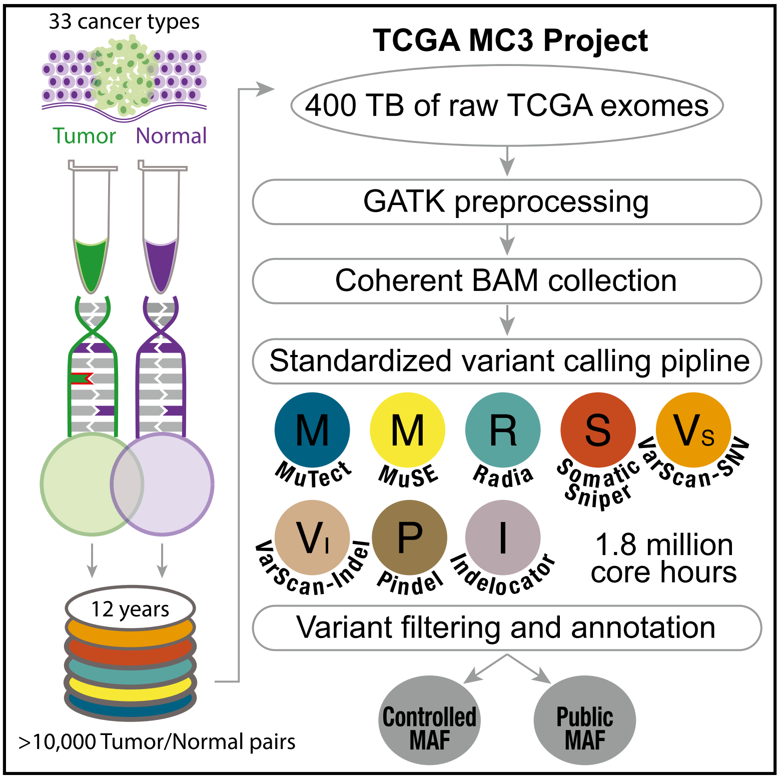
Article

Cell Systems

Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines

Graphical Abstract



Highlights

- Exome sequencing-based variant calls from 10,000 individuals
- Samples from 33 cancer types
- Variants from: MuTect, MuSE, VarScan2, Radia, Pindel, Somatic Sniper, Indelocator

Authors

Kyle Ellrott, Matthew H. Bailey, Gordon Saksena, ..., Li Ding, MC3 Working Group, and The Cancer Genome Atlas Research Network

Correspondence

ellrott@ohsu.edu

In Brief

The MC3 is a variant calling project of over 10,000 cancer exome samples from 33 cancer types. Over three million somatic variants were detected using seven different methods developed from institutions across the United States. These variants formed the basis for the PanCan Atlas papers.

Ellrott et al., 2018, Cell Systems *6*, 271–281
March 28, 2018 © 2018 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2018.03.002

Cell Systems
Article

OPEN ACCESS
CellPress

Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines

Kyle Ellrott,[1,10,11,*] Matthew H. Bailey,[2] Gordon Saksena,[3] Kyle R. Covington,[4] Cyriac Kandoth,[5] Chip Stewart,[3] Julian Hess,[3] Singer Ma,[7] Kami E. Chiotti,[1] Michael McLellan,[2] Heidi J. Sofia,[6] Carolyn Hutter,[6] Gad Getz,[3,8,9] David Wheeler,[4] Li Ding,[2] the MC3 Working Group, and The Cancer Genome Atlas Research Network

[1]Biomedical Engineering, Oregon Health and Science University, Portland, OR 97239, USA
[2]Department of Medicine, McDonnell Genome Institute, Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO 63110, USA
[3]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA
[4]Department of Molecular and Human Genetics, Baylor College of Medicine Human Genome Sequencing Center, 1 Baylor Plaza, Houston, TX 77030, USA
[5]Marie-Josée and Henry R. Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY 10021, USA
[6]National Human Genome Research Institute (NHGRI), NIH, Bethesda, MD 20892, USA
[7]DNAnexus, 1975 W EL Camino Real, Suite 204, Mountain View, CA 94040, USA
[8]Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA 02129, USA
[9]Harvard Medical School, Boston, MA 02115, USA
[10]Twitter: @kellrott
[11]Lead Contact
*Correspondence: ellrott@ohsu.edu

https://doi.org/10.1016/j.cels.2018.03.002

SUMMARY

The Cancer Genome Atlas (TCGA) cancer genomics dataset includes over 10,000 tumor-normal exome pairs across 33 different cancer types, in total >400 TB of raw data files requiring analysis. Here we describe the Multi-Center Mutation Calling in Multiple Cancers project, our effort to generate a comprehensive encyclopedia of somatic mutation calls for the TCGA data to enable robust cross-tumor-type analyses. Our approach accounts for variance and batch effects introduced by the rapid advancement of DNA extraction, hybridization-capture, sequencing, and analysis methods over time. We present best practices for applying an ensemble of seven mutation-calling algorithms with scoring and artifact filtering. The dataset created by this analysis includes 3.5 million somatic variants and forms the basis for PanCan Atlas papers. The results have been made available to the research community along with the methods used to generate them. This project is the result of collaboration from a number of institutes and demonstrates how team science drives extremely large genomics projects.

INTRODUCTION

The cost of sequencing is dropping rapidly while the costs of computing and data storage are dropping more slowly in comparison (Stein, 2010), making it difficult to deploy core analysis on raw data in genomics cohorts. It is often too expensive for individual labs to each use a one-off method on all their data. A more efficient approach is to design, test, and develop cohort-wide analysis by multi-lab consortiums with results that can be shared with a larger group of analysts. Scaling computational systems and genomic analysis to work for these large datasets requires the coordination of many institutions, many experiments, and many computational techniques. Aside from logistical problems, there are several technical issues that encumber large-scale analyses, revealing unmet needs: (1) deployment of reproducible computing methods in new computing environments; (2) the ability to deploy methods without manual intervention; (3) the biases of single methods and the need for consensus; and (4) the large amount of noise and false-positives that come from data including both germline sequencing, heterogeneous tumor sequencing, and low variant allele fraction of observed reads.

There are a number of cancer genomics projects working to do analysis on increasingly large datasets (Table 1) (Barretina et al., 2012; Brunner and Graubert, 2018; Campbell et al., 2017; Hartmaier et al., 2017; Turnbull, 2018; Project GENIE, 2017). The Cancer Genome Atlas (TCGA), for example, was a massive effort in multi-center cooperation, computational tool development, and collaborative science. However, the protocols and tools for identifying and characterizing tumor sequence variants evolved over time and were not uniformly applied across the project. When somatic variant callers were first compared—early in the TCGA timeline (2012)—a surprisingly large number of unique calls were identified for each method (Kim and Speed, 2013). To address some of these preliminary issues, TCGA organized Multi-Center Mutation Calling (MC2), which focused on consensus call sets of calling efforts from the Broad Institute, UCSC, Washington University, and Baylor. By the conclusion of the MC2 effort simply moving these data from one site to another became a daunting task—let alone correcting for

Table 1. Large Cohort Cancer Genomics Projects

Project	Method	Sample Count (Approx.)
TCGA MC3	exome	10,000
GENIE	44 gene panel	19,000
ICGC PCAWG	whole genome	2,800
100,000 Genomes Project	whole genome	projected: 100,000
CCLE	exome	950
Target	exome	700
Foundation medicine	306 gene panel	18,000

potential batch effects or caller-specific biases. Although the MC2 produced high-quality calls within each tumor-specific analysis working group (AWG), there were still differences in the callers, parameters, and filters used from project to project. Another effort of large-scale sequencing aggregation was implemented at the Broad Institute, in the effective deployment of the Firehose system (https://gdac.broadinstitute.org/), which automatically ran a suite of tools, designed at the Broad, to perform variant calling on all TCGA samples. While these data addressed consistency across tumor types, these data were not amenable to custom design by groups outside of the Broad. In 2014, the International Cancer Genome Consortium-TCGA Somatic Mutation Calling DREAM challenge (Ewing et al., 2015) created an open leaderboard to benchmark variant calling methods from groups around the world. The DREAM challenge identified methods with a large variety of techniques and performance profiles. However, no large-scale genomic calling effort had yet deployed a robust set of these methods in a uniform fashion.

To drive analysis outside of these silos, TCGA organized the Multi-Center Mutation Calling in Multiple Cancers (MC3) project, which has developed pipelines to uniformly apply many mature tools across the TCGA sequencing project. The combination of cloud computing power, policy changes, and improved variant calling software made this effort possible. The result is an open science collaboration across multiple institutions, designed to translate brittle custom-coded methods deployed at individual sequencing centers into portable, robust methods that enable reproducibility, transparency, and shareability with the broader research community. The software methods for this endeavor have been made publicly available, along with the datasets that it created.

In this paper, we describe the various challenges and considerations of building standardized genomic analysis pipelines that can be deployed in mass to tens of thousands of samples, we also highlight some lessons learned, and considerations of performance when looking across widely varied cohorts. The resulting dataset, compiled in Mutation Annotation Format ([MAF], https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification), represents several million core-hours of computational time on over 400 TB of short-read data using the current state-of-the-art variant calling and filtering methods. The MAF file represents over 20 million variants produced across approximately 10,000 tumor-normal pairs from 33 cancer types using 7 variant callers. This form of collaborative science, driven by a consortium of researchers across multiple institutions, is needed as the amount of genomic data continues to increase. The data generated by this work has formed the basis of the somatic exome variant analysis presented in the other papers from the TCGA PanCanAtlas project. More detailed analysis of the characteristics of the data and their biological implications will be discussed in other papers, such as “Comprehensive Characterization of Cancer Driver Genes and Mutations” (Bailey et al., 2018). “The Immune Landscape of Cancer” (Thorsson et al., 2018), and the “Genomicand Molecular Landscape of DNA Damage Repair Deficiency across the CancerGenome Atlas” (Knijnenburg et al., 2018).

RESULTS

Cloud Deployment and Reproducibility

The MC3 project in support of the TCGA PanCanAtlas is the result of a number of institutions collaborating to provide resources and methods. In many cases, the project was able to utilize newly developed systems to deploy compute in ways that were not previously possible. These systems included custom-written management scripts, institutional work management platforms, and cloud-based systems. Alignment, The Genome Analysis Toolkit (GATK) processing, and variant calling for MuTect (Cibulskis et al., 2013) and Indelocator (Chapman et al., 2011) were run on the Broad’s Firehose system. Additional GATK Indel realignment and base quality score recalibration was done on over 1,000 tumor normal pairs on the University of California Santa Cruz cluster. Processed files were stored at the CGHub system. Over a 4-week period, using almost 1.8 million core-hours, 400 TB of data was processed for variant calling using the Pindel (Ye et al., 2009, 2015), MuSE (Fan et al., 2016), Radia (Radenbaugh et al., 2014), Varscan (Koboldt et al., 2012), and SomaticSniper (Larson et al., 2012) pipelines on the DNAnexus systems. OxoG scores for samples were calculated on the Institute for Systems Biology Cancer Genomics Cloud, and validation data were processed using the Broad Firecloud platform.

