nature medicine

Anchored multiplex PCR for targeted next-generation sequencing

Zongli Zheng[1,2], Matthew Liebers[1], Boryana Zhelyazkova[1], Yi Cao[1], Divya Panditi[1], Kerry D Lynch[1], Juxiang Chen[1,3], Hayley E Robinson[1], Hyo Sup Shim[1,4], Juliann Chmielecki[5], William Pao[5], Jeffrey A Engelman[6], A John Iafrate[1,6] & Long Phi Le[1,6]

We describe a rapid target enrichment method for next-generation sequencing, termed anchored multiplex PCR (AMP), that is compatible with low nucleic acid input from formalin-fixed paraffin-embedded (FFPE) specimens. AMP is effective in detecting gene rearrangements (without prior knowledge of the fusion partners), single nucleotide variants, insertions, deletions and copy number changes. Validation of a gene rearrangement panel using 319 FFPE samples showed 100% sensitivity (95% confidence limit: 96.5–100%) and 100% specificity (95% confidence limit: 99.3–100%) compared with reference assays. On the basis of our experience with performing AMP on 986 clinical FFPE samples, we show its potential as both a robust clinical assay and a powerful discovery tool, which we used to identify new therapeutically important gene fusions: *ARHGEF2-NTRK1* and *CHTOP-NTRK1* in glioblastoma, *MSN-ROS1*, *TRIM4-BRAF*, *VAMP2-NRG1*, *TPM3-NTRK1* and *RUFY2-RET* in lung cancer, *FGFR2-CREB5* in cholangiocarcinoma and *PPL-NTRK1* in thyroid carcinoma. AMP is a scalable and efficient next-generation sequencing target enrichment method for research and clinical applications.

Next-generation sequencing has been instrumental in the advancement of genomic research and clinical molecular diagnostics in recent years. Although the cataloguing of complete genomes and their variation is an important endeavor for reference building and discovery, the use of whole-human-genome sequencing outside of this context is impractical with respect to cost and efficiency[1]. Certain applications such as cancer genotyping for somatic mutations require selective deep sequencing to achieve the desired analytical sensitivity for clinical utility[2]. At the present time, clinical sequencing is most feasible for assays based on targeted gene panels. The emerging need for a rapid and focused confirmation sequencing strategy to validate variants also remains to be addressed. Currently, there is need for a rapid and efficient technique for gene rearrangement detection by next-generation sequencing.

For clinical molecular diagnostics, we developed AMP to address the escalating demand for gene rearrangement testing of the *ALK* (encoding anaplastic lymphoma receptor tyrosine kinase), *RET* (encoding ret proto-oncogene) and *ROS1* (encoding ROS proto-oncogene 1) genes, all of which are associated with response to targeted therapy in lung cancer[3–5]. Fluorescence *in situ* hybridization (FISH) lacks scalability for high-volume multitarget testing and requires diagnostic expertise. Immunohistochemistry is used to detect expressed fusion genes as a surrogate marker for gene rearrangements; however, the technique relies on the availability of good-quality antibodies and on qualitative scoring. Neither FISH nor immunohistochemistry provide fusion partner breakpoint precision, which may partially explain heterogeneous treatment responses[3,6,7]. Reverse-transcription PCR may yield such information but requires knowledge of all fusion partner variants for primer design and demonstrates limited scalability in the setting of multiple heterologous partners and their involved exons. For example, *ROS1* rearrangements in lung cancer pose a challenge due to potential involvement with at least eleven different fusion partners and variable splicing[8].

RESULTS

Targeted RNA sequencing

Our initial motivation for designing AMP was to tackle the current deficiencies of clinical gene rearrangement detection noted above by employing a targeted RNA sequencing (RNA-seq) strategy. AMP is in theory similar to the technique known as rapid amplification of cDNA ends (RACE)[9], specifically in its ability to uncover unknown sequences adjacent to a known DNA sequence. Briefly, double-stranded cDNA undergoes end repair, adenylation and ligation, as previously described[10–12], with a new universal half-functional adapter. The resulting half-functional library by itself is insufficient for downstream bridge amplification, emulsion PCR or sequencing. The library is rendered fully functional at the end of two rounds of nested low-cycle PCR, which represent the core steps for target enrichment. The second PCR step uses nested primers that are 5′ tagged with a common sequencing adapter. In combination with the first half-functional universal adapter, the resulting target amplicons are functionalized for clonal amplification (for example, emulsion PCR or bridge PCR) and sequencing. Nontarget fragments remain

[1]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [3]Department of Neurosurgery, Shanghai Changzheng Hospital, Shanghai, China. [4]Department of Pathology, Yonsei University College of Medicine, Seoul, Korea. [5]Vanderbilt University Medical Center, Nashville, Tennessee, USA. [6]Cancer Center, Massachusetts General Hospital, Boston, Massachusetts, USA. Correspondence should be addressed to A.J.I. (aiafrate@partners.org) or L.P.L. (lple@partners.org).

