frontiers in Genetics

ORIGINAL RESEARCH
published: 19 July 2019
doi: 10.3389/fgene.2019.00643

Depletion of Hemoglobin Transcripts and Long-Read Sequencing Improves the Transcriptome Annotation of the Polar Bear (*Ursus maritimus*)

*Ashley Byrne[1,2], Megan A. Supple[3], Roger Volden[2,4], Kristin L. Laidre[5], Beth Shapiro[3,6] and Christopher Vollmers[2,4]**

[1] Department of Molecular, Cellular, and Developmental Biology, University of California, Santa Cruz, CA, United States, [2] Genomics Institute, University of California, Santa Cruz, CA, United States, [3] Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA, United States, [4] Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, United States, [5] Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle, WA, United States, [6] Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA, United States

OPEN ACCESS

Edited by:
Ishaan Gupta,
Cornell University, United States

Reviewed by:
Rui Chen,
Baylor College of Medicine, United States
Wei Xu,
Texas A&M University United States

***Correspondence:**
Christopher Vollmers
vollmers@ucsc.edu

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *22 January 2019*
Accepted: *18 June 2019*
Published: *19 July 2019*

Citation:
Byrne A, Supple MA, Volden R, Laidre KL, Shapiro B and Vollmers C (2019) Depletion of Hemoglobin Transcripts and Long-Read Sequencing Improves the Transcriptome Annotation of the Polar Bear (Ursus maritimus). Front. Genet. 10:643. doi: 10.3389/fgene.2019.00643

Transcriptome studies evaluating whole blood and tissues are often confounded by overrepresentation of highly abundant transcripts. These abundant transcripts are problematic, as they compete with and prevent the detection of rare RNA transcripts, obscuring their biological importance. This issue is more pronounced when using long-read sequencing technologies for isoform-level transcriptome analysis, as they have relatively lower throughput compared to short-read sequencers. As a result, long-read based transcriptome analysis is prohibitively expensive for non-model organisms. While there are off-the-shelf kits available for select model organisms capable of depleting highly abundant transcripts for alpha (HBA) and beta (HBB) hemoglobin, they are unsuitable for non-model organisms. To address this, we have adapted the recent CRISPR/Cas9-based depletion method (depletion of abundant sequences by hybridization) for long-read full-length cDNA sequencing approaches that we call Long-DASH. Using a recombinant Cas9 protein with appropriate guide RNAs, full-length hemoglobin transcripts can be depleted *in vitro* prior to performing any short- and long-read sequencing library preparations. Using this method, we sequenced depleted full-length cDNA in parallel using both our Oxford Nanopore Technology (ONT) based R2C2 long-read approach, as well as the Illumina short-read based Smart-seq2 approach. To showcase this, we have applied our methods to create an isoform-level transcriptome from whole blood samples derived from three polar bears (*Ursus maritimus*). Using Long-DASH, we succeeded in depleting hemoglobin transcripts and generated deep Smart-seq2 Illumina datasets and 3.8 million R2C2 full-length cDNA consensus reads. Applying Long-DASH with our isoform identification pipeline, Mandalorion, we discovered ~6,000 high-confidence isoforms and a number of novel genes. This indicates that there is a high diversity of gene isoforms within *U. maritimus* not yet reported. This reproducible and straightforward approach

has not only improved the polar bear transcriptome annotations but will serve as the foundation for future efforts to investigate transcriptional dynamics within the 19 polar bear subpopulations around the Arctic.

Keywords: polar bear (*Ursus maritimus*), R2C2, long-read high throughput sequencing, ONT, Oxford Nanopore Technologies, transcriptome annotation

INTRODUCTION

Accurate isoform-level differential expression analysis of transcriptomes is essential for interpreting gene regulation under different biological, environmental, or physiological conditions. RNA transcript isoforms—which are often unique among different cell types, tissues, developmental stages, and organisms (Kalsotra et al., 2008; Wang et al., 2008; Zhang et al., 2016)—are defined by the use of alternative transcription start sites (TSSs), polyA sites, and splice sites. Use of alternative isoforms is highly regulated and thought to contribute to cellular and organismal diversification within higher eukaryotes (Graveley, 2001), adaptation, and speciation (Harr and Turner, 2010; Shi et al., 2012) and can also reflect certain disease states (Busslinger et al., 1981; Andreadis, 2005; Ilagan et al., 2015).

To perform this type of differential expression analysis on the isoform level requires both short- and long-read sequencing technology. Short-read RNA-seq provides the read depth necessary for gene expression quantification but requires accurate and exhaustive isoform-level transcriptome annotations for its analysis. However, existing transcriptome annotations of non-model organisms are often incomplete or inaccurate (Ungaro et al., 2017) because they cannot rely on labor-intensive efforts like GENCODE, which are working to exhaustively annotate the isoform-level transcriptomes of human and mouse. While short-read RNA-seq data can itself be used for transcriptome annotation, it fails at annotating transcriptomes on the isoform-level because it cannot recapitulate full-length transcripts. This inability to define full-length transcripts is due to the fragmentation of RNA, or their cDNA copies, prior to sequencing, making it difficult to computationally re-assemble reliably (Grabherr et al., 2011; Bankevich et al., 2012; Pertea et al., 2015). To provide an accurate isoform-level transcriptome annotation for non-model organisms, long-read sequencing technology is required to sequence full-length cDNA molecules.

The ability to perform combined short- and long-read transcriptome analysis on non-model organisms is further complicated by sample availability. In contrast to the organs and tissues of model organisms which can be easily acquired, availability of samples from non-model organisms is often more limited. In rare circumstances, sampling can be performed through fat- and muscle-tissue biopsies (Khudyakov et al., 2017), but the current gold standard still relies on whole blood RNA samples, especially for large non-model organisms (Du et al., 2015). This is particularly true for protected and endangered species (Huang et al., 2016; Hernández-Fernández et al., 2017). While whole blood samples can be easily acquired and provide a wealth of information regarding physiological or disease states in surrounding tissues (Liew et al., 2006), polyadenylated RNA extracted from whole blood can be comprised of >50% hemoglobin transcripts (Mastrokolias et al., 2012; Shin et al., 2014). In any high-throughput sequencing-based assay, these highly abundant transcripts will compete for a limited number of sequencing reads and, as a result, will be sequenced over and over again without generating any new information. This would waste valuable reads which could otherwise detect less abundant transcripts.

Currently, there is no approach to deplete hemoglobin transcripts from whole blood RNA while enabling downstream analysis of the depleted RNA/cDNA with both short- and long-read sequencing. Commercially available hemoglobin depletion kits—including GLOBINclear (Ambion) or Ribo-Zero (Illumina)—are specifically designed for human samples and rely on hemoglobin RNA pull-down methods (Field et al., 2007). Even if they would succeed in depleting hemoglobin from non-model organism samples, which is far from guaranteed (Choi et al., 2014), these pull-down approaches use harsh conditions and high temperatures during long incubation steps which contribute to RNA fragmentation and introduce unwanted technical variability (Debey et al., 2004). While fragmented RNA is suitable as input into short-read RNA-seq, it is not suitable for long-read full-length cDNA sequencing.

To perform a comprehensive analysis of non-model organism transcriptomes from whole-blood with short- and long-read technologies, we require a new approach that can deplete highly abundant transcripts like hemoglobin from whole-blood samples of a wide range of organisms without fragmenting transcripts. To this end, we chose to adapt the powerful, recently published DASH (depletion of abundant sequences by hybridization) (Gu et al., 2016) method which utilizes a recombinant Cas9 to perform *in vitro* depletion using sequence-specific sgRNAs. Our adapted method which we will refer to as Long-DASH also takes advantage of the CRISPR/Cas9 system to selectively deplete hemoglobin HBA and HBB transcripts but targets full-length cDNA instead of fragmented short-read Illumina sequencing libraries like regular DASH. By depleting full-length cDNA prior to any library preparation, this allows the user the choice to use any short- or long-read sequencing platform.

As a proof of concept, we evaluated three hemoglobin-depleted and non-depleted polar bear whole blood transcriptomes using our ONT-based R2C2 (Volden et al., 2018) full-length cDNA sequencing method and an Illumina-based modified Smart-seq2 method. We found that by incorporating Long-DASH, we successfully depleted hemoglobin transcripts without non-specifically affecting the rest of the cDNA pool. Finally, we generated ~3.8 million ONT-based R2C2 consensus reads, dramatically refining the polar bear transcriptome annotations.

