AACR Project GENIE: Powering Precision Medicine through an International Consortium

The AACR Project GENIE Consortium

ABSTRACT   The AACR Project GENIE is an international data-sharing consortium focused on generating an evidence base for precision cancer medicine by integrating clinical-grade cancer genomic data with clinical outcome data for tens of thousands of cancer patients treated at multiple institutions worldwide. In conjunction with the first public data release from approximately 19,000 samples, we describe the goals, structure, and data standards of the consortium and report conclusions from high-level analysis of the initial phase of genomic data. We also provide examples of the clinical utility of GENIE data, such as an estimate of clinical actionability across multiple cancer types (>30%) and prediction of accrual rates to the NCI-MATCH trial that accurately reflect recently reported actual match rates. The GENIE database is expected to grow to >100,000 samples within 5 years and should serve as a powerful tool for precision cancer medicine.

SIGNIFICANCE: The AACR Project GENIE aims to catalyze sharing of integrated genomic and clinical datasets across multiple institutions worldwide, and thereby enable precision cancer medicine research, including the identification of novel therapeutic targets, design of biomarker-driven clinical trials, and identification of genomic determinants of response to therapy. *Cancer Discov; 7(8); 818–31. ©2017 AACR.*

See related commentary by Litchfield et al., p. 796.

INTRODUCTION

With significant decreases in the cost of sequencing, and numerous commercial and cancer center–driven initiatives, genomic profiling is increasingly becoming routine across multiple cancer types. It is expected that millions of cancer patients will have their tumors sequenced over the next decade. Nonetheless, cancer profiling efforts are frequently siloed in individual institutions, and data are frequently available only to individual researchers within a single institution or members of a paid consortium. Such exclusivity can make it difficult, if not impossible, to analyze data across multiple institutions, and severely limits statistical power when analyzing specific patient subsets, rare cancer types, or rare variants across multiple cancer histologies. Broad-based sharing of genomic and clinical data is therefore critical to realize the full potential of precision oncology (1), particularly as the scientific community evaluates the overall impact of genomic profiling on patient outcome and on clinical trial enrollment (2–5), and as the clinical community better leverages "big data" and machine learning approaches to improve patient care (6, 7). Several "big data" initiatives, including the Genomics, Evidence, Neoplasia, Information, Exchange (GENIE) project described here, have been launched in recent years to address the challenges of large-scale sharing of genomic and clinical data and to accelerate progress in identifying both effective and ineffective therapies to treat cancer

(8). Indeed, data sharing emerged as a top priority of the recent Blue Ribbon Panel report from the National Cancer Institute, in response to the Cancer Moonshot initiated in 2016 by then Vice President Joe Biden, underscoring the urgency to make real progress (9).

Recognizing the immediate and urgent need for broad data sharing across cancer centers and with the wider scientific community, the American Association for Cancer Research (AACR) in partnership with eight global academic leaders in clinical cancer genomics (Table 1) initiated the AACR Project GENIE. The AACR Project GENIE is a multiphase, multiyear, international data-sharing project that aims to catalyze precision cancer medicine (Box 1; Fig. 1). The GENIE platform is built to integrate and link clinical-grade cancer genomic data with clinical outcomes data for tens of thousands of cancer patients treated at multiple institutions worldwide.

Corresponding Authors: Ethan Cerami, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215. Phone: 617-632-6440; E-mail: cerami@jimmy.harvard.edu; and Charles L. Sawyers, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065. E-mail: sawyersc@mskcc.org

Table 1. Founding members of the GENIE consortium

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

BOX 1. GOALS OF THE AACR PROJECT GENIE

AACR Project GENIE is a multiphase, multiyear, international data-sharing project that aims to catalyze precision oncology by:

- Sharing integrated clinical-grade genomic and clinical data across multiple U.S. and international cancer centers.
- Making all deidentified data publicly available to the entire scientific community.
- Developing harmonized standards for sharing genomic and clinical data.
- Initiating new translational research projects, which specifically leverage the depth and breadth of data available across GENIE consortium members.

The project fulfills an unmet need in oncology by providing a data-sharing platform to enable scientific and clinical discovery, including the identification of novel therapeutic targets, design of new biomarker-driven clinical trials, and deeper understanding of patient response to therapy. Ultimately, the platform can improve clinical decision-making and increase

the likelihood that the cancer treatments patients receive are beneficial. At the societal level, this approach has immense potential to maximize the value of care delivery.

GENIE Consortium

The primary focus of GENIE is to link genotypes with clinical phenotypes and make such data widely available to the entire scientific community. The database currently contains approximately 19,000 genomic and clinical records generated in Clinical Laboratory Improvement Amendments (CLIA)/ International Organization for Standardization (ISO)–certified laboratories obtained at multiple international institutions, and will continue to grow as additional patients are treated at each of the participating centers and as more centers join the consortium. Data from the first 19,000 patients were released to the scientific community on January 5, 2017 (10). Because each of the current participating centers is a tertiary referral center within its community, the platform is enriched in samples of late-stage disease.

Each of the participating centers has extensive clinical data characterizing individual patients via Electronic Health Record (EHR) systems, and GENIE is therefore uniquely positioned to integrate genomic data with clinical data and harmonize such data across multiple cancer centers. To accomplish this, the consortium members have defined
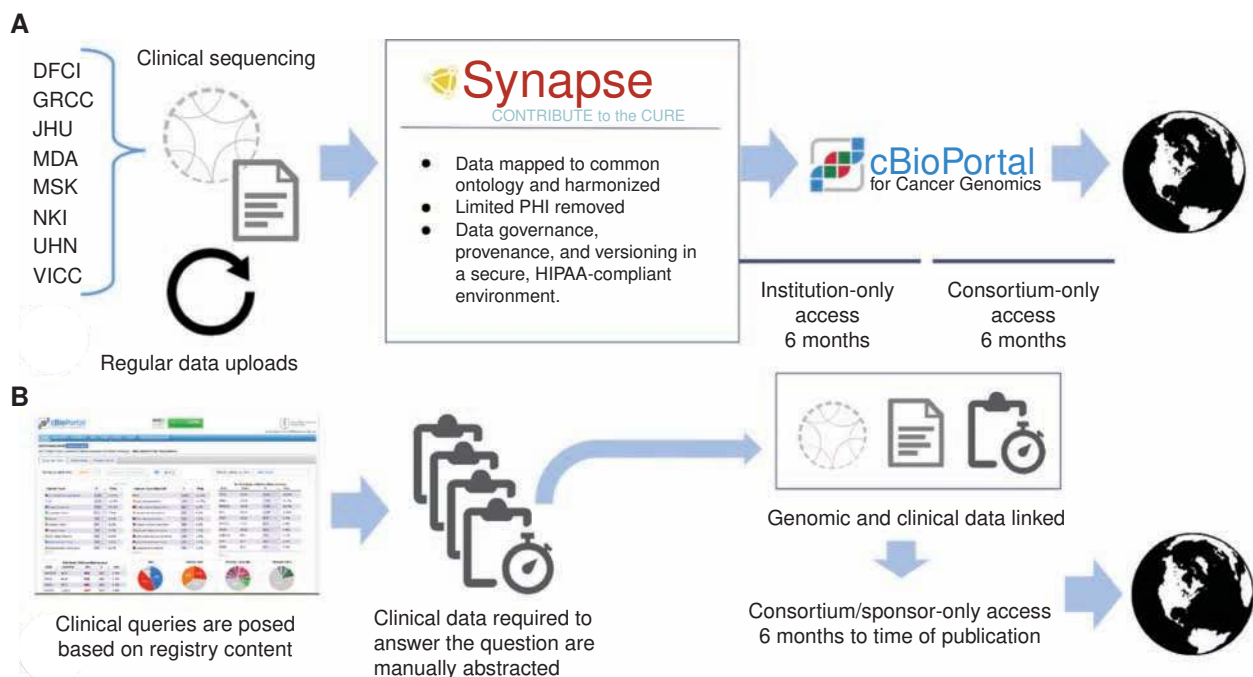

