AACR
American Association
for Cancer Research
FINDING CURES TOGETHER℠

PROJECTGENIE
Genomics Evidence Neoplasia Information Exchange

AACR GENIE Data Guide

About this Document

This document provides an overview of the first public release of American Association for Cancer Research (AACR) GENIE data.

Version of Data

AACR GENIE Project Data: Version 5.0-public

AACR Project GENIE data versions follow a numbering scheme derived from semantic versioning, where the digits in the version correspond to: major.patch-release-type. “Major” releases are public releases of new sample data. “Patch” releases are corrections to major releases, including data retractions. “Release-type” refers to whether the release is a public AACR Project GENIE release or a private/consortium-only release. Public releases will be denoted with the nomenclature “X.X-public” and consortium-only private releases will be denoted with the nomenclature “X.X-consortium”.

Data Access

AACR GENIE Data is currently available via two mechanisms:

- Sage Synapse Platform: http://synapse.org/genie
- cBioPortal for Cancer Genomics: http://www.cbioportal.org/genie/

Terms of Access

All users of the AACR Project GENIE data must agree to the following terms of use; failure to abide by any term herein will result in revocation of access.

- Users will not attempt to identify or contact individual participants from whom these data were collected by any means.
- Users will not redistribute the data without express written permission from the AACR Project GENIE Coordinating Center (send email to: info@aacrgenie.org).

When publishing or presenting work using or referencing the AACR Project GENIE dataset please include the following attributions:

- Please cite: *The AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine Through An International Consortium, Cancer Discov. 2017 Aug;7(8):818-831* and include the version of the dataset used.
- The authors would like to acknowledge the American Association for Cancer Research and its financial and material support in the development of the AACR Project GENIE registry, as well as members of the consortium for their commitment to data sharing. Interpretations are the responsibility of study authors.

Posters and presentations should include the AACR Project GENIE logo.

Introduction to AACR GENIE

The AACR Project Genomics, Evidence, Neoplasia, Information, Exchange (GENIE) is a multi-phase, multi-year, international data-sharing project that aims to catalyze precision cancer medicine. The GENIE platform will integrate and link clinical-grade cancer genomic data with clinical outcome data for tens of thousands of cancer patients treated at multiple international institutions. The project fulfills an unmet need in oncology by providing the statistical power necessary to improve clinical decision-making, to identify novel therapeutic targets, to understand of patient response to therapy, and to design new biomarker-driven clinical trials. The project will also serve as a prototype for aggregating, harmonizing, and sharing clinical-grade, next-generation sequencing (NGS) data obtained during routine medical practice.

The data within GENIE is being shared with the global research community. The database currently contains CLIA-/ISO-certified genomic data obtained during the course of routine practice at multiple international institutions (Table 1), and will continue to grow as more patients are treated at additional participating centers.

Table 1: AACR GENIE Contributing Centers.

Center Abbreviation	Center Name
CRUK	Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, England
DFCI	Dana-Farber Cancer Institute, Boston, MA, USA
GRCC	Institut Gustave Roussy, France
JHU	Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD, USA
MDA	The University of Texas MD Anderson Cancer Center, Houston, TX, USA
MSK	Memorial Sloan Kettering Cancer Center, New York, NY, USA
NKI	Netherlands Cancer Institute, on behalf of the Center for Personalized Cancer Treatment, The Netherlands
UCSF	University of California-San Francisco (UCSF Helen Diller Family Comprehensive Cancer Center), San Francisco, California, USA
UHN	Princess Margaret Cancer Centre, University Health Network, Toronto, Canada

VICC	Vanderbilt-Ingram Cancer Center, Nashville, TN, USA
WAKE	Wake Forest University Health Sciences (Wake Forest Baptist Medical Center), Winston-Salem, NC, USA

Human Subjects Protection and Privacy

Protection of patient privacy is paramount, and the AACR GENIE Project therefore requires that each participating center share data in a manner consistent with patient consent and center-specific Institutional Review Board (IRB) policies. The exact approach varies by center, but largely falls into one of three categories: IRB-approved patient-consent to sharing of de-identified data, captured at time of molecular testing; IRB waivers and; and IRB approvals of GENIE-specific research proposals. Additionally, all data has been de-identified via the HIPAA Safe Harbor Method. Full details regarding the HIPAA Safe Harbor Method are available online at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/.

Summary of Data by Center

The first data release includes genomic and clinical data from eight cancer centers. Tables 2-3 summarize genomic data provided by each of the eight centers, followed by descriptive paragraphs describing genomic profiling at each of the participating GENIE center.

Table 2: Genomic Data Characterization by Center.

	Specimen Types	Specimen Tumor Cellularity	Assay Type	Coverage				Platform		Calling Strategy	Alteration Types				
	Formalin-fixed, paraffin-embedded (FFPE) v. Fresh Frozen (Fresh Froz)	Tumor Cellularity Cutoff	Hybridization Capture v. PCR	Hotspot Regions	Coding Exons	Introns (selected)	Promoters (selected)	Illumina	Ion Torrent	Unmatched (Tumor-only) v. Matched (Tumor-Normal)	Single Nucleotide Variants (SNV)	Small Indels	Gene-Level CNA	Intragenic CNA	Structural Variants
CRUK	Fresh Froz	>10%	Hybridization	x				X			x	x	x	x	

DFCI	FFPE	>20%	Capture		X	X		x		Tumor Only	x	x	x		[1]
GRCC	Fresh Froz	>10%	PCR	x					x	Tumor Only	x	x			
JHU	FFPE	>10%	PCR	x					x	Tumor Only	x	x			
MDA	FFPE	>20%	PCR	x					x	Tumor Only	x	x	[1]		
MSK	FFPE	>10%	Capture		X	X	x	x		Tumor-Normal	x	x	x	x	x
NKI	FFPE	>10%	PCR	x				x		Tumor Only	x	x			
UCSF	FFPE Fresh Froz[2]	>25%	Capture		X	X	x	x		Matched & Tumor-only	x	x	x	x	x
UHN-solid	FFPE	>10%	PCR	X	X			X	X	Tumor-Normal or Tumor Only	X	X			
UHN-myeloid	FFPE	>10%	PCR	X	X			X	X	Tumor only	X	X			
VICC	FFPE	>20%	Capture		X	X		x		Tumor Only	x	x	x		x
VICC-solid/myeloid	FFPE	>10%	PCR	X						Tumor Only	X	X			
WAKE	FFPE, Fresh	>20%	Capture		X			X		Tumor Only	X	X			

[1] Structural variants or copy number events are identified and reported but have not been transferred to GENIE.