MATERIALS AND METHODS

Sample Collection/RNA Extraction From Whole Blood

Permits for field operations and animal care were provided by the Government of Greenland (permit numbers 2015-110281 and 2017-5446). Polar bear whole blood samples were collected in PAXgene Blood RNA Tubes (PreAnalytiX GmbH, BD Biosciences, Mississauga, ON, Canada). Total RNA was isolated from whole blood (2.5 ml) thawed at room temperature for 2 h prior to using the PAXGene RNA Extraction Kit (Qiagen, Chatsworth, CA, USA) according to manufacturer's protocol. All samples were DNAse (QIAGEN) treated and eluted in 50 µl. The RNA yield and purity were accessed using a NanoDrop 8000 UV Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). RNA quantities ranged from 110 to 310 ng/µl, and the A260/280 ratio values were ≥ 2.0. Human whole blood RNA was purchased from Zyagen Labs (NC1453913).

Full-Length cDNA Generation

RNA was reverse transcribed (RT) using SMARTScribe Reverse Transcriptase (Clonetech). We generated full-length cDNA using a modified Smart-seq2 approach (Cole et al., 2018). During the RT reaction, a template-switch oligo and an oligodT primer were used to select for polyA+ RNA (**Table S2**). The RT reaction was performed in 10 µl reactions with an input of 70 ng of RNA and took place at 42°C for 1 h. After cDNA synthesis, 1 µl of 1:10 dilutions of RNAse A (Thermo Fisher) and Lambda Exonuclease (NEB) were added and incubated at 37°C for 30 min. Following the incubation, an amplification step was performed in 25-µl final volumes using KAPA HiFi ReadyMix 2X (Kapa Biosystems) containing 1 µl of the ISPCR primer (10 µM) primer. Samples were incubated at 95°C for 3 min, followed by 12 cycles of (98°C for 20 s, 67°C for 15 s, and 72°C for 4 min), with a final extension of 72°C for 5 min. Samples were purified using Agencourt AMPure XP SPRI beads (Beckman Coulter) and eluted at 25 ml. The final cDNA product was then visualized on an agarose gel to confirm distribution (**Figure 2**).

In Vitro Preparation of CRISPR/Cas9

SpCas9-2xNLS was purified based on the protocol described in (Jinek et al., 2012). Briefly, a plasmid encoding His_6MBP-SpCas9-2xNLS (Addgene plasmid #69090) was transformed into Rosetta2(DE3) *Escherichia coli* cells. Cultures were grown at 37°C in 2YT medium with shaking until they reached an OD_{600} of ~0.6, and then placed on ice for 5 min before adding IPTG to a final concentration of 0.25 mM; cultures were then grown overnight at 18°C with shaking. Cell pellets were harvested by centrifugation, and then lysed in an AVESTIN cell extruder in Ni-A buffer (20 mM Tris pH 8.0, 500 mM NaCl, 5% vol/vol glycerol, 25 mM imidazole) with EDTA-free protease inhibitors (Pierce). Clarified supernatants were purified by gravity column on Ni-NTA agarose (QIAGEN) using Ni-A buffer to load and wash, and Ni-B buffer (20 mM Tris pH 8.0, 500 mM NaCl, 5% vol/vol glycerol, 250 mM imidazole) to elute. Peak fractions were concentrated in an Amicon Ultra spin concentrator with a 30-kDa molecular weight cut-off at 4°C, and then loaded onto a 50-ml HiPrep Desalting Column (GE Healthcare) pre-equilibrated with 17% IEX-B (IEX-A buffer: 20 mM HEPES pH 7.5, 150 mM KCl, 5% vol/vol glycerol; IEX-B 20 mM HEPES pH 7.5, 1 M KCl, 5% vol/vol glycerol). The flow through was then loaded onto a 2-ml HiTrap SP column (GE Healthcare) in 17% IEX-B buffer. After thoroughly washing the column in 17% IEX-B, the protein was eluted with a linear gradient from 17% to 50% IEX-B. Peak fractions were pooled and loaded onto a Superdex 200 16/60 column (GE Healthcare) pre-equilibrated in 20 mM HEPES pH 7.5, 150 mM KCl, 1 mM DTT, and 10% vol/vol glycerol. Peak fractions were concentrated in an Amicon Ultra spin concentrator with a 30-kDa molecular weight cut-off at 4°C until a concentration of 40 µM, which was estimated using the calculated molar extinction coefficient of 120,575 M^{-1} cm^{-1}. The protein was aliquoted into small volumes (10 µL), quick frozen in liquid nitrogen, and stored at −80°C.

sgRNA Design and Construction

Other studies have shown that sgRNAs designed between 17 and 20 bp showed increased efficacy (Fu et al., 2014; Ren et al., 2014). As a result, the sgRNAs were designed between 17 and 20 bp in length. sgRNAs were designed to target hemoglobin transcripts in human and polar bear. A multi-sequence alignment was performed on the human and polar bear annotated HBA and HBB gene transcripts to find conserved regions using the Clustal Omega tool (Sievers et al., 2011; McWilliam et al., 2013; Li et al., 2015) (**Figure S3**). Regions with high homology were chosen for sgRNA design. sgRNAs that did not share complete homology were designed to contain degenerate bases to ensure compatibility across species using the same sgRNA (**Figure S4**). sgRNA specificity was determined by using BLAST (Altschul et al., 1990). One sgRNA was designed even though the N-GG (PAM motif) had been lost in the human but was still kept in the pool for the polar bear depletion (**Figure S3**). A total of 16 sgRNAs were designed to target HBA and HBB hemoglobin transcripts. The target oligos were then constructed into sgRNAs as previously described (Ren et al., 2014). Single-stranded oligos were designed to contain a T7 promoter attached to each sgRNA sequence (IDT) followed by the first 22 bases of the tracrRNA sequence (**Figure S3**). The complementary tracrRNA and single-stranded oligo were annealed and extended to form a dsDNA product containing the T7-sgRNA and tracrRNA templates. The template was then used for *in vitro* transcription using the HiScribe T7 High Yield RNA Synthesis Kit (NEB). The *in vitro* transcription reaction was carried out at 37°C for 16 h. The *in vitro* transcribed RNA was then purified using MEGAclear Transcription Clean-Up Kit (Invitrogen). The final sgRNA product was then checked for purity and quantified using NanoDrop 8000 UV Spectrophotometer (Thermo Fisher). All sgRNAs were then pooled at equal molar concentrations and stored in single-use aliquots at −80°C.

CRISPR/Cas9 Treatment

Since it has been predicted that human whole blood samples can contain up to ~50–80% of hemoglobin transcripts of the

total sample (Field et al., 2007; Mastrokolias et al., 2012), we calculated the ratio of sgRNA and Cas9 molar amounts to sample based upon this assumption. According to the DASH protocol, it was determined that 150-fold of Cas9 and 1,500-fold of sgRNA should be sufficient (Gu et al., 2016). All cDNA samples were quantified by qubit using the dsDNA HS Assay Kit (Thermo Fisher) to calculate the molar amounts. To calculate the expected molar amounts we use the following formula:

$$nM = [\mathrm{DNA}(ng/\mu l)] \div (660\,\mathrm{g}/mol \times size\ of\ hemoglobin\ transcript\ in\ bp)$$

Once the molar amounts were determined, the ribonucleoprotein (RNP) complex was formed by adding the 150-fold Cas9 and 1,500-fold sgRNA excess amount with 1.0 µl of the 10X Cas9 Buffer (final concentration 50 mM Tris pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$, and 1 mM TCEP) and incubated for 25°C for 10 min. Following the 25°C incubation, the calculated cDNA amount was then added (final volume of 10 µl) and incubated at 37°C for 4 h to overnight. After the Cas9 depletion, 1 µl of proteinase K and RNAse A were added to inactivate the Cas9 and remove excess sgRNAs from the reaction and incubated at 37°C for 15 min and 95°C for 15 min. It is critical that the proteinase K is deactivated properly as the samples are immediately used for amplification. Treated samples were PCR amplified [95°C for 3 min, followed by 13 cycles of (98°C for 20 s, 67°C for 15 s, and 72°C for 4 min) followed by a final extension of 72°C for 5 min]. PCR was performed using KAPA HiFi ReadyMix 2X (Kapa Biosystems) and 1 µl of the (10 µM) ISPCR primer. The amplified product was then purified using SPRI beads to remove everything below 500 bp. Selecting against cDNA below 500 bp ensured that all cut hemoglobin products were removed before making the Tn5 libraries. The depleted cDNA product was visualized on a 1–2% agarose gel to confirm depletion. Once confirmed, the depleted cDNA product was then prepped for either Illumina or Nanopore sequencing.

