Direct long-read RNA sequencing uncovers functional variation affecting transcript production and RNA modifications

Réal, Aline; Brown, Andrew; Yung, Gisella Puga; Borel, Christelle; Lykoskoufis, Nikolaos; Seebach, Jörg

DOI:
10.21203/rs.3.rs-4613444/v1

Publication date:
2024

Direct long-read RNA sequencing uncovers functional variation affecting transcript production and RNA modifications

Aline Réal

areal@nygenome.org

University of Geneva

Andrew Brown

State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China. https://orcid.org/0000-0002-8087-3033

Gisella Puga Yung

University of Geneva and University Hospitals and Medical Faculty https://orcid.org/0000-0002-2283-7798

Christelle Borel

University of Geneva Medical School

Nikolaos Lykoskoufis

University of Geneva

Jörg Seebach

University of Geneva and University Hospitals and Medical Faculty

Emmanouil Dermitzakis

University of Geneva

Anna Ramisch

University of Geneva

Ana Viñuela

University of Dundee https://orcid.org/0000-0003-3771-8537

Biological Sciences - Article

Keywords: ONT, long-read RNA, LCL, m6A-QTL, trQTL, eQTL

Posted Date: July 8th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4613444/v1

Additional Declarations: Yes there is potential Competing Interest. Emmanouil T. Dermitzakis is currently an employee of GSK. His contribution to the work presented in this manuscript was performed before he joined GSK. All other authors declare no competing interests.

Direct long-read RNA sequencing uncovers functional variation affecting transcript production and RNA modifications.

Aline Réal[1,2*], Andrew Brown[3], Gisella Puga Yung[2], Christelle Borel[1], Nikolaos M. R. Lykoskoufis[1], Jörg D. Seebach[2], Emmanouil T. Dermitzakis[1‡], Anna Ramisch[1,4‡*] and Ana Viñuela[3,5‡*]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [2]Division of Immunology and Allergology, University Hospitals and Medical Faculty, Geneva, Switzerland. [3]Population Health and Genomics, University of Dundee, Dundee, United Kingdom. [4]Department of Basic Neuroscience, University of Geneva Medical School, Geneva, Switzerland. [5]Biosciences Institute, International Centre for Life, Newcastle University, Newcastle upon Tyne, UK.

[‡]These authors contributed equally

* Corresponding authors: Aline Réal (areal@nygenome.org), Anna Ramisch (anna.ramisch@unige.ch), Ana Viñuela (ana.vinuela@newcastle.ac.uk),

& Current address: Population Health and Genomics, University of Dundee, Dundee, United Kingdom.

Keywords

ONT, long-read RNA, LCL, m6A-QTL, trQTL, eQTL.

Abstract

The production of multiple transcripts per gene is a process regulated by inherited genetic variants and epitranscriptomic modifications, and plays a prominent role in modulating complex traits and diseases. To simultaneously characterize the effect of genetic variants on transcript abundance and N6-methyladenosine (m6A) modifications, we produced long-read native poly(A) RNA-seq data for 60 genetically different lymphoblastoid cell lines (LCLs) from the 1000 Genomes/Geuvadis project. We identified a high diversity of both annotated (31%) and unannotated (61%) transcripts, with only a small proportion expressed across individuals (35% and 7%, respectively). In a genome-wide genetic analysis on transcripts, we identified 105 trQTLs, of which 76 were not detected as eQTLs using a larger published short-read RNAseq dataset (317 samples). A population wide characterization of m6A methylation DRACH motifs identified an average of 40.1 m6A modifications on 6,222 genes. Genetic association analysis of highly variable modifications from 1,155 genes identified m6A modification quantitative trait loci (m6A-QTLs) for 16 transcripts. Colocalization analysis of trQTL and m6A-QTLs, identified 33 candidate transcripts mediating GWAS traits, with 46.4% of the colocalized trQTLs implicating novel risk transcripts. Overall, the simultaneous characterization of transcripts and post-transcriptional modifications identified genetic effects on transcription often missed when using other sequencing technologies.

Introduction

Alternative splicing (AS) is a molecular mechanism that produces a diversity of mRNA molecules from a single gene [1]. By selecting different combinations of exons, genes produce multiple transcripts, which in turn can increase protein diversity [2-4]. Disruption of the splicing process has been associated with a large number of human diseases [5-7], while the regulation of splicing by naturally occurring genetic variants is reported to play a prominent role in modulating complex traits and diseases [1, 8-11]. However, our knowledge about how an individual's genetic background may affect the production of specific transcripts after splicing is limited due to the lack of full transcript measurements in population-based studies. Therefore, it is necessary to generate population level transcript data. However, the most used sequencing technology for assaying mRNA fractions the transcripts before generating short-read sequences, meaning a computational reconstruction of transcripts is required which is prone to errors [12-14]. The process also requires an mRNA to cDNA conversion before sequencing, which removes all marks of post-transcriptional RNA modifications. These modifications influence the stability, dynamics, translation, and cellular location of RNA molecules, with implications for disease development [15]. Long-read sequencing can overcome these limitations by allowing the sequencing of transcripts in their native form, reconstructing their precise structures [14, 16], identifying novel transcripts [17, 18], and studying the allele-specific effects on transcript abundance and structure [19, 20] as well as identifying putative transcripts mediating genetic variant effects on disease risk.

Here, we investigated the influence of human genetic variation on gene expression using direct long-read RNA sequencing of transcripts in their native form, from a population of 60 lymphoblastoid cell lines (LCLs) from the 1000 Genomes project [21]. Whole genome sequence was available from the New York Genome Centre website and short-read RNA sequencing data were available from the Genetic European Variation in Disease project (Geuvadis) [21-23]. With this data, we investigated *(i)* the distribution of annotated and novel isoforms across the population; *(ii)* the effect of genetic variants on directly measured transcript abundance, while *(iii)* characterizing some of the mechanisms by which SNPs affect gene and transcript expression. Moreover, (*iv*) we investigated the influence of human genetic variation on N6-methyladenosine (m6A) modifications in RNA molecules. This modification, present mostly in DRACH motifs (D–A, G, or U, R–A or G while H is A, C or U), is the most common

RNA modification in humans, with reports of up to one-quarter of transcripts in the human heart exhibiting this type of modification [24]. Our findings have the potential to identify novel candidate transcripts mediating GWAS trait regulation that would be missed using traditional sequencing methods. Moreover, we could also study underlying processes influencing human complex traits, such as m6A modifications that often impact splicing, RNA structure, and translation.

Dataset overview

Total RNA from LCLs from 60 unrelated individuals of European ancestry [21] were sequenced using a direct RNA sequencing (dRNA-seq) protocol to generate gene level and transcripts level quantification **(Supplementary Figure 1A-E)**. We quantified total gene expression for 11,929 protein-coding genes and lncRNA genes that were expressed in at least 50% of the samples. This was a smaller number than the 14,712 genes detected using expression data from previously published short-read cDNA derived RNA-seq data (cDNA-RNAseq) from the same 60 LCL samples (**Figure 1A**) [21, 23]. This was likely driven by the lower sequencing yield produced by long-read sequencing (**Supplementary Figure 1F**), as shown by previous studies comparing short and long-read transcriptomics data for both cDNA-RNAseq and dRNAseq [19, 25]. Most of the 3,170 additional genes discovered with cDNA-RNAseq were found to have low expression (**Figure 1A-B**). However, we were able to quantify the expression of 387 genes exclusively with dRNA-seq. These genes were significantly shorter than those detected across both technologies (pvalue = $6.8e^{-122}$, **Figure 1C**), in line with previous reports of bias towards better detection of longer genes by short-read sequencing [26]. Finally, despite the described differences, we observed a good correlation of expression across samples for the 11,542 genes in common (median Spearman correlation 0.61 - 0.79, **Figure 1D**), in line with previous reports [19].

dRNA-seq identifies novel transcripts across individuals

Long-read RNA-seq data from multiple donors can identify the expression of individual-specific transcripts and thousands of novel transcripts by quantification of full transcripts with minimal computational reconstruction. Using FLAIR [27], we quantified the expression of 44,993 transcripts of 12,599 protein-coding genes and lncRNAs with a high mapping quality (>10) (**Table 1**). Of these, 61% were novel. On average, we detected 2.9 transcripts per gene

across individuals, with 2% of genes expressing more than 5 novel transcripts (**Figure 1E**). Annotated transcripts had a significantly higher expression than non-annotated transcripts (Wilcoxon pvalue < $1e^{-296}$, **Supplementary Figure 2A**), with higher expressed genes showing a larger number of transcripts, both annotated and novel (**Supplementary Figures 2B-D**). This suggests that transcript annotations may be biased towards highly expressed transcripts. Moreover, we observed that a small number of transcripts were expressed in all individuals (23.3%, n=10,101), with fewer novel transcripts (16%) expressed across all 60 individuals compared to annotated transcripts (35%). We also found 2,371 genes (18.8%) which only expressed novel transcripts (**Supplementary Figure 3A-B**). Of these, 2,220 (94%) reported detectable expression in the previously published short-read cDNA-RNAseq gene-level summary quantification of the same LCL samples [23]. Our findings suggest that many novel transcripts lacked annotations due to detection limitations and donor diversity. Moreover, they highlight the importance of generating and investigating transcripts quantifications in a population setting.

A known property of multi-exon genes expresses multiple transcripts, is that often one transcript dominates their expression in a specific cell or tissues [29, 30]. By defining genes with dominant transcripts as those with more than 90% of all the reads associated with one transcript, we found that 47% (5,940) of all expressed genes had a dominant transcript, and 2,080 genes expressed only one transcript. We observed that genes with a dominant transcript had a higher mean expression than genes without dominant transcripts (Wilcoxon pvalue < $2.2e^{-16}$, **Supplementary Figure 3C**). This was in agreement with previous reports showing that highly expressed protein-coding genes tend to code for one main protein isoform [29]. We also observed that the majority of dominant transcripts were annotated (71.5%), suggesting that a high proportion of transcripts lacking annotation were not detected before due to low expression compared to known transcripts.

eQTL and trQTLs discovery using dRNA-seq data

To detect genetic associations with gene expression, we performed a genome-wide cis-eQTL analysis (see Methods) and discovered 34 significant eQTLs (FDR = 5%, **Figure 2A and Supplementary Table 1**). This is a considerably lower number than the 428 cis-eQTLs discovered using the exact same 60 LCL samples and short-read cDNA-RNAseq data previously published [23]. The SNPs associated with dRNAseq eQTLs (eSNPs) were located

in close proximity to the transcription start site (TSS) of genes (**Figure 2A**), as expected from other eQTL studies [31, 32]. Genes with mapped eQTLs (eGenes) had a higher median expression than genes without an eQTL (**Figure 2B, Supplementary Figure 3D-F)**, suggesting the lower sequencing yield of the newer technology limited the discovery power. dRNAseq eGenes were also significantly shorter compared to the other genes (Wilcoxon pvalue < $1.82e^{-05}$), supporting our findings that long-read technologies provide some improvement to detect expression from shorter genes (**Figure 2C**). All in all, the lower sequencing yield obtained presents a strong limitation for population studies, however, newer technological developments are expected to overcome this.

