Alignment errors in RNA-sequencing produce false positives in association analyses

ashissaha123
alexisjbattle

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

[1]Department of Computer Science, Johns Hopkins University, [2]Department of Biomedical Engineering, Johns Hopkins University.

[1] Motivation: Reads originated from Gene A may incorrectly map to Gene B because of sequence similarity between the genes, leading to false positive co-expression. Consequently, a true cis expression quantitative trait locus (eQTL) variant of the Gene A may appear as a false positive trans-eQTL of Gene B.

[2] Method: We propose a metric to quantify the potential for mapping error between pairs of genes. We define ***cross-mappability*** from gene A to gene B as the number of k-mers from gene A that map to gene B, allowing for a maximum of 2 mismatches. We use 75-mers from exons, and 36-mers from UTRs.

[3] False positive co-expression: A. Comparison of co-expression between randomly drawn 10k cross-mappable and 10k not cross-mappable gene pairs in five GTEx (v7) tissues. B. Fraction of top co-expressed genes that are cross-mappable and thus potential false positives.

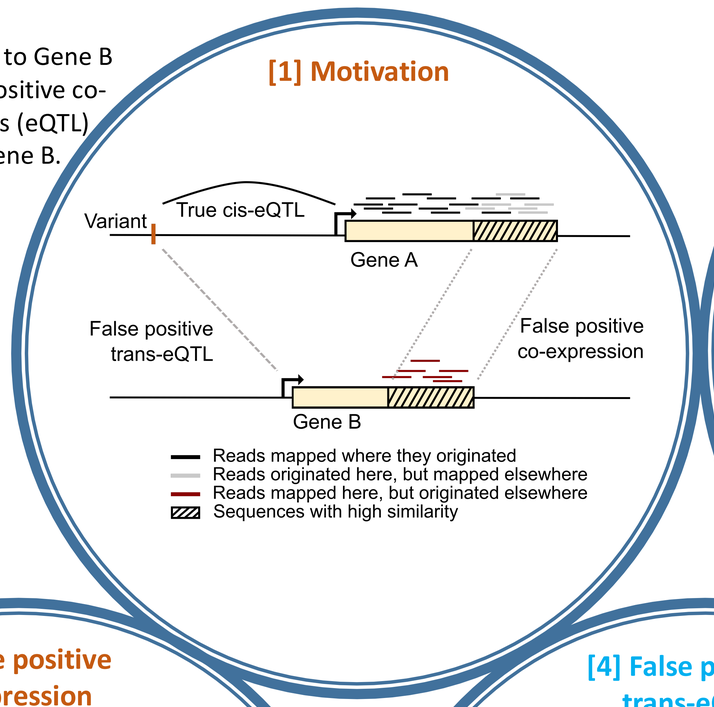
[4] False positive trans-eQTLs: 19,348 unique trans-eQTLs (variant-gene pairs) were detected at FDR ≤ 0.05 from five GTEx tissues. Majority (75.14%) of them were cross-mappable i.e. some gene near the variant is cross-mappable to the trans-eGene. Figure shows that cross-mappable trans-eQTLs are highly abundant among top eQTLs.

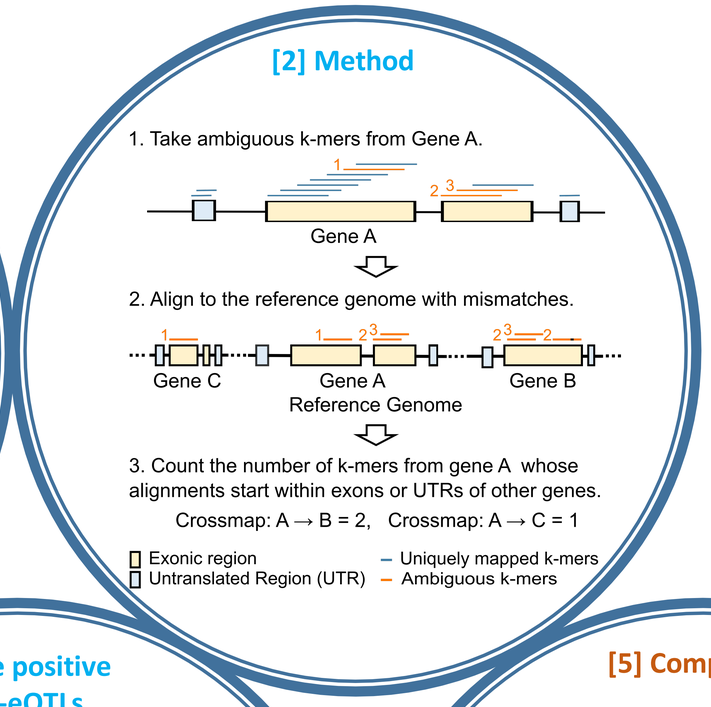
[5] Composition: Abundance of pseudogenes with sequences similar to their parent genes among trans-eQTL target genes (eGene) indicates false positives. A. Representation of gene types among trans eGenes, categorized by cross-mappability. B) Proportion of cross-mappable eQTLs categorized by gene type.