R2C2 Library Preparation and ONT Sequencing

To prepare R2C2 libraries, ~30 ng of the depleted cDNA was used. The R2C2 libraries were made as previously described (Volden et al., 2018). Briefly, an equal concentration of splint to cDNA were combined [30 ng of depleted cDNA and 30 ng of our (~200 bp) DNA splint]. The full-length cDNA was then circularized using the 2X NEBuilder HiFi DNA Assembly Mix (NEB). The reaction took place at 50°C for 1 h per manufacturer protocol. Once the full-length cDNA was circularized, linear ssDNA and dsDNA were digested by adding 3 µl each of Lambda Exonuclease, Exonuclease I, and Exonuclease III (all NEB) and incubated at 37°C overnight. We performed the longer incubation overnight to ensure complete digestion. After digestion, the sample was further purified using SPRI beads and eluted in 30 µl of ultrapure water. Thirty microliters of sample was then split into three reactions containing 10 µl each for the Phi29 amplification. The Phi29 amplification took place in a reaction volume of 50 µl containing 5 µl of 10X Buffer, 2.5 µl of 10 µM each dNTPs, 2.5 µl of random hexamers (10 µM), 29 µl of ultrapure water, and 1 µl of Phi29 polymerase. The Phi29 reactions were incubated at 30°C for 16 h and 65°C for 15 min and held at 4°C. All three samples were pooled together, and ultrapure water was added to make up the final volume to 300 µl. The product was purified using SPRI beads with a 1:0.5 sample-to-bead ratio. This ratio was chosen as it removed all fragments <2,000 kb. The sample was then eluted in 90 µl of ultrapure H20, 10 µl of NEB2 Buffer (NEB), and 3 µl of T7 endonuclease (NEB) and incubated at 37°C for 2 h to ensure complete debranching of the Phi29 product. The eluted sample was again purified using SPRI beads with a 1:0.5 sample-to-bead ratio. The product was eluted in 30 µl and quantified using Qubit dsDNA HS Kit (Thermo Fisher). The length distribution was verified on a 1% agarose gel prior to performing the ONT library prep.

For the library preparation, ~1–2 µg of the final R2C2 product was converted into a ONT compatible library using the SQK-LSK109 kit according to ONT instructions with minor modifications. First, during the end repair and A-tailing reaction, we performed incubations for 30 min each at 20°C and 65°C instead of 5 min each. Second, we adjusted the ligation reaction time to 30 min at room temperature instead of 10 min per the protocol. We also found that loading between ~200 and 500 ng of the final library onto the flowcell was the most optimal. Loading more library resulted in severe loss in throughput as can be seen for the R2C2 runs PB3_depleted_R1 and PB19_depleted_R1 (**Table S2**). R2C2 libraries were sequenced on a MinION device using the 48-h sequencing protocol using the FLO-Min106 R9.4 Rev D chemistry flowcells. All reads were basecalled with Albacore v2.1.3.

Smart-seq2 Library Preparation and Illumina Sequencing

Illumina libraries of the depleted and non-depleted samples were prepared using a tagmentation-based method using our own Tn5 (Picelli et al., 2014). The Tn5 enzyme was custom loaded with Tn5ME-A/R and Tn5ME-B/R adapters (**Table S2**). The Tn5 reaction contained 5 µl of the full-length cDNA product, 1 µl of the loaded Tn5 enzyme, 10 µl of ultrapure water, and 4 µl of the 5X TAPS-PEG buffer and incubated at 55°C for 7 min. After incubation, 5 µl of 0.2% of sodium dodecyl sulfate (SDS) was added to the product to inactivate the Tn5 enzyme. Due to the Tn5-generating gaps, 5 µl of the Tn5 product had to be nick translated at 72°C for 5 min. The Tn5 product was then amplified using KAPA HiFi Polymerase (KAPA) with 10 cycles of PCR using (98°C for 10 s, 63°C for 30 s, 72°C for 2 min) with a final extension at 72°C for 5 min. The final reaction volume was 25 µl and contained 0.5 µl KAPA HiFi Polymerase (KAPA), 5 µl of 5X Buffer, 0.8 µl of dNTPs (10 mM each), 11.7 µl of ultrapure water, 5 µl of the nick-translated product, and 1 µl each of Nextera_Primer_A and Nextera_Primer_B primers (**Table S2**). The amplified Tn5 libraries were then size selected from 300 to 800 bp on a 2% EX E-gel (Thermo Fisher) and purified using QIAquick Gel Extraction Kit (Qiagen). The libraries were then pooled at equal concentrations and ran on a HiSeq X 2×151 bp run.

R2C2 Read Processing and Isoform Analysis

R2C2 consensus reads were generated from raw reads using the C3POa pipeline (https://github.com/rvolden/C3POa). C3POa identifies subreads in the raw reads and uses poaV2 (Lee et al., 2002) and racon (Vaser et al., 2017) to determine a more accurate consensus of these subreads. The consensus reads were then aligned to the polar bear genome (Liu et al., 2014) using minimap2 (Li, 2018) using standard setting and the *-ax splice*'flag. The resulting SAM files are converted to PSL files using SAMtools (Li et al., 2009) and jvarkit SamToPsl utility (Lindenbaum, 2015).

The resulting PSL, SAM, and FASTA files of all depleted samples were merged and used as input into the Mandalorion (https://github.com/rvolden/Mandalorion-Episode-II) pipeline to determine isoforms. To accommodate issues regarding RNA degradation and genomic DNA contamination, we integrated two new optional filter into Mandalorion. We implemented the filtering of isoforms that are entirely contained within one other isoform, which indicates degraded input RNA molecules, and the filtering of unspliced isoforms which might stem from DNA contamination.

Accuracy of R2C2 reads and Mandalorion isoforms were determined using alignments in SAM format containing MD strings and a custom script that calculates

$$Accuracy = Matches / (Matches + Mismatches + Indels)$$

Smart-seq2 Read Processing

Paired FASTQ files were downloaded from BaseSpace and aligned to the polar bear genome using STAR with standard settings. The STAR index for the polar bear genome was built without a transcriptome reference because the GFF file provided by (Liu et al., 2014) did not conform to GFF standard (no "exon" features) and could therefore not be used. Read alignments in ordered BAM format were converted to PSL as described above.

Hemoglobin and Gene Expression Quantification

Hemoglobin content was determined through a kmer-based counting method using a custom script. In short, all possible 10nt kmers were extracted from the sequence of hemoglobin HBA and HBB transcripts. The presence of these kmers were then determined in each R2C2 or Smart-seq2 read from depleted and undepleted samples. Cut-offs for read assignments to hemoglobin were then determined by also analyzing R2C2 and Illumina reads of the GM12878 cell line which does not express hemoglobin.

Gene expression was determined using Smart-seq2 (Illumina) read alignments in PSL format and a custom script. Reads aligning to hemoglobin loci were not counted toward total aligned reads in the RPM calculations.

Both script are available at https://github.com/christopher-vollmers/

Data Visualization

Schematics were prepared using Inkscape (https://inkscape.org). All others figures were prepared using python/matplotlib/numpy/scipy (Jones et al., 2001; Hunter, 2007; Millman and Aivazis, 2011; van der Walt et al., 2011).

RESULTS

Long-DASH Depletes Hemoglobin Transcripts From Full-Length cDNA

We used a modified Smart-seq2 protocol (Picelli et al., 2014; Cole et al., 2018; Volden et al., 2018) to reverse transcribe and amplify full-length cDNA from 70 ng of whole blood RNA of three polar bears (PB3, PB19, PB21). We then performed a targeted depletion of hemoglobin transcripts by incubating the full-length cDNA with Cas9 protein loaded with 16 guide RNAs (sgRNAs) specific to hemoglobin transcripts—eight sgRNAs targeting the HBA transcripts and eight sgRNAs targeting the HBB transcripts. The sgRNAs were selected to deplete hemoglobin transcripts from human and polar bear samples. The sgRNAs were chosen based upon sequence homology between these two species to eventually allow the removal of abundant of hemoglobin transcripts in whole blood from both human and polar bear samples using the same sgRNAs (Field et al., 2007) (**Figure S1**). These 16 sgRNA probes we designed may allow for the depletion of samples of other vertebrates although sequence similarity should be checked before this is attempted.

