Title

Generation of an isoform-level transcriptome atlas of macrophage activation.

Authors

Vollmers, Apple Cortez
Mekonen, Honey E
Campos, Sophia
et al.

Publication Date

2021

DOI

10.1016/j.jbc.2021.100784

Peer reviewed

JBC RESEARCH ARTICLE

Generation of an isoform-level transcriptome atlas of macrophage activation

Received for publication, March 2, 2021, and in revised form, May 5, 2021 Published, Papers in Press, May 14, 2021,
https://doi.org/10.1016/j.jbc.2021.100784

Apple Cortez Vollmers[1], Honey E. Mekonen[2], Sophia Campos[2], Susan Carpenter[1,*], and Christopher Vollmers[2,*]
From the [1]Department of Molecular, Cellular, and Developmental Biology, [2]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, California, USA

Edited by Peter Cresswell

RNA-seq is routinely used to measure gene expression changes in response to cell perturbation. Genes upregulated or downregulated following some perturbation are designated as genes of interest, and their most expressed isoform(s) would then be selected for follow-up experimentation. However, because of its need to fragment RNA molecules, RNA-seq is limited in its ability to capture gene isoforms and their expression patterns. This lack of isoform-specific data means that isoforms would be selected based on annotation databases that are incomplete, not tissue specific, or do not provide key information on expression levels. As a result, minority or nonexistent isoforms might be selected for follow-up, leading to loss in valuable resources and time. There is therefore a great need to comprehensively identify gene isoforms along with their corresponding levels of expression. Using the long-read nanopore-based R2C2 method, which does not fragment RNA molecules, we generated an Isoform-level transcriptome Atlas of Macrophage Activation that identifies full-length isoforms in primary human monocyte-derived macrophages. Macrophages are critical innate immune cells important for recognizing pathogens through binding of pathogen-associated molecular patterns to toll-like receptors, culminating in the initiation of host defense pathways. We characterized isoforms for most moderately-to-highly expressed genes in resting and toll-like receptor–activated monocyte-derived macrophages, identified isoforms differentially expressed between conditions, and validated these isoforms by RT-qPCR. We compiled these data into a user-friendly data portal within the UCSC Genome Browser (https://genome.ucsc.edu/s/vollmers/IAMA). Our atlas represents a valuable resource for innate immune research, providing unprecedented isoform information for primary human macrophages.

The use of RNA-seq is a primary strategy in biomedical research to identify genes involved in biological processes of interest and how gene expression is impacted upon gene editing or use of chemical or biological agonists. Notably, short-read sequencing technology has reliably been used to quantify changes in gene expression levels or the inclusion level of individual exons and splice junctions. However, because short-read RNA-seq relies on fragmenting RNA molecules before sequencing, even advanced computational tools fail at leveraging this ubiquitous data type into isoform-level information (1–3). Short-read RNA-seq ultimately falls short in providing comprehensive and accurate full-length isoform structures as well as the level of expression of each isoform under specific conditions.

More recently, long-read technologies from Pacific Biosciences and Oxford Nanopore Technologies (ONT) have been used for sequencing and analyzing full-length cDNA molecules at the transcriptome scale (4–7). In contrast to RNA-seq, this technology can determine which isoforms, down to the exact transcription start and poly(A) sites, are expressed at what level by each gene.

The comprehensive transcriptome scale isoform information these technologies provide has the potential to remove the need for targeted and work intensive methods like RT-PCR and 5'/3' RACE to identify and characterize transcript and/or protein isoforms expressed by a gene. Therefore, comprehensive transcriptome scale isoform information is bound to simplify and improve the outcome of single gene focused follow-up studies which include knock-down and knock-out experiments, overexpression assays, Western Blots, ELISAs, pull-downs, and many more. This is because of the fact that these assays rely on prior knowledge of what isoform(s) the gene of interest actually expresses in the condition and experimental system being investigated. Finally, detailed knowledge of transcription start sites (TSSs) for each expressed gene in a cell type will also improve the use of CRISPR interference technology to knock down genes because guide RNAs can be targeted to TSSs with greater accuracy.

To build on our previous work (8, 9) and further push the limits of long-read technology to provide a resource for the innate immune research community, we set out to generate an isoform-level transcriptome atlas of macrophage activation by determining (1) what isoform of a given gene is expressed, (2) at what level, and (3) how isoform and gene expression change following toll-like receptor (TLR) activation.

Macrophages are a key cellular component of the innate immune system which represents the first line of host defense against infection and is critical for the development of adaptive immunity (10, 11). Macrophages recognize conserved

* For correspondence: Susan Carpenter, sucarpen@ucsc.edu; Christopher Vollmers, vollmers@ucsc.edu.

structures of microbial-derived molecules or pathogen associated molecular patterns using TLRs. The regulation of this TLR repertoire fundamentally alters the response to infection (12). TLR activation induces the expression of hundreds of genes that encode inflammatory response genes including cytokines, type I interferons, antimicrobial proteins, and regulators of metabolism and regeneration; these molecules in turn mediate inflammation, antimicrobial immunity, and tissue regeneration.

Here, we investigated transcriptional responses of human macrophages treated with lipopolysaccharide (LPS), Pam3CSK4 (PAM), R848, and poly(I:C) which activate TLR4, TLR1/2, TLR7/8, and TLR3, respectively. Using our ONT-based R2C2 method, we then generated a total of ∼15 million full-length cDNA reads at a median accuracy >99% (Q20) and processed this data into isoforms which we characterized in depth and provide alongside deep Smart-seq2 short-read RNA-seq data as a UCSC Genome Browser session for easy exploration.

Results

Experimental setup

To generate a comprehensive isoform-level transcriptome atlas of TLR-dependent macrophage activation, we collected peripheral blood mononuclear cells from two individuals (Rep1 and Rep2) from which we isolated monocytes. From these monocytes, we generated monocyte-derived macrophages (MDMs). We treated these MDMs with TLR ligands LPS, PAM, R848, or poly(I:C) and included a no treatment (NoStim) control. After 6 h, we collected the stimulated and nonstimulated MDMs and proceeded to extract RNA from each sample. We reverse transcribed the poly(A) fraction of this RNA using a modified oligo(dT) primer and a template switch oligo to generate full-length cDNA

