High-Throughput Genomic Profiling of Adult Solid Tumors Reveals Novel Insights into Cancer Pathogenesis

Ryan J. Hartmaier1, Lee A. Albacker1, Juliann Chmielecki1, Mark Bailey1, Jie He1, Michael E. Goldberg1, Shakti Ramkissoon1, James Suh1, Julia A. Elvin1, Samuel Chiacchia1, Garrett M. Frampton1, Jeffrey S. Ross1,2, Vincent Miller1, Philip J. Stephens1, and Doron Lipson1

Abstract

Genomic profiling is widely predicted to become a standard of care in clinical oncology, but more effective data sharing to accelerate progress in precision medicine will be required. Here, we describe cancer-associated genomic profiles from 18,004 unique adult cancers. The dataset was composed of 162 tumor subtypes including multiple rare and uncommon tumors. Comparison of alteration frequencies to The Cancer Genome Atlas identified some differences and suggested an enrichment of treatment-refractory samples in breast and lung cancer cohorts. To illustrate novelty within the dataset, we surveyed the genomic landscape of rare diseases and identified an increased frequency of NOTCH1 alterations in adenoid cystic carcinomas compared with previous studies. Analysis of tumor suppressor gene patterns revealed disease specificity for certain genes but broad inactivation of others. We identified multiple potentially druggable, novel and known kinase fusions in diseases beyond those in which they are currently recognized. Analysis of variants of unknown significance identified an enrichment of SMAD4 alterations in colon cancer and other rare alterations predicted to have functional impact. Analysis of established, clinically relevant alterations highlighted the spectrum of molecular changes for which testing is currently recommended, as well as opportunities for expansion of indications for use of approved targeted therapies. Overall, this dataset presents a new resource with which to investigate rare alterations and diseases, validate clinical relevance, and identify novel therapeutic targets. Cancer Res; 77(9): 2464–75. ©2017 AACR.

Introduction

Large-scale genomic datasets, such as The Cancer Genome Atlas (TCGA), have increased our understanding of the molecular mechanisms driving tumorigenesis (1–3). Public availability of this data has facilitated novel discoveries, validated rare findings, and allowed for the incorporation of genomic features into clinical trial design (4). As more cancer patients undergo clinical genomic profiling, sharing these data with the broader research community is critical for accelerating precision medicine (4).

The value of genomic data is apparent in multiple cancers where molecular alterations define distinct clinical groups. For example, glioblastomas can be divided into four molecular subtypes, each with distinct survival and response rates to standard therapies (5, 6). Insights from cancer genomics analyses have also led to the development and validation of targeted treatment options against molecularly matched alterations that can be more effective and less toxic than traditional chemotherapeutic regimens. For example, in non–small cell lung cancer (NSCLC), alterations in 8 genes are associated with sensitivity to targeted inhibitors and genomic analyses of these targets is now recommended in treatment guidelines (7). In addition, broad genomic features, such as total tumor mutational burden, have been proposed as potential biomarkers of sensitivity for immune checkpoints inhibitors (8). Genomic information can also inform clinical trial design to better identify patients likely to respond to targeted inhibitors. Multi-arm umbrella trials such as NCI-MATCH (NCT02465060) and LUNG-MAP (NCT02154490) are using genomic features to select appropriate treatment arms. Basket trials, selecting patients based on molecular alterations versus tumor type have also been developed for alteration-specific inhibitors (9). Across cancers, patients enrolled in molecularly matched clinical trials have demonstrated superior survival versus random trial enrollment (10).

The sharing of genomic data is critical to further our understanding of molecular drivers and to develop effective therapies. Here, we present genomic profiles from 18,004 unique adult solid tumors that underwent targeted genomic profiling as part of routine clinical care. This collection represents a vast diversity of tumor subtypes, including many rare diseases not profiled as part of large-scale efforts. High-level analysis identified novel alterations in common diseases and confirmed...
the prevalence of alterations that were underrepresented previously in small cohort studies. In addition, we highlight the spectrum of clinically relevant alterations with established roles in determining drug sensitivity. By making this data available to the broader research community, it is anticipated that this information will serve as a source of discovery and validation for projects aimed at improving cancer treatments and outcomes.

Materials and Methods

Sample profiling

This study was reviewed and approved by the Western Institutional Review Board. Samples were submitted between 2012–2014 to a Clinical Laboratory Improvement Amendment (CLIA)-certified, New York State-accredited, and CAP-accredited laboratory (Foundation Medicine) for hybrid capture followed by next-generation sequencing (NGS). The pathologic diagnosis of each case was reviewed and tumor content determined from hematoxylin and eosin (H&E)-stained slides. All samples that advanced to DNA extraction contained a minimum of 20% tumor cells. The majority of samples contain < 50% tumor content (Supplementary Fig. S1A). DNA was extracted from formalin fixed paraffin embedded 10-μm sections. Adaptor-ligated DNA underwent hybrid capture for all coding exons of 287 cancer-related genes plus select introns from 19 genes frequently rearranged in cancer (Supplementary Table S1). Captured libraries were sequenced to a median exon coverage depth of > 600 X using Illumina HiSeq sequencing technology (Supplementary Fig. S1B). To protect against inadvertent reidentification, samples in ultra-rare disease types (< 5 samples) were removed. In cases where multiple samples from the same individual were tested (determined via a SNP-genotype based approach), only one sample was included. No further selection was performed and all samples meeting these criteria are included in this dataset.

