AACR Project GENIE 20.0-public Data Guide

AACR

2026-07-07

Table of contents

About this Document 3

Version of Data 3

Data Access 3

Terms Of Access 3

Introduction to AACR GENIE 4

Quick Start Guide for GENIE Data Users 5
- Downloading the Data . . . 5
- Need Help? . . . 6

Human Subjects Protection and Privacy 6

Data Hamonization & QC Process **6**
Sample Filters . . . 7
Overview of Filtering Process . . . 7
Quality and Completeness Filters . . . 8
Privacy Protection Filters . . . 8
Temporal and Consent Filters . . . 9
Variant Data Filters . . . 10
Processing Transformations . . . 11
Post-Release Quality Checks . . . 12

Summary of Sequence Pipeline **13**

Genomic Profiling at Each Center **32**
Children's Hospital of Philadelphia (CHOP) . . . 32
Herbert Irving Comprehensive Cancer Center, Columbia University (COLU) . . . 33
Cancer Research UK Cambridge Centre, University of Cambridge (CRUK) . . . 33
Dana-Farber Cancer Institute (DFCI) . . . 36
Duke Cancer Institute (DUKE) . . . 36
Institut Gustave Roussy (GRCC) . . . 37
Johns Hopkins Sidney Kimmel Comprehensive Cancer Center (JHU) . . . 38
The University of Texas MD Anderson Cancer Center (MDA) . . . 38
Memorial Sloan Kettering Cancer Center (MSK) . . . 39
Netherlands Cancer Center, The Netherlands (NKI) . . . 39
Providence Health & Services Cancer Institute (PROV) . . . 40
Swedish Cancer Institute (SCI) . . . 41
The University of Chicago (UCHI) . . . 41
University of California-San Francisco (UCSF Helen Diller Family Comprehensive Cancer Center) (UCSF) . . . 42
Princess Margaret Cancer Centre, University Health Network (UHN) . . . 43
University of Miami - Sylvester Comprehensive Cancer Center (UMIAMI) . . . 45
Vall d'Hebron Institute of Oncology (VHIO) . . . 47
Vanderbilt-Ingram Cancer Center (VICC) . . . 49
Wake Forest University Health Sciences, Wake Forest Baptist Medical Center (WAKE) . . . 50
Yale University, Yale Cancer Center (YALE) . . . 50

Description of Data Files **51**

Description of Clinical Data Fields **53**
Linking clinical data to genomic data . . . 57
Center Strategies for OncoTree Assignment . . . 58

Abbreviations and Acronym Glossary **59**

About this Document

```
This document provides an overview of 20.0-public release
of American Association for Cancer Research (AACR) GENIE data.
```

Version of Data

```
AACR Project GENIE Data: Version 20.0-public
```

AACR Project GENIE follows a semantic versioning scheme where version numbers are formatted as major.patch-release_type. Major releases contain new sample data, while patch releases address corrections and data retractions (patches for public releases are deprecated as of the 17 series). The release type indicates whether data is publicly available or restricted to consortium members: public releases use "X.X-public" nomenclature, and consortium-only releases use "X.X-consortium".

Data Access

AACR Project GENIE Data is currently available via two mechanisms:

- Synapse Platform (Sage Bionetworks): https://genie.synapse.org/
- cBioPortal for Cancer Genomics (MSK): https://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: genieinfo@aacr.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: Participating Centers

Center Abbreviation	Center
NKI	Netherlands Cancer Institute, on behalf of the Center for Personalized Cancer Treatment, Amsterdam, Netherlands
DFCI	Dana-Farber Cancer Institute, Boston, MA, USA
GRCC	Institut Gustave Roussy, Paris, France
JHU	Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD, USA
MSK	Memorial Sloan Kettering Cancer Center, New York, NY, USA

Continued on next page

Table 1: Participating Centers

Center Abbreviation	Center
UHN	Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
MDA	The University of Texas MD Anderson Cancer Center, Houston, TX, USA
VICC	Vanderbilt-Ingram Cancer Center, Nashville, TN, USA
CRUK	Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, England
CHOP	Children's Hospital of Philadelphia, Philadelphia, PA, USA
DUKE	Duke Cancer Institute, Duke University Health System, Durham, NC, USA
COLU	The Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA
PROV	Providence Health & Services Cancer Institute, Portland, OR, USA
SCI	Swedish Cancer Institute, Seattle, WA, USA
UCSF	University of California, San Francisco, CA, USA
VHIO	Vall d' Hebron Institute of Oncology, Barcelona, Spain
WAKE	Wake Forest Baptist Medical Center, Wake Forest University Health Sciences, Winston-Salem, NC, USA
YALE	Yale Cancer Center, Yale University, New Haven, Connecticut, USA
UCHI	University of Chicago Comprehensive Cancer Center, Chicago, IL, USA
UMIAMI	University of Miami Sylvester Comprehensive Cancer Center, Miami, FL, USA

Quick Start Guide for GENIE Data Users

This section provides essential information to begin working with GENIE data.

Downloading the Data

From Synapse:

1. Create a Synapse account at synapse.org
2. Navigate to the GENIE project: https://genie.synapse.org/
3. Accept Terms of Use
4. Download files programmatically using Python or R or via web interface: https://docs.synapse.org/synapse-docs/downloading-data-programmatically

Key Files for Most Analyses:

- `data_clinical_patient.txt` - Patient demographics and survival
- `data_clinical_sample.txt` - Sample information and cancer types
- `data_mutations_extended.txt` - Somatic mutations (large file!)
- `genomic_information.txt` - Panel coverage information
- `assay_information.txt` - Sequencing panel specifications

Need Help?

- **Data Guide**: This document
- **cBioPortal Help**: https://www.cbioportal.org/faq
- **Synapse Help**: https://help.synapse.org/
- **GENIE Email**: genieinfo@aacr.org

Human Subjects Protection and Privacy

Protection of patient privacy is paramount, and the AACR Project GENIE 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/.

Data Hamonization & QC Process

The GENIE dataset is generated from data curated by multiple cancer centers. Each center prepares their data by uploading their files onto the Synapse platform.

Required files:

- Clinical patient and sample data
- Variant data (MAF or VCF file)
- Assay information
- Genomic regions data (BED file)

Optional files:

- Structural variant data
- Segmented data
- Discrete copy number data
- Mutations in cis data
- Sample and/or patient retraction file

Only validated files that adhere to the submission guidelines will be pushed to the release. The files that do not pass these validation checks will not be processed and an email notifying any errors or warnings will be sent to each center to correct. Each accepted file type has its own submission guidelines the centers must follow. Some examples of these validation checks are given below:

- assay information file: `SEQ_ASSAY_ID`s must start with the center abbreviation.
- BED file: `Start_Position` must only be integers and is the 2nd column.
- MAF file: `T_ALT_COUNT` must be an integer.

Every month the center files are merged and processed into the release files. During processing, specific sample and variant filters and transformations are applied to the each center's validated files. After processing, the release files go through some manual and automated review processes.

For more information on the filters and review processes, see below.

Sample Filters

The GENIE dataset undergoes several filtering steps to ensure data quality, protect patient privacy, and maintain consistency across releases. This section explains which samples are included or excluded, and why.

Overview of Filtering Process

Filters are applied during data processing:

1. **Quality Filters**: Ensure technical data completeness (e.g. BED files are present, valid oncotree codes)
2. **Privacy Filters**: Protect patient identities (e.g. age redaction)
3. **Temporal Filters**: Define release scope (e.g. sequencing date cutoffs)
4. **Retraction Filters**: Honor site retractions for patient data

Below is a detailed explanation of each filter and its impact on the dataset.

Quality and Completeness Filters

Sequencing Panel Documentation (BED File Requirement)

What it means: Samples are only included if their associated sequencing panel has a corresponding BED file that specifies the genomic regions that were sequenced.

Why it matters: Without a BED file, we cannot determine which variants were within the sequenced regions vs. which regions were simply not covered. This is critical for interpreting the absence of mutations.

Impact: Samples that are associated with a panel lacking a submitted BED files are excluded from the release.

For data users: All samples in GENIE have documented coverage. When you don't see a mutation in a gene, it could mean either (1) no mutation was present, or (2) that region wasn't sequenced. Check the genomic_information.txt file to determine coverage.

OncoTree Code Validation

What it means: Cancer type classifications (ONCOTREE_CODE) must use current, non-deprecated codes from the OncoTree ontology.

Why it matters: OncoTree is regularly updated as cancer classification evolves. Deprecated codes may represent outdated cancer classifications or codes that have been merged into other categories.

What happens: Samples with deprecated codes are removed from the release. Centers are notified to update their cancer type assignments.

For data users: You can trust that all ONCOTREE_CODE values are valid for this release's OncoTree version oncotree_2021_11_02.

Privacy Protection Filters

Age Information Redaction (HIPAA Safe Harbor)

What it means: To comply with HIPAA Safe Harbor de-identification rules, ages and dates are masked for patients over 89 years old. Pediatric patients under 18 years old may also be masked at individual sites' discretion. Note: Universal masking of patients under 18 was deprecated as of the 19.4-consortium release.

Why it matters: Extreme ages combined with rare cancer types could potentially be used to re-identify patients.

How it works:

- If age at sequencing is <18 years (~6,570 days) or >89 years (~32,485 days):
  - Age fields show: "<18" or ">89"
  - Date intervals show: "<6570" or ">32485"
  - Year fields show: "withheld" or "cannotReleaseHIPAA"
  - Affected fields: BIRTH_YEAR, YEAR_CONTACT, YEAR_DEATH, INT_CONTACT, INT_DOD, AGE_AT_SEQ_REPORT

For data users:

- When calculating survival curves, be aware that values of ">32485" and "<6570" represent censored ages
- Consider excluding these samples from age-stratified analyses or treating them as separate categories
- When the difference between BIRTH_YEAR and YEAR_CONTACT or BIRTH_YEAR and YEAR_DEATH is greater than 89 years, those values will be redacted

Temporal and Consent Filters

Sequencing Date Cutoff

What it means: Each release includes samples sequenced up to a specific date.

Why it matters: Ensures consistent release schedules and allows centers time for quality control.

Example: Release 5.0 includes samples sequenced before January 2018 (not including January 2018).

For data users: Check the release version to understand the temporal scope. Newer releases include more recent samples.

Patient and Sample Retractions

What it means: Samples or patients may be removed from releases in two ways:

1. **Implicit Retraction**: When a center stops uploading data for a patient/sample, it automatically disappears from future releases
   - Common reasons: Data no longer meets quality standards, discovered technical issues
2. **Explicit Retraction**: Centers submit a retraction file to formally remove specific patients/samples.
   - Common reasons: Patient withdrew consent, data use agreement expired, re-identification risk

Note: After series 19 sites should submit a reason for the retraction if known.

For data users:

- Sample IDs may disappear between releases
- Always note which release version you're using in publications
- If your analysis depends on specific samples, verify they're still present in the latest release

Variant Data Filters

See the panel-level coverage section in the Quick Start Guide for information on how to account for different sequencing panels when analyzing mutation data. Also see Summary of Sequence Pipeline section for more information on how panels are organized in GENIE.

Germline Filter

- All germline variants are filtered out in all releases.
  - We added a whitelist of GRCh37 loci where pathogenic somatic events are known to occur. The whitelist can be viewed here.
- Genome Nexus reports gnomAD AFs by querying the Genome Aggregation Database to determine common variants. An additional filter named common_variant is also appended if allele count across at least one ExAC subpopulation is >10 (this default cutoff can be changed when running vcf2maf). If any variant has a max gnomAD AF from any subpopulation that is over 0.0005, the variant will be filtered out. 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.
  - So if you're handling somatic variants, the common_variant tag means this is likely a false-positive. It is less likely to be a legit somatic variant at a site that ExAC classifies as germline or artifact.
- While the GENIE data available from Sage contains all mutation data, the following mutation types are automatically altered upon import into the cBioPortal: Silent, Intronic, 3' UTR, 3' Flank, 5' UTR, 5' Flank and Intergenic region (IGR).
- Most of the 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-identication. To minimize this risk, the GENIE consortium developed a stringent germline filtering pipeline, and applied it uniformly to all variants across all centers.