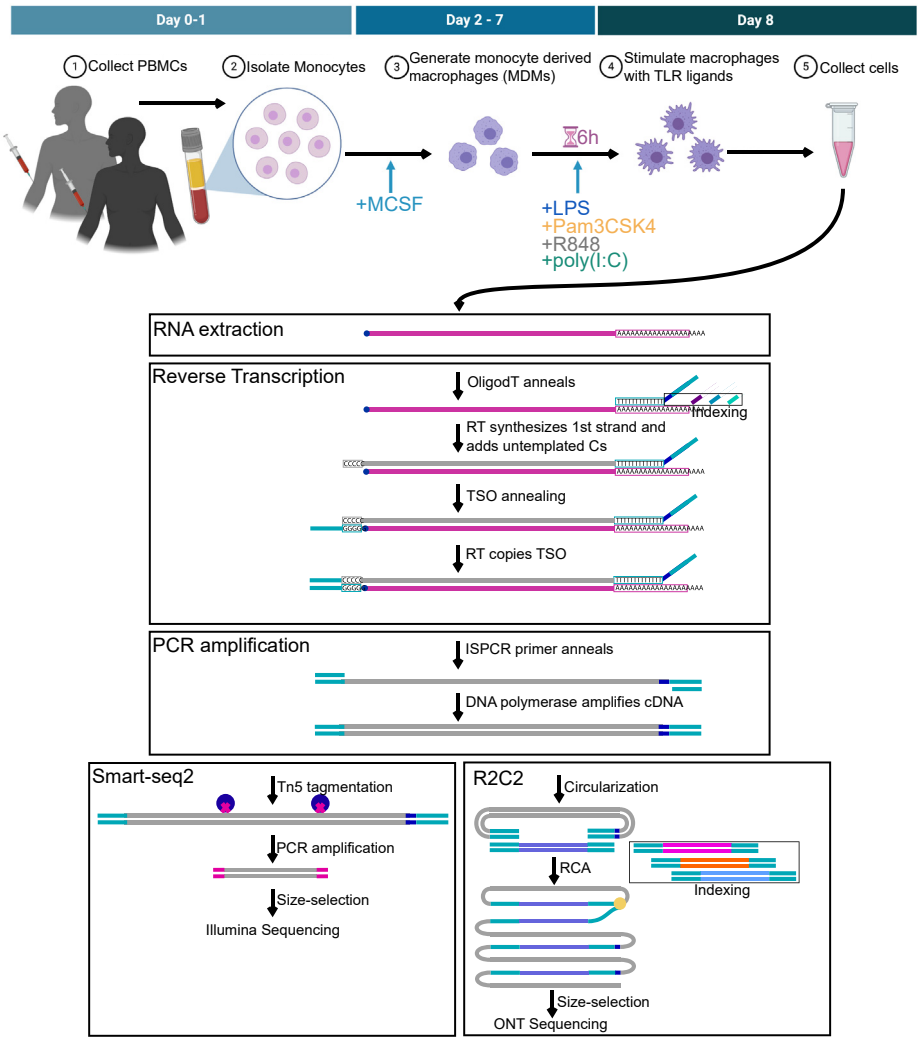

Figure 1. Experimental Design. A schematic of macrophage differentiation and activation is shown on *top*. A workflow for data generation is shown at the *bottom*. *Top*, we generated monocyte-derived macrophages (MDMs) from the peripheral blood mononuclear cells of two individuals (Rep1 and Rep2) by first isolating monocytes and treating them with macrophage colony-stimulating factor (M-CSF). The resulting macrophages were stimulated with TLR ligands for 6 h and then collected. *Bottom*, we extracted RNA and synthesized full-length cDNA which we then processed to generate Smart-seq2 and R2C2 libraries for Illumina and Oxford Nanopore Technologies (ONT) sequencing, respectively. We then performed gene and isoform level analysis of the resulting sequencing data. TLR, toll-like receptor.

with known sequences on both ends. We then amplified this cDNA using PCR and used the resulting double-stranded full-length cDNA as input for both Illumina-based Smart-seq2 (13) and ONT-based R2C2 (14) sequencing protocols (Fig. 1).

Smart-seq2 based gene-level differential expression analysis

To identify genes differentially expressed upon TLR activation following treatment with LPS, PAM, R848, or poly(I:C), we performed Illumina-based Smart-seq2 (15) sequencing as previously described (16, 17) (Fig. S1, see Experimental procedures). We generated approximately 15 to 30 million reads per sample (Table S1) and processed the resulting data using a standard workflow which includes STAR (18), featureCounts (19), and DEseq2 (20). Taking advantage of our biological replicates, we individually compared the LPS, PAM, R848, and poly(I:C) conditions to the NoStim control. Each ligand caused the differential expression (DE) of 1000 to 2000 genes (Tables S2 and S3). Between all conditions, 454 genes were shared, a varying number overlapped between three and two stimulants, and some were exclusive to one stimulant (Fig. 2*A*, Table S3). Using Panther Gene Ontology analysis (21, 22), we observed that genes shared between all stimulants were strongly enriched for biological processes including "response to cytokine" and inflammatory response" (Table S4). Notably, the genes exclusively responding to R848 were enriched for the "response to organic cyclic compound" biological process, likely because R848 is an organic cyclic compound.

R2C2 based isoform-level analysis

To supplement this gene-level DE analysis, we sequenced the same full-length double-stranded cDNA using our ONT-based R2C2 protocol (14, 17, 23) (Fig. 1). R2C2 circularizes cDNA and then amplifies the resulting circular DNA to generate long linear DNA molecules containing concatemeric copies of the initial cDNA. The resulting long DNA is then sequenced and computationally separated into subreads. We then combine these subreads to generate accurate consensus reads of the initial cDNA molecules. To make the creation of this macrophage isoform atlas feasible, we introduced several improvements to the R2C2 method.

First, we enabled multiplexing of cDNA samples by introducing 10 nt sample barcodes into the oligo(dT) primers as well as using highly distinct DNA splints for cDNA circularization (Fig. 1). Throughout the study, we used these two different

