Supplementary Materials for

transcriptogramer: an R/Bioconductor package for transcriptional analysis based on protein-protein interaction

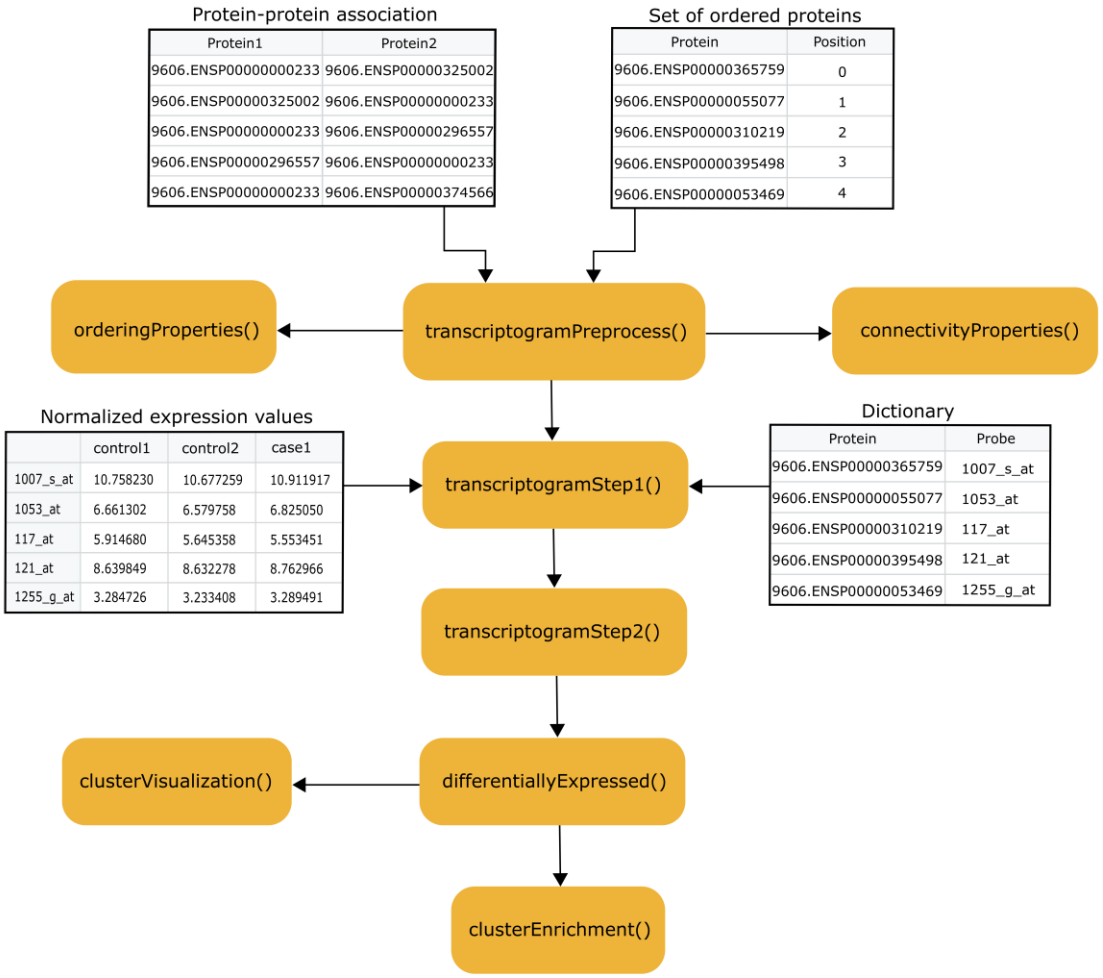
Diego A. A. Morais[1], Rita M. C. Almeida[2] and Rodrigo J. S. Dalmolin[1,3]

