The Journal of Molecular Diagnostics, Vol. 21, No. 2, March 2019

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

Somatic Tumor Variant Filtration Strategies to Optimize Tumor-Only Molecular Profiling Using Targeted Next-Generation Sequencing Panels

Mahadeo A. Sukhai,* Maksym Misyura,* Mariam Thomas,* Swati Garg,* Tong Zhang,* Natalie Stickle,* Carl Virtanen,* Philippe L. Bedard,†‡ Lillian L. Siu,†‡ Tina Smets,§ Gert Thijs,§ Steven Van Vooren,§ Suzanne Kamel-Reid,*¶‖** and Tracy L. Stockley*¶‖

From the Advanced Molecular Diagnostics Laboratory, Princess Margaret Cancer Centre, the Division of Medical Oncology and Hematology,† and the Department of Clinical Laboratory Genetics,¶ Laboratory Medicine Program, University Health Network, Toronto, Ontario, Canada; the Departments of Medicine,‡ Laboratory Medicine and Pathobiology,‖ and Medical Biophysics,** University of Toronto, Toronto, Ontario, Canada; and Aglient Technologies,§ Santa Clara, California*

Accepted for publication September 28, 2018.

Address correspondence to Tracy L. Stockley, Ph.D., F.C.C.M.G., F.A.C.M.G., Department of Clinical Laboratory Genetics, University Health Network, 200 Elizabeth St., 11E-454, Toronto, ON M5G 2C4. E-mail: tracy.stockley@uhn.ca.

A common approach in clinical diagnostic laboratories to variant assessment from tumor molecular profiling is sequencing of genomic DNA extracted from both tumor (somatic) and normal (germline) tissue, with subsequent variant comparison to identify true somatic variants with potential impact on patient treatment or prognosis. However, challenges exist in paired tumor-normal testing, including increased cost of dual sample testing and identification of germline cancer predisposing variants. Alternatively, somatic variants can be identified by *in silico* tumor-only variant filtration precluding the need for matched normal testing. The barrier to tumor-only variant filtration is defining a reliable approach, with high sensitivity and specificity to identify somatic variants. In this study, we used retrospective data sets from paired tumor-normal samples tested on small (48 gene) and large (555 gene) targeted next-generation sequencing panels, to model algorithms for tumor-only variants classification. The optimal algorithm required an ordinal filtering approach using information from variant population databases (1000 Genomes Phase 3, ESP6500, ExAC), clinical mutation databases (ClinVar), and information on recurring clinically relevant somatic variants. Overall the tumor-only variant filtration strategy described in this study can define clinically relevant somatic variants from tumor-only analysis with sensitivity of 97% to 99% and specificity of 87% to 94%, and with significant potential utility for clinical laboratories implementing tumor-only molecular profiling. *(J Mol Diagn 2019, 21: 261—273; https://doi.org/10.1016/j.jmoldx.2018.09.008)*

A significant challenge in tumor-only molecular profiling is defining somatic tumor-specific variants in the background of germline variants also detected during sequencing. To distinguish somatic tumor variants from germline variants, two major approaches have emerged: parallel testing of a normal germline DNA sample, typically from blood or adjacent normal tissue from formalin-fixed, paraffin-embedded (FFPE) samples, with subtraction of germline variants from the total variants identified in the tumor tissue to produce a list of somatic tumor-only variants[1,2] or tumor-only analysis, with *in silico* variant filtration using available databases and other resources to prioritize variants likely to be somatic.[3]

In the clinical molecular diagnostic laboratory, parallel testing of germline samples to classify somatic tumor variants by tumor-normal comparison raises issues for practice. The additional expense of testing both the germline and tumor sample from each patient is significant. In addition,

Supported by the Princess Margaret Foundation (S.K.-R., T.L.S.) and Genome Canada (S.K.-R., T.L.S.).

Disclosures: G.T., S.V.V., and T.S. were employees of Agilent Technologies at the time of manuscript preparation, and G.T. and S.V.V. hold stock options in Agilent Technologies.

Current address of M.A.S., Department of Research, Canadian National Institute for the Blind, Toronto, Ontario, Canada.

https://doi.org/10.1016/j.jmoldx.2018.09.008

the potential identification of inherited germline cancer predisposing variants in normal samples requires appropriate management, including pretest consent, germline variant–specific investigations, and appropriate return of results with genetic counseling support.[4]

Tumor-only testing for identification of somatic variants involves *in silico* filtering of variants against multiple information sources, including laboratory or online databases that contain information on somatic or germline variants, published literature, and protein prediction tools. Collectively, the information in these databases is limited by variable accuracy, hindering the approach.[5–7] For example, germline polymorphic variant databases, such as Single Nucleotide Polymorphism Database (dbSNP), contain pathogenic germline variants,[8–10] whereas somatic variant databases, such as Catalogue of Somatic Mutations in Cancer (COSMIC), contain germline variants.[5] A recent publication assessed whole exome sequencing data filtering to classify variants with the use of paired and unpaired approaches. The study defined an improved set of criteria for successful discrimination of somatic and germline variants; however, optimal sensitivity and specificity were achieved by use of a paired normal sample.[11]

To minimize the potentially significant impact on patient care of inaccurately defined somatic variants[1,2] and to improve *in silico* assessment of somatic variants, we present development and validation of an optimized tumor-only variant filtration strategy for targeted panel tumor molecular profiling. The variant filtration algorithm was developed by using variant results from a medium-sized (48 gene) targeted next-generation sequencing (NGS) panel tested on 1120 tumor FFPE and matched blood samples. The tumor-only filtering algorithm was also tested on a variant data set from 53 pairs of samples (tumor FFPE and blood samples from the same patients) tested on a large (555 genes) targeted NGS panel. The tumor-only filtering algorithm was found to have 99% sensitivity and 94% specificity for detection of somatic variants from medium NGS panels, and 97% sensitivity and 87% specificity from large NGS panels, without the need for testing of matched normal samples. The impact of including commonly used germline and somatic databases and other tools was also modeled on variant filtration, and the optimal variant filtration approach defined for targeted NGS panel tumor-only variant analysis.

Materials and Methods

Samples and NGS Profiling

Variant data were generated from DNA samples extracted from FFPE tumor specimens (biopsies or surgical resections) or peripheral blood lymphocytes (PBLs). Samples were collected under a University Health Network Research Ethics Board–approved study as previously described.[12] Tumor regions of FFPE specimens were acceptable if tumor cellularity was ≥20%, and tumor isolated by 1 to 2 × 1 mm punch from FFPE blocks or macrodissection of unstained material from 15 to 20 slides (4 to 7 μm sections), and DNA extracted from FFPE or PBLs as described.[12] DNA samples (n = 1120) from FFPE tumor tissue (Supplemental Figure S1) (described in Stockley et al[12]) and matched DNA samples from PBLs (ie, 2240 samples from 1120 participants) were tested (250 ng of DNA for library preparation) with the targeted hotspot TruSeq Amplicon Cancer Panel (TSCAP; Illumina, San Diego, CA) that covered regions of 48 genes[12] on the MiSeq sequencer (Illumina). The TSACP was selected because it is a well-used and well-validated commercially available targeted panel that was found to yield high-quality sequencing data.[12] The UHN Hi5 was also similarly designed and validated as a high-quality clinical panel.

One hundred six matched FFPE-PBL DNA samples (from 53 patients) were also tested by using a validated custom capture 555-gene panel (UHN Hi5) that covered exons and minimum of 10 bp of flanking intronic region of 555 cancer-related genes (Supplemental Table S1). Libraries were constructed from 250 ng of DNA sheared by sonication (Covaris, Woburn, MA), with end repair and ligation with barcoded sequencing adaptors, followed by hybrid capture with RNA baits (SureSelect; Agilent, Santa Clara, CA) and sequencing on the NextSeq (Illumina).

