German Medical Data Sciences: A Learning Healthcare System
U. Hübner et al. (Eds.)


doi:10.3233/978-1-61499-896-9-217

# Using RNA-Seq Data for the Detection of a Panel of Clinically Relevant Mutations

*Alexander WOLFF[a], Júlia PERERA-BEL[a], Hans-Ulrich SCHILDHAUS[c], Kia HOMAYOUNFAR[d], Bawarjan SCHATLO[e], Annalen BLECKMANN[a,b] and Tim BEISSBARTH[a,1]*

[a]*Department of Medical Statistics, University Medical Center Göttingen*
[b]*Department of Hematology and Oncology, University Medical Center Göttingen*
[c]*Department of Pathology, University Medical Center Göttingen*
[d] *Department of General, Visceral and Pediatric Surgery, University Medical Center Göttingen*
[e] *Department of Neurosurgery, University Medical Center Göttingen*

**Abstract.** Somatic single nucleotide variants (SNVs) are genomic events with increasing implications in cancer treatment. The clinical standard for SNVs detection is whole genome/exome sequencing (WGS/WES) in matched tumor-normal samples. Yet, this is a very costly approach both economically and biologically and very often only tumor samples are sequenced. On the other hand, RNA sequencing (RNA-Seq) is the most popular technology to study gene expression, and has also the potential for a cost-effective identification of SNVs as an alternative to tumor-only WES. Here we present a method for the identification of SNVs in tumor-only RNA-Seq data putting a special focus on a small panel of clinically relevant SNVs. For evaluation purposes, we analyzed matched tumor-normal WES and tumor-only RNA-Seq data from 14 cancer patients. We compared SNVs detected in i) RNA-Seq by our method, ii) WES tumor-only by Mutect2 and iii) WES matched tumor-normal by Mutect2. We did a detailed evaluation for a reduced panel of clinically relevant SNVs and reliably identified in RNA-Seq data a subset of mutations for which we had pathological annotation. Hence, RNA-Seq rises as a cost-effective option to detect in parallel gene expression as well as a small panel of clinically relevant SNVs in research.

**Keywords:** RNA-Seq, SNVs, Mutect2, variant calling, GATK

## 1. Introduction

Somatic single nucleotide variants (SNVs) are genomic events known to drive cancer. Whole genome and exome sequencing (WGS, WES) in matched tumor-normal samples are the clinical standard for detecting somatic SNVs. There are many tools for identifying SNVs on WGS or WES data, thoroughly compared in different contexts [1-3]. According to these studies, two tools outperform the rest: Mutect [4] and Varscan2 [5]. The first performs better at identifying SNVs with low allele frequencies, whereas the latter detects the highest number of SNVs and outperforms any tool at positions with high coverage. On the other hand, RNA sequencing (RNA-Seq) has become the most popular technology -after replacing microarrays- to study gene expression. Unlike microarrays, RNA-Seq can easily be used to detect alternative splicing, RNA editing, fusion genes, other RNA species, and, potentially, SNVs. Calling somatic SNVs in RNA-Seq data has been done in some studies by applying tools specific for WES/WGS data [6-8]. Besides

[1] Corresponding Author, Tim Beißbarth, University Medical Center Göttingen, Humboldtallee 32 D-37073 Göttingen, Germany; E-mail: Tim.Beissbarth@ams.med.uni-goettingen.de.

obvious false negatives produced in regions with low or no expression, these studies reported false positive SNV calls in RNA-Seq data mainly due to: PCR cycle bias, strand bias, RNA editing and difficulty to align the transcriptome to the reference genome due to splicing. Sheng and colleagues tried to address some of these issues both on DNA and RNA-Seq [9]. An added problem is the fact that clinical samples are usually limited to tumor-only profiling. Detection of somatic SNVs in WES tumor-only samples is challenging and has been addressed with machine learning approaches [10] or the use of whitelists and blacklists as in Mutect [4]. However, the same has not yet been attempted for RNA-Seq data. All in all, its cheaper cost compared to WES/WGS together with all its possible applications makes RNA-Seq a technology with high interest for clinical use (e.g. parallel detection of SNVs and functional activation of genes). It seems worthwhile developing a method to call SNVs in RNA-Seq data optimized for a panel of well-known SNVs.In this study we present a method to call SNVs in RNA-Seq tumor-only samples. We assess its performance putting special focus on optimizing the method for a panel of known SNVs with high clinical interest. We compare our method's performance on a matched dataset comprising RNA-Seq and WES data. We chose Mutect2 to detect SNVs in WES data.We compare the SNVs detected in RNA-Seq data by our method to tumor-only and tumor-normal results by Mutect2.

