Application Note / Short Communication / Brief Report

The Colon Transcriptome Explorer (CoTrEx) 2.0: a Reference Web-Based Resource for Exploring Population-Based Normal Colon Gene Expression

Virginia Díez-Obrero [1,2,3,4], Ferran Moratalla-Navarro [1,3,4], Christopher Dampier [5,6], Matthew Devall [5,6], Robert Carreras-Torres [1,2,3], Graham Casey[5,6] and Victor Moreno [1,2,3,4,*]

1 Oncology Data Analytics Program, Catalan Institute of Oncology (ICO). L'Hospitalet de Llobregat, Barcelona, Spain.
2 Colorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL). L'Hospitalet de Llobregat, Barcelona, Spain.
3 Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Spain.
4 Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain.
5 Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA.
6 Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA.

* Correspondence: v.moreno@iconcologia.net; Tel.: +34 932 607 434

Abstract: Gene expression data is key for the functional annotation of single nucleotide polymorphisms (SNPs) identified in genome-wide association studies (GWAS). Expression and splicing quantitative trait loci (e/sQTLs) in normal colon tissue, such as those from the University of Barcelona and University of Virginia RNA sequencing project (BarcUVa-Seq) and the Genotype-Tissue Expression project (GTEx), are required to gain biological insight of colon-related diseases risk loci. Moreover, transcriptome-wide association studies (TWAS) rely on reference gene expression imputation panels in the tissue of interest to nominate susceptibility genes. Also, it is of high interest to study the relationships between genes in a network framework. For facilitating these analyses, we have updated and expanded the scope of the Colon Transcriptome Explorer (CoTrEx) to the version 2.0. This web-based resource provides exhaustive visualization and analysis of transcriptome-wide gene expression profiles of normal colon tissue from BarcUVa-Seq and GTEx. In addition to the integration of new datasets, CoTrEx 2.0 provides additional e/sQTLs sets, as well as gene expression prediction models and regulatory and co-expression networks. It is freely available at https://barcuvaseq.org/cotrex/. Overall, it is of high interest for researchers aiming to investigate the genetic susceptibility to colon-related complex traits and diseases.

Keywords: RNA-Seq; bioinformatics; web application; gene expression; alternative splicing; visualization; molecular epidemiology

1. Introduction

Datasets of both blood DNA genotyping and RNA sequencing (RNA-Seq) of biopsy samples from a large number of healthy individuals are valuable resources for studies in molecular epidemiology. For example, they provide expression and splicing quantitative trait loci (e/sQTLs) for the annotation of genome-wide association studies (GWAS)-identified risk single nucleotide polymorphisms (SNPs) and gene expression prediction models for transcriptome-wide association studies (TWAS). In this sense, the University of Barcelona and University of Virginia genotyping and sequencing project (BarcUVa-Seq) provided gene expression and alternative splicing profiles of normal (i.e. non-neoplastic, without lessons) colon biopsies from ascending (N=138), transverse (N=143) and descending (N=164) subsites. The expression profiles and their association statistics with

 doi:10.20944/preprints202109.0349.v1

germline genetic variants, i.e. e/sQTLs, were recently reported and included in the initial version of the Colon Transcriptome Explorer (CoTrEx) [1]. Additionally, the Genotype-Tissue Expression (GTEx) project provided normal colon e/sQTLs from transverse (N=368) and sigmoid (N=318) colon samples from corpses [2]. Although the gene expression and related information is provided as supplementary material or deposited in public online repositories, it is often difficult and time consuming for researchers to access the data and analyze and visualize their gene of interest, especially for non-bioinformaticians.

In this article we present the CoTrEx 2.0, an interactive web resource that facilitates the exhaustive visualization and analysis of normal colon gene expression and alternative splicing data from BarcUVa-Seq and GTEx projects. This version, in addition to incorporating GTEx colon datasets and new customization options, provides additional e/sQTL sets, a SNP annotation tool, prediction models statistics for gene expression imputation, and regulatory and gene co-expression networks.

2. Description of CoTrEx 2.0

CoTrEx 2.0 is a web-based resource that includes normal colon gene expression data from BarcUVa-Seq and GTEx projects (see schema in Figure 1). Its main components are divided in the "Expression", "QTLs", "Prediction models", and "Networks" tabs.

On the "Expression" tab, users can search for a gene of interest, select one or more associated transcripts and visualize their expression in multiple ways. On the left panel, the following options are available: i) filter the input data by sex, age and colon anatomic location, ii) select specific visualization features (e.g. heatmap, PCA plot), and iii) group transcripts by relative abundance according to a selected expression threshold (i.e. if 0.05 is selected, the lowest expressed 5% of transcripts is grouped in a single category labeled "Other transcripts"). On the main panel, a customizable stripchart and a barplot are displayed. For example, points in the stripchart can be colored by covariates of interest, and transcript expression can be hidden to show only the expression of selected genes. Annotation by covariate is also available for heatmaps and PCA plots.

