BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 15 2014, pages 2224–2226
doi:10.1093/bioinformatics/btu169

Gene expression

Advance Access publication April 1, 2014

PRADA: pipeline for RNA sequencing data analysis

Wandaliz Torres-García[1,†], Siyuan Zheng[1,†], Andrey Sivachenko[2,†], Rahulsimham Vegesna[1], Qianghu Wang[1], Rong Yao[1], Michael F. Berger[3], John N. Weinstein[1], Gad Getz[2,*] and Roel G.W. Verhaak[1,*]

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, [2]The Eli and Edythe L. Broad Institute of Harvard University and MIT, Cambridge, MA 02142 and [3]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, NY 10015, USA

Associate Editor: Gunnar Ratsch

ABSTRACT

Summary: Technological advances in high-throughput sequencing necessitate improved computational tools for processing and analyzing large-scale datasets in a systematic automated manner. For that purpose, we have developed PRADA (Pipeline for RNA-Sequencing Data Analysis), a flexible, modular and highly scalable software platform that provides many different types of information available by multifaceted analysis starting from raw paired-end RNA-seq data: gene expression levels, quality metrics, detection of unsupervised and supervised fusion transcripts, detection of intragenic fusion variants, homology scores and fusion frame classification. PRADA uses a dual-mapping strategy that increases sensitivity and refines the analytical endpoints. PRADA has been used extensively and successfully in the glioblastoma and renal clear cell projects of The Cancer Genome Atlas program.

Availability and implementation: http://sourceforge.net/projects/prada/

Contact: gadgetz@broadinstitute.org or rverhaak@mdanderson.org

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

Received on June 28, 2013; revised on February 21, 2014; accepted on March 24, 2014

1 INTRODUCTION

Transcriptome sequencing provides insights into the quantity, structure and composition of RNA molecules in a biological sample. Analytical tools for analysis of RNA sequencing data are available (Kim and Salzberg, 2012; McPherson *et al.*, 2011), but those tools generally focus on single end points, such as quantitation of expression levels or identification of fusion transcripts. As the technology becomes more accessible, there is an increased need for computational pipelines that can process large numbers of raw RNA-sequencing datasets quickly, accurately and comprehensively. For that purpose, we have developed PRADA (Pipeline for RNA-Sequencing Data Analysis). PRADA was designed to be modular in the functional sense that different modules output different types of information on the transcripts. It implements resource management structures such as LSF and PBS, allowing quick scale-up for processing of thousands of RNA-seq samples.

2 METHODS

PRADA was designed for processing paired-end sequencing data in fastq, Sequence Alignment/Map (SAM) format or the compressed binary version of SAM (BAM) (Li *et al.*, 2009). The processing module applies an alignment strategy in which reads are mapped to a combined genome and transcriptome reference, allowing reads to align to known transcript sequences, including exon junctions and unannotated mRNAs. The mapping strategy has previously been described in Berger *et al.* (2010). The appropriate reference files are available for download at http://bioinformatics.mdanderson.org/Software/PRADA/. This strategy retrieves all best alignments per read from the dual reference file using BWA (Li and Durbin, 2009). After initial mapping, the alignments of reads that map to multiple locations (both transcriptomic and genomic) are collapsed into single genomic coordinates, including reads that span exon junctions. Once mapped, reads are filtered out if their best placements are not mapped to multiple genomic coordinates. Quality scores are recalibrated using the Genome Analysis Toolkit (GATK) framework (McKenna *et al.*, 2010), index files are generated using Samtools (Li *et al.*, 2009) and duplicate reads are flagged using Picard (http://picard.sourceforge.net/).

For expression and quality control metrics, PRADA's expression module calls the java executable of RNA-SeQC DeLuca *et al.*, 2012). RNA-SeQC is a publicly available tool that produces data quality metrics of three types: mapped read counts, coverage and correlation. The read count metrics include total number of reads, duplicates, uniquely mapped reads and reads per kilobase per million mapped (RPKM). The coverage metrics include GC bias, 3′/5′ bias and mean number of bases per read. Expression correlation is reported when multiple samples are analyzed.

The fusion module aims to detect chimeric transcripts through identification of discordant read pairs and fusion-spanning reads. Discordant read pairs are paired read-ends that map uniquely (i.e. mapping quality equal to 37) to different protein-coding genes with orientation consistent with formation of a sense–sense chimera. Mitochondrial genes and clone IDs are ignored. If a read maps to overlapping genes, the genes are split up as two different instances. Further evidence for transcript fusion is sought through evaluation of putative fusion junction spanning reads. They are detected in PRADA by the construction of a sequence database that holds all possible exon–exon junctions that match the 3′ end of one gene fused to the 5′ end of a second gene. All hypothetical exon junctions are created using the Ensembl transcriptome reference. Then, unmapped reads aligned to the database of hypothetical exon junctions. Only reads of which the mate pair maps to either of the two fusion partner genes are considered. Each fusion is

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

annotated by sample name, 5′ and 3′ gene name, chromosome location and blastn homology scores (see below).

