Identification of cancer drivers from tumor-only RNA-seq with RNA-VACAY

Jon Akutagawa[1], Allysia J Mak[1], Julie L Aspden[2], Angela N Brooks[1*]

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, California, 95064, USA
[2]School of Molecular and Cell Biology, Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT,UK
*anbrooks@ucsc.edu

Abstract

Detecting somatic mutations is a cornerstone of cancer genomics and clinical genotyping; however, there has been little systematic evaluation of the utility of RNA sequencing (RNA-seq) for somatic variant detection and driver mutation analysis. Variants found in RNA-Seq are also expressed, reducing the identification of passenger mutations and would not suffer from annotation bias observed in whole-exome sequencing (WES). We developed RNA-VACAY, a containerized pipeline that automates somatic variant calling from tumor RNA-seq data, alone, and evaluated its performance on simulated data and 1,349 RNA-seq samples with matched whole-genome sequencing (WGS). RNA-VACAY was able to detect at least 1 putative driver gene in 15 out of 16 cancer types and identified known driver mutations in 5' and 3' UTRs. The computational cost and time to generate and analyze RNA-seq data is lower than WGS or WES, which decreases the resources necessary for somatic variant detection. This study demonstrates the utility of RNA-seq to detect cancer drivers.

Introduction

The detection of somatic variants through next generation sequencing (NGS) has enabled researchers and clinicians to associate genetic mutations and disease phenotypes. The rapid improvement and falling costs of these technologies have led to the discovery of a whole host of crucial cancer-driving mutations and have opened new doors for targeted therapies in many cancers. Discovering *EGFR* mutations in lung cancer[1] and *BRAF* mutations in melanoma[2] have led directly to novel treatments[3,4] that have redefined standard of care options for eligible cancer patients. Continued advances in NGS technology, particularly whole exome sequencing (WES), whole genome sequencing (WGS), and RNA sequencing (RNA-seq), have allowed researchers to generate massive amounts of NGS data and better understand the genetic basis of cancer. RNA-seq is commonly employed for gene expression and alternative splicing analyses, which has given researchers an opportunity to uncover the transcriptional and post-transcriptional phenotypes of cancer cells.

Existing variant callers are designed primarily to handle WES or WGS data and have demonstrated the ability to detect novel somatic variants[5]. Variant calling tools are built to differentiate somatic mutations from inherited or *de novo* germline mutations, neutral polymorphisms, and artifacts derived from misalignments, sequencing errors, or PCR errors[6–9]. WES utilizes probes designed for exonic regions, which both introduces annotation biases[10] and results in the omission of variants within UTRs. Variant detection using RNA-seq data includes such regions of transcripts and therefore represents a more comprehensive search space for variants. Variants found in these transcripts are known to be expressed and are more likely to have a functional impact on the cell phenotype. UTRs are responsible for gene expression regulation and while their connections to disease are still being understood, mutations in these regions have been previously linked to cancer formation[11,12]. Despite these advantages, the transcriptome's inherent complexity can prove to be technically challenging when detecting variants. RNA-seq data often contain reads that span intronic regions or harbor variants in genes with low expression, which pose problems for many of these current variant calling tools[13]. The identification of somatic variants in RNA-seq data has previously been achieved[14–16], but there has yet to be an integrated pipeline with its results validated by a matched whole-genome variant list.

Here we present RNA-VACAY (**RNA-seq Va**riant **Ca**lling Pipeline), an integrated pipeline for detecting somatic variants in RNA-seq data (**Fig. 1)**. We evaluated the performance of several existing variant callers that can handle RNA-seq and determined that Platypus performed the best in terms of speed and ability to find variants. We incorporated Platypus into our RNA-VACAY somatic variant calling pipeline and compared the RNA-seq variants with a consensus list of somatic variants detected in matched whole genome sequencing samples. After extensive filtering of the raw candidate variants produced by Platypus, we demonstrate that our pipeline can effectively use RNA-seq data to reliably generate somatic calls found in whole exome and genome sequencing. Unlike other existing variant calling pipelines, it does not require a matched normal RNA-Seq sample. Many RNA-seq datasets lack matched normal samples because it requires biopsying adjacent tissue and is often difficult to obtain. Additionally, we detected recurrent mutations in 5' and 3' untranslated regions that are typically

not detected in WES. Our pipeline also identified potential somatic driver mutations that are consistent with previous reports. RNA-VACAY is a fast and low-cost somatic variant calling pipeline that demonstrates that cancer drivers can be discovered using tumor-only RNA-seq data.

Results

To test the performance of existing variant callers, we first created a synthetic RNA-seq dataset, using normal RNA-seq aligned reads from Pan-Cancer Analysis of Whole Genomes Consortium (PCAWG) of the International Cancer Genome Consortium (ICGC). We spiked 300 somatic SNVs into 20 of these samples, from diverse cancer types, to generate our synthetic RNA-seq dataset. We then surveyed multiple open-source variant callers (**Extended Data Table 1**) and cataloged their relevant features. Platypus[17], GATK[5], VarDict[18], and FreeBayes[19] were all evaluated for their ability to detect variants in RNA-seq data. RNA-MuTect[16], an existing RNA-seq somatic variant caller, was not evaluated because it requires a matched normal WES/WGS sample. FreeBayes was quickly eliminated as an option due to its massive requirements for both time and computational resources (**Fig. 2a**). Platypus, GATK, and VarDict all have multithreading options, allowing the user to decrease the total time necessary to run each tool when using a multicore system. Platypus had the best combination of recall and positive predictive value (PPV) of these three tools when analyzing this dataset (**Fig. 2b**). VarDict had the highest recall, but also detected a large number of false positives. We further curated a small subset of 20 samples with matched tumor and normal RNA-seq data from 8 tumor types within PCAWG to measure the performance of the tools with real world data. The variants detected in the normal RNA-seq were used to identify germline calls and potential false positives due to artifacts of sequencing and alignment. We compared RNA-seq-based variants with the consensus WGS variant calls from the same samples to measure recall. As expected, we saw higher recall at higher levels of expression and coverage across all tools (**Fig. 2c**). GATK had the highest recall in this analysis, which is likely due to GATK being a major component of the PCAWG WGS variant calling pipeline. Platypus had the next best performance after GATK and was again significantly faster and less resource intensive. As a result, Platypus was chosen to be incorporated into a new pipeline to detect somatic mutations in RNA-seq data.

Since PCAWG samples have both whole genome and RNA-seq data, we used this data as an opportunity to benchmark the RNA variant callers by comparing variant calls from RNA-seq data with known somatic variant calls from the whole genome sequencing data. We first broadly examined the performance of Platypus, alone, for detecting variants in known cancer-associated genes within a specific cancer type. We found that a large number of variants reported by Platypus did not replicate the consensus variants found in the matched WGS data. Starting with Platypus variants, we further filtered existing common variants found in dbSNP[20] and known RNA editing sites from REDIportal[21]. We also generated a panel of normal variants from samples from the Genotype-Tissue Expression[22] (GTEx) project as another baseline filter. Further analysis of the candidate variants revealed significant amounts of alignment artifacts, particularly around insertions and deletions and in specific regions of the genome. Gene

ontology analysis was performed on potential false positive variants and a striking number were found in immunoglobulin (Ig) and human leukocyte antigen (HLA) genes. Transcripts from these genes often feature extreme levels of diversity, making accurate mapping to these regions difficult for most tools[23,24]. Previous studies also applied similar filters, such as removing variants found at known RNA editing sites and near splice junctions[15]. Other potential false positives were found nearby homopolymer tracts and on reads with multiple variants in close proximity (within 50bp). These events were deemed to be likely sequencing or alignment artifacts and were removed from the candidate variant list.

The above filtering steps were incorporated into our final RNA-VACAY pipeline (**Extended Data Fig. 1**). We found that the majority of variants reported by RNA-VACAY matched those identified in their whole genome counterparts and significantly lowered the number of potential false positives. For example, in lung squamous cell carcinoma (LUSC), we detected 3,577,489 variants across the entire cohort using Platypus alone. Our pipeline delivered 7,326 candidate variants; 4,319 candidate variants were found in the WGS data. When we subsetted for only cancer-related genes in this cohort, we found 241 candidate variants were found in both RNA-seq and WGS data. 177 variants found in the WGS data were not detected by RNA-VACAY, showing a marked increase in recall. This finding was mirrored across all tissue types in the study (**Extended Data Fig. 2**). We specifically looked at the performance of the pipeline in two genes, *NOTCH1* and *NFE2L2*, that have been linked to cancer formation in LUSC (**Fig. 3a**). While the variants reported by Platypus alone point to a false hotspot mutation, RNA-VACAY largely replicated the WGS mutations found in *NOTCH1*. The *NFE2L2* R34 hotspot mutation was detected with RNA-VACAY with no false positives across the rest of the gene. Of the missed variants, many resulted in truncations or stop codon creation (**Extended Data Fig. 3**), which in turn commonly lead to degradation of the transcript by nonsense-mediated decay[25]; therefore, expression of these variants is low and subsequently were not detected by our pipeline. For example, truncating mutations in *TP53* reported in lung adenocarcinoma (LUAD) WGS data were not identified by RNA-VACAY. Many driver genes are often highly expressed[26] and therefore our pipeline detects these high impact variants with confidence.

Using this finalized pipeline, we detected 161,809 single nucleotide somatic variants in all 1,349 RNA-seq samples from the PCAWG dataset[27]. We surveyed several known cancer-associated genes with published hotspot mutations[28] (*KRAS* G12[29], *BRAF* V600[2], *PIK3CA* H1047[30], etc.) in multiple cancer types to assess the pipeline's performance in all of the cancer types analyzed by PCAWG. We measured all WGS variants found in these genes and found that 78% of these variant calls were also detected by RNA-VACAY, demonstrating the pipeline's ability to detect these variants in cancer-related genes while analyzing only RNA-seq data across different tumors (**Fig. 3b**). RNA-VACAY somatic mutation calls in *TP53*, the most frequently mutated gene in this study, recapitulated the WGS mutational frequency profile and demonstrated similar high mutational frequencies in liver cancer, colon adenocarcinoma, bone cancer, and breast adenocarcinoma. Similarly, mutation frequencies and counts in *KRAS*, *MYC*, *CREBBP*, and *SOCS1* were very similar in both RNA-seq and WGS data. Both *KMT2D* and *ARID1A* surprisingly had a larger share of RNA-seq only variants. After individual confirmation with the WGS aligned reads, the variants were present in both datasets, suggesting that these

particular WGS variants were removed during the consensus variant calling process in WGS analysis. Many of these removed variants were flagged as having a strand bias; our pipeline also identifies strand bias, but does not filter these variants due to the existence of stranded RNA-Seq library preparation methods .