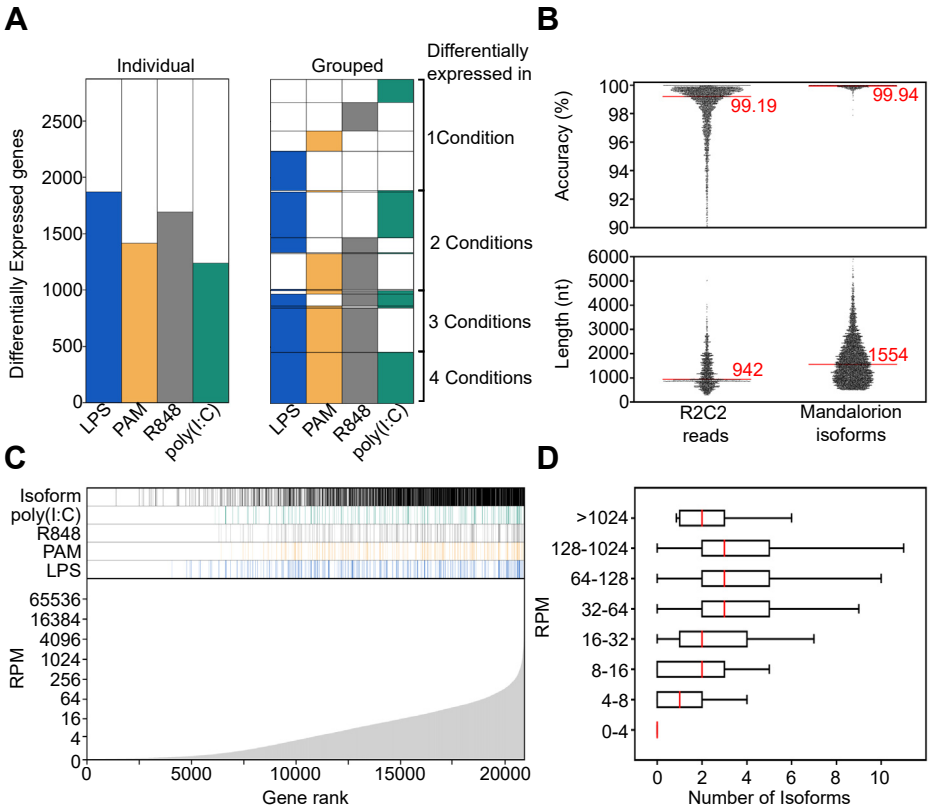

Figure 2. Gene and isoform level analysis. *A*, on the *left*, the numbers of genes differentially expressed between nonstimulated macrophages and macrophages stimulated with the indicated TLR ligands. On the *right*, genes are grouped if they were differentially expressed in more than one condition. For example, the 454 genes differentially expressed in all four conditions are shown at the *bottom*. Above those, the number of genes differentially expressed in LPS, PAM, and R848, but not poly(I:C) are shown. *B*, accuracy and length of individual R2C2 reads and Mandalorion isoforms are shown as swarmplots, with median values indicated by *red lines* and numeric values. *C*, different characteristics of genes ordered by expression are shown in this panel. From the *bottom* to *top,* this panel shows the average gene expression in reads per million (RPM) across all conditions, whether a gene is differentially expressed following LPS, Pam3CSK4 (PAM), R848, or poly(I:C) stimulation and whether we identified an isoform for the gene. *D*, the number of isoforms we detect for genes is shown as box plots for genes with different expression levels. TLR, toll-like receptor.

indexing strategies to separate technical and biological replicates (Table S5). Indexing samples allowed us to sequence samples in pools on the same ONT flow cells, thereby achieving equal sequencing coverage between samples and minimizing batch effects. Including pilot experiments that sequenced regular and size-selected cDNA of NoStim and LPS samples, we generated 14,961,450 R2C2 reads at a median length of 942 nt across multiple ONT MinION flow cells (Fig. 2*B*).

Second, after performing the pilot experiments, we improved R2C2 per read accuracy by increasing the raw read length of R2C2 libraries. We accomplished this by developing a gentle agarose gel extraction protocol that doubled the raw read length of our R2C2 libraries from ∼5 kb to ∼11 kb. Combined with a new ONT basecaller, this increased the median per base accuracy of our R2C2 libraries. Previously, this base accuracy was 97.9% (16). Here, our most recent sequencing runs show an increased base accuracy of 99.45% (Fig. S1). Overall, and including less accurate pilot experiments, the ∼15 million reads generated for this study had a median accuracy of 99.19% (Q21).

Third, to take advantage of this improved accuracy and refine the identification of isoforms from the R2C2 reads we generated, we developed a new version of our Mandalorion pipeline (Episode 3.5 - Rogue Isoform) that, among several changes, includes improved consensus generation using the Medaka polishing tool and improved handling of isoform ends. When applying this pipeline on the combined ∼15 million read data set, we identified 29,637 high confidence isoforms with a median length of 1554 nt and a per base accuracy of 99.94% (Q32) which matches the current state-of-the-art for ONT-only consensus accuracy (24) (Fig. 2*B*).

Enriching differential expression data set with isoform-level information

Next, we determined which genes these 29,637 isoforms were transcribed from and to what extent isoform identification was dependent on gene expression levels. Our Smart-seq2–based analysis identified 20,915 expressed genes (Average reads per million [RPM] across all conditions and replicates >0.05). Our R2C2-based analysis identified at least one isoform for 9688 (46%) of these genes.

Our ability to identify at least one isoform for a gene increased if the gene was expressed at higher levels, *i.e.*, had a higher average RPM (Fig. 2*C*). We identified at least one isoform for 13% of genes between 0.05 and 3 RPM. This percentage increased with increasing RPM to 49% (3–5 RPM), 64% (5–10 RPM), 80% (10–20 RPM), and 91% (>20 RPM).

Because differentially expressed genes tended to have higher average expression (Fig. 2*C*), we identified at least one isoform for 1986 of the 2873 (69%) genes differential expressed in any condition, and 363 of 454 (80%) genes differential expressed in all conditions.

The number of isoforms identified per gene also increased with average RPM (Fig. 2*D*). However, the median number of isoforms per gene did not exceed three even in very highly expressed genes. This is likely because of the Mandalorion filtering settings we used which discarded isoforms with less than 1% of all the R2C2 reads at a locus and thereby excluded minority isoforms.

Identifying genes with differentially expressed isoforms

Next, we established a pipeline to—in addition to gene-level DE—detect genes whose isoforms were differentially expressed between conditions (Fig. 3*A*). While DE pipelines like DEseq2 could be applied to this problem, they would likely just detect isoforms from differentially expressed genes. To detect genes whose relative isoform usage differed between conditions (*e.g.*, Gene A isoform 1 decreases, while Gene A isoform 2 increases following stimulation), we applied a Chi-squared contingency table test.

For each gene, we first determined all the isoforms transcribed from that gene. Then, we calculated the relative usage (%) for each isoform in each condition by dividing the number of R2C2 reads associated with that isoform by the number of R2C2 reads associated with all the isoforms in that condition. We then applied the Chi-squared contingency test to the resulting isoform-by-condition relative usage table. To reduce noise, we only tested the 1872 genes with at least 50 R2C2 reads in at least two experimental conditions. After Bonferroni multiple testing correction, this stringent test produced 47 genes with differential isoform usage with $\alpha = 0.05$ (Table S6). Twenty-five of these 47 (53%) genes were not identified as differentially expressed by the Smart-seq2 short-read workflow.

The 47 genes are likely to contain isoforms whose usage varies strongly between conditions. To confirm this, we determined the standard deviation of this isoform usage between conditions for each isoform in each gene (Fig. 3*B*). By sorting the 1872 genes we tested by the largest standard deviation among their isoforms, we showed that this Chi-square contingency table test did indeed identify genes with isoforms that have highly variable usage between conditions. To further evaluate the robustness of this approach, we validated the changes in the relative usage of alternative individual exons for the top ten genes with significant differential isoform usage following LPS treatment using macrophages from an additional donor. Using RT-qPCR, we found that in all ten cases we tested, relative exon usage in the third donor changed in the same direction (upregulated or downregulated), with many showing fold change similar to that determined by R2C2 in the two original donors (Fig. S2, Table S7).

The majority of the 47 genes featured alternative TSSs which in several cases were located in different and sometimes unannotated first exons (Fig. 3*C*, Table S6).

Classifying isoforms

To evaluate how frequent the use of unannotated exons is across all isoforms and how this affects their coding potential, we first categorized the 29,637 isoforms we identified. To this end, we used the SQANTI (25) algorithm which associates

