viewing and scoring alignments for structural rearrangements

Eitan Halper-Stromberg

August 21, 2013

Abstract

This package is designed to evaluate structural rearrangment calls from a candidate list, the output for tools such as HYDRA [2],GASV [3],VariationHunter [1], etc. The user should have a text file with one row per candidate structural rearrangment. For each candidate rearrangement, read-pairs from the two loci will be read in from a bam file and realigned three different ways. One of these realignments supports the structural variant, with readpairs realigned to a sequence representing the rearranged sequence (the sequence of the two loci concatenated together). The other two realignments support no structural rearrangement, with readpairs realigned to the two sequences representing contiguous fragments of the reference genome taken from each of the two loci.

>

1 Quick Start Example

We start with a text file containing 20 candidate deletion junctions and the bam file containing our read alignments from a whole genome sequencing experiment (for the purposes of this vignette the bam file contains only reads aligning within the regions that we will interrogate). The text file contains the loci allegedely involved in each deletion, one deletion per row. For each row we will load reads aligning to the two loci involved in the alleged deletion from our bam file, realign these reads, and calculate our likelihood score.

```
> path <- system.file("extdata", package="targetSeqView")
> ## This method utilizes the foreach package for parallelization, set nodes to however many cpus are
> ## available.
> nodes=1
> registerDoMC(nodes)
> ## create an instance of the candidates class
> candidateDels<-new('candidates')
> ## set the path where bam files are located (if not in the currect working directory)
> bamFilePath(candidateDels)<-path
> ## set the name of the text file containing candidate SVs (full path if not in the working directory)
> candidatesFileName(candidateDels)<-file.path(path,'wholeGenomeDeletionCandidates.txt')
> ## set the build of the (human) genome
> build(candidateDels) <- 'hg19'
> ## set the read length
> readLength(candidateDels) <- 101
> ## set the mismatch rate for each position along the read length
> mmRate(candidateDels) <- precomputedWholeGenome101bpMMRate()
> ## set the indel rate for reach position along the read length
> indelRate(candidateDels) <- precomputedWholeGenome101bpIndelRate()
```

note: mismatch and indel rates may be calculated based upon reads from a bam file containing normal alignments, the bamFile argument should contain the full path with the bam name if the bam is not in the current directory. The following 3 lines are unevaluated in this vignette

```
>  normalBam <- 'Path/To/Normal/bamfile.bam'
> errorRates<-getErrorRate(normalBam)
> mmRate(candidateDels) <- errorRates[['mmRate']]
> indelRate(candidateDels) <- errorRates[['indelRate']]
```

We first obtain likelihood scores for candidates without performing full smith-waterman realignment on all reads for all 3 alignement configurations. We instead use alignment information in the cigar strings and md tags in our bam file to obtain mismatches and indels for the alignments supporting SVs. In addition, We forgo, for the moment, returning a data.frame formatted for our plot function. This should take a few (1-5) seconds per candidate

```
> candidateDels<- quickScore(candidateDels,verbose=TRUE)

Of 20 events now working on  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

> ## view values returned
> print(candidateDels@quickScore)

          1           2           3           4           5
1280.591846  896.235436  881.603501  793.243114  556.464547
          6           7           8           9          10
 844.628480  741.041974  461.248676 1065.448055  749.677869
         11          12          13          14          15
-252.563470  -77.754819  -41.597421  -35.834722  -33.225283
         16          17          18          19          20
 -32.736277  -28.492610  -25.970714  -25.165426   -7.992096

> ### In this case we have validation data for these candidates
> indexOfvalidated <-1:10
> validated<-candidateDels@quickScore[indexOfvalidated]
> failedvalidation<-candidateDels@quickScore[-indexOfvalidated]
```

Figure 1: The distribution of those that validated and those that did not

```
> boxplot(list(validated=validated,failed=failedvalidation,
+     all=candidateDels@quickScore),ylab='log likelihood score')
```

