Accurate Detection of Insertions and Deletions

Eric F. Tsung[1], Jonathan M. Manning[1], Vasisht Tadigotla[1], Caleb J Kennedy[1], Vrunda Sheth[1], Minita Shah[1], A. John Iafrate[3], Long Phi Le[3], Stephen F. McLaughlin[1], Sowmi Utiramerur[2], Asim S. Siddiqui[2], Clarence C. Lee[1], Heather E. Peckham[1], Fiona C. Hyland[2]
[1]Life Technologies, Beverly, MA, [2]Life Technologies, Foster City, CA, [3]Translational Research Laboratory, Massachusetts General Hospital, Boston, MA

ABSTRACT

We present algorithmic solutions available in the upcoming LifeScope™ Genomic Analysis Software, for indel finding using a ligation based next-generation sequencing platform and Exact Call Chemistry (ECC) available in the SOLiD™ 5500 platform. We demonstrate the algorithm by annotating variants and comparing those results with dbSNP. We show that with the additional ECC information, we gain 33% more indel calls in a HuRef 75x35 paired-end (PE) run, and a large gain of 114% more indel calls in an southern African Bushman, KB1 [ref. 1] mate-pair (MP) run. We also demonstrate that insertions of up to size 29 and deletions up to size 500 are possible with this split-read approach. Furthermore, we show in a targeted sequencing approach of 30 amplicons, known deletions of 15bp and 18bp were detected, and because of the sensitivity of this deep sequencing effort, these variants were possibly detected in previously unknown samples. Finally, we illustrate how split read indels occur in context of other structural variations.

