Gu et al. Genome Biology (2016) 17:41
DOI 10.1186/s13059-016-0904-5

Genome Biology

METHOD Open Access

Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications

W. Gu[1†], E. D. Crawford[2,3†], B. D. O'Donovan[4], M. R. Wilson[5], E. D. Chow[6], H. Retallack[7] and J. L. DeRisi[2,3*]

Abstract

Next-generation sequencing has generated a need for a broadly applicable method to remove unwanted high-abundance species prior to sequencing. We introduce DASH (Depletion of Abundant Sequences by Hybridization). Sequencing libraries are 'DASHed' with recombinant Cas9 protein complexed with a library of guide RNAs targeting unwanted species for cleavage, thus preventing them from consuming sequencing space. We demonstrate a more than 99 % reduction of mitochondrial rRNA in HeLa cells, and enrichment of pathogen sequences in patient samples. We also demonstrate an application of DASH in cancer. This simple method can be adapted for any sample type and increases sequencing yield without additional cost.

Keywords: Cas9, CRISPR, Depletion, Sequencing, RNA-Seq, Infectious Disease, Cancer, Diagnostics

Background

The challenge of extracting faint signals from abundant noise in molecular diagnostics is a recurring theme across a broad range of applications. In the case of RNA sequencing (RNA-Seq) experiments specifically, there may be several orders of magnitude difference between the most abundant species and the least. This is especially true for metagenomic analyses of clinical samples like cerebrospinal fluid (CSF), whose source material is inherently limited [1], making enrichment or depletion strategies impractical or impossible to employ prior to library construction. The presence of unwanted high-abundance species, such as transcripts for the 12S and 16S mitochondrial ribosomal RNAs (rRNAs), effectively increases the cost and decreases the sensitivity of counting-based methodologies.

The same issue affects other molecular clinical diagnostics. In cancer profiling, the fraction of the mutant tumor-derived species may be vastly outnumbered by wild-type species due to the abundance of immune cells or the interspersed nature of some tumors throughout normal tissue. This problem is profoundly exaggerated in the case of cell-free DNA/RNA diagnostics, whether from malignant [2, 3], transplant [4], or fetal sources [5, 6], and relies on brute force counting by either sequencing or digital PCR (dPCR) [7] to yield a detectable signal. For these applications, a technique to deplete specific unwanted sequences that is independent of sample preparation protocols and agnostic to measurement technology is highly desired.

CRISPR (clustered regularly interspaced short palindromic repeats) and Cas (CRISPR associated) nucleases, such as Cas9, function in bacterial adaptive immune systems to remove incoming phage DNA from the host without harm to the bacteria's own genome. The CRISPR-Cas9 system has attained widespread adoption as a genome editing technique [8–11]. When coupled with single guide RNAs (sgRNAs) designed against

* Correspondence: Joe@derisilab.ucsf.edu
W. Gu and E. D. Crawford are co-first authors.
†Equal contributors
[2]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA
[3]Howard Hughes Medical Institute, Chevy Chase, MD, USA
Full list of author information is available at the end of the article

targets of interest, *Streptococcus pyogenes* Cas9 binds to 3' NGG protospacer adjacent motif (PAM) sites and produces double-stranded breaks if the sgRNA successfully hybridizes with the adjacent target sequence (Fig. 1a). In vitro, Cas9 may be used to cut DNA directly, in a manner analogous to a conventional restriction enzyme, except that the target sequence (outside of the PAM site) may be programmed at will and massively multiplexed without significant off-target effects. This affords the unique opportunity to target and prevent amplification of undesired sequences, such as those that are generated during next-generation sequencing (NGS) protocols.

In this paper, we have exploited the unique properties of Cas9 to selectively deplete unwanted high-abundance sequences from existing RNA-Seq libraries. We refer to this approach as Depletion of Abundant Sequences by Hybridization (DASH). Employing DASH after transposon-mediated fragmentation but prior to the following amplification step (which relies on the presence of adaptor sequences on both ends of the fragment) prevents amplification of the targeted sequences, thus ensuring they are not represented in the final sequencing library (Fig. 1b). We show that this technique preserves the representational integrity of the non-targeted sequences while increasing overall sensitivity in cell line samples and human metagenomic patient samples. Further, we demonstrate the utility of this system in the context of cancer detection, in which depletion of wild-type sequences increases the detection limit for oncogenic mutant sequences. The DASH technique may be used to deplete specific unwanted sequences from existing Illumina sequencing libraries, PCR amplicon libraries, plasmid collections, phage libraries, and virtually any other existing collection of DNA species.

Existing specific sequence enrichment techniques — such as pull-down methods [5, 12–14], amplicon-based methods [3, 15], molecular inversion methods [16–18], COLD-PCR [19], competitive allele-specific TaqMan PCR (castPCR) [20], and the classic method of using restriction enzyme digestion on mutant sites [21] — can effectively enrich for targets in sequencing libraries, but these are not useful for discovery of unknown or unpredicted sequences. Brute force counting methods also exist, such as dPCR [3, 7], but they are not easy to multiplex across a large panel. While high-throughput sequencing of select regions can be highly multiplexed to detect rare and novel mutations, and barcoded unique identifiers can overcome sequencing error noise [22], it is costly since the vast majority of the sequencing reads map to non-informative wild-type sequences. A number of sequence-specific RNA depletion methods also currently exist. Illumina's Ribo-Zero rRNA Removal Kit and Ambion's GLOBINclear Kit pull rRNAs and globin mRNAs, respectively, out of total RNA samples using sequence-specific oligos conjugated to magnetic beads. RNAse H-based methods, such as New England BioLab's NEBNext rRNA Depletion Kit similarly mark abundant RNA species with sequence-specific DNA oligos, and then subject them to degradation by RNAse H, which digests RNA/DNA hybrid molecules [23]. These methods are all employed prior to the start of library prep, and are limited to samples containing at least 10 ng to 1 μg of RNA. DASH, in contrast, depletes abundant species after complementary DNA (cDNA) amplification, and thus can be utilized for essentially any amount of input sample.

Results

We demonstrate deletion of unwanted mitochondrial rRNA using DASH first on HeLa cell line RNA (Fig. 2) and then on CSF RNA from patients with pathogens in their CSF (Fig. 3), in order to increase sequencing bandwidth of useful data. Selection of rRNA sgRNA targets was based on examining coverage plots for standard RNA-Seq experiments on HeLa cells as well as on several patient CSF samples. Coverage of the 12S and 16S mitochondrial rRNA genes was consistently several orders of magnitude higher than the rest of the mitochondrial and non-mitochondrial genes (Figs. 2c and 3). We chose 54 sgRNA target sites within this high-coverage region of the mitochondrial chromosome, situated approximately every 50 bp over a 2.5 kb region (sequences listed in Additional file 1). sgRNA sites are indicated by red arrowheads in Fig. 2b. sgRNAs for these sites were generated as described in the "Materials and methods" section.

To calculate the input ratio of Cas9 and sgRNA to sample nucleic acid, we estimated that 90 % of each sample was comprised of the rRNA regions that we targeted; thus, our potential substrate makes up 4.5 ng of a 5 ng sample. This corresponds to a target site concentration of 13.8 nM in the 10 μL reaction volume. To assure the most thorough Cas9 activity possible, and given that Cas9 is a single-turnover enzyme in vitro [24], we used a 100-fold excess of Cas9 protein and a 1000-fold excess of sgRNA relative to the target. Thus, each 10 μL sample of cDNA generated from a CSF sample contained a final concentration of 1.38 μM Cas9 protein and 13.8 μM sgRNA. In the case of HeLa cDNA, we used only 1 ng per sample, and therefore decreased the Cas9 and sgRNA concentrations by fivefold. However, since mitochondrial rRNA sequences represented only approximately 60 % of the HeLa samples (compared with approximately 90 % for CSF), the HeLa samples contained 150-fold Cas9 and 1500-fold sgRNA. To examine dose response, we processed additional 1 ng HeLa samples treated with 15-fold Cas9 and 150-fold sgRNA. Both concentrations were done in triplicate (Additional file 2).