Variant Analysis

For TSCAP, sequence alignment and variant calling for all samples were performed with MiSeq Reporter software version 2.3.1 (Illumina), and variants were reviewed with the Integrative Genomics Viewer (Broad Institute, Cambridge, MA). Somatic variants included for analysis passed MiSeq Reporter quality filter and met laboratory-defined thresholds of ≥250× read depth and >5% variant allele fraction (VAF) in DNA from FFPE tissue. Germline variants included for analysis met thresholds of >50× read depth in DNA from PBLs. Three genes with read depth consistently <250× on TSACP (*GNAS*, *HRAS*, *CDKN2A*) were not included in the analysis.

For UHN Hi5, a custom bioinformatics pipeline was used, including alignment to the human genome reference (build GRCh37/hg19) with the use of the Burrows-Wheeler Aligner,[13] marking duplicate reads by using Picard Mark Duplicates, and application of the Genome Analysis Toolkit base quality score recalibration and insertion and deletion realignment algorithms according to Genome Analysis Toolkit Best Practices recommendations.[13,14] Somatic single nucleotide variant discovery was performed with MuTect.[15] Somatic variants included for analysis met thresholds of ≥250× read depth and ≥5% VAF in DNA from FFPE tissue.

Variant Filtration Algorithm Design

Tumor-only variant analysis used Alissa Interpret (Agilent; formerly Cartagenia Bench NGS; version 4.2) with a custom

designed decision tree filtration algorithm. Alissa Interpret's decision tree label approach was used, with label-based data set filtration applied after analysis to allow for counting of variants in each phase of the algorithm.

The tumor-only filtration algorithm used on TSACP data was designed in three phases (Figure 1). Phase I included quality filtration of candidate variants, with variants retained for further analysis only if they were nonsynonymous variants located within exons or the intronic 2 bp encoding canonical splice sequence, with a quality pass flag as generated by MiSeq Reporter (a high-quality variant with no strand bias), a read depth of $\geq$250$\times$ (somatic) at the specific variant, and $\geq$5% VAF. We also constructed a clinically relevant custom managed variant list (CR-MVL) within Alissa Interpret. The CR-MVL was derived from manual curation of variants identified from paired tumor-normal sequencing of >2500 solid tumor samples, using the Illumina TSACP, within the UHN AMDL from 2012 to 2016. Classification of variants as clinically relevant was performed by using our previously published approach and classification scheme.[16] The CR-MVL was used in phase I to retain any clinically actionable variant with <5% VAF.

In phase II, variants retained after phase I were labeled as germline if they were present in any of the following four germline population variant databases (PVDs) at a minor allele frequency (MAF) of $\geq$1%: 1000 Genomes phase 3 (release version 5.20130502[17,18]); Exome Sequencing Project (ESP; ESP6500SI-V2 data set of Exome Variant Server, National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project, Seattle, WA; *http://evs.gs.washington.edu/EVS*, last accessed August 2016), Exome Aggregation Consortium version 0.3 (ExAC)[19] or dbSNP build 141 (GRCh37.p13).[10] Variants not found in any of the PVDs at the $\geq$1% MAF threshold were labeled as somatic. Also in phase II, variants were labeled as somatic if they were found in the COSMIC database release version 71,[20] with two or more occurrences, even if also found in a germline population database. Finally, in phase III, variants were labeled germline if they were present as a benign or likely benign variant in either the Human Gene Mutation Database (HGMD; HGMD Professional Database 2015.1; *http://www.hgmd.cf.ac.uk/ac/index.php*)[21] or in ClinVar (National Center for Biotechnology Information ClinVar; 20150504; *https://www.ncbi.nlm.nih.gov/clinvar*).[22]

All variant classifications (germline or somatic) from the tumor-only filtration algorithm were compared with matched tumor-normal sample analysis output, whereby variants occurring in DNA from both the FFPE tumor and the matched normal PBLs were considered germline and

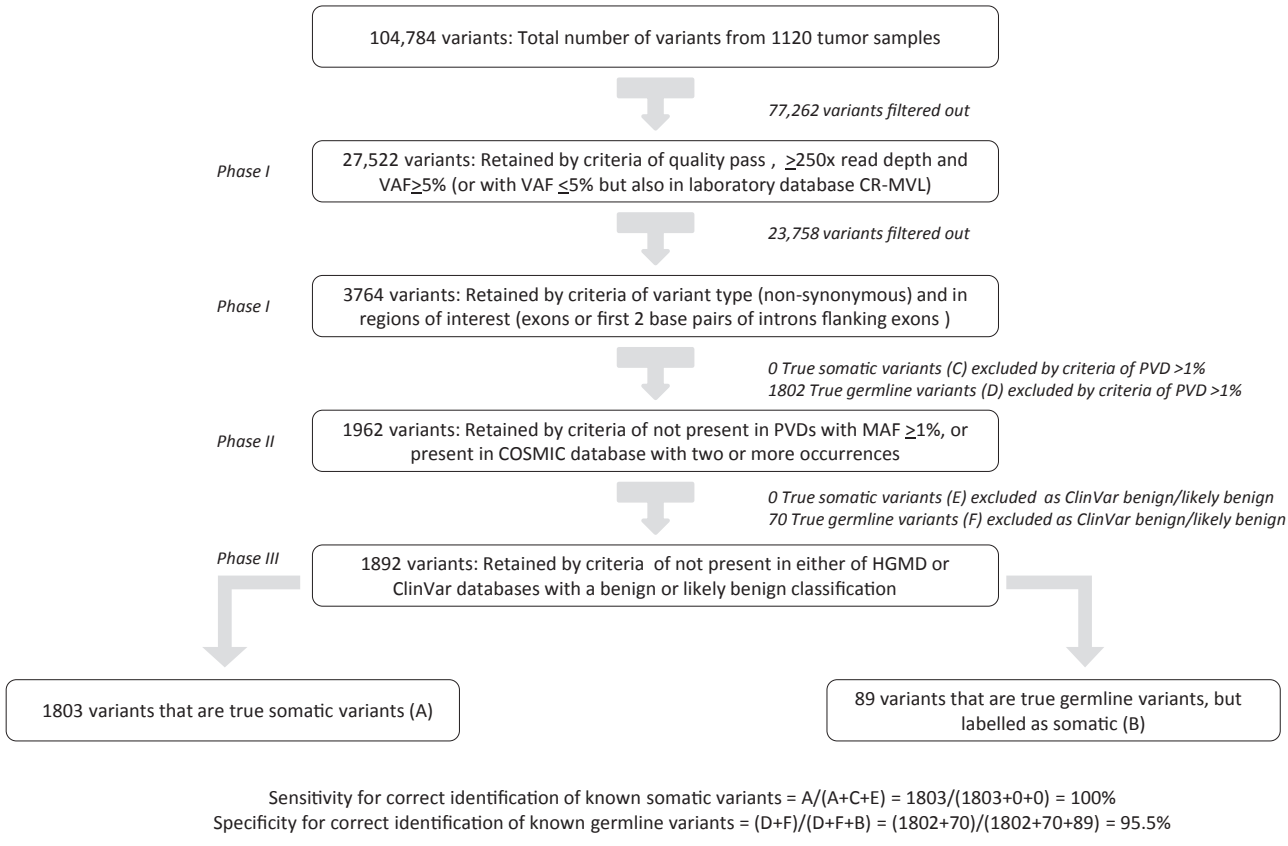


Figure 1 Flowchart illustrating the use of paired tumor-normal variant analysis to verify the *in silico* prediction of somatic and germline variants via tumor-only variant analysis in Alissa Interpret. ClinVar, ClinVar variant database; COSMIC, Catalogue of Somatic Mutations in Cancer; CR-MVL, laboratory clinically relevant managed variant list; HGMD, human gene mutation database; MAF, minor allele frequency in population variant databases; PVD, population variant database; VAF, variant allele frequency.

removed from the tumor variant list. Tumor-normal analysis was conducted within Alissa Interpret by using a second custom filtration algorithm tree that compared variants from the TSACP variant call files from the PBL and FFPE samples for each case (tumor-normal matched sample algorithm not shown).

Comparison of the results from both analyses enabled classification of variants from the tumor-only analysis as true somatic call (TSC), true germline call (TGC), false somatic call (FSC), and false germline call (FGC) in the context of somatic variant classification by the tumor-only variant filtration algorithm (Table 1). Sensitivity, specificity, positive predictive value, and negative predictive value (NPV) (Table 1) were used for performance evaluation of the tumor-only variant filtration algorithm for classification of somatic variants.

