Improved transcriptome annotation and read-through transcript identification with TIF-Seq2

Jingwen Wang[1,7], Bingnan Li[1,7], Sueli Marques[1], Lars M. Steinmetz[2,3,4], Wu Wei[3,5,6], Vicent Pelechano[1]*

[1] SciLifeLab, Department of Microbiology, Tumor and Cell Biology. Karolinska Institutet, Solna, Sweden.
[2] European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany.
[3] Stanford Genome Technology Center, Stanford University, Palo Alto, CA, USA.
[4] Department of Genetics, School of Medicine, Stanford University, Stanford, CA, USA.
[5] CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[6] Center for Biomedical Informatics, Shanghai Engineering Research Center for Big Data in Pediatric Precision Medicine, Shanghai Children's Hospital, Shanghai Jiao Tong University, Shanghai 200040, China
[7] These authors contributed equally to the work.
* Correspondence may be addressed to: vicente.pelechano.garcia@ki.se,

Abstract

Eukaryotic transcriptomes are complex involving thousands of overlapping transcripts. The interleaved nature of the transcriptome complicates its annotation, limits our ability to identify regulatory regions and, in some cases, can lead to misinterpretation of gene expression. To improve the annotation of complex genomes, we have developed an optimized method, TIF-Seq2, able to sequence simultaneously the 5' and 3' end of individual mRNA molecules at single-nucleotide resolution. We investigate the transcriptome of a well characterized human cell line (K562) and identify thousands of new transcript isoforms. By focusing on RNAs challenging to investigate with RNA-seq, we accurately define boundaries of lowly expressed intergenic and read-through transcripts. We validate those novel features in cell lines and in Chronic Myeloid Leukaemia patients. Our results demonstrate that TIF-Seq2 improves the annotation of complex genomes facilitating the assignment of promoters to genes and the identification of transcriptionally fused proteins.

Main

Eukaryotic transcriptomes are complex, involving thousands of coding and non-coding RNA isoforms differing in transcription start sites (TSSs), poly(A) sites (PASs) and splicing. Overlapping isoforms can have divergent functional consequences, changing the encoded protein[1,2] or affecting mRNA post-transcriptional life (e.g. translation, localization and stability)[3]. However, the interleaved nature of the transcriptome complicates genome annotation and limits the accurate identification and quantification of alternative isoforms[4,5]. This can lead to incomplete or inaccurate annotations, which cause misinterpretation of gene expression data[6] and limits our ability to link regulatory regions with genes[7] (and thus to genetically manipulate them). Particularly challenging is the correct identification of transcript boundaries of overlapping isoforms[8], even if we know that alternative TSS and PAS drive most isoforms variations across human tissues[9]. Standard RNA-seq can identify transcribed regions and splicing events, however, it cannot distinguish RNA fragments originating from alternative overlapping features. Single-end approaches such as CAGE or poly(A) site sequencing have been key to define the boundaries of the transcriptomes[10,11]. However, those approaches cannot study the combination between TSS and PAS. Long-read sequencing technologies promise to reveal transcription complexity on the genome-wide scale[12]. Nevertheless, the high cost it entails combined with low throughput and limited resolution in the 5' and 3' transcript regions are still major limitations[13].

To bridge the gap between short-read and long-read technologies, we have developed an optimized **T**ranscript **I**so**f**orm **Seq**uencing (TIF-Seq2) especially well-suited for the interrogation of complex transcriptomes. TIF-Seq2 allows to sequence simultaneously the 5' and 3' end of individual RNA molecules at single-nucleotide resolution. To demonstrate its utility, we dissected the overlapping transcriptome of a chronic myeloid leukaemia cell line (*i.e.* K562) in response to Imatinib treatment. We identify thousands of known and unannotated transcript isoforms and validate them using alternative short-read and targeted long-read sequencing approaches. We focus on those transcript isoforms particularly challenging to investigate with RNA-seq. We accurately define the boundaries of lowly expressed intergenic and read-through transcripts. We show that our improved transcriptome annotation improves the detection and analysis of complex overlapping transcripts in clinical samples involving gene-promoter rewiring and the generation of transcriptionally fused proteins.

Results

Development and validation of TIF-Seq2.

We and others have previously developed approaches able to link the 5' and 3' regions of individual transcripts[14–16]. Our work in *S. cerevisiae* demonstrated the existence of a complex overlapping transcriptome even in a simple eukaryote with limited splicing[14]. However, the applicability of our original protocol[17] to the study of complex transcriptomes was limited due to the variability in length of the mappable 5´ and 3´ tags and its modest throughput. To increase the length of the boundary tags, we designed a new sequencing strategy decoupling the region required for bridge amplification from the Illumina sequencing primers (Fig. 1a, Supplementary Fig. 1-2). This decreases the need to perform stringent library size-selection, allows sequencing from the exact 5' and 3' ends of each mRNA molecule and generates longer reads suitable for the study of complex genomes (see Methods). In addition, we performed extensive enzymatic optimizations to maximize the length and complexity of the full-length cDNA libraries, as well as introduced early sample pooling, unique molecular identifiers (UMI) and barcodes to control for the formation of chimeras (Supplementary Fig. 1). All these modifications improved the number of informative reads, their length, quality and thus allowed its application to complex genomes. To demonstrate its utility, we investigated the transcriptome of a well-characterized chronic myeloid leukaemia cell line (K562) in response to the tyrosine kinase inhibitor Imatinib (Supplementary Fig. 3). After quality control and PCR deduplication (see Methods and Supplementary Fig. 4a), we obtained over 14 million pair tags uniquely mapped to the human genome at single-nucleotide resolution (Supplementary Table 1). We clustered adjacent transcription start sites (TSSs) and polyadenylation sites (PASs) and obtained 32,631 TSS and 31,187 PAS clusters (see Methods). Identified clusters are narrow, and 80% of them are smaller than 15 nt (TSS) and 10 nt (PAS) (Supplementary Fig. 5a-b). To validate the accuracy of TIF-seq2, we performed 3´T-fill[18] with the same samples and compared transcript boundaries with independent published datasets for TSS (CAGE[10]) and PAS measurements[11]. In both cases, those analysis demonstrates our accurate detection of transcript boundaries (Fig. 1b and Supplementary Fig. 6). Among high-confidence (normalised counts > 10) TIF boundaries, 85% of TSS and 92% of PAS clusters are located within 100 bp next to the TSSs and PASs from the published datasets. The main advantage of our approach is that it allows to link the identified TSS and PAS clusters (Fig 1c, Supplementary Fig. 4). Doing so, we identified 49,847 unique combinations of TIF-seq2 linked TSS-PAS clusters (referred here as **TIF**s, *boundary* transcript isoforms) supported by at least 4 independent molecular events across all samples. TIF-Seq2-derived transcript boundaries have good agreement with the curated Gencode v28 annotation (Supplementary Fig. 7). At our current sequencing coverage, TIFs overlap 9006 annotated genes (7210 covered by more than one TIF) (Supplementary Fig. 8). Even through TIF-Seq2 TSS and PAS are in good agreement with annotated transcript boundaries (Supplementary Fig. 7), 60% of the TIFs support non-annotated isoforms with