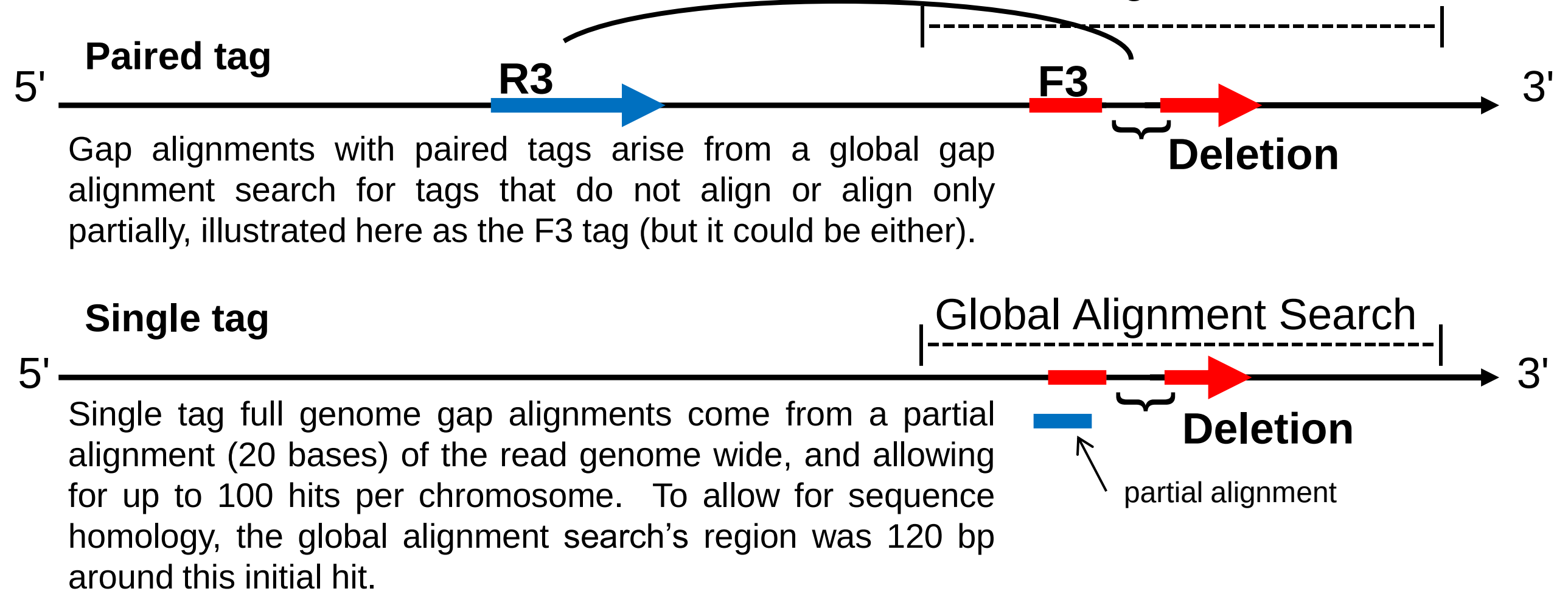
INTRODUCTION

Small complex variants, such as insertions, deletions, and substitutions of a small number of base pairs, although being second only to SNPs in frequency, still pose challenges in performing accurate detection. These challenges come from the need for accurate alignments which are problematical due to the increased degree of freedom of allowing for complex variants, multifaceted zygosity with potentially different allele sequences or gap sizes at the same loci, possible biases imposed by small targeted regions, as well as the possibility of larger scale structural variants such as high copy number. Further challenges include sequencing characteristics that are particular to different platforms.

MATERIALS AND METHODS

The SOLiD™ instrument using technologies available in the 5500, provided the raw DNA sequence, and BioScope™ 1.2/1.3 and a pre-release version of LifeScope™ software, provided the alignment and indel variant calling data used for this analysis.

