High-Throughput Gene Expression Profiles to Define Drug Similarity and Predict Compound Activity

Hans De Wolf,¹ Laure Cougnaud,² Kirsten Van Hoorde,²
An De Bondt,¹ Joerg K. Wegner,¹
Hugo Ceulemans,¹ and Hinrich Göhlmann¹

¹Janssen Research & Development, A Division of Janssen
Pharmaceutica NV, Computational Sciences,
Discovery Sciences, Beerse, Belgium.

²Open Analytics, Antwerpen, Belgium.

ABSTRACT

By adding biological information, beyond the chemical properties and desired effect of a compound, uncharted compound areas and connections can be explored. In this study, we add transcriptional information for 31K compounds of Janssen's primary screening deck, using the HT L1000 platform and assess (a) the transcriptional connection score for generating compound similarities, (b) machine learning algorithms for generating target activity predictions, and (c) the scaffold hopping potential of the resulting hits. We demonstrate that the transcriptional connection score is best computed from the significant genes only and should be interpreted within its confidence interval for which we provide the stats. These guidelines help to reduce noise, increase reproducibility, and enable the separation of specific and promiscuous compounds. The added value of machine learning is demonstrated for the NR3C1 and HSP90 targets. Support Vector Machine models yielded balanced accuracy values $\geq 80\%$ when the expression values from DDIT4 & SERPINE1 and TMEM97 & SPR were used to predict the NR3C1 and HSP90 activity, respectively. Combining both models resulted in 22 new and confirmed HSP90-independent NR3C1 inhibitors, providing two scaffolds (i.e., pyrimidine and pyrazolo-pyrimidine), which could potentially be of interest in the treatment of depression (i.e., inhibiting the glucocorticoid receptor (i.e., NR3C1), while leaving its chaperone, HSP90, unaffected). As such, the initial hit rate increased by a factor 300, as less, but more specific chemistry could be screened, based on the upfront computed activity predictions.

Keywords: L1000, machine learning, transcriptional similarity

INTRODUCTION

High-throughput screening (HTS) is the initial step by which the pharmaceutical industry identifies hit compounds for a desired activity.¹ This process can be either approached in a targeted manner or by using phenotypic HTS where molecules are identified, which induce a desired phenotypic change, while the concrete mechanism of action (MOA) is initially not known. However, compounds can reveal a broad activity spectrum, beyond the desired phenotypic change they induce. Hence, counter screens, selectivity assays, and toxicity assessments are set up to prevent further development of those molecules that lack a therapeutic window (i.e., limited dose range in which the desired activity is measured with no or limited undesired effects). It is, however, desirable to devise strategies that uncover these adverse effects at early stages, reducing compound attrition rates that are still high,² Smietana_2016. As a result, pharmaceutical companies have been investigating novel ways to add compound annotation beyond the chemical structure and the on-target, desired, effect. The utility of observed phenotypic activity and other induced biological effects is explored to select and/or flag compounds at the very early steps in the drug discovery process.

High-throughput characterization of transcriptional effects in cells that are exposed to small molecules is one of the currently explored methodologies.^{3–5} Genes that are differentially expressed in this compound exposure setting can be translated into gene signatures, which can be used to discover new connections among compounds, pathways, diseases, and/or phenotypic states in general.⁶ To this end, Lamb^{6,7} generated the genome-wide gene expression profile in four human cell lines treated with 1309 compounds at different concentrations, using the Human Affymetrix U133 platform and termed this the Connectivity Map (i.e., CMap).

Since the generation of the CMap, technological advances have been made, which could revolutionize high-throughput transcriptional profiling in drug discovery and development. Using combined multiplex ligation mediated amplification with the Luminex FlexMap (Luminex, Austin, TX) optically addressed and barcoded microsphere, and flow cytometric

detection technology, Peck *et al.*⁸ developed, for instance, a cost-effective transcriptional profiling methodology in which up to 100 transcripts could be measured in HTS mode. This technology was further developed at the Broad institute to accommodate the direct measurement of 978 carefully selected human gene transcripts and infer the expression values of the remainder, based on computational models that were trained on historical gene expression profiles. Known as L1000 and commercialized by Genometry, Inc. (Cambridge, MA), this automated assay can now be used to screen and measure the transcriptionally induced effect(s) of thousands of compounds per day at a cost far below conventional transcriptomic techniques like microarrays.⁹ As such, L1000 enabled the generation of more than a million transcriptional profiles for over 19811 small molecule drugs, tool compounds, and screening library compounds across 3–77 cell lines in the NIH-funded and publicly available LINCS program.⁹

Transcriptional similarity between a query gene signature, comprising upregulated and downregulated genes, and the reference CMap/LINCS profiles can be computed by the connectivity score proposed by Lamb.⁶ The latter estimates whether the upregulated and downregulated signature query genes are, respectively, correlated or anticorrelated with the reference profiles (*i.e.*, rank based on correlation with query). This methodology has been widely used to recognize drugs with a common MOA,¹⁰ discover new MOA,⁶ flag adverse effects,¹¹ and/or reposition existing drugs.¹² Meanwhile, a variety of similarity-based methods have been proposed for the large-scale analysis of transcriptional drug response profiles, as reviewed by Cha *et al.*¹³ Based on a selection of 29 drugs, Cha *et al.*¹³ showed that the connectivity score is slightly outperformed by the connection methodology proposed by Zhang and Gant,¹⁴ which will be adopted in this article.