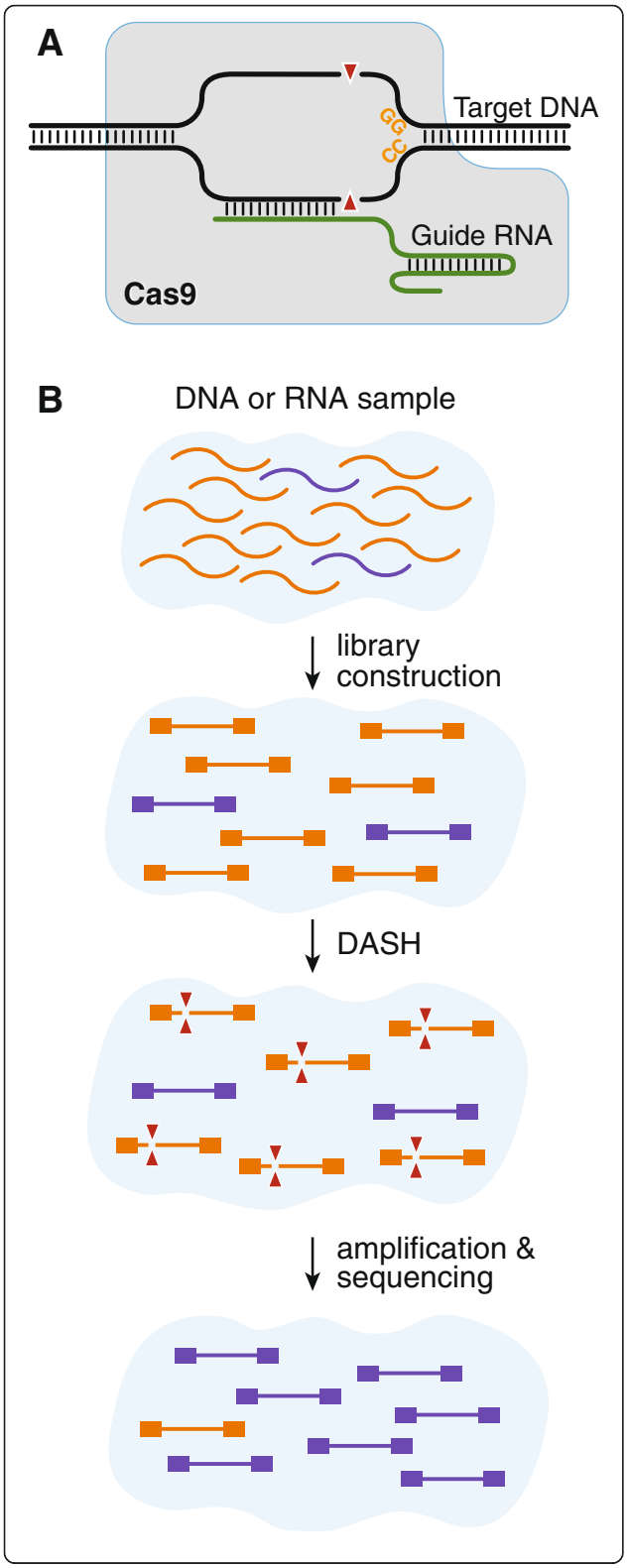

Fig. 1 **a** *S. pyogenes* Cas9 protein binds specifically to DNA targets that match the 'NGG' protospacer adjacent motif (PAM) site. Additional sequence specificity is conferred by a single guide RNA (sgRNA) with a 20 nucleotide hybridization domain. DNA double strand cleavage occurs three nucleotides upstream of the PAM site. **b** Depletion of Abundant Sequences by Hybridization (DASH) is used to target regions that are present at a disproportionately high copy number in a given next-generation sequencing library following tagmentation or flanking sequencing adaptor placement. Only non-targeted regions that have intact adaptors on both ends of the same molecule are subsequently amplified and represented in the final sequencing library

Reduction of unwanted abundant sequences in HeLa samples

We first demonstrate the utility and efficacy of our approach using sequencing libraries prepared from total RNA extracted from HeLa cells. In the untreated samples, reads mapping to 12S and 16S mitochondrial rRNA genes represent 61 % of all uniquely mapped human reads. After DASH treatment, these sequences are reduced to only 0.055 % of those reads (Fig. 2a, b). Comparison of gene-specific fragments per kilobase of transcript per million mapped reads (fpkm) values between treated and untreated samples reveals mean 82-fold and 105-fold decreases in fpkm values for 12S and 16S rRNA, respectively, in the samples treated with 150-fold Cas9 and 1500-fold sgRNA (Fig. 2c). Similarly, the samples treated with 15-fold Cas9 and 150-fold sgRNA show 30- and 45-fold reductions in 12S and 16S fpkm values, respectively, indicating a dose-dependent response to DASH treatment (Additional file 2).

Enrichment of non-targeted sequences and analysis of off-target effects in HeLa samples

This profound depletion of abundant 12S and 16S transcripts increases the available sequencing capacity for the remaining, untargeted transcripts. We quantify this increase by the slope of the regression line fit to the remaining genes, showing a 2.38-fold enrichment in fpkm values for all untreated transcripts. An R^2 coefficient of 0.979 for this regression line indicates strong consistency between replicates with minimal off-target effects (Fig. 2c).

To confirm that our depletion was specific to only the targeted mitochondrial sequences, we calculated the changes in fpkm values across all genes in the treated and untreated samples and identified those genes that were significantly diminished (>2 standard deviations) relative to their control values. To overcome issues with stochastic variation at low gene counts/fpkm, we eliminated those genes that, between the three technical replicates at each Cas9 concentration, showed standard deviations in fpkm values greater than 50 % of the mean. All of the genes meeting this criterion were present at