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

MAF in BED

Any variants in the mutation file that isn't described by the bed file will be filtered out

Mutations-in-cis

Samples that have variants that could be merged together will be filtered out. This filter looks for close proximity variants (within 6bp) and difference of variant allele frequency between two variants (<0.05). Sites review the variants that are filtered out and select one of the three options for this filter:

- ON: Sage filters out all samples with variants flagged
- OFF: If a center already does this filter prior to upload, the filter can be turned off
- FLAG: The samples are not removed, but the variants are annotated in the maf file.

Genome Nexus Annotation Status

Mutation data that fails to be annotated by Genome-Nexus is excluded from the release. For example, this occurs when the allele extracted from the VCF or MAF file does not match the reference allele, or when the VEP tool that Genome-Nexus uses is unable to annotate the data.

Processing Transformations

Genome Nexus Annotation Pipeline: https://genie.genomenexus.org/

- Contributing GENIE centers provided mutation data in Variant Call Format (VCF) or Mutation Annotation Format (GDC MAF v1.0.0) with additionalfields for read counts supporting variant alleles, reference alleles, and total depth. Some "MAF-like" text files with minimal required columns were also received from the participating centers. These various input formats were converted into a complete tab-separated MAF format, with Genome Nexus.
- The GENIE dataset is annotating all variants with Genome Nexus starting from the 9.1-consortium release (instead of vcf2maf).

Gene symbol harmonization

All submitted HUGO gene symbols will be harmonized against GRCh37 by checking the coordinates against the annotation.

- Every single row of the bed file will be matched against the gene database. If the submitted symbol matches a row in the gene database and there is an overlap from the submitted bed region, the submitted symbol will be returned.
- If the submitted symbol does not exist in the database or there is not any overlap, an attempt to find any gene that completely encapsulates the submitted bed region is made. If the bed region is contained completely inside one gene, then it is labelled as that gene. If a bed region is enclosed completely in more than one gene, and the submitted symbol doesn't match any of the genes returned, NULL will be returned.
- If the submitted region isn't enclosed in any gene, then calculations are made to check if the bed region resides at least 90% in a gene. If there is more than one gene returned while doing 90% boundary calculations, NULL is returned.

CNA Value Harmonization

On a per site basis, if there are more two rows that are the same gene, the values are updated using the following logic:

- If there is one value, keep that value
- If there are two values (e.g. 2 or 0), we take the non-zero value.
- If there are two values that aren't zero (e.g. 1 and 2), it is NA.
- If there are more than two values (e.g. 0 or 1 or 2), it is NA.

Clinical Tier Release Scope Filters

Some parts of the clinical data is masked from the public releases. This is based on a clinical tier release scope of release document determined by the consortium.

Post-Release Quality Checks

GENIE-ArtifactFinder

Finds potential genomic artifacts by filtering unique variants for which counts in a given panel >= 10, and counts aggregated from all other panels is < 10 for which there are >= 3 different panels that cover the variant in question.

The resulting set is annotated with a fisher p-value for the variants that pass this filtering:

Number of variants called in panel	Number of samples for this panel
Number of variants called in other panels with coverage for this variant	Number of samples for panels with coverage for this variant

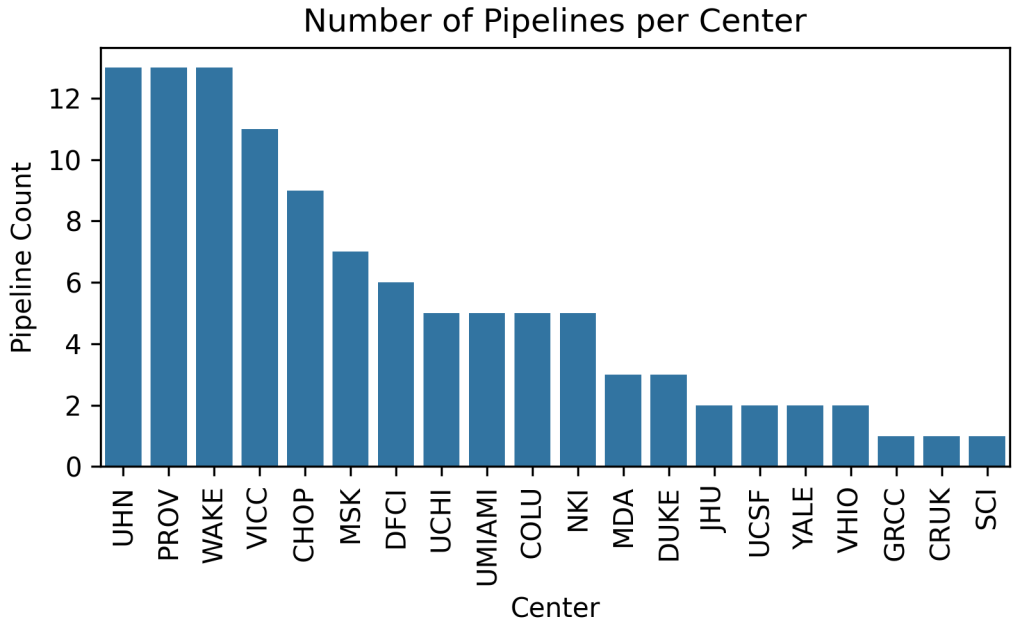
Release Report

After release, each center performs a manual review of their data based on an auto-generated dashboard that summarizes some key elements. Some examples of items sites review are given below:

- Confirm sample count.
- Confirm variant count (flagged mutations). Most of these variants are potential artifacts flagged by manual review of cBioPortal. Suggestions for variants that should be part of this list or any variant shouldn't be part of this list are welcome.
- Fix BED and assay information discrepancies.
- Confirm top 5 most frequently mutated genes per pipeline for all non-synonymous mutations.
- Remove blacklisted variants.
- Confirm clinical attribute distributions/values are valid.
- Confirm patients and samples that were marked for retraction.
- Review failed annotations flagged by Genome-Nexus.

Summary of Sequence Pipeline

Traditionally, the SEQ_ASSAY_ID was used as an institution's identifier for their assays when each assay had one associated gene panel. As GENIE grew, we wanted to support an assay having multiple gene panels. SEQ_ASSAY_ID was repurposed to be an identifier for a center's assay OR panel. For those centers that have multiple panels per assay, we introduced SEQ_PIPELINE_ID (pipeline), which encompasses multiple SEQ_ASSAY_ID (panel).



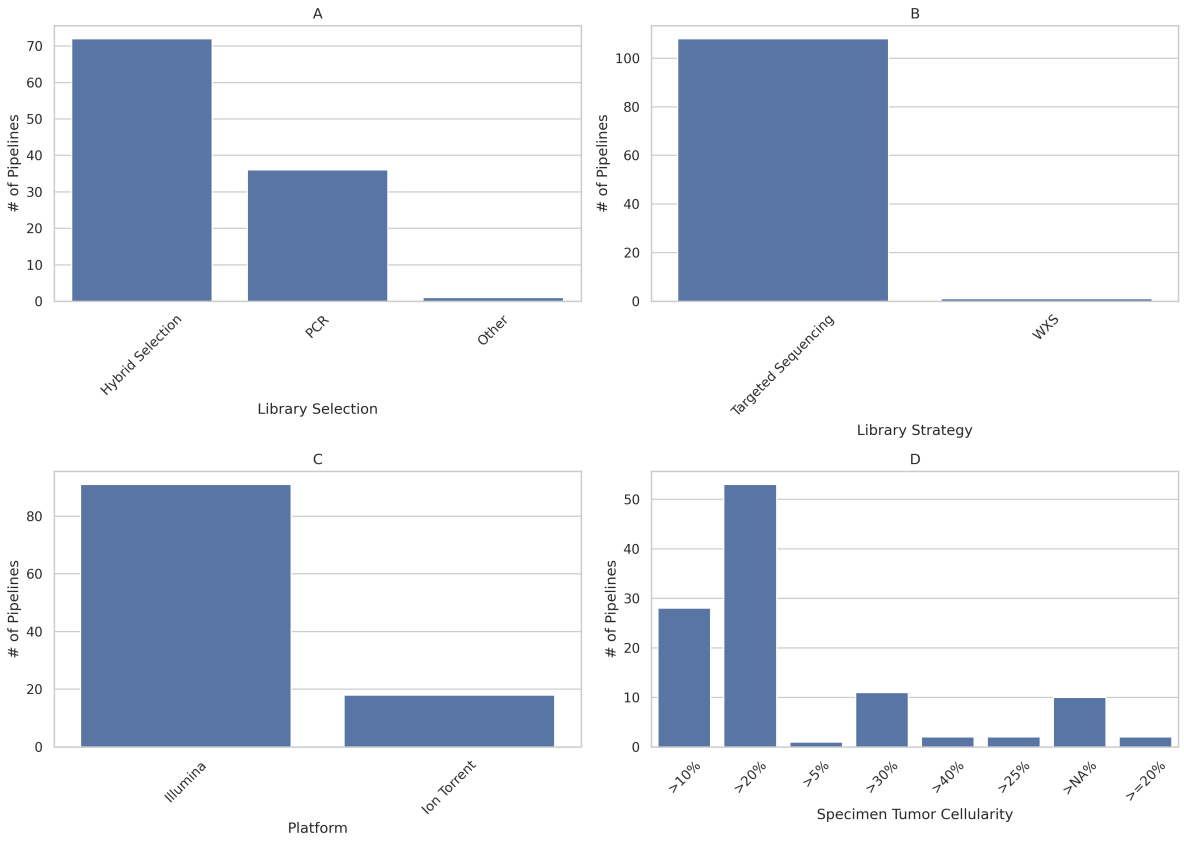


Table 3: Coverage per panel

	hotspot_regions	coding_exons	introns	promoters
CRUK-TS	X			
GRCC	X			
MDA-46-V1	X			
UCHI-ONCOHEME55-V1		X		
UCHI-ONCOSCREEN50-V1		X		
WAKE-CA		X		
WAKE-CLINICAL-R2D2		X		
WAKE-CLINICAL-T5A		X		
WAKE-CLINICAL-T7		X		
YALE-HSM-V1	X			
YALE-OCP	X	X	X	X
DFCI-ONCOPANEL-1		X	X	
DFCI-ONCOPANEL-2		X	X	

Continued on next page

Table 3: Coverage per panel

	hotspot_regions	coding_exons	introns	promoters
DFCI-ONCOPANEL-3		X	X	
NKI-TSACP-MISEQ-NGS	X	X	X	X
MSK-IMPACT341		X	X	X
MSK-IMPACT410		X	X	X
MSK-IMPACT468		X	X	X
UHN-48-V1	X	X		
UHN-50-V2	X	X		
UHN-54-V1	X	X		
UHN-555	X	X		
VICC-01-MYELOID	X			
VICC-01-SOLIDTUMOR	X			
VICC-01-T5A		X	X	
VICC-01-T7		X	X	
MDA-50-V1	X			
UHN-OCA-V3	X	X		
CHOP-STNGS	X	X	X	X
CHOP-HEMEP	X	X	X	X
COLU-CCCP-V1	X	X	X	X
COLU-TSACP-V1		X	X	
DUKE-F1-DX1		X	X	
DUKE-F1-T5A		X	X	
DUKE-F1-T7		X	X	
SCI-PMP68-V1	X			
JHU-50	X			
JHU-500STP	X			
VHIO-CUSTOM	X	X		
NKI-PATH-NGS	X	X	X	X
NKI-CHPV2-NGS	X	X	X	X
NKI-CHPV2-SOCV2-NGS	X	X	X	X
NKI-CHP-V2-PLUS	X			
MDA-409-V1	X			
UCSF-NIMV4		X	X	X
COLU-CSTP-V1		X	X	
CHOP-FUSIP	X	X	X	X
UCSF-IDTV5		X	X	X
PROV-FOCUS-V1		X		
PROV-TRISEQ-V2		X		