Table 3: Gene Panels Submitted by Each Center.

Panel File (all files are prepended as: data_gene_panel_XXX)	**Panel Type** (PCR/Capture)	**All Exons v. Hotspot Regions**	**# of Genes**
CRUK-TS.BED	Custom	Hotspot Regions	173
DFCI-ONCOPANEL-1.TXT	Custom	All Exons	275
DFCI-ONCOPANEL-2.TXT	Custom	All Exons	300

DFCI-ONCOPANEL-3.TXT	Custom	All Exons	447
MSK-IMPACT341.TXT	Custom	All Exons	341
MSK-IMPACT410.TXT	Custom	All Exons	410
MSK-IMPACT468.TXT	Custom	All Exons	468
GRCC-CHP2.TXT	Ion AmpliSeq Cancer Hotspot Panel v2	Hotspot Regions	50
GRCC-MOSC3.TXT	Ion AmpliSeq Cancer Hotspot Panel v2	Hotspot Regions	74
GRCC-MOSC4.TXT	Ion AmpliSeq Cancer Hotspot Panel v2 + Custom	Hotspot Regions	82
JHU-50GP-V2.TXT	Ion AmpliSeq Cancer Hotspot Panel v2	Hotspot Regions	50
MDA-46-V1.TXT	Custom, based on Ion AmpliSeq Cancer Hotspot Panel v1	Hotspot Regions	46
MDA-50-V1.TXT	Ion AmpliSeq Cancer Hotspot Panel v2	Hotspot Regions	50
MDA-409-V1.TXT	Ion AmpliSeq Comprehensive Cancer Panel	All Exons	409
NKI-TSACP-MISEQ-NGS.TXT	TruSeq Amplicon Cancer Panel	Hotspot Regions	48
UCSF-NIMV4.TXT	Custom	Coding exons, select promoters (TERT, SDHD only) and intronic/UTR regions (47 genes)	481
UHN-48-V1.TXT	TruSeq Amplicon Cancer Panel	Hotspot Regions	48
UHN-50-V2.TXT	Ion AmpliSeq Cancer Hotspot Panel v2	Hotspot Regions	50
UHN-2-V1.TXT	Sequenom MassArray	Hotspot Regions	2

UHN-54-V1.TXT	TruSight Myeloid Sequencing Panel	Hotspot Regions – 39 genes, Full gene- 15 genes	54
VICC-01-T5A.TXT	Foundation Medicine	All Exons	322
VICC-01-T7.TXT	Foundation Medicine	All Exons	429
VICC-01-SOLIDTUMOR	Custom	Hotspot Regions	31
VICC-01-MYELOID	Custom	Hotspot Regions	37
WAKE-CA-01.BED	Caris	All Exons	32
WAKE-CA-NGSQ3.BED	Caris	All Exons	577
WAKE-CLINICAL-R2D2.BED	Foundation Medicine	All Exons	234
WAKE-CLINICAL-T5A.BED	Foundation Medicine	All Exons	70
WAKE-CLINICAL-T7.BED	Foundation Medicine	All Exons	308

Genomic Profiling at Each Center

Cancer Research UK Cambridge Centre, University of Cambridge (CRUK)
Sequencing data (SNVs/Indels):

DNA was quantified using Qubit HS dsDNA assay (Life Technologies, CA) and libraries were prepared from a total of 50 ng of DNA using Illumina's Nextera Custom Target Enrichment kit (Illumina, CA). In brief, a modified Tn5 transposase was used to simultaneously fragment DNA and attach a transposon sequence to both end of the fragments generated. This was followed by a limited cycle PCR amplification (11 cycles) using barcoded oligonucleotides that have primer sites on the transposon sequence generating 96 uniquely barcoded libraries per run. The libraries were then diluted and quantified using Qubit HS dsDNA assay.

Five hundred nanograms from each library were pooled into a capture pool of 12 samples. Enrichment probes (80-mer) were designed and synthesized by Illumina; these probes were designed to enrich for all exons of the target genes, as well for 500 bp up- and downstream of the gene. The capture was performed twice to increase the specificity of the enrichment. Enriched libraries were amplified using universal primers in a limited cycle PCR (11 cycles). The quality of the libraries was assessed using Bioanalyser (Agilent Technologies, CA) and quantified using KAPA Library Quantification Kits (Kapa Biosystems, MA).

Products from four capture reactions (that is, 48 samples) were pooled for sequencing in a lane of Illumina HiSeq 2,000. Sequencing (paired-end, 100 bp) of samples and demultiplexing of libraries was performed by Illumina (Great Chesterford, UK).

The sequenced reads were aligned with Novoalign, and the resulting BAM files were preprocessed using the GATK Toolkit. Sequencing quality statistics were obtained using the GATK's DepthOfCoverage tool and Picard's CalculateHsMetrics. Coverage metrics are presented in Supplementary Fig. 1. Samples were excluded if <25% of the targeted bases were covered at a minimum coverage of 50 × .

