Comprehensive biological interpretation of gene signatures using semantic distributed representation

Short title: Biological interpretation of gene signatures using NLP

Yuumi Okuzono*, Takashi Hoshino

Immunology Unit, Pharmaceutical Research Division, Takeda Pharmaceutical Company Limited, Fujisawa, Kanagawa, Japan

*Corresponding author

E-mail: yuumi.okuzono@takeda.com

Abstract

Recent rise of microarray and next-generation sequencing in genome-related fields has simplified obtaining gene expression data at whole gene level, and biological interpretation of gene signatures related to life phenomena and diseases has become very important. However, the conventional method is numerical comparison of gene signature, pathway, and gene ontology (GO) overlap and distribution bias, and it is not possible to compare the specificity and importance of genes contained in gene signatures as humans do.

This study proposes the gene signature vector (GsVec), a unique method for interpreting gene signatures that clarifies the semantic relationship between gene signatures by incorporating a method of distributed document representation from natural language processing (NLP). In proposed algorithm, a gene-topic vector is created by multiplying the feature vector based on the gene's distributed representation by the probability of the gene signature topic and the low frequency of occurrence of the corresponding gene in all gene signatures. These vectors are concatenated for genes included in each gene signature to create a signature vector. The degrees of similarity between signature vectors are obtained from the cosine distances, and the levels of relevance between gene signatures are quantified.

Using the above algorithm, GsVec learned approximately 5,000 types of canonical pathway and GO biological process gene signatures published in the Molecular Signatures Database (MSigDB). Then, validation of the pathway database BioCarta with known

biological significance and validation using actual gene expression data (differentially expressed genes) were performed, and both were able to obtain biologically valid results. In addition, the results compared with the pathway enrichment analysis in Fisher's exact test used in the conventional method resulted in equivalent or more biologically valid signatures. Furthermore, although NLP is generally developed in Python, GsVec can execute the entire process in only the R language, the main language of bioinformatics.

Introduction

The recent rise of microarray and next-generation sequencing (NGS) in genome-related fields has made it possible to easily acquire gene expression data at the whole gene level. As a result, interpretation of life phenomena and diseases is advancing [1].

To identify the gene population involved in a phenotype, gene expression data for comparison between healthy subjects and subjects with diseases as well as treated and untreated groups can be obtained. Based on the correlation between the representative expression value of the gene signature and the phenotype, the gene signature of genes related to the phenotype can be identified, and the biological interpretation of gene signatures can be performed.

To interpret a gene signature identified in this data-driven manner, it is necessary to avoid bias due to the large number of genes that must be interpreted and comprehensiveness

and completeness of human knowledge. Therefore, interpretation is commonly performed by comparing the gene signature, such as differentially expressed genes and gene modules, against a biological gene signature database (such as pathway and GO) and identifying an objective association from a biological perspective [2].

Numerous methodologies for association with pathways have been proposed. Common examples include Fisher's exact test, which is a classical statistical test for the specific overlap of genes; over-representation analysis and gene set enrichment analysis [3], which statistically process the number of overlapping genes and ranking bias by incorporating randomization; and modular enrichment analysis and EnrichNet with graph-based statistics of biological networks [4, 5].

However, these comparisons are numerical, and it is thus not possible to compare the semantic nuances of the included genes as humans do. After performing the aforementioned analyses, to interpret the gene signature it is necessary to perform a number of comprehensive judgments to identify whether the genes overlapped by humans are specific to the pathway, determine their biological importance, and establish whether they contain genes that have similar meaning without direct overlap.

In this study, we propose a novel method for associating biological gene signatures by applying a distributed representation model [6] of documents to facilitate the interpretation of gene signatures (see Fig 1). The distributed representation of documents is a technology for

vectorizing documents of any length that was developed in the field of natural language processing (NLP). Words (that are included in documents) and documents (that are sets of words) are vectorized to convert a semantic expression of the words and documents into a mathematical expression that can be easily processed by a computer. As the first step, feature extraction is performed at the word level, and as the second step, feature extraction is performed at the document level as a set of words. Because feature vectors can be semantically compared by performing mathematical comparisons, they are used in a variety of real-world situations, such as sentence classification, content recommendation, sentiment analysis, and spam filtering [7].

Fig 1. Research concept.

Beginning with Doc2Vec [8], which used a distributed representation of words, innovative techniques related to the distributed expression of a large number of sentences have been proposed in the past several years, and the accuracy of document interpretation has improved [9]. Typical methods of distributed representation of documents include statistical semantic extraction methods [10], methods that combine distributed representations of words [11] into document representations [12], methods that directly compress word and document IDs [8], methods of summing word vectors by multiplying the topics and specificities in the

documents [9].

There are other NLP methods for the distributed representation of documents; however, the methods applicable to bioinformatics are limited, as there are differences in assumptions between NLP and bioinformatics. A meaningful component in bioinformatics is a gene, which corresponds to a word in NLP. In addition, a gene signature, which is a set of genes, corresponds to a document, which is a set of words. However, in NLP, the order and context of words are important, whereas the order of gene signatures compiled in a general pathway database often has no meaning.

In this study, we developed an original method for the interpretation of gene signatures by applying a distributed expression algorithm. The algorithm extracts semantic features of genes and their biological gene signatures and reveals specific relationships by comparing the abovementioned signatures with the gene signatures to be interpreted (e.g., differentially expressed genes, gene modules). As training data, a gene signature was used, which has a clear biological meaning in the molecular signatures database (MSigDB) [13] used in the conventional pathway enrichment analysis. Furthermore, Python is the primary programming language used for machine learning and NLP analysis; however, our proposed method can execute the entire process in R, which is the primary language in the analysis domain of bioinformatics. Therefore, the proposed method can be immediately used without further modification for bioinformatics analysis. Combining our proposed method of biological gene

110 signature vectorization with conventional enrichment analysis can allow for more intuitive
111 and reliable interpretation of gene signatures.
112

113 # Results

114 ## Construction of gene and gene signature feature vectors by distributed representation

115 In this study, we developed a method for creating gene signature feature vectors and
116 clarifying semantic similarity by applying methodology from the field of NLP (Fig 2). We
117 defined proprietary functions using the packages published on the Comprehensive R Archive
118 Network that can be used in the R language. In addition, we executed an original algorithm
119 for creating a unique gene signature feature vector based on the sparse composite document
120 vectors (SCDV) [9] method from NLP using only R language operations.
121

122 **Fig 2. Algorithm and workflow of GsVec.** GsVec is divided into preparation and analysis
123 parts. In the preparation part, gene vectors are created from training data with a clear
124 biological meaning (e.g., GO, Pathway, and Hallmark genes) using BOW and Word2Vec, and
125 the probability of each cluster calculated by GMM is multiplied. Furthermore, gene-topic
126 vectors are created by multiplying the inverse signature factor and averaging for each gene
127 included in the gene signature. In the analysis part, validation data, which are not biologically
128 interpreted gene signatures (e.g., differentially expressed genes and gene modules) are

converted into signature vectors from the gene-topic vector created using training data. The cosine similarity between the validation and training data is calculated to obtain the association with biological meaning. gv_i represents the vector of an arbitrary gene g_i, c represents the cluster, K represents the number of clusters, S_N represents the total number of gene signatures, $Sf(g_i)$ represents the number of gene signatures including gene g_i, and $\oplus$ represents the concatenation.

The training data used 5,456 gene signatures (i.e., C2: Canonical pathway and C5: GO biological process), which were used in conventional pathway/GO enrichment.

A gene × gene signature matrix was created for the gene signature (equivalent to the Bags of Word step in NLP), and gene features were expressed using a distributed word representation algorithm [14] to create gene vectors (equivalent to Word2Vec processing in NLP). The clusters corresponding to the topics present in these gene signatures were extracted by soft clustering of the gene vectors with a Gaussian mixture model (GMM) [15, 16]. The probabilities that each word contributes to each cluster were multiplied for each cluster to obtain the abovementioned gene vectors, and those vectors were combined for each gene to obtain gene-cluster vectors.

Simultaneously, by dividing the total number of gene signatures by the number of gene signatures for each gene that contain that gene, the scores for reducing the weight of the gene

appearing in various signatures were calculated. Hereinafter, this is referred to as the inverse signature factor, which is equivalent to IDF in NLP [17]. Gene-topic vectors were obtained by multiplying the abovementioned gene-cluster vectors by those scores. The signature vectors were obtained by averaging the gene-topic vectors for the genes included in each individual gene signature. The signature vectors which were feature vectors in the genes and the abovementioned gene signatures were used as training data in the subsequent analysis.

Evaluation of gene signature vector (GsVec) performance using gene signatures with known biological interpretation

In this study, 5,242 gene signatures, excluding gene signatures derived from the BioCarta database from the C2 canonical pathway and C5 GO biological process, were used as training data. BioCarta's 214-gene signatures were used as validation data with a known biological meaning. Furthermore, in both the training and validation data, we selected gene signatures related to immune function by human selection and tagged them. The number of applicable gene signatures was 405.

Signature vectors of the validation data were created from the gene-topic vectors created during learning in the same way as the training data, and the degrees of relevance with the training data were evaluated by the cosine similarity score with the learning set (hereinafter, the result of the cosine similarity calculated by a series of operations is referred to as *GsVec*).

First, the GsVec results (i.e., similarity relationship between signature vectors) were visualized by two-dimensionally projecting them using t-distributed Stochastic Neighbor Embedding (t-SNE) [18] (Fig 3[A]). There was a tendency for immune-related signatures to be consolidated in one location. The abovementioned tendency was observed not only in the training data, but also in the validation data, and the meaning of the validation data was correctly predicted by GsVec. These results demonstrate that although GsVec using NLP is an entirely different approach from conventional methods, it can identify groups with similar meanings.

Fig 3. GsVec accuracy evaluation and comparison with Fisher. A) Matrix of the cosine distance between signature vectors of training and validation data projected in two dimensions with t-SNE. The size of the circle represents the number of genes included in each signature. Green color refers to training data related to immunity, purple to training data not related to immunity, blue to validation data related to immunity, and red to validation data not related to immunity. B) Scatter plot by the -log10 P-value of Fisher and the cosine distance between the signature vectors of the training and validation data. A histogram of the distribution of each value is also illustrated. The R in the upper right of the scatter plot indicates the Pearson correlation coefficient between GsVec and Fisher.

Next, the P-value of the overlap between Fisher's exact test training data and validation

187 data was converted to a -log10 value (hereinafter, −log10 (P-value) obtained by Fisher's exact
188 test is referred to as *Fisher*) and compared with the GsVec results. The correlation by Pearson
189 coefficient score between GsVec and Fisher was 0.453, and the relationship with a score of
190 0.75 or higher in GsVec was a significant score less than 0.001 (-log10 (4)) in Fisher. In
191 addition, the degree of relevance was concentrated in the Fisher distribution near 0, which
192 was not significant, whereas in GsVec, the distribution had a long tail. These results suggest
193 that GsVec was able to reflect robust results that were significant in Fisher, and could also
194 associate gene signatures with interpretation that were difficult to interpret in Fisher (Fig
195 3[B]).

196 In addition, the relationship between the gene signatures of individual validation data
197 and training data was compared between GsVec and Fisher. The top 10 results were the same
198 for GsVec and Fisher for gene signatures, with a large number of genes included in the gene
199 signature in the training data and a large number of overlaps. Furthermore, in GsVec, even if
200 a gene included in the gene signature in the training data was small and the number of
201 overlapped genes was small, if the overlapped genes were characteristic genes in the signature
202 vector space, there was a tendency to display high relevance. Typical results are presented in
203 Fig 4.

204

Fig 4. Comparison between GsVec and Fisher on top 15 results with high relevance between training data and validation data. Each top 15 result that was highly correlated with the gene signature and GsVec of the validation data or that was highly significant with Fisher is presented in the left and right tables. The column *Training gene signature name* indicates the name of the corresponding training data, and the biological data that directly match the validation data are highlighted in yellow. The *genes* column contains the total number of genes included in the gene signature of the corresponding training data, the *overlap* column contains the number of genes overlapped between the corresponding validation data and training data, and the *Fisher's* column represents the -log10 (P-value) of the significance of the corresponding training data and validation data by Fisher's exact test. The *GsVec* column exhibits similarities according to the cosine distance of the signature vectors of the corresponding training and validation data.

