Sample processing obscures cancer-specific alterations in leukemic transcriptomes

Heidi Dvinge[a,b], Rhonda E. Ries[c], Janine O. Ilagan[a,b], Derek L. Stirewalt[c], Soheil Meshinchi[c,d], and Robert K. Bradley[a,b,1]

[a]Computational Biology Program, Public Health Sciences Division, [b]Basic Sciences Division, and [c]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109; and [d]Division of Pediatric Hematology/Oncology, School of Medicine, University of Washington, Seattle, WA 98195

Edited* by Robert N. Eisenman, Fred Hutchinson Cancer Research Center, Seattle, WA, and approved October 14, 2014 (received for review July 14, 2014)

Substantial effort is currently devoted to identifying cancer-associated alterations using genomics. Here, we show that standard blood collection procedures rapidly change the transcriptional and posttranscriptional landscapes of hematopoietic cells, resulting in biased activation of specific biological pathways; up-regulation of pseudogenes, antisense RNAs, and unannotated coding isoforms; and RNA surveillance inhibition. Affected genes include common mutational targets and thousands of other genes participating in processes such as chromatin modification, RNA splicing, T- and B-cell activation, and NF-κB signaling. The majority of published leukemic transcriptomes exhibit signals of this incubation-induced dysregulation, explaining up to 40% of differences in gene expression and alternative splicing between leukemias and reference normal transcriptomes. The effects of sample processing are particularly evident in pan-cancer analyses. We provide biomarkers that detect prolonged incubation of individual samples and show that keeping blood on ice markedly reduces changes to the transcriptome. In addition to highlighting the potentially confounding effects of technical artifacts in cancer genomics data, our study emphasizes the need to survey the diversity of normal as well as neoplastic cells when characterizing tumors.

leukemia | RNA splicing | nonsense-mediated decay | batch effects

Recent years have seen rapid growth in the large-scale characterization of tumors by consortia such as The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium, and the Therapeutically Applicable Research to Generate Effective Treatments project. Although the high-throughput assays used by these consortia are precise, analyses of the resulting data can be confounded by artifacts arising from both biological and technical variability. Cancer studies are susceptible to both "sample-intrinsic" (arising from sample handling itself, including specimen collection, sample transfer and storage, and isolation of material) and "assay-intrinsic" (associated with specific assays, such as microarrays or high-throughput sequencing, or differences in equipment, reagents, or personnel) artifacts. Sample-intrinsic artifacts depend on the tissue assayed, may distinguish otherwise similar biological specimens, and typically cannot be eliminated without discarding affected biospecimens. In contrast, assay-intrinsic artifacts are agnostic to the biological specimen and frequently can be experimentally or statistically mitigated once recognized (1).

Although assay-intrinsic artifacts have received widespread attention in the context of genomics data (1–5), the impact of sample-intrinsic artifacts on high-throughput studies has been less well explored. Nonetheless, previous reports suggest that sample-intrinsic artifacts are likely important to consider when interpreting genomics data. For example, blood collection procedures and the choice of anticoagulant can impact subsequent clinical chemistry assays or affect hematological parameters such as cell number and morphology (6, 7). Similarly, blood shipping temperature and the timing of sample processing can affect levels of protein-based biomarkers and other analytes in plasma or serum, alter the transcription (e.g., *IL8*, *IL10*, *CCR2*, *SOCS2*, or *JUN*) or splicing (e.g., *NF1*, *PTEN*, or *ATM*) of specific genes, or cause globally decreased/increased transcription or accelerated RNA degradation, depending upon conditions (8–15).

Genomic studies of leukemias may be particularly susceptible to sample-intrinsic artifacts in comparison with similar profiling of solid tumors. For many solid tumors, patient-matched normal samples can be acquired from adjacent uninvolved tissue, and the matched tumor/normal samples can be subsequently handled similarly. In contrast, the circulating nature of leukemic cells renders the acquisition of patient-matched controls less straightforward, so samples from unrelated healthy donors are typically used to generate reference normal transcriptomes (Fig. 1*A*). Leukemic samples are commonly collected by the treating physician and then shipped for a variable length of time as whole blood or bone marrow to a research center for subsequent processing (Fig. 1*B*). In contrast, control samples are collected expressly for research rather than clinical use and therefore may be more rapidly processed by the collecting research institution or company.

The substantial literature documenting transcriptional changes caused by ex vivo incubation suggests that comparing the transcriptomes of differentially handled leukemic and normal samples is likely problematic. However, the impact of ex vivo incubation on the transcriptome has not been systematically assessed with high-throughput sequencing, and the potential extent of incubation-induced artifacts within large-scale leukemia studies has not been measured. Here, we identify transcriptional and post-transcriptional changes caused by ex vivo sample incubation and search for signs of these artifacts in published leukemia studies (Table S1). Our results show that widespread biological changes caused by ex vivo sample incubation can confound the identification of cancer-specific alterations in many leukemia genomics studies.

Significance

An important goal of cancer biology is to identify molecular differences between normal and cancer cells. Accordingly, many large-scale initiatives to characterize both solid and liquid tumor samples with genomics technologies are currently underway. Here, we show that standard blood collection procedures cause rapid changes to the transcriptomes of hematopoietic cells. The resulting transcriptional and posttranscriptional artifacts are visible in most published leukemia genomics datasets and hinder the identification and interpretation of cancer-specific alterations.