The majority of the pipelines built for this project were designed to be deployed in multiple computing environments. To ensure reproducibility, the methods described in this paper have been implemented using modern workflow technologies, which are showing rapid adoption. In this model, the workflow is described using: (1) a software container—a packaging system that simplifies deployment of the runtime environment, includes exact software dependencies and all features to run the program; (2) the tool wrappers—for each tool utilized, the command line argument to be invoked is described as a set of defined inputs, outputs, and parameters that can be used by a workflow engine to be scheduled and managed; (3) a pipeline description—a document that describes how all the tools fit together, the different parameters that should be modified, and required inputs. For distribution, the MC3 pipeline is described in the Common Workflow Language format with the required software packages deployed using Docker software containerization technology. Docker provides methods to package a program and all of its dependencies. These container images can be shipped to any Linux machine, whether cloud based or bare metal. Then the packaged tools can be easily run in new environments with minimal configuration. This workflow implementation is written using open standards which are easy to distribute and allow other researchers to replicate, modify, and extend this

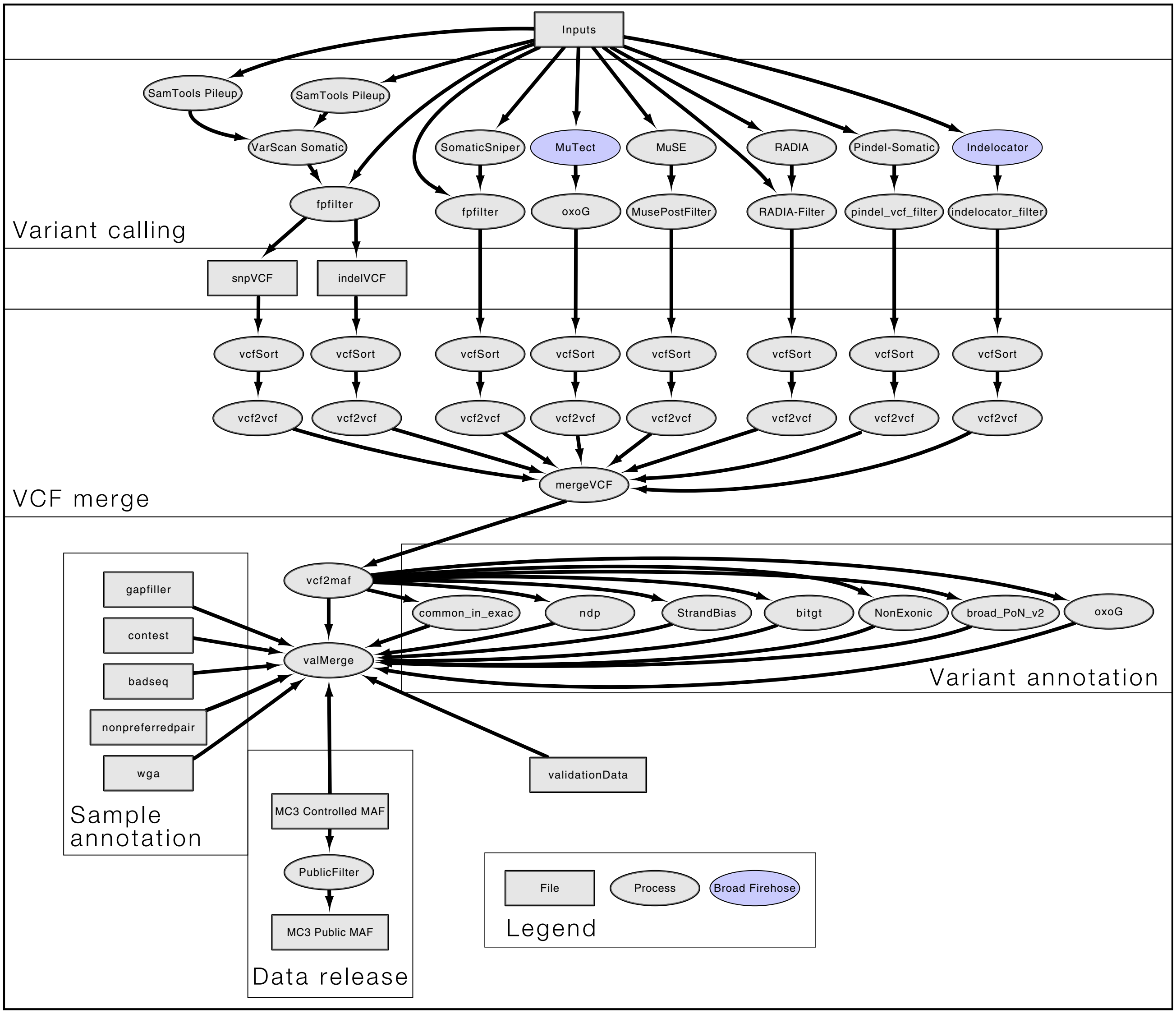


Figure 1. Workflow for Mutation Detection and Filtering

This workflow diagram reflects the internal design of the mutation calling pipeline. Squares in the flowchart represent files, and circles indicate processes. When colored, analysis was performed using the BROAD Firehose pipeline. Aligned input files were analyzed by seven different variant callers using author-recommended parameters to generate VCF files. All VCF files were merged and variant effect predictor annotated using vcf2maf tool. Processes flanking vcf2maf processes illustrate when filters were integrated. Finally, a separate set of annotation files were included and considered for variant and sample selection in the controlled and the public release of the annotated mutations file.

analysis to their own data. Results are publically available from the National Cancer Institute (NCI) Genomic Data Commons and include protected Variant Call Format (VCF) file releases, as well as a filtered, open-access TCGA MC3 MAF release that contains only the highest-confidence somatic mutations. These data will enable further PanCanAtlas efforts and, more generally, cancer research on TCGA data.

MC3 Variant Calling Strategy and Comparison with AWG MAFs

The MC3 effort used seven variant calling methods with proven performance (Figure 1) including Indelocator, MuSE, MuTect, Pindel, RADIA, SomaticSniper, and VarScan (VarScan calls both indels and SNPs). In addition, a collection of filtering methods were applied. These methods were applied to 10,510 tumor/normal pairs from 33 cancer types in the TCGA collection of whole exome sequencing data. This produced nearly 20 million variants. Definitions of controlled and open-access release of genomic variants for the TCGA data allows somatic variants that occur in exonic regions in open-access files (https://tcga-data.nci.nih.gov/docs/publications/tcga/datatype.html). Variants called in non-exonic regions, such as introns, 5' or 3' UTR are restricted to controlled-access release. In addition, somatic variants at sites that lacked sufficient normal depth coverage, or variants found in the panel of normals, were filtered from open-access since they were considered to be possible germline variants. Using these

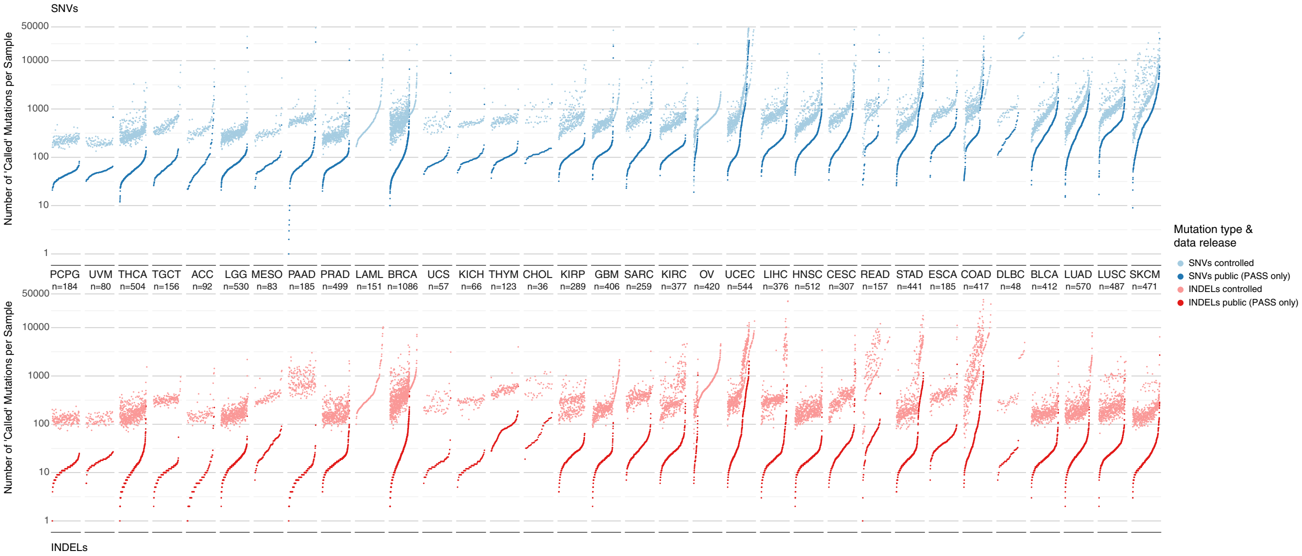


Figure 2. Distribution of Mutations in Controlled and Open-Access Mutation Files

Two panels show mutation load for each sample in the dataset for SNVs (above) and indels (below). Each dot of the sorted scatterplots shows the total number of mutations pre- and post-filtering per sample. Total mutation counts are separated by total number SNVs (blue) and indels (red) per samples. Lighter colors indicate pre-filtered mutations from the controlled-access MAF, and deeper colors indicate post-filtered (PASS only) mutations from the open-access MAF. Cancers are ordered by the median number of post-filtered SNVs per tissue. Furthermore, samples are sorted by increasing number of total mutation count for SNV and indel plots, respectively. Samples removed during post-filtering are also shown, i.e., LAML and OV in lighter colors without an accompanying pair and are sorted accordingly. The total number of samples for each cancer type is displayed under each cancer label. Finally, the y axis limits were placed from 0 to 50,000 for clarity. This resulted in the removal of 14 hypermutator samples from SNV plot and 10 hypermutator samples from the indel plot.

criteria, the full set of variants was narrowed down to an open-access file of around four million variants. A majority of downstream PanCanAtlas analyses was based on this subset of variants.