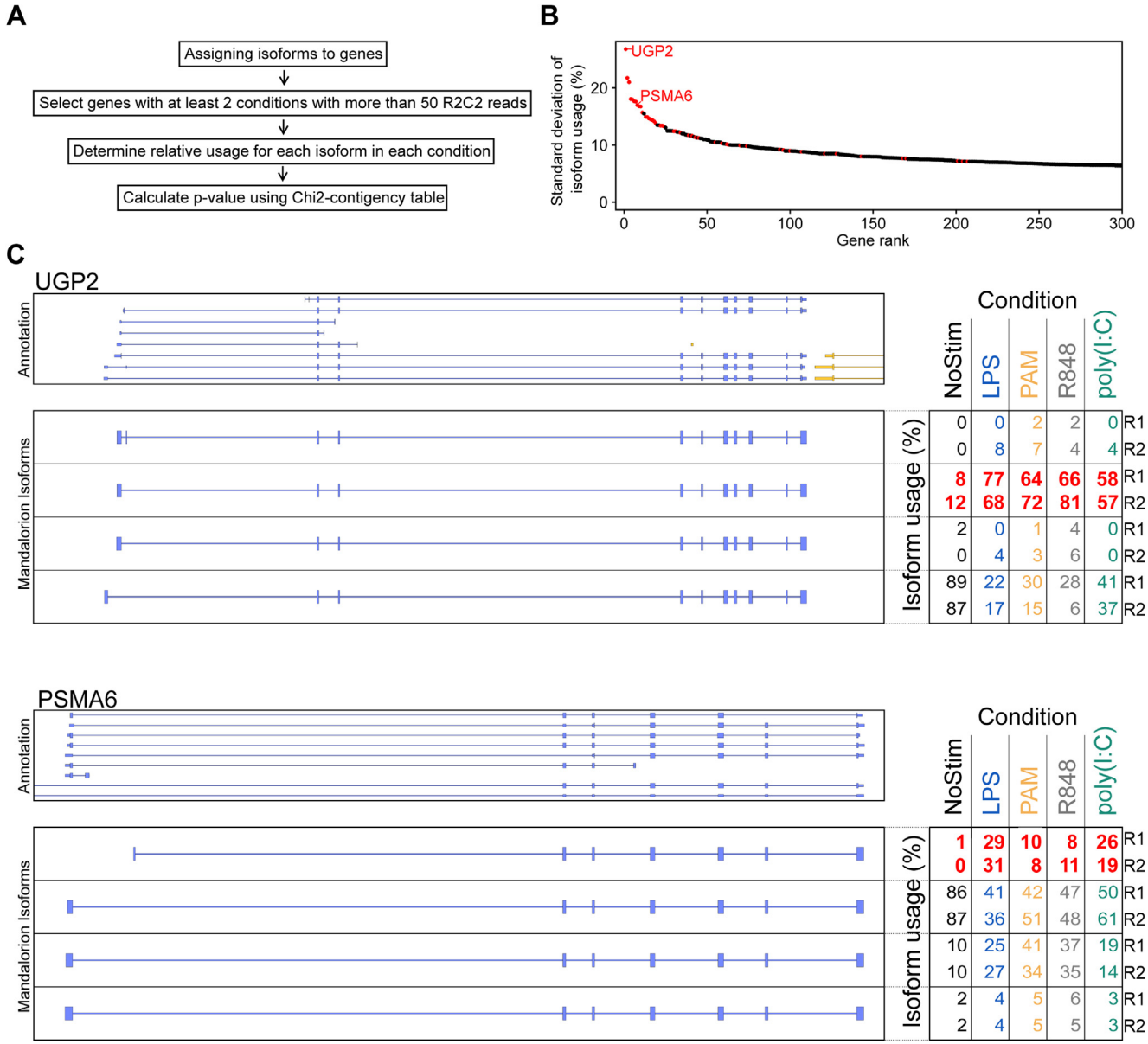

Figure 3. Differential isoform expression. *A*, workflow of differential isoform expression analysis. Genes were subselected based on expression, and differential expression was determined using a Chi-squared contingency table. *B*, genes are sorted by the maximum standard deviation of relative isoform usage among its isoforms. This maximum standard deviation is plotted (*red* if the gene has been identified as containing differentially expressed isoforms.) *C*, on the *left*, a Genome Browser view of the indicated genes is shown with GENCODE v34 annotation on *top* and identified isoforms below. On the *right*, the relative usage of each isoform in each replicate and condition is shown. Relative usage of the most variable isoform for each gene is highlighted in *red*.

isoforms with genes and categorizes them as full-splice matches (FSM), novel in catalog (NIC), novel not in catalog (NNC), incomplete splice-matches (ISM), and other less abundant categories (Fig. 4*A*). FSM and ISM isoforms are defined as fully (FSM) or incompletely (ISM) matching the splice-junction chain of an annotated GENCODE transcript. NIC are defined as isoforms that use annotated splice sites in unannotated configurations. NNCs are defined as isoforms that use at least one unannotated splice site (Fig. 4*A*).

FSM isoforms represented 65% (19,161), NIC isoforms represented 19% (5576), NNC isoforms represented 7% (2136), and ISM isoforms represented 6% (1878) of the 29,637 isoforms we identified (Table S8). If they were associated with a protein-coding gene, isoforms of different categories had different likelihoods to contain a full coding sequence (CDS) of the gene they were associated with. 94% of FSM, 58% of NIC, 47% of NNC, and 36% of ISM isoforms, which contained more than one exon, contained a full CDS of the protein-coding gene they were transcribed from (Fig. 4*B*).

FSM isoforms which did not contain a full CDS likely had to differ from the GENCODE transcript they matched in their TSS and polyA site positions. Indeed, the 5' and 3' ends of FSM isoforms varied in their distance to the annotated TSS and polyA sites of the GENCODE transcript they were associated with (Fig. 4*C*). Taking into account that 5' and 3' ends of a FSM isoform may be closer to the TSS or polyA