## 2. Materials and Methods

### *2.1. Databases*

The panel of known SNVs with high clinical interest was based on the Clinical Interpretation of Variants in Cancer (CIViC, version from 01/06/2017) [11] and Cancer Genome Interpreter (CGI, last updated 02/08/2017) [12]. In both cases, we filtered for SNVs predictive of drug response. Genomic coordinates were transformed from hg19 to hg38 built using the rtracklayer R package. Both databases were merged by aggregating duplicate entries. The panel of actionable variants contains information on 442 variants in 92 genes.

### *2.2. Collection of Patient Samples*

Tissue samples were collected by the surgery departments of the University Medical Center Göttingen. The collected tissues are from seven metastatic brain and seven metastatic liver tumours with origin from either colorectal or breast cancer. Fresh frozen tissue samples for WES and RNA-Seq were separated. From these 14 patients we collected EDTA-blood for WES as well. EDTA-blood samples served as control samples to differentiate between germline and somatic mutations. In total 42 samples were sequenced, 3 samples per patient. The study is approved by the Ethics Committee of the University Medical Centre Göttingen, application number 21/3/11 and 14/10/05.

### *2.3. Data Preprocessing and Analysis*

WES and RNA-Seq reads were quality assessed using fastqc. All WES brain samples showed high duplication levels and a drop in quality at the end of the reads due to high levels of contamination with nextera adapters. Hence, TrimGalore-0.4.3 was applied to all brain samples. Between 11% and 38.4% of base pairs were trimmed. WES paired-end reads were aligned against the reference genome of Homo sapiens version GRCh38 with bowtie2 (version 2.2.3). Samtools was used to create bam files, and Picard (version

2.0.1) to mark duplicates. Then, GATK (v3.8.0) best practices were followed to perform read realignment (IndelRealigner) and base recalibration (BaseRecalibrator). RNA-Seq single-end reads were aligned against the reference genome of Homo sapiens Ensembl Version GRCh38.91 with the splice-aware aligner STAR (v2.5.2b). Picard (version 2.0.1) was used to remove duplicates. We used Mutect2 (GATK v3.8.0, beta version) to detect somatic variants in WES data using matched tumor-normal samples (referred to as "clinical standard" in the text) as well as only tumor samples ("tumor-only" mode). Cosmic (version 83, Coding and Non Coding vcf files) and dbSNP (version 138) were provided as input to Mutect2 to adjust the threshold for evidence of a variant in the normal sample. To confirm germline mutations detected by tumor-only samples, GATK Haplotypecaller (v3.8.0) with dbSNP (version 138) was used.

## 3. Results

### *3.1. Detection of SNVs in RNA-Seq Data*

We implemented a variation of pileuping nucleotide bases at each position in the transcriptome, using mpileup from samtools. In case of a transcriptome-wide analysis, mpileup of all base positions in the RNA-Seq data is performed. Afterwards, the distribution of bases in each position is annotated with supplemental database information (CIVIC, ClinVAR, Cosmic). In case the -panelmode flag is selected, mpilup is called only for the 442 curated SNVs. A minimum of 3 reads or 10% of the reads supporting the alternative variant are used as default thresholds. The output comprises the distribution of bases at each position, the decision whether the position contains an SNV, and clinical annotations from CIViC and CGI.

