METHOD ARTICLE

REVISED False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors [version 2; peer review: 3 approved]

Ashis Saha [1], Alexis Battle [1,2]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, 21218, USA
[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, 21218, USA

v2 **First published:** 28 Nov 2018, **7**:1860
https://doi.org/10.12688/f1000research.17145.1
Latest published: 08 Apr 2019, **7**:1860
https://doi.org/10.12688/f1000research.17145.2

Abstract

Sequence similarity among distinct genomic regions can lead to errors in alignment of short reads from next-generation sequencing. While this is well known, the downstream consequences of misalignment have not been fully characterized. We assessed the potential for incorrect alignment of RNA-sequencing reads to cause false positives in both gene expression quantitative trait locus (eQTL) and co-expression analyses. Trans-eQTLs identified from human RNA-sequencing studies appeared to be particularly affected by this phenomenon, even when only uniquely aligned reads are considered. Over 75% of trans-eQTLs using a standard pipeline occurred between regions of sequence similarity and therefore could be due to alignment errors. Further, associations due to mapping errors are likely to misleadingly replicate between studies. To help address this problem, we quantified the potential for "cross-mapping" to occur between every pair of annotated genes in the human genome. Such cross-mapping data can be used to filter or flag potential false positives in both trans-eQTL and co-expression analyses. Such filtering substantially alters the detection of significant associations and can have an impact on the assessment of false discovery rate, functional enrichment, and replication for RNA-sequencing association studies.

Keywords

Mappability, Cross-mappability, Co-expression, Trans-eQTL, RNA-sequencing, Alignment

Open Peer Review

Reviewer Status ✓ ✓ ✓

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 08 Apr 2019	✓ report		
version 1 28 Nov 2018	✓ report	✓ report	✓ report

1 **Michael I. Love**, The University of North Carolina at Chapel Hill, Chapel Hill, USA

2 **Rob Patro**, Stony Brook University, Stony Brook, USA

3 **Aaron R. Quinlan**, University of Utah, Salt Lake City, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Ashis Saha (ashis@jhu.edu), Alexis Battle (ajbattle@jhu.edu)

Author roles: Saha A: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Battle A**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: A.B. is supported by NIH grant 1R01MH109905, NIH grant R01HG008150 (NHGRI; Non-Coding Variants Program), and NIH grant R01MH101814 (NIH Common Fund; GTEx Program).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Saha A and Battle A. **False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors [version 2; peer review: 3 approved]** F1000Research 2019, **7**:1860 https://doi.org/10.12688/f1000research.17145.2

First published: 28 Nov 2018, **7**:1860 https://doi.org/10.12688/f1000research.17145.1

REVISED **Amendments from Version 1**

We updated the manuscript to address several important points raised by the reviewers. The changes are as follows:

- We added a new section ("Impact of alternative quantification and parameter settings") to discuss the impact of expectation maximization (EM) based quantification methods, and various parameter choices of our methods. Two supplementary figures (Supplementary Figure 10 and Supplementary Figure 11) accompanied the discussion.
- We included a short description of the gene expression quantification pipeline of GTEx v7 and DGN.
- We also briefly discussed the gene model to generate k-mers, and the difference between mappability and cross-mappability.
- Supplementary Figure 4C was added to show the composition of different types of pseudogenes in trans-eQTLs.
- We uploaded several versions of cross-mappability resources computed with different parameters.

See referee reports

Introduction

Sequence similarity among distinct genomic regions makes alignment of short sequencing reads difficult[1,2]. Genomes, including the human genome, contain diverse classes of elements with sequence similarity across regions, ranging from large segmental duplications to pseudogenes to transposable elements. Alignment-based quantification of genomic phenotypes such as gene expression or epigenetic signal is less reliable for such regions[3–6].

Despite attention to the importance of alignment errors, the full range of consequences is not always considered in downstream analyses. Here, we focus on evidence that sequence similarity between pairs of genes and resulting alignment errors between them may lead to false positives in association studies from RNA-sequencing (RNA-seq) data, specifically in expression quantitative trait locus (eQTL) and co-expression analyses. eQTL studies, revealing associations between genetic variants and gene expression levels, have contributed to a greater understanding of gene regulation and genetics of complex traits[7–9]. Trans-eQTLs, where the genetic variant is distant or on a different chromosome from the associated gene, are of particular interest, but have proven challenging to identify in human data due to power, confounders, small effect sizes, and other challenges[10,11]. Given that a trans-eQTL analysis performs genome-wide tests, it is more prone to be affected by systematic errors between genomic regions than a cis-eQTL analysis where only variants close to the target gene are considered. Here, we discuss the impact of alignment errors on RNA-seq association studies. Figure 1A illustrates a cartoon example, where all reads truly originate from transcripts of Gene A, but due to sequence similarity between Gene A and Gene B, some of the reads incorrectly map to Gene B, causing it to erroneously appear to be expressed in the sample. The number of reads misaligned to Gene B across samples may be directly proportional to the number of reads for Gene A, or may be determined by genetic variation creating sequence mismatches with the correct region. In either case, spurious associations can then arise. In Figure 1A, the two genes incorrectly appear to be co-expressed. In addition, a variant associated with expression of Gene A may also appear to be associated with Gene B, giving rise of a false positive trans-eQTL. We note that such errors are not entirely mitigated by filtering multi-mapped reads—some alignment errors may remain between similar regions even among uniquely aligned reads due to genetic variation, errors in the reference genome, and other complications.

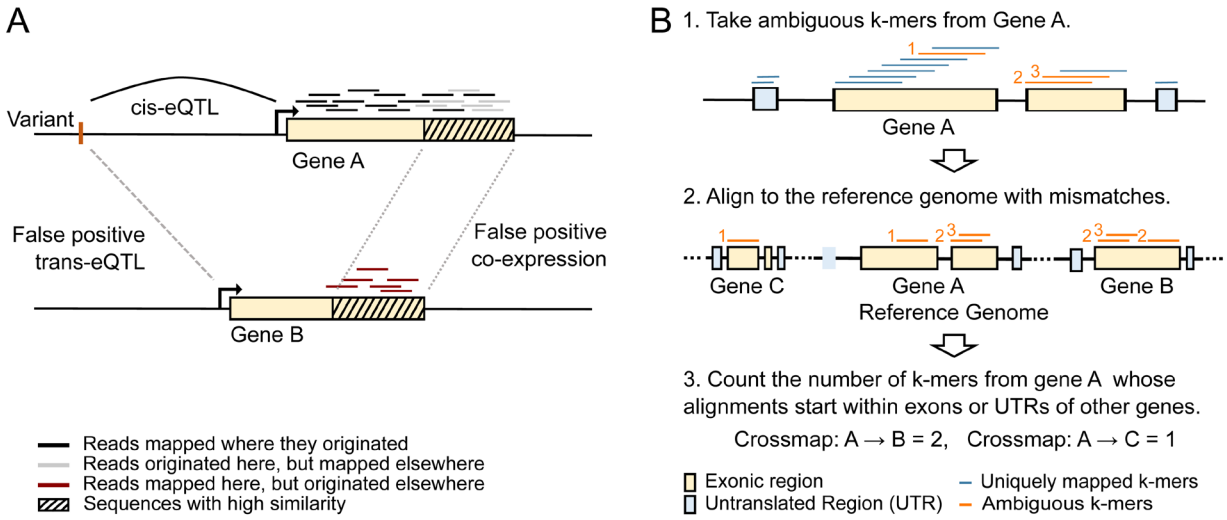

Figure 1. Overview of cross-mappability. A) Some of the reads generated from Gene A are incorrectly mapped to Gene B because of sequence similarity between the genes, leading to false positive co-expression. Consequently, a variant which is a true cis-eQTL of Gene A appears as a false positive trans-eQTL of Gene B. **B**) We align the ambiguous (orange) *k*-mers (75-mers from exons and 36-mers from UTRs) from Gene A to the reference genome using Bowtie and count how many *k*-mers from Gene A map to each other gene to compute cross-mappability. Here, the number beside each ambiguous (orange) *k*-mer represents the identifier for the ambiguous *k*-mer based on its position in Gene A.

Previous studies have shown that uniqueness of sequence in genomic regions should be considered in an analysis of sequencing data[4,5,12]. Karimzadeh *et al.* showed that a differential methylation analysis can identify false signals due to poor mappability[5]. We have previously filtered trans-eQTLs based on sequence similarity as part of the Genotype-Tissue Expression (GTEx) project[10] and the Depression Genes and Networks (DGN) study[13]. Pickrell *et al.*[14] also suggested that the most significant distant eQTL in their RNA-seq study was likely an artifact arising due to sequencing reads originating from a gene near the SNP mapping to another distant gene. Related effects were also discussed in greater depth for microarrays, where probes intended for one gene may cross-hybridize to other genes[11,15]. In microarray studies, one could identify and replace probes displaying poor specificity, but in RNA-seq, any region of sequence similarity between genes can cause alignment errors. Previous studies have not presented a systematic analysis of alignment-related false positives in RNA-seq association testing.

Here, we report the prevalence of potential false positives in trans-eQTL and co-expression analyses arising from alignment errors. We present a method to assess the potential for mapping error between pairs of genes, which can then be used to filter or flag associations that could arise from these errors. We introduce a new metric, "cross-mappability", representing the extent to which reads from one gene may be mapped to another gene. Using gene expression data from GTEx[10] and DGN[13], we demonstrate the impact of misalignment on both trans-eQTL detection and co-expression analysis in real data. Notably, we show that over 75% of trans-eQTLs detected in any GTEx tissue using a naive pipeline are potential false positives, emphasizing that it is critical to consider these errors. To support future studies, we have published codes in Github[16] and also made cross-mappability resources publicly available for the human genome (hg19 and GRCh38)[17].

Methods

Mappability and cross-mappability

We developed a new metric, cross-mappability, to quantify the potential for incorrect read alignment where reads originating from one gene may incorrectly map to another gene. Based on annotated transcripts for each gene, we evaluated *k*-mers from exonic and untranslated regions (UTRs) of the reference genome that serve as a proxy for reads in an RNA-seq experiment. We defined cross-mappability from Gene A to Gene B, crossmap(A, B), as the number of Gene A's *k*-mers whose alignment, allowing mismatches, start within exonic or untranslated regions of Gene B. Notably, existing *mappability* scores[4,5] correspond to a single region (or gene) describing uniqueness of the sequence of the region in the genome, our *cross-mappability* score corresponds to a pair of genes describing similarity between the sequences of the genes.