Author contributions: H.D., R.E.R., S.M., and R.K.B. designed research; H.D. and J.O.I. performed research; D.L.S. contributed new reagents/analytic tools; H.D., R.E.R., S.M., and R.K.B. analyzed data; and H.D. and R.K.B. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

Data deposition: The RNA-sequencing data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GSE58335 and GSE61410).

[1]To whom correspondence should be addressed. Email: rbradley@fhcrc.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1413374111/-/DCSupplemental.

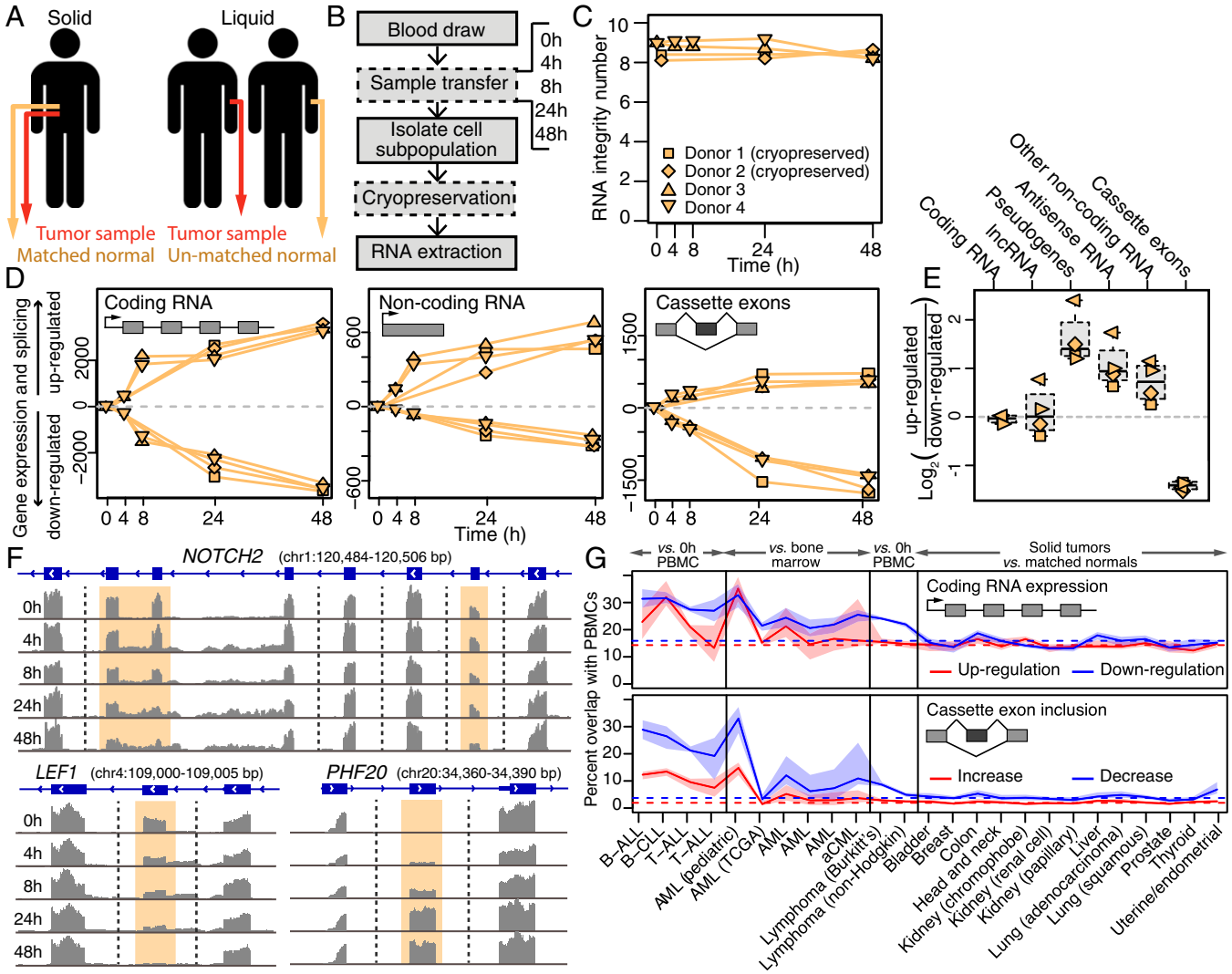

Fig. 1. Ex vivo blood incubation causes rapid transcriptional and posttranscriptional changes. (*A*) Biospecimen collection for solid and liquid tumors. (*B*) Blood sample processing. Whole blood samples frequently incur an ex vivo incubation between collection and processing, which we mimicked in the indicated time series. (*C*) RNA integrity numbers (RINs) from four individual healthy PBMC donors. For the cryopreserved samples (donors 1 and 2), the first time point was at 1 h rather than 0 h. (*D*) Numbers of differentially expressed protein-coding and noncoding transcripts and differentially spliced cassette exons relative to the first (0 or 1 h) time point. Legend is as in *C*. (*E*) Log_2 ratio of numbers of up- vs. down-regulated transcripts or cassette exons with increased vs. decreased inclusion at 48 h. Legend is as in *C*. (*F*) RNA-seq coverage along excerpts of *NOTCH2*, *LEF1*, and *PHF20* (donor 4). Introns are truncated at the vertical dashed lines. The inclusion of specific exons or introns (orange boxes) is time-dependent. (*G*) Overlap between differential gene expression or splicing in PBMCs (0 vs. 48 h; intersection of donors 3 and 4) and tumors vs. normal controls. Solid lines, median overlap per dataset; shading, first and third quartiles of the overlap; dashed lines, median across all solid tumors. Differential gene expression or splicing was computed for each tumor sample individually; the illustrated quantiles were computed over all tumor samples for each dataset. The control samples are as follows: lymphoid leukemias, $t = 0$h PBMCs; myeloid leukemias, median of four normal bone marrow samples; lymphomas, 0h PBMCs; B-ALL, B-cell acute lymphocytic leukemia; B-CLL, B-cell chronic lymphocytic leukemia; T-ALL, T-cell acute lymphocytic leukemia; AML, acute myeloid leukemia; aCML, atypical chronic myeloid leukemia. From left to right, datasets are from: lymphoid leukemias (20–23), myeloid leukemias [Database of Genotypes and Phenotypes (dbGaP) study no. 2447; refs. 22, 24–27], lymphomas (29, 30), and solid tumors (TCGA).