[1]Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte, Natal, RN, Brazil

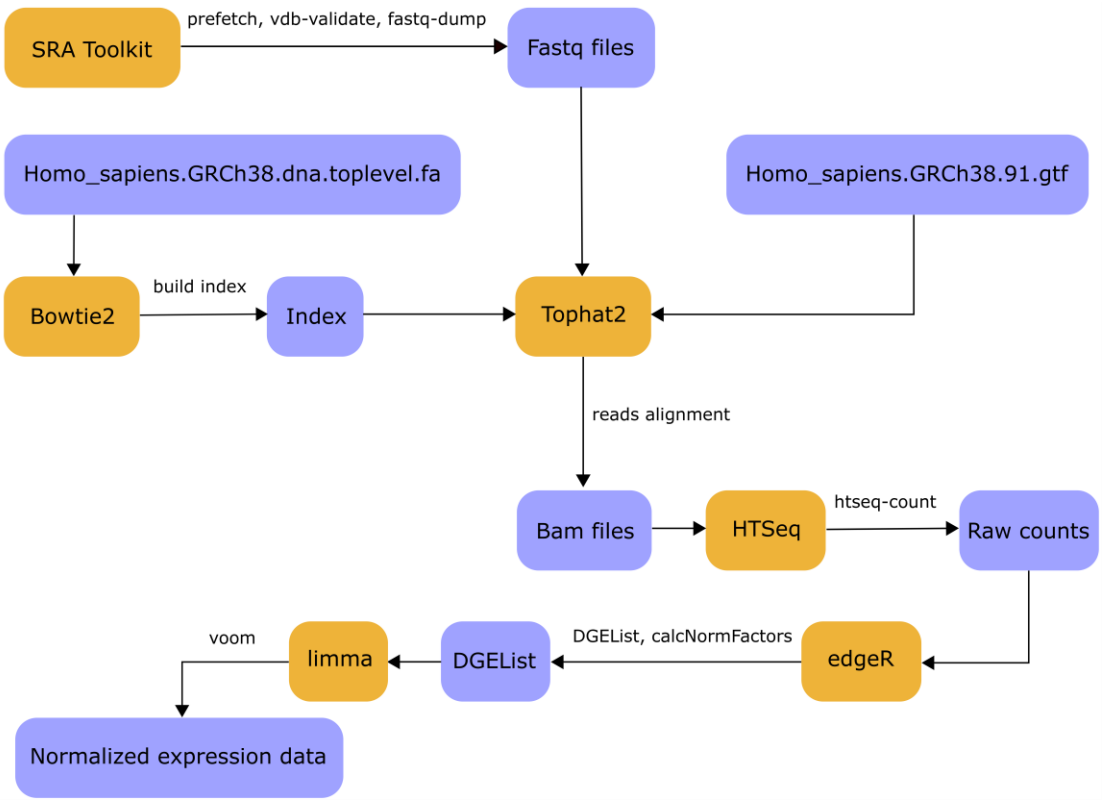
[2]Institute of Physics and National Institute of Science and Technology: Complex Systems, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

[3]Department of Biochemistry – CB, Federal University of Rio Grande do Norte, Natal, RN, Brazil

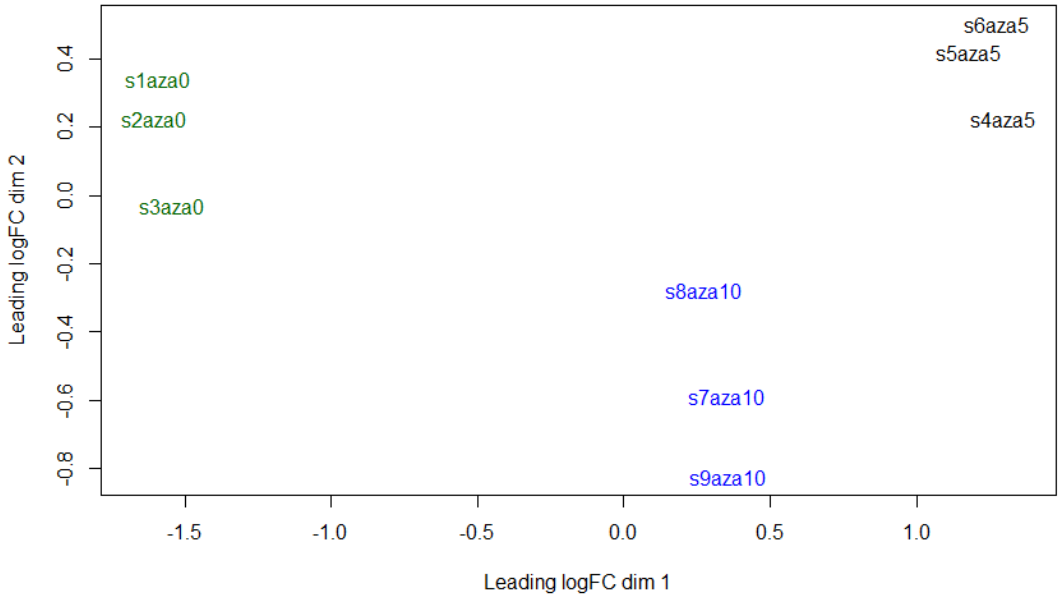
Contents



Supplementary Figure S1. Analysis schematic workflow. Methods are represented as orange rounded squares and inputs are represented as squares. Despite the input data.frame like appearance, other input formats are accepted by the methods. The outputs are provided to the user as Transcriptogram objects, RedPort objects, figures, and data.frames.