Here, we examined whether GsVec was able to produce biologically meaningful insights using three representative examples of immune-related pathways. In the Interleukin-1 receptor (IL1R) pathway, the direct gene signature name *IL1* or *Interleukin 1* could be identified in both GsVec and Fisher. Similar results were observed in the B cell receptor and Toll pathways, including *Toll-like receptor*. In addition, gene signature names indicating broader concepts, such as *cytokine signaling* and *pattern recognition receptor*, were observed in the

IL1R and Toll pathways, respectively. These were extracted by GsVec and are associated with biologically relevant gene signatures.

Association of biological meaning with GsVec using differentially expressed genes (DEGs) from real data

Data-driven gene signatures, such as DEGs in the affected tissues and normal tissues of diseases published in the Expression Atlas [19] and DEGs extracted from The Cancer Genome Atlas (TCGA) [20] data, were analyzed with GsVec and Fisher (Fig 5). In all datasets, DEGs were used; they were upregulated in the affected tissue compared to the normal tissue. However, because in cancer there were too many DEGs (approximately 1,000–2,000) for the pathway enrichment analysis, only the results of other diseases were interpreted. Additionally, only in multiple sclerosis (MS), the number of DEGs was low; thus, we analyzed not only upregulated (up) but also downregulated (down) DEGs. The individual analysis results are presented below.

Fig 5. List of DEGs used for verification. The table illustrates the DEG data of affected tissues and control tissues of the major public diseases used as data-driven signatures, not signatures compiled in terms of biological meaning. The *up* column contains the number of DEGs that were higher than the control for the disease, while the *down* column contains the DEG numbers that were lower than the control for the disease. The red frames represent

subsequent interpretations.

Multiple sclerosis (MS) cerebrospinal fluid (CSF; E-MTAB-69). In MS (up), B-cell-related signatures were identified in GsVec and Fisher (Fig 6). Although the involvement of T cells in the pathology of MS is well known, the participation of B cells has also attracted attention in recent years, and the possibility of therapeutic drugs targeting B cells has also been investigated [21, 22]. In contrast, in MS (down), the gene signature names *locomotion*, *taxis*, or *migration*, reminiscent of cell migration, were highly ranked. In this study, the dataset consisted of CSF-derived samples. The results indicated that the involvement of immune cells in the peripheral and central nervous system (CNS) was captured. Notably, the gene signature name, *nervous system development*, was highly ranked only in GsVec. In other words, when considering the pathological condition of MS, GsVec captured not only immunological but also neuronal aspects and was able to extract more biologically valid signatures.

Fig 6. Comparison between GsVec and Fisher of top 15 results with high relevance in DEGs of multiple sclerosis (MS). Using the DEGs of MS cerebrospinal fluid (E-MTAB-69) as validation data, each top 15 result that was highly correlated with GsVec or highly significant with Fisher is presented in the left and right tables. The *genes* column contains the

263 total number of genes included in the gene signature of the corresponding training data, the
264 *overlap* column contains the number of genes overlapped between the corresponding
265 validation and training data, and the *Fisher's* column represents -log10 (P-value) of the
266 significance of the corresponding training data and validation data by Fisher's exact test. The
267 *GsVec* column exhibits similarities according to the cosine distance of the signature vectors of
268 the corresponding training and validation data.
269
270 **Crohn's disease (CD) colon (E-MEXP-2083, E-GEOD-59071).** We examined a dataset of
271 fresh frozen ileum mucosal tissue from CD patients, and gene signatures, such as *interferon*
272 *gamma*, *cell adhesion*, and *leukocyte migration* were highly ranked in GsVec (see Fig 7). For
273 inflammatory bowel diseases, such as CD and ulcerative colitis, it has been reported that
274 intestinal immune cell trafficking has been identified as a central event in the pathogenesis of
275 diseases. Additionally, cell adhesion is a pivotal step in several aspects of immune cell
276 trafficking [23]. In Fisher, *interferon gamma* was highly ranked, and the broader concept
277 *cytokine* was also highly ranked. However, the gene signature *asthma*, which may not be
278 directly related to CD, was also highly ranked in Fisher. It has been reported that there are
279 many common risk factors for the association between asthma and CD, including genetic and
280 environmental factors [24]. Thus, GsVec and Fisher in CD displayed a similar trend but
281 identified different characteristics on the whole. GsVec can extract more biologically

meaningful signatures.

Fig 7. Comparison between GsVec and Fisher of top 15 results with high relevance in DEGs of Crohn's disease (CD). Using the DEGs of CD colon (E-MEXP-2083, E-GEOD-59071) as validation data, each top 15 result that was highly correlated with GsVec or highly significant with Fisher is presented in the left and right tables. The details of each column are identical to those in Fig 6.

Type 2 diabetes (T2D) islet of Langerhans (E-MTAB-5060). In this study, we used a dataset consisting of islet of Langerhans tissue from healthy donors and T2D patients. In both GsVec and Fisher, gene signatures related to inflammation such as *inflammatory response*, *migration*, and *wound* were highly ranked and demonstrated a similar trend (see Fig 8). Insulin resistance and beta cell dysfunction are well known in T2D pathologies, and inflammation is related to the pathogenesis of these conditions [25]. In addition, injury and wound healing processes associated with the term *wounding* are known to alter responses to growth factors and cytokines in addition to tissue remodeling through cell migration and proliferation [26, 27]. These scientific reports examined the effect of macrophages on T2D-related ulcers and skin wounds, but not the islets themselves. The direct cause-and-effect relationship is unknown; however, based on the GO term definition, the relationship was

linked to a gene signature involved in damaged tissue and tissue repair.

Fig 8. Comparison between GsVec and Fisher of top 15 results with high relevance in DEGs of Type 2 diabetes (T2D). Using the DEGs of T2D islet of Langerhans (E-MTAB-5060) as validation data, each top 15 result that had high correlation with GsVec or high significance with Fisher is presented in the left and right tables.

Duchenne muscular dystrophy (DMD) skeletal muscle (E-GEOD-3307). We then examined DMD, and the results demonstrated that GsVec and Fisher displayed similar trends (Fig 9). Specifically, gene signatures such as *extracellular structure organization* related to the extracellular matrix and *ossification* related to the bone were highly ranked. DMD is an inherited muscular disorder known to be caused by an abnormality in dystrophin, a cytoskeletal protein, and has been linked to extracellular matrix-related molecules [28]. In addition, a relationship between bone morphogenetic proteins signals and this disease has been reported, and several biological features have been extracted [29].

Fig 9. Comparison between GsVec and Fisher of top 15 results with high relevance in DEGs of Duchenne muscular dystrophy (DMD). Using the DEGs of DMD skeletal muscle (E-GEOD-3307) as validation data, each top 15 result that was highly correlated with GsVec

320 or highly significant with Fisher is presented in the left and right tables.

321

322 **Systemic lupus erythematosus (SLE) whole blood (E-GSEOD-72509).** A whole blood
323 dataset from SLE patients was analyzed as representative of autoimmune disease. The data
324 were extracted from a heterogeneous population, including those with high and low interferon
325 signature values. However, both GsVec and Fisher fully identified the signatures of
326 immune-related genes with similar results (see Fig 10).

327

328 **Fig 10. Comparison between GsVec and Fisher of top 15 results with high relevance in**
329 **DEGs of systemic lupus erythematosus (SLE).** Using SLE whole blood (E-GSEOD-72509)
330 DEGs as validation data, each top 15 result that was highly correlated with GsVec or highly
331 significant with Fisher is presented in the left and right tables.

332

333 **Sarcoidosis lung tissue (E-GEOD-16538).** We analyzed a dataset derived from lung tissue
334 from sarcoidosis patients. The results indicated that the features of the gene signatures
335 identified by GsVec and Fisher were partially different (Fig 11). In GsVec, gene signatures
336 related to kinase activity, including *MAPK* (mitogen-activated protein kinase), were highly
337 ranked, whereas Fisher identified several gene signatures related to *chemokine*. The original
338 dataset was intended for the investigation of gene regulation of granulomatous sarcoidosis,

and it was reported that the gene network associated with the Th1-type response was overexpressed mainly in lung tissue derived from sarcoidosis [30]. In another paper, not only were Th1 cytokines increased in sarcoidosis, but MAPK, especially p38 activation, was found in cells of bronchoalveolar lavage fluid from patients with sarcoidosis [31]. Chemokines were also reported [32]. In summary, sarcoidosis-related gene signatures were identified; however, the two algorithms GsVec and Fisher exhibited different characteristics.

Fig 11. Comparison between GsVec and Fisher of top 15 results with high relevance in DEGs of sarcoidosis. Using the DEGs of sarcoidosis lung tissue (E-GEOD-16538) as validation data, each top 15 result that was highly correlated with GsVec or highly significant with Fisher is presented in the left and right tables.

Schizophrenia (SCZ) brain Brodmann area 24 (GEOD-78936). In examining SCZ, we used a dataset derived from postmortem brain tissue. In SCZ GsVec, gene signatures related to *hormone* were highly ranked (Fig 12). *Hormone* was a common but a unique result, which only ranked in GsVec, which suggests the involvement of dopamine in the limbic system [33]. Further, a gene signature, *nervous*, was also highly ranked. In contrast, in Fisher, gene signatures that suggested the involvement of other CNS cells, such as *glia cell* and *astrocyte*, were highly ranked and had different features.

Fig 12. Comparison between GsVec and Fisher of top 15 results with high relevance in DEGs of schizophrenia (SCZ). Using the DEGs of SCZ brain Brodmann area 24 (GEOD-78936) as validation data, each top 15 result that was highly correlated with GsVec or highly significant with Fisher is presented in the left and right tables.

Discussion

In this paper, we propose a method for associating gene signatures by feature extraction using an NLP method. This method is entirely different from traditional pathway enrichment analysis for gene signature interpretation. Biologically reasonable results were obtained both in the verification of the pathway database (BioCarta) with known biological significance and in the verification using DEGs extracted from the actual gene expression. Compared to conventional pathway enrichment analysis by Fisher's exact test (Fisher), the proposed algorithm (GsVec) can identify a signature that is equivalent or more biologically relevant than Fisher.

Among diseases known to be related to immunity, GsVec tends to differentiate well between the biological features of autoimmune diseases. In MS, GsVec extracted more biologically valid signatures from immunological and neuronal aspects. Additionally, in CD, the signature called *interferon gamma* and other characteristics (e.g., *cell trafficking*) can be

377 extracted in GsVec. Moreover, in SCZ, a unique signature, *hormone*, was highly ranked only
378 in GsVec. Thus, GsVec captured the signatures related to periphery and CNS more
379 specifically than Fisher.

380 In this study, there were many reasons for selecting the SCDV-based method among the
381 many methods related to distributed representation of documents in NLP. In the advance
382 analysis, in comparing multiple methods, BOW and TFIDF did not consider gene similarity
383 that was not directly overlapping, resulting in a high correlation with Fisher; thus, the
384 advantage of using NLP methods was low. In the averaging of word2vec (gene vector), the
385 specificity of a gene signature was not taken into account, and thus did not meet our purpose.
386 In addition, as a result of assuming a general bioinformatics analysis environment (e.g., R
387 language, PC specifications) as the potential of this research development, methods with a
388 large amount of computation using deep learning were excluded from the candidates, as well
389 as methods that were difficult to implement in the R language. Based on these considerations,
390 the SCDV method was considered to be an optimal method that could be executed in a
391 general bioinformatics analysis environment while capturing the characteristics of gene
392 signatures.

393 However, there are several problems with this approach. First, it is difficult to determine
394 whether approximately 5,000 gene signatures are sufficient as training data. In NLP, tens of
395 thousands of data are generally used as training data. However, inadvertently mixing different

gene signatures to increase training data (e.g., other non-curation-based collections published in MSigDB) can adversely affect the quality of the signature vector. Further enhancement of pathway data with clear biological meaning is thus necessary.

Second, NLP can identify many words that appear in a specific document as important, but gene signatures do not duplicate genes in signatures; thus, the weight of important genes may be insufficient. This may create a discrepancy with human intuition regarding the key gene in the gene signature (pathway).

The third problem is a general problem in machine learning and artificial intelligence [34]. The relationship between signatures indicated by GsVec has strong elements that cannot be expressed by direct gene duplication; thus, it may be difficult to specify the rationale. Therefore, it may be desirable to combine GsVec with a well-grounded Fisher or other statistical method instead of using it on its own.

