nature medicine

Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine

Eliezer M Van Allen[1,2,8], Nikhil Wagle[1,2,8], Petar Stojanov[1,2], Danielle L Perrin[2], Kristian Cibulskis[2], Sara Marlow[1,2], Judit Jane-Valbuena[1,2], Dennis C Friedrich[2], Gregory Kryukov[2], Scott L Carter[2], Aaron McKenna[2,3], Andrey Sivachenko[2], Mara Rosenberg[2], Adam Kiezun[2], Douglas Voet[2], Michael Lawrence[2], Lee T Lichtenstein[2], Jeff G Gentry[2], Franklin W Huang[1,2], Jennifer Fostel[2], Deborah Farlow[2], David Barbie[1], Leena Gandhi[1], Eric S Lander[2], Stacy W Gray[1], Steven Joffe[1,4], Pasi Janne[1], Judy Garber[1], Laura MacConaill[1,5], Neal Lindeman[1,5], Barrett Rollins[1], Philip Kantoff[1], Sheila A Fisher[2], Stacey Gabriel[2,9], Gad Getz[2,6,7,9] & Levi A Garraway[1,2,9]

Translating whole-exome sequencing (WES) for prospective clinical use may have an impact on the care of patients with cancer; however, multiple innovations are necessary for clinical implementation. These include rapid and robust WES of DNA derived from formalin-fixed, paraffin-embedded tumor tissue, analytical output similar to data from frozen samples and clinical interpretation of WES data for prospective use. Here, we describe a prospective clinical WES platform for archival formalin-fixed, paraffin-embedded tumor samples. The platform employs computational methods for effective clinical analysis and interpretation of WES data. When applied retrospectively to 511 exomes, the interpretative framework revealed a 'long tail' of somatic alterations in clinically important genes. Prospective application of this approach identified clinically relevant alterations in 15 out of 16 patients. In one patient, previously undetected findings guided clinical trial enrollment, leading to an objective clinical response. Overall, this methodology may inform the widespread implementation of precision cancer medicine.

Massively parallel sequencing approaches such as WES have elucidated the landscape of genetic alterations in many tumor types and revealed biological insights relevant to clinical contexts[1]. The increased practical availability and decreased cost of tumor genomic profiling has generated opportunities to test the 'precision medicine' hypothesis in clinical oncology[2]. In principle, knowledge of alterations in the coding regions of all genes may inform immediate treatment choices and further therapeutic discovery efforts[3].

Most prospective clinical genotyping efforts have used 'hotspot' genotyping[4–6] or targeted sequencing panels of clinically relevant genes using either fresh frozen or formalin-fixed, paraffin-embedded (FFPE) tissue[7–9]. Pilot studies that apply research-grade massively parallel sequencing technology in focused clinical settings have also been reported[7,10–12], although production-scale efforts have not been demonstrated. Multiple challenges to widespread clinical WES implementation remain. One challenge involves rapidly generating high-quality WES data from archival FFPE tumor material[13]. Another involves clinically interpreting WES data for prospective use that maximizes clinical and biological exploration. A third involves developing a system to interrogate plausibly actionable variants of uncertain significance. Overcoming these challenges should allow rigorous assessment of the value of WES to guide clinical decision making and inform selected experimental follow-up.

Here, we describe an approach to generate high-quality WES data from archival tumor material and validate WES data from FFPE tumor samples with corresponding WES data from frozen samples. We also present a heuristic algorithm that interprets the resulting data for clinical oncologists and establish the clinical applicability of this interpretation algorithm in a retrospective cohort of 511 cases. Prospective application of this platform in patients with a range of tumor types indicates that this approach can be used for both biological discovery and clinical trial enrollment. This approach may therefore facilitate widespread application of WES for precision cancer medicine studies.

RESULTS

WES of FFPE samples

To produce WES data for clinical use, robust sequencing data must frequently be generated from small quantities of archival FFPE tissue. To test this, we extracted DNA from 99 FFPE samples using the FFPE

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [4]Children's Hospital Boston, Boston, Massachusetts, USA. [5]Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA. [6]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. [7]Cancer Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [8]These authors contributed equally to this work. [9]These authors jointly supervised this work. Correspondence should be addressed to L.A.G. (levi_garraway@dfci.harvard.edu) or G.G. (gadgetz@broadinstitute.org).

Received 19 April 2013; accepted 14 October 2013; published online 18 May 2014; doi:10.1038/nm.3559

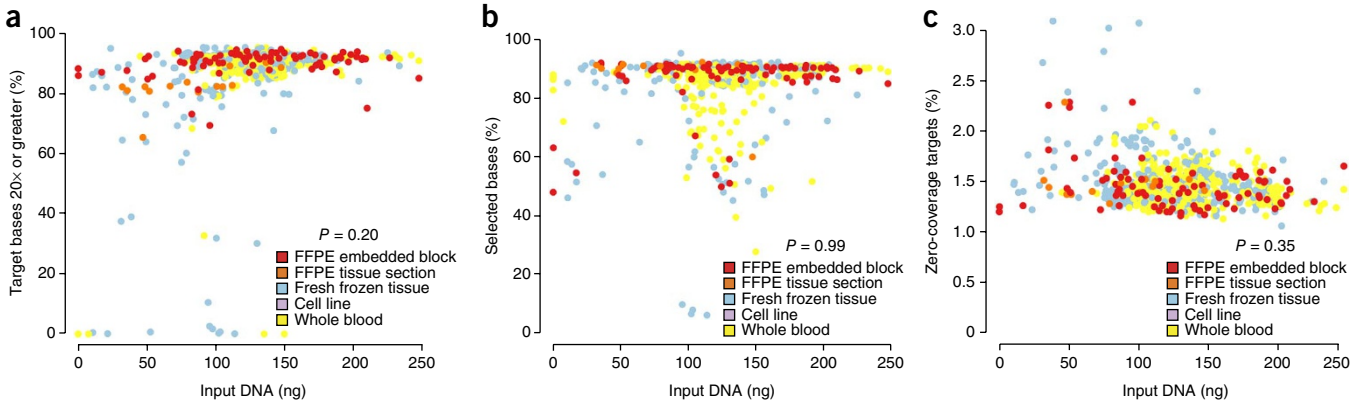


Figure 1 FFPE and frozen sample sequencing metrics. (**a**–**c**) The percentage of target bases covered at 20× (**a**), percentage of selected bases (**b**) and percentage of zero-coverage targets (**c**) in FFPE ($n = 99$) and non-FFPE ($n = 768$) tissue. Additional quality control metrics for all 867 cases are available in **Supplementary Table 1**. No statistically significant difference between FFPE and non-FFPE tissue was observed in these three metrics ($P > 0.05$; two-sided Mann-Whitney *U*-test).

extraction protocol (**Supplementary Table 1** and Online Methods). A comparison of standard WES metrics[14] with 768 non-FFPE samples (394 whole blood, 367 frozen, 7 cell line) sequenced in parallel demonstrated no significant differences independent of input DNA quantity ($P > 0.05$, Mann-Whitney *U*-test; **Fig. 1a–c** and **Supplementary Table 1**). Our lowest successful WES attempts were achieved with 13.6 ng and 16 ng DNA derived from non-FFPE and FFPE tissue, respectively.

Moreover, improvements in process design (Online Methods) combined with the 'with-bead' approach[14] yielded a time to exome data delivery of 17.4 ± 2.2 d (median ± s.d.; 25th and 75th percentiles 14.3 and 18.6, respectively) for FFPE samples received as DNA and 20.1 ± 2.4 d (median ± s.d.; 25th and 75th percentiles 17.5 and 21.2, respectively) for samples received as FFPE tissue blocks (**Supplementary Table 2**). This turnaround time is compatible with several clinical oncology applications.