Next to similarity-based approaches, one can also use the rich high-dimensional transcriptional compound data to train computational models through machine learning algorithms. The latter aims to predict in which assay a novel compound is likely to be (in)active. Those can either be targeted assays (*e.g.*, biochemical protein binding assays) or phenotypic assays (*e.g.*, genotoxicity assays). Using these predictions, at early stages, one can immediately prioritize compounds and evaluate which compounds need which type of experimental follow-up. As such, computational methods are gaining momentum in the field of drug discovery, as they promise to accelerate the discovery of new therapeutic compounds, while reducing attrition rates.^{15,16}

Against this background, we profiled 31K compounds of the primary Janssen screening deck in MCF7 cells using the L1000

platform. We investigated under which conditions the transcriptional connection score, proposed by Zhang and Gant,¹⁴ is best used when ranking compounds, based on a query gene signature. The 31K, L1000 profiled compounds were subsequently used to predict their activity against the glucocorticoid receptor (GR, gene symbol: *NR3C1* - nuclear receptor subfamily 3, group C, member 1) and heat shock protein (*HSP90*) genes, using machine learning-based predictive models. These models were trained on Monte Carlo-based sampling of 115 known *NR3C1* and/or *HSP90* (in) active L1000 profiled compounds. Both *NR3C1* and *HSP90* play a crucial role in GR signaling pathways, which in turn regulate genes that control, among others, cognition (memory), stress response, metabolism, immune response, and fetal development. In the absence of glucocorticoids, the GR is packaged into multiprotein complexes consisting of heat-shock proteins (*e.g.*, *HSP90*, *HSP70*, and *HSP23*) and immunophilins (*e.g.*, *FKBP5*, also known as *FKBP51*), which bind to the receptor and keep it sequestered and inactive in the cytoplasm.¹⁷ A growing body of evidence shows that many depressed patients consistently exhibit hyperactivity of the hypothalamus–pituitary–adrenal axis, which results in increased levels of the glucocorticoid hormone cortisol.¹⁸ Hence, it has been proposed that blocking the GR with an antagonist may reduce the effects of high glucocorticoid levels on the brain, and thus represent a potential treatment strategy for depression.^{19,20}

MATERIALS AND METHODS

L1000 Assay

The experimental setup is graphically represented in Figure 1. Compounds were transcriptionally profiled in the MCF7 human breast epithelial adenocarcinoma cell line (ATCC HTB-22). After the start of each new cell culture, one flask of MCF7 cells was tested for possible mycoplasma contamination. Five thousand cells of the MCF7 were plated per well in a 384-well cell culture plate (*i.e.*, 40 μ L/well). The cells were incubated for 24 h at 37°C. When the monolayer was approximately 75% confluent, compound, positive controls (*i.e.*, LY-294002, Estradiol, tanespimycin, and trichostatin) or vehicle samples (*i.e.*, DMSO) were added to a final volume of 50 μ L/well, yielding a compound concentration of 1 μ M and a positive control concentration of 30, 0.1, 1, and 0.1 μ M respectively. The 384-well plates were stored in an incubator for 7 h at 37°C, after which 45 μ L of a lysis buffer (*i.e.*, Qiagen; 1031576) was added to each well. The plates were subsequently kept at room temperature for 30' and sealed, after which they were stored at –80°C. The lysates were shipped on dry ice to Genometry, Inc., where the L1000 assay is run. The data were preprocessed by Genometry, Inc., following their

Fig. 1. Graphical representation of the experimental setup. L1000 profiling of 31K compounds of Janssen primary screening deck in MCF7 cells and data analysis flow for computing compound similarity ("c" Zhang connection score), and machine learning-based compound activity predictive models. For clarity reasons, only five L1000 genes and two compound treatments are visualized, indicated by different compound and transcript coloring.

standard procedure of gene assignment, invariant set normalization within each sample, quantile normalization across samples within each 384-well plate, and a $\log_2$ transformation of the signal.

Transcriptional Similarity: Zhang Connection Score

The transcriptional effects for each of the 31761 samples were assessed by a Limma differential gene expression analysis.²¹ The t - and P -values from this test were computed on the $\log_2$ -transformed signal for each of the 978 measured genes versus the vehicle sample distribution within the same plate. A total of five genes were filtered out of the analysis, because the t -statistics could not be computed, due to unique signal intensities across all samples (i.e., *CCDC86*, *SLC25A14*, *NSDHL*, *ATP6 V1D*, and *CDC45L*). The P -values were moderated across all tested genes, with the variances of the vehicle samples pooled across all genes. The eBayes function of the R limma package was used for this purpose at a 5% significance level. The Benjamini-Hochberg algorithm was applied to correct for multiple testing. As such, each compound could be rewritten as a vector of t - and p -statistics. The magnitude of the compound effect is given by the P -value, while the direction of the effect is denoted by the sign of the t -value. Based on these two parameters, the 973 genes could be ranked and signed for each of the 31,761 samples.

Transcriptional similarity between a gene signature (i.e., query s) and a transcriptional profile of a compound (i.e., reference R) was computed for all genes “ g ” in the signature using the standardized ranked Zhang connection score¹⁴ c that ranges between +1 and -1 (Fig. 1):

$$c = \frac{\sum_{i=1}^m R(g_i)s(g_i)}{C_{\max}(N, m)}$$

With $C_{\max}$ the maximum possible connection score, where m is the number of genes in the query and N is the number of reference genes,

$$C_{\max}(R, s) = \sum_{i=1}^m (N - i + 1)(m - i + 1)$$

Zhang connection scores were computed for all 15456 pairwise comparisons between compounds and positive controls, affecting significantly at least one gene, using five different gene inclusion criteria for the query compound:

- All 978 signed ranked genes were used ($m=978$)
- The top five highest ranked genes were used ($m=5$)
- The significant signed ranked genes were used (m = number of significant genes) at a 0.1% significance threshold
- The significant signed ranked genes were used (m = number of significant genes) at a 0.05% significance threshold

- The significant signed ranked genes were used (m = number of significant genes) at a 0.01% significance threshold

Since approaches (b)–(e) will often result in nonreciprocal Zhang scores between two compounds, depending on the choice of query and reference compound, and since the number of genes that are included in the analysis may affect the magnitude of the Zhang connection score, a methodology was developed to estimate the significance of the obtained transcriptional connection scores.^{6,14}

The methodology used to estimate P -values based on the Zhang score¹⁴ was extended to confidence intervals. A total of 1×10^5 random gene signatures were created, with $m=5$, on a subset of the compound data ($N=5682$). The resulting distribution of the Zhang scores obtained from the random gene signatures, with 5 genes for each of the 5682 samples, could be approximated by a normal distribution. By varying m , the effect of the gene signature size on the population metrics of the randomized Zhang connection score distribution was assessed. The generalized mean and standard deviation from these different distributions were computed to determine a confidence interval of an obtained Zhang connection score, for a given false positive rate.