Continued on next page

Table 3: Coverage per panel

	hotspot_regions	coding_exons	introns	promoters
PROV-TST170-V1	X	X		
MSK-IMPACT505		X	X	X
MSK-IMPACT-HEME-400		X	X	X
PROV-TSO500HT-V2	X	X		
UHN-TSO500-V1	X	X		
WAKE-CLINICAL-CF2		X		
WAKE-CLINICAL-CF3		X		
WAKE-CLINICAL-DX1		X		
WAKE-CLINICAL-AB1		X		
WAKE-CLINICAL-AB2		X		
WAKE-CLINICAL-AB3		X		
VICC-01-D2		X	X	
VICC-01-T4B		X	X	
VICC-01-T6B		X	X	
VICC-01-DX1		X	X	
CHOP-COMPT	X	X	X	X
COLU-CSTP-V2		X	X	
MSK-IMPACT-HEME-468		X	X	X
MSK-ACCESS129		X	X	X
VICC-02-XTV2	X	X	X	X
VICC-02-XTV3	X	X	X	X
VICC-02-XTV4	X	X	X	X
COLU-CCCP-V2	X	X	X	X
WAKE-CLINICAL-R2		X		
WAKE-CLINICAL-D2		X		
VHIO-300	X	X		
CHOP-STNGS-V2	X	X	X	X
CHOP-HEMEP-V2	X	X	X	X
CHOP-COMPT-V2	X	X	X	X
WAKE-CLINICAL-DX2		X		
PROV-FOUNDATIONONELIQUIDCDX	X	X		
UHN-HM-V1	X	X		
UHN-HM-V2	X	X		
UCHI-ONCOPLUS147-V1		X		
UCHI-ONCOPLUS154-V1		X		
DFCI-F1-AB1	X	X		
DFCI-F1-CF3	X	X		

Continued on next page

Table 3: Coverage per panel

	hotspot_regions	coding_exons	introns	promoters
DFCI-F1-CF2	X	X		
UMIAMI-CA-HYBRID-WES		X		
UHN-WGS-V1	X	X		
UHN-HPV3	X	X		
UHN-HPV4	X	X		
UHN-MELV1.1	X	X		
UHN-CSTV1	X	X		
PROV-MPNC	X	X		
PROV-AMLC	X	X		
PROV-MPNT	X	X		
UCHI-ONCOPLUS119-V1		X		
UMIAMI-GD-360		X		
UMIAMI-FD-F1CDX		X		
UMIAMI-FD-F1H		X		
UMIAMI-FD-F1L		X		
CHOP-STNGS-V3	X	X	X	X
CHOP-COMPT-V3	X	X	X	X
PROV-CARIS-MIP		X		
PROV-TEMPUS-XT		X		
PROV-GUARDANT360-CDX	X	X		
PROV-GUARDANT360-V1	X	X		
PROV-GUARDANT360-V2	X	X		

Table 4: Alteration Types per Panel/Pipeline

	snv	small_indels	gene_level_cna	intragenic_cna	structura
CRUK-TS	X	X	X		
GRCC	X	X			
MDA-46-V1	X	X			
UCHI-ONCOHEME55-V1	X	X			
UCHI-ONCOSCREEN50-V1	X	X			
WAKE-CA	X	X			
WAKE-CLINICAL-R2D2	X	X	X		
WAKE-CLINICAL-T5A	X	X	X		
WAKE-CLINICAL-T7	X	X	X		
YALE-HSM-V1	X	X			

Continued on

Table 4: Alteration Types per Panel/Pipeline

	snv	small_indels	gene_level_cna	intragenic_cna	structura
YALE-OCP	X	X	X		
DFCI-ONCOPANEL-1	X	X	X		X
DFCI-ONCOPANEL-2	X	X	X		X
DFCI-ONCOPANEL-3	X	X	X		X
NKI-TSACP-MISEQ-NGS	X	X			
MSK-IMPACT341	X	X	X	X	X
MSK-IMPACT410	X	X	X	X	X
MSK-IMPACT468	X	X	X	X	X
UHN-48-V1	X	X			
UHN-50-V2	X	X			
UHN-54-V1	X	X			
UHN-555	X	X			
VICC-01-MYELOID	X	X			
VICC-01-SOLIDTUMOR	X	X			
VICC-01-T5A	X	X	X		X
VICC-01-T7	X	X	X		X
MDA-50-V1	X	X			
UHN-OCA-V3	X	X			X
CHOP-STNGS	X	X			
CHOP-HEMEP	X	X			
COLU-CCCP-V1	X	X	X	X	X
COLU-TSACP-V1	X	X	X		
DUKE-F1-DX1	X	X			X
DUKE-F1-T5A	X	X	X		X
DUKE-F1-T7	X	X			X
SCI-PMP68-V1	X	X			
JHU-50	X	X			
JHU-500STP	X	X			
VHIO-CUSTOM	X	X			
NKI-PATH-NGS	X	X			
NKI-CHPV2-NGS	X	X			
NKI-CHPV2-SOCV2-NGS	X	X			
NKI-CHP-V2-PLUS	X	X			
MDA-409-V1	X	X			
UCSF-NIMV4	X	X	X	X	X
COLU-CSTP-V1	X	X	X		
CHOP-FUSIP					X

Continued on

Table 4: Alteration Types per Panel/Pipeline

	snv	small_indels	gene_level_cna	intragenic_cna	structura
UCSF-IDTV5	X	X	X	X	X
PROV-FOCUS-V1	X	X			
PROV-TRISEQ-V2	X	X			
PROV-TST170-V1	X	X	X		X
MSK-IMPACT505	X	X	X	X	X
MSK-IMPACT-HEME-400	X	X	X	X	X
PROV-TSO500HT-V2	X	X	X		X
UHN-TSO500-V1	X	X			X
WAKE-CLINICAL-CF2	X	X			
WAKE-CLINICAL-CF3	X	X			
WAKE-CLINICAL-DX1	X	X	X		
WAKE-CLINICAL-AB1	X	X			
WAKE-CLINICAL-AB2	X	X			
WAKE-CLINICAL-AB3	X	X			
VICC-01-D2	X	X	X		X
VICC-01-T4B	X	X	X		X
VICC-01-T6B	X	X	X		X
VICC-01-DX1	X	X	X		X
CHOP-COMPT	X	X			
COLU-CSTP-V2	X	X	X		
MSK-IMPACT-HEME-468	X	X	X	X	X
MSK-ACCESS129	X	X	X	X	X
VICC-02-XTV2	X	X	X	X	X
VICC-02-XTV3	X	X	X	X	X
VICC-02-XTV4	X	X	X	X	X
COLU-CCCP-V2	X	X	X	X	X
WAKE-CLINICAL-R2	X	X			
WAKE-CLINICAL-D2	X	X			
VHIO-300	X	X	X		
CHOP-STNGS-V2	X	X			
CHOP-HEMEP-V2	X	X			
CHOP-COMPT-V2	X	X			
WAKE-CLINICAL-DX2	X	X	X		
PROV-FOUNDATIONONELIQUIDCDX	X	X	X		X
UHN-HM-V1	X	X			
UHN-HM-V2	X	X			
UCHI-ONCOPLUS147-V1	X	X			

Continued on

Table 4: Alteration Types per Panel/Pipeline

	snv	small_indels	gene_level_cna	intragenic_cna	structura
UCHI-ONCOPLUS154-V1	X	X			
DFCI-F1-AB1	X	X	X		X
DFCI-F1-CF3	X	X	X		X
DFCI-F1-CF2	X	X	X		X
UMIAMI-CA-HYBRID-WES	X	X			
UHN-WGS-V1	X	X			
UHN-HPV3	X	X			
UHN-HPV4	X	X			
UHN-MELV1.1	X	X			
UHN-CSTV1	X	X			
PROV-MPNC	X				
PROV-AMLC	X				
PROV-MPNT	X				
UCHI-ONCOPLUS119-V1	X	X			
UMIAMI-GD-360	X	X			
UMIAMI-FD-F1CDX	X	X	X	X	X
UMIAMI-FD-F1H	X	X			
UMIAMI-FD-F1L	X	X			
CHOP-STNGS-V3	X	X			
CHOP-COMPT-V3	X	X			
PROV-CARIS-MIP	X	X	X		X
PROV-TEMPUS-XT	X	X	X		X
PROV-GUARDANT360-CDX	X	X	X		X
PROV-GUARDANT360-V1	X	X	X		X
PROV-GUARDANT360-V2	X	X	X		X

Table 5: Preservation Techniques per Panels/Pipelines

	FFPE	fresh_frozen
CRUK-TS		X
GRCC		X
MDA-46-V1	X	
UCHI-ONCOHEME55-V1		X
UCHI-ONCOSCREEN50-V1	X	
WAKE-CA	X	X
WAKE-CLINICAL-R2D2	X	X

Continued on next page

Table 5: Preservation Techniques per Panels/Pipelines

	FFPE	fresh_frozen
WAKE-CLINICAL-T5A	X	X
WAKE-CLINICAL-T7	X	X
YALE-HSM-V1	X	
YALE-OCP	X	
DFCI-ONCOPANEL-1	X	
DFCI-ONCOPANEL-2	X	
DFCI-ONCOPANEL-3	X	
NKI-TSACP-MISEQ-NGS	X	
MSK-IMPACT341	X	
MSK-IMPACT410	X	
MSK-IMPACT468	X	
UHN-48-V1	X	
UHN-50-V2	X	
UHN-54-V1	X	
UHN-555	X	
VICC-01-MYELOID	X	
VICC-01-SOLIDTUMOR	X	
VICC-01-T5A	X	
VICC-01-T7	X	
MDA-50-V1	X	
UHN-OCA-V3	X	
CHOP-STNGS	X	X
CHOP-HEMEP	X	X
COLU-CCCP-V1	X	X
COLU-TSACP-V1	X	
DUKE-F1-DX1	X	
DUKE-F1-T5A	X	
DUKE-F1-T7	X	
SCI-PMP68-V1	X	
JHU-50	X	
JHU-500STP	X	
VHIO-CUSTOM	X	
NKI-PATH-NGS	X	
NKI-CHPV2-NGS	X	
NKI-CHPV2-SOCV2-NGS	X	
NKI-CHP-V2-PLUS	X	
MDA-409-V1	X	

