

Widespread intronic polyadenylation diversifies immune cell transcriptomes

Irtisha Singh^{1,2}, Shih-Han Lee³, Mehmet K. Samur⁴, Yu-Tzu Tai⁴, Nikhil C. Munshi⁴, Christine Mayr^{3#} and Christina S. Leslie^{1#}

¹Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY

²Tri-I Program in Computational Biology and Medicine, Weill Cornell Graduate College, New York, NY

³Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, NY

⁴Lebow Institute of Myeloma Therapeutics and Jerome Lipper Multiple Myeloma Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA

#Correspondence: cleslie@cbio.mskcc.org, mayrc@mskcc.org

Christina S. Leslie

Memorial Sloan Kettering Cancer Center

1275 York Ave

New York, NY 10065

Phone: 646-888-2762

Christine Mayr

Memorial Sloan Kettering Cancer Center

1275 York Ave

New York, NY 10065

Phone: 646-888-3115

Abstract

Alternative cleavage and polyadenylation (ApA) can generate mRNA isoforms with differences in 3'UTR length without changing the coding region (CDR). However, ApA can also recognize intronic polyadenylation (IpA) signals to generate transcripts that lose part or all of the CDR. We analyzed 46 3'-seq and RNA-seq profiles from normal human tissues, primary immune cells, and multiple myeloma (MM) samples and created an atlas of 4,927 high confidence IpA events. Up to 16% of expressed genes in immune cells generate IpA isoforms, a majority of which are differentially used during B cell development or in different cellular environments, while MM cells display a striking loss of IpA isoforms expressed in plasma cells, their cell type of origin. IpA events can lead to truncated proteins lacking C-terminal functional domains. This can mimic ectodomain shedding through loss of transmembrane domains or alter the binding specificity of proteins with DNA-binding or protein-protein interaction domains, thus contributing to diversification of the transcriptome.

Introduction

ApA is generally viewed as the selection of ApA signals in the 3'UTR, leading to the expression of different 3'UTR isoforms that code for the same protein. Recent computational analyses of 3'-end sequencing data have characterized the nature and extent of ApA in mammalian 3'UTRs^{1, 2, 3, 4, 5, 6, 7}. For example, analysis of the first human ApA tissue atlas established that half of human genes express multiple 3'UTRs, enabling tissue-specific post-transcriptional regulation of ubiquitously expressed genes¹. However, ApA events can also occur in introns rather than 3'UTRs, generating either non-coding transcripts or transcripts with truncated coding regions that lead to loss of C-terminal domains in the protein product.

The most famous example of cell-type specific usage of an IpA signal occurs in the immunoglobulin M heavy chain (*IGHM*) locus^{8, 9}. In mature B cells, recognition of the polyA signal in the 3'UTR produces the full-length message, including two terminal exons that encode the transmembrane domain of the plasma membrane-bound form of IgM (**Fig. 1a**). In plasma cells, usage of an IpA signal instead results in expression of an IpA isoform lacking these two terminal exons, leading to loss of the transmembrane domain and secretion of IgM antibody. Many additional IpA-generated truncated proteins have been described,^{10, 11} including the soluble forms of EGF and FGF receptors and a truncated version of the transcription factor NFI-B¹². One of the first examples of IpA was the generation of two isoforms of the interferon-induced anti-viral enzyme OAS1¹³. Whereas the IpA and the full-length mRNA transcripts encode enzymes with comparable enzymatic activity, the shorter transcript generates a hydrophobic C-terminus and the longer transcript an acidic C-terminus. This suggests that the two isoforms may interact with different co-factors or cellular structures¹³. Other examples include the transcription factor SREPF, whose IpA isoform can act as a developmental switch during spermatogenesis¹⁴.

In the splicing literature, isoforms generated through recognition of an IpA signal are often described as 'alternative last exon' events¹⁵. It is thought that genes that generate IpA isoforms harbor competing splicing and polyA signals. When splicing outcompetes

polyadenylation, a full-length mRNA is generated, and otherwise a truncated mRNA is made¹⁶. As the defining event is the recognition of an lpA signal, we call these transcripts lpA isoforms. Only through the analysis of 3' end sequencing data has it been possible to recognize the widespread expression of lpA isoforms, and here we present a systematic analysis of lpA isoforms across diverse cell types.

We adapted our previous computational analysis pipeline for 3'-seq to identify robust ApA events that occur in introns and to quantify expression of lpA isoforms across human tissues, immune cells, and in MM patient samples. We focused on immune cells because it is feasible to obtain pure cell populations of primary cells and because in our previous tissue atlas we found B cells to express the largest number of differential 3'UTR isoforms¹. Through integration with RNA-seq profiles in B lineage and MM cells as well as external data sets and annotation databases, we assembled an atlas of confident lpA isoforms that are either supported by independent data sources or are very highly expressed in at least one cell type in our data set. We found that lpA isoforms are widely expressed, most prevalently in blood-derived immune cells, and that generation of lpA isoforms is regulated during B cell development, between cellular environments and in cancer. lpA events in immune cells are enriched at the start of the transcription unit, leading to lpA isoforms that retain none or little of the CDR and hence represent a novel class of robustly expressed non-coding transcripts. The majority of lpA events that occur later in transcription units can lead to truncated proteins often lacking repeated C-terminal functional domain, and thus contribute to the diversification of the transcriptome.

Results

Computational analysis of 3'-seq reveals widespread intronic polyadenylation

To assemble an atlas of lpA isoforms in human tissues and immune cells, we used our previously published 3'-seq data set from normal human tissues (ovary, brain, breast, skeletal muscle, testis), cell types (embryonic stem (ES) cells, naïve B cells from peripheral blood (blood NB)), and cell lines¹ and combined it with a newly generated data set from normal and malignant primary immune cells. The new immune cell profiles

(n = 29) were all performed with biological replicates and included lymphoid tissue-derived naïve B cells (NB), memory B cells (MemB), germinal center B cells (GCB) and CD5+ B cells (CD5+B), blood T cells and plasma cells (PC) and MM derived from bone marrow aspirates (**Supplementary Tables 1, 2**). We adapted our previously described computational pipeline to process 3'-seq libraries and detect and quantify ApA events, including intronic as well as 3'UTR events, while removing technical artifacts (see **Methods**)¹. All subsequent analyses were restricted to protein coding genes. For additional evidence in support of lpA isoforms, we performed RNA-seq profiling in the same normal and malignant B cell types, where possible for the same samples (**Supplementary Table 3**).

We confirmed from both 3'-seq and RNA-seq data that the lpA isoform of *IGHM* is highly expressed in PC while the full-length transcript, encoding membrane-bound IgM, is the dominant isoform in NB cells (**Fig. 1b**). Analysis of 3'-seq also revealed novel putative lpA isoforms, including in the locus of *GTF2H1*, encoding a subunit of general transcription factor II H, and *RAB10*, encoding a member of the Ras oncogene family of small GTPases (**Fig. 1c**). Like 3'UTR isoforms, lpA isoforms display differential expression across tissues and cell types. For example, the lpA isoform of *GTF2H1* is well expressed in skeletal muscle and all immune cell types assayed, and indeed the lpA isoform is the only isoform expressed from this gene in PC, blood NB and T cells; these three cell types are also the only ones to express the lpA isoform of *RAB10*. To validate transcriptome-wide the lpA events identified by 3'-seq, we used RNA-seq data from the same cell types. We expect intronic read coverage upstream but not downstream of the lpA event, as is visible in RNA-seq coverage in PCs flanking the intronic 3'-seq peak in *GTF2H1* (**Fig. 1d**). Formally, we can test if RNA-seq read counts are significantly higher in intronic windows chosen upstream compared to downstream of lpA events (see **Methods**)¹⁷. Significantly differential coverage could be confirmed for 29% (n = 1,670) of lpA events from our 3'-seq peak calls (FDR-adjusted $p < 0.1$), whereas almost no significant read count differences were found relative to randomly chosen positions in introns (**Supplementary Fig. 1a**). To assemble an atlas of highly confident lpA events for further analysis, we compared each intronic peak detected in

the 3'-seq data against external annotation and data sources to find additional evidence in support of the lpa isoform (see **Methods**, **Supplementary Fig. 1b, c**). Briefly, lpa events that overlapped with the last exon of annotated isoforms in RefSeq, UCSC and Ensembl were first added to the atlas (2,241 events); unannotated lpa events that satisfied the test for differential upstream vs. downstream RNA-seq coverage were added next (907 events); unannotated lpa events without differential RNA-seq coverage but supported in data sets from other 3' end sequencing protocols were then added to the atlas (1,332 events)¹⁸. We then added lpa events that lacked the previous sources of evidence but had RNA-seq support of the cleavage event – i.e. reads overlapping untemplated adenosines in the polyA tail (124 events). Finally, events with high expression in at least one cell type were also included in the atlas (323 events). 13% (n = 743) of lpa events could not be validated by any of these methods and thus were not included in the atlas for further analysis (**Supplementary Fig. 1c**). Overall, the atlas contains 4,927 confident lpa events in 3,431 protein coding genes, and 55% of atlas events are unannotated in RefSeq, UCSC and Ensembl (**Supplementary Fig. 1c**). We also found that similar proportions of annotated and unannotated lpa isoforms were validated by various kinds of supporting evidence (**Fig. 1e**); for example, only a slightly larger fraction of annotated vs. unannotated lpa events are supported by RNA-seq coverage (34% vs. 26%).

lpa isoforms are robustly expressed in circulating immune cells

Using our atlas of lpa isoforms, we determined the prevalence of lpa across normal tissues and cell types by computing the fraction of genes expressing at least one lpa isoform out of all expressed genes in each cell type (**Fig. 1f**). Blood T cells had the highest fraction of genes with lpa isoforms (0.16) while ovary had the lowest fraction (0.04). Whereas 6-16% of genes expressed in immune cells generate lpa isoforms, complex tissues produce lpa isoforms in only 4-8% genes. Notably, blood NB cells expressed 1,114 lpa isoforms compared to only 721 lpa isoforms for tissue-derived NB cells, suggesting that the cellular environment has a strong effect on lpa isoform expression.

lpA isoforms are robustly expressed, with median expression of the same order of magnitude as the median expression of full-length isoforms (\log_2 TPM of 3.71-3.83 for lpA isoforms in PC, blood NB, and blood T as compared to \log_2 TPM of 4.53-5.15 for full-length transcripts; **Fig. 1g**). Therefore, lpA isoforms are not ‘transcriptional noise’ produced from recognition of ‘cryptic’ sites, but rather represent major mRNA isoforms generated from alternative mRNA processing.

Figure 1h shows the tissue-specific expression of lpA isoforms. The mean expression level of the lpA isoform across the replicates had to be more than 5 TPM to be considered as an expressed lpA isoform. The heatmap shows that a majority of the lpA isoforms with reproducible expression patterns are expressed in immune cell types ($n = 3,365$), and almost all of these are expressed in at least two immune cell types. Non-immune tissues like testis and ES cells express tissue-specific lpA isoforms, but the majority of these isoforms are produced from tissue-specific genes.