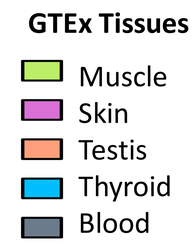
[6] Example: A likely false positive trans association between the variant chr5:149826526 and the gene RP11- 343H5.4. The coverages (reads per million, RPM) of the trans-eGene RP11-343H5.4 (top) and its cross-mapping gene RPS14 (bottom) in Skeletal Muscle are shown along with their annotations ,and mappabilty of 75-mers. Reads map only to non-unique regions of the trans eGene.

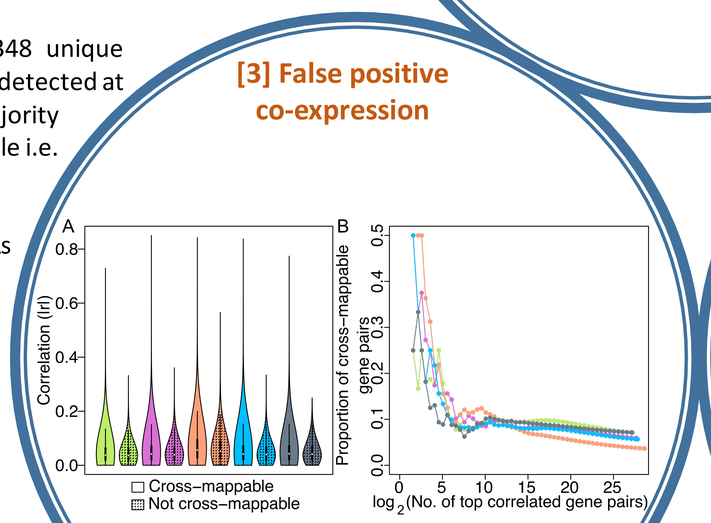
[7] Replication: A) Q-Q plot, replication p-values from DGN for variant-gene pairs discovered in GTEx Whole Blood. B) The fraction of significant eQTLs in each GTEx tissue (row) replicated in another tissue (column) at FDR ≤ 0.05, for cross-mappable (left) and not cross-mappable eQTLs (right). High replication of cross-mappable eQTLs across datasets and across tissues is driven not by biological regulation, but by alignment artifacts.

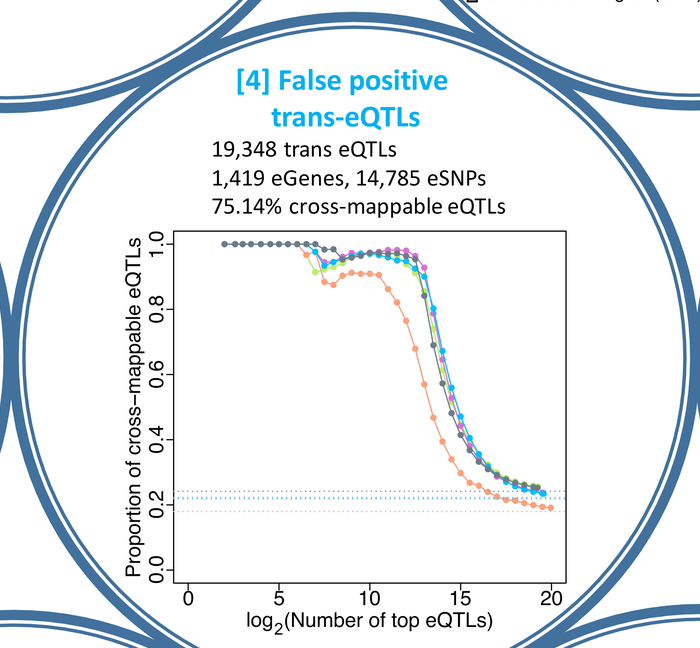
[8] Conclusion: Misalignment of reads should be considered as a potential source of false positives in association studies, particularly for trans-eQTL analysis. Our cross-mappability data can help to flag such potential false positives.