Sequencing analysis and deposition of data

Resultant sequences were analyzed for short variants [base substitutions, insertions/deletions (indels)], copy number (CN) alterations [local amplifications and homozygous deletions], and select gene fusions using the hg19 reference genome, as described previously (11). As tumor samples were sequenced without a genotype based approach), only one sample was included. No further selection was performed and all samples meeting these criteria are included in this dataset.

Analysis of TCGA data

TCGA data (2016-01-28), analyzed via Firehose, was accessed via the ‘firehose_get’ tool. See the Supplementary Methods for additional details. Survival analysis was determined via log-rank test.

Mutation hotspot caller

We performed a hotspot analysis of missense and nonsense mutations to predict hotspot changes within a given gene. See Supplementary Methods for additional details.

Results

Overview of genomic and clinical characteristics

The Foundation Medicine (FM) adult dataset consists of 18,004 unique tumor samples that underwent genomic profiling as part of standard clinical care. Each sample was assigned a detailed diagnosis; these detailed subtypes were then grouped into broader disease categories. In total, 16 broad disease categories were created with tumors from 162 unique disease subtypes (Fig. 1A and B, Supplementary Figs. S3–S11). The most common disease categories are thoracic cancers (20.7%, 12 subtypes), gastrointestinal (GI) cancers (17.1%, 15 subtypes), breast cancers (14.3%, 8 subtypes), gynecologic cancers (8.3%, 25 subtypes), and hepatopancreatobiliary cancers (7.1%, 15 subtypes). The remaining 32.5% of samples include genitourinary, CNS, neuroendocrine, head and neck, melanoma, and additional less common cancer types. Samples for which detailed pathologic information was unavailable were grouped into general “not otherwise specified” (nos) categories within each disease category. Among all the specific and “nos” categories, the most common disease subtypes were lung adenocarcinoma (13.0%), colon adenocarcinoma (9.6%), breast carcinoma (nos; 6.6%), breast invasive ductal carcinoma (6.5%), and lung non-small cell carcinoma (nos; 3.4%). Of the 162 detailed diagnostic subtypes, 63% were comprised of 50 or fewer samples, including multiple rare diseases (Supplementary Table S2). All detailed subtypes had at least 5 unique tumors. In addition, 13.0% of the disease subtypes contained 200 samples or greater (range 203–2,345) allowing for robust statistical analyses in common diseases. Genes frequently altered across the dataset included the TP53 (54%), KRAS (21%), CDKN2A (19%), PIK3CA (14%), and CDKN2B (12%; Fig. 1C).

Gender information was available from all but 9 samples and showed a slight bias toward females (56.7%) versus males (43.3%; Fig. 1D). This bias can be explained in part by the large number of breast and GYN cancer samples within the dataset (Fig. 1A). The average age of patients at the time of genomic profiling was 57.7 years (median: 59 years, range: 19–88 years; Fig. 1D). Patients 89 years old or older at the time of testing were excluded to comply with privacy guidelines. Information about disease recurrence and previous treatment histories was not available for the cohort.

Comparison of alteration frequencies to known datasets

In contrast to other large-scale genomic profiling efforts that employed multiple technologies (i.e., DNA-seq, CN arrays, and RNA-seq) to analyze genomic alterations, samples within this cohort were analyzed for all classes of genomic alterations on a single uniform platform. Due to the clinical setting in which these samples were profiled, no selection criteria (e.g., fresh vs. archival tissue, primary vs. metastatic tissues, or pre- vs. posttreatment samples) was applied. Therefore, we hypothesized that the spectrum and frequency of genomic alterations would vary compared with those from other large-scale profiling efforts (i.e., TCGA) that utilized multiple technologies and applied stringent selection criteria. We analyzed these differences using TCGA samples for which both CN (GISTIC2) and mutation data (MutSig2CV) existed and for which disease subtypes could be mapped readily between the two datasets. Finally, we included only diseases for which at least 200 samples existed in both cohorts to avoid sampling bias. (Fig. 2; Supplementary Table S3).
We first investigated the impact of methodologic differences. Differences in CN alteration frequencies were observed across many tumors (Fig. 2, green bars). These discrepancies may be explained in part by different technologies used to measure these events by TCGA (SNP arrays) and FM (NGS-based modeling), differences in sample input requirements (i.e., >70% tumor content vs. ≥20% tumor content), annotation thresholds (FM requires focal segments for most genes and >6–8 copies depending on tumor ploidy), or functional status (plotted FM data excludes variants of unknown significance, VUSes). However, CN differences appeared to be tumor type specific as few discrepancies were observed in breast and bladder cancer cohorts, whereas multiple differences were observed in ovarian cancers, head and neck squamous cell carcinomas, and lung adenocarcinomas. These CN differences were not due to the high proportion of metastatic samples in the FM dataset as restricting the FM dataset to local disease produced similar results (Supplementary Figs. S12 and S13). Although the low cellularity typical of FM samples could potentially explain the lower rate of CN detection, the high sequencing depth applied allows for high CN sensitivity even at 20–30% cellularity (12). Thus, the differences in copy number observed are most likely due to differences in technology, annotation, and/or sample differences (i.e., differences in proportion of tumor subtypes).

