Supplementary Material

Laraib Malik, Fatemeh Almodaresi, and Rob Patro

Methods

Likelihood values are calculated using the following equations, adopted from [1]:

$$\ell_0 = \sum_c \left[\left(X_i^c \cdot \log\left(r_{ij}\mu_j^c\right)\right) - \left(r_{ij}\mu_j^c\right)\right] + \left[\left(X_i^c \cdot \log\left(\mu_j^c\right)\right) - \left(\mu_j^c\right)\right] \tag{1}$$

$$\ell_1 = \sum_c \left[\left(X_i^c \cdot \log\left(r_{ij}^c X_j^c\right)\right) - \left(r_{ij}^c X_j^c\right)\right] + \left[\left(X_i^c \cdot \log\left(X_j^c\right)\right) - \left(X_j^c\right)\right] \tag{2}$$

where

$$r_{ij}^c = \frac{X_i^c}{X_j^c},\ r_{ij} = \frac{\sum_c X_i^c}{\sum_c X_j^c},\ \text{ and } \mu_j^c = \frac{X_i^c + X_j^c}{1 + r_{ij}},$$

and X_i^c denotes the number of reads mapping to contig c_i under the j^{th} condition (summed over all replicates of a condition for simplicity). Edges with value $2\left(\ell_1 - \ell_0\right) > 20$ are removed from the graph.

Datasets

We tested our tool on datasets from four organisms. The first dataset we used is from human, *Homo sapiens*, (SRA accessions SRR493366-SRR493371) primary lung fibroblast samples, with and without a small interfering RNA (siRNA) knock down of HOXA1 [2]. The second sample is from yeast, *Saccharomyces cerevisiae*, grown under batch and chemostat conditions (SRA accessions SRR453566 to SRR453571) [3]. Differential expression testing was done on these two datasets. The third set is from dendritic cells in mice, *Mus musculus*, (SRA accession SRR203276) [4], and the last from deep sequencing of Asian rice, *Oryza sativa*, (SRA accessions SRR037735-SRR037738) [5]. The *de novo* assemblies were generated using *Trinity* and the "true" clustering for each dataset was obtained by running *BLAST* against the respective genomes. The genome versions used for each of the species were hg19 for human, R64-1-1 for yeast, GRCm38 for mouse and IRGSP-1.0 for Asian rice. To test the labeling module, annotated, closely related species were used. Macaque (*Macaca mulatta*, assembly MMUL 1), chimp (*Pan troglodytes*, assembly CHIMP2.1.4), orangutan (*Pongo abelii*, assembly PPYG2), gorilla (*Gorilla gorilla gorilla*, assembly gorGor3.1), and gibbon (*Nomascus leucogenys*, assembly Nleu1.0) were used for human; rat (*Rattus norvegicus*, assembly Rnor 6.0) was used for mouse; red rice (*Oryza punctata*,assembly AVCL00000000) and wild rice (*Oryza barthii*, assembly ABRL00000000) were used for Asian rice. All the genomes and annotations for the related species were obtained from the Ensembl database [6].

"True" clustering

A genome based analysis is used to determine the "true" clustering of contigs in the *de novo* assembly. The mapping between the *de novo* assembly and reference genome is done using nucleotide level *BLAST*. The best possible mappings are filtered such that the percent match is at least 98 and the length of the match is at least 200 bases. If there are multiple such mappings, the longest one is selected and ties are broken randomly. Then, contigs with the same gene label are said to come from the same "true" cluster. Any contig not included in the truth set is excluded from all the results shown. The precision and recall values are calculated with respect to this clustering. A true positive is counted when two contigs with the same label under the genome based labeling are put in the same cluster by the tool. Similarly, a false positive is counted when two contigs are put in the same cluster but do not have the same label in the truth set. In the DE tests, this clustering is used to aggregate the quantification estimates to the gene level for detecting truly differentially expressed genes.