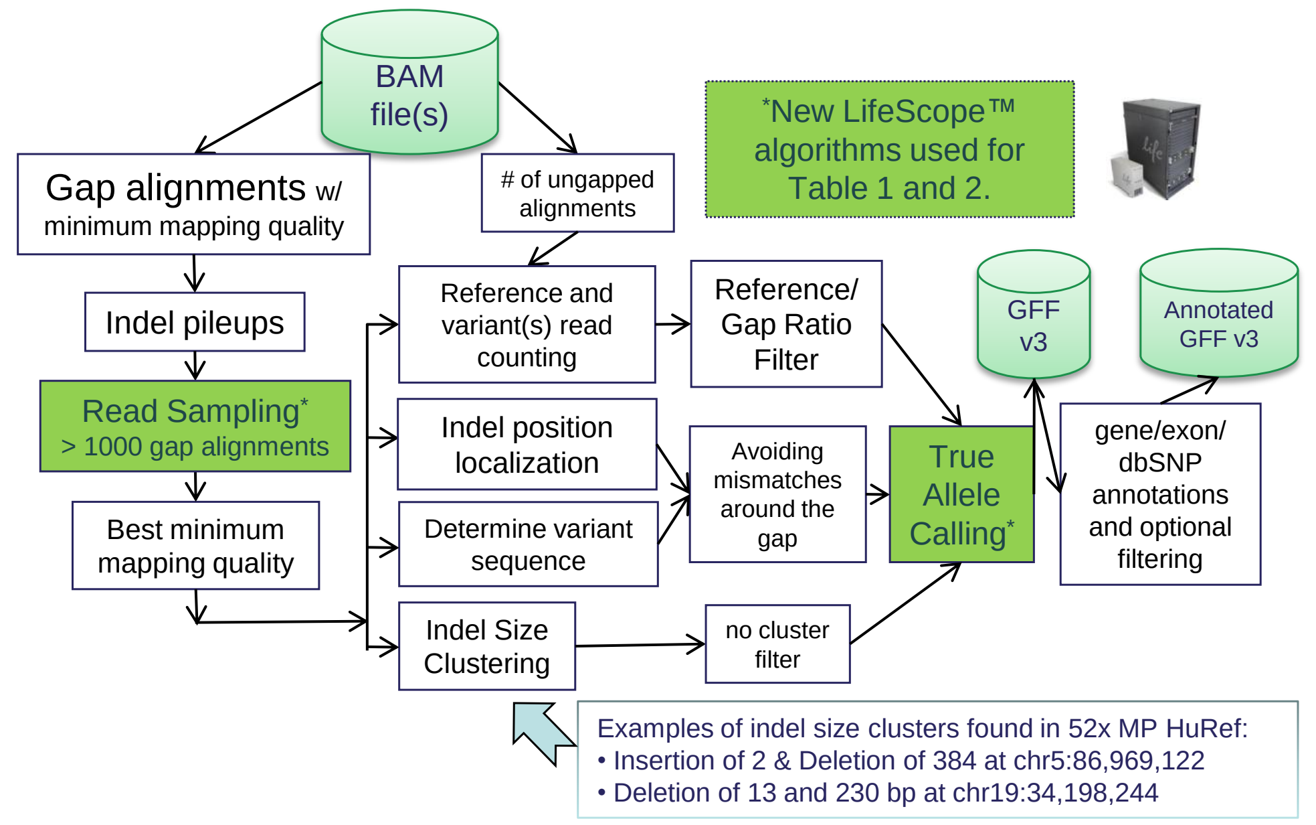
Figure 1. Gap Alignment Strategies



Gap alignments with paired tags arise from a global gap alignment search for tags that do not align or align only partially, illustrated here as the F3 tag (but it could be either).

Single tag full genome gap alignments come from a partial alignment (20 bases) of the read genome wide, and allowing for up to 100 hits per chromosome. To allow for sequence homology, the global alignment search's region was 120 bp around this initial hit.

Figure 2. Algorithms of the split-read (small) indel caller


BAM file(s)
Gap alignments w/ minimum mapping quality
# of ungapped alignments
*New LifeScope™ algorithms used for Table 1 and 2.
Indel pileups
Read Sampling* > 1000 gap alignments
Best minimum mapping quality
Reference and variant(s) read counting
Indel position localization
Determine variant sequence
Indel Size Clustering
Reference/ Gap Ratio Filter
Avoiding mismatches around the gap
no cluster filter
True Allele Calling*
GFF v3
Annotated GFF v3
gene/exon/ dbSNP annotations and optional filtering

Examples of indel size clusters found in 52x MP HuRef:
- Insertion of 2 & Deletion of 384 at chr5:86,969,122
- Deletion of 13 and 230 bp at chr19:34,198,244

This figure describes the algorithms employed here to make accurate indel calls from the alignments contained in the BAM file.

Substantially Greater Indel Detection Sensitivity using ECC

Figure 3. Exact Call Chemistry (ECC): Greater indel calling performance in Mate-Pairs
KB1, 14x 50mer MP library