We next investigated differences in mutation frequencies. The largest mutation frequency difference between the two datasets was observed for LRPIB in lung adenocarcinomas (6.5% vs. 40.9%). Large discrepancies for LRPIB are also seen in melanomas and head and neck squamous cell carcinomas. These discrepancies are largely explained by the many VUSes in this gene (excluded from comparison in Fig. 2) that result from the high background mutation rate in these tumor types (Supplementary Fig. S14). In contrast, the higher TP53 alteration frequency in breast and uterine/endometrial cancers could not be accounted for by filtered VUSes or differences in the types of alterations reported, as rates of missense, nonsense, splice, and frameshift

Figure 1.
Clinical and genomic characteristics of samples within the dataset. A, Samples were grouped into 16 broad disease categories describing their tissue of origin. Not shown on the graph (grouped into the other category for visualization) are four additional diseases (unclassified, 0.9%; nonmelanoma skin cancers, 0.8%; mesenchymal cancers, 0.6%; and germ cell tumors, 0.2%). B, In addition, each sample was assigned a detailed label that represented its pathologic diagnosis. Subtype distributions for the top three broad categories (thoracic, GI, and breast cancers) are depicted in the smaller surrounding charts. Thoracic cancers included samples from 12 disease subtypes (the top 5 are depicted in the small chart). GI cancers included 15 disease subtypes (the top 10 are depicted in the small chart). Breast cancers were comprised of samples from 8 subtypes (the top four are depicted in the small chart). See Supplementary Figs. S3–S11 and Supplementary Table S2 for further details about subtypes. C, Long tail distribution of alterations across the entire dataset. SV, short variants (includes missense mutations and indels); CN, copy number alterations; RE, rearrangements; Mult, multiple events. D, Clinical characteristics of samples within the dataset, including age and gender distributions.
mutations were similar between the two cohorts (Supplementary Fig S1A). Similarly, tissues sites suspected to yield lower quality DNA (i.e., pleural fluid) did not show evidence of distinct mutation detection (Supplementary Fig S1B). Furthermore, within the FM dataset, TP53 alterations are enriched in local disease, indicating the discrepancy is not due to the high proportion of metastatic samples (see Supplementary Table S4 for genes altered at different rates in local vs. metastatic disease). Therefore, differences in mutation frequencies may represent sample differences between the datasets.

To explore biological differences between the dataset, we investigated specific molecular trends associated with advanced refractory tumors. For example, EGFR short variants (point mutations and indels) were slightly more frequent within FM lung adenocarcinomas versus TCGA (20.3% vs. 14.5%, P = 0.003). However, the EGFR T790M mutation, associated with acquired resistance to targeted inhibitors, was observed at a much higher frequency within FM samples compared with the treatment-naïve TCGA dataset (4.1% vs. 1.2%, P = 5.3 × 10^-7). A similar trend was also observed in breast cancers, where the higher frequency of ESR1 alterations in the FM dataset (9.5% vs. 3.7%, P = 2.8 × 10^-5) suggested a selection for samples with acquired resistance to endocrine therapies. These data suggest an enrichment of treatment refractory breast and lung cancer samples within the FM collection.

Collectively, these comparisons confirm differences between TCGA and FM. Methodologic differences between the datasets are likely contributing a minor role to the observed discrepancies. However, the enrichment of resistance-associated alterations in breast and lung cancers suggests a biological difference between the tumors within our cohort and TCGA. Although molecular signatures of therapeutic resistance are less well understood in diseases not treated routinely with targeted therapies, one can infer that these tumors are likely from advanced stage cancers.

