

Using supervised learning methods for gene selection in RNA-Seq case-control studies

1 **Stephane Wenric^{1,2*}▪, Ruhollah Shemirani³▪**

2 ¹University of Liège, GIGA-Research, Laboratory of Human Genetics, Liège, Belgium

3 ²The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount
4 Sinai Hospital, New York, NY, USA

5 ³Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA

6 *** Correspondence:**

7 Stephane Wenric

8 stephane.wenric@mssm.edu

9 ▪ Both authors contributed equally to this work

10 **Keywords: RNA-Seq, supervised learning, random forests, variational autoencoders, gene**
11 **selection, feature selection, transcriptomics, gene expression**

12 **Abstract**

13 Whole transcriptome studies typically yield large amounts of data, with expression values for all genes
14 or transcripts of the genome. The search for genes of interest in a particular study setting can thus be a
15 daunting task, usually relying on automated computational methods. Moreover, most biological
16 questions imply that such a search should be performed in a multivariate setting, to take into account
17 the inter-genes relationships.

18 Differential expression analysis commonly yields large lists of genes deemed significant, even after
19 adjustment for multiple testing, making the subsequent study possibilities extensive.

20 Here, we explore the use of supervised learning methods to rank large ensembles of genes defined by
21 their expression values measured with RNA-Seq in a typical 2 classes sample set. First, we use one of
22 the variable importance measures generated by the random forests classification algorithm as a metric
23 to rank genes. Second, we define the EPS (extreme pseudo-samples) pipeline, making use of VAEs
24 (Variational Autoencoders) and regressors to extract a ranking of genes while leveraging the feature
25 space of both virtual and comparable samples.

26 We show that, on 12 cancer RNA-Seq data sets ranging from 323 to 1210 samples, using either a
27 random forests based gene selection method or the EPS pipeline outperforms differential expression
28 analysis for 9 and 8 out of the 12 datasets respectively, in terms of identifying subsets of genes
29 associated with survival.

30 These results demonstrate the potential of supervised learning-based gene selection methods in RNA-
31 Seq studies and highlight the need to use such multivariate gene selection methods alongside the widely
32 used differential expression analysis.

33

34 1 Introduction

35 Transcriptomics studies making use of RNA-Seq usually produce large amounts of data, namely one
36 expression value for each gene or transcript of each sample assessed [Wang2009, Mortazavi2008].

37 Searching for genes of interest or prioritizing genes in the context of case-control studies related to
38 diseases or other experimental conditions constitutes an important task ascribed to RNA-Seq
39 experiments [Trapnell2009, Garber2011, Love2014, Wenric2017].

40 Current methods often make use of differential expression analysis, to select genes of interest and
41 assign them a p-value related to a statistical test assessing changes in expression between different
42 conditions.

43 Most commonly used software packages performing differential expression analysis make use of the
44 negative binomial distribution to model read counts for each gene. This distribution, which is an
45 extension of the Poisson distribution, has two parameters: the mean and the dispersion, which allows
46 modeling of more general mean–variance relationships than Poisson. The dispersion parameter allows
47 to take into account the biological variability arising in RNA-Seq data [Love2014, Huang2015].

48 However, even though software packages like DESeq2 model relationships between genes by
49 assuming that genes of similar average expression have a similar dispersion, the statistical test
50 conducted to assess significance is a univariate test performed independently for each gene. Albeit
51 providing particularly useful and usually accurate information regarding disruptions of gene expression
52 between conditions, these methods thus do not take into account the potential correlation and
53 concordant or discordant effect between groups of genes. However, such gene-gene interactions are
54 present in most tissues and conditions and they are known to play key roles in said conditions, with
55 groups of genes which might have a significant effect as a group but not when each gene is considered
56 independently [Kanehisa2000, Joshitope2005, Phillips2008, Vidal2011].

57 Here, we explore the use of multivariate classifiers to rank genes in a case-control RNA-Seq
58 experiment. Namely, we're using the permutation importance of the random forests classifier to rank
59 genes, and a newly developed method (EPS) making use of Variational Autoencoders.

60 Machine learning methods are progressively being applied to problems arising in genomics related
61 fields and the idea of using importance measures generated by the random forests algorithm to extract
62 a ranking of features has already been explored with several different data sets, although, to our
63 knowledge, this has never been done with RNA-Seq data sets [Schridder2018, Freres2016, Yao2015,
64 Duro2012, Anaissi2013].

65 Aside from random forests, we also introduce a technique called Extreme Pseudo-Sampling (EPS)
66 allowing to create case and control pseudo-samples lying on the two extremes of the sample space.
67 This method uses Variational Autoencoders (VAE) [Kingma2013] to create new pseudo-samples that
68 are not present in the original datasets but closely imitate their statistical properties, in that they share
69 the properties of independent and identically distributed samples from the same distribution as the real
70 data.