Figure 1. AACR Project GENIE at a glance. **A,** Variant calls and a limited clinical dataset from patients treated at each of the participating centers are sent to the Synapse platform, developed by Sage Bionetworks, where the data are harmonized and protected health information (PHI) removed in a secure Health Insurance Portability and Accountability Act (HIPAA)–compliant environment that provides data governance. Once harmonized, these data are viewed and analyzed in the cBioPortal for Cancer Genomics. Value is provided to both the data generators and the consortium by establishing 6-month periods of exclusivity to each prior to the data becoming available to the broader research community. **B,** Once data are available in the cBioPortal, clinical research projects are proposed and vetted by the project steering committee. Clinical teams are then assembled to define the clinical attributes required to answer the approved research question; these data are then manually curated from the relevant medical records and deposited in an electronic data capture system. The detailed clinical data are then transferred to Synapse where they are linked with the appropriate genomic and limited clinical data and are viewable and analyzable in the cBioPortal platform. Again, value is created by providing a period of at least 6 months' exclusivity to both the consortium and sponsors, where relevant. The primary data are made public at the time of publication.

a parsimonious set of harmonized clinical data elements and outcome endpoints. The GENIE platform will enable researchers to better understand clinical actionability across cancer types, assess the clinical utility of genomic sequencing, define clinical trial enrollment rates to genotype-specific clinical trials, validate genomic biomarkers, reposition or repurpose already-approved drugs, expand existing drug labels by addition of new mutations, and identify new drug targets. Importantly, researchers will also be able to compare and cross-validate the clinically derived datasets generated by GENIE with other publicly available datasets, including The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC; ref. 11).

An essential component to assembling a functional consortium is to provide the infrastructure, funding, and governance necessary to operate as a unified entity. In the case of GENIE, the AACR fulfills these roles not only as a trusted third party, but also as an active participant. The consortium is assembled through two legal constructs: a master participation agreement (MPA) and a data use agreement (DUA). The MPA also requires that each institution share data in a manner consistent with patient consent and center-specific Institutional Review Board (IRB) policies. The exact approach varies by institution, but largely falls into one of three categories: IRB-approved prospective patient consent to sharing; retrospective IRB waivers; and IRB approvals of GENIE-specific research proposals.

Patient consents within member institutions of GENIE enable data sharing for research purposes, and deidentified GENIE data have therefore been made available to the entire scientific community (including academic institutions, government agencies, and industry). Further, deidentified data generated by GENIE-sponsored research projects (see below) are not exclusive to the commercial sponsors and will also be shared with the entire scientific community.

To enable broad-based sharing, the AACR GENIE project has partnership agreements in place with Sage Bionetworks and the cBioPortal for Cancer Genomics, both of which have significant prior experience in similar projects and have developed established and accepted data-sharing platforms within the community. The Synapse platform from Sage specifically provides a secure, Health Insurance Portability and Accountability Act (HIPAA)–compliant infrastructure that enables data versioning and provenance (12), and the cBioPortal provides visualization and analysis features for exploring large-scale, deidentified cancer genomic datasets (13, 14).

Data-Sharing Principles

The data-sharing principles of GENIE are designed to enable a scalable informatics infrastructure for integrating and sharing genomic and clinical outcomes data with specific safeguards to maintain patient privacy. Following IRB approval at each member institution, genomic and clinical data are submitted to Sage Bionetworks through a secure web-based platform (Synapse). Genomic data include high-confidence variant calls for single-nucleotide variants, insertions/deletions (indels), copy-number variations, and structural changes (when available) for tumor sequencing.

Use of clinical-grade genomic sequencing data generated in CLIA/ISO-certified and experienced molecular pathology laboratories ensures high-quality variant calls without the need

for reanalysis. As sequencing of tumor specimens without matched normal tissue may result in identification of germline alterations (15), a stringent germline filtering pipeline is applied to all mutation data, to minimize risk of patient reidentification (see Supplementary Methods and Supplementary Fig. S1). Metadata are captured and versioned for every genomic record and include information regarding the sequencing platform and analytic pipeline used for variant calling. All identifiable protected health information (PHI) is removed via the HIPAA Safe Harbor method and a deidentified dataset is made available at Synapse (https://synapse.org/genie) and the cBioPortal for Cancer Genomics (http://cbioportal.org/genie/).

Individual member institutions are provided with exclusive access to their institutional data for 6 months followed by an additional 6-month period for controlled member-only access to the entire consortium dataset, providing opportunity for analysis and publication before the wider public release. Member institutions maintain exclusive intellectual property (IP) rights to all data provided to the consortium. The GENIE consortium is also committed to further sharing of data via the NCI Genome Data Commons (GDC; ref. 1) and the Cancer Gene Trust, led by the Global Alliance for Genome and Health (GA4GH).

Data Standardization

To ensure consistency across centers, all members of the GENIE consortium have agreed to core data elements, data definitions, and ontologies. For genomic data, all centers provide mutation data in MAF or VCF format, and all centers are required to provide BED files for each assay panel reported. To comply with patient consent agreements at each institution and to ensure patient privacy, raw BAM files are not shared within GENIE, but each center's clinical sequencing pipeline is described within the GENIE Data Guide (Supplementary File S1), enabling researchers to more carefully compare datasets across centers.

For clinical data, core patient-level and specimen-level data elements have been identified and defined. This comprises a set of minimum clinical data attributes (tier 1), which includes sex, race, ethnicity, birth year, age at sequencing, primary cancer diagnosis, and sample type (primary/metastatic). Primary cancer diagnosis is reported using the OncoTree cancer type ontology, initially developed at MSK, which also provides mappings to other widely used cancer type taxonomies, including SNOMED and ICD-9/10 codes (16). Additional clinical information, including prior therapies, overall survival, and disease-free survival, is being defined, and the consortium is currently evaluating the feasibility of extracting such data for all patients and specific subsets of patients. As the project evolves, strategies for automated extraction of clinical outcome data from electronic medical records at member institutions will be developed, including curation and remapping of data attributes where required.