Target Activity Predictions: Machine Learning Algorithms

When L1000 profiled compounds have activity data for a specific target, a cross-validated predictive gene signature can be constructed for that target, using machine learning algorithms (Fig. 1). For this purpose, the available target activity data were categorized into active and nonactive compounds and split into a training (67% of the data) and a test data set (remaining 33% of the data). A Monte Carlo cross-validation sampling procedure was adopted to generate eight of these training and test data sets, using a modified version of the function from the nlcv package in Bioconductor as to accommodate (a) prediction scores (votes) for the diagonal linear discriminant analysis (DLDA), (b) parallelization for random forest (RF), (c) the use of training samples only, in the training step for RF, and (d) add-ons in current and creation of additional visualizations.

The 973 genes were ranked in each of the 8 training sets by means of an RF procedure. Three diverse algorithms: DLDA, RF, and support vector machine (SVM) were subsequently used to create a gene signature from a gene set that contained an increasing number of the top featured genes. The three algorithms and selected genes were then used to make predictions in the corresponding test sample data set, assigning the test compounds to the active and nonactive classes. The quality of the gene signature was assessed by comparing the predicted versus the observed compound classification, using the balanced accuracy and

Kappa statistic from the different models in the 8 different test sets. All quality metrics were computed with the confusionMatrix function in the caret package.²²

$$\text{Balanced accuracy} = \left\{ \frac{\left(\frac{TP}{TP+FN} \right) + \left(\frac{TN}{TN+FP} \right)}{2} \right\}$$

$$\text{Kappa} = \frac{p_{\text{obs}} - p_{\text{exp}}}{1 - p_{\text{exp}}}$$

with p_{obs} the observed agreement, p_{exp} the expected agreement by chance alone, TP the true positive rate, TN the true negative rate, FP the false positive rate, and FN the false negative rate.

The optimal number of features is determined for each classification algorithm tested as the lowest number of features achieving a performance within one standard deviation of the highest performance (average Kappa across runs). Considering this optimal number of features, the algorithm with a Kappa difference less than 0.01 than the maximum performance obtained and using less than 25% of the total number of features is selected.

The target activity status for newly L1000 profiled compounds with unknown activity can now be predicted and the prediction strength was computed by the margin of victories, class probabilities, and probabilities computed from a logistic distribution in the DLDA, RF, and support vector machine models, respectively. (In)activity data for *NR3C1* and *HSP90* were available for 114 and 110 L1000 profiled compounds, respectively. Initial and confirmation activity data were obtained by the *NR3C1* and *HSP90* assay as described below:

NR3C1

The translocation of the GR, encoded by the *NR3C1* gene, was measured using the brain epithelial neuroglioma cell line H4. GR was first driven out of the nucleus by incubating the cells for 24 h at 37°C in a medium containing charcoal stripped serum. The cells were seeded and stored overnight in an incubator at 37°C. Compound was subsequently added and the plates were preincubated for 1 h at 37°C. After compound addition and preincubation, hydrocortisone was added to the cells that induce the translocation of GR. After 1 h of incubation at 37°C, the cells were fixed with 5% formaldehyde and permeabilized with 0.3% triton. By measuring the fraction of GR in the nucleus versus that in the entire cell by immunofluorescence imaging, using the CV7000, the nuclear translocation of GR can be determined.

HSP90

HSP90 inhibition was measured using the CellSensor heat shock responsive element (HSE)-bla HeLa Cell line (Invitro-

gen) that contains the β -lactamase receptor gene, controlled by an HSE-responsive element and a promotor cloned in front of the reporter gene. In each well, 5000 cells were seeded in 20 μ L 1% Dulbecco's fetal calf serum containing the medium and incubated in a 5% CO₂ incubator at 37°C for 16 h. After incubation, the cells were challenged for 5 h using a heat shock inducing tool compound at 1 μ M. This heat shock response induces the activation of quiescent heat shock factor (HSF) monomers in the cytosol. Released HSF's trimerize and translocate to the nucleus of the cell, where they bind to their HSE-responsive elements and drive the transcription of β -lactamase. Finally, the CCF4 β -lactamase substrate was added to the cells for 4 h at room temperature. The expression of β -lactamase could subsequently be quantified using the ratio of the 460/535 nm signal output, measured by FRET technology, on Envision (Perkin Elmer), since the cleavage of the substrate by β -lactamase induces a shift in fluorescent signal

RESULTS

To get an overview of the overall transcriptional effects²¹ of all tested compounds, including positive controls, *Figure 2A* shows the distribution of the number of significantly differentially expressed genes for the 31761 samples. Most compounds affect only a limited number of the 978 measured genes (*Fig. 2A*). Indeed, the number of significantly affected genes ranges from 0 to 768, with an average of 29 and a median of 4 across the entire compound range (*Fig. 2A*). The number of compounds affecting each gene is represented in detail for the first 20 most specific compound classes (*i.e.*, ranked by increasing number of affected genes) (*Fig. 2B*). Hence, each bar represents the compounds that affect the same number of genes. Within a bar, genes are ranked from the bottom to the top in decreasing order of appearance (*Fig. 2B*). It is clear from *Figure 2B* that no single gene or set of specific genes dominate this distribution. The most frequently affected gene (*i.e.*, *INSIG1*) is significantly affected at maximum 1.4% of all samples of a category. When a small number of genes are affected significantly by a compound, they tend to be clearly separated from the remaining gene cloud, as expressed by the difference of the absolute median *t* statistic values from both the significant and nonsignificant gene distributions (*Fig. 2C*). This absolute median *t* statistic difference decreases when the number of significantly affected genes per compound increases (*Fig. 2C*). A total of 330 positive control measurements were analysed in this data set as well, yielding 88, 66, 89, and 87 replicate observations for LY-294002, estradiol, tanespimycin, and trichostatin, respectively. The median number of significantly affected genes across the control replicates,

Fig. 2. Number of significant genes. **(A)** Number of significantly differentially expressed genes for each of the 31,761 profiled samples. **(B)** Representation of genes that are commonly significantly affected across several samples, split by the number of genes affected; bars represent the first 20 compound classes, with genes ranked by the number of samples affecting them across all bars from *bottom* (i.e., redundant genes) to the *top* (i.e., less frequent genes) of the bar. **(C)** Number of significant genes versus specificity with the distance (i.e., minimal difference between the distribution) between the clouds of significant genes versus the clouds of nonsignificant genes, represented in the y-axis.

ranged from 3 in estradiol, 40 in tanespimycin, 49 in LY-294002, to 278 in trichostatin.

