Supplementary Note for the paper

Frattini et al "A metabolic function associated with FGFR3-TACC3 gene fusions", Nature 2017.

Evaluation of resampling techniques for ranked list generation in imbalanced datasets

The highly imbalanced ratio between FGFR3-TACC3-fusion positive and negative samples (9 and 535, respectively) introduces severe limitations in identifying differentially expressed genes that can be used in enrichment analyses. Any test statistic aimed at detecting gene-wise differential expression level recapitulates the differences between two groups of samples[1]. However, the specific method of enrichment analysis can affect the interpretation of results. Therefore, we performed an extensive set of experiments on simulated data to compare accuracy and sensitivity of detection of significant gene sets in a ranked list of genes in order to validate the method adopted in the paper. The basic resampling techniques for imbalanced data can be divided in two main classes: under-sampling of the majority class or *easy ensemble* and oversampling of the minority class or synthetic minority over-sampling technique (SMOTE)[2]. We used several methods to detect the gene-wise differential expression, and tested the two main approaches to perform resampling.

To comparatively evaluate the performance of the test statistics as the metrics to summarize the functional differences between the minority set and the majority set, we considered the GSEA method (KS-GSEA[3]) and the Mann-Whitney-Wilcoxon test (MWW-GST), grouping the genes in the gene set versus those outside, as reported in the Methods section of the

manuscript (Table 1). We considered CAMERA[1] which performs both differential expression and enrichment analysis.

<table>
<tr><th>Resampling method</th><th>Differential Expression statistics</th><th>Enrichment Analysis</th></tr>
<tr><td rowspan="2">Undersampling of the minority class (easy ensemble)
Oversampling of the minority class (SMOTE)</td><td>t-test
Mann-Whitney-Wilcoxon (MWW)
Moderated t-test[4] (VOOM[5])
edgeR[6] (exact)
edgeR (glm)</td><td>Kolmogorov Smirnov Gene Set Enrichment Analysis[3]
MWW Gene Set Enrichment Analysis[7]</td></tr>
<tr><td colspan="2">CAMERA[1]</td></tr>
</table>

Table 1. Tested methods for ranked list generation and enrichment analysis in imbalanced datasets.

We considered a specific scenario of highly imbalanced data. If m is the number of samples associated with a phenotype of interest, and n is the total number of samples, then the ratio $\frac{n-m}{n}$ is a measure of imbalance between the two groups of samples. In imbalanced datasets, $m \ll n$. The subset of the m samples is called the minority set, whereas the other $n-m$ samples define the majority set. In our simulations, we explore various levels of imbalance. We generated a controlled dataset (see below) enriched for known biological classes with different levels of imbalance and for each combination of resampling method, differential expression statistics and enrichment analysis, we computed a measure of accuracy based on the Area Under the ROC (AUC).

Synthetic data generation

We generated a standard RNA-Seq profile by averaging 10,411 TCGA samples from 33 tumor types. This averaged profile defines an empirical distribution function that we assume as the theoretical model. We generated a data matrix X of n samples by first assigning to each sample the same standard expression profile. Then the counts, x_{ij}, in the matrix are independently perturbed as $y_{ij} = x_{ij}(1 + 2u)$, where u is a random value drawn by a uniform random variable on the interval $(0,1)$, $i = 1, \ldots, n \;\; j = 1, \ldots, p$ where p is the number of genes $(p = 20{,}072)$. In this perturbation, any y_{ij} value is increased by a percentage of counts ranging from 0% to 200%. Within the new matrix, Y, we then increase the level of a subset of genes included in selected collections of MSigDB hallmark gene sets: *spermatogenesis, HEME metabolism, apical surface, pancreas beta cells, KRAS signaling down, apoptosis, fatty acid metabolism, apical junction, NOTCH signaling, DNA repair.* We refer to these ten gene sets as "true positive" gene sets. This was an ideal set of categories since they have different sizes, some gene sets overlap each other, whereas others have no genes in common.

To increase the count values of the genes in the true positive gene set, we draw a uniform random value v in $(0, \alpha)$ and replace each gene count y_{ij} by $y_{ij}(1 + v)$. In other words, we increase the counts of the genes in the true positive gene sets by a percentage between 0 and $100\alpha\%$. We chose $\alpha = 0.1$ and $\alpha = 0.05$.