In order to assess our pipeline's ability to detect cancer driver genes, we used oncodriveFML[31] to compare the driver mutation profiles of matched RNA-seq and WGS samples in multiple cancer types. OncodriveFML predicts which genes harbor driver mutations using functional impact scores derived from the Combined Annotation Dependent Depletion (CADD) tool. The mean functional impact (FI) score of the mutations within a gene are compared with the distribution of mean functional impact scores of randomly generated mutations. Genes with significant differences in FI scores are likely to be driver genes. We used single nucleotide variants in coding regions from the RNA-seq data across all tumor types to generate driver gene profiles and compared these profiles to their matched WGS samples. The driver mutation profile of RNA-seq variants called by Platypus alone initially resulted in a multitude of potential driver genes, which can be attributed to the inclusion of germline or false positive variants (**Fig. 4a**). However, RNA-seq variants called from our RNA-VACAY pipeline are often consistent with their WGS equivalents (**Fig. 4a,b and Extended Data Fig. 4**). Known driver genes were identified such as *TP53*, *KMT2D*, *CDKN2A, and NFE2L2* in both LUSC RNA-seq and WGS data. *NOTCH1*, another cancer-related gene, was also predicted to be a driver gene using RNA-VACAY variants, but not WGS. Similarly, *TP53* and *KDM6A* were reported to have driver mutations in bladder adenocarcinoma RNA-seq and WGS data. *SPTAN1* and *KMT2D* were also predicted to be driver genes from RNA-VACAY variants, but not WGS. The mutations in *NOTCH1*, *SPTAN1* and *KMT2D* detected by RNA-VACAY were also found in the WGS consensus calls. However, additional synonymous mutations found in the WGS lower the mean FI score of those genes. Interestingly, the mutations in *NOTCH1* and *TP53* not found in the RNA-seq data are either synonymous or missense mutations, suggesting that the variants are not expressed and may not be functional. *TP53*, the gene with the most recurrent mutations, was most commonly reported as being a driver gene in 13 cancer types using WGS variants. RNA-VACAY delivered the same finding in 11 cancer types. In chronic lymphocytic leukemia (CLLE) and renal cell carcinoma (RECA), there were significantly more driver gene candidates found in the RNA-VACAY variants than the WGS variants (**Extended Data Fig. 4**). Upon inspection, there was a significantly higher number of variants found on the same reads and in close proximity to one another, which may point to either a technical artifact introduced during sample preparation or alignment. However, in cancer types with a high mutation frequency such as skin cutaneous melanoma (SKCM), we saw less overlap between the RNA-seq and WGS data; multiple genes from the protocadherin alpha gene cluster (*PCDHA6, PCDHA10, PCDHA7, PCDHA1,* etc.) were reported as potential driver using the WGS data (**Extended Data Fig. 5a**). These genes are lowly expressed in this cancer type, which could explain why RNA-VACAY was unable to detect these variants (**Extended Data Fig. 5b**). Overall, 16 cancer types reported one or more driver genes using WGS variants and RNA-VACAY was able to detect at least 1 matching gene in 13 of them. RNA-VACAY described 1 or more potential driver genes in 15 of 16 cancer types that were also listed in the Cancer Gene Census, a curated database of mutations implicated in cancer[32]. The driver gene profiles generated from somatic variants

detected by RNA-VACAY largely match the driver gene profiles generated from variants found in the corresponding WGS data, demonstrating the ability to use RNA-seq alone to find driver genes.

Previous PCAWG studies identified recurrent noncoding point mutations in multiple genes as being strong candidate drivers[33]. As RNA-seq captures both 5' and 3' untranslated regions (UTRs), we decided to test RNA-VACAY's ability to detect these same UTR mutations. Somatic variants in the 5' UTR of *MTG2* and 3' UTR of *TOB1* and *NFKBIZ* were detected by RNA-VACAY (**Fig. 5**). RNA-VACAY was unable to detect 5' UTR mutations in *PTDSS1* and *DTL*, as the vast majority of RNA-seq samples had virtually no aligned reads in the specified region of those genes. This may be because somatic variants in this region can often downregulate or upregulate the expression of these genes, particularly in a cancer context[34]. An alternative explanation is that RNA-seq data can exhibit a 3' end coverage bias due to the cDNA amplification process, resulting in reduced 5' UTR coverage. Provided there is satisfactory coverage, RNA-VACAY is successfully able to detect recurrent UTR variants.

Discussion

WES and WGS continue to be the main sources of genomic data for identifying cancer-associated somatic variants, but the function and cost of RNA-seq make it an increasingly attractive option for characterizing tumors. In situations where no WES or WGS data are available, existing RNA-seq data collected for differential gene expression or gene fusion analysis can also be used for somatic variant detection. We applied RNA-VACAY to over 1,300 RNA-seq samples and were able to detect somatic variants with high recall. Across all samples, the median recall was 0.25, but increases to 0.48 when looking specifically at cancer-related genes (**Extended Data Fig. 2**). Our study demonstrates that RNA-seq data can function both as a supplement and as a substitute for WES and WGS data when detecting somatic variants. These variants were detected in actively expressed regions, so they are more likely functionally relevant and significant. RNA-VACAY has demonstrated its ability to detect somatic variants in RNA-seq that match the driver gene profiles of variants detected in WGS.

Using RNA-seq also allows for the discovery of somatic variants in the 5' and 3' UTR, allowing for further discovery of the functional impact of these noncoding variants. While we currently filter out previously identified RNA editing sites, future applications of our pipeline could also be to measure the RNA editing profile of a transcriptome or detect novel RNA editing sites. However, our pipeline is also limited by the biological underpinnings of RNA-seq. Variants in lowly expressed genes or that decrease expression are difficult to detect. Genes with tissue-specific expression can also make variant discovery challenging.

Our pipeline does not currently detect insertions and deletions. The preprocessing step with Opossum has only been evaluated in the context of single nucleotide variant detection. Platypus has been reported to detect indels in WES and WGS data, so extending the scope to detect indels would be a natural goal for updated versions of this pipeline. A consensus strategy, incorporating multiple variant callers into the pipeline, could also be used to increase both recall and PPV.

Tumor-only sequencing can also misidentify germline variants as being somatic variants. Our filtering approach utilizing multiple public variant and mutation databases is designed to

minimize this scenario. Sequencing adjacent normal tissue can decrease the number of inaccurately defined somatic variants. Our driver analyses of cancer cohorts also decrease the chances of a rare germline mutation being identified as a significant somatic mutation.

RNA-VACAY is capable of harnessing existing RNA-seq data and provides a cost-effective and reliable option for the validation of variants found through other methods (**Extended Data Fig. 6**). For example, for the size of the PCAWG study of 1,349 samples we estimate that the cost of detecting somatic mutations from RNA-seq of only tumor samples to be $592,487, while the cost of WGS is estimated at $1,472,926[35,36]. Using RNA-seq would also cut the runtime from 68,326 hours to 16,275 hours. Next generation sequencing technologies continue to enter into the clinic and have become the gold standard in the genetic diagnosis of cancer and other genetic diseases. The importance of RNA-seq as a clinical diagnostic tool requires robust and straightforward pipelines to automate analysis of this data.

Methods

Aligned reads processing

We used data from The Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium[37] of the International Cancer Genome Consortium (ICGC). We downloaded 1,349 RNA-seq samples from the PCAWG data portal[37]. This dataset features 30 cancer types. 161 of these samples originate from normal solid tissue or tissue adjacent to the tumor. These reads were aligned with STAR (v2.4.0i) and hs37d5 or Gencode (release 19) as the reference gene annotation. Matched WGS data for these samples were used to evaluate pipeline performance. To generate a synthetic dataset, 300 randomly selected somatic single-nucleotide variants (SNVs) were manually added to aligned reads from 20 PCAWG normal tissue RNA-seq samples to simulate tumor RNA-seq data. All variants were located in the coding regions of the genome and had random allele frequencies. A test dataset of 8 donors with matched tumor and normal RNA-seq from 8 tumor types were curated and used to evaluate the performance of the variant calling tools.

Normal tissue samples were downloaded from the Genotype-Tissue Expression (GTEx) project[22] portal and re-aligned with identical parameters. Variants detected in these samples were then used to generate a panel of normal variants. 20 samples from 11 different tissue types were incorporated into this panel (**Supplementary Table 1**).

Pipeline description

The RNA-VACAY pipeline is a modular workflow built on Python 2.7 that automates task assignment, downloading and preprocessing data, tool execution, and variant analysis. Each task is completed within a Docker container. Source code can be found at: https://github.com/BrooksLabUCSC/RNA-VaCay.

1. Data retrieval

   This pipeline is built to specifically handle RNA-seq reads aligned with STAR (**Supplementary Note 1**) currently stored in either PCAWG or TCGA repositories. The download module includes the recommended tools by each consortium. RNA-VACAY

accepts file manifests generated by these data repository portals and automates downloading. It can also accept user-generated file manifests to call variants in either previously downloaded data or RNA-seq data generated by the user.

2. Preprocessing data

Aligned reads are sorted and indexed by samtools[38] if necessary and then preprocessed with Opossum (v0.2)[39]. Opossum prepares RNA-seq data for variant calling by Platypus, GATK, and other callers. It splits reads mapped across splice junctions and ensures that minimal information is lost at read ends by merging overlapping reads and modifying base qualities at the edges of these reads. Opossum also eliminates duplicate reads.