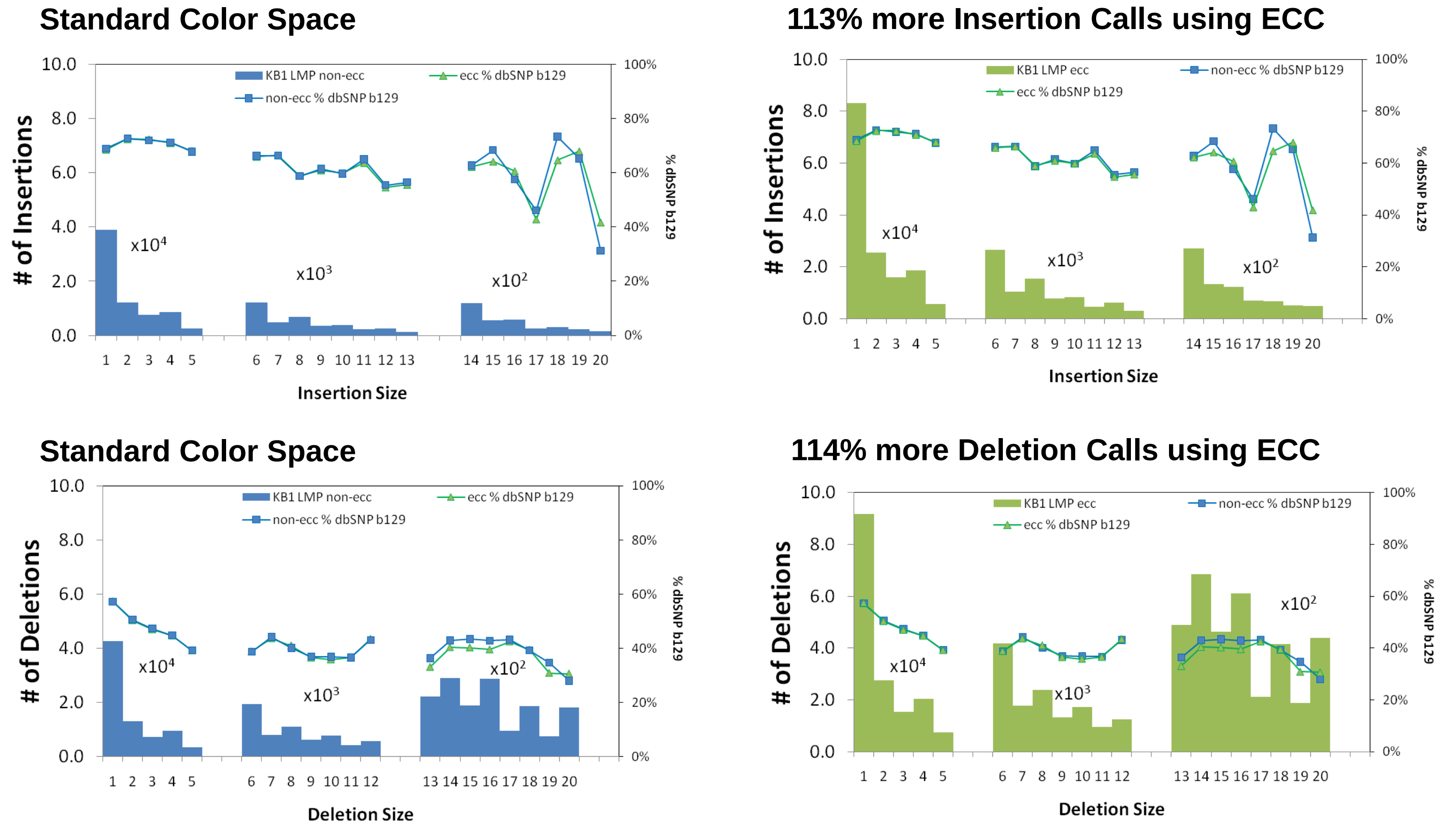
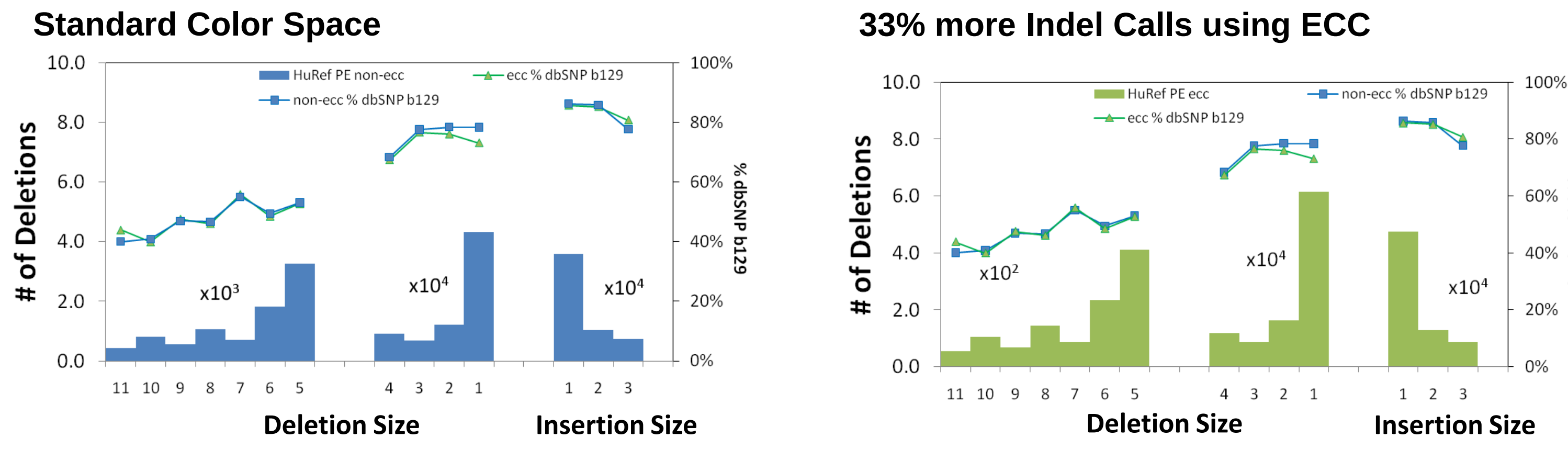