Results

Having fixed $n = 500$, we performed six experiments corresponding to $m = 3$, 5, 10, $\alpha = 0.05$ and $\alpha = 0.1$ for each of the 22 combinations of Table 1. Each experiment was run 50 times and the AUC values were collected. The final value of the performance index is the average of the AUC across the 50 experiments. The ranked gene list obtained from each experiment was then analyzed by the two enrichment methods (KS-GSEA and MWW-GST). In order to remove the bias of the different metrics that are generated by the two differential expression tests, which can ultimately influence KS-GSEA, we transformed any ranked list as in Barbie *et al.*[8]. The computation of the ranked list based on the easy ensemble method consists of 200 random under-sampling of the majority set. We implemented each methods according to the author's description. The results for each of the six experiments are reported in Table 2. According to Kendall's tau test of agreement of the rankings[9], the level of agreement among the six rankings is 95%; the null hypothesis of no agreement is rejected with a *p*-value $< 10^{-6}$.

Table 2a. $m = 3, \alpha = 0.05$

Minority set size	Sampling	Test method	GSEA method	AUC	sd	rank
3	SMOTE	MWW	MWW-GST	0.7911	0.0870	1
3	easy ensemble	MWW	MWW-GST	0.7882	0.0857	2
3	easy ensemble	VOOM	MWW-GST	0.7744	0.0911	3
3	SMOTE	tTest	MWW-GST	0.7707	0.0907	4
3	SMOTE	MWW	KS-GSEA	0.7681	0.0828	5
3	easy ensemble	tTest	MWW-GST	0.7630	0.0924	6
3	easy ensemble	MWW	KS-GSEA	0.7584	0.0833	7
3	SMOTE	tTest	KS-GSEA	0.7459	0.0922	8
3	easy ensemble	CAMERA	CAMERA	0.7446	0.0872	9
3	SMOTE	CAMERA	CAMERA	0.7426	0.0913	10

3	easy ensemble	VOOM	KS-GSEA	0.7413	0.0920	11
3	easy ensemble	tTest	KS-GSEA	0.7378	0.0880	12
3	SMOTE	VOOM	MWW-GST	0.6921	0.0945	13
3	easy ensemble	edgeR (GLM)	MWW-GST	0.6792	0.0481	14
3	SMOTE	VOOM	KS-GSEA	0.6723	0.1022	15
3	SMOTE	edgeR (exact)	MWW-GST	0.6116	0.1047	16
3	SMOTE	edgeR (exact)	KS-GSEA	0.5889	0.0806	17
3	easy ensemble	edgeR (GLM)	KS-GSEA	0.4901	0.0640	18
3	easy ensemble	edgeR (exact)	KS-GSEA	0.4063	0.1079	19
3	SMOTE	edgeR (GLM)	MWW-GST	0.3901	0.0752	20
3	SMOTE	edgeR (GLM)	KS-GSEA	0.3734	0.0685	21
3	easy ensemble	edgeR (exact)	MWW-GST	0.2898	0.0776	22

Table 2b. $m = 3, \alpha = 0.10$

Minority set size	Sampling	Test method	GSEA method	AUC	sd	rank
3	SMOTE	MWW	MWW-GST	0.8561	0.0830	1
3	easy ensemble	MWW	MWW-GST	0.8509	0.0841	2
3	easy ensemble	VOOM	MWW-GST	0.8446	0.0853	3
3	easy ensemble	CAMERA	CAMERA	0.8419	0.0994	4
3	SMOTE	tTest	MWW-GST	0.8418	0.0868	5
3	easy ensemble	tTest	MWW-GST	0.8401	0.0880	6
3	SMOTE	CAMERA	CAMERA	0.8359	0.1002	7
3	SMOTE	MWW	KS-GSEA	0.8328	0.0728	8
3	easy ensemble	MWW	KS-GSEA	0.8273	0.0790	9
3	easy ensemble	tTest	KS-GSEA	0.8186	0.0815	10
3	easy ensemble	VOOM	KS-GSEA	0.8186	0.0798	11
3	SMOTE	tTest	KS-GSEA	0.8172	0.0783	12