Novel variants and disease enrichment patterns

We next applied an internally developed hotspot-calling algorithm (see Supplementary Methods) in silico prediction tools (MutationAssessor and PolyPhen-2; refs. 12, 13) to search for novel functional variants of unknown significance (VUS). Using the hotspot caller, we identified several mutation hotspot in SMAD4, a tumor suppressor gene (TSG) previously implicated in colorectal cancers (14). The hotspots identified were mostly nonsense mutations, consistent with a tumor suppressor role (Fig. 3A). However, we did observe several hotspot missense mutations including known hotspots such as D351A/G/V/Y, R361C/H/S, P356H/L/R/S, D537A/E/G/H/V/Y (15, 16), and novel hotspots such as A118V, E330K, G419M/R/W, and D493H/N, and W524C/G/R (Fig. 3A). We also observed hotspot mutations in the related pathway components SMAD2 (T303 and S464) and TGFBR2 (R495 and R528; Fig. 3B and C). At the TGFβ/SMAD pathway is a tumor suppressor pathway, we analyzed hotspot mutations, truncations, and homozygous deletions (collectively identifying TGFβ/SMAD pathway altered samples). Alterations in this pathway were most frequent in GI and pancreato-hepatobiliary cancers with 15.8% (487/3076) and 13.6% (175/1284) of total cases, respectively (Fig. 3D). Alterations in this pathway were nonoverlapping although significance could not be addressed due to the small sample size (Fig. 3E). We did observe a significant cooccurrence with KRAS alterations (OR = 1.70, P = 9.6 × 10^-4) and a significant mutual exclusivity with TP53 alterations (OR = 0.62, P = 2.34 × 10^-5; Fig. 3F). To demonstrate clinical relevance, we incorporated survival data from TCGA samples harboring TGFβ/SMAD pathway alterations (14). Colon adenocarcinomas were separated into TGFβ/SMAD pathway altered (n = 75; hotspot mutation/truncation/homozygous deletion) versus TGFβ/SMAD pathway unaltered (n = 254) groups. The TGFβ/SMAD-altered group exhibited reduced progression free survival (PFS) that trended toward significance (P = 0.06; log rank test; Fig. 3G).

Figure 2.

Comparison of alteration frequencies between TCGA and FM datasets. Frequencies of alterations in analogous tumor subtypes were compared between FM and TCGA datasets. FM data excludes VUSes, neighboring gene was used to determine CN status at this locus (**), significant differences in alteration frequencies (P < 0.05); BRCA, breast invasive carcinoma; BLCA, bladder urothelial carcinoma; OVCA, ovarian carcinoma; UCEC, uterine corpus endometrial carcinoma; GBM, glioblastoma; COAD, colon adenocarcinoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; LUAD, lung adenocarcinoma; and HNSC, head and neck squamous cell carcinoma.
To capture rare alterations, we also analyzed the functional impact of VUS point mutations occurring in at least 5 samples using MutationAssessor and PolyPhen-2 tools (12, 13). A merge of these outputs identified 23 unique point mutations in 11 genes with predicted functional impact across both algorithms (Supplementary Table S5). All mutations also reached significance using the hotspot caller described above. In addition to the mutations described above, this VUS analysis identified multiple alterations in EBRB3 (T355I and T389I), BRIPI (R762C and R251C), KEAP1 (G523V, R413C, and G419W), and SMARCA4 (E882K, P913L, R973W, R1135Q, and R1192H) and intriguing variants in PTEN (D24H), FLT1 (E432K), STK11 (P179R), LRP1B (G401E), ESR1 (A361V), and CDKN2A (G101V) predicted to have functional impact. Interestingly, these variants tended to occur in tumor types associated with alterations in that gene. For example, 3 of 6 FLT1 E432K alterations were in melanomas, 21/32 of the various SMAD4 alterations were found G1 and hepatobiliary cancers, and 9 of 15 of the KEAP1 alterations were found in thoracic cancers. The nonrandom distribution of these VUSes strongly implicates them as functional alterations. Of note, as matched germline DNA was not tested, we cannot completely discount the possibility that some of these alterations may be rare germline alterations. However, somatic/germline status was predicted with a novel, internally developed algorithm that assesses germline status based on allele frequency and tumor purity/ploidy (Sun and colleagues, in review 2016). For example, in a sample with moderate cellularity (~50%), in copy neutral, diploid regions, somatic alteration allele frequencies will be impacted by cellularity and have allele frequencies approximately 25%. In contrast, heterozygous germline variants will not be impacted by cellularity and will have allele frequencies approximately 50%. These estimates suggest that >90% of the VUSes predicted to have a functional impact are somatic events (data not shown).
Genomic analysis of rare diseases uncovers higher frequencies of NOTCH1 and BCOR alterations in adenoid cystic carcinomas