The identities of those samples with copy number array data available were confirmed by analyzing the samples' genotypes at loci covered by the Affymetrix SNP6 array. Genotype calls from the sequencing data were compared with those from the SNP6 data that was generated for the original studies. This was to identify possible contamination and sample mix-ups, as this would affect associations with other data sets and clinical parameters.

To identify all variants in the samples, we used MuTect (without any filtering) for SNVs and the Haplotype Caller for indels. All reads with a mapping quality <70 were removed prior to calling. Variants were annotated with ANNOVAR using the genes' canonical transcripts as defined by Ensembl. Custom scripts were written to identify variants affecting splice sites using exon coordinates provided by Ensembl. Indels were referenced by the first codon they affected irrespective of length; for example, insertions of two bases and five bases at the same codon were classed together.

To obtain the final set of mutation calls, we used a two-step approach, first removing any spurious variant calls arising as a consequence of sequencing artefacts (generic filtering) and then making use of our normal samples and the existing data to identify somatic mutations (somatic filtering). For both levels of filtering, we used hard thresholds that were obtained, wherever possible, from the data itself. For example, some of our filtering parameters were derived from considering mutations in technical replicates (15 samples sequenced in triplicate). We compared the distributions of key parameters (including quality scores, depth, VAF) for concordant (present in all three replicates) and discordant (present in only one out of three replicates) variants to obtain thresholds, and used ROC analysis to select the parameters that best identified concordant variants.

SNV filtering

- Based on our analysis of replicates, SNVs with MuTect quality scores <6.95 were removed.

- We removed those variants that overlapped with repetitive regions of MUC16 (chromosome 19: 8,955,441–9,044,530). This segment contains multiple tandem repeats (mucin repeats) that are highly susceptible to misalignment due to sequence similarity.

- Variants that failed MuTect's internal filters due to 'nearby_gap_events' and 'poor_mapping_regional_alternate_allele_mapq' were removed.

- Fisher's exact test was used to identify variants exhibiting read direction bias (variants occurring significantly more frequently in one read direction than in the other; FDR=0.0001). These were filtered out from the variant calls.

- SNVs present at VAFs smaller than 0.1 or at loci covered by fewer than 10 reads were removed, unless they were also present and confirmed somatic in the Catalogue of Somatic Mutations in Cancer (COSMIC). The presence of well-known PIK3CA mutations present at low VAFs was confirmed by digital PCR (see below), and supported the use of COSMIC when filtering SNVs.

- We removed all SNVs that were present in any of the three populations (AMR, ASN, AFR) in the 1,000 Genomes study (Phase 1, release 3) with a population alternate allele frequency of >1%.

- We used the normal samples in our data set (normal pool) to control for both sequencing noise and germline variants, and removed any SNV observed in the normal pool (at a VAF of at least 0.1). However, for SNVs present in more than two breast cancer samples in COSMIC, we used more stringent thresholds, removing only those that were observed in >5% of normal breast tissue or in >1% of blood samples. The different thresholds were used to avoid the possibility of contamination in the normal pool affecting filtering of known somatic mutations. This is analogous to the optional 'panel of normals' filtering step used by MuTect in paired mode, in which mutations present in normal samples are removed unless present in a list of known mutations61.

Indel filtering

- As for SNVs, we removed all indels falling within tandem repeats of MUC16 (coordinates given above).

- We removed all indels deemed to be of 'LowQual' by the Haplotype Caller with default parameters (Phred-scaled confidence threshold=30).

- As for SNVs, we removed indels displaying read direction bias. Indels with strand bias Phred-scaled scores >40 were removed.

- We downloaded the Simple Repeats and Microsatellites tracks from the UCSC Table Browser, and removed all indels overlapping these regions. We also removed all indels that overlapped homopolymer stretches of six or more bases.

- As for SNVs, indels were removed if present in the 1,000 Genomes database at an allele frequency >1%, or if they were present in normal samples in our data set. Thresholds were adjusted as for SNVs if the indel was present in COSMIC. The same thresholds for depth and VAF were used.

Microarray data (Copy number):

DNA was hybridized to Affymetrix SNP 6.0 arrays per the manufacturer's instructions. ASCAT was used to obtain segmented copy number calls and estimates of tumour ploidy and purity. Somatic CNAs were obtained by removing germline CNVs as defined in the original METABRIC study3. We defined regions of LOH as those in which there were no copies present of either the major or minor allele, irrespective of total copy number. Recurrent CNAs were identified with GISTIC2, with log2 ratios obtained by dividing the total number of copies by tumour ploidy for each ASCAT segment. Thresholds for identifying gains and losses were set to 0.4 and (−)0.5, respectively; these values were obtained by examining the distribution of log2 ratios to identify peaks associated with copy number states. A broad length cut-off of 0.98 was used, and peaks were assessed to rule out probe artefacts and CNVs that may have been originally missed.

Dana-Farber Cancer Institute (DFCI)

DFCI uses a custom, hybridization-based capture panel (OncoPanel) to detect single nucleotide variants, small indels, copy number alterations, and structural variants from tumor-only sequencing data. Three (3) versions of the panel have been submitted to GENIE: version 1 containing 275 genes, version 2 containing 300 genes, version 3 containing 447 genes. Specimens are reviewed by a pathologist to ensure tumor cellularity of at least 20%. Tumors are sequenced to an average unique depth of coverage of approximately 200x for version 1 and 350x for version 2. Reads are aligned using BWA, flagged for duplicate read pairs using Picard Tools, and locally realigned using GATK. Sequence mutations are called using MuTect for SNVs and GATK SomaticIndelDetector for small indels. Putative germline variants are filtered out using a panel of historical normals or if present in ESP at a frequency >= .1%, unless the variant is also present in COSMIC. Copy number alterations are called using a custom pipeline and reported for fold-change >1. Structural rearrangements are called using BreaKmer. Testing is performed for all patients across all solid tumor types. Version 3 includes the exonic regions of 447 genes and 191 intronic regions across 60 genes targeted for rearrangement detection. 52 genes present in previous versions were retired in the v3 test.