3. Variant calling

Platypus is a Bayesian haplotype-based variant caller that uses local *de novo* assembly and realigns sequences to detect variants. Platypus also shares variant information between multiple samples, increasing the confidence of calls that are weakly supported in one sample, but strongly supported in related samples.

4. Filtering

Raw variant calls from Platypus are first filtered with a custom panel of normal variants generated from RNA-seq samples from the GTEx repositories. Subsequent filters incorporate a combination of preexisting common and normal variant databases - dbSNP[20], gnomAD[40], and REDIportal (RNA editing sites)[21]. Variants with low quality scores or sequencing depth (<7) were filtered. Variants found in certain locations, such as known decoy regions, and repeat regions were excluded. Variants found in human leukocyte antigen genes, immunoglobulin genes, and pseudogenes were also excluded. Variants found within 50 bases of other variants with similar allele frequencies or within 10 bases adjacent to homopolymer tracts of 5+ bases were also excluded. An optional filter will prevent removal of variants found in known cancer hotspots, regardless of call quality. Normal variants from matched normal RNA-seq samples, if available, can also be incorporated as an optional filter.

5. Annotation and analysis

Filtered variants were annotated with SnpEff (v4.3t)[41]. SnpEff categorizes the variants based on their genomic locations and predicts the coding effects of these variants. These candidate variants were then analyzed using custom Python scripts. Driver analysis was performed by oncodriveFML. OncodriveFML calculates a profile of somatic mutations in specific genomic regions and identifies genes that have a higher mutational frequency compared to their background mutation rate. All calls outside of the coding region and any non-single nucleotide variants were filtered before running oncodriveFML.

Initial variant caller evaluation

Four open-source variant callers previously reported to be compatible with RNA-seq data were evaluated – Platypus (v0.8.1.1) , GATK (v4.1.9), VarDict (v1.5.5), and FreeBayes (v1.1). We ran the tools using default recommended options and recommended preprocessing steps for Platypus (Opossum[39]) and GATK (SplitNCigarReads). We measured the speed, recall, PPV, and

resource requirements of the four variant callers processing 10 RNA-seq samples, comparing the results between the 5 pairs of normal and tumor samples.

PCAWG RNA-seq data analysis

RNA-seq samples were downloaded as cohorts based on cancer type. RNA-VACAY was run on multiple OpenStack instances in parallel. Custom python scripts were developed to handle and aggregate results.

Single gene variant comparison

The variants in particular genes were visualized as stickplots with cBioPortal[42,43]. Known cancer-related genes in specific cancer types were chosen and plotted using the MutationMapper tool. Custom python scripts were written to analyze the overlap between the RNA-VACAY and WGS variant sets.

Cancer type variant comparison

The RNA-VACAY and WGS mutational frequencies of the 25 most mutated cancer-related genes were compared across each PCAWG tumor type using a custom python scripts (https://github.com/BrooksLabUCSC/RNA-VaCay).

Driver mutation profiling

OncodriveFML (v2.2.0) was used to identify genes with potential driver mutations. We ran oncodriveFML with default settings on filtered variants, using the whole-exome sequencing option.

5' and 3' UTR mutation confirmation and visualization

Custom python scripts were written to uncover variants detected by RNA-VACAY that matched previously published genes with recurrent 5' and 3' UTR mutations. We used Integrated Genomics Viewer (v2.8.3)[44] to visually confirm the variants.

Data availability

Data associated with this research will be submitted to the Genomic Data Commons (GDC) and European Genome-Phenome Archive (EGA). A list of data files used for filters and analysis is provided in **Supplementary Note 2**. To access information that could potentially identify participants, such as the underlying sequencing data and sequencing variants at https://dcc.icgc.org/PCAWG, researchers will need to apply to the TCGA data access committee via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the dataset, and to the ICGC data access compliance office (http://icgc.org/daco) for the ICGC portion of the dataset.

Code availability

All code and software used in this study is available through GitHub (https://github.com/BrooksLabUCSC/RNA-VaCay/). Commands associated with the pipeline are also provided in **Supplementary Note 3**.

Acknowledgments

We would like to thank Cameron Soulette for his valuable support during the genesis of this project and the members of the Brooks and Vaske Lab for their helpful feedback. We would also like to thank Yoseph Barash for helpful suggestions. This work was supported by the University of California Cancer Research Coordinating Committee CRN-19-586125 (A.N.B.). J.A. was partially supported by an NIH training grant 5T32HG008345 and the UCSC STARS Re-entry Scholarship. J.L.A. was funded by the University of Leeds (University Academic Fellow scheme). Partial support was also provided by the following awards: NIH/NCI 1R21CA238600-01(A.N.B.) and Pew Charitable Trusts (A.N.B.).

Author contributions

J.A., A.J.M., J.L.A., and A.N.B. conceived and designed the study. J.A. and A.J.M. collected and assembled data. J.A., A.J.M., and A.N.B. developed the figures and tables. All authors were involved in the critical review of the manuscript and approved the final version.

Competing interests

A.N.B. is a consultant for Remix Therapeutics, Inc. All other authors have declared no competing interests.

1. Rusch, V. *et al.* Differential expression of the epidermal growth factor receptor and its ligands in primary non-small cell lung cancers and adjacent benign lung. *Cancer Res.* **53**, 2379–2385 (1993).
2. Long, G. V. *et al.* Prognostic and clinicopathologic associations of oncogenic BRAF in metastatic melanoma. *J. Clin. Oncol.* **29**, 1239–1246 (2011).
3. Fukuoka, M. *et al.* Multi-Institutional Randomized Phase II Trial of Gefitinib for Previously Treated Patients With Advanced Non–Small-Cell Lung Cancer. *J. Clin. Orthod.* **21**, 2237–2246 (2003).
4. Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E

mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).

5. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
6. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
7. Xu, F. *et al.* A fast and accurate SNP detection algorithm for next-generation sequencing data. *Nat. Commun.* **3**, 1258 (2012).
8. Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–501 (2011).
9. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
10. Barbitoff, Y. A. *et al.* Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci. Rep.* **10**, 2057 (2020).
11. Signori, E. *et al.* A somatic mutation in the 5′UTR of BRCA1 gene in sporadic breast cancer causes down-modulation of translation efficiency. *Oncogene* **20**, 4596–4600 (2001).
12. Liu, L. *et al.* Mutation of the CDKN2A 5’ UTR creates an aberrant initiation codon and predisposes to melanoma. *Nat. Genet.* **21**, 128–132 (1999).
13. Quinn, E. M. *et al.* Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. *PLoS One* **8**, e58815 (2013).
14. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651 (2013).

15. García-Nieto, P. E., Morrison, A. J. & Fraser, H. B. The somatic mutation landscape of the human body. *Genome Biol.* **20**, 298 (2019).
16. Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, (2019).
17. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
18. Lai, Z. *et al.* VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, (2016).
19. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).
20. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
21. Picardi, E., D'Erchia, A. M., Lo Giudice, C. & Pesole, G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* **45**, D750–D757 (2017).
22. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
23. Watson, C. T. & Breden, F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun.* **13**, 363–373 (2012).
24. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
25. Amrani, N. *et al.* A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature* **432**, 112–118 (2004).
26. Ohshima, K. *et al.* Integrated analysis of gene expression and copy number identified

potential cancer driver genes with amplification-dependent overexpression in 1,454 solid tumors. *Sci. Rep.* **7**, 641 (2017).

27. PCAWG Transcriptome Core Group *et al.* Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).
28. Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
29. Riely, G. J. *et al.* Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma. *Clin. Cancer Res.* **14**, 5731–5734 (2008).
30. Mandelker, D. *et al.* A frequent kinase domain mutation that changes the interaction between PI3Kalpha and the membrane. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 16996–17001 (2009).
31. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 1–13 (2016).
32. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
33. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
34. Lim, Y. *et al.* Multiplexed functional genomic analysis of 5’ untranslated region mutations across the spectrum of prostate cancer. *Nat. Commun.* **12**, 4217 (2021).
35. Yung, C. K. *et al.* Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments. *bioRxiv* 161638 (2017) doi:10.1101/161638.
36. DNA Technologies Core. UC Rate — UC Davis and other UC campuses. *DNA Technologies Core* https://dnatech.genomecenter.ucdavis.edu/uc-prices/ (2022).

37. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

39. Oikkonen, L. & Lise, S. Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. *Wellcome Open Res* **2**, 6 (2017).

40. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

41. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

42. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, l1 (2013).

43. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

44. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

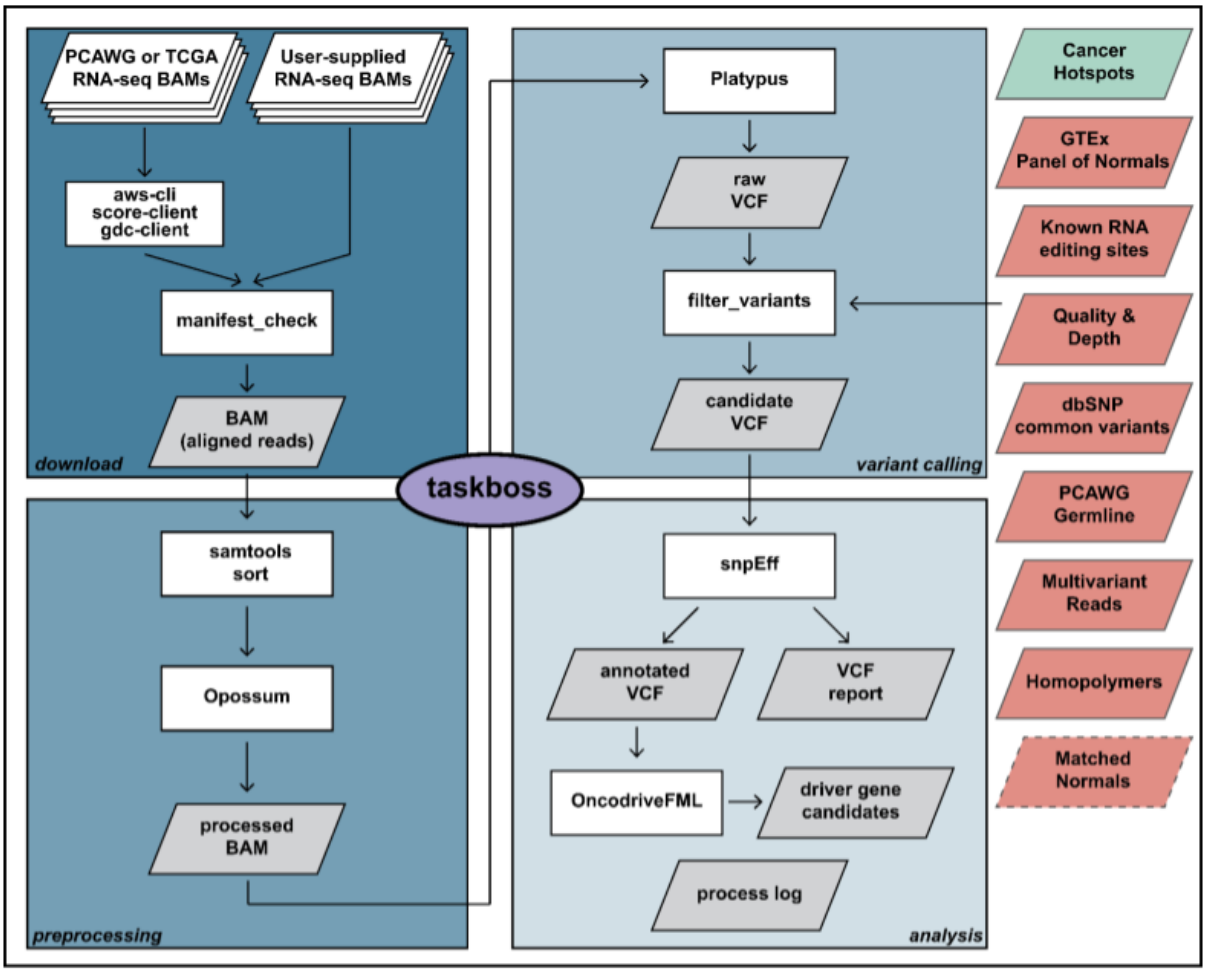


Fig. 1 | A scalable variant calling pipeline to detect somatic variants in RNA-seq data. Schematic of RNA-VACAY, a somatic variant calling pipeline designed for RNA-seq data. The pipeline consists of 4 main modules - data download, preprocessing, variant calling, and analysis. Taskboss, a controller module, assigns tasks according to available resources and can assist in resuming pipeline operations after interruptions. Multiple filtering options (green and red) can be toggled to keep known cancer mutations and remove likely false positives. Variants from matched normal samples, if available, can also be used to filter common variants. White boxes refer to components of the pipeline and gray parallelograms refer to outputs generated by the pipeline.

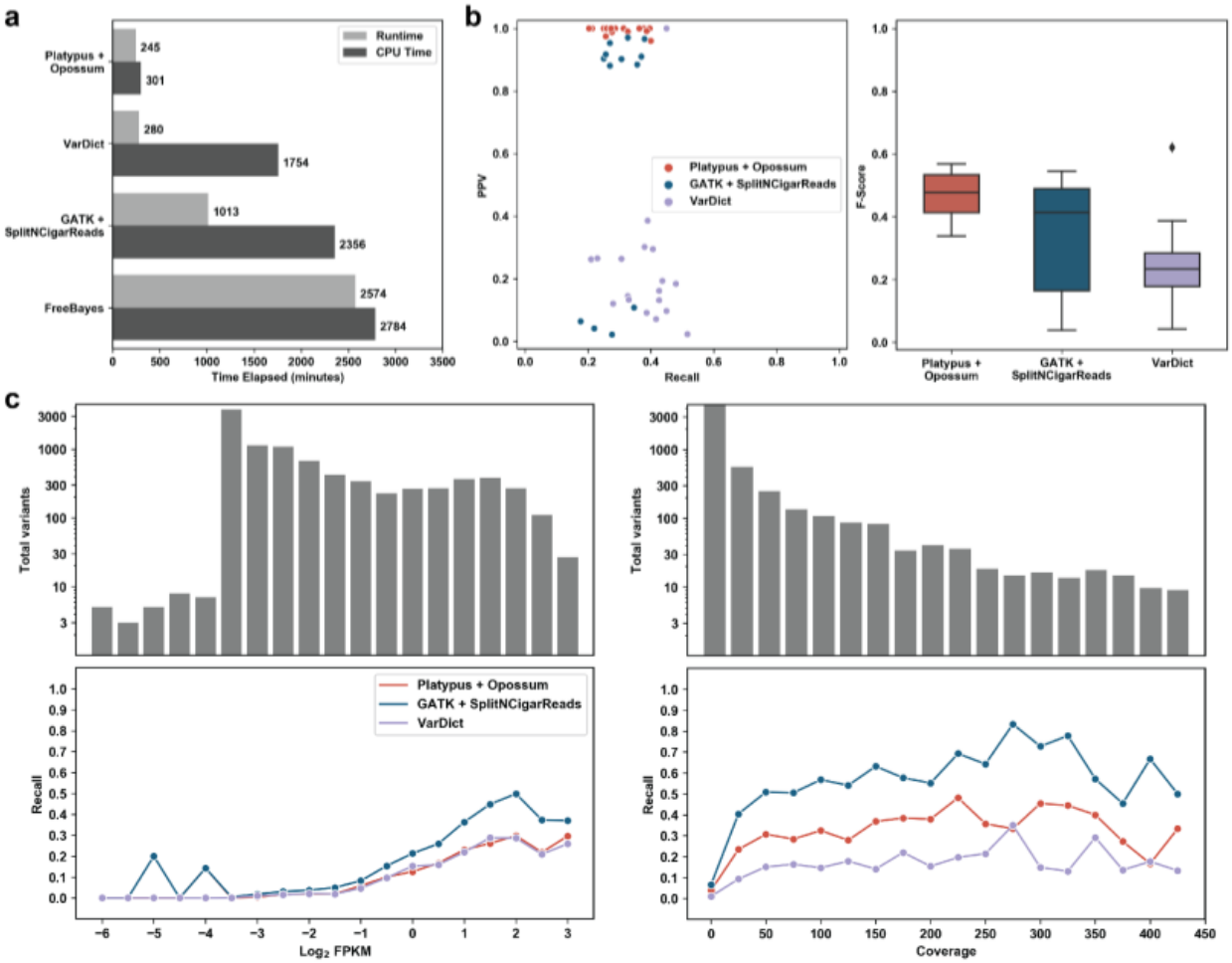


Fig. 2 | Platypus is the tool best fit for a fast and sensitive variant calling pipeline. a, Four variant callers were evaluated for runtime using 5 synthetic RNA-seq samples. The mean time was reported. Together Platypus and Opossum - a preprocessing step - were significantly faster and consumed less resources than all other variant callers. **b,** Three variant callers were evaluated for recall and positive predictive value (PPV) with a synthetic RNA-seq dataset of aligned reads with spiked-in somatic mutations. Platypus had the highest median F-score. **c,** The three tools were then evaluated with a small test dataset (12 samples from diverse cancer types) curated from PCAWG. All variant callers generally were more sensitive in genes with higher expression (log_2 FPKM) and variants with higher coverage.