The depletion process using the Cas9 system should cut the ~700–800 bp transcripts at different sites allowing us to do two things. First, we can re-amplify the sample, thereby only enriching for full-length molecules since the cut cDNA molecules no longer contain two priming sites required for exponential amplification during PCR amplification (**Figure 1**). Second, we can remove the cut transcripts by performing a SPRI-bead-based size selection whereby only transcripts > 500 bp are retained. Indeed, prior to any depletion, we observed very strong bands located at ~700–800 bp in our agarose gels indicating the presence of a substantial amount of HBA and HBB hemoglobin transcripts (**Figure 2**). After depletion, reamplification, and size selection, the full-length cDNA product was visualized again to reveal the removal of the putative hemoglobin bands (**Figure 2**). After hemoglobin depletion is confirmed, the cDNA is ready to be converted into ONT- and Illumina-based libraries, with each protocol using the same input cDNA (**Figure 1**).

Long-DASH Is Compatible With Smart-Seq2 Library Preparation and Does not Distort cDNA Composition

Next, we aimed to validate whether Long-DASH truly depletes hemoglobin transcripts in the cDNA pool and can be used for Illumina's short-read RNA-seq sequencing platform. To show this, we prepared independent Tn5-based Smart-seq2 sequencing libraries for each depleted and undepleted cDNA pool. We then sequenced the Smart-seq2 libraries on a multiplexed Illumina

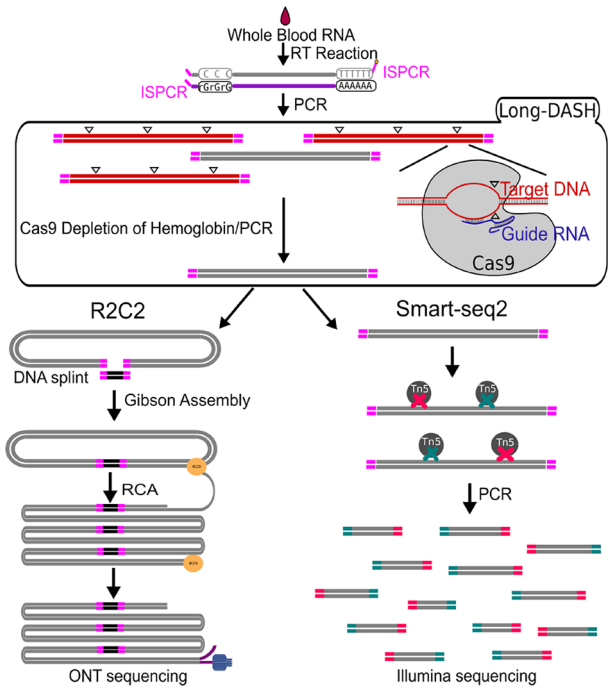

FIGURE 1 | Schematic of Long-DASH, (A) Whole blood RNA is extracted, and full-length cDNA is generated with the first half of the Smart-seq2 protocol. The cDNA is then depleted of hemoglobin transcripts using the recombinant S. pyogenes Cas9 protein bound to hemoglobin-specific sgRNA, which cuts hemoglobin cDNA molecules by introducing double-strand breaks (△) in a sequence-specific manner. The cut molecules can no longer be exponentially amplified with PCR, so a subsequent PCR step is performed to enrich for complete non-hemoglobin cDNA molecules. The resulting hemoglobin-depleted cDNA pool is then sequenced using the ONT-based R2C2 library prep and the Illumina-based Smart-seq2 library prep.

HiSeq X 2 × 151 bp run. We generated ~20 million reads for depleted and ~60 million reads for undepleted samples. By sequencing the undepleted samples deeper, we reasoned that the non-hemoglobin genes should receive equivalent read coverage in depleted and undepleted samples. This allowed us to perform a side-by-side comparison of the depleted and non-depleted samples to ensure no off-target effects.

First, we analyzed the resulting sequencing data using a custom kmer-based approach to estimate the number of reads originating from hemoglobin transcripts. In the undepleted cDNA pools, 48–68% of reads were scored as originating from hemoglobin transcripts. In depleted samples, this was reduced to 1–4% reads (**Figure 3A**). As a consequence, at the same read depth, reads per million (RPM) values for non-hemoglobin genes increased by a factor of 3 on average.

Second, to show that the depletion of hemoglobin transcripts did not distort the rest of the cDNA pool, we aligned the reads to the polar bear genome and quantified the expression of all previously annotated genes. We observed that when reads aligning to the hemoglobin loci were included in the analysis, the reads aligning to the few hemoglobin loci in our undepleted samples skewed the RPM calculations. By sequencing undepleted samples to great depth, this allowed us to exclude hemoglobin from quantification of gene expression while matching non-hemoglobin read depth of depleted samples. This analysis showed that the overall gene expression patterns were not dramatically distorted between depleted and undepleted samples. The three polar bear samples showed a Pearson r-value of 0.97–0.98 (**Figure 3B**) when the gene expression values of depleted and undepleted samples were compared, and reads aligning to hemoglobin loci were discarded.

Next, we checked for genes whose expression was systematically affected by depletion. No genes were downregulated more than

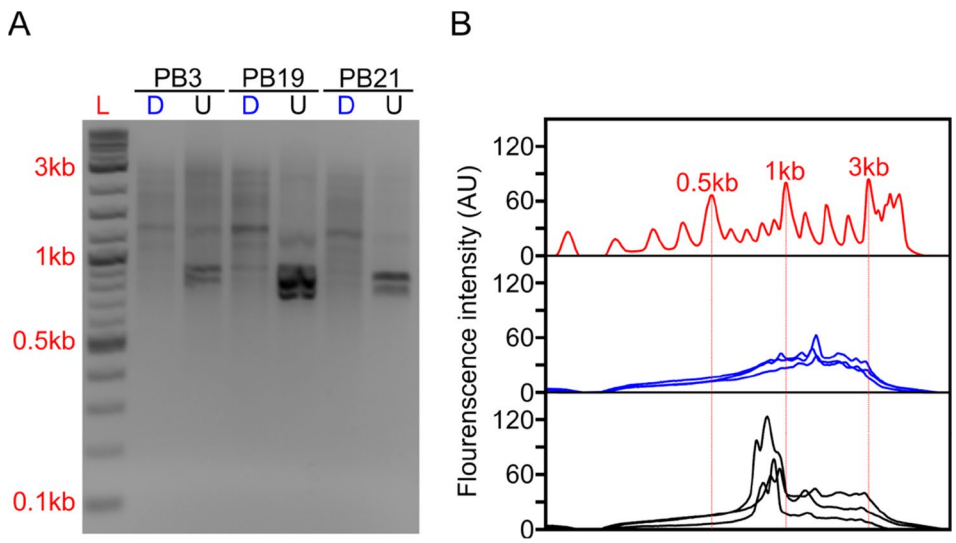

FIGURE 2 | Long-DASH depletes hemoglobin from full-length cDNA. **(A)** Depleted (D) or undepleted (U) cDNA was visualized on a 2% Agarose gel. DNA ladder (L) suggests highly abundant cDNA species—putatively hemoglobin around ~700 bp. **(B)** Fluorescence analysis of the gel by ImageJ (Rasband, 2011) further emphasizes the difference between depleted (blue) and undepleted (black) cDNA pool. Select size markers in the DNA ladder (red) are indicated.