Modeling of Variant Filtration Algorithm

To model the ideal tumor-only variant detection algorithm, the accuracy of variant classification was assessed by databases within the tumor-only filtration algorithm via recursive partitioning using the party package (*https://cran.r-project.org/web/packages/party/index.html*, last accessed February 23, 2017) in the R coding environment. Recursive partitioning considered the following independent dichotomous variables: PVD MAF ($\geq$1% or $<$1%), present in COSMIC (two or more occurrences) or not present in COSMIC (one or no occurrences), present or absent in the Clinical Interpretations of Variants in Cancer (CIViC) database,[23] reported as benign/likely benign or as pathogenic/likely pathogenic in ClinVar, present or not in HGMD, predicted as benign or predicted as damaging in missense mutation protein effect prediction algorithms (SIFT[24,25]; PolyPhen2 HumDiv or HumVar[26]; Mutation Taster[27]; Mutation Assessor[28]; LRT[29]; PROVEAN[30]), and present or not in our internal laboratory CR-MVL. The performance of full (all variables) and minimal (three selected variables) filtration algorithms was evaluated to determine the most optimal approach for building tumor-only classification filtration approaches.

Results

Development of Tumor-Only Variant Filtration Algorithm

To evaluate the tumor-only variant filtration algorithm, an existing data set of 104,784 candidate variants from 1120 matched tumor FFPE and normal PBL sample DNA pairs

Table 1 Definitions as Applied to Variants Classified by the Tumor-Only Filtration Algorithm and Approach to Calculations of Sensitivity and Specificity

Term	Definition	Classification from tumor-normal variant comparison	Classification from tumor-only filtration algorithm
A = True somatic call (TSC)	Variants classified as somatic by tumor-only filtration algorithm and known to be true somatic variants from tumor-normal comparison	Somatic	Somatic
B = True germline call (TGC)	Variants classified as germline by tumor-only filtration algorithm and known to be true germline variants from tumor-normal comparison	Germline	Germline
C = False somatic call (FSC)	Variants classified as somatic by the tumor-only filtration algorithm but known to be true germline variants from tumor-normal comparison	Germline	Somatic
D = False germline call (FGC)	Variants classified as germline by the tumor-only filtration algorithm but known to be true somatic variants from tumor-normal comparison	Somatic	Germline

Term	Definition	Formula
Sensitivity	Proportion of true somatic variants (A) correctly identified by the tumor-only filtration algorithm as somatic out of all true somatic variants (A + D)	A/A + D
Specificity	Proportion of true germline variants (B) correctly identified as germline by the tumor-only filtration algorithm out of all true germline variants (B + C)	B/B + C
Positive predictive value (PPV)	Proportion of true somatic variants (A) correctly identified by the tumor-only filtration algorithm as somatic out of all variants called somatic by the algorithm (A + C)	A/A + C
Negative predictive value (NPV)	Proportion of true germline variants (B) correctly identified as germline by the tumor-only filtration algorithm out of all variants called germline (B + D)	B/B + D

was used and tested by a targeted 48-gene NGS panel (TSACP[12]) (Figure 1). Variants were classified as true somatic or true germline to evaluate the sensitivity and specificity of the tumor-only variant filtration algorithm for identification of somatic variants from tumor-only testing. After candidate variants with low VAF and low coverage were excluded, 27,522 variants remained, of which 3764 met the retention criteria of nonsynonymous variants in regions of interest (exons or first 2 bp of introns flanking exons). The tumor-only variant filtration algorithm classified 1892 variants as somatic and 1802 as germline. The 1892 variants classified as somatic contained 100% of the true somatic variants (1803 variants; TSC) known from the tumor-normal (FFPE-PBL) variant comparison, but over-called true germline variants as somatic, with 4.7% of true germline variants (89 of 1892; FSC) misclassified as somatic. Conversely, of 1961 true germline variants, 95.5% (1872 of 1961; TGC) were correctly identified as germline by the tumor-only variant filtration algorithm, with the remaining 4.5% (89 of 1961; FSC) of true germline variants misclassified as somatic. Overall, the tumor-only variant filtration algorithm provided a sensitivity for correct classification of true somatic variants as somatic of 100% (1803 of 1803), and a specificity for correct classification of true germline variants as germline of 95.5% (1872 of 1961).

Modeling Application of PVDs to Variant Triage

Because the initial tumor-only variant filtration algorithm contained a specific order of filters and an MAF threshold of 1% from population databases, improvements to the algorithm model were evaluated by altering its parameters (Figure 2).

Germline PVDs were first assessed for their contribution as exclusionary filters singly and in combination. ExAc,

