Application note

TRES predicts transcription control in embryonic stem cells.

Christopher Pooley[1], David Ruau[2], Patrick Lombard[2], Berthold Gottgens[2,*] and Anagha Joshi[1,*]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 8GR, UK.

[2]Department of Haematology, Wellcome Trust and MRC Cambridge Stem Cell Institute, Cambridge Institute for Medical Research, Cambridge University, Cambridge, UK.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

***ABSTRACT**

Summary: Unraveling transcriptional circuits controlling embryonic stem cell maintenance and fate has great potential for improving our understanding of normal development as well as disease. To facilitate this, we have developed a novel web tool called 'TRES' that predicts the likely upstream regulators for a given gene list. This is achieved by integrating transcription factor (TF) binding events from 187 ChIP-sequencing and ChIP-on-chip datasets in murine and human ES cells with over 1000 mammalian TF sequence motifs. Using 114 TF perturbation gene sets, as well as 115 co-expression clusters in ES cells, we validate the utility of this approach.

Availability and implementation: TRES is freely available at http://www.tres.roslin.ed.ac.uk.

Contact: Anagha.Joshi@roslin.ed.ac.uk.

1 INTRODUCTION

Embryonic stem (ES) cells have limitless self-renewal capability and the ability to differentiate into multiple cell types. This makes them a good model system to enhance our understanding of normal development, and also for potential clinical applications. They facilitate *in vitro* studies of early developmental events as well as differentiation into crucial cell types, such as hematopoietic or neuronal cells (Keller, 2005). Most importantly they have proven to be a highly valuable resource in regenerative medicine, where ES cells show great potential in tissue repair following disease or injury (Balber, 2011). These therapeutic applications have strengthened the push towards understanding how stem cells are programmed during self-renewal and differentiation.

Transcription factors (TFs) are key players in driving cellular programming (Takahashi and Yamanaka, 2006). Many TFs have been identified with crucial roles in ES-cell biology (Young, 2011), and their genome-wide putative targets have been mapped using ChIP-sequencing or ChIP-on-chip technology (Xu *et al.*, 2013). Moreover, it is also now feasible to obtain binding sequence preferences for hundreds of mammalian TFs (Jolma *et al.*, 2013). To utilize the complementary information from both these resources to aid novel hypotheses generation, we have developed a web tool called 'Transcription Regulation in Embyonic Stem Cells', or TRES for short, to link gene sets to likely upstream regulators in ES cells.

*To whom correspondence should be addressed.

2 THE TRES PIPELINE

Figure 1A shows a screenshot of the web tool where users can paste a query gene list (human or mouse) or upload it from a file. The gene list is interrogated against four databases and enrichment is calculated using the optimized methods for each database (see supplementary data for details). The output is displayed as a ranked list of TFs generated using rank aggregation of results from four databases as well as four tables each listing significantly associated TFs, the source database, the gene overlap and corrected P values sorted from lowest to highest within each database (Figure 1B). The results can also be emailed to the user when the calculation is finished. The enrichment calculation protocols for the four databases are (supplementary information for full details):

(1) Based on genome-wide binding patterns for 97 murine and 49 human TFs (Martello *et al.*, 2012; Dunham *et al.*, 2012), the significance of overlap between each ChIP-sequencing dataset and the gene list is calculated using a weighted approach, which considers the number of binding events at each gene locus (Joshi *et al.*, 2013).

(2) Based on ChIP-on-chip data for 41 TFs (Xu *et al.*, 2013), enrichment for each TF is calculated by assigning weights to genes proportional to the number of binding sites at a given gene locus normalized by gene length.

(3) Based on 684 sequence motifs for mammalian TFs from the JASPAR database (Bryne *et al.*, 2008), motif enrichment is calculated using all genomic regions in a gene locus bound by at least one transcription factor from the ChIP-sequencing compendium (1) using Centrimo (Bailey and Machanick, 2012).

(4) Based on binding sites for 550 human TFs determined using high-throughput SELEX and ChIP sequencing. (Jolma *et al.*, 2013), motif enrichment is calculated using the same method as in (3).

TRES analysis of cancer gene sets from MsigDB (Liberzon *et al.*, 2011) associated Myc and E2F family members (Figure 1B, left) to genes up regulated in multiple cancers, such as breast, bladder, liver and lung. On the other hand, PRC complex components Ezh2, Ring1b and Suz12 (Figure 1B, right) were enriched in genes silenced by methylation across multiple cancer datasets (supplemetary table 1).

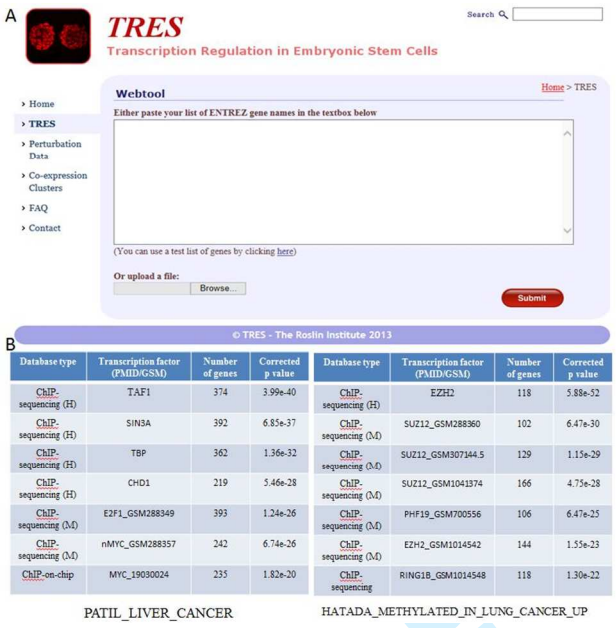

Fig.1. **(A)** Screenshot of the web tool with an option to either paste a user defined gene list or upload it from a file. **(B)** Top table of TRES output for two cancer gene sets from MsigDB. The results show Myc and E2F family members associated with genes up-regulated in cancer (left), and Ezh2, Suz12 and Ring1b associated with genes methylated in cancer (right).