Cell types with frequent lpA retain less coding region in lpA isoforms and express shorter 3'UTRs

To begin to assess the impact of lpA isoform expression, we computed the fraction of retained CDR for each lpA isoform based on the position of the lpA event relative to the full-length annotated CDR. The histogram of retained CDR fraction across the lpA atlas showed a uniform distribution across the CDR except for a substantial overrepresentation of lpA isoforms that lose all or almost all of the CDR (**Fig. 2a**). However, an examination of histograms of retained CDR fraction across individual tissues and cell types revealed a more nuanced picture (**Fig. 2b**), where lpA events near the start of the transcription unit dominate in blood and bone marrow-derived immune cells, while brain and ES cells preferentially generate lpA events close to the end of transcription units. In testis and tissue-derived B cells, we found an intermediate pattern.

We observed a significant negative correlation across tissues between the frequency of lpA isoform expression and length of retained CDR ($r = -0.86$, **Fig. 2c**). We further

observed that cell types with a tendency to produce longer 3'UTRs also prefer to produce lpA isoforms that are located at the 3' ends of transcription units ($r = 0.60$, **Fig. 2d**).

We use the term 5'lpA for lpA isoforms that retain less than 25% of the CDR and 3'lpA for the remainder. Both 5'lpA and 3'lpA events occur in introns that are significantly longer than the introns from the same genes that contain no lpA events or from genes that only express full-length transcripts (one-sided Wilcoxon rank-sum test, $p < 10^{-20}$ for all three comparisons, **Fig. 2e**)¹⁹. Similarly, 5'lpA and 3'lpA isoforms are expressed from significantly longer transcription units than genes that only express full-length transcripts (one-sided Wilcoxon rank-sum test, $p < 10^{-20}$ for both comparisons, **Fig. 2f**). 3'lpA atlas events have higher conservation by PhastCons in the sequence surrounding the polyA signal compared to 5'lpA atlas events (one-sided Wilcoxon signed-rank test, $p < 10^{-66}$; **Fig. 2g**); however, 5'lpA events still show higher conservation than randomly chosen intronic polyA signals with no 3'-seq coverage (one-sided Wilcoxon signed-rank test, $p < 10^{-68}$; **Fig. 2g**)²⁰.

Previously, U1 snRNP expression and the presence of U1 snRNP motifs early in the transcription unit were found to play a crucial role in preventing premature cleavage and polyadenylation^{21, 22}. Consistent with these observations, we found that genes that express lpA isoforms contain a higher frequency of polyA signals within their transcription unit and are depleted for U1 snRNP signals, as compared to genes that only express 3'UTR isoforms (**Fig. 2h, i**). Therefore, genomic architecture and sequence composition may facilitate lpA isoform expression.

lpA isoform expression is associated with moderate downregulation of full-length mRNAs

Next we used a generalized linear model (GLM) approach to determine significant changes in the relative expression of lpA isoforms compared to full-length transcripts (usage of lpA) across normal immune cells (see **Methods**)^{1, 17, 23}. The usage of the majority of expressed lpA isoforms differed significantly when we compared NB cells

from lymphoid tissue and blood T cells (950/1308, **Fig. 3a**, FDR-adjusted $p < 0.05$). Within the B cell lineage, we also found differential usage of lpA between cell types, with PC in particular showing strikingly increased usage of lpA sites compared to tissue-derived NB cells (**Fig. 3a**). However, surprisingly, when we compared NB cells from lymphoid tissue and blood, we found even more significant changes in lpA usage (720/1113) than we observed between different B cell types (**Fig. 3a, b**). This indicates that lpA isoform expression is not only cell type-differential but also highly dynamic, changing between different cellular environments. Genes with differential usage of lpA isoforms between immune cell types were most strongly enriched for annotations including zinc finger domains, bromodomains, and the ubiquitin-like conjugation pathway (**Fig. 3c**)²⁴.

We then asked if cells might use lpA signals in a switch-like fashion to ‘turn off’ expression of the full-length transcript by ‘turning on’ the expression of lpA isoform. In **Fig. 3d**, we plot the expression change of the lpA isoform against that of the full-length transcript and show lpA genes that differentially increase (red points) or decrease (blue points) usage of their lpA isoforms in blood-derived compared with tissue-derived NB cells. If a gene had multiple lpA isoforms, then the one with the most significant differential lpA usage is shown. Genes that increase the usage of the lpA isoform in blood- versus tissue-derived NB cells significantly reduced expression of their full-length transcripts compared to genes without significant change in lpA usage (**Fig. 3e**, one sided KS test, $p < 10^{-5}$). The decrease in full-length isoform expression was modest, but significant, indicating that lpA usage does not predominantly result in a ‘switch-like’ change between full-length and lpA isoform expression.

lpA diversifies the transcriptome through loss of C-terminal functional domains

Next, we investigated the potential functional consequences of lpA. We observed that lpA genes encode full-length proteins that are significantly larger and contain more domains than genes that do not produce lpA isoforms (**Fig. 4a**, median number of amino acids 588 vs. 432; **Fig. 4b**, median 5 vs. 4 domains). Notably, most lpA-generated truncated proteins still retain functional protein domains, suggesting that lpA

may contribute to the diversification of the transcriptome (**Fig. 4b**, median 2 domains). IpA genes preferentially encode proteins with RNA- or DNA-binding or protein-protein interaction (PPI) domains, but they avoid membrane proteins. Proteins encoded by IpA genes are also enriched in repeated domains (**Supplementary Fig. 2a, Fig. 4c**), and in a majority of these we observed that IpA results in partial loss of the repeated domains. For example, the full-length protein encoded by *NFKBID* has six ankyrin domains, while the IpA-generated truncated protein retains four of them. Similarly, the full-length protein of the transcription factor PATZ1 has seven zinc finger domains, while different IpA isoforms are predicted to result in proteins with either four or five zinc fingers (**Fig. 4d**). The partial loss of DNA-binding domains has the potential to change DNA-binding specificity and therefore the set of target genes regulated by transcription factors. Similarly, the partial loss of PPIs can change the binding affinity to interaction partners of a protein. Taken together, these observations support the hypothesis that IpA contributes to diversification of the transcriptome and proteome.

We next investigated if there are specific protein domains that are preferentially lost or retained through IpA. Within the group of genes with a single IpA event and whose IpA isoform retains least one protein domain ($n = 1,405$), we found that IpA results in a preferential loss of DNA-binding or PPI domains but avoids the loss of active sites (**Supplementary Fig. 2b, c** and see **Methods**). Active sites of enzymes are the regions where substrate binding and catalysis take place. Loss of an active site would make an enzyme dysfunctional, but IpA appears to avoid this outcome. IpA genes encode diverse proteins with enzymatic functions, including protein kinases, DNA or RNA helicases or motor proteins, as shown in **Fig. 4e**. While the active sites are retained in the IpA-generated truncated proteins, these enzymes lose PPI domains, which may enable the enzyme to participate in different protein complexes or to change the substrate. For example, the protein kinase RIPK1 exists as full-length protein containing a C-terminal death domain, which is not included in RIPK1 IpA (**Fig. 4e**). BAZ1B, also called WSTF, is a multi-functional protein that contains an N-terminal protein kinase domain but has the option to also include C-terminal located coiled-coil, zinc finger and

bromodomains. Also, helicases, including DDX21, DDX49 and DHX15, as well as motor proteins such as KIF20B retain their enzymatic function but generate proteins lacking interaction domains.

Membrane proteins are characterized by the presence of transmembrane domains (TMDs) and are significantly depleted among lpa genes (**Fig. 4c, Supplementary Fig. 2a**). However, we still found 673 lpa isoforms from 499 genes that encode transmembrane proteins and retain at least one protein domain. Among them, 207 lpa isoforms from 152 genes completely retained their TMDs, whereas 220 lpa isoforms from 175 genes lost their TMDs. Interestingly, lpa isoforms that retain the TMDs often encode intracellular membrane proteins that localize to mitochondria. In contrast, lpa isoforms that lose their TMDs are significantly enriched in signal peptides that are predominantly present in plasma membrane proteins (FDR, $p < 9.1 \times 10^{-29}$, **Fig. 4f, see Methods**). Many of them encode cytokine receptors, integrins, or growth factor receptors. Notably, regardless of the position of the TMD, the truncated protein generated by lpa usually terminates immediately before the TMD (**Fig. 4g**). As all of these candidates contain signal peptides at the N-terminus, the lpa isoform produces a secreted form of the cytokine or growth factor receptor.

5'lpa can produce robustly expressed non-coding RNAs

A large fraction of lpa isoforms that are differentially used in at least one pairwise comparison among normal immune cell types are in fact 5'lpa isoforms (487 out of 1,281). Through *de novo* RNA-seq assembly, we were able to resolve the transcript structure for 954 of the 5'lpa isoforms in our atlas (see **Methods**). Using the transcript structure we found that 469 of these have low predicted coding potential with open reading frames that are predicted to encode less than 100 amino acids²⁵. Therefore, they are likely to either generate micropeptides or represent non-coding RNAs (**Fig. 5, Supplementary Fig. 3**)^{26, 27, 28, 29}. To assess potential functional consequences of expression of non-coding transcripts, we examined if RNA-binding proteins may preferentially bind to the exonized intronic sequences upstream of the lpa cleavage site.

As can be seen for the examples shown in **Fig. 5**, the new exons are enriched for CLIP-seq peaks for RNA-binding proteins such as FUS, ELAVL1, PUM2, TAF15, and TIAL1 (binomial $Z > 10$, see **Methods**), which are typically enriched in the 3'UTRs of coding transcripts. As the newly exonized intronic sequences did not bind RNA-binding proteins usually bound to introns, our analysis supports the exonic nature of the predicted non-coding transcripts.

Multiple myeloma displays a widespread loss of plasma cell lpa isoforms

As alternative 3'UTR isoform expression can be altered in cancer cells^{6, 30, 31}, we investigated whether lpa is also dysregulated in cancer. Since PCs express the highest number of lpa isoforms among the tissue-derived B cells, we compared lpa isoform expression between normal and malignant PCs, derived from MM patients ($n = 15$). As MM is a heterogeneous disease, we used hierarchical clustering based on lpa isoform expression to define three patient subgroups (**Supplementary Fig. 4 and Table 2**). We then performed GLM modeling as described above to determine the differential relative expression of lpa isoforms versus full-length isoforms for each MM group compared to normal PCs. Whereas one patient group had an lpa profile comparable to normal PCs, two MM patient groups showed widespread loss of usage of PC lpa events (groups 1 and 2, **Fig. 6a**). We found that 44% of all PC-expressed lpa isoforms (480/1088) are lost in at least one patient group, while only 15 lpa sites show increased usage (FDR-adjusted $p < 0.05$). The significant events in patient group 1 largely represent a superset of those in group 2 (**Fig. 6b**).