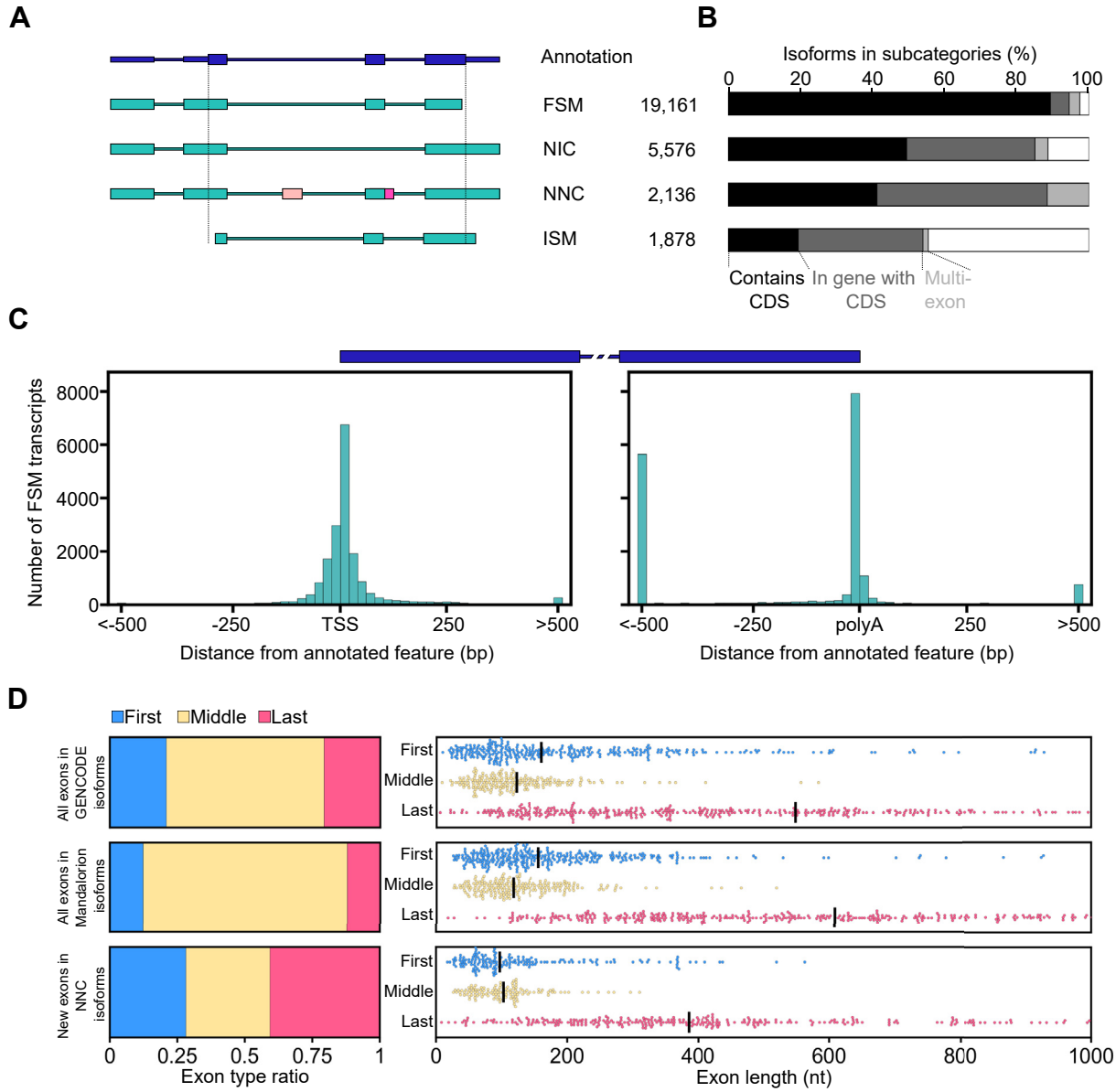

Figure 4. Isoform characterization. *A*, on the *left*, models of different isoform categories are shown. The models shown for full-splice matches (FSM), novel in catalog (NIC), novel not in catalog (NNC), and incomplete splice-matches (ISM) isoforms all do not contain the CDS of the annotation shown on *top*. The NNC model contains a new exon (*light pink*) and an extension of an annotated exon (*dark pink*). On the *right*, the numbers of identified isoforms that fall into each category are shown. *B*, the percentage of isoforms in the different categories that contain more than one exon fall within a gene that has a CDS and contain a CDS of that gene are shown as nested bar plots. *C*, the distance of 5′ and 3′ ends of FSM isoform to the TSS and poly(A) site of the transcript they are associated with is shown as a histogram. A transcript model is shown on *top* to give context to the histograms. *D*, on the *left*, the ratio of first, middle, and last exons within GENCODE isoforms, all isoforms identified by Mandalorion, and newly identified exons in NNC isoforms are shown. On the *right*, the lengths of first, middle, and last exons within these isoform groups are shown as swarm plots with *black bars* indicating the median. CDS, full coding sequence; TSS, transcription start site.

site of another GENCODE transcript in their respective gene, we determined that of the 19,161 FSM isoforms we identified, 276 had 5′ end and 2068 had a 3′ end more than 500 nt away from any annotated TSS or polyA site. The larger number of distant 3′ ends compared with 5′ ends in FSM isoforms could at least in part be explained by last exons being much longer on average than first exons (Fig. 4*D*).

NIC isoforms, which did not contain a full CDS of the gene they were associated with, likely contained a new splice junction between Start and Stop codons which would modify a CDS. NNC isoforms which did not contain a full CDS of the gene they were associated with might differ from the CDS by a few base pairs to encode a slightly different splice-junction or contain entirely new exons (Fig. 4*A*).

Identifying new exons in NNC isoforms

Next, we focused on NNC isoforms to annotate new exons. We defined a new exon as a part of a transcript whose genomic location does not overlap with a known exon at all (Fig. 4*A*). In the 2136 NNC isoforms we identified, 721 new exons of which 203 were first and 294 were last exons. If new exons were distributed equally

among the exons of the 29,637 isoforms we identified, we would expect 89 new first and last exons, indicating that first and last exons are overrepresented in this set of new exons (Fig. 4*D*, left).

Further, these new exons were shorter than exons in the GENCODE annotation (v34), or all exons identified by Mandalorion (Fig. 4*D*, right). Importantly, the length of new first, middle, and last exons followed the trend of annotated exons with last exons being 2 to 3× longer than first and middle exons.

Finally, the vast majority of new exons could be validated with short read Smart-seq2 data. Splice junctions leading into these exons (one for first/last exons, two for internal exons) were present in Smart-seq2 reads generated from the same cDNA pool in 665 of 721 (92%) exons. This established that these new exons are highly likely to be present in the cDNA we generated.

Capturing macrophage-specific long noncoding RNA isoforms

In addition to identifying new exons, NNC isoforms were particularly helpful in redefining long noncoding RNA (lncRNA) loci. 7% of NNC isoforms were associated with lncRNA genes as compared with 3% for both FSM and NIC isoforms. This is likely because of the fact that lncRNA are often expressed at lower levels in a more tissue specific manner than protein-coding genes which complicates their comprehensive annotation (26). Our data set enables the investigation of these lncRNAs and their role in macrophage activation.

In addition to NNC isoforms, isoforms falling into the SQANTI "intergenic" category are also likely to represent noncoding transcripts since protein-coding genes have been exhaustively mapped in the human genome. In total, 184 isoforms were categorized as "intergenic" defined as not overlapping any locus in the Gencode (v34) annotation (Table S8). Of these 184 isoforms, 57 contained more than one exon. These 57 multiexon isoforms in turn grouped into distinct 38 loci. Only six of these 38 loci overlapped with putative lncRNA loci assembled from deep short-read data (27). This showed that investigating specific cell types under different treatments has the potential to identify new previously unobserved lncRNA loci.

Enabling easy data exploration

While annotations like GENCODE are indispensable for any genomic experiment, they are insufficient when aiming to experimentally follow-up potential hits from a screen or an RNA-seq experiment. Our data set addresses this by providing information on which of the potentially numerous isoforms present in (or absent from) the Gencode annotation is actually expressed by a gene of interest and at what level it is expressed compared with other isoforms of that gene. To enable this type of exploration of the data set is available as a custom UCSC genome browser session (https://genome.ucsc.edu/s/vollmers/IAMA). This session contains gene expression information, isoform models, and quantification, as well as R2C2 and Smart-seq2 read tracks.

To demonstrate the user interface of the Isoform-level transcriptome Atlas of Macrophage Activation browser session, we highlight an overlapping pair of lncRNA—LINC01181 and BAALC-AS—which both appear to be differentially upregulated by all TLR ligands based on short-read RNA-seq data. Inspecting these overlapping loci in the genome browser shows that only a very small number of R2C2 reads align to BAALC-AS and that short nondirectional RNA-seq reads assigned to BAALC-AS are therefore likely derived from LINC01181 cDNA. Inspecting the isoforms based on these R2C2 reads also shows no match for BAALC-AS but several for LINC01181 (Fig. 5). Of the eight spliced isoforms in the LINC01181 locus, none match but some fully contain the annotated LINC01181 transcript. Hovering over the isoforms shows that Isoform_136603_75 is expressed the highest in the LPS condition, taking up 48% of R2C2 reads in that condition. The reference-corrected sequence of Isoform_136603_75 can then be retrieved for downstream analysis by clicking its model.