Figure 4. Exact Call Chemistry (ECC) : Greater indel calling performance in Paired-Ends
HuRef, 12x 75x35mer PE library



For indels, shown here to size 20 in KB1 [ref. 1] and HuRef, Exact Call Chemistry (ECC) substantially increases sensitivity of indel calling without sacrificing specificity. Here, the same sequencing run is analyzed using just color space, and then using ECC enhanced color calls during alignment and indel calling. Other factors may be in play here, but one reason for the increase is that accurate gap aligning requires reads with less mismatches, and they are in greater abundance when using ECC.

Targeted Deep Sequencing

Table 1. Accurate Detection of chr7, EGFR, Exon 11 deletions

Sample*	# Reads w/ deletion	# Reads w/ reference	Ref/ Variant Ratio
deletion of 15 in EGFR Exon 19			
# 1	15,465	44,257	2.86
# 2	18	50,946	2,830.30
# 5	88	47,983	545.30
# 6	20	27,798	1,389.90
# 7	24	21,757	906.50
# 8	24	22,514	938.10
deletion of 18 in EGFR Exon 19			
# 4	2,053	49,624	24.17**
# 6	2	20,465	10232.50

*Samples were provided by the Translational Research Laboratory at Massachusetts General Hospital.
** This ratio is above the default setting of the caller's filter, but well under any of the probable false positives found with the filter off. The complete histogram of the noise:

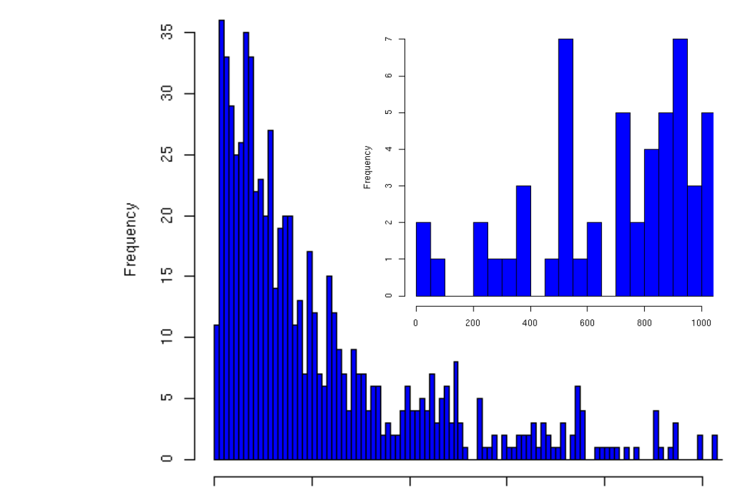


Table 2. All indels found with Ref/Variant ratio filter off

Indel Size	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9
1	91	78	77	67	81	71	59	57	63
2	7	5	7	6	4	2	5	2	3
3	1	2	1	2	1	1	2	3	3
4	0	1	0	0	1	0	0	0	0
7	0	0	0	0	0	0	0	0	1
15	1	1	0	0	1	1	1	1	0
18	0	0	0	1	0	1	0	0	0