3	SMOTE	VOOM	MWW-GST	0.7641	0.0857	13
3	SMOTE	VOOM	KS-GSEA	0.7583	0.0900	14
3	easy ensembl	edgeR (GLM)	MWW-GST	0.6874	0.0536	15
3	SMOTE	edgeR (exact)	KS-GSEA	0.5854	0.1048	16
3	SMOTE	edgeR (exact)	MWW-GST	0.5639	0.0921	17
3	easy ensemble	edgeR (GLM)	KS-GSEA	0.4646	0.0721	18
3	SMOTE	edgeR (GLM)	MWW-GST	0.4399	0.0776	19
3	SMOTE	edgeR (GLM)	KS-GSEA	0.4357	0.0747	20
3	easy ensemble	edgeR (exact)	KS-GSEA	0.4274	0.1286	21
3	easy ensemble	edgeR (exact)	MWW-GST	0.2512	0.0838	22

Table 2c. $m = 5, \alpha = 0.05$

Minority set size	Sampling	Test method	GSEA method	AUC	sd	rank
5	SMOTE	MWW	MWW-GST	0.8102	0.1113	1
5	easy ensemble	MWW	MWW-GST	0.8060	0.1111	2
5	easy ensemble	VOOM	MWW-GST	0.8013	0.1144	3
5	easy ensemble	tTest	MWW-GST	0.7956	0.1131	4
5	SMOTE	tTest	MWW-GST	0.7927	0.1162	5
5	SMOTE	MWW	KS-GSEA	0.7885	0.1075	6
5	easy ensemble	VOOM	KS-GSEA	0.7839	0.1054	7
5	easy ensemble	tTest	KS-GSEA	0.7780	0.1053	8
5	easy ensemble	MWW	KS-GSEA	0.7770	0.1097	9
5	easy ensemble	CAMERA	CAMERA	0.7763	0.0932	10
5	SMOTE	CAMERA	CAMERA	0.7734	0.0984	11
5	SMOTE	tTest	KS-GSEA	0.7726	0.1093	12
5	SMOTE	VOOM	KS-GSEA	0.7009	0.1137	13
5	SMOTE	VOOM	MWW-GST	0.6877	0.1130	14
5	easy ensemble	edgeR (GLM)	MWW-GST	0.6825	0.0408	15
5	SMOTE	edgeR (exact)	MWW-GST	0.5787	0.1011	16
5	SMOTE	edgeR (exact)	KS-GSEA	0.5432	0.1208	17

5	easy ensemble	edgeR (GLM)	KS-GSEA	0.4781	0.0637	18
5	SMOTE	edgeR (GLM)	KS-GSEA	0.4403	0.0824	19
5	SMOTE	edgeR (GLM)	MWW-GST	0.4230	0.0742	20
5	easy ensemble	edgeR (exact)	KS-GSEA	0.4060	0.0856	21
5	easy ensemble	edgeR (exact)	MWW-GST	0.2734	0.0578	22

Table 2d. $m = 5, \alpha = 0.10$

Minority set size	Sampling	Test method	GSEA method	AUC	sd	rank
5	easy ensemble	CAMERA	CAMERA	0.8972	0.0651	1
5	SMOTE	MWW	MWW-GST	0.8896	0.0600	2
5	easy ensemble	MWW	MWW-GST	0.8893	0.0590	3
5	easy ensemble	tTest	MWW-GST	0.8882	0.0604	4
5	easy ensemble	VOOM	MWW-GST	0.8869	0.0586	5
5	SMOTE	CAMERA	CAMERA	0.8834	0.0620	6
5	SMOTE	tTest	MWW-GST	0.8789	0.0589	7
5	SMOTE	MWW	KS-GSEA	0.8760	0.0633	8
5	easy ensemble	MWW	KS-GSEA	0.8745	0.0648	9
5	easy ensemble	VOOM	KS-GSEA	0.8680	0.0689	10
5	easy ensemble	tTest	KS-GSEA	0.8653	0.0697	11
5	SMOTE	tTest	KS-GSEA	0.8647	0.0643	12
5	SMOTE	VOOM	KS-GSEA	0.8006	0.0698	13
5	SMOTE	VOOM	MWW-GST	0.7792	0.0719	14
5	easy ensemble	edgeR (GLM)	MWW-GST	0.7150	0.0597	15
5	SMOTE	edgeR (exact)	KS-GSEA	0.5971	0.1136	16
5	SMOTE	edgeR (GLM)	KS-GSEA	0.5093	0.1000	17
5	SMOTE	edgeR (exact)	MWW-GST	0.5073	0.0908	18
5	SMOTE	edgeR (GLM)	MWW-GST	0.5048	0.0951	19
5	easy ensemble	edgeR (GLM)	KS-GSEA	0.4636	0.0639	20
5	easy ensemble	edgeR (exact)	KS-GSEA	0.3938	0.1040	21