Loss of lpa isoform expression in patient group 1 resulted in a significant increase of full-length mRNA expression (**Fig. 6c, d**). As with differential lpa site usage between normal cell types, the genes that display differential relative lpa isoform expression in MM versus PC are enriched for annotations such as bromodomain, transcriptional regulation, and ubiquitin-like conjugation pathway (data not shown). In the majority of patient samples profiled (11 out of 15), the MM transcriptome is characterized by the loss of 480 lpa isoforms that are normally expressed in PCs. This is in contrast to

3'UTR regulation, where we found shortening of 3'UTRs in 126 and lengthening of 3'UTRs in 215 multi-UTR genes (MM group 1; data not shown).

Interestingly, one of the genes that displays loss of lpA isoform expression is the transcription factor *IKZF1*, a key gene in MM biology and also the target of lenalidomide, a recent MM therapeutic (**Fig. 6e**)³². The lpA isoform of *IKZF1* results in the loss of all zinc finger domains encoded by the full-length isoform, potentially leading to expression of a dysfunctional truncated protein isoform as it only contains 53 amino acids and no known domain. While the lpA isoform is the dominant isoform in PCs, with only minimal expression of the full-length transcript, in MM group 1 patients, expression of the lpA isoform is almost completely lost, and the full-length transcript is instead aberrantly expressed. Similarly, the gene encoding *IQGAP1*, a GTPase-activating scaffold protein that plays a role in cell proliferation in MM, largely loses expression of its lpA isoform in MM³³. This isoform, which lacks its Ras-GTP domain as well as most functional domains, is either non-coding or at best produces a truncated protein with only a fraction of the N-terminal calponin homology domain, an actin-binding domain (**Fig. 6e**). Finally, in PCs, *GSK3B* predominantly expresses an lpA isoform that truncates the kinase domain and loses the catalytic site. In MM, full-length transcript expression is restored, presumably rescuing GSK3B-mediated signaling. This may contribute to MM biology as GSK3B functions as a pro-survival factor in MM^{34, 35, 36}.

Discussion

We used 3'-seq and RNA-seq analyses to demonstrate the widespread expression of lpA isoforms across human tissues. lpA isoform expression has been observed and experimentally validated previously but has primarily been viewed as a form of alternative splicing involving alternative last exon usage^{15, 37}. However, as the defining event is the usage of an intronic alternative polyA signal, we instead use the term lpA. We performed comprehensive lpA analyses using 46 samples and identified 4,927 high-confidence lpA events. The majority of the lpA events described here have not been annotated thus far. Our study showed that lpA is unexpectedly widespread and especially common among normal human immune cells. As lpA isoforms are often

highly expressed, they represent a normal component of the expression program in human cells and are therefore not 'cryptic' or 'transcriptional noise'. Instead, their expression is regulated across normal cells and dysregulated in cancer.

The widespread nature of lpA isoform expression has escaped attention thus far, as RNA-seq analysis alone is unable to accurately identify mRNA 3' ends. However, combining 3'-seq and RNA-seq analyses enabled us to resolve transcript structure and to identify hundreds of new non-coding RNAs as well as truncated mRNAs that are predicted to generate proteins with alternative C-termini. The lpA-generated truncated mRNAs are expressed at the same order of magnitude as full-length mRNAs and are not subject to degradation by nonsense mediated decay, since their stop codons are not premature as they are followed by conventional mRNA 3' ends.

It is currently thought that the expression of lpA isoforms is regulated by a competition between splicing and cleavage-polyadenylation reactions^{10, 16}. Consistent with this model, we found that lpA genes have distinct structural and sequence properties compared to genes that generate only full-length transcripts that may predispose them toward lpA recognition. In particular, lpA genes have longer introns, longer transcription units, an enrichment of polyA signals, and a depletion of U1 snRNP signals relative to genes that only express full-length transcripts^{21, 22}. Nevertheless, the tissue-specific differential expression of many lpA isoforms also suggests more complex regulation of the production or stability of these transcripts. It is possible that degradation factors, including components of the RNA exosome, are downregulated in immune cells, resulting in a more frequent occurrence of lpA³⁸. Alternatively, there may be an expression change of splicing factors, such as hnRNP C and U2AF65, which have been associated with the regulation of Alu exonization, one of the mechanisms known to control the expression of intronic exons^{39, 40}. The fact that more prevalent use of lpA signals, shorter lpA isoforms, and shorter 3'UTRs are all correlated across tissues

suggests that the abundance of the same global co-transcriptional factors may be partially responsible for all three properties.

The finding that long introns and long transcription units are more susceptible to lpA suggests that the processivity of the co-transcriptional machinery may also play a role in lpA expression. Intron retention has also been observed as a prevalent feature of blood cell transcriptomes^{41, 42}. If we view splicing and cleavage-polyadenylation as competing processes carried out by different co-transcriptional complexes, the tendency to retain certain introns through intron retention may provide the polyadenylation machinery time to recognize lpA signals for cleavage and 3' end processing. Interestingly, we did find a statistically significant co-occurrence of introns with lpA events and retained introns (**Supplementary Fig. 5a**). In particular, prevalence of intron retention correlates with prevalence of lpA across cell types (**Supplementary Fig. 5b**) and retained introns are enriched for lpA in each cell type examined (**Supplementary Fig. 5c**). However, lpA events in introns with no evidence of intron retention have higher lpA site usage than those in retained introns (**Supplementary Fig. 5d**). Therefore, it is unclear from our data whether intron retention is a necessary mRNA processing step prior to lpA recognition and 3' end formation, or whether lpA recognition can occur independently of intron recognition.

One of the surprising findings of our study was an enrichment of lpA isoforms located at the 5' end of the transcription unit that predominantly occur in immune cells. We observed 378 robustly expressed lpA isoforms that occurred in introns in 5'UTRs and thus are predicted to be non-coding transcripts. Using less stringent criteria and allowing the generation of 100 amino acids, we identified 469 of 5'lpA isoforms. These lpA isoforms are either non-coding or represent a novel source of micropeptides^{26, 27, 28, 29}. The cellular function of non-coding RNAs generated through lpA is unclear. There are reports of promoter-associated RNAs that initiate upstream of transcription start sites and that regulate transcript expression through RNA interference or interaction

with epigenetic modifying enzymes^{43, 44, 45, 46}. However, in our matching RNA-seq data, we did not find read evidence upstream of transcription start sites indicating that the predicted non-coding RNAs that we observed in our study originate at the annotated transcription start sites. Additionally, we demonstrated through analysis of CLIP sequencing data that the exonized intronic sequence of the 5'lpA isoforms contains binding sites for RBPs; potentially, these non-coding RNAs serve as scaffolds for RBPs and thereby exert a regulatory role in *trans* on other RNAs.

The vast majority of lpA isoforms (n = 2,667), however, are predicted to generate truncated proteins that retain at least one domain. Notably, lpA genes encode larger proteins that contain significantly more domains than proteins generated from non-lpA genes. As the lpA-generated truncated proteins still retain a median number of two domains, the majority of proteins have the potential to be functional, thus suggesting that lpA isoforms are an important source of diversification of the transcriptome. This is supported by a significant enrichment of repeated protein domains among the proteins encoded by lpA genes. Strikingly, in the majority of cases, the repeated domains are only partially lost, thus modulating but not losing overall protein function. lpA-induced transcriptome diversification is also supported by the finding that the active sites of enzymes are retained in lpA-generated truncated proteins. Again in these cases, lpA results in proteins with similar function as the full-length proteins but different affinity or different binding partners. This suggests that the cell-type specific expression of truncated proteins generated through lpA may be a widely used mechanism to diversify the proteome, not a peculiarity of a few well-known examples like IgM.

lpA can also generate C-terminal truncations of membrane proteins. Although lpA genes avoid membrane proteins overall, lpA can mimic ectodomain shedding in transmembrane proteins. For example, metalloproteinases such as ADAM10 or ADAM17 are known to release the ectodomains of several surface receptors, including TNF α , L-selectin, TGF α or CD40^{47, 48}. For several of these molecules, the soluble

ligands act as agonists or antagonists of the membrane-bound ligands. In particular, while membrane-bound Fas ligand kills T lymphocytes, soluble Fas ligand blocks this activity⁴⁹. In the vast majority of the investigated cases, proteolytic cleavage occurs close to the plasma membrane, cutting at a site near the TMD, thus releasing the extracellular domain of the growth factor receptor or cytokine. Intriguingly, we found that IpA is another mechanism to produce soluble versions of membrane-bound receptors – very similar to the soluble ligands generated through proteolytic cleavage – as IpA-generated truncations also occur close to the TMD. This demonstrates that developmental regulation of membrane-bound versus secreted molecules – which was first described for IgM – is widespread and can mimic proteolytic cleavage carried out by proteases. The soluble proteins generated through IpA include CD274 IPA, encoding a truncated PD-L1. PD-L1 is an important immune checkpoint receptor ligand, and serum levels of soluble PD-L1 correlate with poor prognosis in B cell lymphomas⁵⁰. IpA also generates a soluble tumor necrosis factor receptor 1 (TNFR1), which has been shown to block TNF activity and is associated with multiple sclerosis⁵¹.

As IpA isoform expression changes in different stages of B cell development as well as after environmental changes, the balance of membrane-bound versus soluble forms can change for many cytokine- or surface receptors. Importantly, IpA isoform expression is not only dynamic during development but is also dysregulated in cancer. A majority of the MM patients that we profiled showed a striking loss of IpA isoforms normally expressed in PCs. Thus, it seems that hundreds of genes avoid the generation of full-length proteins in PCs and the majority of them are re-expressed in MM samples. One gene that displays a switch-like loss of IpA isoform expression and rescue of full-length transcript expression in MM is *IKZF1*, the gene encoding IKAROS, a transcription factor and chromatin remodeler and a key therapeutic target in this malignancy³². Another gene with biological relevance for MM is GSK3B which has pro-survival activity in MM^{34, 36}. Whereas PCs express the full-length mRNA as well as the IpA isoform of GSK3B, MM samples exclusively express the full-length mRNA. As a group, genes that lose IpA expression in MM samples compared to PCs also upregulate full-length transcript

expression, presumably rescuing the function of the full-length protein to varying degrees. Interestingly, not all cancer cells show depletion of IpA isoforms as we found increased rather than reduced IpA isoform expression in another B cell malignancy, chronic lymphocytic leukemia, which is presented elsewhere.

Methods

3'-seq computational analyses

Preprocessing of 3'-seq libraries, read alignment (hg19)⁵², identification and quantification of peaks were performed as described by Lianoglou et al. (2013)¹. The peaks were assigned to genes using RefSeq annotations. To obtain an atlas of robust cleavage events in 3'UTRs and introns, we started with all the peaks that were detected by peak calling of all the pooled samples and then followed a series of steps to filter lowly expressed peaks and the ones that potentially originate from different artifacts.