We next assessed WES data using even smaller amounts of input DNA. Here, we achieved >80% of targeted nucleotides from the hybrid selection reaction, even when we used only 1 ng of input DNA; we saw equivalent results with DNA derived from FFPE and non-FFPE tissue. However, to meet our metrics of ≥80% targets with at least 20× coverage and ≥100× mean target coverage across the exome, a disproportionate amount of additional sequencing was required owing to an increase in the fraction of duplicate molecules in the library.

FFPE and fresh frozen samples yield comparable WES results

Next, we sought to compare WES data generated from FFPE and frozen material. We assessed WES data from 11 lung adenocarcinomas for which tumor and adjacent normal tissue were available from matched FFPE (aged ≤5 years, **Supplementary Table 3** and **Supplementary Figs. 1** and **2**) and frozen (**Fig. 2a**) samples. First, we applied our standard mutation detection pipeline on the tumor-normal pairs (Online Methods) and considered the concordance of mutation calls observed in FFPE tumors that we observed in frozen tumors and vice versa. We did not expect identical data, given tumor heterogeneity[15] and low allelic fraction nucleotide transition artifacts induced by the FFPE fixation process[16–18]. Moreover, the mean target coverage achieved for the FFPE tumor and adjacent tissue samples was 1.5–2 times that for the corresponding fresh frozen samples (**Supplementary Fig. 3**); as a result, we had increased power to detect mutations in FFPE samples compared to the fresh frozen samples[19]. Therefore, we considered the subset of observed exonic mutations in FFPE tumor cases where the depth of coverage afforded sufficient power (>95%) to detect the mutation in two or more reads in the matched frozen tumor case and vice versa. For sufficiently powered sites, 91.5% (2,923/3,194, 95% confidence interval (CI) ± 0.97) of mutations in FFPE samples were validated in patient-matched frozen samples. Similarly, 91.0% (3,399/3,735, 95% CI ± 0.92) frozen mutations were validated in sufficiently powered FFPE samples ($P = 0.47$) (**Fig. 2a–c** and **Supplementary Table 4**). Because the mean target coverage in the FFPE cases was higher than in their fresh frozen counterparts, we then obtained a random subset of reads from each case such that all sites had a maximum coverage of 90× ('downsampling'[19]) and repeated the cross-validation exercise. In this scenario, our validation rates for FFPE to fresh frozen and fresh frozen to FFPE for sufficiently powered sites were 92.6% (2,811/3,036, 95% CI ± 0.93) and 91.5% (3,340/3,651, 95% CI ± 0.90), respectively (**Supplementary Fig. 4a**,**b** and **Supplementary Table 4**).

In both FFPE and fresh frozen cases from each patient, we observed mutations for which there was insufficient power to detect that mutation in the validation cohort after downsampling (**Supplementary Fig. 4c** and **Supplementary Table 4**). Demonstrative examples of mutations in FFPE samples that could not be validated in fresh frozen counterparts are provided in **Supplementary Figure 5a–c**. Overall, these results suggested that the ability to detect base mutations that were sufficiently powered was equivalent regardless of whether frozen or FFPE tissue–derived genomic DNA was used for WES.

We also examined the chromosomal copy number patterns evident in WES data from frozen and FFPE tumor DNA in the 11 lung adenocarcinomas. In one representative patient, copy ratios for matching exons in FFPE and frozen sample data correlated ($r^2 = 0.89$, $P < 0.0001$, Pearson's correlation; **Fig. 2d**,**e**). This correlation held across all 11 cases, representing 1,338,859 exons ($r^2 = 0.79$, $P < 0.0001$, Pearson's correlation; **Fig. 2f**). Thus, WES data obtained from FFPE tumor DNA are comparable to fresh frozen sample WES data and may equally be used to measure global chromosome copy number information.

Clinical analysis and interpretation of WES data

Having demonstrated robust WES using FFPE tumor–derived DNA, we next sought to integrate this methodology into a broader framework for clinical interpretation of somatic alterations. We reasoned that a heuristic (rule-based) approach that incorporated prior clinical and scientific knowledge might offer a useful set of organizing principles. By using primary literature, manual curation and expert opinion,

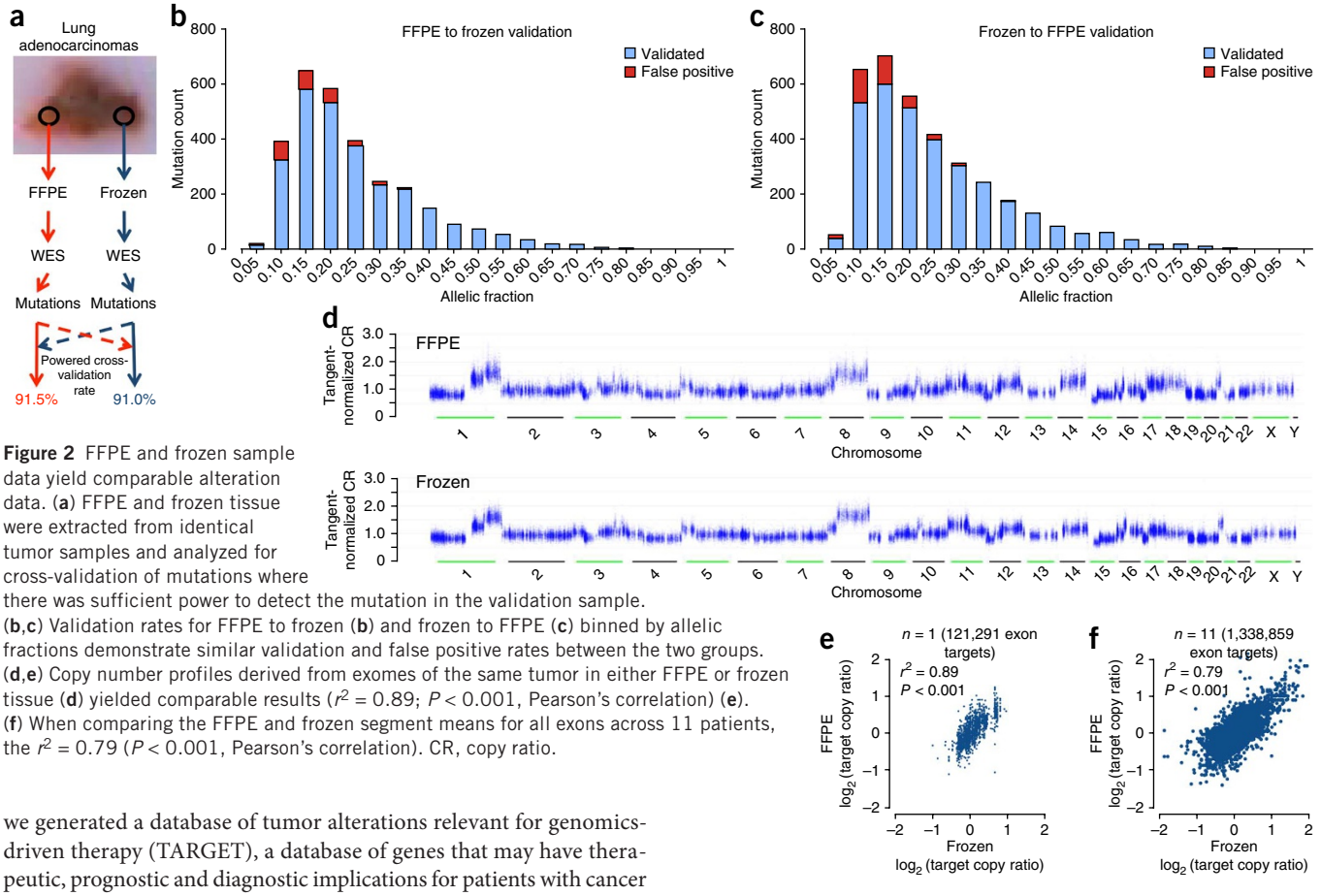


Figure 2 FFPE and frozen sample data yield comparable alteration data. (**a**) FFPE and frozen tissue were extracted from identical tumor samples and analyzed for cross-validation of mutations where there was sufficient power to detect the mutation in the validation sample. (**b**,**c**) Validation rates for FFPE to frozen (**b**) and frozen to FFPE (**c**) binned by allelic fractions demonstrate similar validation and false positive rates between the two groups. (**d**,**e**) Copy number profiles derived from exomes of the same tumor in either FFPE or frozen tissue (**d**) yielded comparable results ($r^2 = 0.89$; $P < 0.001$, Pearson's correlation) (**e**). (**f**) When comparing the FFPE and frozen segment means for all exons across 11 patients, the $r^2 = 0.79$ ($P < 0.001$, Pearson's correlation). CR, copy ratio.