Continued on next page

Table 5: Preservation Techniques per Panels/Pipelines

	FFPE	fresh_frozen
UCSF-NIMV4	X	X
COLU-CSTP-V1	X	
CHOP-FUSIP	X	
UCSF-IDTV5	X	X
PROV-FOCUS-V1	X	
PROV-TRISEQ-V2	X	
PROV-TST170-V1	X	
MSK-IMPACT505	X	
MSK-IMPACT-HEME-400	X	
PROV-TSO500HT-V2	X	
UHN-TSO500-V1	X	
WAKE-CLINICAL-CF2	X	X
WAKE-CLINICAL-CF3	X	X
WAKE-CLINICAL-DX1	X	X
WAKE-CLINICAL-AB1	X	X
WAKE-CLINICAL-AB2	X	X
WAKE-CLINICAL-AB3	X	X
VICC-01-D2	X	
VICC-01-T4B	X	
VICC-01-T6B	X	
VICC-01-DX1	X	
CHOP-COMPT	X	X
COLU-CSTP-V2	X	
MSK-IMPACT-HEME-468	X	
MSK-ACCESS129	X	
VICC-02-XTV2	X	
VICC-02-XTV3	X	
VICC-02-XTV4	X	
COLU-CCCP-V2	X	X
WAKE-CLINICAL-R2	X	X
WAKE-CLINICAL-D2	X	X
VHIO-300	X	
CHOP-STNGS-V2	X	X
CHOP-HEMEP-V2	X	X
CHOP-COMPT-V2	X	X
WAKE-CLINICAL-DX2	X	X
UHN-HM-V1	X	

Continued on next page

Table 5: Preservation Techniques per Panels/Pipelines

	FFPE	fresh_frozen
UHN-HM-V2	X	
UCHI-ONCOPLUS147-V1	X	
UCHI-ONCOPLUS154-V1	X	
UMIAMI-CA-HYBRID-WES	X	
UHN-WGS-V1	X	
UHN-HPV3	X	
UHN-HPV4	X	
UHN-MELV1.1	X	
UHN-CSTV1	X	
UCHI-ONCOPLUS119-V1	X	
UMIAMI-GD-360	X	
UMIAMI-FD-F1CDX	X	
UMIAMI-FD-F1H	X	
CHOP-STNGS-V3	X	X
CHOP-COMPT-V3	X	X
PROV-CARIS-MIP	X	
PROV-TEMPUS-XT	X	

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
CHOP-COMPT	tumor_normal	238	Custom GENIE-CHOP-COMPT Panel - 238 Genes
CHOP-COMPT-V2	tumor_normal	239	Custom GENIE-CHOP-COMPT Panel - 239 Genes
CHOP-COMPT-V3	tumor_normal	240	Custom GENIE-CHOP-COMPT Panel - 240 Genes
CHOP-FUSIP	tumor_only	111	Custom GENIE-CHOP-FUSIP Panel - 111 Genes
CHOP-HEMEP	tumor_only	118	Custom GENIE-CHOP-HEMEP Panel - 118 Genes
CHOP-HEMEP-V2	tumor_only	119	Custom GENIE-CHOP-HEMEP Panel - 112 Genes
CHOP-STNGS	tumor_only	238	Custom GENIE-CHOP-STNGS Panel - 238 Genes

Continued on next page

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
CHOP-STNGS-V2	tumor_only	239	Custom GENIE-CHOP-STNGS Panel - 239 Genes
CHOP-STNGS-V3	tumor_only	240	Custom GENIE-CHOP-STNGS Panel - 240 Genes
COLU-CCCP-V1	tumor_only	465	Custom CCCPv1 Panel
COLU-CCCP-V2	tumor_only	586	Custom CCCPv2 Panel
COLU-CSTP-V1	tumor_only	45	TruSeq Amplicon Cancer Panel
COLU-CSTP-V2	tumor_only	48	Pillar oncoReveal
COLU-TSACP-V1	tumor_only	49	TruSeq Amplicon Cancer Panel
CRUK-TS	tumor_only	174	Unknown
DFCI-F1-AB1	plasma_normal	324	FoundationOne Liquid CDx
DFCI-F1-CF2	plasma_normal	324	FoundationOne Liquid CDx
DFCI-F1-CF3	plasma_normal	324	FoundationOne Liquid CDx
DFCI-ONCOPANEL-1	tumor_only	275	Custom GENIE-DFCI OncoPanel - 275 Genes
DFCI-ONCOPANEL-2	tumor_only	300	Custom GENIE-DFCI Oncopanel - 300 Genes
DFCI-ONCOPANEL-3	tumor_only	447	Custom GENIE-DFCI Oncopanel - 447 Genes
DFCI-ONCOPANEL-3.1	tumor_only	447	Custom GENIE-DFCI Oncopanel - 447 Genes
DUKE-F1-DX1	tumor_only	324	FoundationOne CDx Panel
DUKE-F1-T5A	tumor_only	244	Foundation Medicine T5a Panel - 244 Genes
DUKE-F1-T7	tumor_only	322	Foundation Medicine T7 Panel - 322 Genes
GRCC-2022569	tumor_only	146	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2022572	tumor_only	2	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2030859	tumor_only	166	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2040388	tumor_only	44	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2040390	tumor_only	138	Ion AmpliSeq Cancer Hotspot Panel v2

Continued on next page

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
GRCC-2058737	tumor_only	82	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2061930	tumor_only	161	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2063107	tumor_only	150	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2067719	tumor_only	54	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2067720	tumor_only	44	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2105390	tumor_only	44	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2106321	tumor_only	44	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2112732	tumor_only	151	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-2176207	tumor_only	151	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-CHP2	tumor_only	50	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-CP1	tumor_only	40	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-MOSC3	tumor_only	75	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-MOSC4	tumor_only	75	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-OCAV3	tumor_only	75	Ion AmpliSeq Cancer Hotspot Panel v2
GRCC-SAFIR02	tumor_only	75	Ion AmpliSeq Cancer Hotspot Panel v2
JHU-500STP	tumor_only	760	Illumina NGS instruments
JHU-50GP	tumor_only	50	Ion AmpliSeq Cancer Hotspot Panel v2
MDA-409-V1	tumor_only	409	Ion AmpliSeq Comprehensive Cancer Panel

Continued on next page

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
MDA-46-V1	tumor_only	46	Custom AmpliSeq Cancer Hotspot GENIE-MDA Augmented Panel v1 - 46 Genes
MDA-50-V1	tumor_only	50	Ion AmpliSeq Cancer Hotspot Panel v2
MSK-ACCESS129	tumor_normal	129	Custom MSK ACCESS Panel - 129 Genes
MSK-IMPACT-HEME-400	tumor_normal	399	Custom MSK IMPACT HEME Panel - 400 Genes
MSK-IMPACT-HEME-468	tumor_normal	467	Custom MSK IMPACT HEME Panel - 468 Genes
MSK-IMPACT341	tumor_normal	341	Custom MSK IMPACT Panel - 341 Genes
MSK-IMPACT410	tumor_normal	410	Custom MSK IMPACT Panel - 410 Genes
MSK-IMPACT468	tumor_normal	468	Custom MSK IMPACT Panel - 468 Genes
MSK-IMPACT505	tumor_normal	505	Custom MSK IMPACT Panel - 505 Genes
NKI-CHP-V2-PLUS	tumor_only	52	Ion AmpliSeq Cancer Hotspot Panel v2
NKI-CHPV2-NGS	tumor_only	50	Ion AmpliSeq Cancer Hotspot Panel v2
NKI-CHPV2-SOCV2-NGS	tumor_only	52	Ion AmpliSeq Cancer Hotspot Panel v2 plus SOCv2
NKI-PATH-NGS	tumor_only	32	PATH (Predictive analysis for therapy) panel
NKI-TSACP-MISEQ-NGS	tumor_only	48	TruSeq Amplicon - Cancer Panel
PROV-AMLC	plasma_normal	28	Archer VARIANTPlex Myeloid
PROV-CARIS-MIP	tumor_only	869	Caris MI Profile (Agilent SureSelect Exome V7 + 720G hybrid capture + Archer FusionPlex Solid Tumor)
PROV-FOCUS-V1	tumor_only	46	Oncomine Focus Assay, AmpliSeq Library

Continued on next page

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
PROV-FOUNDATIONONELIQUIDCDX	plasma_normal	324	FoundationOne Liquid CDx
PROV-GUARDANT360-CDX	plasma_normal	74	Guardant360 CDx (Guardant Health, FDA-approved)
PROV-GUARDANT360-V1	plasma_normal	85	Guardant360 LDT (Guardant Health, pre-2024 versions)
PROV-GUARDANT360-V2	plasma_normal	745	Guardant360 LDT (Guardant Health, 2024+ Smart Liquid Biopsy)
PROV-MPNC	plasma_normal	36	Archer VARIANTPlex Myeloid
PROV-MPNT	plasma_normal	3	Archer VARIANTPlex Myeloid
PROV-TEMPUS-XT	tumor_normal	691	Tempus xT (versions V4 + ONCO-V1 union)
PROV-TRISEQ-V2	tumor_normal	331	xGen Exome Research Panel v2
PROV-TSO500HT-V2	tumor_only	523	TruSight Oncology 500 High-Throughput V2
PROV-TST170-V1	tumor_normal	160	TruSight Tumor 170
SCI-PMP68-V1	tumor_only	68	TruSeq Amplicon Cancer Panel
UCHI-ONCOHEME55-V1	tumor_only	55	Kappa
UCHI-ONCOPLUS119-V1	tumor_only	119	Roche SeqCap EZ supplemented with IDT xGen Lockdown Probes
UCHI-ONCOPLUS147-V1	tumor_only	147	Roche SeqCap EZ supplemented with IDT xGen Lockdown Probes
UCHI-ONCOPLUS154-V1	tumor_only	154	Roche SeqCap EZ supplemented with IDT xGen Lockdown Probes
UCHI-ONCOSCREEN50-V1	tumor_only	55	Kappa

Continued on next page

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
UCSF-IDTV5-TN	tumor_normal	531	Custom GENIE-UCSF-IDTV5 Panel - 531 Genes
UCSF-IDTV5-TO	tumor_only	531	Custom GENIE-UCSF-IDTV5 Panel - 531 Genes
UCSF-NIMV4-TN	tumor_normal	478	Custom GENIE-UCSF-NIMV4 Panel - 478 Genes
UCSF-NIMV4-TO	tumor_only	478	Custom GENIE-UCSF-NIMV4 Panel - 478 Genes
UHN-48-V1	tumor_normal	48	TruSeq Amplicon Cancer Panel
UHN-50-V2	tumor_only	50	Ion AmpliSeq Cancer Hotspot Panel v2
UHN-54-V1	tumor_only	54	TruSight Myeloid Sequencing Panel
UHN-555-BLADDER-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-BREAST-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-GLIOMA-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-GYNE-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-HEAD-NECK-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-LUNG-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-MELANOMA-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-PAN-GI-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-PROSTATE-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-RENAL-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-V1	tumor_only	556	Custom SureSelect GENIE-UHN Panel - 555 Genes
UHN-555-V2	tumor_only	564	Custom SureSelect GENIE-UHN Panel - 555 Genes