Though cross-mappability is a straightforward metric, its computation is non-trivial due to the size of the genome. We followed a systematic approach to compute genome-wide cross-mappabilities in practice. Following Derrien *et al.*[4], we define mappability of a *k*-mer as $\frac{1}{C_k}$, where C_k is the number of positions where the *k*-mer maps to the genome with a tolerance of up to 2 mismatches. We computed exon- and UTR-mappability of a gene as the average mappability of all *k*-mers in exonic regions and untranslated regions, respectively. We used a collapsed gene model to generate *k*-mers where overlapped exons and overlapped UTRs were merged to form exonic and UTR regions, respectively. Then, mappability of a gene is computed as the weighted average of its exon- and UTR-mappability, weights being proportional to the total length of exonic regions and UTRs, respectively. Importantly, we only have to compute cross-mappability from genes with mappability < 1, as no *k*-mer from a gene with mappability = 1 will map to other regions of the genome (i.e. these will all result in cross-mappability of 0). Moreover, we need to consider only *k*-mers with mappability < 1 from a gene, as uniquely mapped *k*-mers will not map to other genes. So, we align all such *k*-mers from exonic and untranslated regions of each gene to the reference genome using Bowtie v1.2.2[18], tolerating up to 2 mismatches, and then count the number of *k*-mers whose alignment start within exonic or untranslated regions of every other gene to compute cross-mappability with each gene genome-wide (Figure 1B).

The length *k* may be tuned to match particular read length or alignment method. Here, if the value of *k* is not mentioned for *k*-mers, the default value of *k* is 75 for exons and 36 for UTRs. We used a smaller *k* for UTRs than for exons because UTRs are generally shorter than exons. Mappability of a gene and cross-mappability to/from a gene is undetermined if all the exons of the gene are shorter than 75 bp and all the UTRs are shorter than 36 bp.

We computed genome-wide mappability and cross-mappability for human genome hg19 using annotations from Gencode v19[19]. 26,200 (out of 57,820) genes had at least one *k*-mer cross-mapping to/from another gene. There were 31,167,448 gene pairs (0.93%) that were cross-mappable (cross-mappability > 0). Supplementary Figure 1A shows the cross-mappability distribution. We found that 2.45–4.92% of gene pairs expressed and quantified in five tissues of the GTEx v7 data were cross-mappable (Supplementary Figure 1B). We also computed the same set of resources for human genome GRCh38 using annotations from Gencode v26, all of which are publicly available[17].

Data

We downloaded fully processed, filtered and normalized gene expression data used in GTEx eQTL analysis from the GTEx portal (www.gtexportal.org). For this study, we focused on gene expression data from 5 tissues: whole blood, skeletal muscle, thyroid, sun-exposed skin, and testis. We also obtained covariates including 3 genotype PCs representing ancestry, sex, genotyping platform, and PEER factors[20] as released in GTEx v7. GTEx aligned 76-bp paired-end reads to the reference genome with STAR v2.4.2a[21], quantified gene expression levels with RNA-SeQC v1.1.8[22] using uniquely mapped reads aligned in proper pairs and fully contained within exon boundaries where each alignment must not contain more than six non-reference bases. We downloaded genotype data from GTEx release v7 from dbGaP (accession number: phs000424.v7.p2).

We also collected genotype, processed RNA-seq, and covariate data for the DGN cohort, which is available through the National Institute of Mental Health (NIMH) Center for Collaborative Genomic Studies on Mental Disorders. DGN aligned the reads to the reference genome using TopHat[23] and quantified gene expression levels using HTSeq[24]. Latent factors inferred from the expression data have already been regressed out of the processed DGN data to address hidden confounders, as described in 13. Gene symbols were mapped to Ensembl gene ids using Gencode v19.

We downloaded the list of trans-eQTLs in 33 cancer types detected by PancanQTL[25] from http://bioinfo.life.hust.edu.cn/PancanQTL. For consistency with our study, we used trans-eQTLs where the variant and the gene were on different chromosomes, and the gene symbols were mapped to unique Ensembl gene ids according to Gencode v19.

Trans-eQTL detection

For trans-eQTL analysis, we selected autosomal variants with MAF ≥ 0.05 that did not fall in a repeat region as annotated by the UCSC RepeatMasker track[26]. We tested trans-eQTL association for each inter-chromosomal variant-gene pair using Matrix-eQTL's linear model test[27]. For GTEx, three genotype PCs, genotyping platform, sex, and PEER covariates estimated by GTEx were used as covariates in Matrix-eQTL. We computed the false discovery rate using the Benjamini-Hochberg method within each tissue. The covariates used for trans-eQTL replication in DGN were three genotype PCs, sex and age, as the expression data already had latent factors regressed out.

Co-expression analysis

We quantified co-expression of a pair of genes as the absolute Pearson correlation ($|r|$) between expression levels of the genes across all available samples. For GTEx, we regressed out all covariates including PEER factors before co-expression analysis. For DGN, we used the corrected data which also regresses out latent factors.

Results

Effect of cross-mappability on trans-eQTL detection

To investigate the effects of alignment errors on trans-eQTL detection, we performed a standard trans-eQTL analysis using data from the GTEx project for five human tissues. For this study, we categorized an eQTL as "cis" if the variant is within 1Mb of the transcription start site (TSS) of the gene, and "trans" if they are on different chromosomes, approximating the regions where cis and trans mechanisms are likely to occur. We call a trans-eQTL "cross-mappable" if any gene within 1Mb of the identified trans-eQTL variant cross-maps to the trans-eQTL target gene. The cross-mappable trans-eQTLs represent suspicious hits that could potentially arise simply due to alignment errors, although cross-mappability does not definitively establish that any individual trans-eQTL is a false positive.

We identified 19,348 unique trans-eQTLs (variant-gene pairs) at FDR ≤ 0.05 from five tissues corresponding to 14,785 unique SNPs and 1,419 unique genes. Notably, a large majority (75.14%) of these statistically significant trans-eQTLs were cross-mappable.Furthermore, the cross-mappable eQTLs tended to be the most highly significant (ordered by increasing p-value, Figure 2A). In GTEx tissues, 90.8–97.3% of top 1000 trans-eQTLs were cross-mappable, compared to a background rate of 19.1–25.6% (based on all tested variant-gene pairs). The fraction of cross-mappable trans-eQTLs is very high even when we restrict our analysis to protein-coding genes or to genes with mappability ≥ 0.8 (Supplementary Figure 2A–C).

We observed a similar pattern in the trans-eQTLs reported from RNA-seq data of 33 cancer types[25] (Supplementary Figure 2D). We also observed that randomly selected variant-gene pairs susceptible to cross-mapping yield more trans-eQTLs than randomly selected pairs with no cross-mapping potential (Supplementary Figure 3). Overall, the high fraction of cross-mappable eQTLs among the top associations in multiple tissues and multiple datasets indicates that alignment errors could be a major source of artifacts, dominating legitimate trans-eQTLs. It is also important to note that filtering such prevalent potential false-positives necessitates re-assessing FDR. For example, while 4,809 trans-eQTLs with no evidence of cross-mapping (corresponding to 969 unique genes) were among the 19,348 hits from the original scan of GTEx, only 2,456 (corresponding to 228 unique genes) would appear significant if FDR were reassessed after filtering cross-mapping hits.

When we further analyzed the composition of the 19,348 significant naive trans-eQTLs, we observed a majority (>70%) of cross-mappable eQTLs corresponded to pseudogene targets. The non-cross-mappable eQTLs contained far fewer pseudogene targets (30%, Supplementary Figure 4). Likewise, we observed that more than 85% of eQTLs corresponding to pseudogenes were cross-mappable. Due to sequence similarity between pseudogenes and their corresponding parent genes, this is not surprising and could be due to alignment errors. One simple preventative measure in trans-eQTL studies would be to simply exclude pseudogenes entirely. However, 42.4% of eQTLs corresponding to protein-coding genes were also cross-mappable, which still exceeded expectation, and the top hits remained enriched for cross-mapping errors as noted above.

We investigated one GTEx trans-eQTL in greater detail for illustration – variant: chr5:149826526 and gene: RP11-343H5.4 (ENSG00000224114) – which was significant in each of 5 GTEx tissues. RP11-343H5.4 is a pseudogene on chromosome 1. In the coverage plots of the gene, we noticed that reads were aligned to only a fraction of the exonic region of the gene; if the gene were truly expressed, we would expect reads being mapped across the whole exon (Figure 2B). RP11-343H5.4 is cross-mappable with RPS14 (ENSG00000164587), a protein-coding gene in chromosome 5 near the putative trans-eQTL variant. There was also a cis-association between the variant and RPS14. *k*-mers from RPS14 indeed map to the region within RP11-343H5.4, where we observed a non-zero number of reads. Interestingly, in this case, read mapping appears to be genotype-dependent - the variant at chr5:149826526 alters sequence such that it would lead to reads from RPS14 uniquely, but likely incorrectly, mapping to RP11-343H5.4.

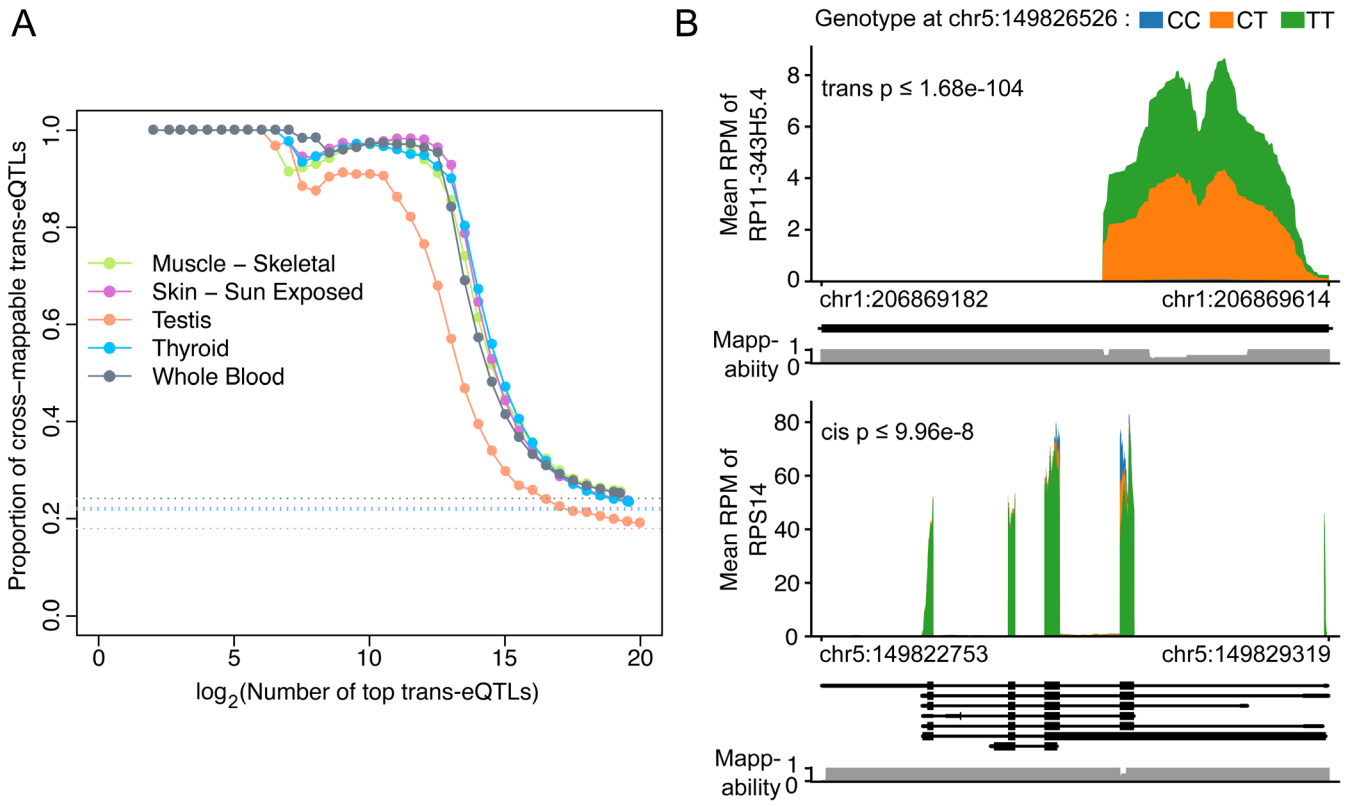

Figure 2. Effect of cross-mappability on trans-eQTLs in GTEx. A) Fraction of cross-mappable trans-eQTLs among the top significant variant-gene pairs (ordered by increasing FDR) in each tissue (color). Each dotted horizontal line represents the background cross-mappable rate in a given tissue. **B**) An example of likely false positive trans association between the variant chr5:149826526 and the gene RP11-343H5.4. The coverages (reads per million, RPM) of the trans-eGene RP11-343H5.4 (top) and its cross-mapping gene RPS14 (bottom) in Thyroid are shown along with their exons and UTRs (black lines below the coverage plot), and mappability of 75-mers. The regions of mappability less than 1.0 have sequence similar between the two genes.