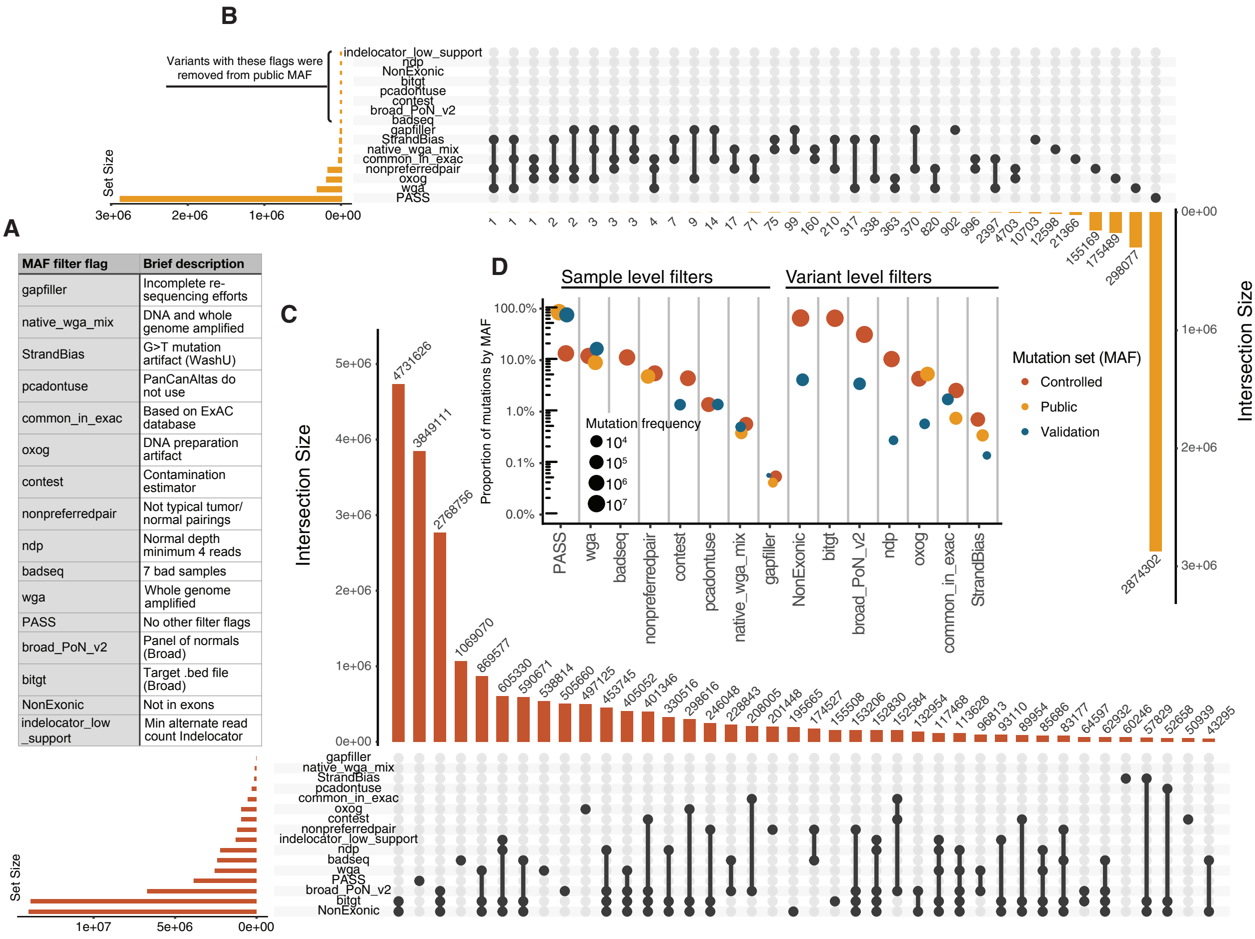
To gauge complementarity with previous efforts of calling mutations across many of these same tumor types, we compared the new set of calls with the MAF published as part of the first TCGA PanCan12 project for 12 tumor cancer types in 2013 (http://www.nature.com/tcga/). The PanCan12 MAF was created by collecting the variants from each separate TCGA AWG without any attempt at unification and includes data from a number of TCGA projects beyond the original PanCan12 set, including pancreatic adenocarcinoma (PAAD) and skin cutaneous melanoma (SKCM). We found that the new MC3 MAF had 1,079,216 variants in the PanCan12 MAF set of samples, while the PanCan12 MAF has 804,571. Among these calls, 717,326 variants are shared between the two sets (Figure S1). Thus, the MC3 project captured 89.5% of the original calls while increasing the size of the call set by 25%. The largest deviation was the PAAD project, which only saw 33% of the original variants and is likely due to poor tumor purity (see the PAAD marker paper for more details about somatic mutation calling efforts for this cancer type (Cancer Genome Atlas Research Network, 2017). Conversely, head-neck squamous cell carcinoma, SKCM, breast cancer, urothelial bladder carcinoma, colon adenoma/rectal adenoma, and uterine corpus endometrial carcinoma (UCEC) had greater than 90% of the original variants rediscovered by the MC3 effort (Figure S2).

For some cancer types, tumor cells profuse into the normal cells, causing issues in the identification of somatic variants. The best example of this is acute myeloid leukemia (LAML), which affects blood and bone marrow. Normal tissue samples (skin biopsies) from LAML patients often contain blood enriched with tumor cells. This can cause variant calling programs to mislabel somatic mutations as germline. The MC3 pipeline is conservative, attempting to remove all false-positive germline calls. Much of the original MAF created by the TCGA LAML AWG was derived by manual interventions, including Sanger sequencing data not included as part of the TCGA data catalog, to recover variants that would have otherwise been uncalled. As a result, the open-access MC3 call set only recovered 44% of the variants called in the original MAF (Figure S1).

Effects of Somatic Filtering for Open-Access Release

To conform to release guidelines for open-access data in TCGA, the MC3 efforts took significant steps to remove potential germline calls as well as non-exonic variants. To accomplish this, filters were used against the flags that marked low normal depth coverage, non-exonic sites, sites outside of capture kit, sites marked by the Broad Panel of Normals, samples marked as being contaminated by ContEst, and variants that were only called by a single caller. The controlled-access MAF file contains 22,485,627 variants from 10,510 tumor samples and is comprised of 13,044,511 single-nucleotide variant (SNV) events and 9,441,116 indels. The open-access MAF file contains 3,600,963 variants from 10,295 tumors with 3,427,680 SNV events and 173,283 indels. We observed that skin and lung cancers (SKCM, lung squamous cell carcinoma, and lung adenocarcinoma) had the largest median number of SNVs per sample, consistent with previous publications (Akbani et al., 2015; Collisson et al., 2014; Hammerman et al., 2012) (Figure 2).

We plotted the proportion at which each of the different filters were found on variants in the three different datasets (the full call set, the open-access dataset, and the set of variants used for



MAF filter flag	Brief description
gapfiller	Incomplete re-sequencing efforts
native_wga_mix	DNA and whole genome amplified
StrandBias	G>T mutation artifact (WashU)
pcadontuse	PanCanAltas do not use
common_in_exac	Based on ExAC database
oxog	DNA preparation artifact
contest	Contamination estimator
nonpreferredpair	Not typical tumor/ normal pairings
ndp	Normal depth minimum 4 reads
badseq	7 bad samples
wga	Whole genome amplified
PASS	No other filter flags
broad_PoN_v2	Panel of normals (Broad)
bitgt	Target .bed file (Broad)
NonExonic	Not in exons
indelocator_low _support	Min alternate read count Indelocator

Figure 3. Description of the Filters Implemented in Controlled and Open-Access Mutation Files

(A) Filter flags (as displayed in MAF) and a brief description of their purpose.

(B) Variant counts in the open-access MAF by filter were processed using an UpSetR plot (Conway et al., 2017). The following filters were globally applied to the Open-access MAF: ndp, NonExonic, bitgt, pcadontuse, contest, broad_PoN_v2, and badseq. Thus, zero variants in the open-access MAF were annotated with these flags. The inverted bar chart allows for the interpretation co-occurring filters at the variant level. For example, 304,602 variants were labeled with wga alone, whereas 2,455 variants were annotated with both wga and common_in_exac. The connected dots indicate which of filter flags are assessed.

(C) UpSetR plot indicates the co-occurrence of filters with variants of the controlled MAF, same as in (B).

(D) The proportion and frequency of filters for both the open and controlled datasets are displayed. In addition, validation flag counts and proportions are shown. The set of validation calls has a higher percentage of PASS calls, reflecting its bias toward higher-quality variant calls. Filter flags are separated into samples level filters and variant level filters. See also Figure S4.

validation) to show the reasons for differences in variant counts in the different sets (Figure 3). The most notable shift is the number of variants (over 60%) found in the full call set that were marked by the NonExonic and bitgt filters, which remove variants by genomic regions rather than technical reasons. These sites do not qualify for open-access release and may not be equally covered by all of the variant calling methods. In addition, the Broad Panel of normals flagged almost 30% of the calls in the full set, which were also removed in accordance with TCGA data release policies.

To further illustrate the importance of filtering on biological findings, we performed significantly mutated gene (SMG) analysis using both MutSig2CV (Lawrence et al., 2013) and MuSiC2 (Dees et al., 2012) for all KIRC variants present in the controlled-access MAF compared with those present in the open-access MAF and marked as PASS in the annotation. Typically this method of SMG analysis, using raw mutation calls, is performed in order to quickly identify sequencing and technical artifacts. Using the stringent p value cutoff for both tools, MutSig2CV ($p < 3.5 \times 10^{-5}$) and MuSiC2 ($p < 1 \times 10^{-7}$) each identified 10 SMGs using PASS variants from the open-access MAF. Seven of these genes overlapped between MutSig2CV and MuSiC2, *TP53*, *PTEN*, *VHL*, *SETD2*, *PBRM1*, *BAP1*, and *MTOR.* MutSig2CV uniquely identified *TCEB1, PIK3CA,* and *ATM*, and MuSiC2 uniquely identified *ERBB4*, *SLITRK6,* and *KDM5C* after long gene filtering. The complete set of variants from the controlled MAF yielded many more SMG hits (MutSig2CV = 1,203, and MuSiC2 = 321). The noise introduced by the unfiltered variant calls made the identification of real SMG signals very difficult.