A total of 238,887,936 pairwise Zhang connection scores were computed, based on five settings, using (a) all genes, (b) the 5 most affected genes, (c) per treatment all significant affected genes at a 0.1% significance threshold, (d) per treatment all significant affected genes at a 0.05% significance

threshold, or (e) per treatment all significant affected genes at a 0.01% significance threshold, in the query gene signature. The Zhang scores for these different conditions seemed to be normally distributed, and not centered around zero (Fig. 3A). When the number of genes present in the query gene signature decreased, the Zhang connection score distribution was flattened, as more compounds revealed transcriptional (anti)

Fig. 3. Effect of the gene signature length on the transcriptional similarity. **(A)** Distribution of Zhang scores based on the 973 genes, the 5 most strongly affected genes or significant genes (thresholds: 0.01; 0.05 and 0.1) of the query only. **(B)** Progression of the number of pairs of samples with a Zhang score higher than a given threshold.

similarity with higher absolute Zhang scores. This was also illustrated when the Zhang connection scores for pairwise compound comparisons were plotted in function of different possible (di)similarity thresholds (Fig. 3B). Indeed, focusing on a subset of genes in the query led to a higher number of compound pairs that were picked up at higher Zhang connection score thresholds (Fig. 3B).

Likewise, the replicate similarity in each of the four controls decreased when all genes were used in the query gene signature (Fig. 4). The highest transcriptional reproducibility was obtained (across samples of the same positive control) when the query gene signature made use of the significant genes only, regardless of which of the four controls (Fig. 4). Nonetheless, there was still a big difference in reproducibility among the four controls, with estradiol (positive control with lowest signal) being the only control lacking good reproducibility among its replicates (Fig. 4). The Zhang connection score-based heatmaps also illustrated the loss of the symmetry of the connection score between two replicates, when a subset of genes was considered in the query sample (Fig. 4). Symmetry is, indeed, lost when both replicates affect a different number of differentially expressed genes.

Despite the caveat of similarity loss, the clearest transcriptional connection (Fig. 3A) and highest reproducibility (Fig. 4) were obtained when the gene signature query was

based on the significant genes only. However, the chance of obtaining a high connection score, using this gene signature query criterion, seemed to decrease as the promiscuity of compounds increased. The latter is illustrated by the funnel-shaped distribution of the differences between the Zhang score when only the five highest ranked genes or the significant genes at various significance thresholds are considered, versus all 973 genes (Fig. 5). All five distributions vary around zero in function of the number of significant genes, which is indicative of the promiscuity of compounds (Fig. 5).

Since the number of significant genes, present in the query gene signature, affected the magnitude of the Zhang connection score, there was a need to introduce confidence intervals for this similarity parameter. To this end, Zhang connection scores were computed, based on 1×10^5 random gene signatures. When varying the number of m , the mean of the different random distributions varied, as expected, around zero (Fig. 6A). However, the standard deviation decreased exponentially when the number of genes in the query gene signature increased (Fig. 6B). The function " $1/\sqrt{m}$ " was fitted to estimate the standard deviation for a specific number of genes, yielding an intercept of 0.009 and a slope of 0.602 (Fig. 6B). Hence, the randomized Zhang connection score distribution has a mean of zero and a standard deviation of $sd_{nGenes} = 0.009 + 0.602 \cdot (1/\sqrt{m})$. The confidence interval can now be rewritten as follows:

Fig. 4. Effect of the gene signature length on the transcriptional reproducibility. Heatmaps representing the Zhang connection similarity among replicates from the same control, plotted in columns (LY-294002 $n=88$; Estradiol $n=66$; Tanespimycin $n=89$; Trichostatin $n=87$), using all 973 genes, significant genes only at different significant thresholds (0.1; 0.05; 0.01) and the five most strongly affected genes plotted in rows. For ease of reading, the figure can be viewed online at www.liebertpub.com/adt

$$qnorm\left(1 - \left(\frac{\text{threshold}}{2}\right), \text{mean}=0, \text{sd}=1\right) * sd_{n\text{Genes}}$$

with the threshold set at $1/N$ when N is the number of considered compounds, controlling the expected number of false positive connections at 1. If zero falls inside the Zhang confidence interval, there would be not enough evidence to reject the H_0 of nonsignificance, irrespective of the query gene signature size, while allowing a user-defined false positive rate. The radius or half width of the Zhang confidence intervals is tabulated for different gene signature widths (rows) and compound sample sizes (columns), when a false positive

connection score of 1 is adopted (Table 1). The confidence intervals can easily be computed for all row x column conditions as the Zhang score $\pm$ the radius. The gray cells indicate experimental conditions in which a Zhang connection score will never attain significance, since the absolute Zhang connection score, which is limited to 1, must exceed the radius for this to happen.

NR3C1

The number of available L1000 profiled compounds, which were also screened in the NR3C1 assay ($n=114$), is presented, along with the activity threshold in Figure 7A. The data were unbalanced with 87 compounds grouped in the inactive and

Fig. 4. Continued.

27 compounds grouped in the active *NR3C1* compound class. The discriminative power (as expressed by the Kappa statistic) for diagonal linear discriminant analysis (DLDA), RF, and support vector machine (SVM) algorithms across the different test runs (using 10 different gene set sizes ranging from 2 to 35) is plotted for *NR3C1* in Figure 7B. RF and DLDA performed equally well, irrespective of the number of gene features that were used in the models. The discriminative power for SVM dropped rapidly, once more than seven gene features were included. When considering the best Kappa statistic for the least number of gene features used, SVM slightly outperformed the other two machine learning algorithms (Fig. 7B). The mean balanced accuracy of the final SVM model across test sets using two gene features (i.e., *DDIT4* and *SERPINE1*) was 0.81 ± 0.07 and the Kappa was 0.69 ± 0.15 across the test sets (i.e., “substantial” agreement according to Viera and Garrett²³). The compound class assignment for the entire data set, using SVM with *DDIT4* and *SERPINE1* as input features, is visualized in Figure 7C. All 114 compounds are plotted in function of their measured and predicted *NR3C1* activity. Only 7 compounds were wrongly assigned by the SVM model (Fig. 7C). Finally, the actual expression values for *DDIT4* and *SERPINE1* are plotted in Figure 7D, in which each compound is colored, based on its observed *NR3C1* activity. Compounds with higher upregulation of both genes tend to show higher *NR3C1* activity (Fig. 7D).