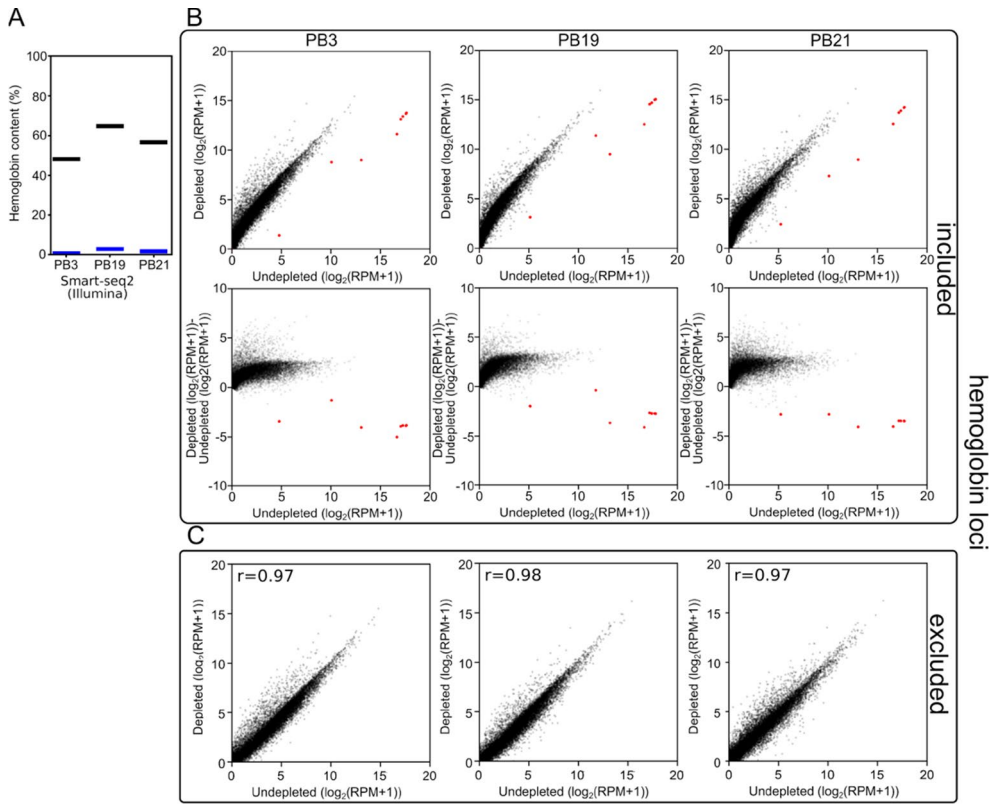

FIGURE 3 | Long-DASH depletes hemoglobin from cDNA. **(A)** Hemoglobin content was measured in Smart-seq2 (Illumina) libraries of depleted (blue) or undepleted (black) cDNA pools. **(B)** (Top) and **(C)** Scatterplots comparing gene expression in undepleted and depleted Smart-seq2 libraries of PB3, PB19, and PB21 with reads aligning to hemoglobin loci (red) (Liu et al., 2014) either included **(B)** or excluded **(C)**. **(B)** (Bottom) Scatterplots showing log2(fold-change) between depleted and undepleted cDNA pools as calculated by (depleted [log2(RPM+1)]-undepleted [log2(RPM+1)]) with hemoglobin loci included in the RPM normalization.

4-fold in all three polar bear samples suggesting that there were no strong systematic off-target effects using the Cas9-based depletion. We did, however, find 151 genes out of ~12,000 expressed genes to be upregulated by at least fourfold in all three polar bears suggesting that Cas9-based depletion and subsequent second PCR amplification have had a systematic impact on a number of genes. We then investigated whether this effect would affect differential expression analysis between depleted samples. To this end, we calculated gene expression differences for each pair of polar bears twice, once pre- and once post-depletion. We then compare the pre- and post-depletion gene expression differences and found that, while depletion does introduce differences in the upregulated genes, these effects appear to be small, random in direction, and similar to a random selection of genes with similar expression levels (**Figure S2**).

Overall, this indicates that the depletion of hemoglobin from full-length cDNA pools was successful, thereby freeing up the vast majority of sequencing reads to analyze the rest of the polar bear transcriptome. Although, the data suggests that a number of genes were systematically affected by depletion and additional PCR steps, further experiments including several technical replicates should enable differential expression analysis between depleted samples.

Long-DASH Is Compatible With Full-Length cDNA Sequencing Methods

Having established the compatibility of Long-DASH with the short-read RNA-seq assay, we investigated whether we could generate a long-read data set from the depleted cDNA using our R2C2 approach. By incorporating R2C2, we can generate error-corrected full-length cDNA reads using long-read ONT sequencers. We used five partially multiplexed flowcells to generate ~3.8 million R2C2 consensus reads of five depleted cDNA pools—two Long-DASH replicates (R1 and R2) for PB3 and PB19 as well as a single Long-DASH run for PB21. The R2C2 reads we generated had a median accuracy of 94%, which is between 8 and 10% more accurate than standard ONT cDNA sequencing protocols (**Table S1**).

We also generated ~5,000 R2C2 consensus reads of undepleted cDNA from one polar bear, which allowed us to compare hemoglobin content and consensus read length distributions between depleted and undepleted samples (**Figure 4**). In the undepleted sample, the majority of R2C2 reads were of two distinct lengths, both around 700 bp, likely representing the 79.3% of hemoglobin transcripts present in that sample. The five depleted samples showed a much more evenly distributed read length with a median hemoglobin content of 1.2% (0.6–8.3%)

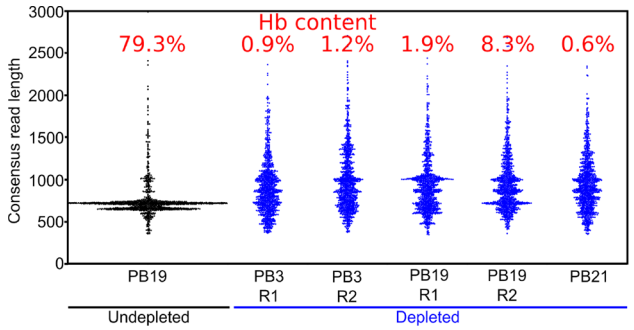

FIGURE 4 | Long-DASH depletes hemoglobin from cDNA. Length distribution of R2C2 consensus reads is shown as swarmplots in the indicated samples. Independent Long-DASH replicates (R1 and R2) were performed for samples PB3 and PB19 but not PB21. Percent of hemoglobin reads as determined with a kmer approach is given in red on top.

(**Figure 4**). Higher hemoglobin levels for R2C2 compared to Smart-seq2 based library preps (1–4%) using the same cDNA might be explained with R2C2 being somewhat biased toward transcripts between 500 and 1,000 bp.

The median read length of the depleted samples was slightly below 1 kb, which is in line with cDNA read length distributions published to date (Workman et al., 2018). This means that despite the less than ideal conditions for RNA integrity given difficult field conditions and the lag time between sample collection and processing, the analyzed RNA molecules were largely intact.

R2C2 Reads of Depleted Full-Length cDNA Can Refine Transcriptome Annotations

Next, we generated high confidence isoform-level information from our full-length cDNA to refine the currently available polar bear transcriptome annotation. To this end, we analyzed our 3.8 million R2C2 consensus reads using the Mandalorion pipeline we previously developed (Byrne et al., 2017). We aligned the R2C2 reads to the polar bear genome sequence (Liu et al., 2014) using minimap2. These alignments, together with previously known individual splice sites (Genomic Resources Development Consortium et al., 2014; Liu et al., 2014), then serve as input into our Mandalorion pipeline which processes read alignments into high-confidence isoforms.

The Mandalorion pipeline first complements known splice sites with new splice sites; it identifies *de novo* from R2C2 read alignments. It then groups R2C2 reads based on the splice sites they use. Pairs of TSSs and polyA sites are then determined for each group to identify full-length isoforms. Two additional processing steps were performed whereby isoforms were excluded if they were fully contained within longer isoforms or unspliced. This was to ensure removal of any non-full-length isoforms that may result from RNA degradation, as well as isoforms potentially caused by DNA contamination, respectively. In total, this analysis produced 5,831 high-confidence isoforms with a median accuracy of 99.1%.

We then classified these 5,831 high-confidence-spliced isoforms using the SQANTI algorithm (Tardaguila et al., 2018) that determines what relationship an experimentally determined isoform has to genes and isoforms in a reference annotation (**Figure 5**). As a reference, we downloaded 28,880 known and predicted mRNA sequences from NCBI by selecting "RefSeq" and "mRNA" filters in the NCBI Nucleotide database most of which are based on the NCBI *Ursus maritimus* Annotation Release 100 catalog of polar bear mRNA sequences (Pruitt et al., 2014).

Out of the 5,831 Mandalorion isoforms, 1,239 were classified as "novel_not_in_catalog" (NNC), which means that they overlapped a known gene but contained at least one unannotated splice site. In-depth analysis of this NNC group found that they contained a total of 521 new exons. In addition to R2C2 read coverage, Smart-seq2 read coverage was elevated in these new exons providing additional evidence for their inclusion in transcripts. Further, 1,301 isoforms were classified as "novel_in_catalog" (NIC), which means that they overlapped a known gene and used only annotated splice sites but at least once as part of a previously unannotated splice junction. In total, we observed 2,540 (1,239 NNC and 1,301 NIC) new isoforms with unannotated exon configurations. An additional 1,893 isoforms were classified and "full_splice_match" (FSM), which means that their splice junctions matched an annotated isoform completely, but it doesn't mean that TSS and polyA sites also matched this isoform. In-depth analysis of the putative full-length NNC, NIC, and FSM isoform groups identified 2,885 new TSSs and 1,817 new polyA sites. R2C2 read coverage declined rapidly at TSSs and polyA sites providing clear evidence for their validity. Smart-seq2 read coverage was elevated inside TSS and polyA sites but declined slowly toward the respective features, which is characteristic for standard short-read Illumina data (**Figure 5**). This is not surprising as short-read-based protocols have to be specifically designed to capture those features (Ruan and Ruan, 2011; Salimullah et al., 2011; Cole et al., 2018). So, while these data validate the existence of these features, they cannot be used for confirming their exact location.

