OMICS A Journal of Integrative Biology
Volume 26, Number 6, 2022
Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2022.0026

A New Approach to Drug Repurposing with Two-Stage Prediction, Machine Learning, and Unsupervised Clustering of Gene Expression

Yi Cong, Misaki Shintani, Fuga Imanari, Naoki Osada, and Toshinori Endo

Abstract

Drug repurposing has broad importance in planetary health for therapeutics innovation in infectious diseases as well as common or rare chronic human diseases. Drug repurposing has also proved important to develop interventions against the COVID-19 pandemic. We propose a new approach for drug repurposing involving two-stage prediction and machine learning. First, diseases are clustered by gene expression on the premise that similar patterns of altered gene expression imply critical pathways shared in different disease conditions. Next, drug efficacy is assessed by the reversibility of abnormal gene expression, and results are clustered to identify repurposing targets. To cluster similar diseases, gene expression data from 262 cases of 31 diseases and 268 controls were analyzed by Uniform Manifold Approximation and Projection for Dimension Reduction followed by *k*-means to optimize the number of clusters. For evaluation, we examined disease-specific gene expression data for inclusion, body myositis, polymyositis, and dermatomyositis (DM), and used LINCS L1000 characteristic direction signatures search engine (L1000CDS^2) to obtain lists of small-molecule compounds that reversed the expression patterns of these specifically altered genes as candidates for drug repurposing. Finally, the functions of affected genes were analyzed by Gene Set Enrichment Analysis to examine consistency with expected drug efficacy. Consequently, we found disease-specific gene expression, and importantly, identified 20 drugs such as BMS-387032, phorbol-12-myristate-13-acetate, mitoxantrone, alvocidib, and vorinostat as candidates for repurposing. These were previously noted to be effective against two of the three diseases, and have a high probability of being effective against the other. That is, inclusion body myositis and DM. The two-stage prediction approach to drug repurposing presented here offers innovation to inform future drug discovery and clinical trials in a variety of human diseases.

Keywords: drug repurposing, big data, bioinformatics, machine learning, drug research and OMICS, drug development

Introduction

Drug repurposing is a method of developing new targets for existing drugs, that is, discovering new efficacy for a previously approved drug, for which safety and pharmacokinetics have been demonstrated in humans. While *de novo* drug development typically takes 6–9 years and costs 2–3 billion dollars, drug repurposing will lead directly to preclinical testing and clinical trials (Rapicavoli et al., 2022), thereby significantly reducing the time, cost, and side effects, leading

Laboratory of Information Biology, Information Science and Technology, Hokkaido University, Sapporo, Japan.

potentially to a higher success rate for introduction in clinical practice because the compounds have already been tested for safety and pharmacokinetics in humans (Jourdan et al., 2020).

Drug repurposing has broad importance in planetary health for therapeutics innovation in infectious diseases as well as common or rare chronic human diseases. Drug repurposing has also proved important to develop interventions against the COVID-19 pandemic. In recent years, drug repurposing has reached 30% of new drugs and vaccines approved by the U.S. Food and Drug Administration (FDA) (Kwon et al., 2019). However, the main questions often arise from how to customize or optimize the repurposing methods into efficient drug repurposing pipelines (Jin and Wong, 2014). Developing promising and affordable approaches for the effective treatment of complex diseases is difficult without prior knowledge of the complete drug-target network. This is now the greatest challenge to advancing drug repurposing technology (Zeng et al., 2020).

The present study presents a unique two-stage approach to drug repurposing that (1) harnessed machine learning (ML) to identify significantly altered gene expression profiles based on comparative data under diseased and normal conditions, and (2) analyzed the data on gene expression changes due to drug treatment, and (3) estimated the expected normalization of expression changes caused by a disease. To fully validate this approach, we analyzed Gene Ontology (GO) for a group of disease-variant genes.