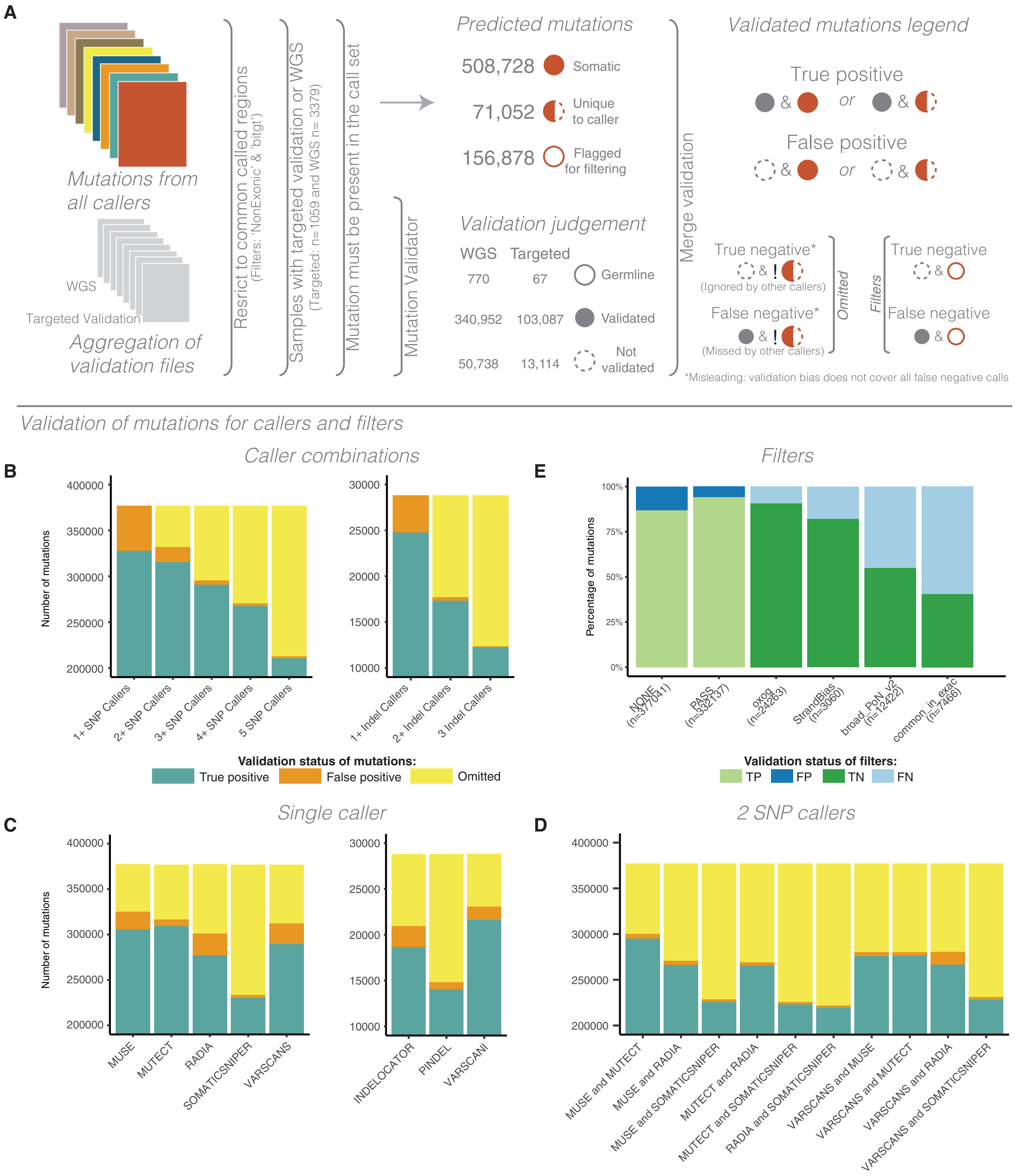


Figure 4. Validation Statistics of Mutations Calls

While these results reflect validation of resequenced samples, technical artifacts may still be present because orthogonal technology was not implemented.

(A) Overview of the mutations validation process. Symbols are used to illustrate how mutations predictions were assessed. Values shown in under Predicted mutations are not mutually exclusive. Exclamation marks under true-negative and false-negative denotes the logical negation or not.

(B) The composition of variants with overlapping callers. Starting with any caller and increasing to require more callers to agree on a site. This is done for both SNVs (left) and indels (right).

(legend continued on next page)

Performance Evaluation of MC3 Variant Calling by Experimental Validation

To evaluate calling performance, the TCGA project performed targeted deep sequencing on select variants for the purpose of validation for individual cancer papers. Selection of these variants were made by the original tumor-specific AWGs, and was not performed specifically to validate MC3 efforts This targeted sequencing included 3,128 samples with validation of a wide range of selected genes and was used for MC3 validation. This set of sequences included 33 samples with more than 500 targets genes and a median of 4 genes per sample. Variants from UCEC comprised almost 28% of the sites and esophageal carcinoma 23% of the sites in this targeted validation dataset (Figure S3). In addition, whole genome sequencing (WGS) was also run providing additional orthogonal data to use for validation. WGS data was available for a subset of 1,059 samples, and provided a median of 126 validation sites per sample. Some methods, such as MuTect deployed by the Broad Firehose, only called variants within a region defined by the sequence capture kit definitions, even if additional sequencing was available. Because of this, sites marked by the bitgt filter, which marked non-common-capture regions, were removed from the validation dataset to provide consistent statistics when comparing across methods.

Because sites for targeted validation were selected from the most likely SMG candidates in the TCGA AWGs, rather than a random sampling of data, the validation data does not represent a robust benchmarking dataset. Every site involved in the targeted validation was called at least once by one of the variant calling methods. Because there is no background sampling, such as random sites not called by any of the methods, the false-negative rate neglects sites not called by any method. Sites related to false-positive germline signals would have been filtered before validation selection, and also not been part of validation efforts. In addition, validation sites would be biased toward less-complex and smaller events, which would impact performance evaluation of sites that are more difficult to characterize using targeted sequencing. We were able to partially manage this effect by including additional validation sites from samples where orthogonal WGS had also been performed. We should also note that the majority of validation data was generated using a similar sequence technology, therefore systematic errors such as those that several of the filters attempt to address will appear as erroneous filtering events. This particularly affects PoN filters. When comparing the subset of sites validated by targeted sequencing against WGS-based validation, the rate of these types of events does not seem to be very large. Given the profiles of filters among the datasets we see in Figure 4, the validation data do not mirror the characteristics of the full call sets. Despite these limitations, the validation dataset does provide extensive data about the relative performance of callers and filters (Table S2).

As seen in Figure 4, meta calling methods, such as two caller intersection, are able to quickly eliminate false-positives. This has been noted previously in other studies (Goode et al., 2013). The two-caller rule for the set of five SNP callers finds more valid sites than any specific combination of two callers (Table S3). This draws on the wisdom of crowd principle (Costello and Stolovitzky, 2013). The two-caller intersection is much less effective for indel calling methods, as it causes an increase in false negatives due to its conservative nature. We see general trends, such as MuTect and MuSE, detecting the largest number of true-positive sites among the validation variants surveyed. Somatic Sniper had the lowest number of detected sites, omitting the largest number of validated sites, but, at the same time, it had the smallest number of false-positive validated sites.

We observed many tool-specific patterns pertaining to mutation identification (Figure 5). Most calls that passed all the filters were supported by all five callers. For SNP calls, MuSE and MuTect have the highest pairwise agreement. They each also have the largest number of unique calls. For indel callers, Pindel makes the most calls, but over 130,000 of the variants were found in two samples, suggesting there may be characteristics of these samples that skew the numbers. Only a small fraction of indel calls are made by all three callers.

DISCUSSION

The previous paradigm of genomic research was that groups downloaded data, ran methods on their own, and then provided results to the community, representing a results-oriented approach. Under this model, it became extremely difficult for external groups to reproduce calculations or apply new methods to new datasets. However, with the advent of cloud technologies, such as computational virtualization and containerization systems, there is now the ability to capture computational methods in a way that can be run on external compute systems. This change allows for a methods-oriented strategy in which the collaborating institutions provide shareable algorithms to be run on the data, rather than processing it themselves. The MC3 is a showcase for a methods-oriented project, collecting reproducible codes for methods from collaborators and deploying them uniformly to data on the cloud.

Through collaboration, open science, and improved resources, the MC3 effort overcame lingering artifacts from previous cancer-type-specific analyses and reflects a true PanCancer set of somatic mutations. Many lessons were learned, or re-confirmed, while leveraging multi-institutional expertise: (1) while many methods have a public facing software on GitHub or clouds resources, default parameters were often insufficient. Achieving the best performance required additional input from the original authors. (2) Some tumor types, such as liquid tumors, require different strategies of variant calling and filtering to obtain an optimal set of mutations. (3) Providing annotation generated by various filtering methods, as opposed to generating files with fixed removal of possible artifacts allows for flexible usage of the mutation call set. (4) Using reproducible code- and methods-based approaches are essential as datasets increase

(C) The composition of validation status for calls from each independent caller for both SNPs (left) and indels (right).

(D) The composition of validation status for pairs of callers. (B), (C), and (D) all have a truncated y axis, all values below indicate true-positives mutation status. Omitted, as illustrated in (A), reflects the limitations of assessing mutation predictions when validations does incorporate all possible events.

(E) The composition of validation status for each of the filter flags. See also Figure S3. See also Figure 3 and Tables S2 and S3.

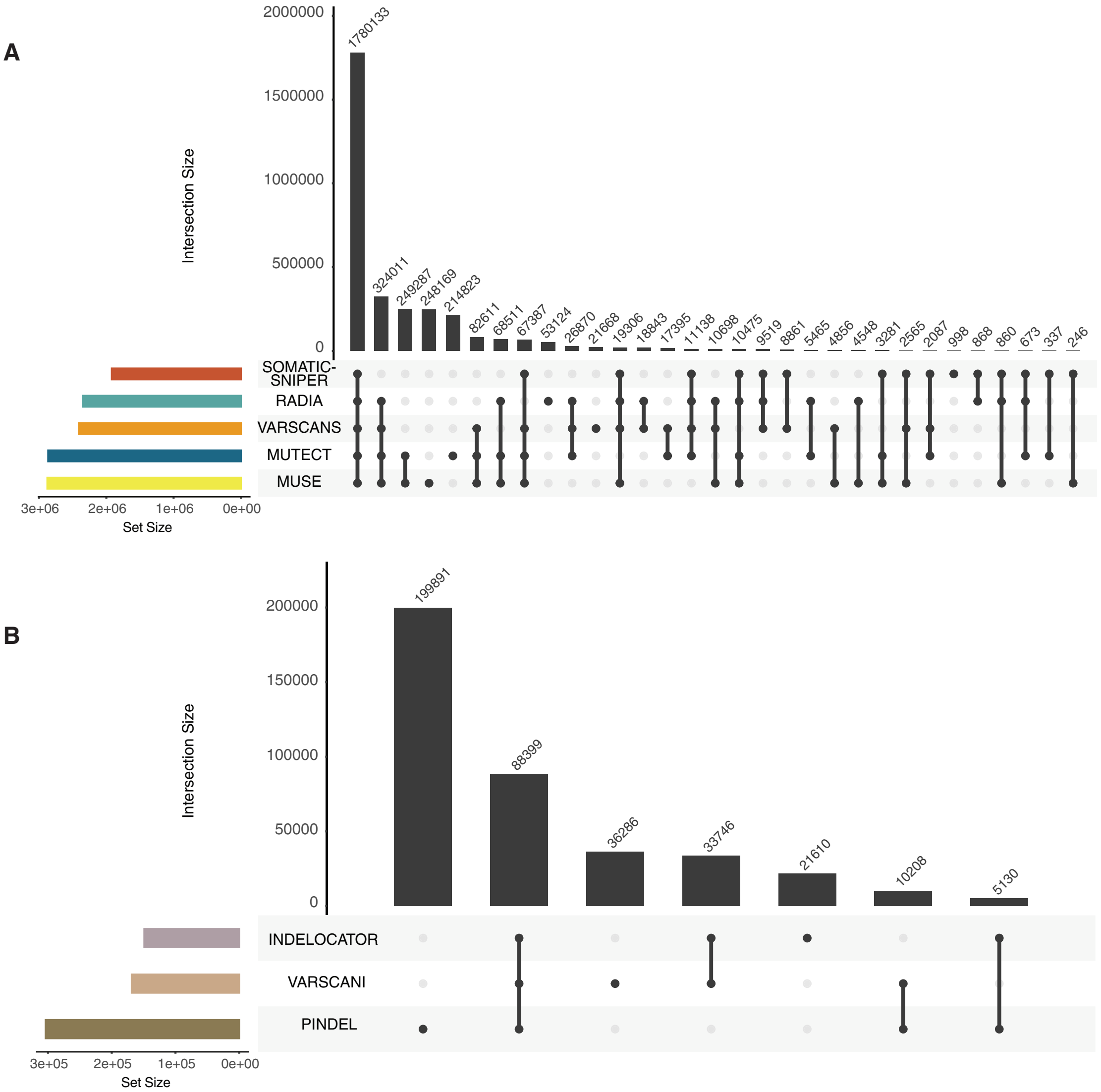


Figure 5. Intersection of Mutation Calls across Variant Prediction Software

The top bar-plot indicates intersection size. More specifically, one or more tools called each variant. This plot provides the number of variants that are uniquely called by one tool (a single point) or the numbers of variants called by many tools (two or more points). The bottom left plot indicates the set size. The linked points below display the intersecting sets of interest or which tools called variants.