71 The idea of using autoencoders to classify and examine genomics datasets is not new [Tan2015].
72 However, VAEs differ from other autoencoders in that they can create a meaningful latent
73 representation space where one can choose a new vector in the latent space and create a valid,

74 previously unseen sample in real space that closely follows the real samples (the aforementioned
75 pseudo-samples).

76 Additionally, although autoencoders have been used as an auxiliary tool in the classification of existing
77 datasets, no attempt has been made to extract the knowledge learnt by the autoencoders in this process
78 to trace the analysis and results back to the actual gene expression values and their relationships. Here,
79 we suggest a way to make use of that information [Tan2015].

80 2 Materials and Methods

81 2.1 Data sets

82 Several data sets from the TCGA database have been selected to validate both methods
83 [Weinstein2013].

84 Only the data sets containing 30 healthy samples (denoted as “Solid Tissue Normal” in the TCGA
85 database) or more have been selected. All read counts produced by HTSeq as well as the clinical data
86 have been downloaded with the TCGABiolinks R/Bioconductor package [Colaprico2016].

87 The data sets selected are summarized in Table 1.

88

89

| Name | Cancer type | N (tumors) | n (healthy) | Median age | Age range |
|-----------|---------------------------------------|------------|-------------|------------|-----------|
| TCGA-BRCA | Breast invasive carcinoma | 1097 | 113 | 59.07 | 26-90 |
| TCGA-LUAD | Lung adenocarcinoma | 582 | 59 | 66.88 | 33-88 |
| TCGA-UCEC | Uterine Corpus endometrial carcinoma | 559 | 35 | 64.24 | 31-90 |
| TCGA-KIRC | Kidney renal clear cell carcinoma | 535 | 72 | 61.16 | 26-90 |
| TCGA-HNSC | Head and neck squamous cell carcinoma | 528 | 44 | 61.14 | 20-90 |
| TCGA-THCA | Thyroid carcinoma | 507 | 58 | 46.92 | 15-89 |
| TCGA-LUSC | Lung squamous cell carcinoma | 504 | 49 | 68.66 | 39-90 |
| TCGA-PRAD | Prostate adenocarcinoma | 498 | 52 | 61.99 | 42-78 |
| TCGA-COAD | Colon adenocarcinoma | 460 | 41 | 68.88 | 31-90 |
| TCGA-STAD | Stomach adenocarcinoma | 443 | 32 | 67.56 | 30-90 |
| TCGA-LIHC | Liver hepatocellular carcinoma | 377 | 50 | 61.53 | 16-88 |
| TCGA-KIRP | Kidney renal papillary cell carcinoma | 291 | 32 | 62.03 | 28-88 |

90

91

92

Table 1. TCGA data sets used in this study.

93 2.2 Methodology

94 For each data set, the methodology illustrated in Fig. 1 has been applied:

- 95 • All samples are normalized with the DESeq2 software package [Love2014].
- 96 • The samples are split into a training set and a validation set. The training set contains all the
97 healthy samples of the original data set (n) and the same number of tumor samples as healthy
98 samples (n). The validation set contains the remaining tumor samples ($N - n$).
- 99 • Differential expression analysis is performed on the training set with the DESeq2 software
100 package, using default parameters and options. A ranking of genes, based on their adjusted p-
101 value relative to the differential expression test, is obtained.
- 102 • A random forests classifier is built on the training set with the ranger R package, using 100000
103 trees and a value for the m_{try} parameter of 236 (equal to the square root of the total number of
104 features) [Wright2015]. A ranking of genes based on their permutation importance values is
105 obtained (the permutation importance is computed by randomly permuting the values of the
106 feature of interest and measuring the resulting increase in error).
- 107 • The Extreme Pseudo-Sampling method (see 2.3) is applied on the training set(s) to extract a
108 ranking of genes.
- 109 • Let RF denote the random forests based gene ranking, DE the differential expression based
110 gene ranking and EPS the extreme pseudo-samples based gene ranking. RF_i denotes the i -th
111 gene of the random forests based gene ranking. Similarly, DE_i denotes the i -th gene of the
112 differential expression based gene ranking and EPS_i denotes the i -th gene of the extreme
113 pseudo-samples based gene ranking.
- 114 • For both rankings, 20 gene signatures are generated, including an incremental number of genes.
115 Let $sigRF_i$ denote the i -th gene signature based on the random forests ranking, $sigDE_i$ denote
116 the i -th gene signature based on the differential expression ranking and $sigEPS_i$ the i -th gene
117 signature based on the extreme pseudo-samples ranking. The signatures are formally defined
118 as:
 - 119 ○ $sigRF_i = \{RF_1, \dots, RF_i\}$, for $i = 1, \dots, 20$
 - 120 ○ $sigDE_i = \{DE_1, \dots, DE_i\}$, for $i = 1, \dots, 20$
 - 121 ○ $sigEPS_i = \{EPS_1, \dots, EPS_i\}$, for $i = 1, \dots, 20$
- 122 • For each signature,
 - 123 ○ A Cox proportional hazard model was built using all genes of the signature
 - 124 ○ The samples of the validation set were split into two groups (higher and lower survival),
125 based on the median of the Cox proportional hazard model.
 - 126 ○ A log-rank test was performed to compare the survival of the two groups.