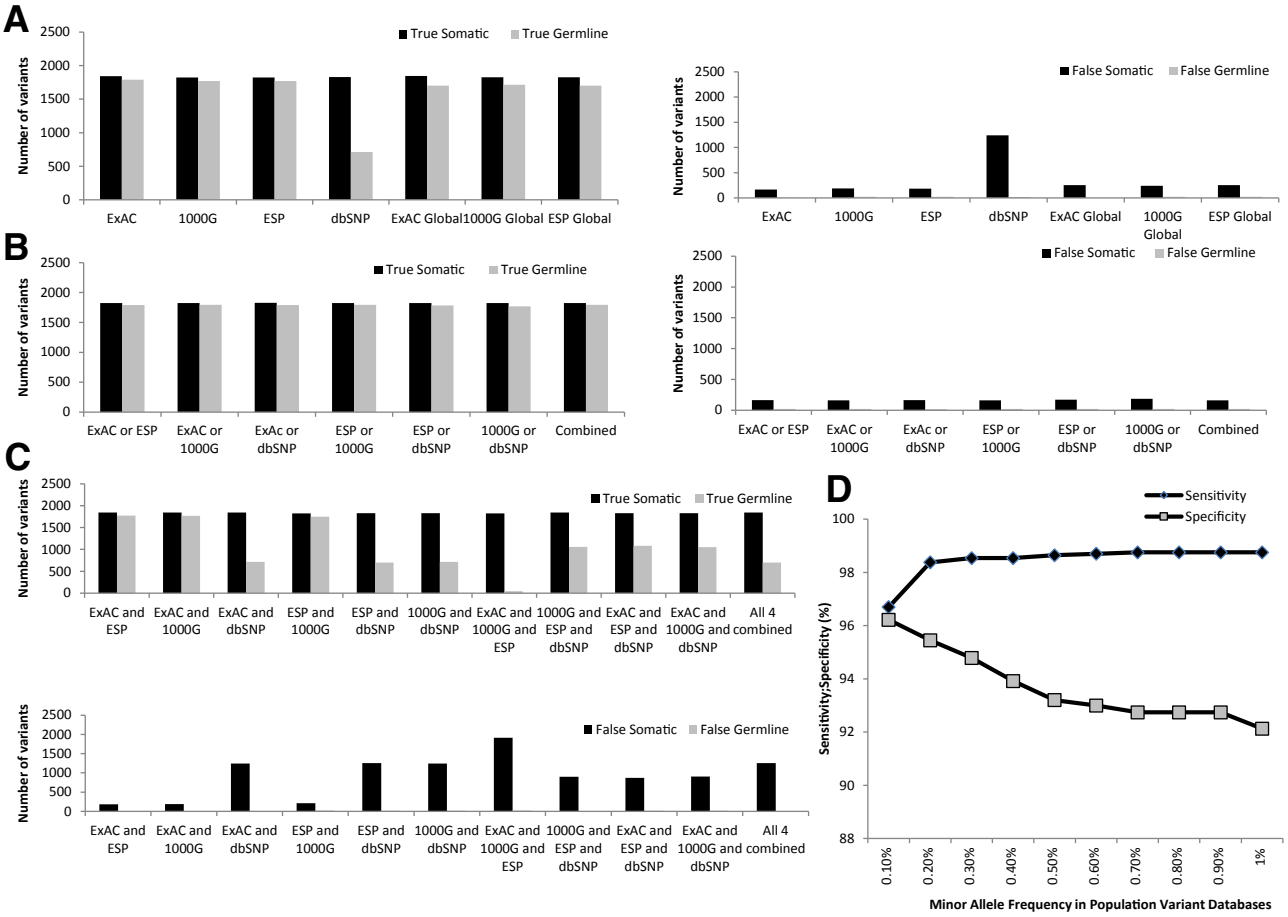


Figure 2 Application of population variant databases as exclusionary filters in variant triage. **A:** Application of population variant databases as exclusionary filters in variant triage, where presence of a variant in any subpopulation data set for ≥1% minor allele frequency in Exome Aggregation Consortium (ExAC), 1000 genomes phase 3 (1000G), and Exome Sequencing Project 6500 (ESP), was examined. Both global data sets and subpopulation data sets were evaluated for ExAC, 1000 Genomes Phase 3, and ESP6500 (no subpopulation data sets for available for dbSNP). **B:** Application of ExAC, 1000G, ESP, and Database of Single Nucleotide Polymorphisms (dbSNP) as exclusionary filters, when used in pairwise combination (OR condition) or applied simultaneously as a collective of four databases (Combined). Variants were excluded if they were present at minor allele frequency (MAF) ≥1% in at least one of the databases indicated for each group. **C:** Application of ExAC, 1000G, ESP, and dbSNP as exclusionary filters, when used in pairwise combination (AND condition) or applied simultaneously as a collective of four databases (Combined). Variants were excluded if they were present at MAF ≥ 1% in all of the databases indicated for each group. **D:** Evaluation of impact on sensitivity and specific by use of population variant database MAF threshold values from 0.1% to 1%. Sensitivity and specificity were calculated as described in *Materials and Methods*.

1000 Genomes Phase3, and ESP6500, when used alone with a threshold of 1% MAF, had high sensitivity in identifying TSCs (98.8% to 99.8%) and identified minimal FSC (90.4% to 91.6% specificity) (Figure 2A and Supplemental Table S2). However, dbSNP performed poorly with a high number of FSCs (specificity 36.4%). The effect of including versus excluding subpopulation analyses, available within 1000 Genomes Phase 3, ESP6500 and ExAC, was also considered in population variant triage. Excluding subpopulation analysis yielded minor gains in sensitivity (a benefit of 0.2% to 0.6%) but with a more significant adverse effect on specificity (specificity reduced by 2.7% to 4.5%) (Figure 2A and Supplemental Table S2).

The effect of combining PVDs, in groups of two, three, or four, while retaining a MAF threshold of $\geq$1%, was examined. Combinations in which variants were considered germline if identified in at least one database tested (Figure 2B) and combinations in which variants were considered germline if identified in every database tested in the group (Figure 2C) were evaluated. Combining two or more PVDs using OR conditions yielded similar results, regardless of combination tested (Figure 2B). The two-way combination of 1000 Genomes Phase 3 and ESP6500 or the two-way combination of ExAC and 1000 Genomes Phase 3 yielded the highest sensitivity and specificity (sensitivity: 98.8%; specificity: 91.7%) (Figure 2B and Supplemental Table S2). Strikingly, AND conditions only slightly improved sensitivity (98.8% to 99.9%) (Figure 2C and Supplemental Table S2) but with a significant reduction in specificity (2.30% to 90.5%) compared with OR conditions (Figure 2B and Supplemental Table S2). Because of the relatively poor ability of dbSNP to correctly classify germline variants (Figure 2A), combinations excluding dbSNP performed best. Taken together, these analyses (Figures 2, A–C) demonstrated that the PVD filter most effective at segregating TSCs from TGCs was defined by variant presence in at least one of ExAC, ESP6500, or 1000 Genomes phase 3, with subpopulation data sets included for analysis.

Finally, the effect of altering the MAF threshold (0.1% to 5%) in PVDs (ie, a variant was considered germline if it met the assigned MAF threshold in at least one of four PVDs) was tested. A MAF cutoff of 5% was established in variant interpretation guidelines for triaging benign population variants in the germline context.[31] Applying a 5% MAF threshold to classification of somatic variants yielded a sensitivity of 98.9%, with a specificity of 87.7% (not shown). A MAF cutoff of 1%, commonly accepted as the threshold for triage of benign population variants,[32] yielded minimal difference in comparison with the 5% MAF threshold (98.8% sensitivity, 92.1% specificity) (Figure 2D and Supplemental Table S2). Decreasing the MAF threshold to 0.2% had no significant impact on sensitivity (98.8%), whereas specificity improved to 95.5%. Sensitivity decreased to 96.7% at 0.1% MAF, although specificity improved (96.2%) (Figure 2D and Supplemental Table S2). Therefore, using a PVD MAF threshold of 0.2% yielded optimal specificity for correct classification of TGCs while maintaining sensitivity in classifying TSCs.

Modeling Application of PFPAs to Variant Triage

Protein functional prediction algorithms (PFPAs; eg, SIFT, PolyPhen HumDiv, PolyPhen HumVar, Mutation Taster, Mutation Assessor, LRT, and PROVEAN) are commonly used in variant assessment. Although the initial tumor-only variant filtration algorithm did not include their use, the utility of PFPAs singly and in combination in classifying benign versus deleterious missense variants in a tumor-only context was tested. An inclusionary analysis was conducted to identify variants that would be kept for further analysis as potential somatic variants (Supplemental Figure S2A). An exclusionary analysis to identify variants labeled as benign, which would be excluded from further analysis as potential germline variants, was also performed (Supplemental Figure S2B).

During the inclusionary analysis all seven PFPAs, when used alone, had poor sensitivity [2.4% (Mutation Assessor) to 39.4% (Mutation Taster)] (Supplemental Figure S2A and Supplemental Table S2) but relatively good specificity [71.3% (SIFT) to 100% (Mutation Assessor)]. Given the poor performance of PFPAs, the effect of agreement among combinations of five of seven, six of seven, or seven of seven PFPAs was tested in the inclusionary analysis. This approach did not yield better results than use of individual PFPAs; although specificity remained high (98.3% to 100%), sensitivity decreased: 21.7% (agreement among five of seven algorithms) to 14.6% (six of seven algorithms) to 1.5% (seven of seven algorithms) (Supplemental Figure S3 and Supplemental Table S2).

Conversely, during the exclusionary analysis, all seven algorithms when used alone had good performance to classify TGC [87.0% (PROVEAN) to 97.4% (Mutation Taster)] but poor specificity [14.8% (Mutation Assessor) to 44.8% (PROVEAN)] (Supplemental Figure S2B and Supplemental Table S2). A combinatorial approach did not yield better results than use of individual PFPAs. Although sensitivity remained high (96.2% to 99.1%), specificity was poor when assessing agreement among five of seven, six of seven, or seven of seven PFPAs (9.9% to 14.1%).

Modeling Application of Publicly Available Clinically Relevant Variant Databases to Variant Triage

The utility of applying information from disease-associated variant databases to the tumor-only variant filtration algorithm was also evaluated. Each of the following were assessed as filters for classification of somatic variants: presence of a variant in COSMIC (at least two occurrences), absent in HGMD, or absent in ClinVar as a benign or likely benign variant (Figure 3). The ClinVar exclusionary filter performed best in enabling the identification of somatic

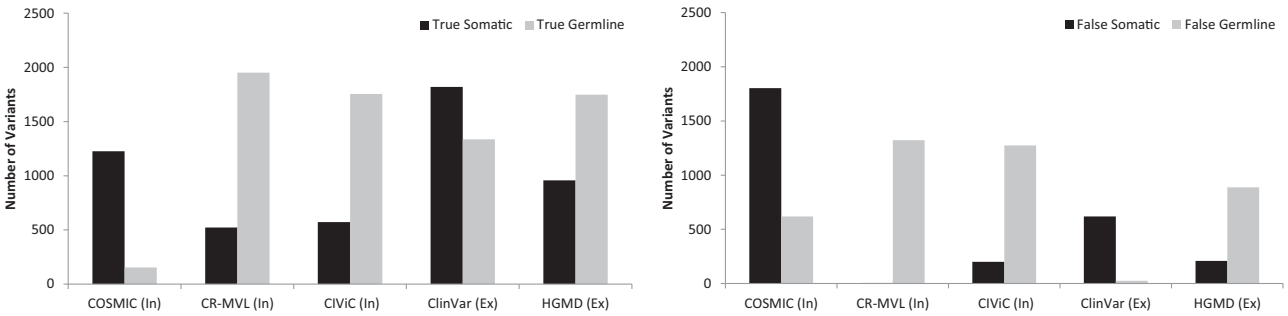


Figure 3 Application of publicly available variant databases within variant triage. Cancer-associated variant databases of Catalogue of Somatic Variants in Cancer (COSMIC), Clinical Interpretations of Variants in Cancer (CIViC), or an internal laboratory variant database (clinically relevant custom managed variant list; CR-MVL) were used as inclusionary filters (In), whereas germline variant databases [human gene mutation database (HGMD); ClinVar] were used as exclusionary filters (Ex).