Finally, we found that cross-mappable eQTLs, which we believe to be enriched for false-positives, are highly replicable between datasets. This misleading replication occurs because it is driven by the underlying sequence of the genome, and similar alignment errors frequently occur regardless of tissue and study. We showed this by measuring the replication between the significant trans-eQTLs detected at FDR ≤ 0.05 from whole blood from GTEx and whole blood data from the DGN study[13]. To avoid the effects of linkage disequilibrium, we tested for trans-association in DGN only for the best variant per GTEx trans-eQTL gene (with the lowest p-value in GTEx), where both the variant and the gene were present in the DGN data. At FDR ≤ 0.05, only 10.71% (3 out of 28) non-cross-mappable trans-eQTLs were replicated in DGN while 31.25% (5 out of 16) cross-mappable trans-eQTLs were replicated. The Q-Q plot in Figure 3A shows that cross-mappable trans-eQTLs were more likely to be replicated compared to non-cross-mappable ones. We observed the same phenomenon when we attempted to replicate significant trans-eQTLs detected from one GTEx tissue in other GTEx tissues. On average, 63.0% (range: 50.3–70.2%) and 16.3% (range: 7.6–25.1%) of cross-mappable and non-cross-mappable trans-eQTLs, respectively, were replicated (Figure 3B). This suggests that replication of a trans-eQTL does not necessarily indicate a true positive. Overall, we suggest that regardless of replication, cross-mappable trans-eQTLs require further investigation to establish that they arise from biological regulation rather than alignment artifacts.

Effect of cross-mappability in co-expression analysis

Next, we evaluated evidence that alignment errors between genes can cause spurious correlation between gene expression levels (co-expression). If alignment errors did not affect co-expression analysis, we would expect that the distribution of pairwise correlation between cross-mappable genes would not deviate from that between non-cross-mappable genes. To test this, we used the gene expression data in five tissues from GTEx v7 after correction for covariates and latent confounders (see Methods). For each tissue, we selected a random set of 10,000 non-cross-mappable gene pairs and a random set of 10,000 cross-mappable gene pairs chosen with probability proportional to their cross-mappabilities (sampling probability proportional to cross-mappability ensures sampling from the whole cross-mappability range, as opposed to just from the massive number of low cross-mappability pairs). Then we computed the absolute Pearson correlation ($|r|$) between expression levels of the genes in each randomly selected pair. We found that expression levels of cross-mappable genes were more correlated than expression levels of non-cross-mappable genes (median p across tissues ≤ 4.7×10^{-5}, Wilcoxon rank-sum test, Figure 4A). The difference was more significant when uncorrected data were used (median $p \leq 1.3 \times 10^{-74}$,

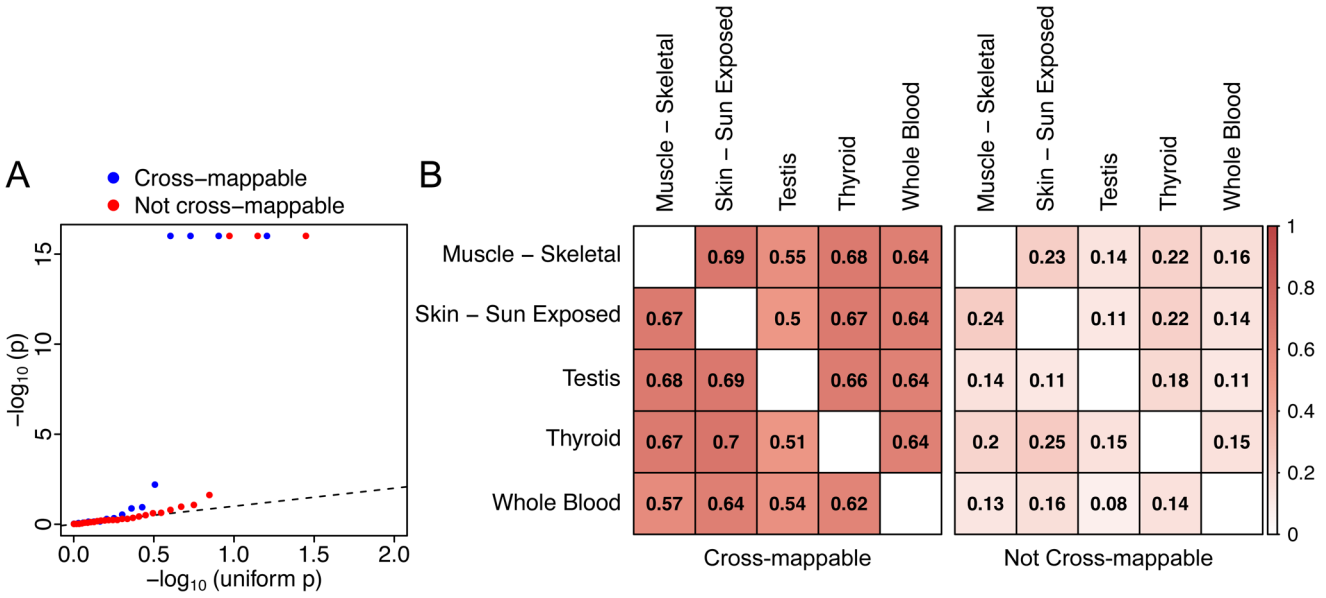

Figure 3. Trans-eQTL replication. (**A**) Q-Q plot, replication p-values from DGN for variant-gene pairs discovered in GTEx Whole Blood, grouped by cross-mappability. (**B**) The fraction of significant eQTLs in each GTEx tissue (row) replicated in another tissue (column) at FDR ≤ 0.05, for cross-mappable eQTLs (left) and not cross-mappable eQTLs (right).

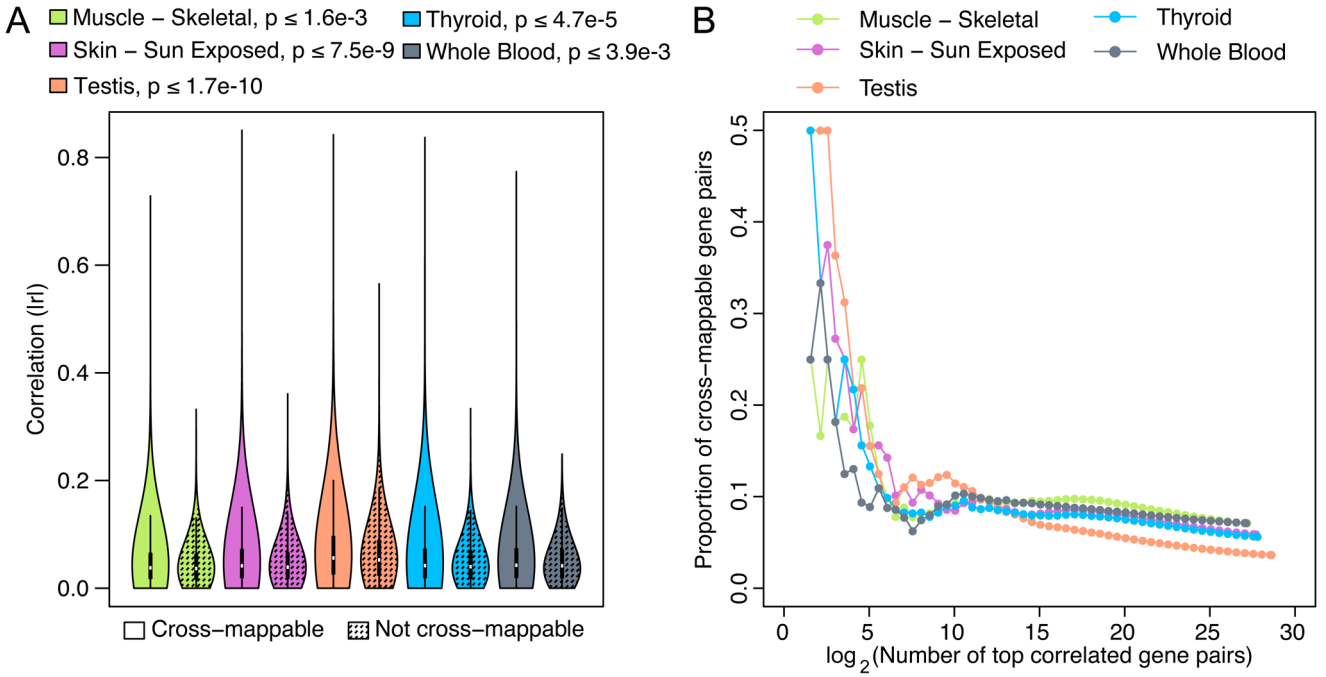

Figure 4. Effect of cross-mappability on co-expression. (**A**) Comparison of co-expression between randomly drawn pairs of cross-mappable genes and not cross-mappable genes. Each violin plot shows the distribution of the absolute Pearson correlation (y-axis) between corrected gene expression levels of randomly drawn 10,000 gene pairs in a tissue (color). P-value of the Wilcoxon test to determine whether cross-mappable genes are more correlated than not cross-mappable genes in each tissue is shown in the legend. (**B**) Fraction of top co-expressed genes that are cross-mappable and thus potential false positives.

Supplementary Figure 5). We also observed that the correlation coefficient tends to increase with increasing levels of cross-mappability between genes (Supplementary Figure 6), indicating a high rate of false co-expression in the most highly cross-mappable genes. The increased correlation between cross-mappable genes was observed even after discounting genes from same gene family (Supplementary Figure 7), somewhat alleviating concerns that our observations were due to exclusively true functional relationships. We observed a similar pattern using data from an independent RNA-seq study, DGN (Supplementary Figure 8).

