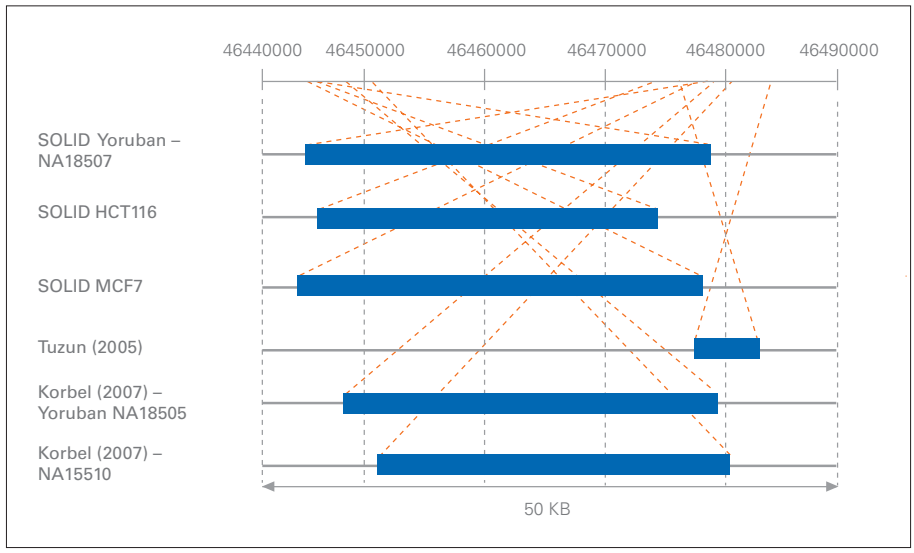

Figure 5. Evidence of chromosomal rearrangements on Chromosome 10, demonstrated in multiple cell lines.

Printed in the USA, 1/2008, Publication 139AP06-01

Headquarters
850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.345.5224
www.appliedbiosystems.com

International Sales
For our office locations please call the division headquarters or refer to our Web site at
www.appliedbiosystems.com/about/offices.cfm