Running tools

All results were generated using Python 2.7.12. The *de novo* assemblies in all tests were generated using *Trinity* v2.2.0, run with default parameters. *Salmon* v0.8.2 was also run with 4 threads in all tests, using the flags `--dumpEq` to write *equivalence classes* to a file that can then be processed by *Grouper*, `--writeOrphanLinks` to write orphan reads to a file that can optionally be read by *Grouper* to improve the mapping ambiguity graph, `--discardOrphans` which ignores orphan reads while estimating expression levels of the contigs, and setting `--incompatPrior` to 0 in order to ignore reads that disagree with the specified or inferred library format. The `--discardOrphans` option improves results of *Grouper* since the number of orphan reads is much higher in *de novo* assemblies and affects the overall quantification estimates that *Grouper* uses for various graph filtering steps. *Corset* v1.05 was also run with default parameters, using 4 threads. The alignment files for *Corset* were not processed concurrently in the analysis, since we observed this gave significantly worse accuracy results even though it was much faster. Bowtie was used to align reads to the reference for *Corset*, and was run with 4 threads using parameters suggested in the *Corset* paper. We use *MCL*, an off-the-shelf method, for clustering the mapping ambiguity graph. We performed multiple tests using other clustering methods, with *MCL* giving the best results. All experiments were performed on a 64-bit Linux server, running Ubuntu 14.04, with 4 hexacore Intel Xeon E5-4607 v2 CPUs (with hyper-threading) running at 2.60GHz and 256GB of RAM. Wall-clock time was recorded using the Unix `time` command.

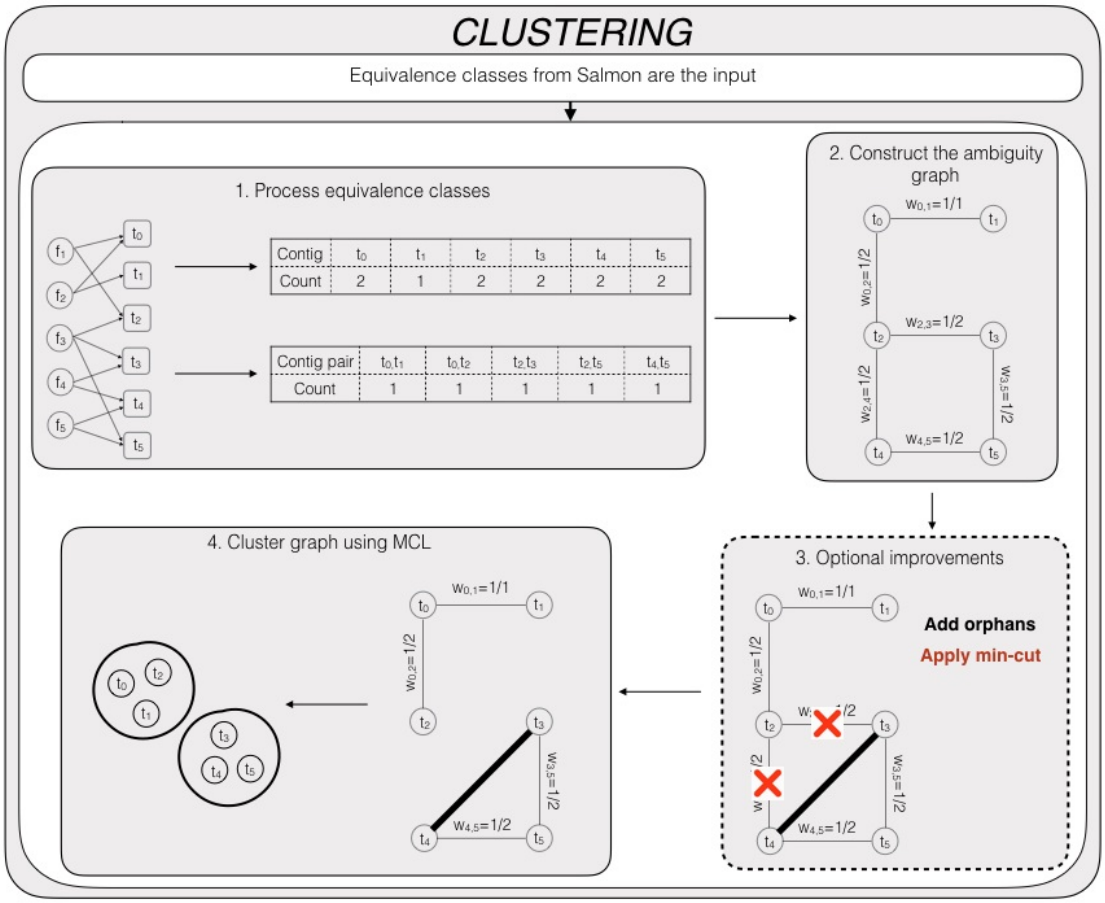

Figure 1: Overview of the clustering module in *Grouper*.