we generated a database of tumor alterations relevant for genomics-driven therapy (TARGET), a database of genes that may have therapeutic, prognostic and diagnostic implications for patients with cancer (**Fig. 3b**, **Supplementary Table 5** and Online Methods). We integrated the resulting 121 TARGET genes with existing open-source resources to create a series of rules that (i) sort each somatic variant by clinical and biological relevance, (ii) link TARGET genes with additional biologically significant pathways and gene sets and (iii) demote variants of uncertain significance. Thus, the resulting analytical algorithm used precision heuristics for interpreting the alteration landscape (PHIAL) (**Fig. 3a–d** and Online Methods). Beyond annotating variants, PHIAL applies rules that rank variants on the basis of clinical and biological relevance to computationally sort a patient's somatic variants.

We assessed the functionality of PHIAL using 511 patient cases from six prior WES studies[20–25]. Analysis tools (Online Methods) yielded 258,226 somatic alterations in protein-coding genes, of which 135,903 were nonsynonymous. Of these, PHIAL identified 1,842 somatic alterations in genes linked to clinical actions (TARGET genes) for 80% (408/511) of the patients (**Fig. 3e**). Additional descriptive statistics regarding altered genes per patient, stratified by inclusion in databases explored in PHIAL, are available in **Supplementary Table 6**. PHIAL identified known and highly recurrent actionable findings across this patient cohort. It also revealed a long tail of TARGET gene alterations present in small patient subsets that did not reach statistical significance in the individual cohort studies but may have immediate clinical ramifications for individual patients (**Fig. 3f**). Specifically, 39% (201/511) of the cases had alterations in at least one TARGET gene that was somatically altered in <2% of the overall cohort. This finding was reminiscent of similar long-tail alteration distributions observed for driver genes in cancer[1].

As a major near-term goal of precision cancer medicine is to use genetic information to inform clinical trial enrollment, we also systematically queried ClinicalTrials.gov, a centralized registry of publicly and privately supported clinical studies worldwide, for oncology clinical trials linked to TARGET genes. The number of clinical trials including a TARGET gene in the title, the strictest means of identifying clinical trials with a genomic emphasis, grew steadily between 2005 and 2012 (**Fig. 3g**).

WES and clinically actionable events across cancers

To pilot prospective sequencing and clinical interpretation, we performed WES and PHIAL in 16 patients with a range of advanced cancers (**Fig. 4a**). WES data for 3 of these 16 patients predated the WES protocol described herein but were included to assess PHIAL output. WES data from all patients in the rapid sequencing protocol met our quality control parameters irrespective of tissue processing type (**Supplementary Table 7**). By completion of the pilot period, time from sample receipt through data delivery was 16 d.

For these 16 patients, PHIAL revealed 29 unique TARGET genes in the 'Investigate Clinical Relevance' category (median 2, range 0–5). Although, by definition, alterations in TARGET genes may have implications for clinical decision making, their actual clinical relevance requires case-by-case evaluation in real time. To facilitate this, we manually curated every alteration ranked as Investigate Clinical Relevance by PHIAL to include up-to-date knowledge from databases, literature and computational algorithms. We generated a standardized, structured annotation for each alteration (**Supplementary Note**) and assigned a level of evidence to each potential clinical action based on that alteration. These levels of evidence (**Supplementary Table 8**) included predictive, prognostic and diagnostic categories and encompassed validated indications, preclinical evidence and analytical associations.