Removing artifacts. The peaks potentially resulting from different artifacts were identified and removed: i) peaks overlapping blacklisted regions of human genome (n = 2,841; 0.16%) (<https://sites.google.com/site/anshulkundaje/projects/blacklists>)⁵³; ii) internally primed peaks (Lianoglou et al. 2013) (n = 662,562; 36%); and iii) antisense peaks (n = 289,340; 16%).

Removing the immunoglobulin peaks. Our dataset included plasma cells which are fully differentiated B cells that secrete antibodies. As plasma cells produce massive quantities of antibodies, a large fraction of 3'-seq reads mapped to immunoglobulin loci on chromosomes 2 and 14. It was essential to account for this skewed expression of specific genomic regions in order to get a reasonable quantification for the expression of other genes. Thus, peaks (n = 11) overlapping with parts of the genome coding for immunoglobulins were removed. Even after this correction, one sample of plasma cells had a high number of intergenic reads (PC2). Thus, this sample was not used for identification of robustly expressed isoforms but only to quantify them. The library size

was reduced accordingly for plasma cell samples, since the peaks described above result either from sequencing artifacts or from skewed expression of specific genomic regions.

Removing ambiguous peaks. Some genes in the genome overlap with each other. In such cases, it is difficult to assign 3'-seq reads to the genes accurately, and thus such genes (n = 336) were removed from further analysis. This resulted in the removal of 8,437 peaks from the atlas. Since we were interested in investigating the lpA isoforms of protein coding genes, peaks falling in introns that potentially originated from microRNAs, small nucleolar RNAs and retrotransposons were also removed (n = 4,722). Genes that were on the opposite strand but had a 3'UTR end in the intron (100 nt) of a convergent gene can create artifactual antisense peaks in the intron. Thus, peaks in introns that were close to the end of an opposite strand 3'UTR were also removed (n = 2,091). This corresponded to discarding peaks in the introns of 630 genes. There are genes where the end of the 3'UTR might fall in the intron of the downstream gene on the same strand. This would also create peaks in introns that are contributed by the preceding gene. Therefore, peaks in the intron that were within 5000 nt of the 3' end of the 3'UTR of the previous gene were discarded (n = 2,079); the discarded peaks came from the introns of 785 genes.

Identification of robust isoforms. The expression levels of lpA and 3'UTR ApA isoforms were quantified by tags per million (TPM) falling in 3'-seq peaks, i.e. the read count of the peak regions was normalized by the library size of the respective sample. A gene can have many cleavage events with adequate expression levels. To examine cleavage events that represented one of the major isoforms with respect to all isoforms with a 3' end in a given gene, these isoforms were filtered by usage. Usage is a statistic that gives an estimate of the relative expression of the isoform. As different 3'UTR ApA isoforms create the same protein irrespective of the 3'UTR length, the usage of lpA isoforms was calculated with respect to the total expression of 3'UTR ApA isoforms. lpA

isoforms that end in different introns result in distinct protein isoforms (when translated), therefore, their usage was calculated relative to the total expression of both lpA isoforms and 3'UTR isoforms.

As we were interested in analyzing functionally relevant isoforms, we filtered for robustly expressed isoforms by imposing TPM and usage cutoffs. For the 3'UTR ApA isoforms, an isoform that was expressed with at least 3 TPM and with usage of 0.1 or more became part of the atlas. To focus on the most confident lpA isoforms, an lpA isoform was considered to be robustly expressed only when it was expressed with 5 TPM or more and had 0.1 usage in at least one sample. The interquartile range of the start position of the reads was also required to be 5 or more for the peaks in that particular sample to be defined as a real lpA isoform to eliminate peaks originating from PCR duplicates. These criteria helped to filter the lowly expressed isoforms as well as any possible known artifacts. Filtering for these expression criteria shrunk the atlas from 410,404 peaks to 46,923. As we were interested in lpA and 3'UTR ApA isoforms that would have different functional consequences, peaks that were within 200 nt were clustered to represent a single 3' cleavage event. Clustering reduced the number of peaks to 40,105. After following the steps above, the atlas comprised 27,927 peaks in 15,670 genes for cleavage events of the 3' UTRs and 3' ends of 5,957 lpA isoforms in 3,945 genes. For downstream analysis, we only focused on the lpA isoforms ($n = 5,670$) of protein coding genes ($n = 3,768$).

Validation and independent read evidence of lpA isoforms

We tried to corroborate the robustly expressed lpA events described thus far ($n = 5,670$) with external sources of evidence as described below:

1. External annotation: As annotated, we consider mRNA isoforms present in RefSeq, UCSC or Ensembl. Last exons of all the existing transcripts of the hg19 annotation for Refseq, UCSC and Ensembl were obtained. These last exons were resized to include a region 100 nt downstream of the annotated end. If the 3' end of the lpA isoform detected

by our 3'-seq analysis overlapped with an expanded last exon, then it was considered to be substantiated by an external annotation. 39.52% ($n = 2,241$) of all the lpA events fell in the vicinity of annotated 3' end (using the previous definition) based on an external annotation.

2. RNA-seq GLM: RNA-seq read coverage is expected only over the exons and not over the introns, since the splicing machinery splices out introns during co-transcriptional processing of the pre-mRNA. However, if there is an lpA isoform that ends in an intron, then there should be RNA-seq read coverage before the 3' end of the lpA isoform and no read coverage after the 3' end (**Fig. 1d**). To test whether the upstream read coverage was significantly higher than the downstream read coverage, two windows of 100 nt separated by 51 nt upstream and downstream of the lpA 3' end were defined. These two windows served as replicate bin counts for upstream and downstream coverage. As this was done within each single RNA-seq sample, library size normalization was not required (i.e. the size factor was set as 1 for every comparison). Significant differential expression upstream vs. downstream using DESeq¹⁷ was then tested (FDR-adjusted $p < 0.1$). Not all lpA isoforms could be tested by DESeq. lpA isoforms where the defined windows overlapped with an annotated exon were excluded from further analysis. In total, 4,802 events were tested. As a control for this analysis, random introns of expressed genes that did not contain 3' end peaks were sampled and analyzed as described above. DESeq analysis returned p values consistent with the null hypothesis (**Supplementary Fig. 1a**). The RNA-seq validation was applied over all the RNA-seq samples. If an lpA event was validated in any sample, then it was considered be supported by RNA-seq data. Of all the lpA isoforms, 29% ($n = 1,670$) could be validated by this approach.

3. Other 3'-end sequencing protocols: If lpA isoforms detected by our 3'-seq protocol were also found by other 3'-end sequencing methods¹⁸, we include the lpA event in our atlas of high confidence lpA events. This led to the inclusion of 1,332 lpA isoforms. The

peaks reported by Gruber et al. (2016) were resized to be 75 nt width (25 nt upstream of the original start and 50 nt downstream of the original start). Overall 70% (n = 3,999) of lpA events were supported by other 3'-end sequencing protocols.

4. Untemplated adenosines from RNA-seq reads (RNA-seq, polyA reads): In RNA-seq data, some reads may overlap the 3' end of the templated transcript and the start of the polyA tail; these reads contain untemplated adenosines and thus fail to map to the genome. Reads that did not map to the human genome were therefore used to get additional support for the lpA 3' ends. To make sure that these reads were at the 3' end, the reads ending with 4 or more As were trimmed. Only reads that were greater than 21 nt in length after trimming were retained. Unmapped reads from all RNA-seq samples were trimmed, and all reads with untemplated As were pooled. These reads were then aligned to the human genome. Using the aligned BAM file, all the reads that were possible PCR duplicates were further filtered out. The uniquely mappable reads that overlapped with the lpA peak (20 nt extended upstream and downstream) were counted. If an lpA isoform was supported by four or more trimmed RNA-seq polyA reads together with the presence of one of the polyA signals (AAUAAA and its variants)⁵⁴, then the lpA isoform was considered to be corroborated by polyA RNA-seq reads.

5. Highly expressed lpA isoforms: Since many of our cell types have not been previously assayed by other 3'-end sequencing methods and are also not represented well in existing RNA-seq data sets, we rescued highly expressed cell type-specific lpA isoforms by using a stringent expression cutoff (10 TPM and 0.1 usage). We also required the presence of an upstream polyA signal (AAUAAA and its variants)⁵⁴. This step enabled us to include 323 lpA events in the atlas of highly confident lpA events.

Expression cut-offs used for lpA and full-length mRNA expression

A gene is considered to be expressed if either the lpA isoform (≥ 5 TPM) or the full-length isoform (≥ 5.5 TPM) were expressed in 75% of the samples of the particular cell type.

Conservation analysis

We obtained phastCons 46-way conservation scores²⁰ for 200 nts upstream and downstream of the 3' ends of lpA isoforms to compare the mean conservation score of the 3' ends of lpA isoforms against random introns containing polyA signals, but without lpA site usage. The random introns ($n = 5,000$) were chosen from lpA genes but we selected introns without lpA events, but with at least one polyA signal (AAUAAA). One of these polyA signals was randomly selected, and we obtained the phastCons 46-way conservation score for 200 nts upstream and downstream of this polyA signal.

Identification of differentially used lpA sites

lpA site usage was calculated as the fraction of reads that map to the lpA site compared to all the reads that map to the 3'UTR of each gene. This translates into the relative expression of the truncated protein compared with the full-length protein. To identify the statistically significant changes in the usage of pA signals, we used a GLM, where we model the read counts of all isoforms across conditions by negative binomial distributions and we test for the significance of an interaction term between isoform and condition. This form of modeling approach was adapted from DEXSeq, which is formulated for testing the differential usage of exons²³. If a gene has multiple lpA isoforms, then the relative expression of each lpA isoform as well as the pooled full-length mRNA expression were tested independently, since different lpA isoforms are translated into different protein isoforms.

Gene ontology enrichment analysis

Functional annotation enrichment was performed on the genes with significant differential usage of lpa sites (**Fig. 3a**) using DAVID with the expressed genes as the background²⁴. Functional annotation enrichment by DAVID was also performed for the genes that lose TMDs and retain TMDs with all the genes that have TMDs expressing lpa isoforms as background.