### *3.2. Comparison of WES and RNA-Seq Data in Detecting SNVs*

We generated WES matched tumor-normal and RNA-Seq tumor-only data from 14 cancer patients. We applied a standard pipeline to detect somatic SNVs in WES matched tumor-normal samples, referred to as clinical standard. We also applied a tumor-only mode to WES tumor samples, referred to as tumor-only. Finally, we applied our method to detect SNVs in RNA-Seq tumor samples. SNVs detection was focused on 442 cancer-specific variants with clinical interest. As shown in Figure 1A we found 109 SNVs in all samples by the three methods: 10 in the gold standard, 104 in tumor-only and 73 in the RNA-Seq (average of 0.7, 7.4 and 5.2 SNVs/sample, respectively). The 10 SNVs detected by the clinical standard were also detected in WES tumor-only and RNA-Seq. We found a higher overlap between WES tumor-only and RNA-Seq (68 SNVs) than between the two methods on WES (10 SNVs). This finding is explained by the fact that our panel includes germline mutations and polymorphisms; the clinical standard is optimized to reliably detect mutations only present in the tumor sample by filtering out any mutation present in the normal sample. Accordingly, the clinical standard only detected SNVs known to be somatic. Nonetheless, all somatic SNVs were also detected by tumor-only and RNA-Seq (Figure 1B).

The variants uniquely detected by WES tumor-only (36 SNVs) could be explained in the majority of the cases due to low expression in the RNA-Seq data. The only exception presenting high expression was the *MGMT* promoter SNP rs16906252. Yet, this SNP is known to be associated with low MGMT expression, leading to allele specific expression [13]. On the other hand, only 5 SNVs were exclusively detected by RNA-

Seq. Two of them - *TPMT* Y240C and *TPMT* A154T - are known to be an haplotype of the TPMT enzyme (TPMT*3A) [14] and were indeed found in the same patient (BM4). These haplotype was not confirmed by WES tumor-only due to high duplication levels, which did not pass Mutect2 filters. The other three mutations (*ETS2* mutation in patient BM1, *NQO1* in patient BM7 and *XRCC1* mutation in patient LM3) were not found in WES tumor-only also due to the same reason. As a matter of fact, these 5 germline polymorphisms detected exclusively in RNA-Seq data could be confirmed by a germline SNV caller (Haplotypecaller) in normal samples. Last but not least, 7 out of the 8 pathologically validated mutations in *BRAF*, *KRAS*, *NRAS* and *PIK3CA* were consistently detected by the three methods (Figure 1). *PIK3CA* E545K mutation in patient LM5 was not detected by any method.

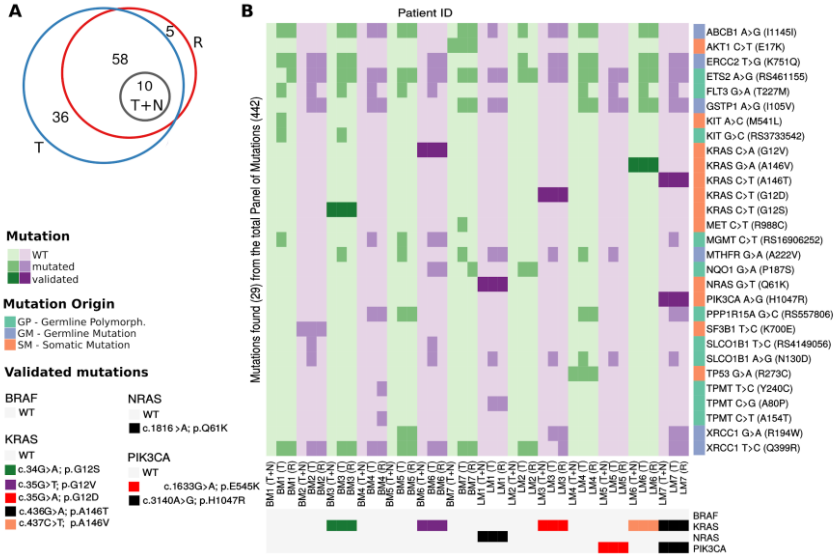


**Figure 1:** A) Venn diagram depicting the number of SNVs identified by each method across all samples (T+N: Mutect2, Tumor+Normal samples, T: Mutect2, Tumor samples, R: Wileup RNA-Seq). B) Heatmap visualization of 29 unique SNVs which were found by at least one of the methods in any of the 14 patients. Wild Type (WT) mutations are shown light green and purple, mutations found by the methods are in green and purple, mutations agreeing with the pathological annotation (validated) are marked in dark green and purple. The details of the pathological mutations are described in the annotation bars at the bottom of the figure. The origin of the mutation is annotated in the bar at the right sight of the heatmap.