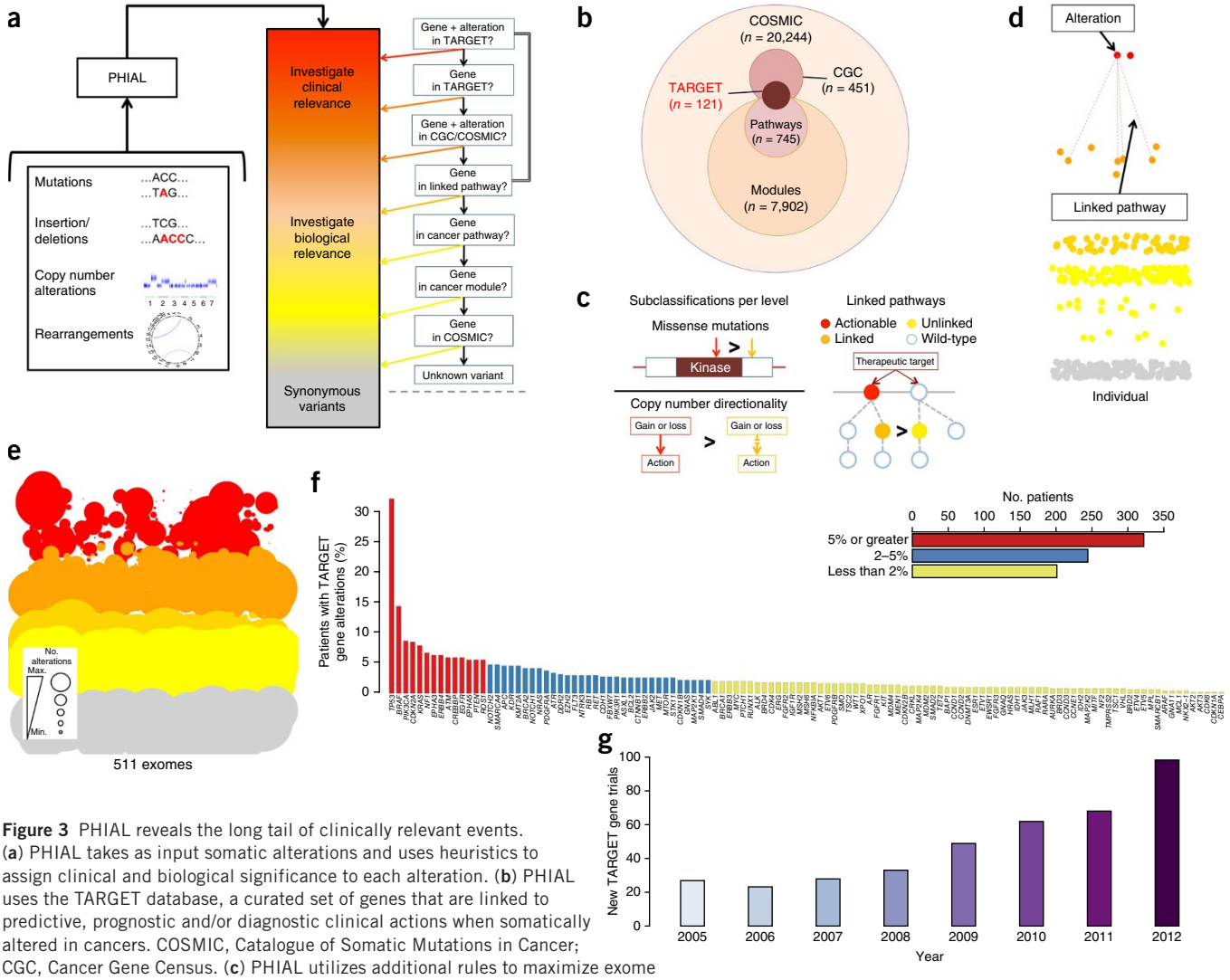


Figure 3 PHIAL reveals the long tail of clinically relevant events. (**a**) PHIAL takes as input somatic alterations and uses heuristics to assign clinical and biological significance to each alteration. (**b**) PHIAL uses the TARGET database, a curated set of genes that are linked to predictive, prognostic and/or diagnostic clinical actions when somatically altered in cancers. COSMIC, Catalogue of Somatic Mutations in Cancer; CGC, Cancer Gene Census. (**c**) PHIAL utilizes additional rules to maximize exome data for individuals, including knowledge about kinase domains, copy number directionality and two-hit pathway events. (**d**) The resulting data were visualized for individual or cohort-level information with this demonstrative PHIAL 'gel'. Each alteration is a point sorted by PHIAL score (top are of highest clinical relevance) and color coded by potential clinical relevance (red), biological relevance (orange), pathway relevance (yellow) or synonymous variants (gray). (**e**) A PHIAL gel for 511 patient exomes spanning six different disease types (n = 258,226 total somatic alterations). The size of the point is proportional to the number of times a given gene arises at that PHIAL score level. (**f**) This approach highlights the long tail of potentially clinically relevant alterations in TARGET genes (n = 121) that may be present in an individual patient but does not occur sufficiently to be labeled a biological driver across a cohort. The majority of events occur in genes that individually are altered in less than 2% of the overall cohort. (**g**) New cancer clinical trials with TARGET genes specifically integrated into the study per ClinicalTrials.gov over a 7-year period.

Following curation and assignment of levels of evidence, we identified 41 clinically relevant alterations in 15 out of 16 patients. These included standard-of-care findings, such as an *EGFR*^{L858R} mutation in lung adenocarcinoma linked to epidermal growth factor receptor (EGFR) inhibitors (predictive for US Food and Drug Administration (FDA)-approved therapies, level A), and *PIK3CA* alterations that are entry criteria for clinical trials (predictive for therapies in clinical trials, level A). 46.3% (19/41) of these alterations were based on preclinical evidence for the association of the alteration with response or resistance to FDA-approved therapies or therapies in clinical trials (level D) (**Fig. 4b** and **Supplementary Table 9**).

We identified multiple unexpected clinically relevant findings in genes not well characterized for the corresponding tumor type. For instance, we observed *CRKL* amplification in a patient with metastatic urothelial carcinoma (**Supplementary Fig. 6**); this alteration has been predicted to confer resistance to EGFR inhibitors[26] and sensitivity to Src inhibitors[27] in preclinical studies but had not previously been described in urothelial carcinoma. To accommodate new TARGET genes emerging with future findings, we have made TARGET publically available online (http://www.broadinstitute.org/cancer/cga/target) and encourage community contributions.

The use of WES in clinical decision making

We used the prospective WES framework for clinical decision making in one demonstrative case. A patient with metastatic lung adenocarcinoma underwent standard clinical genetic testing that revealed wild-type *EGFR*, *KRAS* (codon 12 and 13) and *ALK* status. Mass spectrometry testing of 471 alterations in 41 genes[5] revealed an *STK11* frameshift deletion. We started the patient on carboplatin, paclitaxel and bevacizumab (**Fig. 5a**). In parallel, we applied the clinical WES platform on the FFPE metastatic tumor sample and germline peripheral blood. PHIAL nominated a *KRAS*^{A146V} mutation