To demonstrate the impact of this pattern on a realistic genome-wide co-expression analysis, we evaluated how many of the top-most correlated gene pairs in each GTEx tissue suffer from cross-mappability. We observed that cross-mappable pairs of genes are over-represented among the top hits, with gene pairs ordered by the absolute Pearson correlation after excluding pairs of genes whose genomic coordinates actually overlap (Figure 4B, Supplementary Figure 9). Overall, the impact of cross-mappability on co-expression appears to be less than on trans-eQTL analysis, but the phenomenon may still require consideration when examining specific co-expressed gene pairs or enrichment patterns.

Impact of alternative quantification and parameter settings

We have made several versions of our cross-mappability resources publicly available for the human genome (hg19 and GRCh38)[17], and also published code in Github[16]. Researchers should carefully choose settings according to the study design and goals. Genome version and gene annotations can be directly matched, but other parameter choices such as k and the maximum number of mismatches allowed in alignment may affect the detection of false positives. Small values of k will produce more conservative cross-mappability scores, but large k may not correctly handle small exons or UTRs. For example, if 75-mers (instead of 36-mers) were used from UTRs, a smaller proportion of trans-eQTLs (67.2% instead of 75.14%) would appear as cross-mappable in GTEx, although cross-mappable trans-eQTLs would still tend to be most highly significant (Supplementary Figure 10A). Similarly, increasing the number of mismatches allowed in k-mer alignment results in an increased number of cross-mappable trans-eQTLs (Supplementary Figure 10B). For convenience, k and the number of mismatches are configurable in our software so that, if needed, one can compute cross-mappability scores with settings appropriate for a given study.

We also note that utilization of improved alignment and quantification methods to generate gene expression data may also be helpful to avoid false positives. For example, quantification of gene expression levels using RSEM[28], an expectation maximization based quantification tool, results in a smaller fraction of false positive trans-eQTLs (60.17%) than that using RNA-SeQC (75.14%). However, potential false positives due to cross-mappability still remain abundant in both trans-eQTL and co-expression studies (Supplementary Figure 11).

Discussion

Misalignment of short sequencing reads has the potential to induce false positives in association studies. For RNA-seq, both trans-eQTL and co-expression analyses are susceptible to these artifacts, related to false positives in microarray analysis due to probe cross-hybridization. This is readily apparent from the enrichment of processed pseudogenes among the top hits for such association studies, but misalignment can affect protein-coding genes as well. Our results demonstrate that trans-eQTL associations in a standard pipeline are dominated by potential false-positives due to sequence similarity and replication rates between studies may be artificially inflated due to this pattern. Additionally, genes with sequence similarity display more correlated expression levels, and mapping errors should be considered in co-expression analysis as well.

Our results do not imply that all instances of co-expression or trans-eQTL associations arising from genes with sequence similarity are in fact false positives. Genes with sequence similarity also sometimes have true functional relationships. Pseudogene transcripts may interact with coding transcripts, and some associations with pseudogene expression may reflect true regulatory relationships[29]. Furthermore, the background (random) rate of sequence similarity between any two regions in the human genome is above zero; that is, a hit may occur between regions of sequence similarity by chance, even when no actual misalignment of reads has taken place. However, we believe the exceedingly high fraction of cross-mappable regions among trans-eQTLs from a naive analysis warrants suspicion that these hits are predominantly false positives. Researchers should consider their particular application and tolerance for false negatives and false positives when applying filters targeting alignment errors. Other information, such as base-level coverage plots and outside functional information can help disambiguate particular cases of interest.

Extensions, modifications, and other approaches related to this problem should also be considered. First, specifics of study design, and in particular sequencing read length, should be taken into account when using our data to filter potential false positives. If read length is much shorter or longer than our k-mer setting, our existing data may be insufficient and new mappability and cross-mappability estimates should be derived. In the initial resource provided, we used k-mer alignment to the genome, which does not directly handle splice junctions in transcriptomic data (and also limits appropriate k-mer length even for studies with longer reads). Alignment to the transcriptome or splice-aware alignment may offer future improvements, but computational cost and inaccuracies due to incorrect annotation will have to be evaluated. Our observations and methods may be relevant to analyses of other functional genomic data as well, including detection of interactions from HI-C, and detection of associations with data types such as ATAC-seq or ChIP-seq. Other approaches, such as filtering reads themselves before quantification can also be applied if raw reads rather than quantified data are available and tractable[12].

Our evaluation provides evidence that misalignment of reads should be considered as a potential source of false positives in

association studies, particularly for trans-eQTL analysis. The resources we provide can be used directly to filter potential false positives, or the ideas presented may be tuned and adapted to new studies and data types.

Data availability

Underlying data

Pre-computed cross-mappability resources for human genomes (hg19 and GRCh38) are available on figshare, DOI: 10.6084/m9.figshare.c.4297352.v4[17]. GTEx (v7) expression and covariate data are available from www.gtexportal.org. GTEx (v7) genotype data are available from dbGap (accession number: phs000424.v7.p2). DGN data are available by application through NIMH. Other data, including annotations and intermediate results, required to reproduce analyses in the manuscript are available on figshare, DOI: 10.6084/m9.figshare.7309625.v2[30].

Extended data

Supplementary Figure 1–Supplementary Figure 11 are available on figshare, DOI: 10.6084/m9.figshare.7359539.v2[31].

Supplementary Figure 1. Cross-mappability statistics. (A) Distribution of cross-mappability between cross-mappable pairs of genes, restricted to gene pairs with cross-mappability > 0, using Gencode v19 annotations on human genome hg19. (B) Background percentage of cross-mappable gene pairs between all available expressed genes in GTEx data, categorized by tissue. For both panels, directed gene pairs were used; i.e., (Gene A, Gene B) and (Gene B, Gene A) pairs were considered different.

Supplementary Figure 2. Cross-mappability among top trans-eQTLs. Detected (A) using protein-coding genes in GTEx, (B) using genes with mappability ≥ 0.8 in GTEx, (C) using protein-coding genes with mappability ≥ 0.8 in GTEx, and D) by PancanQTL where unique eQTLs were ordered by lowest p-value across all cancer types.

Supplementary Figure 3. Large number of trans-eQTLs among random cross-mappable gene pairs. We tested for trans-eQTLs taking the same number of random variant-gene pairs in 3 different categories: 1) Not cross-mappable, 2) Cross-mappable, and 3) Cross-mappable (Top). In the first category, we randomly selected 1,000 not cross-mappable gene pairs (g1, g2) where g1 and g2 were on different chromosome and there was at least one variant near g1 (within 1Mb of the TSS of g1), then selected the best cis-variant s (with lowest p-value) for g1, and finally tested for trans-association between s and g2. Variant-gene pairs for other two categories were selected in a similar way as the first category except that the gene pairs were cross-mappable (crossmap(g1, g2) > 0) in the second category, and highly cross-mappable (among top 10,000 cross-mappable pairs) in the third category. The above plot shows the number of significant trans-eQTLs (y-axis) detected at a given FDR (x-axis) in each category (line marker) in each tissue (color).

Supplementary Figure 4. Composition of trans-eQTLs. (A) Representation of gene types among trans-eQTL target genes, categorized by cross-mappability. (B) Proportion of cross-mappable eQTLs categorized by gene type. Only the four most frequent gene types in trans-eQTL hits are shown. (C) Among trans-eQTLs with a pseudogene target gene, quantification of different pseudogene sub-types, categorized by cross-mappability. Pseudogene sub-types were identified from the Gencode v26 annotation, as subtypes are not available in Gencode v19. The five most frequent types among trans-eQTL hits are shown.

Supplementary Figure 5. Correlation between random gene pairs using uncorrected data. Each violin plot shows the distribution of absolute Pearson correlation (y-axis) between uncorrected gene expression levels of 10,000 randomly drawn gene pairs in a tissue (color). The p-value of the Wilcoxon test to determine whether cross-mappable genes are more correlated than not cross-mappable genes in each tissue is shown in the legend.

Supplementary Figure 6. Correlation between random gene pairs increases with cross-mappability. Gene pairs available in each tissue were categorized into 22 groups (x-axis) based on quantiles. A quantile group "$q_1 - q_2(n)$" represents gene pairs of (q_1 * 100, q_2 * 100]-th percentile of cross-mappability with a total of n pairs. In order to visualize the impact of the highest range of cross-mappability, the rightmost nine quantile groups were selected in such a way that each contains about a certain number of pairs: (from right) 2,000, 2,000, 2,000, 2,000, 2,000, 5,000, 10,000, 25,000, 50,000. The leftmost quantile group "0" represents gene pairs which are not cross-mappable. From each group, 1,000 gene pairs were randomly selected where the probability of drawing a pair was proportional to its cross-mappability. Each violin plot shows the distribution of absolute Pearson correlation (y-axis) between corrected expressions of the genes in each pair.

Supplementary Figure 7. Increased correlation between cross-mappable genes is not exclusively due to sequence similarity between genes from same gene family. Here, two genes in the same HGNC gene family were artificially excluded from cross-mappable pairs. We computed the absolute Pearson correlation between gene pairs within different groups as described in Figure 4A and Supplementary Figure 6. Note: gene family information was downloaded from www.genenames.org. (A–B) Comparison of co-expression between 10,000 randomly drawn pairs of cross-mappable and not cross-mappable genes in Muscle – Skeletal (A) and Whole Blood (B). (C–D) Random correlation between genes in Muscle – Skeletal (C) and Whole Blood (D).

Supplementary Figure 8. Co-expression analysis using gene expression data from DGN. (A) Comparison of co-expression between 10,000 randomly drawn cross-mappable and non-cross-mappable gene pairs. (B) Random correlation between genes in DGN increases with cross-mappability.

Supplementary Figure 9. Fraction of gene pairs with cross-mappability ≥ 100 among the top co-expressed genes, categorized by GTEx tissues.

Supplementary Figure 10. Effects of varying k-mer length and the number of mismatches allowed. Cross-mappability among the top GTEx trans-eQTLs when (A) 75-mers (instead of 36-mers) from UTRs were used, (B) a maximum of 3 (instead of 2) mismatches were allowed. 67.2% and 76.1% of the significant trans-eQTLs remain cross-mappable in (A) and (B), respectively, compared to 75.14% using 75-mers from exons and 36-mers from UTRs with 2 mismatches in the original analysis. In both cases, cross-mappable trans-eQTLs still tend to be the most highly significant.