- 127 • For $i = \{1, \dots, 20\}$, the p-value of the log-rank tests obtained with $sigDE_i$, $sigRF_i$, $sigEPS_i$
128 are compared.

129

130 For each data set, correlation coefficients have been computed between the expression values of the
131 50% most expressed genes; a hierarchical clustering of the 50% most expressed genes was performed,
132 to assess if multicollinearity played a role in the performance of the RF based method (multicollinearity
133 denotes the presence of non-independent features such that the relationship between each of these
134 features and the model output is influenced by the relationships between the non-independent features).
135 A hierarchical clustering of all samples was also performed, with the 50% most expressed genes.
136 Enrichment analysis was performed on gene lists from both methods.

137 The correlation coefficient between each top-ranked gene from both list and the 50% most expressed
138 genes has been computed for each data set.

139 Globally, the correlation between the overall survival at 5 years of all cancer types, and the performance
140 of the presented methods was computed.

141 **2.3 Extreme Pseudo-Sampling**

142 It is worth noting that, in most data sets considered in this study, the samples from both classes reside
143 in a high dimensional space and are tightly coordinated together, such that a linear classifier cannot
144 separate them at all. The low count of normal samples compared to the total sum of samples also
145 contributes to the failure of linear classifiers; which tend to receive bias from such unbalance of class
146 membership statistics.

147 We decided to use a dimensionality reduction technique in order to both address the *curse of*
148 *dimensionality* and find a representation in which these samples lay in a linearly-separable subspace.

149 Autoencoders have shown to be able to create such latent representations better than their linear
150 counterparts such as PCA [Tan2015, Danaee2016]. However, such representations do not provide us
151 with useful, actionable knowledge about genes due mainly to their non-linear activation functions.

152 Moreover, Normal Autoencoders are not generative, i.e. while it is possible to come up with useful
153 latent representations for classification purposes, one cannot generate new samples similar to the real
154 samples by slightly modifying their latent representation values and feeding the result into the decoder
155 network.

156 A new type of Autoencoder, called the Variational Autoencoder, however, can succeed in this task
157 [Kingma2013]. VAEs are fundamentally different from other AEs in that they are generative models:

158 Each point x in real space will be associated with distribution $P(z|x)$. For the purpose of this
159 methodology, we assumed this distribution to be normal. Getting latent representation z_l from sample
160 x_l , thus, would be equal to drawing a sample from distribution $\mathcal{N}(\mu_l, \sigma_l)$, where μ_l, σ_l are learned from
161 the training data.

162 The training VAE comprises 9 layers, having 30000, 15000, 10000, 2000, 500, 2000, 10000, 15000,
163 30000 perceptrons respectively. The training process of these layers requires fine-tuning approximately
164 5 billion parameters. Given that the performance of this fine-tuning process increases with the number

165 of samples, in addition to the training set extracted from the studied TCGA dataset, a random selection
166 of samples from the 11 other training sets is used in the VAE training process.

167 After the training step, each dataset D_c is transformed to its latent representation L_c . Said latent
168 representation allows to linearly separate the normal samples from cancerous ones with almost 100%
169 accuracy for both testing and training datasets. Considering the linear separator, let us denote the
170 furthest populated areas on both sides of the separator, called N_c for the normal side of the linear
171 separator and C_c for the cancerous side. If we consider a point z_n in one of these areas, we know it has
172 been randomly drawn from distribution $\mathcal{N}(\mu_n, \sigma_n)$.

173 While selecting z_n is a random process, once a z_n has been drawn from any of the distributions,
174 reconstructing $\hat{x}_n \approx x_n$ from z_n is a deterministic process done by the decoder. However, every point in
175 the close proximity of z_n can be drawn from the same distribution. Due to the deterministic features of
176 the decoder, each of these points would end up generating a different \hat{x}_n . Although different, every
177 possible \hat{x}_n should resemble the original x_n closely and should also follow the general statistical
178 characteristics of all x 's in the dataset.

179 We then drew 400 random points in areas N_c and C_c of the latent space L_c , on both sides of the linear
180 separator and generated new “virtual” or “pseudo” samples of both cancerous and normal classes, a
181 process that we call Extreme Pseudo Sampling (EPS). The amount of random points drawn (400) was
182 chosen using cross validation on the training data. It was the smallest number of samples that ended up
183 in a successful regression process.

184 While real samples cannot be divided using a linear separator and suffer from unbalance of class
185 member counts; we were able to generate new pseudo samples that can be divided linearly in real space
186 due to their exaggerated cancerous/normal features. These samples also are of equal count. The later
187 trait enables the dividing regression lines to be less biased towards a specific class. Thus, said
188 regression lines maintain the same distance from both classes.

189 Finally, since all sample features have been normalized in the process, weight coefficients in the line
190 formula can be translated into importance factors for classifying extreme pseudo samples. The larger
191 a coefficient, the more important its related feature is in determining class membership. Thereby, we
192 are able to extract an importance ranking for all genes, in each data set.