We investigated the genomics of rare tumors given that many of the subtypes within the FM dataset were excluded from large-scale analyses (i.e., TCGA) or profiled only as part of small cohorts. Interesting results were observed in adenoid cystic carcinomas (ACC) of the head and neck region (n = 156 total), including head and neck ACCs (n = 78), salivary gland ACCs (n = 57), tracheal ACCs (n = 7), and unknown primary ACCs (n = 14). In agreement with recent findings of 36 recurrent and metastatic ACCs (17), the most frequent alterations occurred in NOTCH1 (23%; Fig. 4A). NOTCH1 alterations were clustered in the C-terminal PEST domain of the protein (Fig. 4B), and were significantly enriched within this disease (Supplementary Table S6). NOTCH1 PEST domain alterations are weakly activating by themselves (18). However, they are synergistic with HD domain mutations in cis and can significantly increase the activity of this protein when this combination of alterations is present. Interestingly, 7 samples harbored mutations in both the PEST and HD domains, suggesting a potential mechanism through which activity of this gene could be altered (Fig. 4B). Unfortunately, phasing of the mutations was impossible to determine definitively due to the distance between mutations and the length of the sequencing reads. However, two tumors contained a third alteration (E794* and N390fs*243) toward the 5’ end of NOTCH1, likely disrupting one allele and suggestive that the other two alterations are in cis. Collectively, other genomic studies have investigated 111 total ACC tumors, including 28 samples also represented within this dataset, and reported NOTCH1 missense and nonsense alterations in 5–10% of samples (19–21). This analysis confirms findings from smaller studies that NOTCH1 is the most commonly altered gene in ACCs at approximately 24% and extends it by providing multiple examples of cooccurring PEST and HD domain alterations. Further work to evaluate the effect of these alterations is warranted as multiple inhibitors of this protein are currently in clinical trials (22).

The second most common alterations in ACCs occurred in the TSG BCOR (17%; Fig. 4A). All variants are predicted to inactivate this protein (Fig. 4C). These results agree with previous findings where 4/36 ACCs were found to have truncating mutations in BCOR. Together, these results establish BCOR inactivation as a signature event in ACC. BCOR alterations have been described in...
myelodysplastic syndromes where they are associated with poor prognosis and shorter overall survival (23). They have also been described in multiple pediatric tumors where they are thought to play a role in chromatin modification (24). Both BCor and NOTCH1 alterations cooccurred with other events (Supplementary Fig. S16).

Identification of novel fusion events

We investigated the spectrum of fusions involving 8 clinically relevant kinases (ALK, BRAF, EGFR, NTRK1, PDGFRA, RAF1, RET, and ROS1) with established druggability. In total, we identified 19 novel fusions with structures similar to known oncogenic fusion proteins and multiple known fusions in diseases different from those in which they were reported originally.

Nine novel fusions involved the serine threonine kinases BRAF or RAF1 (Fig. 5). These included fusions with four novel fusion partners in diseases known to be driven by these events, such as astrocytoma, melanoma, and prostatic acinar cell carcinoma (Fig. 5A). RAF1 fusions were also observed in 5 other disease types (Fig. 5B). Interestingly, the PRKAR2A-RAF1 fusion was recurrent and observed in both a lung adenocarcinoma and an unknown primary melanoma. In addition, we also observed 7 novel tyrosine kinase fusions involving ALK, RET, ROS1, and NTRK1 (Fig. 5C) in non–small cell lung cancers and thyroid cancer. Novel fusions involving ALK, RET, ROS1, and NTRK1 were also observed in a colon adenocarcinoma, a GI neuroendocrine tumor, and a uterine endometrial carcinoma (Fig. 5D).

Although kinase fusions can be hallmarks of certain cancers, we observed known fusions outside of the diseases in which they were identified originally. CLTC-ALK and STRN-ALK have been observed in aggressive thyroid cancers and soft tissue malignancies (25, 26). Within the FM cohort, these fusions were observed in a lung adenocarcinoma and an epithelioid peritoneal mesothelioma, respectively (Supplementary Fig. S17). Diagnosis of the epithelioid peritoneal mesothelioma is supported by IHC staining (positive for cytokeratin 7, calretinin and vimentin; negative for CEA, B72.3, and TF1). While a subset of lung adenocarcinomas is known to be driven by ALK fusions, CLTC-ALK has yet to be described in this disease. In contrast, kinase fusions have not been noted in mesothelioma; STRN-ALK represents a novel, yet rare (1/184, 0.5%), driver event in this disease. RET fusions in thyroid and lung cancers are well characterized (27–28). Here, two breast cancers were found to harbor the oncogenic Ccdc6-Ret fusion (Supplementary Fig. S17). Two similar oncogenic RET fusions (Ncoa4 RET and Kif5b-RET) were also identified in a rare liver cholangiocarcinoma and an ovarian epithelial carcinoma, respectively (Supplementary Fig. S17). BRAF fusions have been described in melanoma, thyroid cancers, and pediatric brain cancers (29). A single thyroid papillary carcinoma was found with a Mad1l1-Braf fusion identified previously in melanoma (30). We also observed the known Tmem106b-Ros1 fusion (31) in a liver cholangiocarcinoma. Interestingly, Gopc(Fig)-Ros1 has been reported in a glioblastoma cell line (32), in a lung adenocarcinoma (33), and in rare biliary tract carcinomas (34). Here, we observed this fusion in a small intestine adenocarcinoma and confirm its presence in a human glioblastoma sample (Supplementary Fig. S17). Finally, we observed an imatinib sensitive hypereosinophilic syndrome fusion, Fip1l1-Pdgfra (35), in a glioblastoma (Supplementary Fig. S17).