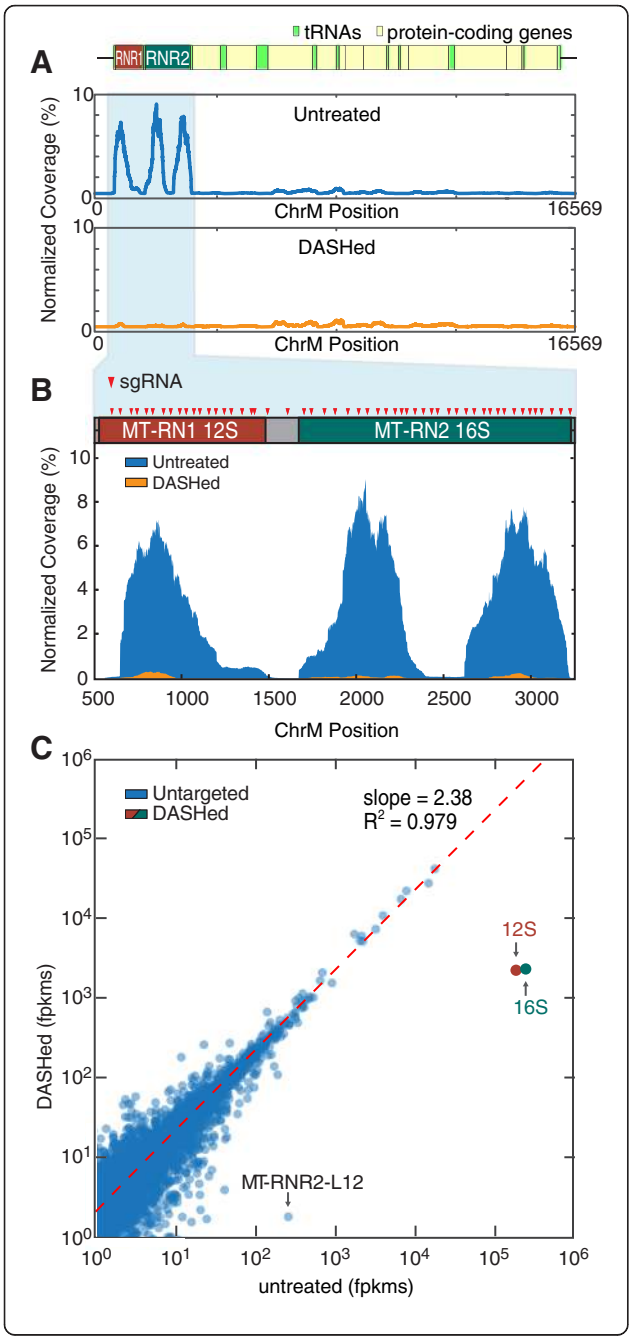

Fig. 2 Depletion of Abundant Sequences by Hybridization (DASH) targeting abundant mitochondrial ribosomal RNA in HeLa RNA extractions. **a** Normalized coverage plots showing alignment to the full-length human mitochondrial chromosome. Before treatment, three distinct peaks representing the 12S and 16S ribosomal subunits characteristically account for a large majority of the coverage (>60 % of total mapped reads). After treatment, the peaks are virtually eliminated — with 12S and 16S signatures reduced 1000-fold to 0.055 % of mapped reads. **b** Coverage plot of Cas9-targeted region with 12S and 16S gene boundaries across the top. Each *red arrowhead* represents one sgRNA target site. We chose 54 target sites, spaced approximately 50 bp apart. **c** Scatterplot of the log of fragments per kilobase of transcript per million mapped reads (log-fpkm) values per human gene in the control versus treated samples illustrate the significant reduction in reads mapping to the targeted 12S and 16S genes. DASH treatment results in 82- and 105-fold reductions in coverage for the 12S and 16S subunits, respectively. The slope of the regression line (*red*) fit to the untargeted genes indicates a 2.38-fold enrichment in reads mapped to untargeted transcripts. R-squared (R^2) value of the regression line (0.979) indicates minimal off-target depletion. Between replicates, the R^2 coefficient between fpkm values across all genes is 0.994, indicating high reproducibility (three replicates). Notably, one gene, MT-RNR2-L12 (MT-RNR2-like pseudogene), shows significant depletion in the DASHed samples compared with the control

less than 15 fpkm. Of the remaining genes, only one non-targeted human gene, MT-RNR2-L12, showed significant depletion when compared with the un-treated samples (Fig. 2c). MT-RNR2-L12 is a pseudogene and shares over 90 % sequence identity with a portion of the 16S mitochondrial rRNA gene. Out of the 24 sgRNA sites within the homologous region, 16 of them retain intact PAM sites in MT-RNR2-L12. Of these, seven have perfectly matching 20mer sgRNA target sites, and the remaining nine each have between one and four mutations (Additional file 3). Depletion of this gene is, therefore, an expected consequence of our sgRNA choices.

Reduction of unwanted abundant sequences in CSF samples

We next tested the utility of our method when applied to clinically relevant samples. In the case of pathogen detection in patient samples, the microbial transcripts are typically low in number and become greatly outnumbered by human host sequences. As a result, sequencing depth must be drastically increased to confidently detect such small minority sequence populations. We reasoned that depletion of unwanted high-abundance sequences from patient libraries could result in increased representation of pathogen-specific sequence reads. We thus integrated the DASH method with our in-house metagenomic deep sequencing diagnostic pipeline for patients with meningeal inflammation (i.e., meningitis) or brain inflammation (i.e., encephalitis) likely due to an infectious agent or pathogen. Figure 3 and Table 1 summarize the results of this analysis. In all three cases, the DASHed and untreated samples have a similar number of reads (1.8–3.4

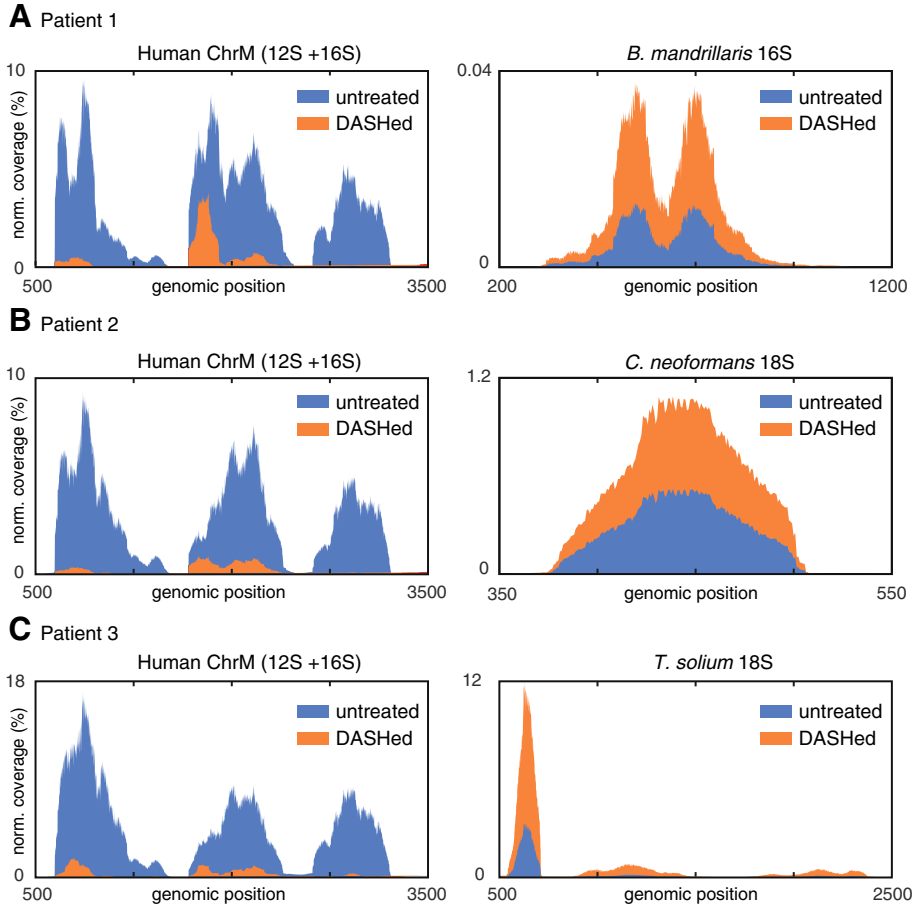

Fig. 3 Normalized coverage plots of DASH-treated (*orange*) and untreated (*blue*) libraries generated from patient cerebrospinal fluid (CSF) samples with confirmed infections. Targeted mitochondrial rRNA genes (*left*) and representative genes for pathogen diagnosis (*right*) are depicted for the following: patient 1, *Balamuthia mandrillaris* (**a**), patient 2, *Cryptococcus neoformans* (**b**), patient 3, *Taenia solium* (**c**). Across all cases, the DASH technique significantly reduced the coverage of human 12S and 16S genes by an average of 7.5-fold while increasing the coverage depth for pathogenic sequences by an average 5.9-fold. See Table 1 for relevant data

million), but DASHing reduces the number of duplicate reads, indicating an increase in library complexity.

In the case of a patient with meningoencephalitis whose CSF was previously shown to be infected with the amoeba *Balamuthia mandrillaris* [25] (patient 1), diagnosis was originally made by identification of a small fraction (<0.1 %) of reads aligning to specific regions of the *B. mandrillaris* 16S mitochondrial gene. After DASH treatment, human mitochondrial 12S and 16S genes were reduced by more than an order of magnitude, and

Table 1 Summary of depletion/enrichment results in DASH-treated clinical CSF samples

Pathogen	Read count (percentage duplicates)		Targeted genes (fpkm)				Representative pathogenic gene[a] (fold change)		R^2 non-targeted genes, untreated versus DASHed
			12S		16S				
	Un-treated	DASHed	Un-treated	DASHed	Un-treated	DASHed	Un-treated	DASHed	
B. mandrillaris	1.81 M (26 %)	2.54 M (15 %)	298,922	28,005	380,073	93,164	0.028 %	0.102 % (3.6×)	0.992
C. neoformans	2.95 M (27 %)	3.43 M (11 %)	361,501	37,168	342,857	93,703	1.5 %	15.4 % (10.3×)	0.986
T. solium	2.38 M (33 %)	1.89 M (30 %)	451,044	46,993	317,640	43,257	12.0 %	44.3 % (3.7×)	0.994

[a] Representative genes are 16S for *Balamuthia mandrillaris* and 18S for *Cryptococcus neoformans* and *Taenia solium*

sequencing coverage of the *B. mandrillaris* 16S fragment increased 3.6-fold. Notably, *B. mandrillaris* is a eukaryotic organism, yet depletion of the human 16S gene by DASH did not have off-target effects on the 16S *B. mandrillaris* mitochondrial gene. Similarly, patient CSF samples with confirmed *Cryptococcus neoformans* (fungus; patient 2) and *Taenia solium* (pork tapeworm; patient 3) infections showed 2- and 3.9-fold increases in coverage of the 18S genes of *C. neoformans* and *T. solium*, respectively, the detection of which was crucial in the initial diagnoses. The observed increases in relative signal can be translated into either a sequencing cost savings or a higher sensitivity that may be useful clinically for earlier detection of infections.