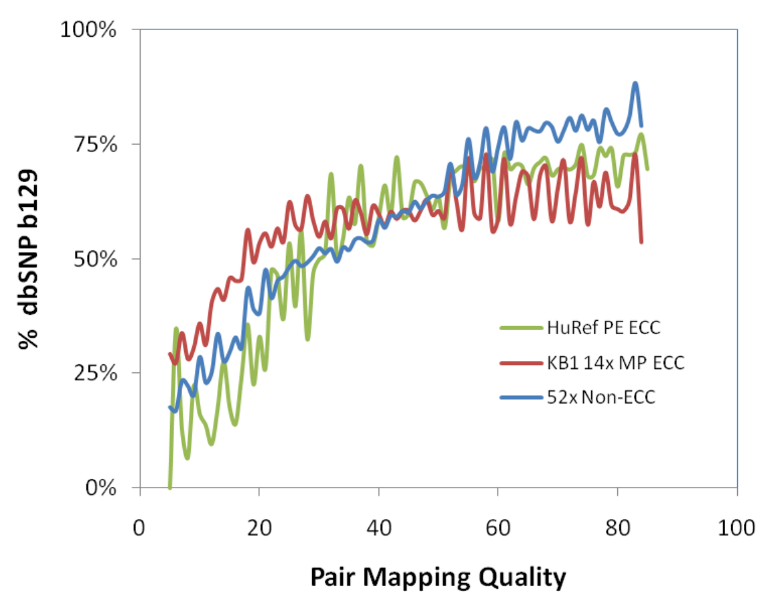
🟩 Known indel 🟦 Possible Indel 🟥 Probable False Positive

9 samples were sequenced using 50mer fragments. Using the primer sequences, 30 amplicon regions ranging from 75-244bp with a total length of 3,679bp were inferred from hg18, and reads were aligned using only these regions. This resulted in an average of 37,700x coverage for each sample with all amplicons fully covered.

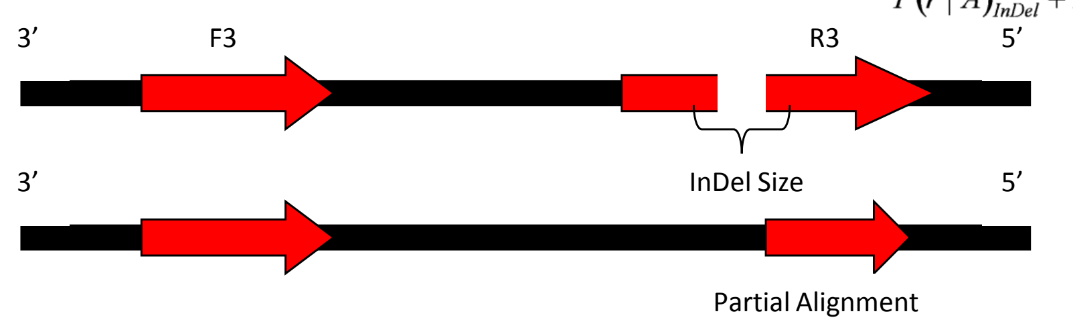
With these alignments, a pre-released version of the LifeScope™ indel caller found, in chr7, EGFR Exon 19, the known deletions of size 15 in sample #1, and size 18 in sample #4, both in green (table 1). The same deletion was found in the other samples, however the reference over variant ratio (figure 2) fell well into the range typically associated with false positives. (orange in table 1). Closer examination of these indels with high reference counts (table 2) reveled that the vast majority had indel sizes of 3 or less (red), and the only larger indels detected (blue) were the same as those in the known samples (green).

Pair Tag's Mapping Quality

Figure 5. Higher pairing quality corresponds to higher dbSNP agreement



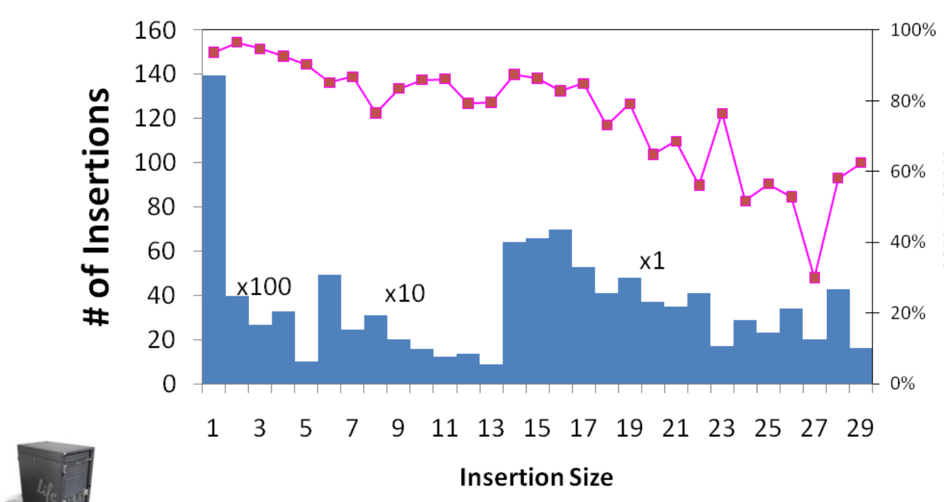
$$P(A\mid r)_{InDel} = \frac{P(r\mid A)_{InDel}}{P(r\mid A)_{InDel} + P_{PartialAlignment}}$$