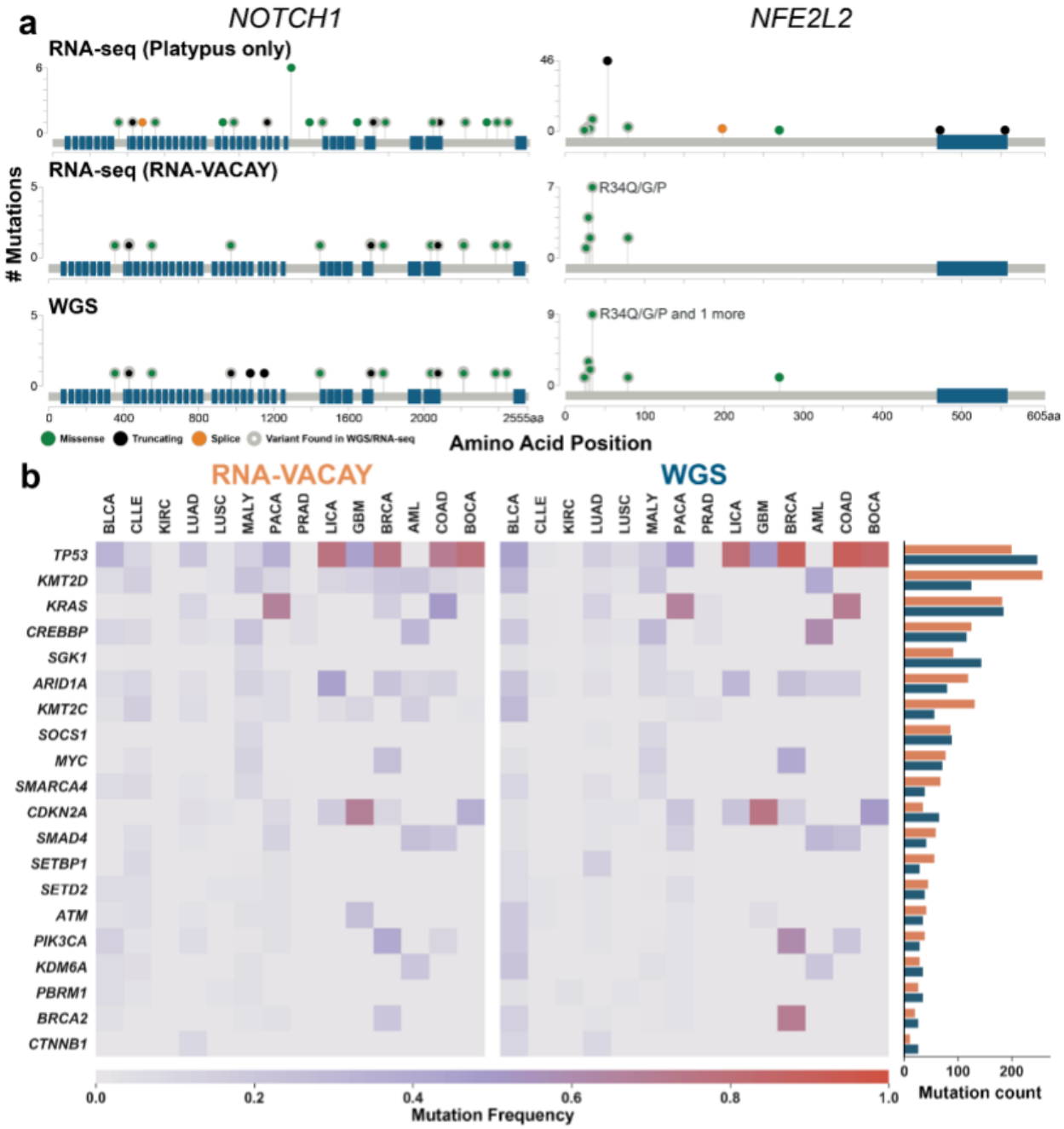


Fig. 3 | RNA-VACAY identifies cancer-related variants found in WGS data.
a, Stickplot of RNA-seq and WGS mutations found in the genes *NOTCH1* and *NFE2L2* of lung squamous cell carcinoma samples. *NFE2L2* R34* is a known hotspot mutation. **b,** Heatmap compares frequency of samples in a tissue type containing mutations in known cancer-associated genes. Barplot displays total number of mutations found in each gene - RNA-VACAY variants are orange and WGS variants are blue.

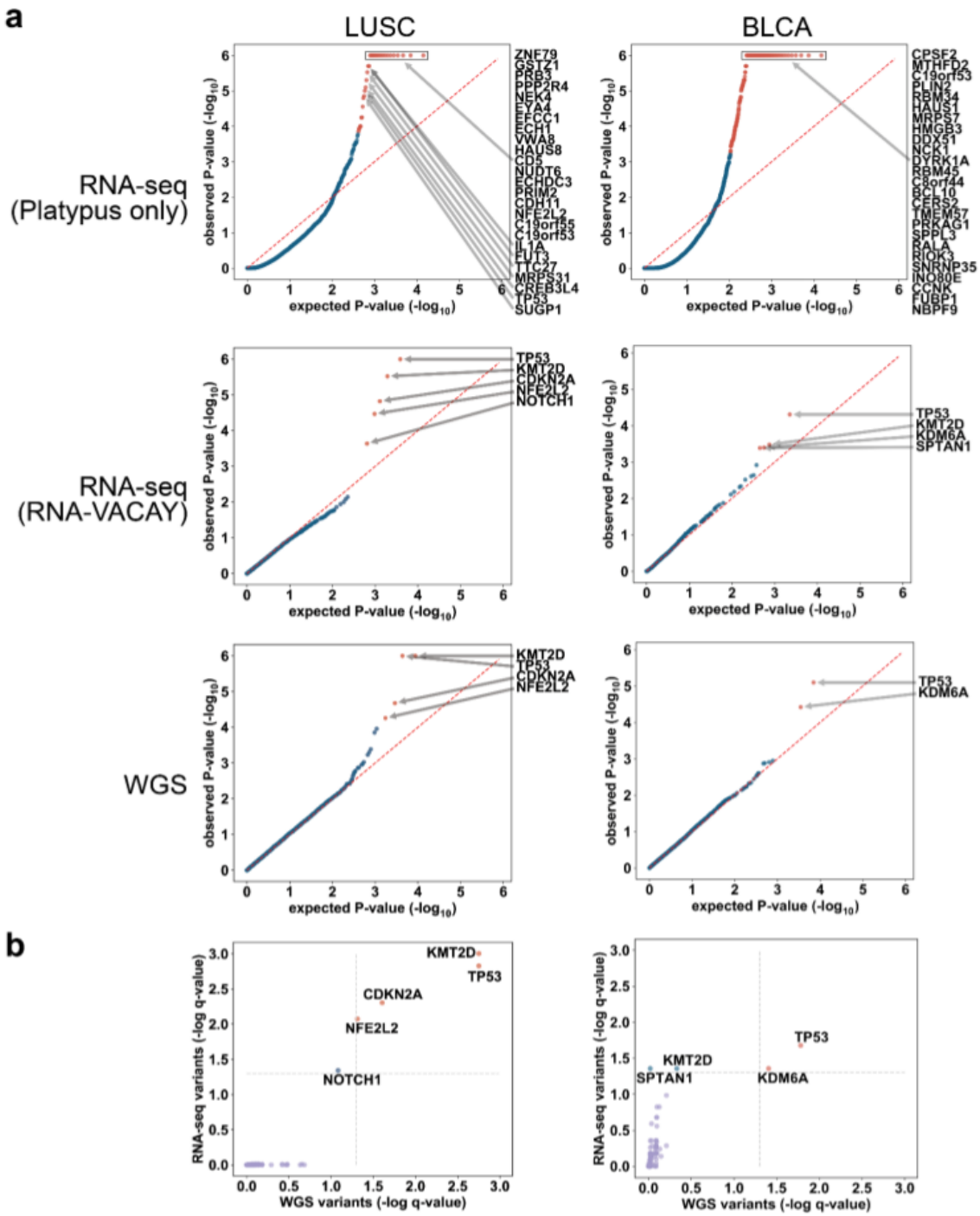


Fig. 4 | Driver mutation profiles from RNA-VACAY variants match WGS. a, Quantile-quantile (QQ) plots of p-values generated from oncodriveFML. Red line displays where observed and expected p-values match. Blue dots represent genes with Q-values ≥0.1, red <0.1. Genes with Q-values <0.1 are indicated. **b,** Plots of q-values from genes with detected RNA-seq and WGS variants. Quadrants show significance in RNA-seq, WGS, or both.

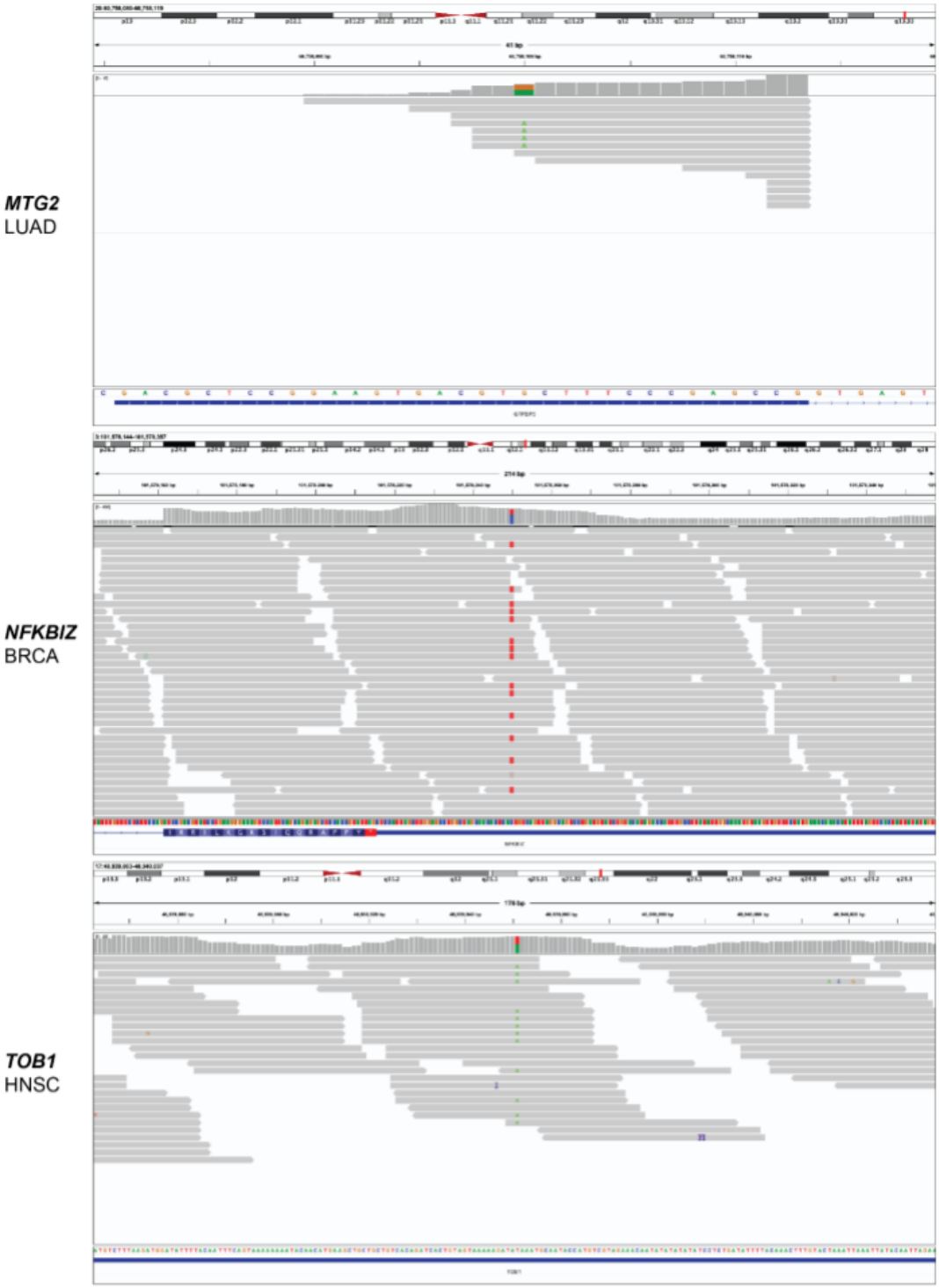
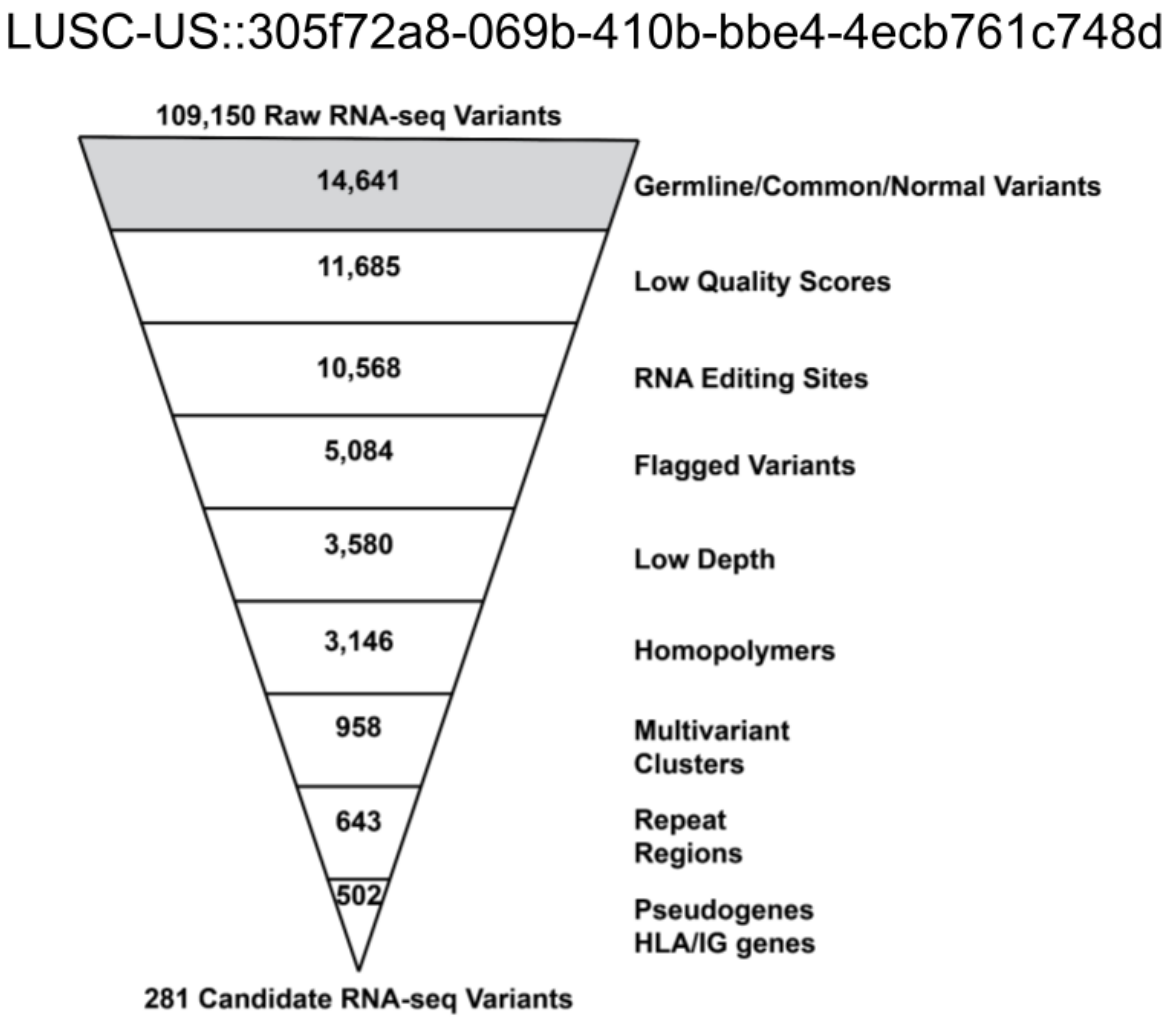


