Genome Biology

OPINION Open Access

Bridging the gap between reference and real transcriptomes

Antonin Morillon[1] and Daniel Gautheret[2*]

Abstract

Genetic, transcriptional, and post-transcriptional variations shape the transcriptome of individual cells, rendering establishing an exhaustive set of reference RNAs a complicated matter. Current reference transcriptomes, which are based on carefully curated transcripts, are lagging behind the extensive RNA variation revealed by massively parallel sequencing. Much may be missed by ignoring this unreferenced RNA diversity. There is plentiful evidence for non-reference transcripts with important phenotypic effects. Although reference transcriptomes are inestimable for gene expression analysis, they may turn limiting in important medical applications. We discuss computational strategies for retrieving hidden transcript diversity.

Reference transcriptomes: the making of

Reference transcriptomes (RefTs) aim to provide a comprehensive picture of transcripts produced by an organism. Early RefTs were produced at the turn of the century based on sanger sequencing of full-length cDNAs (flcDNA) [1–3]. Later on, projects such as ENCODE, modENCODE, and FANTOM5 harnessed the power of massively parallel cDNA sequencing (RNA-seq) to accelerate transcript discovery in multiple species and tissues. Due to limited RNA-seq read size (approximately 100 nucleotides), these efforts had to include additional technologies to guarantee accurate full-length transcript assembly. For instance, the FANTOM5 RNA-seq based human cDNA collection was assembled with assistance of the CAGE technology to identify RNA 5′ ends, ENCODE transcript sets were based on RNA-seq and rapid amplification of cDNA ends (RACE) technologies [4], and the fly and *Caenorhabditis elegans* ModENCODE sets combined RNA-seq, RACE, and expressed sequence tag (EST) sequencing [5, 6]. In yeast, major transcriptomics efforts have involved CAGE, TIF-seq, high coverage paired-end RNA-seq (both total and poly(A)+) and 3′-end tags, covering both stable and cryptic transcripts [7–10] . A third generation of transcriptomics projects now combines single-molecule, long-read sequencing technologies with short-read sequencing. Long-read-based datasets are now available for human [11, 12] and several plants [13, 14] and new sets of high-quality full-length transcripts are expected for all model species

Major genome databases integrate sequence data from the above sources into non-redundant, curated transcript datasets (Fig. 1). RefSeq [16] and Ensembl [15] are pan-species databases that implement a homogenous computational annotation workflow combining assembled high-throughput data and manually curated transcripts when available. Specialized RefTs such as Gencode for human and mouse [17, 22], Wormbase for *C. elegans* [18], Flybase for *Drosophila* [19, 23], and Araport for *Arabidopsis* [20], are produced through a combination of manual curation of full-length transcript collections from various origins and dedicated short-read assembly software. The Saccharomyces Genome Database [21] does not provide a set of full-length transcript sequences; however, RefSeq and Ensembl provide RefTs for yeast.

The most striking lessons drawn from large-scale transcript sequencing have been the widespread expression of long non-coding RNA genes and the abundance of alternative transcripts. This is well reflected in the number of genes and transcripts in current genome annotations (Fig. 1). For instance the human Gencode RefT now harbors 58,721 genes (that is, three times more than coding genes) and a transcript-to-gene ratio of 3.52.

Enter direct RNA-seq assembly

While current transcript counts in RefTs may seem impressive, these datasets have actually grown relatively slowly, constrained by their rigorous curation process. For instance, Gencode has grown from 161,000 human

* Correspondence: daniel.gautheret@u-psud.fr
[2]Institute for Integrative Biology of the Cell, CEA, CNRS, Université Paris-Sud, Université Paris Saclay, Gif sur Yvette, France
Full list of author information is available at the end of the article

