Improved detection of genetic effects on promoter usage with augmented transcript annotations

Andreas Vija[1,2], Kaur Alasoo[1]

[1]Institute of Computer Science, University of Tartu, Tartu, Estonia

[2]STACC OÜ, Tartu, Estonia

Abstract

Disease-associated non-coding variants can modulate their target genes by disrupting multiple mechanisms, including regulating total gene expression level, splicing, alternative polyadenylation or promoter usage. Quantifying promoter usage from standard RNA sequencing data is challenging due to incomplete reference transcriptome annotations and low read coverage observed at the ends of transcripts. We previously developed the txrevise tool (https://github.com/kauralasoo/txrevise) to quantify promoter usage events from RNA-seq data using reference transcriptome annotations. Here, we augment the txrevise promoter event annotations with experimentally identified Cap Analysis of Gene Expression (CAGE) promoters from the FANTOM5 project. Applying the new annotations to RNA-seq data from 358 individuals, we found that augmented promoter event annotations increased the power to detect promoter usage quantitative trait loci (puQTLs) by ~30%. However, concordance between puQTLs inferred from RNA-seq data and those directly measured using CAGE remained low, suggesting that additional experimental and computational improvements are needed to capture the full range of regulatory effects of non-coding variants.

Introduction

Genetic variants regulating promoter usage can play an important role in human complex traits (Alasoo et al., 2019; Garieri et al., 2017; Kubota and Suyama, 2022). Promoter usage can be directly quantified using experimental techniques that capture 5' ends of transcripts such as Cap Analysis of Gene Expression (CAGE) (Shiraki et al., 2003), but currently only one such population-level human dataset exists (Garieri et al., 2017). Alternatively, promoter usage can be quantified from standard bulk RNA sequencing (RNA-seq) data (Figure 1a). The advantage of this approach is that bulk RNA-seq data is readily available from thousands of individuals and over a hundred different cell types or tissues (Kerimov et al., 2021; The GTEx Consortium, 2020).

However, quantification of promoter usage from standard RNA-seq data is complicated by multiple factors. First, read coverage is much lower at the ends of transcripts (Love et al., 2016; Roberts et al., 2011), which makes it difficult to precisely detect the location of each transcription start site (TSS) (Pertea et al., 2015). Consequently, it is often hard to ascertain more than one TSS per gene from RNA-seq data (Adiconis et al., 2018). Secondly, transcripts contain overlapping exons which means that most reads cannot be uniquely assigned to a specific transcript (Figure 1a). Thus, methods for quantifying promoter usage from RNA-seq such as txrevise (Alasoo et al., 2019) and proActiv (Demircioğlu et al., 2019) rely heavily on pre-existing promoter annotations (Figure 1a) from databases such as Ensembl (Howe et al., 2021). This means that even if a genetic effect on promoter usage is clearly visible from RNA-seq read coverage track, it might not be detected by existing methods due to incomplete promoter annotations (Figure 1c).

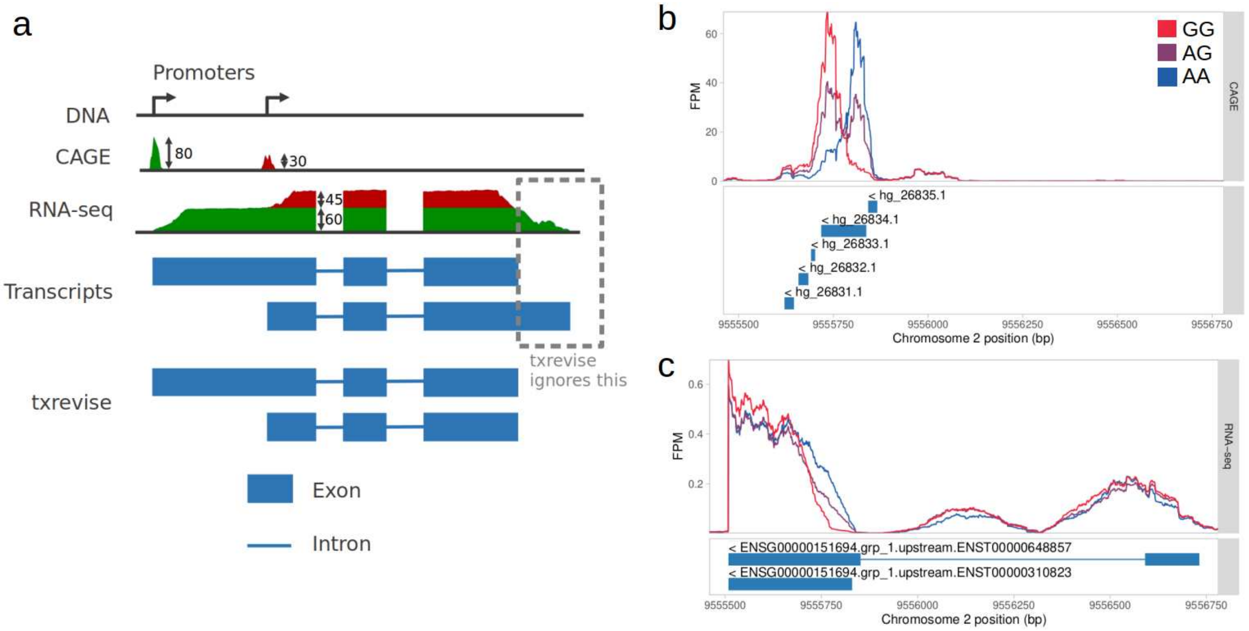


Figure 1. Comparison of RNA sequencing methods for promoter usage quantification. (**a**) Fictional gene with two alternative promoters. CAGE directly sequences the 5' ends of expressed transcripts and thus detects two peaks corresponding to the two alternative promoters. Promoter usage is quantified by counting the reads overlapping the two peaks. RNA-seq captures reads across the full length of the expressed transcripts with a notable drop at the beginning. Most reads originating from this gene are compatible with both known transcripts. Promoter usage can be quantified by estimating the relative expression of the two transcripts that best explains the observed read coverage pattern across the gene. However, transcript annotations often couple alternative promoters with unrelated splicing or 3' end events. Txrevise overcomes this by constructing new annotations corresponding to independent promoter usage events. (**b**) Promoter usage QTL affecting *ADAM17* (ENSG00000151694). The CAGE read coverage stratified by the

genotype of the lead puQTL variant (rs12692386) shows strong genotype-dependent shift in promoter usage. (**c**) RNA-seq read coverage signal captures similar change, but this does not correspond to any annotated *ADAM17* promoters.