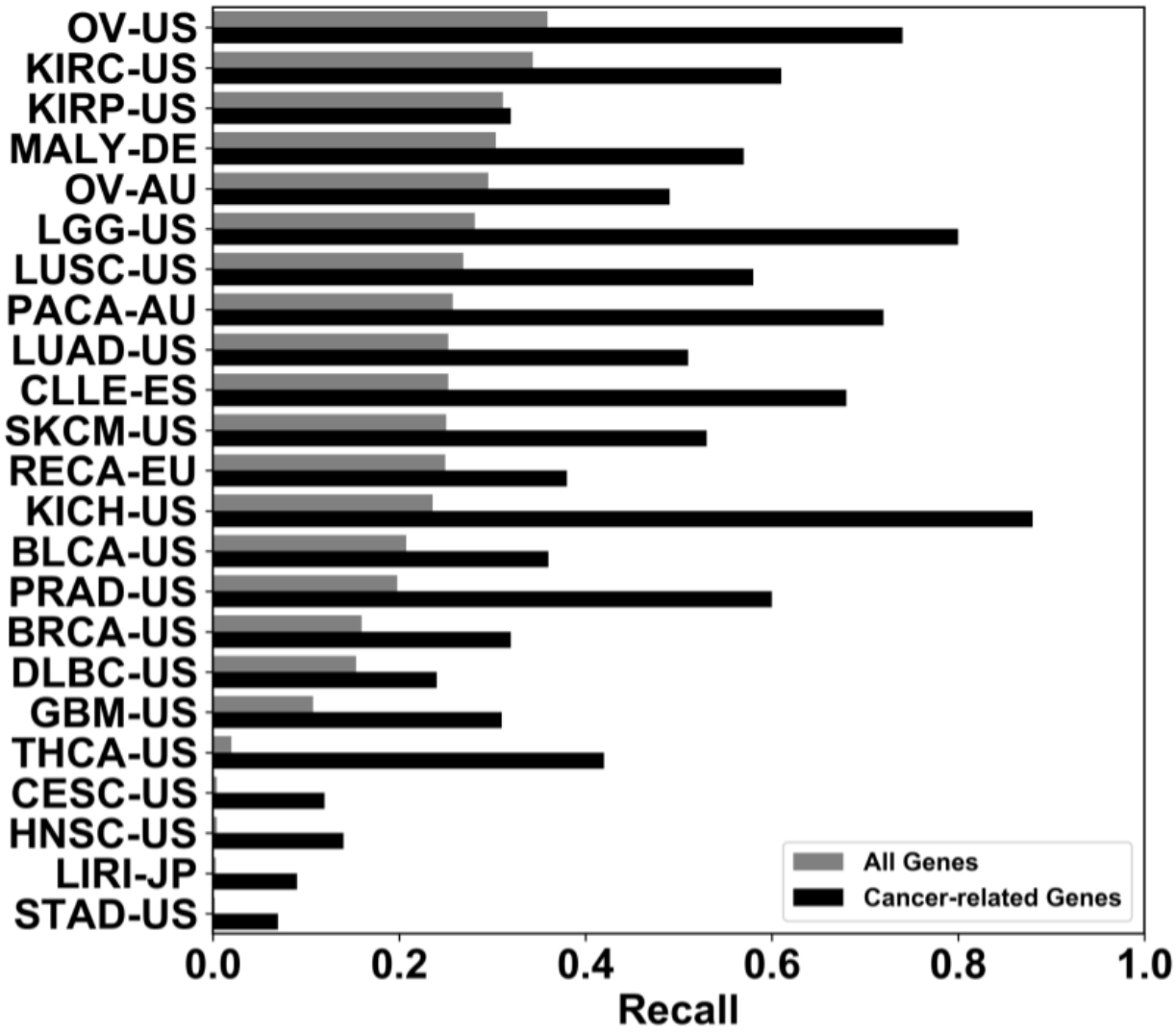
Fig. 5 | RNA-VACAY detects 5' and 3' UTR driver mutations. RNA-VACAY detected previously discovered candidate driver mutations in the 5' UTR of *MTG2* and and 3' UTR of *TOB1* and *NFKBIZ*.

	Requires DNA	Matched Normal	Preprocessing
GATK	No	No	Yes
Platypus	No	No	Yes
VarDict	No	No	No
FreeBayes	No	No	No
Samtools	No	No	No
Strelka2	No	Yes	No
SNPiR	No	No	No
RADIA	Yes	Yes	No
VarScan2	No	Yes	No

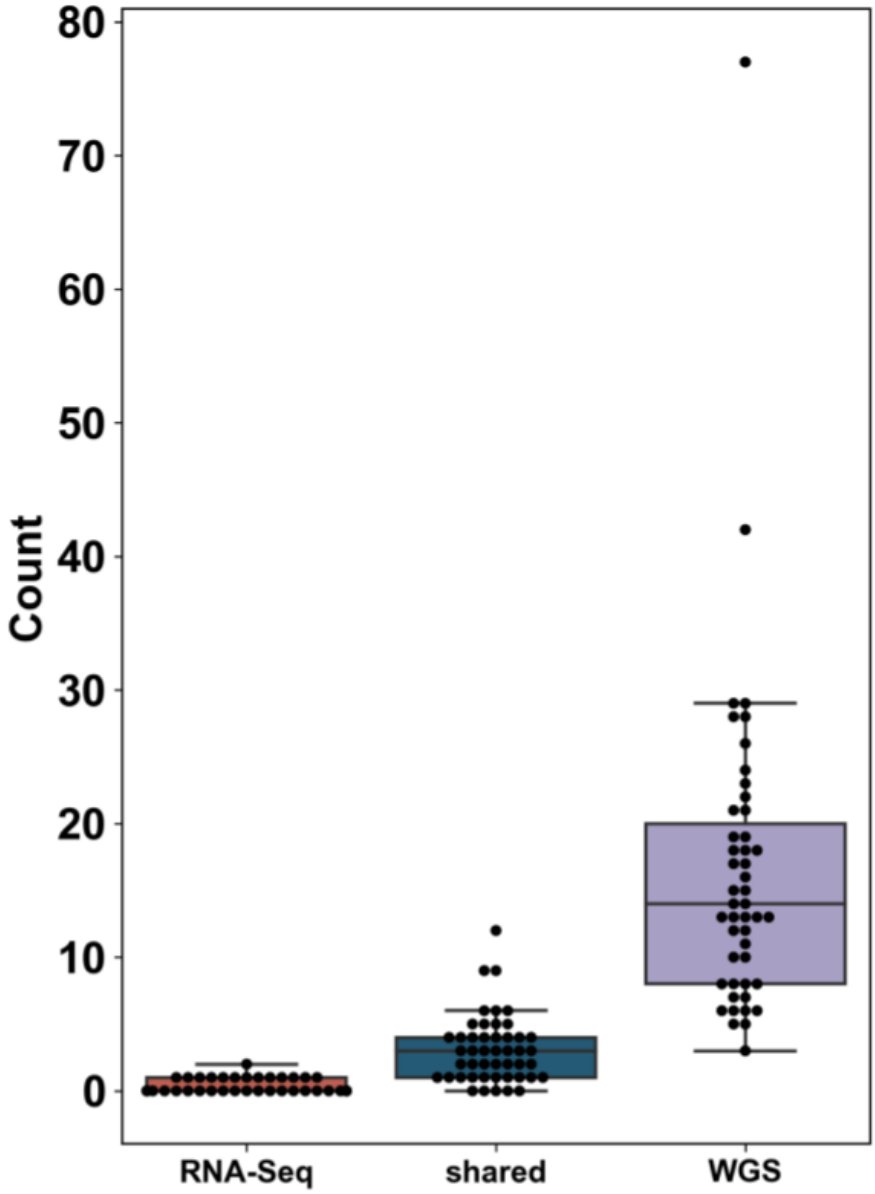
Extended Data Table 1 | Existing variant calling tools. These 9 variant callers have been previously used to call somatic variants in RNA-seq data.



