BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 18 2013, pages 2353–2354
doi:10.1093/bioinformatics/btt385

Sequence analysis

Advance Access publication July 3, 2013

MMuFLR: missense mutation and frameshift location reporter

Susan K. Rathe[1,*], James E. Johnson[2,*], Kevin A.T. Silverstein[2], Jesse J. Erdmann[2], Adrienne L. Watson[1,3,4], Flavia E. Popescu[5], John R. Ohlfest[5] and David A. Largaespada[1,3,5]

[1]Masonic Cancer Center, [2]Supercomputing Institute for Advanced Computation Research, [3]Department of Genetics, Cell Biology and Development, [4]Brain Tumor Program and [5]Department of Pediatrics, University of Minnesota, Minneapolis, MN 55455, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Cancer researchers seeking immunotherapy targets in cancer cells need tools to locate highly expressed proteins unique to cancer cells. Missense mutation and frameshift location reporter (MMuFLR), a Galaxy-based workflow, analyzes next-generation sequencing paired read RNA-seq output to reliably identify small frameshift mutations and missense mutations in highly expressed protein-coding genes. MMuFLR ignores known SNPs, low quality reads and poly-A/T sequences. For each frameshift and missense mutation identified, MMuFLR provides the location and sequence of the amino acid substitutions in the novel protein candidates for direct input into epitope evaluation tools.

Availability: http://toolshed.g2.bx.psu.edu/

Contact: rath0096@umn.edu or johns198@umn.edu

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

Received on March 12, 2013; revised on June 14, 2013; accepted on July 1, 2013

1 INTRODUCTION

Next-generation sequencing (NGS) holds promise for individualized targeted treatment, particularly in the realm of cancer immunotherapy (Meinke *et al*., 2005). Identifying aberrant transcripts from a patient's tumor cells may provide candidate target epitopes, which can be used to generate patient-specific vaccines. Given the quantity of data generated by NGS, it is important to have an efficient and repeatable process to deconstruct the data and present the most relevant vaccine targets, which includes missense and small frameshift mutations located in protein-coding regions of highly expressed genes. We offer missense mutation and frameshift location reporter (MMuFLR) as a process to analyze NGS RNA-seq paired read output to identify a limited set of peptides that may be targets for immunotherapy. MMuFLR is implemented as a workflow in Galaxy (Blankenberg *et al*., 2010). Galaxy is a server-based workflow system designed to run a series of applications, presented to the user as tools, in an easy-to-use repeatable manner. MMuFLR uses a few custom-developed Galaxy tools and those in the standard distribution of Galaxy. In this pilot project, MMuFLR's ability to identify potential vaccine targets was tested in RNA-seq datasets derived from two benign human meningiomas. However, the use of MMuFLR is not limited to cancer immunotherapy. It can also be used to locate prominent mutations responsible for cancer development and for other maladies, such as genetic diseases, developmental defects and drug resistance.

*To whom correspondence should be addressed.

2 WORKFLOW

2.1 Map read sequences to reference genome

MMuFLR starts with the NGS RNA-seq paired read output in the form of two fastq-formatted files. Each read pair is individually mapped to the reference genome using TopHat (Langmead *et al*., 2009), resulting in a bam formatted file. The standard Galaxy distribution includes TopHat in its toolset. The custom Galaxy tool 'Filter SAM or BAM' invokes the SAMtools (Li *et al*., 2009) view command to filter TopHat bam output by the MAPQ score.

2.2 Refine the mapping

TopHat, as is typical with short read mappers, may find multiple mappings for individual read pairs, which are equally valid. The mapped reads can be analyzed in aggregate to improve upon the initial mapping of individual reads. The MMuFLR workflow uses software from the genome analysis toolkit (GATK) (DePristo *et al*., 2011; McKenna *et al*., 2010) along with Picard-tools (http://picard.sourceforge.net) to perform this collective analysis and realign the individual reads. GATK requires bam input to be labeled with read group information and be sorted in reference order. Several Picard-tools ('Add or Replace Groups', 'Paired Read Mate Fixer', 'Mark Duplicate Reads' and 'Reorder SAM/BAM') add the read groups, filter out unmapped pairs, remove exact duplicate reads and sort the resulting bam output into reference order. Two GATK tools, 'Realigner Target Creator' and 'Indel Realigner', realign the reads based on aggregate information from all reads. Duplicate reads are again identified and removed with 'Mark Duplicate Reads'.

2.3 Multiple alignment

The realigned reads are then assembled into a multiple alignment using the Galaxy tool 'MPileup', which invokes SAMtools MPileup. The custom Galaxy tool, 'Pileup to VCF', filters the pileup results and converts them to the standard variant call format (VCF). Filter parameters to the 'Pileup to VCF' tool

include minimum base quality, minimum coverage depth and minimum frequency of a specific allele.