HSP90

The number of available L1000 profiled compounds that were also screened in the *HSP90* assay ($n = 110$) is presented along with the activity threshold in Figure 7E. The data were unbalanced with 75 compounds grouped in the inactive and 35 compounds grouped in the active *NR3C1* compound class. The discriminative power (as expressed by the Kappa statistic) for DLDA, RF, and SVM across the different runs (using 10 different gene set sizes ranging from 2 to 35) is plotted for *HSP90* in Figure 7F. Support Vector Machine and DLDA outperformed RF when a smaller number of gene features (less than 7) were picked. When considering the best Kappa statistic for the least number of gene features used, SVM performed better than DLDA and outperformed RF (Fig. 7F). The mean balanced accuracy of the final SVM model using two gene features (i.e., *SPR* and *TMEM97*) was 0.80 ± 0.06 with a Kappa value of 0.65 ± 0.12 across the test sets. The compound class assignment for the entire data set, using SVM with *SPR* and *TMEM97* as input features is visualized in Figure 7G. All 110 compounds are plotted in function of their measured and predicted *HSP90* activity. Only 15 compounds were wrongly assigned by the SVM model (Fig. 7G). Finally, the actual expression values for *SPR* and *TMEM97* are plotted in Figure 7H, in which each compound is colored, based on its observed *HSP90* activity.

Fig. 5. Effect of the gene signature length on the Zhang score difference. **(A)** Average difference of Zhang score when only the significant genes (threshold:0.01) in the query versus all 973 genes are considered. **(B)** Average difference of Zhang score when only the significant genes (threshold:0.05) in the query versus all 973 genes are considered. **(C)** Average difference of Zhang score when only the significant genes (threshold:0.1) in the query versus all 973 genes are considered. **(D)** Average difference of Zhang score when only the five most strongly affected genes in the query versus all 973 genes are considered.

Compounds with higher upregulation of these genes tend to have higher *HPS90* activity (Fig. 7H).

DISCUSSION

Ever since the publication of CMap,⁶ the LINCS library, and the commercialization of the L1000 platform, the use of high-throughput transcriptomics and query gene signatures has been successfully applied in the drug discovery and devel-

opment process by a variety of research groups (For a review²⁴). In the last decade, new therapeutic targets have been found,^{4,6,25,26} existing drugs have been repurposed or repositioned,^{12,27,28} and side effects have been flagged.¹¹ To this end, a variety of similarity metrics have been proposed and their gene pattern matching success has been investigated by Cha *et al.*¹³ The connectivity score proposed by Lamb⁶ and the connection approach proposed by Zhang and Gant¹⁴ seemed

Fig. 6. Mean and standard deviation fitting for random generated Zhang connection score distributions. **(A)** Mean Zhang connection score (y-axis) for random generated gene signatures with varying number of genes m in the query (x-axis) for the subset of 5682 compounds. **(B)** Standard deviation of the Zhang connection scores (y-axis) for random generated gene signatures with varying number of genes m in the query (x-axis) for the subset of 5682 compounds; red line represents the fitted function $1/\sqrt{m}$.

to be the most promising metrics. Unlike the connectivity score,⁶ the Zhang connection score¹⁴ has, however, the advantage of treating similarly upregulated and downregulated genes to the same effect (i.e., equal symmetric weight). Moreover, the Zhang connection score¹⁴ combines the ranks of all genes present in the query signature, instead of retaining the rank differences across all genes that are regulated in the same direction, allocating weights that are proportional to the induced differential expression. In addition, the signed ranked based connection score¹⁴ can be extended to an unordered gene signature, when information on the magnitude of the significance for the genes is not available.

The value and symmetry of the obtained connection scores largely depend on the number of genes that are present in the query gene signature. The connection symmetry between two compounds is kept when all genes are used for the query, although noise is added to the metric under these circumstances. Indeed, using all genes adversely affects the reproducibility of the connection score, as demonstrated for the replicates in each of the positive controls. Hence, despite the loss of connection symmetry, a query gene signature derived from a subset of genes will increase the transcriptional reproducibility and decrease existing noise. As a matter of fact, the connectivity score proposed by Lamb⁶ incorporated this idea already and made use of a fixed number of upregulating and downregulating genes when constructing gene signatures. However, connection scores that are

computed from a fixed number of genes, derived from a very specific or a promiscuous compound, are highly similar. The difference between both types of compounds can only be appreciated when a significance gene inclusion threshold is set. Given that most compounds in this data set affect only a few genes, it is even more important to correctly compute the connection score for specific, rather than promiscuous compounds. When setting such a significance gene entry threshold, promiscuous compounds are less likely to pick up high connections with other compounds. The proposed confidence interval and the flexibility to set a user-defined false positive rate can, however, amend this caveat. Hence, our simulations indicate that a Zhang connection score should always be interpreted in the context of its confidence interval, when making compound (di)similarity calls.

In addition to finding transcriptional (di)similarity among compounds, it was investigated if this L1000 data set could also be used in the context of machine learning. The latter was achieved by taking the activity status of less than 115 L1000 profiled compounds, tested in the *NR3C1* and *HSP90* assay, into account. Surprisingly, only four genes, *SERPINE1*, *DDIT4*, *TMEM97*, and *SPR*, were needed to reach a sufficient predictive power for both targets. The small number of predictive genes is most likely related to the fact that the 978 genes, measured on the L1000 platform, have been identified to capture most of the information contained in the entire transcriptome (www.lincscloud.org/l1000). As such, these