Supplementary Figure 11. Effects of EM-based quantification methods. We downloaded transcript-level quantifications based on RSEM[28] from GTEx and derived gene-level TPMs using the tximport package[32] in R. We used same set of genes and samples as available in the regular gene-level quantifications used in our main GTEx eQTL anaylysis. Following the GTEx pipeline, we normalized gene expression values between samples using TMM as implemented in edgeR[33] and then performed an inverse normal transformation of expression values for each gene across all samples. (A) We computed the number of cis-eGenes on two chromosomes (chr7 and chr14) for different numbers of PEER factors, estimated from RSEM-quantifed gene-level data for each tissue. As in the standard GTEx Consortium analysis, the number of cis-eGenes tended to increase with the number of PEER factors. With no clear difference in behavior, we opted to use the same number of PEER factors as used in the standard analysis, along with three genotype PCs, genotyping platform and sex. (B) The plot shows trans-eQTL p-values using RSEM-quantified data (y-axis) vs. standard RNA-SeQC GTEx data (x-axis), for each significant trans-eQTL (point) in whole blood. Here, the color represents whether the eQTL is cross-mappable or not, and the symbol represents whether the target gene is a pseudogene or not. The majority of points lie close to the diagonal line, indicating the two quantification methods give mostly similar results, regardless of gene type and including the majority of cross-mappable hits. A few points along the horizontal line at y=0 shows that a small fraction SNP-gene pairs are no longer significant with RSEM-quantified gene expression and most of them are cross-mappable pseudogenes. Thus RSEM offers some modest improvement, but does not resolve the majority of problematic hits. (C) We computed trans-eQTLs genome-wide using RSEM-quantified data. A total of 27,035 trans-eQTLs were detected at FDR ≤ 0.05, 60.17% of which were cross-mappable compared to 75.14% with RNA-SeQC-quantified data. The plot shows the fraction of cross-mappable trans-eQTLs among the top significant variant-gene pairs (ordered by increasing FDR) in each tissue (color). Again, we observe a modest improvement from RSEM. (D) Fraction of top co-expressed genes that are cross-mappable and thus potential false positives. Cross-mappable gene pairs still appear abundant in most correlated genes of multiple tissues.

Software availability

GitHub repository to compute cross-mappability: https://github.com/battle-lab/crossmap.

Archived code at time of publication: https://doi.org/10.5281/zenodo.2602096[16].

License: GPL-3.

GitHub repository to replicate analyses in the manuscript: https://github.com/battle-lab/crossmap_analysis.

Archived code at time of publication: https://doi.org/10.5281/zenodo.2602170[34].

License: GPL-3.

Grant information

A.B. is supported by NIH grant 1R01MH109905, NIH grant R01HG008150 (NHGRI; Non-Coding Variants Program), and NIH grant R01MH101814 (NIH Common Fund; GTEx Program).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We thank Sara Mostafavi, Xiaowei Zhu, and Barbara Engelhardt for discussions and feedback. We thank Brian Jo for running *k*-mer mappability for hg38. We also thank François Aguet for the generating the coverage plots. We also thank Yuan He and Princy Parsana for reviewing codes. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplements to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The genotype data used for this manuscript were obtained from dbGaP accession number phs000424.v7.p2.

References

1. Kahles A, Behr J, Rätsch G: **MMR: a tool for read multi-mapper resolution.** *Bioinformatics.* 2016; **32**(5): 770–772.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Johnson NR, Yeoh JM, Coruh C, *et al.*: **Improved Placement of Multi-mapping Small RNAs.** *G3 (Bethesda).* 2016; **6**(7): 2103–11.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Robert C, Watson M: **Errors in RNA-Seq quantification affect genes of relevance to human disease.** *Genome Biol.* 2015; **16**(1): 177.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Derrien T, Estellé J, Sola SM, *et al.*: **Fast computation and applications of genome mappability.** *PLoS One.* 2012; **7**(1): e30377.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Karimzadeh M, Ernst C, Kundaje A, *et al.*: **Umap and Bismap: quantifying genome and methylome mappability.** *bioRxiv.* 2017; 095463.
Publisher Full Text

6. Degner JF, Marioni JC, Pai AA, *et al.*: **Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.** *Bioinformatics.* 2009; **25**(24): 3207–12.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Grundberg E, Small KS, Hedman ÅK, *et al.*: **Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins.** *Nat Genet.* 2012; **44**(10): 1084–1089.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Nica AC, Dermitzakis ET: **Expression quantitative trait loci: present and future.** *Philos Trans R Soc Lond B Biol Sci.* 2013; **368**(1620): 20120362.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Albert FW, Kruglyak L: **The role of regulatory variation in complex traits and disease.** *Nat Rev Genet.* 2015; **16**(4): 197–212.
PubMed Abstract | Publisher Full Text

10. The GTEx Consortium: **Genetic effects on gene expression across human tissues.** *Nature.* 2017; **550**(7675): 204–213.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Westra HJ, Peters MJ, Esko T, *et al.*: **Systematic identification of *trans* eQTLs as putative drivers of known disease associations.** *Nat Genet.* 2013; **45**(10): 1238–43.
PubMed Abstract | Publisher Full Text | Free Full Text

12. van de Geijn B, McVicker G, Gilad Y, *et al.*: **WASP: allele-specific software for robust molecular quantitative trait locus discovery.** *Nat Methods.* 2015; **12**(11): 1061–1063.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Battle A, Mostafavi S, Zhu X, *et al.*: **Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.** *Genome Res.* 2014; **24**(1): 14–24.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Pickrell JK, Marioni JC, Pai AA, *et al.*: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature.* 2010; **464**(7289): 768–772.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Reilly C, Raghavan A, Bohjanen P: **Global assessment of cross-hybridization for oligonucleotide arrays.** *J Biomol Tech.* 2006; **17**(2): 163–72.
PubMed Abstract | Free Full Text

16. Saha A, Battle A: **battle-lab/crossmap: Github repository to compute cross-mappability (release 1.2)**. 2019.
http://www.doi.org/10.5281/zenodo.2602096

17. Saha A, Battle A: **Pre-computed cross-mappability resources for human genomes (hg19 and grch38)**. 2019.
http://www.doi.org/10.6084/m9.figshare.c.4297352.v4

18. Langmead B, Trapnell C, Pop M, *et al.*: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol.* 2009; **10**(3): R25.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Harrow J, Frankish A, Gonzalez JM, *et al.*: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res.* 2012; **22**(9): 1760–74.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Stegle O, Parts P, Piipari M, *et al.*: **Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses.** *Nat Protoc.* 2012; **7**(3): 500–7.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text

22. DeLuca DS, Levin JZ, Sivachenko A, *et al.*: **RNA-SeQC: RNA-seq metrics for quality control and process optimization.** *Bioinformatics.* 2012; **28**(11): 1530–2.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics.* 2009; **25**(9): 1105–1111.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics.* 2015; **31**(2): 166–169.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Gong J, Mei S, Liu C, *et al.*: **PancanQTL: systematic identification of *cis*-eQTLs and *trans*-eQTLs in 33 cancer types.** *Nucleic Acids Res.* 2018; **46**(D1): D971–D976.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Casper J, Zweig AS, Villarreal C, *et al.*: **The UCSC Genome Browser database: 2018 update.** *Nucleic Acids Res.* 2017; **46**(D1): D762–D769.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Shabalin AA: **Matrix eQTL: ultra fast eQTL analysis via large matrix operations.** *Bioinformatics.* 2012; **28**(10): 1353–8.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics.* 2011; **12**(1): 323.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Pink RC, Wicks K, Caley DP, *et al.*: **Pseudogenes: pseudo-functional or key regulators in health and disease?** *RNA.* 2011; **17**(5): 792–8.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Saha A, Battle A: **Data required to analyze effects of cross-mappability in trans-eqtl and co-expression studies**. 2018.
http://www.doi.org/10.6084/m9.figshare.7309625.v2

31. Saha A, Battle A: **False positives in trans-eqtl and co-expression analyses arising from rna-sequencing alignment errors (supplementary)**. 2019.
http://www.doi.org/10.6084/m9.figshare.7359539.v2

32. Soneson C, Love MI, Robinson MD: **Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved].** *F1000Res.* 2015; **4**: 1521.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol.* 2010; **11**(3): R25.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Saha A, Battle A: **battle-lab/crossmap_analysis: Github repository to analyze effects of cross-mappability in trans-eqtl and co-expression studies (release 1.4)**. 2019.
http://www.doi.org/10.5281/zenodo.2602170

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 15 July 2019

https://doi.org/10.5256/f1000research.20603.r46860

✔ **Michael I. Love**
Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

In their revision, the authors addressed a comment in my initial report that EM-based quantification methods may, to some degree, alleviate the cross-mapping problem. In their Supplementary Figure 11, they show that using RSEM to estimate expression removes some of the false positive trans eQTL (points on the x-axis), but that the majority of FP trans eQTL persist (points on the diagonal). They have therefore addressed all of my initial comments/concerns. The added distinction between mappability and cross-mappability is also useful.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Developer of methods for estimating gene and transcript expression, and statistical testing of expression across samples.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 27 December 2018

https://doi.org/10.5256/f1000research.18744.r41238

✔

Aaron R. Quinlan
Department of Human Genetics, USTAR (Utah Science Technology and Research) Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA

Most assays that leverage high-throughput DNA sequencing ultimately involve counting molecules that align to genome loci. The observed counts are, in turn, used as proxies for some underlying cellular phenomenon. If the process of sequence alignment is fundamentally biased, so too will be the observations and potential conclusions drawn.

In this manuscript, Saha and Battle explore the important role of sequence similarity between genes ("cross-mapping") on the spurious detection of trans eQTLs and gene co-expression. In particular, they argue that up to 75% of trans e-QTLs detected with standard methods could be spurious owing to systematic mapping (and thus both downstream read-counting and subsequent expression measures) artifacts caused by sequence similarity. This manuscript is well-structured, easy to read, and provides yet another warning that properly accounting for mapping and alignment artifacts is critical genomic research. The approach and analyses are sound overall. However, there are a few area that would benefit from clarification, and some technical aspects that warrant greater detail and perhaps further analysis.

1. Choice of k-mer size. While the choice of k=75 for exons and k=36 for UTRs is certainly reasonable, one wonders whether these choices provide the greatest power to detect and quantify cross-mappability. Were analyses conducted to choose these k-mer sizes empirically? It would also be nice to add methods describing from which gene models the k-mers were derived. Furthermore, 5' UTRs are, on average, only slightly smaller than the typical exon, and 3' UTRs are actually larger on average (https://gist.github.com/arq5x/0d44bae195fc6260984ee01fd253712c). Therefore, it is not clear that k=36 for UTRs versus k=75 for exons is well-justified.

2. When reading the manuscript, one naturally wonders whether the effects revealed by cross-mappability could be detected via existing "mappability" scores from ENCODE and others. It would helpful to have an explicit discussion and/or analysis of how this metric differs.

3. By my reading and interpretation of Figure 1B, k-mers spanning exon/exon junctions were not modeled. This is perhaps critical given the results describing an enrichment of cross-mappable eQTLs corresponding to processed pseudogenes.

4. Echoing the point raised by Rob Patro, it is important to understand how the gene expression qualification was done for GTeX and discuss how these methods impact cross-mappability analysis.