Received 31 July 2013; accepted 29 July 2014; published online 10 November 2014; doi:10.1038/nm.3729

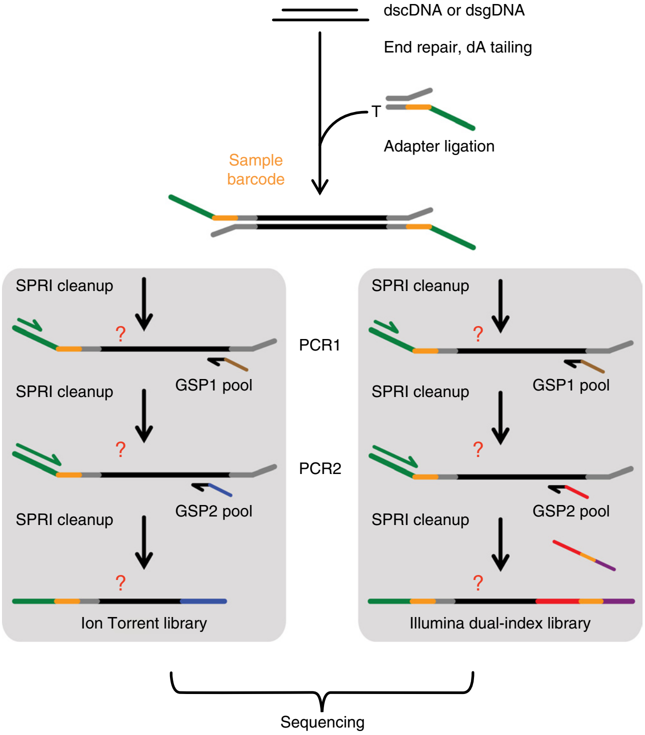

Figure 1 AMP for targeted RNA and DNA sequencing. Double-stranded cDNA (dscDNA) synthesis starts with total nucleic acid or RNA from fresh or FFPE material without ribosomal RNA or genomic DNA (gDNA) depletion. Solid-phase reversible immobilization (SPRI)-cleaned double-stranded cDNA or fragmented genomic DNA is processed with end repair and dA tailing, directly followed by ligation with a half-functional adapter. SPRI-cleaned, ligated fragments are amplified with 10–14 cycles of multiplex PCR1 using gene-specific primers (GSP1 pool) containing a PCR multiplexing tag (brown) and a primer complementary to a portion of the universal ligated adapter (short green). An unknown fusion partner or target sequence is indicated by a question mark. SPRI-cleaned PCR1 amplicons are amplified with a second round of 10-cycle multiplex PCR2 using a combination of GSP2 pool nested gene-specific primers (3′ downstream of GSP1), which are tagged with the second adapter sequence specific for Ion Torrent (black-blue) or Illumina (black-red, subsequently tagged with an indexing primer (red-orange-purple)), and a second nested primer against the ligated universal adapter (long green). After a final SPRI cleanup, the target amplicon library is ready for quantitation, downstream clonal amplification and sequencing.

half-functional (inconsequential) and need not be eliminated from the library. Libraries are quantitated and processed for Illumina or Ion Torrent sequencing (**Fig. 1**).

One powerful advantage of AMP is the anchoring of target-specific nested primers on one side while the other end is randomly ligated with the half-functional universal adapter. In contrast to other routine PCR techniques, AMP enables enrichment of a target region with knowledge of only one of its ends. We exploited this feature of AMP for targeted RNA-seq detection of gene rearrangements using total nucleic acid derived from clinical FFPE material (**Supplementary Fig. 1**). Gene fusion detection via targeted RNA-seq offers several advantages over genomic DNA sequencing, including expressed fusion transcript sequence information (such as in-frame status), a smaller target window, potentially easier unique alignment and confident fusion calls with deeper coverage.

We designed a single-tube 23-plex AMP panel to amplify the kinase domains of *ALK*, *RET*, *ROS1*, *MUSK* (encoding muscle, skeletal, receptor tyrosine kinase) and the *CTBP1* housekeeping gene (encoding C-terminal binding protein 1) as an internal control (**Supplementary Table 1**). We tested FISH-positive *ROS1* fusion cases using AMP and found multiple partners and alternative splicing events (**Fig. 2a**). Similarly, this assay successfully detected *EML4-ALK* and *KIF5B-RET* fusions in lung cancer and *CCDC6-RET* and *NOCA4-RET* fusions in thyroid cancer (**Fig. 2a**).

We developed a 137-amplicon expanded rearrangement AMP panel to detect possible fusions with 14 additional receptor tyrosine kinase genes, targeting from both the 5′ and 3′ ends of the kinase domains in a single-tube format. The results revealed an *FGFR3-TACC3* fusion in glioblastoma and an *ETV6-NTRK3* fusion in secretory breast carcinoma (**Supplementary Table 1**). We discovered two novel fusions in glioblastoma as a result of screening 115 brain tumors: one case with an in-frame fusion involving exon 21 of *ARHGEF2* (encoding Rho/Rac guanine nucleotide exchange factor 2) and exon 10 of *NTRK1* (encoding neurotrophic tyrosine kinase, receptor, type 1) and two cases with in-frame fusions involving exon 5 of *CHTOP* (encoding the chromatin target of PRMT1) and exon 10 of *NTRK1* (**Fig. 2b–e**). We confirmed these fusions by FISH (**Fig. 2f**) and RT-PCR, respectively, and they represent potential therapeutic targets for small-molecule inhibitors[13]. The two *CHTOP-NTRK1* cases harbored *IDH1* R132 mutations previously tested by the SNaPshot assay[14]. The *ARHGEF2-NTRK1* case showed *EGFR* (encoding epidermal growth factor receptor) amplification by FISH (data not shown). Additional novel fusion findings with potential therapeutic importance include *MSN-ROS1* (confirmed by a hybrid capture assay[15]), *TRIM4-BRAF*, *VAMP2-NRG1*, *TPM3-NTRK1* and *RUFY2-RET* in lung cancer, *FGFR2-CREB5* in cholangiocarcinoma and *PPL-NTRK1* in thyroid carcinoma (**Supplementary Table 2**).