In calculating the pairing quality (PQV) for gapped alignments, we test against alternative hypothesis of partial un-gapped alignments against having gapped alignments. The read with the gapped alignment is treated as two partial reads on either side of gap. In the end, the read with significant length of alignment on either side of the indel starting point will be assigned a higher PQV.

Accuracy in Insertions to size 29 and Deletions to size 500

Figure 6. Larger Insertions up to size 29* found
HuRef 12x 75x35 PE, using only F3 tag.



Insertion that added an ACTAG repeat detected:

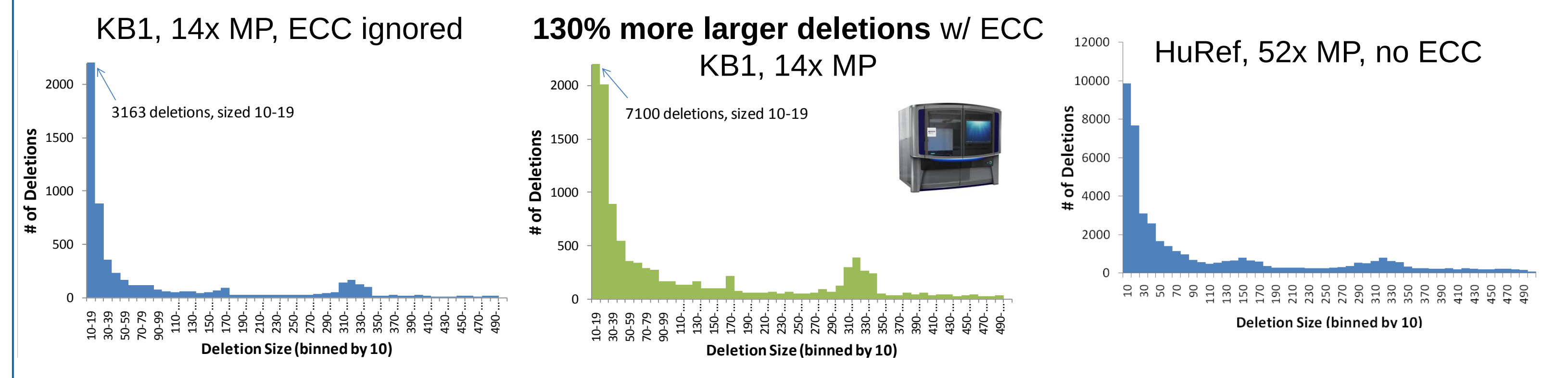
```
chr5:173801687-173801694 (hg18)
ref:       a------------------------actagac
Leftmost:  AACTAGCCATATGCAGAAAACTTAAACTAGAC (after a)
ref:       aactag------------------------ac
Rightmost:AACTAGCCATATGCAGAAAACTTAAACTAGAC (after g)
```

Contained in Indel Caller's GFF:
allele-call-pos=173801687;reference=-;allele-call=-/ACTAGCCATATGCAGAAAACTTAA; zygosity=HEMIZYGOUS-REF;
rightmost-allele-call-pos=173801692;rightmost-reference=-;rightmost-allele-call=-/CCATATGCAGAAAACTTAAACTAG;
context-pos=173801688;context-reference-seq=actag;context-variant-seq=ACTAGCCATATGCAGAAAACTTAAACTAG;

