Incorporating multiple NGS read features enables detection of transposon insertions across the genome

Anjali Zimmer, Alex Robertson, Ziga Mahkovec, Gilad Mishne, Lawrence Hon, Jeroen van den Akker
Color Genomics, Burlingame, California, USA

Introduction

Structural variants (SV) are believed to account for at least 10% of pathogenic mutations[1,2], with large deletions and duplications constituting the bulk of these events. To date, only a handful of clinically relevant inversions have been reported in hereditary cancer, with the Boland inversion in *MSH2*[3,4] being the most prominent example. Transposon insertions have been reported in multiple hereditary cancer genes, most frequently in *ATM* and *BRCA2*, including the Portuguese founder mutation (*BRCA2* c.156_157insAlu)[5,6]. Furthermore, insertions have been implicated in dozens of additional diseases[7]. However, the detection of these variations by short read next generation sequencing (NGS) based methodologies has been historically challenging[8]. Here we describe the methodologies that we employ to detect two clinically important structural variants: inversions and insertions.

Methods

Large inversions and insertions induce a major disruption to the genomic sequence, which can impede hybridization between targeting probes and sample DNA. Nonetheless, as long as probes efficiently target regions adjacent to a breakpoint, partial NGS reads up- and downstream of this breakpoint can typically be aligned uniquely. Inversions can then be detected based on split read as well as paired read algorithms such as LUMPY[9]. However, large insertions (especially transposons) impose additional challenges: (i) the inserted sequence is generally longer than NGS reads, so no reads span the inserted sequence; (ii) a transposon inserted sequence is present many times in the genome, causing alignment problems; (iii) transposon insertions often include A-rich linker sequences which cause sequencing artifacts.

To cope with these challenges, we developed a novel algorithm where only a single breakpoint is required to identify an insertion. Since this approach does not rely on a predefined library of mobile element sequences, it's also suitable to detect insertions other than Alu elements. Our algorithm is based on reads where one end aligns to the region of interest, but the other does not (soft-clipped reads). For positions with a significant number of soft-clipped reads, we compute several quality features of the reads and the locus that can be used to distinguish true insertions from NGS artifacts. The most important features are: number of soft-clipped reads, fraction of soft-clipped reads, sequence agreement of the soft-clipped reads, GC content of the locus, sequence uniformity of the locus, number of reads supporting a nearby large indel, and number of inverted reads. Low-quality calls are retained only if the inserted sequence has a high match with Alu elements, which are associated with frequent human gene rearrangements[10]. A second step uses local assembly to estimate the inserted sequence, which is typically sufficient to uniquely identify the exact element in public databases. This method has been validated by sequencing selected samples from the 1000 Genomes data set. Here we present inversions and insertions that have been detected in clinical samples and classified as VUS, LP or P.