Finally, 769 isoforms were classified as "incomplete_splice_match" (ISM), which means that they contain a subset of splice junctions of an annotated isoform. While these isoforms could represent real, shorter transcripts, they might also represent experimental artifacts so we excluded them from TSS and polyA analysis.

Considering RefSeq mRNA sets are in part based on deep short-read data and computational annotation, we did not expect to discover many entirely new gene loci. However, 509 of the 5,831 isoforms were classified as "intergenic" (IG), which means that they did not overlap with any annotated gene locus. By determining which of these isoforms overlapped with each other, we identified 176 new gene loci.

Overall, this analysis dramatically refined our isoform-level knowledge of the whole blood polar bear transcriptome (**Figure 5**). To make this knowledge straightforward to use for future analysis, we have generated a GTF annotation file containing RefSeq mRNA entries merged with our R2C2/Mandalorion isoforms.

How these new isoforms and isoform features have improved the current annotation can be seen clearly in these three

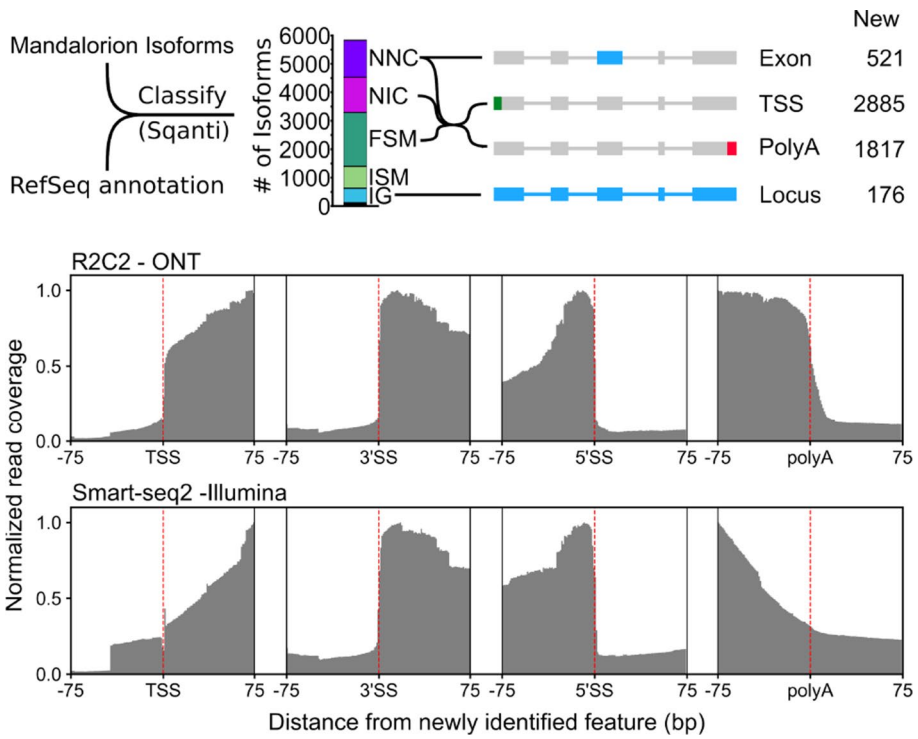

FIGURE 5 | R2C2/Mandalorion isoforms identify new features in the polar bear transcriptome. (Top) General workflow for comparing RefSeq mRNAs and Mandalorion isoforms is shown on the left. RefSeq mRNAs were aligned to the polar genomes using minimap2 and converted to GTF format to create a reference annotation. Isoforms determined by Mandalorion were then classified using this reference annotation using the sqanti_qc algorithm. Isoforms were classified as Novel_not_in_catalog (NNC), Novel_in_catalog (NIC), Full_splice_match (FSM), Incomplete_splice_match (ISM), and Intergenic (IG). New transcriptome features were then determined based on the minimap2 alignments of isoforms in the indicated categories. (Bottom) R2C2 and Smart-seq2 read coverage around newly identified TSS, the splice sites (3' and 5') of newly identified exons, and newly identified polyA sites.

following examples. In the RBX1 gene, we discovered 10 new isoforms containing new TSSs and polyA sites, several of which were associated with new terminal first or last exons (**Figure 6A**). In the GMFG gene, we similarly identified new isoforms containing unannotated internal and terminal exons, intron retention events, TSSs, and polyA sites (**Figure 6B**). Finally, we discovered a new gene locus that contains two isoforms and is entirely absent in the polar bear RefSeq mRNA set. However, aligning the two isoforms to the Panda genome (Li et al., 2010) resulted in unique matches to the CCDC72 gene (**Figure 6C**).

DISCUSSION

To better understand how humans and environmental perturbations impact threatened or endangered species, it is critical to understand changes in transcriptome dynamics. Fluctuations at the molecular and cellular level are sensitive indicators of environmental change (Kim et al., 2011; Brown et al., 2017); they are analogous to veterinary medicine where blood transcriptomes serve as proxies for identifying health status, disease, and exposures to environmental toxicants (Burgess et al., 2012; McLoughlin et al., 2014; Watson et al., 2017; Lv et al., 2018). Changes at the transcriptome level may also be useful indicators of ecological specialization, and therefore useful to design strategies for species management and conservation (Supple and Shapiro, 2018). However, existing approaches to generate transcriptome data from whole blood RNA are either specifically designed for short-read sequencing (DASH) or human samples (commercial hemoglobin depletion kit like GLOBINclear) and therefore lack a cost-effective approach for analyzing isoform-level transcriptomes of non-model organisms.

Any study investigating whole blood transcriptomes using short- or long-read sequencing will greatly benefit from the Long-DASH method. Long-DASH effectively and economically depletes hemoglobin transcripts from whole blood full-length cDNA, which can then be sequenced with short- or long-read sequencing. We validated Long-DASH by depleting hemoglobin transcripts from polar bear whole blood cDNA pools and generated deep short-read Smart-seq2 RNA-seq data as well as 3.8 million R2C2 full-length cDNA consensus reads. We processed the 3.8 million full-length R2C2 reads to identify close to 6,000 high confidence isoforms which we then used to refine and improve the polar bear whole blood transcriptome annotation.

In addition to polar bear hemoglobin transcripts, the sgRNAs designed for this study will also target human hemoglobin transcripts making them useful for basic research as well as

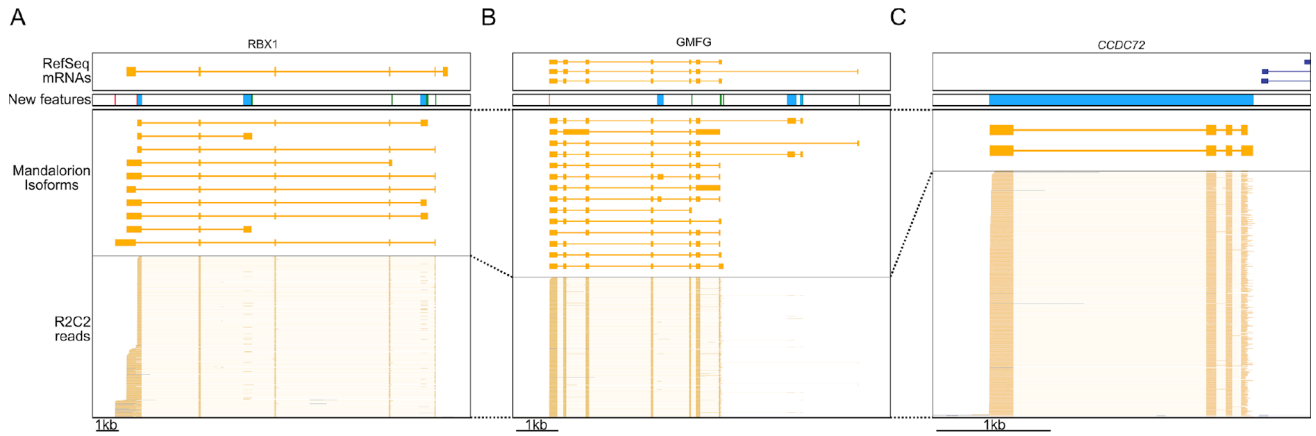

FIGURE 6 | R2C2/Mandalorion refine transcriptome annotations. Genome Browser views of the RBX1 locus **(A)**, the GMFG locus **(B)**, a locus likely corresponding to the CCDC72 gene not yet included in the RefSeq mRNA set **(C)**. From top to bottom, (1) RefSeq mRNA alignments, (2) new features based on Mandalorion isoforms (green: TSS, red: polyA site, blue: new exon or locus), (3) Mandalotion isoforms, and (4) R2C2 reads. Plus strand alignment are in blue, minus strand alignments in orange.