Conceptual background on the study

Screening of drugs for repurposing includes a variety of methods, such as target-based, knowledge-based, signature-based, pathway- or network-based, and mechanism-targeted methods (Rapicavoli et al., 2022). The oldest case of drug repurposing is sildenafil for erectile dysfunction, which is usually considered accidental (Roundtable on Translating Genomic-Based Research for Health et al., 2014). However, serendipity-based discoveries cannot be expanded, and various methods have been used to enable systematic discovery of the off-label effects of drugs.

Phenotypic screening is a relatively more proactive and controlled method (Ciallella and Reaume, 2017), but it is not usually systematic and comprehensive enough, although it can occasionally successfully identify lead compounds. Since a disease usually emerges as a complex interaction between multiple genetic variants (Hirschhorn and Daly, 2005), computer-based approaches must be used to achieve systematic or comprehensive repurposing.

With the advent of high-throughput technologies for exploring biological systems, an impressive amount of data awaits computational analysis and mining tools to be explored and harnessed (Rapicavoli et al., 2022). Systems biology approaches to drug repurposing utilize pathophysiological mapping of diseases to identify targets that modify them, and potential compounds that can hit those targets (Turanli et al., 2018). In addition to computer simulations and target docking using algorithmic solutions other than medicinal chemistry, data mining based on gene expression provides clues to active pharmaceutical ingredients that may be formulated into clinically viable drugs (Chen et al., 2017).

The signature-based drug repurposing approach relies on the use of genetic signatures derived from disease-wide data such as microarrays, RNA-seq, which can identify unknown off-targets or unknown disease mechanisms. Since the required information may be difficult to obtain from the existing literature, obtaining genetic signatures for these diseases from publicly available genomic data becomes the best option (Rapicavoli et al., 2022).

In this context, artificial intelligence (AI) tools such as ML are powerful because they can identify patterns at scale. The history of AI can be briefly described in terms of three paradigms: good old-fashioned artificial intelligence (GOFAI) (1950s–1960s), expert systems (late 1970s–1980s), and ML (2010–present). GOFAI focused on the creation of general logic systems and led to the development of fundamental techniques such as heuristic searches. Expert systems narrowed the focus from general intelligence to human experts in specific fields, such as chemistry and medicine, and attempted to replicate their knowledge and decision-making processes. This led to the first major medical AI systems such as MYCIN (Garvey, 2018). While these yielded some practical results, none of these AI paradigms became ''thinking machines.''

However, the current ML paradigm has overcome some of the hurdles associated with the real world, thanks to the ever-increasing amount of human-generated data, the massive increase in computing power, and the renaissance of neural networks and other ML algorithms. These ''learning'' algorithms can be ''trained'' to infer patterns from human-generated data, and therefore do not require explicit representation of knowledge by the programmer (Garvey, 2018).

Although AI is still in its infancy in drug development, AI and ML algorithms have unprecedented potential to accelerate the discovery of effective new drugs. DSP-1181 is reportedly the first off-targets drug created using AI to enter clinical trials. Exscientia, which developed it, noted that it took <12 months from initial screening to the end of preclinical testing, compared with 4 years using traditional methods (Farghali et al., 2021). To date, many computational methods for drug repurposing using ML techniques are continuously being proposed and improved as new problems arise. Over the past several decades, computer tools, such as quantitative structure-activity relationship modeling, were developed to identify potential bioactive molecules quickly and inexpensively from great numbers of candidate compounds.

As ML approaches evolve into deep learning approaches, they become more powerful and efficient way to deal with the massive amounts of data generated from modern drug discovery approaches (Farghali et al., 2021). For example, to address the challenge of how to derive drug repurposing from drug-disease interactions, a methodological approach that focuses primarily on drug properties has been established (Napolitano et al., 2013). These ML approaches involving data integration are established by integrating information from different layers based on similarities such as chemical structures, molecular targets, and induced gene expression characteristics, and they focus on predicting the therapeutic class of United States Food and Drug Administration (FDA)-approved compounds without considering data about the disease.