Previous research has demonstrated that detection of alternative polyadenylation events can be significantly improved by incorporating experimental annotations such as data from 3' RNA-seq experiments (Ha et al., 2018; Shah et al., 2021). Here, we augment txrevise promoter annotations using experimental CAGE (Shiraki et al., 2003) data from the FANTOM5 project (FANTOM Consortium and the RIKEN PMI and CLST et al., 2014). We demonstrate that incorporating experimentally detected promoter annotations improves concordance between CAGE and RNA-seq data and increases the number of detected promoter usage quantitative trait loci (puQTLs) by around 30%. Nevertheless, overall concordance between puQTLs detected by CAGE and RNA-seq remains low, suggesting that the two approaches have distinct strengths and weaknesses.

Results

Concordance of puQTLs detected by CAGE and RNA-seq

To assess the concordance between the puQTLs detected by CAGE and RNA-seq, we re-analysed two transcriptomic datasets generated from lymphoblastoid cell lines. The Garieri_2017 (Garieri et al., 2017) dataset contained CAGE data from 154 individuals from the 1000 Genomes (1000 Genomes Project Consortium et al., 2015) and GENCORD (Gutierrez-Arcelus et al., 2013) cohorts. The GEUVADIS (Lappalainen et al., 2013) datasets contained RNA-seq data from 358 individuals from the 1000 Genomes cohort. All individuals were of European ancestries and 78 individuals were shared between the two datasets.

We first compared CAGE against RNA-seq using Ensembl reference promoter annotations. Alternative promoter annotations were extracted from reference transcriptome with txrevise (Alasoo et al., 2019). We found that using the same +/- 200 kb *cis* window and 5% false discovery rate (FDR) sigificance threshold, CAGE detected at least one significant puQTL for more genes than txrevise (1145 vs 979 genes), even though the RNA-seq dataset was more than two times larger (154 vs 358 samples) (Table 1). When varying the gene expression threshold, we found that CAGE was more sensitive for lowly expressed genes whereas txrevise found more associations at higher gene expression thresholds (Table 1, Supplementary Figure 1). Nevertheless, the agreement between the two datasets was small with only 307 genes having a significant puQTL in both datasets (Figure 2a). An example puQTL for *ADAM17* detected by CAGE and missed by txrevise is illustrated on Figure 2b. Close inspection of the corresponding read coverage plots (Figure 1b-c) revealed that while the puQTL signal was clearly visible from the RNA-seq data, the association was missed because neither of the two annotated promoters (corresponding to transcripts ENST00000648857 and ENST00000310823) overlapped the downstream alternative promoter captured by CAGE.