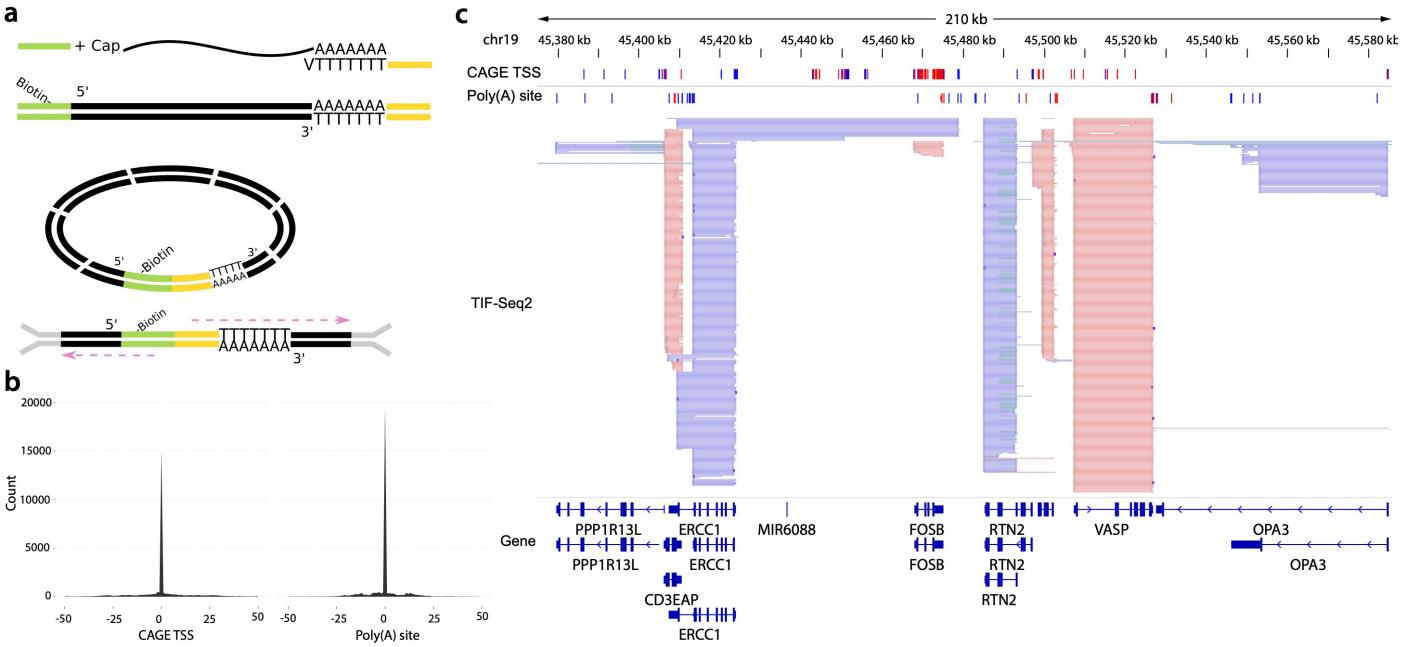

Figure 1. Genome-wide measurement of transcript isoforms with TIF-Seq2. a, TIF-Seq2 protocol; b, Transcript boundaries agree with the transcription start sites (TSSs) defined by CAGE[10] and the poly(A) sites defined by 3' sequencing[11]; c, Transcription landscape of K562 cells without drug treatment. CAGE TSS and poly(A) site (PAS) tracks (red on the forward strand and blue on the reversed strand) are on the top. Individual transcripts molecules captured by TIF-Seq2 are represented by pink (forward strand) or blue (reversed strand) lines connecting the identified TSS and PAS. Annotated transcript isoforms in Gencode v38 are depicted in the bottom.

alternative TSSs, PASs, or both (Supplementary Fig. 9). Current genome annotation is based mainly on classical cloning strategies, RNA-seq and CAGE[5,10], which limit its ability to distinguish between overlapping isoforms. However, TIF-Seq2 is able to pinpoint interdependence of TSSs and PASs across overlapping transcripts (Supplementary Fig. 10). Hence, we decided to focus on isoforms that are more challenging to annotate by traditional approaches. In particular, we focused on lowly expressed intergenic transcripts and read-through transcripts.

Defining the boundaries of lowly expressed intergenic transcripts.

Lowly expressed transcripts are in general challenging to detect. Long-read approaches lack the throughput, and even when combined with capture-based enrichment, they require a prior definition of the regions of interest[19]. On the other hand, short-read RNA-seq approaches with much higher throughput distribute their sequencing power along the whole transcribed region, and need to be linked with independent TSS or PAS dataset to accurately infer their putative boundaries[8]. On the contrary, TIF-seq2 focus its sequencing power on the transcript boundaries (TSS and PAS), allowing to link them. This simplifies the annotation of lowly expressed transcripts and makes it easy to distinguish them from background noise. Using this approach, we identified 1034 TIFs in 426 novel non-overlapping transcribed regions, defined as novel genes (Supplementary Table 2). Although the main application of TIF-Seq2 is the annotation of complex regions, it also allows the quantification of differential expression of the same isoforms across samples. To test that, we treated the K562 cells with 1µM imatinib for 24h. Doing so we identified 879 TIFs that are up/down-regulated (Supplementary Fig. 11 and Supplementary Table 3). And among those, 60 TIFs in 36 novel non-overlapping transcripts were differential expressed (in agreement with our estimation by 3´T-fill, Supplementary Fig. 12 and Supplementary Table 4). We found evidence for their transcription and expression pattern using independent RNA-seq datasets[20] (Supplementary Table 4). About 24% of novel transcripts are significantly regulated (adjusted p-value < 0.05) after Imatinib treatment. In order to validate the identified isoforms and