Results

Ex Vivo Incubation of Blood Causes Widespread Genomic Alterations. We collected whole blood from two male and two female healthy donors in anticoagulant blood collection tubes, which conforms to standard practice. We left this whole blood at room temperature for defined lengths of time (0-48 h), isolated peripheral blood mononuclear cells (PBMCs), and extracted RNA (Table S2). For two donors, we additionally tested the effect of cryopreservation by introducing a liquid nitrogen freeze–thaw cycle before RNA extraction. RNA Integrity Number (RIN) measurements were stable across the incubation period for all four donors, suggesting that RNA quality is maintained during whole blood incubation or cryopreservation (Fig. 1*C*). We measured genome-wide RNA abundance with the Illumina HiSeq 2500, and found that transcript abundance was highly reproducible between the different donors at each time point (Fig. S1).

Simply counting total numbers of differentially expressed coding and noncoding genes and alternatively spliced cassette exons relative to the 0h time point demonstrated that rapid and widespread changes affected virtually all levels of the gene expression process (Fig. 1*D*). We observed highly similar time-dependent changes for samples that were or were not cryopreserved, suggesting that transcriptional and posttranscriptional changes introduced during ex vivo incubation are relatively unaffected by long-term storage of samples. We henceforth describe data obtained from the fresh (not cryopreserved) PBMC time course,

but our conclusions were unaffected when we instead used the cryopreserved time course.

Transcriptional and posttranscriptional alterations that occurred during incubation were not random. For example, pseudogenes, antisense RNAs, and other abnormal noncoding RNAs were preferentially up-regulated, and cassette exons were preferentially excluded rather than included (Fig. 1*E*). We observed widespread isoform switches, wherein isoforms that were rare or even undetectable at 0 h became the major isoform after 8–24 h of incubation (Fig. 1*F*). Incubation-induced isoform switches occurred in genes that are reportedly misspliced in leukemic relative to normal hematopoietic cells [*NOTCH2* (16)], that have been used as prognostic markers in leukemia [*LEF1* (17, 18)], or that participate in cancer-relevant biological pathways such as NF-κB signaling [*PHF20* (19)].

Leukemic Transcriptomes Exhibit Gene Expression Signatures of Incubation. We next sought to determine whether the effects of sample incubation were detectable in published leukemic transcriptomes. We identified differentially expressed genes and alternatively spliced cassette exons in four lymphoid (20–23) and six myeloid (22, 24–27) leukemia studies. For the lymphoid leukemias, we compared with our 0h PBMC samples as a normal control; for the myeloid leukemias, we compared with mononuclear cells isolated from four commercially purchased normal bone marrow samples (28). We then computed the overlap between putative cancer-dysregulated coding genes and cassette exons—e.g., differentially expressed genes or differentially spliced cassette exons in lymphoid/myeloid leukemias vs. normal PBMCs/bone marrow—with coding genes and cassette exons that were altered by incubation (Fig. 1*G*). All leukemic transcriptomes exhibited substantial overlap with incubation-induced alterations, particularly for genes and cassette exons with decreased expression or inclusion in tumors. The magnitude of this effect varied substantially both within and between datasets. We next identified cancer-dysregulated genes and cassette exons across a broad panel of lymphoma (29, 30) and solid tumor samples (TCGA), which are typically obtained from flash frozen or otherwise rapidly stabilized biopsies. In contrast to leukemias, lymphomas and solid tumors exhibited relatively low levels of overlap with incubation-induced changes in RNA expression and splicing, as well as little inter- or intradataset variability (Fig. 1*G*). We conclude that sample incubation likely contributes to interdataset as well as intradataset variability in leukemic transcriptomes.

To determine which biological pathways are sensitive to sample incubation, we identified enriched Gene Ontology (GO) terms among differentially expressed genes after 24 h (Datasets S1 and S2). Many of these terms correspond to cancer-relevant biological processes, such as immune cell activation, cytokine production, NF-κB signaling, chromatin modification, and RNA splicing. These processes all exhibited differential expression after only 4 h of incubation, our shortest time point, but the magnitude of the response continued to increase throughout the entire time course (Fig. 2*A*). Differentially expressed genes in leukemic relative to normal transcriptomes were similarly enriched for these biological pathways (Fig. 2*B*). Processes such as chromatin modification and NF-κB signaling play important roles in many leukemias, suggesting that these gene expression signals represent true cancer biology. However, it is difficult to confirm cancer-specific pathway activation in the presence of incubation as an uncorrected source of variability. For example, genes that are differentially expressed upon incubation include common mutational targets in leukemia (*IDH1*, *EZH2*, *TP53*, *SRSF2*, and *U2AF1*), genes that are frequently affected by chromosomal translocations (*MLL*), and targets of cancer therapeutics (*HDAC1*).