Spectrum of known clinically relevant alterations across diseases

Recent publications have suggested a broad spectrum of genomic changes in clinically relevant targets (29, 36–38), and expanded analyses are warranted to identify more patients predicted to be sensitive or resistant to targeted therapies. Furthermore, the identification of drug sensitivity and resistance biomarkers across multiple indications suggests that targeted agents may have broader utility beyond that for which they were approved originally. We surveyed the spectrum of known clinically relevant genomic changes to identify (i) the spectrum of these changes in indications for which testing is currently recommended and (ii) potential opportunities for broader utility of approved targeted agents.

In colorectal cancers (CRC), activating mutations in KRAS are predictive of poor response to cetuximab. Alterations in hotspot...
amplifications and deletions (indels) in carcinomas (29.8%). In gallbladder adenocarcinomas (11.5%) and liver cholangiomas (46.3%) and pancreatic cancers (89%) and a low frequency (60.3%). MAPK alterations in cancers of the biliary tree (64.5%) as neighboring small intestine (56.6%) and CRCs adenocarcinomas had a similar rate of MAPK alterations (39). These extended RAS testing guidelines captured an additional 188 CRC samples (9.5%) within our dataset. Beyond these extended guidelines, we also observed activating KRAS and NRAS amplifications and mutations at codons 14 and 22 in an additional 29 CRC samples (1.5%; Fig. 6A; refs. 40, 41). These data provide robust estimates of RAS alterations in CRCs, utilizing both ASCO guidelines and the current state of knowledge, which is critical when considering EGFR targeted therapies.

As mutations in the MAPK pathway are known drivers of multiple GI cancers, we performed a survey of KRAS, NRAS, HRAS, and BRAF alterations across cancers of the GI and hepato-biliary tracts (Fig. 6B). Gastroesophageal junction adenocarcinoma was unique in that it had a high proportion of KRAS amplifications without a KRAS mutation (16.3%). Appendix adenocarcinomas had a similar rate of MAPK alterations (64.5%) as neighboring small intestine (56.6%) and CRCs (60.3%). MAPK alterations in cancers of the biliary tree (see Fig. 6B) included a high frequency in bile duct adenocarcinomas (46.3%) and pancreatic cancers (89%) and a low frequency in gallbladder adenocarcinomas (11.5%) and liver cholangiocarcinomas (29.8%).

We next investigated the spectrum of clinically relevant insertions and deletions (indels) in EGFR (EGFRvIII rearrangements were also identified and are discussed in a subsequent section). The most prevalent EGFR exon 19 indel was the canonical E746-A750 deletion, although the length of deletions at amino acids 746 and 747 ranged from 3–7 amino acids (Fig. 6C). We also observed rare deletions of 8 amino acids at positions 751 and 752. In contrast, drug-resistant EGFR exon 20 insertions lacked a single dominant location and length, but the vast majority occurred between amino acids 769–774 and inserted 1–3 residues within this region (Fig. 6D). Analogous ERBB2 insertions primarily consisted of a 4 amino acid duplication between residues 772–775, however, these insertions ranged from 1–4 amino acids between residues 772 and 780 (Fig. 6E). The diversity of these mutations has implications for robust diagnostic detection and developing drugs to target this heterogeneous set of insertions (37).

Current data suggest that ERBB2 amplifications, oncogenic point mutations, and activating insertions can confer clinical sensitivity to ERBB2 targeted agents (42). We observed recurrent activating ERBB2 alterations across 15 tumor types (Fig. 7A), including new trends such as ERBB2 amplifications in cervical cancer and skin squamous cell carcinomas. ERBB2 point mutations have been described in cervical cancers (43), but amplifications suggest an alternate mechanism through which the gene can be activated in this disease. Consistent with previous studies, ERBB2 activating mutations in cervical cancers did not cooccur with copy number alterations (Supplementary Fig. S18). To our knowledge, ERBB2 amplifications in skin squamous cell carcinoma represent a novel therapeutic target in this disease.

Amplification of the CDK4/6 locus has been associated with response to the CDK4 inhibitor palbociclib in breast cancers and liposarcomas (44, 45). Within this cohort, we observed CDK4/6 amplification across 43 tumor types (Fig. 7B). Novel findings

---

**Figure 6.** Diversity of clinically relevant alterations across the dataset. **A.** Distribution of clinically relevant KRAS alterations in colorectal adenocarcinomas. **B.** Distribution pattern of all RAS alterations in GI cancers. **C–E.** Indel alterations in EGFR (C and D) and ERBB2/HER2 (E) can vary in length. Numbers on the left side of the graphs correspond to the codon positions; heatmaps display the number of samples.
Hartmaier et al.