5. Mappability is defined as 1/C_k where C_k is the number of genomic loci to which the k-mer aligns with <= 2 mismatches. Given sequencing errors, polymorphism, and the fact that RNA-seq is actually cDNA-seq, one might be concerned that allowing more mismatches (and/or INDEL errors) would more properly model cross-mappability. While perhaps a minor concern, it would be nice to know how much the scores and resulting cross-mappability table differ with tolerance for larger edit distances.

6. In the results describing an enrichment of spurious trans eQTLs associated with cross-mapping pseudogenes, it is unclear whether the authors are referring to all types of pseudogenes or solely processed pseudogenes. Crossmapping could arise from both "classic" (i.e., duplicated owing to NAHR) and "processed" (i.e., mature RNAs duplicated/inserted into the genome). However, the effect of excluding k-mers that span exon/exon boundaries from the cross-mappability would have

a differential effect on these two types of pseudogenes. It would be helpful to be explicit in which types of pseudogenes are being discussed.

7. SNPs were excluded from evaluation if they overlapped annotations in the RepeatMasker track. I would argue that SNPs should (and typically would be) excluded if they lie within segmental duplications. Such regions also harbor genes with a high degree of paralogy (because of the ancestral duplication) and if excluded, may have a substantial impact on the results. For example, I suspect many "classic" pseudogenes would also be excluded were SNPs in segmental duplications excluded. While I have never conducted a trans eQTL analysis, I suspect that removal of segdups is standard practice owing to the inherently high paralogy between duplicated regions.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Human genomics, computational genomics, structural variation, genome evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Apr 2019

Ashis Saha, Johns Hopkins University, Baltimore, USA

Thank you for reviewing our paper. Please see our responses inline in bold and italic font following your comments in standard font.

Most assays that leverage high-throughput DNA sequencing ultimately involve counting molecules that align to genome loci. The observed counts are, in turn, used as proxies for some underlying cellular phenomenon. If the process of sequence alignment is fundamentally biased, so too will be the observations and potential conclusions drawn.

In this manuscript, Saha and Battle explore the important role of sequence similarity between

genes ("cross-mapping") on the spurious detection of trans eQTLs and gene co-expression. In particular, they argue that up to 75% of trans e-QTLs detected with standard methods could be spurious owing to systematic mapping (and thus both downstream read-counting and subsequent expression measures) artifacts caused by sequence similarity. This manuscript is well-structured, easy to read, and provides yet another warning that properly accounting for mapping and alignment artifacts is critical genomic research. The approach and analyses are sound overall. However, there are a few area that would benefit from clarification, and some technical aspects that warrant greater detail and perhaps further analysis.

>> Thank you for your positive assessment of our approach.

1. Choice of k-mer size. While the choice of k=75 for exons and k=36 for UTRs is certainly reasonable, one wonders whether these choices provide the greatest power to detect and quantify cross-mappability. Were analyses conducted to choose these k-mer sizes empirically? It would also be nice to add methods describing from which gene models the k-mers were derived. Furthermore, 5' UTRs are, on average, only slightly smaller than the typical exon, and 3' UTRs are actually larger on average (https://gist.github.com/arq5x/0d44bae195fc6260984ee01fd253712c). Therefore, it is not clear that k=36 for UTRs versus k=75 for exons is well-justified.

>> We appreciate this consideration. To alleviate such concerns regarding the value of k, we have provided options to select appropriate k for exons or UTRs in our code for users that would prefer a different setting. Our choice of k=75 was originally guided by the RNA-seq read-length of the GTEx data, and a reduced length for UTRs based on manual investigation of our top hits suggesting that k=75 for UTRs left too many questionable instances unfiltered. However, because there is no gold standard for which trans-eQTLs are truly false positives, we did not think of a way to tune this more closely.

To provide further insight, we have also now computed cross-mappability using 75-mers from both exons and UTRs and shared these resources publicly. We also shared resources with 50-mers from exons and 36-mers from UTRs. Using 75-mers from both exons and UTRs, 67.2% of unique trans-eQTLs appeared as cross-mappable, compared to 75.1% when 75-mers from exons and 36-mers from UTRs were used. We believe it is likely missing some false-positives at this larger k, but we cannot prove this for certain. Cross-mappable eQTLs still tend to be the most highly significant. We have added discussion of the issue in the manuscript along with Supplementary Figure 10A.

- ***"Researchers should carefully choose settings according to the study design and goals. Genome version and gene annotations can be directly matched, but other parameter choices such as k and the maximum number of mismatches allowed in alignment may affect the detection of false positives. Small values of k will produce more conservative cross-mappability scores, but large k may not correctly handle small exons or UTRs. For example, if 75-mers (instead of 36-mers) were used from UTRs, a smaller proportion of trans-eQTLs (67.2% instead of 75.14%) would appear as cross-mappable in GTEx, although cross-mappable trans-eQTLs would still tend to be most highly significant (Supplementary Figure 10A)"***

We also clarified that we used a collapsed gene model to generate k-mers where overlapped exons and overlapped UTRs were merged to form exonic and UTR regions, respectively. A locus that overlaps both an annotated UTR and an annotated exon will be considered as both.

2. When reading the manuscript, one naturally wonders whether the effects revealed by cross-mappability could be detected via existing "mappability" scores from ENCODE and others. It would helpful to have an explicit discussion and/or analysis of how this metric differs.

>> Thank you for raising the point. Existing "mappability" scores (described by either Derrien et al. or Karimzadeh et al.) correspond to a single genomic region (e.g., gene). In contrast, our "cross-mappability" score corresponds to a pair of genes. While mappability score tells about how unique the sequences of a gene (or a region) is, it does not talk about how similar the sequences of a given pair of genes are. A gene with mappability=1 (mappability = 1 for every k-mer of the gene) will have cross-mappability = 0 from/to any other gene (we directly used this property to optimize our computation of cross-mappability genome-wide). However, cross-mappability between two genes with mappability < 1 cannot be revealed by mappability. We now described the difference between these two metrics in the revised manuscript. Genes with mappability < 1 may indeed cross-map with other genes, or may have sequence similarity with non-transcribed regions of the genome, in which case they would not appear cross-mappable in our analysis.

- ***"Notably, existing mappability scores [4, 5] correspond to a single region (or gene) describing uniqueness of the sequences of the region in the genome, our cross-mappability score corresponds to a pair of genes describing similarity between the sequences of the genes."***

3. By my reading and interpretation of Figure 1B, k-mers spanning exon/exon junctions were not modeled. This is perhaps critical given the results describing an enrichment of cross-mappable eQTLs corresponding to processed pseudogenes.

>> We appreciate your concern. We do think accounting for k-mers spanning exon-exon junctions would offer improvements of cross-mappability, and we plan to investigate this further. Please see our response to Rob Patro for details.

4. Echoing the point raised by Rob Patro, it is important to understand how the gene expression qualification was done for GTeX and discuss how these methods impact cross-mappability analysis.

>> We agree. We performed additional analyses using RSEM-quantified gene expressions and we still observed potential false positives in our analyses. Please see our response to Mike Love for details along with new Supplementary Figure 11.

5. Mappability is defined as 1/C_k where C_k is the number of genomic loci to which the k-mer aligns with <= 2 mismatches. Given sequencing errors, polymorphism, and the fact that RNA-seq is actually cDNA-seq, one might be concerned that allowing more mismatches (and/or INDEL errors) would more properly model cross-mappability. While perhaps a minor concern, it would be nice to know how much the scores and resulting cross-mappability table differ with tolerance for larger edit distances.

>> This is a good point. The number of mismatches is a configurable parameter of our code, and one can compute cross-mappabilities with any desired value. The

cross-mappability scores may increase if the allowed number of mismatches is increased. We have now also added an additional analysis with <=3 mismatches (Bowtie v1 allows max 3 mismatches) along with new Supplementary Figure 10B. Maximum 3 mismatches results in a slightly higher percentage of cross-mappable trans-eQTLs than maximum 2 mismatches (76.1% vs. 75.14%).

- ***"Similarly, increasing the number of mismatches allowed in k-mer alignment results in an increased number of cross-mappable trans-eQTLs (Supplementary Figure 10B)"***

6. In the results describing an enrichment of spurious trans eQTLs associated with cross-mapping pseudogenes, it is unclear whether the authors are referring to all types of pseudogenes or solely processed pseudogenes. Crossmapping could arise from both "classic" (i.e., duplicated owing to NAHR) and "processed" (i.e., mature RNAs duplicated/inserted into the genome). However, the effect of excluding k-mers that span exon/exon boundaries from the cross-mappability would have a differential effect on these two types of pseudogenes. It would be helpful to be explicit in which types of pseudogenes are being discussed.

>> This is a good suggestion. It would be nice to see how different types of pseudogenes contribute to cross-mappability. Gencode v19 annotation does not distinguish among different types of pseudogenes, so we mapped the types of pseudogenes from Gencode v26 annotation, which should be sufficiently accurate for aggregate characterization. Our results show that processed pseudogenes are most common source of cross-mappability, while other pseudogenes do also contribute to cross-mappability. We added an extra plot in Supplementary Figure 4 showing the representation of different types of pseudogenes in trans-eQTLs with pseudogene targets.

7. SNPs were excluded from evaluation if they overlapped annotations in the RepeatMasker track. I would argue that SNPs should (and typically would be) excluded if they lie within segmental duplications. Such regions also harbor genes with a high degree of paralogy (because of the ancestral duplication) and if excluded, may have a substantial impact on the results. For example, I suspect many "classic" pseudogenes would also be excluded were SNPs in segmental duplications excluded. While I have never conducted a trans eQTL analysis, I suspect that removal of segdups is standard practice owing to the inherently high paralogy between duplicated regions.

>> This is a good point. We do mask SNPs based on RepeatMasker, but this only covers a subset of segdups. In general for our false positives, the SNP is actually not within a segdup itself (segdup SNP eQTLs display a slightly higher rate of cross-mappability than background - 83% vs 75%) . The classic pattern is a SNP in the cis-region of one gene appearing associated with a distant region similar to the cis-gene (potentially a segdup in some cases), but the region of similarity does not include the SNP locus. Excluding pseudogenes and/or segdups from candidate genes in a trans-eQTL study is certainly one option. Making SNPs or excluding genes in segdups does not appear to be standard practice to the best of our knowledge, nor does it cover every likely false positive we observe, but is a change worth considering to standard trans-eQTL pipelines.

Competing Interests: No competing interests were disclosed.

Reviewer Report 21 December 2018

https://doi.org/10.5256/f1000research.18744.r41237

Rob Patro
Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

In this manuscript, Saha and Battle describe how errors in the genomic alignment of RNA-seq data can confound specific downstream analyses. Specifically, the focus on the discovery of trans-eQTLs and on co-expression analysis. Surprisingly, they find that, when a "naive" pipeline is used for trans-eQTL discovery, up to 75% of the trans-eQTL events detected may be false positives resulting from cross-mappability (the type of alignment error they discuss and characterize). In addition to describing this phenomenon, and demonstrating its effect on trans-eQTL and co-expression analysis, the authors also propose a new "cross-mappability" score, which allows one to map out which genes in a reference are likely to suffer from the types of spurious alignments, and subsequently, spurious correlations, that are described. The idea of cross-mappability seems a useful and logical extension of the mappability concept, where one is instead interested in which pairs or groups of genes share homologous sequence. The authors also provide pre-computed cross-mappability scores for hg19 and GRCh38.