Despite the aforementioned problems, the proposed method demonstrates results that are equivalent or superior to those of conventional methods, and has high potential. Training data improvements, feature vectorization and topicalization methods, and identification of important genes are examples of potential improvements.

In the future, if the pathway database is generalized considering the direction of the regulatory relationship of genes, NLP methods that focus on context and learn to sequence from the beginning of sentences can also be applied in this field. Several NLP platforms are

already able to graph regulatory relationships between genes (e.g., IBM Watson for Drug Discovery). Improvements in the accuracy of these platforms will increase the value of methods that use NLP and are capable of biological interpretation close to human performance.

Methods

Preparation of training and validation data

A total of 5,254 gene signatures were used as a learning set whose genes fell within the range of 10 to 500 genes from KEGG and REACTOME in the C2 canonical pathway and C5 GO biological process sets in MSigDB. Similarly, 214 gene signatures of BioCarta in the C2 curated gene set were used as the validation set. For these signatures, only relevant gene signatures were extracted from the gtm file provided by MSigDB and saved in the same gtm format as MSigDB.

The following operations were performed using the R language integrated environment Microsoft R version 3.5 and R studio version 1.1.463. The created gtm file was converted to data.frame listing each gene signature's unique ID, name, description, number of genes, and gene symbol, and was output as a text file. The above operations can be executed in one step as the original R functions *make_train.data* (for training data) and *make_validation.data* (for validation data).

434

435 **Creation of gene vectors**

436 Gene feature vectors (gene vectors) were created using the R fastText package [35]. The
437 number of characters used for subwording and the number of preceding and following words
438 analyzed as related words was set to 10,000, and gene vectors were created by the
439 co-occurrence of the entire gene signature without using functions for subwording, preceding
440 and following words. First, a 1/0 matrix (one-hot vector) of genes × gene signatures based on
441 the presence/absence of corresponding genes was created. Then, a large 0/1 matrix of
442 combinations of the number of genes and gene signatures was formed. The matrix was
443 compressed to a low dimension by the skip-gram model using negative sampling for the
444 co-occurrence probability of genes that appeared simultaneously in the gene signature. It was
445 then designated as a gene vector. Fig 2 presents gv_i, a vector of an arbitrary gene g_i.

446 The number of dimensions to be compressed (i.e., gene vector length) and the number of
447 learning iterations (epoch number) had to be adjusted according to the number of vocabularies
448 in the NLP analysis. Because of the advanced analysis, the number of dimensions was set to
449 150 and the number of epochs to 100 (see S1 and S2 Figs). The above operations can be
450 executed in one step as an original R function, *gs.train_genevec*.

451

452 **Creation of gene-topic vectors**

453 Topics (clusters) included in these gene signatures were extracted from the created gene
454 vector by soft clustering using the Gaussian mixture model (GMM). For the GMM analysis,
455 *mclust* of the R mclust package [16] was used, and the number of clusters was estimated using
456 the *mclustBIC* function. The GMM analysis was performed by the *mclust* function with the
457 number of clusters determined by the Bayesian information criterion (BIC), and the
458 probability that each gene contributed to each cluster was calculated. This probability was
459 multiplied for each cluster by the previous gene vector to obtain a gene-cluster vector. Fig 2
460 demonstrates $gcv_{ik} = gv_i \times P(c_k \mid g_i)$, where c represents a cluster and K represents the
461 number of clusters.

462 Separately, a score to reduce the weight of genes that appeared in various signatures was
463 calculated by determining the value obtained by dividing the total number of signatures for
464 each gene by the number of signatures that contained the gene from the one-hot vector of the
465 genes × gene signatures (hereinafter referred to as the *inverse signature factor*). The following
466 equation was expressed as a function of R as a countermeasure to infiniteness; when the gene
467 was 0, the weights were normalized.

468 $$isf(g_i) = log\frac{s_N}{sf(g_i)} + 1$$

469 Here, S_N represents the total number of gene signatures, while $Sf(g_i)$ represents the
470 number of gene signatures including the gene g_i.

471 By multiplying this value by the previous gene-cluster vector, gene-topic vectors

reflecting the height of gene contribution to each topic and the gene specificity weight were calculated. Fig 2 displays $gtv_i = isf(g_i) \times \bigoplus_{k=1}^{K} gcv_{ik}$, where $\bigoplus$ represents concatenation.

In addition, the *estimate_cluster_size* function for estimating the number of clusters from the gene vector and the *gs.train_topicvec* function for creating the gene-cluster and gene-topic vectors were created, making executable in one step. The validation results of the parameters for estimating the number of clusters are presented in Supplemental Fig 3.

Creation of signature vector

Training data or validation data were input to the generated gene-topic vectors, and the gene-topic vectors were averaged for the genes included in each gene signature. As a result, each signature-specific feature vector (hereinafter referred to as the signature vector) was created, taking into account the gene specificity and the relevance of the gene to the topic in the training data. Fig 2 demonstrates $sv = \sum_{i=0}^{j} gtv_i$. The above operation can be performed with an original function, *predict_GsVec_from.TopicVec*.

It should be noted that the original SCDV method of NLP, which is the basis of this method, can increase the speed and accuracy using the sparse method [9]. However, in gene signature analysis, the number of genes corresponding to the number of vocabularies is overwhelmingly small compared to natural language; thus, this step was excluded because the

491 above procedure neither increased speed nor improved accuracy.

492

493 **Association between signature vectors by GsVec**

494 The association between the signature data of the training and validation data was
495 calculated based on the cosine similarity score of the signature vector. Depending on the
496 combination of training and validation data, a large amount of computation is required; thus,
497 the existing cosine distance function of the R package was not used, and high-speed program
498 code was created by original matrix computation.

499 This operation can be performed with the original functions *similarity_vectors* and
500 *GSVEC*. However, while the former outputs minimal results, the latter is a comprehensive
501 function with various options, such as adding annotations of the original signature and
502 simultaneously outputting the results of Fisher.

503

504 **Conventional pathway enrichment analysis by Fisher's exact test**

505 Fisher's exact test created and implemented its own function, *gs.enrich_fisher*, to
506 perform comprehensive processing between gene signatures using the *fisher.test* function of
507 the R stats package.

508

509 **Visualization of GsVec results with tSNE**

510 To visualize the similarity by the cosine distance between the signature vectors of
511 GsVec, the cosine distance matrix was first linearly compressed with principal component
512 analysis (PCA) using the *prcomp* function of the R stats package. The top 95% of the
513 principal components were projected in two dimensions using the *Rtsne* function in the R
514 Rtsne package [36] and visualized using the R ggplot2 package [37]. This series of operations
515 can be executed with the original function *pca.tsne_GsVec*.
516
517 **Extraction of DEGs from public gene expression data**
518 The DEGs of representative diseases were selected from several datasets in which the gene
519 expression of the appropriate disease site was used, and which was a representative disease
520 from various disease areas among the already calculated DEGs published in the Expression
521 Atlas [19]. With regard to cancer, the Expression Atlas did not provide an appropriate dataset;
522 therefore, cancer tissue and matched normal tissue datasets of major cancer types were taken
523 from the TCGA database [20]. TCGA data were normalized from the RNA-seq count data
524 using the *voom* method in the R limma package, and statistically tested by the experimental
525 Bayes method. DEGs with a false discovery rate -adjusted P-value of 0.001 or less and a fold
526 change of ±2 or more were extracted.
527
528 **Publishing program codes**

529 The program code for the GsVec analysis developed in this study is freely available
530 from https://github.com/yuumio/GsVec
531

532 ## Acknowledgments

533 We would like to show our greatest appreciation to Prof. Masami Hagiya, Dr. Rei
534 Kawakami, and Dr. Toshiaki Nakazawa of the University of Tokyo AI Data Frontier Course
535 who taught us about AI, machine learning, and NLP. We would like to offer special thanks to
536 Dr. Shinichi Kondo and Dr. Shuji Sato for supporting this research. The authors would like to
537 thank Enago (www.enago.jp) for the English language review.
538

539 ## References

540 1. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression
541 and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10.
542 2. Jin L, Zuo XY, Su WY, Zhao XL, Yuan MQ, et al. Pathway-based Analysis Tools for
543 Complex Diseases: A Review. Genomics Proteomics Bioinformatics. 2014
544 Oct;12(5):210-20.
545 3. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al.
546 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide
547 expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50.

4. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". Nucleic Acids Res. 2009 Jan;37(1):1-13.

5. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. Bioinformatics. 2012 Sep 15;28(18):i451-i457.

6. Quoc Le, Tomas Mikolov. Distributed Representations of Sentences and Documents. PMLR. 2014 32(2):1188-96.

7. Jey Han Lau, Timothy Baldwin. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. arXiv:1607.05368v1. (Submitted on 19 Jul 2016) Available from: https://arxiv.org/abs/1607.05368v1

8. Andrew M Dai, Christopher Olah, Quoc V Le. Document Embedding with Paragraph Vectors. arXiv:1507.07998v1. (Submitted on 29 Jul 2015) Available from: https://arxiv.org/abs/1507.07998v1

9. Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, Harish Karnick. SCDV: Sparse Composite Document Vectors using soft clustering over distributional representations. arXiv:1612.06778v3. (Submitted on 20 Dec 2016 (v1), last revised 12 May 2017 (this version, v3)) Available from: https://arxiv.org/abs/1612.06778v3

10. Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation. 2004 October 60(5):503-20.

11. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781v3. (Submitted on 16 Jan 2013 (v1), last revised 7 Sep 2013 (this version, v3)) Available from: https://arxiv.org/abs/1301.3781v3

12. Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, et al. Recursive deep models for semantic compositionality over a sentiment treebank. Empirical methods in NLP, 1631, 1642. 2013.

13. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23;1(6):417-25.

14. Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. Enriching Word Vectors with Subword Information. arXiv:1607.04606v2. (Submitted on 15 Jul 2016 (v1), last revised 19 Jun 2017 (this version, v2)) Available from: https://arxiv.org/abs/1607.04606v2

15. Chris Fraley, Adrian E Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. Journal of Classification. 2007, 24, Issue 2, pp 155–81.

16. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. 2016 Aug;8(1):289-317.

17. Ashok Koujalagi. Determine Word Relevance in Document Queries Using TF-IDF.

International Journal of Scientific Research. 2015, 4: 8.

18. van der Maaten, LJP, Hinton, GE. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research. 2008. 9(nov), 2579-2605.

19. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 2016 Jan 4;44(D1):D746-52.

20. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct 4;490(7418):61-70.

21. Li R, Patterson KR, Bar-Or A. Reassessing B cell contributions in multiple sclerosis. Nat Immunol. 2018 Jul;19(7):696-707.

22. Greenfield AL, Hauser SL. B-cell Therapy for Multiple Sclerosis: Entering an era. Ann Neurol. 2018 Jan;83(1):13-26.

23. Zundler S, Becker E, Schulze LL, Neurath MF. Immune cell trafficking and retention in inflammatory bowel disease: mechanistic insights and therapeutic advances. Gut. 2019 Sep;68(9):1688-1700.

24. Kuenzig ME, Barnabe C, Seow CH, Eksteen B, Negron ME, et al. Asthma Is Associated With Subsequent Development of Inflammatory Bowel Disease: A Population-based Case-Control Study. Clin Gastroenterol Hepatol. 2017 Sep;15(9):1405-12.e3.

25. Eguchi K, Nagai R. Islet inflammation in type 2 diabetes and physiology. J Clin Invest.

2017 Jan 3;127(1):14-23.

26. Falanga V. Wound healing and its impairment in the diabetic foot. Lancet. 2005 Nov 12;366(9498):1736-43.

27. Kimball AS, Davis FM, denDekker A, Joshi AD, Schaller MA, et al. The Histone Methyltransferase Setdb2 Modulates Macrophage Phenotype and Uric Acid Production in Diabetic Wound Repair. Immunity. 2019 Aug 20;51(2):258-71.e5.

28. Ogura Y, Tajrishi MM, Sato S, Hindi SM, Kumar A. Therapeutic potential of matrix metalloproteinases in Duchenne muscular dystrophy. Front Cell Dev Biol. 2014 Apr 1;2:11.