5	easy ensemble	edgeR (exact)	MWW-GST	0.2171	0.0788	22

Table 2e. $m = 10, \alpha = 0.05$

Minority set size	Sampling	Test method	GSEA method	AUC	sd	rank
10	SMOTE	MWW	MWW-GST	0.8717	0.0623	1
10	easy ensemble	MWW	MWW-GST	0.8663	0.0629	2
10	easy ensemble	VOOM	MWW-GST	0.8617	0.0660	3
10	easy ensemble	tTest	MWW-GST	0.8593	0.0693	4
10	SMOTE	tTest	MWW-GST	0.8563	0.0660	5
10	easy ensemble	MWW	KS-GSEA	0.8501	0.0618	6
10	easy ensemble	CAMERA	CAMERA	0.8496	0.0757	7
10	SMOTE	CAMERA	CAMERA	0.8442	0.0721	8
10	SMOTE	MWW	KS-GSEA	0.8429	0.0693	9
10	easy ensemble	VOOM	KS-GSEA	0.8412	0.0703	10
10	easy ensemble	tTest	KS-GSEA	0.8405	0.0708	11
10	SMOTE	tTest	KS-GSEA	0.8271	0.0698	12
10	SMOTE	VOOM	KS-GSEA	0.7367	0.0781	13
10	SMOTE	VOOM	MWW-GST	0.7072	0.0772	14
10	easy ensemble	edgeR (GLM)	MWW-GST	0.6938	0.0438	15
10	SMOTE	edgeR (exact)	KS-GSEA	0.6243	0.0913	16
10	SMOTE	edgeR (exact)	MWW-GST	0.5675	0.0924	17
10	easy ensemble	edgeR (GLM)	KS-GSEA	0.4653	0.0632	18
10	SMOTE	edgeR (GLM)	MWW-GST	0.4447	0.0969	19
10	SMOTE	edgeR (GLM)	KS-GSEA	0.4348	0.0930	20
10	easy ensemble	edgeR (exact)	KS-GSEA	0.3819	0.0827	21
10	easy ensemble	edgeR (exact)	MWW-GST	0.2452	0.0535	22

Table 2f. $m = 10, \alpha = 0.10$

Minority set size	Sampling	Test method	GSEA method	AUC	sd	rank
10	easy ensemble	CAMERA	CAMERA	0.8887	0.0798	1
10	SMOTE	MWW	MWW-GST	0.8834	0.0697	2
10	easy ensemble	MWW	MWW-GST	0.8819	0.0723	3
10	easy ensemble	tTest	MWW-GST	0.8797	0.0727	4
10	easy ensemble	VOOM	MWW-GST	0.8766	0.0717	5
10	SMOTE	MWW	KS-GSEA	0.8762	0.0661	6
10	easy ensemble	tTest	KS-GSEA	0.8755	0.0670	7
10	SMOTE	tTest	MWW-GST	0.8724	0.0697	8
10	easy ensemble	VOOM	KS-GSEA	0.8712	0.0621	9
10	SMOTE	CAMERA	CAMERA	0.8698	0.0797	10
10	easy ensemble	MWW	KS-GSEA	0.8659	0.0702	11
10	SMOTE	tTest	KS-GSEA	0.8629	0.0685	12
10	easy ensemble	edgeR (GLM)	MWW-GST	0.7790	0.0531	13
10	SMOTE	VOOM	KS-GSEA	0.7709	0.0751	14
10	SMOTE	VOOM	MWW-GST	0.7088	0.0704	15
10	SMOTE	edgeR (exact)	KS-GSEA	0.6651	0.0960	16
10	SMOTE	edgeR (GLM)	KS-GSEA	0.5825	0.0857	17
10	SMOTE	edgeR (GLM)	MWW-GST	0.5741	0.0831	18
10	SMOTE	edgeR (exact)	MWW-GST	0.4513	0.0891	19
10	easy ensemble	edgeR (GLM)	KS-GSEA	0.4500	0.0635	20
10	easy ensemble	edgeR (exact)	KS-GSEA	0.3907	0.0922	21
10	easy ensemble	edgeR (exact)	MWW-GST	0.1611	0.0673	22