AMP demonstrated superior performance with respect to clinical sensitivity and specificity when compared to our standard clinical FISH assays. Using total nucleic acid extracted from 319 FFPE samples, the 23-plex targeted RNA-seq assay detected 56 of 56 positive cases detected by FISH, which represents a clinical sensitivity of 100% (95% confidence limit: 96.5–100%). All 273 negative cases detected by FISH were negative by the targeted RNA-seq assay, resulting in a clinical specificity of 100% (95% confidence limit: 99.3–100%) (**Table 1**). One case was clinically reported as indeterminate by FISH on the basis of an unusual but abnormal pattern of tandem *ROS1* gene copies, with one copy showing an imbalance between the 5′ and 3′ signals (**Supplementary Fig. 2**). AMP definitively detected a *CD74-ROS1* rearrangement in this case, suggesting that targeted RNA-seq may be more specific than FISH. Unusual *ALK* and *ROS1* FISH cases showing individual 5′ probe signals (green only) present a diagnostic challenge for the clinical lab. Evaluation of these indeterminate cases with our assay did not show any typical *ALK* and *ROS1* fusion gene transcripts (data not shown), indicating the possibility of atypical structural changes detected by FISH that are not related to the known crizotinib-sensitive *ALK* and *ROS1* gene rearrangements. Of note, the green-only *ALK* FISH case (**Supplementary Fig. 3**) did not show increased ALK expression by immunohistochemistry (typically expected in an *ALK*-rearranged tumor) but instead harbored a *KRAS* mutation (usually mutually exclusive of *ALK* rearrangement). Similarly, the green-only *ROS1* FISH case (data not shown) showed no response to crizotinib after 8 weeks of treatment, indicating the biological absence of a canonical *ROS1* rearrangement.

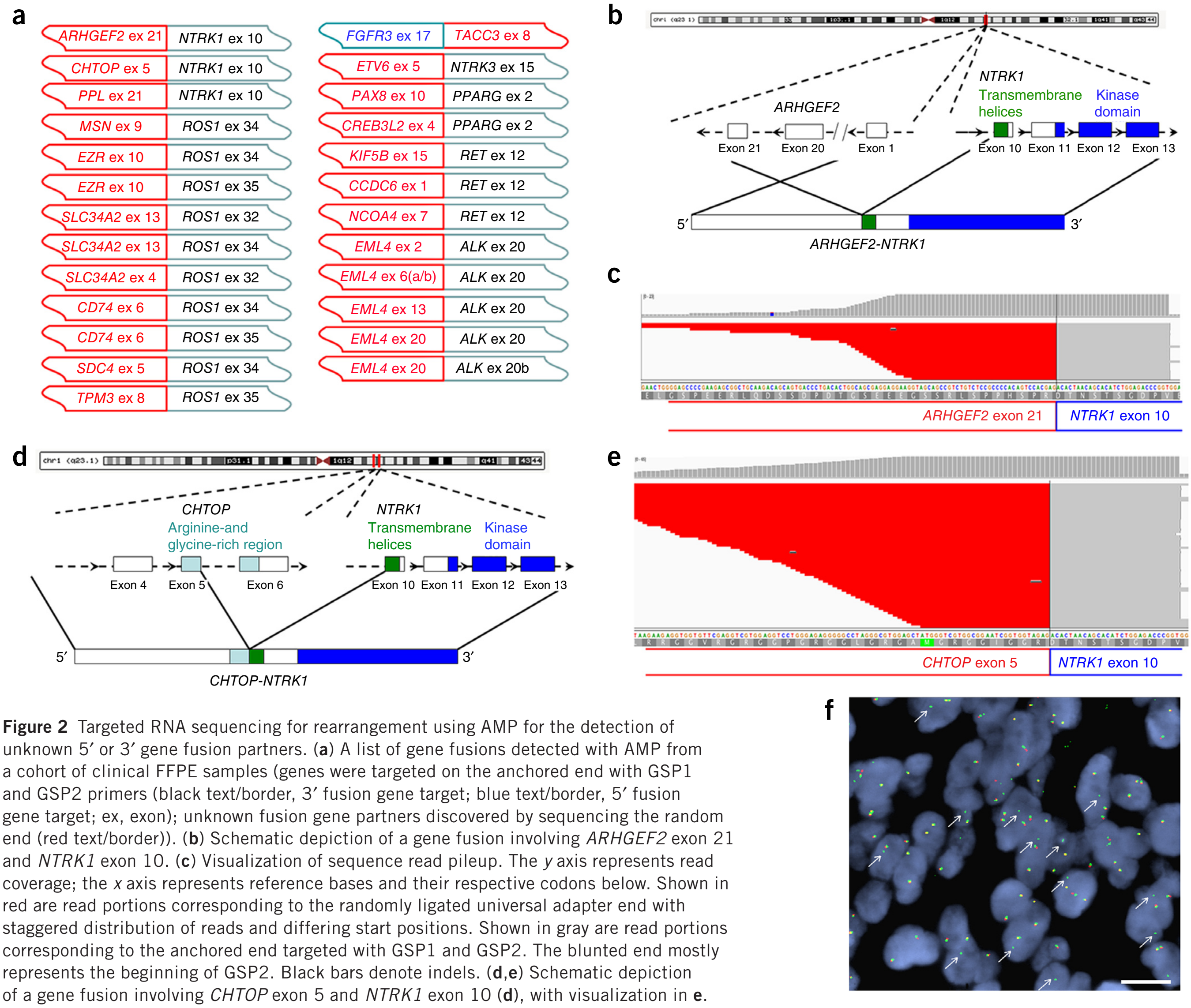

Figure 2 Targeted RNA sequencing for rearrangement using AMP for the detection of unknown 5′ or 3′ gene fusion partners. (**a**) A list of gene fusions detected with AMP from a cohort of clinical FFPE samples (genes were targeted on the anchored end with GSP1 and GSP2 primers (black text/border, 3′ fusion gene target; blue text/border, 5′ fusion gene target; ex, exon); unknown fusion gene partners discovered by sequencing the random end (red text/border)). (**b**) Schematic depiction of a gene fusion involving *ARHGEF2* exon 21 and *NTRK1* exon 10. (**c**) Visualization of sequence read pileup. The *y* axis represents read coverage; the *x* axis represents reference bases and their respective codons below. Shown in red are read portions corresponding to the randomly ligated universal adapter end with staggered distribution of reads and differing start positions. Shown in gray are read portions corresponding to the anchored end targeted with GSP1 and GSP2. The blunted end mostly represents the beginning of GSP2. Black bars denote indels. (**d**,**e**) Schematic depiction of a gene fusion involving *CHTOP* exon 5 and *NTRK1* exon 10 (**d**), with visualization in **e**. (**f**) FISH confirmation of the glioblastoma case harboring the *ARHGEF2-NTRK1* rearrangement showed individual 5′ (green only, white arrows) probe signals that represent a region upstream of the *NTRK1* gene along with normal paired green and red signals (scale bar, 10 μm).