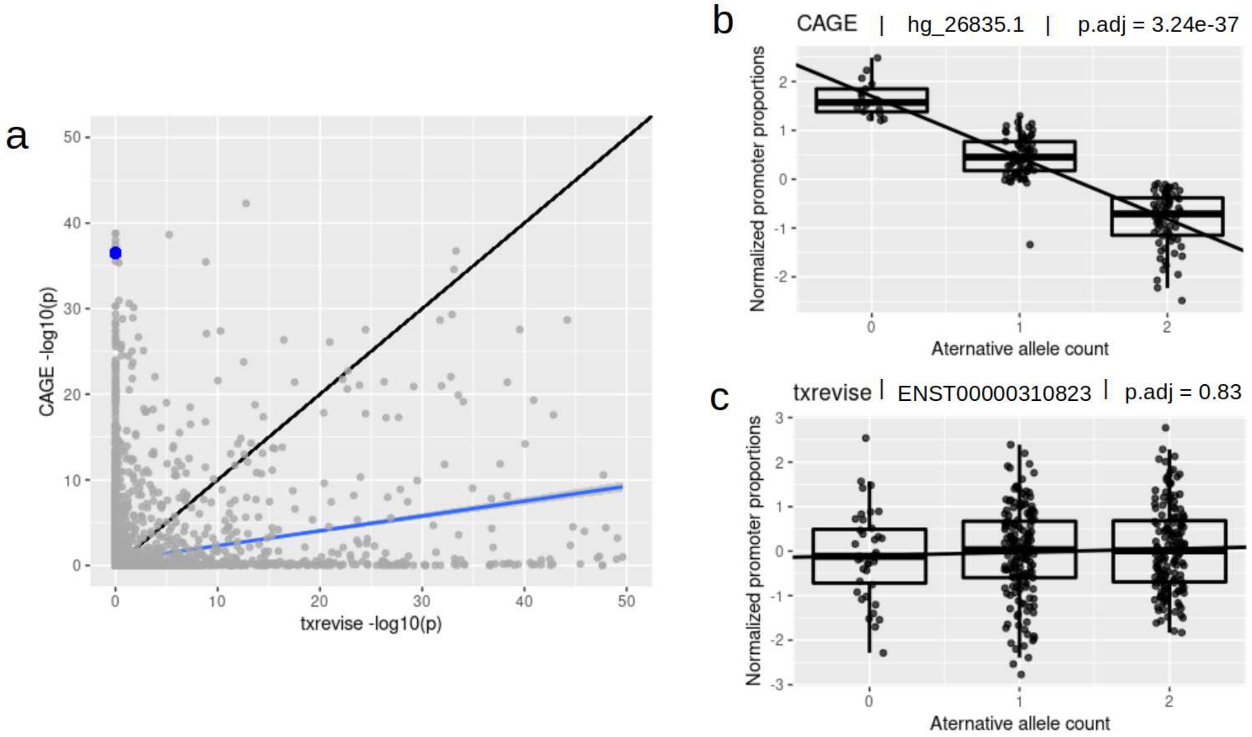


Figure 2. Concordance of promoter usage QTLs detected by CAGE and txrevise. (**a**) Scatterplot of the lead variant p-values of each gene from CAGE and reference-based txrevise analysis. The identity line (black) diverges from the regression line (blue), indicating low concordance between the two methods. The blue dot corresponds to the lead puQTL variant of *ADAM17*, rs12692386. The variant was significantly associated with *ADAM17* promoter usage in CAGE analysis but not in txrevise analysis. (**b**) Normalised usage of the *ADAM17* CAGE promoter hg_26835.1 stratified by the genotype of the puQTL lead variant (rs12692386). (**c**) Normalised usage of the *ADAM17* ENST0000031082 txrevise reference promoter stratified by the genotype of the puQTL lead variant (rs12692386).

Augmenting promoter annotations using CAGE data

		Number of puQTLs detected			
Min TPM threshold	**% of genes above threshold**	**CAGE**	**txrevise (reference only)**	**txrevise (augmented)**	**puQTL % increase**
None	100%	1208	1073	1393	29.8%
0.1	79.4%	1235	1051	1359	29.3%
1	62.9%	1145	979	1287	31.5%
10	31.5%	564	663	868	30.9%
100	4.8%	87	152	201	32.2%

Table 1. Number of puQTLs (FDR < 0.05) detected at various gene expression thresholds by CAGE, reference-based txrevise and augmented txrevise. As transcripts per million (TPM) > 1 allowed filtering out many lowly expressed genes without a major drop in the number of puQTLs detected, it was chosen as the threshold.

To overcome the limitation of incomplete promoter annotations in txrevise analysis, we obtained a list of experimentally detected promoters from the FANTOM5 project (Abugessaisa et al., 2017) and used a simple heuristic approach to construct novel transcript annotations based on these promoters (Figure 3) (see Methods). Briefly, we used the FANTOM5 promoters to construct new alternative first exons if the novel promoters were in an existing exon or 1000 bp upstream of one and if the newly constructed promoter was at least 20 nucleotides away from any existing promoter. This process increased the number of txrevise alternative promoter annotations from 72,292 to 114,768 (37%). Furthermore, 1650 genes that previously had only one annotated promoter now had more than one, thus enabling promoter usage quantification for those genes.