Institut Gustave Roussy (GRCC)

Gustave Roussy Cancer Centre submitted data includes somatic variants (single nucleotide variants and small indels) identified with Cancer Hotspot Panel v2 from tumor-only sequence data. Several versions of the panel have been used: CHP2 covering hotspots in 50 genes, MOSC3 covering hotspots in 74 genes and MOSC4 covering 89 genes. Tumors are sequenced to an average unique depth of coverage of >500X. The sequencing data were analyzed with the Torrent Suite^{TM} Variant Caller 4.2 and higher and reported somatic variants were compared with the reference genome GRCh37 (hg19). The variants were called if >5 reads supported the variant and/or total base depth >50 and/or variant allele frequency >1% was observed. All the variants identified were visually controlled on .bam files using Alamut v2.4.2 software (Interactive Biosoftware). All the germline variants found in 1000 Genomes Project or ESP (Exome Sequencing Project database) with frequency >0.1% were removed. All somatic mutations were annotated, sorted, and interpreted by an expert molecular biologist according to available databases (COSMIC, TCGA) and medical literature.

The submitted data set was obtained from selected patients that were included in the MOSCATO trial (MOlecular Screening for CAncer Treatment Optimization) (NCT01566019). This trial collected on-purpose tumour samples (from the primary or from a metastatic site) that are immediately fresh-frozen, and subsequently analyzed for targeted gene panel sequencing. Tumour cellularity was assessed by a senior pathologist on a haematoxylin and eosin slide from the same biopsy core to ensure tumor cellularity of at least 10%.

The University of Texas MD Anderson Cancer Center (MDA)

The University of Texas MD Anderson Cancer Center submitted data in the current data set includes sequence variants (small indels and point mutations) identified using an amplicon-based targeted hotspot tumor-only assay, and sequence variants/gene level amplifications identified on an amplicon-based exonic gene panel which incorporates germline variant subtraction (MDA-409). Two different amplicon pools and pipeline versions are included for the hotspot tumor-only assays: a 46-gene assay (MDA-46) corresponding to customized version of AmpliSeq Cancer Hotspot Panel, v1 (Life Technologies), and a 50-gene assay (MDA-50) corresponding to the AmpliSeq Hotspot Panel v2. The exonic assay with germline variant subtraction and amplification detection corresponds to the AmpliSeq Comprehensive Cancer Panel. DNA was extracted from unstained sections of tissue paired with a stained section that was used to ensure adequate tumor cellularity (human assessment > 20%) and marking of the tumor region of interest (macrodissection). Sequencing was performed on an Ion Torrent PGM (hotspot) or Proton (exonic). Tumors were sequenced to a minimum depth of coverage (per amplicon) of approximately 250X. Bioinformatics pipeline for MDA-46 was executed using TorrentSuite 2.0.1 signal processing, basecalling, alignment and variant calling. For MDA-50, TorrentSuite 3.6 was used. Initial calls were made by Torrent Variant Caller (TVC) using low-stringency somatic parameters. For MDA-50, TorrentSuite 3.6 was used. For MDA-409, TorrentSuite 4.4 was used. For MDA-409, TorrentSuite 4.4 was used. Initial calls were made by Torrent Variant Caller (TVC) using low-stringency somatic parameters.

All called variants were parsed into a custom annotation & reporting system, OncoSeek, with a back-end SQL Server database using a convergent data model for all sequencing platforms used by the laboratory. Calls were reviewed with initial low stringency to help ensure that low effective tumor cellularity samples do not get reported as false negative samples. Nominal variant filters (5% variant allelic frequency minimum, 25 variant coverage minimum, variant not present in paired germline DNA for the exonic assay) can then be applied dynamically. Clinical sequencing reports were generated using OncoSeek to transform genomic representations into HGVS nomenclature. To create VCF files for this project, unfiltered low stringency VCF files were computationally cross checked against a regular expressions-based variant extract from clinical reports. Only cases where all extracted variants from the clinical report were deterministically mappable to the unfiltered VCF file and corresponding genomic coordinates were marked for inclusion in this dataset. This method filters a small number of cases where complex indels may not have originally been called correctly at the VCF level. Testing is performed for patients with advanced metastatic cancer across all solid tumor types.

Memorial Sloan Kettering Cancer Center (MSK)

MSK uses a custom, hybridization-based capture panel (MSK-IMPACT) to detect single nucleotide variants, small indels, copy number alterations, and structural variants from matched tumor-normal sequence data. Three (3) versions of the panel have been submitted to GENIE: version 1 containing 341 genes, version 2 containing 410 genes, version 3 containing 468 genes. Specimens are reviewed by a pathologist to ensure tumor cellularity of at least 10%. Tumors are sequenced to an average unique depth of coverage of approximately 750X. Reads are aligned using BWA, flagged for duplicate read pairs using GATK, and locally realigned using ABRA. Sequence mutations are called using MuTect, VarDict, and Somatic indel detector, and reported for >5% allele frequency (novel variants) or >2% allele frequency (recurrent hotspots). Copy number alterations are called using a custom pipeline and reported for fold-change >2. Structural rearrangements are called using Delly. All somatic mutations are reported without regard to biological function. Testing is performed for patients with advanced metastatic cancer across all solid tumor types.