Figure 7.
Spectrum of druggable alterations across cancers. Distribution of alterations across the top 15 disease types for ERBB2/HER2 (A), amplifications in CDK4/6 (B), amplifications in MET (C), and activating mutations in AKT1 (D).

Included amplification of these genes in gallbladder carcinomas (11.5%) and oligodendrogliomas (6.6%). Preclinical models have suggested that this event may contribute to oligodendroglioma formation (46).

MET amplification was observed in glioblastoma samples at an appreciable frequency (2.1%; Fig. 7C). We explored whether this finding correlated with overexpression of MET mRNA by incorporating expression data from TCGA glioblastoma samples. Although few TCGA GBM cases exist with MET amplification, amplification was associated with increased expression (Supplementary Fig. S19). Interestingly, at least one case report has demonstrated clinical sensitivity to the MET inhibitor, crizotinib, in this disease (47). A similar observation was confirmed in AKT1 E17K-mutant colorectal adenocarcinomas (Fig. 7D). While this mutation has been observed previously in this disease, frequencies varied from 0%–8.2% (14, 48, 49). Our data confirm the rare occurrence of this alteration in approximately 1.0% of routine CRC samples. Interestingly, in contrast to a previous study (50), we observe AKT1 E17K-mutant colorectal samples to be enriched for KRAS alterations \((P = 0.02)\) but not BRAF alterations (Supplementary Fig. S20).

In addition to potentially targetable alterations that occur across diseases, we also observed striking patterns of disease specificity for certain alterations. EGFR/III rearrangements and extracellular activating mutations were found almost exclusively in glioblastomas while activating indels in EGFR occurred almost exclusively in NSCLCs. Although we also saw an appreciable rate of EGFR indels in unknown primary adenocarcinomas, it is likely that these represent NSCLC samples for which incomplete pathology information was available (J. Ross, personal communication). A similar trend was also observed for ROS1 rearrangements. These events were observed primarily in NSCLC and a small proportion of glioblastomas. Both of these rates are consistent with published reports, and suggest that ROS1 fusions show tissue specificity.

While many oncogenic alterations cluster within diseases and are targetable directly, some inhibitors rely on the status of TSGs as biomarkers of response. For example, deleterious alterations in BRCA1/2 are associated with sensitivity to PARP inhibitors (51), and multiple trials require intact p53 (TP53) as enrollment criteria (NCT01760525, NCT02143635, NCT02264613). Therefore, we investigated patterns of TSG alterations. Unsupervised clustering of alterations within a curated list of TSGs (Supplementary Table S7) identified unique patterns of inactivation across solid tumors (Supplementary Fig. S21). Genes including TP53 and CDKN2A/B were altered uniformly across multiple solid tumors. In contrast, other TSGs displayed disease-specific clustering, such as APC alterations within GI cancers. Multiple novel disease-gene associations were also present. For example, alterations in BCOR, NOTCH1, KDM6A, CREBBP, and KMT2D clustered primarily in ACCs. Collectively, these data suggest patterns of TSG inactivation that may be disease specific, similar to patterns for some oncogenes. Further research to understand this tissue selectivity is warranted.

Discussion
We describe herein a dataset of 18,004 unique adult solid tumors that underwent genomic profiling as part of routine clinical care. This collection represents 'real world' specimens that were not selected for any features prior to sequencing. The dataset was composed of 162 disease subtypes, including many rare and unusual tumors not included previously as part of larger...
variants. Comparison of alteration frequencies to TCGA, where possible, identified some significant differences, mostly in CN frequencies. Detailed examination suggests that both technical (i.e., platform, annotation) and sample differences underlie between the discrepancies between the two datasets.

We also observed an enrichment of treatment refractory samples in FM breast and lung cancer cohorts based on an increased frequency of alterations associated with acquired resistance to targeted therapies in these diseases. To exemplify novelty within the FM dataset, we surveyed the genomic landscape of rare diseases and identified NOTCH1 alterations in ACCs at a higher frequency compared with previous studies. We also identified multiple potentially druggable novel kinase fusions as well as known fusions in diseases beyond those in which they are currently recognized. Analysis of VUSes identified a clinically significant enrichment of SMAD4 alterations in colon cancer, as well as multiple other rare alterations predicted to have functional impact. A survey of clinically relevant alterations highlighted the spectrum of molecular changes for which testing is recommended as well as opportunities for expansion of approved targeted therapies. Clustering of alterations in TSGs revealed patterns of disease specificity for certain genes, but broad inactivation of others. This dataset is rich with discovery potential and presents a new resource in which to investigate rare alterations and diseases, validate clinical relevance, and identify novel therapeutic targets.