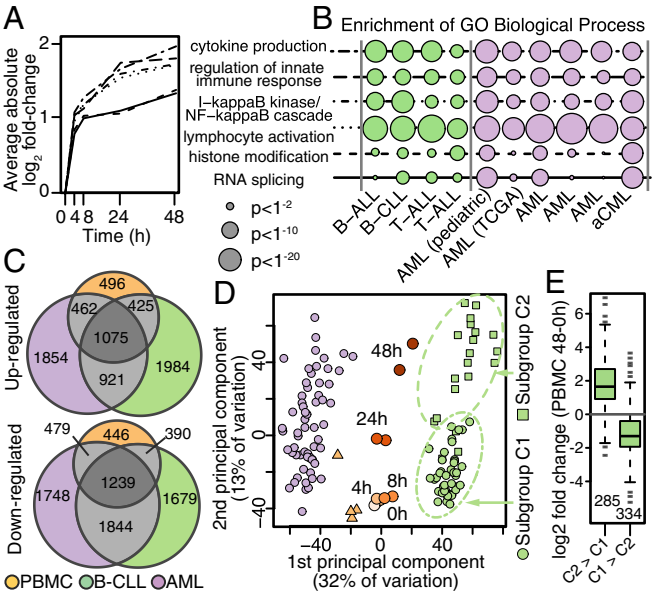

Fig. 2. Sample incubation affects the interpretation of leukemic gene expression. (*A*) Average absolute $\log_2$ fold change of differentially expressed genes (intersection of donors 3 and 4) associated with select GO terms enriched during ex vivo sample incubation. (*B*) Enrichment of the GO terms from *A* in differentially expressed genes in tumors vs. normal controls across a panel of lymphoid (green) and myeloid (purple) leukemias. Dataset order and references are as in Fig. 1*G*. Only genes differentially expressed in >25% of samples within each dataset were included in the GO enrichment analysis. (*C*) Overlap between up- and down-regulated genes, calculated as in *B*. Orange, PBMCs (0 vs. 48 h); green, B-CLL (21); purple, AML (dbGaP study no. 2447). (*D*) Principal components analysis of PBMCs (orange hues), B-CLL (green) (21), AML (purple), and normal bone marrow samples (28) (orange triangles). Clustering was performed by using 6,756 protein-coding genes with normalized expression more than five transcripts per million in ≥90% of samples. B-CLL samples are marked according to the proposed molecular subgroups C1 and C2 (31). (*E*) $\log_2$ fold change after 48 h of incubation (intersection of donors 3 and 4) of genes differentially expressed between subgroups C1 and C2 (31), divided according to group with the highest expression level. Numbers indicate genes within each subtype.

The effects of sample incubation could be especially apparent when performing pan-cancer analyses. To test this hypothesis, we identified differentially expressed genes in 61 pediatric acute myeloid leukemia (AML) samples relative to normal bone marrow mononuclear cells, as well as in 61 chronic lymphocytic leukemia (CLL) (31) samples relative to normal 0h PBMCs. Because AML and CLL represent distinct cell types and were analyzed with respect to different control datasets, random overlap in gene expression patterns is unlikely. For both of these studies, approximately one-third of the genes that were either up- or down-regulated in the leukemic samples were likewise differentially expressed upon sample incubation, indicating that incubation causes similar changes in specific genes, even in distinct hematopoietic cell types (Fig. 2*C*). The majority of these incubation-affected genes were differentially expressed in both the AML and CLL data, suggesting that artifacts of sample incubation are particularly evident in pan-leukemia analyses.

We next performed an unsupervised principal components analysis of these AML and CLL data with our PBMC time series and four normal bone marrow samples (Fig. 2*D*). The bone marrow samples were commercially purchased as cryopreserved mononuclear cells in a previous study (28) and subject to an incubation period of unknown length. The first principal component separated the four datasets, as expected. The second principal component corresponded to the axis of time in our

PBMC time series. Three of the four normal bone marrow samples were placed with our 0h PBMC samples along the second principal component, potentially due to a relatively short incubation period. In contrast, most of AML or CLL samples were placed above our 8h time point, suggesting that they exhibit gene expression signatures of sample incubation.

In addition to highlighting likely temporal differences between cases and controls, our principal components analysis illustrates how intradataset variability due to sample incubation can be difficult to separate from true biological differences. The second principal component of our unsupervised analysis, which corresponds to incubation time of our PBMC samples, separated two recently described molecular subdivisions of CLL (31) (Fig. 2*D*). Furthermore, the published genes that distinguish these two CLL subdivisions correlated with incubation time in our PBMC time series (Fig. 2*E*). In contrast, the majority of published biomarkers from seven studies of AML subgroups and prognostic signatures were relatively stable over time, but a substantial minority changed significantly during incubation (Fig. S2). Excluding incubation-responsive genes from subgroup or prognostic analyses may improve the accuracy of leukemic gene signatures.