Targeted DNA-seq

Because AMP works with any double-stranded input nucleic acid, we next determined its utility for targeted sequencing of genomic DNA. We evaluated assay performance (on-target specificity rate and minimum coverage across target bases) using two assays: a 626-amplicon assay for the coding regions of 18 important tumor suppressor genes, and a 96-amplicon assay for cancer hotspot mutations and the coding regions of three tumor suppressor genes, *TP53* (encoding p53), *PTEN* (encoding phosphatase and tensin homolog) and *CDKN2A* (encoding cyclin-dependent kinase inhibitor 2A) (**Table 2** and **Fig. 3a–d**). To avoid conventional PCR, we segregated the plus and minus primers into two reactions.

We assessed different PCR conditions using the 626-amplicon assay on a non-tumor FFPE sample (**Table 2**). PCR amplification bias among high-GC-content amplicons is greatly improved by using a slower ramping rate during PCR[16]. Therefore, we set a 20% ramping

Table 1 Clinical sensitivity and specificity of AMP for gene rearrangement detection compared to FISH

		ALK FISH			*ROS1* FISH			*RET* FISH		*PPARG* FISH	
		+	–	Indeterminate	+	–	Indeterminate	+	–	+	–
Targeted RNA-seq	+	19	0	0	23	0	1[a]	10	0	4	0
	–	0	88	1[b]	0	172	1[b]	0	0	0	0

A cohort of 319 archived clinical FFPE samples with available FISH results were tested with an AMP RNA-seq assay targeting the receptor tyrosine kinase domains of *ALK*, *RET*, *ROS1* and *PPARG*, showing 56 AMP-positive cases out relative to 56 FISH-positive cases and 273 AMP-negative cases out relative to 273-FISH negative cases (FISH indeterminate cases not included in calculation).

[a]One case was a clinical *ROS1* FISH case that was reported as indeterminate based on an unusual probe pattern (**Supplementary Fig. 2**). On the AMP assay, it was confirmed to have a *CD74-ROS1* gene fusion. [b]One *ALK* FISH case and one *ROS1* FISH case showed indeterminate green probe only results and tested negative by the targeted RNA-seq assay.

Table 2 Anchored multiplex PCR enrichment metrics and variant detection

Condition, input DNA (ng)[a]	Total reads	Post-trimming reads[b]	% aligned reads using BWA + BLAT[c]	% on-target BWA + BLAT[c]	Target 100× minimum coverage	Target 500× minimum coverage	Clinical testing results	AMP results[d] (dbSNP filtered; min. variant freq.: 5%)
AMP optimization with different polymerases and TMAC additive (626-amplicon gDNA panel[d])								
Platinum Taq, 500	2,610,297	2,608,220	99.5	87.8	97.1	93.4	Not applicable	Not applicable
Platinum Taq + TMAC, 500	2,516,838	2,513,743	99.5	89.2	97.1	94.3	Not applicable	Not applicable
OneTaq, 500	3,506,335	3,483,793	95.8	10.8	85.5	43.5	Not applicable	Not applicable
Phusion, 500	2,670,374	2,666,972	99.1	66.6	96.0	91.5	Not applicable	Not applicable
AMP performance on clinical FFPE samples with known genotypes (96-amplicon panel[e])								
200	642,338	632,186	99.0	83.4	98.8	93.5	Wild type	None
200	567,238	558,367	98.7	83.2	99.2	93.8	Wild type	None
200	529,623	520,241	99.0	84.2	99.7	94.6	Wild type	None
200	737,942	715,553	98.4	86.0	99.9	92.1	*KRAS* c.34G>A	*KRAS* c.34G>A, 1.04%; *CTNNB1* c.171G>A, 6.4%; *CTNNB1* c.206G>A, 6.8%; *FGFR3* c.746C>T 6.3%
200	450,450	442,746	99.0	81.2	96.9	90.6	*CTNNB1* c.98C>T, *EGFR* Ex19 15-bp del[c]	*CTNNB1* c.98C>T, 10.7%; *EGFR* Ex19 15-bp del, 16.9%[c]
200	478,867	468,913	99.0	81.2	98.1	91.3	*ERBB2* Ex20 3-bp ins	*ERBB2* Ex20 3-bp ins, 21.3%
200	553,994	541,422	98.8	80.8	98.3	94.2	*ERBB2* Ex20 12-bp ins[c]	*ERBB2* Ex20 12-bp ins, 95%[c]
200	513,991	501,669	99.0	87.0	99.4	86.6	*EGFR* amplification	See **Figure 3d**
AMP performance with low input amounts of a clinical FFPE sample (96-amplicon panel[e])								
200	589,234	579,032	98.4	85.4	100.0	95.5	Not applicable	Not applicable
100	549,797	536,978	98.4	86.7	100.0	95.9	Not applicable	Not applicable
50	986,319	963,424	98.6	86.0	100.0	97.3	Not applicable	Not applicable
10	217,989	213,371	99.1	87.0	91.0	75.2	Not applicable	Not applicable
5	142,132	138,422	99.2	86.9	82.0	60.8	Not applicable	Not applicable