Next, we performed a genome-wide cis transcript quantitative trait loci (trQTLs) analysis to identify genetic regulation of transcripts that were expressed in at least 50% of the samples (n = 33,840 from 11,561 genes). The analysis included annotated and novel transcripts and all SNPs in a 1Mb cis-window around each gene. To control for the lack of independence on the expression of transcripts from the same gene, we first performed a gene-level multiple testing correction across all transcripts and SNPs tested followed by a genome-wide correction (see Methods). Reporting the most significant transcript-SNP pair per gene, we detected 105 trQTLs (FDR = 5%) in close proximity to the TSS of the genes (**Figure 2D, Supplementary Table 2**). We found more associations than for the gene quantification analysis, despite the higher multiple testing burden. Only 13 genes had both eQTLs and trQTLs, with 9 involving the same SNP (n=7) or SNPs in high LD (R^2>0.95, **Supplementary Table 3 and Supplementary Figure 4A**). Transcripts with trQTLs had a significantly higher expression than transcripts without any trQTL (Wilcoxon pvalue = $1.3e^{-10}$, **Figure 2E)**, and their genes had on average a higher number of transcripts than genes without significant transcript associations (7.11 and 2.9, respectively; **Figure 2F**). Our results indicate that transcripts qualifications provide a more informative phenotype to characterize genetic effects on gene expression and can identify genetic associations not observable by standard eQTL analyses.

The identification of more trQTLs compared to gene-level cis-eQTLs may be driven by transcript specific effects which are missed when using gene-level summary phenotypes. To investigate this, we explored how often the same trSNP was associated with other transcripts from the same gene and how these compared to eQTLs from short and long reads quantifications derived from the sum of reads of all transcripts. A re-evaluation of the transcript

158 significance within each gene (see Methods) found that half of our trSNPs (52 out of 105) were
159 significantly associated with two (n = 27) or more (n = 25) transcripts per gene
160 (**Supplementary Figure 4B**). For the majority of these 52 trSNPs (57.7%) the direction of the
161 effect was opposite on at least one of the transcripts (**Supplementary Figure 4 C-D**). This may
162 explain the lower resolution to identify genetic associations using gene-level quantification, as
163 opposite genetic effects where a variant caused an increase in one transcript's abundance and
164 a decrease in another would to some degree cancel out when considering a summary, gene
165 level measure. For example, we found that the trSNP rs35736654 was significantly associated
166 with two out of seven transcripts of the *CAST* gene (**Figure 3A**) with opposite directions of
167 effect (**Figure 3B-C**). The effect of rs35736654 on *CAST* gene-level quantifications was not
168 significant (pvalue = 0.66 dRNA-seq, **Figure 3D-E**), but it followed the same general direction
169 of effect as the most significant trQTL which affected one of the less abundant transcripts
170 (**Figure 3B**). On the other hand, the trSNP rs35251247 was significantly associated with two
171 annotated transcripts of the gene *HSD17B12,* also with opposite direction (**Figure 3F-H**). The
172 gene level quantifications with short reads recapitulated the effects observed for the most
173 significant of the trQTLs, the dominant transcript (ENST00000278353.4, **Figure 3I-J**), while
174 they missed the association with a second less expressed transcript (ENST00000395700.4,
175 **Figure 3H**). Genes with dominant transcripts were involved in 34 of the 52 trSNPs showing
176 opposite directions of effects (**Figure 2G-H**), but only 5 trSNPs associated with a dominant
177 transcript itself. This indicates that significant trQTLs often involve non-dominant transcripts,
178 contributing to the difficulty in detection of genetic associations using gene quantifications.

179 One reason why an allele could have an opposite direction of effect on different transcripts
180 from the same gene is by affecting splice events through a splice QTL (sQTL). sQTLs have
181 been reported closer to the gene body than eQTLs [33], changing the abundance of reads
182 associated with a splice event and the abundance of specific transcripts without necessarily
183 altering the overall expression of the gene. We found that trSNPs were slightly closer to the
184 TSS than eSNPs (54Kbp vs 78Kbp, **Figure 2A and 2D**), suggesting trQTLs may be the result
185 of genetic regulation of splice events. Using splice-QTLs detected by the GTEx LCLs dataset
186 (n = 147, [32]), we investigated how many of the 1,607 genes expressed in both studies had a
187 trQTL. After multiple testing correction on all the transcript-SNP pairs from genes with sQTLs,
188 we found that 54 (51.4%) trQTLs were also significant sQTLs (FDR = 0.05). This suggests
189 that alternative splicing as the underlying biological process for many, but not all of the trQTLs.

Our results showed that the combination of QTL results from short and long-read technologies can contribute to characterizing biological processes underlying eQTLs, such as alternative splicing or effects on specific transcripts. However, given the sample size of the dRNAseq dataset, low for genetic studies, we could only detect genetic signals affecting a small minority of genes, even though most genes are expected to be affected by genetic variation [31, 32]. To overcome this limitation, and further explore the molecular processes underlying eQTL effects, we used 3,917 previously reported pairs of significantly associated cis-eQTLs using gene summary quantifications from the full dataset of 317 LCL samples sequenced with short-reads [27] In contrary to our findings from the genome-wide analysis, we found a slightly higher number of significant eQTLs (n = 277) than trQTLs (n = 259) using the long-reads dataset (FDR = 5%), but this approach increased the number of significant associations detected. These QTLs involved 414 unique genes, of which only 29.4% had significant associations for both types of QTLs, finding again only a limited overlap between eQTLs and trQTLs (**Supplementary Figure 4E**). For example, a transcript for the lncRNA gene *FLVCR1-DT* was associated with rs2279692 (**Figure 4A)**, a SNP also significantly associated with the gene in the short- and long-read gene quantifications (**Figure 4B-C**). Moreover, we found again that genes for which a trQTLs was found (n = 137 out of 259) had a higher number of transcripts compared to 155 out of 277 genes with an eQTL (Wilcoxon test pvalue = $7.93e^{-06}$, **Supplementary Figure 4F**). Since these comparisons were limited by the differences in the sample size of the short- and long-read studies (317 vs 60), we also estimated the proportion of associations from the alternative hypothesis (π_1), ranging from 0 to 1, with 1 being an estimation that all tests involved the alternative hypothesis of associations between variant and long read phenotype [34, 35]. Here, we estimated that of the short-reads eQTLs 24% ($\pi_1 = 0.24$) had evidence of being and eQTL ascertained with long read phenotypes, while for only 14% ($\pi_1 = 0.14$) was there evidence of acting as trQTLs.

TrQTLs provided a better characterization of the genetic regulatory effects on gene expression. Of the 105 trSNPs identified by the genome-wide analysis, 98 were not significantly associated in any of the gene-level eQTL analyses. This could be driven by trSNPs of opposite direction of effect between transcripts of the same gene, pleiotropic effects from SNPs in linkage disequilibrium (LD), and effects on lower expressed and nondominant transcripts, harder to detect using gene-level quantification. Of the examples of trQTLs identified using the targeted, short read eQTLs approach, we find associations with the *OAS1* gene. We detected significant

associations between rs1154970 and six out of seven expressed transcripts (**3 representative examples in Figure 4D-F, Supplementary Figures 5A and 6A**), but no significant eQTL using long reads (**Supplementary Figures 5B-C**). The trSNP increased the abundance of one transcript while decreasing the abundance of the others, but the overall gene quantification was not different between genotype groups with the sample size available. SNPs in partial LD with an effect on expression may also mask the effect of trSNP making it harder to detect transcript effects at the gene level. For example, rs4796398 was significantly associated with the expression of only one out of ten detected transcripts of the *EIF5A* gene (pvalue = $4.25e^{-09}$, ENST00000336458.8, **Figure 4G, Supplementary Figures 5D and 6B**). Neither the short-read nor the long-read eQTL analysis for the corresponding SNP-gene pair was significant (**Figure 4H-I**). However, another SNP (rs28636077) in LD with the trSNP (R^2 =0.992) was significantly associated with the gene-level expression (pvalue=$7.99e^{-58}$). More difficult to detect were genetic associations involving lowly expressed non-dominant transcripts. An example involves the *ARPC2* gene with 17 expressed transcripts, including an annotated dominant transcript, i.e. the transcript with more than 90% of all reads associated with that gene **(Supplementary Figure 5E and 6C)**. We detected two novel transcripts associated with rs2271541 (**Figure 4L-M)**, but there was no cis-eQTL detected since the gene-level quantification mainly summarised the expression of the dominant transcript, which was not significantly associated with the trSNP (**Supplementary Figures 5F-G**). These examples demonstrate the advantage of QTL analyses on transcript- rather than gene-level quantifications and exemplify the complicated processes involved in gene expression regulation.

Genetic regulation of m6A RNA modifications abundance

N6-methyladenosine (m6A) modifications of RNA molecules are known to regulate pre-mRNA processing and mRNA stability. Previous studies have shown that human genetic variation regulates the abundance of RNA modifications by identifying m6A modifications quantitative trait loci (m6A-QTLs) using m6A sequencing (m6A-seq). However, dRNA-seq technologies also allows the identification of m6A modifications without additional experiments. Therefore, we identify RNA modifications on transcripts using m6Anet [36], which reports the ratio of reads with modified compared to unmodified bases per transcript on DRACH motifs (D–A, G, or U, R–A or G while H is A, C or U). After quality assessments, we

identified 255,014 m6A RNA modification events on 18 motifs detected in at least one sample (**Figure 5A-B, Supplementary Figure 7A**). These modifications were detected on transcripts from 6,222 unique genes, with a mean of 40.1 modifications per gene. Most modifications were located in introns, exons, or 3'UTR regions of genes (**Figure 5B**). Although not directly comparable, these numbers were in line with those reported by m6A-seq experiments in human tissues, including LCLs [37, 38].