Discussion

For investigators interested in the in-depth analysis of a gene's expression and function in a specific cell-type under defined experimental conditions, current efforts to annotate and quantify transcriptomes fall short. This is because of two major limitations of these efforts: (1) the use of short-read RNA-seq methods which precludes the identification and quantification of isoforms and (2) efforts using long-read methods are mostly limited to the analysis of a limited number of cell-types, almost always at baseline, which will miss isoforms specific to a different cell-type and under different experimental conditions.

The data set and exploration options we present here will be of real-world use to researchers investigating human macrophages at both baseline and after TLR activation and could provide a blueprint for future studies combining short and long-read transcriptome analysis.

The analysis of the transcriptome we generated shows that differential isoform expression between conditions exists but is limited and most often associated with the differential usage of TSSs which is similar to observations we have previously made in mouse and human macrophages (9). This shows that the splicing of genes itself is very similar between the conditions we investigated.

We further show that most isoforms we identify match the splice-junction chain of an annotated isoform exactly but often not its TSS and polyA sites. We also detect hundreds of new exons, enriched for first and last exons. The absence of these exons from the Gencode annotation may be caused by technical or biological limitations of previous studies and annotation efforts. Short-read RNA-seq, which is most often used as the foundation for annotation, is characteristically struggling with capturing transcript ends. Further, these new exons might only be expressed in naive or activated macrophages and even then at fairly low levels.

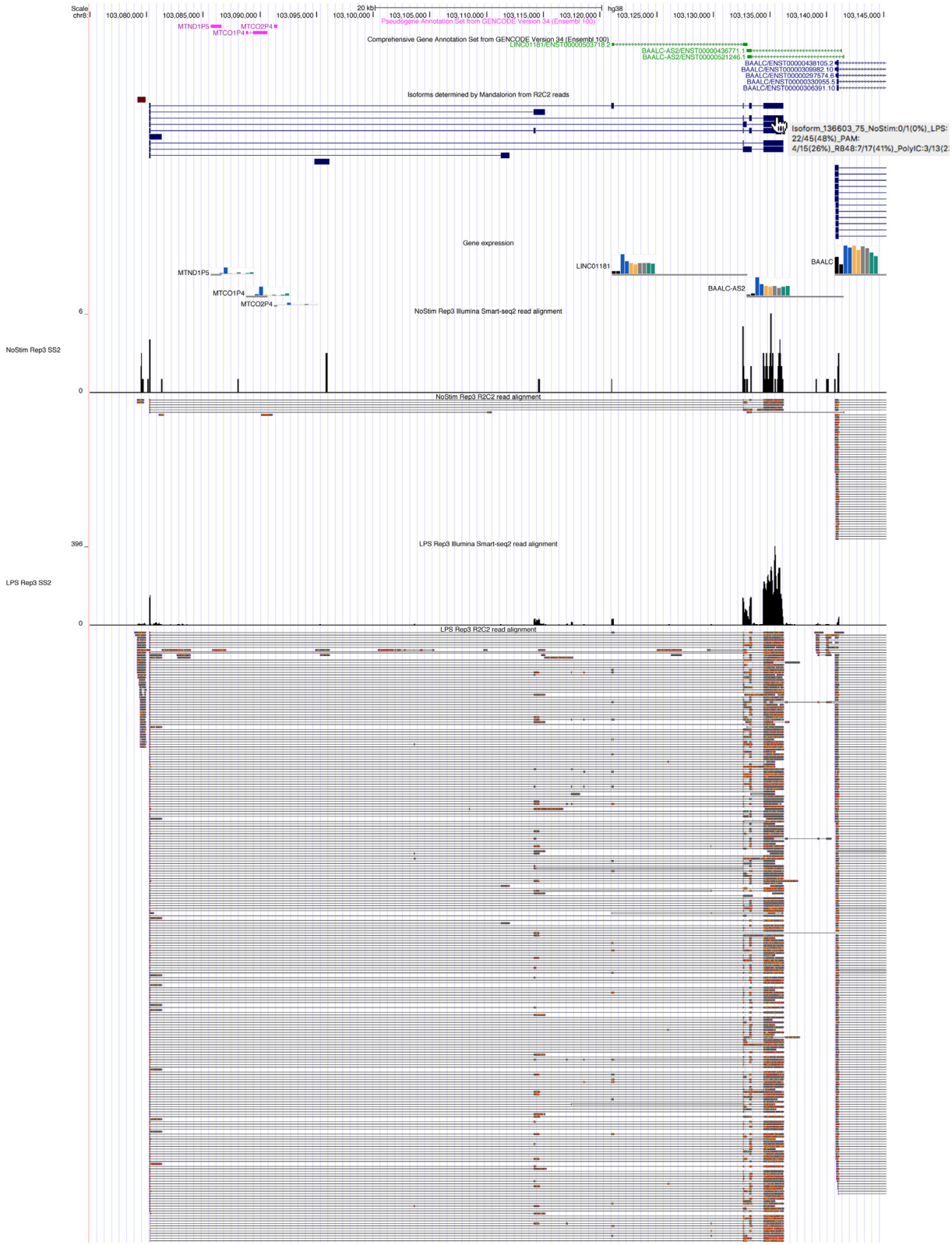

Figure 5. Data exploration. A screenshot of the IAMA session in the UCSC genome browser is shown. From the *top*, GENCODE annotation, Mandalorion Isoforms, Smart-seq2 based gene expression (bar graphs), Smart-seq2 (histogram), and R2C2 reads. Highlighted are Smart-seq2 and R2C2 reads for just one replicate and two conditions to demonstrate the IAMA browser session and for space-saving purposes. The complete IAMA session for both replicates and all TLR-activated conditions are available here (https://genome.ucsc.edu/s/vollmers/IAMA). IAMA, Isoform-level transcriptome Atlas of Macrophage Activation; TLR, toll-like receptor.

Conclusions

While we believe the entire data set presents a unique window into macrophage biology, the main purpose in creating it was to enable researchers to better understand their gene or genes of interest. We hope that researchers that, for example, might be interested in a particular gene expressed in macrophages after LPS activation will select the isoform with the highest expression in that condition from our data set to synthesize its exact sequence for follow-up studies or to, for example, locate its promoter for CRISPR interference experiments.

In most cases, the isoform selected from our data set would be different from an isoform picked arbitrarily from an annotation database. Even if the most abundant isoform of a gene was a FSM isoform, it would likely match one of several annotated isoforms for that gene in the GENCODE annotation and might have different TSS and poly sites than the annotated isoform it is matching. Further, in 1028 genes in this data set, the most abundant isoform was a NIC or NNC isoform which by definition are not present in the GENCODE annotation. Ultimately, the isoform-level transcriptome we generated here should contain valuable information for the vast majority of medium and highly expressed genes in macrophages.

Finally, while long-read full-length cDNA sequencing is still too expensive to replace routine RNA-seq, the additional isoform-level information supplied by methods like Pacific Biosciences sequencing, ONT-based R2C2, or other error-correcting approaches (6) should make it a valuable addition to target identification and characterization workflows.