Continued on next page

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
UHN-CST-11	tumor_only	11	Custom GENIE-UHN-CSTv1
UHN-CST-2	tumor_only	2	Custom GENIE-UHN-CSTv1
UHN-CST-3	tumor_only	3	Custom GENIE-UHN-CSTv1
UHN-CST-43	tumor_only	43	Custom GENIE-UHN-CSTv1
UHN-CST-7	tumor_only	7	Custom GENIE-UHN-CSTv1
UHN-CST-8	tumor_only	8	Custom GENIE-UHN-CSTv1
UHN-HM-26	tumor_only	26	Custom OGT GENIE-UHN-HM-V1
UHN-HM-29	tumor_only	29	Custom OGT GENIE-UHN-HM-V1
UHN-HM-V1	tumor_only	49	Custom OGT GENIE-UHN-HM-V1
UHN-HM-V2	tumor_only	49	Custom OGT GENIE-UHN-HM-V2
UHN-HP-1	tumor_only	1	Custom GENIE-UHN-HPv4
UHN-HP-2	tumor_only	2	Custom GENIE-UHN-HPv4
UHN-HP-8	tumor_only	8	Custom GENIE-UHN-HPv4
UHN-HPV3-19	tumor_only	19	Custom GENIE-UHN-HPv4
UHN-HPV4-19	tumor_only	19	Custom GENIE-UHN-HPv4
UHN-MEL-11	tumor_only	11	Custom GENIE-UHN-MELv1.1
UHN-MEL-12	tumor_only	12	Custom GENIE-UHN-MELv1.1
UHN-MEL-13	tumor_only	13	Custom GENIE-UHN-MELv1.1
UHN-MEL-34	tumor_only	34	Custom GENIE-UHN-MELv1.1
UHN-MEL-7	tumor_only	7	Custom GENIE-UHN-MELv1.1
UHN-OCA-V3	tumor_only	146	Ion Oncomine Comprehensive Assay v3
UHN-TSO500-V1	tumor_only	523	TruSight Oncology 500
UHN-WGS-V1	tumor_normal	19533	Not Applicable
UMIAMI-CA-HYBRID-WES	tumor_only	736	MI Cancer Seek Sequencing Kit
UMIAMI-FD-F1CDX	tumor_only	324	FoundationOne CDx
UMIAMI-FD-F1H	tumor_only	406	FoundationOne Heme
UMIAMI-FD-F1L	plasma_normal	324	FoundationOne Liquid CDx
UMIAMI-GD-360	tumor_only	736	MI Cancer Seek Sequencing Kit
VHIO-300	tumor_only	435	SureSelect XT, Agilent
VHIO-BILIARY-V01	tumor_only	59	VHIO Custom Amplicon panel-hotspots
VHIO-BRAIN-V01	tumor_only	57	VHIO Custom Amplicon panel-hotspots

Continued on next page

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
VHIO-BREAST-V01	tumor_only	60	VHIO Custom Amplicon panel-hotspots
VHIO-BREAST-V02	tumor_only	62	VHIO Custom Amplicon panel-hotspots
VHIO-COLORECTAL-V01	tumor_only	60	VHIO Custom Amplicon panel-hotspots
VHIO-ENDOMETRIUM-V01	tumor_only	60	VHIO Custom Amplicon panel-hotspots
VHIO-GASTRIC-V01	tumor_only	63	VHIO Custom Amplicon panel-hotspots
VHIO-GENERAL-V01	tumor_only	56	VHIO Custom Amplicon panel-hotspots
VHIO-HEAD-NECK-V01	tumor_only	61	VHIO Custom Amplicon panel-hotspots
VHIO-KIDNEY-V01	tumor_only	59	VHIO Custom Amplicon panel-hotspots
VHIO-LUNG-V01	tumor_only	58	VHIO Custom Amplicon panel-hotspots
VHIO-OVARY-V01	tumor_only	58	VHIO Custom Amplicon panel-hotspots
VHIO-PANCREAS-V01	tumor_only	60	VHIO Custom Amplicon panel-hotspots
VHIO-PAROTIDE-V01	tumor_only	58	VHIO Custom Amplicon panel-hotspots
VHIO-SKIN-V01	tumor_only	60	VHIO Custom Amplicon panel-hotspots
VHIO-URINARY-BLADDER-V01	tumor_only	61	VHIO Custom Amplicon panel-hotspots
VICC-01-D2	tumor_only	507	Foundation Medicine HemeComplete Panel - 507 Genes
VICC-01-DX1	tumor_only	324	Foundation Medicine DX1 Panel - 324 Genes
VICC-01-MYELOID	tumor_only	37	Custom Myeloid GENIE-VICC Panel - 37 Genes

Continued on next page

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
VICC-01-SOLIDTUMOR	tumor_only	31	Custom Solid Tumor GENIE-VICC Panel - 34 Genes
VICC-01-T4B	tumor_only	214	Foundation Medicine Clinical Panel - 214 Genes
VICC-01-T5A	tumor_only	323	Foundation Medicine T5a Panel - 322 Genes
VICC-01-T6B	tumor_only	411	Foundation Medicine HemeOnc Panel - 411 Genes
VICC-01-T7	tumor_only	429	Foundation Medicine T7 Panel - 429 Genes
VICC-02-XTV2	tumor_normal	596	Custom Tempus XT Probe Set
VICC-02-XTV3	tumor_normal	648	Custom Tempus XT Probe Set
VICC-02-XTV4	tumor_normal	648	Custom Tempus XT Probe Set
WAKE-CA-01	tumor_only	32	Caris
WAKE-CA-NGSQ3	tumor_only	591	Caris
WAKE-CLINICAL-AB1	tumor_only	223	Foundation Medicine AB1 Panel
WAKE-CLINICAL-AB2	tumor_only	9	Foundation Medicine AB2 Panel
WAKE-CLINICAL-AB3	tumor_only	10	Foundation Medicine AB3 Panel
WAKE-CLINICAL-CF2	tumor_only	4	Foundation Medicine CF2 Panel
WAKE-CLINICAL-CF3	tumor_only	11	Foundation Medicine CF3 Panel
WAKE-CLINICAL-D2	tumor_only	42	Foundation Medicine D2 Panel
WAKE-CLINICAL-DX1	tumor_only	313	Foundation Medicine DX1 Panel
WAKE-CLINICAL-DX2	tumor_only	285	Foundation Medicine DX2 Panel
WAKE-CLINICAL-R2	tumor_only	30	Foundation Medicine R2 Panel
WAKE-CLINICAL-R2D2	tumor_only	266	Foundation Medicine R2D2 Panel

Continued on next page

Table 6: Sequence Assay Genomic Information

Sequencing Assay	Calling Strategy	Number of genes	Target Capture Kit
WAKE-CLINICAL-T5A	tumor_only	70	Foundation Medicine T5a Panel - 322 Genes
WAKE-CLINICAL-T7	tumor_only	261	Foundation Medicine T7 Panel - 429 Genes
YALE-HSM-V1	tumor_only	50	Ion AmpliSeq Cancer Hotspot Panel v2
YALE-OCP-V2	tumor_normal	134	Ion Oncomine Comprehensive Assay v2
YALE-OCP-V3	tumor_normal	146	Ion Oncomine Comprehensive Assay v3

Genomic Profiling at Each Center

Children's Hospital of Philadelphia (CHOP)

The CHOP Comprehensive Solid Tumor Panel and Comprehensive Hematologic Cancer Pane include sequence and copy number analyses of 238 and 117 cancer genes, respectively, genotyping of two genes associated with cancer pharmacogenomics, and a Fusion Panel targeting over 700 exons of 117 cancer genes.

Next generation sequencing (NGS) and data analysis: Nucleic acid is extracted from the patient's sample following standard DNA and RNA extraction protocols. Extracted DNA is fragmented and tagged using SureSelect QXT target enrichment to generate adapter-tagged libraries. Biotin-labeled probes specific to the targeted regions are used for capture hybridization. Libraries are enriched for the desired regions using streptavidin beads. Enriched libraries are then indexed and pooled for sequencing. Libraries are subject to sequence analysis on Illumina NovaSeq 6000 system for 150 bp paired end reads. All coding exons and the flanking intron sequences of targeted genes in the panel are sequenced, and selected promoter regions and known intronic variants are also evaluated. Sequence data are analyzed using the home brew software ConcordS V4.0.0 and NextGENe V2 NGS Analysis Software. Sequence variants within exons and 5 bp flanking intron sequences are annotated. Copy number variation (CNV) analysis for gross deletions and duplications are evaluated using NGS data. RNA sequencing libraries are prepared using Archer Universal RNA Reagent Kit with CHOP fusion panel custom-designed primers with target specific molecular barcode. Sequencing data are analyzed using ArcherTM Analysis for fusion genes. Clinically significant variants including single nucleotide variants (SNVs), indels, CNVs and fusion genes are confirmed by Sanger sequencing, MLPA, Real-Time PCR, or ddPCR only when necessary.

Variant categorization and reporting: Sequence variants, copy number variants, and gene fusions are evaluated based on the currently available information from relevant resources, such as professional guidelines, clinical and population variant databases, tumor specific databases, and the scientific literature. Sequence variants are reported according to HGVS nomenclature [den Dunnen 2016, PMID: 26931183]. Somatic variants are classified using criteria consistent with those recommended by the Association for Molecular Pathology (AMP), American Society of Clinical Oncology (ASCO), and College of American Pathologists (CAP) [Li 2017, PMID: 27993330], as described below. Tier 1-3 variants are reported. Tier 4 variants are not reported.

TPMT and NUDT15 genotyping: NUDT15 diplotypes *1 through* 9, and TPMT diplotypes *1,* 2, *3A,* 3B, and *3C, are assessed [Relling 2019, PMID: 30447069; CPIC Guideline for Thiopurines and TPMT and NUDT15].

Herbert Irving Comprehensive Cancer Center, Columbia University (COLU)

Columbia University Irving Medical Center uses the Illumina TruSeq Amplicon –Cancer Panel (TSACP) to detect known cancer hotspots.DNA is extracted from unstained sections of FFPE tissue paired with an H&E stained section that is used to ensure adequate tumor cellularity (human assessment $>$ 30%) and marking of the tumor region of interest (macrodissection). Extraction for FFPE tissue is performed on the QiaCube instrument (Qiagen). 50-250ng of genomic DNA is used as input.Tumors are sequenced to an average depth of at least 1000X. Alignment (to hg19) and variant calling is performed using NextGENe v2.4.2software. Variants lower than 1% allele frequency in all three control populations (White, African American, Asian) of the Exome Variant Server database, the 1000 genome project database are retained, and annotation of variants is performed using a custom pipeline. All cases are reviewed and interpreted by a molecular pathologist.

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 50x.

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 $\geq$.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 exonicregions 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.

Duke Cancer Institute (DUKE)

Foundation medicine panels: Duke uses Illumina hybridization-based capture panels from Foundation Medicine to detect single nucleotide variants, small indels, copy number alterations and structural variants from FFPE, tumor-only sequencing data. Three gene panels were

used: Panel 1 (T5a bait set), covering 326 genes, Panel 2 (T7 bait set), covering 434 genes, and Panel 3 (DX1 bait set), covering 324 genes. The clinical sequencing data were analyzed by Foundation Medicine-developed pipelines. Briefly: 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 $\geq$ 5% (MAF $\geq$ 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.

Institut Gustave Roussy (GRCC)

Gustave Roussy Cancer Centre submitted data includes somatic variants (single nucleotide variants and small indels) identified with CancerHotspot 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 SuiteTMVariant 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%.

Johns Hopkins Sidney Kimmel Comprehensive Cancer Center (JHU)

Johns Hopkins submitted genomic data from the Ion AmpliSeqCancer 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 adequatetumor 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 withcustom 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.

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 anamplicon-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 wasextracted from unstained sections of tissue paired with astained 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% variantallelic 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 (a pool of normals is used for a small subset of samples with a missing normal). 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.

Netherlands Cancer Center, The Netherlands (NKI)

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 isused 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.

Providence Health & Services Cancer Institute (PROV)

PROV has submitted data from two assays: PROV-focus-v1 and PROV-triseq-v1. For the PROV-focus-v1 we have employed the Thermo Fisher Oncomine Focus Assay for amplification of 52 genes from DNA extracted from macro-dissected FFPE samples taken from Pathologist specified tumor regionsof interest (ROIs), with 20% minimal tumor cellularity. Samples include primary and metastatic tumor ROIs. The assay is a tumor only assay, no paired "normal" DNA is extracted from each case.

The PROV-focus-v1 BED file describes the positions of the genome assayed by the PROV-focus-v1 panel relative to hg19.

Amplification products are sequenced on the Life Technology Ion Torrent platform to an average read depth of coverage greater than 500X average per base coverage.

The TMAP aligner developed by Life Technology for the ION torrent sequencing platform was used to align reads to hg19 using the manufacture suggested settings. Variants are called with the Torrent Suite Variant Caller 4.2 software plug-in.