To our knowledge, this dataset represents the largest collection of tumors to date profiled on a single uniform platform. The high unique sequencing coverage (>600×) across all targets enables accurate detection of all classes of genomic variants, even in impure clinical specimens. Previous validation has optimized this assay for sensitive and specific detection for all classes of variants down to low mutant allele frequencies (11). The samples within this dataset lack sequencing of patient-matched normal tissue, but multiple steps have been taken to enrich for significant cancer-associated variants (see Supplementary Methods). These include inclusion of (i) all truncation events in TSGs, (ii) known pathogenic germline events, and (iii) uncharacterized alterations reported previously in cancer (Supplementary Fig. S2). To minimize the number of benign germline variants, those variants not meeting the criteria above were filtered through online databases (ExAC and 1000 Genomes) to remove events recognized currently as benign polymorphisms. Collectively, the uniformity of the data and the stringent filtering to enrich for cancer-associated alterations facilitate comparisons and enhance the discovery potential for variants contributing to tumorigenesis.

The dataset can be used by basic researchers to identify novel findings for validation and to validate previous observations, especially those involving rare diseases and rare variants. Our preliminary analyses exemplify how hypothesis-generating discoveries within this cohort can be integrated with existing data. For example, the identification of novel SMAD4 genomic alterations in colorectal adenocarcinomas was expanded using the TCGA cohort to investigate survival differences among patients. In addition, MET amplifications in glioblastoma were shown to correlate with mRNA overexpression of this target in TCGA, a finding that may have been unappreciated in the past due to small datasets and the rarity of the event. A data collection of this size also allows for pan-cancer analyses to better understand tissue-specific patterns of alterations, such as TSG inactivation. These findings can be used to plan thoughtful functional follow-up experiments. This resource also has applicability to clinical oncology and drug development. We highlight the spectrum of clinically relevant molecular markers and show that a wide variety of alterations exist in targets for which established routine clinical tests exist. This resource can also be used to explore opportunities for drug expansion with new or approved targeted agents for whom biomarkers of therapeutic sensitivity or resistance are known. For example, we highlight multiple alterations, including MET amplification, ERBB2 amplification, and activating point mutations, and amplification of CDK4/6, that occur across multiple diseases. In contrast, we observed that ALK rearrangements and EGFR activating alterations are confined to specific diseases and very rarely observed outside of those tissues.

The lack of clinical data is a limitation of this dataset. Information about previous exposure to previous treatments, survival, and response rates was unavailable. As these samples were profiled on a clinical platform, and not as part of a research study, genomic information is only available for those 295 genomic targets deemed to have clinical relevance today. While the role of genomic changes in cancer development and treatment response is well studied, it is likely that other changes in methylation, expression, and noncoding DNA regions may have implications and would not be captured within this dataset.

The National Cancer Moonshot Initiative has emphasized that data sharing is essential to accelerate progress in oncology. Academic, private, and public sectors have an obligation to patients, researchers, and clinicians to share data, knowledge, and insight across the field. Large-scale sequencing projects have profiled many common tumors but often lack robust sample numbers for rare diseases and variants. The public availability of large genomic datasets, such as the one described herein, enables the broad use of this data across multiple disciplines, and is designed to remove barriers to progress. The insights gleaned from this data release will be instrumental in accelerating research and development efforts for targeted agents and immunotherapies.

Disclosure of Potential Conflicts of Interest

R.J. Hartmaier has received speakers’ bureau honoraria from Bio-Rad Laboratories and has ownership interest in Foundation Medicine. L. Albaker has ownership interest in Foundation Medicine. J. Chmielecki is a current employee at Foundation Medicine. M. Bailey has ownership interest in Foundation Medicine, Inc. J. He is a senior manager at Foundation Medicine. M. Goldberg is a clinical data analyst and has ownership interest in Foundation Medicine, Inc. J.A. Elvin has ownership interest in Foundation Medicine. G.M. Frampton reports receiving a commercial research grant and has ownership interest in Foundation Medicine. J.S. Ross is a medical director at Foundation Medicine and has ownership interest in Foundation Medicine. V. Miller is a chief medical officer at Foundation Medicine, Inc. P.J. Stephens is a chief scientific officer at Foundation Medicine, D. Lipson is a vice president and has ownership interest in Foundation Medicine. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions
Conception and design: R.J. Hartmaier, L.A. Albaker, J. Chmielecki, J. He, J.S. Ross, P.J. Stephens, D. Lipson
Development of methodology: R.J. Hartmaier, L.A. Albaker, J. He, M.E. Goldberg, G.M. Frampton, J.S. Ross, D. Lipson
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J. He, S. Ramkissoon, J. Suh, J.A. Elvin, G.M. Frampton, J.S. Ross

www.aacrjournals.org Cancer Res; 77(9) May 1, 2017 2473
Published OnlineFirst February 24, 2017; DOI: 10.1158/0008-5472.CAN-16-2479
Downloaded from cancerres.aacrjournals.org on October 22, 2017. © 2017 American Association for Cancer Research.
References


28. Gotlib J, Coils J, Malone JM, Schrier SI, Gilliland DG, Coutre SE. The HIP1L1-PTDGRalfa fusion tyrosine kinase in hypereosinophilic...


