Tropical Plant Biology manuscript No.
(will be inserted by the editor)

Sampling strategies for sugarcane using either clonal replicates or diverse genotypes can bias the conclusions of RNA-seq studies

Victor Hugo Mello[1] · **Ana Letycia Basso Garcia**[1] · **Fernando Henrique Correr**[1] · **Guilherme Kenichi Hosaka**[1] · **Monalisa Sampaio Carneiro**[2] · **Gabriel Rodrigues Alves Margarido**[1]

Received: date / Accepted: date

[1] Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Avenida Páadua Dias, 11, Piracicaba, SP 13418-900, Brazil; [2] Departamento de Biotecnologia e Produção Vegetal e Animal, Centro de Ciências Agrárias, Universidade Federal de São Carlos, Rodovia Washington Luis km235, Araras, SP 13565-905, Brazil

G. R. A. Margarido
E-mail: gramarga@usp.br

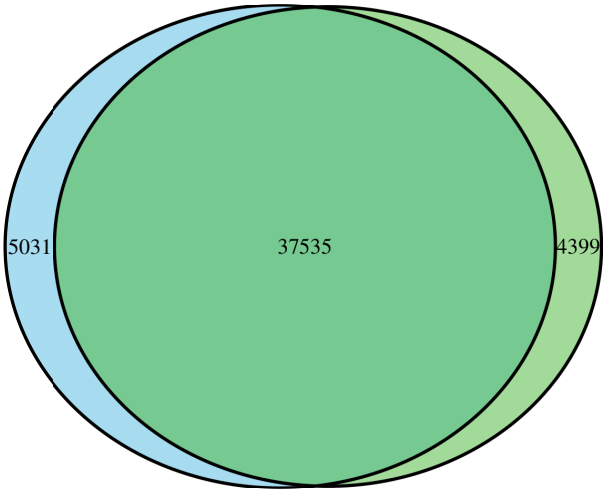

Fig. 1 Number of genes kept after filtering out lowly expressed genes. The strategy based on clones (blue) had 42,566 genes analyzed for differential gene expression, while the strategy based on diverse genotypes (green) had 41,934. 11.8% of SBC and 10.5% of SBDG gene datasets were exclusive.

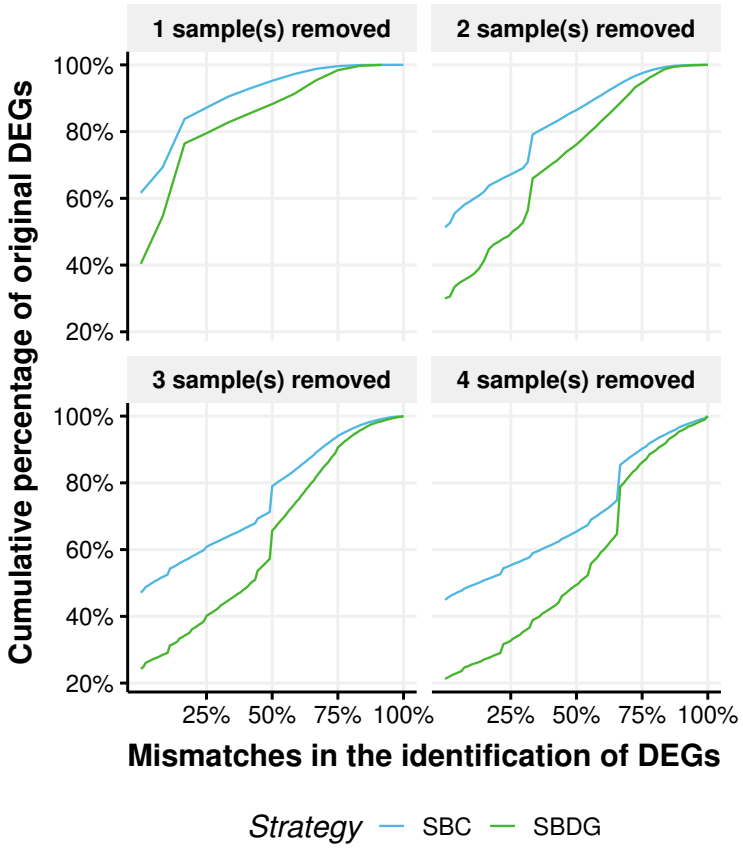

Fig. 2 Cumulative distribution of identified mismatches. The total number of combinations of subsets of samples are: $n_1 = 12$, $n_2 = 54$, $n_3 = 108$, and $n_4 = 81$, in which the index i represents the number of removed samples. Thus, the original DEGs and the subsamples may match from 0 to n_i times. We consider a match when a given iteration has the same differential expression result for a gene. Hence, the curves indicate the percentage of DEGs which showed the minimum mismatch rate in the x-axis. For example, with a single sample removed for the SBDG (top left panel), 40% of genes did not have a single mismatch for all twelve combinations, and 80% of genes presented a maximum mismatch rate of 25%.

