Single Base-pair Breakpoint Resolution Map of Small To Large Indels Using a Split Read Technique with High-throughput Mate-pair and Fragment Sequencing

Eric F. Tsung*, Heather E. Peckham*, Yutao Fu*, Caleb J. Kennedy*, Swati S. Ranade¶, Clarence C. Lee*, Christopher R. Clouser*, Jonathan M. Manning*, Cynthia L. Hendrickson*, Lei Zhang*, Eileen T. Dimalanta*, Tanya D. Sokolsky*, Jeffrey K. Ichikawa*, Jason B. Warner*, Mike W Laptewicz*, Brittney E Coleman*, Fiona C Hyland¶, Francisco M De La Vega¶, Alan P Blanchard*, Gina L Costa* and Kevin J. McKernan***

Applied Biosystems, *Beverly, MA, USA, ¶Foster City, CA, USA

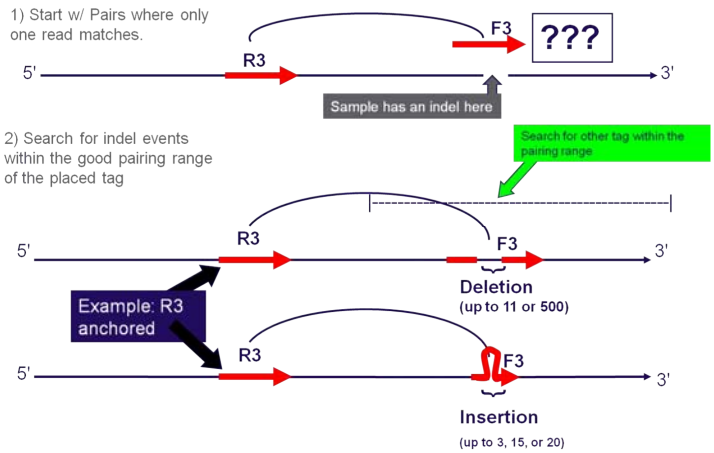
ABSTRACT

Insertions and deletions are often classified as small events detectable within a read and large events detectable between mate-pair reads. We demonstrate that we are now able to detect mid-size variants, previously unattainable by either approach, by using 2x50-bp mate-pair libraries with the Applied Biosystems SOLiD™ System. Specifically, we are able to detect variants of up to 500 bp in length throughout the genome with single-base pair resolution using split-reads.

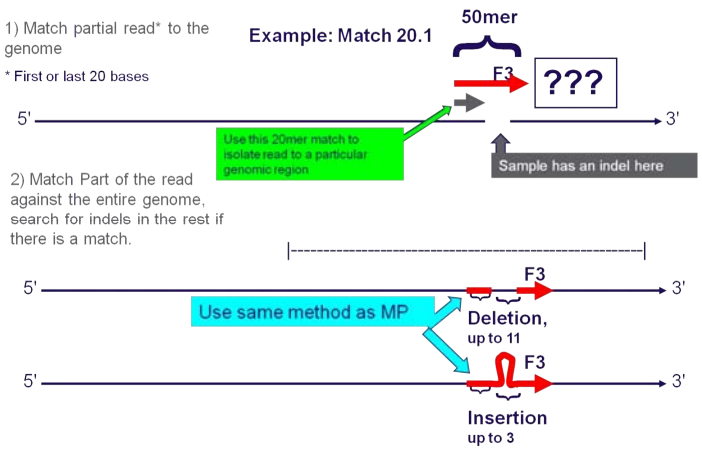
With modest sequence coverage of 8x for the HapMap individual NA18507, we are able to detect over 100,000 small insertions of size 1 to 19, over 100,000 small deletions of size 1 to 20, and 1,156 deletions of sizes 21 to 500 bp, of which 50.9% have not been previously identified in the Venter, Watson, or YanHuang genomes.

Furthermore, we have called indels in several CEPH individuals and show that we are able to detect small indels with a single read critical at low sequence coverage (~0.7x). This is done by aggregating several low coverage samples together, enabling the discovery of substantially greater number of indels. Finally we demonstrate that we can represent indel alleles without ambiguity in the placement while preserving the local short repeat structure that may be present.

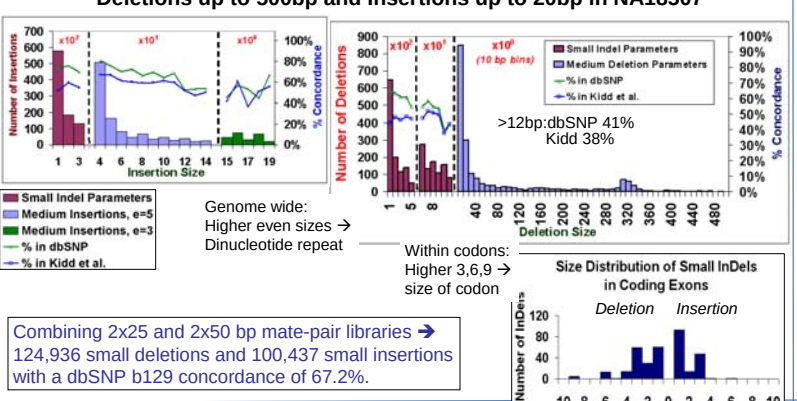
Method of Finding Gap Alignments in Mate Pair Libraries



Method of Finding Gap Alignments in Fragment Libraries



Deletions up to 500bp and Insertions up to 20bp in NA18507



Combining 2x25 and 2x50 bp mate-pair libraries → 124,936 small deletions and 100,437 small insertions with a dbSNP b129 concordance of 67.2%.