To identify m6A-QTLs, we studied the 30% most variable modifications which were present in at least 50% of the samples (**Supplementary Figure 7B-C**): 33,933 modifications from 1,155 unique genes (31.5 mean modifications per gene). To allow comparison with other QTLs, we tested all SNPs in the same 1Mb window around the TSS of the genes, controlling for multiple testing across modifications per gene and reporting the best modification-SNP association per gene. After genome-wide multiple testing correction, we detected 16 significant m6A-QTLs (FDR 5%) (**Supplementary Table 4**). The significantly associated SNPs (m6A-SNPs) were in 43% of cases downstream gene variants according to the VEP database [39], in 14% of cases they were intronic variants (18%) (**Supplementary Table 5**). m6A-SNPs were closer to the TSS than either eSNPs or trSNPs (mean distance = 41.37 Kbp, vs >54Kbp, **Figure 5C**), likely because they were often close to the m6A modifications in coding regions (mean distance = 73.94 Kbp). Given the proximity of the m6A-SNPs to the actual motifs and to discard possible false positives, we investigated if SNP were often located in the modified motifs. Among all the RNA modifications detected (n = 257,910) only on 1.4% had a SNP as part of their motif (3,767) and only 1.4% of the motifs tested for m6A-QTLs (472 out of 33,933) (**Supplementary Table 6)**. None of the 16 motifs involved in m6A-QTLs contain a SNP in the sequence. Finally, we observed that transcripts with significant m6A-QTLs were coded by genes with a higher number of transcripts, both novel and annotated, compared to genes without significant m6A-QTLs (Wilcoxon pvalue = $6.19e^{-106}$ annotated, $4.21e^{-84}$ novel and $8.49e^{-158}$ all transcripts, **Supplementary Figure 7D-F**). Our results indicate that long-read direct RNAseq is suitable for identifying genetic effects modulating the proportion of transcripts with m6A modifications.

To better understand the processes mediating m6A-QTLs, we investigated how often significant eQTL and trQTLs acted as m6A-QTLs. We used the 7,657 genes with eQTLs in the larger short-reads data (317 samples), of which 528 included motifs evaluated for m6A-

QTLs and found no significant eSNP after multiple testing corrections (FDR < 5%) in the m6a-QTL analysis ($\pi_1 = 0$). Likewise, of the 11,529 genes with sQTLs reported in GTEx LCLs analyses, none of the 713 genes with motifs had a significant m6A-QTL [32]. We found one transcript-SNP pair as a significant m6A-QTL for the canonical and most abundant transcript of the *POLE4* gene, a DNA Polymerase Epsilon Subunit 4 with seven annotated transcripts. The rs12366-T allele increased the expression of the most abundant transcript and decreased the expression of the less abundant transcript of the gene. The same allele decreased the ratio of reads with m6A modifications in the GGACC motif, suggesting rs12366 may not influence the RNA modification process itself, but simply increase the number of transcripts produced that were not modified (**Figure 5C-D**). Overall, direct-RNAseq allowed for the identification of genetic effects influencing the relative abundance of m6A modification on transcripts, without the need of additional experiments, contributing to our understanding of the underlying processes regulating gene expression.

GWAS colocalization

Genetic effects on gene expression are often used to identify genes mediating the activity of genetic variants on complex traits identified using GWAS. To evaluate the possible improvement on identifying genes mediating GWAS loci activity using quantifications from transcripts and m6A modifications, we investigated the overlap of significant SNPs with the GWAS catalog [40, 41] and performed a colocalization analysis with 14 traits. Of the 105 trSNPs, 13 trSNPs (12.38%) were reported as lead variants for 29 GWAS traits from the catalog and 14 colocalized with signals from 9 GWAS (COLOC probability > 0.9, **Supplementary Tables 7 and 8**), making a total of 28 trQTLs implicated with GWAS traits. Of these, 13 (46.4%) were associated with the expression of novel transcripts. Among the 16 SNPs involved in m6A-QTLs, three were previously identified as lead GWAS variants for five traits, and two colocalized with two traits. For example, a trQTL involving rs35251247 and a transcript of the *HSD17B12* gene (**Figure 3G)**, was found associated with Type 2 Diabetes in a previously published study [42, 43] (**Supplementary Figure 8A-C**). The same SNP was also detected as a eQTL, and our results now suggest that such an effect may be mediated by the expression of a specific transcript. In another example, the SNP rs55936281, lead variant for a m6A-QTL on a AGGCT motif, colocalized (COLOC probability = 0.93) with a previously reported GWAS association for the metabolite cis-4-decenoyl carnitine (**Figure 5E**) [34, 44]. The m6A-QTL

pointed to a transcript for the *PPID* gene, encoding for the peptidylprolyl isomerase D (cyclophilin D), an enzyme involved in cellular processes such as protein folding. All in all, the use of population based data for the identification of genetic effects on transcripts and RNA modifications allow us to identify novel candidate genes and transcripts mediating the activity of GWAS traits. Moreover, in some cases it may suggest an underlying biological process, such as alternative splicing or RNA modifications, to be involved in this mediation.

Discussion

In this study, we performed sequencing of mRNA molecules in their native form using long-read direct RNA sequencing (dRNA-seq) from 60 of the 1000 Genomes project samples. We report similar read length distributions, and quality to previous studies using long read sequencing of cDNA reads from mRNA (cDNA-RNAseq) [25] or short reads in the same LCL samples [23]. Our population-based dataset, together with DNA sequence, and eQTLs from short-read RNA-sequencing, allowed us to study the effect of human genetic variation on transcript abundance and RNA modification ratios. A previous study using 90 samples from 14 GTEx tissues evaluated the effects of human genetic variation on transcripts across tissues. However, due to their limited number of donors (<5), only allelic specific expression (ASE) was considered, meaning that while the presence of cis genetic effects could be inferred, these genetic effects could not be mapped [32]. Moreover, the long-read sequencing used in that study employed cDNA-RNAseq. Their experimental design allowed them to detect genetic effects and a higher diversity of novel transcripts (77% vs 61% in this study), likely driven by a larger sequencing yield and the use of samples from 14 different tissues. However, some of these additional transcripts may be PCR-derived artefacts produced by cDNA conversion, an experimental step not required for the sequencing of native molecules. Our study, on the other hand, we were able to investigate the diversity of transcripts across individuals reporting many novel transcripts expressed across 60 donors. Further studies using this and similar theologies, will likely identify many more transcripts.

Direct RNA sequencing however, did provide lower sequencing coverage, which meant that fewer eQTLs were found compared to short-read RNA-seq. Lower coverage has been reported before for this technology [25, 45], but newer developments are already providing better quality data and higher resolution for population studies [46]. All in all, even with this limited

coverage, we were able to discover genetic effects on gene expression that short-read technology missed. Just as eQTL studies using exon quantifications typically find more eQTLs than gene-level studies [34, 47], transcript quantifications and isoform centric analyses [48] are able to identify genetic effects missed using the aggregate gene-level expression. Examples of these advantages include the identification of genetic effects acting exclusively on the abundance of lowly expressed transcripts, or those with opposite directions on multiple transcripts from the same gene. The later example would be driven by genetic effects influencing splicing and the ratio of transcripts produced, e.g.: exon skipping events, and would be detected using transcripts ratios as phenotypes [19]. Ultimately, we were able to report that 8% of significant *cis*-eQTLs identified with short-reads and a larger sample size were the results of changes in the expression of specific transcripts (trQTLs), helping to untangle the nature of the genetic regulation.

An additional advantage of direct RNA-seq is the ability to directly evaluate the influence of post-transcriptional chemical modifications of RNA molecules without the need of any additional experiment. We detected a comparable number of modifications reported in different tissues [49] and by other sequencing methods [37, 38]. Their distribution along the genes were also in agreement, with the vast majority of them falling in introns, exons, and 3'UTR regions of genes, highlighting their relevance for transcription regulation and transcript splicing and production. A genetic association analysis detected that SNPs associated to m6A modifications were located closer to the gene's TSS than eQTL and trQTL, pointing out to a possible role on splicing regulation. However, the lack of overlap between m6A-QTLs and other forms of genetic associations (QTLs) reported here and in previous studies [37, 38], suggest these genetic effects are missed using traditional sequencing and multiple testing methods [48]. The only example we identified with a SNP that acted both as trQTL and a m6A-QTL, pointed to a mechanism where genetic variation influences transcripts abundance without interfering m6A addition or removal. Recent research indicates that m6A may have its strongest effects on decay or translation in differentiating cells or those undergoing stimulation [50, 51]. Therefore, a significant future direction is to map m6A-QTLs across various disease-related cellular and physiological contexts which could provide insights into new mechanisms by which genetic variants can influence disease risk.

Long-read sequencing, in particular using direct RNA, can help to identify novel candidate genes that mediate the activity of GWAS variants. These new findings are mostly provided by the better biological resolution and additional information provided by the data generated. Transcripts provide a more direct characterization of gene products, resulting in a larger number of genetic associations, just as has been reported before for exon and other transcript aware methods and technology [31, 48]. As a consequence, we were able to detect trQTLs colocalizing with GWAS loci. We noted that half of those signals were associated to novel transcripts, highlighting not only the improvement in identifying genes mediating GWAS variants activity but also the importance of better molecular phenotype annotations. We and others have reported thousands of novel transcripts using long-read sequencing methods [19, 40, 46]. Some of these novel transcripts have been attributed to artifacts introduced by the RNA-to-cDNA transformation used by some or as result of mRNA molecules being fragmented during sequencing. However, there is strong evidence that many of the novel transcripts were simply not annotated, and many novel transcripts are observed in all individuals in our study. Future work will greatly benefit from novel datasets using these long read technologies on native molecules, as well annotation initiatives that aim to harmonize current databases, such as Matched Annotation from NCBI and EMBL-EBI (MANE) collaboration [40].

Material and Methods

LCL samples

The 60 LCLs were from the 1000 Genomes Project cohort with European ancestry and from unrelated individuals [21]. All the samples are part of the NHGRI sample repository for human genetic research. All LCLs came from the Coriell Institute for Medical Research (Camden, New Jersey, USA). LCLs were grown under identical conditions in RPMI 1640 media supplemented with 1% penicillin-streptomycin, 1% L-glutamine, and 10% fetal bovine serum (FBS). LCLs were cultured for at least 3 to 4 weeks until their exponential growth phase and had a total concentration of at least 4×10^7 LCLs and a constant high viability (~98%). All cells were tested for mycoplasma contamination (Lonza, MycoAlert mycoplasma detection kit) before being used for the following steps.