193 The R and Python scripts used to perform the aforementioned analyses are available online:
194 https://github.com/stephwen/ML_RNA-Seq & <https://github.com/roohy/Extreme-Pseudo-Sampler>

195 **3 Results**

196 For each data set, 60 log-rank tests have been performed on the validation set, using gene signatures
197 $sigDE_i$, $sigRF_i$, and $sigEPS_i$ with $i = \{1, 2, \dots, 20\}$ which contain from 1 to 20 genes out of the gene
198 ranking derived from differential expression analysis, the gene ranking derived from the random forests
199 classifier, and the gene ranking derived from the Extreme Pseudo-Sampling method respectively. The
200 p-values of these tests have been compared two by two.

201 Table 2 summarizes the results and shows the number of gene signatures where the random forests
202 based gene ranking outperforms the differential expression based gene ranking and where the Extreme-
203 Pseudo Sampling method outperforms the differential expression based gene ranking.

204

| Name | Cancer type | Random forests | Extreme pseudo-samples |
|-----------|---------------------------------------|----------------|------------------------|
| TCGA-BRCA | Breast invasive carcinoma | 5 | 19 |
| TCGA-LUAD | Lung adenocarcinoma | 14 | 14 |
| TCGA-UCEC | Uterine Corpus endometrial carcinoma | 16 | 9 |
| TCGA-KIRC | Kidney renal clear cell carcinoma | 13 | 10 |
| TCGA-HNSC | Head and neck squamous cell carcinoma | 14 | 15 |
| TCGA-THCA | Thyroid carcinoma | 15 | 15 |
| TCGA-LUSC | Lung squamous cell carcinoma | 5 | 0 |
| TCGA-PRAD | Prostate adenocarcinoma | 12 | 19 |
| TCGA-COAD | Colon adenocarcinoma | 11 | 18 |
| TCGA-STAD | Stomach adenocarcinoma | 13 | 19 |
| TCGA-LIHC | Liver hepatocellular carcinoma | 19 | 8 |
| TCGA-KIRP | Kidney renal papillary cell carcinoma | 10 | 19 |

205

206 *Table 2. The random forests column denotes the number of random forests based signatures having*
 207 *a lower log-rank p-value than their corresponding differential expression based signatures. The*
 208 *extreme pseudo-samples column denotes the number of extreme pseudo-samples based signatures*
 209 *having a lower log-rank p-value than their corresponding differential expression based signatures.*
 210 *The 3 colors (green, yellow, red) refer to cases where the proposed methods have a higher number,*
 211 *the same number, and a lower number of best-performing gene signatures than DESeq2,*
 212 *respectively.*

213

214 For 9 out of the 12 data sets analyzed (lung adenocarcinoma, uterine corpus endometrial carcinoma,
 215 kidney renal clear cell carcinoma, head and neck squamous cell carcinoma, thyroid carcinoma, prostate
 216 adenocarcinoma, colon adenocarcinoma, stomach adenocarcinoma, liver hepatocellular carcinoma),
 217 the random forests based gene ranking outperforms the differential expression based gene ranking in
 218 terms of identifying subsets of genes associated with survival. For 8 out of the 12 datasets (breast
 219 invasive carcinoma, lung adenocarcinoma, head and neck squamous cell carcinoma, thyroid
 220 carcinoma, prostate adenocarcinoma, colon adenocarcinoma, stomach adenocarcinoma, kidney renal
 221 papillary cell carcinoma), the extreme pseudo-samples based gene ranking outperforms the differential
 222 expression based gene ranking. For one data set (kidney renal papillary cell carcinoma), both the
 223 DESeq2 and the random forests based gene rankings share the same number of best performing
 224 signatures. For one data set (kidney renal clear cell carcinoma), both the DESeq2 and the extreme
 225 pseudo-samples based gene rankings share the same number of best performing signatures. For 2 out

226 of the 12 data sets (breast invasive carcinoma, lung squamous cell carcinoma), the differential
227 expression based gene ranking outperforms the random forests based gene ranking. For 3 out of the 12
228 data sets (uterine corpus endometrial carcinoma, lung squamous cell carcinoma, liver hepatocellular
229 carcinoma), the differential expression based gene ranking outperforms the extreme pseudo-samples
230 based gene ranking.

231 Figure 2 shows the log-rank p-values for the 3 different methods (DESeq2, random forests, extreme
232 pseudo-samples) and their respective gene signatures ranging from 1 to 20 genes, for the 4 largest data
233 sets (TCGA-BRCA, TCGA-LUAD, TCGA-UCEC, TCGA-KIRC). Similar figures for the 8 other data
234 sets are available as supplementary data. The log-rank p-values for the 20 gene signatures related to
235 the 3 rankings for each dataset and the genome wide ranking of genes based on the permutation
236 importance computed by the random forests classifier and on the extreme pseudo-samples method can
237 be found in Supplementary Table 1 and Supplementary Table 2 respectively.

