BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 20 2011, pages 2903–2904
doi:10.1093/bioinformatics/btr467

Genome analysis

Advance Access publication August 11, 2011

ChimeraScan: a tool for identifying chimeric transcription in sequencing data

Matthew K. Iyer[1,2], Arul M. Chinnaiyan[1,2,3,4,5] and Christopher A. Maher[1,2,3,*]

[1]Michigan Center for Translational Pathology, [2]Center for Computational Medicine and Biology, [3]Department of Pathology, [4]Howard Hughes Medical Institute and [5]Department of Urology, University of Michigan Medical School, Ann Arbor, MI 48109, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Next generation sequencing (NGS) technologies have enabled *de novo* gene fusion discovery that could reveal candidates with therapeutic significance in cancer. Here we present an open-source software package, ChimeraScan, for the discovery of chimeric transcription between two independent transcripts in high-throughput transcriptome sequencing data.

Availability: http://chimerascan.googlecode.com

Contact: cmaher@dom.wustl.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 4, 2011; revised on July 26, 2011; accepted on August 3, 2011

1 INTRODUCTION

High-throughput transcriptome sequencing (RNA-Seq) facilitates detection of aberrant, chimeric RNAs (Maher *et al.*, 2009a; Maher *et al.*, 2009b). Methods for chimera detection have already uncovered recurrent classes of clinically relevant gene fusions in prostate (Palanisamy *et al.*, 2010) and lymphoid cancers (Steidl *et al.* 2011). Therefore, the continued development of accurate and efficient software tools for chimera discovery is of major clinical significance. To this end, we have developed a chimera discovery methodology, or ChimeraScan, and offer it as open-source software package for the community to utilize for their own sequencing efforts. ChimeraScan includes features such as the ability to process long (>75 bp) paired-end reads, processing of ambiguously mapping reads, detection of reads spanning a fusion junction, integration with the popular Bowtie aligner (Langmead *et al.*, 2009), supports the standardized SAM format and generation of HTML reports for easy investigation of results. Overall, we believe that the ChimeraScan will facilitate the discovery of additional gene fusions that may serve as clinically relevant targets in cancer.

2 METHODS

Initial paired-end alignment: ChimeraScan uses Bowtie to align paired-end reads to a combined genome-transcriptome reference. An indexing program creates the combined index from genomic sequences (FASTA format) and transcript features (UCSC GenePred format). Paired alignments within the fragment size range (default: 0–1000) are referred to as concordantly mapping reads (Fig. 1A). ChimeraScan uses these alignments to estimate the insert size distribution of the library, which will later be used to filter out likely false positive chimeras.

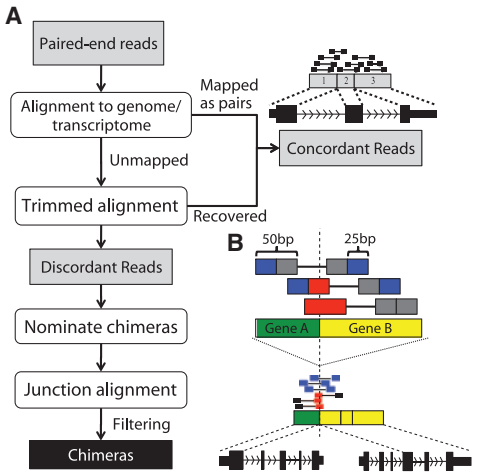

Fig. 1. ChimeraScan flowchart. (**A**) Paired-end reads failing an initial alignment step are segmented and realigned to detect discordant reads. Discordant reads that pass filter criteria are realigned across putative chimeric junctions. (**B**) Chimera with encompassing (blue) and spanning (red) segments detected during realignment.

Trimmed paired-end alignment: read pairs that could not be aligned concordantly are trimmed into smaller segments (default = 25 bp) and realigned. Trimming increases the chance that neither read alignment spans a chimeric junction, thereby improving sensitivity for nominating chimeras.

Nomination of chimera candidates: the trimmed alignments are scanned for evidence of discordant read pairs, or reads that align to distinct references or distant genomic locations (as determined by the fragment size range) of the same reference. Reads aligning to overlapping transcripts are not considered discordant. ChimeraScan clusters the discordant reads and produces a list of putative 5′–3′ transcript pairs that serve as chimera candidates.

Detection of reads spanning the chimeric junction: ChimeraScan builds a new reference index from the set of putative chimeric junction sequences, and realigns candidate junction-spanning reads to this index. Candidate spanning reads are either (i) discordant reads with trimmed alignments bordering a junction or (ii) unmapped reads whose mates align to a predicted chimera (Fig. 1B). A read that spans a junction by more than a minimum 'anchor' length is denoted as a 'spanning' read. We compute the required 'anchor' length separately for each chimera by insisting that the number of bases overlapping its junction be greater than number of homologous bases between the 5′ and 3′ genes at the breakpoint plus the number of mismatches allowed.

