*Ontoscope:* Determining “identity-defining” transcription factors for various cell types

Vignette authors: Kartikay Chadha, Shivani Kamdar, Burton Mendonca, Anam Qudrat

University of Toronto

Department of Biochemistry

Compiled: April 27, 2016

**Contents**

**1 Introduction 2**

**2 Processing overview 2**

**3 Submodules and workflow**  **z**

3.1 Normalization z

3.2 FANTOM expression data import z

3.3 Background derivation from cell ontology z

3.4 Protein and transcription factor network analysis z

3.4.1 STRING

3.4.2 TRRUST

3.4.3 Regnet

3.5 Differential gene expression analysis z

3.6 Integration of network- and expression-based influence scores z

3.7 Selection of transcription factors needed for conversion z

**4 Post-processing and analysis z**

4.1 Visualization z

4.2 Literature-based confidence score z

4.3 Validation from published experimental data z

**5 Acknowledgments z**

**6 Session info z**

**7 References z**

**1 Introduction**

This document offers an overview of the R package *Ontoscope*, which uses publicly available expression and protein-protein interaction data to computationally determine a list of transcription factors able to facilitate conversion from one cell type to another. Following determination of this transcription factor list, it is also designed to assign confidence scores to each factor based on literature search and validation against published experimental data and/or other computational prediction software.

The *Ontoscope* package is designed to independently validate the findings of Owen Rackham and colleagues’ *Mogrify* ([www.mogrify.net](http://www.mogrify.net))1. Its workflow is based on the published *Mogrify* protocol, with modifications including the usage of updated transcription factor definitions and the inclusion of different regulatory networks (specifically, the TRRUST and Regnet protein-protein interaction datasets).

This vignette contains an overview of the package workflow, defining initialization parameters and walking through each submodule with examples and usage instructions provided.

**2 Processing overview**

*Ontoscope* determines transcription factors required for conversion based on calculation of network- and expression-based influence scores for differentially expressed factors in the target cell line as compared to the source. Target and source cells are identified by FANTOM consortium cell line IDs, and are restricted to those cell types for which FANTOM gene expression data from CAGE-seq is available.

Calculation involves six primary phases:

1. *Import of expression data:* The first step is to define source and target cell lines, reading in raw FANTOM expression count data for each cell line. Transcription factors are defined from published literature and databases, and all gene names are normalized to HGNC IDs in order to ensure accuracy of comparison between cell lines and datasets. Download of FANTOM count data is automatic upon input of cell line IDs, and does not require further input from the user.
2. *Calculation of background:* Something
3. *Differential expression-based influence score assignment:* Something
4. *Interaction-based influence score assignment:* Something
5. *Integration and binning:* Something
6. *Visualization and validation:* Something

**3 Submodules and workflow**

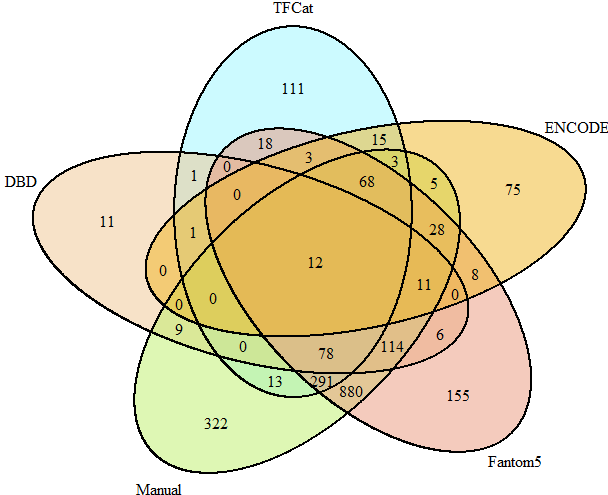
This section outlines examples and functions underlying the usage of *Ontoscope* to generate lists of transcription factors for conversion.

(An outline of how the user is supposed to input source and target IDs will go here).

**3.1 Normalization**

Normalization works through doing (x, y, z).

> samplecode(x)



**Figure 1: Sample figure.** Sample legend about figure. Generated by the VennDiagrams package: draw.quintuple.venn(sample, cat.col=rep(“black”))

**3.2 FANTOM expression data import**

**3.3 Background derivation from cell ontology**

**3.4 Protein and transcription factor network analysis**

3.4.1 STRING

3.4.2 TRRUST

3.4.3 Regnet

**3.5 Differential gene expression analysis**

**3.6 Integration of network- and expression-based influence scores**

**3.7 Selection of transcription factors needed for conversion**

**4 Post-processing and analysis**

**4.1 Visualization**

**4.2 Literature-based confidence score**

**4.3 Validation from published experimental data**

**5 Acknowledgments**

This package was developed as part of the BCB420/JTB2020 bioinformatics class as part of the Department of Biochemistry at the University of Toronto, Canada, with the support of Dr. Boris Steipe. We wish to acknowledge all co-authors of *Ontoscope* package modules, who have equally contributed to the development of this project: Julian Mazzitelli, Dmitry Horodetsky, Christopher Li, Kartikay Chadha, Fupan Yao, Shivani Kamdar, Dan Litovitz, Howard Wu, Anam Qudrat, Jialun Tom Chen, Ryoga Li, Eugenia Barkova, Burton Mendonca, Phil Fradkin, and Boris Steipe.

**6 Session Info**

> sessionInfo(package=NULL)

(The output of the above command needs to be filled in by Dr. Steipe after integrating all the code and running it).

**7 References**

[1] Rackham OJ, Firas J, Fang H, Oates ME, Holmes ML, Knaupp AS, FANTOM Consortium, Suzuki H, Nefzeger CM, Daub CO, Shin JW, Petretto E, Forrest AR, Hayashizaki Y, Polo JM, Gough J. A predictive computational framework for direct reprogramming between human cell types. *Nat Genet,* 48(3): 331-335, 2016.