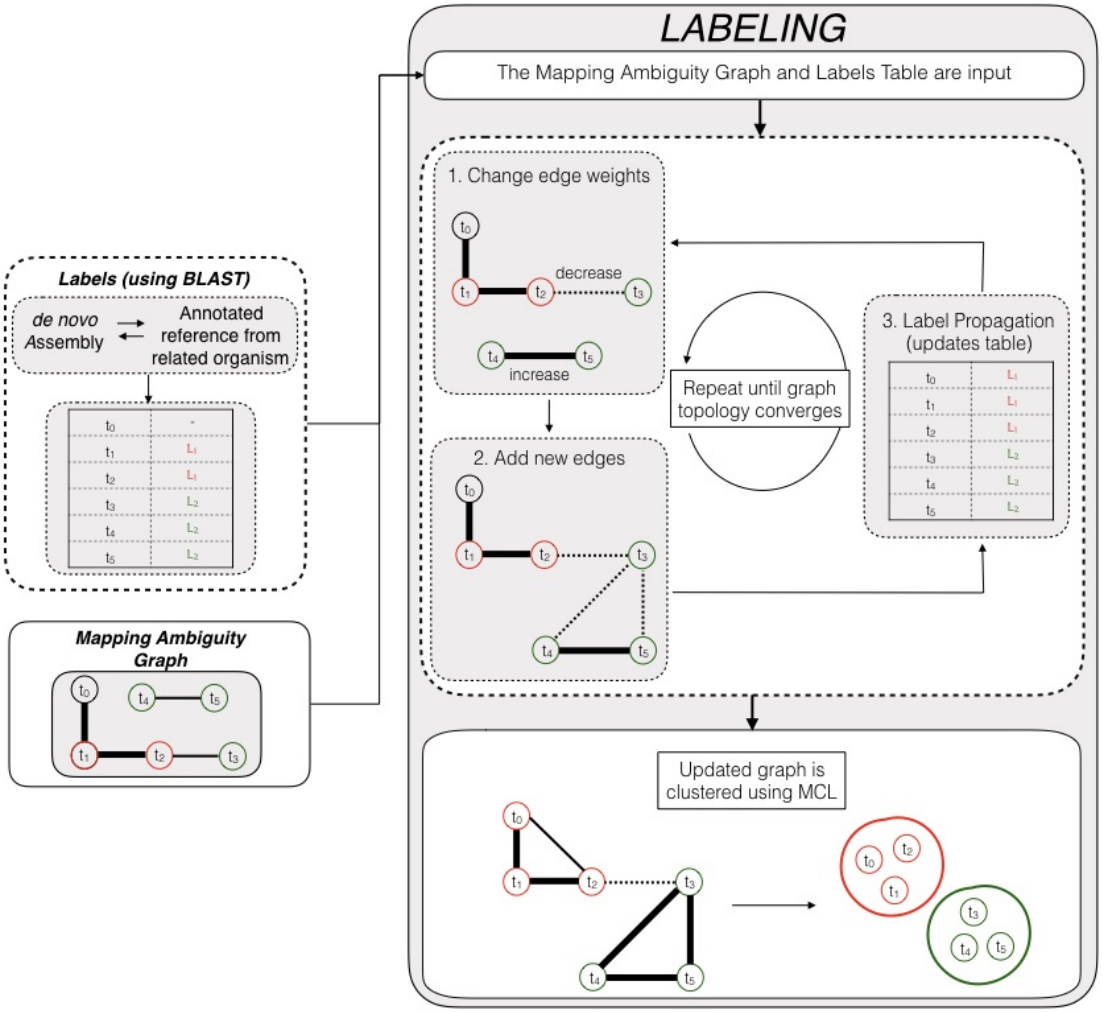

Figure 2: Overview of the labeling module in *Grouper*.