Table 1. Radius (i.e., Half Width CI) for Different Combinations of Gene Signature Lengths (i.e., Rows) and L1000 Profiled Compound Data Base Sizes (i.e., Columns) with a False Positive Connection Score Threshold (t) Set at 1

	t:1/2	t:1/5	t:1/10	t:1/50	t:1/100	t:1/500	t:1/1000	t:1/1500	t:1/2000	t:1/2500	t:1/5000	t:1/10000	t:1/30000	t:1/50000	t:1/100000	t:1/200000
nGenes:1	0.412	0.783	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
nGenes:2	0.293	0.557	0.715	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
nGenes:3	0.241	0.457	0.587	0.830	0.919	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
nGenes:4	0.209	0.398	0.511	0.722	0.799	0.959	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
nGenes:5	0.188	0.357	0.458	0.648	0.718	0.861	0.917	0.948	0.970	0.986	1.000	1.000	1.000	1.000	1.000	1.000
nGenes:6	0.172	0.327	0.420	0.594	0.657	0.788	0.840	0.868	0.888	0.903	0.949	0.993	1.000	1.000	1.000	1.000
nGenes:7	0.160	0.304	0.390	0.551	0.610	0.732	0.780	0.806	0.825	0.839	0.881	0.922	0.983	1.000	1.000	1.000
nGenes:8	0.150	0.285	0.366	0.517	0.572	0.687	0.731	0.756	0.774	0.787	0.827	0.865	0.922	0.948	0.982	1.000
nGenes:9	0.142	0.269	0.346	0.489	0.541	0.649	0.691	0.715	0.731	0.744	0.781	0.817	0.872	0.896	0.928	0.959
nGenes:10	0.135	0.256	0.329	0.465	0.515	0.617	0.657	0.680	0.695	0.707	0.743	0.777	0.829	0.852	0.882	0.912
nGenes:15	0.111	0.211	0.271	0.384	0.425	0.509	0.543	0.561	0.574	0.584	0.613	0.641	0.684	0.703	0.728	0.753
nGenes:20	0.097	0.185	0.237	0.335	0.371	0.445	0.474	0.490	0.501	0.510	0.536	0.560	0.598	0.614	0.636	0.658
nGenes:25	0.088	0.166	0.214	0.302	0.334	0.401	0.427	0.442	0.452	0.460	0.483	0.505	0.539	0.554	0.574	0.593
nGenes:30	0.081	0.153	0.196	0.278	0.307	0.369	0.393	0.406	0.415	0.423	0.444	0.464	0.495	0.509	0.527	0.545
nGenes:35	0.075	0.143	0.183	0.259	0.286	0.344	0.366	0.378	0.387	0.394	0.414	0.433	0.461	0.474	0.491	0.508
nGenes:40	0.071	0.134	0.172	0.243	0.270	0.323	0.344	0.356	0.364	0.370	0.389	0.407	0.434	0.446	0.462	0.478
nGenes:45	0.067	0.127	0.163	0.231	0.256	0.307	0.326	0.338	0.345	0.351	0.369	0.386	0.412	0.423	0.438	0.453
nGenes:50	0.064	0.121	0.156	0.220	0.244	0.292	0.311	0.322	0.329	0.335	0.352	0.368	0.393	0.403	0.418	0.432
nGenes:60	0.059	0.112	0.143	0.203	0.225	0.269	0.287	0.297	0.303	0.309	0.324	0.339	0.362	0.372	0.385	0.398
nGenes:70	0.055	0.104	0.134	0.189	0.210	0.252	0.268	0.277	0.283	0.288	0.303	0.317	0.338	0.347	0.360	0.372
nGenes:80	0.052	0.098	0.126	0.179	0.198	0.237	0.253	0.261	0.267	0.272	0.286	0.299	0.319	0.327	0.339	0.350
nGenes:90	0.049	0.093	0.120	0.170	0.188	0.225	0.240	0.248	0.254	0.258	0.271	0.284	0.303	0.311	0.322	0.333
nGenes:100	0.047	0.089	0.115	0.162	0.179	0.215	0.229	0.237	0.243	0.247	0.259	0.271	0.289	0.297	0.308	0.318
nGenes:150	0.040	0.075	0.096	0.136	0.151	0.181	0.193	0.200	0.204	0.208	0.218	0.228	0.243	0.250	0.259	0.268
nGenes:200	0.035	0.067	0.086	0.121	0.134	0.161	0.171	0.177	0.181	0.184	0.194	0.203	0.216	0.222	0.230	0.238
nGenes:300	0.030	0.057	0.073	0.103	0.114	0.137	0.146	0.151	0.154	0.157	0.165	0.172	0.184	0.189	0.195	0.202
nGenes:400	0.027	0.051	0.065	0.092	0.102	0.122	0.130	0.135	0.138	0.140	0.147	0.154	0.164	0.169	0.175	0.181
nGenes:500	0.025	0.047	0.060	0.085	0.094	0.113	0.120	0.124	0.127	0.129	0.135	0.142	0.151	0.155	0.161	0.166

The computed Zhang connection score is significant when its value exceeds the radius for the specific experimental conditions.

Fig. 7. Machine learning-based target activity predictions for NR3C1 and HSP90. **(A)** Distribution of the NR3C1 bioassay activity values for all compounds profiled on L1000, with the activity threshold indicated with a vertical red line ($n = 114$). **(B)** The rate of misclassification and standard deviation as expressed by the Kappa statistic averaged across the eight different test runs, varying the size of the input gene feature set for the different algorithms that were used: DLDA, RF, and SVM. **(C)** Representation of the compound classification for NR3C1 activity by SVM machine, using DDIT4 and SERPINE 1 as signature (y-axis), compared to experimental measurements (x-axis). **(D)** Actual gene expression values (i.e., DDIT4 and SERPINE 1) for all profiled compounds colored based on their measured NR3C1 activity. **(E)** Distribution of the HSP90 bioassay activity values for all compounds profiled on L1000, with the activity threshold indicated with a vertical red line ($n = 110$). **(F)** The rate of misclassification and standard deviation as expressed by the Kappa statistic averaged across the eight different test runs, varying the size of the input gene feature set for the different algorithms that were used, DLDA, RF, and SVM. **(G)** Representation of the compound classification for HSP90 activity by SVM machine, using SPR and TMEM97 as signature (y-axis), compared to experimental measurements (x-axis). **(H)** Actual gene expression values (i.e., SPR and TMEM97) for all profiled compounds colored based on their measured HSP90 activity. DLDA, diagonal linear discriminant analysis; RF, random forest; SPR, sepiapterin reductase (7,8-dihydrobiopterin:NADP+ oxidoreductase); SVM, support vector machine.