Variant filters requiring a total read depth of greater than 100X, variant allele coverage of greater than 10X, and a variant allele frequency for substitutions of greater than or equal to 0.03 are applied. Also, the specimen variant must not be annotated as "COMMON" (a variant allele frequency for substitutions of ≥ 0.05) in dnSNP. VCF files were created for upload to GENIE 6.1 by further filtering all detected variants to only those reported after expert review by clinicians.

For the PROV-triseq-v1 we used DNA extracted from macro-dissected FFPE samples taken from Pathologist specified tumor regions of interest (ROIs), with 20% minimal tumor cellularity for extraction of tumor DNA, and whole peripheral blood for extraction of normal DNA. Tumor samples include both primary and metastatic tumor ROIs.

The PROV-triseq-v1 BED file describes the positions of the genome assayed by the PROV-triseq-v1 panel relative to hg19.

Libraries are prepared using the KAPA for Illumina reagents protocols. Indexed libraries are pooled for exome capture on the xGen V1.0 panel (https://www.idtdna.com/). Sequencing is performed on Illumina 2500, 4000, or Novaseq platforms.

Raw sequencing data in the form of BCL files are uploaded to the Providence secure computing cloud environment maintained by Amazon Web Services. Following upload, raw files are converted to unaligned reads in FASTQ format using the software program bcl2fastq2, and resultant FASTQ files are aligned to the hg19 human reference genome using the Burrows-Wheeler Aligner (BWA). Aligned reads in the SAM format are subsequently converted to binary BAM format using the samtools software package, and aligned reads are processed for single-nucleotide variants (SNVs) and short insertions and deletions (indels) using our custom variant calling pipeline (see below). FASTQ and aligned BAM files are analyzed with FastQC and Picard metrics for Molecular Genomics Lab staff run-level and sample-level review.

The Providence variant calling pipeline includes multiple variant calling algorithms including VarScan2, SomaticSniper, Mutect2 and Strelka. Variant filters requiring a total read depth of greater than 100X, variant allele coverage of greater than 10X, and a variant allele frequency for substitutions of greater than or equal to 0.03 are applied. Calls with low-quality variants, silent mutations, and germline variants are also filtered. Annotations from SnpEff, ClinVar, ExAC, 1000 Genomes, ANNOVAR, and COSMIC are incorporated for each call. Finally, all common variants, with non-zero allele frequencies the ExAC or 1000 Genomes databases, are removed.

GENIE 6.1 VCF files containing annotated calls from Mutect2 were created for upload to GENIE 6.1.

Swedish Cancer Institute (SCI)

SCI uses CellNetixPMP gene panel to detect hotspot mutations in known cancer genes from solid tumor DNA (Formalin-fixed, paraffin-embedded tissue). The hotspot gene panel covers 68 genes. Tumor cell content is greater than 10% verified by pathologist. Tumor DNA is sequenced to >200x on average (Variant that allele frequency is less than 10% requires more than 400X) on Illumina MiSeq (TruSeq Amplicon) platform, and data is analyzed in MiSeq Reporter 2.5. Reads are aligned to hg19 reference genome by the BWA (v0.6.1-r104-tpx) aligner adapted by the MiSeq Reporter Software (v2.4.1 or v2.5) using the manufacture suggested settings. MiSeq Reporter provided Somatic Variant Caller (v2.1.12) is run on the aligned .bam files to identify variants present in DNA samples. Detailed stepsplease refer toIllumina MiSeq Reporter User Guide. Variants are filtered for allele frequency greater than 3% except for actionable mutations. Variants that are observed in $\geq$ 75% samples on the same run, or common variant with population frequency of $>$ 50%, or average population frequency >5% reported in the 1000 genome and/or in ExAc. are filtered.

The University of Chicago (UCHI)

The University of Chicago submitted data in the current data set includes somatic variants (SNVs and INDELS) identified using two amplicon-based targeted tumor-only assays. The

ONCOSCREEN assay is tumor only targeted gene panel covering 50 genes. The ONCOHEME assay is a tumor only targeted gene panel covering 53 genes. Both ONCOSCREEN and ONCOHEME assays follow the same DNA extraction, data processing, alignment and variant calling steps. Only ONCOSCREEN assay has sample replicates (The samples were prepped and sequenced in duplicate and the variant calls were reported only if the variant was detected in both replicates). DNA is extracted from unstained sections of FFPE tissue paired with an H&E stained section that is used to ensure adequate tumor cellularity (human assessment > 20%) and marking of the tumor region of interest (macrodissection). No paired "normal" DNA is extracted from each case. The Novoalign v3.02.07 aligner (NovoCraft, Selangor, Malaysia) was used to align to hg19 using automatic adapter and primer trimming options for paired end 2x152bp reads sequencing for ONCOSCREEN assay and paired end 2x255bp reads sequencing for ONCOHEME assay on the Illumina Miseq platform. The input is a .fastq file and the output is a .sam file. Before alignment, off-target reads are removed from the FASTQ files by matching the primer pairs of the amplicons to the reads. After alignment, we use picard tools v1.92 to convert the .sam file to .bam file. samtools v0.1.19 mpileup is run on the aligned .bam file. Custom python scripts process the output of samtools mpileup for each sample to identify SNV & INS/DEL. Details in here. The filtered FASTQ files were analyzed for the presence of insertion and deletion mutations >5bp using the reference-independent Amplicon Indel Hunter. Variant filtering criteria was applied. For ONCOSCREEN assay, the cutoff is as follows: Read Depth with Phred score (>Q30) >= 200 and Variant allele frequency - SNV >= 0.05, INS/DEL >= 0.05. The reported depth is that of bases with Phred quality score >30. Depth and allele frequency values of the replicates were averaged for reporting. For ONCOHEME assay the cutoff is as follows: Read Depth with Phred score (>Q30) >= 100 and Variant allele frequency - SNV >= 0.05, INS/DEL >= 0.05. The reported depth is that of bases with Phred quality score >30. No filters for data uploaded to GENIE for both assays.

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. Two versions of the panel have been submitted to GENIE: UCSF-NIMV4 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. UCSF-IDTV5 consists of 529 genes, retains TERT and SDHD promoter coverage, and expands intronic or UTR region coverage to 73 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 prior to October 2020, and NovaSeq6000 after. 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 approximately 30 pooled normal samples

Variants are removed if present with frequency $\geq 1\%$ in ESP6500, 1000 Genomes, or ExAC datasets. 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 (6) panels to sequence samples - UHN-48-V1, UHN-50-V2, UHN-54-V1, UHN-555-V1, UHN-555-V2 and UHN-OCAv3. 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:

- 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%.
- Data analysis of tumor only processed by UHN_TSACP_tumorONLY_v2_workflow: MiSeq fastq were aligned using (MiSeqReporter 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.

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.

UHN Custom 555 Gene Panel (UHN-555-V1 and UHN-55-V2): Princess Margaret Cancer Centre used a custom Sure Select (Agilent) 555 gene panel to detect single nucleotide variants and small indels in DNA from tumour tissue. Specimens are reviewed by a pathologist to ensure tumor cellularity of at least 10% (usually >20%). Tumors are sequenced to an average unique depth of coverage of approximately 500x. NextSeq fastq were aligned using bwa-mem and the BAMs were processed using GATK best practices. Variants were called using VarScan v2.3.8. VCF files were filtered to remove known artifacts. Variants that are not on a manually

curated list of hotspots that have depth less than 50, or if depth 50 to 100 with frequency <10%, or if depth was at least 100 with frequency <5%, were removed.

Ion Oncomine Comprehensive Assay v3 (UHN-OCAv3): Princess Margaret Cancer Centre used the Thermofisher Oncomine Comprehensive Assay v3 to detect relevant single nucleotide variants, small indels, copy number changes and gene fusions. Genomic DNA and RNA were co-isolated and then sequencing was performed on the Ion S5 XL System. Variant calls were generated using Torrent Browser 5.12 (ThermoFisher) with alignment to genome build GRCh37/hg19. Filtering of reportable variants uses the Oncomine Variant Annotator plugin to annotate variants in Ion Reporter 5.12, with variant criteria and thresholds as supplied. The reportable range is 5%-100% variant allele frequency. Minimum acceptable coverage for all reported genomic regions is >200x. Copy number gains and losses are assessed using the Oncomine Variant Annotator plugin, with thresholds of <0.5 or >4 copies reported. Gene fusions are assessed using the Oncomine Variant Annotator plugin, with variant criteria and thresholds as supplied (quality threshold of >500,000 reads reported as 'pass' for gene fusion sequencing). The assay detects recurrent and novel gene partners with expressed gene fusions of the genes listed and minimum acceptable read count for all reported fusions is >2000.

University of Miami - Sylvester Comprehensive Cancer Center (UMIAMI)

Caris Life Sciences

Direct sequence analysis was performed on genomic DNA isolated from a micro-dissected tumor sample using Illumina NovaSeq 6000 sequencers. Sequencing was conducted through Caris Life Sciences, utilizing a hybrid pull-down panel of baits to enrich more than 700 clinically relevant genes along with > 20,000 other genes. Base calls are detected from the signal intensity measurements during each cycle. Base Call (BCL) files generated from the Illumina NovaSeq systems are stored on a secure High-Performance Computing server on rapid solid-state hard drives.

The initial sequencing metrics for the flow cells are reviewed by the NGS bioinformatics pipeline, which includes the following run validity criteria and quality control checks: Reads Passing Filter, Yield, and Q30 scores for each read 1 and read 4 (forward and reverse). The pipeline initiates automatic conversion and de-multiplexing using the FPGA-adapted version of bcl2fastq via Illumina DRAGEN devices to convert the Illumina-generated BCL files to FASTQ file formats for downstream analysis. Demultiplexing is performed on the TNA reads, creating sample-specific FASTQ files. Custom software is applied to differentiate RNA and DNA molecules using the synthetic labels added during cDNA synthesis.

Guardant Health

The Guardant360 assay uses next-generation sequencing–based digital sequencing (DS) to profile 73 cancer-related genes, including clinically actionable and cancer-specific biomarkers.

Cell-free DNA (cfDNA) is extracted from stabilized blood (using QIAamp Circulating Nucleic Acid Kit), labeled with molecular barcodes, enriched through hybrid capture, and sequenced (on Illumina platforms NextSeq 500/ HiSeq 2500) in a CLIA-certified, CAP-accredited laboratory. Sequencing data are used to reconstruct each cfDNA molecule with high accuracy through double-stranded consensus sequencing. This process achieves 60–85% molecule recovery and highly uniform coverage across the gene panel.

DS significantly reduces sequencing errors—by over three orders of magnitude compared with standard NGS—enabling reliable detection of variants at very low allele fractions (0.01% for indels and 0.04% for SNVs and fusions). Over half of reconstructed molecules derive from both DNA strands, enhancing accuracy and specificity beyond other sequencing methods that recover both strands in fewer than 12% of molecules.

Variant analysis used the locked Guardant360 bioinformatics pipeline, with thresholds pre-established using independent training cohorts. Reads were demultiplexed, quality-trimmed, and aligned to hg19, then used to build double-stranded consensus molecules.