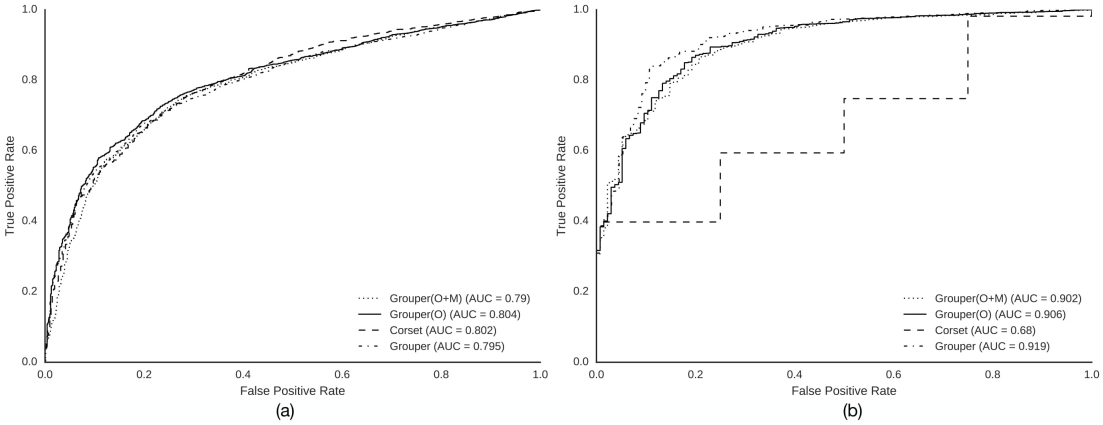

Figure 3: DGE results: The curve represents the accuracy (true positives against false positives) of calling differentially expressed genes using *RSEM* counts as ground truth, represented by the clusters generated by each method in the human (a) and yeast (b) datasets.

Table 1: Number of contigs clustered by each method in the *de novo* assemblies. Note that although both *Corset* and *Grouper* have the same underlying principle of removing low count contigs, the quantification method itself is different in the tools. This is also why the various filters in *Grouper* do not significantly alter the count of contigs.

	Corset	*Grouper*	*Grouper* (O)	*Grouper* (O+M)
Human	69,107	80,034	80,441	80,441
Yeast	4,145	4,417	4,421	4,421
Mouse	36,627	36,943	37,106	37,106
Rice	73,082	71,853	72,152	72,152

Table 2: Number of contigs clustered by each method in the transcriptomes.

	Corset	*Grouper*	*Grouper* (O)	*Grouper* (O+M)
Human	109,232	125,672	125,695	125,695
Yeast	3,592	6,725	6,725	6,725
Mouse	54,931	53,199	53,219	53,219
Rice	33,340	31,022	31,025	31,025

Table 3: Memory usage of each clustering method using the *de novo* assemblies and the reference transcriptomes (in Mb).

	de novo assembly				Transcriptome			
	Corset	*Grouper*	*Grouper* (O)	*Grouper* (O+M)	*Corset*	*Grouper*	*Grouper* (O)	*Grouper* (O+M)
Human	3561	210	364	363	4765	482	1149	1733
Yeast	897	77	84	84	746	80	90	91
Mouse	8672	123	174	174	8943	189	310	310
Rice	1779	163	253	253	1522	126	162	162

Table 4: Alignment time of each method on the *de novo* assemblies and the transcriptomes (in minutes).

	De novo assemblies		Transcriptomes	
	Bowtie	*Salmon*	Bowtie	*Salmon*
Human	196.2	12.85	354.8	20.31
Yeast	34.33	3.67	32.11	4.38
Mouse	58.3	6.24	178.95	6.82
Rice	24.24	6.72	112.16	7.5

References

[1] Nadia M Davidson and Alicia Oshlack. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome biology*, 15(7):410, 2014.

[2] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):46–53, 2013.

[3] Intawat Nookaew, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen. A comprehensive comparison of rna-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in saccharomyces cerevisiae. *Nucleic acids research*, 40(20):10084–10097, 2012.

[4] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.

[5] Guojie Zhang, Guangwu Guo, Xueda Hu, Yong Zhang, Qiye Li, Ruiqiang Li, Ruhong Zhuang, Zhike Lu, Zengquan He, Xiaodong Fang, et al. Deep rna sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome research*, 20(5):646–654, 2010.

[6] Javier Herrero, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J Vilella, Stephen MJ Searle, Ridwan Amode, Simon Brent, et al. Ensembl comparative genomics resources. *Database*, 2016:bav096, 2016.