Protein domain analysis

The information about protein domains was obtained from the UCSC UniProt annotation table (spAnnot) via the Bioconductor package-rtracklayer. Only the domains with annotation type 'active site', 'domain', 'transmembrane region', 'repeat', 'zinc finger region', 'compositionally biased region', 'DNA-binding region', 'region of interest', 'lipid moiety-binding region', 'short sequence motif', 'calcium-binding region', 'nucleotide phosphate-binding region', 'metal ion-binding site' and 'topological domain' from UniProt were used for analysis. These domains were further categorized into more broad categories: i) Active site – active site and catalytic sites, ii) DNA-binding domains - C2H2-type, PHD-type, C3H1-type, KRAB, Bromo, Chromo, DNA-binding, C4-type, CHCR, A.T hook, bZIP, bHLH, CCHC-type, CHCH, Bromodomain-like, CH1, C6-type, A.T hook-like, C4H2-type and CHHC-type, iii) Protein-protein interaction domains (PPI) - WD, ANK, TPR, LRR, HEAT, Sushi, EF-hand, ARM, PDZ, PH, SH3, RING-type, LIM zinc-binding, WW, SH2, BTB, FERM, CH, Rod, Coil 1A, MH2, WD40-like repeat, t-SNARE coiled-coil homology, Coil 1B, Cbl-PTB, Coil, CARD, SH2-like, DED, IRS-type PTB, SP-RING-type, EF-hand-like, RING-CH-type, v-SNARE coiled-coil homology, Arm domain, LIM protein-binding, GYF, PDZ domain-binding and PDZD11-binding. Also, if a region in protein was annotated with 'Interaction with' then that region was considered a PPI domain, iv) RNA-binding domains - RRM, SAM, KH, DRBM, RBD, Piwi, PAZ, S1 motif, Pumilio and THUMP, v) Transmembrane domains (TMDs) - transmembrane region, ABC transmembrane type-1, ABC transporter and ABC transmembrane type-2, and, vi) Repeated - Any domains that were repeated in the protein were considered repeated domains. If a gene had multiple protein isoforms, then the longest isoform was

used in the analysis. The protein lengths were obtained from <http://www.uniprot.org/> for *Homo sapiens*.

Distance of lpA from TMDs

lpA isoforms for which there was positional information about the start of first TMD and those that retained at least one domain were used for this analysis. Further, we focused on lpA isoforms that completely lost all their TMDs due to the cleavage event in the intron. The distance of the retained CDR (in amino acids) by lpA from the first TMD was determined as: (Upstream CDR from lpA – Upstream CDR from the intron before the first TMD)/3.

De novo transcript assembly

The complete transcript structure was obtained through the following steps. i) We used StringTie, an improved method for more accurate *de novo* assembly of transcripts from RNA-seq data⁵⁵. *De novo* assembly was performed on every RNA-seq sample with default settings using the hg19 RefSeq annotation. ii) Transcripts from multiple assemblies were subsequently unified using CuffCompare, which removes redundant transcripts and provides a set of unique transcript structures⁵⁶. iii) For each individual gene, we obtained the transcripts that overlapped the gene's coordinates. We gave preference to multi-exon transcripts over single exon transcripts. For single exon transcripts, we allowed the start/end to be within 100 nt of the TSS. We gave this advantage to the single exon transcripts because the direction of transcription for these transcripts is not certain. iv) Finally, using the 3' ends of lpA isoforms (from our 3'-seq data), we assigned transcripts with the nearest ends to these lpA isoforms.

Firstly, we identified transcripts that ended within 50 nt of 3'-seq events. If there were several assembled transcripts meeting this criterion, we chose the transcript that had

the maximum number of exons. If there was a tie in the number of exons, then we chose the transcript that started closest to annotated TSS. For the remaining 3'-seq events, we assigned the nearest ending transcript. Finally, using the above defined criteria for selecting transcript structures, we determined which lpa isoforms corresponded to these assembled transcripts. If the 3' end of the lpa isoform was within 500 nt of the defined transcript end, then we assumed that this particular transcript represented the full structure of the lpa isoform. For some lpa isoforms we observed usage of different polyA signals within the same intron. Thus, to account for such cases, for the lpa events that did not fall within 500 nt of a transcript end, we determined if it overlapped a transcript that ended within 5000 nt. If this was the case, then we assigned this transcript to that 3' end. We were able to define the transcript architecture for $n = 954$ lpa isoforms (both annotated and unannotated). If the transcripts ends differed from the lpa 3'-seq events, then we defined the 3' end determined from 3'-seq to be the real end. This was done as 3'-seq identifies 3' ends of polyadenylated mRNAs at single nucleotide resolution and thus is more accurate than transcript ends obtained from short read assembly.

Coding potential prediction - To determine the probability that the 5'lpa events represented non-coding transcripts, we made use of CPAT, a tool that predicts the coding potential of the transcript based on four sequence features: open reading frame size, open reading frame coverage, Fickett TESTCODE statistic, and hexamer usage bias²⁵. For our analysis, we considered non-coding lpa isoforms to be the ones that had coding potential probability less than 0.3, had retained coding sequence less than 25%, and had ORF ≤ 300 nt ($n = 469$).

Binding site enrichment of RNA-binding proteins in exonized introns

We used available CLIP (cross-linking immunoprecipitation) sequencing data of RNA-binding proteins from dorina⁵⁷. As the majority of CLIP studies was performed in HEK293 cells, we we focused on non-coding lpa isoforms expressed in HEK293 and

only included lpa isoforms in the analysis whose exonized intron was larger than 50 nt (lpa isoforms = 62, genes = 58).

We determined if the exonized part of the intron was enriched for binding sites of RNA-binding proteins compared to other regions (introns, coding exons, 3'UTRs) of the transcription units, called 'background' here. We calculated the expected number of binding sites in the exonized introns using each background and compared it to the observed number of binding sites in the exonized introns. This enabled us to calculate a binomial Z-score of each CLIP experiment and each background region. We observed enrichment of binding sites of RNA-binding proteins in the exonized introns compared with introns and coding exons but no enrichment compared to 3'UTRs. The RNA-binding proteins with Z-scores ≥ 10 compared to introns or coding exons are PUM2, FUS, ELAVL1, TIAL1 and TAF15.

Identification of retained introns

Retained introns were identified using a modified version of the IRFinder algorithm⁴¹. To avoid genes with a complex genomic architecture, we removed genes that overlap with other genes in either the sense or antisense strand. An intron was categorized as retained if it satisfied the following criteria. i) There should be at least three reads spanning both the upstream and downstream exon-intron junction. ii) At least 50% of the intron length should be covered by 3 or more unique reads. Mappability of introns could be a limitation in this case, and thus we focused only on introns that had at least 50% uniquely mappable sequence relative to its complete length. iii) To ensure adequate expression of the flanking exons, the median coverage over the flanking exons was required to be 10 reads or more. iv) Since the introns should have more coverage than background noise, we considered introns to be retained if the ratio of median coverage over the intron to median coverage of the upstream and downstream exons was at least 10%.

An intron was annotated as retained if it fulfilled all criteria in at least 66% of the RNA-

seq samples of a given cell type. Introns retained in 33% or fewer samples were flagged as not retained while the introns that were retained in more than 33% samples but less than 66% of RNA-seq samples were removed from the analysis. For a 3' end of an lpa isoform to occur in a particular intron, the intron must contain a polyA signal or one of its variants ⁵⁴.

Our data showed that some genes had very high coverage over almost all the introns of the gene, presumably due to sequencing artifacts. We determined the (median coverage over all the introns)/(median coverage over all the exons), and if this ratio was ≥ 0.2 then these genes were flagged for removal.

The number of introns that would have lpa and IR simultaneously by chance were calculated as – Probability of lpa \times Probability of IR \times Number of expressed introns with polyA signal

References

1. Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27**, 2380-2396 (2013).
2. Derti A, *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**, 1173-1183 (2012).
3. Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761-772 (2011).
4. Hoque M, *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods* **10**, 133-139 (2013).
5. Martin G, Gruber AR, Keller W, Zavolan M. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep* **1**, 753-763 (2012).
6. Fu Y, *et al.* Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res* **21**, 741-747 (2011).
7. Beck AH, *et al.* 3'-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One* **5**, e8768 (2010).
8. Early P, *et al.* Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**, 313-319 (1980).
9. Rogers J, *et al.* Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell* **20**, 303-312 (1980).
10. Edwalds-Gilbert G, Veraldi KL, Milcarek C. Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res* **25**, 2547-2561 (1997).
11. Di Giammartino DC, Nishida K, Manley JL. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**, 853-866 (2011).
12. Kruse U, Sippel AE. The genes for transcription factor nuclear factor I give rise to corresponding splice variants between vertebrate species. *J Mol Biol* **238**, 860-865 (1994).

13. Benech P, Mory Y, Revel M, Chebath J. Structure of two forms of the interferon-induced (2'-5') oligo A synthetase of human cells based on cDNAs and gene sequences. *EMBO J* **4**, 2249-2256 (1985).
14. Wang H, Sartini BL, Millette CF, Kilpatrick DL. A developmental switch in transcription factor isoforms during spermatogenesis controlled by alternative messenger RNA 3'-end formation. *Biol Reprod* **75**, 318-323 (2006).
15. Taliaferro JM, *et al.* Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Mol Cell* **61**, 821-833 (2016).
16. Peterson ML. Regulated immunoglobulin (Ig) RNA processing does not require specific cis-acting sequences: non-Ig RNA can be alternatively processed in B cells and plasma cells. *Mol Cell Biol* **14**, 7891-7898 (1994).
17. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
18. Gruber AJ, *et al.* A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res* **26**, 1145-1159 (2016).
19. Tian B, Pan Z, Lee JY. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res* **17**, 156-165 (2007).
20. Siepel A, *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005).
21. Kaida D, *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664-668 (2010).
22. Berg MG, *et al.* U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**, 53-64 (2012).
23. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008-2017 (2012).
24. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

25. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**, e74 (2013).
26. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5**, e106 (2007).
27. Slavoff SA, *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9**, 59-64 (2013).
28. Anderson DM, *et al.* A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **160**, 595-606 (2015).
29. Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, Kageyama Y. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol* **9**, 660-665 (2007).
30. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673-684 (2009).
31. Xia Z, *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**, 5274 (2014).
32. Kronke J, *et al.* Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. *Science* **343**, 301-305 (2014).
33. Gocke CB, *et al.* IQGAP1 Scaffold-MAP Kinase Interactions Enhance Multiple Myeloma Clonogenic Growth and Self-Renewal. *Mol Cancer Ther* **15**, 2733-2739 (2016).
34. Busino L, *et al.* Fbxw7 α - and GSK3-mediated degradation of p100 is a pro-survival mechanism in multiple myeloma. *Nat Cell Biol* **14**, 375-385 (2012).
35. M GA, *et al.* Regulation of myeloma cell growth through Akt/Gsk3/forkhead signaling pathway. *Biochem Biophys Res Commun* **297**, 760-764 (2002).
36. Zhou Y, Uddin S, Zimmerman T, Kang JA, Ulaszek J, Wickrema A. Growth control of multiple myeloma cells through inhibition of glycogen synthase kinase-3. *Leuk Lymphoma* **49**, 1945-1953 (2008).