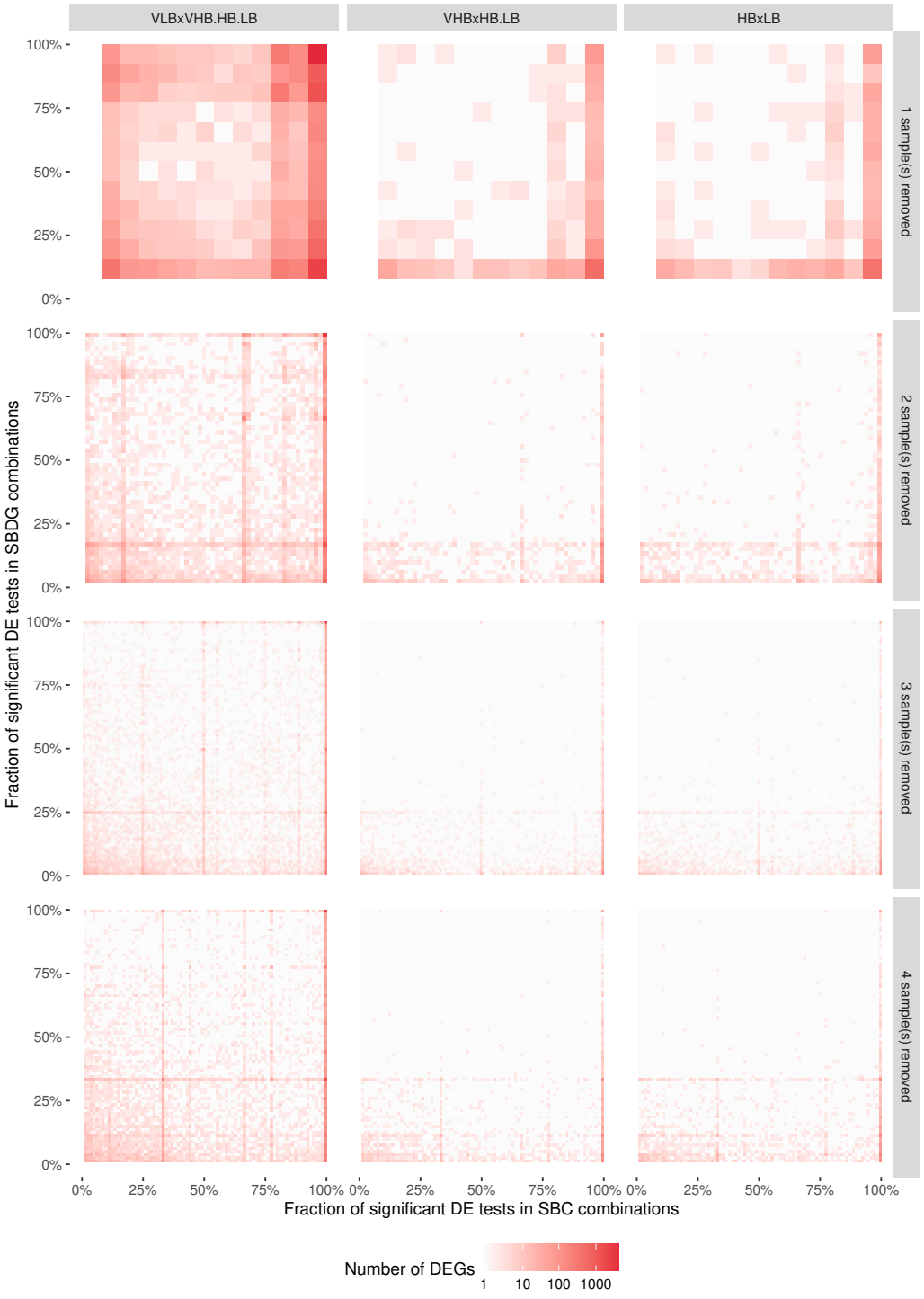

Fig. 3 Correspondence between differentially expressed genes found by SBC and SBDG per contrast and number of removed samples. The number of combinations for i removed sample(s) is n_i, such that each gene ranges from 0 to n_i chances to be called as differentially expressed. The heatmap shows the number of DEGs identified by both strategies, for at least one combination of samples, normalized by n_i.