29. Shi S, de Gorter DJ, Hoogaars WM, 't Hoen PA, ten Dijke P. Overactive bone morphogenetic protein signaling in heterotopic ossification and Duchenne muscular dystrophy. Cell Mol Life Sci. 2013 Feb;70(3):407-23.

30. Crouser ED, Culver DA, Knox KS, Julian MW, Shao G, Abraham S, et al. Gene expression profiling identifies MMP-12 and ADAMDEC1 as potential pathogenic mediators of pulmonary sarcoidosis. Am J Respir Crit Care Med. 2009 May 15;179(10):929-38.

31. Rastogi R, Du W, Ju D, Pirockinaite G, Liu Y, Nunez G, Samavati L. Dysregulation of p38 and MKP-1 in response to NOD1/TLR4 stimulation in sarcoid bronchoalveolar cells. Am J Respir Crit Care Med. 2011 Feb 15;183(4):500-10.

624 32. Grunewald J, Spagnolo P, Wahlström J, Eklund A. Immunogenetics of Disease-Causing
625 Inflammation in Sarcoidosis. Clin Rev Allergy Immunol. 2015 Aug;49(1):19-35.
626 33. McCutcheon RA, Abi-Dargham A, Howes OD. Schizophrenia, Dopamine and the
627 Striatum: From Biology to Symptoms. Trends Neurosci. 2019 Mar;42(3):205-220.
628 34. Ruth Fong, Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful
629 Perturbation. arXiv:1704.03296v3. (Submitted on 11 Apr 2017 (v1), last revised 10 Jan
630 2018 (this version, v3)) Available from: https://arxiv.org/abs/1704.03296v3
631 35. Schwendinger F. fastTextR: An Interface to the 'fastText' Library. R CRAN 2017
632 36. Krijthe, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut
633 Implementation. 2015. Available from: https://github.com/jkrijthe/Rtsne
634 37. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
635 2016
636 38. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. limma powers differential
637 expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res.
638 2015 Apr 20;43(7):e47.

639

640 ## Supporting information

641 **S1 Fig. Examination of dimensionality condition in gene vectors.** Gene vectors were
642 created for five vector sizes of the gene vectors: 50, 100, 150, 200, and 250. A) Similarity

643 with the validation data was calculated. The distribution is illustrated in a violin plot, while
644 the Fisher result is presented in a scatter plot. B) The Pearson correlation coefficient with
645 Fisher for each vector size is presented as a bar graph.
646
647 **S2 Fig. Examination of epoch number in gene vectors.** Gene vectors were created for 16
648 levels: 1, 5, 10, 15, 20, 30, 40, 50, 75, 100, 250, 500, 1,000, 1,500, and 2,000. A) Similarity
649 with the validation data was calculated. The distribution is presented in a violin plot, while the
650 Fisher result is presented in a scatter plot. B) The Pearson correlation coefficient with Fisher
651 for each vector size is illustrated as a bar graph.
652
653 **S3 Fig. Validation of parameters for estimating the number of clusters as topics in gene**
654 **signatures.** The output of the *mclustBIC* function of the R mclust package was visualized by
655 the *plot* function. The Bayesian information criterion in the mclust package is 2 × log
656 likelihood. Thus, the largest value was selected as the optimal cluster.
657

	Natural Language	Bioinformatics
Element	Word	Gene
Aggregation	Documents	Gene Signatures
Utilization	Understanding and classifying the meaning of documents	**Understanding the mechanisms of life phenomena and diseases**

Natural Language Processing

Figure 1

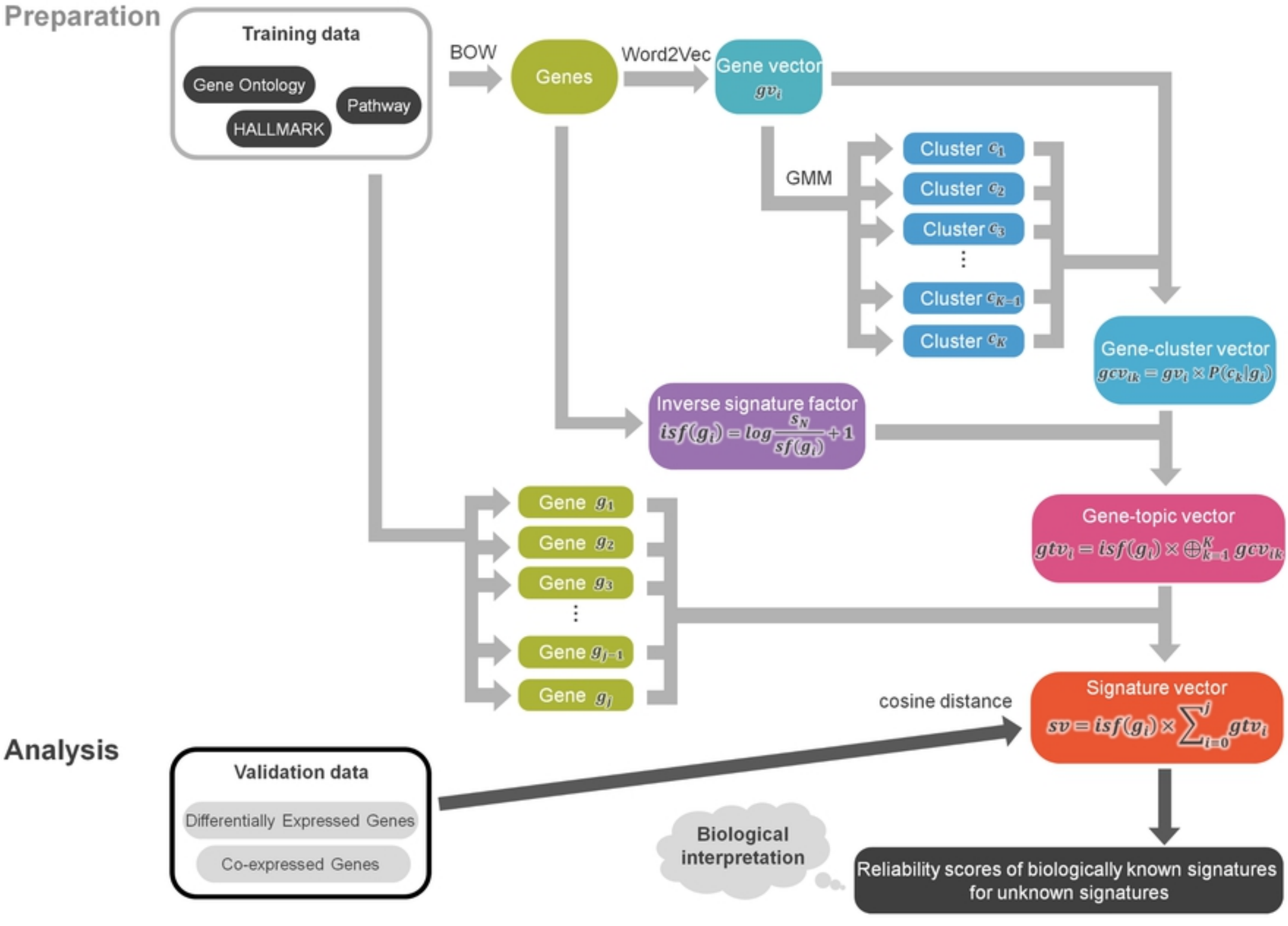

Preparation
Training data
Gene Ontology
Pathway
HALLMARK
BOW
Genes
Word2Vec
Gene vector
gv_i
GMM
Cluster c_1
Cluster c_2
Cluster c_3
Cluster c_{K-1}
Cluster c_K
Gene-cluster vector
$gcv_{ik} = gv_i \times P(c_k|g_i)$
Inverse signature factor
$isf(g_i) = log\frac{s_N}{sf(g_i)} + 1$
Gene g_1
Gene g_2
Gene g_3
Gene g_{j-1}
Gene g_j
Gene-topic vector
$gtv_i = isf(g_i) \times \oplus_{k=1}^{K} gcv_{ik}$
Signature vector
$sv = isf(g_i) \times \sum_{i=0}^{j} gtv_i$
cosine distance
Analysis
Validation data
Differentially Expressed Genes
Co-expressed Genes
Biological interpretation
Reliability scores of biologically known signatures for unknown signatures

Figure 2

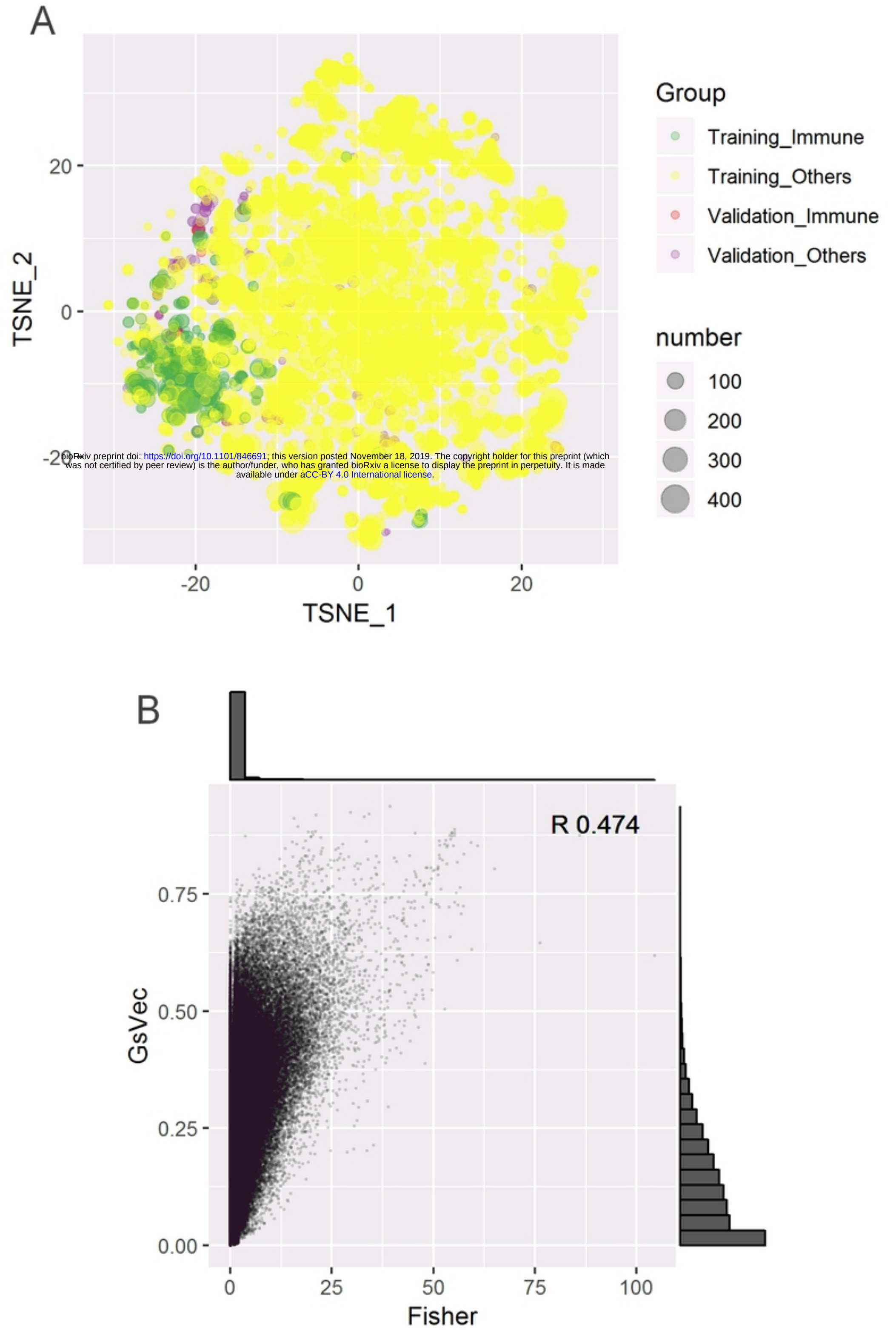


Figure 3

BIOCARTA_IL1R_PATHWAY (33 genes)