Landscape of the First Integrated GENIE Consortium Cohort

The first integrated GENIE dataset (version 1.1) provides genomic and limited clinical data for 18,804 genomically profiled samples across 18,324 patients at 8 academic medical
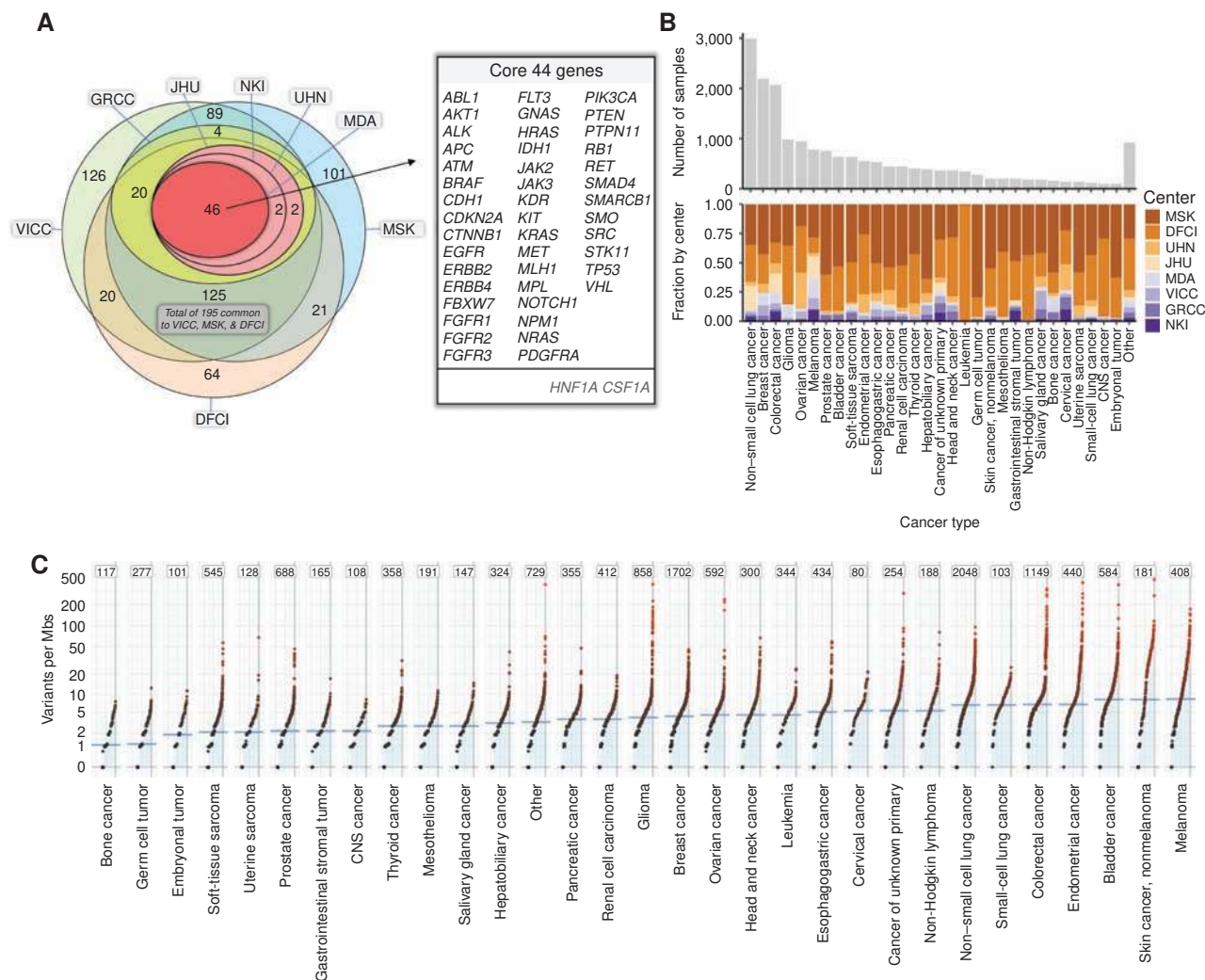
Figure 2. Landscape overview of GENIE dataset. **A,** The degree of overlap at the gene level across the contributing centers' genomic assays is shown. A core set of 44 genes (listed in the inlay) is represented across all genomic assays in the GENIE dataset. The 2 additional genes listed in the bottom right of the inlay in gray are genes that were common to the smaller panels, not present in some of the previous versions of the larger panels but are present on the most recent version of all panels. **B,** Total sample counts by tumor type and contributing center. The contribution of samples for each tumor type across the institutions in shown within each bar of the lower stacked barplot. **C,** Mutations (all nonsilent substitutions and small insertions/deletions reported) per coding megabase (Mbs) sequenced for each sample, stratified by tumor type, and ordered by median mutation rate in those tumor types. The data are shown as empirical cumulative distributions (blue-shaded area) with individual samples shown as points colored black to red for low to high mutation burden, respectively. These data are limited to the 14,310 samples analyzed by the larger gene panels used at centers DFCI, MSK, and VICC.

Genomics Overview

centers, each of which utilized genomic strategies tailored to best support their local clinical programs. These strategies include highly targeted, amplicon-based panels covering mutation hotspots from approximately 50 genes, designed to cover current clinically actionable mutations and clinical trials, as well as broader, custom panels (275–429 genes) utilizing hybrid-capture to isolate all exons and some introns to support discovery as well as clinical research projects. In addition, each center's approach has evolved, such that the GENIE dataset contains 12 different gene panels that were used in at least 50 samples. A total of 44 genes were included on all 12 of these panels. The larger hybrid-capture gene panels included all of the genes on the smaller gene panels and added 125 genes common to all of the larger panels and an additional 134 genes common to up to 2 of these larger panels (Fig. 2A).

Genomic data within GENIE include mutation data (all centers), copy number (three centers), and structural rearrangement data (two centers). Two centers implemented paired tumor/normal sequencing, whereas all other centers conducted tumor-only sequencing (Supplementary Tables S1 and S2). The majority of the samples came from MSK ($n = 7,341$) and DFCI ($n = 6,137$), with the other 6 institutions each contributing 505 to 1,296 samples. Clinical data are currently limited to cancer types based on the OncoTree ontology, whether the material sequenced came from a primary or metastatic tumor (if known), age at date of genomic sequencing, sex, and race. The complete dataset can be downloaded from the Sage Synapse platform (https://synapse.org/genie) or visualized via the cBioPortal for Cancer Genomics (http://cbioportal.org/genie/).

The spectrum of tumor types across the consortium is shown in Fig. 2B. The most highly represented tumor types across the GENIE consortium tend to be those where genomic data are currently used to guide standard treatment decisions, such as non–small cell lung cancer ($n = 2,985$) and colorectal cancer ($n = 2,081$) along with melanoma ($n = 785$). The contributing institutions of the GENIE consortium also had varying approaches to patient selection for genomic profiling. For example, some centers performed genomic profiling on all patients and all cancer types, whereas others have chosen to focus on only a few select tumor types (Fig. 2B). Three sites, DFCI, MSK, and VICC, submitted 14,310 samples with sequencing data from relatively large, 275+ gene panels which we used to investigate the per-sample mutational burden (Supplementary Fig. S2).

As previously shown (17), plotting the distribution of the number of mutations/Mb for each sample by tumor type (Fig. 2C) demonstrates a wide variance of mutation rates both within and among tumor types. As expected, tumors with strong mutagenic backgrounds such as melanoma and non–small cell lung cancer have a high median mutation burden across the centers. Endometrial cancers and colorectal carcinomas have a wide within-tumor mutation burden distribution, reflecting the inclusion of both MSI/POLE-positive and MSI/POLE-negative patients. Some surprises were also identified, perhaps due to the uniqueness of this dataset; for instance, we believe the wide distribution of mutation burden in glioma, which has not been seen previously, is likely due to inclusion of patients who received temozolomide.

Although a rigorously defined cut-off for a mutation burden that will respond to checkpoint inhibition or other immune modulation has not been identified (18), the GENIE data demonstrate that almost all tumor types have at least some samples with a mutation burden above the 90th percentile of all samples tested on the larger sequencing panels (12.3 mutations per Mb). This includes carcinomas of unknown primary, of which 17% are in the top 10% of all samples tested on the larger sequencing panels. Carcinomas of unknown primary currently present clinical quandaries, and the relatively large proportion of samples with high mutation burden suggests that checkpoint inhibition may be variably, but widely, applicable in many cancer types, including some difficult to treat tumors.