Table 1 Processing of sequencing data. Raw data represents the number of reads after downsampling of the SBC dataset. Quality control consisted of trimming and removing low quality paired reads. Mapping rates indicate alignment of reads after quality control against the *de novo* assembled transcriptome. SBC - Strategy based on clones; SBDG - Strategy based on diverse genotypes.

Sample	Raw reads	Reads after quality control	Mapping rate to transcriptome
SBC			
SP80-3280 R1	43966734	40114252	75.12%
SP80-3290 R2	37891704	34496514	75.46%
SP80-3280 R3	41403168	37820702	75.03%
R570 R1	41005116	37893862	84.22%
R570 R2	44286020	40690986	75.20%
R570 R3	40290944	37096532	79.19%
F36-819 R1	43363828	39809118	76.32%
F36-819 R2	46032304	42445412	76.61%
F36-819 R3	42181506	38925260	78.64%
IN84-58 R1	48151512	44359660	75.04%
IN84-58 R2	48957312	45338270	79.26%
IN84-58 R3	42272278	38980420	81.70%
Mean	42818054	39394770	76.47%
SBDG			
SP80-3280	50481854	46337774	75.12%
White Mauritius	41749608	38642584	75.86%
RB835486	38829088	35247810	77.62%
R570	45443304	42257284	84.27%
RB92579	45191582	41952538	76.33%
White Transparent	37865194	35088926	76.15%
F36-819	45598224	42274354	76.62%
Criolla Rayada	40096612	40093304	76.30%
IJ76-317	20048306	37184558	76.71%
IN84-58	43422964	40264200	81.70%
Krakatau	40349586	37372702	76.92%
SES205A	43589964	40383156	76.54%
Mean	43323910	40178752	76.58%

Table 2 Summary statistics of the *de novo* transciptome assembly and its annotation with Gene Ontology terms. The terms here expressed as *gene* and *transcript* refer to the output of Trinity. Therefore, *gene* comprehends a set of transcripts clustered together in a common de Bruijn graph, sharing sequence identity.

Assembly statistics	**Transcriptome**	**Genes kept in SBC**	**Genes kept in SBDG**
Number of genes	262281	42566	41934
Number of transcripts	598874	262099	261547
Genes with 1 transcript	64.3%	13.5%	12.8%
Genes with 2 transcripts	13.7%	11.8%	11.4%
Genes with 3 transcripts	5.9%	9.3%	9.3%
Genes with >4 transcripts	16.1%	65.4%	66.5%
Mean transcript length	932.63 nt	1573.62 nt	1567.19 nt
Median transcript length	495 nt	1239 nt	1230 nt
Contig N10	4295 nt	-	-
Contig N20	3277 nt	-	-
Contig N30	2618 nt	-	-
Contig N40	2124 nt	-	-
Contig N50	1687 nt	-	-
Total assembled bases	558526632 nt	-	-
Number of identified GO terms	1187	1169	1164
Number of annotated genes in the transcriptome	18667	12364	11979

File 1 Differential gene expression results. The table contains the outcome of the likelihood ratio test after p-value correction for all contrasts in both strategies. The columns *logCPM* and *logFC* contain the average expression levels among all the samples included in the contrast and the fold-change of the comparison, respectively, both in log scale 2. *DE* summarizes the final result assigned by the test as not significant, upregulated, or downregulated (NotSig, Up, or Down).

File 2 List of annotated genes in the transcriptome. The table shows Gene Ontology terms with a brief description of their function and the list of assembled genes associated with them. The column *Ontology* groups each term by the categories molecular function, biological process, or cell compartment.

File 3 Gene Ontology enrichment results. The test evaluates whether a GO category is overrepresented in the set of differentially expressed genes against the full transcriptome annotation. The p-values were corrected using the false discovery rate (FDR) method. *numDEInCat* and *numInCat* represent the number of genes for each category in the set of differentially expressed genes and the whole transcriptome, respectively.

File 4 Gene Ontology enrichment for the high confidence set. We selected a group of genes with the same result for the differential expression test in more than 95% of the sample removal iterations. GO categories changed not only in the number of genes but also in proportion when compared to the original data. The structure of the file matches the description given for **File 3**.