2 Scoring and Viewing

In this section we will view read alignments at the junctions of 3 candidate structural variants. We will use a sequencing dataset taken from a target-capture exerpiment. As with the last section, we start with a text file containing the candidate structural variants (in this case 1 inversion and 2 chromosomal translocations) and a bam file containing read alignments.

```
> ## create an instance of the candidates class
> candidateSVs<-new('candidates')
> bamFilePath(candidateSVs) <- path
> candidatesFileName(candidateSVs) <- file.path(path,'targetCaptureSVs.txt')
> build(candidateSVs) <- 'hg19'
> readLength(candidateSVs) <- 100
> mmRate(candidateSVs) <- precomputedTargetCapture100bpMMRate()
> indelRate(candidateSVs) <- precomputedTargetCapture100bpIndelRate()
> ## fullScoreAndview will perform full smith-waterman realignment for all reads in the 3
> ## configurations. In addition, if the input text file contains a SplitsSample column,
> ## the function will look for split-reads within the bam file specified by the column 'SplitsSample'
> candidateSVs<-fullScoreAndView(candidateSVs,verbose=TRUE,findSplitReads=TRUE)
```

```
[1] "Working on event 1 of 3"
[1] "primary alignment for event 1 done"
[1] "secondary alignment (1 of 2) for event 1 done"
[1] "secondary alignment (2 of 2) for event 1 done"
[1] "Working on event 2 of 3"
[1] "primary alignment for event 2 done"
[1] "secondary alignment (1 of 2) for event 2 done"
[1] "secondary alignment (2 of 2) for event 2 done"
[1] "Working on event 3 of 3"
[1] "primary alignment for event 3 done"
[1] "secondary alignment (1 of 2) for event 3 done"
[1] "secondary alignment (2 of 2) for event 3 done"
```

The scores for these 3 events:

```
> print(candidateSVs@fullScore)

[1]   1.182166 957.484838 501.008999
```

Figure 2: A chromosomal translocation that failed to validate

PCR validation informs us that the first event is negative and the other two are positive. Let's View the negative (the flipLeftandRight argument is just a style preference, in this case putting the chr14 junction on the left and the chr15 junction on the right)

```
> plotSV(candidateSVs,indices=1,flipLeftandRight=TRUE,pdfname='fig1.pdf')
```

.

Figure 3: A validated inversion

Let's view the first positive. Read-pair alignments supporting the SV look good, read-pair alignments supporting contiguous fragments do not look good because in both contiguous fragment alignment pictures, one read from each pair has many mismatches/indels Figure 3)

```
> plotSV(candidateSVs,indices=2,pdfname='fig2.pdf')
```

.

Figure 4: A validated chromosomal translocation

Lets view the second positive. Read-pair alignments supporting the SV look good, albeit for a different reason than the first positive. In this picture we have some reads aligning well to the chr14 side and their partners aligning across the junction of the SV (i.e split-reads). The split-reads align well to both sides. There are a few mismatches right at the junction for these split reads but otherwise they match the reference. The contiguous fragment alignments do not look good, as we would expect. Again, the flipLeftandRight option is a style preference, putting the chr14 junction on the left and the chr18 junction on the right. Figure 4

```
> plotSV(candidateSVs,indices=3,flipLeftandRight=TRUE,pdfname='fig3.pdf',width=10)
```

.

```
> toLatex(sessionInfo())
```

- R version 3.0.1 Patched (2013-05-16 r62754), `x86_64-unknown-linux-gnu`

- Locale: LC_CTYPE=en_US.iso885915, LC_NUMERIC=C, LC_TIME=en_US.iso885915, LC_COLLATE=C, LC_MONETARY=en_US.iso885915, LC_MESSAGES=en_US.iso885915, LC_PAPER=C, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.iso885915, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, stats, utils
- Other packages: BSgenome 1.28.0, BSgenome.Hsapiens.UCSC.hg19 1.3.19, BiocGenerics 0.6.0, Biostrings 2.28.0, GenomicRanges 1.12.3, IRanges 1.18.1, Rsamtools 1.12.3, doMC 1.3.0, foreach 1.4.0, ggplot2 0.9.3.1, iterators 1.0.6, targetSeqView 0.99
- Loaded via a namespace (and not attached): MASS 7.3-26, RColorBrewer 1.0-5, bitops 1.0-5, codetools 0.2-8, colorspace 1.2-2, compiler 3.0.1, dichromat 2.0-0, digest 0.6.3, gtable 0.1.2, labeling 0.1, munsell 0.4, plyr 1.8, proto 0.3-10, reshape2 1.2.2, scales 0.2.3, stats4 3.0.1, stringr 0.6.2, tools 3.0.1, zlibbioc 1.6.0