Reduction of wild-type background for detection of the *KRAS* G12D (c.35G>A) mutation in human cancer samples

Specific driver mutations known to promote cancer evolution and at times to make up the genetic definition of malignant subtypes are important for diagnosis and targeted therapeutics (i.e., precision medicine). In complex samples isolated from biopsies or cell-free body fluids such as plasma, wild-type DNA sequences often overwhelm the signal from mutant DNA, making the application of traditional Sanger sequencing challenging [2, 3, 26]. For NGS, detection of minority alleles requires additional sequencing depth and therefore increases cost. We reasoned that the DASH technique could be applied to increase mutation detection from a PCR amplicon derived from a patient sample. We chose to focus on depletion of the wild-type allele of *KRAS* at the glycine 12 position, a hotspot of frequent driver mutations across a variety of malignancies [27–29]. This is an ideal site for DASH because all codons encoding the wild-type glycine residue contain a PAM site (NGG), while any mutation that alters that residue (e.g., c.35G>A, p.G12D) ablates the PAM site and is thus uncleavable by Cas9 (Fig. 4a). This will be true of any mutation that changes a glycine (codons GGA, GGC, GGG, and GGT) or a proline (codons CCA, CCC, CCG, and CCT) to any other amino acid. Furthermore, it is relevant to the ubiquitous C>T nucleotide change found in germline mutations as well as somatic cancer mutations [30]. Targeting of other mutations will likely be possible in the near future with reengineered CRISPR nucleases or those that come from alternative species and have different PAM site specificities [31, 32].

The sequence of the sgRNA designed to target the *KRAS* G12D PAM site is listed in Additional file 1, as is the non-human sequence used for the negative control sgRNA. Both were transcribed from a DNA template by T7 RNA polymerase, purified, and complexed with Cas9 as described in the "Materials and methods" section. Samples were prepared by mixing sheared genomic DNA from a healthy individual (with wild-type *KRAS* genotype confirmed with dPCR) and *KRAS* G12D genomic DNA to achieve mutant to wild-type allelic ratios of 1:10, 1:100, and 1:1000, and 0:1. For each mixture, 25 ng of a DNA was incubated with 25 nM Cas9 pre-complexed with 25 nM of sgRNA targeting *KRAS* G12D. This concentration is high relative to the concentration of target molecules, but empirically we found it to be the most efficient ratio. We hypothesize that this may be due to non-cleaving Cas9 interactions with the rest of the human genome [24], which effectively reduce the Cas9 concentration at the cleavage site.

Samples were subsequently heated to 95 °C for 15 min in a thermocycler to deactivate Cas9 ("Materials and methods"). Droplet digital PCR (ddPCR) was used to count wild-type and mutant alleles using the primers and TaqMan probes depicted in Fig. 4a and described in the "Materials and methods" section. All samples were processed in triplicate. Samples incubated with or without Cas9 complexed to a non-human sgRNA target show the expected percentages of mutant allele: approximately 10 %, 1 %, and 0.1 % for the 1:10, 1:100, and 1:1000 initial mixtures respectively (Fig. 4b). With addition of Cas9 targeted to *KRAS*, the wild-type allele count drops nearly two orders of magnitude (purple bars in Fig. 4b), while virtually no change is observed in number of mutant alleles (blue bars). This confirms the high specificity of Cas9 for the NGG of the PAM site.

With the addition of DASH targeted to *KRAS* G12, the percentage of mutant allele jumps from 10 % to 81 %, from 1 % to 30 %, and from 0.1 % to 6 % (Fig. 4c). This corresponds to 8.1-fold, 30-fold and 60-fold representational increases for the mutant allele, respectively. As expected, there was virtually no detection of mutant alleles in the wild-type-only samples both with and without DASH treatment (one droplet in one of three no DASH wild-type-only samples).

Discussion

In this paper we have introduced DASH, a technique that leverages in vitro Cas9 ribonucleoprotein (RNP) activity to deplete specific unwanted high-abundance sequences, which results in the enrichment of rare and less abundant sequences in NGS libraries or amplicon pools.

While the procedure may be easily generalized, we developed DASH to address current limitations in metagenomic pathogen detection and discovery, where the sequence abundance of an etiologic agent may be present as a minuscule fraction of the total. For example, infectious encephalitis is a syndrome caused by well over 100 pathogens ranging from viruses, fungi, bacteria and parasites. Because of the sheer number of diagnostic possibilities and the typically low pathogen load present in CSF, more than half of encephalitis patients never have an etiologic agent identified [33]. We have demonstrated that NGS is a powerful tool for identifying infections, but as