so-called landmark genes are believed to represent different clusters of similarly behaving genes. The drawback of measuring landmark genes in the context of target activity prediction modeling is that, it can be difficult to assess the relevance of the chosen predictive genes in the context of the target activity they are aiming to predict (*i.e.*, *DDIT4* & *SERPINE1*; *NR3C1*; *TMEM97* & *SPR*; *HSP90*). Nonetheless, *DDIT4* has been reported to be regulated by glucocorticoids.²⁹ Kumari *et al.*³⁰ demonstrated that changes in the expression of *DDIT4* can, for instance, be used as an excellent surrogate biomarker for the glucocorticoid receptor blockade in skeletal muscle. Sautron *et al.*³¹ found a significant *DDIT4* upregulation in pigs after they were injected with 222 µg of a synthetic adrenocorticotrophic hormone, which triggered the production of cortisol. In another study, involving 25 overweight and obese subjects, the expression of *SERPINE1* was negatively correlated with serum cortisol and 24-h urinary free cortisol levels.³² *TMEM97* is a known regulator of cellular cholesterol homeostasis.³³ Since cholesterol is the precursor of pregnenolone,³⁴ from which steroid hormones, including the glucocorticoids are derived,³⁵ it can be implicated in the formation of cortisol. A direct link between *TMEM97* and *HSP90* is, however, not available. Likewise, *SPR*, which plays a role in the formation of dopamine and serotonin,³⁶ cannot be linked directly to *HSP90*. Nevertheless, despite the caveat that chosen landmark genes may be difficult to explain in the context of the target they aim to predict, these findings illustrate that they can be used to build predictive models. Indeed, despite the unbalanced design, different algorithms yielded robust predictive models for both *NR3C1* and *HSP90*, with SVM being picked as the top model in both cases, with balanced accuracy values of ≥80% for both targets. Moreover, based on both models, 275 newly L1000 profiled compounds were predicted to be active and tested for their *NR3C1* and *HSP90* (in)activity in both assays. From these 275 compounds, 22 revealed the desired *NR3C1* and *HSP90* activity. All 22 compounds represented either a pyrimidine or a pyrazolo-pyrimidine scaffold in linear or macrocycle constellations. The latter nicely illustrates the chemical diversity, which can be picked up when investigating chemical structures, based on a common transcriptional effect they induce. Finding different scaffolds in this exercise underlines the scaffold hopping potential of L1000 and its usefulness when optimizing chemical series. As a matter of fact, this L1000 exercise, performed on a subset of Janssen entire screening deck, outperformed the initial, classical HTS GR campaign, which was performed at Janssen in 2016, by more than 300 times.

Given that L1000 can find compound connections and/or predict their activity on a variety of targets, it could be used to construct more biodiverse screening libraries. The obtained

L1000 information could enable the construction of compound screening libraries composed of chemical matter that spans a maximum of diverse biological effects rather than diversity that aims to cover a large portion of the chemical space. In this context, Petrone *et al.*³⁷ have demonstrated that strategies, which optimize for target diversity, outperform chemically diverse screening decks both with respect to hit rate as well as the chemical diversity in identified compound scaffolds among the hits. Along the same lines, L1000 could be used to identify compounds that have never shown activity even though they have been tested in several different biological assays. This so-called dark chemical matter³⁸ has been shown to be much more selective when deorphanized as they interact with fewer targets, while their biological activity was shown to be similar as other compounds.³⁹ Given the advancements made in this article in combination with our strategic investment into generating L1000 profiles and target deconvolution strategies based on high content imaging, we are establishing the means to rapidly propose and test hypotheses for the likely mechanism underlying any biological activity that may be observed for Janssen's dark chemical matter. These capabilities provide the tools to unravel mechanistic insights into both the desired activities as well as any undesired activities⁴⁰ of the compounds we process through the Janssen drug discovery and development pipeline.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

1. Settleman J, Cohen RL: Communication in drug development: Translating scientific discovery. *Cell* 2016;164:1101–1104.
2. Smietana K, Siatkowski M, Möller M: Trends in clinical success rates. *Nat Rev Drug Disc* 2016;15:379–380.
3. Verbist B, Klambauer G, Vervoort L, *et al.*: Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project. *Drug Discov Today* 2015;20:505–513.
4. Senkowski W, Jarvius M, Rubin J, *et al.*: Large-scale gene expression profiling platform for identification of context-dependent drug responses in multicellular tumor spheroids. *Cell Chem Biol* 2016;23:1428–1438.
5. De Wolf H, De Bondt A, Turner H, Göhlmann WH: Transcriptional characterization of compounds: Lessons learned from the public LINCS data. *Assay Drug Dev Technol* 2016;14:252–260.
6. Lamb J: The connectivity map: Using gene-expression signatures to connect small molecules genes, and disease. *Science* 2006;313:1929–1935.
7. Lamb J: The connectivity map: a new tool for biomedical research. *Nat Rev Canc* 2007;7:54–60.
8. Peck D, Crawford ED, Ross KN, *et al.*: A method for high-throughput gene expression signature analysis. *J Gen Biol* 2006;7:R61.
9. Subramanian A, Narayan R, Corsello SM: A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;171:1437–1452.
10. Lv C, Wang YC, Liu RH, Zhang WD: Application of connectivity map database to research on Chinese Materia Medica. *Chin Herbal Med* 2016;8:117–120.
11. Wang K, Weng Z, Sun L, Sun J, Zhou SF, He L: Systematic drug safety evaluation based on public genomic expression (connectivity map) data: Myocardial and