Incubation Induces Novel Coding Isoform Expression. Widespread and biased changes in alternative splicing occurred rapidly during sample incubation, such as preferential skipping of cassette exons and an increased number of unspliced introns (Fig. 1 *D* and *F*). These changes were not limited to annotated alternative splicing events. We anecdotally noticed that many time-dependent changes in splicing gave rise to isoforms encoding unspliced or abnormally spliced isoforms of coding genes (Fig. S3*A*). We tested whether this effect occurred genome-wide by searching for abnormal alternative splicing of ~160,000 splice junctions annotated as constitutive in the UCSC Genome Browser (32). We found that ~5% of normally constitutively spliced junctions produced increased levels of unspliced or misspliced products after only 4 h of incubation, rising to >20% after 48 h (Fig. 3*A*).

We next tested whether such abnormal isoforms appear in public leukemic datasets, potentially reflecting artifacts caused

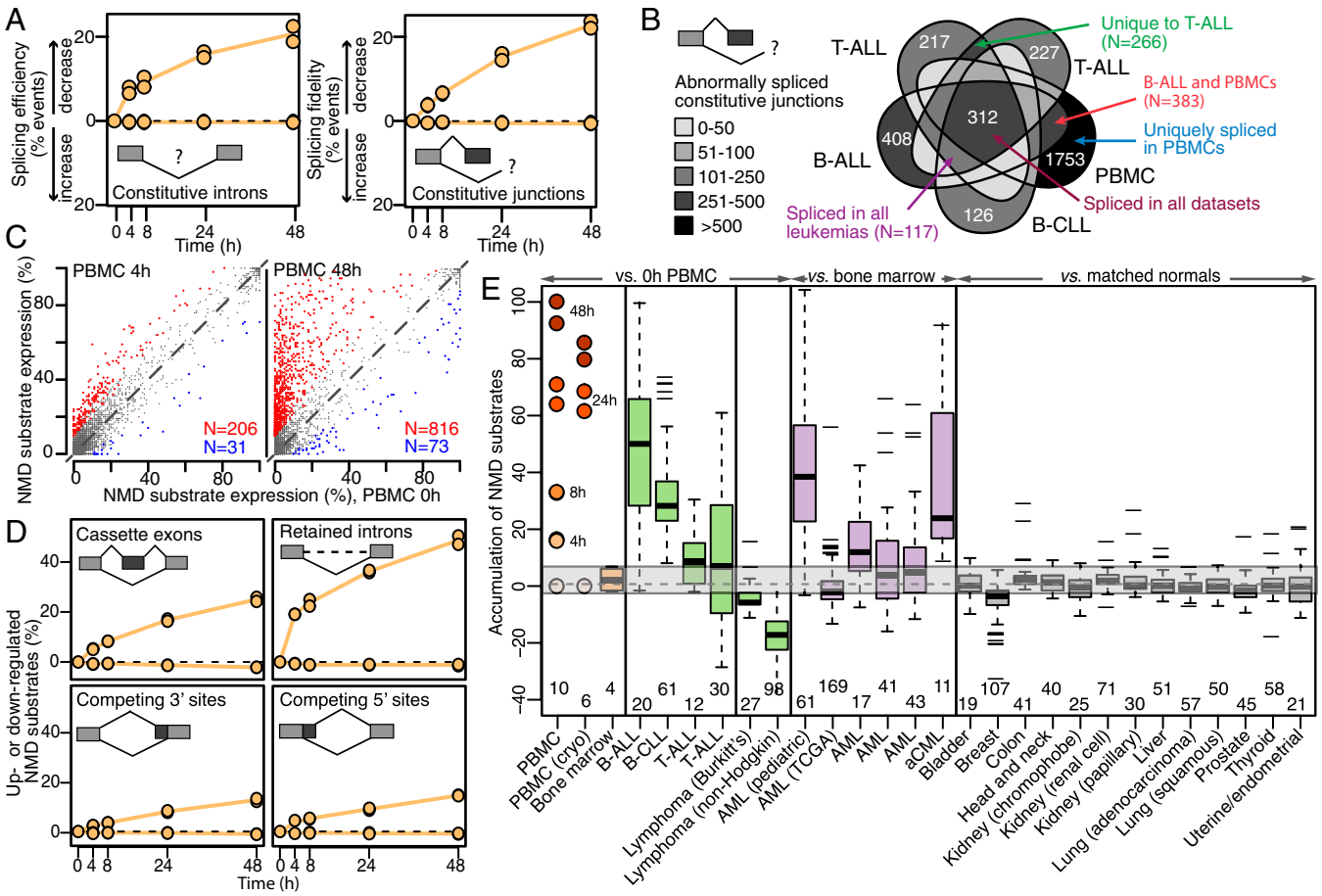

Fig. 3. Sample incubation dysregulates RNA processing and surveillance. (*A*) Increases (upper line) and decreases (lower line) in alternative splicing and intron retention of annotated constitutive splice junctions. (*B*) Overlap between alternatively spliced constitutive junctions in PBMCs after 48 h of incubation and lymphoid leukemias, both with respect to 0h PBMCs (20–23). Color hue indicates number of putative novel isoforms. For leukemia datasets, only events that are alternatively spliced in >25% of samples are included. (*C*) Scatter plot of NMD substrates resulting from alternative splicing of cassette exons. Each dot represents the isoform ratio of a predicted NMD substrate created by inclusion or exclusion of a cassette exon. Red/blue, differentially spliced isoforms exhibiting changes in isoform ratio ≥10% relative to 0 h. (*D*) Increases (upper line) and decreases (lower line) in isoform ratios of isoforms that are predicted NMD substrates, subdivided by the type of splicing event. Numbers are normalized to the detected number of alternatively spliced events of each type that introduce or remove premature termination codons. (*E*) Accumulation of NMD substrates resulting from alternative splicing of cassette exons in normal PBMCs or bone marrow (orange), lymphoid leukemias (green), lymphomas (green), myeloid leukemias (purple), and solid tumors (gray). Dataset order and references are as in Fig. 1*G*. Numbers indicate tumor samples within each dataset. NMD accumulation is $\log_2$ of the ratio of increased vs. decreased NMD substrates, multiplied by the total number of alternatively spliced NMD substrates and normalized to the number of detected events. Positive ratios correspond to a decrease in NMD efficiency. Shaded box, changes in normal bone marrow samples.