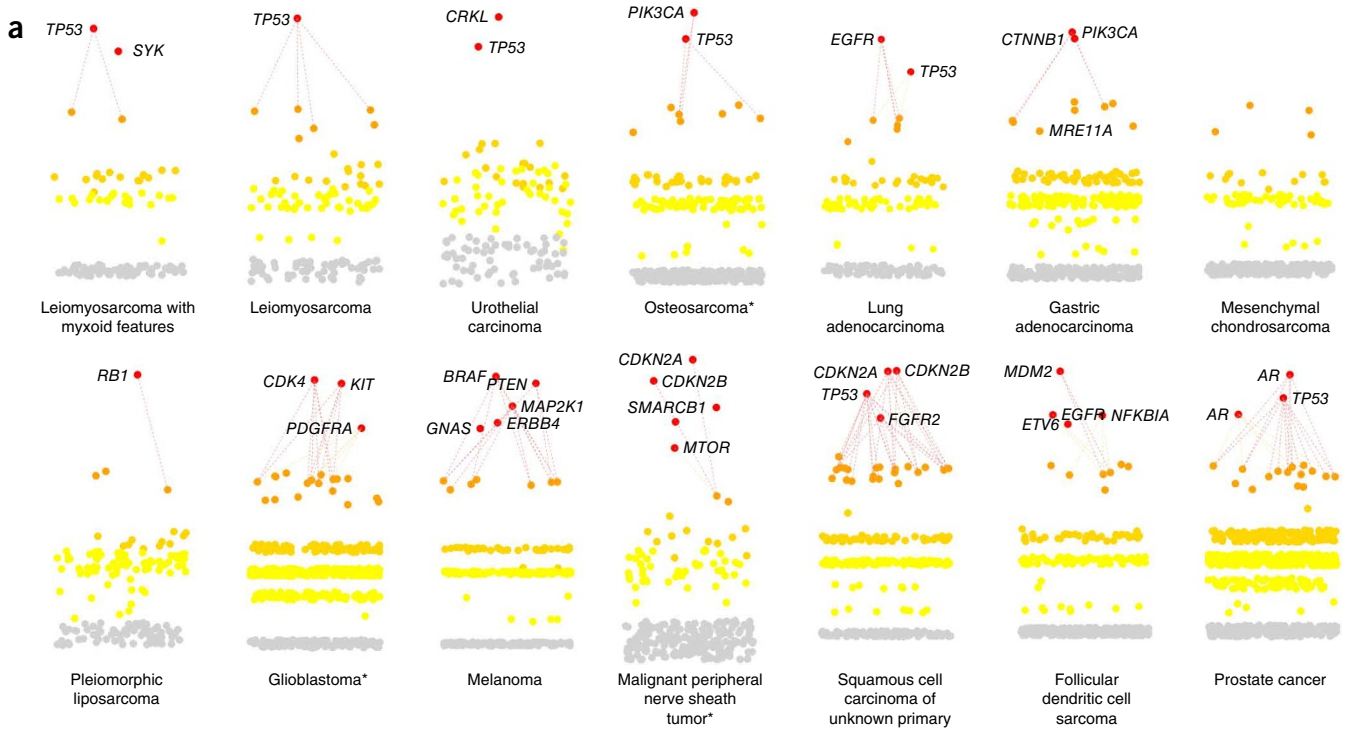


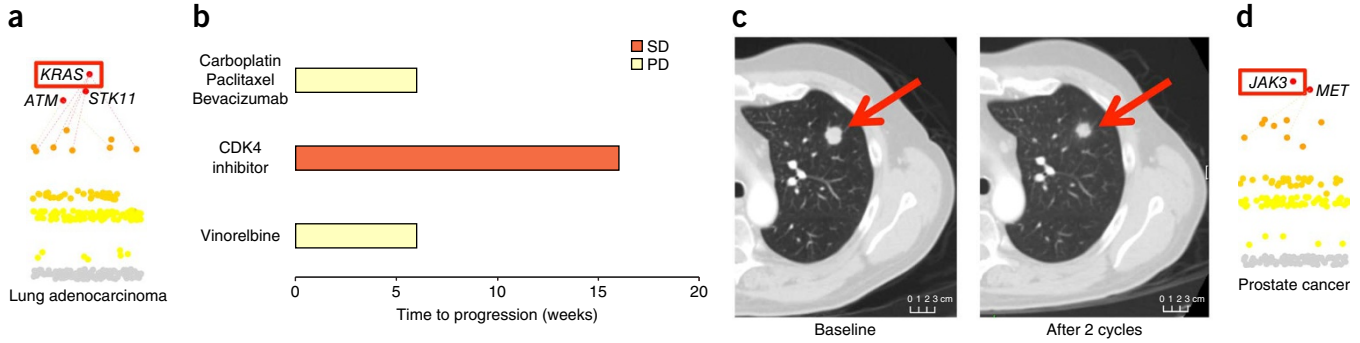
Figure 4 Clinically relevant findings from individual patients. (**a**) PHIAL results for 14 patients with a spectrum of malignancies, highlighting nominated clinically actionable alterations in 13 of 14 patients. Asterisks denote patient sequencing data that predated the rapid WES protocol. (**b**) Using the level of evidence schematic (**Supplementary Table 8**), all nominated alterations for patients in this study were manually curated and assigned a level of evidence (**Supplementary Table 7**).

as clinically relevant, along with alterations in *STK11* (identical to other testing) and *ATM* (**Fig. 5a** and **Supplementary Table 9**). *KRAS*^{A146V} is a known activating mutation, although it is possibly less potent than the codon 12 and 13 mutations[28]. Although activating *KRAS* mutations are found in 15–30% of all patients with non–small-cell lung cancer (NSCLC) and commonly in conjunction with *STK11* loss[29], this specific *KRAS* alteration has not been reported in NSCLC[20,30–32]. We confirmed *KRAS*^{146V} using the same FFPE tumor sample in a clinical lab (Online Methods) and then returned the data to the patient's oncologist. After rapidly progressing on combination chemotherapy (**Fig. 5b**), the patient was enrolled in a phase 1 clinical trial of a cyclin-dependent kinase 4 (CDK4) inhibitor (LY2835219) on the basis of preclinical data (level D) implicating a synthetic lethal relationship between activated *KRAS* and CDK4 (ref. 33). The patient achieved stable disease (per response evaluation criteria in solid tumors (RECIST) 1.1 criteria; 7.9% reduction in tumor volume compared to baseline) and was on therapy for 16 weeks (**Fig. 5b**,**c**). Of note, this was the patient's best and only clinical response to any cancer-directed therapy.

To maximize the potential of clinical WES, we also implemented a procedure to generate experimental evidence for selected level E (inferential association) alterations. An exemplary case involved WES in a patient with metastatic castration-resistant prostate cancer that harbored an R870W missense mutation in the gene encoding Janus kinase 3 (*JAK3*) (**Fig. 5d**). Activating mutations in *JAK3* have been described in hematological malignancies[34], and *JAK3* inhibitors are available clinically, including the FDA-approved agent tofacitinib. *JAK3*^{R870W} has not been previously identified in cancer, and the function of this mutation is unknown.

The crystal structure of JAK3 demonstrates that the arginine at residue 870 directly coordinates the phosphate group of the primary activating tyrosine phosphorylation site (pTyr981)[35] (**Fig. 5e**). This interaction is expected to pull JAK3 into the active conformation. Indeed, residue 870 is conserved as an arginine or lysine in virtually all JAKs. Given the functional importance of this residue, we hypothesized that this alteration could, in principle, be activating. Thus, we categorized this alteration as level E (**Supplementary Table 9**).

We used a Ba/F3 system to examine the activity of *JAK3*^{R870W} as compared to wild-type *JAK3* and a known activating mutation in *JAK3*, A572V[36]. Ba/F3 cells are mouse hematopoietic cells dependent on interleukin-3 (IL-3) for survival. Expression of some oncoproteins substitutes for IL-3 signaling, allowing for the growth of Ba/F3 cells in the absence of IL-3. This system has been used extensively to characterize activating mutants of *JAK3* in prior studies[36]. Ba/F3 cells expressing *JAK3*^{R870W} did not achieve IL-3–independent growth following complete IL-3 withdrawal, in contrast to cells expressing a known JAK3 activating mutation (JAK3^{A572V}) or those growing in the presence of IL-3 (**Fig. 5f**). This suggested that *JAK3*^{R870W} is unlikely to be an activating mutation and that *JAK3* inhibitors are unlikely to benefit this patient.



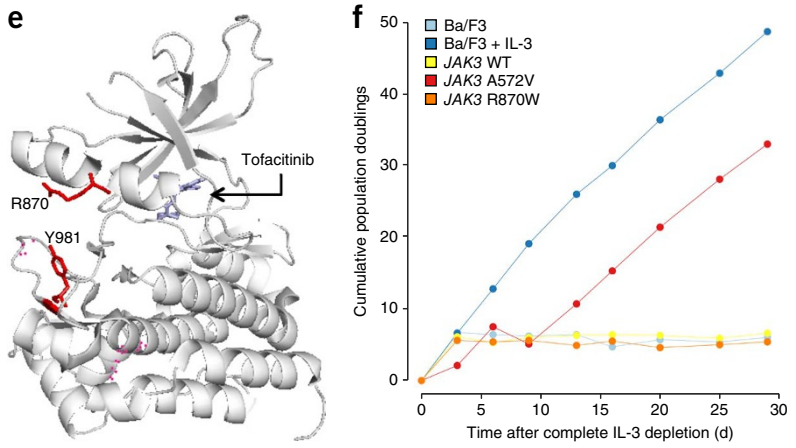


Figure 5 Clinical sequencing informs clinical trial enrollment and experimental discovery. (**a**) The PHIAL output and treatment course for a patient with metastatic lung adenocarcinoma is shown, with the integration of clinical WES occurring during the patient's first-line therapy allowing subsequent clinical trial enrollment. (**b**) The patient's time-to-relapse data for the three treatment regimens received. (**c**) Computed tomography radiographic imaging of a representative metastatic focus for the patient on the CDK4 inhibitor trial after two cycles of therapy (measurement is 1.7 × 1.5 cm for baseline mass and 1.3 × 1.3 cm for 2-month interval scan of the same mass). Per RECIST criteria, overall tumor reduction was 7.9%. (**d**) For another patient, PHIAL nominated a *JAK3* missense mutation, and given its location in the kinase domain near alterations previously defined as activating, was considered to have inferential evidence (level E) for being clinically actionable. (**e**) The crystal structure of JAK3 highlighting the arginine at residue 870 which directly coordinates the phosphate group of the primary activating tyrosine phosphorylation site. (**f**) Experimental follow-up of this alteration was performed in a Ba/F3 system compared to wild-type or a known activating JAK3 mutation (A572V).

DISCUSSION

This study demonstrates that rapid WES can be applied to FFPE clinical samples and that robust WES analysis and interpretation can prospectively inform clinical trial enrollment. This approach incorporates new algorithms to identify clinically relevant alterations among numerous somatic events. Furthermore, real-time curation of nominated alterations assigns levels of evidence to the corresponding clinical actions for that alteration in that tumor type. In a proof-of-concept application, we identified at least one clinically relevant alteration in 15 of 16 patients and showed how such findings can lead to clinical trial enrollment and biological discovery.

Targeted sequencing of clinically relevant gene panels (containing hundreds of genes) have recently become possible from FFPE tumor samples[7] and are increasingly used clinically. However, there are numerous advantages to clinical WES over targeted sequencing. First, as the spectrum of clinically actionable alterations grows[2], targeted sequencing of particular genes is likely to be incomplete: the rapid pace of drug development linked to a growing number of clinically relevant genes will probably outpace the ability to alter targeted sequencing approaches in real time, while, as the same time, performing clinical WES becomes more facile and cost efficient. The completeness of clinical WES also enables longitudinal queries if new clinical trials open for previously unrecognized cancer genes not acted on therapeutically during the initial evaluation.