variants, with fewest FGCs compared with the inclusionary COSMIC and exclusionary HGMD filters. However, HGMD performed best in identifying germline variants, with fewest FSCs compared with ClinVar or COSMIC. Because the exclusionary ClinVar filter performed the best in this analysis, it was hypothesized that use of COSMIC and HGMD would not be beneficial in an optimized variant filtration algorithm. Modeling a range of COSMIC occurrence thresholds did not yield improved sensitivity or specificity as evaluated by receiver operating characteristic analysis (Supplemental Figure S4). A true positive rate of >90% resulted in a false positive rate of >55% and was deemed insufficiently stringent compared with other filtering strategies and databases under evaluation.

Modeling Application of a Laboratory-Curated Actionable Variants List to Variant Triage

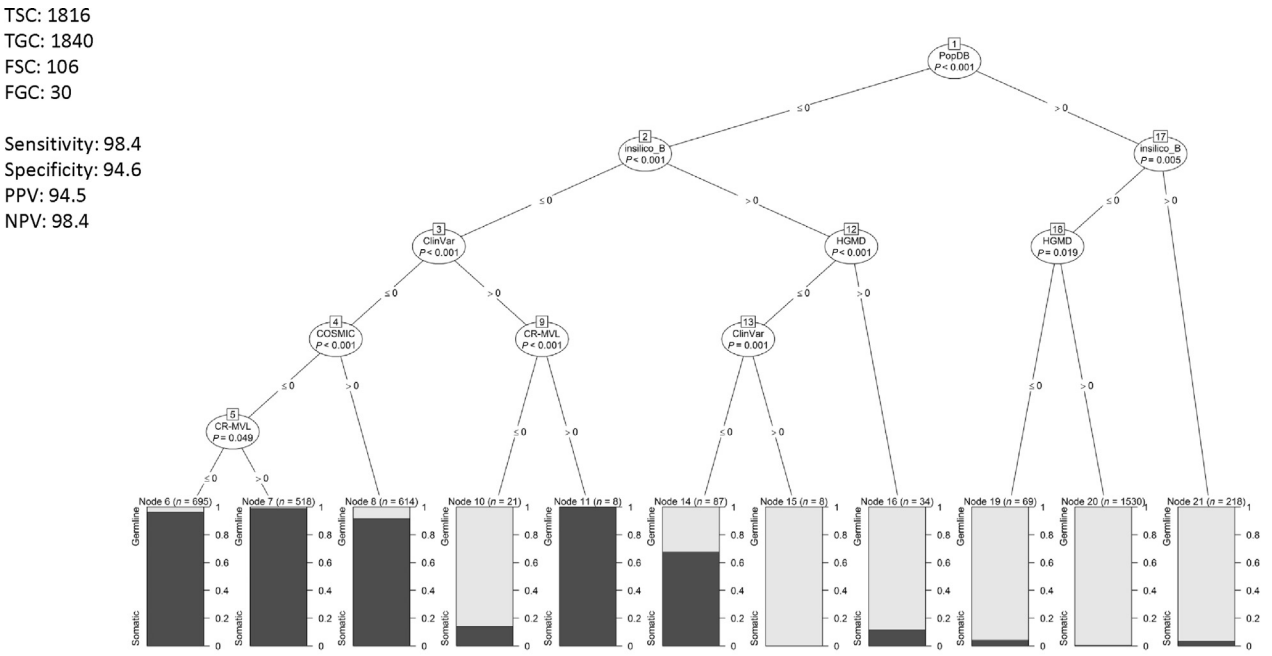
Many clinical laboratories accumulate private data on clinically relevant variants identified in previously tested cases. Alternatively, they use publicly available data sets of clinically relevant variants such as My Cancer Genome (*mycancergenome.org*[33]) or CIViC[23] to inform interpretation. In initial algorithm development, an approach that did not use such a data set, which is more reflective of laboratories beginning NGS tumor-only diagnostic testing, was assessed. The potential benefit of using curated lists of clinically relevant variants as a component of the tumor-only variant triage was determined. Two approaches were examined. First, the benefit of CIViC was assessed in the variant triage algorithm; variants present in CIViC were retained as potential somatic variants. Addition of CIViC had a sensitivity of 31.0% to identify TSCs and a specificity of 89.7% to exclude TGCs (Figure 3 and Supplemental Table S2). The sensitivity was low because CIViC contained a relatively limited number of somatic variants, and the specificity, although high, was <100% because of the occurrence of known benign variants in the CIViC data set.

Second, the utility of a list of recurrent known actionable variants (diagnostic, prognostic, predictive of response to therapy and/or druggable) previously identified in testing of solid tumors with the use of the TSACP panel (CR-MVL; *Materials and Methods*) in the tumor-only filtration algorithm was examined. Addition of the CR-MVL for classifying TSCs had a sensitivity of 28.3% and a specificity of 99.7% to exclude TGCs (Figure 3 and Supplemental Table S2). The sensitivity was slightly lower than CIViC, given the relative sizes of the variant lists (CR-MVL compared with CIViC). The specificity was significantly higher than with CIViC but was <100% because of the real occurrence of known actionable somatic variants that are also pathogenic in the germline context. For example, certain variants in *TP53* can occur as somatic tumor-only variants, but also occur as germline inherited variants that cause Li-Fraumeni syndrome.

Proposed Filtration Algorithm for Triage of Variants Identified in Tumor-Only Analysis

A series of sequential filter combinations were tested to identify the optimal algorithm for classification of somatic variants from tumor-only NGS profiling data (Figure 4). It was important to evaluate filter order to minimize the probability of FSC/FGC variants when starting with the largest data set.

The optimal combination of tools to classify TSCs used the internal laboratory CR-MVL (inclusionary), presence in ClinVar as a benign or likely benign germline variant (exclusionary), and presence in one of three global germline PVDs (1000 Genomes Phase 3, ESP6500, ExAC; exclusionary) (optimal combination modeled in Figure 5A). Inclusion of COSMIC and/or HGMD and/or PFPAs negatively affected sensitivity and specificity or, at best, provided no benefit (Figure 4). It was next assessed whether filtration order was relevant, by calculating TSC, TGC, FSC, and FGC for all six ordered combinations of the three most useful filters. The three best combinations of filtration tools yielded a sensitivity of 98.5% (1819 TSCs of 1819 TSCs + 27 FGCs) (Figure 5B) and a specificity of 93.7% (1834 TGCs of 1834 TGCs + 122 FSCs) (Figure 5B). Filtration order was relevant; some combinations performed worse than the optimal three combinations (Figure 5B). The



Review	YES	YES	YES	NO	YES	YES	NO	NO	NO	NO	NO
TSC	672	513	564	NA	8	49	NA	NA	NA	NA	NA
TGC	NA	NA	NA	18	NA	NA	8	30	66	1518	210
FSC	23	5	50	NA	0	28	NA	NA	NA	NA	NA
FGC	NA	NA	NA	3	NA	NA	0	4	3	12	8

Figure 4 Analysis of filter combinations for variant filtration algorithms. Decision/classification trees were constructed in the R coding environment by using recursive partitioning and using all available variables [minor allele frequency, Human Gene Mutation Database (HGMD), Catalogue of Somatic Variants in Cancer (COSMIC), ClinVar, *in silico* prediction algorithms, and internal laboratory variant database; *Materials and Methods*]. True somatic call (TSC), true germline call (TGC), false somatic call (FSC), and false germline call (FGC) values were calculated for each branch of the tree, as shown, as well as collectively, depending on whether variants in that branch were flagged for review based on the combination of the established filtering criteria. CR-MVL, clinically relevant custom managed variant list; NA, not applicable; NPV, negative predictive value; PopDB, population variant database; PPV, positive predictive value.

distribution of TSCs, FSCs, TGCs, and FGCs by gene is shown in Supplemental Figure S5; FSCs and FGCs were not grouped within specific patients.