RNA extraction, library preparation, and sequencing

From 4×10^7 mycoplasma-free LCLs we obtained total RNA using Trizol Reagent (Invitrogen). Cells were washed twice with 1 × phosphate-buffered saline (Invitrogen) to remove all the media and 1ml of Trizol was added per 5-10 $\times 10^6$ cells in each sample, incubated for 5min at room temperature (RT), and transferred to Eppendorf tubes. The rest of the protocol followed the manufacturer's guidelines with the addition of 200µl of chloroform for every ml of Trizol for the phase separation followed by mixing and centrifuging for 15min at 2000 × *g* at 4°C. The phase containing RNA was recovered and transferred in new Eppendorf tubes for RNA precipitation with 500µl of isopropanol for every 1ml of Trizol and incubation at RT for 15min followed by centrifugation 20min at 2000 × *g* at 4°C. After RNA precipitation, 70% ethanol was used to wash the pellet and centrifuged 5min at 2000 × *g* at 4°C, the supernatant was discarded, and the RNA pellet was air dried for 5 -10min. The pellet was solubilized in 20µl of 0.5% SDS in RNase-free water. No DNAse treatment was applied.

The total RNA was quantified using Qubit Fluorometer 2.0 with the Qubit RNA Broad Range (BR) Assay kit according to the manufacturer's instructions (Thermo Fisher) and Nanodrop to exclude the presence of alcohol and protein contaminants that could interfere with the sequencing, keeping RNA samples with a ratio $OD_{260/280}$ of at least 1.9 and ratio $OD_{260/230}$ > 1.5. Agilent Bioanalyzer RNA 6000 Nano Kit (Agilent) was used to assess the quality and

integrity of RNA. Only samples with RNA Integrity Number (RIN) >8.9 were used for the following steps. The total RNA was then poly-A^+ tailed before the library preparation using the Dynabeads™ mRNA Purification Kit (Thermo Fisher). The poly-A^+ tailed capture step was repeated re-using the same Dynabeads, which increased the enrichment of poly-A^+ tailed RNA and improved the consequent elimination for ribosomal RNA. The final quantification of poly-A^+ tailed RNA was performed by the TapeStation with High Sensitivity RNA ScreenTape (Agilent) to ensure the quasi-total elimination of the 28S and 16S ribosomal peaks.

For the library preparation, we used 500ng poly-A^+ tailed RNA in a total volume of 9µl and followed all the steps of the ONT protocol for the Direct RNA Sequencing Kit (updated version 27/12/2019 nanoporetech.com, cat# SQK-RNA002). The quantification of the library RNA was performed using the Qubit fluorometer DNA HS assay (Thermo Fisher) - recovery aim of ~200ng. Then, before loading the RNA into the FLOW-MIN106D flow cell, the numbers of pores and properly primed pores were checked according to the manufacturer's instructions. Finally, the sequencing was carried on for 72h on the GridION Mk1 sequencing device (ONT) that allows sequencing of a maximum of five samples in each run, one per flow cell.

Pre-processing of RNA sequencing data

Base-calling was performed using the Guppy software (from ONT, v 3.2.10) in the *high accuracy mode*. Guppy used the *fast5* files generated by the ONT Device Control software (MinKNOW), embedded in the GridION sequencing device, as input to *(i)* generate a *fastq* file for each *fast5* file containing the base-called sequences; *(ii)* create base-called *fast5* files. *(iii)* classify *fastq* and *fast5* files into pass/fail folders according to the average quality score of each read (above 7.0), and (*iv*) make summary files for every flow cell sequenced. We applied the specific options suggested in the Direct RNA sequencing protocol taking into consideration the reversed direction of the sequencing (3′→5′), the presence of uracil instead of thymine, and an optimized strategy for trimming the adapter's raw signal. We used only the passing reads for the following analysis.

Mapping sequences. We concatenated the *fastq* files obtained from base-calling into a single *fastq* file. These *fastq* files were then mapped to the reference human genome GRCh37 (hg19_chr_only_and_herpes.fa) using minimap 2 v2.12 [52]. The calling was performed in a splicing-aware manner with the following options: *minimap2 -a -x splice -k14 -uf* (-a -x splice:

splice alignment mode; -uf: force minimap2 to consider the forward transcript strand only; -k14: small k-mer to increase sensitivity to the first or the last exons). Alignment files from minimap2 were converted to *bam* format, sorted, and indexed using samtools v1.6 [53].

Data quality control (QC). We applied Nanoplot (v1.33.0) [54] to produce QC graphs displaying multiple aspects of sequencing raw data; while NanoStat (v1.4.0) was used to obtain a statistical data summary [54]. The pycoQC tool (v2.5.2) [55] served to generate an interactive QC report from the base caller's datasets. Specifically, pycoQC uses the sequencing summary file generated by Guppy and the *bam / sam* file to generate a pre / post-alignment QC report.

Gene quantification

We used *featureCounts* to provide the gene-level counts (from Subread v1.6.0) to the genome alignments using exons as the feature type [56]. We used the *-L* argument of *featureCounts* to enable the long-read mode with a minimum overlap of 10 bases *(--minOverlap 10*) and we used GENCODE v19 as the reference annotation [57]. We converted counts to RPKM (Reads Per Kilobase of transcript, per Million mapped reads) using the *rpkm* function of the edgeR package (v3.9) [58].

RNA-seq data and genotype data from external datasets

Curated short-read Illumina RNA-seq data and genotype data of 60 LCLs were used as described in Delaneau et al. [23]. Briefly, gene expression was quantified using QTLtools (v1.3.3) [59] with GENCODE v19 [57] as the reference gene annotation. Genes were filtered to retain only protein-coding genes and long non-coding RNAs (lncRNAs) expressed in more than 90% of the samples. The gene expression was quantified using RPKM units for the gene expression quantification in the ONT dataset. The genotype data for these samples, available from either the 1000 Genomes project or the Illumina Human OMNI 2.5M SNP array, were filtered using standard procedures to remove low-quality SNPs. Moreover, the resulting genotype matrix of 317 individuals and 9,255,024 variants was imputed from the 1000 Genomes phase 3 reference panel [21], and poorly imputed variants were removed [23].

Gene expression correlation between short-reads and dRNA long-reads seq dataset

Pair-wise gene expression Spearman correlations were computed between Illumina short-reads and direct long-read RNA sequencing ONT from the same 60 LCLs samples using the *rcorr* function in the corrplot R package (v0.92) [60]. Both gene expression datasets included protein-coding genes and lncRNAs expressed in at least 50% of their samples, which were found in both sets (n = 11,542).

Transcript detection and characterization

To identify transcripts from the native RNA sequences we used FLAIR v1.5 (https://github.com/BrooksLabUCSC/flair) [27]. For the analysis, *bam* files obtained using the minimap2 aligner were converted to *bed* format using the bam2bed12.py script provided with FLAIR. FLAIR-*correct* was used to correct the splice-site boundaries of reads. It corrected misaligned splice sites using genome annotations from GENCODE v19 and GRCh37 as the reference genome. Next, the FLAIR-*collapse* command processed the corrected reads, generating a first-pass transcripts set. To do this, FLAIR-*collapse* grouped reads on their splice junction chains and only kept transcripts supported by at least 10 reads and mapping quality >10. At this this step, the first-round alignments were split by chromosome due to computational limitations. FLAIR-*quantify* was used to determine transcript levels in all samples where reads aligned to annotated transcripts (GENCODE v19). Transcripts with intron chains not matching any transcripts in the reference annotation (GENCODE v19) were defined as *'novel isoforms'*. The 36,782 transcripts not aligning with any gene in the GENCODE v19 annotation were excluded from the following analyses. Reads were normalized using transcripts per million (TPM) normalization. Moreover, mitochondrial transcripts as well as transcripts supported by less than 10 reads and expressed with less than five TPMs in at least one sample were excluded.

Molecular quantitative trait loci

For each molecular phenotype, gene abundance, and transcript abundance, we identify QTLs using the QTLtools software package (v1.3.1) [59]. Shortly, all genetic variants within +/– 1Mb of the transcription start site were associated with the phenotypes, and the best-associated SNP (i.e., with the smallest nominal pvalue) was retained. After that, the nominal pvalues were

adjusted for the number of variants being tested using 1,000 permutations. This is implemented in the *cis* mode of the QTLtools software package (v1.3.3) [59]. Multiple testing correction across phenotypes was done using the qvalue package in R (version 2.18.0) [35] to identify all significant phenotype-variant pairs at 5% False Discovery Rate (FDR). For gene-eQTL analysis, we tested genes expressed in at least 50% of the samples (n = 13,997). For the transcripts-eQTL analysis, we tested annotated transcripts expressed in at least 50% of the samples (n = 14,447) which corresponded to 9,364 unique genes. For transcript-eQTL analysis, we used the option *-grp* to correct and account for multiple phenotypes (transcripts) per gene. This option performs a permutation pass at the gene group level across all phenotype-SNP pairs per gene to discover gene-level trQTLs. All QTL analyses included the following covariates: sex, the first three principal components (PCs) from genotypes, and three and one PCs from expression genes and transcripts, respectively.

Short-reads eQTL recapitulation in dRNA long-reads dataset

The eQTLs already identified using 317 samples from the Illumina RNA-seq data described in Delaneau et al. [23] were used. From these 7,658 significant eQTLs in the Illumina dataset, only 4,169 involved genes and transcripts expressed in at least 50% of the samples that were kept as part of the dRNA long-reads dataset. To detect how many of the significant short-reads eQTL were also detected in long-read native RNA-seq, we extracted the pvalues from the same phenotype-SNP pair associations and calculated π_1 using the *q*-value package in R [35]. The qvalue estimation for false discovery rate control R package (v2.18.0) used a $\lambda = 0.05$ and FDR = 0.05.

To detect how many of the 105 significant trQTLs we discovered have a significant effect on multiple transcripts produced by the same gene, we extracted the nominal pvalues from the same phenotype-SNP pair associations and calculated π_1 using the qvalue package in R (v2.18.0) [35]. The qvalue estimation for false discovery rate control R package (version 2.18.0) used a $\lambda = 0.05$ and FDR = 0.05.

