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 2.0.0

GENIE data versions follow a numbering scheme derived from semantic versioning, where the digits in the version correspond to: major.minor.patch. “Major” releases are public releases of new sample data. “Minor” releases are internal releases, available only within the GENIE consortium. “Patch” releases are corrections to major or minor releases, including data retractions.

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
DFCI	Dana-Farber Cancer Institute, USA
GRCC	Institut Gustave Roussy, France
JHU	Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, USA
MDA	MD Anderson Cancer Center, USA
MSK	Memorial Sloan Kettering Cancer Center, USA
NKI	Netherlands Cancer Institute, on behalf of the Center for Personalized Cancer Treatment, The Netherlands
UHN	Princess Margaret Cancer Centre, University Health Network, Canada
VICC	Vanderbilt-Ingram Cancer Center, 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 eight centers.

Table 2: Genomic Data Characterization by Center.

		DFCI	GRCC	JHU	MDA	MSK	NKI	UHN-solid	UHN-myeloid	VICC	VICC-solid/myeloid
Specimen Types	Formalin-fixed, paraffin-embedded (FFPE) v. Fresh Frozen (Fresh Froz)	FFPE	Fresh Froz	FFPE	FFPE	FFPE	FFPE	FFPE	FFPE	FFPE	FFPE
Specimen Tumor Cellularity	Tumor Cellularity Cutoff	>20%	>10%	>10%	>20%	>10%	>10%	>10%	>10%	>20%	>10%
Assay Type	Hybridization Capture v. PCR	Capture	PCR	PCR	PCR	Capture	PCR	PCR	PCR	Capture	PCR
Coverage	Hotspot Regions		x	x	x		x	X	X		X
	Coding Exons	x				x		X	X	x	
	Introns (selected)	x				x				x	
	Promoters (selected)					x					
Platform	Illumina	x				x	x	X	X	x	
	Ion Torrent		x	x	x			X	X		
Calling Strategy	Unmatched (Tumor-only) v. Matched (Tumor-Normal)	Tumor Only	Tumor Only	Tumor Only	Tumor Only	Tumor-Normal	Tumor Only	Tumor-Normal	Tumor only	Tumor Only	Tumor Only
Alteration	Single Nucleotide	x	x	x	x	x	x	X	X	x	X

Types	Variants (SNV)										
	Small Indels	x	x	x	x	x	x	X	X	x	X
	Gene-Level CNA	x			[1]	x				x	
	Intragenic CNA					x					
	Structural Variants	[1]				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
DFCI-ONCOPANEL-1.txt	Custom	All Exons	275
DFCI-ONCOPANEL-2.txt	Custom	All Exons	300
MSK-IMPACT341.txt	Custom	All Exons	341
MSK-IMPACT410.txt	Custom	All Exons	410
GRCC-CHP2.txt	Ion AmpliSeq Cancer Hotspot Panel v2	Hotspot Regions	50
GRCC-MOSC3.txt	Ion AmpliSeq Cancer Hotspot Panel v2	Hotspot Regions	74
JHU-50GP-V1.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
NKI-TSACP.txt	TruSeq Amplicon Cancer Panel	Hotspot Regions	48
UHN-48-V1.txt	TruSeq Amplicon Cancer Panel	Hotspot Regions	48

UHN-50-V2.txt	PCR - Ion AmpliSeq Cancer Panel	Hotspot Regions	50
UHN-54-V1.txt	PCR - 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	34
VICC-01-myeloid	Custom	Hotspot Regions	37

Genomic Profiling at Each Center

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. Two versions of the panel have been submitted to GENIE: version 1 containing 275 genes, and version 2 containing 300 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.

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. Two versions of the panel have been used: CHP2 covering hotspots in 50 genes, and MOSC3 covering hotspots in 74 genes. Tumors are sequenced to an average unique depth of coverage of >500X. The sequencing data were analyzed with the Torrent Suite Variant Caller 4.2 software and reported somatic variants were compared with the reference genome 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-01 trial (MOlecular Screening for CAncer Treatment Optimisation). This trial collected on-purpose tumor samples (from the primary or from a metastatic site) that are immediately fresh-frozen, and subsequently analyzed for targeted gene panel sequencing. Tumor 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%.