238 No significant difference in the average absolute correlation coefficient obtained between the 50%
239 most expressed genes was found between the different cohorts whose DE based signatures performed
240 better than the RF and EPS signatures and the cohorts whose RF or EPS based signatures performed
241 better than the DE ones. No significant difference in terms of the number of clusters of samples
242 obtained with a hierarchical clustering with the 50% most expressed genes when using a constant
243 height cutoff value of $h = 2 \cdot 10^6$ was found between the different cohorts whose DE based signatures
244 performed better than the RF and EPS signatures and the cohorts whose RF or EPS based signatures
245 performed better than the DE ones. No significant difference in terms of the number of clusters of
246 genes obtained with a hierarchical clustering with the 50% most expressed genes when using a constant
247 height cutoff value of $h = 10^5$ was found either. No significant difference was found between the
248 correlation between the top-ranked genes selected with both methods and the 50% most expressed
249 genes. No correlation was found between the overall survival at 5 years of the different cancer types
250 and the performance of either method (measured as the ratio of $n/20$ top-performing signatures). There
251 is, however, a loose correlation (Pearson correlation coefficient: 0.627, p-value: 0.029) between the
252 number of best-performing DE based signatures among the 20 signatures of each data set and the
253 number of differentially expressed genes (adjusted p-value < 0.05) in each data set. Correlation
254 coefficients and numbers of clusters are present, for all data sets, in Supplementary Table 3.

255 4 Discussion

256 Highlighting genes of interest has always been a part of transcriptomics studies and the advent of
257 RNA sequencing technologies has but further emphasized this endeavor. Traditionally, genes of
258 interest, in case-control studies where one had access to their expression values, were genes where
259 said expression varied greatly from one class to the other. This definition has led to the development
260 of numerous methods making use of diverse statistical models and tests, achieving impressive results
261 in a lot of different use cases. However, these methods often implicitly neglected the importance of
262 gene-gene relationships, by only looking at univariate changes.

263 Here, we propose a paradigm shift, by directing the search for genes of interest towards the use of
264 machine learning methods originally conceived to predict the membership of a sample in a class, as
265 these methods intrinsically model the inter-variable relationships (*i.e.* the previously overlooked
266 gene-gene links).

267 An obvious kind of data sets which should theoretically benefit from this are cancers, as these
268 pathologies are known to involve several genes in a multistep process, with different mechanisms
269 implicating intricate relationships between said genes [Vogelstein2013, Yates2012].

270 By using 12 data sets containing samples of various cancers, we have shown that supervised
271 classification algorithms could be used to extract a meaningful ranking of genes. Namely, the
272 permutation importance (also known as Mean Decrease in Accuracy) generated by the random
273 forests algorithm and the weights coefficients used in the extreme pseudo-samples provided a
274 ranking of genes which outperformed classical methods in most data sets.

275 The permutation importance is not the only variable importance generated by the random forests
276 classifier, as the Gini importance (or Mean Decrease in Impurity) is also available. However, using
277 the Gini importance to classify the genes of these data sets yielded slightly worse results than the
278 results obtained with the permutation importance. Using a combination of both variable importances,
279 as in [Frères2016], also produced worse results than when using the permutation importance alone.

280 Given the fact that neither the random forests based gene ranking nor the extreme pseudo-samples
281 based one outperformed the differential expression based one for all of the 12 data sets, one might
282 wonder if using both a supervised learning based gene selection technique in conjunction with
283 differential expression would not yield better results. However, using the supervised learning based
284 gene selection method after the differential expression one (*i.e.* using only the genes with a
285 significant differential expression adjusted p-value as input features of the random forests classifier
286 or the EPS method) also produced worse results than when using the random forests gene ranking or
287 the EPS gene ranking alone.

288 Using survival analysis as a way to validate gene lists coming from cancer data sets whose average
289 survival differs greatly might spark questions, however there does not seem to be a link between the
290 overall survival (OS) of these cancers and the performance of the proposed methods. Survival
291 information constitutes a quantifiable and relatively easily available information for different data
292 sets. However, using the presumed relationship between the expression values of a gene and the
293 survival of a patient as a proxy for the role of said gene in the selected disease relies on a strong
294 hypothesis whose validity might vary across data sets. Therefore, other gene ranking validation
295 methods should be further explored to assess the performance of a random forests based gene ranking
296 method and the EPS method in a wider range of RNA-Seq experiments.