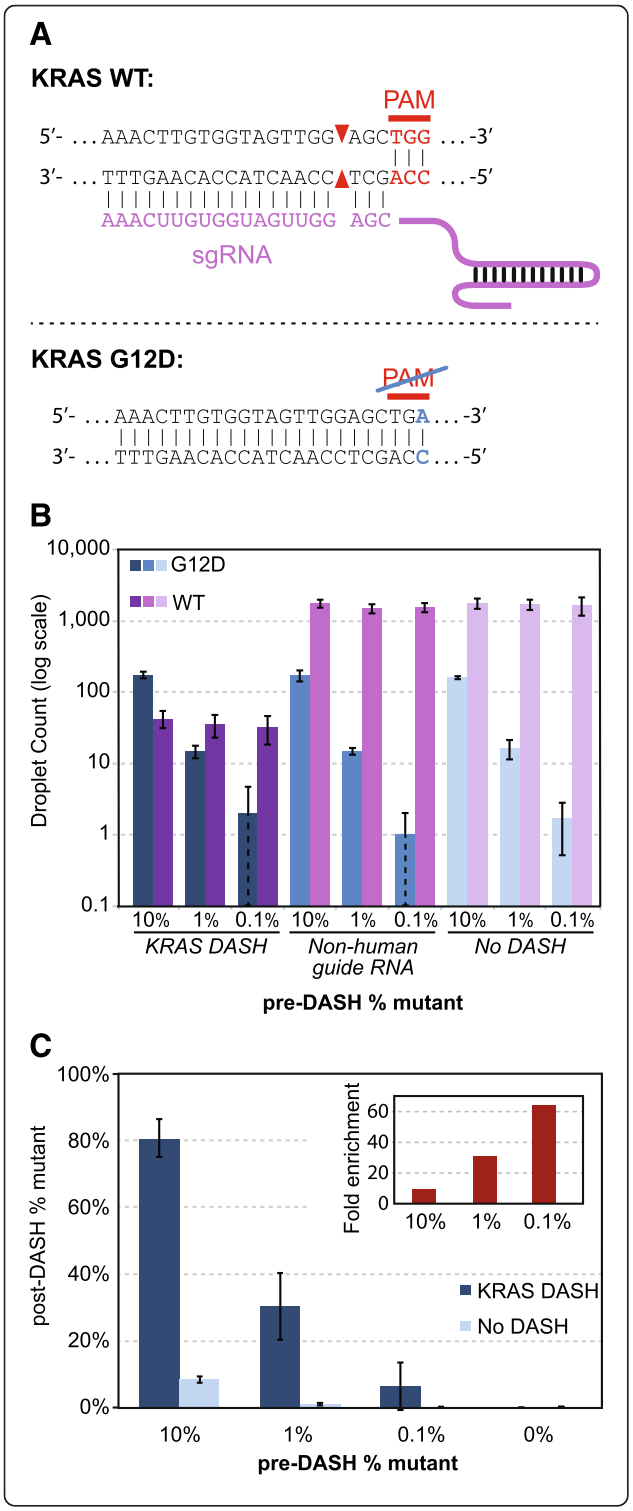

Fig. 4 **a** DASH is used to selectively deplete one allele while keeping the other intact. An sgRNA in conjunction with Cas9 targets a wild-type (*WT*) *KRAS* sequence. However, since the G12D (c.35G>A) mutation disrupts the PAM site, Cas9 does not efficiently cleave the mutant *KRAS* sequence. Subsequent amplification of all alleles using flanking primers, as in the case of digital PCR, Sanger sequencing, or high-throughput sequencing, is only effective for non-cleaved and mutant sites. **b** Three human genomic DNA samples with varying ratios of wild-type to mutant (G12D) *KRAS* were treated either with *KRAS*-targeted DASH, a non-human control DASH, or no DASH. Counts of intact wild-type and G12D sequences were then measured by droplet digital PCR (ddPCR). **c** Same data as in (**b**), presented as percentage of mutant sequences detected. *Inset* shows fold enrichment of the percentage of mutant sequences with *KRAS*-targeted DASH versus no DASH. For both (**b**) and (**c**), values and error bars are the average and standard deviation, respectively, of three independent experiments

the *B. mandrillaris* meningoencephalitis case demonstrates, the vast majority of sequence reads are "wasted" re-sequencing high abundance human transcripts. In this case, we have shown that DASH depletes with incredible specificity the small number of human rRNA transcripts that comprise the bulk of the NGS library, thereby lowering the required sequencing depth to detect non-human sequences and enriching the proportion of non-human (*Balamuthia*) reads in the metagenomic dataset. In this study, we have targeted mitochondrial rRNA species because we have consistently observed them to be the most abundant sequences in these CSF-derived RNA samples. For other types of tissues, alternative programming of DASH for removal of nuclear rRNA species or essentially any other abundant sequences would be warranted.

In the case of infectious agents, it is possible to directly enrich rare sequences by hybridization to DNA microarrays [34] or beads [12]. However, these approaches rely on sequence similarity between the target and the probe and therefore may miss highly divergent or unanticipated species. Furthermore, the complexity and cost of these approaches will continue to increase with the known spectrum of possible agents or targets. In contrast, the identity and abundance of unwanted sequences in most human tissues and sample types has been well described in scores of previous transcriptome profiling projects [23], and therefore optimized collections of sgRNAs for DASH depletion are likely to remain stable.

A number of methods for depleting ribosomal RNA from RNA-Seq libraries exist in the form of commercially available kits. We assert that DASH is equally effective or better than these methods on four metrics: (1) input requirements, (2) performance, (3) programmability, and (4) cost. These can be assessed based on information available on company websites or in publications for three major competing techniques: Illumina's Ribo-Zero and Thermo Fisher's RiboMinus, which both use biotinylated capture

probes for depletion; and New England Biolab's NEBNext rRNA depletion kit, which uses RNAse H for depletion.

Input requirements

Illumina recommends 1 μg of total RNA as input for Ribo-Zero, but also has a low-input protocol requiring only 100 ng [35]. ThermoFisher recommends 2–10 μg of total RNA for its standard RiboMinus protocol [36], and 100 ng to 1 μg for its Low Input RiboMinus Eukaryote System v.2 [37]. NEB recommends 10 ng to 1 μg total RNA input for the NEBNext rRNA Depletion Kit [38]. The reason for these stringent amount requirements is that these three methods all deplete samples at the RNA stage. DASH, in contrast, avoids the need to delicately manipulate the original sample. Instead, DASH is employed after cDNA synthesis and library generation; thus, it can be performed on any library, without regards to starting total RNA amount, or the manner in which the library was constructed (tagmentation or otherwise). For scarce and precious samples, such as patient CSF, often less than 10 ng of total cDNA is available even after NuGEN Ovation amplification; prior to this work, no commercial depletion method was available for these samples.

Performance

All commercial rRNA depletion methods promise at least 85 % reduction in reads of the sequences they target. Illumina states that the Ribo-Zero technique can achieve between 85 % and >99 % reduction in the rRNA sequences it targets [35]; RiboMinus states 95–98 % reduction [39]; and NEBNext states 95–99 % reduction [38]. Adiconis et al. [23] compared several RNA-Seq methods and reported on many metrics, including depletion of rRNA sequences. Ribosomal RNA sequences comprised 84.7 % of reads in their un-depleted sample (100 ng total RNA from K-562 cells), while Ribo-Zero reduced this to 11.3 % (an 86.7 % reduction), and RNAse H reduced it to 0.1 % (a 99.9 % reduction). In this paper, we show that DASH decreases the mitochondrial rRNA reads in HeLa total RNA from 61 % to 0.055 % (99.9 % reduction). Adiconis et al. obtained similar numbers from 1 μg total RNA samples from formalin-fixed paraffin-embedded (FFPE) kidney tissue (78.2 % and 99.9 % reduction for Ribo-Zero and RNAse H, respectively) and pancreas tissue (73.0 % and 99.7 % reduction for Ribo-Zero and RNAse H, respectively). This is comparable to DASH reduction in three patient CSF samples (82.1 %, 81.4 % and 88.2 % reduction). However, it is important to note again that Adiconis et al. used 1 μg total RNA from tissue samples, while the DASHed CSF samples consisted of only 5 ng of NuGEN Ovation-amplified cDNA (total RNA content in the original CSF samples was too low to accurately quantify).

Another important measure of performance is maintenance of relative abundances of non-targeted sequences, such as the human transcriptome. Correlation coefficients for samples with and without DASH treatment ranged from $R^2 = 0.979$ to 0.994 in this study (Fig. 2; Additional file 4), slightly higher than those found by Adiconis et al. for all methods [23].

Programmability

DASH can be adapted to target any sequence containing a PAM site; construction of new sgRNAs is facile and inexpensive (see "Materials and methods" section). Because it is employed after sequencing adapter addition, DASH's utility is not limited to RNA-Seq; it can be applied to any library type. Examples include ATAC-Seq libraries, in which desired nuclear DNA is contaminated with a significant amount of mitochondrial DNA sequences, and microbiome sequencing, where it may be desirable to eliminate a particularly abundant species in order to better sample the underlying diversity. Since Ribo-Zero, RiboMinus and NEBNext are all proprietary kits, they cannot easily be re-programmed by the user to target other sites.

Cost

Based on current publicly available list prices of the most economical kit sizes, the per-sample costs (in US dollars) of the kits discussed here are $82.00 (Ribo-Zero Gold Kit H/M/R) [35], $93.67 (RiboMinus Human/Mouse Transcriptome Isolation Kit) [36] and $45.00 (NEBNext rRNA Depletion Kit H/M/R) [38]. In contrast, we calculate the cost of DASH at less than $4 per sample when Cas9 and T7 RNA polymerase are made in-house — a very sensible solution for labs that are already spending large amounts of money on NGS. Where Cas9 production is not possible, DASH can still be carried out using commercially available Cas9 protein.

DASH may also enhance the detection of rare mutant alleles that are important for liquid biopsy cancer diagnostics. Allelic depletion with DASH increases the signal (oncogenic mutant allele) to noise (wild-type allele) by more than 60-fold when studying the *KRAS* hotspot mutant p.G12D. Other approaches for enriching low-abundance mutations exist, such as restriction enzyme digestion and COLD-PCR. However, these methods are limited when large mutation panels are required. Here we have described a single application for DASH in cancer, but the utility of this method will be fully realized by multiplexing large panels of mutation sites, using guide RNAs and PAM sites as a way to essentially create programmable restriction enzymes that can be used in a single pool. With the rapidly growing number of oncologic therapies that target particular cancer mutations,

sensitive and non-invasive techniques for cancer allele detection are increasingly relevant for optimizing patient care [26]. These same techniques are also becoming increasingly important for diagnosis of earlier stage (and generally more curable) cancers as well as the detection of cancer recurrence without needing to re-biopsy the patient [2, 14, 36–38].

