Methods 83 (2015) 118–127

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer

Timothy D. O'Brien [a,b], Peilin Jia [b], Junfeng Xia [b], Uma Saxena [c], Hailing Jin [d], Huy Vuong [b], Pora Kim [b], Qingguo Wang [b], Martin J. Aryee [c], Mari Mino-Kenudson [c], Jeffrey A. Engelman [e], Long P. Le [c], A. John Iafrate [c], Rebecca S. Heist [e], William Pao [d], Zhongming Zhao [b,f,g,*]

[a] Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, TN 37232, United States
[b] Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37203, United States
[c] Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, United States
[d] Department of Medicine/Division of Hematology-Oncology, Vanderbilt University School of Medicine, Nashville, TN 37232, United States
[e] Department of Medicine, Division of Hematology and Oncology, Massachusetts General Hospital, Boston, MA 02114, United States
[f] Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37232, United States
[g] Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, United States

ARTICLE INFO

Article history:
Received 21 February 2015
Received in revised form 16 April 2015
Accepted 16 April 2015
Available online 23 April 2015

Keywords:
Single nucleotide variants
Whole exome sequencing
RNA-Seq
Somatic mutations
Allele frequency
RNA editing

ABSTRACT

Whole exome sequencing (WES) and RNA sequencing (RNA-Seq) are two main platforms used for next-generation sequencing (NGS). While WES is primarily for DNA variant discovery and RNA-Seq is mainly for measurement of gene expression, both can be used for detection of genetic variants, especially single nucleotide variants (SNVs). How consistently variants can be detected from WES and RNA-Seq has not been systematically evaluated. In this study, we examined the technical and biological inconsistencies in SNV detection using WES and RNA-Seq data from 27 pairs of tumor and matched normal samples. We analyzed SNVs in three categories: WES unique – those only detected in WES, RNA-Seq unique – those only detected in RNA-Seq, and shared – those detected in both. We found a small overlap (average ~14%) between the SNVs called in WES and RNA-Seq. The WES unique SNVs were mainly due to low coverage, low expression, or their location on the non-transcribed strand in RNA-Seq data, while the RNA-Seq unique SNVs were primarily due to their location out of the WES-capture boundary regions (accounting ~71%), as well as low coverage of the regions, low coverage of the mutant alleles or RNA-editing. The shared SNVs had high locus-specific coverage in both WES and RNA-Seq and high gene expression levels. Additionally, WES unique and RNA-Seq unique SNVs showed different nucleotide substitution patterns, e.g., ~55% of RNA-Seq unique variants were A:T → G:C, a hallmark of RNA editing. This study provides an important evaluation on the inconsistencies of somatic SNVs called in WES and RNA-Seq data.

1. Introduction

Single nucleotide variants (SNVs) are the most abundant form of genetic variation in genome sequences and somatic SNVs play critical roles in disease [1]. The discovery of many driver SNVs has led to new targets for therapeutic treatments and preventive measures. Examples include vemurafenib for the BRAF V600 mutations in melanoma [2,3] and gefitinib, erlotonib, and afatinib for EGFR mutations in lung cancer [4]. The recent advances in next-generation sequencing (NGS) technologies, especially whole exome sequencing (WES) and whole transcriptome sequencing (RNA-Seq), have helped investigators generate a massive amount of NGS data, from which genetic variants, including SNVs, are

* Corresponding author at: Department of Biomedical Informatics, Vanderbilt University School of Medicine, 2525 West End Avenue, Suite 600, Nashville, TN 37203, USA. Fax: +1 615 936 8545.

E-mail addresses: timothy.obrien@vanderbilt.edu (T.D. O'Brien), peilin.jia@vanderbilt.edu (P. Jia), jfxia@ahu.edu.cn (J. Xia), stqa8350@gmail.com (U. Saxena), hailing.jin@Vanderbilt.Edu (H. Jin), huy.vuong@gmail.com (H. Vuong), pora.kim@vanderbilt.edu (P. Kim), josephw10000@gmail.com (Q. Wang), Aryee.Martin@mgh.harvard.edu (M.J. Aryee), MMINOKENUDSON@partners.org (M. Mino-Kenudson), JENGELMAN@partners.org (J.A. Engelman), LPLE@partners.org (L.P. Le), AIAFRATE@partners.org (A.J. Iafrate), RHEIST@partners.org (R.S. Heist), william.pao@Vanderbilt.Edu (W. Pao), zhongming.zhao@vanderbilt.edu (Z. Zhao).

http://dx.doi.org/10.1016/j.ymeth.2015.04.016
1046-2023/

detected. Many tools are now available for the detection of somatic SNVs from NGS data [5].

Both whole genome sequencing (WGS) and WES have been applied to detect SNVs in large scale cancer studies. While WGS can detect the full spectrum of variants (SNVs, insertions/deletions (indels), copy number variations (CNVs), and structural variants (SVs) across the whole cancer genome, WES is more cost-effective in detecting SNVs and indels located in the 1–2% of the genome that encodes for functional proteins [6]. There is good evidence that SNVs within the exome are responsible for many diseases, so WES has been applied extensively in research and clinically [6–8]. RNA-Seq is commonly used for the measurement of gene expression levels, detection of gene fusions, and identification of splicing events. Because RNA-Seq is based on direct sequencing of cDNA, the product of the mRNA through reverse transcription, it is practically feasible to detect SNVs from RNA-Seq data [9,10]. This is a unique feature that is different from the traditional microarray-based gene expression. RNA-Seq also has the ability to detect RNA editing, which is a post-transcriptional process that modifies RNA transcripts. One of the most common mechanisms of RNA editing is the deamination of adenosine to inosine by the protein Adenosine Deaminase Acting on RNA (ADAR). The inosine is interpreted in a similar way to guanosine and, thus, results in an adenosine to guanine (A → G) change [11].