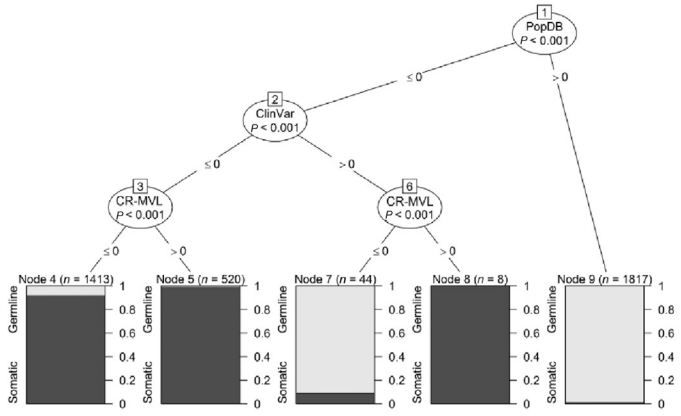
Performance of Filtration Algorithm on Larger NGS Panel

To test whether the tumor-only filtration algorithm was extendable to larger NGS panels, its performance was analyzed on variants from a cohort of 53 tumor and matched blood samples (106 samples) tested by a laboratory-developed 555-gene panel (UHN Hi5). A sensitivity of 97.3% (854 TSCs of 854 TSCs + 24 FGCs) for detecting true somatic variants, but a reduced specificity of 67.7% (2950 TGCs of 2950 TGCs + 1405 FSCs) (Figure 6) were obtained.

Our filtration algorithm was adapted to include a more stringent population variant threshold (MAF = 0.2%) and to remove technical artifacts. In the revised algorithm, panel-specific artifacts (ie, variants recurring in >20% of samples tested on this panel in our laboratory) and not clinically relevant (absent from our CR-MVL) were excluded from further analysis. This improved specificity to 86.9% (2950 TGCs of 2950 TGCs + 443 FSCs) (Figure 6), achieved by excluding 962 of 1405 FSCs, while still producing a sensitivity of 97.1% (808 TSCs of 808 TSCs + 24 FGCs). The remaining 443 FSCs were manually investigated for germline characteristics independent of the filtration algorithm, specifically high allele count in ExAC, absence in COSMIC, VAF, coverage, and for potential artifacts. Only 22 of 443 variants (14 unique) were present in COSMIC with ≥2 occurrences, with the remainder (421 of 443 variants) absent from COSMIC or present only once. Conversely, using the revised filtration algorithm, 46 known somatic variants were excluded, of which 29 were excluded as artifacts and 17 because of the presence in PVDs with MAF ≥0.2% (Supplemental Table S3). None are currently known to be clinically relevant in the somatic context.

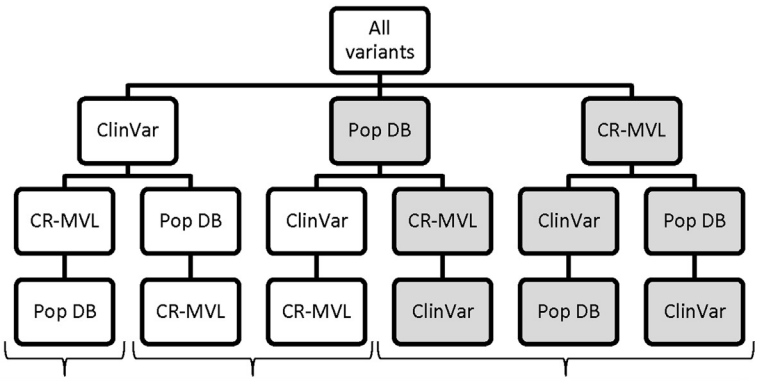
A significant number of FSCs on the Hi5 panel were technical artifacts present in data from both FFPE and PBL samples and were considered as known germline. Of note,

A



Review	YES	YES	NO	YES	NO
TSC	1296	515	NA	8	NA
TGC	NA	NA	40	NA	1794
FSC	117	5	NA	NA	NA
FGC	NA	NA	4	NA	23

B



TSC	1811	515	1819
FSC	122	5	122
TGC	1834	1951	1834
FGC	35	1331	27
Sensitivity	98.1	27.9	98.5
Specificity	93.7	99.7	93.8
PPV	93.7	99.0	93.7
NPV	98.1	59.5	98.6

Figure 5 Optimal variant filtration algorithms. Decision/classification trees were constructed in the R coding environment by using recursive partitioning and using minor allele frequency, ClinVar, and laboratory-curated list (*Materials and Methods*). **A:** True somatic call (TSC), true germline call (TGC), false somatic call (FSC), and false germline call (FGC) values were calculated for each branch of the tree, as shown, as well as collectively, depending on whether variants in that branch were flagged for review based on the combination of the established filtering criteria. **B:** Modeling of the effect of variant filtration order on identification of true somatic variants. The number of TSC, TGC, FSC, and FGC variants were assessed when varying the order of the three optimal variant filtration criteria: ClinVar benign (ClinVar; exclusionary filter); population variant databases (PopDB; exclusionary filter); internal laboratory variant knowledgebase [clinically relevant custom managed variant list (CR-MVL); inclusionary filter]. Of the six possible ordinal combinations of filters, the three combinations with gray fill yielded the highest number of true somatic call variants. NA, not applicable; NPV, negative predictive value; PPV, positive predictive value.

24 known somatic variants were misclassified as germline (FGCs) by the use of the filtration criteria of present in ClinVar within the category of a benign/likely benign variant. Sixteen of the 24 variants were present in PVDs or subpopulations with a MAF of >1% and may represent clinically neutral somatic changes (Supplemental Table S4). Five of the 24 variants were known or suspected artifacts on other hybrid capture panels within our laboratory and may represent sequencing artifacts present in population databases.[9,34] Only 3 of the 24 variants were present in COSMIC, which may indicate clinically relevant somatic variants misclassified as germline. Therefore, the sensitivity of the filtration algorithms on this data set was more accurately 99.6% for both algorithms [initial algorithm 854/(854 + 3); revised algorithm 808/(808 + 3)].

Discussion

This study presents an optimized tumor-only variant filtration algorithm for distinguishing tumor-specific somatic variants from background germline variants in panel-based NGS testing of solid tumors. It was demonstrated that this algorithm is capable of defining somatic variants in a