GsVec

Training gene signature name	Genes	Overlap	GsVec
PID_IL1_PATHWAY	34	19	0.833
REACTOME_IL1_SIGNALING	39	16	0.824
REACTOME_SIGNALING_BY_ILS	107	18	0.781
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	102	21	0.733
KEGG_LEISHMANIA_INFECTION	72	16	0.710
REACTOME_JNK_C_JUN_KINASES_PHOSPHORYLATION_AND_ACTIVATION_MEDIATED_BY_ACTIVATED_HUMAN_TAK1	16	6	0.701
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	270	20	0.692
KEGG_RIG_I_LIKE_RECEPTOR_SIGNALING_PATHWAY	71	13	0.689
REACTOME_ACTIVATED_TAK1_MEDIATES_P38_MAPK_ACTIVATION	18	8	0.683
GO_INTERLEUKIN_1_MEDIATED_SIGNALING_PATHWAY	13	6	0.668
GO_CYTOPLASMIC_PATTERN_RECOGNITION_RECEPTOR_SIGNALING_PATHWAY	33	10	0.660
REACTOME_ACTIVATED_TLR4_SIGNALLING	93	18	0.645
GO_CYTOKINE_MEDIATED_SIGNALING_PATHWAY	452	15	0.640
GO_LIPOPOLYSACCHARIDE_MEDIATED_SIGNALING_PATHWAY	31	7	0.634
REACTOME_TRIF_MEDIATED_TLR3_SIGNALING	74	14	0.633

Fisher

Training gene signature name	Genes	Overlap	Fisher
PID_IL1_PATHWAY	34	19	44.145
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	102	21	38.021
REACTOME_IL1_SIGNALING	39	16	33.915
REACTOME_MYD88_MAL_CASCADE_INITIATED_ON_PLASMA_MEMBRANE	83	18	32.558
REACTOME_ACTIVATED_TLR4_SIGNALLING	93	18	31.573
REACTOME_TRAF6_MEDIATED_INDUCTION_OF_NFKB_AND_MAP_KINASES_UPON_TLR7_8_OR_9_ACTIVATION	77	17	30.765
KEGG_MAPK_SIGNALING_PATHWAY	267	22	30.734
REACTOME_SIGNALING_BY_ILS	107	18	30.377
REACTOME_TOLL_RECEPTOR_CASCADES	118	18	29.552
KEGG_LEISHMANIA_INFECTION	72	16	28.893
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	270	20	26.602
REACTOME_INNATE_IMMUNE_SYSTEM	279	20	26.311
KEGG_APOPTOSIS	88	15	25.092
REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIGNALING_REPERTOIRE	72	14	24.211
REACTOME_TRIF_MEDIATED_TLR3_SIGNALING	74	14	24.029

BIOCARTA_BCR_PATHWAY (37 genes)

GsVec

Training gene signature name	Genes	Overlap	GsVec
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY	75	24	0.767
PID_BCR_5PATHWAY	65	24	0.733
KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	108	18	0.673
PID_FCER1_PATHWAY	62	18	0.608
REACTOME_ANTIGEN_ACTIVATES_B_CELL_RECEPTOR_LEADING_TO_GENERATION_OF_SECOND_MESSENGERS	29	14	0.608
KEGG_VEGF_SIGNALING_PATHWAY	76	16	0.606
PID_NFAT_3PATHWAY	54	11	0.569
PID_AVB3_OPN_PATHWAY	31	7	0.564
KEGG_GNRH_SIGNALING_PATHWAY	101	16	0.562
PID_ERBB2_ERBB3_PATHWAY	44	13	0.557
KEGG_FC_EPSILON_RI_SIGNALING_PATHWAY	79	16	0.543
KEGG_MAPK_SIGNALING_PATHWAY	267	20	0.541
PID_VEGFR1_2_PATHWAY	69	8	0.540
GO_CALCIUM_MEDIATED_SIGNALING	90	10	0.530
PID_TCR_CALCIUM_PATHWAY	29	7	0.526

Fisher

Training gene signature name	Genes	Overlap	Fisher
PID_BCR_5PATHWAY	65	24	49.372
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY	75	24	47.563
GO_FC_EPSILON_RECEPTOR_SIGNALING_PATHWAY	142	23	37.698
GO_IMMUNE_RESPONSE_REGULATING_CELL_SURFACE_RECEPTOR_SIGNALING_PATHWAY	323	27	37.115
PID_FCER1_PATHWAY	62	18	33.929
GO_FC_RECEPTOR_SIGNALING_PATHWAY	206	23	33.742
KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	137	20	31.337
REACTOME_ANTIGEN_ACTIVATES_B_CELL_RECEPTOR_LEADING_TO_GENERATION_OF_SECOND_MESSENGERS	29	14	29.904
KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	108	18	29.076
KEGG_VEGF_SIGNALING_PATHWAY	76	16	27.438
KEGG_FC_EPSILON_RI_SIGNALING_PATHWAY	79	16	27.141
PID_CD8_TCR_DOWNSTREAM_PATHWAY	65	15	26.315
KEGG_MAPK_SIGNALING_PATHWAY	267	20	25.286
KEGG_GNRH_SIGNALING_PATHWAY	101	16	25.283
PID_ERBB2_ERBB3_PATHWAY	44	13	24.290

BIOCARTA_TOLL_PATHWAY (37 genes)

GsVec

Training gene signature name	Genes	Overlap	GsVec
REACTOME_TOLL_RECEPTOR_CASCADES	118	31	0.853
REACTOME_ACTIVATED_TLR4_SIGNALLING	93	28	0.817
GO_MYD88_DEPENDENT_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	32	16	0.784
GO_PATTERN_RECOGNITION_RECEPTOR_SIGNALING_PATHWAY	109	24	0.769
REACTOME_MYD88_MAL_CASCADE_INITIATED_ON_PLASMA_MEMBRANE	83	26	0.760
GO_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	85	21	0.741
REACTOME_TRAF6_MEDIATED_INDUCTION_OF_NFKB_AND_MAP_KINASES_UPON_TLR7_8_OR_9_ACTIVATION	77	23	0.733
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	102	29	0.724
REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIGNALING_REPERTOIRE	72	22	0.719
REACTOME_TRAF6_MEDIATED_INDUCTION_OF_TAK1_COMPLEX	14	8	0.708
GO_I_KAPPAB_KINASE_NF_KAPPAB_SIGNALING	70	16	0.707
REACTOME_IKK_COMPLEX_RECRUITMENT_MEDIATED_BY_RIP1	10	7	0.707
REACTOME_JNK_C_JUN_KINASES_PHOSPHORYLATION_AND_ACTIVATION_MEDIATED_BY_ACTIVATED_HUMAN_TAK1	16	8	0.701
REACTOME_TRIF_MEDIATED_TLR3_SIGNALING	74	19	0.700
REACTOME_ACTIVATED_TAK1_MEDIATES_P38_MAPK_ACTIVATION	18	9	0.698

Fisher

Training gene signature name	Genes	Overlap	Fisher
REACTOME_TOLL_RECEPTOR_CASCADES	118	31	60.853
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	102	29	57.292
REACTOME_ACTIVATED_TLR4_SIGNALLING	93	28	55.790
REACTOME_MYD88_MAL_CASCADE_INITIATED_ON_PLASMA_MEMBRANE	83	26	51.755
REACTOME_INNATE_IMMUNE_SYSTEM	279	31	48.165
REACTOME_TRAF6_MEDIATED_INDUCTION_OF_NFKB_AND_MAP_KINASES_UPON_TLR7_8_OR_9_ACTIVATION	77	23	44.558
GO_PATTERN_RECOGNITION_RECEPTOR_SIGNALING_PATHWAY	109	24	43.070
REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIGNALING_REPERTOIRE	72	22	42.688
GO_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	85	21	38.336
GO_ACTIVATION_OF_INNATE_IMMUNE_RESPONSE	204	25	38.197
GO_POSITIVE_REGULATION_OF_INNATE_IMMUNE_RESPONSE	246	25	36.060
REACTOME_TRIF_MEDIATED_TLR3_SIGNALING	74	19	34.770
GO_MYD88_DEPENDENT_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	32	16	34.669
GO_REGULATION_OF_INNATE_IMMUNE_RESPONSE	357	25	31.880
GO_POSITIVE_REGULATION_OF_DEFENSE_RESPONSE	364	25	31.664

Figure 4

Disease	Tissue	Dataset	up	down
Multiple sclerosis (MS)	Cerebrospinal fluid	E-MTAB-69	15	59
Crohn Disease (CD)	Colon	E-MEXP-2083, E-GEOD-59071	55	6
Type2 diabetes (T2D)	islet of Langerhans	E-MTAB-5060, E-MEXP-1878	80	15
Duchenne muscular dystrophy (DMD)	skeletal muscle	E-GEOD-3307	153	13
Systemic lupus erythematosus (SLE)	whole blood	E-GEOD-72509	606	89
Sarcoidosis	lung	E-GEOD-16538	165	34
Schizophrenia (SCZ)	Brain Brodmann area 24	E-GEOD-78936	224	295

Figure 5

MS (up)

GsVec

Training gene signature name	Genes	Overlap	GsVec
GO_POSITIVE_REGULATION_OF_B_CELL_ACTIVATION	80	3	0.711
GO_B_CELL_RECEPTOR_SIGNALING_PATHWAY	54	3	0.699
GO_REGULATION_OF_B_CELL_ACTIVATION	121	4	0.694
GO_PHAGOCYTOSIS_ENGULFMENT	38	2	0.678
GO_PHAGOCYTOSIS_RECOGNITION	34	2	0.673
GO_MEMBRANE_INVAGINATION	48	2	0.651
GO_B_CELL_MEDIATED_IMMUNITY	99	4	0.636
GO_HUMORAL_IMMUNE_RESPONSE_MEDIATED_BY_CIRCULATING_IMMUNOGLOBULIN	69	4	0.632
GO_POSITIVE_REGULATION_OF_CELL_ACTIVATION	311	4	0.616
GO_COMPLEMENT_ACTIVATION	76	4	0.613
GO_ADAPTIVE_IMMUNE_RESPONSE_BASED_ON_SOMATIC_RECOMBINATION_OF_IMMUNE_RECEPTORS_BUILT_FROM_IMMUNOGLOBULIN_SUPERFAMILY_DOMAINS	154	4	0.612
GO_HUMORAL_IMMUNE_RESPONSE	187	5	0.612
GO_LYMPHOCYTE_MEDIATED_IMMUNITY	147	4	0.606
GO_ADAPTIVE_IMMUNE_RESPONSE	288	5	0.594
GO_LEUKOCYTE_MEDIATED_IMMUNITY	189	4	0.581

Fisher

Training gene signature name	Genes	Overlap	Fisher
GO_IMMUNE_RESPONSE_REGULATING_CELL_SURFACE_RECEPTOR_SIGNALING_PATHWAY	323	6	6.426
GO_HUMORAL_IMMUNE_RESPONSE_MEDIATED_BY_CIRCULATING_IMMUNOGLOBULIN	69	4	6.291
GO_HUMORAL_IMMUNE_RESPONSE	187	5	6.138
GO_COMPLEMENT_ACTIVATION	76	4	6.121
GO_PROTEIN_ACTIVATION_CASCADE	99	4	5.66
GO_B_CELL_MEDIATED_IMMUNITY	99	4	5.66
GO_REGULATION_OF_B_CELL_ACTIVATION	121	4	5.312
GO_ADAPTIVE_IMMUNE_RESPONSE	288	5	5.217
GO_LYMPHOCYTE_MEDIATED_IMMUNITY	147	4	4.976
GO_ADAPTIVE_IMMUNE_RESPONSE_BASED_ON_SOMATIC_RECOMBINATION_OF_IMMUNE_RECEPTORS_BUILT_FROM_IMMUNOGLOBULIN_SUPERFAMILY_DOMAINS	154	4	4.896
GO_B_CELL_RECEPTOR_SIGNALING_PATHWAY	54	3	4.727
GO_LEUKOCYTE_MEDIATED_IMMUNITY	189	4	4.546
GO_PHAGOCYTOSIS	190	4	4.537
GO_ACTIVATION_OF_IMMUNE_RESPONSE	427	5	4.39
GO_REGULATION_OF_CELL_ACTIVATION	484	5	4.131

MS (down)