Figure 1A: Bioinformatics pipeline for detection of inversions and insertions

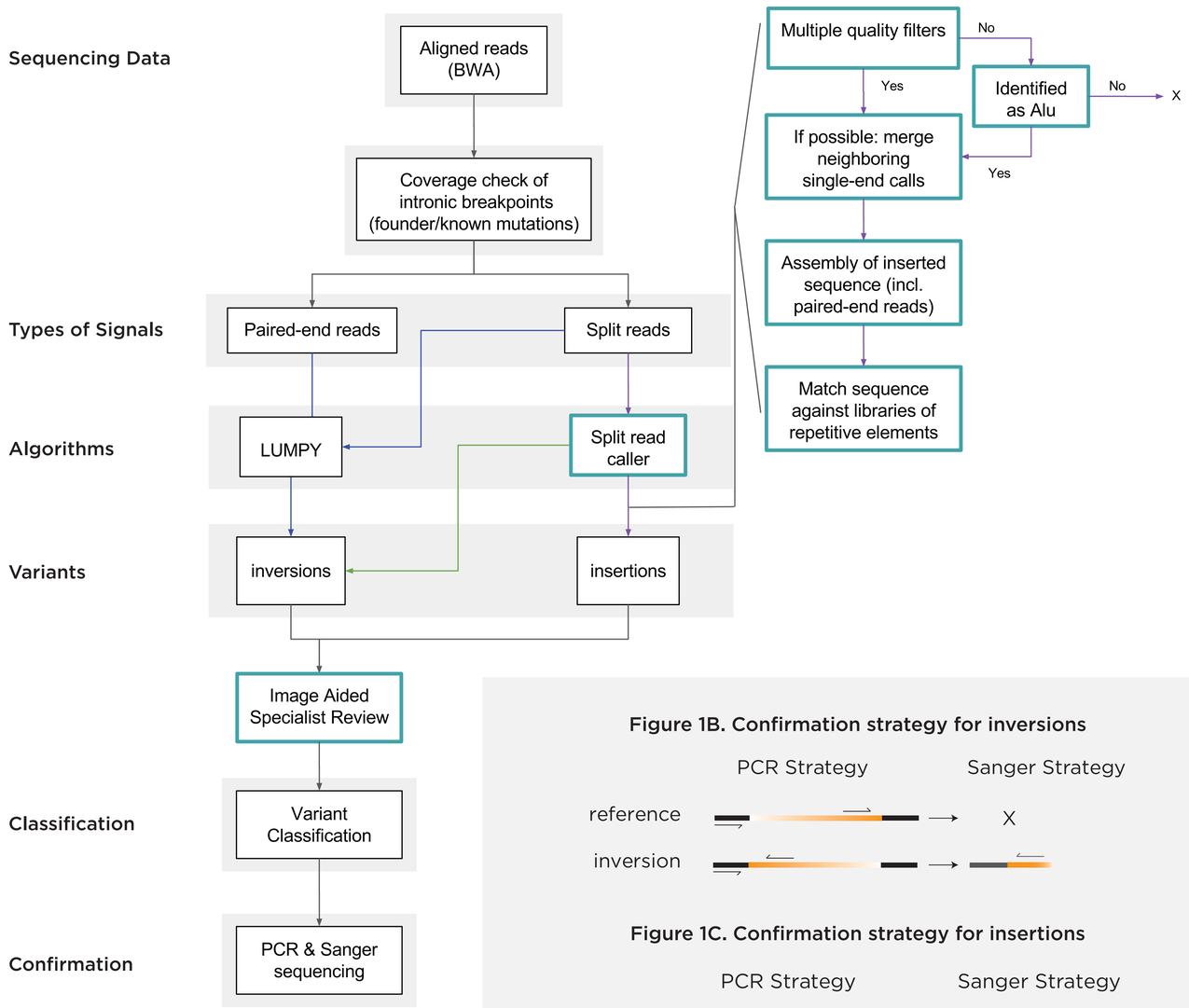

Figure 1B. Confirmation strategy for inversions

Figure 1C. Confirmation strategy for insertions

Overview of typical strategies for confirmation of inversions (Figure 1B) and insertions (Figure 1C) with known breakpoints by Sanger sequencing. Broken arrows represent primers, 'X' indicates no PCR product. Figure 1B shows the strategy for an inversion (orange gradient), and Figure 1C shows the strategy for an insertion (blue).

Results

Figure 2. Identification of an inversion in NGS data

Example of a novel inversion identified in *PMS2*, c.-89564_23+1221inv (exon 1), showing both the left and right breakpoint. Note that approximately half of reads are soft-clipped reads that only partially align to the reference genome (mismatches are shown as primary colored boxes) at both breakpoints.

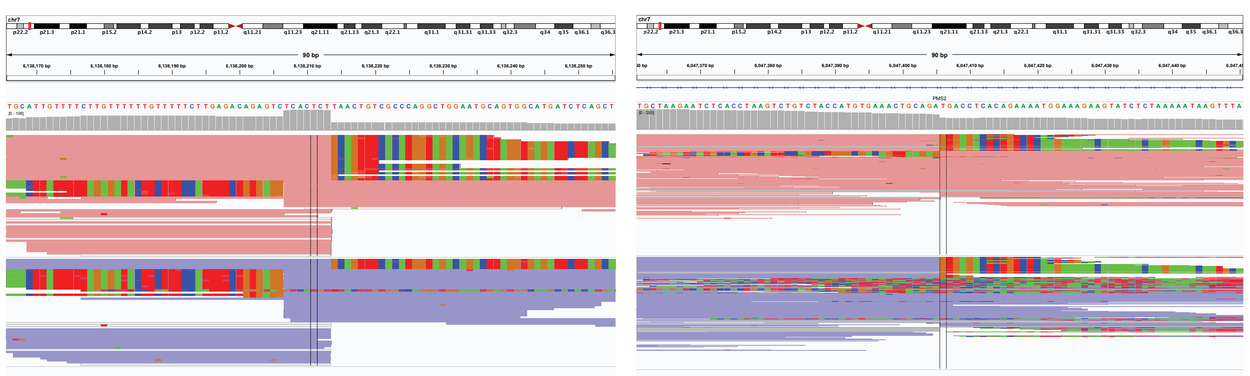

Figure 3. Identification of an insertion in NGS data

Example of a novel Alu insertion identified in *BRCA2*, c.7665_7666insAluYa5 (exon 16). As described, only one breakpoint is necessary to identify an insertion. Note the soft-clipped reads that do not align to reference genome, and the characteristic Alu poly-A tail (shown as a run of misaligned adenines in green).