Concordance across GENIE Institutions

Despite the differences by which the eight contributing centers implemented genomic testing of these tumors, the results from the top three most prevalent tumor types in the GENIE dataset (Fig. 3A–C) were largely concordant across centers. The smaller targeted amplicon-based gene panels (assays from MDA, NKI, UHN, JHU, and GRCC) detected the majority of the higher frequency mutations, whereas the larger gene panels (assays from DFCI, MSK, and VICC) detected multiple additional genes with mutations that occurred at lower frequencies. Importantly, the clinical benefit associated with detecting these genomic alterations is not necessarily related to the frequency of the genomic alteration, as can be seen for example with *ALK* rearrangements which occur in only 3% to 7% of non–small cell lung cancer but are of high clinical importance. Furthermore, the larger panels in aggregate detected approximately 500 more genes with lower

frequency alterations (beyond for example what is shown in Fig. 3A–C) that may prove to be of high clinical value in the future (see Supplementary Table S3). In addition, gene panels that differ in the fraction of coding regions sequenced in a given gene can lead to different conclusions. For example, a decreased number of *APC* mutations is observed in colorectal cancer when a smaller panel is used due to the limited regions analyzed of the *APC* gene, 532–1,367 base pairs (bps) for the smaller amplicon panels as compared with 8,622–8,936 bps for the larger gene panels that covered all coding gene regions (Fig. 3C).

Comparison with TCGA

The gene mutation rates across centers in the entire GENIE dataset are comparable with those reported by TCGA and other databases for the tumor types examined (Fig. 3A–C; refs. 19–22). However, some important differences are evident. In particular, the GENIE dataset has an increased prevalence of *EGFR* mutations in the context of non–small cell lung cancer compared with TCGA (19), likely driven by the referral of *EGFR*-mutant patients to the large academic centers of the GENIE consortium for clinical care and potential clinical trials. In support of this supposition, when we examined the specific *EGFR*-mutant variants observed in the GENIE dataset in comparison with TCGA, we observed that *EGFR* p.T790M represented 11.3% (83/737) of *EGFR* mutations in the GENIE dataset but only 2.2% (3/137) of *EGFR* mutations in TCGA. This is most likely due to an increased proportion of recurrent/relapsing tumors in the GENIE dataset as compared with TCGA.

We also systematically compared mutation hotspot frequencies in GENIE with those from cancerhotspots.org, a dataset derived from TCGA (ref. 23; Supplementary Figs. S3 and S4). In this analysis, a binomial distribution test for each hotspot found an enrichment for *KRAS* p.G12 mutations in the GENIE cohort, likely indicative of a higher fraction of patients with late-stage, metastatic disease, and a different distribution of tumor types. Although hotspot mutation frequencies in GENIE are similar to those reported in cancerhotspots.org, the exact prevalence of lower frequency variants will require increased sample numbers, which will be facilitated by participation of additional centers in the GENIE project.

Finally, GENIE data exhibit similar patterns of mutual exclusivity observed in TCGA. For example, in non–small cell lung cancer, mutations in *KRAS* (27%) are mutually exclusive of mutations in *EGFR* (19%; $P < 0.001$); in breast cancer, *PIK3CA* mutations (35%) are mutually exclusive of *AKT1* (5%) and *PTEN* (8%) mutations ($P = <0.001, 0.03$); and in colorectal cancer, *KRAS* mutations (47%) are mutually exclusive of *BRAF* mutations (11%; $P < 0.001$).

Assessing Clinical Actionability: For Treatment Decisions with Approved Drugs and for Clinical Trial Eligibility

Recent commentaries have questioned the clinical utility of matching patients to drugs based on tumor molecular profiling (3–5), largely based on the low frequency of patients matched to current targeted therapy trials and a lack of data from clinical trials assessing the added benefit of molecular profiling. Our collection of genomic and clinical records
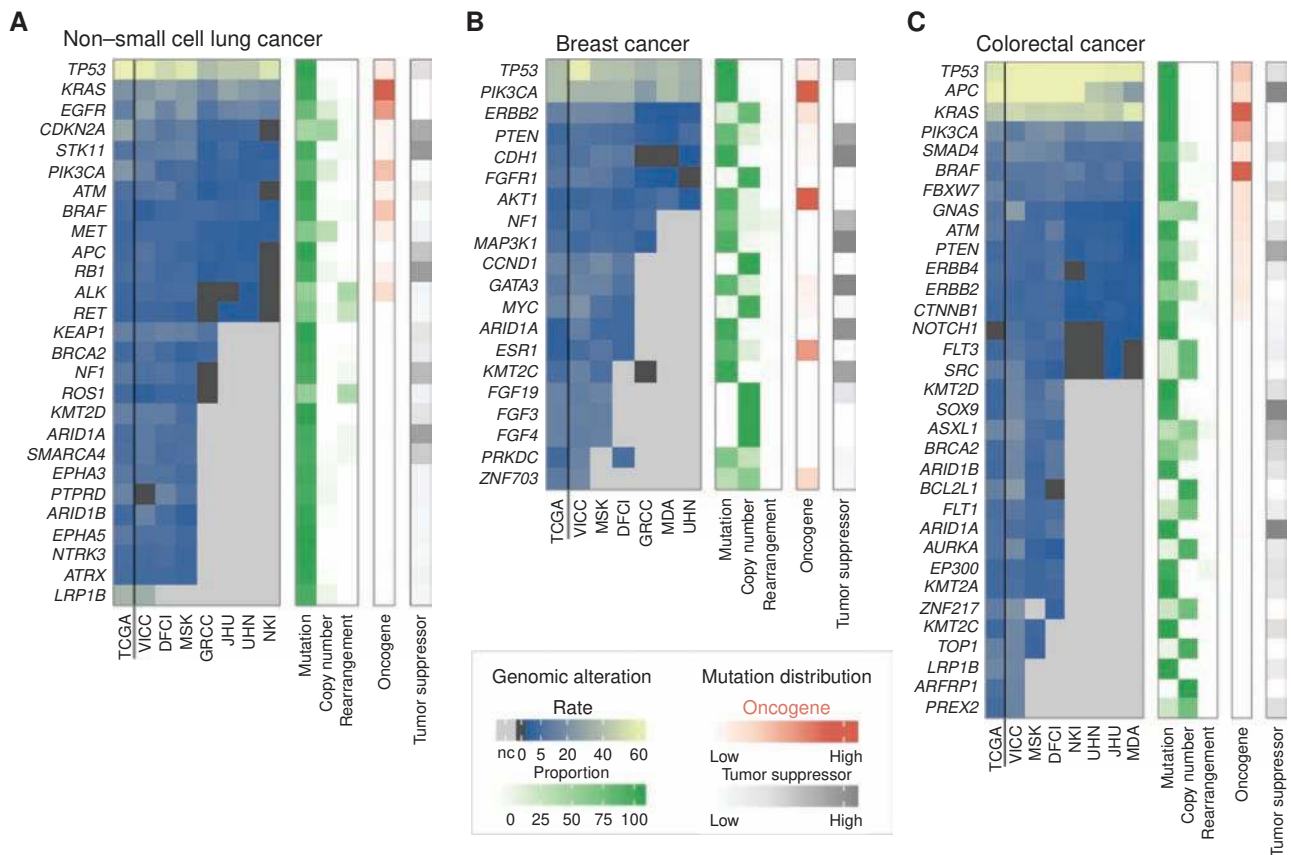
Figure 3. Genomic alterations in non–small cell lung cancer, breast cancer, and colorectal cancer. **A–C,** The genomic alteration rate (including mutation, copy number, and rearrangement) aggregated to the gene level across the cohort for the top three most common tumor types is shown: non–small cell lung cancer, colorectal cancer, and breast cancer (**A–C,** respectively). Data for each center are shown as percentage of samples from that center with genomic alterations in a given gene. Directly adjacent to the main heat map is the proportional breakdown of the types of genomic alterations observed, and characterization of the mutation distribution observed in a given gene as oncogene and tumor suppressor, based on the normalized entropy ($\log_2(N)$-$\sum p_i \log_2(p_i)$, where N is the number of unique mutations in a given gene and p_i is the proportion of mutations accounted for by a given unique mutation of a given gene) in the mutation spectrum and the prevalence of truncating and frameshift mutations, respectively. These data are limited to the gene with either: (i) 15% genomic alteration rate in at least one center, (ii) 5% genomic alteration rate in at least three centers, and (iii) OncoKB level 1 or 2A evidence for the tumor types shown. The "nc" designation in the colorbar legend indicates no coverage.

from nearly 19,000 cancer patients provides a large dataset to begin to address these questions.