(A) PASS only mutations from the controlled MAF are shown.

(B) Tools designed to call indels are displayed in a similar fashion to plot (A). Only indels with greater than three supporting reads are displayed in this plot. In addition, two samples were removed from these plots that represent extreme hypermutators (TCGA-D8-A27V and TCGA-EW-A2FV).

in size and complexity. (5) Meta-calling methods, which utilize the results of multiple methods, can provide more robust results than single methods. (6) Multiple precautions and filters were needed to protect potential germline leakage of patient data into public facing, open-access data. These lessons learned allowed for customizable strategies based on algorithmic objectives or biological inquiries.

This organization of coherent variant calling for 10,000 genomes was a multi-year process. However, there were a number of technical advances that occurred during this time frame, and these technologies will make utilization of cloud resources much more accessible for researchers going forward. While this effort was informed by the DREAM challenge (Ewing et al., 2015), many of the methods selected were based on best practices of the original TCGA AWGs. Ideally, future variant calling and filtering efforts should use a robust benchmarking effort to scan the various combinations of callers, filters, and parameters, and evaluate which callers and filters are optimal for different tumor types and contexts. The lessons learned from this project should inform the design of a new somatic mutation calling pipeline having an end-to-end FASTQ-to-filtered-MAF file workflow with complete containerization in a single cloud. Resources such as the TCGA catalog form the backbone of reference datasets that can be used as a point of comparison in new research projects. But those comparisons are only useful if the analysis is applied consistently. Thus, when pipelines are applied to large datasets, the methods should be made available alongside the resultant data so that other groups can apply them to their own experimental data.

The PanCanAtlas project encompasses many research goals. For this reason, a one-size-fits-all approach would not cover the different types of analyses. An example of this would be the problems of driver gene discovery versus heterogeneity analysis. A high-confidence caller with lower false-positive profiles is better geared for driver gene discovery, because the removal of false-positive noise helps to better identify significant recurring patterns. Once the significant driver genes have been identified, a second pass over the mutation set can find lower confidence calls that could provide additional examples of the gene of interest. In contrast, heterogeneity analysis, which looks for variants that occur in fractions of the population, works much better with very sensitive algorithms because these variants, with potentially low variant allele fractions, may be filtered out by more stringent methods. Therefore, it was appropriate to include called variants and provide mechanisms for doing additional filtering that was appropriate to the analysis. These steps, in accordance with the TCGA open-access release guidelines, resulted in the collection of three mutation annotation format (MAF) files: a controlled-access MAF, an open-access MAF, and a validation MAF. Each of these MAF files has distinct properties that are compared and contrasted here.

The MC3 effort reflects three objectives of large-scale data generation in an age of open science: collaboration, consensus, and consistency. First, multi-center collaboration combined efforts and expertise from multiple academic institutions. Second, mutation calling was performed using an array of seven mutation-callers developed by the adopted by different TGCA analysis centers. We show consensus calling outperforms single algorithms in both sensitivity and validation status. Finally, the use of consistent methods for calling across multiple-cancers enhances the utility of this resource in future efforts to contrast the molecular makeup across tumors. The results of this effort provide integral components necessary for future efforts in somatic variant calling.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - ○ Sample List Creation
  - ○ Variant Calling and Filtering Strategies
  - ○ Merger of Mutation Calls
  - ○ Workflow Deployment
  - ○ SMG Performance Analysis
  - ○ Mutation Validation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Effects of Cancer Type on Mutation Callers
  - ○ OxoG Events
  - ○ Per Gene Filtering Effects
  - ○ Indel Realignment and BQSR
  - ○ Variant Calling
  - ○ Variant Merger
  - ○ Panel-of-Normals Filter
  - ○ Restricting to Target/Coding Exons
  - ○ Minimum 3 Supporting Reads for Pindel Indel Calls
  - ○ Minimum Indelocator Indel Depth
  - ○ Annotation
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and three tables and can be found with this article online at https://doi.org/10.1016/j.cels.2018.03.002.

ACKNOWLEDGMENTS

The authors would like to acknowledge contributions of all of the members of the TCGA network. This work was funded by grants from to the UC Santa Cruz Genomics Institute and supported by the NIH NHGRI (grant no U54HG007990), NCI (U24CA143858), and NCI ITCR (grant no R01CA180778); Oregon Health and Science University from the NCI (U24CA210957, U24CA143799); M.L.A. grants from NCI (R01CA183793, U24CA210950, U24CA210949, U24CA143883, CA150252, and U24CA143845) and Keck Center of the Golf Coast Consortia for the Cancer Biology Training Program CPRIT (RP140113); NCI and NHGRI grants to McDonnell Genome Institute at Washington University (U24CA211006, U01HG006517, and U54HG003079); NCI grants to the Institute of Systems Biology (U24CA143835); NHGRI grants to the Broad Institute (U54HG003067); NCI and NHGRI grants to the Baylor College of Medicine (U54HG003273 and U24CA143843); NCI grants to MSKCC (U24CA143840); NCI grants to M.L.A. (U24CA143845, P30CA016672, and U24CA143883); NCI grants to University of North Carolina (U24CA143848); NCI grants to the Van Andel Research Institute (U24CA143882); NCI grants to Partners Healthcare (U24CA144025); NCI grants to Harvard (U24CA143867); and NCI grants to BC Cancer Foundation (U24CA143866). In addition, the author would like to thank DNAnexus for the compute processes they provided for the variant calling; OxoG calculations in this study was performed on the Institute for Systems Biology-Cancer Genomics Cloud (ISB-CGC), a pilot project of the National Cancer Institute (under contract number HHSN261201400007C) and the Broad Institute for validation data processing. The author would like to acknowledge the many individuals that contributed to the success of the MC3 project, including Liu Xi, Walker Hale, Katayoon Kasaian, Ignaty Leshchiner, Yifei Men, Sheila M. Reynolds, Gordon B. Mills, John N. Weinstein, Rehan Akbani, Wenyi Wang, Yu Fan, Melpomeni Kasapi, Adam Struck, Alex Buchanan, Allison Creason, John Letaw, Myron Peto, Pavana Anur, Amie Radenbaugh, Christopher K. Wong, David Haussler, Joshua M. Stuart, Beifang Niu, Dave Larson, Steven Foltz and Kai Ye.

AUTHOR CONTRIBUTIONS

K.E. led the dockerization of variant calling tools and wrote the manuscript. M.H.B. provided analysis and generated figures. K.R.C. and C.K. developed the MAF conversion and variant annotation pipelines and provided analysis support. G.S. built sample sets, developed pipeline tools, and provided analysis support. C.S. provided analysis for performance and validation data. J.H. provided Panel of Normals filter and description of its method. M.McL. conducted data analysis. L.D., G.G., and D.W. developed project plan and provided guidance on data analysis and interpretation. H.J.S. and C.H. managed the project. All authors revised the manuscript.

DECLARATION OF INTERESTS

Michael Seiler, Peter G. Smith, Ping Zhu, Silvia Buonamici, and Lihua Yu are employees of H3 Biomedicine, Inc. Parts of this work are the subject of a patent application: WO2017040526 titled "Splice variants associated with neomorphic sf3b1 mutants." Shouyoung Peng, Anant A. Agrawal, James Palacino, and Teng Teng are employees of H3 Biomedicine, Inc. Andrew D.

Cherniack, Ashton C. Berger, and Galen F. Gao receive research support from Bayer Pharmaceuticals. Gordon B. Mills serves on the External Scientific Review Board of Astrazeneca. Anil Sood is on the Scientific Advisory Board for Kiyatec and is a shareholder in BioPath. Jonathan S. Serody receives funding from Merck, Inc. Kyle R. Covington is an employee of Castle Biosciences, Inc. Preethi H. Gunaratne is founder, CSO, and shareholder of NextmiRNA Therapeutics. Christina Yau is a part-time employee/consultant at NantOmics. Franz X. Schaub is an employee and shareholder of SEngine Precision Medicine, Inc. Carla Grandori is an employee, founder, and shareholder of SEngine Precision Medicine, Inc. Robert N. Eisenman is a member of the Scientific Advisory Boards and shareholder of Shenogen Pharma and Kronos Bio. Daniel J. Weisenberger is a consultant for Zymo Research Corporation. Joshua M. Stuart is the founder of Five3 Genomics and shareholder of NantOmics. Marc T. Goodman receives research support from Merck, Inc. Andrew J. Gentles is a consultant for Cibermed. Charles M. Perou is an equity stock holder, consultant, and Board of Directors member of BioClassifier and GeneCentric Diagnostics and is also listed as an inventor on patent applications on the Breast PAM50 and Lung Cancer Subtyping assays. Matthew Meyerson receives research support from Bayer Pharmaceuticals; is an equity holder in, consultant for, and Scientific Advisory Board chair for OrigiMed; and is an inventor of a patent for EGFR mutation diagnosis in lung cancer, licensed to LabCorp. Eduard Porta-Pardo is an inventor of a patent for domainXplorer. Han Liang is a shareholder and scientific advisor of Precision Scientific and Eagle Nebula. Da Yang is an inventor on a pending patent application describing the use of antisense oligonucleotides against specific lncRNA sequence as diagnostic and therapeutic tools. Yonghong Xiao was an employee and shareholder of TESARO, Inc. Bin Feng is an employee and shareholder of TESARO, Inc. Carter Van Waes received research funding for the study of IAP inhibitor ASTX660 through a Cooperative Agreement between NIDCD, NIH, and Astex Pharmaceuticals. Raunaq Malhotra is an employee and shareholder of Seven Bridges, Inc. Peter W. Laird serves on the Scientific Advisory Board for AnchorDx. Joel Tepper is a consultant at EMD Serono. Kenneth Wang serves on the Advisory Board for Boston Scientific, Microtech, and Olympus. Andrea Califano is a founder, shareholder, and advisory board member of DarwinHealth, Inc. and a shareholder and advisory board member of Tempus, Inc. Toni K. Choueiri serves as needed on advisory boards for Bristol-Myers Squibb, Merck, and Roche. Lawrence Kwong receives research support from Array BioPharma. Sharon E. Plon is a member of the Scientific Advisory Board for Baylor Genetics Laboratory. Beth Y. Karlan serves on the Advisory Board of Invitae.