Johns Hopkins Sidney Kimmel Comprehensive Cancer Center (JHU)

Johns Hopkins submitted genomic data from the Ion AmpliSeq Cancer Hotspot Panel v2, which detects mutations in cancer hotspots from tumor-only analysis. Data from the JHU_50GP_V2 panel covering frequently mutated regions in 50 genes was submitted to GENIE. Pathologist inspection of an H&E section ensured adequate tumor cellularity (approximately 10% or greater). DNA was extracted from the macro-dissected FFPE tumor region of interest. Tumors are sequenced to an average unique read depth of coverage of greater than 500X. For alignment the TMAP aligner developed by Life Technology for the Ion Torrent sequencing platform is used to align to hg19/GRCh37 using the manufacturer's suggested settings. Tumor variants are called with a variety of tools. Samtools mpileup is run on the aligned .bam file and then processed with custom perl scripts (via a naive variant caller) to identify SNV and INS/DEL. Specimen variant filters have a total read depth filter of >= 100, a variant allele coverage of >= 10, variant allele frequency for substitutions >= 0.05, variant allele frequency for small (less than 50 base pair) insertions or deletions >= 0.05, and "strand bias" of total reads and of variant alleles are both less than 2-fold when comparing forward and reverse reads. Additionally, variants seen in greater than 20% of a set of non-neoplastic control tissues (>3 of 16 samples) with the same filter criteria are excluded. Finally, variants documented as “common” in dbSNP and not known to COSMIC are excluded. The cohort includes both primary and metastatic lesions and some repeated sampling of the same patient.

Netherlands Cancer Center (NKI), The Netherlands

NKI uses Illumina TruSeq Amplicon – Cancer Panel (TSACP) to detect known cancer hotspots from tumor-only sequencing data. A single gene panel, NKI-TSACP covering known hotspots in 48 genes with 212 amplicons has been used. Specimens are reviewed by a pathologist to ensure tumor cellularity of at least 10%. Tumors are sequenced to an average unique depth of coverage of approximately 4000x. The sample plate and sample sheet are made using the Illumina Experiment Manager software before running the sample on the MiSeq Sequencing System (Illumina, SY-410-1003) and MiSeq Reporter (v2.5) is used for data analysis. Reads are aligned using Banded Smith Waterman (v2.5.1.3), and samtools is used to further sort and index the BAM files. Variant calling is performed via the Illumina somatic variant caller (v3.5.2.1). Further detailed variant analysis (e.g. removal of known artifacts, known benign SNPs and variants with read depth < 200 or VAF < 0.05 and manual classification) is performed via Cartagenia BenchLab (https://cartagenia.com/). Testing is performed for all patients across all solid tumor types.

University of California-San Francisco (UCSF Helen Diller Family Comprehensive Cancer Center) (UCSF)

- UCSF uses a custom, hybridization-based capture panel (UCSF500) to detect single nucleotide variants, small indels, copy number alterations, and structural variants from both matched tumor-normal and tumor-only specimens. The current version of the assay consists of 481 genes and includes coverage of select promoter regions (*TERT* and *SDHD*) as well as the intronic or UTR regions of 47 genes for the detection of structural rearrangements. Testing is performed for patients with solid or hematological malignancies. Specimens are reviewed by a pathologist to ensure tumor cellularity of greater than 25%. Tumor DNA is extracted from sections of FFPE tissue; for uveal melanoma cases, frozen fresh fine needle aspirates are accepted. Normal DNA can be extracted from peripheral blood draw, buccal swab, or micro-dissected non-lesional areas. Hybridization capture is performed with SeqCap EZ target enrichment kit; sequencing platform is the HiSeq2500. Tumors are sequenced to an average unique depth of coverage of approximately >500X. FASTQC is run on unaligned sequencing reads to collect read-level summary statistics for downstream quality control; additionally, a suite of Picard tools are also run to assess quality metrics from sequencing runs. BWA-MEM aligner is used to align sequencing reads from each sample to the reference genome (hg19). The following bioinformatic workflows are used for variant calling:

- SNV callers:
    - Tumor sample: FreeBayes, GATK UnifiedGenotyper, Pindel
    - Normal sample: FreeBayes, GATK HaplotypeCaller, Pindel
    - Matched pairs: FreeBayes, Mutect, GATK SomaticIndelDetector
- Structural variant callers:
    - DELLY
    - Pindel calls larger than 100bp are treated as structural variants
- Copy Number Calls:
    - CNVkit using a reference profile for normalization of 30 pooled normal samples

  Variants are removed if present with frequency >= 1% in ESP6500 or 1000 Genomes datasets, or >=5% in ExAC. Known sequencing artifacts are removed. Variants with < 50x total coverage in the tumor sample are removed.

Princess Margaret Cancer Centre, University Health Network (UHN)
Princess Margaret Cancer Centre used three (3) panels to sequence samples for the GENIE 3.0.0 release - UHN-48-V1, UHN-50-V2, and UHN-54-V1. Each panel is described below:

Illumina TruSeq Amplicon panel (UHN-48-V1): Princess Margaret Cancer Centre used the TruSeq Amplicon Cancer Panel (TSACP, Illumina) to detect single nucleotide variants and small indels from matched tumor-normal sequencing data. Specimens are reviewed by a pathologist to ensure tumor cellularity of at least 20%. Tumors are sequenced to an average unique depth of coverage of approximately 500x and normal blood samples to 100x. Data was processed using one of two workflows:

1. Data analysis of tumor-normal pairs processed by UHN_TSACP_workflow_v2: MiSeq fastq were aligned using (MiSeq Reporter v2.4.60 and the corresponding default version of hg19) followed by local realignment and BQSR using GATK v3.3.0. Somatic sequence mutations were called, using MuTect (v1.1.5) for SNVs and Varscan (v2.3.8) for indels, using both normal and tumor data. Data were filtered to ensure there are no variants included with frequency of 3% or more in the normal sample. Results were filtered to keep only those with tumor variant allele frequency of at least 10%.

2. Data analysis of tumor only processed by UHN_TSACP_tumorONLY_v2_workflow: MiSeq fastq were aligned using (MiSeq Reporter v2.4.60 and the corresponding default version of hg19) followed by local realignment and BQSR using GATK v3.3.0. Sequence mutations (SNV and indel) were called using Varscan (v2.3.8). Results were filtered to keep only those with tumor variant allele frequency of at least 10%.