To determine the frequency of potentially actionable mutations across tumor types, we mapped all mutations to variant interpretations merged from three knowledge bases: My Cancer Genome (http://mycancergenome.org), OncoKB (24), and Personalized Cancer Therapy (http://pct.mdanderson.org). A total of 7.3% of tumors in GENIE contained a Level 1 or 2A alteration indicative of treatment with an FDA-approved drug or standard care, as defined by the National Comprehensive Cancer Network (NCCN) or other guidelines (Fig. 4). An additional 6.4% of tumors contained Level 3A alterations, i.e., those with clinical evidence for response to investigational therapies in the same disease (Fig. 4; see Supplementary Table S4 for all Level 1–3 annotations). Furthermore, 6.7% of tumors had Level 2B alterations (alterations that are Level 1 or 2A in other tumor types), and 11.1% had Level 3B alterations (Level 3A in other tumor types).

Collectively, this suggests an overall potential actionability rate >30%. These frequencies varied widely across disease, from

highly recurrent and druggable mutations in gastrointestinal stromal tumor (GIST; 66%, almost all Level 1 or 2A mutations of *KIT* and *PDGFRA*) to tumor types with few actionable alterations, such as renal cell, prostate, or pancreatic cancer. Breast cancer is the disease with the highest fraction of patients who might benefit from existing investigational targeted therapies (Level 3A), due to frequent mutations of *AKT1, ERBB2*, and *PIK3CA*, accounting for 38% of patients. We anticipate one of the benefits of GENIE will be an increased power for delineating the clinical significance of somatic mutations (particularly new indications for approved drugs) as well as data-driven selection of high-yield tumors likely to contain actionable mutations for clinical trials.

To evaluate the potential for using GENIE data for assessing clinical trial feasibility and theoretical match rates, we curated somatic mutations as biomarker inclusion criteria for 18 of the 24 substudies that comprise the NCI-MATCH trial (Supplementary Methods; Fig. 5A and B). Using these criteria, 2,516 patients matched 2,885 times against 17 of 18 substudies within NCI-MATCH. We then compared
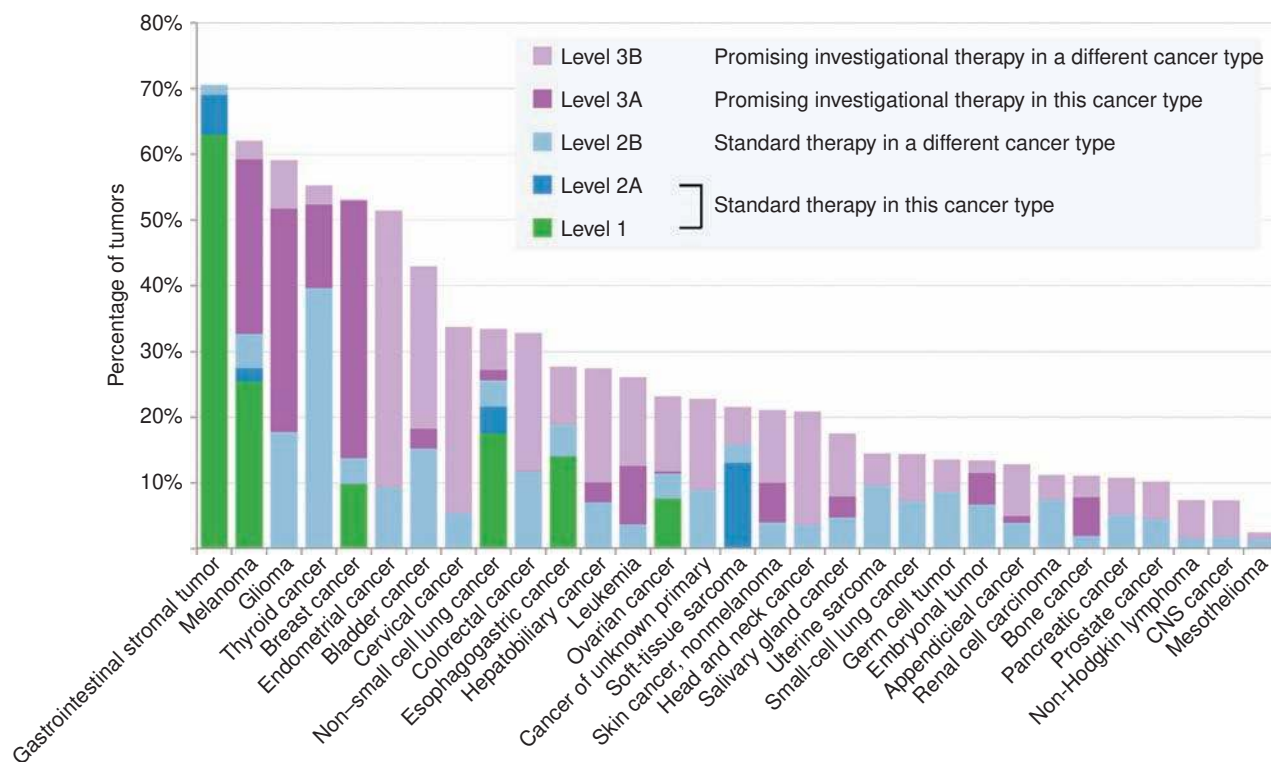
Figure 4. Potential clinical actionability. Tumor types are shown by decreasing overall frequency of actionability. Actionability was defined by the union of three knowledge bases: My Cancer Genome (http://mycancergenome.org), OncoKB (http://oncokb.org), and the Personalized Cancer Therapy knowledge base (http://pct.mdanderson.org). For each tumor sample, the highest level of actionability of any variant was considered. Only tumor types with 100 or more samples were included in this analysis.

these theoretical match rates against real-world match rates reported by an interim analysis (25) of 645 patients profiled for the NCI-MATCH trial (Fig. 5C). Outside of substudies S1 (*NF1* inactivating mutations) and Z1B (amplifications of *CCND1/2/3*), there was high concordance between real-world NCI-MATCH and theoretical GENIE match rates ($P < 10^{-4}$, $P = 8.1 \times 10^{-4}$ with two outlier trials included). This concordance demonstrates the utility of the GENIE cohort to accurately forecast genomic match rates and to serve as a valuable tool to guide design of new clinical trials as the dataset grows. Furthermore, substudies A and S2 had zero reported matches by the interim NCI-MATCH analysis, whereas the larger GENIE cohort observed 7 (A) and 11 (S2) matches for each (~0.1% match rate). Overall, the GENIE cohort will only grow in power as additional data are added to the knowledge base, enabling similar comparisons with ongoing clinical trials.