BWA, Burrows-Wheeler Aligner; BLAT, BLAST-like alignment tool.
[a]DNA input corresponds to total DNA used across two assay tubes (for example, 200 ng indicates 100 ng double-stranded DNA input per tube). [b]Post-adapter-trimming reads shorter than 50 bases were discarded. [c]Unmapped reads from BWA were further mapped with BLAT. Large insertion (*ERBB2* 12 bp) and deletions (*EGFR* 15 and 18 bp) were detected with BLAT mapping. [d]Only oncogene mutations listed. [e]The assay employs a bi-template, two-tube (plus and minus strands, each in a separate tube) sequencing approach when targeting exons longer than 200 bp. Effectively, the 626-amplicon assay is 313-plex per tube, and the 96-amplicon assay is 48-plex per tube.

rate for all PCR amplifications. We experimented with tetramethyl ammonium chloride (TMAC) to improve amplification of AT-rich targets[17,18] and evaluated three DNA polymerases. Platinum Taq polymerase resulted in the highest rate of mapped (99.5% to human genome) and on-target (88%) reads, with 97% of targeted bases sequenced at more than 100-fold coverage and 94% at more than 500-fold coverage (**Fig. 3a**). The majority of targeted bases showed even coverage: 93.1% within fivefold above and below the average (25-fold range) and 83.9% within 3.2-fold above and below the average (tenfold range). OneTaq and Phusion HF yielded poorer performance, whereas addition of TMAC to Platinum Taq showed a minor improvement in uniformity (94.3% coverage within 25-fold and 86.5% within tenfold) (**Table 1** and **Fig. 3a**). These enrichment metrics were achieved with one attempt at primer design, synthesis and pooling without any optimization.

Testing of archived FFPE material of varying yields and quality is challenging for any PCR-based method. For the targeted RNA-seq assay, we have implemented quality metrics (Online Methods) that allow us to confidently determine cases that fail due to poor RNA quantity or quality while minimizing false negative results. With more assay experience in production, the sample failure rate in our most recent 3 months was 3% (9 of 313 cases) (**Supplementary Fig. 4**). For targeted DNA-seq, we analyzed the sequencing results of AMP libraries generated with input amounts of 200, 100, 50, 10 and 5 ng of an FFPE tumor DNA sample divided across a two-tube reaction. AMP optimally requires 25 ng minimum of DNA input per reaction (**Supplementary Fig. 5**).

Next, we evaluated the ability of AMP to detect known single nucleotide variants, indels and copy number changes previously genotyped with our clinical laboratory assays[14]. Using the 96-amplicon AMP hotspot mutation assay, all expected single nucleotide variants, insertions (3 and 12 bp in *ERBB2*, encoding v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2) and deletions (15 and 18 bp in *EGFR*) were identified (**Table 2** and **Fig. 3b**). Sequencing read coverage for *PTEN* was relatively even and complete despite the presence of the pseudogene (**Fig. 3c**). An *EGFR* gene amplification was identified by its excessive read coverage relative to both the intersample average *EGFR* read depth and the intrasample coverage for *BRAF*, which is also located on chromosome 7 (**Fig. 3d**). We further evaluated AMP with a more clinically relevant 168-target cancer panel and tested a wide range of samples ($n = 63$). The results showed excellent concordance for expected single nucleotide variants ($n = 24$), indels ($n = 18$), and copy number amplifications ($n = 11$, **Supplementary Table 3**).

DISCUSSION

Various target enrichment methods for next-generation sequencing have been described and compared, each associated with its own advantages and disadvantages[1]. Microdroplet-based PCR may achieve a high level of multiplexing[19] but requires special instrumentation and a large amount of input template (>1 μg DNA), which is often unavailable in clinical specimens. The molecular inversion probe approach employs an initial long hybridization step and suffers from poor evenness, with only 58% of targets within a tenfold abundance range[20]. Hybridization-capture–based target enrichment demonstrates high scalability from hundreds of genes[21] to the entire human exome[22]. Yet this method generally requires a long hybridization time, a relatively large amount of starting material and specialized bait design, synthesis and optimization. Methods such as AmpliSeq[23], TruSeq Amplicon[24],