Received: August 14, 2017
Revised: January 21, 2018
Accepted: March 1, 2018
Published: March 28, 2018

REFERENCES

Akbani, R., Akdemir, K.C., Aksoy, B.A., Albert, M., Ally, A., Amin, S.B., Arachchi, H., Arora, A., Auman, J.T., Ayala, B., et al. (2015). Genomic classification of cutaneous melanoma. Cell *161*, 1681–1696.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., et al. (2016). The Ensembl gene annotation system. Database (Oxford) *2016*, https://doi.org/10.1093/database/baw093, baw093.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, https://doi.org/10.1016/j.cell.2018.02.060.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Brunner, A.M., and Graubert, T.A. (2018). Genomics in childhood acute myeloid leukemia comes of age. Nat. Med. *24*, 7–9.

Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O., and Stein, L.D.; - ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net (2017). Pan-cancer analysis of whole genomes. bioRxiv. https://doi.org/10.1101/162784, 162784.

Cancer Genome Atlas Research Network, Raphael, B.J., Hruban, R.H., Aguirre, A.J., Moffitt, R.A., Yeh, J.J., Stewart, C., Robertson, A.G., Cherniack, A.D., Gupta, M., Getz, G., et al. (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. Cancer Cell *32*, 185–203.e13.

Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.P., Ahmann, G.J., Adli, M., et al. (2011). Initial genome sequencing and analysis of multiple myeloma. Nature *471*, 467–472.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Collisson, E.A., Campbell, J.D., Brooks, A.N., Berger, A.H., Lee, W., Chmielecki, J., Beer, D.G., Cope, L., Creighton, C.J., Danilova, L., et al. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.

Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938–2940.

Costello, J.C., and Stolovitzky, G. (2013). Seeking the wisdom of crowds through challenge-based competitions in biomedical research. Clin. Pharmacol. Ther. *93*, 396–398.

Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. *41*, e67.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res. *22*, 1589–1598.

Ewing, A.D., Houlahan, K.E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T.N., Bare, J.C., P'ng, C., Waggott, D., Sabelnykova, V.Y., et al. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat. Methods *12*, 623–630.

Fan, Y., Xi, L., Hughes, D.S.T., Zhang, J., Zhang, J., Futreal, P.A., Wheeler, D.A., and Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol. *17*, 178.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. *43*, D805–D811.

Goode, D.L., Hunter, S.M., Doyle, M.A., Ma, T., Rowley, S.M., Choong, D., Ryland, G.L., and Campbell, I.G. (2013). A simple consensus approach improves somatic mutation prediction accuracy. Genome Med. *5*, 90.

Hammerman, P.S., Lawrence, M.S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E.S., Gabriel, S., et al. (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519–525.

Hartmaier, R.J., Albacker, L.A., Chmielecki, J., Bailey, M., He, J., Goldberg, M.E., Ramkissoon, S., Suh, J., Elvin, J.A., Chiacchia, S., et al. (2017). High-throughput genomic profiling of adult solid tumors reveals novel insights into cancer pathogenesis. Cancer Res. *77*, 2464–2475.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kim, S., and Speed, T.P. (2013). Comparing somatic mutation-callers: beyond Venn diagrams. BMC Bioinformatics *14*, 189.

Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Cherniack, A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., et al. (2018).

Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Reports *23*, https://doi.org/10.1016/j.celrep.2018.03.076.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics *28*, 311–317.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome Biol. *17*, 122.

Project GENIE. (2017). Project GENIE goes public. Cancer Discov. *7*, 118.

Radenbaugh, A.J., Ma, S., Ewing, A., Stuart, J.M., Collisson, E.A., Zhu, J., and Haussler, D. (2014). RADIA: RNA and DNA integrated analysis for somatic mutation detection. PLoS One *9*, e111516.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Stein, L.D. (2010). The case for cloud computing in genome informatics. Genome Biol. *11*, 207.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Yang, T.-H.O., Porta-Pardo, E., Gao, G., Plaisier, C.L., Eddy, J.A., et al. (2018). The Immune Landscape of Cancer. Immunity *48*, https://doi.org/10.1016/j.immuni.2018.03.023.

Turnbull, C. (2018). Introducing whole genome sequencing into routine cancer care: the genomics England 100,000 genomes project. Ann. Oncol. https://doi.org/10.1093/annonc/mdy054.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Ye, K., Wang, J., Jayasinghe, R., Lameijer, E.W., McMichael, J.F., Ning, J., McLellan, M.D., Xie, M., Cao, S., Yellapantula, V., et al. (2015). Systematic discovery of complex insertions and deletions in human cancers. Nat. Med. *22*, 97–104.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
MC3 Files		https://gdc.cancer.gov/about-data/publications/mc3-2017
Software and Algorithms		
MuTect		https://github.com/broadinstitute/mutect
Pindel		https://github.com/genome/pindel
Radia		https://github.com/aradenbaugh/radia
VarScan2		http://dkoboldt.github.io/varscan/
SomaticSniper		https://github.com/genome/somatic-sniper
MuSE		https://github.com/danielfan/MuSE
Indelocator		http://archive.broadinstitute.org/cancer/cga/indelocator
Maf2Vcf		https://github.com/covingto/vcf2maf/

CONTACT FOR REAGENT AND RESOURCE SHARING

All data associated with this project will be made available via the NCI's GDC data portal, source code will be made available on GitHub and docker containers on the Docker and Quay docker repos. Questions can be directed to the contact author at ellrott@ohsu.edu

METHOD DETAILS

Sample List Creation

The MC3 sample list was extensively verified to make sure that poor quality samples were removed, and that for every donor the best tumor and normal samples were paired. To this end, a number of rules were applied to remove samples and identify appropriate sequence data which BAM files fit pipeline specifications as well as identify samples with available sequencing information that required preprocessing prior running analysis.

The list of rules applied included:

1.*Exclude redacted samples* - A number samples in the TCGA had been removed or flagged over the course of the TCGA project for various reasons.
2.*Exclude non-HG19 aligned files* - Earlier samples from the TCGA project were aligned with older genome builds, including HG17 and HG18. Rather than attempt to back-port variant calling platforms to older genomes and lift-over the variants to new genome builds, these samples were eliminated from the resource pool when building the sample list. In many cases the data from these files had been realigned by the Broad Firehose platform as part of their efforts in various tumor specific working groups.
3.*Preferentially select Broad genome build* - In cases where a sample's sequencing data had been run through multiple alignment pipelines, the Broad pipeline was preferentially selected to eliminate variance. In most cases when there were multiple pipeline runs, the Broad pipeline was run to update the alignments to an HG19 genome build.
4.*Ensure GATK co-cleaning/BQSR* - Co-cleaning refers to the process of applying the GATK IndelRealignment to both the tumor and normal samples of an individual. This process is also accompanied by running Base Quality Score Recalibration (BQSR). While complete realignment of sequences was not required for inclusion in the MC3 analysis, it was required that the GATK co-cleaning process has been applied. Because this step was part of the Broad pipeline, any sample selected fit this requirement, thus the previous rule. In cases where a sample was not co-cleaned and had not already been realigned as part of the Broad pipeline, the co-cleaning was done and the new sequences stored in a special project at CGHub.
5.*Exclude non-Illumina sequenced samples* -A small number of samples in the TCGA cohort had been sequenced with other technologies including ABI SOLiD and 454 for validation sequencing. To reduce artifacts and maintain consistency, these sample were eliminated from the list.
6.*Exclude FFPE samples* - Most of the TCGA samples were derived from fresh frozen samples, but a subset of samples were derived from Formalin-Fixed Paraffin-Embedded samples. These samples may have experienced more DNA damage and had different error profiles in mutation calling. This rule results in the removal of 97 samples.
7.*Matched genome build string* - While HG19 alignment was required for sample inclusion, there was in fact a number of different genome versions, including 'HomoSapien19' 'WustlBuild1' and others. These genome build were all based on HG19,

but contained various patches. Genome patches add additional sequencing information to the assembly, without disrupting the chromosome coordinates. But while these multiple patches were allowed, for a tumor and normal sample to be matched the genome build title had to match, to eliminate the possibility of sequence patches being misidentified as somatic mutations.
8.*Prefer Native DNA pairs over WGA pairs over Native+WGA* - There is a number of earlier TCGA samples which were sequences with Whole Genome Amplification. Because of the technical artifacts associated with this technique, in cases where there was sequencing done without WGA, those samples were preferentially selected.
9.*Prefer samples with matching RNA-Seq* - We selected samples that had quality measures based on RNA-Seq.
10.*Usually prefer latest plate* - Operating on the principle that any later sequencing effort would have been triggered by issues in the earlier runs, the latest run from a sample was chosen.
11.*Prefer pairs sequenced at the same center* - Sometime tumor normal pairs were sequenced at multiple centers. We selected for samples sequenced at the same center. This step was not adjusted based on ContEst or OxoG scores.
12.*Tumor contamination estimates using ContEst* - Samples were removed if the ContEst score estimated more than 4% contamination from another participant.
13.*Spurious sequence artifacts: BadSeq* - 6 samples were removed because they appeared to be affected by systematic sequencing artifacts. Systematic insertions or deletions were identified at the same base pair location in each of the reads in the both forward and reverse strands. These artifacts have been previously reported(Ye et al., 2015).

Given these rules, the sample selection algorithm is as follows:

1)Pick best bam within aliquot + original sequencing center. This involves applying all hard filters and picking samples with a preference toward BAMs processed via the Broad pipeline or the MC3 secondary co-cleaning pipeline.
2)Pick best set of BAMs within an individual. First selecting the most ‘‘popular’’ build, using Broad-aligned or number-of-native as tiebreakers, and avoid selecting WGA samples. Some overrides were applied in these step, ie selecting Baylor-aligned native samples vs Broad-aligned WGA samples.
3)Pare back the aliquots within the individual. First drop non-paired samples and select one aliquot per sample.

The final white list consisted of 11,069 tumor-normal pairs for 10,486 participants. In cases where more than one pair was selected for a participant, all of the pairs were analyzed for mutations, but all but one were tagged as ’nonpreferredpair’, based on criteria like preferring a primary to a metastatic tumor sample, and for solid tumor types preferring a blood to a tissue normal sample.