clinical applications in cancer biology and disease diagnosis (**Figure S1**) (Valk et al., 2004; Borovecki et al., 2005; Gervasoni et al., 2008; Morey et al., 2016). Further, the sgRNA sequences used in Long-DASH can be easily adapted to any organisms or gene. The ease and adaptability place Long-DASH at an advantage over "as-is" commercial kits like GLOBINclear (Ambion), which promises >95% of depletion of human and mouse hemoglobin transcripts, but fails to efficiently deplete hemoglobin transcripts from pig whole blood RNA samples (Choi et al., 2014).

Since cDNA can be generated from femtogram levels of polyA-RNA, Long-DASH requires very little RNA input compared to RNA pull-down methods. This allows the investigator to gather small samples, or only process small aliquots of existing samples, thereby maximizing the usefulness of each sample collection and minimizing harm to animals. In its current state, depletion by Long-DASH is still somewhat variable, resulting in hemoglobin levels from 0.6% to 8.3%. While still a large improvement compared to the undepleted samples, future work on the method will center on removing this variability through either longer incubation times or higher number or concentration of sgRNA probes and the Cas9 protein. It may also be beneficial to measure depletion success by qPCR before sequencing a depleted cDNA pool.

Going forward, the Long-DASH depletion method and the R2C2 long-read sequencing method will form a very powerful combination for transcriptome analysis and annotation from whole blood samples and beyond. The transcriptomes of many tissues contain several highly abundant transcripts that represent >50% of all transcript molecules (Mure et al., 2018). A set of sgRNAs targeting any abundant transcripts can be easily generated, making Long-DASH conducive for surveying other tissues as well. Specifically, depleted full-length cDNA libraries can be sequenced using our R2C2 method, which currently represents the most powerful combination of throughput and accuracy in the long-read sequencing field. Our most recent R2C2 run emphasizes this by generating ~1,000,000 R2C2 reads at a median accuracy of 97.5% on a single ONT MinION flowcell at a cost of ~$650 (**Table S1**). This represents an increase in accuracy of >10% over standard ONT cDNA sequencing and 10 times more complete reads than the PacBio Sequel at the same cost. Combining our Long-DASH and R2C2 methods therefore brings the exhaustive annotation of non-model organisms within reach.

DATA AVAILABILITY

All Illumina and ONT raw read data are available at SRA under Bioproject accession PRJNA514749.

ETHICS STATEMENT

Permits for field operations and animal care were provided by the Government of Greenland (permit numbers 2015-110281 and 2017-5446).

AUTHOR CONTRIBUTIONS

AB developed the Long-DASH method. AB, MS, KL, BS, and CV conceived and designed the research. KL collected polar bear samples. AB and MS processed the samples. AB, RV, and CV analyzed the data. AB and CV wrote the manuscript draft. AB, MS, RV, KL, BS, and CV edited the manuscript.

FUNDING

We acknowledge funding by the National Human Genome Research Institute/National Institute of Health Training Grant 1T32HG008345-01 (to AB and RV), the 2017 Hellman Fellowship (to CV), National Science Foundation grant DEB 1754451 (to

BS), funding for the field work provided by Environmental Protection Agency (Ministry of Environment and Food of Denmark) DANCEA Program, and the Greenland Institute of Natural Resources.

ACKNOWLEDGMENTS

We thank Professor Rebecca Dubois and her lab at UC Santa Cruz for producing and providing us with the recombinant Tn5 enzyme used in this study and Professor Carrie Partch and her lab at UC Santa Cruz for producing and providing us with the recombinant Cas9 enzyme used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00643/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Andreadis, A. (2005). Tau gene alternative splicing: expression patterns, regulation and modulation of function in normal brain and neurodegenerative diseases. *Biochim. Biophys. Acta* 1739, 91–103. doi: 10.1016/j.bbadis.2004.08.010

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Borovecki, F., Lovrecic, L., Zhou, J., Jeong, H., Then, F., Rosas, H. D., et al. (2005). Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc. Natl. Acad. Sci. U. S. A.* 102, 11023–11028. doi: 10.1073/pnas.0504921102

Brown, T. M., Hammond, S. A., Behsaz, B., Veldhoen, N., Birol, I., and Helbing, C. C. (2017). De novo assembly of the ringed seal (Pusa hispida) blubber transcriptome: a tool that enables identification of molecular health indicators associated with PCB exposure. *Aquat. Toxicol.* 185, 48–57. doi: 10.1016/j.aquatox.2017.02.004

Burgess, S. T. G., Greer, A., Frew, D., Wells, B., Marr, E. J., Nisbet, A. J., et al. (2012). Transcriptomic analysis of circulating leukocytes reveals novel aspects of the host systemic inflammatory response to sheep scab mites. *PLoS One* 7, e42778. doi: 10.1371/journal.pone.0042778

Busslinger, M., Moschonas, N., and Flavell, R. A. (1981). Beta + thalassemia: aberrant splicing results from a single point mutation in an intron. *Cell* 27, 289–298. doi: 10.1016/0092-8674(81)90412-8

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., et al. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8, 16027. doi: 10.1038/ncomms16027

Choi, I., Bao, H., Kommadath, A., Hosseini, A., Sun, X., Meng, Y., et al. (2014). Increasing gene discovery and coverage using RNA-seq of globin RNA reduced porcine blood samples. *BMC Genomics* 15, 954. doi: 10.1186/1471-2164-15-954

Cole, C., Byrne, A., Beaudin, A. E., Forsberg, E. C., and Vollmers, C. (2018). Tn5Prime, a Tn5 based 5′ capture method for single cell RNA-seq. *Nucleic Acids Res.* 46 (10), e62–e62. doi: 10.1093/nar/gky182

Debey, S., Schoenbeck, U., Hellmich, M., Gathof, B. S., Pillai, R., Zander, T., et al. (2004). Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. *Pharmacogenomics J.* 4, 193–207. doi: 10.1038/sj.tpj.6500240

Du, L., Li, W., Fan, Z., Shen, F., Yang, M., Wang, Z., et al. (2015). First insights into the giant panda (Ailuropoda melanoleuca) blood transcriptome: a resource for novel gene loci and immunogenetics. *Mol. Ecol. Resour.* 15, 1001–1013. doi: 10.1111/1755-0998.12367

Field, L. A., Jordan, R. M., Hadix, J. A., Dunn, M. A., Shriver, C. D., Ellsworth, R. E., et al. (2007). Functional identity of genes detectable in expression profiling assays following globin mRNA reduction of peripheral blood samples. *Clin. Biochem.* 40, 499–502. doi: 10.1016/j.clinbiochem.2007.01.004

Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M., and Joung, J. K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* 32, 279–284. doi: 10.1038/nbt.2808

Genomic Resources Development Consortium, Coltman, D. W., Davis, C. S., Lunn, N. J., Malenfant, R. M., and Richardson, E. S. (2014). Genomic resources notes accepted 1 August 2013-30 September 2013. *Mol. Ecol. Resour.* 14, 219. doi: 10.1111/1755-0998.12190

Gervasoni, A., Monasterio Muñoz, R. M., Wengler, G. S., Rizzi, A., Zaniboni, A., and Parolini, O. (2008). Molecular signature detection of circulating tumor cells using a panel of selected genes. *Cancer Lett.* 263, 267–279. doi: 10.1016/j.canlet.2008.01.003

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107. doi: 10.1016/S0168-9525(00)02176-4

Gu, W., Crawford, E. D., O'Donovan, B. D., Wilson, M. R., Chow, E. D., Retallack, H., et al. (2016). Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* 17, 41. doi: 10.1186/s13059-016-0904-5

Harr, B., and Turner, L. M. (2010). Genome-wide analysis of alternative splicing evolution among Mus subspecies. *Mol. Ecol.* 19 Suppl 1, 228–239. doi: 10.1111/j.1365-294X.2009.04490.x