Table 3. Detection of repetitive sequences possible
All positional concordant insertions of size 24 with rsIds

	Position	rsId	Allele match?	Inserted Sequence Detected
chr5	173801687	56012798	Yes	/ACTAGCCATATGCAGAAAACTTAA
chr6	6349078	66501261	Yes	/TAATCCAAACTCTCTCCACGATGA
chr7	3569046	2614943	No	/GAAGCAAATACTTAAATATGAAGA
chr7	39581702	70996815	No	/GCTCTGACTCACTGAGAGCGCTA
chr7	155371648	72305936	Yes	/TCATTCATTCACTCATTCATTCAC
chr8	29259377	66831162	Yes	/TATTTGCAAGCAGAAAGGCTATTT
chr10	116342777	67044911	Yes	/TAGAAAGTAGACAGGGAGAAGAGG
chr11	74516192	71942194	No	/TCTTCTCTTTCTCTCCTCCTCTTTCTT
chr13	72228314	72322477	Yes	/AATAATTCATATGTACCCTTAGGA
chr13	73615474	11267168	Yes	/CAAATACAATTTCATTCTCAATTC
chr13	111649954	71131448	Yes	/ACAACCTCTCCCCTGGCACCTGCC
chr14	103010572	66509506	Yes	/CTACACCTTCCCCTGTCTACTCAA
chr16	84974850	57604273	Yes	/GAAGGAAGCAAGGAAGGAAAGGAA
chr21	43477820	66472242	Yes	/GTCCCACCAAGTTAGAGGAGGGAA
chr22	33896604	72249138	Yes	/GGGAAAGTTTCTTGGTGGGAAGTG

Figure 7. Large Deletions up to size 500* found, ECC again increases sensitivity.



We demonstrate here the accurate detection of larger insertions to size 29 (figure 6) and larger deletions to size 500 (figure 7). Concordance for insertions was by position only, but upon closer examination of all concordant insertions of size 24, the vast majority matched the allele sequence found in dbSNP (table 3). Here, the larger insertions were found in HuRef 12x 75x35 paired-end (PE) library, using only the forward tag. The larger deletions shown here were found in various mate pair (MP) libraries, in KB1 and HuRef. Exact Call Chemistry (ECC) here yielded a 130% increase in the number of large deletions detected.

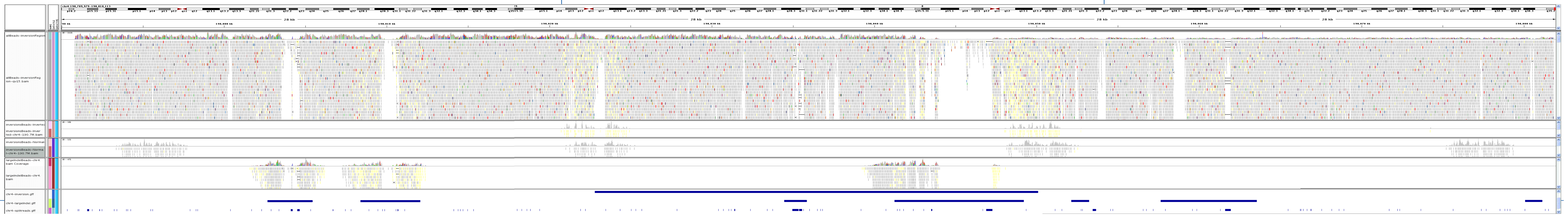
* *LifeScope™ software also has a stretched-mates algorithm capable of finding insertions up to 1.2 kB and deletions up to 100 kB. [ref. 2]*

REFERENCES

1. Schuster, Stephan C., et. al. *Nature* **463**, 943-947 (18 February 2010)
2. McKernan, Kevin J., et. al. *Genome Res. 2009. 19: 1527-1541* (22 June 2009)

Indel calls in an inverted region in HuRef, 52x MP

(end of chr. 4)



Life Technologies • 5791 Van Allen Way • Carlsbad, CA 92008 •
www.lifetechnologies.com