Variant Calling and Filtering Strategies

For the variant calling step, seven methods were applied, five covering Single Nucleotide Variant (SNV) calling and three covering short Insertion Deletion (INDEL) events, with Varscan 2 providing both types of analysis. Parameters used for these tools are found in Table S1.

1.MuTect (SNV) - This method at the Broad Institute(Cibulskis et al., 2013) uses a Bayesian classifier that allows it to identify low-read/low-allele fraction somatic mutations, while maintaining a high specificity. It was one of the top performing methods in the SMC-DNA DREAM challenge(Ewing et al., 2015).
2.Varscan 2 (SNV/INDEL) - Developed by Daniel Koboldt, Washington University, the algorithm uses heuristic and statistical approaches in its algorithm to detect germline, somatic and loss of heterozygosity. It can calculate SNV, Indel and CNA events(Koboldt et al., 2012).
3.Indelocator (INDEL) - Developed by the Broad team(Chapman et al., 2011) uses read count and alignment quality information to detect indel events found in tumor alignments.
4.Pindel (INDEL) - Developed by Kai Ye et al. at Washington University is used to identify medium size insertion and large deletion events. Pindel also generates complex variant calls that accurately reflect the genomic alterations even around substitution sites(Ye et al., 2009, 2015).
5.SomaticSniper (SNV) - Developed by David Larson et al. at Washington University, this method compares the tumor and normal bams to find differences using the samtools MAQ genotype likelihood model to make alteration calls(Larson et al., 2012).
6.RADIA (SNV) - Developed by Radenbaugh et al at University of California in Santa Cruz, RADIA stands for RNA and DNA Integrated Analysis. It augments it mutation calls using RNA-Seq samples from the same tumor making it possible to make mutation calls when there is lower DNA allelic frequencies. RADIA was applied using matched RNA when available(Radenbaugh et al., 2014).
7.MuSE (SNV) - Developed by Fan et al at Baylor College of Medicine and MD Anderson (Fan et al., 2016), uses a markov substitution model which characterizes the evolution of the allelic composition of the tumor and normal tissues at each reference base and is tuned for sensitivity. It further adopts a sample-specific error model that reflects the underlying tumor heterogeneity to improve the overall accuracy. Uses a markov substitution model to calls mutations. MuSE was another method that scored very well in the SMC-DNA DREAM Challenge.

Default parameters for programs were used as much as possible, however in a number of instances non-default parameters for particular programs were used, based on discussions with tool authors or analysis that had utilized the tool in institutional pipelines (Table S1). These selections were based on empirical knowledge gained by observing effects on small cohorts.

In the process of sample collection, DNA amplification and short read sequencing, there are a number of events that could induce noise and create false mutation patterns. Though callers are tuned to remove some classes of systematic sequence error, it is often necessary to impose additional post hoc filters. In some cases the techniques are already embedded in some of the mutation calling programs themselves, but to maintain consistency these filters were applied across all calls uniformly. We applied several common filters employed by major sequencing centers during the TCGA. The filtering steps do not increase sensitivity, they only remove calls, so sensitivity can only be decreased in this phase. Since false positive somatic events can be highly misleading for downstream research, maintaining high specificity of the call set using post hoc filters is crucial. The final call set was filtered to identify cohort level artifacts and was subject to extensive variant, subject, and cohort level QC. In sum, 22,485,627 putative variants were identified and 2,907,335 high confidence mutations were retained after filtering.

To provide filtering, 8 methods were utilized. The final two filtering methods are not necessarily designed to increase accuracy, Some of the variant calls marked by these methods may be correct, but were removed from the public open-access release in accordance with TCGA data access tiers.

1.Broad PON V2 - (MAF tag: broad_PoN_v2) One of the most effective filters of false-positive, contamination, and germline variants is a Panel-of-Normals (PoN) (Hess et al., unpublished data) filter. This filter postulates that if a variant is called or detected in a set of control (often non-tumor normal samples) then it is very unlikely that the variant is actually a somatic variant in any given tumor sample.
2.Common In ExAC - The Exome Aggregation Consortium (ExAC) publishes germline variants and recurrent artifacts seen in exome sequencing of over 60K unrelated individuals from across seven subpopulations(Lek et al., 2016). As implemented in vcf2maf v1.6.11, this filter tags variants with a non-reference allele count >16 in at least one subpopulation of the non-TCGA subset of ExAC v0.3.1, unless ClinVar flags it as pathogenic. AC=16 (for SF3B1:K700) was the highest value observed among known somatic events detected in the normal blood of older individuals due to clonal hematopoiesis.
3.OxoG - (MAF tag: oxog) The 8-Oxoguanine (OxoG) DNA lesion is a common sequence artifact caused by excessive oxidation during sequence library preparation(Costello et al., 2013). The DetOxoG tool was used to identify and flag likely OxoG error variants.
4.ContEst - (MAF tag: contest) This program predicts levels of contamination. Contamination coefficient produced by this method is used as a coefficient in the MuTect pipelines, and samples with a value greater than 4% were removed from the analysis.
5.StrandBias - (MAF tag: StrandBias) Implemented post MAF production and more appropriately identified as a mutation bias artifact, the StrandBias filter tags low-VAF G>T from samples sequenced at Washington University such that the number of untagged G>T variants equals the number of C>A variants within a sample. VAF cutoffs are set on a sample by sample basis such that the number of tagged G>T variants (with lowest VAF) maintains balanced untagged G>T mutations and C>A variant counts. This was implemented because of strong disparities between G>T and C>A mutation counts in samples sequenced at Washington University.
6.Normal Depth - (MAF tag: ndp) To avoid miscalling germline variants at least 8 reads in the normal sample in non-dbSNP sites and at least 19 reads in dbSNP sites.
7.Capture Kit - (MAF tag: 'bitgt') The filter represented a simple process of intersecting all mutations calls with the subset of the genome that intersected with all of the capture kits used by the different sequencing centers. During PCR small fractions of non-targeted sequences could be amplified and during alignment reads could have been placed in incorrect locations in the genome. This leads to low read coverage areas in non-targeted section of the genome to be included in the BAM file. If the variant calling program sweeps across of the the reads, it may produce calls using these off target reads, and create calls.
8.NonExomic - (MAF tag: NonExonic) As part of the NCI/NHGRI mutation release process, non-exonic mutations must be verified with orthogonal sequencing before they can be released publicly. The exon definitions were derived from the GAF 4.0 definition, which was based on Gencode 19 Basic.

Merger of Mutation Calls

Mutations were called by each of the callers and stored in VCF format. Following initial calling, variants from each caller were merged by allele with the exception of calls from Pindel. For alleles not involving Pindel, we extracted and averaged coverage metrics across all callers asserting the presence of a mutation and combined the various callers into the calling center column in the resulting MAF file. As Pindel generates complex variant calls we allowed Pindel to supersede allele representations from other callers. Any allele intersecting a Pindel call by position was discarded and the Pindel call was modified to add the other caller to the calling center column. We annotated these by placing a "star" next to the caller ID to signify that the caller may not have represented the allele in the same way.

Workflow Deployment

The various components of this part of the MC3 computational task took place at multiple sites using different technologies and computational resources.

1.UCSC NCI Cluster - A computational cluster, associated with the CGHub, was utilized to perform GATK co-cleaning on a subset of sequence files that had not been previously processed. This dataset represent approximately 1600 BAM files. The results of this run were stored on CGHub until its close in July 2016.
2.DNAnexus - The primary set of computations, related to running the core set of variant calling pipelines as run on DNAnexus's cloud platform. Over a four-week period approximately 1.8 million core-hours of computational time were used to process 400 TB of data on the DNAnexus Platform to yield reproducible results. This resulted in the 500GB of VCF files representing all detected variants.
3.Broad Firehose - The Broad Firehose is a system to deploy automated pipeline analysis on all the TCGA data. The somatic variant calling pipeline includes ContEst, MuTect, and Indelocator, and was run using an SGE cluster of 200 nodes. In addition, the OxoG filter was applied at this stage, and were also later applied to the calls from the other callers .
4.Institute of Systems Biology. These validation runs were deployed on the Institute of Systems Biology Cloud pilot. One this system, the OxoG variant filtering step was run on all variant data. Also, the WheelJack validation data genotyping algorithm was run on all samples with available validation data.

SMG Performance Analysis

MutSig2CV and MuSiC2 were performed on subsets of the data based on different filtering criteria. The results of this analysis resulted in drastically different results when taking filtered for raw variant calls. KIRC was selected because of its unique set of driver mutations compared to other tissues (*PBRM1* and *VHL*) and it is often associate with few SMGs. Variants for the raw variants were assembled for the unfiltered MAF. MutSig2CV consists of 3 statistical tests, including mutation abundance, local clustering, and conservation. SMGs from MutSig2CV were defined as genes with a q-value <= 0.1. MuSiC2 analysis calculates SMGs using mutation abundance compared to background mutations rate calculations. Convolutions of multiple transition and transversion rates were used to calculate p values. Strict p value cutoffs of 1e-7 were used in defining SMGs for MuSiC2. SMGs were further filtered using the MuSiC2 long gene filter. This is a MuSiC2 specific long gene filter systematically increases the p value threshold for larger genes until it no longer indicates a correlation between p value and gene size. If the larger gene doesn't reach the new threshold it is subsequently removed from the SMG list. This was not applied to MutSig2CV output. Filtered variants were processed using "pass-only" variants from the public facing, open-access MAF. The same parameters from the above were applied resulting in a reduced number of SMGs in KIRC. No hypermutators were removed for this analysis.

Mutation Validation

The Broad 'Mutation Validator' pipeline was used to identify validation evidence at variant sites using alternate sequencing runs. Mutation Validator provides validation evidence at sites of candidate SNVs or INDELs from read pileups across multiple data-types including WXS, WGS, Targeted Validation, and RNA. The algorithm for each validation followed the step:

1.Collect pileup for each allele (A<C<G<T, INS,DEL) at candidate sites from each validation data type.
2.Parse normal sample for each data type to estimate maximum noise alternate allele fraction. If datatype has no normal sample (eg. RNA-seq) then use exome to estimate noise. Use binomial conditional distribution field to calculate the 99% upper limit alt count in the tumor at this noise allele fraction. This upper limit is the threshold validation read count "min_val_count" in the tumor sample. The minimum "min_val_count" for any data type is 2.
3.The power to validate the mutation is calculated using the hypergeometric cumulative probability distribution to project the observed tumor alternate allele fraction from the exome onto the coverage of the validation data type with a minimum of "min_val_count" alt supporting reads. If power is less than 0.95, disregard this site+data type as unpowered.
4.If the normal sample for a given validation datatype has an allele fraction exceeding 0.2 for SNVs or 0.1 for INDELS, label the site+data type as "validation_judgement"=2 (germline).
5.If not germline, and if the tumor validation datatype alternate read count is at least "min_val_count" (from step 2) then label the site+datatype as "validation_judgement=1 (somatic).
6.Otherwise, set "validation_judgement"=0 (not validated).