ThermoFisher Ion AmpliSeq Cancer Panel (UHN-50-V2): Princess Margaret Cancer Centre also used the TruSeq Amplicon Cancer Panel (TSACP, Illumina) to detect single nucleotide variants and small indels from matched tumor-normal sequencing data. Specimens were reviewed by a pathologist to ensure tumor cellularity of at least 20%. Tumors were sequenced to an average unique depth of coverage of approximately 500x and normal blood samples to 100x. Ion Torrent data was converted to fastq and sequences were aligned using NextGENe Software v2.3.1. NextGENe Software v2.3.1 provides a version of hg19 (Human_v37_3_dbsnp_135_dna). NextGENe was used to call SNV and indels. Results were then filtered to keep all with VAF of at least 10% and total coverage of at least 100x.

Sequenom MassArray Panel (UHN-2-V1): Princess Margaret Cancer Centre also used the Sequenom MassArray Solid Tumor Panel v1.0 assay to detect variants from tumor samples. Mutation calling was determined using the TyperAnalyzer software and results were filtered to include only postions listed in the bed file.

Illumina TruSeq Myeloid Sequencing Panel (UHN-54-V1): Princess Margaret Cancer Centre also used the TruSeq Myeloid Sequencing Panel (Illumina) to detect single nucleotide variants and small indels in DNA from bone marrow or peripheral blood samples from patients with acute leukemia, myelodysplastic syndrome, or myeloproliferative neoplasms. The diagnosis of each patient was confirmed by hematopathologist using the 2016 revision of the World Health Organization classification system for myeloid neoplasms. Tumors were sequenced to an average unique depth of coverage of approximately 500x. MiSeq fastq were aligned using (MiSeq Reporter v2.4.60 and the corresponding default version of hg19). MiSeq Reporter was then used to call variants. In the "Illumina Experiment Manager", "TruSeq Amplicon Workflow – specific settings" were adjusted as follows: “Export to gVCF – MaxIndelSize” from default “25” to “55”. Results were then filtered to keep only those with tumor variant allele frequency of at least 10% and a depth of coverage greater than 500x.

Vanderbilt-Ingram Cancer Center (VICC)

Foundation medicine panels: VICC uses Illumina hybridization-based capture panels from Foundation Medicine to detect single nucleotide variants, small indels, copy number alterations and structural variants from tumor-only sequencing data. Two gene panels were used: Panel 1 (T5a bait set), covering 326 genes and; and Panel 2 (T7 bait set), covering 434 genes. DNA was extracted from unstained FFPE sections, and H&E stained sections were used to ensure nucleated cellularity ≥ 80% and tumor cellularity ≥ 20%, with use of macro-dissection to enrich samples with ≤20% tumor content. A pool of 5’-biotinylated DNA 120bp oligonucleotides were designed as baits with 60bp overlap in targeted exon regions and 20bp overlap in targeted introns with a minimum of 3 baits per target and 1 bait per SNP target. The goal was a depth of sequencing between 750x and 1000x. Mapping to the reference genome was accomplished using BWA, local alignment optimizations with GATK, and PCR duplicate read removal and sequence metric collection with Picard and Samtools. A Bayesian methodology incorporating tissue-specific prior expectations allowed for detection of novel somatic mutations at low MAF and increased sensitivity at hotspots. Final single nucleotide variant (SNV) calls were made at MAF≥ 5% (MAF≥ 1% at hotspots) with filtering for strand bias, read location bias and presence of two or more normal controls. Indels were detected using the deBrujn approach of de novo local assembly within each targeted exon and through direct read alignment and then filtered as described for SNVs. Copy number alterations were detected utilizing a comparative genomic hybridization-like method to obtain a log-ratio profile of the sample to estimate tumor purity and copy number. Absolute copy number was assigned to segments based on Gibbs sampling. To detect gene fusions, chimeric read pairs were clustered by genomic coordinates and clusters containing at least 10 chimeric pairs were identified as rearrangement candidates. Rare tumors and metastatic samples were prioritized for sequencing, but ultimately sequencing was at the clinician's discretion.

VICC also submitted data from 2 smaller hotspot amplicon panels, one used for all myeloid (VICC-01-myeloid) tumors and 1 used for some solid tumors (VICC-01-solidtumor). These panels detect point mutations and small indels from 37 and 31 genes, respectively. Solid tumor H&E were inspected to ensure adequate tumor cellularity (>10%). Sections were macrodissected if necessary, and DNA was extracted. Tumors were sequenced to an average depth greater than 1000X. Reads were aligned to hg19/GRCh37 with novoalign, and single nucleotide variants, insertions and deletions greater than 5% were called utilizing a customized bioinformatic pipeline. Large (15bp and greater) FLT3 insertions were called using a specialized protocol and were detected to a 0.5% allelic burden.

Wake Forest University Health Sciences (Wake Forest Baptist Medical Center) (WAKE)
We utilized the sequencing analysis pipelines from Foundation Medicine and Caris to analyze clinical samples and support. Enrichment of target sequences was achieved by solution-based hybrid capture with custom biotinylated oligonucleotide bases. Enriched libraries were sequenced to an average median depth of >500× with 99% of bases covered >100× (IlluminaHiSeq 2000 platform using 49 × 49 paired-end reads). The clinical sequencing data were analyzed by Foundation Medicine and Caris developed pipelines. Sequenced reads were mapped to the reference human genome (hg19) using the Burrows-Wheeler Aligner and the publicly available SAM tools, Picard, and Genome Analysis Toolkit. Point mutations were identified by a Bayesian algorithm; short insertions and deletions determined by local assembly; gene copy number alterations identified by comparison to process-matched normal controls; and gene fusions/rearrangements determined by clustering chimeric reads mapped to targeted introns. Following by computational analysis with tools such as MutSig and CHASM , the driver mutations can be identified which may help the selection of treatment strategy. In addition, the initial report of the analysis of 470 cases has been published and highlighted on the cover of the journal Theranostics in 2017.