Furthermore, we expect the volume of inferentially actionable or unknown-significance alterations will rise as more patient exomes emerge clinically. Clinical WES allows the generation of deeply annotated genomic data (linked to outcomes and responses) that could be mined to inform TARGET entries. We recognize that the pace of cancer discovery will necessitate continual TARGET updates to ensure its relevance, and we encourage input from the clinical and scientific community to expand and update its content for all to benefit. Methods to aggregate such data in a systems biology approach[37] are being developed to foster functional and clinical follow-up[38,39].

There are ways to improve upon the framework. Efforts to further minimize the input DNA requirement and predict which samples yield successful WES will improve production-level sequencing. This process will be enhanced by pathology review of clinical samples to enrich tumor DNA selection. Improvements in exome-derived copy number algorithms will better distinguish homozygous from heterozygous deletions in stromally admixed tumor samples. Integration of additional profiling technologies (for example, transcriptome profiling) will provide increasingly complex views of an individual's cancer and incorporate other changes (for example, epigenetic) that may have clinical relevance. In parallel, efforts to demonstrate the utility of massively parallel sequencing platforms in larger prospective clinical settings are underway.

PHIAL is heuristic based; a probabilistic model that assesses alteration clonality with preclinical data may better inform the functional impact of WES findings for individual patients. Even with predictive models, sequencing will frequently identify previously uncharacterized alterations in known genes. Furthermore, relevant information about known genomic alterations is constantly changing, and the availability of new therapies and clinical trials is in rapid flux. Because of this, alteration interpretation presently requires real-time manual curation, which requires dedicated and skilled resources that would

benefit from crowdsourcing efforts such as those we are establishing with TARGET and PHIAL.

Finally, rapid experimental validation of level E alterations to understand their clinical relevance will require large-scale innovations of scale to accelerate functional follow-up. Our experimental efforts described here establish a priority biological evaluation system for one type of functional assessment. A flexible experimental follow-up system to comprehensively assess any alteration will need to be developed.

With the 'start-to-finish' approach for clinical WES described here, it may be possible to implement these methods widely and facilitate routine WES in clinical oncology. Once implemented, this will enable the prospective study of patients in trials to determine whether large-scale genomic profiling improves patient care and, ultimately, outcomes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. WES BAM files for exome sequencing data from this study are deposited in dbGaP with accession codes phs000488.v1.p1 for lung adenocarcinoma cases and phs000694.v1.p1 for clinical cases.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the patients and clinicians for their participation in this project. We thank the Broad Genomics Platform (specifically K. Anderka, A. Cheney, E. Wheeler, T. Mason and C. Crawford). We thank C. Sougnez for facilitating data deposition. We also thank A. Lane and A. Yoda (Dana-Farber Cancer Institute), for their contributions to the *JAK3* experimental work: A. Lane provided Ba/F3 cells, and A. Yoda provided murine stem cell virus–puromycin vector. G.G. is partially funded by a Paul C. Zamecnik, MD, Chair in Oncology at Massachusetts General Hospital. This work was supported by the Starr Cancer Foundation (L.A.G.), the Prostate Cancer Foundation (E.M.V.A. and L.A.G.), US National Institutes of Health (NIH) NHGRI Clinical Sequencing Exploratory Research grant 1U01HG006492 (L.A.G.), the NIH National Cancer Institute grant 1U24CA126546 (L.A.G. and E.S.L.), the US Department of Defense (L.A.G.), NIH U24CA143845 grant (G.G.) and the Dana-Farber Leadership Council (E.M.V.A.).

AUTHOR CONTRIBUTIONS

All authors contributed extensively to the work presented in this paper. E.M.V.A. and N.W. contributed equally to this work. S.G., G.G. and L.A.G. contributed equally to this work. D.L.P., D.C.F., J.F., E.S.L., S.A.F., E.M.V.A. and S.G. contributed to FFPE sample sequencing protocols and analysis of sequencing metrics. E.M.V.A., P.S., D.F., K.C., G.K., S.L.C., A.M., A.S., A.K., D.V., M.L., L.T.L., J.G.G., M.R. and G.G. contributed to computational analyses for FFPE versus frozen sample comparisons and analysis of WES data generally. E.M.V.A., N.W., G.K., F.W.H., S.W.G., S.J., P.J., J.G., L.M., N.L., B.R., P.K. and L.A.G. contributed to clinical analysis and interpretation methods and application. N.W., S.M., J.J.-V. and L.A.G. contributed to experimental follow-up of the *JAK3* mutation. D.B. and L.G. contributed clinical input for the patient case. All authors discussed the results and implications and commented on the manuscript at all stages.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Garraway, L.A. & Lander, E.S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
2. Garraway, L.A. Genomics-driven oncology: framework for an emerging paradigm. *J. Clin. Oncol.* **31**, 1806–1814 (2013).
3. Garraway, L.A. & Janne, P.A. Circumventing cancer drug resistance in the era of personalized medicine. *Cancer Discov.* **2**, 214–226 (2012).
4. Thomas, R.K. *et al.* High-throughput oncogene mutation profiling in human cancer. *Nat. Genet.* **39**, 347–351 (2007).
5. MacConaill, L.E. *et al.* Profiling critical cancer gene mutations in clinical tumor samples. *PLoS ONE* **4**, e7887 (2009).
6. Dias-Santagata, D. *et al.* Rapid targeted mutational analysis of human tumours: a clinical platform to guide personalized cancer medicine. *EMBO Mol. Med.* **2**, 146–158 (2010).
7. Wagle, N. *et al.* High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov.* **2**, 82–93 (2012).
8. Lipson, D. *et al.* Identification of new *ALK* and *RET* gene fusions from colorectal and lung cancer biopsies. *Nat. Med.* **18**, 382–384 (2012).
9. Beltran, H. *et al.* Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity. *Eur. Urol.* **63**, 920–926 (2013).
10. Roychowdhury, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.* **3**, 111ra121 (2011).
11. Craig, D.W. *et al.* Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. *Mol. Cancer Ther.* **12**, 104–116 (2013).
12. Kerick, M. *et al.* Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genomics* **4**, 68 (2011).
13. Goetz, L., Bethel, K. & Topol, E.J. Rebooting cancer tissue handling in the sequencing era: toward routine use of frozen tumor tissue. *J. Am. Med. Assoc.* **309**, 37–38 (2013).
14. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
15. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
16. Pao, W., Ladanyi, M. & Miller, V.A. Erlotinib in lung cancer. *N. Engl. J. Med.* **353**, 1739–1741 (2005).
17. Williams, C. *et al.* A high frequency of sequence alterations is due to formalin fixation of archival specimens. *Am. J. Pathol.* **155**, 1467–1471 (1999).
18. Spencer, D.H. *et al.* Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. *J. Mol. Diagn.* **15**, 623–633 (2013).
19. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
20. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
21. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
22. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
23. Barbieri, C.E. *et al.* Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
24. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
25. Lohr, J.G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci. USA* **109**, 3879–3884 (2012).
26. Cheung, H.W. *et al.* Amplification of *CRKL* induces transformation and epidermal growth factor receptor inhibitor resistance in human non-small cell lung cancers. *Cancer Discov.* **1**, 608–625 (2011).
27. Natsume, H. *et al.* The *CRKL* gene encoding an adaptor protein is amplified, overexpressed, and a possible therapeutic target in gastric cancer. *J. Transl. Med.* **10**, 97 (2012).
28. Janakiraman, M. *et al.* Genomic and biological characterization of exon 4 *KRAS* mutations in human cancer. *Cancer Res.* **70**, 5901–5911 (2010).
29. Carretero, J. *et al.* Integrative genomic and proteomic analyses identify targets for Lkb1-deficient metastatic lung tumors. *Cancer Cell* **17**, 547–559 (2010).
30. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
31. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075 (2008).
32. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121–1134 (2012).
33. Puyol, M. *et al.* A synthetic lethal interaction between K-Ras oncogenes and Cdk4 unveils a therapeutic strategy for non-small cell lung carcinoma. *Cancer Cell* **18**, 63–73 (2010).
34. Levine, R.L. JAK-mutant myeloproliferative neoplasms. *Curr. Top. Microbiol. Immunol.* **355**, 119–133 (2012).
35. Boggon, T.J., Li, Y., Manley, P.W. & Eck, M.J. Crystal structure of the Jak3 kinase domain in complex with a staurosporine analog. *Blood* **106**, 996–1002 (2005).
36. Malinge, S. *et al.* Activating mutations in human acute megakaryoblastic leukemia. *Blood* **112**, 4220–4226 (2008).
37. Gonzalez-Angulo, A.M., Hennessy, B.T. & Mills, G.B. Future of personalized medicine in oncology: a systems biology approach. *J. Clin. Oncol.* **28**, 2777–2783 (2010).
38. Yeh, P. *et al.* DNA-mutation Inventory to Refine and Enhance Cancer Treatment (DIRECT): a catalogue of clinically relevant cancer mutations to enable genome-directed cancer therapy. *Clin. Cancer Res.* **19**, 1894–1901 (2013).
39. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