297 In conclusion, we have shown that using the permutation importance internally computed by the
298 random forests algorithm, when said algorithm is used to build a classifier based on gene expression
299 values of a case-control RNA-Seq data set, allowed to obtain a ranking of genes; Variational
300 Autoencoders could be used to generate pseudo-samples mimicking the properties of real samples,
301 albeit with extreme localizations in latent space; Using the feature weights of said pseudo-samples
302 allowed to obtain a ranking of genes. These rankings were compared with the results of a differential
303 expression analysis, with all three gene rankings being evaluated through survival analysis on a
304 validation cohort different from the cohort used to generate both rankings. The results have shown
305 that the random forests based method and the extreme pseudo-samples outperformed the differential
306 expression based method for 9 and 8 out of the 12 data sets analyzed, respectively. Although the
307 genes selected by both methods are different, there is no significant difference in the number of
308 highly correlated genes between both methods. Although the goal of this research is not to supersede
309 differential expression analysis to select genes of interest in RNA-Seq studies, we have shown that

310 differential expression analysis might miss out on important genes, and a supervised learning based
311 gene selection method should be used alongside.

312 As the field of machine learning contains many different supervised classification and feature
313 selection algorithms, it would be of interest to extend this work by testing the performance of other
314 methods for gene selection in the context of case-control RNA-Seq data sets.

315

316 5 References

- 317 1. Wang2009: Wang, Zhong, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary
318 tool for transcriptomics." *Nature reviews genetics* 10, no. 1 (2009): 57.
- 319 2. Mortazavi2008: Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and
320 Barbara Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature*
321 *methods* 5, no. 7 (2008): 621.
- 322 3. Trapnell2009: Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. "TopHat: discovering
323 splice junctions with RNA-Seq." *Bioinformatics* 25, no. 9 (2009): 1105-1111.
- 324 4. Garber2011: Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Garber, Manuel,
325 Manfred G. Grabherr, Mitchell Guttman, and Cole Trapnell. "Computational methods for
326 transcriptome annotation and quantification using RNA-seq." *Nature methods* 8, no. 6 (2011):
327 469.
- 328 5. Love2014: Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of
329 fold change and dispersion for RNA-seq data with DESeq2." *Genome biology* 15, no. 12
330 (2014): 550.
- 331 6. Wenric2017: Wenric, Stephane, Sonia ElGuendi, Jean-Hubert Caberg, Warda Bezzaou,
332 Corinne Fasquelle, Benoit Charlotteaux, Latifa Karim et al. "Transcriptome-wide analysis of
333 natural antisense transcripts shows their potential role in breast cancer." *Scientific reports* 7,
334 no. 1 (2017): 17452.
- 335 7. Huang2015: Huang, Huei-Chung, Yi Niu, and Li-Xuan Qin. "Differential Expression
336 Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software:
337 Supplementary Issue: Sequencing Platform Modeling and Analysis." *Cancer informatics* 14
338 (2015): CIN-S21631.
- 339 8. Freres2016: Frères, Pierre, Stéphane Wenric, Meriem Boukerroucha, Corinne Fasquelle,
340 Jérôme Thiry, Nicolas Bovy, Ingrid Struman et al. "Circulating microRNA-based screening
341 tool for breast cancer." *Oncotarget* 7, no. 5 (2016): 5416.
- 342 9. Yao2015: Yao, Dengju, Jing Yang, Xiaojuan Zhan, Xiaorong Zhan, and Zhiqiang Xie. "A
343 novel random forests-based feature selection method for microarray expression data
344 analysis." *International journal of data mining and bioinformatics* 13, no. 1 (2015): 84-101.

- 345 10. Duro2012: Duro, Dennis C., Steven E. Franklin, and Monique G. Dubé. "Multi-scale object-
346 based image analysis and feature selection of multi-sensor earth observation imagery using
347 random forests." *International Journal of Remote Sensing* 33, no. 14 (2012): 4502-4526.
- 348 11. Anaissi2013: Anaissi, Ali, Paul J. Kennedy, Madhu Goyal, and Daniel R. Catchpoole. "A
349 balanced iterative random forest for gene selection from microarray data." *BMC*
350 *bioinformatics* 14, no. 1 (2013): 261.
- 351 12. Colaprico2016: Colaprico, Antonio, Tiago C. Silva, Catharina Olsen, Luciano Garofano,
352 Claudia Cava, Davide Garolini, Thais S. Sabedot et al. "TCGAbiolinks: an R/Bioconductor
353 package for integrative analysis of TCGA data." *Nucleic acids research* 44, no. 8 (2015): e71-
354 e71.
- 355 13. Weinstein2013: Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna R. Mills
356 Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M. Stuart,
357 and Cancer Genome Atlas Research Network. "The cancer genome atlas pan-cancer analysis
358 project." *Nature genetics* 45, no. 10 (2013): 1113.
- 359 14. Wright2015: Wright, Marvin N., and Andreas Ziegler. "ranger: A fast implementation of
360 random forests for high dimensional data in C++ and R." *arXiv preprint arXiv:1508.04409*
361 (2015).
- 362 15. Vogelstein2013: Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin
363 Zhou, Luis A. Diaz, and Kenneth W. Kinzler. "Cancer genome landscapes." *science* 339, no.
364 6127 (2013): 1546-1558.
- 365 16. Yates2012: Yates, Lucy R., and Peter J. Campbell. "Evolution of the cancer genome." *Nature*
366 *Reviews Genetics* 13, no. 11 (2012): 795.
- 367 17. Kanehisa2000: Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes
368 and genomes." *Nucleic acids research* 28, no. 1 (2000): 27-30.
- 369 18. Joshitope2005: Joshi-Tope, G., Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther
370 Schmidt, Bernard de Bono, Bijay Jassal et al. "Reactome: a knowledgebase of biological
371 pathways." *Nucleic acids research* 33, no. suppl_1 (2005): D428-D432.
- 372 19. Phillips2008: Phillips, Patrick C. "Epistasis—the essential role of gene interactions in the
373 structure and evolution of genetic systems." *Nature Reviews Genetics* 9, no. 11 (2008): 855.
- 374 20. Vidal2011: Vidal, Marc, Michael E. Cusick, and Albert-László Barabási. "Interactome
375 networks and human disease." *Cell* 144, no. 6 (2011): 986-998.
- 376 21. Kingma2013: Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes."
377 *arXiv preprint arXiv:1312.6114* (2013).
- 378 22. Tan2015: Tan, Jie, Matthew Ung, Chao Cheng, and Casey S. Greene. "Unsupervised feature
379 construction and knowledge extraction from genome-wide assays of breast cancer with
380 denoising autoencoders." In *Pacific Symposium on Biocomputing Co-Chairs*, pp. 132-143.
381 2014.