Table 2. Results of simulation experiments for each setting of minority set m and expression increase percentage α. $n = 500$ in all experiments. **a,** $m = 3, \alpha = 0.05$ **b,** $m = 3, \alpha = 0.10$ **c,** $m = 5, \alpha = 0.05$ **d,** $m = 5, \alpha = 0.10$ **e,** $m = 10, \alpha = 0.05$ **f,** $m = 10, \alpha = 0.10$

A summary of the six rankings is reported in Table 3 where the score is the mean of the ranks from each sub-experiment (the higher, the better). According to our simulations, MWW-GST is the best performing method, followed by CAMERA. The top-ranking test methods are Mann-Whitney-Wilcoxon test and the VOOM procedure. Overall, the SMOTE oversampling and easy ensemble coupled with MWW differential expression seem to perform better than all other methods. When the configuration of the samples is less critical ($\alpha = 0.1$), the easy ensemble version of CAMERA has the best performance. On the other side, more difficult and imbalanced configurations ($m = 3$ or $\alpha = 0.05$) are dominated by the combination of the Mann-Whitney-Wilcoxon test (both SMOTE and easy ensemble) and the MWW-GST. In summary, the generation of a ranked list with both oversampling and undersampling based on MWW, *ee-MWW* and *SMOTE-MWW* respectively, show the best and most coherent performance among all the evaluation experiments on synthetic data. Therefore, we tested both methods on the GBM TCGA dataset to discover signatures that characterize the molecular functions associated with FGFR3-TACC3 fusion-positive samples.

Sampling method	**Test method**	**GSEA method**	**score**
SMOTE	MWW	MWW-GST	19.67
easy ensemble	MWW	MWW-GST	18.83
easy ensemble	VOOM	MWW-GST	18.00
easy ensemble	CAMERA	CAMERA	17.33
SMOTE	tTest	MWW-GST	16.83
easy ensemble	tTest	MWW-GST	16.50
SMOTE	CAMERA	CAMERA	15.17
SMOTE	MWW	KS-GSEA	14.67
easy ensemble	MWW	KS-GSEA	13.67

easy ensemble	tTest	KS-GSEA	12.83
easy ensemble	VOOM	KS-GSEA	12.00
SMOTE	tTest	KS-GSEA	11.83
SMOTE	VOOM	MWW-GST	10.33
SMOTE	VOOM	KS-GSEA	10.00
easy ensemble	edgeR (GLM)	MWW-GST	9.33
SMOTE	edgeR (exact)	KS-GSEA	8.00
SMOTE	edgeR (exact)	MWW-GST	6.50
easy ensemble	edgeR (GLM)	KS-GSEA	6.33
SMOTE	edgeR (GLM)	MWW-GST	5.33
SMOTE	edgeR (GLM)	KS-GSEA	5.17
easy ensemble	edgeR (exact)	KS-GSEA	3.67
easy ensemble	edgeR (exact)	MWW-GST	1.00

Table 3. Summary of the six sub-experiments in Table 2. The score is the mean of the ranks from the sub-experiments.

Evaluation of the stability of resampling methods

Due to the underlying random samplings, an important question is the stability of the ranked list when it is used to generate a signature for unsupervised clustering. To this aim, we generated 100 signatures through independent runs that were compared each other with the Kendall's tau correlation coefficient. The ee-MWW resulted in much more stable and coherent results across different runs compared with SMOTE-MWW (Figure 1a). The lower stability of SMOTE was evident even when the ranked lists were used to generate signatures by selecting the top and bottom extreme values (Figure 1b, c).