References

[1] Fereydoun Hormozdiari, Iman Hajirasouliha, Phuong Dao, Faraz Hach, Deniz Yorukoglu, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. Next-generation variationhunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, 26(12):i350–i357, 2010.

[2] Aaron R. Quinlan, Royden A. Clark, Svetlana Sokolova, Mitchell L. Leibowitz, Yujun Zhang, Matthew E. Hurles, Joshua C. Mell, and Ira M. Hall. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research*, 20(5):623–635, 2010.

[3] Suzanne Sindi, Elena Helman, Ali Bashir, and Benjamin J. Raphael. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25(12):i222–i230, 2009.

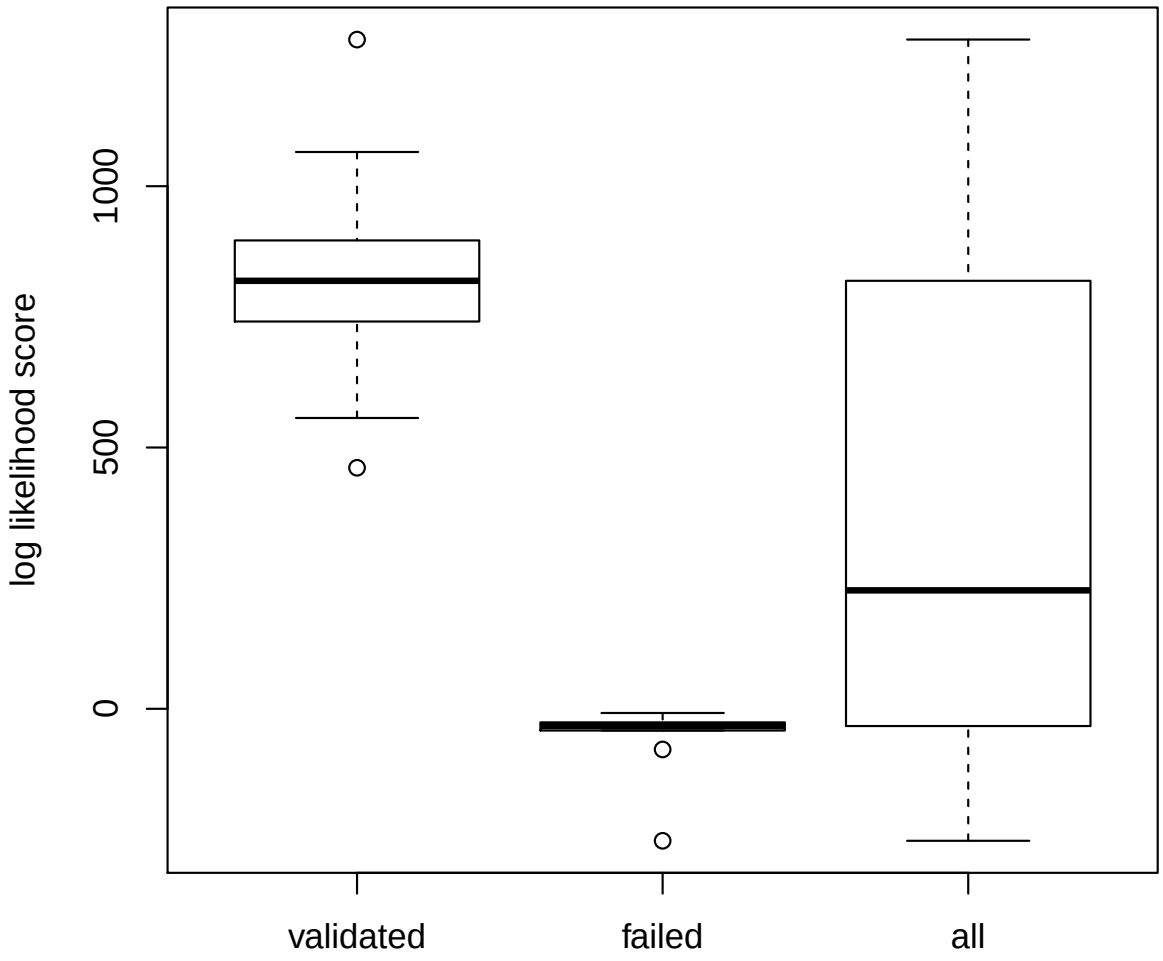

Figure 1: Distribution of 20 candidate deletions taken from a whole-genome sequencing dataset, broken down by validation status