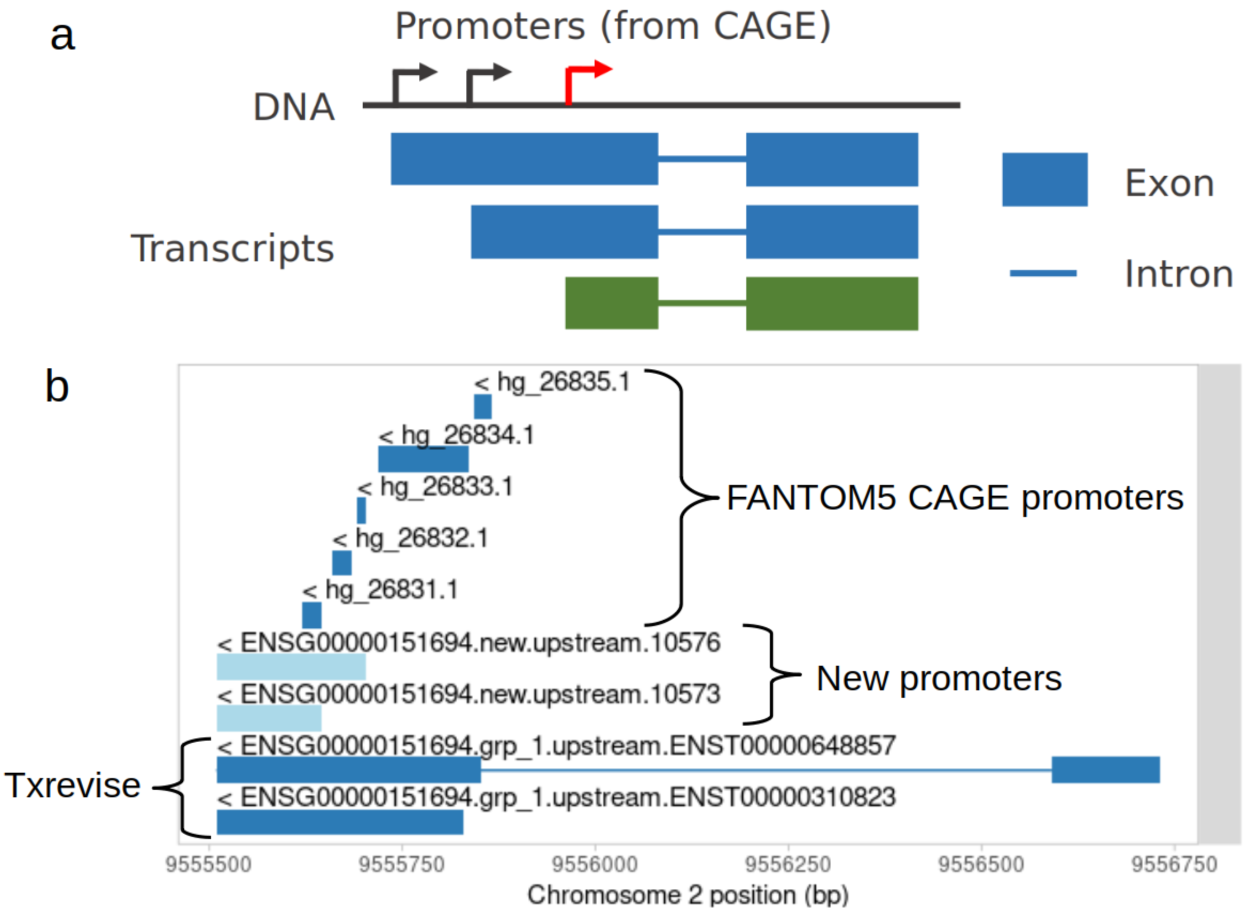

a
Promoters (from CAGE)
DNA
Transcripts
Exon
Intron
b
< hg_26835.1
< hg_26834.1
< hg_26833.1
< hg_26832.1
< hg_26831.1
FANTOM5 CAGE promoters
< ENSG00000151694.new.upstream.10576
< ENSG00000151694.new.upstream.10573
New promoters
< ENSG00000151694.grp_1.upstream.ENST00000648857
< ENSG00000151694.grp_1.upstream.ENST00000310823
Txrevise
9555500
9555750
9556000
9556250
9556500
9556750
Chromosome 2 position (bp)

Figure 3. Constructing new transcript annotations based on CAGE peaks. (**a**) Fictional gene with two alternative promoters (black) corresponding to two transcripts (blue) starting from those promoters, but also an additional promoter (red) that has no existing transcript annotation but for which a hypothetical transcript (green) could be constructed based on existing transcripts. (**b**). Two new txrevise promoter annotations (light blue) constructed for *ADAM17* by augmenting existing txrevise promoter annotations (dark blue, bottom) with CAGE promoters (dark blue, top) from FANTOM5.

Impact of augmented promoter annotations on puQTL detection

Next, we re-analysed the GEUVADIS dataset using the augmented txrevise promoter annotations. We found that augmented annotations increased puQTL yield by ~30% at all gene expression level thresholds (Table 1, Figure 4a). Similarly, the number of shared puQTLs genes detected by both CAGE and txrevise increased from 307 to 397 (Figure 4b). One such example was the previously missed *ADAM17* gene, but even with augmented annotations, the association detected from RNA-seq data (Figure 4c-d) was much weaker compared to the CAGE signal (Figure 2b). Similarly, the overall concordance between CAGE and txrevise still remained relatively low (Figure 4b).

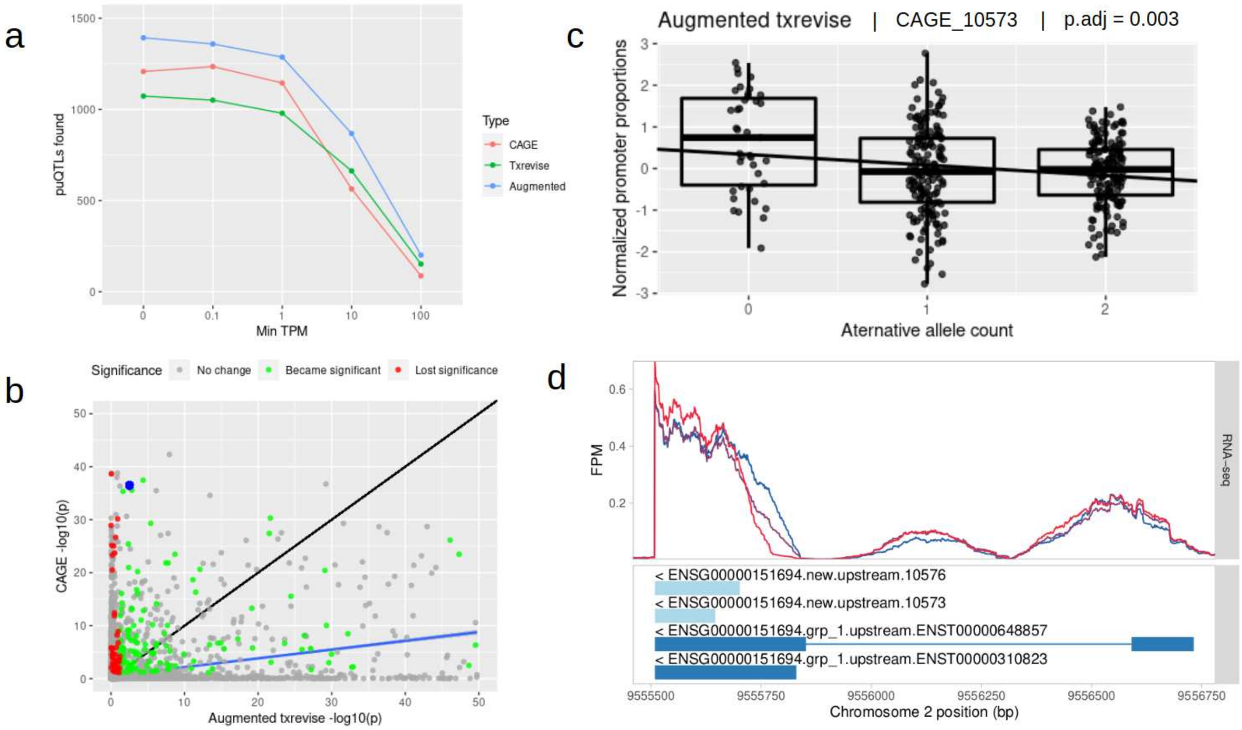


Figure 4. Impact of augmented promoter annotations on puQTL detection. (**a**) Number of genes with at least one significant puQTL (y-axis) as a function of gene expression level threshold (x-axis) (TPM - transcripts per million). (**b**) The lead variant p-values of each gene for CAGE and txrevise with an identity line (black) and a regression line (blue). Green dots represent genes for which a significant puQTL was detected only after promoter augmentation, red dots represent genes that lost a significant puQTL after augmentation. The blue dot corresponds to the lead variant of *ADAM17*, rs12692386. (**c**) Normalised usage of the newly added ENSG00000151694.new.upstream.10573 (CAGE_10573) txrevise promoter stratified by the genotype of the puQTL lead variant (rs12692386) (**d**) RNA-seq read coverage at the *ADAM17* promoter stratified by the genotype of the puQTL lead variant (rs12692386). The GG genotype (red line) is associated with a shift towards an upstream promoter relative to the AA genotype (blue line). The newly added promoter annotation (ENSG00000151694.new.upstream.10573, light blue) can better capture this shift compared to the two existing reference-based promoters (dark blue).