The potential applications of DASH are manifold. Currently, DASH can be customized to deplete any set of defined PAM-adjacent sequences by designing specific libraries of sgRNAs. Given the popularity and promise of CRISPR technologies, we anticipate the adaptation and/or engineering of CRISPR-associated nucleases with more diverse PAM sites [31, 32, 40]. A portfolio of next-generation Cas9-like nucleases would further enable DASH to deplete large and diverse numbers of arbitrarily selected alleles across the genome without constraint. We envision that DASH will be immediately useful for the development of non-invasive diagnostic tools, with applications to low input samples or cell-free DNA, RNA, or methylation targets in body fluids [4, 6, 41–45].

Many other NGS applications could also benefit from depletion of specific sequences, including hemoglobin mRNA depletion for RNA-Seq of blood samples [46] and tRNA depletion for ribosome profiling studies. Depletion of pseudogenes or otherwise homologous sequences by small but consistent differences in sequences is also theoretically possible, and may serve to remove ambiguities in clinical high-throughput sequencing. Using DASH to enrich for minority variations in microbial samples may enable early discovery of pathogen drug resistance. Similarly, the application of DASH to the analysis of cell-free DNA may augment our ability to detect early markers of drug resistance in tumors [26].

Conclusions

Here, we have demonstrated the broad utility of DASH to enhance molecular signals in diagnostics and its potential to serve as an adaptable tool in basic science research. While the degree of regional depletion of mitochondrial rRNA was sufficient for our application, the depletion parameters were not maximized: we used only 54 sgRNA target sites out of about 250 possible *S. pyogenes* Cas9 sgRNA candidates in the targeted mitochondrial region. Future studies will explore the upper limit of this system while elucidating the most effective sgRNA and CRISPR-associated nuclease selections, which will likely differ based on target and application. Irrespective, depletion of unwanted sequences by DASH is highly generalizable and may effectively lower costs and increase meaningful output across a broad range of sequence-based approaches.

Materials and methods

Generation of cDNA from HeLa cell line and clinical samples

CSF samples were collected under the approval of the institutional review boards of the University of California San Francisco and San Francisco General Hospital. Samples were processed for high-throughput sequencing as previously described [1, 25]. Briefly, amplified cDNAs were made from randomly primed total RNA extracted from 250 μL of CSF or 250 pg of HeLa RNA using the NuGEN Ovation v.2 kit (NuGEN, San Carlos, CA, USA) for low nucleic acid content samples. A Nextera protocol (Illumina, San Diego, CA, USA) was used to add on a partial sequencing adapter on both sides.

In vitro preparation of the CRISPR/Cas9 complex

The Cas9 expression vector, containing an N-terminal MBP tag and C-terminal mCherry, was kindly provided by Dr. Jennifer Doudna. The protein was expressed in BL21 Rosetta cells for three hours at 18 °C. Cells were pelleted and frozen. Upon thawing, cells from a 4 L culture preparation were resuspended in 50 mL of lysis buffer (50 mM sodium phosphate pH 6.5, 350 mM NaCl, 1 mM TCEP (tris(2-carboxyethyl)phosphine), 10 % glycerol) supplemented with 0.5 mM EDTA, 1 μM PMSF (phenylmethanesulfonyl), and a single Roche complete EDTA-free protease inhibitor tablet (Roche Diagnostics, Indianapolis, IN, USA) and passed through an HC-8000 homogenizer (Microfluidics, Westwood, MA, USA) five times. The lysate was clarified by centrifugation at 20,000 rpm for 45 min at 4 °C and then filtered through a 0.22 μm vacuum filtration unit. The filtered lysate was loaded onto three 5 mL HiTrap Heparin HP columns (GE Healthcare, Little Chalfont, UK) arranged in series on a GE AKTA Pure system. The columns were washed extensively with lysis buffer, and the protein was eluted with a gradient of lysis buffer to buffer B (lysis buffer supplemented with NaCl up to 1.5 M). The resulting fractions were analyzed by Coomassie gel, and those containing Cas9 (centered around the point on the gradient corresponding to 750 mM NaCl) were combined and concentrated down to a volume of 1 mL using 50 K MWCO Amicon Ultra-15 Centrifugal Filter Units (EMD Millipore, Billerica, MA, USA) and then fed through a 0.22 μm syringe filter. Using the AKTA Pure, the 1 mL of filtered protein solution was then injected onto a HiLoad 16/600 Superdex 200 size exclusion column (GE Healthcare, Little Chalfont, UK) pre-equilibrated with buffer C (lysis buffer supplemented with NaCl up to 750 mM). Resulting fractions were again analyzed by Coomassie gel, and those containing purified Cas9 were combined, concentrated, supplemented with glycerol up to a final concentration of 50 %, and frozen at −80 °C until use. Protein concentration was determined by BCA

assay. Yield was approximately 80 mg from 4 L of bacterial culture.

sgRNA target sites were selected as described in the main text. DNA templates for sgRNAs based on an optimized scaffold [47] were made with a similar method to that described in [48]. For each chosen target, a 60mer oligo was purchased including the 18-base T7 transcription start site, the targeted 20mer, and the first 22 bases of the tracr RNA (5′-TAATACGACTCACTATAGNN NNNNNNNNNNNNNNNNNGTTTAAGAGCTATG CTGGAAAC-3′). This was mixed with a 90mer representing the 3′ end of the sgRNA on the opposite strand (5′- AAAAAAAGCACCGACTCGGTGCCACTTTTTC AAGTTGATAACGGACTAGCCTTATTTAAACTTGC TATGCTGTTTCCAGCATAGCTCTTA-3′). DNA templates for T7 sgRNA transcription were then assembled and amplified with a single PCR reaction using primers 5′- TAATACGACTCACTATAG-3′ and 5′- AAAAAA AGCACCGACTCGGTGC-3′. The resulting 131 base pair (bp) transcription templates, with the sequence 5′- TAATACGACTCACTATAGNNNNNNNNNNNNN NNNNNNNGTTTAAGAGCTATGCTGGAAACAGCA TAGCAAGTTTAAATAAGGCTAGTCCGTTATCAAC TTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′, were pooled (for the mitochondrial rRNA library), or transcribed separately (for the *KRAS* experiments). All oligos were purchased from IDT (Integrated DNA Technologies, Coralville, IA, USA).

Transcription was performed using custom-made T7 RNA polymerase (RNAP) [49, 50] In each 50 μL reaction, 300 ng of DNA template was mixed with T7 RNAP (final concentration 8 ng/μL), buffer (final concentrations of 40 mM Tris pH 8.0, 20 mM $MgCl_2$, 5 mM DTT, and 2 mM spermidine), and Ambion brand NTPs (ThermoFisher Scientific, Waltham, MA, USA) (final concentration 1 mM each ATP, CTP, GTP and UTP), and incubated at 37 °C for 4 h. Typical yields were 2–20 μg of RNA. sgRNAs were purified with a Zymo RNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA, USA), aliquoted, stored at −80 °C, and used only a single time after thawing.

CRISPR/Cas9 treatment

To form the ribonucleoprotein (RNP) complex, Cas9 and the sgRNAs were mixed at the desired ratio with Cas9 buffer (final concentrations of 50 mM Tris pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$, and 1 mM TCEP), and incubated at 37 °C for 10 min. This complex was then mixed with the desired amount of sample cDNA in a total of 20 μL, again in the presence of Cas9 buffer, and incubated for 2 h at 37 °C.

Since Cas9 has high nonspecific affinity for DNA [24] it was necessary to disable and remove the Cas9 before continuing. For the rRNA depletion samples, 1 μL (at >600 mAU/mL) of Proteinase K (Qiagen, Hilden, Germany) was added to each sample which was then incubated for an additional 15 min at 37 °C. Samples were then expanded to a volume of 100 μL and purified with three phenol:chloroform:isoamyl alcohol extractions followed by one chloroform extraction in 2 mL Phaselock Heavy tubes (5prime, Hilden, Germany). We added 10 μL of 3 M sodium acetate pH 5.5, 3 μL of linear acrylamide and 226 μL of 100 % ethanol to the 100 μL aqueous phase of each sample. Samples were cooled on ice for 30 min. DNA was then pelleted at 4 °C for 45 min, washed once with 70 % ethanol, dried at room temperature and resuspended in 10 μL water.