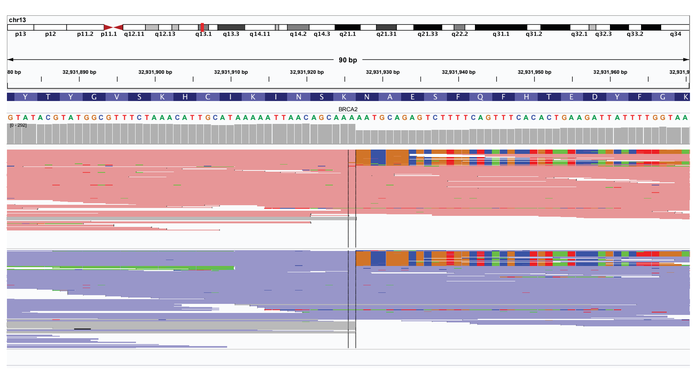

Table 1. Overview of clinically reported inversions

Gene	Inversion	Frequency	Comments
MSH2 (ENST00000233146)	c.-9509220_1277-3164inv (exon 1-7)	1	**Boland founder**
BARD1 (ENST00000260947)	c.1904-6533_*4910157inv (exon 10-11)	1	
PMS2 (ENST00000265849)	c.-89564_23+1221inv (exon 1)	2	

Table 2. Overview of clinically reported insertions

Gene	Insertion	Frequency	Comments
ATM (ENST00000278616)	c.3292_3293insAluYa5 (exon 23)	1	
	c.7388_7389insAluYa5 (exon 50)	2	Identical to 'c.7374_7375insAlu'
BARD1 (ENST00000260947)	c.1568+16_1568+17insAluYa5 (intron 6)	1	
	c.1776_1777insAluYa5 (exon 8)	1	
BRCA1 (ENST00000357654)	c.932_933insAluYb (exon 10)	1	
BRCA2 (ENST00000544455)	c.156_157insAluYa5 (exon 3)	1	**Portuguese founder**
	c.7665_7666insAluYa5 (exon 16)	1	
BRIP1 (ENST00000259008)	c.2505_2506insAluYa (exon 18)	1	
MLH1 (ENST00000231790)	c.588+9_588+10insL1HS (partial)	1	Insertion of ~1.2kb, only sequenced first ~700bp
MSH6 (ENST00000234420)	c.458-19_458-18insAluY (intron 2)	2	
RAD51C (ENST00000337432)	c.945_946insAluYb8 (exon 7)	1	

Conclusions

- While our test has been optimized to detect previously reported inversions (incl. the "Boland" inversion of *MSH2* exon 1-7), our generalized methodology also allowed detection of novel inversions in *PMS2* and *BARD1*.
- Our novel insertion caller is able to identify large insertions based on a single breakpoint and independent of the inserted sequence. In addition to relatively common insertions in *BRCA2* and *ATM*, this approach allowed the detection of Alu insertions in several genes without any published transposon insertions (*BRIP1, MSH6* and *RAD51C*), as well a partial L1HS element in *MLH1*.

References
1. Ewald, I. P. *et al. Genet. Mol. Biol.* **32**, 437–446 (2009). 2. van der Klift, H. *et al. Genes Chromosomes Cancer* **44**, 123–138 (2005). 3. Rhees, J., Arnold, M. & Boland, C. R. *Fam. Cancer* **13**, 219–225 (2014). 4. Wagner, A. *et al. Genes Chromosomes Cancer* **35**, 49–57 (2002). 5. Teugels, E. *et al. Hum. Mutat.* **26**, 284–284 (2005). 6. Peixoto, A. *et al. Breast Cancer Res. Treat.* **114**, 31–38 (2009). 7. Deininger, P. *Genome Biol.* **12**, 236 (2011). 8. De Brakeleer, S., De Grève, J., Lissens, W. & Teugels, E. *Hum. Mutat.* **34**, 785–791 (2013). 9. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. *Genome Biol.* **15**, R84 (2014). 10. Rüdiger, N. S., Gregersen, N. & Kielland-Brandt, M. C. *Nucleic Acids Res.* **23**, 256–260 (1995).