37. Rutledge T, Cosson P, Manolios N, Bonifacino JS, Klausner RD. Transmembrane helical interactions: zeta chain dimerization and functional association with the T cell antigen receptor. *EMBO J* **11**, 3245-3254 (1992).
38. Houseley J, Tollervey D. The many pathways of RNA degradation. *Cell* **136**, 763-776 (2009).
39. Zarnack K, *et al.* Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453-466 (2013).
40. Attig J, *et al.* Splicing repression allows the gradual emergence of new Alu-exons in primate evolution. *Elife* **5**, (2016).
41. Wong JJ, *et al.* Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**, 583-595 (2013).
42. Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res* **44**, 838-851 (2016).
43. Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchiev A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666-670 (2007).
44. Di Ruscio A, *et al.* DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* **503**, 371-376 (2013).
45. Wang KC, *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124 (2011).
46. Wang X, *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* **454**, 126-130 (2008).
47. Peschon JJ, *et al.* An essential role for ectodomain shedding in mammalian development. *Science* **282**, 1281-1284 (1998).
48. Contin C, Pitard V, Itai T, Nagata S, Moreau JF, Dechanet-Merville J. Membrane-anchored CD40 is processed by the tumor necrosis factor- α -converting enzyme. Implications for CD40 signaling. *J Biol Chem* **278**, 32801-32809 (2003).

49. Suda T, Hashimoto H, Tanaka M, Ochi T, Nagata S. Membrane Fas ligand kills human peripheral blood T lymphocytes, and soluble Fas ligand blocks the killing. *J Exp Med* **186**, 2045-2050 (1997).
50. Rossille D, *et al.* High level of soluble programmed cell death ligand 1 in blood impacts overall survival in aggressive diffuse large B-Cell lymphoma: results from a French multicenter clinical trial. *Leukemia* **28**, 2367-2375 (2014).
51. Gregory AP, *et al.* TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. *Nature* **488**, 508-511 (2012).
52. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
53. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
54. Tian B, Hu J, Zhang H, Lutz CS. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**, 201-212 (2005).
55. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295 (2015).
56. Trapnell C, *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578 (2012).
57. Blin K, Dieterich C, Wurmus R, Rajewsky N, Landthaler M, Akalin A. DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* **43**, D160-167 (2015).

Singh Figure 1

Singh, et al., page 34

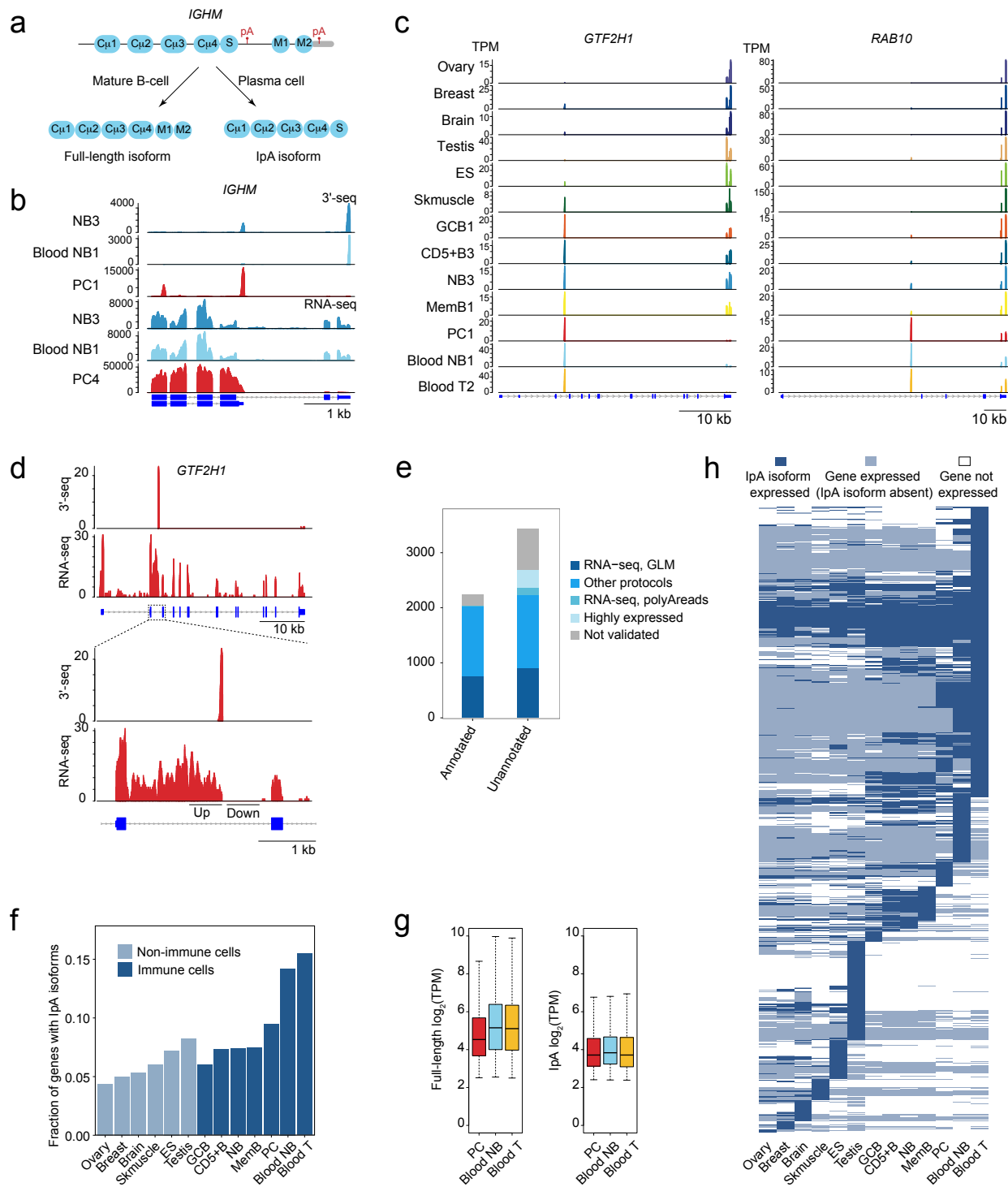


Figure 1: Widespread intronic polyadenylation with robust expression in circulating immune cells.

- a) Schematic representation of full-length and lpA isoform of *IGHM* expressed in mature B cells and plasma cells (PC).
- b) 3'-seq (tags per million, TPM) and RNA-seq (read coverage) tracks showing expression of the lpA and full-length mRNA isoforms of *IGHM* (ENSG00000211899), encoding the immunoglobulin mu heavy chain, IgM. The full-length isoform is expressed in NB from blood and lymphoid tissue and includes two exons encoding the C-terminal transmembrane domain of membrane-bound IgM. The lpA isoform is expressed in PCs obtained from bone marrow. It lacks the transmembrane domain which leads to expression of soluble IgM.
- c) 3'-seq tracks showing lpA isoform expression for two genes across human tissues and immune cell types.
- d) RNA-seq coverage of intronic regions flanking lpA sites. A GLM-based test is used to validate the lpA isoforms. An isoform is considered validated if there is a significant difference (FDR-adjusted $p < 0.1$) in read counts in windows located up- and downstream of the putative lpA site.
- e) The fraction of lpA isoforms validated by read evidence from independent data sets is shown for annotated and unannotated lpA isoforms. lpA isoforms present in RefSeq, UCSC genes and Ensembl databases are considered to be annotated.
- f) The fraction of expressed genes that generate lpA isoforms is shown for each cell type.
- g) Expression levels (\log_2 TPM) for full-length mRNAs and lpA isoforms are shown as boxplots for in PCs, blood NB and T cells. lpA isoforms are robustly expressed as full-length mRNA expression is 4.53, 5.15 and 5.11, respectively, compared to 3.71, 3.83 and 3.71 for lpA isoforms.
- h) Tissue-specific expression of lpA isoforms. Each row represents a gene. Dark blue, lpA isoform is expressed (≥ 5 TPM); light blue, lpA isoform is not expressed, but full-length mRNA is expressed (≥ 5.5 TPM); and white, gene is not expressed.

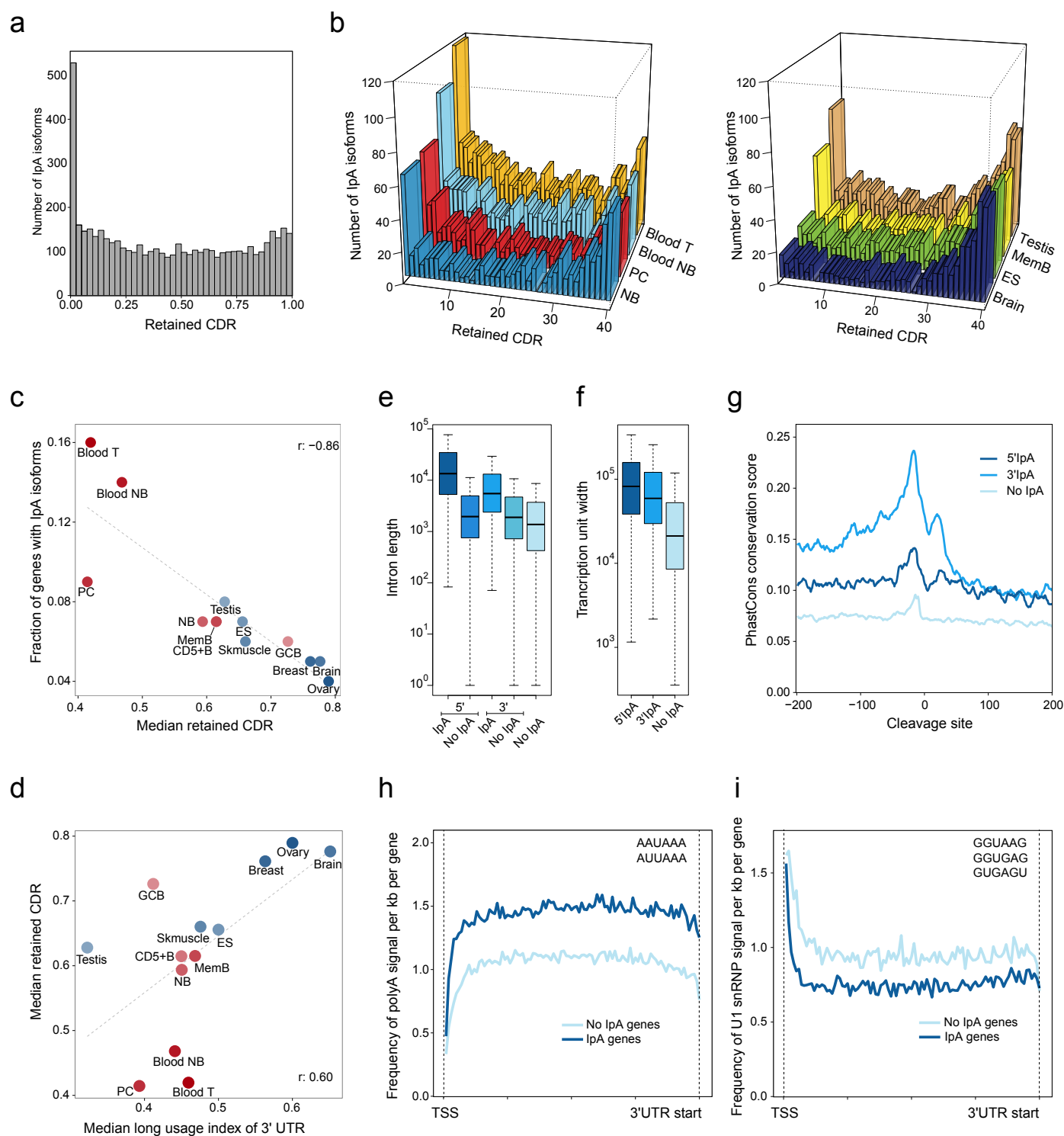


Figure 2: Enrichment of IpA sites at the start of transcription units.