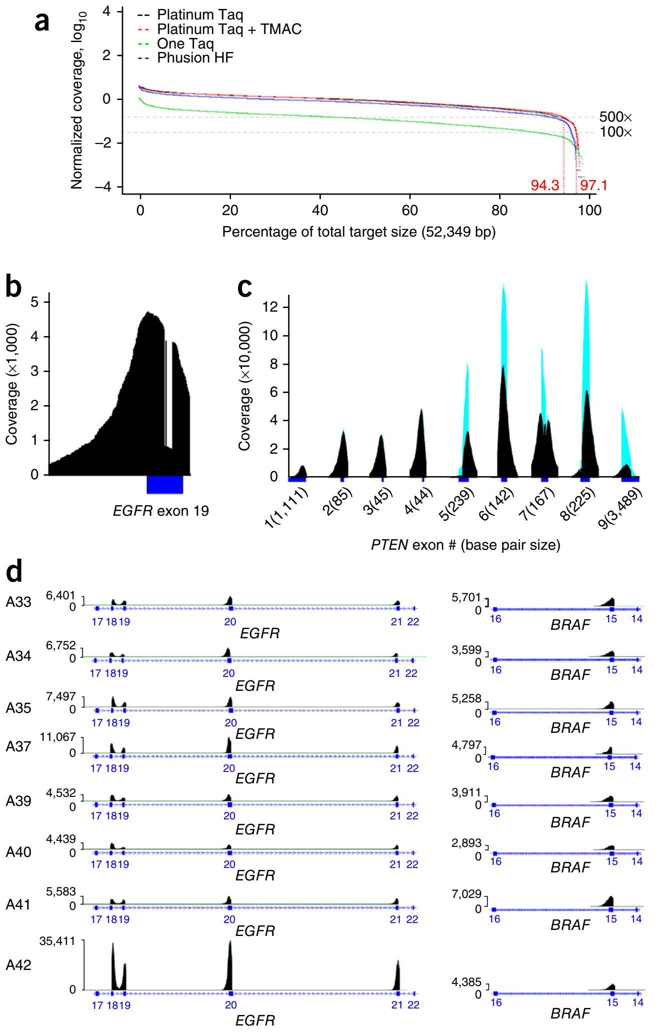

Figure 3 Targeted DNA sequencing using AMP for the detection of single nucleotide variants, insertions, deletions and copy number variants. (**a**) AMP optimization for genomic DNA sequencing. Four conditions were tested for AMP using a 626-amplicon assay targeting 18 tumor suppressor genes (370 exons total) in one experiment: Platinum Taq polymerase alone, Platinum Taq polymerase with TMAC, OneTaq polymerase and Phusion HF polymerase. Normalized coverage relative to the mean coverage of the Platinum Taq polymerase alone condition ($\log_{10}$ scale) is plotted against the percentage of covered total target. (**b**) AMP detection of a clinical deletion variant. An example read pileup from one experimental run showing an *EGFR* exon 19 18-bp deletion targeted in the 96-amplicon cancer panel. (**c**) Example illustration of the *PTEN* gene showing even coverage of on-target exons (black) and off-target pseudogene regions (cyan) across the entire coding sequence. (**d**) A clinical glioblastoma case (A42) showing *EGFR* amplification as represented by overabundant read coverage (*y* axis).

HaloPlex[25] and Nested Patch PCR[26] are all disadvantaged by strategies targeting the two ends flanking a region of interest, yielding sequencing read pileups that are blunted on both ends. The lack of unique sequencing start sites associated with these methods may introduce systematic errors, prohibit read deduplication and preclude confident variant calling based on random sequence sampling.

We have described AMP as a rapid enrichment method for targeted RNA and DNA next-generation sequencing. We demonstrate its robust utility for detection of gene fusions, point mutations, insertions, deletions and copy number changes from low amounts of clinical FFPE RNA and DNA samples. A unique advantage of AMP compared to other PCR methods is its ability to assess for unique reads. As a result, AMP may be used to assess sequence read complexity based on random start sites, in contrast to other PCR-based enrichment techniques described above. By targeting sequences with a one-sided nested primer approach, AMP offers the distinct ability to agnostically detect gene rearrangements by simply targeting one of the consistently involved fusion partners. Based on a core set of standard molecular biology reagents, AMP utilizes primers that may be quickly designed and synthesized as part of a facile, custom targeted sequencing solution wherein library construction can be completed in 1–2 days (**Supplementary Fig. 6**). Our method is economical (**Supplementary Table 4**) and practical for applications such as confirmation sequencing for larger-scale methods like whole-exome or genome sequencing. We believe that AMP is scalable for targeted applications in RNA-seq, genomic DNA sequencing and clinical genotyping.

METHODS

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

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

ACKNOWLEDGMENTS

This study was supported by a US National Institutes of Health grant (R21CA161590) to A.J.I. and a postdoctoral fellowship from the Swedish Research Council (350-2012-368) to Z.Z.

AUTHOR CONTRIBUTIONS

Z.Z., M.L., B.Z., Y.C., D.P., K.D.L., J.C., H.E.R., H.S.S., J.C., W.P. & J.A.E. conducted the experiments; Z.Z. and L.P.L. conducted the data analyses; Z.Z. and L.P.L. wrote the manuscript; A.J.I. supervised the project.

COMPETING FINANCIAL INTERESTS

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

1. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).
2. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
3. Kwak, E.L. *et al.* Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.* **363**, 1693–1703 (2010).
4. Bergethon, K. *et al.* *ROS1* rearrangements define a unique molecular class of lung cancers. *J. Clin. Oncol.* **30**, 863–870 (2012).
5. Drilon, A. *et al.* Response to Cabozantinib in patients with RET fusion-positive lung adenocarcinomas. *Cancer Discov.* **3**, 630–635 (2013).
6. Heuckmann, J.M. *et al.* Differential protein stability and ALK inhibitor sensitivity of EML4-ALK fusion variants. *Clin. Cancer Res.* **18**, 4682–4690 (2012).
7. Crystal, A.S. & Shaw, A.T. Variants on a theme: a biomarker of crizotinib response in ALK-positive non-small cell lung cancer? *Clin. Cancer Res.* **18**, 4479–4481 (2012).
8. Takeuchi, K. *et al.* RET, ROS1 and ALK fusions in lung cancer. *Nat. Med.* **18**, 378–381 (2012).
9. Frohman, M.A., Dush, M.K. & Martin, G.R. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA* **85**, 8998–9002 (1988).
10. Zheng, Z. *et al.* Titration-free 454 sequencing using Y adapters. *Nat. Protoc.* **6**, 1367–1376 (2011).
11. Zheng, Z. *et al.* Titration-free massively parallel pyrosequencing using trace amounts of starting material. *Nucleic Acids Res.* **38**, e137 (2010).
12. Neiman, M. *et al.* Library preparation and multiplex capture for massive parallel sequencing applications made efficient and easy. *PLoS ONE* **7**, e48616 (2012).