Translational Research Projects

As new medicines are developed to treat small, well-defined patient subpopulations harboring specific genetic variants, clinical trial design has shifted from randomized trials to single-arm studies wherein all eligible patients receive the study drug. In this context, it is beneficial for study sponsors to understand the natural history of the disease in patients with the genetic variant who are naïve to the study drug in comparison with those patients lacking the variant. These are

research studies that GENIE is uniquely positioned to enable, as one can use the GENIE platform to identify genomically defined patient cohorts and then return to the respective EHRs to curate the detailed clinical data necessary to answer important medical questions about the population under study (Fig. 1B).

To date, GENIE has successfully entered into two sponsored research agreements to provide the analysis for two rare populations in breast cancer, the platform's second largest cohort with approximately 2,200 samples. The first of these studies seeks to define the clinicopathologic features and outcomes of patients with metastatic breast cancer harboring known pathogenic variants in *ERBB2* as compared with *ERBB2* wild-type patients. A second study is examining similar parameters in *AKT1* E17K mutant metastatic breast cancer. *ERBB2* and *AKT1* E17K mutations are relatively rare in breast cancers, and it is only by pooling samples across multiple institutions that such studies are feasible. In addition to potentially accelerating the pace of drug approval, sponsored studies are a critical mechanism for covering the costs associated with data-sharing projects, because such research efforts are not typically supported by traditional grant mechanisms.

Lessons Learned and Future Challenges

The long-term goal of GENIE is to create a large, high-quality clinical cancer genomics database and to make it
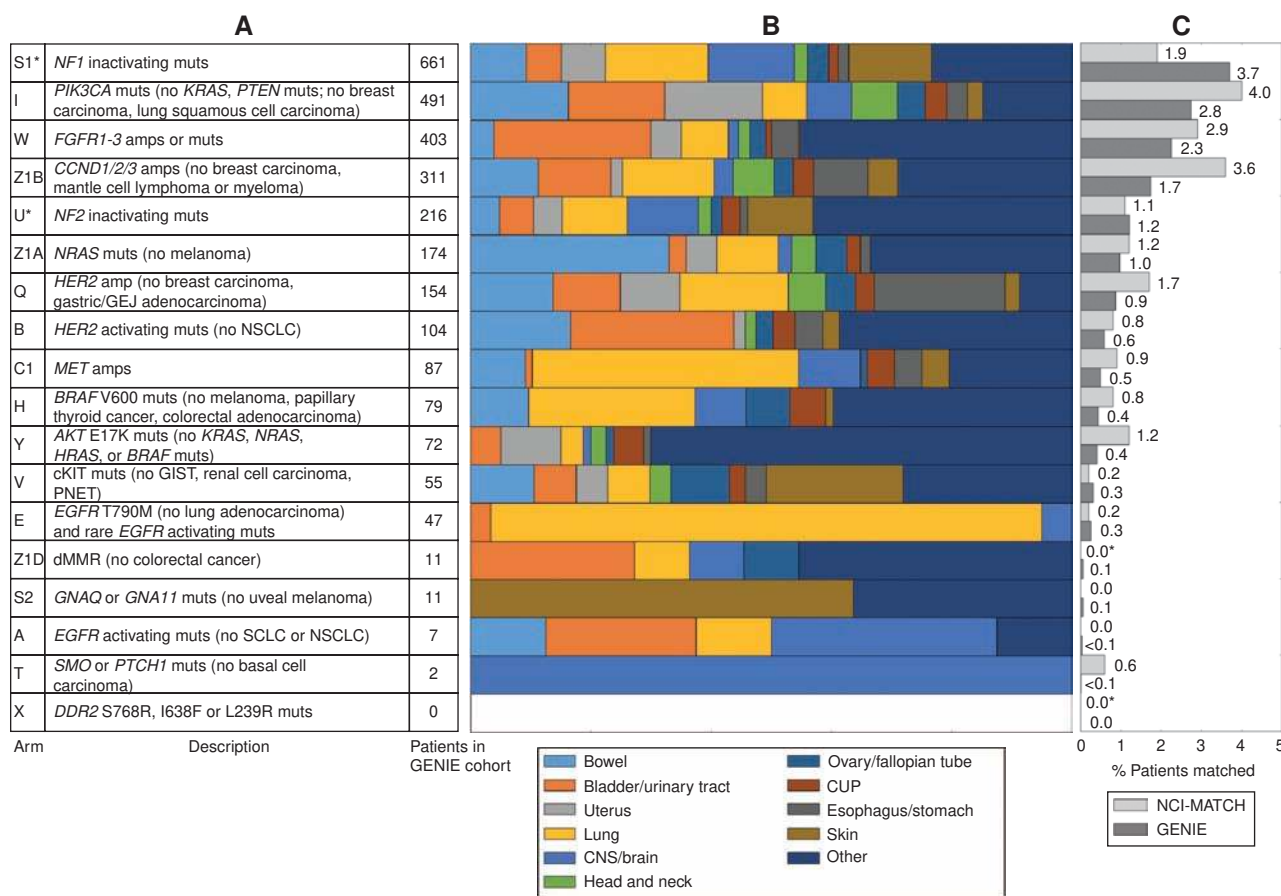
Figure 5. Clinical trial matching. Overview of GENIE samples matched to NCI-MATCH, based on genomic and cancer type criteria. Each patient with a reported sequencing date in 2014 or later was matched against 18 arms of the study that use somatic mutations or copy-number alterations for enroll-ment. Arms with fusion criteria were excluded because only two of the eight contributing GENIE centers provided fusion data. **A,** Information regarding 18 arms of the NCI-MATCH trial, including a summary of genomic trial eligibility, and the total count of GENIE samples matched. For arms S1 and U (indicated with an asterisk), the exact set of inactivating mutations was not specified in the NCI protocol, and all mutations were therefore considered matches. **B,** Proportion of the matches attributed to the top 10 most frequently matched cancer types. The categories are the top-level OncoTree codes. **C,** Comparison of the observed matching rate in the GENIE cohort with the reported rates observed by the first 645 patients by the NCI-MATCH group. Substudies X and Z1D had not reported interim rates.

widely accessible to the global cancer research community. In doing so, GENIE aims to catalyze precision medicine research across the entire cancer community, providing critical infra-structure for a "learning healthcare system," capable of using integrated genomics and clinical data to improve patient out-comes (26). Under this broad vision, GENIE specifically aims to spur the development of clinical and genomic data stand-ards, promote best practices for clinical genomic sequencing, and further encourage broad-based sharing between cancer centers. In developing standards and collaborating with other initiatives, including the NCI GDC and the GA4GH, GENIE further aims to be a critical component of the recently pro-posed Cancer Moonshot National Cancer Data Ecosystem (9), ensuring that cancer data are widely shared across the entire scientific community.

We have intentionally described the organizational prin-ciples of the GENIE consortium in this report, as well as an initial analysis of the first ~19,000 patients, to provide lessons for other institutions contemplating similar data-

sharing efforts. One significant barrier to participation in such consortia is concern about protecting patient privacy. This was largely overcome by adhering to HIPAA safe harbor deidentification policies, developing a unified germline filter-ing pipeline, and making data available under specific terms of access, which prohibit patient reidentification and data redistribution. GENIE has also adopted a "federated model," whereby the primary genomic and clinical data reside at the participating institution with the agreement that additional data elements can be accessed by the consortium in response to specific queries. This model also alleviates concerns of local investigators about compromising their individual academic interests, through controlled access to longitudinal clinical data as well as defined periods of institutional exclusivity within the consortium.