Discussion

We performed puQTL analysis in lymphoblastoid cell lines using two complementary technologies: CAGE that directly sequences 5' ends of transcripts and txrevise that can leverage reference transcript annotations to capture promoter usage events from full-length RNA-seq data. We found that the concordance in the puQTLs detected with the two approaches was generally low. Augmenting reference transcript annotations with novel FANTOM5 promoters increased the ability of txrevise to detect puQTLs by 30%, but concordance with CAGE puQTLs still remained low. We believe that the discordance is primarily due to differences in technology with CAGE being able to better distinguish promoters of lowly expressed genes and having higher signal-to-noise ratio due to more direct measurement. However, since the CAGE and RNA-seq data were generated by two independent studies, we cannot rule out that other technical factors might contribute to the observed differences.

Our results indicate that a significant proportion of alternative promoter annotations are still missing from the Ensembl database. Consequently, we found that augmenting reference transcripts with experimentally determined promoters from the FANTOM5 project significantly increased the number of puQTLs detectable from RNA-seq data. As a result, using augmented promoter annotation to re-process publicly available RNA-seq eQTL datasets by projects such as the eQTL Catalogue (Kerimov et al., 2021) has a great potential to increase the number of puQTLs detected. Future studies can explore if incorporating additional experimentally derived promoter annotations such as those generated by the RAMPAGE project (Moore et al., 2021) or transcriptome assembly methods (Kubota and Suyama, 2022) can further improve the ability to detect puQTLs from RNA-seq data.

Methods

FANTOM5 promoter annotations

We downloaded annotations of 210,250 human promoters from the FANTOM5 database (Abugessaisa et al., 2017; FANTOM Consortium and the RIKEN PMI and CLST et al., 2014), which was constructed based on CAGE peaks. Of these, 96,562 promoter annotations associated with an autosomal gene were kept. The name of the associated gene was mapped to an Ensembl id using the eQTL Catalogue gene metadata files (https://doi.org/10.5281/zenodo.3366011). Of the remaining annotations, 93,663 promoters from 20,201 genes had a gene name which mapped to exactly one Ensembl id. Of these, 93,554 promoters from 20,193 genes were mapped to a unique chromosome and were thus used in further analysis.

GAGE data processing

We downloaded the raw CAGE sequencing data from the Garieri_2017 (Garieri et al., 2017) study from ArrayExpress (E-MTAB-5835). CAGE reads were mapped to the GRCh38 human reference genome using Burrows-Wheeler Aligner v0.7.12 (Li and Durbin, 2009) and multi-mapping reads were discarded. The number of CAGE reads corresponding to each promoter was counted using featureCounts (Liao et al., 2014). On average, 46.0% of all CAGE reads overlapped with the TSS of some FANTOM5 promoter.

Based on these read counts, we omitted from further analysis all CAGE promoters with zero total mapped reads across all samples. After this, 90,003 promoters from 18,546 genes remained.

To quantify promoter usage and not general gene expression in CAGE data, the number of reads assigned to each promoter was divided by the total number of reads assigned to all promoters of the same gene. Missing promoter usage values were replaced by the mean calculated across all individuals. Finally, rank-based inverse normal transformation was used to enable more robust use of linear models (McCaw et al., 2019).

Genotype data processing

We re-analyzed genotype data from 154 individuals of European descent from the Garieri_2017 study, 86 of which were from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) and 68 from the GENCORD project (Gutierrez-Arcelus et al., 2013). Genotypes from the GENCORD project were imputed to the 1000 Genomes reference panel as described previously (Kerimov et al., 2021). Whole genome sequencing genotypes for the 1000 Genomes Project samples were downloaded from the 1000 Genomes Project FTP server (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/). The 9.2 million genetic variants shared by these two datasets were used in the rest of the analysis.

RNA-seq data processing

Txrevise promoter usage events (both original and augmented) were quantified with the eQTL-Catalogue/rnaseq workflow and normalised with the eQTL-Catalogue/qcnorm workflow as described previously (Kerimov et al., 2021). Only the 15,275 genes present in both txrevise transcript annotations and the filtered FANTOM5 annotations were used for downstream analysis.

Genes with very low expression levels are likely to be biologically insignificant and their expression estimates are vulnerable to noise. Gene expression levels were calculated based on the output of featureCounts generated during the execution of the eQTL-Catalogue/rnaseq pipeline on the RNA-seq data. For various TPM values, only genes with at least 5% of individuals exhibiting at least that much expression were considered.

Promoter usage QTL analysis

To map puQTLs, we used the eQTL-Cataloge/qtlmap Nextflow workflow built on top of fastQTL (Ongen et al., 2016) and QTLtools (Delaneau et al., 2017) by the eQTL Catalogue project (Kerimov et al., 2021). For every gene, only genetic variants within the +/- 200kb *cis* window centred around the canonical Ensembl promoter of the gene were considered. The coordinates of genes were obtained from the eQTL Catalogue gene metadata files (https://doi.org/10.5281/zenodo.3366011). Multiple testing correction was performed as described previously (Kerimov et al., 2021).