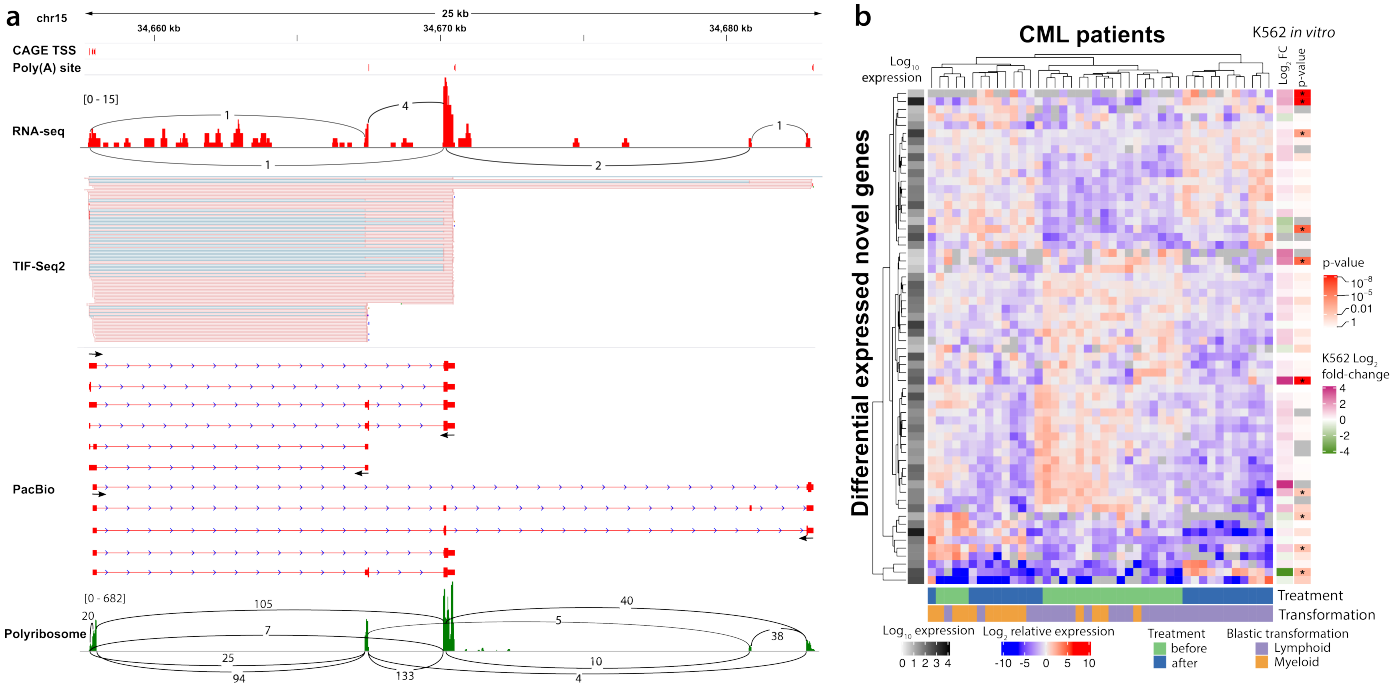

Figure 2. **Novel lowly expressed intergenic transcripts can be detected in CML patients.** a, TIF-Seq2 improve the genome annotation by identifying lowly expressed transcripts, for example, a novel gene on chr15: 34657737- 34683026. TIF-Seq2 read pairs are in pink. CAGE TSS and poly(A) sites (as in Fig. 1c) from public repositories [10,11] and an independent K562 RNA-seq experiment[20] validates the expression of this gene. It is downregulated after drug treatment in K562 cells. Primer designed for target long-read Pac Bio validation are marked in arrows. Polyribosome-associated RNA-seq[28] track is in the bottom, with expression marked in green and splicing junction in black lines, suggests its coding potential. b, expression of new transcription features in CML patient[21]. Each row represents a differentially expressed gene (adjusted p-value < 0.005) in CML patient. Each column represents a sample before or after drug treatment. Patients were classified as lymphoid blast crisis (LBC) or myeloid blast crisis (MBC) according to the types of their blastic transformations. Each cell represents log_2-scale ratio of gene expression in each sample to the average expression of the gene across all samples. The average expression (in log_{10} scale) of genes is depicted on the left. On the right side, we present the log_2-scale fold change of those genes in an independent *in vitro* experiment exposing K562 cells to imatinib[20] and their significant levels of differential expression after treatment. Genes in K562 cells with adjusted p-value < 0.01 are labelled in asterisks.

determine their internal structure, we used the TIF-Seq2 derived boundaries of 28 differentially expressed novel genes for designing amplicon-based enrichment of full-length isoforms followed by long-read sequencing (Supplementary Fig. 13). We confirmed the existence of 25 candidates and predicted the existence of short (29-179 aa) open reading frames in 17 cases (Fig. 2a and Supplementary File 1), suggesting their coding potential.

Once the novel intergenic transcript in K562 were annotated, we investigated their potential relevance in chronic myeloid leukaemia (CML) patients. We used the TIF-Seq2 derived annotation to re-analyse RNA-seq data from a cohort of CML patients before and after exposure to therapy, and focus on those who stopped responding to treatment and eventually developed blast crisis[21]. We confirm that 365 novel genes identified in K562 are also expressed in at least three CML patients with minimum one count per million reads. 59 of them respond to Imatinib treatment, and 10 are differentially regulated between patient groups that developed myeloid blast crisis (MBC) or lymphoid blast crisis (LBC) (Fig. 2b, adjusted p-value < 0.005). Interestingly transcripts downregulated by Imatinib treatment in K562 are upregulated in patients that stopped responding to treatment and underwent blast crisis. As an example, a K562-specific transcript on chr6:15555763-15559977 (putatively encoding the HTH domain of the Mos1 transposase) was significantly upregulated in patients who developed LBC after therapy, while it is significantly downregulated in K562 cell line after Imatinib treatment (Supplementary Fig. 14a-c). Independently of the potential relevance of the new identified transcripts for disease progression, it is clear that they are present but excluded from most analysis due to incomplete transcriptome annotation.

Identifying read-through transcripts

After showing the ability of TIF-Seq2 to better identify the boundaries of lowly expressed isoforms, we focused on the analysis of transcriptionally fused genes[22–24]. Gene fusions play a key role in oncogenesis[25], and although most known fusion transcripts arise through chromosomal rearrangements, they can also arise via transcription-induced chimeras (*e.g.* read-through or alternative splicing)[22]. Short-read RNA-seq can detect read-through transcripts by using the small proportion of reads connecting two neighbouring genes[23]. However, this becomes extremely challenging when the fusion event involves an intermediate exon not included in the annotations of the involved genes (*e.g.* LHX6-NDUFA8_03, Fig. 3a). Additionally, RNA-seq does not allow to identify which TSS of the upstream gene is connected to a particular PAS of the downstream gene. On the contrary, all TIF-Seq2 reads have the potential to detect such fusion events (not only the small fraction covering a splicing between the annotated genes) and define their complete boundaries (TSS to PAS). We discovered 29 non-annotated read-through transcripts in K562 cell lines, and once defined their boundaries, we were able to identify supporting splicing reads for all of them using RNA-seq[20] (Supplementary Table 5). However, without the additional information provided by TIF-Seq2, those same RNA-seq datasets were not sufficient to confidently classify them as read-through transcripts. Interestingly, during the preparation of this work, a few cases have been investigated in detail and independently reported as fusion transcripts in agreement with our findings[26,27].