by sample incubation. We compared four distinct acute lymphocytic leukemia and CLL datasets to our 0h PBMCs and identified candidate cancer-specific isoforms for each dataset. A total of >400 such candidate cancer-specific isoforms were shared between all four leukemic datasets. Of those, >300 were likewise up-regulated in the 24h PBMC sample, whereas only 117 were present exclusively in the leukemic datasets (Fig. 3*B*).

RNA Surveillance Is Inhibited in Leukemic Transcriptomes. Why do novel isoforms rapidly appear when whole blood is incubated? The vast majority of unspliced or misspliced products arising from constitutive junctions contain premature termination codons. Many such abnormal isoforms are produced at low basal levels in cells due to the stochastic nature of splicing, but are normally degraded by nonsense-mediated decay (NMD), which degrades RNAs containing premature termination codons during the pioneer round of translation (Fig. S3*B*).

To test whether the appearance of novel isoforms is likely due to reduced NMD efficiency, we quantified global levels of predicted NMD substrates generated by cassette exon splicing. We observed a 6.6-fold enrichment for up- vs. down-regulation of NMD substrates after 4 h of sample incubation, rising to 11.2-fold after 48 h (Fig. 3*C*). NMD substrates generated by every class of alternative splicing were similarly up-regulated, including 49% of all NMD-inducing intron retention events (Fig. 3*D*).

To determine whether this NMD inhibition likely affects leukemia genomics studies, we measured NMD substrate levels within four lymphoid (20–23) and six myeloid (22, 24–26) leukemia datasets. Our 0h PBMCs and normal bone marrow exhibited similar levels of NMD substrates. In contrast, almost every leukemia dataset exhibited increased levels of NMD substrates relative to control PBMCs or bone marrow (Fig. 3*E*). Levels of NMD substrates were independent of clinical variables such as the origin of the leukemic cells (bone marrow or peripheral blood), percentage of blasts, and disease state at the time of collection (diagnosis or relapse) (Fig. S3*C*). The TCGA AML dataset was a notable exception, probably because samples were rapidly collected and processed at the same institution, and therefore fewer were subject to artifacts introduced by sample incubation (24). Neither lymphomas (29, 30) nor solid tumors (TCGA) exhibited globally higher levels of NMD substrates relative to normal controls (PBMCs or patient-matched uninvolved tissue).

Splicing Biomarkers Detect and Quantify Incubation. Incubation time is typically not available in the sample metadata associated with published leukemia genomics studies and may not be recorded in the clinical annotation stored by biorepositories. Therefore, we sought to find biomarkers of ex vivo incubation that could be used to quantify the effects of incubation in incompletely annotated blood samples. We identified 27 alternative splicing events that changed monotonically and dramatically during our PBMC time course, such as cassette exons within *PHF20* and *LEF1* (Fig. 1*F*, Fig. S4*A*, and Dataset S3). We used splicing events as biomarkers because alternative splicing measurements are insensitive to total gene expression levels, which are frequently cell type or dataset-specific. For example, *LEF1* gene expression depends upon the lymphocyte fraction of a blood draw, but the inclusion of a specific *LEF1* exon is likely more robust to these differences. We restricted to splicing events that were not predicted to trigger NMD because NMD substrates are typically more difficult to quantify due to their low abundance.

To test the utility of these 27 splicing events as incubation biomarkers, we performed unsupervised clustering of our PBMC time series with 123 lymphoid and 342 myeloid leukemia samples. This clustering correctly ordered the PBMC samples by incubation time and, furthermore, ordered the leukemic samples by the genome-wide level of NMD substrates in each sample (Fig. 4*A*). As we selected our splicing biomarkers to be insensitive to