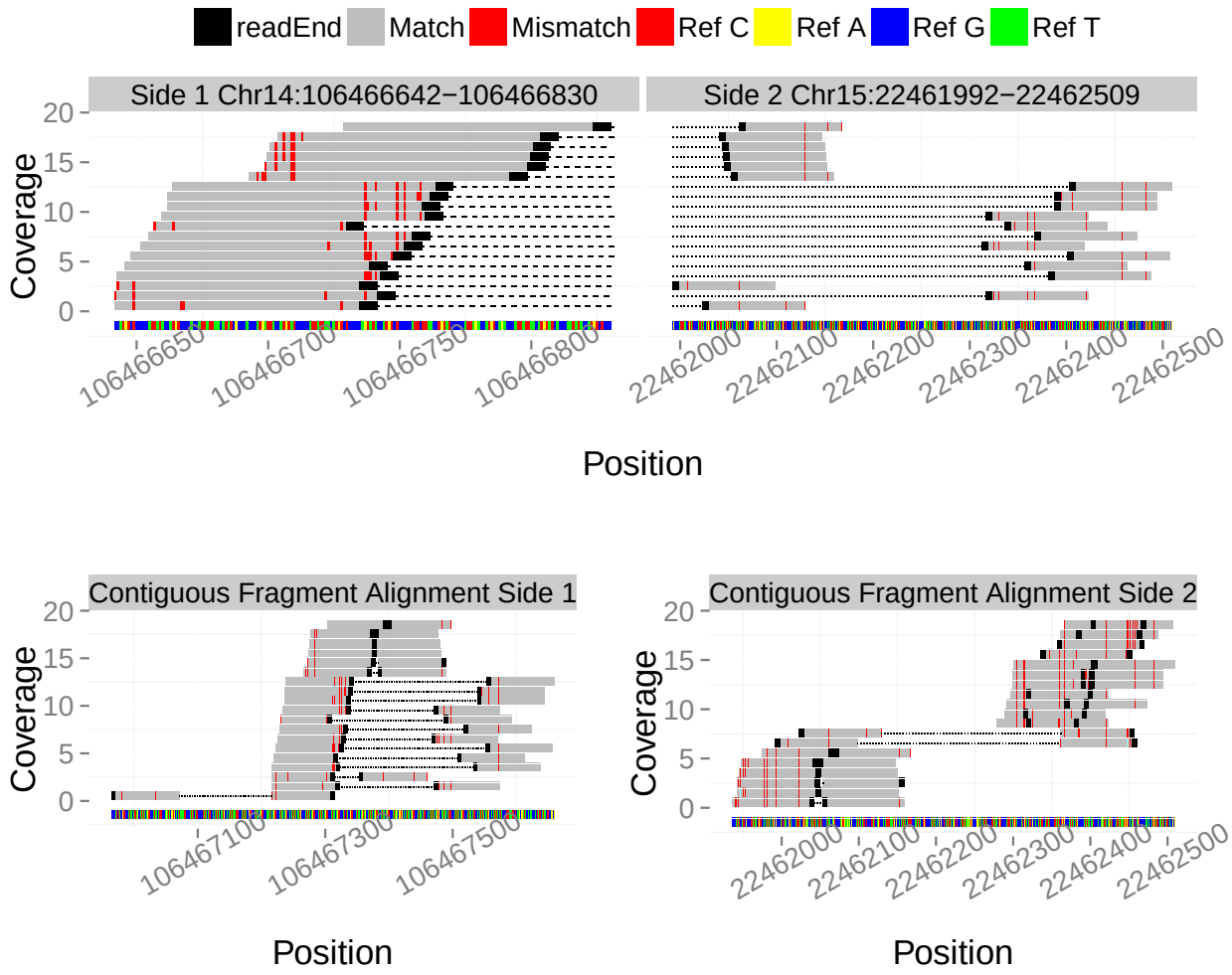

Figure 2: A negative (i.e not real) chromosomal translocation. The top plot shows read-pair alignments supporting the SV and the bottom plots show read-pair alignments supporting contiguous sequences.