Experimental procedures

Cell culture

All experiments abided by the Declaration of Helsinki principles and were approved by the Internal Review Board at the University of California Santa Cruz under protocol HS2614. Human buffy coat samples from healthy blood donors provided by the Stanford Blood Center were completely de-identified. Peripheral blood mononuclear cells were isolated from these buffy coat samples by density gradient centrifugation using Ficoll-Paque PLUS (GE Healthcare) followed by 3X washes in HBSS (Sigma Aldrich, H6648), then resuspended in complete RPMI-1640 (Gibco, 11875093) supplemented with 5 ml pen/strep (100X, Gibco, 15140122), 10% FCS (Gibco, 16140-071), 12.5 ml Hepes (1M, Gibco, 15630-080), 5 ml NEAAs (100X, Life Sciences, SH3023801), 5 ml GlutaMax (100X, Gibco, 35050-061), 5 ml Na-Pyruvate (100 mM, Gibco, 11360-070), 500 µl ciprofloxacin (10 mg/ml, Acros, AC456880050) and plated onto 10-cm tissue culture plated dishes. Nonadherent cells were removed after 2 h of incubation at 37 °C in 5% CO2. The remaining cells were expanded and differentiated into macrophages by culturing cells in the presence of recombinant human macrophage colony-stimulating factor (R&D, 216-MC-025, 50 ng/ml). Cells were cultured for 8 days with the replacement of culture medium every 2 days. Cells were stimulated for 6 h at the following concentrations: LPS (200 ng/ml), Pam3CSK4 (200 ng/ml), poly(I:C) (50 µg/ml), and R848 (1 µg/ml). Unstimulated cells were collected at the same time point for use as control.

Sample preparations

RNA extraction

Total RNA was purified from cells using Direct-zol RNA MiniPrep Kit (Zymo Research, R2072) and TRIzol reagent (Ambion, T9424) according to the manufacturer's instructions. RNA was assessed for purity using a nanodrop spectrometer (Thermo Fisher). RNA was quantified using a Qubit Fluorometer (Thermo Fisher) and Qubit RNA HS Assay Kit (Thermo Fisher, Q32852).

cDNA synthesis

For each of these samples, 100 to 200 ng of total RNA was used to generate full-length cDNA using a modified Smart-seq2 protocol (13). RNA was reverse transcribed using Smartscribe RT (Clontech). For each sample, reverse transcription was primed with a different OligodT primer containing 30 Ts, a 10 nt sample index, and a universal ISPCR priming site. The reverse transcription reaction also contained a template switch oligo (TSO-Smart-seq2) to attach the same universal priming site to the 5' end of transcripts. After reverse transcription, RNA and primer dimers were digested using RNAseA and Lambda Exonuclease (NEB) after which cDNA was amplified using the Kapa Biosystems HiFi HotStart ReadyMix (2X) (KAPA) with the following heat-cycling protocol: 37 °C for 30 min, 95 °C for 30 s followed by 12 cycles of (98 °C 20 s; 67 °C 15 s; 72 °C for 6 min). The reaction was then purified using SPRI beads at a 0.65:1 ratio (to retain cDNA longer than 500 bp) and eluted in H2O.

Smart-seq2

For each of the samples we processed, Smart-seq2 libraries were generated as previously described. In short, full-length cDNA was then tagmented with Tn5 enzyme loaded with Tn5ME-A/R and Tn5ME-B/R adapters. The Tn5 reaction was performed using 50 ng of cDNA in 5 µl, 1 µl of the loaded Tn5 enzyme, 10 µl of H2O, and 4 µl of 5× TAPS-PEG buffer and incubated at 55 °C for 5 min. The Tn5 reaction was then inactivated by the addition of 5 µl of 0.2% sodium dodecyl sulphate, and 5 µl of the product was then nick-translated at 72 °C for 6 min and further amplified using KAPA Hifi Polymerase (KAPA) using a distinct set of indexing primers for each sample and an incubation of 98 °C for 30 s, followed by 13 cycles of (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 resulting Illumina library was sequenced on an Illumina NextSeq500 1 × 75 run.

R2C2

Amplified cDNA was sequenced on the ONT MinION sequencer using the R2C2 method (14, 16, 17, 23). In short, 100 ng of cDNA is circularized using 100 ng of a DNA splint

(Table S9) and 2x NEBuilder HiFi DNA Assembly Master Mix (NEB). This mix was incubated at 50 °C for 60 min. Noncircularized cDNA was digested by adding 5 μl of NEBuffer 2, 3 μl Exonuclease I, 3 μl of Exonuclease III, and 3 μl of Lambda Exonuclease (all NEB) and adjusting the volume to 50 μl using H2O. This reaction was then incubated 37 °C for 16 h followed by a heat inactivation step at 80 °C for 20 min. Circularized DNA was then extracted using SPRI beads with a size cutoff to eliminate DNA <500 bp (0.65 beads:1 sample) and eluted in 40 μl of ultrapure H2O. Circularized DNA was split into four aliquots of 10 μl, and each aliquot was amplified in its own 50 μl reaction containing Phi29 polymerase (NEB) and exonuclease resistant random hexamers (Thermo) [5 μl of 10× Phi29 Buffer, 2.5 μl of 10 μM (each) dNTPs, 2.5 μl random hexamers (10 μM), 10 μl of DNA, 29 μl ultrapure water, 1 μl of Phi29]. Reactions were incubated at 30 °C overnight. T7 Endonuclease was added directly to each reaction which was then incubated at 37 °C for 2 h with occasional agitation. The debranched DNA was then size-selected by either using SPRI beads at a 0.5:1 ratio (pilot experiments) or agarose gel extraction.

For the agarose gel extraction, debranched DNA was pooled and concentrated using DNA Clean & Concentrator-5 columns (Zymo Research), and >5 kb DNA was then excised from a 1% DNA low-melt agarose gel. Agarose was then melted at 65 °C for 10 min, transferred to 42 °C, and digested by immediate addition of 2 μl of beta-agaraseI (NEB) per 300 μl of melted gel and incubation at 42 °C for 1 h. Undigested Agarose was then pelleted by centrifugation (14,000 RPM for 7 min in microcentrifuge), and the DNA in the supernatant was extracted using SPRI beads at a 0.7:1 ratio.

The resulting DNA was sequenced on MinION 9.4.1 Flow Cells. For each run, 1 μg of DNA was prepared using the LSK-109 kit according to the manufacturer's instructions with only minor modifications. End-repair and A-tailing steps were both extended from 5 min to 30 min. The final ligation step was also extended to 30 min. Depending on pore status, runs were DNAseI treated and reloaded after 24 or 48 h

RT-qPCR validation

Total RNA was purified from MDMs derived from an additional donor and stimulated with LPS (200 ng/ml) for 6 h. Unstimulated cells were collected at the same time point for use as control. RNA was quantified and assessed for purity using a nanodrop spectrometer (Thermo Fisher). Equal amounts of RNA (500 ng) were reverse-transcribed using iScript Reverse Transcription Supermix (Biorad, 1708841) followed by qPCR using iQ SYBR Green Supermix reagent (Biorad, 1725122) with the following parameters: 50 °C for 2 min, 95 °C for 2 min followed by 40 cycles of 95 °C for 15 s, 60 °C for 30 s, and 72 °C for 45 s. Oligos used in RT-qPCR analysis were designed using Primer3 Input version 0.4.0 (https://bioinfo.ut.ee/primer3-0.4.0/). Primer sequences are provided in Table S9. Gene expression levels were normalized to *Hprt* as a housekeeping gene.