Another valuable outcome was agreement on standards for harmonizing genomic and clinical data elements from differ-ent platforms, different electronic medical record systems, and different countries. For example, after much discussion,

we converged on the use of OncoTree cancer type taxonomy (rather than ICD or SNOMED) as a preferred method for histopathologic classification of tumors. In addition, the decision to integrate genomic data from panels varying in size from approximately 50 to 500 genes allowed us to incorporate larger numbers of patients across a much broader geographic spectrum than would have been possible with a common platform. This decision comes with obvious limitations, particularly for new target gene discovery, but allowed us to assemble the largest database of its kind that, we hope, will serve as an evidence base for assessing clinical actionability.

A particular challenge for precision oncology is the need for large patient populations to provide sufficient evidence of clinical utility for genomic testing. Indeed, this is a major goal of the GENIE consortium. However, novel discoveries of clinical actionability require, by definition, surveys of large numbers of genes across large numbers of patients. The reluctance of insurance payers to cover large panel sequencing (> 50 genes), with rare exceptions, places the field in a "Catch-22." In the absence of such evidence, there is no coverage of expenses by payers, but in the absence of payer coverage, there will be no evidence generated. Even if this issue were resolved (i.e., through "coverage with evidence development" programs; ref. 27), there is the additional challenge of collecting the associated longitudinal clinical outcomes. GENIE is in the initial phase of generating such data in subsets of patients with defined genomic alterations, but we are facing the challenge of covering the costs associated with clinical curation. Despite the promise of inexpensive, automated curation technology such as natural language processing, manual curation remains the gold standard today, particularly for regulatory-grade registry data. That said, the data curation field is evolving rapidly, and successful application of less expensive, automated technologies in efforts like GENIE could be catalytic for precision medicine. But we will get there only through organized, responsible data-sharing efforts.

Following completion of this initial public release, GENIE is now soliciting membership from other academic and research institutions. Membership in GENIE is open to academic medical centers and research institutions that can contribute at least 500 unique clinical and genomic records generated by CLIA/ISO-certified or equivalent clinical sequencing laboratories per year across multiple cancer types, with the ability to perform curation of clinical data including treatment and outcomes. This will enable the inclusion of additional cancer types that are not well represented in the initial data release, such as pediatric, hematologic, and rare malignancies, as well as inclusion of data from additional international partners.

Based on yearly rates of sequencing at each of the eight founder institutions, the GENIE database is expected to grow by approximately 16,000 samples per year. But, with the addition of new members, it is likely that the GENIE database will grow to >100,000 samples within 5 years. With recent technological advances, we also anticipate that future releases of GENIE data will be enriched for large, targeted DNA-sequencing panels that characterize further sources of genomic variation, including new structural rearrangements and promoter mutations, and integration of additional genomic platforms, including whole-exome and whole-genome DNA sequencing, transcriptome sequencing, methylation, proteomics, and immunoprofiling. In addition, analyses of circulating tumor DNA or circulating tumor cells from blood specimens or other bodily fluids (28) may be included to identify molecular changes in cancer genomes at the time of diagnosis or during therapy as these analyses become included in routine laboratory practice.

Alphabetical List of Authors

Last name	First name	M	Institution
André	Fabrice		Institut Gustave Roussy
Arnedos	Monica		Institut Gustave Roussy
Baras	Alexander	S.	Sidney Kimmel Cancer Center at Johns Hopkins University
Baselga	José		Memorial Sloan Kettering Cancer Center
Bedard	Philippe	L.	Princess Margaret Cancer Centre, University Health Network
Berger	Michael	F.	Memorial Sloan Kettering Cancer Center
Bierkens	Mariska		Netherlands Cancer Institute
Calvo	Fabien		Institut Gustave Roussy
Cerami	Ethan		Dana-Farber Cancer Institute
Chakravarty	Debyani		Memorial Sloan Kettering Cancer Center
Dang	Kristen	K.	Sage Bionetworks
Davidson	Nancy	E.	Fred Hutchinson Cancer Research Center
Del Vecchio Fitz	Catherine		Dana-Farber Cancer Institute
Dogan	Semih		Institut Gustave Roussy
DuBois	Raymond	N.	Medical University of South Carolina
Ducar	Matthew	D.	Dana-Farber Cancer Institute and Brigham and Women's Hospital

(continued)

Alphabetical List of Authors (*Continued*)

Last name	First name	M	Institution
Futreal	P. Andrew		UT MD Anderson Cancer Center
Gao	Jianjiong		Memorial Sloan Kettering Cancer Center
Garcia	Francisco		UT MD Anderson Cancer Center
Gardos	Stu		Memorial Sloan Kettering Cancer Center
Gocke	Christopher	D.	Sidney Kimmel Cancer Center at Johns Hopkins University
Gross	Benjamin	E.	Memorial Sloan Kettering Cancer Center
Guinney	Justin		Sage Bionetworks
Heins	Zachary	J.	Memorial Sloan Kettering Cancer Center
Hintzen	Stephanie		Dana-Farber Cancer Institute
Horlings	Hugo		Netherlands Cancer Institute
Hudeček	Jan		Netherlands Cancer Institute
Hyman	David	M.	Memorial Sloan Kettering Cancer Center
Kamel-Reid	Suzanne		Princess Margaret Cancer Centre, University Health Network
Kandoth	Cyriac		Memorial Sloan Kettering Cancer Center
Kinyua	Walter		UT MD Anderson Cancer Center
Kumari	Priti		Dana-Farber Cancer Institute
Kundra	Ritika		Memorial Sloan Kettering Cancer Center
Ladanyi	Marc		Memorial Sloan Kettering Cancer Center
Lefebvre	Céline		Institut Gustave Roussy
LeNoue-Newton	Michele	L.	Vanderbilt-Ingram Cancer Center
Lepisto	Eva	M.	Dana-Farber Cancer Institute
Levy	Mia	A.	Vanderbilt-Ingram Cancer Center
Lindeman	Neal	I.	Dana-Farber Cancer Institute, Brigham and Women's Hospital, and Harvard Medical School
Lindsay	James		Dana-Farber Cancer Institute
Liu	David		Dana-Farber Cancer Institute
Lu	Zhibin		Princess Margaret Cancer Centre, University Health Network
MacConaill	Laura	E.	Dana-Farber Cancer Institute, Brigham and Women's Hospital, and Harvard Medical School
Maurer	Ian		GenomOncology
Maxwell	David	S.	UT MD Anderson Cancer Center
Meijer	Gerrit	A.	Netherlands Cancer Institute
Meric-Bernstam	Funda		UT MD Anderson Cancer Center
Micheel	Christine	M.	Vanderbilt-Ingram Cancer Center
Miller	Clinton		GenomOncology
Mills	Gordon		UT MD Anderson Cancer Center
Moore	Nathanael	D.	Dana-Farber Cancer Institute
Nederlof	Petra	M.	Netherlands Cancer Institute
Omberg	Larsson		Sage Bionetworks
Orechia	John	A.	Dana-Farber Cancer Institute
Park	Ben	Ho	Sidney Kimmel Cancer Center at Johns Hopkins University
Pugh	Trevor	J.	Princess Margaret Cancer Centre, University Health Network
Reardon	Brendan		Dana-Farber Cancer Institute
Rollins	Barrett	J.	Dana-Farber Cancer Institute, Brigham and Women's Hospital, and Harvard Medical School
Routbort	Mark	J.	UT MD Anderson Cancer Center
Sawyers	Charles	L.	Memorial Sloan Kettering Cancer Center and Howard Hughes Medical Institute Investigator

(continued)

Alphabetical List of Authors (*Continued*)