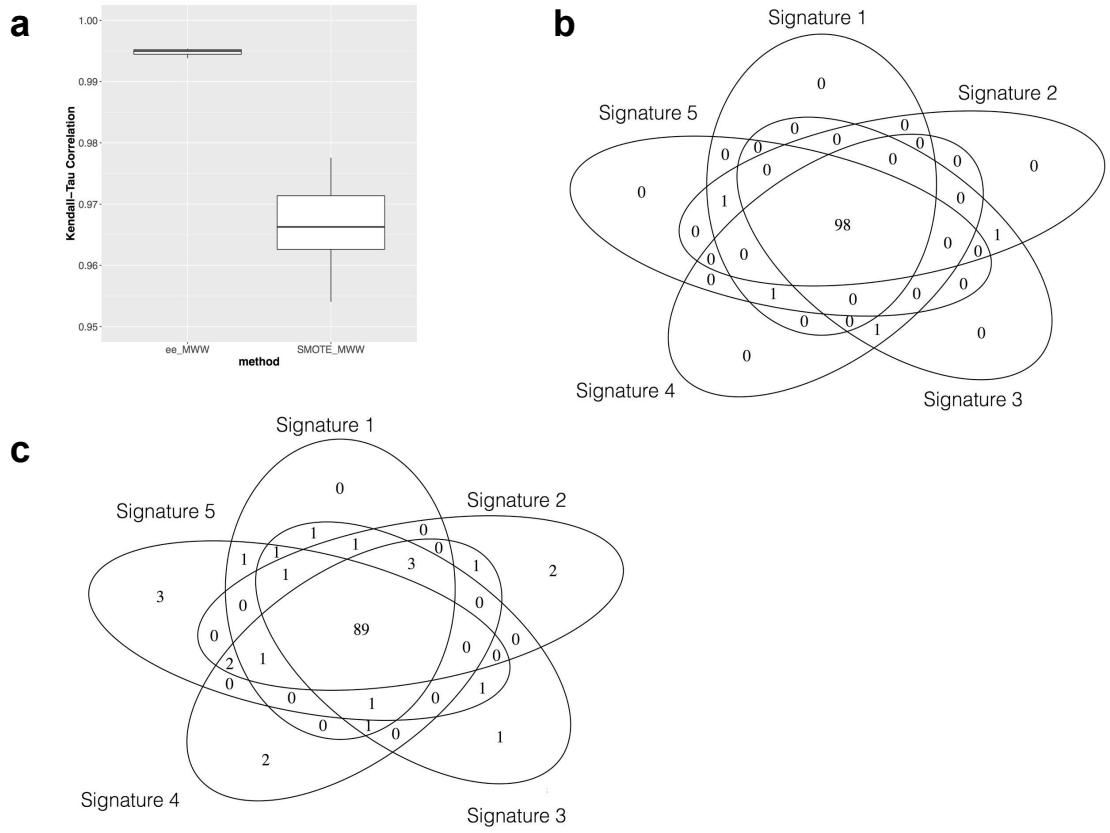

Figure 1. **a,** Boxplot of Kendall's tau rank correlation between lists generated by ee-MWW and SMOTE-MWW. **b-c,** Venn diagram of the overlap between five 100-genes signatures of generated by **b,** ee-MWW and **c,** SMOTE-MWW.

Figure 2 below reports four different clustering based on the signatures generated by SMOTE. The signatures are obtained selecting the top and bottom genes in the SMOTE generated ranked list. In different runs, the oversampling method generates different signatures producing sometimes a non-uniform cluster of the FGFR3-TACC3-positive samples (Figure 2a, b and d). This is due to random generation of synthetic examples used to perform the differential expression.