*To whom correspondence should be addressed.

Filtering false-positive chimeras: after spanning reads are incorporated, ChimeraScan filters chimeras with few supporting reads (default is <3 reads) and chimeras with fragment sizes far outside the range of the distribution (default is >99% of all fragment sizes). When isoforms of the same gene support a fusion ChimeraScan only retains the isoform(s) with highest coverage.

Reporting chimeras: ChimeraScan produces a tabular text file describing each chimera, and optionally generates a user-friendly HTML page with links to detailed descriptions of the chimeric genes.

3 RESULTS

To evaluate the results from ChimeraScan, we applied it to three well-characterized cancer cell lines known to harbor multiple chimeric transcripts: VCaP (prostate cancer, 2×53 bp) (Tomlins *et al.*, 2005), LNCaP (prostate cancer, 2×34 bp) and MCF7 (breast cancer, 2×35 bp) (Hampton *et al.*, 2009; Volik *et al.*, 2006). Sequence data are deposited in GenBank under the accession number GSE29098. We aligned to human genome (VR-hg19) and UCSC known transcripts (December 2010), allowing for up to two mismatches and no >100 alignments per read. The trimmed alignment step was performed with 25 bp segments.

As our initial benchmark, we confirmed that ChimeraScan was able to recapitulate experimentally validated candidates, our 'gold standard' (Supplementary Table 1) (Maher *et al.*, 2009b). ChimeraScan was able to detect 9/10, 4/4 and 12/13 chimeras from VCaP, LNCaP and MCF-7, respectively.

In addition to recapitulating previously reported results, we have identified novel candidates that demonstrate ChimeraScan's ability to identify and prioritize high-quality chimeras. Overall, ChimeraScan nominated 335 novel chimeras (78 in VCaP, 105 in LNCaP and 152 in MCF7) from the three cell lines (Supplementary Table 2–4). Interestingly, we detected an interchromosomal rearrangement *TBL1XR1-RGS17* detected in the MCF-7 cell line. While not originally reported within NGS data (Maher *et al.*, 2009b), *TBL1XR1-RGS17* was previously detected by a paired-end diTag approach and experimentally confirmed (Ruan *et al.*, 2007). Another novel candidate was the intrachromosomal rearrangement, *NDUFAF2-MAST4*, in VCaP that is supported by just two encompassing reads and one spanning reads. The ability to identify a high-quality spanning read that uniquely confirms the fusion junction (Supplementary Table 2), thereby increasing our confidence in *NDUFAF2-MAST4*, demonstrates the sensitivity of ChimeraScan.

We next compared ChimeraScan with publicly available tools deFuse (McPherson *et al.*, 2011), shortFuse (Kinsella *et al.*, 2011) and MapSplice (Wang *et al.*, 2010) using the 10 experimentally validated VCaP chimeras (Supplementary Table 5). While deFuse nominated the fewest chimeras, it only detected 60% of the true positives. In comparison, ChimeraScan detected 90% of the true positives from 78 predicted chimeras. Of the remaining programs, MapSplice nominated 400 chimeras while detecting 60% of the true positives and ShortFuse nominated 245 chimeras while confirming 70% of the true positives. Overall, these results suggest that ChimeraScan is among the more stringent programs while enriching for true positives.

4 CONCLUSION

Here, we present an optimized publicly available chimera discovery methodology for identifying novel therapeutically targetable gene fusions in human cancers. Our results suggest that ChimeraScan produces a stringent list of predictions that are enriched with true positives. Furthermore, due to its trimmed alignment steps we believe ChimeraScan will be scalable when longer reads are available to provide increased coverage of fusion junctions. Overall, we feel that with the existing features ChimeraScan is a user-friendly tool that will enable other research groups to make discoveries within their own RNA-Seq data collections.

Funding: Department of Defense Breast Cancer Predoctoral Grant (to M.K.I.); Prostate Cancer Foundation Young Investigator Award and National Institutes of Health Pathway to Independence (K99 CA149182-01) Award (to C.A.M.); National Institutes of Health, Department of Defense and Early Detection Research Network (to A.M.C.).

Conflict of Interest: none declared.

REFERENCES

Hampton,O.A. *et al.* (2009) A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res.*, **19**, 167–177.

Kinsella,M. *et al.* (2011) Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics*, **27**, 1068–1075.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Maher,C.A. *et al.* (2009a) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.

Maher,C.A. *et al.* (2009b) Chimeric transcript discovery by paired-end transcriptome sequencing, *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.

McPherson,A. *et al.* (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.*, **7**, e1001138.

Palanisamy,N. *et al.* (2010) Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nat. Med.*, **16**, 793–798.

Ruan,Y. *et al.* (2007) Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.*, **17**, 828–838.

Steidl,C. *et al*. (2011) MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature*, **471**, 377–381.

Tomlins,S.A. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.

Volik,S. *et al.* (2006) Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res.*, **16**, 394–404.

Wang,K. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.