GsVec

Training gene signature name	Genes	Overlap	GsVec
GO_POSITIVE_REGULATION_OF_LOCOMOTION	420	9	0.592
GO_POSITIVE_REGULATION_OF_CELL_PROJECTION_ORGANIZATION	303	2	0.590
GO_TAXIS	464	6	0.581
GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT	437	3	0.578
GO_POSITIVE_REGULATION_OF_CELL_DEVELOPMENT	472	4	0.577
GO_SECRETION_BY_CELL	486	8	0.575
GO_NEGATIVE_REGULATION_OF_TRANSPORT	458	8	0.566
GO_POSITIVE_REGULATION_OF_NEURON_DIFFERENTIATION	306	3	0.559
GO_POSITIVE_REGULATION_OF_NEURON_PROJECTION_DEVELOPMENT	232	2	0.557
GO_RESPONSE_TO_GROWTH_FACTOR	475	4	0.555
GO_CELL_MORPHOGENESIS_INVOLVED_IN_NEURON_DIFFERENTIATION	368	4	0.550
GO_NEGATIVE_REGULATION_OF_SECRETION	200	3	0.550
GO_REGULATION_OF_ANATOMICAL_STRUCTURE_SIZE	472	8	0.550
GO_NEURON_PROJECTION_MORPHOGENESIS	402	4	0.548
GO_POSITIVE_REGULATION_OF_SECRETION	370	5	0.548

Fisher

Training gene signature name	Genes	Overlap	Fisher
GO_INFLAMMATORY_RESPONSE	454	11	5.939
PID_INTEGRIN_A9B1_PATHWAY	25	4	5.613
GO_RECEPTOR_MEDIATED_ENDOCYTOSIS	231	8	5.541
PID_INTEGRIN1_PATHWAY	66	5	5.257
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	69	5	5.161
GO_RESPONSE_TO_MOLECULE_OF_BACTERIAL_ORIGIN	321	8	4.501
GO_POSITIVE_REGULATION_OF_LOCOMOTION	420	9	4.496
GO_RESPONSE_TO_OXIDATIVE_STRESS	352	8	4.217
GO_LEUKOCYTE_MIGRATION	259	7	4.209
GO_POSITIVE_REGULATION_OF_ENDOCYTOSIS	114	5	4.103
GO_RESPONSE_TO_INORGANIC_SUBSTANCE	479	9	4.055
GO_RESPONSE_TO_MAGNESIUM_ION	23	3	4.031
PID_TOLL_ENDOGENOUS_PATHWAY	25	3	3.92
GO_POSITIVE_REGULATION_OF_VASCULAR_ENDOTHELIAL_GROWTH_FACTOR_PRODUCTION	26	3	3.868

Figure 6

CD (up)

GsVec

Training gene signature name	Genes	Overlap	GsVec
GO_RESPONSE_TO_INTERFERON_GAMMA	144	9	0.576
GO_REGULATION_OF_HOMOTYPIC_CELL_CELL_ADHESION	307	9	0.563
GO_CELLULAR_RESPONSE_TO_INTERFERON_GAMMA	122	7	0.560
GO_REGULATION_OF_CELL_CELL_ADHESION	380	9	0.558
KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUCTION	48	7	0.557
GO_POSITIVE_REGULATION_OF_CELL_ADHESION	376	8	0.553
GO_LYMPHOCYTE_COSTIMULATION	78	5	0.553
GO_POSITIVE_REGULATION_OF_CELL_CELL_ADHESION	243	8	0.551
GO_INFLAMMATORY_RESPONSE	454	12	0.551
GO_LEUKOCYTE_MIGRATION	259	5	0.543
GO_REGULATION_OF_LEUKOCYTE_MIGRATION	149	6	0.543
GO_REGULATION_OF_IMMUNE_EFFECTOR_PROCESS	424	4	0.541
GO_POSITIVE_REGULATION_OF_RESPONSE_TO_EXTERNAL_STIMULUS	296	7	0.540
GO_NEGATIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS	372	5	0.529

Fisher

Training gene signature name	Genes	Overlap	Fisher
GO_CYTOKINE_MEDIATED_SIGNALING_PATHWAY	452	15	12.983
KEGG_ASTHMA	30	7	12.209
KEGG_ALLOGRAFT_REJECTION	38	7	11.423
KEGG_GRAFT_VERSUS_HOST_DISEASE	42	7	11.096
KEGG_TYPE_I_DIABETES_MELLITUS	44	7	10.945
KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUCTION	48	7	10.665
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	89	8	10.398
KEGG_AUTOIMMUNE_THYROID_DISEASE	53	7	10.348
GO_RESPONSE_TO_INTERFERON_GAMMA	144	9	10.227
GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_OR_POLYSACCHARIDE_ANTIGEN_VIA_MHC_CLASS_II	94	8	10.204
KEGG_LEISHMANIA_INFECTION	72	7	9.383
KEGG_VIRAL_MYOCARDITIS	73	7	9.340
GO_INFLAMMATORY_RESPONSE	454	12	9.214
GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN	177	8	8.007
GO_CELLULAR_RESPONSE_TO_INTERFERON_GAMMA	122	7	7.764

Figure 7

T2D (up)

GsVec

Training gene signature name	Genes	Overlap	GsVec
GO_TAXIS	464	11	0.630
GO_SINGLE_ORGANISM_CELL_ADHESION	459	7	0.612
GO_LEUKOCYTE_MIGRATION	259	7	0.611
GO_POSITIVE_REGULATION_OF_LOCOMOTION	420	15	0.607
GO_REGULATION_OF_RESPONSE_TO_WOUNDING	413	13	0.595
GO_POSITIVE_REGULATION_OF_RESPONSE_TO_EXTERNAL_STIMULUS	296	9	0.582
GO_REGULATION_OF_PEPTIDASE_ACTIVITY	392	11	0.580
GO_REGULATION_OF_INFLAMMATORY_RESPONSE	294	8	0.575
GO_POSITIVE_REGULATION_OF_MAPK_CASCADE	470	10	0.572
REACTOME_DEVELOPMENTAL_BIOLOGY	396	3	0.568
GO_REGULATION_OF_HOMEOSTATIC_PROCESS	447	5	0.562
GO_NEGATIVE_REGULATION_OF_PROTEOLYSIS	329	8	0.561
GO_EPITHELIAL_CELL_DIFFERENTIATION	495	8	0.561
GO_POSITIVE_REGULATION_OF_CELL_DEVELOPMENT	472	8	0.557
GO_INFLAMMATORY_RESPONSE	454	16	0.555

Fisher

Training gene signature name	Genes	Overlap	Fisher
GO_INFLAMMATORY_RESPONSE	454	16	10.343
GO_POSITIVE_REGULATION_OF_LOCOMOTION	420	15	9.785
PID_FRA_PATHWAY	37	6	8.058
PID_INTEGRIN1_PATHWAY	66	7	7.990
GO_REGULATION_OF_LEUKOCYTE_MIGRATION	149	9	7.957
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	267	11	7.888
GO_REGULATION_OF_RESPONSE_TO_WOUNDING	413	13	7.856
GO_POSITIVE_REGULATION_OF_LEUKOCYTE_MIGRATION	109	8	7.785
GO_REGULATION_OF_VASCULATURE_DEVELOPMENT	233	10	7.373
GO_EXTRACELLULAR_STRUCTURE_ORGANIZATION	304	11	7.310
GO_VASCULATURE_DEVELOPMENT	469	13	7.208
GO_WOUND_HEALING	470	13	7.197
GO_POSITIVE_REGULATION_OF_VASCULATURE_DEVELOPMENT	133	8	7.107
GO_REGULATION_OF_NEUTROPHIL_MIGRATION	32	5	6.703
GO_CYTOKINE_MEDIATED_SIGNALING_PATHWAY	452	12	6.478

Figure 8

DMD (up)

GsVec

Training gene signature name	Genes	Overlap	GsVec
GO_EXTRACELLULAR_STRUCTURE_ORGANIZATION	304	26	0.743
GO_OSSIFICATION	251	21	0.700
REACTOME_DEVELOPMENTAL_BIOLOGY	396	9	0.698
GO_MULTICELLULAR_ORGANISM_METABOLIC_PROCESS	93	11	0.676
KEGG_FOCAL_ADHESION	201	12	0.673
GO_CONNECTIVE_TISSUE_DEVELOPMENT	194	10	0.668
REACTOME_AXON_GUIDANCE	251	9	0.668
REACTOME_SIGNALING_BY_PDGF	122	9	0.665
GO_POSITIVE_REGULATION_OF_CELL_DEVELOPMENT	472	11	0.660
GO_MULTICELLULAR_ORGANISMAL_MACROMOLECULE_METABOLIC_PROCESS	79	11	0.656
GO_COLLAGEN_FIBRIL_ORGANIZATION	38	7	0.652
GO_SENSORY_ORGAN_DEVELOPMENT	493	14	0.649
GO_SINGLE_ORGANISM_CELL_ADHESION	459	5	0.649
GO_RESPONSE_TO_MECHANICAL_STIMULUS	210	11	0.648
GO_CARTILAGE_DEVELOPMENT	147	9	0.647

Fisher

Training gene signature name	Genes	Overlap	Fisher
GO_EXTRACELLULAR_STRUCTURE_ORGANIZATION	304	26	19.283
GO_OSSIFICATION	251	21	15.391
PID_INTEGRIN1_PATHWAY	66	12	13.046
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION	87	13	12.936
GO_SKELETAL_SYSTEM_DEVELOPMENT	455	22	11.238
PID_AVB3_INTEGRIN_PATHWAY	75	11	10.927
REACTOME_COLLAGEN_FORMATION	58	10	10.708
GO_MULTICELLULAR_ORGANISMAL_MACROMOLECULE_METABOLIC_PROCESS	79	11	10.673
PID_SYNDECAN_1_PATHWAY	46	9	10.210
GO_MULTICELLULAR_ORGANISM_METABOLIC_PROCESS	93	11	9.883
GO_WOUND_HEALING	470	20	9.262
KEGG_ECM_RECEPTOR_INTERACTION	84	10	9.058
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	69	9	8.563
GO_OSTEOBLAST_DIFFERENTIATION	126	11	8.454
GO_CELLULAR_RESPONSE_TO_AMINO_ACID_STIMULUS	53	8	8.187

Figure 9

SLE (up)

GsVec

Training gene signature name	Genes	Overlap	GsVec
GO_IMMUNE_EFFECTOR_PROCESS	486	53	0.731
GO_INFLAMMATORY_RESPONSE	454	33	0.712
GO_REGULATION_OF_MULTI_ORGANISM_PROCESS	470	43	0.694
GO_NEGATIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS	372	25	0.693
GO_RESPONSE_TO_VIRUS	247	35	0.692
GO_SINGLE_ORGANISM_CELL_ADHESION	459	15	0.685
GO_REGULATION_OF_IMMUNE_EFFECTOR_PROCESS	424	22	0.684
GO_NEGATIVE_REGULATION_OF_CYTOKINE_PRODUCTION	211	18	0.683
GO_WOUND_HEALING	470	10	0.683
GO_NEGATIVE_REGULATION_OF_DEFENSE_RESPONSE	144	12	0.681
GO_REGULATION_OF_CELL_GROWTH	391	15	0.681
GO_LEUKOCYTE_ACTIVATION	414	12	0.680
GO_REGULATION_OF_CELL_ACTIVATION	484	21	0.679
GO_NEGATIVE_REGULATION_OF_RESPONSE_TO_EXTERNAL_STIMULUS	274	16	0.679
GO_POSITIVE_REGULATION_OF_CYTOKINE_PRODUCTION	370	24	0.679

Fisher

Training gene signature name	Genes	Overlap	Fisher
GO_DEFENSE_RESPONSE_TO_VIRUS	164	32	16.093
GO_IMMUNE_EFFECTOR_PROCESS	486	53	14.606
GO_NEGATIVE_REGULATION_OF_VIRAL_GENOME_REPLICATION	49	17	13.246
GO_RESPONSE_TO_VIRUS	247	35	13.106
REACTOME_INTERFERON_SIGNALING	159	28	12.976
GO_RESPONSE_TO_TYPE_I_INTERFERON	68	19	12.783
GO_NEGATIVE_REGULATION_OF_MULTI_ORGANISM_PROCESS	151	27	12.701
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	64	18	12.197
GO_HUMORAL_IMMUNE_RESPONSE	187	29	11.959
GO_REGULATION_OF_SYMBIOSIS_ENCOMPASSING_MUTUALISM_THROUGH_PARASITISM	205	30	11.682
GO_NEGATIVE_REGULATION_OF_VIRAL_PROCESS	89	20	11.482
GO_REGULATION_OF_VIRAL_GENOME_REPLICATION	75	18	10.912
GO_CYTOKINE_MEDIATED_SIGNALING_PATHWAY	452	44	10.523
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	270	32	9.941
GO_REGULATION_OF_MULTI_ORGANISM_PROCESS	470	43	9.439