RNA modifications

The m6A RNA modifications detection was performed using m6anet tool[36] that was specifically trained on dataset sequenced using the dRNA SQK-RNA002 kit. We follow the

general pipeline starting from the alignment to the transcriptome reference using nanopolish tools (v0.14.0) [61], the data prep step, and the inference step. m6Anet performs a sampling process by selecting 20 reads from each candidate site. The probability of modification is averaged over 1000 rounds of sampling and the resulting data contains the probability of modification for each individual read.

Filter on modifications: 261,708 modifications found in at least one transcript. 257,910 were annotated to known genes as reported by Gencode v19 [57]. We remove anything on chromosomes Y and mitochondria, keeping only protein-coding genes and lncRNAs, leaving a total of 255,014 modifications. Next, we remove any modifications with more than 50% missing values. That left 113,110 modifications from 3,586 genes for evaluation.

m6A RNA modification molecular quantitative trait loci

For the RNA modification molecular phenotype, we identify QTLs using the QTLtools software package (v1.3.1) [59]. As for the eQTLs and trQTL, all genetic variants located within a range of ±1Mb from the transcription start site were linked to the phenotypes (RNA modification), and the SNP showing the strongest association (i.e., with the lowest nominal pvalue) was preserved. Following this step, the nominal pvalues underwent adjustment to account for the number of variants being examined, employing 1,000 permutations. This process was carried out using the cis mode feature within the QTLtools software package (v1.3.3) [59]. Multiple testing correction across phenotypes was done using the qvalue package in R (v2.18.0) [35] to identify all significant phenotype-variant pairs at 5% FDR.

For m6aQTL analysis, we tested RNA modifications expressed in at least 50% of the samples (n = 113,110 modifications from 3,586 genes) and we filtered further removing modifications with low variation in the population. We therefore only tested those modifications that were in the top 30% of most variables, leaving a total of 33,933 modifications from 1,155 genes for QTL analysis. As for trQTLs, we employed the "*-grp*" option to correct for and consider multiple phenotypes (modifications) associated with each transcript. This option conducts a permutation process at the transcript group level across all phenotype-SNP pairs per transcript, enabling the identification of transcript-level m6A-QTLs. All QTL analyses included the following covariates: sex, the first three principal component (PCs) from genotypes, and one PCs from expression (RNA modifications).

Short-reads eQTL recapitulation in dRNA long-reads dataset mA6RNA modifications

We used the 7,658 significant eQTLs already identified using 317 samples from the Illumina RNA-seq data described in Delaneau et al. [23] to recapitulate them into dRNA-seq dataset for RNA modifications. Only 558 involved genes coding for transcripts affected by RNA modifications tested in the m6A RNA modification QTL analysis. To identify the overlap between significant short-read eQTLs and long-read native RNA-seq RNA modifications, we retrieved the pvalues associated with the same phenotype-SNP pairs. Subsequently, we computed π_1 using the q-value package in R [35] to estimate the proportion of true discoveries. For false discovery rate (FDR) control, we utilized the qvalue estimation package (v2.18.0) with parameters $\lambda = 0.05$ and FDR = 0.05.

GWAS overlap and colocalization

To identify genetics variants with known GWAS associations we used the GWAS catalog v1.0.2, accessed March 2024 [43] and we overlapped SNPs with the 105 trQTL and 16 m6A-QTL to investigate their possible implication on GWAS-traits. For further colocalization analysis we calculate the probability that a GWAS hit shares the same causal variant as a trQTLs or m6A-QTL using bayesian colocalisation analyses as implemented by COLOC (v5.2.3) [62] and in a subset of GWAS studies. We used GWAS summary statistics from 16 studies listed on **Supplementary Table 9**. SNPs were filtered to keep variants in a 20 kb window around the lead QTL variant. The minor allele frequencies used for the analysis were those from the GWAS summary statistics. We reported probability that both GWAS and QTLs were shared as P(H4') = P(H4) / (P(H3) + P(H4)). Being H3 the probability of both traits to have different causal variants and H4 the probability of both traits sharing the same causal variant.

Data and code availability

Data has been deposited in ENA, under the following accession numbers PRJEB76585. This includes raw fast5 files, base-called aligned reads in BAM files and quantifications derived, as well as m6A modifications information as provided by m6Anet. All summary statistics from genetic associations is being deposited in Zenodo (DOI to be generated). All links be

live at the same time as a MedRxiv preprint that will be submitted in the coming weeks. This paper does not report original code.

Authors contribution statement

Conceptualization: AB, ARa, ETD, AV. Methodology: AB, ARa, AV. Software: ARe, AB, ARa, AV. Validation: ARe, ARa, AV. Formal analysis: ARe, NMRL, ARa, AV. Resources: ARe, GPY, CB. Data curation: ARe, ARa, AV. Writing-original draft: ARe, AB, ARa, AV. Writing- review & editing: ARe, AB, GPY, ARa, AV. Visualization: ARe, ARa, AV. Supervision: JDS, ETD, ARa, AV. Funding acquisition: JDS, ETD, AV.

Competing Interest Statement

Emmanouil T. Dermitzakis is currently an employee of GSK. His contribution to the work presented in this manuscript was performed before he joined GSK. All other authors declare no competing interests.

Funding

This work was supported by the Swiss National Science Foundation (FNS ME10662 and ME11559) to ETD, and the AMS Springboard Award (SBF007\100033) to AV.

References

1. Park, E., et al., *The Expanding Landscape of Alternative Splicing Variation in Human Populations.* Am J Hum Genet, 2018. **102**(1): p. 11-26.

2. Nurk, S., et al., *The complete sequence of a human genome.* Science, 2022. **376**(6588): p. 44-53.

3. Kelemen, O., et al., *Function of alternative splicing.* Gene, 2013. **514**(1): p. 1-30.

4. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.

5. Hou, W., et al., *Aberrant splicing of Ca.* Cell Mol Life Sci, 2024. **81**(1): p. 164.

6. Li, J. and G. Huang, *Insulin receptor alternative splicing in breast and prostate cancer.* Cancer Cell Int, 2024. **24**(1): p. 62.

7. Rogalska, M.E., C. Vivori, and J. Valcárcel, *Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects.* Nat Rev Genet, 2023. **24**(4): p. 251-269.

8. Li, Y.I., et al., *RNA splicing is a primary link between genetic variation and disease.* Science, 2016. **352**(6285): p. 600-4.

9. Raj, T., et al., *Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility.* Nat Genet, 2018. **50**(11): p. 1584-1592.

10. Takata, A., N. Matsumoto, and T. Kato, *Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci.* Nat Commun, 2017. **8**: p. 14519.

11. Caswell, J.L., et al., *Multiple breast cancer risk variants are associated with differential transcript isoform expression in tumors.* Hum Mol Genet, 2015. **24**(25): p. 7421-31.

12. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biol, 2016. **17**: p. 13.

13. Steijger, T., et al., *Assessment of transcript reconstruction methods for RNA-seq.* Nature Methods, 2013. **10**(12): p. 1177-1184.

14. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis.* Genome Biol, 2020. **21**(1): p. 30.

15. Stephenson, W., et al., *Direct detection of RNA modifications and structure using single-molecule nanopore sequencing.* Cell Genom, 2022. **2**(2).

16. Sedlazeck, F.J., et al., *Piercing the dark matter: bioinformatics of long-range sequencing and mapping.* Nat Rev Genet, 2018. **19**(6): p. 329-346.

17. Weirather, J.L., et al., *Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis.* F1000Res, 2017. **6**: p. 100.

18. Anvar, S.Y., et al., *Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing.* Genome Biol, 2018. **19**(1): p. 46.

19. Glinos, D.A., et al., *Transcriptome variation in human tissues revealed by long-read sequencing.* Nature, 2022. **608**(7922): p. 353-359.

20. Tilgner, H., et al., *Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events.* Nat Biotechnol, 2015. **33**(7): p. 736-42.

21. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

22. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.

23. Delaneau, O., et al., *Chromatin three-dimensional interactions mediate genetic effects on gene expression.* Science, 2019. **364**(6439).

24. Berulava, T., et al., *Changes in m6A RNA methylation contribute to heart failure progression by modulating translation.* Eur J Heart Fail, 2020. **22**(1): p. 54-66.

25. Soneson, C., et al., *A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes.* Nat Commun, 2019. **10**(1): p. 3359.

26. Oshlack, A. and M.J. Wakefield, *Transcript length bias in RNA-seq data confounds systems biology.* Biol Direct, 2009. **4**: p. 14.

27. Tang, A.D., et al., *Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns.* Nat Commun, 2020. **11**(1): p. 1438.

28. Schulz, L., et al., *Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts.* Genome Biol, 2021. **22**(1): p. 190.

29. Ezkurdia, I., et al., *Most highly expressed protein-coding genes have a single dominant isoform.* J Proteome Res, 2015. **14**(4): p. 1880-7.

30. Tung, K.F., C.Y. Pan, and W.C. Lin, *Dominant transcript expression profiles of human protein-coding genes interrogated with GTEx dataset.* Sci Rep, 2022. **12**(1): p. 6969.

31. Brown, A.A., et al., *Genetic analysis of blood molecular phenotypes reveals common properties in the regulatory networks affecting complex traits.* Nat Commun, 2023. **14**(1): p. 5062.

32. Consortium, G., *The GTEx Consortium atlas of genetic regulatory effects across human tissues.* Science, 2020. **369**(6509): p. 1318-1330.

33. Li, Y.I., et al., *Annotation-free quantification of RNA splicing using LeafCutter.* Nat Genet, 2018. **50**(1): p. 151-158.

34. Viñuela, A., *Genetic analysis of blood molecular phenotypes reveals regulatory networks affecting complex traits: a DIRECT study*, V.O.P.A.A. Brown, Editor. 2021.

35. Storey, J.D., et al., *qvalue: Q-value estimation for false discovery rate control.* R package version, 2015. **2**(0): p. 10.18129.

36. Hendra, C., et al., *Detection of m6A from direct RNA sequencing using a multiple instance learning framework.* Nat Methods, 2022. **19**(12): p. 1590-1598.

37. Xiong, X., et al., *Genetic drivers of m.* Nat Genet, 2021. **53**(8): p. 1156-1165.

38. Zhang, Z., et al., *Genetic analyses support the contribution of mRNA N.* Nat Genet, 2020. **52**(9): p. 939-949.

39. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.