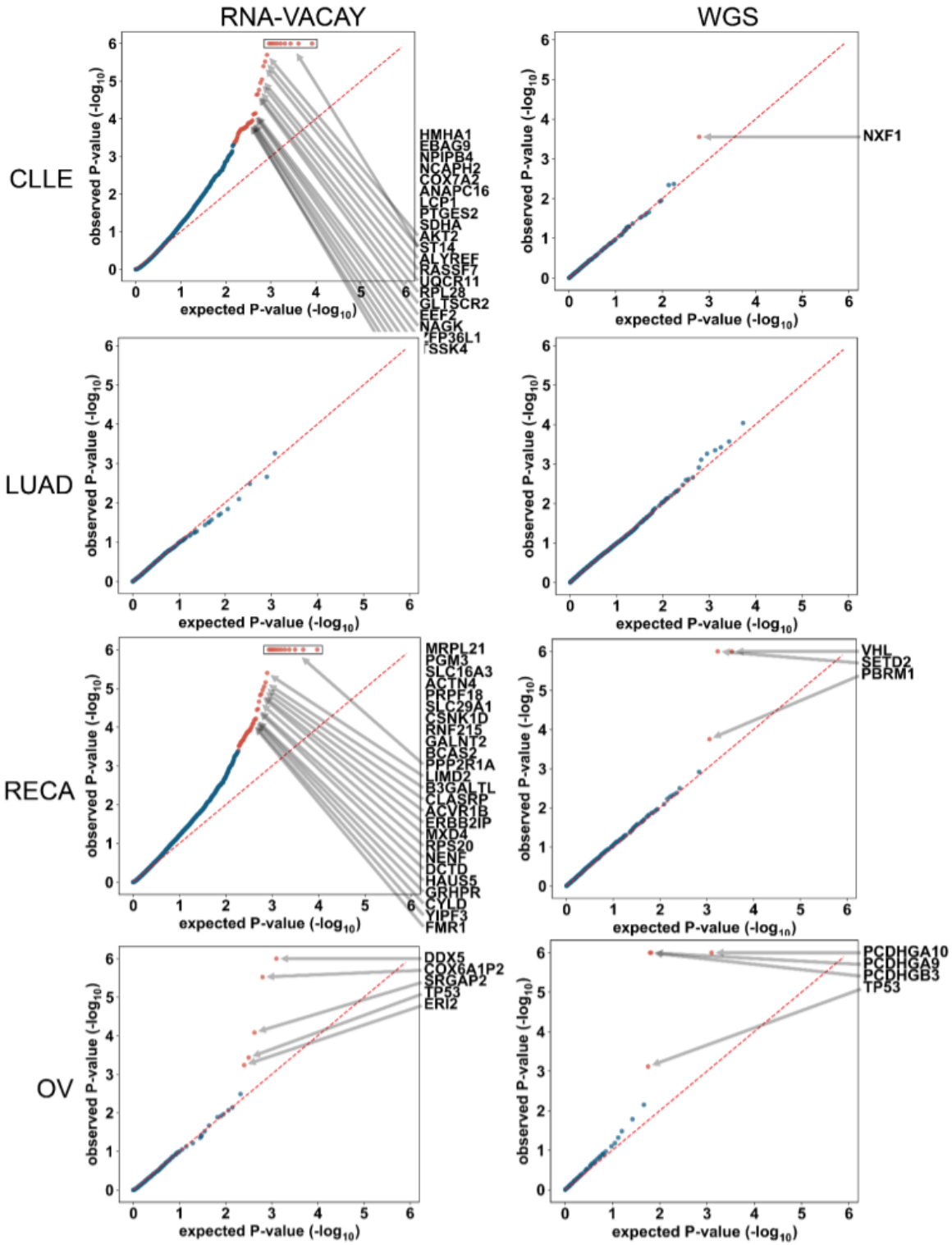
Extended Data Fig. 1 | Filtering strategy for RNA-VACAY. An example filtering strategy for a single lung squamous cell carcinoma sample. RNA-VACAY uses multiple filters to remove false positives from the raw variant candidates.