To investigate up to what degree alternative splicing contributes to the appearance of those read-through transcripts, we further validated 10 candidates using the TIF-Seq2 derived boundaries and amplicon-based enrichment of full-length isoforms followed by long-read sequencing (as before, Supplementary Figure 13). This revealed an interleaved and complex transcriptome organization (Fig. 3a-b and Supplementary Fig. 15). In some cases, read-through transcripts connect the 5' regions of one gene with the coding sequence of the downstream one (e.g. C17orf99_SYNGR2_01, Fig. 3b), suggesting a regulatory rewiring of the downstream gene (*e.g.* putative regulators of C17orf99 could regulate SYNGR2 expression). While in other cases, the read-through transcripts have the potential to encode fusion proteins (*e.g.* SPN_QPRT_04, Supplementary Fig. 15). Our approach can detect even more complex scenarios, as is the case of transcripts initiating in the body of one gene but using the PAS of a downstream gene (e.g. *LHX6-NDUFA8* in Fig. 3a). Or other complex cases where the read-through transcript connect both genes using a non-annotated exon (*e.g.* LHX6-NDUFA8_03-06 or C17orf99_SYNGR2_04 in Fig 3).

To provide additional supporting evidence for the identified read-through transcripts, we investigated their potential to encode fusion proteins. We used RNA-seq from polyribosome associated mRNAs[28] and confirmed that reads connecting the novel splicing sites predicted by our long-read experiments can be identified in 10 read-through genes (Fig. 3a). This suggests that the identified transcripts associate to polyribosomes, and thus have the potential to encode fusion proteins. In addition, we investigated their presence in CML patients by re-analysing clinical RNA-seq data[21]. Even if we had to restrict our analysis to those few reads bridging the gene pairs, we were able to confirm their expression in patients (Supplementary Fig. 16). Thus, by combining an improved annotation using TIF-Seq2 with available clinical short read

dataset, we were able to identify read-through transcripts that would be ideal candidates for in deep molecular characterization in disease models.

Discussion

Here we have presented an improved approach, TIF-Seq2, designed to link transcription boundaries in complex genomes. This approach is in agreement with previous maps of TSS and PAS, and can enrich current genome annotations by clarifying complex arrangement of overlapping transcripts. By focusing its sequencing power on TSS and PAS, it can easily define the complete boundaries of lncRNAs or other lowly expressed transcriptional features. This allows also to link genes with putative regulatory regions and to facilitate its genetic manipulations (*e.g.* using CRISPRi or CRISPRa). Additionally, by defining the whole transcript boundaries, it opens the door for the design of target enrichment strategies to allow their study with other sequencing approaches. TIF-Seq2 is especially useful to dissect those overlapping transcripts challenging to investigate by RNA-seq. Here we focus on the identification of read-through transcripts, linking the usage of particular TSS and PAS from neighbouring genes. This reveals the ability of read-through transcripts to putatively rewire their regulation (*i.e.* usage of alternative shared upstream promoters) and produce fusion proteins. By combining TIF-Seq2 information with RNA-seq, we show how an improved annotation can improve our analysis of available clinical RNA-seq datasets. This will be especially important to prioritize those transcriptional features more relevant for in depth molecular validation. This approach would be also useful to clarify the full boundaries and post-transcriptional consequences of complex overlapping features such as cryptic intragenic TSSs[2] or promoter overlapping transcripts. Application of TIF-Seq2 goes beyond the re-annotation of the human genome, and it could be particularly useful to facilitate the annotation of less studied genomes. Finally, as TIF-Seq2 is based in short-read sequencing, in principle it could be easily combined with traditional target enrichment approaches to link unannotated, but experimentally validated, TSS or PAS sites (*i.e.* using a unique probe targeting TSS or PAS of interested, without the need to enrich the intervening region). By combining the respective advantages of short-read and long-read strategies, TIF-Seq2 has the potential to significantly improve the annotation of complex eukaryotic transcriptomes and thus to facilitate its understanding.

Data availability

All TIF-Seq2, 3'T-fill sequencing and long-read sequencing files were deposited in the Gene Expression Omnibus under accession number (GSE140912).

Code availability

The source code for TIF-Seq2 data analysis is available on GitHub (https://github.com/jingwen/TIFseq2)