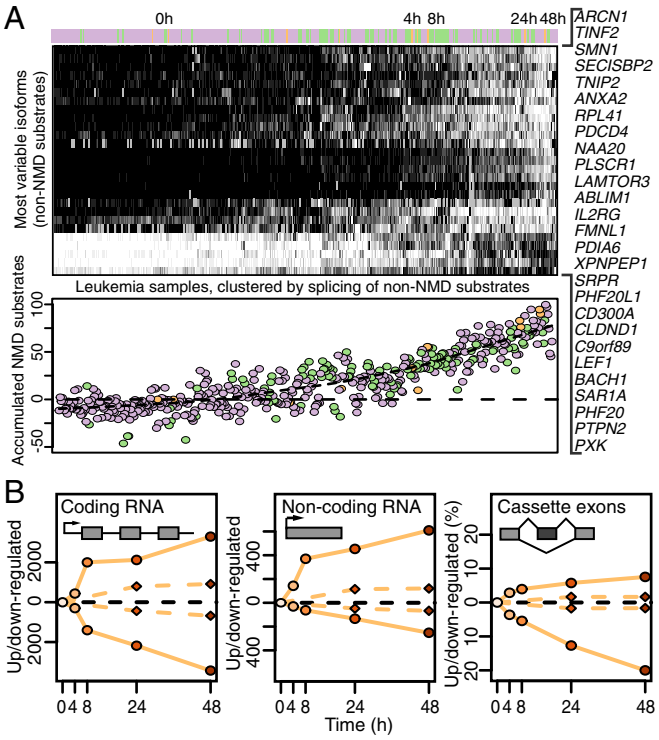

Fig. 4. Sample incubation can be detected with biomarkers and ameliorated by ice. (*A*, *Upper*) Unsupervised clustering of normal PBMCs and bone marrow (orange), lymphoid leukemias (green), and myeloid leukemias (purple), based on a panel of 27 cassette exons with the largest splicing changes after 24 and 48 h (Dataset S3). Shading indicates exon inclusion (white, 0%; black, 100%). (*Lower*) Log_2 accumulation of NMD substrates for the samples indicated in *Upper*. (*B*) Numbers of differentially expressed coding (*Left*) and noncoding (*Center*) genes and percent differentially spliced cassette exons (*Right*). Solid lines, whole blood incubated at room temperature; dashed lines, incubation on ice. Lines are averaged across donors 3 and 4.

NMD, these data suggest that RNA splicing and surveillance change concordantly during incubation. We conclude that a relatively small set of splicing biomarkers is useful for quantifying the effects of sample incubation.

Ice Ameliorates Incubation-Induced Dysregulation. The pleiotropic biases that we describe could be reduced in future studies by rapidly processing samples. However, TCGA's single-center approach for AML is not feasible for many studies. Similarly, commercially available PAXgene collection tubes or similar products rapidly stabilize RNA (9, 15, 33), but also prevent isolation of specific cell populations with a Ficoll gradient or flow sorting. Therefore, we tested the simple expedient of placing whole blood on ice immediately after collection by a phlebotomist. Incubating samples on ice dramatically reduced the rate of time-dependent changes in the transcriptome. The magnitude of differential gene expression and alternative splicing after 48 h on ice was approximately equivalent to changes observed after storage for 4 h at room temperature (Fig. 4*B*).

Discussion

The most surprising aspect of our study is not that changes occur during ex vivo incubation, but rather that these changes affect so many facets of the gene expression process. Pseudogenes, antisense RNAs, novel coding isoforms, and RNA surveillance are of current interest in cancer biology, and all are dysregulated by sample incubation. The molecular origins of incubation-induced

changes in the transcriptome are unclear, but are probably not due to RNA degradation alone. RIN measurements do not change substantially during our time course (Fig. 1*C*), potentially because our samples are stored as whole blood (34). Furthermore, many previously low-abundance isoforms become highly expressed, consistent with de novo transcription (Fig. 3). We used the IPA software to identify potential upstream regulators of genes exhibiting differential expression after 24–48 h of ex vivo incubation. The identified regulatory factors included cytokines, growth factors, and transcriptional regulators, and select downstream transcriptional targets were down- or up-regulated as incubation proceeded (Fig. S4 *B* and *C*). Many gene expression changes may also be byproducts of the widespread inhibition of NMD, which normally degrades many coding RNAs, as well as abnormal RNAs such as transcribed pseudogenes and antisense RNAs. NMD inhibition may in turn be due to translation inhibition or incubation-associated cellular stresses such as hypoxia (35).

In addition to obscuring cancer-specific alterations, sample incubation can confound subgroup analyses or comparisons of different diseases. Sample incubation time may correlate with clinical parameters, such as time to treatment or diagnosis vs. relapse, or systematically differ between studies. For example, studies of rare diseases such as pediatric cancers frequently rely upon a worldwide network of clinics, whereas studies of common disorders may be highly centralized. Our findings may additionally have implications for studies of other diseases that rely upon blood collection, such as infectious or autoimmune diseases.

Because sample incubation alters biologically relevant processes, how can we identify true cancer-specific alterations? Standardizing collection procedures may prove helpful for future studies, but will not aid studies that rely upon existing biorepositories. One possibility is to statistically detect and correct for incubation-induced artifacts, potentially by using our proposed panel of alternatively spliced exons (Fig. 4*B*). Another productive path forward might be to increase the relative resources devoted to characterizing "normal" control tissues. Most large-scale leukemia genomics studies characterize tens or hundreds of cases, but only a few controls. The scarcity of control samples is justified by the assumption that variation between individual tumors is large, whereas variation between normal tissue samples is comparatively small. However, the increasing appreciation of transcriptional variation in healthy tissues, as well as our study's demonstration that perceived intertumor variability can be augmented by disease-irrelevant technical differences, suggest that further characterizing control cells from diverse sources will be productive.

Materials and Methods

Whole blood was collected and PBMCs were isolated using a Ficoll gradient at specified time points. RNA-sequencing (RNA-seq) libraries were generated by using the Illumina TruSeq kit, with modifications. Reads were mapped to all genes and splice junctions from UCSC knownGene and the Ensembl 71 gene annotation. See *SI Materials and Methods* for detailed information about data generation and processing. The RNA-seq data have been submitted to the NCBI Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo/) under accession nos. GSE58335 and GSE61410.