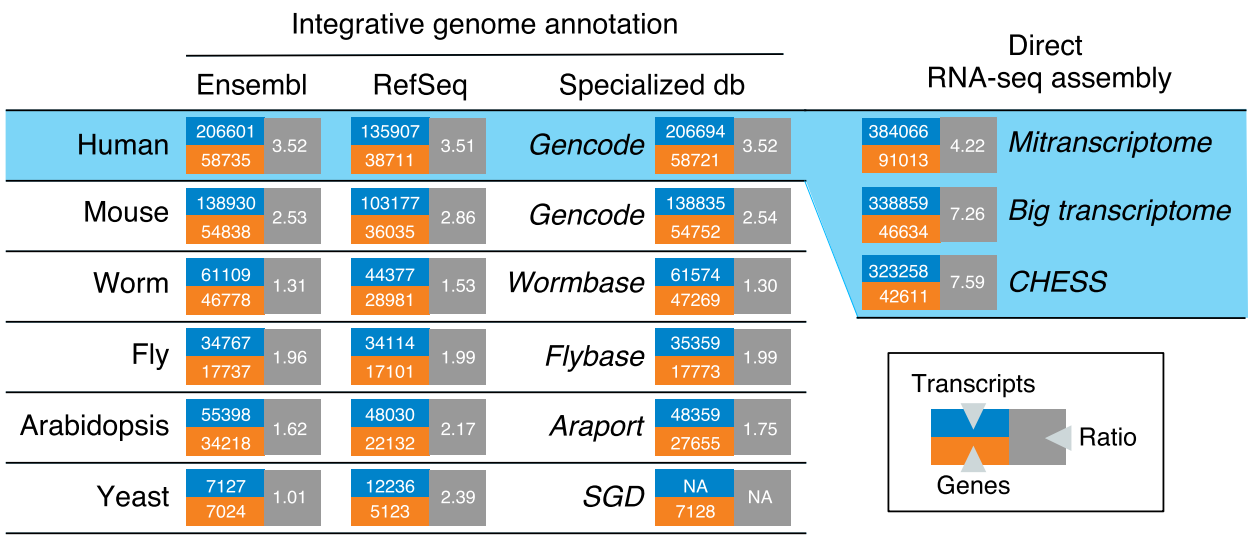

Fig. 1 Contents of major reference transcriptomes for human and model eukaryotes. Versions of databases used: Ensembl [15], V95; RefSeq [16], human rel 109, mouse rel 106, worm rel WS268, fly rel 6.18, arabidopsis rel TAIR10.1, yeast rel R64–2-1; Gencode [17], Human V29, mouse M20; Wormbase [18], WS268; Flybase [19], r6.26; Araport [20], V11; Saccharomyces Genome Database (*SGD*) [21], V20150113. Database URLs and additional curation information are provided in Additional file 1: Table S1

transcripts in 2012 to 207,000 now, i.e., a 29% growth in 7 years. In the meantime, projects generating raw RNA-seq data have exploded. Three projects alone, The Cancer Genome Atlas (TCGA) [24], GTEX [25], and Human Protein Atlas (HPA) [26], have produced 31,000 RNA-seq libraries covering normal and cancerous tissues from thousands of individuals (Additional file 1: Table S2). Raw RNA-seq datasets have been reanalyzed by direct RNA-seq assembly projects such as miTranscriptome [27], BigTranscriptome [28], and CHESS [29]. These computational protocols, which do not implement the strict validation process used for RefTs, led to a 55–85% growth of the number of annotated human transcripts (Fig. 1; Additional file 1: Table S1). Nevertheless, the largest sets used in direct computational assembly are still 40 times smaller than public RNA-seq databases (over 400,000 human libraries in SRA [30] and ENA [31]). This vast wealth of RNA-seq data contains extensive transcript variation that is not yet included in RefTs. Therefore, a deep information gap may be building up between slow moving RefTs and yet undiscovered RNA variants from short read data.

We describe below the different types of transcript variations that may be missing from RefTs. We contend that the information gap between RefTs and high-throughput data is not going to be closed. Based on multiple evidence gathered from medical transcriptome studies, we argue that non-reference transcript information is highly significant and its neglect limits our understanding of genotype–phenotype relationships. This underlines the need for computational methods that can extract non-reference events from RNA-seq data.

Shall we ever reach a complete reference transcriptome?

Each cell of an organism produces a distinct set of transcripts. Transcriptome differences between cells stem from three mechanisms that are potentially cumulative (Fig. 2). First, genetic variation occurs across individuals in a population as well as within each individual through aging and cancer. This includes a vast range of variation, from single nucleotide substitutions and indels to mobile element insertion and large chromosomal rearrangements. Second, transcriptional regulation programs are implemented during organism development and cell differentiation. These comprise all variations of transcription activity, whether in intensity, start site, or strandedness. Third, post-transcriptional regulations, including a wide array of RNA processing, editing, base modification, and cleavage/degradation mechanisms, are specific to cell type, cell compartment (e.g., splicing in the nucleus), and environmental conditions. It is worthy to note that transcriptomic complexity is not limited to higher eukaryotes, as illustrated by the discovery of bi-directional promoters [9, 32] and cryptic transcripts [7] in yeast.