13. Wang, T., Yu, D. & Lamb, M.L. Trk kinase inhibitors as new treatments for cancer and pain. *Expert Opin. Ther. Pat.* **19**, 305–319 (2009).
14. 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).
15. Chmielecki, J. *et al.* Targeted next-generation sequencing of DNA regions proximal to a conserved GXGXXG signaling motif enables systematic discovery of tyrosine kinase fusions in cancer. *Nucleic Acids Res.* **38**, 6985–6996 (2010).
16. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
17. Oyola, S.O. *et al.* Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics* **13**, 1 (2012).
18. Chevet, E., Lemaitre, G. & Katinka, M.D. Low concentrations of tetramethylammonium chloride increase yield and specificity of PCR. *Nucleic Acids Res.* **23**, 3343–3344 (1995).
19. Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat. Biotechnol.* **27**, 1025–1031 (2009).
20. Turner, E.H. *et al.* Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* **6**, 315–316 (2009).
21. 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).
22. Clark, M.J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29**, 908–914 (2011).
23. Yousem, S.A. *et al.* Pulmonary Langerhans cell histiocytosis: profiling of multifocal tumors using next-generation sequencing identifies concordant occurrence of *BRAF* V600E mutations. *Chest* **143**, 1679–1684 (2013).
24. Do, H. *et al.* Reducing sequence artifacts in amplicon-based massively parallel sequencing of formalin-fixed paraffin-embedded DNA by enzymatic depletion of uracil-containing templates. *Clin. Chem.* **59**, 1376–1383 (2013).
25. Johansson, H. *et al.* Targeted resequencing of candidate genes using selector probes. *Nucleic Acids Res.* **39**, e8 (2011).
26. Varley, K.E. & Mitra, R.D. Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Res.* **18**, 1844–1850 (2008).

ONLINE METHODS

Study samples and nucleic acids extraction. Discarded samples from the Diagnostic Molecular Pathology Laboratory, Massachusetts General Hospital (MGH), were used in this study. For initial assay development and validation, biologically relevant samples were used for this study, including lung adenocarcinoma specimens for *ALK*, *RET* and *ROS1* rearrangements, papillary thyroid carcinoma for *RET* and *NTRK1* rearrangements, and glioblastoma for *ROS1* and *NTRK1* translocations. No other selection criteria were applied. No sample restrictions were followed for discovery runs. Total nucleic acids containing total RNA and genomic DNA were extracted from formalin-fixed paraffin-embedded biopsies, using the Agencourt FormaPure Kit for FFPE Tissue (Beckman Coulter, Indianapolis, IN). *ALK*, *RET*, *ROS1* and *PPARG* FISH results were available for comparison based on assays performed as previously described[3,4]. In addition, SNaPshot[14] point mutation results were available for comparison. The study was approved by the Massachusetts General Hospital Institutional Review Board and Partners Healthcare Human Research Committee. A waiver of consent for discarded clinical material was obtained. To compare a new diagnostic assay with reference (FISH) assays, with a positive sample proportion of 0.16 (1 positive and 5 negative) and type I error of 0.05, the sample size needed for an excellent agreement (Kappa 0.9, range 0.8 – 0.99) is 292 (49 positive and 243 negative cases), calculated using the kappaSize R package[27].

Anchored multiplex PCR. Library construction for AMP (**Fig. 1**) starts with RNA or total nucleic acid (DNA and RNA mix) as input, without the need for ribosomal RNA or genomic DNA depletion. First- and second-strand complementary DNA (cDNA) synthesis was performed using a combination of SuperScript III (Life Technologies, Carlsbad, CA), DNA Polymerase I (Enzymatics, Beverly, MA) and RNAse H (Enzymatics). Double-stranded cDNA was cleaned with Ampure XP SPRI beads (Beckman Coulter). Either double-stranded cDNA or, alternatively, genomic DNA underwent end repair (End-Repair Mix, Enzymatics), adenylation (Klenow Exo-, Enzymatics; Taq Polymerase, Life Technologies), and ligation (T4 DNA Ligase, Enzymatics) with a universal half-functional adapter. SPRI-cleaned ligated libraries were subjected to two rounds of nested PCR at 10 to 14 cycles each for target enrichment (Platinum Taq Polymerase, Life Technologies). The first round of PCR was performed using a primer complementary to the universal adapter and a first pool of up to hundreds of target specific primers (Operon, Huntsville, AL). After SPRI cleanup, a second round of PCR is executed using a 3′ nested universal adapter primer downstream of the first adapter primer and a second pool of 3′ nested target specific primers downstream of the respective initial first-pool target primers. These nested primers are each 5′ tagged with a common sequencing adapter which, in combination with the first half-functional universal adapter, creates target amplicons ready for clonal amplification (for example, emulsion PCR or bridge PCR) and sequencing. Libraries are quantitated using quantitative PCR (Kapa Biosystems, Woburn, MA), normalized, and processed respectively for sequencing on the MiSeq (Illumina, San Diego, CA) or Ion Torrent Personal Genome Machine (PGM) (Life Technologies) according to the manufacturers' standard protocol. Upfront library construction before quantitation by qPCR can be typically accomplished in 6–8 h.

Assay quality controls. Testing of archived FFPE material of varying yields and integrity requires upfront quality measures to help interpret the results. We find that standard spectrophotometry and fluorescent-based RNA quantitation methods are not predictive of AMP success or failure. For the targeted RNA-seq AMP assay, we have instead implemented quality control by using a commercial qPCR-based kit (TaqMan GAPDH Control Reagents, Human, Life Technologies) and also by targeting internal housekeeping genes (*GAPDH*, *B2M*, and *CTBP1*) in our assay. From our 6-month experience of clinical testing, all samples that passed the QC kit with a Ct value less than 35 showed >100 cDNA reads for at least one of the housekeeping genes, allowing us to confidently determine cases that fail due to poor RNA quality while minimizing false negative results. Our 6-month clinical (262 samples) and research (405 samples) testing experience showed an approximate overall 6% failure rate and detection of 29 cases positive for fusion transcripts. Ruling out early failures due to testing older, archived specimens with low nucleic acid concentration and also due to technical errors, our more recent 3 month failure rate is approximately 3% (**Supplementary Fig. 4**).