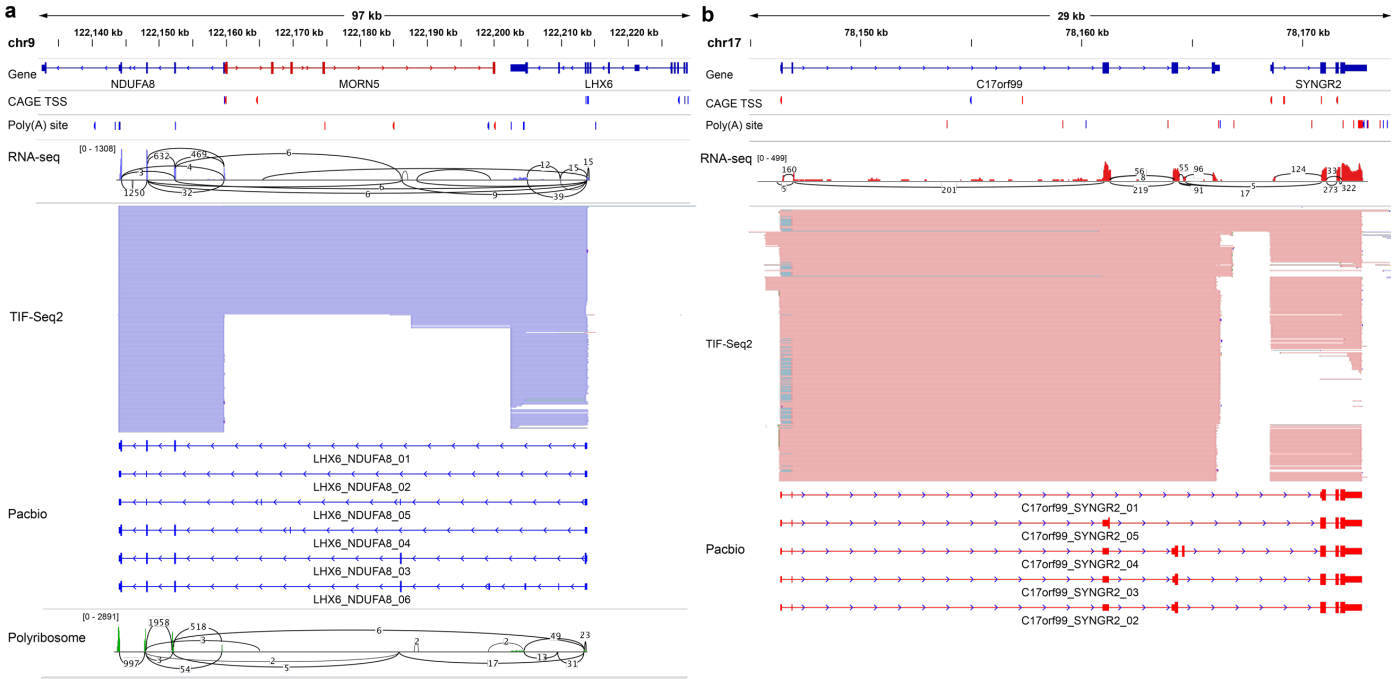

Figure 3. **TIF-Seq2 facilitates the identification of read-through transcripts.** a, *LHX6-NDUFA8* read-through gene; b, *C17orf99-SYNGR2* read-through gene. Annotated genes in Gencode v38 are listed on the top, followed by CAGE TSS and poly(A) site (PAS) tracks[10,11] as in Fig. 1c. RNA-seq validates the presence of splicing junctions (black lines with supporting number of reads) linking two adjacent genes. TIF-Seq2 track as in Fig. 1c shows the transcriptional fusion events between adjacent genes. Primers designed for target long-read PacBio validation are marked in arrows. PacBio Long-read sequencing of target transcripts validates the intergenic splicing events and dissect the transcription model of read-through genes. Polyribosome-associated RNA-seq[28] data are labelled in green, with splicing junction in black lines between two genes, showing the coding potential of two transcript isoforms of *LHX6-NDUFA8*.

Acknowledgements

We thank Donal Barrett for technical assistance and all members of Pelechano, Kutter and Friedländer laboratories for discussion. We thank Aino Järvelin and Judith Zaugg for early discussion. The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure (NGI) and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure. This study was financially supported by the Swedish Research Council (VR 2016-01842), a Wallenberg Academy Fellowship (KAW 2016.0123), the Swedish Foundations' Starting Grant (Ragnar Söderberg Foundation) and Karolinska Institutet (SciLifeLab Fellowship, SFO and KI funds) to V.P.; the National Key R&D Program of China (2017YFC0908405) and National Natural Science Foundation of China (Grant No: 81870187) and National Key R&D Program of China (2017YFC0908405) to W.W.; the US National Institutes of Health (NIH grant P01 HG000205), Deutsche Forschungsgemeinschaft (1422/4-1) and a European Research Council Advanced Investigator Grant to L.M.S., V.P and W.W. acknowledge the support from a Joint China-Sweden mobility grant from STINT (CH2018-7750) and the National Natural Science Foundation of China (Grant No: 81911530167) respectively.

Author Contributions

V.P., B.L., J.W., W.W. and L.S. conceived the study. B.L. and S.M. performed experiments and assisted with data interpretation. J.W., B.L., W.W. and V.P. designed data analysis and interpreted data. J.W. performed computational analysis. J.W., B.L. and V.P. drafted the original manuscript. All authors reviewed and edited the manuscript. V.P. supervised the whole project.

Competing Interests

The authors declare no competing interests.

References

1. Yao, P. *et al.* Coding region polyadenylation generates a truncated tRNA synthetase that counters translation repression. *Cell* **149**, 88–100 (2012).
2. Wei, W. *et al.* Chromatin-sensitive cryptic promoters putatively drive expression of alternative protein isoforms in yeast. *Genome Res.* (2019). doi:10.1101/GR.243378.118
3. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2017).
4. de Klerk, E. & 't Hoen, P. A. C. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet.* **31**, 128–139 (2015).
5. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nature Reviews Genetics* **19**, 535–548 (2018).
6. Chabbert, C. D., Eberhart, T., Guccini, I., Krek, W. & Kovacs, W. J. Correction of gene model annotations improves isoform abundance estimates: the example of ketohexokinase (Khk). *F1000Research* **7**, 1956 (2019).
7. Bertin, N. *et al.* Linking FANTOM5 CAGE peaks to annotations with CAGEscan. *Sci. Data* **4**, 170147 (2017).
8. Hon, C.-C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199–204 (2017).
9. Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46**, 582–592 (2017).
10. FANTOM Consortium and the RIKEN PMI and CLST, D. G. T. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
11. Gruber, A. J. *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).
12. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027 (2017).
13. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci.* 201806447 (2018). doi:10.1073/PNAS.1806447115
14. Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**, 127–31 (2013).
15. Matsumoto, K., Suzuki, A., Wakaguri, H., Sugano, S. & Suzuki, Y. Construction of mate pair full-length cDNAs libraries and characterization of transcriptional start sites and termination sites. *Nucleic Acids Res.* **42**, e125–e125 (2014).
16. Ruan, X. & Ruan, Y. Genome Wide Full-Length Transcript Analysis Using 5' and 3' Paired-End-Tag Next Generation Sequencing (RNA-PET). in *Transcriptional Regulation. Methods in Molecular Biology (Methods and Protocols), vol 809* 535–562 (Springer, New York, NY, 2012). doi:10.1007/978-1-61779-376-9_35
17. Pelechano, V., Wei, W., Jakob, P. & Steinmetz, L. M. Genome-wide identification of transcript start and end sites by transcript isoform sequencing. *Nat. Protoc.* **9**, 1740–59 (2014).
18. Wilkening, S. *et al.* An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res.* **41**, e65–e65 (2013).
19. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* (2017). doi:10.1038/ng.3988
20. Gallipoli, P. *et al.* Glutaminolysis is a metabolic dependency in FLT3ITD acute myeloid leukemia unmasked by FLT3 tyrosine kinase inhibition. *Blood* **131**, 1639–1653 (2018).
21. Branford, S. *et al.* Integrative genomic analysis reveals cancer-associated mutations at diagnosis of CML in patients with high-risk disease. *Blood* **132**, 948–961 (2018).
22. Akiva, P. *et al.* Transcription-mediated gene fusion in the human genome. *Genome Res.* **16**, 30–36 (2006).
23. Kumar, S., Razzaq, S. K., Vo, A. D., Gautam, M. & Li, H. Identifying fusion transcripts using next generation sequencing. *Wiley Interdisciplinary Reviews: RNA* **7**, 811–823 (2016).
24. Hu, X. *et al.* TumorFusions: an integrative resource for cancer-associated transcript fusions. *Nucleic Acids Res.* **46**, D1144–D1149 (2018).
25. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer* **15**, 371–381 (2015).
26. Weber, L. *et al.* Olfactory receptors as biomarkers in human breast carcinoma tissues. *Front. Oncol.* **8**, 33 (2018).
27. Wu, P. *et al.* The Landscape and Implications of Chimeric RNAs in Cervical Cancer. *EBioMedicine* **37**, 158–167