382 23. Danaee2016: Danaee, Padideh, Reza Ghaeini, and David A. Hendrix. "A deep learning
383 approach for cancer detection and relevant gene identification." In *PACIFIC SYMPOSIUM*
384 *ON BIOCOMPUTING 2017*, pp. 219-229. 2017.

385

386 **6 Author Contributions**

387 SW: Conceived and designed the experiments; Performed the random forests analysis; Contributed to
388 the writing of the manuscript. RS: Developed and performed the Extreme-Pseudo Samples analysis;
389 Contributed to the writing of the manuscript

390 **7 Conflict of Interest**

391 The authors declare that the research was conducted in the absence of any commercial or financial
392 relationships that could be construed as a potential conflict of interest.

393 **8 Funding**

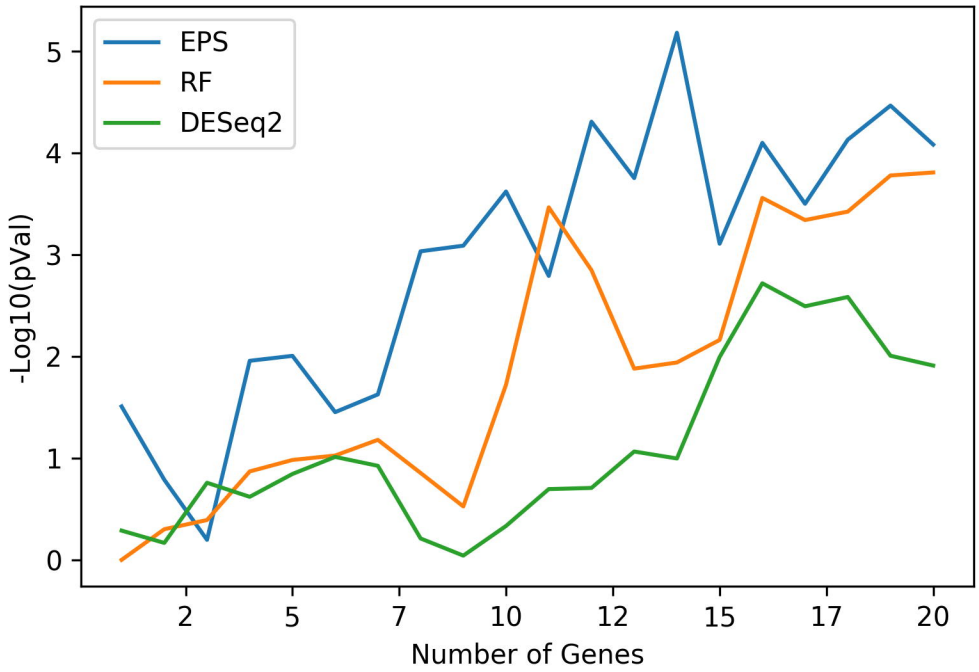
394 S.W. was supported by Wallonia through the following grants: WallInnov2016 - NACATS
395 (1610125), BioWin - TREATBEST - n° 7741, by a Fellowship of the Belgian American Educational
396 Foundation, and a WBI.World Fellowship.