Supplementary Figure S2. Sequence read archives processing schematic workflow. Tools are represented as orange rounded squares and inputs/outputs are represented as blue rounded squares. edgeR package was used to create a DGEList object from raw counts, to filter low expressed counts, and to normalize it by Trimmed Mean of M-values (TMM) normalization method. After filtering, only genes with cpm > 1 in at least three samples remained. The normalized DGEList object was then used to obtain log2-counts-per-million values using the limma voom() function.



Supplementary Figure S3. **RNA-Seq samples multidimensional scaling plot.** Visual representation of the three groups of treated HT-29 cells: 0 µM (aza0), 5 µM (aza5), and 10 µM (aza10) of 5-aza-deoxy-cytidine, after processing, filtering, and trimmed mean of M values (TMM) normalization.

Supplementary Table S1. Package and main dependencies versions.

Tool	Version
transcriptogramer	1.1.21
limma	3.34.9
topGO	2.30.1

Supplementary Table S2. Count-based protocol software versions.

Software	Version
R	3.4.3
tophat2	2.1.1
bowtie2	2.3.4.1
HTSeq	0.9.1
SRA toolkit	2.8.2-1
edgeR	3.20.9

Supplementary Table S3. Differential expression results.

Description	Pipeline	Platform	aza5 x aza0	aza10 x aza0
Number of DEGs[1]	limma-topGO	Microarray	6326	4139
Number of DEGs[1]	limma-topGO	RNA-Seq	6188	4238
Number of DEGs[1]	transcriptogramer	Microarray	6815	2274
Number of DEGs[1]	transcriptogramer	RNA-Seq	6375	3164

[1]Differentially Expressed Genes

Supplementary Table S4. Gene Ontology enrichment analysis.

Description	Pipeline	Platform	aza5 x aza0	aza10 x aza0
Number of unique terms	limma-topGO	Microarray	247	142
Number of unique terms	limma-topGO	RNA-Seq	54	117
Number of unique terms	transcriptogramer	Microarray	3996	2460
Number of unique terms	transcriptogramer	RNA-Seq	2807	2801

Supplementary Table S5. Top 5 terms transcriptogramer Microarray.

	transcriptogramer Microarray			transcriptogramer RNA-Seq		limma-topGO Microarray		limma-topGO RNA-Seq	
Contrast	**Top terms**	**Position**	**Rate[1]**	**Position**	**Rate[1]**	**Position**	**Rate[1]**	**Position**	**Rate[1]**
aza5 x aza0	Chromosome organization	1	0.54	8	0.714	-	-	-	-
	DNA metabolic process	2	0.507	29	0.687	98	0.257	-	-
	Protein modification by small protein conjugation or removal	3	0.472	1	0.76	-	-	-	-
	DNA repair	4	0.65	39	0.761	-	-	-	-
	Cellular response to DNA damage stimulus	5	0.524	28	0.712	-	-	-	-
aza10 x aza0	Transmembrane receptor protein tyrosine kinase signalling pathway	1	0.301	139	0.157	-	-	-	-
	Enzyme linked receptor protein signalling pathway	2	0.22	140	0.123	-	-	-	-
	Endocytosis	3	0.245	141	0.157	-	-	-	-
	Import into cell	4	0.232	143	0.148	-	-	-	-
	Cell surface receptor signalling pathway	5	0.108	105	0.123	-	-	65	0.389

[1]Significant/Annotate

Supplementary Table S6. Top 5 terms transcriptogramer RNA-Seq.

	transcriptogramer RNA-Seq			transcriptogramer Microarray		limma-topGO Microarray		limma-topGO RNA-Seq	
Contrast	**Top terms**	**Position**	**Rate[1]**	**Position**	**Rate[1]**	**Position**	**Rate[1]**	**Position**	**Rate[1]**
aza5 x aza0	Protein modification by small protein conjugation or removal	1	0.76	3	0.472	-	-	-	-
	Cellular protein modification process	2	0.567	68	0.227	-	-	-	-
	Protein modification process	3	0.567	69	0.227	-	-	-	-
	Macromolecule modification	4	0.551	30	0.228	-	-	-	-
	Cellular response to stress	5	0.647	12	0.339	40	0.607	-	-
aza10 x aza0	Chromosome organization	1	0.431	76	0.153	-	-	-	-
	Chromatin organization	2	0.513	106	0.12	-	-	-	-
	DNA metabolic process	3	0.418	74	0.2	45	0.167	-	-
	DNA repair	4	0.55	73	0.31	-	-	-	-
	Nucleic acid metabolic process	5	0.194	78	0.062	63	0.146	-	-