The paper is well-written, and the issues that the authors raise are important ones. It suggests that researchers should be cautious in interpreting the results of eQTL and co-expression analyses, and, importantly, provides them with tools to reassess their data and help control for the strong potential confounding factor of cross-mappability. I believe this is an important contribution.

My main questions and comments about the manuscript concern cross-mappability scores, and the alignment errors that lead to the observed problems.

- Though the authors only explore the effect of alignment errors on eQTL and co-expression analysis, it seems that these types of issues could affect most analyses involving spliced-mapping of RNA-seq data to the reference genome. Specifically, the type of alignment errors illustrated in Figure 1 (A) would affect even basic expression quantification, let-alone co-expression analysis. This is particularly true for reads where this effect persists even when one considers only reads aligned uniquely by the tool. What would cause the aligner to return only a single (incorrect) locus for the read when multiple equally-good alignments should exist? Are the alignments contiguous at one locus but spliced at the other, or is the manner in which the read would align to the "true" and "spurious" locus different? Interestingly, it seems as though the cross-mappability map could act as a sort of homology table 1 that might even be useful for correcting these spurious alignments, or at least suggesting the true locus as an equally-good match.
- Given that cross mappability is computed by mapping specific k-mers back to the genome (allowing up to 2 mismatches) using Bowtie, how does it deal with accounting for k-mers that span splicing junctions? It seems to me that the specific case where reads map to pseudogenes rather than what is presumed to be the true (protein-coding) locus of the read could be explained by regions of the genome that are contiguous (un-spliced) in the pseudogene, but which span a splicing junction in the protein coding gene. If the cross-mappability score doesn't account for the

cross-mappability of k-mer that may span splice junctions, then it seems it might miss such important cases. However, given that the score is computed assuming some known annotation, it would be possible to explicitly extract appropriately-sized contexts around each known splicing junction, and to include them into the reference that Bowtie maps against when computing the cross-mappability scores. How would the cross-mappability scores change if they also accounted for junction-spanning k-mers rather than just genomically contiguous k-mers?

- Related to the above point, but thinking in the other direction, might the cross-mappability scores be too "conservative" in some cases? Specifically, the scores are computed using k-mers that are the length of relatively short reads for exons, and k-mers that are much shorter than typical reads for UTRs. However, the ambiguity of these sequences individually may not be sufficient to induce a false alignment when the reads being processed are paired-end reads. In that case, one would expect misalignments to require that there be sequence ambiguity supporting both ends of the sequencing read in order for the read to align, concordantly, to the wrong locus.
- The GTEx data that was analyzed relies on spliced alignment to the genome using STAR, and I presume the same tools were used to analyse the DGN cohort data (is this correct?). It would be enlightening to see if and how results would change if the reads were instead aligned using HISAT2. I am aware, for example, that HISAT2 takes special measures to avoid aligning reads to pseudogenes when similar quality alignments to transcripts of other biotypes are available (personal communication with Daehwan Kim). This may have some effect on the type and magnitude of false alignments you see (though, certainly, will not account for all of them). Likewise, it would be very interesting to see how the analyses differ if alignment is done directly to the transcriptome using, e.g., Bowtie2 (the authors already mention this possibility in the discussion section).
- I agree with Mike Love's suggestion that it seems important to explore the extent to which this effect may be mitigated by adopting more accurate methods for gene expression quantification.

References

1. Yorukoglu D, Yu Y, Peng J, Berger B: Compressive mapping for next-generation sequencing. *Nature Biotechnology*. 2016; **34** (4): 374-376 Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I design and develop methods and algorithms for processing high-throughput sequencing data, with some specific focus on methods for gene and transcript quantification and de novo transcriptome analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Apr 2019

Ashis Saha, Johns Hopkins University, Baltimore, USA

Thank you for reviewing our paper. Please see our responses inline in bold and italic font following your comments in standard font.

In this manuscript, Saha and Battle describe how errors in the genomic alignment of RNA-seq data can confound specific downstream analyses. Specifically, the focus on the discovery of trans-eQTLs and on co-expression analysis. Surprisingly, they find that, when a "naive" pipeline is used for trans-eQTL discovery, up to 75% of the trans-eQTL events detected may be false positives resulting from cross-mappability (the type of alignment error they discuss and characterize). In addition to describing this phenomenon, and demonstrating its effect on trans-eQTL and co-expression analysis, the authors also propose a new "cross-mappability" score, which allows one to map out which genes in a reference are likely to suffer from the types of spurious alignments, and subsequently, spurious correlations, that are described. The idea of cross-mappability seems a useful and logical extension of the mappability concept, where one is instead interested in which pairs or groups of genes share homologous sequence. The authors also provide pre-computed cross-mappability scores for hg19 and GRCh38.

The paper is well-written, and the issues that the authors raise are important ones. It suggests that researchers should be cautious in interpreting the results of eQTL and co-expression analyses, and, importantly, provides them with tools to reassess their data and help control for the strong potential confounding factor of cross-mappability. I believe this is an important contribution.

>> Thank you, we are appreciate your comments on our contribution.

My main questions and comments about the manuscript concern cross-mappability scores, and the alignment errors that lead to the observed problems.

● Though the authors only explore the effect of alignment errors on eQTL and co-expression analysis, it seems that these types of issues could affect most analyses involving spliced-mapping of RNA-seq data to the reference genome. Specifically, the type of alignment errors illustrated in Figure 1 (A) would affect even basic expression quantification, let-alone co-expression analysis. This is particularly true for reads where this effect persists even when one considers only reads aligned uniquely by the tool. What would cause the aligner to return only a single (incorrect) locus for the read when multiple equally-good alignments should exist? Are the alignments contiguous at one locus but spliced at the other, or is the manner in which the read would align to the "true" and "spurious" locus different? Interestingly, it seems as though the cross-mappability map could act as a sort of homology table 1 that might even be useful for correcting these spurious alignments, or at least suggesting the true locus as an equally-good match.

>> Thank you for this comment. We agree that alignment errors potentially affect quantification of expression including splicing-aware quantifications. We observed cross-mapping errors even when only uniquely-mapped reads were used in the quantification (for e.g., in GTEx v7). There could be several reasons an aligner returns only a single incorrect locus for the read when multiple equally-good alignments should exist including genetic variation, errors in the reference genome, incorrect annotations, ambiguity due to splicing, and sequencing errors. It is difficult to assess the relative contribution of each of these. We briefly mentioned the issues in the Introduction section of the manuscript.

- ***"some alignment errors may remain between similar regions even among uniquely aligned reads due to genetic variation, errors in the reference genome, and other complications."***

Regarding the pattern of alignment errors in "true" and "spurious" loci, we manually inspected suspected false positives, but did not observe any clear pattern of alignment errors. We observed multiple classes of errors including cases where alignment is contiguous at one locus (either true or spurious) but spliced at the other locus, and cases where alignment is contiguous at both loci.

Finally, it is a nice idea to see if alignment methods could directly utilize our cross-mappability resources or related approaches. We would be interested to follow up on this.

- Given that cross mappability is computed by mapping specific k-mers back to the genome (allowing up to 2 mismatches) using Bowtie, how does it deal with accounting for k-mers that span splicing junctions? It seems to me that the specific case where reads map to pseudogenes rather than what is presumed to be the true (protein-coding) locus of the read could be explained by regions of the genome that are contiguous (un-spliced) in the pseudogene, but which span a splicing junction in the protein coding gene. If the cross-mappability score doesn't account for the cross-mappability of k-mer that may span splice junctions, then it seems it might miss such important cases. However, given that the score is computed assuming some known annotation, it would be possible to explicitly extract appropriately-sized contexts around each known splicing junction, and to include them into the reference that Bowtie maps against when computing the cross-mappability scores. How would the cross-mappability scores change if they also accounted for junction-spanning k-mers rather than just genomically contiguous k-mers?

>> Thanks for raising this important issue. We did not account for k-mers spanning splice junctions and thus our current approach might miss some cross-mapping cases. However, we expect that for many genes with reasonable length exons, some k-mers that fall completely within the exon will also cross-map, and the genes will be identified as cross-mapping by our method in this case, but exceptions to this will also occur. Cross-mappability scores would sometimes increase (but not decrease) if junction-spanning k-mers are considered, and we would need to consider whether these are also sometimes spurious. We mention in the manuscript that alignment to the transcriptome or splice-aware alignment might offer future improvements, but the computational cost and potential for some inaccuracies due to incorrect annotation would also have to be evaluated.

In future work, we plan to extend cross-mappability scores using splice-aware alignments. One challenge here is to find all alignments of a k-mer using a spliced aligner. Spliced aligners are generally optimized to find the best alignment, and it is not as fast to find all alignments. Our current plan is to use STAR aligner (instead of Bowtie in the current setting) with a high number of multiple alignments allowed per read. In addition, we have to modify the k-mer generation process to include exon-exon junction spanning k-mers. Our plan is to evaluate every k-mer spanning exon-exon junctions of the annotated transcripts from a gene, along with the k-mers from the current collapsed gene model. As the k-mer generation process and the alignment process are going to be changed, we can no longer use the GEM Library to compute mappability of each k-mer; we will have to implement this part as well. We leave the implementation of this, and evaluation, as future work, but agree it could offer improvements.

- Related to the above point, but thinking in the other direction, might the cross-mappability scores be too "conservative" in some cases? Specifically, the scores are computed using k-mers that are the length of relatively short reads for exons, and k-mers that are much shorter than typical reads for UTRs. However, the ambiguity of these sequences individually may not be sufficient to induce a false alignment when the reads being processed are paired-end reads. In that case, one would expect misalignments to require that there be sequence ambiguity supporting both ends of the sequencing read in order for the read to align, concordantly, to the wrong locus.

>> This is a good observation. Yes, depending on the goal of the analysis, cross-mappability scores may be conservative in some situations. In a simple case, the value of k we used may be too short for some studies. One can easily recompute cross-mappabilities with appropriate k for the sequencing protocol, using our released code. Also, as you mentioned in the comment, misalignment of paired-end reads would require sequence ambiguity in both ends. Thus, our method may conservatively suggest two genes are cross-mappable, when paired-end reads would mostly or completely resolve the ambiguity. To compute cross-mappability incorporating pair-ended k-mers, we would need to model the distribution of fragment lengths in addition to splice junctions, which we have not yet addressed here. For simplicity and usability of our resources across different studies with different sequencing methods, we treat k-mers like single-ended reads. In general, small cross-mappability scores (where only a few k-mers overlap between genes) may not be sufficient to introduce alignment errors. One can easily filter gene pairs more stringently by thresholding the cross-mappability scores to higher numbers, which may weakly approximate requiring both ends of reads to align. Importantly, we do not claim that all instances of co-expression or trans-associations between cross-mappable pairs are actually false positives, some sequence similarity may occur between genes where a true hit also exists. One should consider the baseline rate of cross-mapping compared with that among observed hits, along with parameters of the study that may necessitate a specialized analysis.