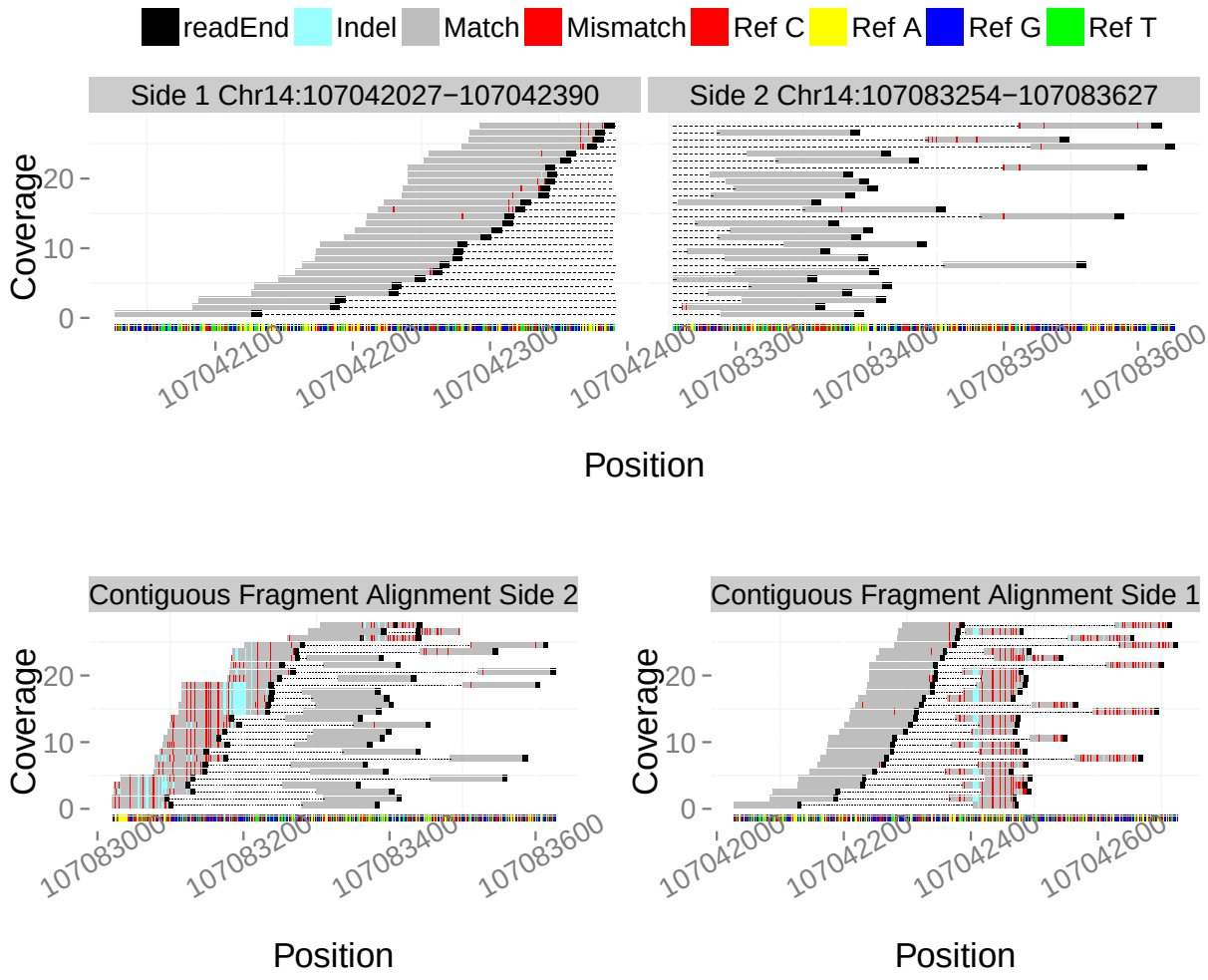

Figure 3: A positive (i.e real) inversion. The top plot shows read-pair alignments supporting the SV and the bottom plots show read-pair alignments supporting contiguous sequences.