Most individual RNA variations do not find their way into RefTs. An analysis of splice junctions in approximately 21,500 human RNA-seq libraries from SRA [33] identified over three million junctions supported by at least 20 reads, which is nine times more than found in Gencode transcripts. Yet, the analysis did not include the restricted access TCGA [24] dataset. Considering the importance of aberrant splicing in cancer [34] and other diseases [35], one may expect RNA-seq data from

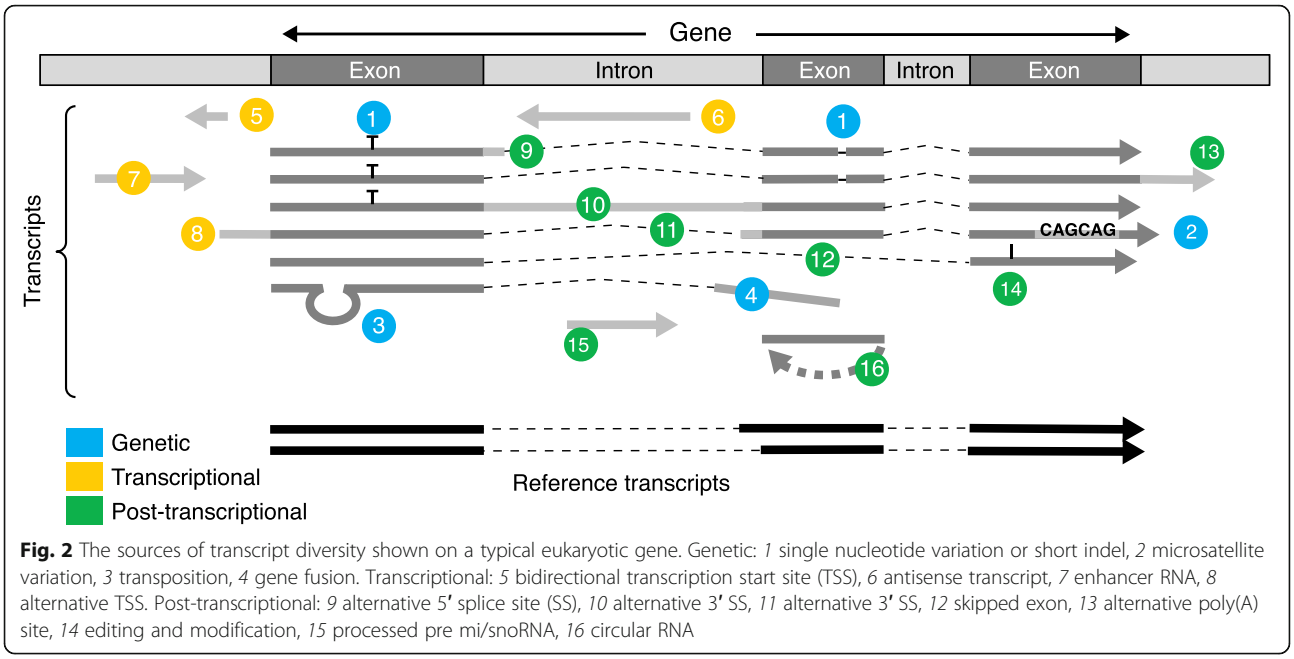

Fig. 2 The sources of transcript diversity shown on a typical eukaryotic gene. Genetic: *1* single nucleotide variation or short indel, *2* microsatellite variation, *3* transposition, *4* gene fusion. Transcriptional: *5* bidirectional transcription start site (TSS), *6* antisense transcript, *7* enhancer RNA, *8* alternative TSS. Post-transcriptional: *9* alternative 5′ splice site (SS), *10* alternative 3′ SS, *11* alternative 3′ SS, *12* skipped exon, *13* alternative poly(A) site, *14* editing and modification, *15* processed pre mi/snoRNA, *16* circular RNA

pathological samples to yield large quantities of novel variations. National medical genomics projects will deliver millions more individual sequence sets, including RNA-seq, raising the question of whether these data should eventually be incorporated into RefTs.