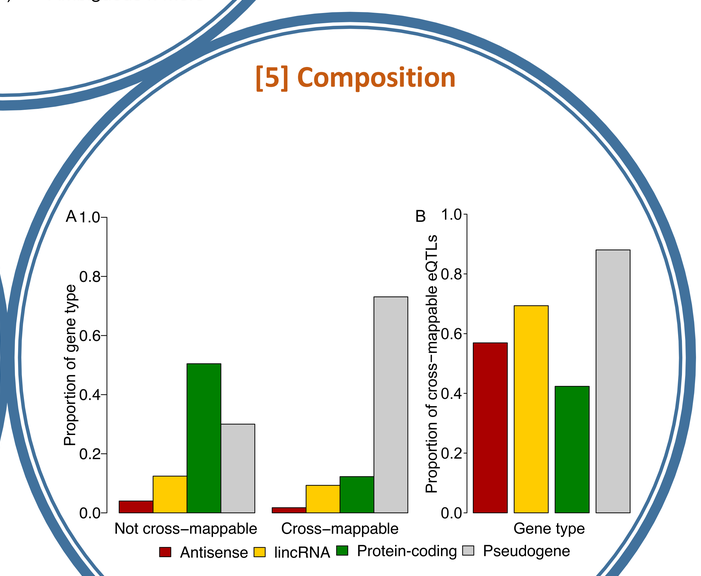


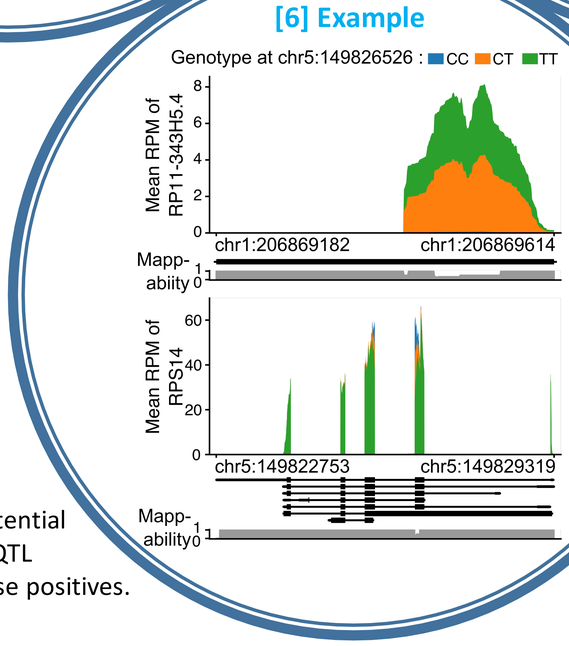


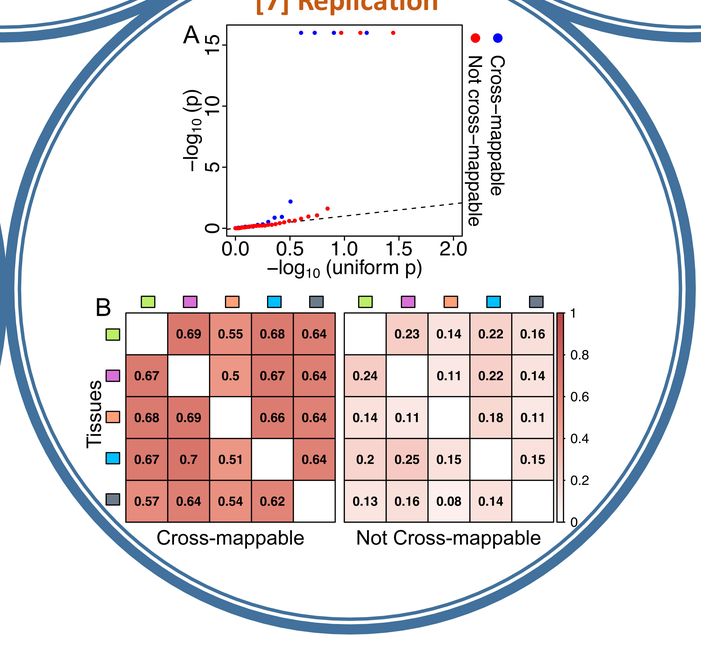












Pre-computed cross-mappability for human genome:
http://bit.ly/mappability

Github repository:
https://github.com/battle-lab/crossmap