Foundation Medicine

University of Miami uses Illumina hybridization-based capture panels to detect single nucleotide variants, small indels, copy number alterations and structural variants from FFPE tumor or cell-free blood DNA sequencing using the following:

Foundation Medicine panels: The gene panels were used for FFPE tumor sequencing are: * One CDx (DX2 bait set) covering 338 genes * One CDx (DX1 bait set) covering 324 genes * One (T5a bait set) covering 328 genes * One (T7 bait set) covering 435 genes * Heme (D2 bait set) covering 508 genes * Heme (R2 bait set) covering 498 genes * Heme (T4b bait set) covering 220 genes * Heme (T6b bait set) covering 412 genes

Foundation Medicine panels: The gene panels used for cell free blood DNA sequencing are: * Liquid CDx (AB1 bait set) covering 324 genes * Liquid (CF3 bait set) covering 70 genes

FoundationOne CDx

FoundationOne®CDx (F1CDx) is performed exclusively as a laboratory service using DNA extracted from formalin-fixed, paraffin-embedded (FFPE) tumor samples. The assay employs CoExtraction, an automated DNA/RNA co-extraction methodology for DNA extraction from routine FFPE biopsy or surgical resection specimens; 50-1000 ng of DNA will undergo whole-genome shotgun library construction and hybridization-based capture of all coding exons from 309 cancer-related genes, one promoter region, one non-coding (ncRNA), and select intronic regions from 34 commonly rearranged genes, 21 of which also include the coding exons. In total, the assay detects alterations in a total of 324 genes. Using the Illumina® NovaSeq 6000, hybrid capture–selected libraries are sequenced to high uniform depth (targeting >500X median coverage with >99% of exons at coverage >100X). Sequence data is then processed using

a customized analysis pipeline designed to detect all classes of genomic alterations, including base substitutions, indels, copy number alterations (amplifications and homozygous gene deletions), and select genomic rearrangements (e.g., gene fusions). Rearrangements in one of the targeted genes may be reported along with their uniquely identified genomic partners, which can be any gene in the genome even if not explicitly targeted by the assay. Additionally, genomic signatures including microsatellite instability (MSI) and tumor mutational burden (TMB) are reported.

FoundationOne Heme

FoundationOne®Heme uses hybrid capture-based next-generation sequencing. It identifies the four main classes of genomic alterations (base substitutions, insertions and deletions, copy number alterations, and rearrangements). This assay sequences complete coding region of 406 genes using DNA, including introns of 31 select genes involved in rearrangements, to a median depth of ~500X unique coverage. The assay also sequences RNA of 265 genes commonly rearranged in cancer to better identify known and novel gene fusions to an average ~6.9 million unique pairs.

FoundationOne Liquid CDx

FoundationOne® Liquid CDx utilizes circulating cell-free DNA (cfDNA) isolated from plasma derived from anti-coagulated peripheral whole blood. The assay employs a single DNA extraction method to obtain cfDNA from plasma from whole blood. Extracted cfDNA undergoes whole-genome shotgun library construction and hybridization-based capture of 324 cancer-related genes. Hybrid-capture selected libraries are sequenced with deep coverage using the NovaSeq® 6000 platform. Sequence data are processed using a custom analysis pipeline designed to detect genomic alterations, including base substitutions and indels, copy number variants, and genomic rearrangements.

Vall d'Hebron Institute of Oncology (VHIO)

Vall d'Hebron institute of Oncology (VHIO) submitted data that includes somatic variants (single nucleotide variants and small indels) identified with VHIO Card Amplicon panels that target frequently mutated regions in oncogenes and tumor suppressors. A total of fifteen panels have been submitted taking different tumor types into consideration. The panels are:

- VHIO-GENERAL-V01: Panel containing 56 oncogenes and tumor suppressor genes
- VHIO-BRAIN-V01 (General + NF1 v1: 57 genes)
- VHIO-BILIARY-V01 (General + *FGFR v1 + **NOTCH v1: 59 genes)
- VHIO-COLORECTAL-V01 (General + RingFingers v1 + **NOTCH v1: 60 genes)

- VHIO-HEAD-NECK-V1 (General + MTOR v1 + **NOTCH v1: 61 genes)
- VHIO-ENDOMETRIUM-V01 (General + RingFingers v1 + *FGFR v1 + NF1 v1: 60 genes)
- VHIO-GASTRIC-V01 (General + RingFingers v1 + MTOR v1 + **NOTCH v1: 63 genes)
- VHIO-PAROTIDE-V01 (General + **NOTCH v1: 58 genes)
- VHIO-BREAST-V01 (General + *FGFR v1 + **NOTCH v1+ GATA3 v1: 60 genes)
- VHIO-OVARY-V01 (General + BRCA v1: 58 genes)
- VHIO-PANCREAS-V01 (General + Ring Fingers v1 + BRCA v1: 60 genes)
- VHIO-SKIN-V01 (General + NF1 v1 + MTOR v1: 60 genes)
- VHIO-LUNG-V01 (General + NF1 v1 + MET v1 + FGFRw7 v1: 58 genes)
- VHIO-KIDNEY-V01 (General + MTOR v1: 59 genes)
- VHIO-URINARY-BLADDER-V01 (General + *FGFR v1 + NF1 v1 + MTOR v1: 61 genes)

*FGFRv1 panel includes extra regions in FGFR1, FGFR2 and FGFR3 genes. **NOTCHv1 panel includes extra regions in FBXW7 and NOTCH1 genes. FGFRw7 v1 panel includes extra regions in FGFR1 gene. MET v1 panel includes intronic regions flanking Exon 14 of MET gene.

Tumor samples are reviewed by a pathologist to ensure tumor cellularity of at least 20%. For the sample loading into tumor-specific panels, we use a FREEDOM EVO 150 Platform from TECAN. Tumors are sequenced in an Illumina MiSeq instrument, to an average depth of coverage of approximately 1000X. Samples are sequenced, and two independent chemistries are performed and sequenced. Sequencing reads are aligned (BWA v0.7.17, Samtools v1.9), base recalibrated, Indel realigned (GATK v3.7.0), and variant called (VarScan2 v2.4.3). A minimum of 7 reads supporting the variant allele is required in order to call a mutation. Frequent SNPs in the population are filtered with the 1000g database (MAF>0.005). The average number of reads representing a given nucleotide in the panel (Sample Average Coverage) is calculated. Manual curation of variants is performed after manual search of available literature and databases, in terms of their clinical significance.

- VHIO-300 panel

DNA from tumor-FFPE sample was obtained (Maxwell® RSC FFPE Plus DNA Kit (Promega)) and a custom gene capture approach (see below) performed (enrichment probes: SureSelect XT, Agilent). The resulting library was sequenced using the Illumina sequencing by synthesis (SBS) technology (2 x 100 PE run).

Sequencing reads were aligned (BWA v0.7.17, Samtools v1.9) against the hg19 reference genome, base recalibrated, indel realigned (GATK v3.7.0, abra2 v2.23) and variant called (VarScan2 v2.4.3, Mutect2 v4.1.0.0). Variants from both callers are reported. A minimum of 5 reads supporting the variant allele were required to call a mutation. The sensitivity of the technique is 5% MAF for SNVs and 10% MAF for INDELs. Frequent single nucleotide polymorphisms (SNPs) in the population were filtered based on the gnomAD database (allele frequency 0.0001) and copy number alterations (CNA) were calculated (CN- Vkit v0.9.6.dev0). Variants were manually curated and classification of identified variants was performed using publicly available databases (COSMIC, cBioPortal, ClinVar, VarSome, OncoKB).

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 $\geq 80\%$ and tumor cellularity $\geq 20\%$, with use of macro-dissection to enrich samples with $\leq 20\%$ 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 $\geq 5\%$ (MAF $\geq 1\%$ 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 thesequencing analysis pipelines from Foundation Medicine and Caristo 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× (IlluminaHiSeq2000 platform using 49 × 49 paired-end reads).The clinical sequencing data were analyzed by Foundation Medicine and Carisdevelopedpipelines.Sequenced readswere 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 numberalterations 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 asMutSigand 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 highlightedon the cover of the journalTheranosticsin 2017.

Yale University, Yale Cancer Center (YALE)

GENIE samples submitted by Yale belong to one of three targeted NGS panels (1) YALE-HSM-V1, (2) YALE-OCP-V2, or (3) YALE-OCP-V3. The first panel corresponds to the Thermofisher Ion AmpliSeq Cancer Hotspot Panel v2, which is designed to assess hotspot variants in 50 of the most frequently mutated genes in cancer, and is performed as a tumor-only analysis. The latter two panels refer to v2 and v3C of the Thermofisher Oncomine Comprehensive Assay, which provides a more comprehensive assessment of somatic alterations including single nucleotide variants, insertions, deletions, copy number alterations (CNAs), and gene fusions across 143 and 161 genes, respectively. Target region design (i.e. full exonic, hotspot only, intronic, promoter) varies based on known relevance of each gene. Pathologist inspection of an H&E section ensured adequate tumor cellularity (approximately 10% or greater). Tumor samples are enriched for malignant cells by manual microdissection of unstained formalin-fixed, paraffin-embedded (FFPE) tissue sections. If available, germline control DNA from the same patient is obtained either from FFPE non-tumor tissue, from the patient's blood, or from a buccal swab. Subsequent libraries are barcoded and sequenced on either an Ion Torrent PGM™ or an Ion S5™ XL next generation sequencer.

Pre-processing and alignment of reads is performed within Torrent Suite, with TMAP serving as the alignment algorithm. Resulting BAM files are uploaded to the Ion Reporter software for variant detection, as well as CNA and gene fusion assessment for Oncomine samples. The bioinformatics pipeline also uses MuTect2 (GATK) and Strelka (Illumina) to assess somatic variants. Variants are initially filtered based on quality metrics; a minimum read depth of 20x and avariant allelic fraction (VAF) of 0.02 is required. All variants passing quality filters are passed through the Ensembl Variant Effect Predictor for variant annotation. Variants that are intronic or synonymous are filtered at this stage; all other variants are manually reviewed for accuracy before submission to the attending pathologist. Variants below a VAF below 0.05 are not typically reviewed unless tumor cellularity estimates are low. CNA assessment is performed using the IonReporter CNV algorithm,as well as an internally developed workflow that uses the DNAcopy R package. Custom visualizations for amplified genes are used to confirm accuracy of CNAs reported by the pipeline. Only CNAs with a ploidy of 5 or higher are reported. Gene fusion assessment is handled by a custom workflow in the Ion Reporter software which aligns cDNA reads to known fusion breakpoints. A fusion read is mapped successfully if there is an overlap of 70% and exact matches of 66.66%.