On the "QTLs" tab, users can explore lists of significant colon e/sQTLs, including summary statistics and customizable plots showing the distribution of gene expression/percent splicing index by SNP genotype. Users can also search for association statistics for SNPs of interest by selecting the "Annotate SNPs" option. The "Prediction models" tab includes elastic net-based gene expression prediction models for the entire colon and by colon subsite. Descriptive statistics of the prediction models and the SNP weights can be obtained for a gene of interest.

On the "Networks" tab, by selecting the "Regulatory network" option, users can explore gene interactions between TFs and regulated target genes in a network. Arrows are directed from TFs to target genes (either TFs or non-TFs). It is possible to explore first and second order step neighbors by selecting the corresponding option. Descriptive and topological network parameters are provided in tables, including the mutual information (MI) values for each interaction, which indicate the strength of an interaction. The weighted correlation network analysis (WGCNA) approach [3] was used for exploring patterns of correlated gene expression in a gene co-expression network framework. This method makes groups of highly interconnected genes called modules. A total of 20 modules with a mean of 777 highly correlated genes per module were defined, each of them labelled with a color name. The gene-module assignments can be downloaded, and hierarchical clusters of all modules can be explored.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0349.v1

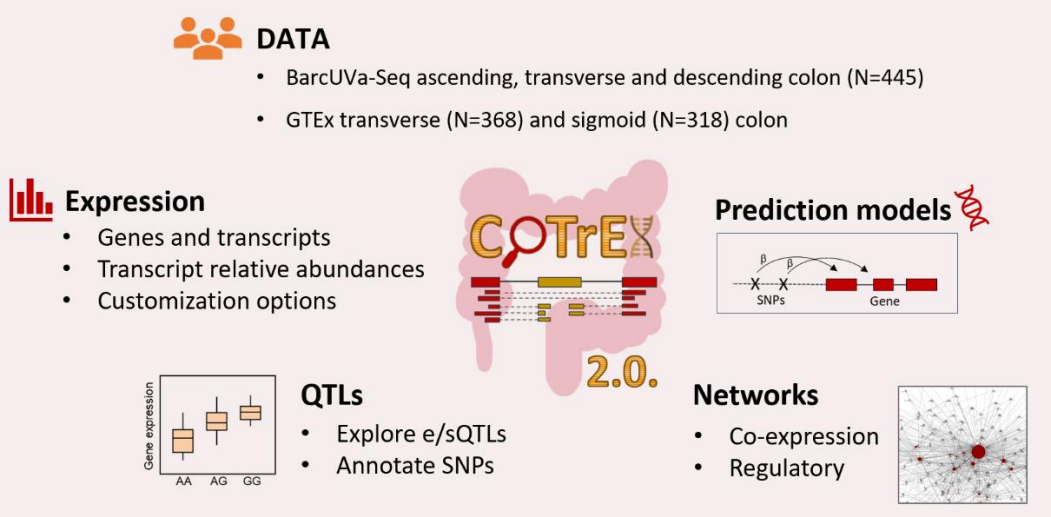

Figure 1. CoTrEx 2.0 schematic.

3. Discussion

We have updated and expanded the scope of CoTrEx to the newest version 2.0, including new data and functionalities. This version includes gene expression and alternative splicing-related data from the GTEx v8 transverse and sigmoid colon. In this version, the genes and transcripts visualized on the Expression tab can be filtered or colored according to the individuals' age and sex. Also, the expression statistics associated with the selected samples can be retrieved. Transcripts can be grouped by relative abundance and hierarchical clustering can be observed in a heatmap. These features are not provided by the GTEx Transcript Browser [4]. In addition, we provide a SNP annotation tool on the QTLs tab where users can provide a list of SNPs of interest to explore associations with genes located up to 1Mb of distance. In contrast, the GTEx eQTL Calculator [5] requires that the users provide the gene ID in addition to the SNP of interest. This is not convenient in cases where the user wants to explore SNP-gene associations of all genes nearby a SNP of interest. Also, the gene expression prediction models can be downloaded from the Prediction models tab, which are useful for investigators interested in performing TWAS and nominating candidate susceptibility genes for a phenotype of interest. A list of complex traits and diseases for which the gene expression prediction models provided in CoTrEx 2.0 are relevant for TWAS is provided elsewhere [1]. Future developments of CoTrEx 2.0 would include additional QTL sets generated, such as regulatory QTLs, associated with changes in interactions between genes.

In conclusion, CoTrEx 2.0 facilitates a quick and centralized access to explore and analyze the most up to date reference gene expression and splicing profiles for non-neoplastic human colon tissue, and their associations with germline genetic variants, which facilitates the understanding of the transcriptomic basis of this tissue. Finally, the CoTrEx 2.0 is a valuable resource for researchers interested in annotating risk loci identified in colon-related GWAS, in performing TWAS for colon-related diseases, and in unraveling the mechanisms underlying inherited susceptibility to colon-related diseases.