One last important factor limiting RefT completeness stems from the nature of RNA libraries analyzed (Additional file 1: Table S3). RefTs are based primarily on poly(A) + libraries, which are far from encompassing all transcripts and present quantitative and qualitative bias related to poly(A) retention efficiency [36]. Alternative RNA selection protocols, including ribo-depleted RNA-seq, nascent RNA-seq, capture-seq, small RNA-seq, M6A-seq, and compartment-specific RNA-seq [37–40], have already revealed large quantities of previously hidden RNAs. The ability to sequence modified RNA bases will add yet another dimension to transcriptomics. As RNA modifications cause abortive reverse transcription, specific protocols are needed to either allow bypass of modified bases or recovery of aborted cDNAs [41]. Alternative strategies involving direct sequencing of modified RNA with the Nanopore technology are still under development.

The above observations are in line with recent studies that have underlined the difficulty of ever completing a mammalian transcriptome. Uszczynska-Ratajczak et al. [42] showed large-scale lncRNAs catalogues are far from converging while Deveson et al. [43] conclude from their analysis of alternative splicing of non-coding exons that "there does not exist a finite list of noncoding isoforms that can be feasibly catalogued".

Ignore non-reference transcripts at your own risks

It may be argued that non-reference transcripts are predominantly transient or expressed at a low level and therefore can be ignored as transcriptional [44] or splicing [45, 46] noise. The function of pervasive, intergenic transcripts has been particularly disputed on this basis [47–49]. Although pervasive transcription is now recognized as a source of de novo gene birth [50, 51] and thus may be important for a species as a whole, it is obviously difficult to speculate or raise much interest about future gene functions. A more sensible approach to establish function is arguably that taken by evolutionary biologists who use negative selection as an evidence for function. Selection measures based either on phylogenetic conservation [52] or allele frequencies in populations [53] are converging towards 4–9% of the human genome under selection, which is to be compared with the 1.5% coding fraction. Predicted functional regions include about 130 Mb which are either expressed (mRNA and lncRNA exons and introns) or potentially expressed (enhancers, transposable elements, pseudogenes) [52]. One can reasonably propose that any transcript variation altering these regions, whether genetic, transcriptional, or post-transcriptional, may impact phenotype.

An alternative way to appreciate the biological impact of non-reference transcripts is to consider transcript alterations in human diseases. The list of disease-causing or disease-related transcripts that are not part of the RefT is a long one (Additional file 1: Table S2). Chimeric transcripts [54] and viral transcripts from integrated or free virus, such as human papillomavirus (HPV) [55],

are important cancer drivers which are not included in RefTs. Aberrant splicing is a source of key drivers in cancer [56] and other diseases [35, 57]. Alternative polyadenylation events contribute to human disease and are connected with development, cell differentiation, and proliferation [58]. Intron retention events are considered as novel disease factors [59, 60]. Reactivated transposable elements and retrotransposed mRNAs are involved in tumorigenesis [61] and Alzheimer's disease [62]. Rearranged T-cell receptor transcripts are used to monitor T-cell clonal expansion in tumors [63]. Both A-to-I RNA editing events and M6A base modifications contribute to cancer progression [37]. Two abundant classes of non-reference RNAs, circular and antisense RNAs, have been involved in gene regulation [64] and used as disease biomarkers [65]. Lastly, genetic polymorphism in transcripts, whether in the form of single-nucleotide variants, short indels, or microsatellite expansion, may strongly impact RNA processing, stability, and translation. An extreme illustration is the CAG repeat expansion in the HD gene at the origin of Huntington's disease [66]. Although sequence polymorphisms are generally ignored in transcriptome studies, taking into account this dimension should lead to a better understanding of the potential impact of transcripts on phenotypes, as the medical community enters the "personal transcriptome" era [35, 67].

RNA-seq analysis in the personal transcriptome era

RNA-seq data analysis commonly involves mapping reads to an annotated genome or a RefT to quantify transcript and gene expression [68]. These protocols do not permit detection of novel transcripts and may lead to inaccurate expression measures due to incomplete transcript annotations [69]. A straightforward improvement to quantification protocols is to replace a RefT with an extended catalogue generated by direct RNA-seq assembly, as available for human [27–29]. This may work satisfyingly when studying datasets similar to those from which the catalogue originated (TCGA, GTEX, etc.). However, these catalogues have shown large divergences [42] and thus do not guarantee that events present in an arbitrary RNA-seq experiment are covered. The only way to ensure this is to implement a RefT-free strategy.