## 4. Discussion

We showed a high overlap between RNA-Seq and tumor-only WES. Previous studies reported high numbers of false positives in RNA-Seq data, however, by using a whitelist of well-defined SNVs we avoid this problem. In this setting, detecting SNVs in RNA-Seq data is a comparable approach to WES tumor-only; yet, in RNA-Seq it is regarded as an extra analysis which can be quickly performed (average of 11-15 min/sample) for no extra cost. More important, RNA-Seq appears to be a reliable approach for detecting the selected panel of clinically relevant SNVs, as confirmed by the pathologically validated data (for somatic variants) and by the analysis of normal WES samples (for germline variants). Of course, the user has to accept false negatives in non expressed

genes, but that is inherent to RNA-Seq data. For future implementations, it would be important to consider RNA editing processes as well as including indels in the analysis.

## 5. Conclusion

We showed that RNA-Seq is a cost-effective option to detect a curated list of SNV in parallel to gene expression. Although WES will remain to be the clinical standard, the method presented here can become an alternative when WES is not available.

## 6. Conflict of Interest

The authors declare no conflict of interest.

## 7. Acknowledgements

This work was performed as part of the collaborative project *MetastaSys* (0316173) with input from the projects *Genoperspektiv* (01GP1402) and *MyPathSem* (031L0024) funded by the German Ministry of Education and Research (BMBF). The Transcriptome and Genome Analysis Laboratory (TAL) of the University Medical Center Göttingen was responsible for RNA and Exome sequencing. Tabea Hugo for preparing the tumor tissue samples and Meike Schaffinski for preparing the blood samples.

## References

[1] T.S. Alioto, I. Buchalterm S. Derdak et al., A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing, *Nat. Commun.* **6** (2015) 10001. doi:10.1038/ncomms10001.
[2] L. Cai, W. Yuan, Z. Zhang et al., In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data, Scientific Reports. **6** (2016) 36540. doi:10.1038/srep36540.
[3] K. Cibulskis, M.S. Lawrence, S.L. Carter et al., Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, Nat Biotechnol. **31** (2013) 213–219. doi:10.1038/nbt.2514.
[4] D.C. Koboldt, Q. Zhang, D.E. Larson et al., VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing, Genome Res. **22** (2012) 568–576. doi:10.1101/gr.129684.111.
[5] X. Xu, K. Zhu, F. Liu et al., Identification of somatic mutations in human prostate cancer by RNA-Seq, Gene. **519** (2013) 343–347. doi:10.1016/j.gene.2013.01.046.
[6] R. Piskol, G. Ramaswami, J.B. Li et al., Reliable Identification of Genomic Variants from RNA-Seq Data, Am J Hum Genet. **93** (2013) 641–651. doi:10.1016/j.ajhg.2013.08.008.
[7] T.D. O'Brien, P. Jia, J. Xia et al., Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer, Methods. **83** (2015) 118–127. doi:10.1016/j.ymeth.2015.04.016.
[8] Q. Sheng, S. Zhao, C.I. Li et al, Practicability of Detecting Somatic Point Mutation from RNA High Throughput Sequencing Data, Genomics. **107** (2016) 163–169. doi:10.1016/j.ygeno.2016.03.006.
[9] Y.-C. Hsu, Y.-T. Hsiao, T.-Y. Kao et al., Detection of Somatic Mutations in Exome Sequencing of Tumor-only Samples, Scientific Reports. **7** (2017) 15959. doi:10.1038/s41598-017-14896-7.
[10] M. Griffith, N.C. Spies, K. Krysiak et al., CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer, Nat Genet. **49** (2017) 170–174. doi:10.1038/ng.3774.
[11] D. Tamborero, C. Rubio-Perez, J. Deu-Pons et al., Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations, Genome Medicine. **10** (2018) 25. doi:10.1186/s13073-018-0531-8.
[12] R.W. Rapkins, F. Wang, H.N. Nguyen et al., The MGMT promoter SNP rs16906252 is a risk factor for MGMT methylation in glioblastoma and is predictive of response to temozolomide, Neuro Oncol. **17** (2015) 1589–1598. doi:10.1093/neuonc/nov064.
[13] C. Szumlanski, D. Otterness, C. Her et al., Thiopurine Methyltransferase Pharmacogenetics: Human Gene Cloning and Characterization of a Common Polymorphism, DNA and Cell Biology. **15** (1996) 17–30. doi:10.1089/dna.1996.15.17.