(2018).
28. Ramirez, O., Kesarwani, A., Abhishek, G., Minella, A. C. & Pillai, M. M. Integrative Analysis of RNA-Interactome and Translatome Reveal Functional Targets of MSI2 in Myeloid Leukemia. *Blood* **128**, 1881–1881 (2016).
29. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
30. Smith, T., Heger, A. & Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
31. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
32. Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: Precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, e51–e51 (2015).
33. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
34. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
35. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
36. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
37. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

Methods

Cell Culture

The human erythroleukemia cell line K562 was obtained from ATCC(ATCC®CCL-243™). Cells were cultured in RPMI 1640 medium supplemented with 10% FBS, 2 mM L-glutamine, 1% pen/strep (Life Technologies) at 37 °C with 5% CO2 in a humidified atmosphere. Cells (3 X 10^5 cells/mL) were exposed to 0.2 to 5µM (0.2, 0.5, 1, 1.7 and 5µM) imatinib for 8h, 24 h and 48h. Aliquots were taken at each time point for assessment of cell viability via Trypan blue staining by EVE™ automated cell counter. Two biological replicates of K562 cells treated with 1µM imatinib for 24h and corresponding DMSO control were used for TIF'Seq2 library preparation.

TIF-Seq2 Library preparation

In brief, capped and polyadenylated mRNA was used as template to generate full-length cDNA. We circularized the cDNA, remove non circularized molecules and fragmented circularized molecules using sonication. Streptavidin magnetic beads were used to purify the fragments spanning the 5' and 3' end of cDNA and then were used for Illumina library preparation (Supplementary Fig. 1). In detail, 2.5 µg total RNA was treated with Turbo DNase (0.12U/µl) (Fisher Scientific) for 20min at 37˚C to prevent genomic DNA contamination. After inactivation of DNase, input RNA was dephosphorylated by incubating with Alkaline Phosphatase, Calf Intestinal (0.3U/µl) (CIP)(NEB) at 37˚C for 30 min. Two rounds of phenol-chloroform extraction followed by ethanol precipitation was performed to remove CIP. After dephosphorylation, input RNA was decapped by incubating with Cap-Clip (0.125U/µl) (CellScript) at 37˚C for 60min. After phenol-chloroform purification and ethanol precipitation, RNA was ligated overnight at 16˚C at 5' with DNA/RNA chimeric oligonucleotide adaptor (TCAGACGTGTGCTCTTCCGATCTrNrNrWrNrNrWrNrN, TIF2-RNA in Supplementary Table 6 using T4 RNA ligase (NEB) in the presence of 10% dimethylsulphoxide (DMSO), RiboLock RNase Inhibitor (Thermo Fisher Scientific EO0382) and 1mM ATP. The chimeric TIF2-RNA adaptor introduced a common anchor sequence for forward primer of subsequent PCR amplification and an 8 mer unique molecular identifier (UMI). Ligated RNA was purified with 1.8:1 volumetric ratio (1.8X) RNA clean XP beads according to the manufacturer instruction and then used as the template for reverse transcription (RT). Ligated mRNAs were reverse transcribed using barcoded oligo-dT primers (i.e., TAGTTCAGTCTTCAGTACCTCGTGCGGCCGCXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTTTTTTTTTTTTTTTVN; where X refers to the specific barcode, TIF2-RT in Supplementary Table 6) which introduced Illumina sequencing primer 1, a 3' index, a Not1 endonuclease digestion site and a 3' common sequence for the subsequent PCR reaction. Specifically, mRNAs were mixed with the corresponding TIF2-RT oligo and dNTPs, denatured at 65˚C for 5min and then put on ice. The sample was mixed with 5x First strand buffer, Trehalose (1.57M) and RiboLock RNase inhibitor and incubated first at 42˚C for 2 min. Finally 2 µl of SuperScript™ III reverse transcriptase (Thermo Fisher Scientific) was added to each reaction and incubated at 42˚C for 50 min, at 50˚C for 30 min, at 55˚C for 30 min and inactivated at 70˚C for 15 min. Used RNA template was removed by incubating with 0.5ul RNase H(5U/µl) and 0.5µl RNase cocktail (Ambion) at 37˚C for 30min. First-strand cDNA was purified with 2X Ampure XP beads according to manufacturer instruction. To avoid saturating the reaction, only half of the obtained cDNA was used for the following PCR, and the rest was stored as a backup. Used cDNA template was further split into two PCR reactions with Terra PCR Direct Polymerase (Takara) with the following program 98˚C for 2min, then 16 cycles of 98˚C for 20s, 60˚C for 30s, 68˚C for 5min (+10s/cycle) and finally 72˚C for 5min. The PCR above used as primers TIF2-Rv: TAGTTCAGTCTTCAGTACCTCGT and TIF2-Fw: TATAGCGGCCGCXXXXXXGTGAC[BtndT]GGAGTTCAGACGTGTGCTCTTCCGATC (where X refers to different barcodes). PCR products were purified with 1X Ampure XP beads according to manufacturer instruction and then quantified with Qubit dsDNA HS assay. PCR products from different samples were then pooled together with equal mass. By pooling full length cDNA containing sample specific barcodes, we were able to detect and estimate the percentage of intermolecular circularization events (i.e. chimeras connecting barcodes originating from different samples). This information was used to select the optimal concentration favouring intramolecular ligation (see below). Pooled PCR products were subjected to 1U/µl Not1 HF (New England BioLabs) endonuclease digestion at 37˚C for one hour. After inactivation of Not1 at 65˚C for 20 min, samples were purified with 1.8x Ampure XP beads according to manufacturer instruction.

To favour intramolecular ligation, PCR products with sticky ends were highly diluted to a final concentration less than 1ng/µl and ligated with a high concentration (66.68U/uL) of T4 DNA ligase (New England BioLabs) at 16˚C for at least 16 hours. To remove the unligated linear PCR products, we added 0.5 µl plasmid-safe for every 100 µl total volume and incubated at 37˚C for 1h 1h in the presence of 1uL ATP 100mM. After inactivation at 70˚C for 30min, the self-circularized cDNA was purified with phenol-chloroform and ethanol precipitation. Circularized cDNA was fragmented by sonication (Covaris ME220; Covaris, Inc) (Duration 240s, peak power 30, duty factor 10, 200