University of Texas MD Anderson Cancer Center (MDA)

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. Two different amplicon pools and pipeline versions are included: 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. DNA was extracted from unstained sections of tissue paired with an 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 using 318 chip, 260 flows. 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.

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) 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 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. Two versions of the panel have been submitted to GENIE: version 1 containing 341 genes, and version 2 containing 410 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 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 a single panel (JHU_50GP_V1) 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.

Princess Margaret Cancer Centre, University Health Network (UHN)
Princess Margaret Cancer Centre used three panels to sequence samples for the GENIE 2.0.0 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 four workflows:

1. Data analysis of tumor-normal pairs processed by UHN_TSACP_workflow: Reads are aligned to either hg19 using BWA mem version 0.7.12 or by MiSeq fastq 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 are called using MuTect (v1.1.5) for SNVs and Varscan (v2.3.8) using both normal and tumor data. Default settings for this version of VarScan result in a minimum variant frequency of 20%. Data are filtered to ensure there are no variants included with frequency of 3% or more in the normal sample. SNV results are filtered to keep only those with tumor variant allele frequency of at least 10%. Testing is performed for patients with advanced disease from select tumor types.

2. 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) using both normal and tumor data. Default settings for this version of VarScan result in a minimum variant frequency of 20%. Data were filtered to ensure there are no variants included with frequency of 3% or more in the normal sample. SNV results were filtered to keep only those with tumor variant allele frequency of at least 10%.

3. Data analysis of tumor only processed by UHN_TSACP_tumorONLY_workflow: Reads were aligned to hg19 using BWA mem version 0.7.12 followed by local realignment and BQSR using GATK v3.3.0. Sequence mutations (SNV and indel) are called using Varscan (v2.3.8). SNV results were filtered to keep only those with tumor variant allele frequency of at least 10%.

4. 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). SNV 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 indel variants filtered to keep all with VAF of at least 10% and total coverage of at least 100x.

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%.

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 60 bp 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 34 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.

Pipeline for Annotating Mutations and Filtering Putative Germline SNPs

Contributing GENIE centers provided mutation data in Variant Call Format (VCF v4.x, samtools.github.io/hts-specs) or Mutation Annotation Format (MAF v2.x, wiki.nci.nih.gov/x/eJaPAQ) with additional fields for read counts supporting variant alleles, reference alleles, and total depth. Some "MAF-like" text files with minimal required columns (github.com/mskcc/vcf2maf/blob/v1.6.12/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 (github.com/mskcc/vcf2maf/blob/v1.6.12/docs/vep_maf_readme.txt) using either vcf2maf or maf2maf v1.6.12 (github.com/mskcc/vcf2maf/tree/v1.6.12), wrappers around the Variant Effect Predictor (VEP v86, gist.github.com/ckandoth/f265ea7c59a880e28b1e533a6e935697). The vcf2maf "custom-enst" option overrode VEP's canonical isoform for most genes, with Uniprot's canonical isoform (github.com/mskcc/vcf2maf/blob/v1.6.12/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/v1.6.12/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://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification
data_CNA.txt	Discretized copy number data. Tab-delimited: rows represent genes, columns represent individual samples.	**-2**: deep loss, possibly a homozygous deletion **-1**: single-copy loss (heterozygous deletion) **0**: diploid **1**: low-level gain

	Note: not all centers contributed copy number data to GENIE.	**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 all platforms contributed to GENIE.	For a description of the BED file format, see: 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	DFCI GRCC JHU MSK NKI UHN MDA VICC	The center submitting the clinical and genomic data.
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).
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".

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)"

Cancer types are reported using the OncoTree ontology (http://oncotree.mskcc.org/oncotree/), originally developed at Memorial Sloan Kettering. Version 2.0 of GENIE uses the OncoTree specification from August 18, 2016, 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
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.
MSK	Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.
NKI	Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.
UHN	Original diagnosis from pathologist was mapped to OncoTree diagnosis by medical oncologist and research manager.
MDA	OncoTree cancer types were mapped from ICD-O codes.
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.

Abbreviations and Acronym Glossary

Abbreviation	Full Term

AACR	American Association for Cancer Research
CNA	Copy number alterations
CNV	Copy number variants
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	M.D. 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
UHN	Princess Margaret Cancer Centre, University Health Network
VICC	Vanderbilt-Ingram Cancer Center