40. Morales, J., et al., *A joint NCBI and EMBL-EBI transcript set for clinical genomics and research.* Nature, 2022. **604**(7905): p. 310-315.

41. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.

42. Vujkovic, M., et al., *Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis.* Nat Genet, 2020. **52**(7): p. 680-691.

43. Catalog, G. https://www.ebi.ac.uk/gwas/.

44. Hysi, P.G., et al., *Metabolome Genome-Wide Association Study Identifies 74 Novel Genomic Regions Influencing Plasma Metabolites Levels.* Metabolites, 2022. **12**(1).

45. Workman, R.E., et al., *Nanopore native RNA sequencing of a human poly(A) transcriptome.* Nature Methods, 2019. **16**(12): p. 1297-1305.

46. Chen, Y., et al., *A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines*. 2024: bioRxiv.

47. Montgomery, S.B., et al., *Transcriptome genetics using second generation sequencing in a Caucasian population.* Nature, 2010. **464**(7289): p. 773-7.

48. LaPierre, N. and H. Pimentel, *Accounting for isoform expression increases power to identify genetic regulation of gene expression.* PLoS Comput Biol, 2024. **20**(2): p. e1011857.

49. Gleeson Josie, et al., *Isoform-level profiling of m6A epitranscriptomic signatures in human brain*. 2024: bioRxiv.

50. Frye, M., et al., *RNA modifications modulate gene expression during development.* Science, 2018. **361**(6409): p. 1346-1349.

51. Wang, Y., et al., *Nanopore sequencing technology, bioinformatics and applications.* Nat Biotechnol, 2021. **39**(11): p. 1348-1365.

52. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.

53. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* Gigascience, 2021. **10**(2).

54. De Coster, W., et al., *NanoPack: visualizing and processing long-read sequencing data.* Bioinformatics, 2018. **34**(15): p. 2666-2669.

55. Leger, *pycoQC, interactive quality control for Oxford Nanopore Sequencing. Journal of Open Source Software, .* 2019.

56. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-30.

57. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.

58. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

59. Delaneau, O., et al., *A complete tool set for molecular QTL discovery and analysis.* Nat Commun, 2017. **8**: p. 15452.

60. Wei, T., & Simko, V., *R package "corrplot": Visualization of a Correlation Matrix* . https://github.com/taiyun/corrplot. 2017.

61. Lee, J.Y., et al., *Comparative evaluation of Nanopore polishing tools for microbial genome assembly and polishing strategies for downstream analysis.* Sci Rep, 2021. **11**(1): p. 20740.

62. Giambartolomei, C., et al., *Bayesian test for colocalisation between pairs of genetic association studies using summary statistics.* PLoS Genet, 2014. **10**(5): p. e1004383.

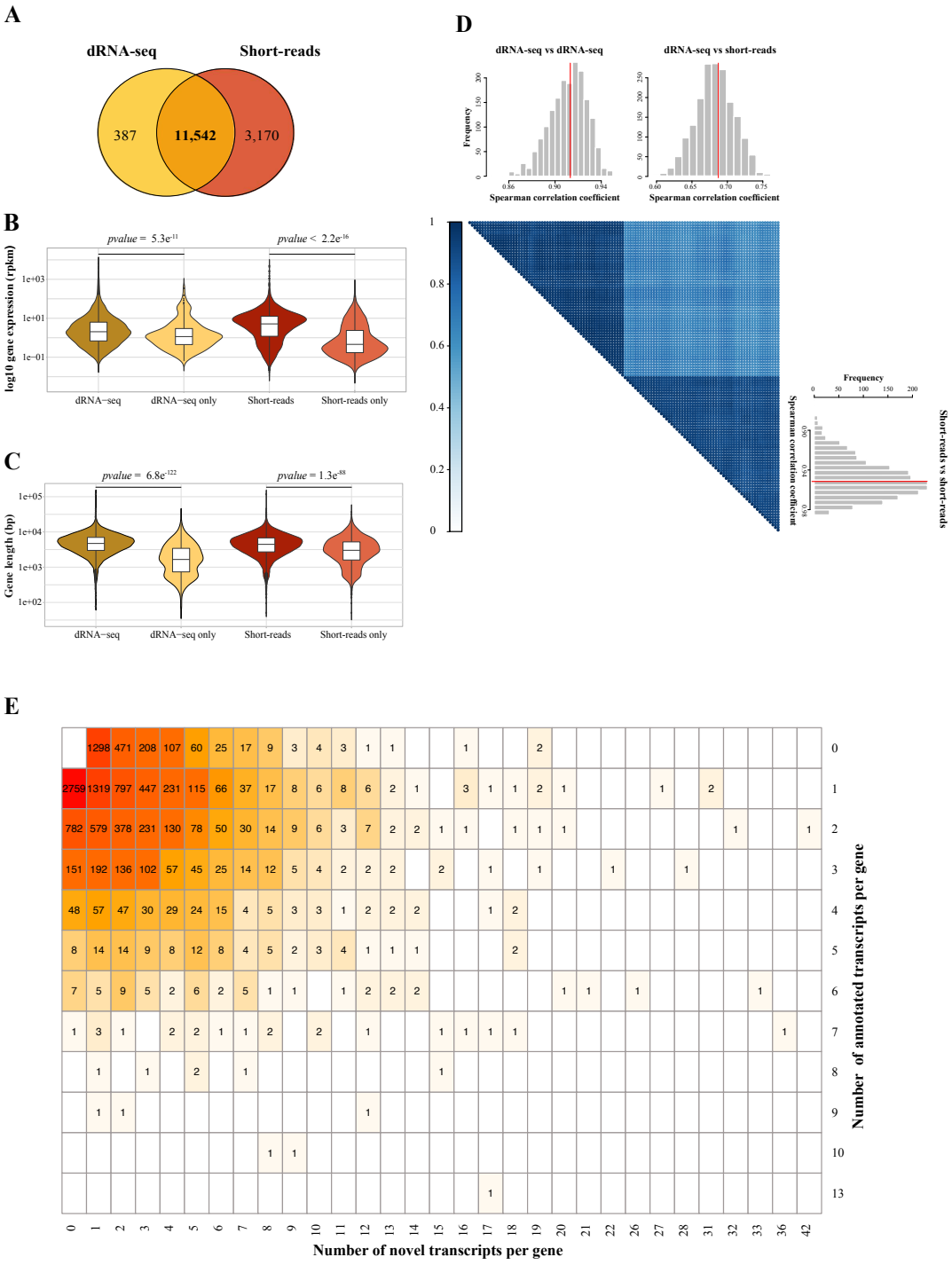

Figure 1. Comparison of protein-coding and lncRNA gene expression between dRNA-seq long-reads and Short-reads (Illumina). (A) Venn diagram showing the protein-coding genes and lncRNA expressed in 50% of the sample by dRNA-seq long-reads and Short-reads technologies. In the samples, the expression of 11,154 genes and lncRNA were detected by both technologies. In dRNA-seq long-reads, 3,170 genes were not detected and 387 were missed by Short-reads (Illumina). **(B)** Violin plots comparing the $\log_{10}$ gene expression (rpkm) between all genes detected using dRNA-seq long-reads (dark yellow) and the one only detected with dRNA-seq (yellow) (Wilcoxon pvalue= $5.3e^{-11}$) and between all gene detected with Short-reads (dark red) and the one only detected with Short-reads and not with dRNA-seq (light red) (Wilcoxon pvalue=$2.2e^{-16}$). **(C)** Violin plots comparing the gene length (bp) between all genes detected using dRNA-seq long-reads (dark yellow) and the one only detected with dRNA-seq (yellow) (Wilcoxon pvalue= $6.8e^{-122}$) and between all genes detected with Short-reads (dark red) and the one only detected with Short-reads and not with dRNA-seq (light red) (Wilxocon pvalue=$1.3e^{-88}$). **(D)** Pair-wise Spearman correlation between the overall gene expression in the same 60 LCLs samples among and between technologies. The correlation observed inside the same technology, median ρ = 0.906 and 0.95 in dRNA-seq and Short-reads, respectively. Whereas the correlation in gene expression between the two technologies was median ρ = 0.695. **(E)** Heatmap representing the number of annotated and novel transcripts per gene quantified in our dataset.