cycles/burst, Avg Power(W) 3. The fragments were then purified with 1X Ampure XP beads according to manufacturer instruction. Biotin labelled fragments which contained the 5' and 3' connecting region was enriched using Dynabeads M280 streptavidin (Invitrogen) and incubating at room temperature for 30 min. Captured fragments were subjected to end repair with End Repair Enzyme Mix (New England Biolabs) and dA tailing with Klenow Fragment exo 5U/uL (New England Biolab). To add the required Illumina grafting sequences, each sample (20µl of resuspended beads) was incubated at room temperature for one hour with a mixture of 10 µl 5X Quick ligation buffer (New England BioLabs), 16µl nuclease-free water (Ambion), 3 µl T4 DNA ligase(2000U/µl) and 1 µl 1µM duplex adaptors. Duplex adaptors were generated by annealing TIF2-forkFw (5'-AATGATACGGCGACCACCGAGATCTACACACACCTGCCGGTCACC*T-3) and TIF2-forkRv (5'-phos-GGTGACCGGCAGGTGTATCTCGTATGCCGTCTTCTGCTTG-3') at 15µM and diluted to 1µM working solution freshly when used each time. After cleaning the beads, the beads resuspended in 20µL EB buffer were used for PCR amplification using 25µl Phusion High-fidelity MasterMix (2X) (New England BioLabs), 4 µl nuclease-free water and 0.5 µl PCRgraftP5 primer (5uM, 5'-AATGATACGGCGACCACCGAGATCTACAC-3') and 0.5 µl PCRgraftP7 primer (5uM, 5'-CAAGCAGAAGACGGCATACGAGAT-3'). Samples were subjected to PCR amplification with the following program: 98˚C for 30s, then 18 cycles of [98˚C for 20s, 65˚C for 30s, 72˚C for 30s] and finally 72˚C for 5min. DNA library then went through two step beads-based size selection (first step 0.35X, second step take the supernatant from first step and add extra 0.45X) with expected size distribution in the range of 300bp-1000bp. Purified DNA library was adjusted to 4nM, then denatured, diluted according to Illumina instruction and sequenced on Nextseq 500 platform.

We used four custom sequencing oligos as follows: SeqR1+15T (5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTTTTTTTTTTTTTT-3'), SeqINDX1 (5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3'), SeqINDX2 (5'-GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3') and SeqR2 (5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'). See Supplementary Figure 2 for details regarding the sequencing oligonucleotide annealing. Two mixtures of sequencing oligos were prepared as follows: mixture 1 contained SeqR1+15T(0.3µM) and SeqR2(0.3µM), mixture 2 composed of SeqINDX1 primer (0.3µM) and SeqINDX2 primer (0.3µM). Mixture 1 was loaded into both positions (#7 and #8 on reagent cartridge in a NextSeq 500 instrument) for custom read1 primer and for custom read2 primer. Mixture2 was loaded into the position for custom index primer (#9 on reagent cartridge). Sequencing was carried out in an Illumina NextSeq 500 instrument with stand-alone configuration and custom sequencing oligos. Paired end sequencing read lengths were set as read1 76bp, read2 76bp, index1 6bp and index2 6bp.

3'T-fill library preparation

We performed 3'T-fill as previously described[18]. In brief, 10µg DNA-free total RNA (16µL) was fragmented by adding 4µl fragmentation buffer(5X)(200mM Tris-acetate, pH 8.1, 500mM Potassium Acetate, 150mM Magnesium acetate) and incubated at 80˚C for 5min. Fragmented RNA was purified with 1.5X Ampure XP beads according to manufacturer instruction and used as the template for reverse transcription(RT). Sample were mixed with biotinylated dT primer (P5_dT16VN 5'-[Btn]AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTTTTTTTTTTTTTTTVN-3') (where V refers to A, C or G) and denatured at 65˚C for 5min. Denatured RNA was then mixed with 4µl 5X first strand buffer, 2µl DTT(0.1M) and 4µl freshly prepared actinomycin D(0.1µg/ul). After incubation at 42˚C for 2min, 0.5µl SuperScript™ II (Invitrogen) was added, incubated at 42˚C for 50 min and finally 72˚C for 15min. cDNA was purified with 1.5X Ampure XP beads according to manufacturer instruction. For second strand synthesis, cDNA was mixed with of 0.5 µl RNase H (5U/µl) (New England BioLabs) and 2 µl DNA Polymerase I (10U/µl) (Thermo Scientific) and incubated at 16˚C for 2.5 hours. cDNA was purified with 0.9X Ampure XP beads according to manufacturer instruction. 20 µl biotin labelled cDNA was mixed with 20µl Dynabeads M280 streptavidin beads (Invitrogen) at room temperature for 15 min. Following end repair and dA tailing, each sample(8 µl) was mixed with 12.5 µl 2X Quick Ligation buffer (New England BioLabs), 2.5µL T4 DNA ligase (2000U/µL, New England BioLabs) 2µl of anneal duplexed adaptors at 2.5 µM(5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3', 5'-Phos-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[AmC7]-3') and incubated for at 20˚C for 20min. Adaptor ligated libraries were washed and amplified by PCR by adding 25 µl 2X Fusion High Fidelity Master Mix, 0.5 µl PE1 primer (10µM, 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3')and 0.5µl PE2_MPX primers (10µM, 5'-CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'; where X refer to the specific barcode). Samples were subjected to the following program: 98˚C for 30 seconds, 18 cycles of 98˚C for 10s, 65˚C for 10s, 72˚C for 10s and finally 72˚C for 5min. Library was purified

with 1.8X Ampure XP beads according to manufacturer instruction.

PacBio long reads sequencing

To validate the existence and structure of novel transcript isoforms, we used the TIF-Seq2 derived information to perform targeted amplification followed by Pacific Biosciences (Pac Bio) sequencing. To decrease the costs associated to library preparation, we performed a pooled reverse transcription with a mix of gene specific primers and then split the obtained cDNAs for individual PCR reactions. We used as input 2 biological replicates of K562 exposed or not to imatinib (4 samples, as in the TIF-Seq2). Each sample was individually labelled during the PCR reaction (see below). We pooled equal amount of all isoform-specific RT primers (Supplementary Table 6) and used total RNA as a template for reverse transcription (RT). Isoform-specific RT primers (PB_RT) were designed with a universal primer sequence (5'-GTGACTGGAGTTCAGACGTGT), plus 8 random nucleotides as unique molecular identifier and a gene specific sequence. RNA was mixed with primer and dNTPs, denatured at 65˚C for 5min and transferred to ice. For the reverse transcription we added first strand buffer, Trehalose (1.57M), RiboLock RNase Inhibitor and the mix was incubated at 42˚C for 2min, and then 2 µl of SuperScript II reverse transcriptase (ThermoFisher Scientific) was added to each reaction. We incubated each reverse transcription reaction at 42˚C for 50min, 50˚C for 30 min, 55˚C for 30min and inactivated at 70˚C for 15min. Obtained cDNA was distributed in 96-well plates containing in each well an isoform-specific forward primer (PB_FW, composed of a common sequence 5'- ACACTCTTTCCCTACACGAC and gene specific sequence) and a common reverse primer containing 8 mer sample identifier (FigS13). cDNA was PCR amplified using the Phusion High Fidelity Master Mix with the following program: denaturation at 98˚C for 2min, then 98˚C 20s, 60˚C 30s, 68˚C 5min (+10s/cycle) for 35 cycles and final extension was performed at 72˚C for 5min. Amplified products were analysed by gel electrophoresis, purified by 1.8X Ampure XP beads, and pooled at similar concentration for sequencing. Pooled PCR products were used for library generation using SMRTbell™ Template Prep Kit 1.0-SPv3 and sequenced on PacBio Sequel system.