Figure 3 presents a selection of RefT-free software pipelines for RNA-seq analysis. As a guide for users, the figure shows whether pipelines are limited to small numbers of initial libraries (here arbitrarily shown as < 20) or can scale to hundreds of libraries. Two other highlighted

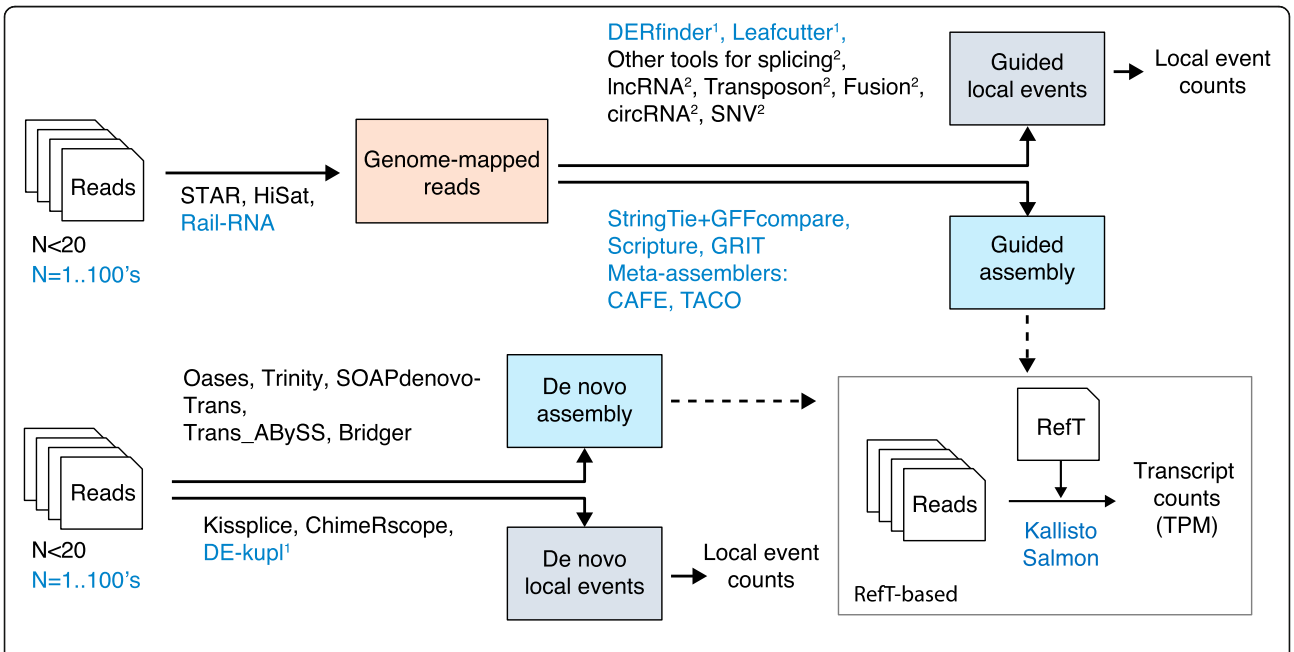

Fig. 3 RNA-seq pipelines for the discovery and quantification of transcripts and processing events, unconstrained by a RefT. Software in *black* are best suited to "small" input datasets (represented by an arbitrary size $N < 20$) while software in *blue* can process large datasets (up to hundreds of libraries). Protocols are subdivided into four combinations of genome-guided versus de novo and assembly-based versus local event discovery. Local events include splice variants, transcribed regions, gene fusions, circular RNAs, sequence polymorphisms (*SNV*) and expressed transposons (Additional file 1: Table S2). Results from assembly software can be used as RefTs in standard quantification pipelines (*inset*)

differences between strategies are (i) whether or not they attempt full-length transcript assembly and (ii) whether they are genome-guided or de novo methods.

Assembly software predict full-length transcripts either de novo from raw RNA-seq data [70–72] or following genome alignment [73–76]. Major motivations for using assembly software are transcript quantification and analysis of protein-coding potential. De novo assembly is computationally demanding and is mostly used with small datasets and when a reference genome is unavailable. On the other hand, genome-guided assemblers can be applied iteratively to hundreds of RNA-seq libraries. However, a major limitation in all assembly processes stems for their reliance on splicing graph analysis, which has a relatively high error rate that grows with the number of reads analyzed [77–79]. As said by Hayer et al. [78], "with more reads, most algorithms find more ways to go wrong". The assembly of large datasets is thus performed stepwise, first by assembling individual libraries and then using meta assemblers [28, 29, 80] to merge results. Of note, some assembly protocols are able to use transcript boundary information from CAGE and 3′-seq data to improve assembly quality [76, 80].