4. Materials and Methods

CoTrEx 2.0 was built with the R platform Shiny [6]. Gene and transcript expression counts and e/sQTLs of GTEx v8 sigmoid and transverse colon were obtained from the database of Genotypes and Phenotypes (dbGaP) at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v8.p2. Genes with at least 6 counts in more than 20% of the samples were provided. Expression counts were transformed to trimmed mean of M-values (TMMs). Gene expression prediction models of GTEx v8 were obtained

from elsewhere [1,7] (see data availability statement). Gene expression prediction models of BarcUVa-Seq were generated for the whole sample size and for subsets of the data according to the anatomic location where the biopsies were collected (ascending, transverse and descending colon). The elastic net-based models were generated following the PredictDB pipeline, which was the one used for GTEx v8 data [7]. Following this pipeline, we considered significant gene models those with a predictive performance $P < 0.05$ and $R2 > 0.1$. Gene expression data was adjusted for sex, sequencing batch, probabilistic estimation of expression residuals [PEER] factors [8] and genetic ancestry (2 principal components).

The BC3net R package [9] was used to generate weighted directed gene regulatory networks between 2,195 transcription factors (TFs) and 8,785 target genes. TFs were chosen according to three GO annotations: GO:0045449 "regulation of transcription", GO:0001071 "Nucleic acid binding transcription factor activity", and GO:0140110 "transcription regulator activity". A total of 1,000 bootstraps were run to get a robust final network. Finally, the weighted correlation network analysis (WGCNA) was performed with the WGCNA R package [3]. A soft thresholding of 6 was selected to approximate to scale free topology.

Author Contributions: Conceptualization, V.M. and V.D.; methodology, V.D, F.M., R.C.; software, V.D.; data curation, V.D., F.M., C.D., M.D.,R.C; writing—original draft preparation, V.D.; writing—review and editing, V.D., F.M., C.D., M.D., R.C., G.C., V.M.; visualization, V.D.; supervision, V.M.; funding acquisition, V.M., G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Institutes of Health, grant numbers R01 CA204279, R01 CA143237 and R01 CA201407; the Agency for Management of University and Research Grants (AGAUR) of the Catalan Government, grant number 2017SGR723; the Instituto de Salud Carlos III, co-funded by FEDER funds –a way to build Europe, grant numbers PI14-00613, PI17-00092; the Spanish Association Against Cancer (AECC) Scientific Foundation, grant number GCTRA18022MORE; and the Centro de investigación biomédica en red. Epidemiología y salud pública (CIBERESP), grant number CB07/02/2005. RCT received funding through the EU H2020 – MSC, grant number 796216; CHD received funding through the National Institutes of Health, grant number T32 5T32CA163177-07; VDO received founding through the Spanish "Ministerio de Educación, Cultura y Deporte", grant number FPU16/00599.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: GTEx v8 sigmoid and transverse colon data were obtained from the database of Genotypes and Phenotypes (dbGaP) at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v8.p2. GTEx v8 gene expression prediction models were obtained from Zenodo, at https://dx.doi.org/10.5281/zenodo.3519321.

Acknowledgments: We thank the CERCA Program, Generalitat de Catalunya, for institutional support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Díez-Obrero, V.; Dampier, C.H.; Moratalla-Navarro, F.; Devall, M.; Plummer, S.J.; Díez-Villanueva, A.; Peters, U.; Bien, S.; Huyghe, J.R.; Kundaje, A.; et al. Genetic Effects on Transcriptome Profiles in Colon Epithelium Provide Functional Insights for Genetic Risk Loci. *Cell Mol Gastroenterol Hepatol* **2021**, doi:10.1016/j.jcmgh.2021.02.003.
2. GTEx Consortium The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* **2020**, 369, 1318–1330.
3. Langfelder, P.; Horvath, S. WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* **2008**, 9.
4. GTEx Transcript Browser Available online: https://gtexportal.org/home/transcriptPage (accessed on 17 May 2021).
5. GTEx eQTL Calculator Available online: https://gtexportal.org/home/testyourown (accessed on 17 May 2021).
6. Chang, W.; Cheng, J.; Allaire, J.J.; Sievert, C.; Schloerke, B.; Xie, Y.; Allen, J.; McPherson, J.; Dipert, A.; Borges, B. Shiny: Web Application Framework for R; 2021.
7. Barbeira, A.N.; Liang, Y.; Bonazzola, R.; Wang, G.; Wheeler, H.E.; Melia, O.J.; Aguet, F.; Ardlie, K.G.; Wen, X.; Im, H.K.; et al. Fine-Mapping and QTL Tissue-Sharing Information Improve Causal Gene Identification and Transcriptome Prediction Performance. *bioRxiv* 2020:2020.03.19.997213. Doi: 10.1101/2020.03.19.997213.

 doi:10.20944/preprints202109.0349.v1

8. Stegle, O.; Parts, L.; Piipari, M.; Winn, J.; Durbin, R. Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses. *Nat. Protoc.* **2012**, 7, 500–507.
9. de Matos Simoes, R.; Emmert-Streib, F. Bagging Statistical Network Inference from Large-Scale Gene Expression Data. *PLoS One* **2012**, 7, e33624.