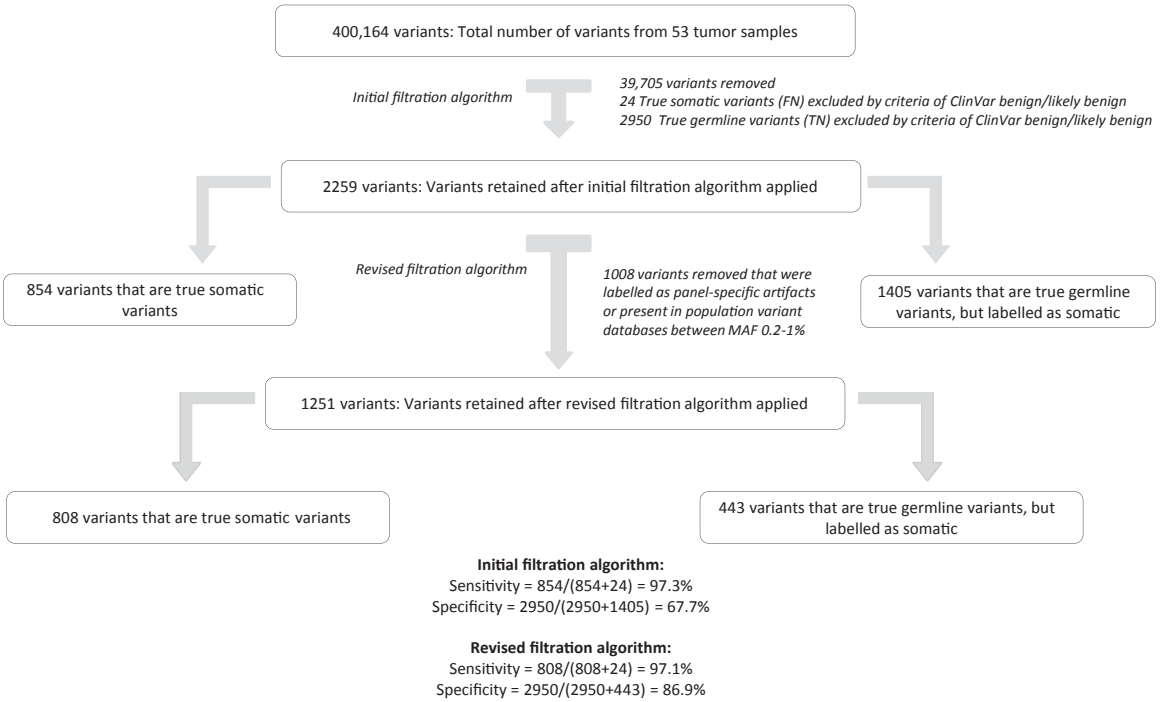


Figure 6 Application of optimized tumor-only filtration algorithm from the TruSeq Amplicon Cancer Panel (TSACP) panel to 53 tumor samples run on the 555-gene panel. Both an initial filtration algorithm and a revised filtration algorithm were used, with the revised algorithm adjusted to exclude variants if they were identified as panel-specific artifacts or were present at the threshold of minor allele frequency (MAF) $\geq$ 0.2% in any subpopulation data set in Exome Aggregation Consortium (ExAC), 1000 genomes, and Exome Sequencing Project 6500, or the global data sets for the three databases.

medium-sized (48 gene) targeted amplicon NGS panel with a sensitivity of 100% and specificity of 95.5%. Specific variant assessment tools were modeled to derive the optimal algorithm, including the order of operations, to minimize errors in classification of somatic and germline variants. With the use of a larger-sized targeted hybridization capture panel (555 genes), sensitivity and specificity are slightly reduced (99.6% and 86.9%, respectively) but with minimal impact on identification of clinically relevant variants in tumor-only analysis. Although the use of specific software (Alissa Interpret; Agilent) is described, the same approach can be manually implemented within clinical laboratories.

PVDs

Optimal results were obtained by de-emphasizing dbSNP as a germline variant classification tool, because a high number of FSCs were identified when using this tool (Figure 2). In addition, OR combinations of PVDs provides for simultaneous input from more than one database, diminishing differences between single databases, whereas AND combinations exacerbate these differences. Use of subpopulation-specific PVDs adds a layer of information in the context of ethnic group–specific polymorphisms with increased specificity and without compromised sensitivity. Subpopulation-specific data sets are especially relevant in geographic regions with large and diverse multi-ethnic groups such as our own Toronto region patient population. However, not all PVDs enabled effective use of this feature, typically because of low population numbers or lack of given ethnicity cohorts in the data.

MAF Thresholds

A key aspect in using germline PVDs to classify variants is the set MAF threshold. Standard practice is to use a threshold between 1% and 2% MAF as the threshold above which a variant is considered a neutral polymorphism, based on the traditional definition of a polymorphism as appearing in >1% of a population. With a targeted NGS panel, a MAF cutoff of as low as 0.2% preserves a high sensitivity and specificity by excluding known polymorphisms found at low frequency in the population. In these data, only private changes (unique to individuals or families) and those not found within the large data sets curated for the PVDs were not excluded and potentially misidentified as somatic instead of germline variants. However, application of other curation tools (eg, literature) identify these as variants of uncertain significance, so there were no circumstances in these data sets in which FSCs were of potential clinical significance or actionable for treatment in the somatic context.

Application of COSMIC

COSMIC's utility to distinguish between somatic and germline variants detected by small targeted NGS panels

was investigated. The analysis suggests that there is no COSMIC occurrence threshold applicable that would ensure that COSMIC performs with high sensitivity and specificity. This is likely because of the contamination of COSMIC with germline variants, reducing its discriminatory power. An improvement of the COSMIC data set, retaining only verified somatic, and excluding potential germline, variants would be a useful tool for tumor-only filtration algorithms. However, COSMIC was a useful tool in eliminating FSCs from consideration when analyzing data from larger panels, because private germline variants and artifacts are unlikely to be recurrent within the COSMIC data set.

Protein Function Prediction Algorithms

PFPAs that predict if missense changes have deleterious effects on protein structure or function are known to have limited sensitivity and specificity in helping to determine pathogenicity of germline variants. PFPAs are similar in the underlying principles used to predict deleterious effects (ie, use of evolutionary conservation and chemical differences). As such, PFPAs are typically used in a redundant and limited manner for germline variant interpretation. This study demonstrates that PFPAs are challenging for use in assessing somatic variants. One reason may be that they do not take into consideration predicted effects for oncogenic variants versus those for tumor suppressive and neomorphic variants. In addition, PFPAs cannot be used to predict the effects of single nucleotide variants on signaling pathway function, an important feature of assessing variant status as benign or deleterious.

Application of Tumor-Only Filtration Algorithm to Large NGS Panel

To evaluate applicability of the tumor-only filtration algorithm to data from a larger NGS panel, the algorithm was trialed on our laboratory-developed UHN Hi5 panel (*Materials and Methods*). Initial results demonstrated a lower sensitivity and specificity (97.3% and 67.7%, respectively) than on the TSACP. To improve the algorithm for the larger panel, a step was added to remove known technical artifacts, and the stringency of excluding variants in PVDs (MAF $\geq$0.2%) increased. After these improvements, the revised algorithm had a sensitivity and specificity (99.6% and 86.9%, respectively), comparable with the data from the TSACP. This additional step was not necessary when analyzing the smaller panel, because smaller panels focus on gene regions known to contain clinically relevant variants, which are less likely to contain private germline variants that may be misclassified as FSCs.

Use of Laboratory-Specific MVLs

We describe the use of two types of laboratory-curated MVLs. An MVL containing actionable somatic variants from previous testing within our laboratory (CR-MVL), and, for the larger panel, one containing potential test or platform specific artifacts were used. These curated lists provided an easy way to enhance variant filtering. For example, our CR-MVL was used to retain potentially clinically relevant variants irrespective of quality or other filtration criteria (Supplemental Figure S2). A surrogate of our CR-MVL that contained clinically relevant or known actionable variants can be readily constructed from information in publicly available databases (eg, from variants listed in *mycancergenome.org*, OncoKB, or CIViC somatic variant databases), whereas other MVLs such as the artifact MVL must be constructed from laboratory experience. In contrast, the artifact MVL was useful in eliminating low VAF variants identified as a result of sequencing or bioinformatics artifacts. The relative footprints of the UHN Hi5 compared with the TSACP necessitated artifact removal, because the UHN Hi5 was approximately 50$\times$ larger in library size and was more likely to accumulate artifacts.

Order of Operations

A key finding of this study is that not all filtering resources available for deployment should be included in the optimal version of the filtration process. In fact, inclusion of certain tools or filters may actually be detrimental rather than beneficial to classification of tumor-only variants, such as the use of HGMD, COSMIC, and PFPAs. Another important distinction in designing filtration algorithms is the use of inclusionary as opposed to exclusionary filtering steps. It is important to understand the type of data existing within each tool and the effect of each inclusion/exclusion step on germline and somatic variants. The ideal process for classification of somatic tumor variants from tumor-only testing will minimize FSCs and FGCs at each step of the process. For larger panels, specificity is decreased, because the likelihood of FSCs increases. In those cases, consideration of the cost–benefit of paired germline testing alongside the tumor sample is warranted. This is consistent with recently published work suggesting that parallel germline and tumor testing were needed with large panel sequencing to avoid misclassification of potentially relevant pathogenic germline mutations.[35] This work is in agreement with another recent study that identified the benefit of a PVD filter in conjunction with a pool of normal samples to tumor-only variant assessment.[36] This work also demonstrates the additional benefit of using clinically relevant variants, a list of sequencing artifacts, and the ClinVar database in tumor-only variant filtration.