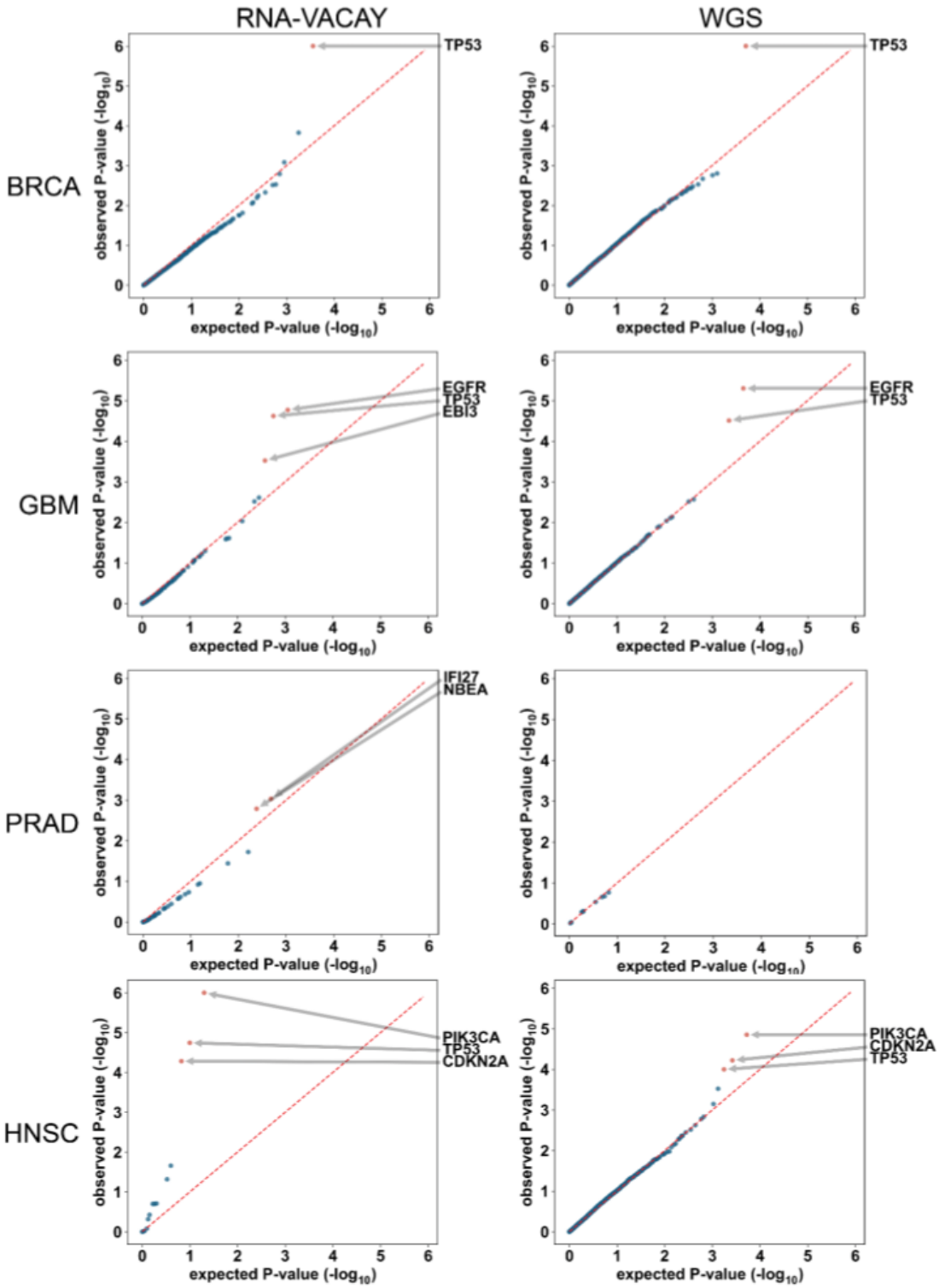


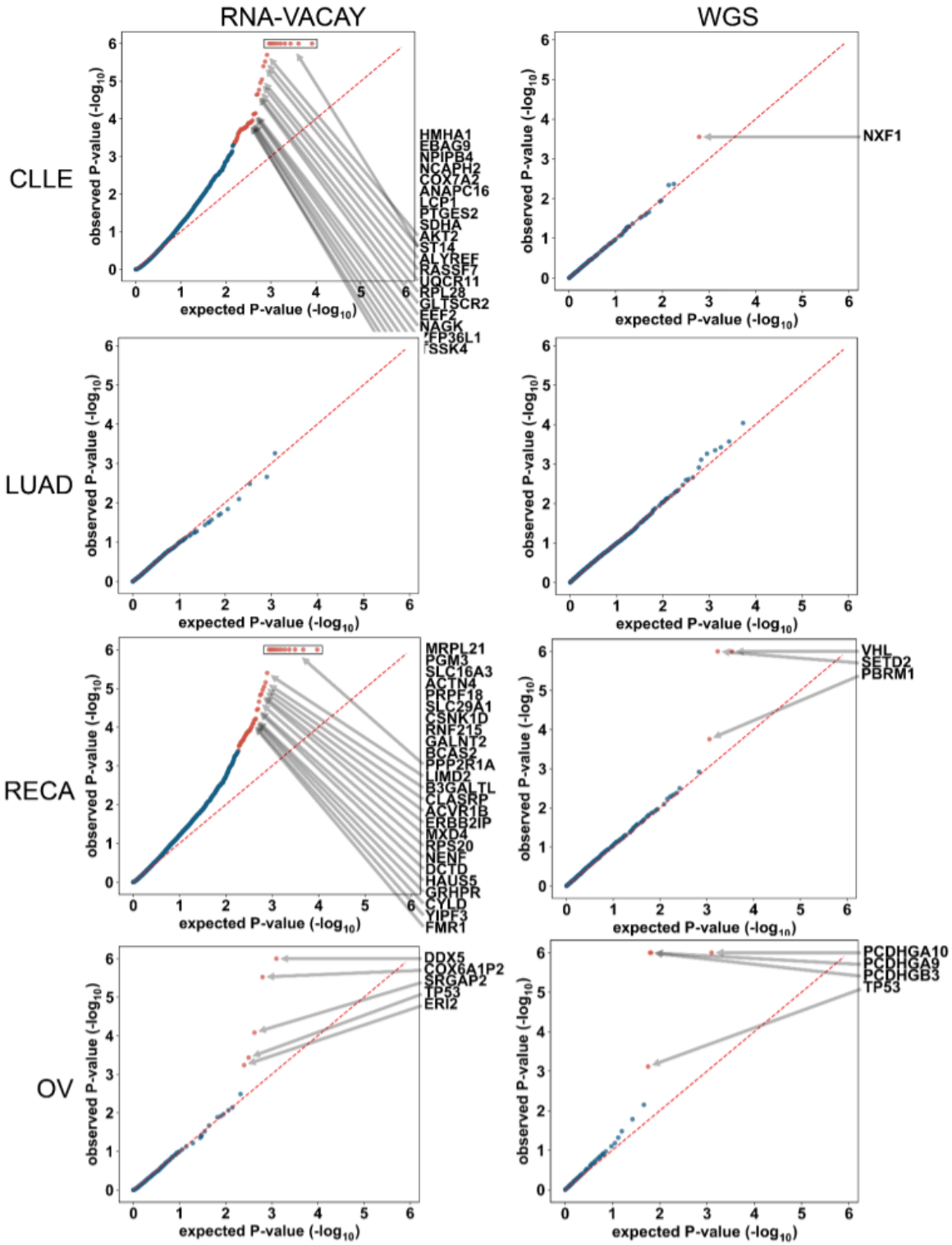
Extended Data Fig. 2 | Increased recall when focused on cancer-related genes. Median recall across all tumor types was 0.25. Focusing on a subset of genes provided by the Cancer Gene Census, median recall increases to 0.48.

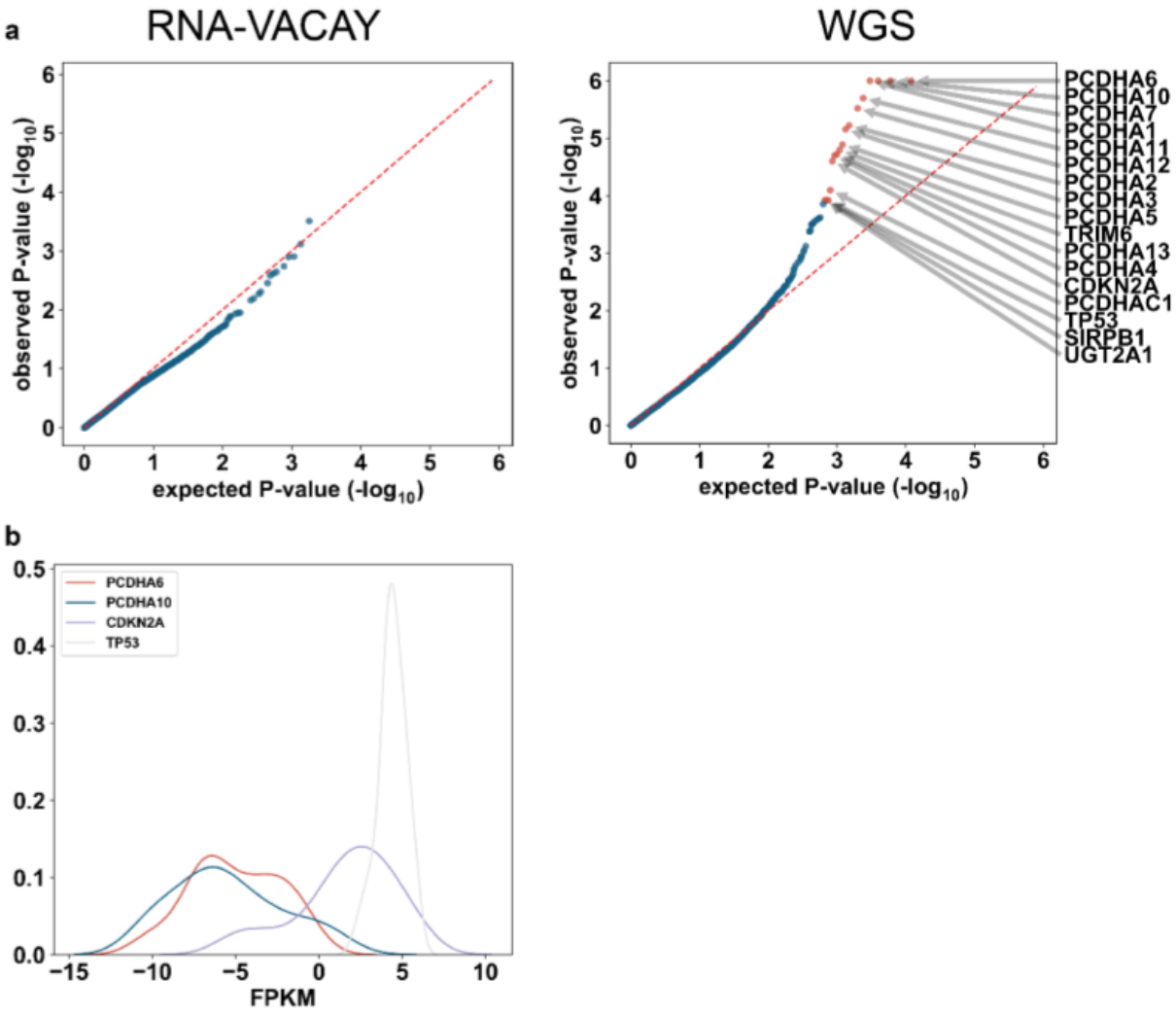


Extended Data Fig. 3 | Potential stop codon creating variants not detected in RNA-seq data. In lung squamous cell carcinoma samples, the median count of stop codon creating variants detected in WGS data alone was 14, while the median count from RNA-seq alone was 0. The median count of stop codon creation variants found in both datasets was 3.

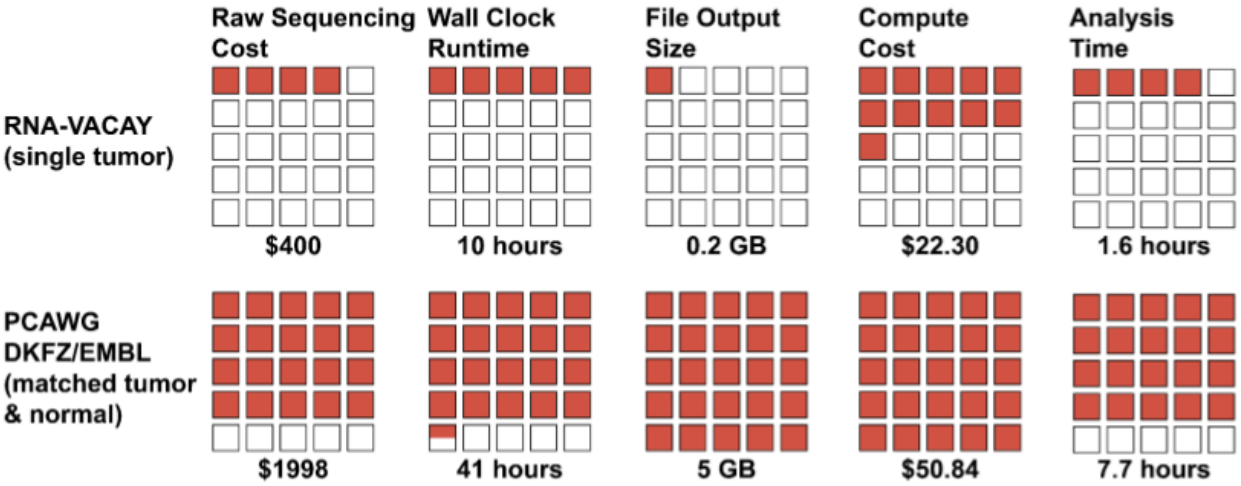








Extended Data Fig. 5 | Potential driver genes contain variants found in lowly expressed genes. a, Quantile-quantile (QQ) plots of p-values generated from oncodriveFML in skin cutaneous melanoma (SKCM). **b,** Density plot of gene expression of top WGS driver gene candidates in SKCM samples.



Extended Data Fig. 6 | RNA-VACAY lowers the cost of variant calling. Variant calling with RNA-seq data significantly lowers time and cost constraints of a randomly selected sample compared to WGS sequencing.