Creating new transcript annotations

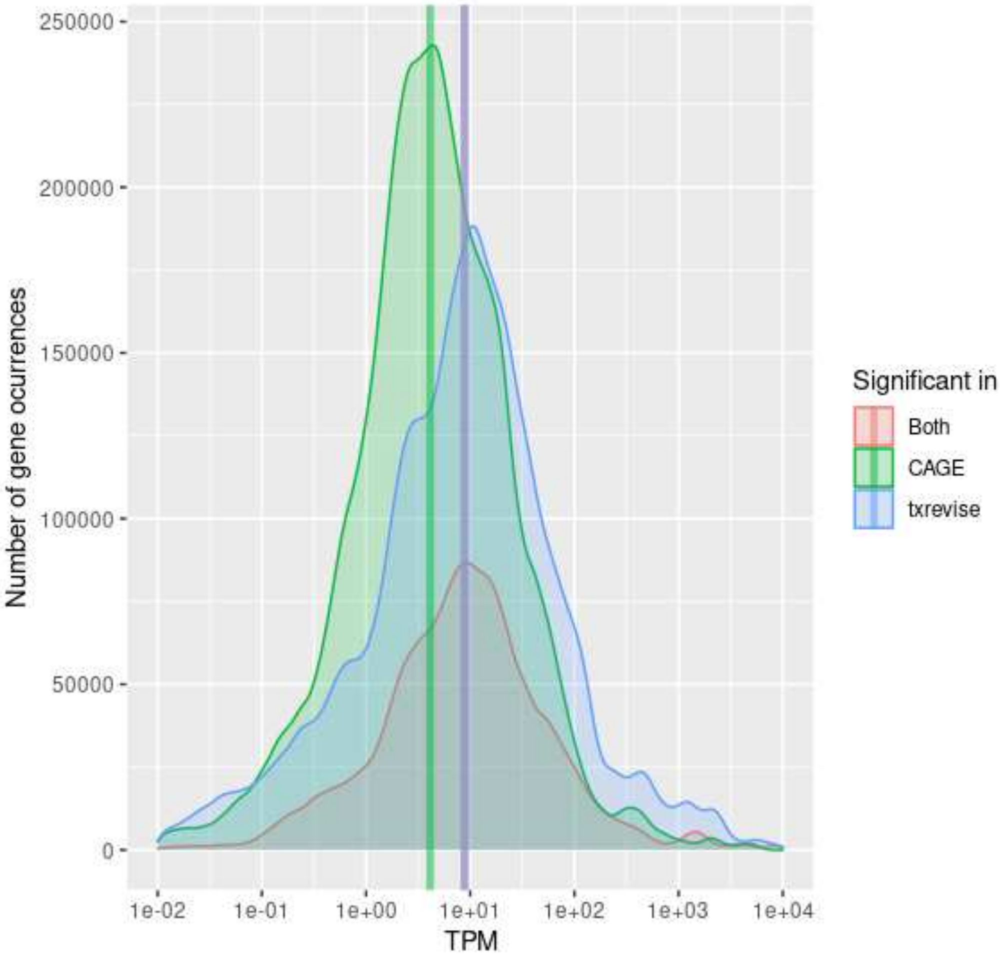
All FANTOM5 promoters meeting the following criteria for different values of N were chosen:

- The promoter is not within N bp of the start of the exon of any existing txrevise transcript (taking into account the strand of the transcript)
- The promoter is not within N bp of any already added promoter
- The promoter overlaps with an exon of an existing txrevise transcript or is within 1000 bp upstream of one (94.4% of all promoters corresponding to a gene possessing a txrevise transcript match this criterion)

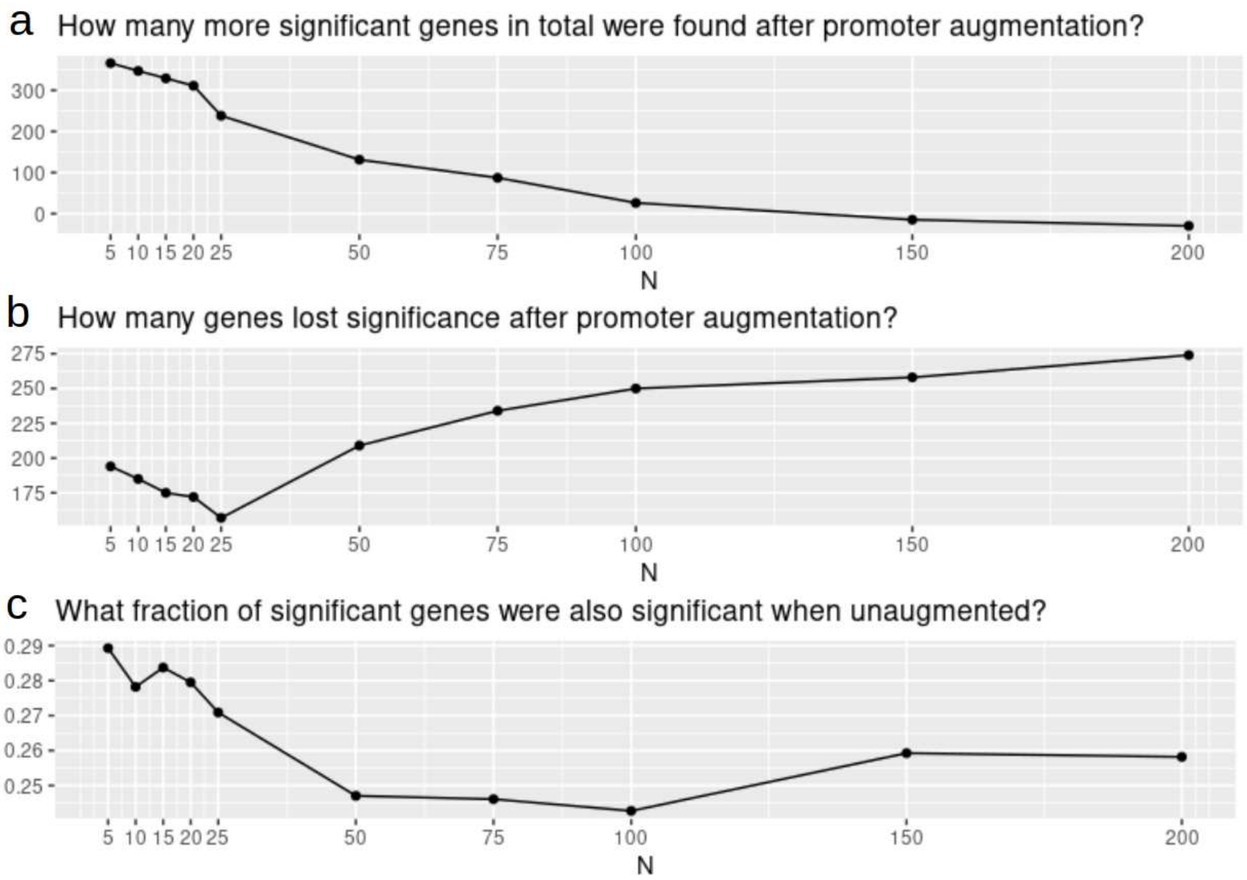
The chosen promoters and the txrevise transcripts whose exon the promoters overlapped or were near upstream to were used to construct a new set of artificial transcript annotations. The first exon of these new transcripts was an artificial exon from the promoter's start coordinate to the end coordinate of the nearest (overlapping) first exon and the remaining exons were all the remaining exons of the existing txrevise transcript (Figure 3).