[1]Significant/Annotated

Supplementary Table S7. Top 5 terms limma-topGO Microarray.

	limma-topGO Microarray			limma-topGO RNA-Seq		transcriptogramer Microarray		transcriptogramer RNA-Seq	
Contrast	**Top terms**	**Position**	**Rate[1]**	**Position**	**Rate[1]**	**Position**	**Rate[1]**	**Position**	**Rate[1]**
aza5 x aza0	Oxidation-reduction process	1	0.292	-	-	338	0.156	925	0.346
	Cellular metabolic process	2	0.235	-	-	26	0.169	866	0.174
	Small molecule metabolic process	3	0.26	-	-	341	0.108	384	0.348
	Small molecule biosynthetic process	4	0.297	-	-	1406	0.117	1760	0.339
	Metabolic process	5	0.233	-	-	62	0.162	1008	0.172
aza10 x aza0	Cell cycle	1	0.173	-	-	95	0.085	193	0.184
	Regulation of cell cycle	2	0.181	-	-	412	0.074	754	0.162
	Cellular metabolic process	3	0.144	-	-	268	0.032	86	0.117
	Cell cycle process	4	0.176	-	-	99	0.093	274	0.19
	Metabolic process	5	0.143	-	-	314	0.031	168	0.114

[1]Significant/Annotated

Supplementary Table S8. Top 5 terms limma-topGO RNA-Seq.

	limma-topGO RNA-Seq			limma-topGO Microarray		transcriptogramer Microarray		transcriptogramer RNA-Seq	
Contrast	**Top terms**	**Position**	**Rate[1]**	**Position**	**Rate[1]**	**Position**	**Rate[1]**	**Position**	**Rate[1]**
aza5 x aza0	SRP-dependent cotranslational protein targeting to membrane	1	0.876	-	-	229	0.977	187	0.977
	Cotranslational protei targeting to membrane	2	0.851	-	-	223	0.978	184	0.978
	Protein targeting to ER	3	0.814	-	-	228	0.948	188	0.946
	Establishment of protein localization to endoplasmic reticulum	4	0.8	-	-	233	0.91	191	0.917
	Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	5	0.778	-	-	234	0.857	192	0.845
aza10 x aza0	Cotranslational protein targeting to membrane	1	0.755	67	0.229	130	0.505	-	-
	SRP-dependent cotranslational protein targeting to membrane	2	0.764	-	-	133	0.489	-	-
	Protein targeting to ER	3	0.711	122	0.215	132	0.469	-	-
	Protein targeting to membrane	4	0.645	48	0.215	137	0.324	-	-
	Establishment of protein localization to endoplasmic reticulum	5	0.7	-	-	134	0.45	-	-

[1]Significant/Annotated

Supplementary Note #1: transcriptogramer and limma-topGO pipelines

Transcriptograms are produced by assigning to each gene the average expression level of its neighbors, according to a window of a given radius. Here, radius = 125 was chosen for all experiments using transcriptogramer package and protein-protein interaction (PPI) data was obtained from STRINGdb 10.5 (PPI of combined score greater than or equal to 800). The number of differentially expressed genes (FDR < 0.05) and unique Gene Ontology biological processes terms (FDR < 0.05) obtained on each experiment were annotated (Supplementary table S3 and S4, respectively). Each Differentially Expressed Gene (DEG) counted in transcriptogramer pipeline represents one window central gene, whose expression level denotes the average expression of its neighbors within the window. A group of DEGs, separated by less gene positions than the radius value, represents a Differentially Expressed Gene Group (DEGG).

Gene Ontology enrichment analysis for both pipelines was performed using topGO classic algorithm, not considering Gene Ontology hierarchy. The transcriptogramer pipeline enriched each DEGG individually and limma-topGO pipeline enriched all DEGs collectively.

Supplementary Note #2: Pipeline results comparison

The number of differentially expressed genes are similar when comparing limma-topGO and transcriptogramer, including Microarray and RNA-Seq results (Supplementary table S3). For Gene Ontology enrichment analysis, the number of unique terms obtained by transcriptogramer pipeline was more than 10 times higher than those obtained by limma-topGO pipeline, indicating that transcriptogramer is more sensitive to identify alterations in cellular systems. Supplementary table S4 to S8 shows that transcriptogramer is able to identify almost all GO terms identified by limma-topGO for both Microarray and RNA-Seq. The supplementary tables show that transcriptogramer GO enrichment analysis identifies more terms. Some of those terms appears in limma-topGO results, but many are uniquely found using transcriptogramer GO enrichment.