Data analysis

Smart-seq2

Data in demultiplexed fastq files were aligned to the human genome (hg38) using STAR and standard settings. Gene expression was determined using featureCounts (-O -g gene_name -a) and GENCODE v34 (28). Differential gene expression was then determined using DESeq2 (20).

R2C2

Data in Fast5 format were basecalled using the bonito research basecaller (version 0.0.5) (https://github.com/nanoporetech/bonito). The resulting fasta files were converted to fastq files by adding a constant Q15 quality score to each base. To generate R2C2 consensus reads, we processed and demultiplexed the resulting fastq files using our C3POa pipeline (https://github.com/rvolden/C3POa).

To identify and quantify isoforms, we combined data from all samples and used the 3.5 version of Mandalorion (-O 0,40,0,40 -r 0.01 -i 1 -w 1 -n 2 -R 5) (https://github.com/rvolden/Mandalorion) which uses the minimap2 (29), racon (30), Medaka (https://github.com/nanoporetech/medaka), and abpoa (31) tools. Isoforms were categorized using the sqanti_qc.py script of the SQANTI (25). To identify new exons, we used the *ProcessSqantiClassification.py* utility of Mandalorion. To investigate protein-coding potential of isoforms, we translated all three possible reading frames of each isoform using BioPython (32) and checked whether these translations contained a CDS sequence provided by GENCODE (28). For differential isoform expression, we used the Chi-squared contingency test as implemented in SciPy (33) on data excluding the pilot experiments. For further analysis and visualization of data, we used both the Numpy (34) and Matplotlib (35) Python libraries.

Ethics approval and consent to participate

We received fully anonymized human blood products for this study from the Stanford Blood Center (SBC). The SBC consents individuals and collects blood products as approved by the Stanford University IRB.

Data availability

ONT raw sequencing data is available from the Sequence Read Archive (SRA) under Bioproject PRJNA639136. Illumina raw sequencing data is available from the SRA under Bioproject PRJNA660772.

Processed data can also be explored as a UCSC Genome Browser session at https://genome.ucsc.edu/s/vollmers/IAMA.

Supporting information—This article contains supporting information.

Acknowledgments—We thank the Georgia Genomics and Bioinformatics Core (GGBC) for generating and sequencing Smart-seq2 libraries.

Author contributions—A. C. V. and Susan Carpenter conceptualized the study design. C. V. conceptualized and implemented data generation and analysis strategies. A. C. V., H. E. M., and Sophia Campos performed experiments. A. C. V and C. V. analyzed the data. A. C. V, Susan Carpenter, and C. V. wrote and edited the manuscript.

Funding and additional information—We acknowledge funding by the National Institute of General Medical Sciences/National Institutes of Health Grant 1R35GM133569 (to C. V.) and R35GM137801 (to Susan Carpenter), NIH Predoctoral Training Grant (T32GM008646), Ford Predoctoral Fellowship, and the Howard Hughes Medical Institute Gilliam Fellowship (to A. C. V.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of interest—The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: CDS, full coding sequence; DE, differential expression; FSM, full-splice matches; ISM, incomplete splice-matches; lncRNA, long noncoding RNA; LPS, lipopolysaccharide; MDM, monocyte-derived macrophage; NIC, novel in catalog; NNC, novel not in catalog; ONT, Oxford Nanopore Technologies; PAM, Pam3CSK4; RPM, reads per million; TLR, toll-like receptor; TSS, transcription start site.

References

1. 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
2. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., *et al.* (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477
3. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652
4. Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E., and Tilgner, H. U. (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4259
5. Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., Zuzarte, P. C., Gilpatrick, T., Payne, A., Quick, J., Sadowski, N., Holmes, N., de Jesus, J. G., Jones, K. L., Soulette, C. M., *et al.* (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305
6. Lebrigand, K., Magnone, V., Barbry, P., and Waldmann, R. (2020) High throughput error corrected nanopore single cell transcriptome sequencing. *Nat. Commun.* **11**, 4025
7. [preprint] Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Zeng, W., Williams, B., Trout, D., England, W., Chu, S., Spitale, R. C., Tenner, A., Wold, B., and Mortazavi, A. (2019) A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*. https://doi.org/10.1101/672931
8. Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., DuBois, R. M., Forsberg, E. C., Akeson, M., and Vollmers, C. (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027
9. [preprint] Robinson, E. K., Jagannatha, P., Covarrubias, S., Cattle, M., Safavi, R., Song, R., Viswanathan, K., Shapleigh, B., Abu-Shumays, R., Jain, M., Cloonan, S. M., Wakeland, E., Akeson, M., Brooks, A. N., and Carpenter, S. (2020) Inflammation drives alternative first exon usage to regulate immune genes including a novel iron regulated isoform of Aim2. *bioRxiv*. https://doi.org/10.1101/2020.07.06.190330
10. Medzhitov, R., and Horng, T. (2009) Transcriptional control of the inflammatory response. *Nat. Rev. Immunol.* **9**, 692–703
11. Carpenter, S., Aiello, D., Atianand, M. K., Ricci, E. P., Gandhi, P., Hall, L. L., Byron, M., Monks, B., Henry-Bezy, M., Lawrence, J. B., O'Neill, L. A. J., Moore, M. J., Caffrey, D. R., and Fitzgerald, K. A. (2013) A long noncoding RNA mediates both activation and repression of immune response genes. *Science* **341**, 789–792
12. Kawai, T., and Akira, S. (2008) Toll-like receptor and RIG-I-like receptor signaling. *Ann. N. Y. Acad. Sci.* **1143**, 1–20
13. 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
14. Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R. J., Green, R. E., and Vollmers, C. (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**, 9726–9731
15. Picelli, S., Björklund, A. K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014) Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040
16. Cole, C., Byrne, A., Adams, M., Volden, R., and Vollmers, C. (2020) Complete characterization of the human immune cell transcriptome using accurate full-length cDNA sequencing. *Genome Res.* **30**, 589–601
17. Byrne, A., Supple, M. A., Volden, R., Laidre, K. L., 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
18. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) Star: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
19. Liao, Y., Smyth, G. K., and Shi, W. (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930
20. Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550
21. Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003) PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141
22. Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P. D. (2010) PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* **38**, D204–D210
23. [preprint] Volden, R., and Vollmers, C. (2020) Highly multiplexed single-cell full-length cDNA sequencing of human immune cells with 10X genomics and R2C2. *bioRxiv*. https://doi.org/10.1101/2020.01.10.902361
24. Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., *et al.* (2020) Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053
25. Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., *et al.* (2018) SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* **28**, 396–411
26. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789

27. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927
28. Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., *et al.* (2012) Gencode: The reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774
29. Li, H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100
30. 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
31. [preprint] Gao, Y., Liu, Y., Ma, Y., Liu, B., Wang, Y., and Xing, Y. (2020) abPOA: An SIMD-based C library for fast partial order alignment using adaptive band. *bioRxiv*. https://doi.org/10.1101/2020.05.07.083196
32. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423
33. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., *et al.* (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272
34. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., *et al.* (2020) Array programming with NumPy. *Nature* **585**, 357–362
35. Hunter, J. D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95