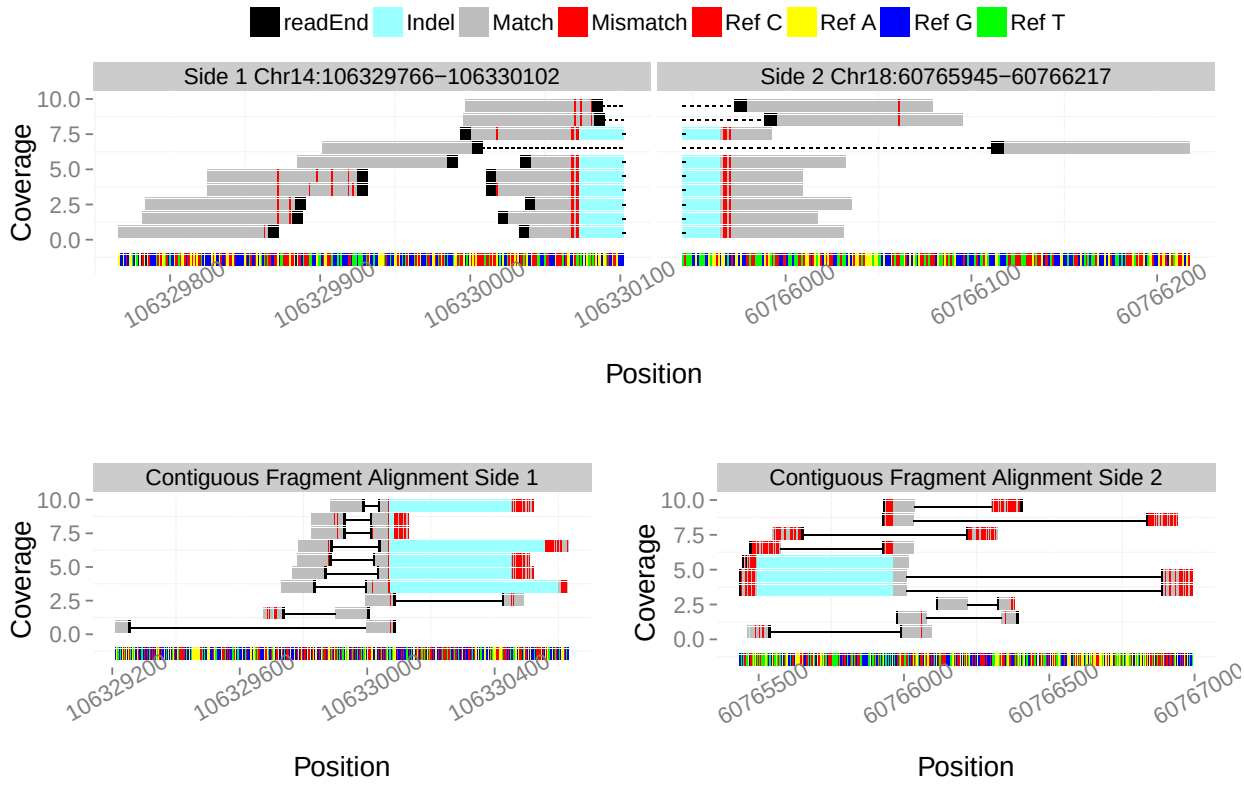

Figure 4: A positive (i.e real) chromosomal translocation. The top plot shows read-pair alignments supporting the SV and the bottom plots show read-pair alignments supporting contiguous sequences.