Description of Data Files

Description on most of the data files can be found in the cBioPortal file formats docs.

data_mutations_extended.txt

- Description: The mutation data file expands the Mutation Annotation Format (MAF) developed by the Cancer Genome Atlas project by including additional annotations for each mutation record.
- Details: MAF format

data_clinical_patient.txt/data_clinical_sample.txt

- Description: The clinical data file is used to capture both clinical attributes and the mapping between patient and sample IDs.
- Details: Clinical format

data_CNA.txt

- Description: The copy-number data file contains values derived from copy-number analysis algorithms (e.g., GISTIC or RAE). GISTIC can be run locally or via the GISTIC 2.0 module on GenePattern. Note: some sequenced samples may not have CNA data; this file includes only samples with available copy-number profiles and does not represent all sequenced samples.
- Details: CNA format

data_sv.txt

- Description: The structural variant file contains information on structural variants. This file format replaces the deprecated data_fusions.txt file format. Note: some sequenced samples may not have SV data; this file includes only samples with available structural variant data and does not represent all sequenced samples.
- Details: Structural Variant format

genomic_information.txt

- Description: The genomic information file describes genomic coordinates covered by all platforms contributed to GENIE. This is used in the inBED filter and to generate gene panel files.
  - `Chromosome`, `Start_Position`, `End_Position`: Gene positions
  - `Hugo_Symbol`: Re-mapped gene symbol based on gene positions
  - `ID`: Center submitted gene symbols
  - `SEQ_ASSAY_ID`: The institutional assay identifier for genomic testing platform. Feature_Type: "exon", "intron", or "intergenic"
  - `includeInPanel`: Used to define gene panel files for cBioPortal.
  - `clinicalReported`: These are the genes that were clinically Reported. Blank means information not provided.

assay_information.txt

- Description: This describes the genomic profile information for each assay and is used to auto write the Summary of Sequence Pipeline section of the data guide.
- Details: This is not a cBioPortal file format
  - `SEQ_ASSAY_ID`: The assay identifier for the genomic testing platform
  - `is_paired_end`, `library_selection`, `library_strategy`, `platform`, `read_length`, `target_capture_kit`, `instrument_model`: defined by GDC read group
  - `number_of_genes`: Number of genes from which variants are called.
  - `variant_classifications`: List of types of variants that are reported for this assay.
  - `gene_padding`: Number of base pairs to add to exon endpoints for the inBED filter.
  - `alteration_types`: List of alteration types.
  - `preservation_technique`: Either FFPE, fresh_frozen, or NA.
  - `specimen_tumor_cellularity`: Tumor Cellularity Cutoff.
  - `calling_strategy`: Tumor only or tumor normal.
  - `coverage`: List of coverage types.
  - `SEQ_PIPELINE_ID`: For those centers that have multiple panels per assay (multiple `SEQ_ASSAY_ID`s)

data_cna_hg19.seg

- Description: A SEG file (segmented data; .seg or .cbs) is a tab-delimited text file that lists loci and associated numeric values. The segmented data file format is the output of the Circular Binary Segmentation algorithm (Olshen et al., 2004).
- Details: SEG format

data_gene_matrix.txt

- Description: This file contains a mapping between `SAMPLE_ID` and `SEQ_ASSAY_ID`. This assumes the type of genomic data extracted for the sample based on the associated SEQ_ASSAY_ID.
- Details: Gene Matrix format

data_gene_panel_* files

- Description: These files specify which genes are assayed on a panel and assign samples and genetic profiles (such as mutation data) to a panel.
- Details: Gene panel format

release_notes.pdf

- Description: Detailed release notes for each release.

meta_* files

- Description: Metadata files required for import into cBioPortal.
- Details: Each cBioPortal file format has an associated meta file.

case_lists files

- Description: Case lists are used to define sample lists that can be selected on the query page. Each case list file has the naming format cases_{type}.txt and is located in the case_lists folder found under each GENIE release.
- Details: Case list format

Description of Clinical Data Fields

data_clinical_patient.txt

`PATIENT_ID`

- Expected Values: GENIE-[CENTER]-[patient identifier]
- Data Description: The unique, anonymized patient identifier for the GENIE project. 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.

SEX

- Expected Values: Female, Male, Other, Transsexual, Not Collected, Unknown
- Data Description: The patient's sex code; this data element derives from NAACCR v24, Element #220.

PRIMARY_RACE

- Expected Values: Asian, Black, Native American, Not Applicable, Not Collected, Other, Unknown, Pacific Islander, White
- Data Description: The primary race recorded for the patient; this data element derives from NAACCR v24, 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 Not Collected.

SECONDARY_RACE [Not available for public releases]

- Expected Values: Asian, Black, Native American, Not Applicable, Not Collected, Other, Unknown, Pacific Islander, White
- Data Description: The secondary race recorded for the patient; this data element derives from NAACCR v24, Element #161. Institutions not collecting race have set this field to Not Collected.

TERTIARY_RACE [Not available for public releases]

- Expected Values: Asian, Black, Native American, Not Applicable, Not Collected, Other, Unknown, Pacific Islander, White
- Data Description: The tertiary race recorded for the patient; this data element derives from NAACCR v24, Element #162. Institutions not collecting race have set this field to Not Collected.

ETHNICITY

- Expected Values: Non-Spanish/non-Hispanic, Spanish/Hispanic, Unknown, Not Collected
- Data Description: Indication of Spanish/Hispanic origin of the patient; this data element derives from NAACCR v24, Element #190. Institutions not collecting Spanish/Hispanic origin have set this column to Not Collected.

BIRTH_YEAR [Not available for public releases]

- Expected Values: [Integer], Unknown, cannotReleaseHIPAA, withheld
- Data Description: The four-digit year corresponding to the patient's birth date.

CENTER

- Expected Values: The center abbreviation (e.g. MSK, DFCI, UHN)
- Data Description: The center submitting the clinical and genomic data.

`INT_CONTACT`

- Expected Values: [Integer], <6570, >32485, Not Collected, Not Released, Unknown
- Data Description: Interval in days from date of birth (DOB) to date of last contact.

`INT_DOD`

- Expected Values: [Integer], <6570, >32485, Not Collected, Not Released, Not Applicable, Unknown
- Data Description: Interval in days from date of birth (DOB) to date of death (DOD).

`YEAR_CONTACT`

- Expected Values: [Integer], <18, >89, Not Collected, Not Released, Unknown
- Data Description: Record of the year the patient is last known to be alive, as determined from electronic health records (EHR), tumor registries, or other relevant systems.

`YEAR_DEATH`

- Expected Values: [Integer], <18, >89, Not Collected, Not Released, Not Applicable, Unknown
- Data Description: Year of death.

`DEAD`

- Expected Values: TRUE, FALSE, Not Collected, Not Released, Not Applicable, Unknown
- Data Description: The patient's vital status.

data_clinical_sample.txt

`SAMPLE_ID`

- Expected Values: GENIE-[CENTER]-[patient identifier]-[sample identifer]
- Data Description: The unique, anonymized sample identifier for the GENIE project. 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.

`AGE_AT_SEQ_REPORT`

- Expected Values: [Integer], <18, >89, Unknown

- Data Description: The age of the patient at the time that the sequencing results were reported.

`AGE_AT_SEQ_REPORT_DAYS` [Not available for public releases]

- Expected Values: [Integer], >32485, <6570, Unknown
- Data Description: The interval in days between the patient's date of birth and the date of the sequencing report that is associated with the sample. The interval is masked for >32485 and <6570.

`ONCOTREE_CODE`

- Expected Values: See full list of values here.
- Data Description: The primary cancer diagnosis, or "main type", classified based on the OncoTree ontology. The version of the OncoTree ontology used for this release is oncotree_2021_11_02.

`SAMPLE_TYPE`

- Expected Values: Primary, Metastasis, Unspecified, Not Applicable or Heme, Not Collected
- Data Description: The specimen's sample type based on its location.

`SAMPLE_TYPE_DETAILED`

- Expected Values: Primary tumor, Metastasis site unspecified, Local recurrence, Lymph node metastasis, Not applicable or hematologic malignancy, Distant organ metastasis, Not Collected, Not otherwise specified
- Data Description: The specimen's detailed sample type based on its location.

`SEQ_ASSAY_ID`

- Expected Values: [Center]-[Panel]
- Data Description: 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.

`CANCER_TYPE`

- Expected Values: Non-Small Cell Lung Cancer (example). See full list of values here.
- Data Description: The primary cancer diagnosis label, or "main type", based on the OncoTree ontology. For example, the OncoTree code of LUAD maps to: "Non-Small Cell Lung Cancer".

`CANCER_TYPE_DETAILED`

- Expected Values: Lung Adenocarcinoma (example). See full list of values here.

- Data Description: The detailed primary cancer diagnosis label based on the OncoTree ontology. For example, the OncoTree code of LUAD maps to the label: "Lung Adenocarcinoma (LUAD)".

`SAMPLE_CLASS`

- Expected Values: Tumor, cfDNA
- Data Description: Specifies whether the sequenced sample is derived from tumor tissue or circulating cell-free DNA (cfDNA). cfDNA samples are publicly available starting from the 17.0-public release and onward.

`SEQ_YEAR` [Not available for public releases]

- Expected Values: [Integer]
- Data Description: The year the sample was sequenced.

Linking clinical data to genomic data

- Link between `data_clinical_patient.txt` and `data_clinical_sample.txt`:
  - Column in `data_clinical_patient.txt`: PATIENT_ID
  - Column in `data_clinical_sample.txt`: PATIENT_ID
- Link between `data_clinical_sample.txt` and `data_mutations_extended.txt`:
  - Column in `data_clinical_sample.txt`: SAMPLE_ID
  - Column in `data_mutations_extended.txt`: Tumor_Sample_Barcode
- Link between `data_clinical_sample.txt` and `data_sv.txt`:
  - Column in `data_clinical_sample.txt`: SAMPLE_ID
  - Column in `data_sv.txt`: Sample_Id
- Link between `data_clinical_sample.txt` and `genie_data_cna_hg19.seg`:
  - Column in `data_clinical_sample.txt`: SAMPLE_ID
  - Column in `genie_data_cna_hg19.seg`: ID
- Link between `data_clinical_sample.txt` and `genomic_information.txt`:
  - Column in `data_clinical_sample.txt`: SEQ_ASSAY_ID
  - Column in `genomic_information.txt`: SEQ_ASSAY_ID

Center Strategies for OncoTree Assignment

Cancer types are reported using the OncoTree ontology originally developed at Memorial Sloan Kettering Cancer Center. This release uses the OncoTree version oncotree_2021_11_02. The centers participating in GENIE applied the OncoTree cancer types to the tested specimens in a variety of methods depending on center-specific workflows. Here is a brief description of how the cancer type assignment process for each center is specified.

- CHOP: Diagnosis assigned by Pathologist, and confirmed through genomic diagnostics. Mapped to oncotree by clinical oncologist and medical geneticist.
- CRUK: Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.
- COLU: Original diagnosis from pathologist was mapped to OncoTree diagnosis by medical oncologist and research manager
- DFCI: Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type
- DUKE: Anatomic and molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.
- GRCC: OncoTree cancer types were mapped from ICD-O codes. If no ICD-O code was available, a staff scientist and an oncologist mapped the diagnosis made by the pathologist to OncoTree cancer type.
- 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.
- PROV: Pathology PhD assigns OncoTree code after reviewing patholoy report. Pathologist (MD) adjusts OncoTree code as needed prior to sign out.
- SCI: Original diagnosis from the pathology report was mapped to OncoTree diagnosis by a research coordinator and molecular pathologist.
- UCHI: The original diagnosis was mapped to OncoTree by molecular pathologists.
- UCSF: The original diagnosis was mapped to OncoTree by molecular pathologists from the Clinical Cancer Genomics Laboratory.
- UHN: The original diagnosis was mapped to OncoTree by a medical oncologist and research manager
- UMIAMI: The original diagnosis and diagnosis site was programmatically mapped to OncoTree, then validated by a medical oncologist.
- VHIO: Original diagnosis from pathologist or medical oncologist was mapped to OncoTree diagnosis by research data curator
- VICC: OncoTree cancer types were mapped from ICD-O codes. If no ICD-O code was available, a research manager mapped the diagnosis to an OncoTree cancer type.
- WAKE: Diagnoses from Foundation Medicine and Caris Diagnostics to ICD-O-3, then mapped from ICD-O-3 to Oncotree.
- YALE: Molecular pathologists assigned diagnosis and mapped to OncoTree cancer type.

Abbreviations and Acronym Glossary

For center abbreviations please see Table 1.

Abbreviation	Full Term
AACR	American Association for Cancer Research, Philadelphia, PA, USA
CNA	Copy number alterations
CNV	Copy number variants
FFPE	Formalin-fixed, paraffin-embedded
GENIE	Genomics, Evidence, Neoplasia, Information, Exchange
HIPAA	Health Insurance Portability and Accountability Act
IRB	Institutional Review Board
MAF	Mutation annotation format
NAACCR	North American Association of Central Cancer Registries
NGS	Next-generation sequencing
PCR	Polymerase chain reaction
SNP	Single-nucleotide polymorphism
SNV	Single-nucleotide variants
VCF	Variant Call Format
SV	Structural variants