2.4 Eliminate common variants

The resulting VCF file will likely contain many commonly known variants, which would not be useful in determining candidate targets. MMuFLR uses the 'annotate' command of SnpSift (Cingolani *et al*., 2012) to merge variants from the samples being evaluated with known variants from the dbSNP database (Sherry *et al*., 2001), as well as a user-provided list. Then the 'filter' command of SnpSift is used to eliminate those annotated variant entries.

2.5 Determine variant effects

MMuFLR next categorizes the effect of each variant using SnpEff (Cingolani *et al*., 2012). The 'filter' command of SnpSift is used to separate the missense and frameshift variants into individual files for reporting.

2.6 Report variants

Finally, a new Galaxy tool, 'SnpEff Ensembl CDS', inserts the variants into transcripts retrieved from Ensembl (Flicek *et al*., 2013) and reports the amino acid sequence of the abnormal peptides. Included in the report are the sequencing depth of reads at the mutation site and the prevalence of the mutation. A parameter can be set to ignore poly-A/T sequences of specified lengths.

3 RESULTS

MMuFLR was executed for two human meningioma RNA-seq datasets, MG1 and MG2. MMuFLR settings restricted the selection of candidate genes with minimum coverage of five reads, with a prevalence >0.32 and a Phred quality score of 20 or more. MMuFLR also ignored poly A/T mononucleotide sequences of ≥5. MMuFLR identified 16 frameshifts and 576 missense mutations in the MG1 sample, and 10 frameshifts and 509 missense mutations in the MG2 sample. The candidates were evaluated for their presence in other samples. Sanger sequencing of nine prominent candidates, not found in other cancer samples, verified the existence of the mutations within the tumor samples, with the MMuFLR prevalence accurately reflecting the relative abundance of the mutation within the total transcripts present (Table 1). Because there were no normal samples available for comparison, these mutations may represent rare germline variants.

Table 1. Frameshift and missense mutations identified by MMuFLR and verified using Sanger sequencing

Sample	Gene	Variant position	Reference	Variant	Prevalence	Depth
Frameshift mutations						
MG1	ZNF812	chr19:9800968	CT	T	0.50	20
MG2	CPNE1	chr20:34215234	G	TG	0.45	98
	MRPL16	chr11:59577372	C	GC	0.37	27
	ZP3	chr7:76071183	T	TG	0.50	14
	DNAJB7	chr22:41257834	A	TA	1.00	6
Missense mutations						
MG1	COL6A2	chr21:47545737	A	G	0.50	899
MG2	APP	chr21:27484329	T	A	0.46	432
	FRMPD2	chr10:49389016	A	G	0.48	227
	CCDC74A	chr2:132288362	T	C	0.97	29

Note: Included in the table are the location of the variant, the reference sequence being replaced, the fraction of reads with the mutation (prevalence) and the number of reads (depth).

4 CONCLUSIONS

One challenge for developing patient-specific cancer vaccines is identifying novel peptides specific to the tumor cells. MMuFLR is an innovative Galaxy-based software tool designed to locate highly abundant transcripts with frameshift or missense mutations within protein-coding sequences. MMuFLR ignores potentially false candidates such as frameshifts located within poly A/T sequences of mononucleotides, which is believed to result from a loss of fidelity during replication due to dissociation within the reverse transcriptase binding site in a phenomenon referred to as 'stuttering'. Because there is always a lag between the identification of SNPs and their incorporation into the dbSNP database, MMuFLR provides the ability to establish a supplementary SNP table in which the researcher can record and thereby ignore SNPs identified during their research efforts. MMuFLR has been successfully tested with both human and mouse samples, but should be easily adaptable to other organisms.

ACKNOWLEDGEMENTS

We thank the BioMedical Genomics Center for providing RNA sequencing, oligo preparation and Sanger sequencing, and the Minnesota Supercomputing Institute, which maintains the Galaxy software, and provides data management services and training.

Funding: This work was funded by The Children's Cancer Research Fund; Children's Tumor Foundation Young Investigator Award Grant 2011-01-018 (to A.L.W.).

Conflict of Interest: Dr. Largaespada is co-owner and advisor to NeoClone Biotechnology, Inc. and Discovery Genomics, Inc. No resources or personnel from either company were involved in this research in any way.

REFERENCES

Blankenberg,D. *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.*, **Chapter 19**, Unit 19.10.1–21.

Cingolani,P. *et al.* (2012) Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.*, **3**, 35.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Flicek,P. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.

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

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.

Meinke,A. *et al.* (2005) Antigenome technology: a novel approach for the selection of bacterial vaccine candidate antigens. *Vaccine*, **23**, 2035–2041.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.