- The GTEx data that was analyzed relies on spliced alignment to the genome using STAR, and I presume the same tools were used to analyse the DGN cohort data (is this correct?). It would be enlightening to see if and how results would change if the reads were instead aligned using HISAT2. I am aware, for example, that HISAT2 takes special measures to avoid aligning reads to pseudogenes when similar quality alignments to transcripts of other biotypes are available (personal communication with Daehwan Kim). This may have some effect on the type and magnitude of false alignments you see (though, certainly, will not account for all of

them). Likewise, it would be very interesting to see how the analyses differ if alignment is done directly to the transcriptome using, e.g., Bowtie2 (the authors already mention this possibility in the discussion section).

>> Good points. Firstly, GTEx and DGN used different alignment and quantification tools. While GTEx v7 used STAR for alignment and RNA-SeQC for quantification, DGN used TopHat for alignment and HTSeq for quantification. False positives are observed for multiple tools. Secondly, we definitely agree avoiding pseudogenes (using --avoid-pseudogene option in HISAT2 or using other tools) would have an effect on the number of false positives. Another option is to simply not map trans-eQTLs for pseudogenes (or other biotypes) at all, which some studies in fact do. As mentioned in the manuscript that 42.4% of eQTLs corresponding to protein-coding genes were also cross-mappable, compared to 75.14% of all eQTLs). This is lower than for pseudogenes, but still above the expected background level, giving us an idea of the impact such a change would have. Of course, most researchers use public datasets without re-aligning them, so our resource shouldn't rely on a specific aligner or parameters. Finally, we agree that alignment to the transcriptome would be an important avenue for future research.

- I agree with Mike Love's suggestion that it seems important to explore the extent to which this effect may be mitigated by adopting more accurate methods for gene expression quantification.

>> We agree. We performed additional analyses using RSEM-quantified gene expressions and we still observed potential false positives in our analyses. Please see our response to Mike Love for details along with Supplementary Figure 11.

Competing Interests: No competing interests were disclosed.

Reviewer Report 12 December 2018

https://doi.org/10.5256/f1000research.18744.r41234

Michael I. Love
Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

The authors present a detailed analysis of how false positives arise in trans-eQTL and co-expression analyses, due to cross-mapping of reads among genes with sequence homology. They present a cross-mappability metric, and provide pre-computed cross-mappability for two human Gencode annotations (v19 and v26). They also provide software for efficiently computing cross-mappability available at a GitHub link. The command line software has detailed instructions online. The software requires genome FASTA files, a GTF file, a mappability bedgraph or bigwig file, Bowtie (v1) and an index, and a few R packages (data.table, intervals, argparser, stats).

The report provides an important warning to the research community, and the pre-computed

cross-mappability and the software will be a valuable resource for groups studying trans-eQTL and co-expression and making use of unique read counts for gene expression quantification.

One of the key points of the article is noting that the cross-mappable and likely spurious trans-eQTLs are highly replicable between datasets, because "*it is driven by the underlying sequence of the genome, and similar alignment errors frequently occur regardless of tissue and study*." This logic also extends to co-expression and gene networks built on gene-gene expression correlations. Another key point was that filtering out of cross-mappable trans-eQTLs necessitates re-assessment of FDR, as the highly significant cross-mappable trans-eQTLs bring down the nominal FDR for other eQTLs.

My main comment on the article is regarding details of the gene expression quantification.

The authors note that,

"The number of reads misaligned to Gene B across samples may be directly proportional to the number of reads for Gene A, or may be determined by genetic variation creating sequence mismatches with the correct region....We note that such errors are not entirely mitigated by filtering multi-mapped reads—some alignment errors may remain between similar regions even among uniquely aligned reads due to genetic variation, errors in the reference genome, and other complications."

How would the analysis change if an expectation maximization (EM) approach were used for quantifying transcript and gene expression, where a latent variable is used for the origin of each read or pair of reads (both across isoforms and gene loci)? Such methods may resolve issues as shown in Figure 1A, because the observed coverage and expression of only gene A may give a higher likelihood than the observed coverage and expression of gene A and gene B. However the degree to which these methods may offer an improvement with respect to the cross-mappability issue depends on the distribution of genetic variation and errors in the reads, and on potential errors in the reference genome/transcriptome and on incomplete gene annotations. Therefore, it would be critical to perform the trans-eQTL analysis with regards to cross-mappability, when an EM algorithm, or a similar method that resolves multi-mapping reads, is used to quantify gene expression. Methods such as RSEM, Kallisto, or Salmon may be used to quantify gene and transcript abundance, which use EM or variational Bayes EM to resolve multi-mapping reads (as well as reads consistent with multiple isoforms of a gene) (disclosure: I am a co-author of the Salmon method).

The authors note that:

"Alignment to the transcriptome or splice-aware alignment may offer future improvements, but computational cost and inaccuracies due to incorrect annotation will have to be evaluated."

One potential solution which may alleviate both the issue of cross-mappability and incorrect annotation, would be to use an isoform discovery method to detect and characterize novel isoforms in a particular dataset (large datasets of rare tissues or sequencing of RNA from populations which have been under-represented in previous studies may very well discover novel isoforms), and then to use an EM or similar method to quantify expression.

The details about how the GTEx (v7) and DGN data were quantified is missing, although these details are critical for understanding how broad the conclusions of the analysis may be. Gene expression was quantified in GTEx v7 using RNA-SeQC v1.1.8 which does not use an EM approach to resolve multi-mapping reads. According to the Methods section of Battle, et al (2014), for the DGN dataset,

HTSeq was used to quantify gene expression, which also does not use an EM approach. I would recommend to add such quantification details of the datasets to this article.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Developer of methods for estimating gene and transcript expression, and statistical testing of expression across samples.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Apr 2019

Ashis Saha, Johns Hopkins University, Baltimore, USA

Thank you for reviewing our paper. Please see our responses inline in bold and italic font following your comments in standard font.

The authors present a detailed analysis of how false positives arise in trans-eQTL and co-expression analyses, due to cross-mapping of reads among genes with sequence homology. They present a cross-mappability metric, and provide pre-computed cross-mappability for two human Gencode annotations (v19 and v26). They also provide software for efficiently computing cross-mappability available at a GitHub link. The command line software has detailed instructions online. The software requires genome FASTA files, a GTF file, a mappability bedgraph or bigwig file, Bowtie (v1) and an index, and a few R packages (data.table, intervals, argparser, stats).

The report provides an important warning to the research community, and the pre-computed cross-mappability and the software will be a valuable resource for groups studying trans-eQTL and co-expression and making use of unique read counts for gene expression quantification.

One of the key points of the article is noting that the cross-mappable and likely spurious

trans-eQTLs are highly replicable between datasets, because "*it is driven by the underlying sequence of the genome, and similar alignment errors frequently occur regardless of tissue and study*." This logic also extends to co-expression and gene networks built on gene-gene expression correlations. Another key point was that filtering out of cross-mappable trans-eQTLs necessitates re-assessment of FDR, as the highly significant cross-mappable trans-eQTLs bring down the nominal FDR for other eQTLs.

>> Thank you for nicely summarizing the contribution of our work.

My main comment on the article is regarding details of the gene expression quantification.

The authors note that,

"The number of reads misaligned to Gene B across samples may be directly proportional to the number of reads for Gene A, or may be determined by genetic variation creating sequence mismatches with the correct region....We note that such errors are not entirely mitigated by filtering multi-mapped reads—some alignment errors may remain between similar regions even among uniquely aligned reads due to genetic variation, errors in the reference genome, and other complications."

How would the analysis change if an expectation maximization (EM) approach were used for quantifying transcript and gene expression, where a latent variable is used for the origin of each read or pair of reads (both across isoforms and gene loci)? Such methods may resolve issues as shown in Figure 1A, because the observed coverage and expression of only gene A may give a higher likelihood than the observed coverage and expression of gene A and gene B. Howeverthe degree to which these methods may offer an improvement with respect to the cross-mappability issue depends on the distribution of genetic variation and errors in the reads, and on potential errors in the reference genome/transcriptome and on incomplete gene annotations. Therefore, it would be critical to perform the trans-eQTL analysis with regards to cross-mappability, when an EM algorithm, or a similar method that resolves multi-mapping reads, is used to quantify gene expression. Methods such as RSEM, Kallisto, or Salmon may be used to quantify gene and transcript abundance, which use EM or variational Bayes EM to resolve multi-mapping reads (as well as reads consistent with multiple isoforms of a gene) (disclosure: I am a co-author of the Salmon method).

>> We appreciate the importance of this concern. We agree that EM-based quantification in principle has the potential to help address the cross-mappability issue to some extent. We performed additional analyses using RSEM-quantified gene expressions and we still observed a high rate of potential false positives in our analyses, particularly for trans-eQTLs. We updated the manuscript along with a supplementary figure (Supplementary Figure 11).

- ***"We also note that utilization of improved alignment and quantification methods to generate gene expression data may also be helpful to avoid false positives. For example, quantification of gene expression levels using RSEM[28], an expectation maximization based quantification tool, results in a smaller fraction of false positive trans-eQTLs (60.17%) than that using RNA-SeQC (75.14%). However, potential false positives due to cross-mappability still remain abundant in both trans-eQTL and co-expression studies (Supplementary Figure 11)."***

We also added a brief description of the quantification pipelines of GTEx v7 and DGN in the revised manuscript.

The authors note that:

"Alignment to the transcriptome or splice-aware alignment may offer future improvements, but computational cost and inaccuracies due to incorrect annotation will have to be evaluated."

One potential solution which may alleviate both the issue of cross-mappability and incorrect annotation, would be to use an isoform discovery method to detect and characterize novel isoforms in a particular dataset (large datasets of rare tissues or sequencing of RNA from populations which have been under-represented in previous studies may very well discover novel isoforms), and then to use an EM or similar method to quantify expression.

>> This is an interesting idea. For the current analysis, we plan to leave the approach as specified, but would like to explore this in the future. We note, however, that EM-based approaches still do not address many cross-mappability errors in general, as shown in response to the previous question.

The details about how the GTEx (v7) and DGN data were quantified is missing, although these details are critical for understanding how broad the conclusions of the analysis may be. Gene expression was quantified in GTEx v7 using RNA-SeQC v1.1.8 which does not use an EM approach to resolve multi-mapping reads. According to the Methods section of Battle, et al (2014), for the DGN dataset, HTSeq was used to quantify gene expression, which also does not use an EM approach. I would recommend to add such quantification details of the datasets to this article.

>> We appreciate your concern. GTEx v7 used only uniquely mapped reads, but did not use an EM approach in the original analysis, though we have added the RSEM version for the revision as well. To make the manuscript self-contained, we briefly described the gene expression quantification pipeline of both GTEx v7 and DGN in the manuscript.

Competing Interests: No competing interests were disclosed.