Amount of input DNA. To access the amount of starting material required for the AMP-targeted DNA-seq, we analyzed the sequencing results of AMP libraries generated with total input amounts of 200, 100, 50, 10 and 5 ng of an FFPE tumor DNA sample divided across a two-tube assay. We applied the 96-amplicon hotspot plus tumor suppressor assay for testing. Alignment percentages and on-target reads were similar for the five libraries. While minimum target coverage for the first three libraries were 100% at >100× and ~96% at >500×, it decreased for the 10-ng library (91% and 75%, respectively) and even further for the 5-ng library (82% and 61%, respectively) (**Table 2**). The raw read pileup for *EGFR* exon 20 showed adequate library complexity with few duplicates for samples with 50-ng or higher input. In contrast, lower library complexity and higher duplication rates were observed for the 10-ng and 5-ng input samples (**Supplementary Fig. 5**). We conclude that using an adapter ligation approach for AMP optimally requires 25 ng of minimal DNA input per reaction.

Data analysis. Paired-end sequence data were processed initially by adapter trimming with a custom shell script. Sequences at the 3′ end that potentially represent whole or partial adapter sequences (minimally one single base) were trimmed. Reads with whole adapter sequence in the middle were trimmed from the adapter to the 3′ end. Processed reads with lengths shorter than 50 bp after adapter trimming were discarded. A combination of BWA (v0.6.2)[28] and BLAT (v35)[29] was used in a hybrid manner to optimize fusion gene alignment and detection. BWA was used first to align nearly perfect reads to the hg19 reference genome. The unmapped reads from BWA, potentially containing chimeric fusion reads, were then aligned using the BLAT algorithm in a 2-step optimized multithreaded process. The first BLAT stage (tileSize 11 and stepSize 16) was used to loosely align the BWA unmapped reads to the hg19 reference genome. The resulting unmapped reads and the mapped reads with greater than 15 nt unmapped overhangs were then realigned more stringently with a second BLAT stage (tileSize 11 and stepSize 9). Gene fusion variant calling was executed with the following three criteria for the BWA-BLAT hybrid mapped reads: (i) fusion partners must show at least 25 non-overlapping mapped bases on their respective ends of a chimeric read; (ii) the two fusion partners must map to different genes; and (iii) the unknown partner upstream or downstream of the targeted gene on the anchored end must be minimally represented with 15 uniquely starting reads (reflecting random ligation of the universal adapter).

To detect point mutations and indels from genomic DNA sequencing, reads were mapped to the human hg19 reference genome using the BWA short read aligner with default parameters. BWA unmapped reads were submitted for BLAT mapping using a minimum score of 50 and a minimum identity of 95% to retain the highest quality mapped reads. The resulting PSL mapping files were converted to SAM and then BAM using SAMtools[28]. The BAM files derived from BWA and BLAT mappings were merged into one single BAM file per sample, processed by SAMtools mpileup, and variant called using VarScan (v2.3.3)[30]. Minimal thresholds of 100× coverage and 5% allelic fraction were applied. The resulting variants were filtered against dbSNP (1000 Genomes Project 20120626 Release) and annotated using the Bioconductor Variant Annotation package. Overall coverage was calculated using a 21-bp window for hotspot point mutation targets (±10 bp) and 5-bp intronic flanks for whole exon targets.

Primer design. A custom primer design engine was developed specifically for efficient AMP primer design based on Primer3[31]. To maintain applicability for fragmented nucleic acids from such samples as FFPE tissue, primers were designed to yield short amplicons of approximately 90 bp. For targeted RNA-seq to detect gene fusions, gene-specific primers were designed in a tiled fashion against the tyrosine kinase domains and near the exon boundary putatively involved in the rearrangement (**Supplementary Fig. 1**). A set of three primers for housekeeping genes (*GAPDH*, *B2M*, *CTBP1*) were included as an internal control for RNA quality check (**Supplementary Table 1**). An initial assay targeting *ALK*, *RET*, and *ROS1* was designed and implemented for gene fusion detection in lung cancer. Subsequently, a panel of 14 receptor tyrosine kinase genes (*ALK*, *ROS1*, *RET*, *MUSK*, *EGFR*, *FGFR1*, *FGFR3*, *INSR*, *INSRR*, *MET*, *NTRK1*, *NTRK2*, *NTRK3* and *PDGFRA*) was also designed for both 3′ and 5′ fusion partner detection (**Supplementary Table 5**). For targeted gDNA sequencing, we initially designed a 96-amplicon panel to cover 40 hotspot cancer mutations (shared in common with our clinical SNaPshot assay, **Supplementary Table 6**) and the entire coding region of three important tumor suppressor genes (*PTEN*, *TP53* and *CDKN2A*). Additionally, a gDNA 626-amplicon panel for 18 tumor suppressor genes (**Supplementary Table 7**) was constructed to demonstrate the scalability of the assay. The gDNA sequencing primers were designed to avoid common single nucleotide variants found in dbSNP and clinically relevant SNPs from the 1000 Genomes Project (20120626 Release). Candidate primers were prioritized to avoid potential homodimerization, heterodimerization, and mispriming with the library construction sequencing adapters and barcodes (IonTorrent 96 barcodes, and Illumina MiSeq 96 forward and 12 reversed indexes). The source code was deposited at bitbucket.org and is available upon request.

27. Donner, A. & Rotondi, M.A. Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. *Int. J. Biostat.* **6**, 31 (2010).
28. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
29. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
30. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
31. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).

doi:10.1038/nm.3729