TIF-Seq2 and 3'T-fill sequencing processing and alignment

We employed bcf2fastq (v2.20.0) for converting raw image to sequence information (FASTQ files) and demultiplex, allowing two mismatches in index 1 and one mismatch in index 2. For TIF-Seq2 data, we collapsed all 5'-end or 3'-end sequence reads in each sample according to the indexes. Cutadapt[29] (v1.16) was utilized to trim TIF-seq2 sequencing primer (-a AGGTGACCGGCAGGTGT) and Illumina TruSeq adapter (-a AGATCGGAAG). After extracting 8 bp of unique molecular identifiers (UMIs) with UMI-tools[30] (v0.5.4) from the 5' ends and removing extra A stretches in the 3' ends caused by poly(A) slippage during PCR amplification, we kept reads over 20 bp for alignment. We used STAR[31] (v2.5.3a) for aligning 5'-end reads and 3'-end reads separately to the human reference genome hg38, supplying Gencode v27 transcripts as splicing junction annotation. Alignment setting was adjusted as below, --alignIntronMax 200000 --alignEndsType Extend5pOfRead1 --alignSJoverhangMin 10. We then linked paired-end reads and kept the uniquely mapped pairs that are on the same chromosome (Supplementary Fig. 4). Furthermore, a customised script adapted from UMI-tools was employed to remove PCR duplicates from the leftover reads, allowing 1 bp mismatch in the UMIs and 1 bp shifting in the transcription start sites (https://github.com/jingwen/TIFseq2/blob/master/dedup.py).

For 3'T-fill sequencing data, we trimmed Illumina TruSeq adapter (-a AGATCGGAAG) and extra A stretches in the 3' ends with cutadapt v1.16. Reads over 20 bp are aligned to hg38 by using STAR v2.5.3a in paired-end mode, supplying Gencode v27 transcripts as splicing junction annotation with adjusted setting, --alignSJDBoverhangMin 1 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --alignEndsType Extend5pOfReads12. Only uniquely mapped reads were kept for downstream analysis.

Transcription boundary determination

A customised python script was employed to extract both boundaries of TIF-Seq2 read pairs and 3' tags of 3'T-fill sequencing data, collapse the boundary tags and calculate the coverage (https://github.com/jingwen/TIFseq2/blob/master/boundary.py). In order to filter out false positive poly(A) sites caused by internal priming, we exclude the 3' end tags with at least 7 As in the downstream 10 nt sequences. Then we employed CAGEr[32] to define the cluster of transcripts 5'- or 3'-end tags of TIF-Seq2 respectively (Supplementary Fig. 4b). Transcription boundary tags were normalized to match a power-law distribution. Low-coverage tags supported by less than 1 normalized counts in more than 1 sample were excluded before clustering. The boundary tags within 10 bp window were spatially clustered together. Clusters with only one boundary tag are kept if the normalized counts are above 1. The tag clusters were further form into non-overlapping consensus clusters across all samples if they are within 10 bp apart. The same strategy was applied for identifying consensus PAS cluster from 3'T-fill sequencing data.

TIF definition, annotation and quantification

We then link the 5'-end TSS and 3'-end PAS clusters according to the supporting read pairs from TIF-Seq2 (Supplementary Fig. 4b). We filtered out pairs with extremely long (> 2Mb) and extremely short (< 300 bp) mate-pair distance. To keep a conservative estimate of novel transcript isoforms identified, we excluded pairs mapping to different chromosomes. Transcript isoform boundaries (TIFs) were defined as connection between TSS clusters and PAS clusters supported by at least 4 read pairs connecting them across all samples (unique molecular events). The TIFs were further assigned to Gencode v28 annotation features based their relative distance to the annotated transcripts (Supplementary Fig. 4c). TSS distance (d1) and PAS distances (d2) are calculated between a TIF and its overlapping annotated transcripts. The TIF is assigned to the transcript with the least sum of d1 and d2 among all overlapping transcripts, further assigned to the gene that harbour the transcript. According to the relative position to their assigned transcripts, the TIFs were classified as 1) annotated transcripts, if both TSS and PAS are within 200 bp away of annotated transcripts boundaries; 2) transcripts with new TSS; 3) transcripts with new PAS; 4) transcripts with new boundaries, if both TSS and PAS not annotated and 5) intergenic TIFs (Supplementary Fig. 5). We measured TIF-Seq2 expression in K562 cells as count of read pairs that link TIF boundaries. We employed DEseq2[33] for normalization and differential expression analysis (before and after drug treatment) with default setting.

Long-read sequencing data analysis

We trimmed the PCR primers from both ends of the PacBio highly accurate consensus sequences using cutadapt (v1.16) and extract UMIs with UMI-tools (v1.0.0), keeping the reads with at least 200 bp in length. Then the reads were aligned to human reference genome using minimap2[34] (v2.16) with the following setting (-ax splice -uf -C5 -O6,24 -B4). We used UMI-tools for removing the PCR duplicates with adjusted setting as --method cluster --spliced-is-unique. We further employed BEDTools[35] (v2.27.1) bamToBed function to convert the alignment reads into BED format and a customised script to convert BED format to GFF format.

Independent RNA-seq validation

RNA-seq data of K562 cells before and after Imatinib treatment (n=4) from study Gallipoli et al.[20] were download from GEO depositories (GSE105161). RNA-seq data of 21 paired CML patient[21] before and after Imatinib treatment from were downloaded from EGA archive (EGAD00001004179). We employed STAR (v2.5.3a) to align paired-end reads to human reference genome hg38, supplying Gencode v28 and long-read sequencing validated transcripts as splicing junction annotation, with adjusted setting (--alignSJDBoverhangMin 1 --alignIntronMax 200000 --alignEndsType Extend5pOfReads12). Polyribosome profiling data of K562 cells[28] (n=3) was downloaded from GEO (GSE93210). We used HISAT2[36] (v2.1.0) for aligning the reads to human reference genome hg38 with Gencode v32 and long-read sequencing validated transcripts as splicing junction annotation, meanwhile adjusting maximum intron length to 200kb.

Gene expression in K562 cells[20] and CML patients[21] from standard RNA-seq was quantified using featureCounts[37] according to Gencode v28 transcript annotation, TIF-Seq2 transcription boundaries and long-read sequencing validated transcript model. We employed DEseq2[33] for normalization and differential expression analysis (before and after drug treatment) with default setting for K562 cells. For CML patient RNA-seq data, we set up multiple factors (paired samples, drug treatment and phenotype) in DESeq2 for differential expression test.