Small Indel Analysis of 10 Individuals from the CEPH population

We sequenced 10 members of the CEPH (CEU) population. Displayed are the number of tags and the number of ungapped alignments found. Effective coverage includes only reads that are non-redundant (removing reads with duplicate F3-R3 locations) and have clone size within the expected distribution.

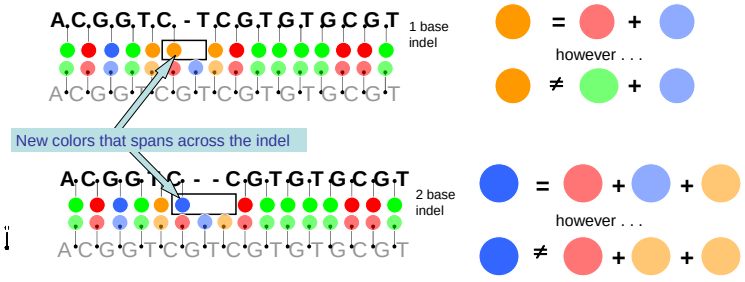
	Coriell ID	Sex	family ID	pop.	Total # of ungapped paired tags (millions)	Effective # of ungapped paired tags (millions)	Total Coverage	Effective Coverage
1	NA06986	Male	13291	CEU	(ask speaker)			
2	NA07000	Female	1340	CEU	163.9	156.5	1.43	1.37
3	NA07357	Male	1345	CEU	200.6	186.2	1.76	1.63
4	NA10851	Male	1344	CEU	(ask speaker)			
5	NA11919	Male	1423	CEU	285.4	269.4	2.50	2.36
6	NA11920	Female	1423	CEU	262.6	230.5	2.29	2.02
7	NA12144	Male	1334	CEU	260.9	242.8	2.28	2.12
8	NA12749	Female	1444	CEU	286.7	268.1	2.51	2.35
9	NA12776	Female	1451	CEU	83.8	80.7	0.73	0.71
10	NA12828	Female	1456	CEU	108.5	102.0	0.95	0.89

Number and dbSNP Concordance of Small Indels (Deletions to 11, Insertions to 4)

	Number of Alignments/ pileups		Called using only the single sample			Found within the aggregate indel calling			Results in the aggregate called with just one read			
Sample:	# of gapped alignments	# of pileups	# indels called	# in dbSNP	% dbSNP b129	# indels called	# in dbSNP	% dbSNP b129	# indels called	# in dbSNP	% dbSNP b129	% found in one read
NA06986	718,982	55,346	48,319	36,777	76.1%	116,991	86,294	73.8%	63,329	45,400	71.7%	**54.1%**
NA07000	517,255	28,637	24,257	18,475	76.2%	84,685	62,917	74.3%	58,456	42,875	73.3%	**69.0%**
NA07357	644,502	35,993	30,115	22,812	75.7%	93,692	68,948	73.6%	61,111	44,196	72.3%	**65.2%**
NA10851	654,721	43,382	37,447	28,636	76.5%	104,235	77,086	74.0%	63,549	45,948	72.3%	**61.0%**
NA11919	907,265	66,529	57,182	43,396	75.9%	131,297	95,881	73.0%	70,206	49,473	70.5%	**53.5%**
NA11920	918,661	47,791	39,984	30,058	75.2%	103,332	74,938	72.5%	55,055	38,603	70.1%	**53.3%**
NA12144	872,601	56,077	47,835	36,171	75.6%	119,016	86,993	73.1%	67,707	48,043	71.0%	**56.9%**
NA12749	797,644	57,334	48,699	36,135	74.2%	109,742	79,369	72.3%	56,614	39,713	70.1%	**51.6%**
NA12776	213,879	11,419	10,002	7,733	77.3%	53,334	40,598	76.1%	42,743	32,351	75.7%	**80.1%**
NA12828	80,400	1,342	1,110	800	72.1%	17,868	13,491	75.5%	16,655	12,600	75.7%	**93.2%**
Totals	6,325,910	403,850	344,950	260,993	75.7%	934,192	686,515	73.5%	555,425	399,202	71.9%	**59.5%**

By pulling all the 2x25 mate pair samples together within the indel caller, we were able to determine more indels found in each of the ten CEPH individuals then by themselves.

Indels have a Distinct Color Space Signature that Differs from Sequencing Errors



Greater confidence in single read results because indels have a distinct color space signature where the color that spans the gap must "color space add" correctly

Short Tandem Repeats at an exact genomic location found within the alignments

Software available at http://solidsoftwaretools.com/gf/project/indel/
Search terms: *solid software tools small indels*

CONCLUSIONS

Using the split-read approach for determining indels on SOLiD™ 50-mer fragment, 2x25 mate pair, and 2x50 mate pair libraries, we were able to find with high confidence a number of small indels. We were also able to determine large deletions up to 500bp, determine STR structure, and determine indels within an aggregated sample with just a single read.

Genome Res. 2009. 19: 1527-1541