Last name	First name	M	Institution
Schrag	Deborah		Dana-Farber Cancer Institute, Brigham and Women's Hospital, and Harvard Medical School
Schultz	Nikolaus		Memorial Sloan Kettering Cancer Center
Shaw	Kenna R Mills		UT MD Anderson Cancer Center
Shivdasani	Priyanka		Dana-Farber Cancer Institute and Brigham and Women's Hospital
Siu	Lillian	L.	Princess Margaret Cancer Centre, University Health Network
Solit	David	B.	Memorial Sloan Kettering Cancer Center
Sonke	Gabe	S.	Netherlands Cancer Institute
Soria	Jean Charles		Institut Gustave Roussy
Sripakdeevong	Parin		Dana-Farber Cancer Institute
Stickle	Natalie	H.	Princess Margaret Cancer Centre, University Health Network
Stricker	Thomas	P.	Vanderbilt-Ingram Cancer Center
Sweeney	Shawn	M.	American Association for Cancer Research
Taylor	Barry	S.	Memorial Sloan Kettering Cancer Center
ten Hoeve	Jelle	J.	Netherlands Cancer Institute
Thomas	Stacy	B.	Memorial Sloan Kettering Cancer Center
Van Allen	Eliezer	M.	Dana-Farber Cancer Institute
Van 'T Veer	Laura	J.	UCSF Helen Diller Family Comp. Cancer Center
van de Velde	Tony		Netherlands Cancer Institute
van Tinteren	Harm		Netherlands Cancer Institute
Velculescu	Victor	E.	Sidney Kimmel Cancer Center at Johns Hopkins University
Virtanen	Carl		Princess Margaret Cancer Centre, University Health Network
Voest	Emile	E.	Netherlands Cancer Institute
Wang	Lucy	L.	Vanderbilt-Ingram Cancer Center
Wathoo	Chetna		UT MD Anderson Cancer Center
Watt	Stuart		Princess Margaret Cancer Centre, University Health Network
Yu	Celeste		Princess Margaret Cancer Centre, University Health Network
Yu	Thomas	V.	Sage Bionetworks
Yu	Emily		UT MD Anderson Cancer Center
Zehir	Ahmet		Memorial Sloan Kettering Cancer Center
Zhang	Hongxin		Memorial Sloan Kettering Cancer Center

Disclosure of Potential Conflicts of Interest

F. André reports receiving commercial research grants from AstraZeneca, Lilly, Novartis, and Pfizer. M. Arnedos has received honoraria from the speakers bureaus of Novartis and AstraZeneca, and is a consultant/advisory board member for Puma. D.M. Hyman reports receiving commercial research grants from AstraZeneca, Loxo Oncology, and PUMA Biotechnology, and is a consultant/advisory board member for Atara Biotherapeutics, Chugai, and CytomX. M.A. Levy is an advisory board member of Personalis, Inc., and receives royalty distribution from GenomOncology. F. Meric-Bernstam reports receiving commercial research grants from Aileron, AstraZeneca, Bayer, Calithera, Curis, CytoMx, Debiopharma, Effective Pharma, Genentech, Jounce, Novartis, PUMA, Taiho, and Zymeworks, and is a consultant/advisory board member for Clearlight Diagnostics, Darwin Health, Dialecta, GRAIL, Inflection Biosciences, and Pieris. G.B. Mills reports receiving commercial research grants from Adelson Medical Research Foundation, AstraZeneca, Breast Cancer Research Foundation, Critical Outcome Technologies, Illumina, Karus, Komen Research Foundation, NanoString, and Takeda/Millennium Pharmaceuticals; has received honoraria from the speakers bureaus of Allostery, AstraZeneca, ImmunoMet, ISIS Pharmaceuticals, Lilly, MedImmune, Novartis, Pfizer, Symphogen, and Tarveda; has ownership interest (including patents) in Catena Pharmaceuticals, ImmunoMet, Myriad Genetics, PTV Ventures, and Spindletop Ventures; and is a consultant/advisory board member for Adventist Health, Allostery, AstraZeneca, Catena Pharmaceuticals, Critical Outcome Technologies, ImmunoMet, ISIS Pharmaceuticals, Lilly, MedImmune, Novartis, Precision Medicine, Provista Diagnostics, Signalchem Lifesciences, Symphogen, Takeda/Millennium Pharmaceuticals, Tarveda, and Tau Therapeutics. T.J. Pugh is a consultant/advisory board member for Dynacare. C.L. Sawyers is a consultant/advisory board member for Novartis. V.E. Velculescu has ownership interest (including patents) in Personal Genome Diagnostics and is a consultant/advisory board member for the same. No potential conflicts of interest were disclosed by the other authors.

One of the Editors-in-Chief is an author on this article. In keeping with the AACR's editorial policy, the peer review of this submission

was managed by a senior member of *Cancer Discovery*'s editorial team; a member of the AACR Publications Committee rendered the final decision concerning acceptability.

Authors' Contributions

REFERENCES

1. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med 2016;375:1109–12.
2. Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. J Clin Oncol 2013;31:1803–5.
3. Tannock IF, Hickman JA. Limits to personalized cancer medicine. N Engl J Med 2016;375:1289–94.
4. Prasad V. Perspective: The precision-oncology illusion. Nature 2016; 537:S63.
5. Joyner MJ, Paneth N, Ioannidis JPA. What happens when underperforming big ideas in research become entrenched? JAMA 2016; 316:1355–6.
6. Shrager J, Tenenbaum JM. Rapid learning for precision oncology. Nat Rev Clin Oncol 2014;11:109–18.
7. Obermeyer Z, Ziad O, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. N Engl J Med 2016;375: 1216–9.
8. Siu LL, Lawler M, Haussler D, Knoppers BM, Lewin J, Vis DJ, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. Nat Med 2016;22:464–71.
9. Blue Ribbon Panel report to the National Cancer Advisory Board [Internet]; 2016 [cited 2013 Nov 9]. Available from: https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/blue-ribbon-panel-report-2016.pdf
10. AACR Project GENIE Publicly Releases Large Cancer Genomic Data Set [Internet]. Available from: http://www.aacr.org/Newsroom/Pages/News-Release-Detail.aspx?ItemID=994#.WJEGZLYrJ0I
11. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature 2010;464:993–8.
12. Derry JMJ, Mangravite LM, Suver C, Furia MD, Henderson D, Schildwachter X, et al. Developing predictive molecular maps of human disease through community-based modeling. Nat Genet 2012;44: 127–30.
13. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6:l1.
14. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401–4.
15. Jones S, Anagnostou V, Lytle K, Parpart-Li S, Nesselbush M, Riley DR, et al. Personalized genomic analyses for cancer mutation discovery and interpretation. Sci Transl Med 2015;7:283ra53.
16. OncoTree [Internet]. Available from: http://www.cbioportal.org/oncotree/
17. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495–501.
18. Miao D, Van Allen EM. Genomic determinants of cancer immunotherapy. Curr Opin Immunol 2016;41:32–8.
19. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet 2016;48: 607–16.
20. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7.
21. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.

22. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015;43:D805–11.

23. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol 2016;34:155–63.

24. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. JCO PO 2017 May 16 [Epub ahead of print].

25. Conley BA. Abstract IA38: NCI MATCH: a national precision medicine trial – Conception, development, and adjustment. Cancer Epidemiol Biomarkers Prev 2016;IA38.

26. Gracia D, Diego G. Institute of Medicine (IOM). The learning healthcare system: workshop summary. Washington, DC: The National Academies Press; 2007. EIDON no 39. 2013.

27. Tunis S, Whicher D. The national oncologic PET registry: lessons learned for coverage with evidence development. J Am Coll Radiol 2009;6:360–5.

28. Haber DA, Velculescu VE. Blood-based analyses of cancer: circulating tumor cells and circulating tumor DNA. Cancer Discov 2014;4:650–61.