The created transcripts were added to txrevise transcripts, given as input to another run of txrevise and put through the eQTL-Cataloge/qtlmap workflow. Supplementary Figure 2 shows that the smaller the N gets, the more statistically significant puQTL genes were found, especially at values of N of 20 and smaller. For values of N larger than 100, the effect of adding annotations was negative, likely because re-running txrevise with N values larger than 25 causes some original txrevise transcripts to be removed. The biggest fraction of new genes that were also significant with CAGE occurred at N values of 5-25 and the least originally found genes were lost when N was 25. For these reasons, we chose 20 as the optimal N value based on our tests. This resulted in 77,869 new transcripts across 11,877 genes.

Supplementary Figures



Supplementary Figure 1. The distribution of transcripts per million (TPM) values across all significant puQTL genes grouped by whether the gene was found to be significant in CAGE or txrevise. Vertical lines show the median TPM of the corresponding group. The genes detected by CAGE have significantly lower mean TPM values than the genes detected by txrevise (Mann-Whitney U test p-value < 2.2e-16).



Supplementary Figure 2. Effect of different values of N on detecting genes with at least one puQTL. (**a**) The effect of N on the number of additional puQTL genes detected. (**b**) The effect of N on the number of puQTL genes lost. (**c**) The effect of N on the agreement between reference and augmented and promoters.

Data availability

The CAGE sequencing data from Garieri_2017 is available from ArrayExpress (E-MTAB-5835). The genotype data from the GENCORD study is available from EGA (EGAD00001000428). The RNA-seq data from the GEUVADIS study is available from ArrayExpress (E-GEUV-1). The GEUVADIS genotype data was downloaded from the 1000 Genomes Project FTP server (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/). The CAGE and txrevise promoter usage QTL summary statistics have been deposited to Zenodo (https://doi.org/10.5281/zenodo.5831090). The pre-computed txrevise annotations using Ensembl 105 transcriptome annotations and FANTOM5 promoters with N=25 parameter have been deposited to Zenodo (https://doi.org/10.5281/zenodo.6499127).

Code availability

Source code of all the analyses presented in the paper is available from GitHub (https://github.com/andreasvija/cage). The updated version of the txrevise software supporting augmenting annotations with FANTOM5 CAGE promoters is available from GitHub (https://github.com/kauralasoo/txrevise). The txrevise promoter usage quantification was performed with the eQTL-Catalogue/rnaseq workflow and puQTL mapping was performed with the eQTL-Catalogue/qtlmap workflow.

Funding

A.V. and K.A. were supported by the Estonian Research Council (grant no. PSG415). A.V. received salary from STACC OÜ.

Acknowledgements

Many of the analyses were performed at the High Performance Computing Center, University of Tartu.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. *Nature* **526**:68–74.

Abugessaisa I, Noguchi S, Hasegawa A, Harshbarger J, Kondo A, Lizio M, Severin J, Carninci P, Kawaji H, Kasukawa T. 2017. FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. *Sci Data* **4**:170107.

Adiconis X, Haber AL, Simmons SK, Moonshine AL, Ji Z, Busby MA, Shi X, Jacques J, Lancaster MA, Pan JQ, Regev A, Levin JZ. 2018. Comprehensive comparative analysis of 5′-end RNA-sequencing methods. *Nat Methods* **15**:505–511.

Alasoo K, Rodrigues J, Danesh J, Freitag DF, Paul DS, Gaffney DJ. 2019. Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *Elife* **8**. doi:10.7554/eLife.41673

Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. 2017. A complete tool set for molecular QTL discovery and analysis. *Nat Commun* **8**:15452.

Demircioğlu D, Cukuroglu E, Kindermans M, Nandi T, Calabrese C, Fonseca NA, Kahles A, Lehmann K-V, Stegle O, Brazma A, Brooks AN, Rätsch G, Tan P, Göke J. 2019. A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. *Cell* **178**:1465–1477.e17.