397 **9 Acknowledgments**

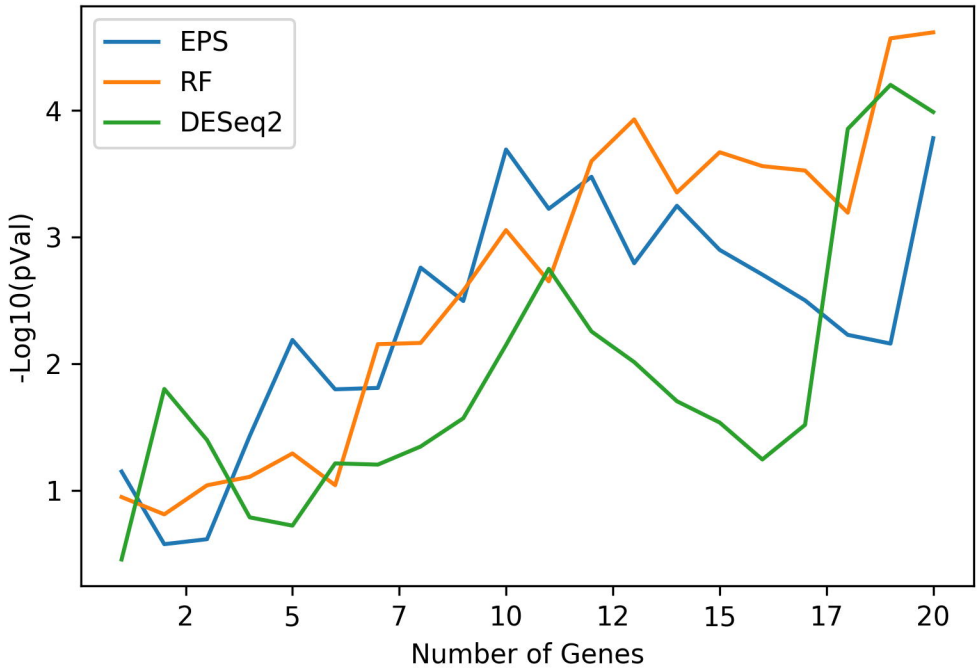
398 We thank Claire Josse, Pierre Geurts, Vincent Botta, Eimear Kenny, Gillian Belbin, Jose-Luis
399 Ambite and Shunsuke Saito. This work was supported in part through the computational resources
400 and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

A

TCGA-BRCA

**B**

TCGA-LUAD



| Name | Cancer type | N (tumors) | n (healthy) | Median age | Age range |
|-------------|---------------------------------------|-------------------|--------------------|-------------------|------------------|
| TCGA-BRCA | Breast invasive carcinoma | 1097 | 113 | 59.07 | 26-90 |
| TCGA-LUAD | Lung adenocarcinoma | 582 | 59 | 66.88 | 33-88 |
| TCGA-UCEC | Uterine Corpus endometrial carcinoma | 559 | 35 | 64.24 | 31-90 |
| TCGA-KIRC | Kidney renal clear cell carcinoma | 535 | 72 | 61.16 | 26-90 |
| TCGA-HNSC | Head and neck squamous cell carcinoma | 528 | 44 | 61.14 | 20-90 |
| TCGA-THCA | Thyroid carcinoma | 507 | 58 | 46.92 | 15-89 |
| TCGA-LUSC | Lung squamous cell carcinoma | 504 | 49 | 68.66 | 39-90 |
| TCGA-PRAD | Prostate adenocarcinoma | 498 | 52 | 61.99 | 42-78 |
| TCGA-COAD | Colon adenocarcinoma | 460 | 41 | 68.88 | 31-90 |
| TCGA-STAD | Stomach adenocarcinoma | 443 | 32 | 67.56 | 30-90 |
| TCGA-LIHC | Liver hepatocellular carcinoma | 377 | 50 | 61.53 | 16-88 |
| TCGA-KIRP | Kidney renal papillary cell carcinoma | 291 | 32 | 62.03 | 28-88 |

Table 1. TCGA data sets used in this study.

| Name | Cancer type | Random forests | Extreme pseudo-samples |
|-----------|---------------------------------------|----------------|------------------------|
| TCGA-BRCA | Breast invasive carcinoma | 5 | 19 |
| TCGA-LUAD | Lung adenocarcinoma | 14 | 14 |
| TCGA-UCEC | Uterine Corpus endometrial carcinoma | 16 | 9 |
| TCGA-KIRC | Kidney renal clear cell carcinoma | 13 | 10 |
| TCGA-HNSC | Head and neck squamous cell carcinoma | 14 | 15 |
| TCGA-THCA | Thyroid carcinoma | 15 | 15 |
| TCGA-LUSC | Lung squamous cell carcinoma | 5 | 0 |
| TCGA-PRAD | Prostate adenocarcinoma | 12 | 19 |
| TCGA-COAD | Colon adenocarcinoma | 11 | 18 |
| TCGA-STAD | Stomach adenocarcinoma | 13 | 19 |
| TCGA-LIHC | Liver hepatocellular carcinoma | 19 | 8 |
| TCGA-KIRP | Kidney renal papillary cell carcinoma | 10 | 19 |

Table 2. The random forests column denotes the number of random forests based signatures having a lower log-rank p-value than their corresponding differential expression based signatures. The extreme pseudo-samples column denotes the number of extreme pseudo-samples based signatures having a lower log-rank p-value than their corresponding differential expression based signatures. The 3 colors (green, yellow, red) refer to cases where the proposed methods have a higher number, the same number, and a lower number of best-performing gene signatures than DESeq2, respectively.