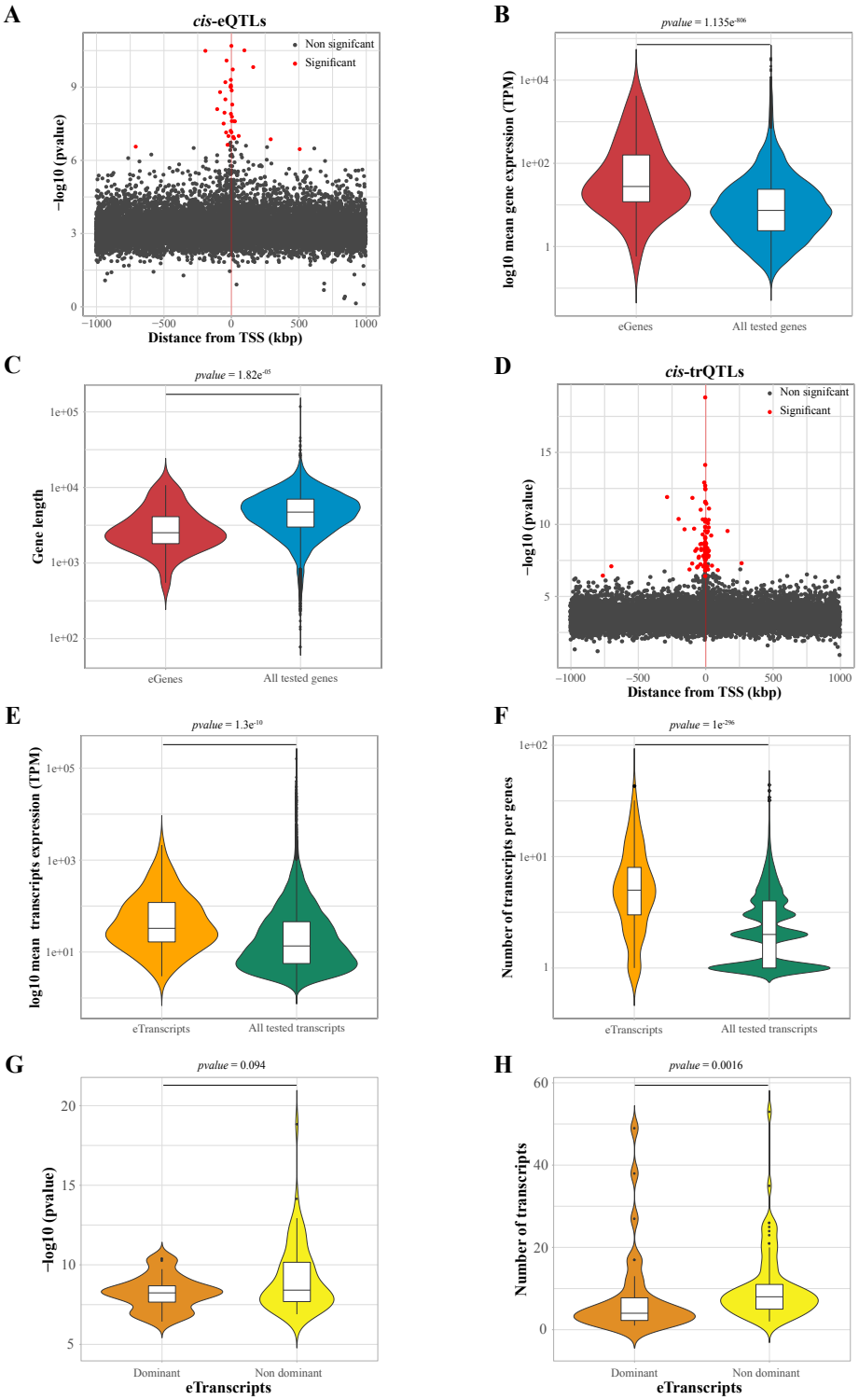

Figure 2. cis-eQTL and trQTLs discovery using dRNA-seq data. (A) Scatter plot representing the distance from the TSS of the gene (Kbp) versus the -$\log_{10}$ *pvalue* for all the SNPs tested in *cis* for gene-SNP associations. In red are highlighted the significant *cis*-eQTLs (34; FDR5%). **(B)** Violin plot comparing the mean gene expression (TPM) between genes with a significant *cis*-eQTL signal (red) and all the genes tested (blue) (Wilcoxon *pvalue* = 1.135e^{-806}). **(C)** Violin plot represents the different gene lengths of genes with a significant *cis*-eQTL signal (red) and all the genes tested (blue) (Wilcoxon *pvalue* = 1.82e^{-05}). **(D)** Scatter plot representing the distance from the TSS of the gene (Kbp) versus the -$\log_{10}$ *pvalue* for all the SNPs tested in *cis* for transcripts-SNP associations. In red are highlighted the significant *cis*-trQTLs (105; FDR 5%). **(E)** Violin plot comparing the mean gene expression (TPM) between transcripts with a significant *cis*-trQTL signal (orange) and all the genes tested (green) (Wilcoxon *pvalue* = 1.3e^{-10}). **(C)** Violin plot represents the different number of transcripts per transcript with a significant *cis*-trQTL signal (orange) and all the tested transcripts (green) (Wilcoxon *pvalue* = 1e^{-296}). **(G)** Violin plot showing the difference of *pvalue* of eTranscripts with dominant (orange) or non-dominant (yellow) transcripts (Wilcoxon *pvalue* 0.094). **(H)** Violin plot showing the difference in number of transcripts for eTranscripts with dominant (orange) or non-dominant (yellow) transcripts (Wilcoxon *pvalue* 0.0016).

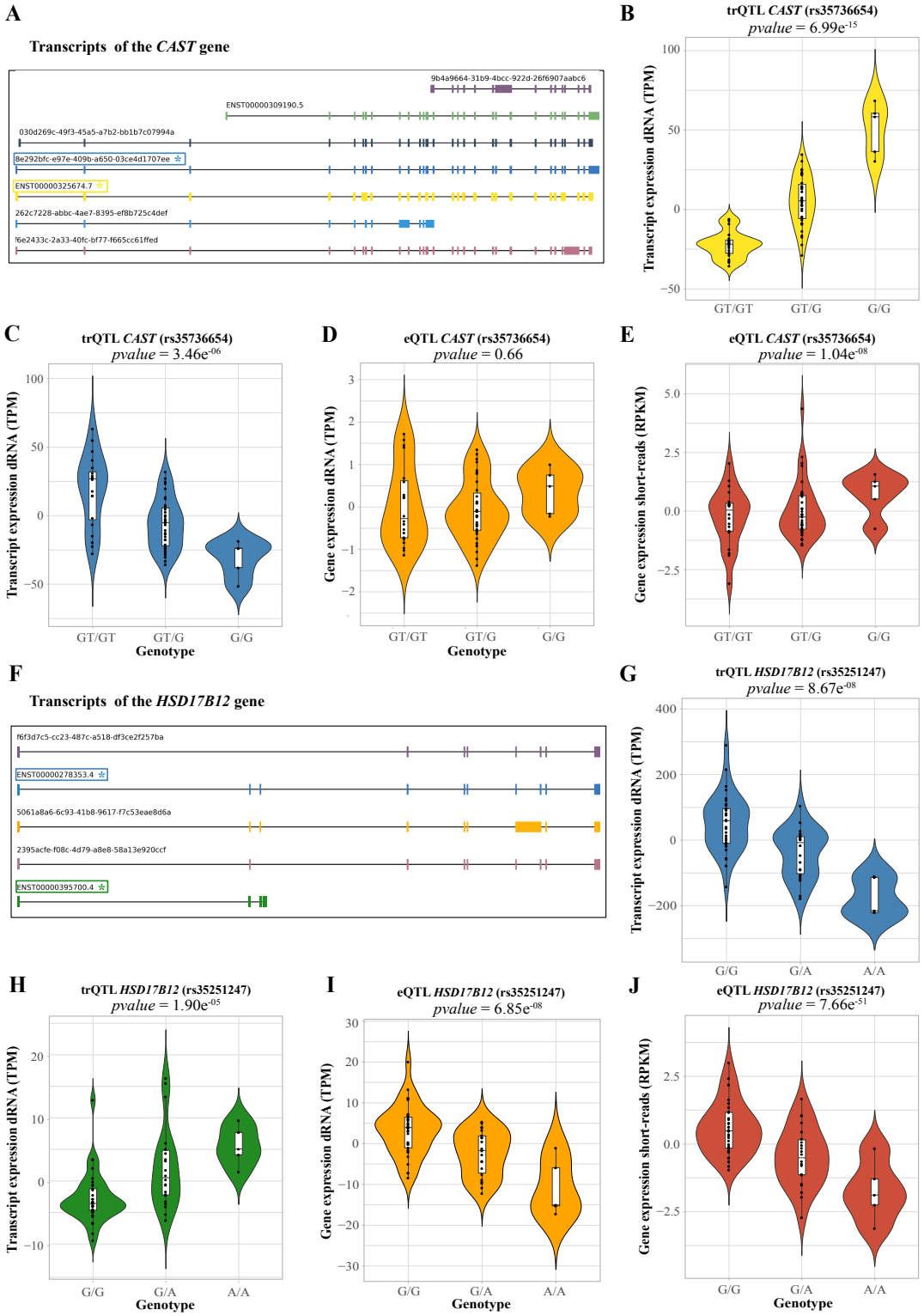

Figure 3. Significant trQTL effects on transcripts of the *CAST* and *HSD17B12* genes. (A) Cartoon summarizing the structure of the transcripts detected with dRNA_seq long-reads for the *CAST* gene. Rectangular boxes and asterisk highlighting the transcripts affected by significant cis-trQTLs (blue and yellow). Violin-plot representing the genotype versus transcript expression expressed transcripts per million (TPM), where outliers are shown as black dots. In **(B)** the violin plot of the effect of the SNP rs35736654 for the ENST00000325674.7 transcripts (*pvalue* = 6.99e^{-15}, yellow). In **(C)** the effect of the SNP rs357366548 for the novel transcript e292bfc-e97e-409b-a650-03ce4d1707ee expression (*pvalue*= 3.46e^{-06}, blue). In **(D** and **E)**, the effect of the SNP rs35736654 on the expression of the *CAST* gene was detected with dRNA-seq long-read and short-reads sequencing respectively (*pvalue*= 0.66 and 1.04e^{-08} respectively, orange and red). **(F)** Cartoon summarizing the structure of the transcript for the transcripts detected with dRNA_seq long-reads for the *HSD17B12* gene. Rectangular boxes and asterisk highlighting the transcripts affected by significant *cis*-trQTLs (blue and green). **(G-H)** Violin-plot showing the effect of the SNP rs35251247 on the ENST00000278353.4 transcript of the *HSD17B12* gene (*pvalue*= 8.67e^{-08}, blue) and the ENST00000395700.4 transcript (*pvalue*=1.90e^{-05}, green). **(I-J)** Effect of the SNP rs35251247 on the *HSD17B12* gene expression detected by dRNA-seq and short-reads sequencing (*pvalue* =6.85e^{-08} and 7.66e^{-51}, orange and red respectively).