ONLINE METHODS

Patient samples. Tumor and germline samples used for this study were obtained under approved protocols from the Dana-Farber/Harvard Cancer Center Institutional Review Board, the Peter MacCallum Cancer Center Ethics Committee and the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects. Written informed consent was obtained from all subjects. Patient characteristics are described in **Supplementary Table 7**.

Rapid formalin-fixed, paraffin-embedded sequencing. Using industrial best practices in workflow design and a value-add approach, the standard exome workflow was modified to minimize touch points, handoffs and wasted process steps. Next, optimizations were made to the library construction and in-solution hybridization protocols to enable a 17-h hybridization reaction, 55 h shorter than the standard 72-h hybridization reaction.

FFPE DNA extraction. Paraffin is removed from FFPE sections and cores using CitriSolv (Fisher Scientific) followed by ethanol washes, and then tissue is lysed overnight at 56 °C. Samples are then incubated at 90 °C to remove DNA crosslinks, and extraction is performed using Qiagen's QIAamp DNA FFPE Tissue Kit.

Library construction. This was performed as previously described[14] with the following modifications: initial genomic DNA input into shearing was reduced from 3 µg to 10–100 ng in 50 µL of solution. For adaptor ligation, Illumina paired-end adapters were replaced with palindromic forked adapters, purchased from Integrated DNA Technologies, with unique 8-base molecular barcode sequences included in the adaptor sequence to facilitate downstream pooling. With the exception of the palindromic forked adapters, the reagents used for end repair, A-base addition, adaptor ligation and library enrichment PCR were purchased from KAPA Biosciences in 96-reaction kits. In addition, during the postenrichment solid phase reversible immobilization (SPRI) bead cleanup, elution volume was reduced to 20 µL to maximize library concentration, and a vortexing step was added to maximize the amount of template eluted from the beads. Any libraries with concentrations below 40 ng/µl, as measured by a PicoGreen assay automated on an Agilent Bravo, were considered failures and reworked from the start of the protocol.

In-solution hybrid selection. In-solution hybrid selection was also performed as previously described[14] with the following modifications to the hybridization reaction. Before hybridization, any libraries with concentrations >60 ng/µL, as determined by PicoGreen, were normalized to 60 ng/µL, and 8.3 µL of library was combined with blocking agent, bait and hybridization buffer. Any libraries with concentrations between 50 and 60 ng/µL were normalized to 50 ng/µL, and 10.3 µL of library was combined with blocking agent, bait and hybridization buffer. Any libraries with concentrations between 40 and 50 ng/µL were normalized to 40 ng/µL, and 12.3 µL of library was combined with blocking agent, bait and hybridization buffer. Regardless of library concentration range, the same volume of blocking agent and bait previously described[14] were used, and hybridization buffer volume was adjusted to equal the combined volume of library, blocking agent and bait. Finally, the hybridization reaction was reduced to 17 h with no changes to the downstream capture protocol.

Preparation of libraries for cluster amplification and sequencing. After postcapture enrichment, libraries were quantified using PicoGreen (automated assay on the Agilent Bravo), normalized to equal concentration on the PerkinElmer MiniJanus and pooled by equal volume on the Agilent Bravo. Library pools were then quantified using quantitative PCR (kit purchased from KAPA Biosystems) with probes specific to the ends of the adapters; this assay was automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 2 nM and then denatured using 0.2 N NaOH on the PerkinElmer MiniJanus. After denaturation, libraries were diluted to 20 pM using hybridization buffer purchased from Illumina.

Cluster amplification and sequencing. Cluster amplification of denatured templates was performed according to the manufacturer's protocol (Illumina) HiSeq v3 cluster chemistry and flowcells, as well as Illumina's Multiplexing Sequencing Primer Kit. Flowcells were sequenced using HiSeq 2000 v3 Sequencing-by-Synthesis Kits and then analyzed using RTA v.1.12.4.2 or later. Each pool of whole-exome libraries was run on paired 76-bp runs, and 8-base index sequencing read was performed to read molecular indices across the number of lanes needed to meet coverage for all libraries in the pool.

DNA assembly and quality control. *Sequence data processing.* Exome sequence data processing was performed using established pipelines at the Broad Institute. A BAM file was produced with the Picard pipeline (http://picard.sourceforge.net/), which aligns the tumor and normal sequences to the hg19 human genome build using Illumina sequencing reads. The BAM was uploaded into the Firehose pipeline (http://www.broadinstitute.org/cancer/cga/Firehose/), which manages input and output files to be executed by GenePattern[40]. Whole-exome sequencing BAM files for data from this study were deposited in dbGAP (phs000488 for lung adenocarcinoma cases; phs000694 number pending for clinical cases).

Sequencing quality control. Quality control modules within Firehose were applied to all sequencing data for comparison of the origin for tumor and normal genotypes and to confirm fingerprinting concordance. Cross-contamination of samples was estimated using ContEst[41] to confirm that neither tumor nor germline sample had >3% contamination. Single nucleotide polymorphism fingerprints from each lane of a tumor and normal pair were cross-checked to confirm concordance, and nonmatching lanes were removed from analysis.

Somatic alteration identification and annotation. The MuTect algorithm[19] was applied to identify somatic single-nucleotide variants in targeted exons. Indelocator (http://www.broadinstitute.org/cancer/cga/indelocator/) was applied to identify small insertions or deletions. Annotation of identified variants was done using Oncotator (http://www.broadinstitute.org/cancer/cga/oncotator/). Rearrangments were identified using dRanger (http://www.broadinstitute.org/cancer/cga/breakpointer/)[42]. Copy ratios were calculated for each hybrid capture bait by dividing the tumor coverage by the median coverage obtained in a set of reference normal samples[43]. The resulting copy ratios were segmented using the circular binary segmentation algorithm[44]. Genes in copy ratio regions with segment means of greater than 2 were evaluated for focal amplifications given the potential clinical significance of a large focal event. Genes in regions with segment means of less than −1 were evaluated for hemizygous or homozygous deletions, as either broad or focal deletions may involve genes with clinical relevance. RefSeq[45] was used to identify the genes that reside in the chromosomal coordinates demarcated by the segment start and end points.