ACKNOWLEDGMENTS. We thank Martin McIntosh for providing RNA-seq data from normal bone marrow, Era L. Pogosova-Agadjanyan and Sommer Castro for assistance with collection of peripheral blood, and Martin McIntosh for comments on the manuscript. The results published here are in part based upon data generated by The Cancer Genome Atlas Research Network (cancergenome.nih.gov). This work was supported by Damon Runyon Cancer Research Foundation Grant DFS 04-12 (to R.K.B.), Ellison Medical Foundation Grant AG-NS-1030-13 (to R.K.B.), NIH/National Cancer Institute (NCI) Grant P30 CA015704 recruitment support (to R.K.B.), Fred Hutchinson Cancer Research Center institutional funds (R.K.B.), and NIH/NCI Training Grant T32 CA009657 (to J.O.I.).

1. Leek JT, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11(10):733–739.
2. Chen C, et al. (2011) Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE* 6(2):e17238.
3. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127.
4. Scharpf RB, et al. (2011) A multilevel model to address batch effects in copy number estimation using SNP arrays. *Biostatistics* 12(1):33–50.
5. Li S, et al. (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol* 32(9):888–895.
6. Bowen RA, Remaley AT (2014) Interferences from blood collection tube components on clinical chemistry assays. *Biochem Med (Zagreb)* 24(1):31–44.
7. Tammen H (2008) Specimen collection and handling: Standardization of blood sample collection. *Methods Mol Biol* 428:35–42.
8. Gillio-Meina C, Cepinskas G, Cecchini EL, Fraser DD (2013) Translational research in pediatrics II: Blood collection, processing, shipping, and storage. *Pediatrics* 131(4):754–766.
9. Rainen L, et al. (2002) Stabilization of mRNA expression in whole blood samples. *Clin Chem* 48(11):1883–1890.
10. Liu Y, Malaviarachchi P, Beggs M, Emanuel PD (2010) PTEN transcript variants caused by illegitimate splicing in "aged" blood samples and EBV-transformed cell lines. *Hum Genet* 128(6):609–614.
11. Thomson SA, Wallace MR (2002) RT-PCR splicing analysis of the NF1 open reading frame. *Hum Genet* 110(5):495–502.
12. Birrell GW, Ramsay JR, Tung JJ, Lavin MF (2001) Exon skipping in the ATM gene in normal individuals: The effect of blood sample storage on RT-PCR analysis. *Hum Mutat* 17(1):75–76.
13. Barnes MG, Grom AA, Griffin TA, Colbert RA, Thompson SD (2010) Gene expression profiles from peripheral blood mononuclear cells are sensitive to short processing delays. *Biopreservation Biobanking* 8(3):153–162.
14. Salway F, Day PJ, Ollier WE, Peakman TC (2008) Levels of 5′ RNA tags in plasma and buffy coat from EDTA blood increase with time. *Int J Epidemiol* 37(suppl 1):i11–i15.
15. Baechler EC, et al. (2004) Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation. *Genes Immun* 5(5):347–353.
16. Adamia S, et al. (2014) NOTCH2 and FLT3 gene mis-splicings are common events in patients with acute myeloid leukemia (AML): New potential targets in AML. *Blood* 123(18):2816–2825.
17. Kühnl A, et al. (2011) Overexpression of LEF1 predicts unfavorable outcome in adult patients with B-precursor acute lymphoblastic leukemia. *Blood* 118(24):6362–6367.
18. Metzeler KH, et al. (2012) High expression of lymphoid enhancer-binding factor-1 (LEF1) is a novel favorable prognostic factor in cytogenetically normal acute myeloid leukemia. *Blood* 120(10):2118–2126.
19. Zhang T, et al. (2013) PHF20 regulates NF-κB signalling by disrupting recruitment of PP2A to p65. *Nat Commun* 4:2062.
20. Meyer JA, et al. (2013) Relapse-specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. *Nat Genet* 45(3):290–294.
21. Quesada V, et al. (2012) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* 44(1):47–52.
22. Macrae T, et al. (2013) RNA-Seq reveals spliceosome and proteasome genes as most consistent transcripts in human cancer cells. *PLoS ONE* 8(9):e72884.
23. Atak ZK, et al. (2013) Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. *PLoS Genet* 9(12):e1003997.
24. Cancer Genome Atlas Research Network (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 368(22):2059–2074.
25. McNerney ME, et al. (2013) CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. *Blood* 121(6):975–983.
26. Wen H, et al. (2012) New fusion transcripts identified in normal karyotype acute myeloid leukemia. *PLoS ONE* 7(12):e51203.
27. Piazza R, et al. (2013) Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nat Genet* 45(1):18–24.
28. Stirewalt DL, et al. (2008) Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer* 47(1):8–20.
29. Schmitz R, et al. (2012) Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature* 490(7418):116–120.
30. Morin RD, et al. (2011) Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* 476(7360):298–303.
31. Ferreira PG, et al. (2014) Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res* 24(2):212–226.
32. Meyer LR, et al. (2013) The UCSC Genome Browser database: Extensions and updates 2013. *Nucleic Acids Res* 41(database issue):D64–D69.
33. Müller MC, et al. (2002) Improvement of molecular monitoring of residual disease in leukemias by bedside RNA stabilization. *Leukemia* 16(12):2395–2399.
34. Gallego Romero I, Pai AA, Tung J, Gilad Y (2014) RNA-seq: Impact of RNA degradation on transcript quantification. *BMC Biol* 12(1):42.
35. Gardner LB (2010) Nonsense-mediated RNA decay regulation by cellular stress: Implications for tumorigenesis. *Mol Cancer Res* 8(3):295–308.