Transcript assembly is not the most adequate route in many situations. First, individual transcript variations such as alternative transcription start sites and splicing/polyadenylation events are under-represented in predicted full-length transcripts [81]. Second, assembled transcripts are especially unreliable with certain RNA classes such as the weakly expressed, highly heterogeneous lncRNAs [82]. Third, certain RNAs, such as fusion or circular RNAs, are generally absent from genome-guided assemblies. Therefore, non-canonical or alternative transcription is often best studied using strategies that bypass assembly altogether and focus solely on specific variations recovered from the genome mapping (BAM) files. This category includes powerful software such as LeafCutter [83] for splice site discovery and DERfinder [84] for the characterization of lncRNAs and alternative mRNA boundaries. Other software tools are able to use partly mapped or unmapped reads for the recovery of gene fusions, circular RNAs, single-nucleotide variants, and expressed transposons (Fig. 3; Additional file 1: Table S4).

Genome-guided procedures assume that all samples under study have the same genetic makeup. This does not hold when RNA-seq data come from individuals with significant genetic divergences or from samples harboring somatic structural variations. Transcripts expressed from variable regions may erroneously map to the reference genome, leading to incorrect transcript assemblies and counts. An emerging class of software, including Kissplice [85], ChimerScope [86], and DE-kupl [87], avoid both genome alignment and transcript reconstruction through direct mining of the k-mer (subsequence of fixed size) contents of the original sequence files. These are promising approaches that apply particularly to cases where a reference genome cannot be relied upon.

Concluding remarks

In spite of continuous updates, RefTs are not catching up on short-read RNA-seq data in their coverage of transcript diversity. Single molecule (long-read) RNA sequencing will help improving RefTs faster than current technologies that require capture of cDNA ends in complement to short reads. However, the combinatorial nature of transcript variation, the higher yield of short-read sequencing, and the huge diversity of tissues, diseases, and transcript classes probed by short-read sequencing make it unlikely that RefTs will ever match the level of diversity observed in short read data.

Of note, limitations of RefTs are in a large part intentional. Indeed, these databases are manually curated to exclude a majority of pervasive transcripts resulting from expressed repeats, pseudogenes, or erroneous splicing. Transcript catalogues computationally generated from thousands of RNA-seq libraries apply less stringent inclusion criteria and are poised to include a large fraction of non-functional and pathological products, as well as incorrect boundaries and exon structures [11, 77].

Well-curated RefTs are essential resources for measuring gene expression. RefT-based gene expression analyzes are now highly efficient [88, 89], provide accurate gene expression measures [90], and can be functionally interpreted via multiple resources for gene ontology and pathway analysis. For these reasons, RefTs will remain a major tool for transcriptomics. Functional analysis of non-reference transcripts is more hazardous as many are non-coding and there is no commonly accepted way to annotate their function. Yet, their impact should not be underestimated. The aforementioned examples taken from human diseases reveal a wide diversity of non-reference transcripts with phenotypic effects. Even though these transcripts might be of low abundance, they can be essential in understanding genotype–phenotype relationships and should not be ignored.

There is no consensus on the most efficient RNA-seq analysis protocols for characterizing and quantifying non-reference transcripts. Strategies focusing on local or regional transcript variations are a powerful way to circumvent limitations related to full-length assembly. Such methods can be combined to conventional RefT-based analysis to achieve a complete description of normal and aberrant transcript forms present in a set of RNA-seq libraries.

Additional file

Additional file 1: Table S1. Overview of major eukaryotic transcriptome databases. **Table S2.** Large-scale RNA-seq projects (human). **Table S3.** Sequencing methods providing insight on specific events shown in Fig 2. **Table S4.** Transcript variations related to cancer and other diseases; and software for retrieving these variations from RNA-seq data. (XLSX 18 kb)

Abbreviation

RefT: reference transcriptome

Funding

This work was funded in part by ANR-18-CE45–0020 "Transipedia" to DG. AM's lab is supported by the European Research Council (ERC CoG - GA616180 - DARK).