- infectious adverse reactions as application cases. *Bioch Biophys Res Commun* 2015;457:249–255.
12. Iorio F, Bosotti R, Scacheri E, et al.: Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA* 2010;107:14621–14626.
 13. Cha K, Kim MS, Oh K, Shin H, Yi GS: Drug similarity search based on combined signatures in gene expression profiles. *Healthcare Inform Res* 2014;20:52.
 14. Zhang SD, Gant TW: A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinform* 2008;9:258.
 15. Jamali AA, Ferdousi R, Razzaghi S, et al.: DrugMiner: Comparative analysis of machine learning algorithms for prediction of potential druggable proteins. *Drug Discov Today* 2016;21:718–724.
 16. Xie L, He S, Wen Y, Bo X, Zhang Z: Discovery of novel therapeutic properties of drugs from transcriptional responses based on multi-label classification. *Sci Rep* 2017;7:7136.
 17. Bledsoe RK, Montana VG, Stanley TB, et al.: Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition. *Cell* 2002;110:93–105.
 18. Pariante CM: Risk factors for development of depression and psychosis. *Ann N Y Acad Sci* 2009;1179:144–152.
 19. Hubler TR, Scammell JG: Intronic hormone response elements mediate regulation of FKBP5 by progestins and glucocorticoids. *Cell Stress Chaperones* 2004;9:243.
 20. Holsboer F: The corticosteroid receptor hypothesis of depression. *Neuropsychopharmacology* 2000;23:473–501.
 21. Ritchie ME, Phipson B, Wu D, et al.: Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl Acids Res* 2015;43:47.
 22. Kuhn M: Building predictive models in R using the caret package. *J Stat Software* 2008;28:123099.
 23. Viera AJ, Garrett JM: Understanding interobserver agreement: The Kappa statistic. *Fam Med* 2005;37:360–363.
 24. Musa A, Ghorai LS, Zhang SD, et al.: A review of connectivity map and computational approaches in pharmacogenomics. *Brief Bioinform* 2017;bbw112.
 25. Liu C, Su J, Yang F, Wei K, Ma J, Zhou X: Compound Signature Detection on LINCS L1000 Big Data. *Mol Bio Syst* 2015;11:714–722.
 26. Won SJ, Wu HC, Lin KT, et al.: Discovery of molecular mechanisms of lignan justicidin a using L1000 gene expression profiles and the library of integrated network-based cellular signatures database. *J Funct Foods* 2015;16:81–93.
 27. Kunkel SD, Suneja M, Ebert S, et al.: MRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell Metab* 2011;13:627–638.
 28. Ramsey JM, Kettyle LMJ, Sharpe DJ, et al.: Entinostat prevents leukemia maintenance in a collaborating oncogene-dependent model of cytogenetically normal acute myeloid leukemia. *Stem Cells* 2013;31:1434–1445.
 29. Britto FA, Begue G, Rossano B, et al.: REDD1 Deletion Prevents Dexamethasone-Induced Skeletal Muscle Atrophy. *Am J Physiol Endocrinol Metab* 2014;307:983–993.
 30. Kumari R, Willing LB, Jefferson LS, Simpson IA, Kimball SR: REDD1 (Regulated in Development and DNA Damage Response 1) expression in skeletal muscle as a surrogate biomarker of the efficiency of glucocorticoid receptor blockade. *Biochem Biophys Res Commun* 2011;412:644–647.
 31. Sautron V, Terenina E, Gress L, et al.: Time course of the response to ACTH in Pig: Biological and transcriptomic study. *BMC Genomics* 2015;16.
 32. Pavlatou MG, Vickers KC, Varma S, et al.: Circulating cortisol-associated signature of glucocorticoid-related gene expression in subcutaneous fat of obese subjects. *Obesity* 2013;21:960–967.
 33. Bartz F, Kern L, Erz D, et al.: Identification of cholesterol-regulating genes by targeted RNAi screening. *Cell Metab* 2009;10:63–75.
 34. Tuckey RC, Atkinson HC: Pregnenolone synthesis from cholesterol and hydroxycholesterols by mitochondria from ovaries following the stimulation of immature rats with pregnant mare's serum gonadotropin and human choriongonadotropin. *Eur J Biochem* 1989;186:255–259.
 35. Macfarlane DP, Forbes S, Walker BR: Glucocorticoids and fatty acid metabolism in humans: Fuelling fat redistribution in the metabolic syndrome. *J Endocrinol* 2008;197:189–204.
 36. Abeling N, Duran M, Bakker H, et al.: Sepiapterin reductase deficiency an autosomal recessive DOPA-responsive dystonia. *Mol Genet Metabol* 2006;89:116–120.
 37. Petrone PM, Wassermann AM, Lounkine E, et al.: Biodiversity of small molecules a new perspective in screening set selection. *Drug Discov Today* 2013;18:674–680.
 38. Wassermann AM, Lounkine E, Hoepfner D, et al.: Dark chemical matter as a promising starting point for drug lead discovery. *Nat Chem Biol* 2015;11:958–966.
 39. Wassermann AM, Tudor M, Glick M: Deorphanization strategies for dark chemical matter. *Drug Discov Today Technol* 2017;23:69–74.
 40. Verbist BMP, Verheyen GR, Vervoort L, et al.: Integrating high-dimensional transcriptomics and image analysis tools into early safety screening: Proof of concept for a new early drug development strategy. *Chem Res Toxicol* 2015;28:1914–1925.

Address correspondence to:

Hans De Wolf, PhD

Janssen Research & Development

A Division of Janssen Pharmaceutica NV, Computational

Sciences, Discovery Sciences

Beerse 2340

Belgium

E-mail: hdwolf@its.jnj.com

Abbreviations Used

ATP6 V1D	= ATPase, H ⁺ transporting, lysosomal 34 kDa, V1 subunit D
CCDC86	= coiled-coil domain containing 86
CDC45 L	= CDC45 cell division cycle 45-like
CMap	= connectivity map
DDIT4	= DNA damage-inducible transcript 4
DLDA	= diagonal linear discriminant analysis
DMSO	= dimethyl sulfoxide
FKBP51	= FK506 binding protein 5
FN	= false negative rate
FP	= false positive rate
FRET	= fluorescence resonance energy transfer
GR	= glucocorticoid receptor
HSE	= heat shock responsive element
HSF	= heat shock factor
HSP	= heat shock protein
HTS	= high-throughput screens
INSIG1	= insulin-induced gene 1
LIMMA	= linear models for microarray and RNA-Seq Data
LINCS	= library of integrated network-based cellular signatures
MCF7	= Michigan cancer foundation-7
MOA	= mode of action
NIH	= national institutes of health
NLCV	= nested loop cross-validation
NR3C1	= nuclear receptor subfamily 3 group C member 1
NSDHL	= NAD(P) dependent steroid dehydrogenase-like
RF	= random forest
SERPINE1	= serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
SLC25A14	= solute carrier family 25 member 14
SPR	= sepiapterin reductase (7,8-dihydrobiopterin:NADP+ oxidoreductase)
SVM	= support vector machine
TMEM97	= transmembrane protein 97
TN	= true negative rate
TP	= true positive rate