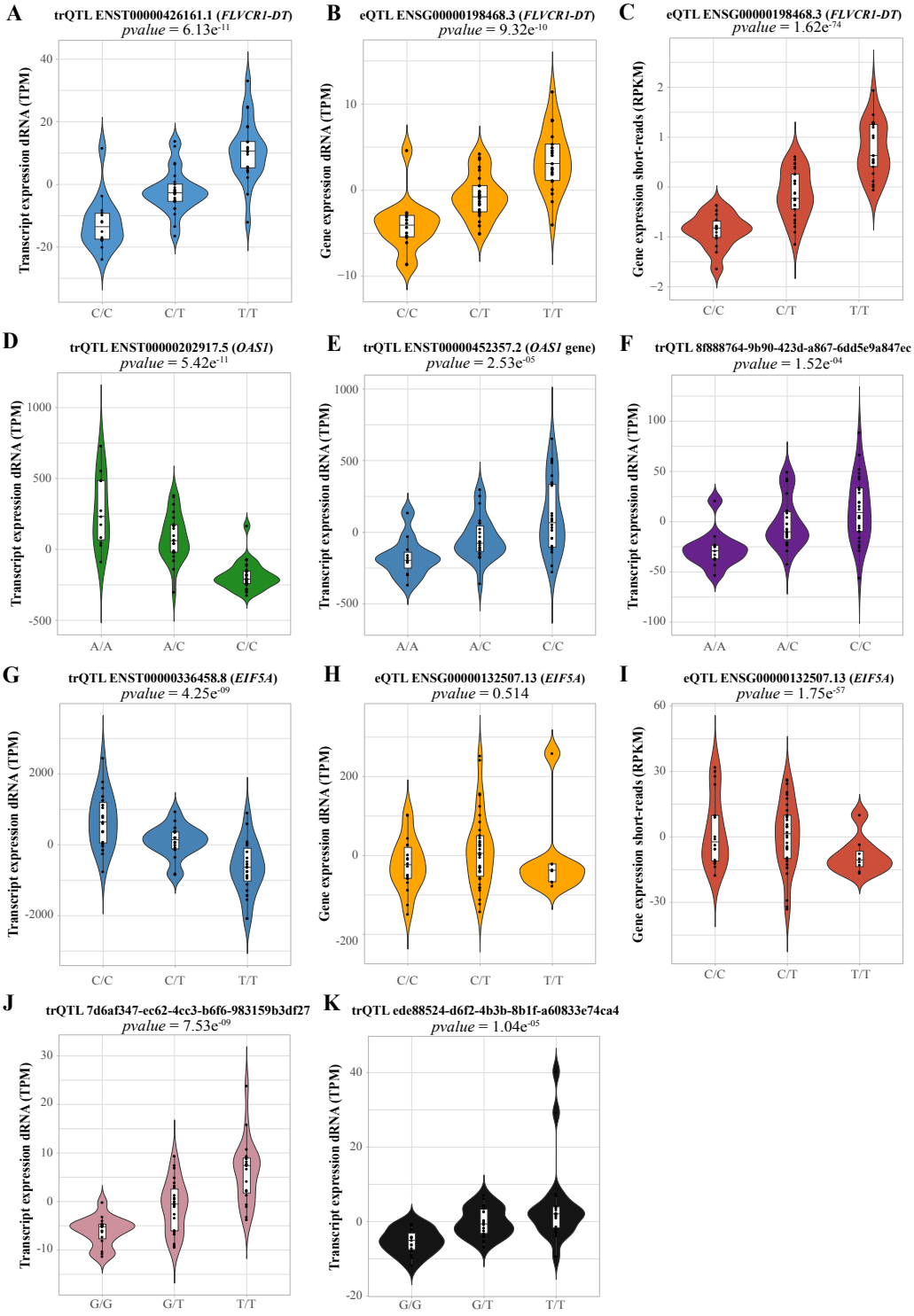

Figure 4. Transcripts-specific effects of trQTL detected by long-reads. (A-C) Violin-plot representing the genotype versus transcript or gene expression in transcripts per million (TPM) or Reads Per Kilobase per Million mapped reads (RPKM), where outliers are shown as black dots. **(A)** Shows the significant effect of the SNP rs2279692 on the ENST00000426161.1 transcript of the *FLVCR1-DT* gene (*pvalue* = 6.13e^{-11}, blue) on the *FLVCR1-DT* gene expression detected with **(B)** dRNA-seq (*pvalue* = 9.32e^{-10}, orange) and **(C)** short-reads (*pvalue* = 1.62e^{-74}, red). **(D-F)** Violin-plot representing the genotype versus transcript expression expressed transcripts per million (TPM), where outliers are shown as black dots. The trQTL effect of the rs1154970 SNP on the expression of three of the transcripts detected for the *OAS1* gene (no dominant transcript) are represented in **(D)** ENST00000202917.5 transcript, **(E)** ENST00000452357.2 and **(F)** 8f888764-9b90-423d-a867-6dd5e9a847ec novel transcript (*pvalue* = 5.42e^{-11}, 2.53e^{-05} and 1.52e^{-04}, green, blue and purple respectively). **(G-I)** Violin-plot representing the genotype versus transcript expression or gene expression in TPM and RPKM respectively, outliers are shown as black dots. **(G)** shows the significant effect of the SNP rs4796398 on the dominant annotated transcript of the *EIF5A* gene ENST00000336458.8 (*pvalue* = 4.25e^{-09}, blue); while non-significant effect was detected with **(H)** dRNA- long reads and **(I)** short-reads (*pvalue*= 0.514 and 1.68e^{-56} , orange and red respectively). **(J-K)** Violin-plot representing the genotype versus transcript expression expressed in TPM, where outliers are black dots. Here the significant effect is shown for two novel transcripts of the *ARPC2* gene for the SNP rs2271541 (*pvalue*= 7.53e^{-09} and 1.04e^{-05}, pink and black respectively).

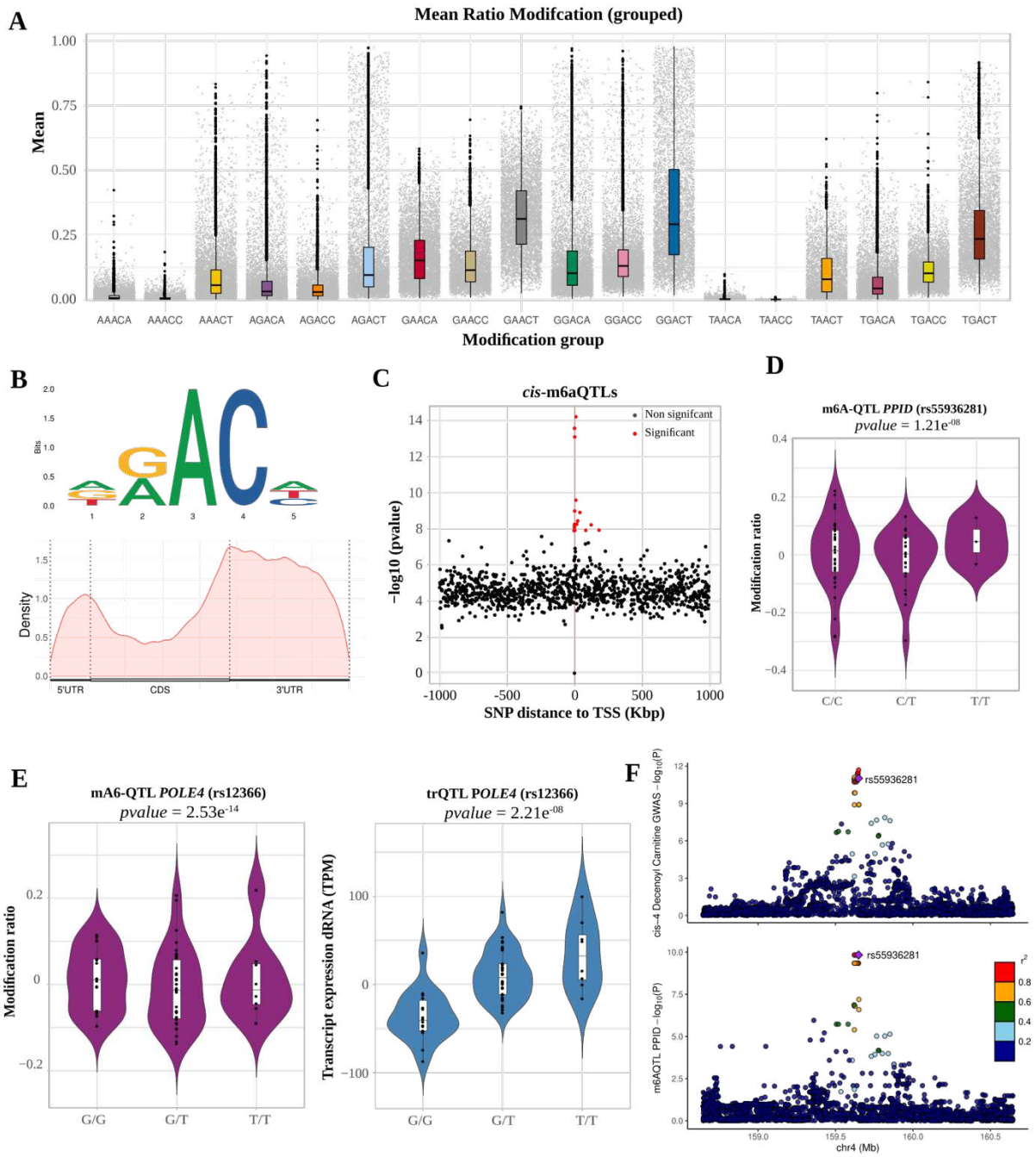

Figure 5: m6aRNA modifications and m6aQTL effects. (A) The boxplot shows the mean ratio of modified versus unmodified reads per modification and across the samples. Values were grouped by each of the 18 motifs that would be identified as having m6A modifications (N= 255,014). (**B**) Motif analysis of the m6A modifications identifies the m6A DRACH consensus motif (D–A, G, or U, R–A or G while H is A, C or U) (upper panel) and guitar plot showing the distribution of m6A modifications along the mRNAs bodies (lower panle). (**C**) Scatter plot representing the distance from the TSS of the gene (kbp) versus the -$\log_{10}$*pvalue* for all the SNPs tested in cis for m6A modification-SNP associations. In red are highlighted the significant cis-eQTLs (16; FDR5%). (**D-E**) Violin-plot showing the genotype versus m6ARNA modification ratio, where outliers are shown as black dots; in (**D**) the significant effect of the SNP rs55936281 on the RNA modification motifs AGGCT of the *PPID* gene (ENSG00000171497 *pvalue*=1,21e^{-8}, purple) in (**E**) the significant effect of the SNP rs12366 on the RNA modification motifs GGACC of the *POLE4* transcript is represented (ENST00000483063.1, *pvalue*= 2.53e^{-14}, purple) and it significant effect on the most abundant transcript of the *POLE4* gene expression (TPM) (ENST00000483063.1, *pvalue*= 2.21e^{-08}, blue). (**F**) Locuszoom plots of 600kb region around the m6A-QTL (rs7477) for the gene *CENPV* that is also a GWAS associated with ALS (no summary statistic released [Kreshnik B Ahmeti, Neurobiol Aging 2013]. On the x-axis the 500kb genomic window around the SNP on chr17 and on the y-axis the -log10 of the m6A-QTL *pvalue*, every dot is a SNP and the color code represents the r square LD with the leading SNP; the box at the bottom have represented the genes mapping the zoomed genomic region. LocusZoom plot generated with the R package LocusCompare [https://www.nature.com/articles/s41588-019-0404-0].

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfigures.docx
- SupplementaryTables.xls