Authors' contributions

AM and DG developed the idea and wrote the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]ncRNA, Epigenetic and Genome Fluidity, CNRS UMR 3244, Sorbonne Université, PSL University, Institut Curie, Centre de Recherche, 26 rue d'Ulm, 75248 Paris, France. [2]Institute for Integrative Biology of the Cell, CEA, CNRS, Université Paris-Sud, Université Paris Saclay, Gif sur Yvette, France.

Published online: 03 June 2019

References

1. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, et al. Functional annotation of a full-length mouse cDNA collection. Nature. 2001;409:685–90.
2. Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, et al. Functional annotation of a full-length Arabidopsis cDNA collection. Science. 2002;296:141–5.
3. Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, et al. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci U S A. 2002;99:16899–903.
4. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794–801.
5. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330:1775–87.
6. The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010;330:1787–97.
7. van Dijk EL, Chen CL, d'Aubenton-Carafa Y, Gourvennec S, Kwapisz M, Roche V, et al. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature. 2011;475:114–7.
8. Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature. 2013;497:127–31.
9. Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature. 2009;457:1038–42.
10. Wery M, Descrimes M, Vogt N, Dallongeville A-S, Gautheret D, Morillon A. Nonsense-mediated decay restricts LncRNA levels in yeast unless blocked by double-stranded RNA structure. Mol Cell. 2016;61:379–92.
11. Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat Genet. 2017;49:1731–40.
12. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013;31:1009–14.
13. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun. 2016;7:11708.
14. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, et al. A survey of the sorghum transcriptome using single-molecule long reads. Nat Commun. 2016;7:11706.
15. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44:D710–6.
16. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42:D756–63.
17. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 2012;22:1760–74.
18. Lee RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, et al. WormBase 2017: molting into a new stage. Nucleic Acids Res. 2018;46:D869–74.
19. St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase Consortium. FlyBase 102 – advanced approaches to interrogating FlyBase. Nucleic Acids Res. 2014;42:D780–8.
20. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J Cell Mol Biol. 2017;89:789–804.
21. Skrzypek MS, Nash RS, Wong ED, MacPherson KA, Hellerstedt ST, Engel SR, et al. Saccharomyces genome database informs human biology. Nucleic Acids Res. 2018;46:D736–42 Former ref 24 removed.
22. Mudge JM, Harrow J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. Mamm Genome. 2015;26:366–78.
23. Matthews BB, dos Santos G, Crosby MA, Emmert DB, St Pierre SE, Gramates LS, et al. Gene model annotations for *Drosophila melanogaster* : impact of high-throughput data. G3 (Bethesda). 2015;5:1721–36.
24. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, KRM S, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.
25. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreservation Biobanking. 2015;13:311–9.
26. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347:1260419.
27. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015;47:199–208.
28. You B-H, Yoon S-H, Nam J-W. High-confidence coding and noncoding transcriptome maps. Genome Res. 2017;27:1050–62.
29. Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang Y-C, et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. 2018;19:208.
30. Leinonen R, Sugawara H, Shumway M, on behalf of the international nucleotide sequence database collaboration. The Sequence Read Archive. Nucleic Acids Res. 2011;39:D19–21.
31. Silvester N, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L, Cleland I, et al. The European nucleotide archive in 2017. Nucleic Acids Res. 2018;46:D36–40.
32. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, et al. Bidirectional promoters generate pervasive transcription in yeast. Nature. 2009;457:1033–7.
33. Nellore A, Jaffe AE, Fortin J-P, Alquicira-Hernández J, Collado-Torres L, Wang S, et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the sequence read archive. Genome Biol. 2016;17:266.
34. Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. 2016;35:2413–27.
35. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9:eaal5209.
36. Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature. 2014;508:66–71.
37. Lian H, Wang Q-H, Zhu C-B, Ma J, Jin W-L. Deciphering the epitranscriptome in cancer. Trends Cancer. 2018;4:207–21.
38. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res. 2014;322:12–20.

39. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322:1845–8.
40. Deveson IW, Hardwick SA, Mercer TR, Mattick JS. The dimensions, dynamics, and relevance of the mammalian noncoding transcriptome. Trends Genet. 2017;33:464–78.
41. Helm M, Motorin Y. Detecting RNA modifications in the epitranscriptome: predict and validate. Nat Rev Genet. 2017;18:275–91.
42. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. Nat Rev Genet. 2018;19:535–48.
43. Deveson IW, Brunck ME, Blackburn J, Tseng E, Hon T, Clark TA, et al. Universal alternative splicing of noncoding exons. Cell Syst. 2018;6:245–255.e5.
44. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most "dark matter" transcripts are associated with known genes. PLoS Biol. 2010;8:e1000371.
45. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet. 2010;6:e1001236.
46. Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, et al. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biol. 2017;18:208.
47. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Response to "The Reality of Pervasive Transcription.". PLoS Biol. 2011;9:e1001102.
48. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, et al. The reality of pervasive transcription. PLoS Biol. 2011;9:e1000625 discussion e1001102.
49. Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5:578–90.
50. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012;487:370–4.
51. Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A molecular portrait of de novo genes in yeasts. Mol Biol Evol. 2018;35:631–45.
52. Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS Genet. 2014;10:e1004525.
53. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 2015;47:276–83.
54. Li Z, Qin F, Li H. Chimeric RNAs and their implications in cancer. Curr Opin Genet Dev. 2018;48:36–43.
55. Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, et al. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. J Virol. 2013;87:8916–26.
56. Singh B, Eyras E. The role of alternative splicing in cancer. Transcription. 2017;8:91–8.
57. Cáceres JF, Kornblihtt AR. Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet. 2002;18:186–93.
58. Elkon R, Ugalde AP, Agami R. Alternative cleavage and polyadenylation: extent, regulation and function. Nat Rev Genet. 2013;14:496–506.
59. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 2014;24:1774–86.
60. Wong JJ-L, Au AYM, Ritchie W, Rasko JEJ. Intron retention in mRNA: no longer nonsense: known and putative roles of intron retention in normal and disease biology. BioEssays News Rev Mol Cell Dev Biol. 2016;38:41–9.
61. Burns KH. Transposable elements in cancer. Nat Rev Cancer. 2017;17:415–24.
62. Lee M-H, Siddoway B, Kaeser GE, Segota I, Rivera R, Romanow WJ, et al. Somatic APP gene recombination in Alzheimer's disease and normal neurons. Nature. 2018;563:639–45.
63. Gong Q, Wang C, Zhang W, Iqbal J, Hu Y, Greiner TC, et al. Assessment of T-cell receptor repertoire and clonal expansion in peripheral T-cell lymphoma using RNA-seq data. Sci Rep. 2017;7:11301.
64. Han B, Chao J, Yao H. Circular RNA and its mechanisms in disease: from the bench to the clinic. Pharmacol Ther. 2018;187:31–44.
65. Day JR, Jost M, Reynolds MA, Groskopf J, Rittenhouse H. PCA3: from basic molecular science to the clinical lab. Cancer Lett. 2011;301:1–6.
66. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell. 1993;72:971–983.
67. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17:257–71.
68. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
69. Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015;16:97.
70. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086–92.
71. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
72. Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, et al. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. 2015;16:30.
73. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31:46–53.
74. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11:1650–67.
75. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010;28:503–10.
76. Boley N, Stoiber MH, Booth BW, Wan KH, Hoskins RA, Bickel PJ, et al. Genome-guided transcript assembly by integrative analysis of RNA sequence data. Nat Biotechnol. 2014;32:341–6.
77. Steijger T, Abril JF, Engström PG, Kokocinski F, Abril JF, Akerman M, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10:1177–84.
78. Hayer KE, Pizarro A, Lahens NF, Hogenesch JB, Grant GR. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics. 2015;31:3938–45.
79. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12:671–82.
80. Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK. TACO produces robust multisample transcriptome assemblies from RNA-seq. Nat Methods. 2017;14:68–70.
81. Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. eLife. 2016;5:e11752.
82. Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature. 2017;543:199–204.
83. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet. 2018;50:151–8.
84. Collado-Torres L, Nellore A, Frazee AC, Wilks C, Love MI, Langmead B, et al. Flexible expressed region analysis for RNA-seq with derfinder. Nucleic Acids Res. 2017;45:e9.
85. Sacomoto GAT, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot M-F, et al. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics. 2012;13(Suppl 6):S5.
86. Li Y, Heavican TB, Vellichirammal NN, Iqbal J, Guda C. ChimeRScope: a novel alignment-free algorithm for fusion transcript prediction using paired-end RNA-Seq data. Nucleic Acids Res. 2017;45:e120.
87. Audoux J, Philippe N, Chikhi R, Salson M, Gallopin M, Gabriel M, et al. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. Genome Biol. 2017;18:243.
88. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
89. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
90. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 2015;16:150.