Conclusions

This study describes the optimization and investigation of a tumor-only filtration algorithm for classification of somatic

variants. Applicability of the tumor-only filtration algorithm to small−medium targeted amplicon (TSCAP, 48 genes) and larger hybrid-capture NGS panels (UHN HI5, 555 genes) with similar sensitivity and specificity is shown, and the optimal use of available tools such as PVDs, MVLs, and somatic variant databases is described. The most optimal filtration approach for the smaller panel (Figure 5B) incorporated a database of clinically relevant variants, exclusion of ClinVar benign variants, and exclusion of variants in PVDs. Modification of the algorithm for a larger panel included an additional step of filtering out artifactual calls was required (Figure 6). The tumor-only filtration algorithm as described may prove useful to clinical molecular laboratories performing tumor molecular profiling as an approach to maximize correct classification of both somatic and germline variants identified from tumor-only approaches to molecular profiling.

Acknowledgment

We thank Ainsley R. Latour for careful review of the manuscript.

Supplemental Data

Supplemental material for this article can be found at *https://doi.org/10.1016/j.jmoldx.2018.09.008*.

References

1. Schrader KA, Cheng DT, Joseph V, Prasad M, Walsh M, Zehir A, Ni A, Thomas T, Benayed R, Ashraf A, Lincoln A, Arcila M, Stadler Z, Solit D, Hyman DM, Zhang L, Klimstra D, Ladanyi M, Offit K, Berger M, Robson M: Germline variants in targeted tumor sequencing using matched normal DNA. JAMA Oncol 2016, 2: 104−111
2. Jones S, Anagnostou V, Lytle K, Parpart-Li S, Nesselbush M, Riley DR, Shukla M, Chesnick B, Kadan M, Papp E, Galens KG, Murphy D, Zhang T, Kann L, Sausen M, Angiuoli SV, Diaz LA Jr, Velculescu VE: Personalized genomic analyses for cancer mutation discovery and interpretation. Sci Transl Med 2015, 7:283ra53
3. Hiltemann S, Jenster G, Trapman J, van der Spek P, Stubbs A: Discriminating somatic and germline mutations in tumor DNA samples without matching normals. Genome Res 2015, 25:1382−1390
4. Robson ME, Bradbury AR, Arun B, Domchek SM, Ford JM, Hampel HL, Lipkin SM, Syngal S, Wollins DS, Lindor NM: American Society of Clinical Oncology Policy Statement Update: genetic and genomic testing for cancer susceptibility. J Clin Oncol 2015, 33: 3660−3667
5. Yen JL, Garcia S, Montana A, Harris J, Chervitz S, Morra M, West J, Chen R, Church DM: A variant by any name: quantifying annotation discordance across tools and clinical databases. Genome Med 2017, 9:7
6. Harrison SM, Dolinsky JS, Knight Johnson AE, Pesaran T, Azzariti DR, Bale S, Chao EC, Das S, Vincent L, Rehm HL: Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. Genet Med 2017, 19:1096−1104
7. Harrison SM, Riggs ER, Maglott DR, Lee JM, Azzariti DR, Niehaus A, Ramos EM, Martin CL, Landrum MJ, Rehm HL: Using ClinVar as a resource to support variant interpretation. Curr Protoc Hum Genet 2016, 89:8.16.1−8.16.23
8. Nishiguchi KM, Tearle RG, Liu YP, Oh EC, Miyake N, Benaglio P, Harper S, Koskiniemi-Kuendig H, Venturini G, Sharon D, Koenekoop RK, Nakamura M, Kondo M, Ueno S, Yasuma TR, Beckmann JS, Ikegawa S, Matsumoto N, Terasaki H, Berson EL, Katsanis N, Rivolta C: Whole genome sequencing in patients with retinitis pigmentosa reveals pathogenic DNA structural changes and NEK2 as a new disease gene. Proc Natl Acad Sci U S A 2013, 110: 16139−16144
9. Arthur JW, Cheung FSG, Reichardt JKV: Single nucleotide differences (SNDs) continue to contaminate the dbSNP database with consequences for human genomics and health. Hum Mutat 2015, 36: 196−199
10. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001, 29:308−311
11. Garofalo A, Sholl L, Reardon B, Taylor-Weiner A, Amin-Mansour A, Miao D, Liu D, Oliver N, MacConaill L, Ducar M, Rojas-Rudilla V, Giannakis M, Ghazani A, Gray S, Janne P, Garber J, Joffe S, Lindeman N, Wagle N, Garraway LA, Van Allen EM: The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. Genome Med 2016, 8:79
12. Stockley TL, Oza AM, Berman HK, Leighl NB, Knox JJ, Shepherd FA, Chen EX, Krzyzanowska MK, Dhani N, Joshua AM, Tsao M-S, Serra S, Clarke B, Roehrl MH, Zhang T, Sukhai MA, Califaretti N, Trinkaus M, Shaw P, van der Kwast T, Wang L, Virtanen C, Kim RH, Razak ARA, Hansen AR, Yu C, Pugh TJ, Kamel-Reid S, Siu LL, Bedard PL: Molecular profiling of advanced solid tumors and patient outcomes with genotype-matched clinical trials: the Princess Margaret IMPACT/COMPACT trial. Genome Med 2016, 8:109
13. Li H, Durbin R: Fast and accurate short read alignment with Burrows−Wheeler transform. Bioinformatics 2009, 25:1754−1760
14. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010, 20:1297−1303
15. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, Meyerson M, Getz G: Oncotator: cancer variant annotation tool. Hum Mutat 2015, 36:E2423−E2429
16. Sukhai MA, Craddock KJ, Thomas M, Hansen AR, Zhang T, Siu L, Bedard P, Stockley TL, Kamel-Reid S: A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer. Genet Med 2016, 18:128−136
17. Siva N: 1000 Genomes project. Nat Biotechnol 2008, 26:256
18. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: A global reference for human genetic variation. Nature 2015, 526:68−74
19. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al: Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016, 536:285−291
20. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ: COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015, 43:D805−D811
21. Krawczak M, Ball EV, Stenson P, Cooper DN: HGMD: the human gene mutation database. Edited by Letovsky S. In Bioinformatics: Databases and Systems. Boston, MA: Springer, 2002. pp. 99−104
22. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR: ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016, 44:D862−D868

23. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al: CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet 2017, 49:170−174
24. Kumar P, Henikoff S, Ng PC: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009, 4:1073−1081
25. Ng PC, Henikoff S: SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003, 31:3812−3814
26. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: A method and server for predicting damaging missense mutations. Nat Methods 2010, 7:248−249
27. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D: MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010, 7:575−576
28. Reva B, Antipin Y, Sander C: Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011, 39:e118
29. Chun S, Fay JC: Identification of deleterious mutations within three human genomes. Genome Res 2009, 19:1553−1561
30. Choi Y, Chan AP: PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 2015, 31:2745−2747
31. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015, 17:405−424
32. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, Tsimberidou AM, Vnencak-Jones CL, Wolff DJ, Younes A, Nikiforova MN: Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn 2017, 19:4−23
33. Swanton C: My cancer genome: a unified genomics and clinical trial portal. Lancet Oncol 2012, 13:668−669
34. Reich DE, Gabriel SB, Altshuler D: Quality and completeness of SNP databases. Nat Genet 2003, 33:457−458
35. Mandelker D, Zhang L, Kemel Y, Stadler ZK, Joseph V, Zehir A, et al: Mutation detection in patients with advanced cancer by universal sequencing of cancer-related genes in tumor and normal DNA vs guideline-based germline testing. JAMA 2017, 318: 825−835
36. Teer JK, Zhang Y, Chen L, Welsh EA, Cress WD, Eschrich SA, Berglund AE: Evaluating somatic tumor mutation detection without matched normal samples. Hum Genomics 2017, 11:22