2.1 Perturbation gene sets: case study I

To validate the TRES approach, we collected differentially expressed gene sets after deletion of 60 TFs (Xu *et al.*, 2013) and after overexpression of 54 TFs (Nishiyama *et al.*, 2009). If the perturbed TF was among the enriched TFs, it was considered a true positive. The TRES output for all perturbation datasets is provided on the website. 22 of the 35 TFs present in the TRES database were correctly associated with the differentially expressed gene sets after deletion of TFs, whereas 19 of 34 TFs were correctly associated with overexpression of TFs. The combination of four different information sources thus provides a much better coverage compared to any individual source at similar recall and precision values (Table 1).

#	Database type	Deletion	Overexpression
1	ChIP sequencing	10 (0.50; 0.033)	5 (0.56; 0.065)
2	ChIP-on-chip	14 (0.61; 0.10)	7 (0.70; 0.094)
3	Jaspar motifs	13 (0.54; 0.015)	13 (0.59; 0.055)
4	Jolma 2013 motifs	4 (0.29; 0.017)	9 (0.47; 0.061)
5	All	22 (0.51; 0.023)	19 (0.58; 0.058)

Table 1:The number of true positives (with corresponding recall and precision values in brackets) for each of the database type and all together (rows) and TF deletion or overexpression sets (columns).

We performed motif enrichment by replacing ChIP-bound regions with promoters (H3K4me3 peaks from (Xiao *et al.*, 2012)) and promoters and enhancers (H3K4me3 or H3K4me1 or H3K27ac peaks from (Xiao *et al.*, 2012)). Both performed much worse than ChIP-bound regions (data not shown). This shows that ChIP-bound regions provide the best repertoire of regulatory regions for motif enrichment analysis.

2.2 Co-expression clusters: case study II

To investigate gene networks operating in ES cells, the FunGenES consortium analyzed the transcriptome of mouse ES cells in 67 experimental conditions and created 115 co-expression clusters (Schulz *et al.*, 2009). 70 of the 115 clusters were associated with one or more TFs as their putative regulators (results available on the website). Sp1 was associated with cluster 10 over-represented for `immune response` function using motif information. Cluster 30, containing genes involved in the formation of the three embryonic germ layers during gastrulation, as well as cluster 15 involved in early mesoderm development, are preferentially bound by Ezh2 and Suz12 transcription repressors in ES cells. Cluster 4, which consists mostly of genes associated with neuronal development and differentiation, is preferentially bound by Tfap2a.

3 CONCLUSION

As next generation sequencing technology evolves, more and more information is being generated about genome-wide TF-targets and motifs. These resources, however, still remain under-explored for hypothesis generation. We have combined predictions from four databases, using the most suited method for each data type, to associate likely upstream regulators. The TRES web tool can help to unravel potential regulatory mechanisms underlying cancer gene sets, thus enables investigations into the mechanisms responsible for the expression of gene sets with diagnostic or prognostic relevance. A web-based implementation of TRES allows user-friendly access for the wider research community, and thus provides a substantial new addition to the bioinformatic toolbox for stem cell gene set analysis.

ACKNOWLEDGEMENTS

This work was supported by a University of Edinburgh Chancellors Fellowship awarded to AJ and strategic funding from the BBSRC. CP was funded by the Scottish Government through the Strategic Partnership for Animal Science Excellence (SPASE). The Gottgens' lab is supported by LLR, the MRC, BBSRC, Cancer Research UK, and Wellcome Trust core support to the Cambridge Institute for Medical Research and Wellcome Trust–MRC Cambridge Stem Cell Institute.

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128–e128.
Balber,A.E. (2011) Concise review: aldehyde dehydrogenase bright stem and progenitor cell populations from normal tissues: characteristics, activities, and emerging uses in regenerative medicine. *Stem Cells Dayt. Ohio*, **29**, 570–575.
Bryne,J.C. *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–106.
Dunham,I. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
Jolma,A. *et al.* (2013) DNA-Binding Specificities of Human Transcription Factors. *Cell*, **152**, 327–339.
Joshi,A. *et al.* (2013) Gene set control analysis predicts hematopoietic control mechanisms from genome-wide transcription factor binding data. *Exp. Hematol.*, **41**, 354–366.e14.
Keller,G. (2005) Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes Dev.*, **19**, 1129–1155.
Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinforma. Oxf. Engl.*, **27**, 1739–1740.
Martello,G. *et al.* (2012) Esrrb is a pivotal target of the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal. *Cell Stem Cell*, **11**, 491–504.
Nishiyama,A. *et al.* (2009) Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell*, **5**, 420–433.
Schulz,H. *et al.* (2009) The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *Plos One*, **4**, e6804.
Takahashi,K. and Yamanaka,S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.
Xiao,S. *et al.* (2012) Comparative epigenomic annotation of regulatory DNA. *Cell*, **149**, 1381–1392.
Xu,H. *et al.* (2013) ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database J. Biol. Databases Curation*, **2013**, bat045.
Young,R.A. (2011) Control of the Embryonic Stem Cell State. *Cell*, **144**, 940–954.