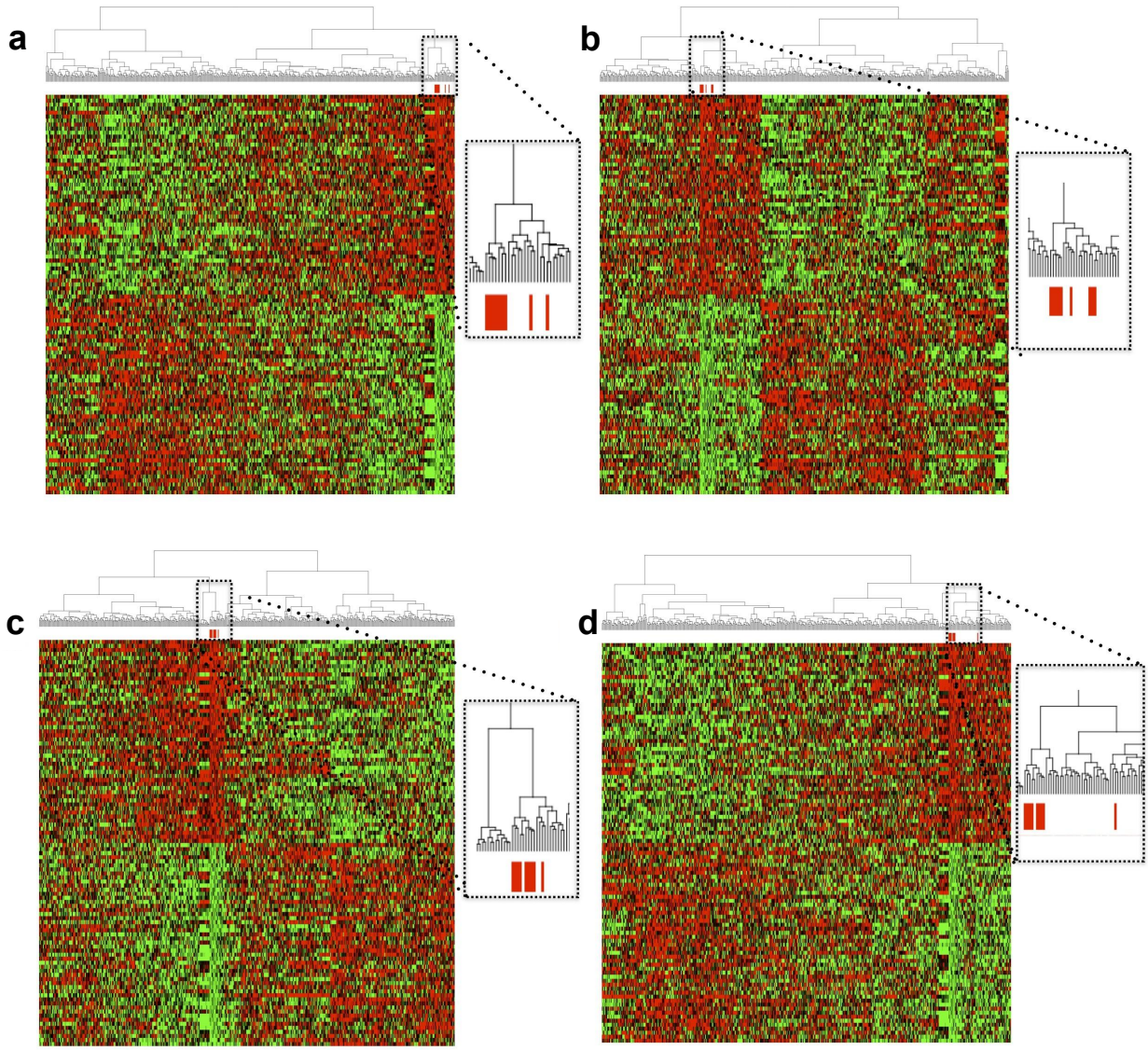

Figure 2. Four different clusters obtained using the signature generated by the SMOTE-MWW differential expression analysis on the TCGA microarray GBM dataset.

The same variability is observed when we analyzed the pan-glioma dataset (Figure 3). However, the ee-MWW resulted in much more reproducible and stable results, especially after increasing the number of resampling steps. Moreover, easy ensemble can handle extreme datasets that include very few examples of the minority class, such as the case of KRAS mutations ($m = 2$) reported in Extended Data Fig. 7a of the manuscript. On the basis of

these observed results, we adopted the easy ensemble method coupled with MWW differential expression analysis and MWW-GST to generate the enrichment analysis reported in the revised manuscript.