Figure 10

Sarcoidosis (up)

GsVec

Training gene signature name	Genes	Overlap	GsVec
GO_POSITIVE_REGULATION_OF_MAPK_CASCADE	470	5	0.666
GO_POSITIVE_REGULATION_OF_KINASE_ACTIVITY	482	6	0.652
GO_POSITIVE_REGULATION_OF_PROTEIN_SERINE_THREONINE_KINASE_ACTIVITY	289	4	0.625
GO_REGULATION_OF_MAP_KINASE_ACTIVITY	319	3	0.625
GO_ACTIVATION_OF_PROTEIN_KINASE_ACTIVITY	279	3	0.617
GO_POSITIVE_REGULATION_OF_CELLULAR_COMPONENT_BIOGENESIS	406	7	0.614
GO_POSITIVE_REGULATION_OF_MAP_KINASE_ACTIVITY	207	3	0.608
GO_RESPONSE_TO_GROWTH_FACTOR	475	5	0.607
GO_WOUND_HEALING	470	7	0.606
GO_REGULATION_OF_PROTEIN_SERINE_THREONINE_KINASE_ACTIVITY	470	5	0.601
GO_TRANSMEMBRANE_RECEPTOR_PROTEIN_TYROSINE_KINASE_SIGNALING_PATHWAY	498	2	0.599
GO_SIGNAL_TRANSDUCTION_BY_PROTEIN_PHOSPHORYLATION	404	5	0.597
GO_NEGATIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUCTION	437	5	0.596
GO_PROTEIN_AUTOPHOSPHORYLATION	192	2	0.594
GO_REGULATION_OF_PROTEIN_SECRETION	389	10	0.592

Fisher

Training gene signature name	Genes	Overlap	Fisher
GO_CYTOKINE_PRODUCTION	120	9	6.043
GO_INFLAMMATORY_RESPONSE	454	16	5.735
GO_CHEMOKINE_MEDIATED_SIGNALING_PATHWAY	72	7	5.568
GO_RESPONSE_TO_INTERFERON_GAMMA	144	9	5.383
KEGG_CHEMOKINE_SIGNALING_PATHWAY	190	10	5.252
REACTOME_G_ALPHA_I_SIGNALLING_EVENTS	195	10	5.153
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	57	6	5.047
GO_CELLULAR_RESPONSE_TO_INTERFERON_GAMMA	122	8	4.996
GO_CELL_CHEMOTAXIS	162	9	4.967
REACTOME_INTERFERON_GAMMA_SIGNALING	63	6	4.793
GO_CYTOKINE_SECRETION	38	5	4.767
GO_LYMPHOCYTE_CHEMOTAXIS	38	5	4.767
GO_REGULATION_OF_CHEMOTAXIS	180	9	4.602
GO_POSITIVE_REGULATION_OF_LEUKOCYTE_MIGRATION	109	7	4.377
GO_LYMPHOCYTE_MIGRATION	49	5	4.220

Figure 11

SCZ (up)

GsVec

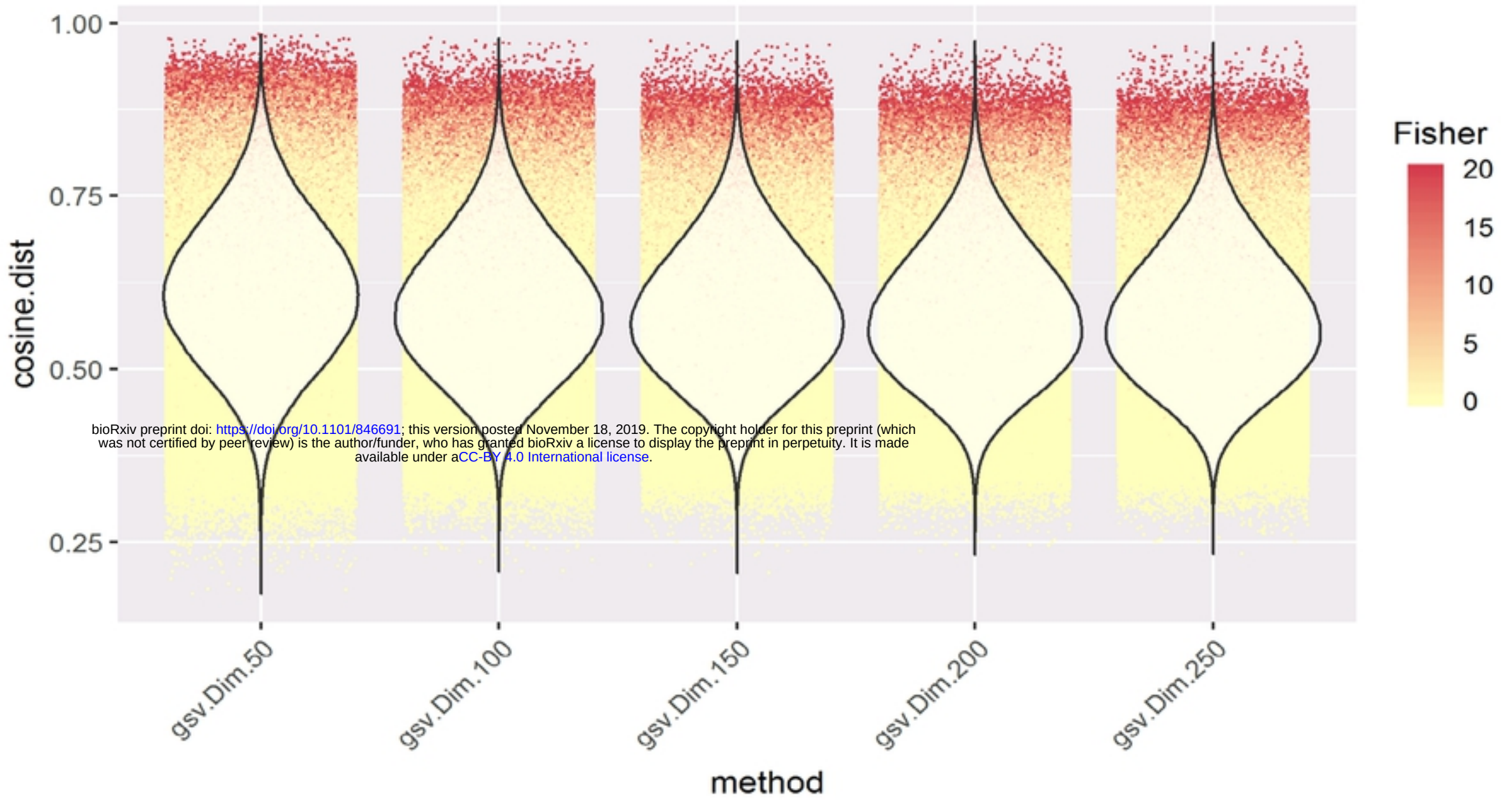
Training gene signature name	Genes	Overlap	GsVec
GO_REGULATION_OF_HORMONE_LEVELS	478	7	0.705
GO_RESPONSE_TO_DRUG	431	10	0.685
GO_RESPONSE_TO_INORGANIC_SUBSTANCE	479	8	0.667
GO_RESPONSE_TO_STEROID_HORMONE	497	8	0.662
GO_POSITIVE_REGULATION_OF_CELL_DEVELOPMENT	472	9	0.658
GO_EPITHELIAL_CELL_DIFFERENTIATION	495	14	0.657
GO_VASCULATURE_DEVELOPMENT	469	15	0.656
GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT	437	8	0.649
GO_AGING	264	6	0.648
GO_RESPONSE_TO_ALCOHOL	362	9	0.647
GO_NEGATIVE_REGULATION_OF_TRANSPORT	458	7	0.647
GO_WOUND_HEALING	470	14	0.646
GO_CELLULAR_RESPONSE_TO_ORGANIC_CYCLIC_COMPOUND	465	8	0.645
GO_RESPONSE_TO_EXTRACELLULAR_STIMULUS	441	6	0.645
GO_SENSORY_ORGAN_DEVELOPMENT	493	20	0.644

Fisher

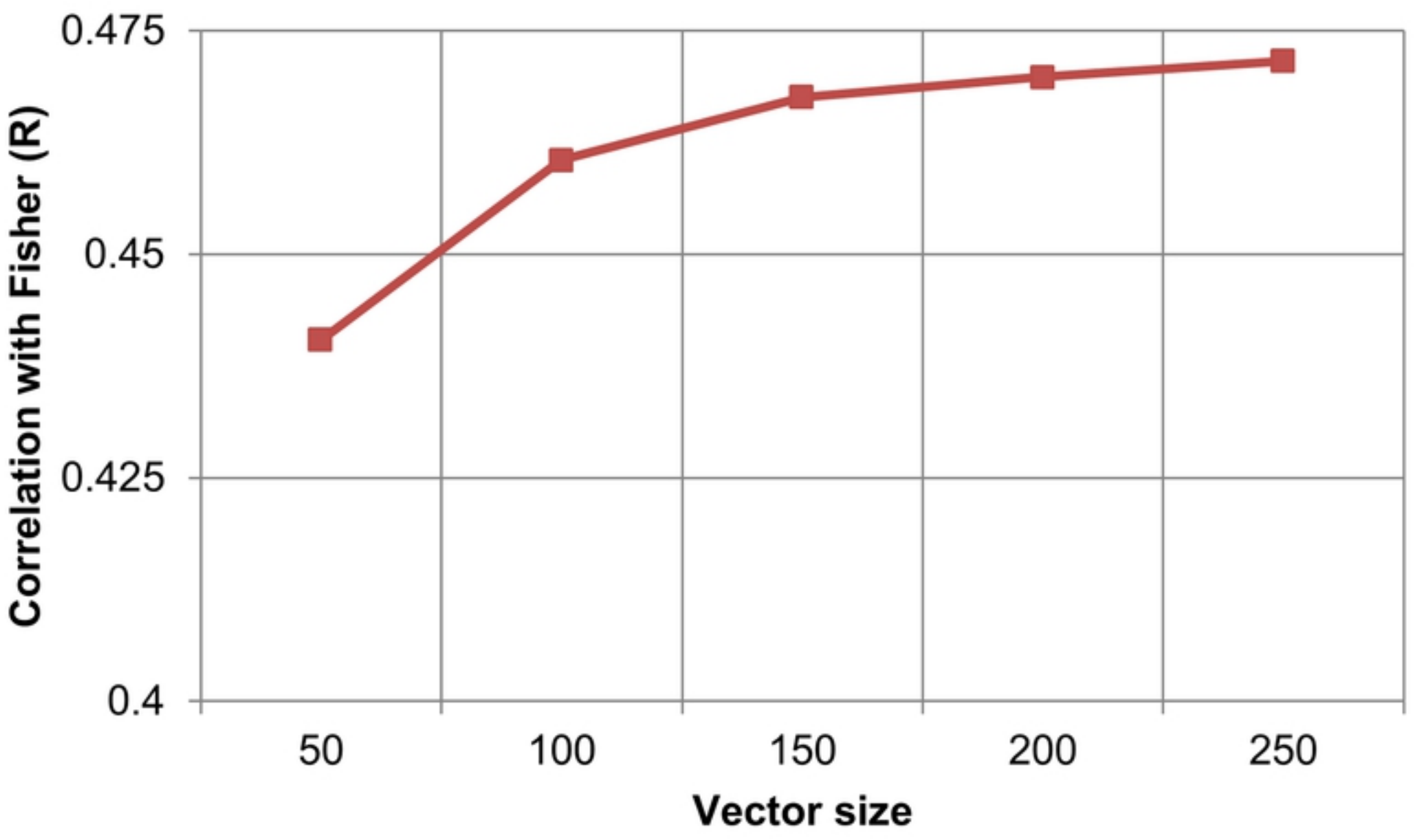
Training gene signature name	Genes	Overlap	Fisher
GO_GLIOGENESIS	175	13	6.889
GO_GLIAL_CELL_DIFFERENTIATION	136	11	6.282
GO_NEPHRON_DEVELOPMENT	115	10	6.050
GO_DORSAL_VENTRAL_PATTERN_FORMATION	91	9	5.970
GO_SENSORY_ORGAN_DEVELOPMENT	493	20	5.907
GO_NOTOCHORD_DEVELOPMENT	18	5	5.828
GO_UROGENITAL_SYSTEM_DEVELOPMENT	299	15	5.660
GO_ASTROCYTE_DIFFERENTIATION	39	6	5.279
GO_SPINAL_CORD_PATTERNING	24	5	5.157
GO_TELENCEPHALON_DEVELOPMENT	228	12	4.843
GO_CELL_FATE_SPECIFICATION	71	7	4.752
GO_BRANCHING_MORPHOGENESIS_OF_AN_EPITHELIAL_TUBE	131	9	4.664
GO_EMBRYONIC_ORGAN_MORPHOGENESIS	279	13	4.637
GO_PROXIMAL_DISTAL_PATTERN_FORMATION	32	5	4.514
GO_RESPONSE_TO_GROWTH_FACTOR	475	17	4.401