RNA-Seq has been extensively applied to genomic and transcriptomic studies, including cancer. For example, a large-scale RNA-Seq study of lung adenocarcinoma identified several cancer driver genes [12], indicating its utility in a transcriptome analysis of cancer samples. This study demonstrated that in addition to identifying fusion genes and differential gene expression, RNA-Seq could detect well-known cancer driver genes. RNA-Seq has also been combined with WGS to better understand the mutational landscape of lung cancer [13,14]. These studies, in addition to showing the standard applications of RNA-Seq in gene expression analysis, highlight its usefulness as a technology platform for SNV detection, though challenges remain [15]. Large consortia such as The Cancer Genome Atlas (TCGA), have applied both WES and RNA-Seq, as well as other platforms, to comprehensively catalog the cancer genome landscape [16]. The combined WES and RNA-Seq of the same tumor samples allow for large-scale examinations of somatic mutations in both the DNA and RNA. By applying these two types of technology together, one can improve the detection of various mutations, including those in the expressed genes with different splicing and expression levels, and those in non-transcribed regions. However, sequencing the same tumor using both platforms is rarely used in real projects due to the cost and analysis issues.

A detailed comparison of SNVs called from WES and RNA-Seq data using the same samples can not only reveal the technical differences of these two technologies, but also help us better understand the underlying biological processes that lead to the ambiguous observations of SNVs at the DNA and RNA levels, respectively. Such a comparison can provide guidance on the utility of WES and RNA-Seq in SNV detection. So far, there have been only a few attempts to unveil the advantages and disadvantages of WES and RNA-Seq in SNV detection. For example, Cirulli et al. [17] recently compared WGS with RNA-Seq in detecting SNVs using peripheral blood mononuclear cells from the same subjects. They highlighted many important aspects for SNV detection such as expression levels and read depth, but its conclusions are yet to be validated due to the limited sample size. Another recent review compared WES and RNA-Seq [18], but it only discussed several global features without a systematic comparison of many detailed features.

In this study, we compared the features of SNVs from WES and RNA-Seq using a collection of 27 lung tumor and matched normal samples from the same patients. Through our systematic analyses, we attempted to unveil the unique features of SNVs from each platform and determined why variants are missed between these platforms. Because of the high false calling rate of indels, we only focused on SNVs. We observed only a small overlap of SNVs between WES and RNA-Seq, and identified multiple technological and biological reasons leading to discrepancies in SNV calling.

2. Materials and methods

2.1. Samples and sequencing

Twenty-seven paired tumor and normal lung cancer samples from patients undergoing lung cancer surgery at Massachusetts General Hospital were used for this analysis. For all 27 paired tumor and normal lung cancer samples, we performed both WES and RNA-Seq experiments. All participants provided written informed consent. Tumor content was assessed with an average of 60% across samples. The exome regions were captured using the Agilent SureSelect Human All Exon kit and then sequenced on an Illumina HiSeq 2000 platform (paired end, 100 bp) in a MGH core. We obtained a total of 3,677,811,274 paired-end reads with an average sequencing depth of 121×. For RNA-Seq, Illumina Tru-Seq v2 RNA-Seq kit was used for enrichment of mRNA, cDNA synthesis, and library construction. Then, RNA sequencing was performed on an Illumina HiSeq 2000 platform in the Vanderbilt Technologies for Advanced Genomics (VANTAGE) core (paired end, 100 bp). We obtained a total of 4,778,766,598 paired end reads with an average of 88,495,678 paired end reads per sample. We used FASTQC to check the quality of reads of all samples (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

2.2. WES data analysis

We mapped the WES reads to the human reference genome hg19 (GRCh37) using BWA (version 0.5.9c) [19]. In order to further process the data, we used Picard (version 1.95) [20] to mark duplicate reads and used GATK (version 1.0.3825) to perform local realignment and recalibration [21,22]. After post-alignment processing of the data, we called SNVs with MuTect (version 1.1.4). To generate mpileup files for each tumor and normal sample, we used the "mpileup" function in Samtools (version 0.1.19) [23]. Read count values were obtained from the mpileup files using VarScan2 (version 2.3.5) [24] with the "readcounts" function. Read count values were split up into categories of values: not covered (NA), single read (1), low coverage (2–7) and high coverage ($\geqslant 8$).

2.3. RNA-Seq data analysis

We used TopHat2 (version 2.0.0) [25] to map RNA-Seq reads to the human reference transcriptome and genome (hg19). TopHat2 firstly attempts to map reads to the reference transcriptome and then for the unmapped reads from the initial transcriptome, it attempts to map them to the human genome reference. As we did for WES data, we called SNVs using MuTect (version 1.1.4). Specifically, we generated mpileup files using Samtools and obtained read count values using VarScan2. We used Cufflinks (version 2.1.1) [26] to obtain gene-based FPKM (fragments per kilobase of exon per million fragments mapped) values for all samples. FPKM values corresponding to degrees of expression were as follows: not covered (NA), no expression (FPKM < 1), very low expression (FPKM 1–5), low to moderate expression (FPKM 5–20), and high expression (FPKM > 20).