In the case of the *KRAS* samples, Cas9 was disabled by heating the sample at 95 °C for 15 min in a thermocycler and then removed by purifying the sample with a Zymo DNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA, USA).

High-throughput sequencing and analysis of sequencing data

Tagmented samples with and without DASH treatment underwent 10–12 cycles of additional amplification (Kapa Amplification Kit, Kapa Biosystems, Wilmington, MA, USA) with dual-indexing primers. A BluePippin instrument (Sage Science, Beverly, MA, USA) was used to extract DNA between 360 and 540 bp. Sequencing libraries were purified using the Zymo DNA Clean & Concentrator-5 kit and amplified again on an Opticon qPCR machine (MJ Research, Waltham, MA, USA) using a Kapa Library Amplification Kit until the exponential portion of the quantitative PCR signal was found. Sequencing libraries were then pooled and re-quantified with a ddPCR Library Quantification Kit (Bio-Rad, Hercules, CA, USA). Sequencing was performed on portions of one lane in an Illumina HiSeq 4000 instrument using 135 bp paired-end sequencing.

All reads were quality filtered using PriceSeqFilter v.1.2 [51] such that only read pairs with less than five ambiguous base calls (defined as Ns or positions with <95 % confidence based on Phred score) were retained. Filtered reads were aligned to the hg38 build of the human genome using the STAR aligner (v.2.4.2a) [52]. The number of mapped reads per gene and fpkm values were calculated using the exon length and sequence information encoded in the Gencode v.23 primary annotations (GTF file). Library complexity was determined by calculating the reduction in library size after clustering using the cd-hit-dup package [53, 54]. Pathogen-specific alignments to 16S and 18S sequences were accomplished using Bowtie2 [55]. Per-nucleotide coverage was calculated from alignment (SAM/BAM) files using the SAMtools suite [56] and analyzed with custom iPython [57] scripts utilizing the Pandas data package. Plots were generated with Matplotlib [58].

dPCR of *KRAS* mutant DNA

KRAS wild-type DNA was obtained from a healthy consenting volunteer. The sample sat until cell separation occurred, and DNA was extracted from the buffy coat with the QIAamp Blood Mini Kit (Qiagen, Hilden, Germany). *KRAS* G12D genomic DNA from the human leukemia cell line CCRF-CEM was purchased from ATCC (Manassas, VA, USA). All DNA was sheared to an average of 800 bp using a Covaris M220 (Covaris, Woburn, USA) following the manufacturer's recommended settings. Cas9 reactions occurred as described above.

A primer/probe pair was designed with Primer3 [59, 60] targeting the relatively common *KRAS* G12D (c.35G>A) mutation. Reactions were themocycled according to manufacturer protocols using a two-step PCR. An ideal 62 °C annealing/extension temperature was determined by a gradient experiment to ensure proper separation of FAM and HEX signals. The PCR primers and probes used were as follows (purchased from IDT): forward 5′- TAGCTG TATCGTCAAGGCAC-3′, reverse 5′-GGCCTGCTGAA AATGACTGA-3′; wild-type probe, 5′-/5HEX/TGCCT ACGC/ZEN/CA<C>CAGCTCCA/3IABkFQ/-3′; mutant probe, 5′-/56-FAM/TGCCTACGC/ZEN/CA<T>CAGCT CCA/3IABkFQ/-3′, with <> denoting the mutant base location, 5HEX and 56-FAM denoting the HEX and FAM reporters, and ZEN and 3IABkFQ denoting the internal and 3′ quenchers. Original samples and those subjected to DASH were measured with the ddPCR assay on a Bio-Rad QX100 Droplet Digital PCR system (Bio-Rad, Hercules, CA, USA), following the manufacturer's instructions for droplet generation, PCR amplification, and droplet reading, and using best practices. Pure CCRF-CEM samples were approximately 30 % G12D and 70 % wild type; all calculations of starting mixtures were made based on this starting ratio.

Ethics

CSF samples, as well as a whole blood sample for the KRAS negative control, were collected under the approval of the institutional review boards of the University of California San Francisco and San Francisco General Hospital (IRB number 13-12236). All experimental methods comply with the Helsinki Declaration.

Availability of data and materials

All sequencing data for human subjects has been deposited to NCBI's database of Genotypes and Phenotypes (dbGaP) and can be accessed at http://www.ncbi.nlm.nih.gov/gap by entering study accession number phs001067.v1.p1. Sequencing data for HeLa samples has been deposited as a separate BioProject in NCBI's Sequence Read Archive (SRA) and can be found at http://www.ncbi.nlm.nih.gov/bioproject by entering study accession number PRJNA311047. Reagents are available upon request from J.L.D.

Additional files

Additional file 1: Table S1. List of all sgRNA sequences used in this paper. (PDF 34 kb)

Additional file 2: Figure S1. Depletion efficiency by Cas9 dosage. (PDF 10688 kb)

Additional file 3: Figure S2. PAM sites in a pseudogene. (PDF 442 kb)

Additional file 4: Figure S3. Scatter plots for patient samples. (PDF 8093 kb)

Abbreviations

bp: Base pair; Cas: CRISPR-associated; cDNA: Complementary DNA; CRISPR: Clustered regularly interspaced short palindromic repeats; CSF: Cerebrospinal fluid; DASH: Depletion of Abundant Sequences by Hybridization; ddPCR: droplet digital PCR; dPCR: Digital polymerase chain reaction; fpkm: Fragments per kilobase of transcript per million mapped reads; NGS: Next generation sequencing; PAM: Protospacer adjacent motif; rRNA: Ribosomal RNA; sgRNA: Single guide RNA.

Competing interests

W.G. is a consultant for Pacgeno and holds patents relevant to microfluidics and non-invasive prenatal diagnostics. E.D.Cr., B.D.O., M.R.W., E.D.Ch., H.R., and J.L.D. have no competing interests.

Authors' contributions

WG and EDCr contributed equally to this work and are co-first authors. WG, EDCr, BDO, MRW, HR, EDCh, and JLD conceived of the project. EDCr, WG, MRW, and JLD designed and performed the experiments. MRW oversaw the encephalitis study and prepared the original NGS libraries used to identify pathogens and determine high-abundance sequences. EDCr produced and optimized the in vitro Cas9 and sgRNA system and directed its usage. EDCh performed sequencing. BDO, WG, EDCr, and JLD analyzed the data. EDCr, WG, BDO, and JLD drafted the manuscript with input and edits from all other authors. All authors read and approved the final manuscript.

Acknowledgements

We thank Derek Bogdanoff and the Center for Advanced Technologies at UCSF for assistance with sequencing and help with equipment; Hannah Sample for assistance with sample collection; Dr. Joe Kliegman for HeLa RNA; Jennifer Mann for administrative support; Dr. Chong Park and the Innovative Genomics Initiative for help with sgRNA production protocols; Lara Pesce Ares for help with Cas9 purification; Drs. Sy Redding, Winston Koh, and Charles Chiu for helpful discussions; and Drs. Jeffrey Gelfand, Luke Strnad and Niraj Shanbhag for patient referrals. We would also like to thank the patients and their families who volunteered to participate in our research program.

Funding

This work was supported by the Sandler Foundation, the William K. Bowes, Jr. Foundation, Howard Hughes Medical Institute (to E.D.Cr. and J.L.D.), and the National Center for Advancing Translational Sciences of the NIH [KL2TR000143] (to M.R.W.). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Author details

[1]Departments of Pathology and Laboratory Medicine, University of California San Francisco, San Francisco, CA, USA. [2]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA. [3]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [4]Integrative Program in Quantitative Biology, Bioinformatics, University of California San Francisco, San Francisco, CA, USA. [5]Department of Neurology, University of California San Francisco, San Francisco, CA, USA. [6]Center for Advanced Technology, Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA. [7]Medical Scientist Training Program, Biomedical Sciences Graduate Program, University of California San Francisco, San Francisco, CA, USA.