Figure 12

A

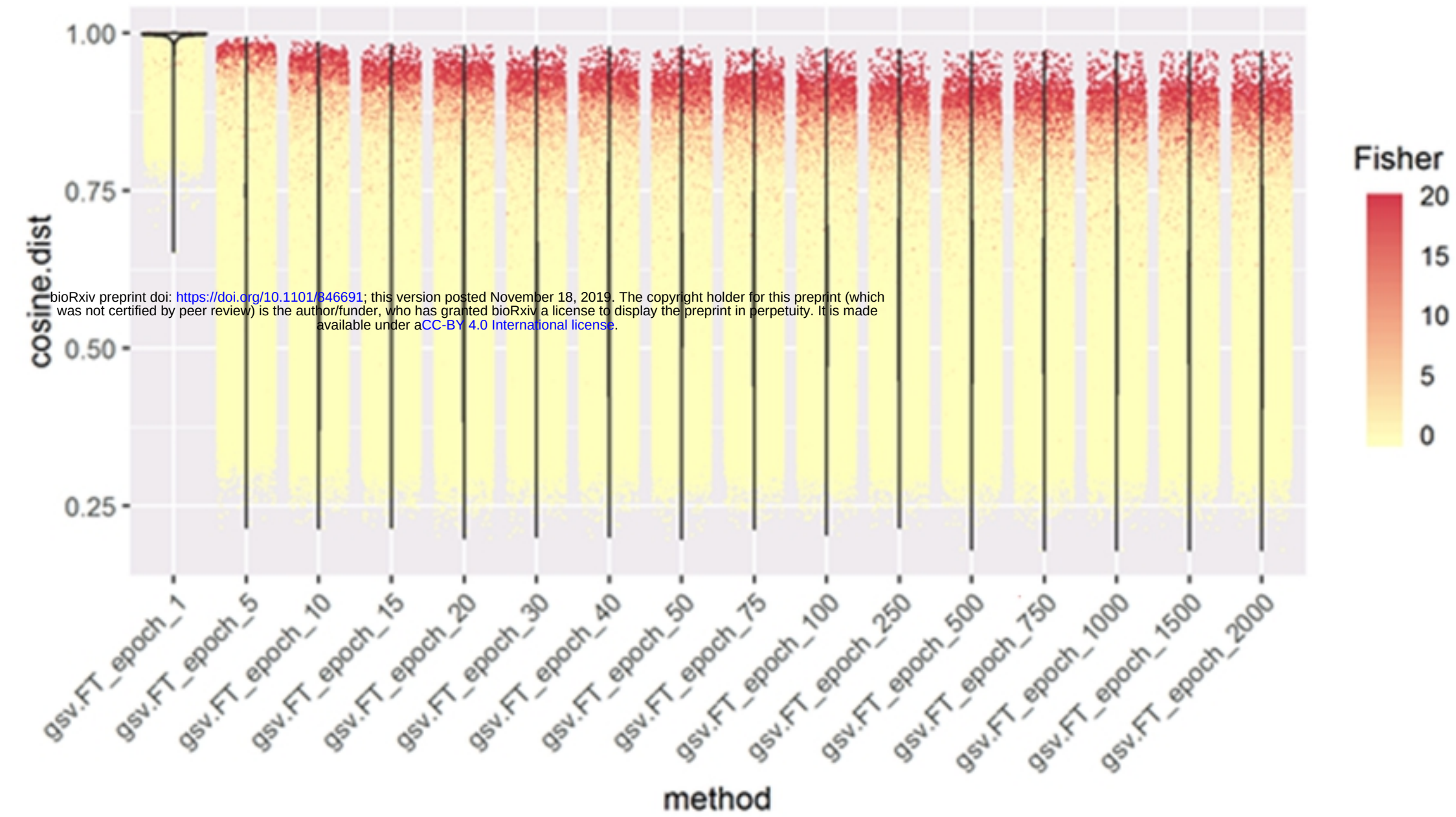


B

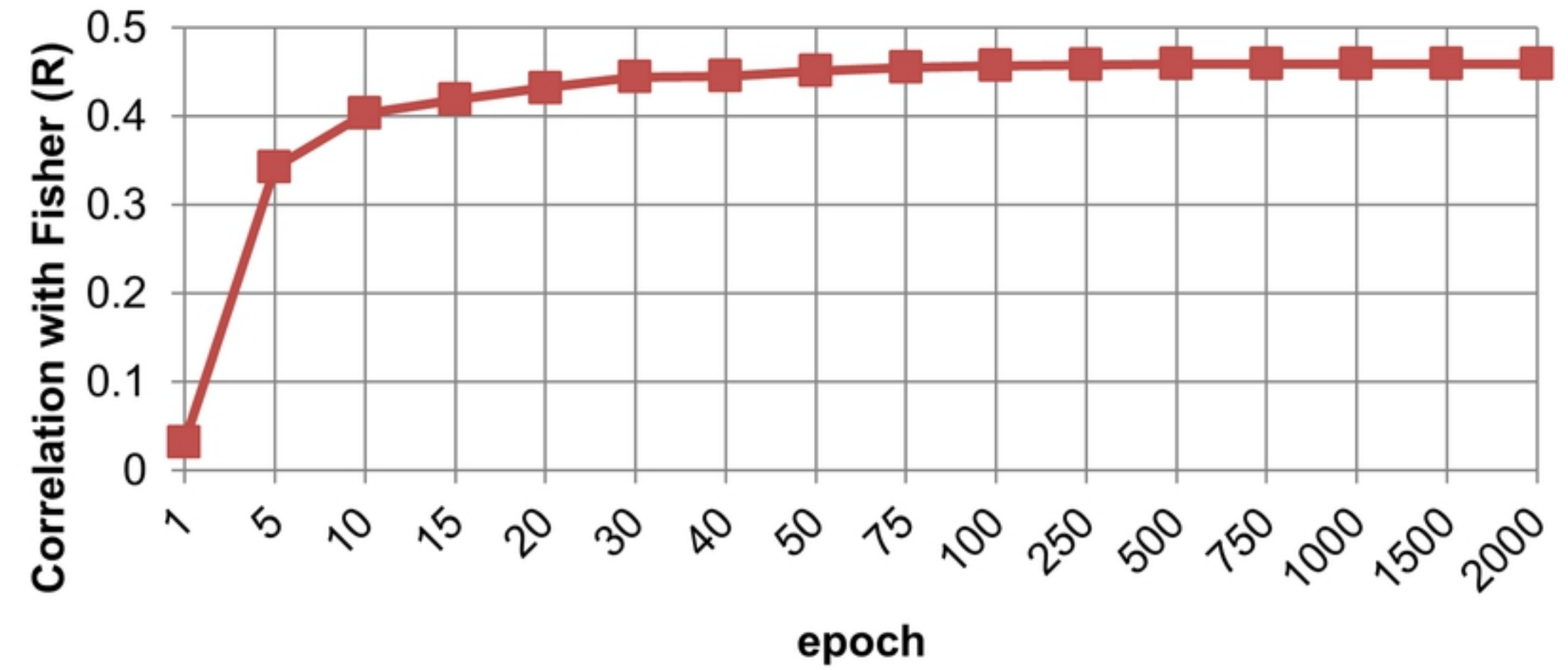


S1 Figure

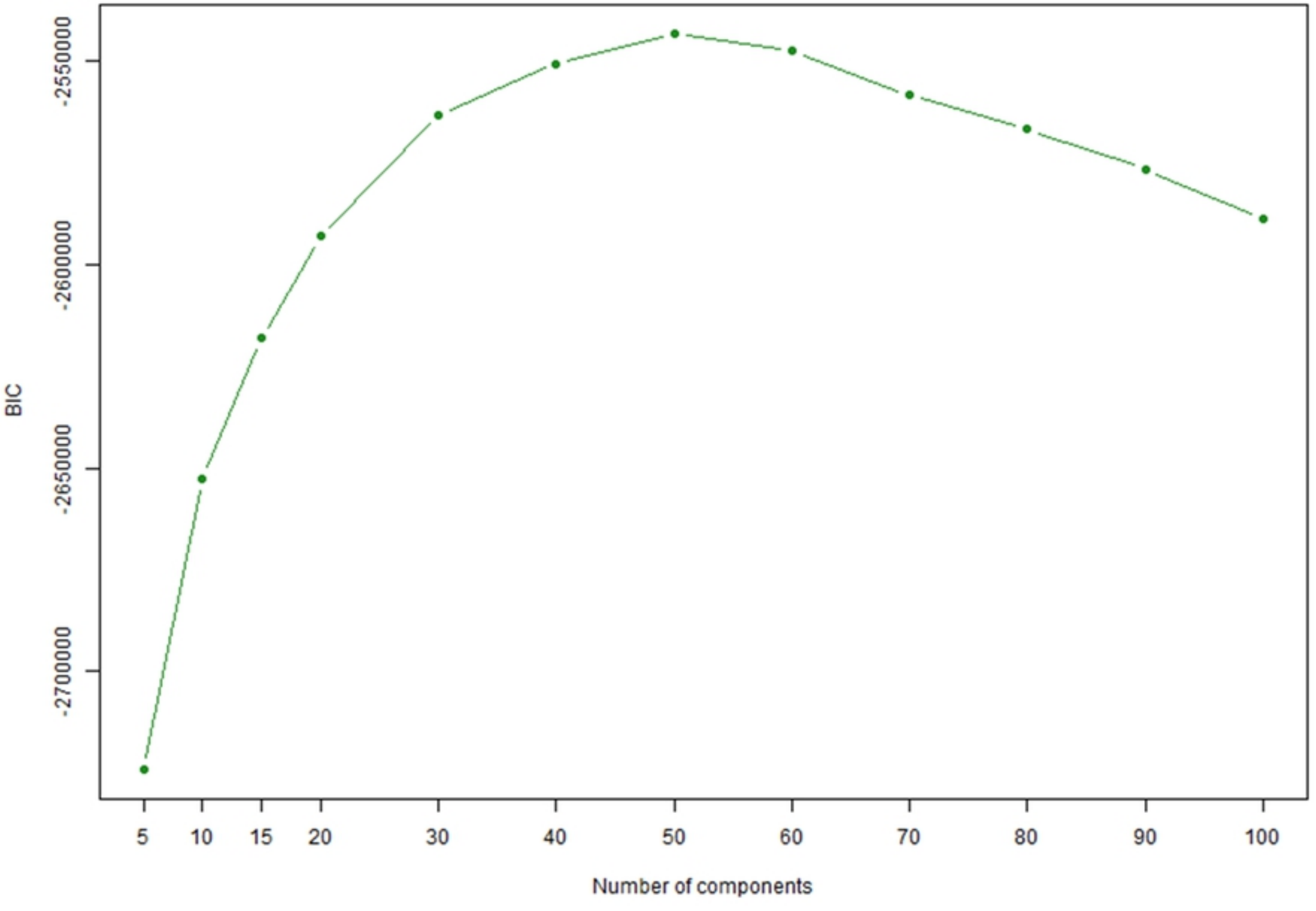
A



B



S2 Figure


BIC
-2550000
-2600000
-2650000
-2700000
5
10
15
20
30
40
50
60
70
80
90
100
Number of components

S3 Figure