Cross-validation of formalin-fixed, paraffin-embedded and fresh frozen mutation data. FFPE sections were received as 15-µm slices (9 per sample), from 2007 to 2009. All FFPE samples were sequenced as described above with 100 ng of input DNA. Frozen tumor samples were sequenced according to established methods[14]. All downstream computational analysis methods for assembly, alignment, mutation and copy number alteration identification were identical to the pipelines described above. For the downsampling experiment, MuTect was rerun on all the cases with the 'downsample-to-coverage' parameter set to 90. Mutations in intronic regions were excluded. For cross-validation of mutations, validation power was defined as the probability to observe at least two alternative allele reads in the validation sample (given the allelic fraction, coverage in validation sample at that site and the assumption that the mutation should be present there).

Clinical gene database (TARGET). The TARGET (tumor alterations relevant for genomics-driven therapy) database included genes that, when altered somatically in cancers, met one of three criteria: (i) alterations in the gene predicted resistance and/or sensitivity to specific therapies, (ii) alterations in the gene had prognostic significance in a cancer type or (iii) alterations in the gene had diagnostic significance in a cancer type.

To build this database, we performed a systematic review of the primary literature, manually curated specific genes based on clinician input and consulted expert opinion. This resulted in a list of 121 genes that met at least one of the

three criteria required for entry into the TARGET database (**Supplementary Table 5**, accessible at http://www.broadinstitute.org/cancer/cga/target).

Somatic heuristic algorithm for interpretation (PHIAL). Each somatic variant was scored individually using a series of rules and then was considered in aggregate to determine relationships between alterations in the same patient (for example, linked pathways). First, variants in TARGET are ranked highest, with scoring modifications for known mutational hotspots (for example, *BRAF* V600E), missense mutations in protein kinase regions and copy number alterations with directionality known to have clinical impact (for example, *PTEN* deletion). To assign maximum granularity between alterations, additional rules assign priority on the basis of presence of recurrent alterations in the Cancer Gene Census[46], presence in the pathway of concurrently altered actionable genes in the same sample using curated cancer pathways from MSigDB[47], presence in known cancer pathways, gene sets or modules identified by MSigDB and finally presence in COSMIC[30]. All code for PHIAL was implemented using the R statistical package language and is available online (http://www.broadinstitute.org/cancer/cga/phial/).

Visual representation. A decision support tool built around the results was developed to allow curation team members and clinicians to engage the data with web-based resources integrated directly into the patient's results. The tool is built to convey effective clinical review with the minimum manual steps so that such a process can be scaled rapidly. The report structure was implemented using the Nozzle R package[48]. All clinically actionable relevant somatic variants were linked to search criteria in ClinicalTrials.gov.

Curation. Somatic alterations nominated by PHIAL as 'investigate clinical relevance' were assigned for curation by a team of oncology and genomics experts charged with answering a series of structured questions pertaining to each nominated variant to facilitate final review (**Supplementary Fig. 6**). A curated alteration required review of published data to determine which level of evidence could be assigned to a clinical action for the alteration (**Supplementary Table 8**).

Clinical trial data analyses. ClinicalTrials.gov (http://clinicaltrials.gov/) was accessed on 19 February 2013, and the search entry 'cancer' was used to extract all cancer-related clinical trials in the database. Duplicated trial entries and trials designated as 'Terminated' or 'Withdrawn' were excluded. Provided trial start-date dates (by year) were used to select all trials that were initiated between 2005 and 2012, and trial titles were queried using string matching in R for those that specifically mention TARGET genes in the title of the trial.

CLIA validation. $KRAS^{146V}$ was confirmed using the same FFPE tumor sample in a clinical lab that met Clinical Laboratory Improvement Amendments (CLIA) standards (Knight Diagnostic Laboratories, Oregon) before being returned to the patient's oncologist.

Ba/F3 experimental methods. *Cell culture.* HEK 293T cells were maintained in DMEM (Gibco) with 10% (vol/vol) FBS (Gibco). Ba/F3 cells (gift from A. Lane at Dana-Farber Cancer Institute) were maintained in RPMI 1640 (Gibco) with 10% FBS and 10 ng/mL mouse interleukin-3 (IL-3; Prospec).

Retroviral infections. The wild-type JAK3 cDNA cloned in the pDONR223 vector was obtained from The Broad Institute RNAi Consortium. JAK3 mutations were generated by site-directed mutagenesis using the QuikChange Lightning Mutagenesis Kit (Stratagene) and verified by full sequencing of the JAK3 cDNA insert. WT and mutant cDNAs were recombined into a Gateway adapted murine stem cell virus (MSCV)-puromycin vector (gift from A. Yoda at Dana-Farber Cancer Institute) using the Gateway LR Clonase kit. Ecotropic viruses were produced by cotransfection of MSCV constructs with pCL-Eco vector (Imgenex) in HEK 293T cells. Ba/F3 cells were plated in 6-well plates at a 30% confluency and spin-infected at 800*g* for 90 min at 33 °C in the presence of 8 μg/mL polybrene (hexadimethrine bromide; Sigma). The same infection protocol was repeated 24 h later. Upon completion, the viral supernatant was removed and fresh medium added. Twenty-four hours after the medium change, Ba/F3 cells were subjected to a 3-d puromycin selection (2 μg/mL) in the presence of IL-3. Expression of ectopic JAK3 protein was verified by immunoblot analysis using a primary antibody against phospho-JAK3 (1:500 dilution; Cell Signaling #5031).

IL-3 Depletion. Ba/F3 cells and Ba/F3 cells expressing WT and mutant forms of JAK3 were seeded in 25-cm^2 vented-cap flasks at 20,000 cells/ml in a total volume of 5 ml in the absence of IL-3 to select IL-3–independent cells. Cells were grown in the absence of IL-3 over several weeks. In parallel, Ba/F3 cells were maintained in 10 ng/ml IL-3 throughout as a positive control. Cell counts were recorded every 4 d using ViCell counter and split as needed.

Statistical analyses. *Statistical analysis of raw sequencing metrics.* All analyses of raw sequencing metrics were performed using the R statistical package. Sample size was established by incorporating all available FFPE samples sequenced under the FFPE sequencing protocol by the time of analysis freeze ($n = 99$). Significance between two means (FFPE and non-FFPE samples for the sequencing metrics) was calculated with the two-tailed Mann-Whitney *U*-test, given the non-normal distribution of values. $P < 0.05$ was considered significant.

Statistical analysis of formalin-fixed, paraffin-embedded and frozen tissue. Two-tailed Fisher's exact test was used to test the statistical significance of the contingency table represented by tissue type (FFPE or frozen) and validation status. Pearson's correlation was performed on $\log_2$(target copy ratio) segment mean data for FFPE and frozen exon targets, and significance was calculated using Pearson's product moment correlation coefficient. Sample size for exons from all 11 cases ($n = 1{,}338{,}859$) greatly exceeded the minimum sample size needed to determine a linear correlation coefficient of 0.8 with power of 0.8 and significance level of 0.05. The variance estimate among FFPE (0.054) and frozen (0.049) copy number signal data was similar. Whole-exome sequencing data for lung adenocarcinoma cases were deposited in dbGaP (phs000488.v1.p1).

40. Reich, M. *et al.* GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).
41. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601–2602 (2011).
42. Drier, Y. *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* **23**, 228–235 (2013).
43. Chiang, D.Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103 (2009).
44. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
45. Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res.* **37**, D32–D36 (2009).
46. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
47. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
48. Gehlenborg, N., Noble, M.S., Getz, G., Chin, L. & Park, P.J. Nozzle: a report generation toolkit for data analysis pipelines. *Bioinformatics* **29**, 1089–1091 (2013).

doi:10.1038/nm.3559