Pipeline for Annotating Mutations and Filtering Putative Germline SNPs

Contributing GENIE centers provided mutation data in Variant Call Format (VCF vcf2maf v1.6.14, samtools.github.io/hts-specs) or Mutation Annotation Format (MAF v2.x, https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

) with additional fields for read counts supporting variant alleles, reference alleles, and total depth. Some “MAF-like” text files with minimal required columns (https://github.com/mskcc/vcf2maf/blob/v1.6.14/data/minimalist_test_maf.tsv) were also received from the participating centers. These various input formats were converted into a complete tab-separated MAF v2.4 format, with a standardized set of additional columns (https://github.com/mskcc/vcf2maf/blob/v1.6.14/docs/vep_maf_readme.txt) using either vcf2maf or maf2maf v1.6.14 (https://github.com/mskcc/vcf2maf/tree/v1.6.14) wrappers around the Variant Effect Predictor (VEP v86, https://github.com/mskcc/vcf2maf/blob/v1.6.14/docs/vep_maf_readme.txt). The vcf2maf “custom-enst” option overrode VEP’s canonical isoform for most genes, with Uniprot’s canonical isoform (https://github.com/mskcc/vcf2maf/blob/v1.6.14/data/isoform_overrides_uniprot).

While the GENIE data available from Sage contains all mutation data, the following mutation types are automatically filtered upon import into the cBioPortal (http://www.cbioportal.org/genie): Silent, Intronic, 3’ UTR, 3’ Flank, 5’ UTR, 5’ Flank and Intergenic region (IGR).

Six of the eight GENIE participating centers performed tumor-only sequencing i.e. without also sequencing a patient-matched control sample like blood, to isolate somatic events. These centers minimized artifacts and germline events using pooled controls from unrelated individuals, or using databases of known artifacts, common germline variants, and recurrent somatic mutations. However, there remains a risk that such centers may inadvertently release germline variants that can theoretically be used for patient re-identification. To minimize this risk, the GENIE consortium developed a stringent germline filtering pipeline, and applied it uniformly to all variants across all centers. This pipeline flags sufficiently recurrent artifacts and germline events reported by the Exome Aggregation Consortium (ExAC, http://exac.broadinstitute.org). Specifically, the non-TCGA subset VCF of ExAC 0.3.1 was used after excluding known somatic events in https://github.com/mskcc/vcf2maf/blob/v1.6.14/data/known_somatic_sites.bed, based on:

- Hotspots from Chang et al. minus some likely artifacts (dx.doi.org/10.1038/nbt.3391).
- Somatic mutations associated with clonal hematopoietic expansion from Xie et al. (dx.doi.org/10.1038/nm.3733).
- Somatically mutable germline sites at MSH6:F1088, TP53:R290, TERT:E280, ASXL1:G645_G646.

The resulting VCF was used with vcf2maf’s “filter-vcf” option, to match each variant position and allele to per-subpopulation allele counts. If a variant was seen more than 10 times in any of the 7 ExAC subpopulations, it was tagged as a “common_variant” (vcf2maf’s “max-filter-ac” option), and subsequently removed. This >10 allele count (AC) cutoff was selected because it tagged no more than 1% of the somatic calls across all MSK-IMPACT samples with patient-matched controls.

Description of Data Files

The following is a summary of all data files available in the release.

Table 4: GENIE Data Files.

File Name	Description	Details
data_mutations_extended.txt	Mutation data. Tab-delimited Mutation Annotation Format (MAF).	For a description of the MAF file format, see: https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/
data_CNA.txt	Discretized copy number data. Tab-delimited: rows represent genes, columns represent individual samples. **Note**: not all centers contributed copy number data to GENIE.	**-2**: deep loss, possibly a homozygous deletion **-1**: single-copy loss (heterozygous deletion) **0**: diploid **1**: low-level gain **2**: high-level amplification.
data_fusions.txt	Structural variant data. Tab-delimited: rows represent individual structural variants identified in samples, columns represent variant details. **Note**: not all centers contributed structural rearrangement data to GENIE.	**HUGO_SYMBOL**: HUGO gene symbol. **CENTER**: GENIE center. **TUMOR_SAMPLE_BARCODE**: GENIE Sample ID. **FUSION**: A description of the fusion, e.g., "TMPRSS2-ERG fusion". **DNA_SUPPORT**: Fusion detected from DNA sequence data, "yes" or "no". **RNA_SUPPORT**: Fusion detected from RNA sequence data, "yes" or "no". **FRAME**: "in-frame" or "frameshift".
genie_combined.bed	Combined BED file describing genomic coordinates covered by	For a description of the BED file format, see:

	all platforms contributed to GENIE.	https://genome.ucsc.edu/FAQ/FAQformat#format1
genie_data_cna_hg19.seg	Segmented copy number data. Tab-delimited: rows represent copy number events within samples, columns represent genomic coordinates and continuous copy number values. **Note**: not all centers contributed segmented copy number data to GENIE.	
data_clinical.txt	De-identified tier 1 clinical data. Tab-delimited: rows represent samples, columns represent de-identified clinical attributes.	See Clinical Data section below for more details.

Clinical Data

A limited set of Tier 1 clinical data have been submitted by each center to provide clinical context to the genomic results (Table 5). Additional clinical data elements, including staging, treatments, and outcomes will be added in the future. When possible the clinical data are collected at the institutions in a fashion that can be mapped to established oncology data specifications, such as the North American Association of Central Cancer Registrars (NAACCR).

Table 5: GENIE Tier 1 Clinical Data Fields.