Hernández-Fernández, J., Pinzón, A., and Mariño-Ramírez, L. (2017). De novo transcriptome assembly of loggerhead sea turtle nesting of the Colombian Caribbean. *Genom. Data* 13, 18–20. doi: 10.1016/j.gdata.2017.06.005

Huang, Z., Gallot, A., Lao, N. T., Puechmaille, S. J., Foley, N. M., Jebb, D., et al. (2016). A nonlethal sampling method to obtain, generate and assemble whole blood transcriptomes from small, wild mammals. *Mol. Ecol. Resour.* 16, 150–162. doi: 10.1111/1755-0998.12447

Hunter, J. D. (2007). Matplotlib: a 2D Graphics Environment. *Comput. Sci. Eng.* 9, 90–95. doi: 10.1109/MCSE.2007.55

Ilagan, J. O., Ramakrishnan, A., Hayes, B., Murphy, M. E., Zebari, A. S., Bradley, P., et al. (2015). U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res.* 25, 14–26. doi: 10.1101/gr.181016.114

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA NEB in adaptive bacterial immunity. *Science* 337, 816–821. doi: 10.1126/science.1225829

Jones, E., Oliphant, T., and Peterson, P., (2001). {SciPy}: Open source scientific tools for {Python}. Available at: http://www.scipy.org.

Kalsotra, A., Xiao, X., Ward, A. J., Castle, J. C., Johnson, J. M., Burge, C. B., et al. (2008). A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proc. Natl. Acad. Sci. U. S. A.* 105, 20333–20338. doi: 10.1073/pnas.0809045105

Khudyakov, J. I., Champagne, C. D., Meneghetti, L. M., and Crocker, D. E. (2017). Blubber transcriptome response to acute stress axis activation involves transient changes in adipogenesis and lipolysis in a fasting-adapted marine mammal. *Sci. Rep.* 7, 42110. doi: 10.1038/srep42110

Kim, J. K., Jung, K. H., Noh, J. H., Eun, J. W., Bae, H. J., Xie, H. J., et al. (2011). Identification of characteristic molecular signature for volatile organic compounds in peripheral blood of rat. *Toxicol. Appl. Pharmacol.* 250, 162–169. doi: 10.1016/j.taap.2010.10.009

Lee, C., Grasso, C., and Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics* 18, 452–464. doi: 10.1093/bioinformatics/18.3.452

Liew, C.-C., Ma, J., Tang, H.-C., Zheng, R., and Dempsey, A. A. (2006). The peripheral blood transcriptome dynamically reflects system wide biology:

a potential diagnostic tool. *J. Lab. Clin. Med.* 147, 126–132. doi: 10.1016/j.lab.2005.10.005

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34 (18), 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lindenbaum, P. (2015). JVarkit: java-based utilities for Bioinformatics. *figshare*. doi: 10.6084/m9.figshare.1425030.v1

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311–317. doi: 10.1038/nature08696

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., et al. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* 157, 785–794. doi: 10.1016/j.cell.2014.03.054

Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., et al. (2015). The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* 43, W580–4. doi: 10.1093/nar/gkv279

Lv, J., Ding, Y., Liu, X., Pan, L., Zhang, Z., Zhou, P., et al. (2018). Gene expression analysis of porcine whole blood cells infected with foot-and-mouth disease virus using high-throughput sequencing technology. *PLoS One* 13, e0200081. doi: 10.1371/journal.pone.0200081

Mastrokolias, A., den Dunnen, J. T., van Ommen, G. B., 't Hoen, P. A. C., van Roon-Mom, W. M. C. (2012). Increased sensitivity of next generation sequencing-based expression profiling after globin reduction in human blood RNA. *BMC Genomics* 13, 28. doi: 10.1186/1471-2164-13-28

McLoughlin, K. E., Nalpas, N. C., Rue-Albrecht, K., Browne, J. A., Magee, D. A., Killick, K. E., et al. (2014). RNA-seq transcriptional profiling of peripheral blood leukocytes from cattle infected with mycobacterium bovis. *Front. Immunol.* 5, 396. doi: 10.3389/fimmu.2014.00396

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., et al. (2013). Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.* 41, W597–600. doi: 10.1093/nar/gkt376

Millman, K. J., and Aivazis, M. (2011). Python for scientists and engineers. *Comput. Sci. Eng.* 13, 9–12. doi: 10.1109/MCSE.2011.36

Morey, J. S., Neely, M. G., Lunardi, D., Anderson, P. E., Schwacke, L. H., Campbell, M., et al. (2016). RNA-Seq analysis of seasonal and individual variation in blood transcriptomes of healthy managed bottlenose dolphins. *BMC Genomics* 17, 720. doi: 10.1186/s12864-016-3020-8

Mure, L. S., Le, H. D., Benegiamo, G., Chang, M. W., Rios, L., Jillani, N., et al. (2018). Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. *Science* 359 (6381), eaao0318. doi: 10.1126/science.aao0318

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* 9, 171–181. doi: 10.1038/nprot.2014.006

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., et al. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* 42, D756–63. doi: 10.1093/nar/gkt1114

Rasband, W. S. (2011). Imagej, US National Institutes of Health, Bethesda, Maryland, USA. http://imagej.nih.gov/ij/. Available at: https://ci.nii.ac.jp/naid/10030139275/.

Ren, X., Yang, Z., Xu, J., Sun, J., Mao, D., Hu, Y., et al. (2014). Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. *Cell Rep.* 9, 1151–1162. doi: 10.1016/j.celrep.2014.09.044

Ruan, X., and Ruan, Y. (2011). "RNA-PET: full-Length Transcript Analysis Using 5′- and 3′-Paired-End Tag Next-Generation Sequencing," in *Tag-Based Next Generation Sequencing*, Wiley-VCH Verlag GmbH & Co. p. 73–90. doi: 10.1002/9783527644582.ch5

Salimullah, M., Sakai, M., Mizuho, S., Plessy, C., and Carninci, P. (2011). NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb. Protoc.* 2011, db.prot5559. doi: 10.1101/pdb.prot5559

Shin, H., Shannon, C. P., Fishbane, N., Ruan, J., Zhou, M., Balshaw, R., et al. (2014). Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. *PLoS One* 9, e91041. doi: 10.1371/journal.pone.0091041

Shi, X., Ng, D. W.-K., Zhang, C., Comai, L., Ye, W., and Chen, Z. J. (2012). Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. *Nat. Commun.* 3, 950. doi: 10.1038/ncomms1954

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi: 10.1038/msb.2011.75

Supple, M. A., and Shapiro, B. (2018). Conservation of biodiversity in the genomics era. *Genome Biol.* 19, 131. doi: 10.1186/s13059-018-1520-3

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* 28 (3), 396–411. doi: 10.1101/gr.222976.117

Ungaro, A., Pech, N., Martin, J.-F., McCairns, R. J. S., Mévy, J.-P., Chappaz, R., et al. (2017). Challenges and advances for transcriptome assembly in non-model species. *PLoS One* 12, e0185020. doi: 10.1371/journal.pone.0185020

Valk, P. J. M., Verhaak, R. G. W., Beijen, M. A., Erpelinck, C. A. J., Barjesteh van Waalwijk van Doorn-Khosrovani, S., Boer, J. M., et al. (2004). Prognostically useful gene-expression profiles in acute myeloid leukemia. *N. Engl. J. Med.* 350, 1617–1628. doi: 10.1056/NEJMoa040465

van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy Array: a Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* 13, 22–30. doi: 10.1109/MCSE.2011.37

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746. doi: 10.1101/gr.214270.116

Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R. J., Green, R. E., et al. (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* 115 (39), 9726–9731. doi: 10.1073/pnas.1806447115

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi: 10.1038/nature07509

Watson, H., Videvall, E., Andersson, M. N., and Isaksson, C. (2017). Transcriptome analysis of a wild bird reveals physiological responses to the urban environment. *Sci. Rep.* 7, 44180. doi: 10.1038/srep44180

Workman, R. E., Tang, A., Tang, P. S., Jain, M., Tyson, J. R., Zuzarte, P. C., et al. (2018). Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv* 459529. doi: 10.1101/459529

Zhang, X., Chen, M. H., Wu, X., Kodani, A., Fan, J., Doan, R., et al. (2016). Cell-Type-specific alternative splicing governs cell fate in the developing cerebral cortex. *Cell* 166, 1147–1162.e15. doi: 10.1016/j.cell.2016.07.025

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.