FANTOM Consortium and the RIKEN PMI and CLST, Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple CA, Ishizu Y, Young RS, Francescatto M, Alam I, Albanese D, Altschuler GM, Arakawa T, Archer JAC, Arner P, Babina M, Rennie S, Balwierz PJ, Beckhouse AG, Pradhan-Bhatt S, Blake JA, Blumenthal A, Bodega B, Bonetti A, Briggs J, Brombacher F, Burroughs AM, Califano A, Cannistraci CV, Carbajo D, Chen Y, Chierici M, Ciani Y, Clevers HC, Dalla E, Davis CA, Detmar M, Diehl AD, Dohi T, Drabløs F, Edge ASB, Edinger M, Ekwall K, Endoh M, Enomoto H, Fagiolini M, Fairbairn L, Fang H, Farach-Carson MC, Faulkner GJ, Favorov AV, Fisher ME, Frith MC, Fujita R, Fukuda S, Furlanello C, Furino M, Furusawa J-I, Geijtenbeek TB, Gibson AP, Gingeras T, Goldowitz D, Gough J, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M, Hara M, Harbers M, Harshbarger J, Hasegawa A, Hasegawa Y, Hashimoto T, Herlyn M, Hitchens KJ, Ho Sui SJ, Hofmann OM, Hoof I, Hori F, Huminiecki L, Iida K, Ikawa T, Jankovic BR, Jia H, Joshi A, Jurman G, Kaczkowski B, Kai C, Kaida K, Kaiho A, Kajiyama K, Kanamori-Katayama M, Kasianov AS, Kasukawa T, Katayama S, Kato S, Kawaguchi S, Kawamoto H, Kawamura YI, Kawashima T, Kempfle JS, Kenna TJ, Kere J, Khachigian LM, Kitamura T, Klinken SP, Knox AJ, Kojima M, Kojima S, Kondo N, Koseki H, Koyasu S, Krampitz S, Kubosaki A, Kwon AT, Laros JFJ, Lee W, Lennartsson A, Li K, Lilje B, Lipovich L, Mackay-Sim A, Manabe R-I, Mar JC, Marchand B, Mathelier A, Mejhert N, Meynert A, Mizuno Y, de Lima Morais DA, Morikawa H, Morimoto M, Moro K, Motakis E, Motohashi H, Mummery CL, Murata M, Nagao-Sato S, Nakachi Y, Nakahara F, Nakamura T, Nakamura Y, Nakazato K, van Nimwegen E, Ninomiya N, Nishiyori H, Noma S, Noma S, Noazaki T, Ogishima S, Ohkura N, Ohimiya H, Ohno H, Ohshima M, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov DA, Pain A, Passier R, Patrikakis M, Persson H, Piazza S, Prendergast JGD, Rackham OJL, Ramilowski JA, Rashid M, Ravasi T, Rizzu P, Roncador M, Roy S, Rye MB, Saijyo E, Sajantila A, Saka A, Sakaguchi S, Sakai M, Sato H, Savvi S, Saxena A, Schneider C, Schultes EA, Schulze-Tanzil GG, Schwegmann A, Sengstag T, Sheng G, Shimoji H, Shimoni Y, Shin JW, Simon C, Sugiyama D, Sugiyama T, Suzuki M, Suzuki N, Swoboda

RK, ’t Hoen PAC, Tagami M, Takahashi N, Takai J, Tanaka H, Tatsukawa H, Tatum Z, Thompson M, Toyodo H, Toyoda T, Valen E, van de Wetering M, van den Berg LM, Verado R, Vijayan D, Vorontsov IE, Wasserman WW, Watanabe S, Wells CA, Winteringham LN, Wolvetang E, Wood EJ, Yamaguchi Y, Yamamoto M, Yoneda M, Yonekura Y, Yoshida S, Zabierowski SE, Zhang PG, Zhao X, Zucchelli S, Summers KM, Suzuki H, Daub CO, Kawai J, Heutink P, Hide W, Freeman TC, Lenhard B, Bajic VB, Taylor MS, Makeev VJ, Sandelin A, Hume DA, Carninci P, Hayashizaki Y. 2014. A promoter-level mammalian expression atlas. *Nature* **507**:462–470.

Garieri M, Delaneau O, Santoni F, Fish RJ, Mull D, Carninci P, Dermitzakis ET, Antonarakis SE, Fort A. 2017. The effect of genetic variation on promoter usage and enhancer activity. *Nat Commun* **8**:1–9.

Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, Falconnet E, Bielser D, Gagnebin M, Padioleau I, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis SE, Dermitzakis ET. 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**. doi:10.7554/eLife.00523

Ha KCH, Blencowe BJ, Morris Q. 2018. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol* **19**:45.

Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, El Houdaigui B, Fatima R, Gall A, Garcia Giron C, Grego T, Guijarro-Clarke C, Haggerty L, Hemrom A, Hourlier T, Izuogu OG, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Gonzalez Martinez J, Marugán JC, Maurel T, McMahon AC, Mohanan S, Moore B, Muffato M, Oheh DN, Paraschas D, Parker A, Parton A, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Steed E, Szpak M, Szuba M, Taylor K, Thormann A, Threadgold G, Walts B, Winterbottom A, Chakiachvili M, Chaubal A, De Silva N, Flint B, Frankish A, Hunt SE, IIsley GR, Langridge N, Loveland JE, Martin FJ, Mudge JM, Morales J, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Cunningham F, Yates AD, Zerbino DR, Flicek P. 2021. Ensembl 2021. *Nucleic Acids Res* **49**:D884–D891.

Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, Samoviča M, Sakthivel MP, Kuzmin I, Trevanion SJ, Burdett T, Jupp S, Parkinson H, Papatheodorou I, Yates AD, Zerbino DR, Alasoo K. 2021. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet* **53**:1290–1299.

Kubota N, Suyama M. 2022. Mapping of promoter usage QTL using RNA-seq data reveals their contributions to complex traits. *bioRxiv*. doi:10.1101/2022.02.24.481875

Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Geuvadis Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen A-C, van Ommen G-J, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**:506–511.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**:923–930.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**:1754–1760.

Love MI, Hogenesch JB, Irizarry RA. 2016. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat Biotechnol* **34**:1287–1291.

McCaw ZR, Lane JM, Saxena R, Redline S, Lin X. 2019. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*. doi:10.1111/biom.13214

Moore JE, Zhang X-O, Elhajjajy SI, Fan K, Reese F, Mortazavi A, Weng Z. 2021. A catalog of transcription start sites across 115 human tissue and cell types. *bioRxiv*. doi:10.1101/2021.05.12.443890

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**:1479–1485.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**:290–295.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12**:R22.

Shah A, Mittleman BE, Gilad Y, Li YI. 2021. Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome Biol* **22**:1–21.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences* **100**:15776–15781.

The GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**:1318–1330.