Data Element	Example Values	Data Description
AGE_AT_SEQ_REPORT	Integer values, <18 or >89.	The age of the patient at the time that the sequencing results were reported. Age is masked for patients aged 90 years and greater and for patients under 18 years.
CENTER	CRUK DFCI GRCC JHU MDA MSK NKI UCSF	The center submitting the clinical and genomic data.

	UHN VICC WAKE	
ETHNICITY	Non-Spanish/non-Hispanic Spanish/Hispanic Unknown	Indication of Spanish/Hispanic origin of the patient; this data element maps to the NAACCR v16, Element #190. Institutions not collecting Spanish/Hispanic origin have set this column to Unknown.
ONCOTREE_CODE	LUAD	The primary cancer diagnosis code based on the OncoTree ontology (http://cbioportal.org/oncotree). The version of Oncotree ontology that was used for GENIE 5.0-public is 2017_06_21.
PATIENT_ID	GENIE-JHU-1234	The unique, anonymized patient identifier for the GENIE project. Conforms to the following the convention: GENIE-CENTER-1234. The first component is the string, "GENIE"; the second component is the Center abbreviation. The third component is an anonymized unique identifier for the patient.
PRIMARY_RACE	Asian Black Native American Other Unknown White	The primary race recorded for the patient; this data element maps to the NAACCR v16, Element #160. For institutions collecting more than one race category, this race code is the primary race for the patient. Institutions not collecting race have set this field to Unknown..
SAMPLE_ID	GENIE-JHU-1234-9876	The unique, anonymized sample identifier for the GENIE project. Conforms to the following the convention: GENIE-CENTER-1234-9876. The first component is the string, "GENIE"; the second component is the Center abbreviation. The third component is an anonymized, unique patient identifier. The fourth component is a unique identifier for the sample that will distinguish between two or more specimens from a single patient.
SAMPLE_TYPE	Primary Metastasis Unspecified	Sample type, e.g. Primary or Metastasis.

SEQ_ASSAY_ID	DFCI-ONCOPANEL-1 DFCI-ONCOPANEL-2 MSK-IMPACT341 MSK-IMPACT410	The institutional assay identifier for genomic testing platform. Components are separated by hyphens, with the first component corresponding to the Center's abbreviation. All specimens tested by the same platform should have the same identifier.
SEX	Female Male	The patient's sex code; this data element maps to the NAACCR v16, Element #220.
CANCER_TYPE	Non-Small Cell Lung Cancer	The primary cancer diagnosis "main type", based on the OncoTree ontology (http://cbioportal.org/oncotree). For example, the OncoTree code of LUAD maps to: "Non-Small Cell Lung Cancer". The version of Oncotree ontology that was used for GENIE 5.0-public is 2017_06_21.
CANCER_TYPE_DETAILED	Lung Adenocarcinoma	The primary cancer diagnosis label, based on the OncoTree ontology (http://cbioportal.org/oncotree). For example, the OncoTree code of LUAD maps to the label: "Lung Adenocarcinoma (LUAD)". The version of Oncotree ontology that was used for GENIE 5.0-public is 2017_06_21.

Cancer types are reported using the OncoTree ontology (http://oncotree.mskcc.org/oncotree/), originally developed at Memorial Sloan Kettering Cancer Center. Version 5.0-public of GENIE uses the OncoTree specification from June 21, 2017, containing diagnosis codes for 524 tumor types from 32 tissues. The centers participating in GENIE applied the OncoTree cancer types to the tested specimens in a variety of methods depending on center-specific workflows. A brief description of how the cancer type assignment process for each center is specified in Table 6.

Table 6: Center Strategies for OncoTree Assignment.

Center	OncoTree Cancer Type Assignments
CRUK	Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.
DFCI	Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.
GRCC	OncoTree cancer types were mapped from ICD-O codes.
JHU	Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.

MDA	OncoTree cancer types were mapped from ICD-O codes.
MSK	Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.
NKI	Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.
UCSF	Original diagnosis from pathologist was mapped to OncoTree cancer type by molecular pathologists from Clinical Cancer Genomics Laboratory.
UHN	Original diagnosis from pathologist was mapped to OncoTree diagnosis by medical oncologist and research manager.
VICC	OncoTree cancer types were mapped from ICD-O codes. If no ICD-O code was available, research manager mapped pathologist and/or medical oncologist diagnosis to OncoTree cancer type.
WAKE	We have mapped Foundation Medicine and Caris Diagnosis to ICD-O-3 with a process then utilized the ICD-O-3/Oncotree mapping.

Abbreviations and Acronym Glossary

Abbreviation	Full Term
AACR	American Association for Cancer Research
CNA	Copy number alterations
CNV	Copy number variants
CRUK	Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, England
DFCI	Dana-Farber Cancer Institute
FFPE	Formalin-fixed, paraffin-embedded
GENIE	Genomics, Evidence, Neoplasia, Information, Exchange
GRCC	Institut Gustave Roussy
HIPAA	Health Insurance Portability and Accountability Act
IRB	Institutional Review Board
JHU	Johns Hopkins Sidney Kimmel Comprehensive Cancer Center

MAF	Mutation annotation format
MDA	The University of Texas MD Anderson Cancer Center
MSK	Memorial Sloan Kettering Cancer Center
NAACCR	North American Association of Central Cancer Registries
NGS	Next-generation sequencing
NKI	Netherlands Cancer Institute
PCR	Polymerase chain reaction
PHI	Protected Health Information
SNP	Single-nucleotide polymorphism
SNV	Single-nucleotide variants
UCSF	University of California-San Francisco (UCSF Helen Diller Family Comprehensive Cancer Center), San Francisco, California
UHN	Princess Margaret Cancer Centre, University Health Network
VCF	Variant Call Format
VICC	Vanderbilt-Ingram Cancer Center
WAKE	Wake Forest University Health Sciences (Wake Forest Baptist Medical Center), Winston-Salem, North Carolina