Received: 20 November 2015 Accepted: 18 February 2016
Published online: 04 March 2016

References

1. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med. 2014;370:2408–17.
2. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014;6:224ra24–4.
3. Pan W, Gu W, Nagpal S, Gephart MH, Quake SR. Brain tumor mutations detected in cerebral spinal fluid. Clin Chem. 2015;61:514–22.
4. De Vlaminck I, Valantine HA, Snyder TM, Strehl C, Cohen G, Luikart H, et al. Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. Sci Transl Med. 2014;6:241ra77–7.
5. Fan HC, Gu W, Wang J, Blumenfeld YJ, El-Sayed YY, Quake SR. Non-invasive prenatal measurement of the fetal genome. Nature. 2012;487:320–4.
6. Gu W, Koh W, Blumenfeld YJ, El-Sayed YY, Hudgins L, Hintz SR, et al. Noninvasive prenatal diagnosis in a fetus at risk for methylmalonic acidemia. Genet Med. 2014;16:564–7.
7. Vogelstein B, Kinzler KW. Digital PCR. Proc Natl Acad Sci U S A. 1999;96:9236–41.
8. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339:819–23.
9. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014;346:1258096.
10. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014;157:1262–78.
11. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816–21.
12. Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, Lipkin WI. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. mBio. 2015;6:1–11.
13. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011;29:908–14.
14. Newman AM, Bratman SV, To J, Wynne JF, Eclov NCW, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014;20:548–54.
15. Zou H, Allawi H, Cao X, Domanico M, Harrington J, Taylor WR, et al. Quantification of methylated markers with a multiplex methylation-specific technology. Clin Chem. 2012;58:375–83.
16. Akhras MS, Unemo M, Thiyagarajan S, Nyrén P, Davis RW, Fire AZ, et al. Connector inversion probe technology: a powerful one-primer multiplex DNA amplification system for numerous scientific applications. PLoS One. 2007;2:e915.
17. Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, Shendure J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res. 2013;23:843–54.
18. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J. Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods. 2009;6:315–6.
19. Li J, Wang L, Mamon H, Kulke MH, Berbeco R, Makrigiorgos GM. Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. Nat Med. 2008;14:579–84.
20. Didelot A, Le Corre D, Luscan A, Cazes A, Pallier K, Emile J-F, et al. Competitive allele specific TaqMan PCR for KRAS, BRAF and EGFR mutation detection in clinical formalin fixed paraffin embedded samples. Exp Mol Pathol. 2012;92:275–80.
21. Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H, et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230:1350–4.
22. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. 2011;108:9530–5.
23. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods. 2013;10:623–9.
24. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014;507:62–7.
25. Wilson MR, Shanbhag NM, Reid MJ, Singhal NS, Gelfand JM, Sample HA, et al. Diagnosing Balamuthia mandrillaris encephalitis with metagenomic deep sequencing. Ann Neurol. 2015;78:722–30.
26. Oxnard GR, Paweletz CP, Kuang Y, Mach SL, O'Connell A, Messineo MM, et al. Noninvasive detection of response and resistance in EGFR-mutant lung cancer using quantitative next-generation genotyping of cell-free plasma DNA. Clin Cancer Res. 2014;20:1698–705.
27. Almoguera C, Shibata D, Forrester K, Martin J, Arnheim N, Perucho M. Most human carcinomas of the exocrine pancreas contain mutant c-K-ras genes. Cell. 1988;53:549–54.
28. Burmer GC, Loeb LA. Mutations in the KRAS2 oncogene during progressive stages of human colon carcinoma. Proc Natl Acad Sci U S A. 1989;86:2403–7.
29. Tam IYS, Chung LP, Suen WS, Wang E, Wong MCM, Ho KK, et al. Distinct epidermal growth factor receptor and KRAS mutation patterns in non-small cell lung cancer patients with different tobacco exposure and clinicopathologic features. Clin Cancer Res. 2006;12:1647–53.
30. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.
31. Kleinstiver BP, Prew MS, Tsai SQ, Nguyen NT, Topkar VV, Zheng Z, Joung JK. Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. Nat Biotechnol. 2015; advance online publication.
32. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015;163:759–71.
33. Granerod J, Tam CC, Crowcroft NS, Davies NWS, Borchert M, Thomas SL. Challenge of the unknown A systematic review of acute encephalitis in non-outbreak situations. Neurology. 2010;75:924–32.
34. Wang D, Urisman A, Liu Y-T, Springer M, Ksiazek TG, Erdman DD, et al. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. 2003;1:e2.
35. Ribo-Zero Gold rRNA Removal Kit (Human/Mouse/Rat). http://www.illumina.com/products/ribo-zero-gold-rrna-removal-human-mouse-rat.html. Accessed 5 Jan 2016.
36. RiboMinus Human/Mouse Transcriptome Isolation Kit. http://www.thermofisher.com/order/catalog/product/K155001. Accessed 5 Jan 2016.
37. Low Input RiboMinus™ Eukaryote System v2 (pub. no. MAN0007160, rev. 2.0). http://tools.thermofisher.com/content/sfs/manuals/MAN0007160_RiboMinus_Eukaryote_V2_LowInput_UG_08Jan2013.pdf. Accessed 5 Jan 2016.
38. NEBNext® rRNA Depletion Kit (Human/Mouse/Rat). https://www.neb.com/products/e6310-nebnext-rrna-depletion-kit-human-mouse-rat. Accessed 5 Jan 2016.
39. Transcriptome enrichment without ribosomal RNA for improved microarray analysis. https://www.thermofisher.com/content/dam/LifeTech/migration/en/filelibrary/nucleic-acid-purification-analysis/pdfs.par.83981.file.dat/f-075051-ribominus-lrf.pdf. Accessed 5 Jan 2016.
40. Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature. 2015;520:186–91.
41. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med. 2014;370:1287–97.
42. Kinde I, Munari E, Faraj SF, Hruban RH, Schoenberg M, Bivalacqua T, et al. TERT promoter mutations occur early in urothelial neoplasia and are biomarkers of early disease and disease recurrence in urine. Cancer Res. 2013;73:7162–7.
43. Li M, Chen W, Papadopoulos N, Goodman S, Bjerregaard NC, Laurberg S, et al. Sensitive digital quantification of DNA methylation in clinical samples. Nat Biotechnol. 2009;27:858–63.
44. Koh W, Pan W, Gawad C, Fan HC, Kerchner GA, Wyss-Coray T, et al. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. Proc Natl Acad Sci U S A. 2014;111:7361–6.
45. Zheng Z, Liebers M, Zhelyazkova B, Cao Y, Panditi D, Lynch KD, et al. Anchored multiplex PCR for targeted next-generation sequencing. Nat Med. 2014;20:1479–84.
46. Shin H, Shannon CP, Fishbane N, Ruan J, Zhou M, Balshaw R, et al. Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. PLoS One. 2014;9:e91041.
47. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li G-W, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell. 2013;155:1479–91.

48. Lin S, Staahl BT, Alla RK, Doudna JA. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. eLife. 2015;3:e04766.
49. Davanloo P, Rosenberg AH, Dunn JJ, Studier FW. Cloning and expression of the gene for bacteriophage T7 RNA polymerase. Proc Natl Acad Sci U S A. 1984;81:2035–9.
50. Zawadzki V, Gross HJ. Rapid and simple purification of T7 RNA polymerase. Nucleic Acids Res. 1991;19:1948.
51. Ruby JG, Bellare P, DeRisi JL. PRICE: software for the targeted assembly of components of (meta) genomic sequence data. G3 Genes Genomes Genetics. 2013;3:865–80.
52. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
53. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
54. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
55. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009;25:2078–9.
57. Pérez F, Granger BE. IPython: A system for interactive scientific computing. Comput Sci Eng. 2007;9:21–9.
58. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9:90–5.
59. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007;23:1289–91.
60. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012;40:e115–5.