- a) The fraction of retained coding region (CDR) was calculated as the nucleotides from the start codon to the end of the exon located upstream of the IpA peak, divided by all coding nucleotides of the longest annotated open reading frame and is shown for all IpA isoforms in the atlas.
- b) As in (a), but shown for individual cell types.
- c) Correlation between the median retained CDR with the fraction of genes that generate IpA isoforms in each sample (Pearson correlation coefficient, $r = -0.86$). Tissues with a higher proportion of IpA isoforms generate IpA isoforms with shorter CDRs.
- d) Correlation between the median retained CDR and the median usage of the distal ApA site in the 3'UTR (Pearson correlation coefficient, $r = 0.60$). Tissues with shorter 3'UTRs have IpA isoforms with shorter CDRs.
- e) IpA isoforms occur in long introns. The introns in which 5'IpA events occur are longer than the other introns of the same genes (one-sided Wilcoxon rank-sum test, $p < 10^{-20}$). Similarly, the introns in which 3'IpA events occur are longer than the remaining introns of those genes (one-sided Wilcoxon rank-sum test, $p < 10^{-20}$). If taken together, then the introns in which IpA events occur are longer than the introns of the genes that only express full-length isoform (one-sided Wilcoxon rank-sum test, $p < 10^{-20}$).
- f) IpA isoforms occur in genes with long transcription units. Genes that express IpA isoforms have longer transcription units compared to genes that only express full-length isoforms (one-sided Wilcoxon rank-sum test, $p < 10^{-20}$).
- g) Higher conservation around the cleavage sites of IpA isoforms. The plot shows PhastCons scores of 200 nt upstream and downstream of IpA cleavage sites ($x = 0$). 5'IpA and 3'IpA events both have significantly higher conservation flanking the cleavage site compared to corresponding regions of randomly selected polyA signals (AAUAAA) in introns lacking IpA events (one-sided Wilcoxon signed-rank test, $p < 10^{-68}$ for both comparisons).
- h) Genes with IpA isoforms ($n = 3,481$) are enriched for polyA sites compared with genes that do not generate IpA isoforms ($n = 12,092$) (one-sided Wilcoxon signed-rank test, $p < 10^{-18}$). The frequency of polyA sites was counted from the TSS to the beginning of the 3'UTR and is shown as the average number of signals per kb per gene.

i) As in (h), but U1 binding sites are shown. IpA genes are depleted for U1 snRNP signals (one-sided Wilcoxon signed-rank test, $p < 10^{-18}$).

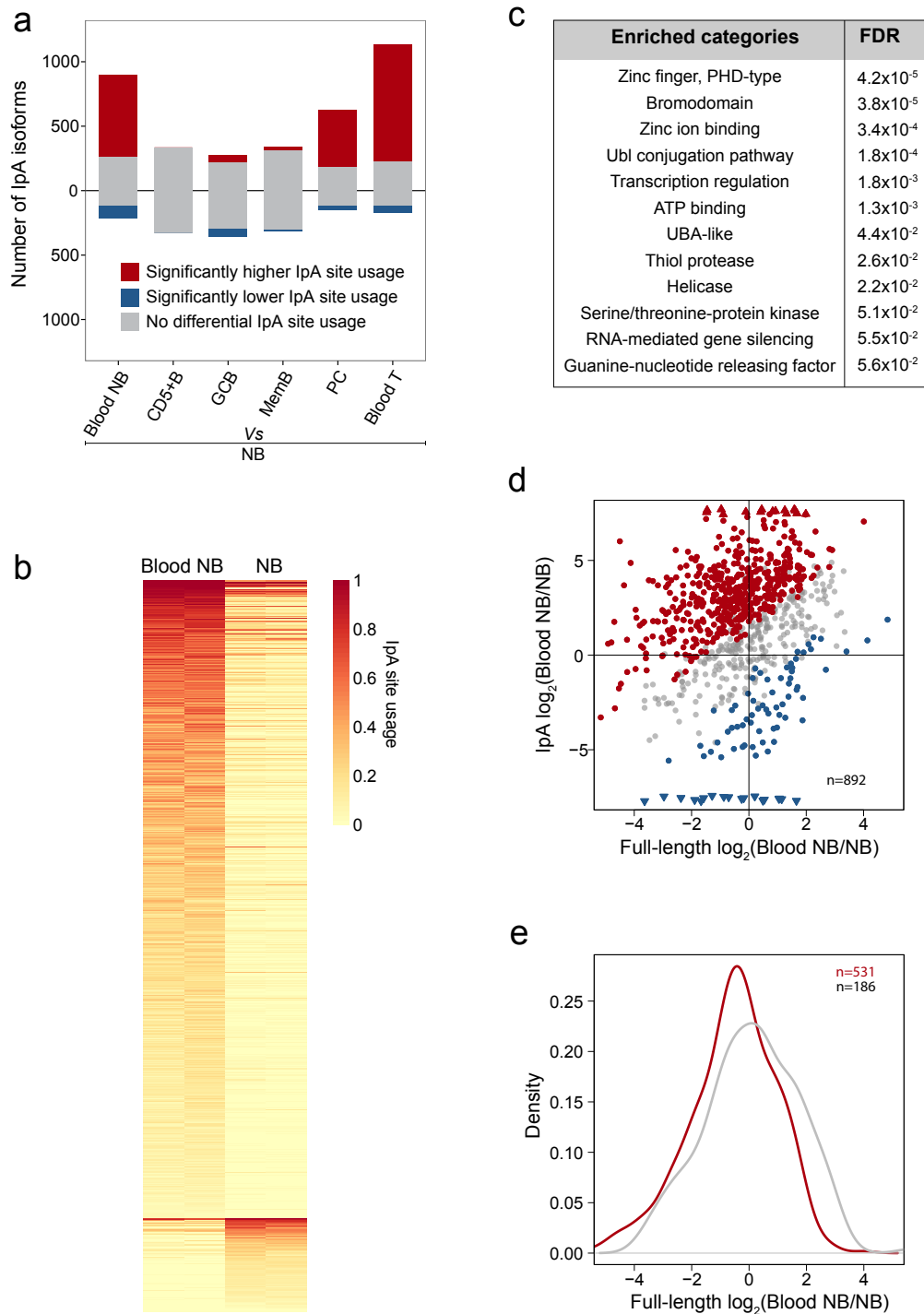


Figure 3: Dynamic expression of lpa isoforms in immune cells.

- a) Number of lpa isoforms with differential usage of lpa sites between NB from lymphoid tissue versus other immune cells (FDR-adjusted $p < 0.05$).
- b) Heatmap showing lpa site usage of lpa isoforms with significantly different usage (FDR-adjusted $p < 0.05$) between NB derived from blood or lymphoid tissue ($n = 720$). Each row indicates an lpa isoform.
- c) Enrichment of gene ontology terms for the genes shown in (b).
- d) Fold change of lpa isoform and full-length mRNA expression in blood versus lymphoid tissue-derived NB by TPM. All the genes that were tested for differential usage are shown ($n = 892$). If a gene had multiple lpa isoforms, then the one with the most significant differential lpa usage is shown. lpa isoforms with significantly different usage (FDR-adjusted $p < 0.05$) are highlighted in red (higher usage) or blue (lower usage).
- e) Significant downregulation of full-length mRNAs in genes with significant lpa isoform expression (one sided KS test, $p < 10^{-5}$). Shown are genes highlighted in red from (d).

Singh Figure 4

Singh, et al., page 41

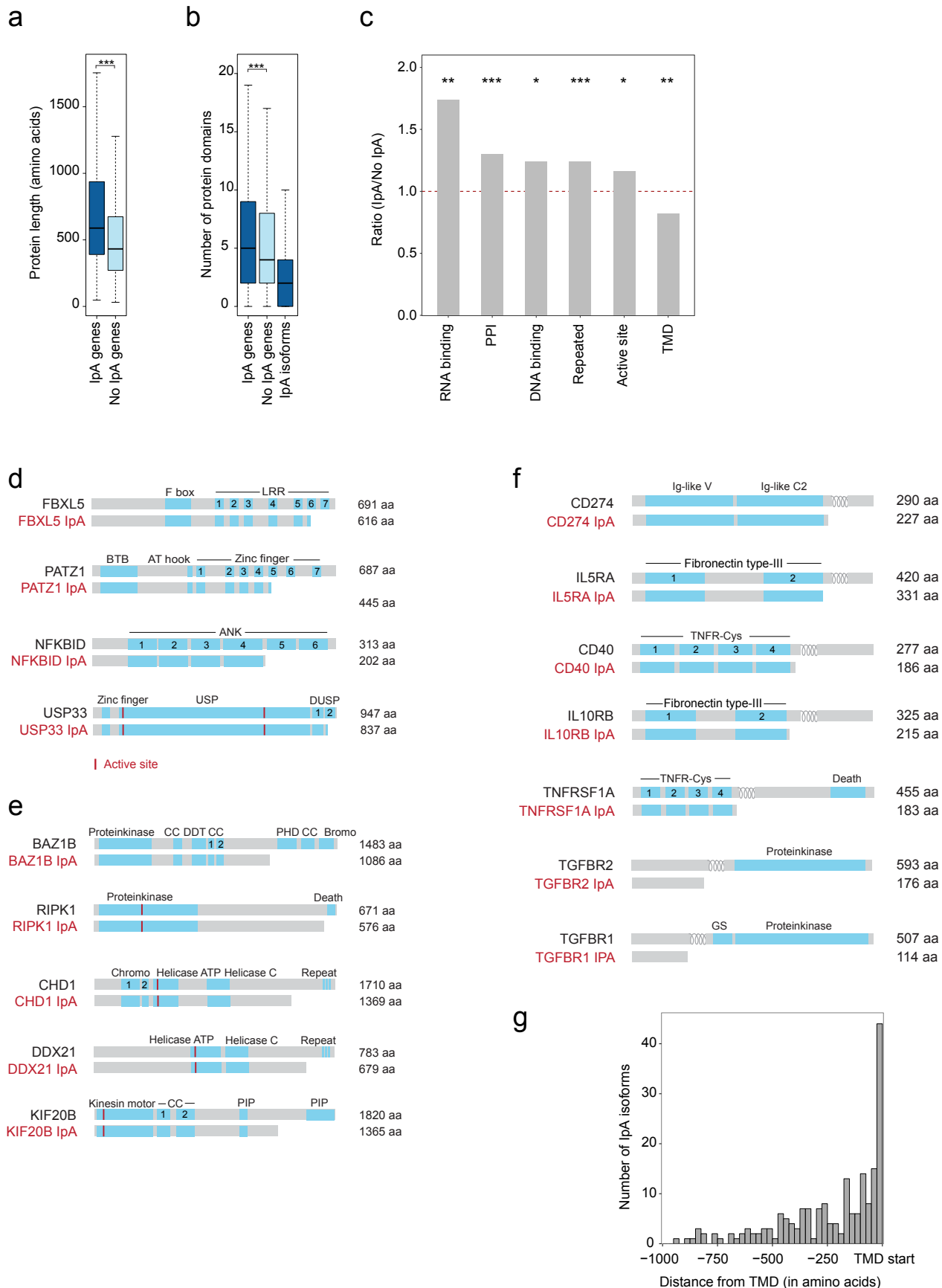


Figure 4: IpA isoforms diversify the proteome.