The supervised fusion screen module *G*eneral *U*ser d*E*fined *S*upervised *S*earch (GUESS) was developed to facilitate rapid detection of a single fusion, e.g. *FGFR3-TACC3* in GBM (Singh *et al.*, 2012). GUESS screens BAM files for the presence of discordant read pairs and fusion-spanning reads of specific genes defined by the user. We have developed two variants of GUESS, one that searches for *f*usion *t*ranscripts involving two given genes (GUESS-ft) and one that searches for *i*ntragenic *f*usions (GUESS-if), such as the *EGFR* vIII variant that deletes exons 2–7.

To allow filtering of homology artifacts from the results of the fusion module and GUESS-ft, the similarity of two fusion partner genes is assessed using BlastN. Metrics provided are bitscore and its associated *E*-value, where an *E*-value of >0.001 is considered to be non-homologous.

The frame module predicts whether a fusion transcript is in frame and thereby capable of producing a functional protein, based on the combinatorics of the transcript(s) in the Ensembl database for the genes involved.

3 RESULTS

3.1 The Cancer Genome Atlas unsupervised fusion results

We used PRADA to process RNA-seq data from 416 renal clear cell carcinoma (ccRCC) samples and 164 glioblastoma multiforme (GBM) samples from The Cancer Genome Atlas (TCGA). Among 84 predicted gene fusions in 416 ccRCCs were 5 *SFPQ-TFE3* transcripts, and the overall validation rate was 85% (Cancer Genome Atlas Research Network, 2013). Fusions found in 164 GBMs ($n = 229$) included recurrent rearrangements such as the previously reported *FGFR3-TACC3* in 2 samples and *EGFR*-associated fusions in 11 samples (Zheng *et al*., 2013). Data from whole genome sequencing, available for a subset of the GBM, validated 41 of 49 predicted fusions (84%). A *TFG-GPR128* fusion was observed in both renal and GBM samples.

3.2 Supervised detection of TFG-GPR128

A germ line copy number variant involving *TFG* and *GPR128* has been described in human population cohorts (Jakobsson *et al*., 2008). Using the GUESS-ft supervised fusion search module, we evaluated the presence of *TFG-GPR128* fusions in 321 TCGA tumor-adjacent normal tissues from 11 cancer types (Supplementary Table S1). *TFG-GPR128* fusion was detected at low levels in 3 of the 321 normal samples (Supplementary Table S1). The matching tumor sample of two of three *TFG-GPR128* harboring normals also expressed this fusion construct, corroborating its germ line status.

3.3 Correlation of RPKM values with U133A microarray expression levels

We tested the RPKM functionality of PRADA's expression module in the context of subtype classification using 164 RNA-seq samples from GBM, comparing its subtype stratification with that based on U133A array data. The comparison showed a high (80.9%) concordance rate in subtype calls for expression data generated by the two platforms (Supplementary Table S2), a similar percentage classified reliably as previously reported (Verhaak *et al*., 2010).

3.4 Comparison of fusion transcript detection by PRADA, Defuse and Tophat-fusion

To evaluate PRADA fusion detection accuracy, we obtained RNA-seq data and whole genome sequencing data of the U87 glioma cell line. PRADA detected 11 fusions, 6 of which related to DNA structural rearrangements, TopHat-fusion (Kim and Salzberg, 2012) predicted 42 fusions of which 12 validated in DNA, while Defuse (McPherson *et al*., 2011) found 51 fusions of which 12 related to DNA lesions (Supplementary Text and Supplementary Table S3).

4 DISCUSSION

The power of PRADA is based on (i) its scalability, (ii) its mapping to both transcriptomic and genome, a distinctive feature of PRADA in comparison with other RNA analysis tools such as Tophat-fusion and Defuse, which rely on alignments of partial reads to identify gene fusions, (iii) its modularity and (iv) its comprehensive repertoire of output information from the incorporated modules. It enables the user to compute multiple analytical metrics using one software package and to do so for large numbers of samples at once in a fully automated fashion. It has been tested on thousands of RNA-seq samples from a wide variety of tumor types and normal tissues in TCGA. PRADA is designed to be run out of the box with little configuration, and is compatible with PBS and LSF compute clusters. A single PRADA tarball, including binaries of the packages it relies on, a comprehensive and detailed manual, and test FASTQ/BAM files, are freely available at http://sourceforge.net/projects/prada/ and through Galaxy at http://toolshed.g2.bx.psu.edu/view/siyuan/prada.

Funding: The content is solely the responsibility of the authors and does not necessarily represent NCI/NIH. Supported in part by NCI grant number CA143883/Chapman Foundation/Dell Foundation.

Conflict of Interest: none declared.

REFERENCES

Berger,M.F. *et al.* (2010) Integrative analysis of the melanoma transcriptome. *Genome Res.* **20**, 413–427.

Cancer Genome Atlas Research Network. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**, 43–49.

DeLuca,D.S. *et al.* (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, **28**, 1530–1532.

Jakobsson,M. *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.

Kim,D. and Salzberg,S.L. (2012) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.*, **12**, R72.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

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

Singh,D. *et al.* (2012) Transforming fusions of FGFR and TACC genes in human glioblastoma. *Science*, **377**, 1231–1235.

Verhaak,R.G.W. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–110.

Zheng,S. *et al.* (2013) A survey of intragenic breakpoints in glioblastoma identifies a distinct subset associated with poor survival. *Genes Dev.*, **27**, 1462–1472.