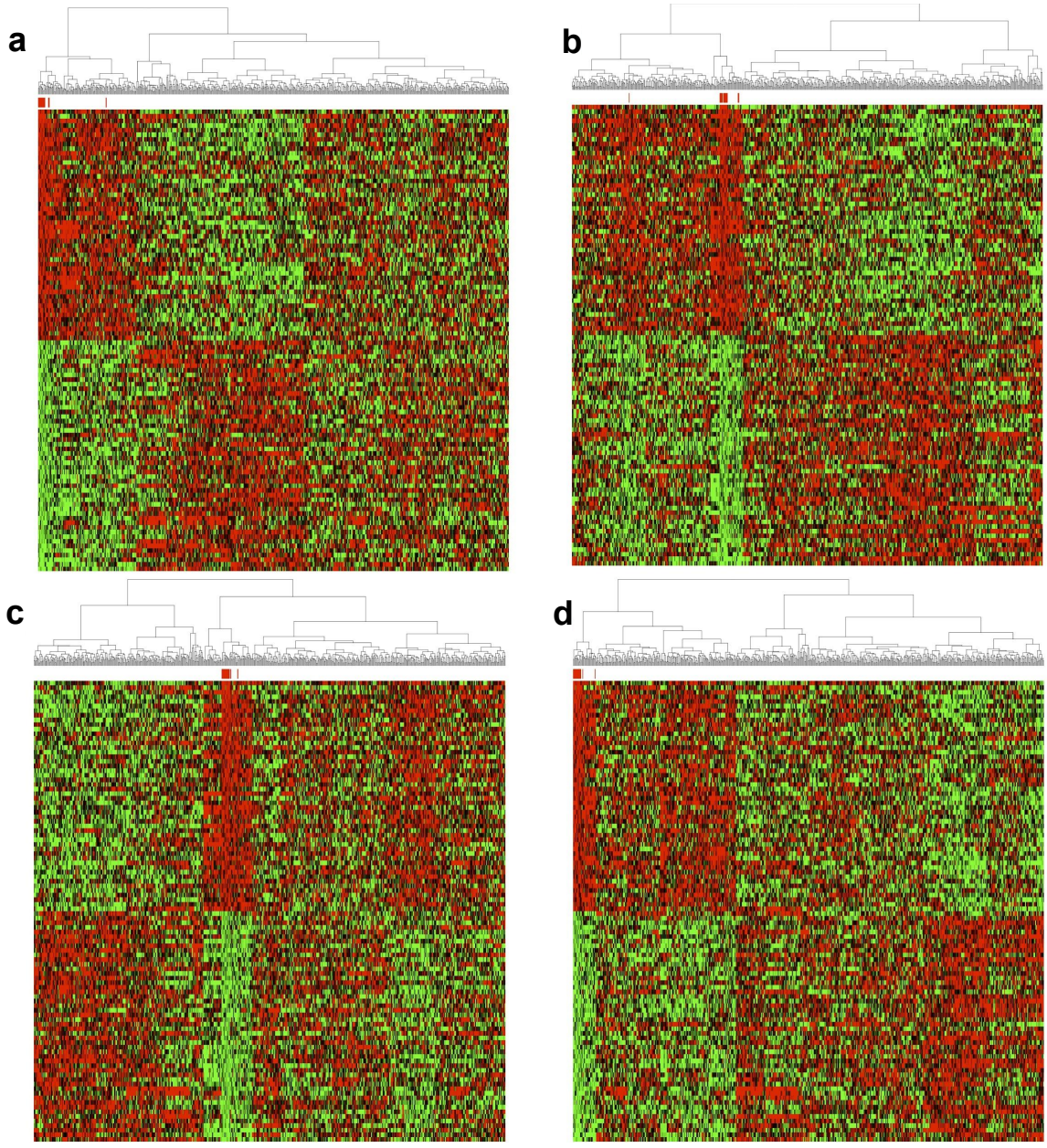

Figure 3. Four different clusters obtained using the signature generated by the SMOTE-MWW differential expression analysis on the TCGA pan-glioma dataset.

Literature Cited

1 Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res* **40**, e133 (2012).

2 He, H. B. & Garcia, E. A. Learning from Imbalanced Data. *Ieee Transactions on Knowledge and Data Engineering* **21**, 1263-1284 (2009).

3 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

4 Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).

5 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).

6 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

7 Michaud, J. *et al.* Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics* **9**, 363 (2008).

8 Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-112 (2009).

9 Kendall, M. G. Rank correlation methods. (1948).