Using this method 7,680,483 candidate variants processed by mutation validator (1,476,028 DEL, 603,637 INS, 5,600,818 SNP). The sites within the target region (bitgt) created a set of 1,352,467 variants having 95% power in either rna, targeted, wgs, or lowpass validation data. Validation rules at sites with power in targeted or wgs data.

QUANTIFICATION AND STATISTICAL ANALYSIS

Effects of Cancer Type on Mutation Callers

When observing the total number of mutations per sample, separated by cancer type, we identified that mutation calling consistency differs by cancer type. Specifically, within single nucleotide events THYM, and PAAD, KICH and UVM tumors varied greatly between sample when compared to the total number of unique variants identified per sample. Such inconsistencies are likely attributable to various pathological reasons that yield low purity biopsies. For instances, when comparing to purity estimates (Figure S2), THYM and PAAD samples had the lowest purity estimates (ABSOLUTE (syn7870168) median 39.0% and 39.7% respectively).

OxoG Events

The oxidation of guanine to 8-oxoguanine, known as the OxoG event, affects a subset of TCGA samples. It can be caused by heat, contamination and physical forces on the DNA. This mutation causes G to T and C to A substitutions in the reads. To filter for this event, an OxoQ score is calculated, which describes the probability of an entire sample being affected by OxoG events. If this OxoQ value is above a threshold, then the sample is run through the OxoG filter which examines the original BAM file reads to determine if G to T and C to A mutations are real or created by the OxoG artifact.

Per Gene Filtering Effects

Per gene counts were generated based on the number of variants found in the MC3 controlled-access and open access files. The genes with the largest disparity of variant counts between the two populations were assessed (Figure S4A). Additionally, significant cancer genes found as part of the original PanCan12 project were highlighted(Kandoth et al., 2013) in this analysis (Figure S4B).

Indel Realignment and BQSR

In order to remove biases in the alignment protocols, a process called ’co-cleaning’ was deployed, as part of the GATK best practices(McKenna et al., 2010), on each tumor normal pair. This processing step is composed of two analysis and adjustments that are run in the BAM files prior to variant calling. The first step, local realignment uses reads from both the tumor and normal, thus the ’co-cleaning’, and utilizes this information to disambiguate potential areas of misalignment. The tumor and normal are co-analyzed so that arbitrary decisions can be made cohesively. Areas with small insertions and deletions in the initial alignments were realigned using all reads from an individual, including reads from both the tumor and normal samples. This additional joint information help to eliminate false positive SNPs caused be misaligned reads, particularly at the 3’ end. There has been a noted performance increase in downstream variant calling process for both indel and SNV calling. Pindel incorporates a similar process internally and thus doesn’t require it, but for consistency all variant calling methods were based on the same co-cleaned BAMs.

The second step of co-cleaning is Base Quality Score Recalibration (BQSR). BQSR tweaks the quality score so that it represents a calibrated probability. This step is especially important for BAMs with a wide range of quality scores, as is common with older sequence data.

Co-cleaning had already been applied to all sequence alignments produced by the Broad since 2012, but for a subset of the TCGA cohort, totaling almost 50% of the pairs, the co-cleaning process was applied on samples already uploaded to the CGHub resource. Approximately 35% of the samples required full realignment. These secondary BAMs represented analysis products of the MC3 effort, and totalled almost 150TB. This processing was carried out at the Broad Institute and UCSC.

Variant Calling

The next phase in the MC3 process was variant calling followed by filtering. In the variant calling step, pairs of BAM files were run through programs developed from multiple institutions and the results of the putative variant calls were stored as Variant Call Format (VCF) files. The filtering steps, with the notable exception of the OxoG filter, use information stored in the VCF files produced by the different callers and produce a secondary filtered mutation file (usually VCF or MAF). This is an important detail for analysis and job scheduling. A pair of TCGA exome BAM files can average 10-30GB, while the average VCF file, filtered for somatic variants is a few hundred kilobytes. Many analysts employ a strategy of calling-then-filtering, ie create a set of putative variant calls and then applying filters as secondary steps downstream to remove false positives. If any information is required from the BAM file, it means that scheduling the analysis on the variant calls on 10K exomes would require accessing over 300TB of files. But if all of the filtering can be done only using the initial VCF file, the data requirements become tractable for doing analysis on a single machine. This strategy allows tuning of filtering methods, parameters and strategies but removes the complexity and logistical issues of obtaining the BAM files.

Variant Merger

We merged variants based on allelic location except in the case of Pindel calls, where we merged variants by proximity to Pindel calls. The additional merge criteria for Pindel calls was required because Pindel generates complex variant calls that other callers are incapable of generating. Complex variants are simultaneous indel and substitution mutations in cis. This merger process created 14,241 complex indel sites that included merged calls from SNP callers in the full MAF file, and 3,611 sites in the filtered open-access file. Finally, in order to generate consensus metrics, such as variant and reference allele counts, we averaged them across all callers that yielded a specific call.

Panel-of-Normals Filter

In the case of systematic false positive variants, as the cohort becomes larger the likelihood that one of the PoN samples will also contain the systematic false positive increases. By statistical chance it is possible to miss germline variants in low coverage regions because the variant is not detected in the normal, the PoN reduces the rate of germline calls because it effectively increases sequence depth at these locations by leveraging the control cohort. Although the PoN filter is an effective way to remove germline variants, most of the variants that it flags are, in fact, recurrent sequencing artifacts.

Across the entire cohort the number of germline SNP events for every site where totals and if a SNP occurs in a number of samples above a threshold, it was determined that it was more likely that a mutational event was not recognized as a germline event, rather than a genuine somatic event.

One of the the most effective filters encoded the expected distribution of alternate allele read counts at every genomic position, based on a large panel of 8000 TCGA normals (PoN). A somatic variant call is tagged by this filter if its observed readcount is consistent with the PoN, based on a likelihood test. This allows calls with many supporting reads to be retained, if they occur at a site with low allele-fraction (AF) sequencing noise in the PoN. To remove germline events or high AF artifacts, all somatic call at a site with recurrently high AF across the PoN are removed.

For each genomic position, the PoN encodes the distribution of alt read counts across all TCGA normals. For each mutation call, we compute a score that its observed read counts are consistent with the PoN; if this score is above a certain threshold, the site gets flagged. Thus, if a site recurrently harbors low-level sequencing noise in the PoN and it is called at low allelic fraction, it is flagged, whereas a call with many supporting reads at the same locus would be left alone. Likewise, a common germline site would have recurrently high allelic fractions across the PoN; if a call at that site has similarly high AF, it gets flagged.

A full description of the PoN filter follows. Each genomic position's histogram comprises six bins:

1: alt read count >= 1 and alt fraction >= 0.1%
2: alt read count >= 2 and alt fraction >= 0.3%
3: alt read count >= 3 and alt fraction >= 1%
4: alt read count >= 3 and alt fraction >= 3%
5: alt read count >= 3 and alt fraction >= 20%
6: alt read count >= 10 and alt fraction >= 20%

For a given position, denote the vector of bin counts $\overrightarrow{h}$. For each variant call, we represent its allelic fraction as a beta distribution parameterized by its alternate and reference read counts (to account for numerical uncertainty when converting read counts to allelic fraction):

$$f \sim \text{beta}(n_{\text{alt}}+1, n_{\text{ref}}+1),$$

and then slicing the beta distribution's PDF according to the alt. fraction bins encoded by the PoN, i.e.

$$\overrightarrow{f} = \left[\int_{0}^{0.1\%} df\, p(f), \int_{0.1\%}^{0.3\%} df\, p(f) \ldots, \int_{20\%}^{100\%} df\, p(f) \right].$$

Finally, we compute a score for this position by weighting each element of \$\vec $\overrightarrow{f}$ by its corresponding histogram bin counts:

$$S = \overrightarrow{f} \cdot \overrightarrow{h}$$

The units of this score are arbitrary. We found empirically that a cutoff of log10(S) $\geq$ -2.5 works well, determined by decreasing the score cutoff (thereby increasing the aggressiveness of the filter) until it started removing recurrently called sites ($\geq$3 patients) listed in the COSMIC database. Because some COSMIC sites are themselves recurrent artifacts, manual review was necessary to exclude those from the list of true positives.

Restricting to Target/Coding Exons

While there are whole genome sequences that are part of the TCGA catalogue, the MC3 project targeted exome sequences. During PCR small fractions of non-targeted sequences could be amplified and during alignment reads could have been placed in incorrect locations in the genome. This leads to low read coverage areas in non-targeted section of the genome to be included in the BAM file. If the variant calling program sweeps across of the reads, it may produce calls using these off target reads, and create calls. To filter these non-target calls out, a BED file of the intersection of capture kit locations and applied to the variant calls to remove variant calls from non-target/non-exon regions. This target filter was applied across all samples, even on samples where other targeting panels may have been used because 1) not all capture kit targeting data were universally available and well annotated to sequences and 2) to simply cohort mutation significance analysis. The disadvantages of the capture kit based filtering strategy was that 170 CDS altering MC3 calls in MSK IMPACT's 410 cancer genes, that fall outside the Broad BED. The key misses are TERT promoter hits, truncations in putative tumor-suppressor CIC, splice alterations in the frequently rearranged CRLF2, and a cluster of events in the 5' end of FOXP1.

The exone definitions were derived from the GAF 4.0 definition, which was based on Gencode 19 Basic. The exome capture was based on the Broad Target Bed.

Minimum 3 Supporting Reads for Pindel Indel Calls

Some of the filtering parameters in Pindel were recently reconfigured to allow it to detect complex indel events. Complex indel events involve both the insertion and deletion of nucleotides in a mutation site(Ye et al., 2015). This increased ability of Pindel resulted in a number of false positive indel being included as part of the initial MC3 call set. To combat this, a minimum of three supporting reads were required to support a Pindel call, otherwise it was filtered out.

Minimum Indelocator Indel Depth

For analyses in this manuscript we restricted Indelocator calls to indels depth of greater than or equal to 3 supporting alternate reads.

Annotation

Additional annotations were added from COSMIC(Forbes et al., 2015), dbGaP(Sherry et al., 2001), ExAC(Lek et al., 2016), and Ensembl(Aken et al., 2016) using Variant Effect Predictor (VEP)(McLaren et al., 2016) and other custom built annotation tools including the normal depth of coverage filter and strand bias filters. The final call set was filtered to identify cohort level artifacts and was subject to extensive variant, subject, and cohort level QC.

DATA AND SOFTWARE AVAILABILITY

Data have been made available at the NCI's Genomic Data Commons. Result MAF files of the MC3 dataset is available in two different versions, the open-access and controlled-access data files. Additionally, intermediate files, such as the original called VCF and annotation marking files have been made available.

All pipelines and software developed as part of this project have been made available in https://github.com/OpenGenomics/mc3

Reference Files and intermediate result files have been made available at https://gdc.cancer.gov/about-data/publications/mc3-2017