- a) Genes that express IpA isoforms encode significantly larger proteins compared to genes that only express full-length mRNAs (one-sided Wilcoxon rank-sum test, $p < 10^{-118}$).
- b) Genes that express IpA isoforms encode proteins with significantly more protein domains than genes that only express full-length mRNAs (one sided Wilcoxon rank-sum test, $p < 10^{-14}$). IpA isoforms retain a median of two domains.
- c) Genes that express IpA isoforms are enriched in proteins encoding RNA- and DNA-binding, PPI repeated domains and active sites compared to genes that only express full-length mRNAs. However, IpA genes are depleted for proteins encoding transmembrane domains (TMDs).
- d) Protein models of full-length and IpA-generated truncated proteins are shown in grey for examples that contain repeated domains. Known protein domains are shown as blue boxes and repeated domains are numbered.
- e) As in (d), but shown for enzymes that retain their active sites but lose PPI domains.
- f) As in (d), but shown for plasma membrane proteins. The TMD is indicated by the loops.
- g) Distance between the IpA event and the start of the TMD in IpA isoforms that completely lose their TMDs ($n = 272$).

Singh Figure 5

Singh, et al., page 43

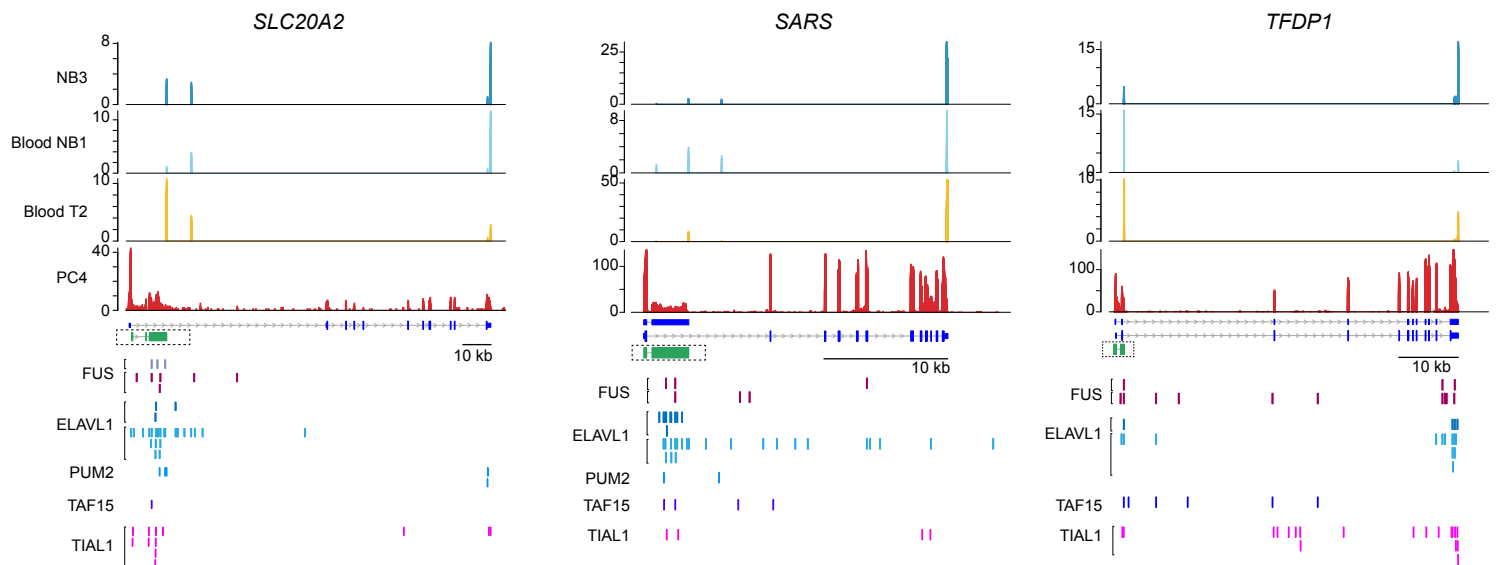


Figure 5: 5'lpA isoforms potentially express non-coding RNAs.

Examples of 5'lpA isoforms are shown as in Fig. 1b. Also shown is the structure of the assembled lpA isoform transcripts in green. Enrichment of CLIP-seq tags over exonized introns of lpA isoforms are shown for the RNA-binding proteins FUS, ELAVL1, PUM2, TAF15 and TIAL1.

Singh Figure 6

Singh, et al., page 45

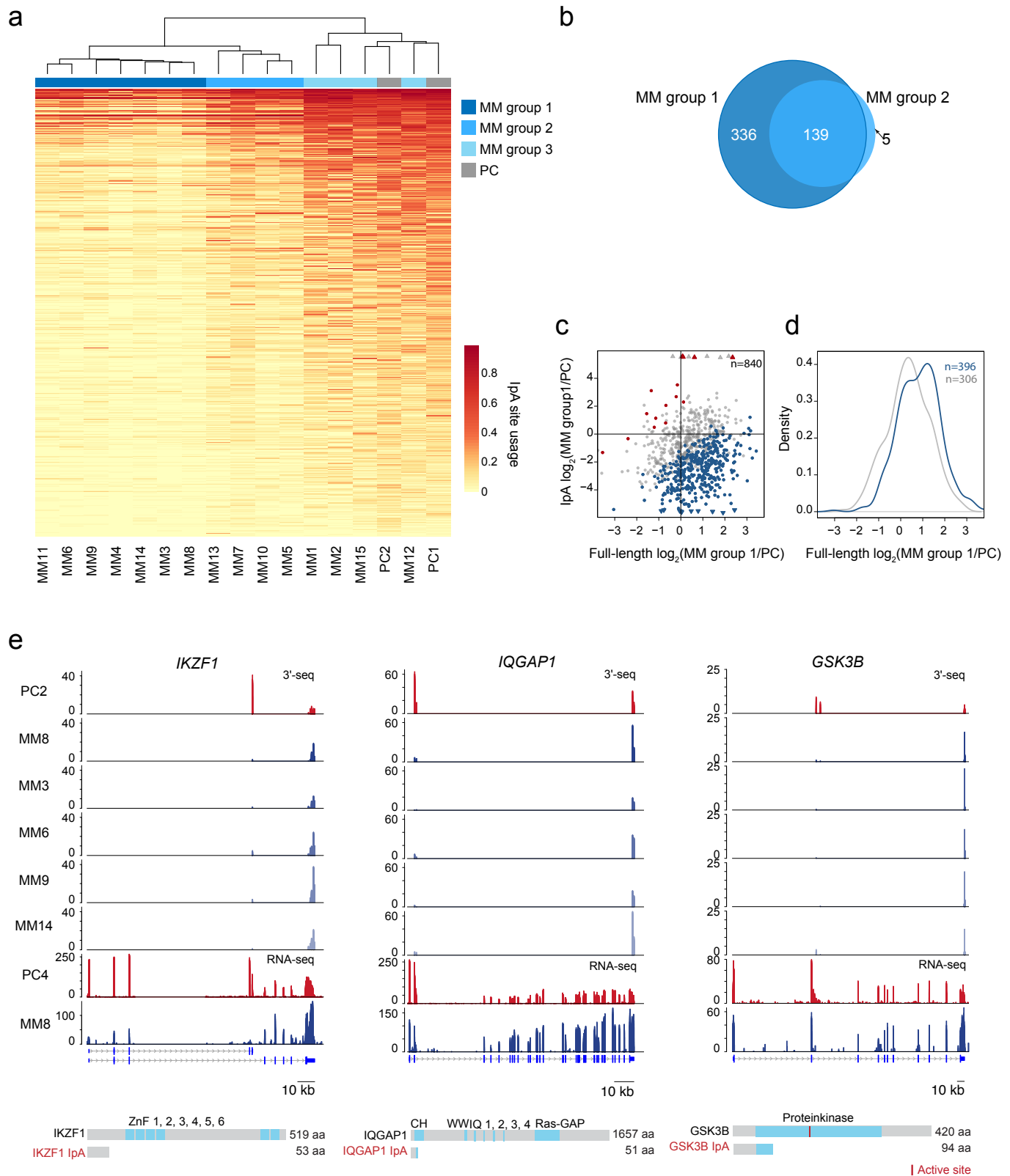


Figure 6: Loss of usage of lpA sites in MM.

- a) Heatmap shown as in Fig. 3b, but for PCs and MM patient samples. MM samples were grouped according to lpA site usage and color-coded, and the lpA isoforms with significantly different usage compared to PCs are shown (FDR-adjusted $p < 0.05$, lower usage of lpA sites in MM, $n = 480$; higher usage of lpA sites in MM, $n = 15$, not shown).
- b) Overlap of significantly lower used lpA sites in MM group 1 and group 2. MM group 3 is not shown as lpA site usage was very similar to PCs and only one lpA isoform was differentially used.
- c) As in Fig. 3d, but for MM group 1 versus PCs. Full-length and lpA isoform expression is shown and significantly different lpA isoforms are color-coded (FDR-adjusted $p < 0.05$).
- d) As in Fig. 3e. Shown is a significant upregulation of full-length mRNA isoform expression (one-sided KS test, $p < 10^{-8}$) of genes highlighted in blue in (c).
- e) Examples of lpA isoforms expressed in PCs, but significantly decreased in MM samples. Shown as in Fig. 1b and Fig. 4e.