The method reclassifies specific drugs and purposefully interprets the parts that do not match the original classification as genuine reclassification opportunities, while showing promising and highly accurate results by integrating different information layers and maximizing their efficacy through the dimensionality reduction-based computational procedures classical multidimensional scaling and principal component analysis (PCA). However, this approach only starts from the direction of drugs, and while it does address problems such as the complexity, variability, and sparsity of data currently available for diseases, the inability to explore specific diseases with strong targeting also largely limits its use.

Another representative modified method, drug repurposing in Alzheimer's disease (DRIAD), combines the use of omics datasets on drug-induced perturbation of neuronal cells with the molecular changes that occur in the brains of individuals suffering from different stages of Alzheimer's disease (AD) to generate a drug-associated gene list (Rodriguez et al., 2021). DRIAD was used to effectively decouple gene set enrichment and predictor performance by filtering the transcriptomic space for genes associated with drugs before model training and predictor evaluation. Simultaneous prefiltering to a limited set of features also addresses issues with overfitting and enables a direct, unbiased quantification of the association between the effects of a drug and AD progression.

However, DRIAD is performed to determine whether drug-induced changes correlate with molecular markers of disease severity by measuring what happens to nerve cells in the human brain when they are treated with drugs, which has some limitations for application in other diseases. Also, the method is based on a specified list of drugs for further analysis of the genes they affect, which means that only the list of drugs specific to the user can be screened, resulting in a local optimum in the list, but still relying to some extent on the user's drug screening strategy.

The *k*-means method, a commonly used clustering method, has good scalability with sample size increase. However, since it relies on computing the distance between the clustering centers given randomly and each sample, it requires a large feature space in high dimension that can result in expensive computation, large memory requirement, and poor clustering performance (Hozumi et al., 2021). Since the gene expression data are inherently highly dimensional, dimensionality reduction should be effective to avoid this problem, and finding a group of diseases sharing a hidden core of abnormality. PCA is often used for such purposes due to its intuitiveness and mathematical simplicity, but as a linear algorithm, PCA performs poorly on the features with nonlinear relationship.

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), an ML algorithm, is a topological data analysis technique based on many-body theory, which provides significant improvements in data localization and preservation of local structure compared with PCA. Also, in capturing similar word vector groups, UMAP has advantages over T-distributed stochastic neighbor embedding for large datasets, because it captures global and topological structures, and the error between two topological spaces will be minimized by optimizing the spectral layout of data in the low-dimensional space (Hozumi et al., 2021; McInnes et al., 2018). However, the method does not provide the boundary of clusters that can often be unapparent. In practice, logical clustering is mandatory for consistency and validity.

In the present study, we propose a five-step method for clustering. First, UMAP is applied to downscale gene expression data from different diseases. Second, data are classified using the *k*-means method with Silhouette analysis to determine *k*, the appropriate number of groupings. Third, those genes with significantly deviated expression for each disease were identified. Forth, L1000CDS^2 knowledge base is searched to identify chemical compounds revert impaired gene expression. Finally, the obtained lists are compared among the disease cluster to highlight possible drug repurposing. The result was examined with gene set enrichment analysis (GSEA) to address possible mechanisms shared across the group of diseases.

We use the data from microarrays to identify genes that may be up-/downregulated in disease microarrays, then searches the literature on drugs known to have opposite effects. By combining the gene expression responses of cell lines caused by diseases with data on drug-induced changes in gene expression, which not only solves the problem of scarce genomic data for diseases and quantifies the association between drug effects and diseases, but also generates a list of candidate drugs for specific diseases and performs analytical calculations by clustering. Therefore, compared with other existing methods, this method is considered to have wider applicability in practical applications and largely alleviates the bias caused by subjective screening by users.

Materials and Methods

Transformation of gene expression data

From the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/), gene expression datasets were obtained from 262 patients, and 268 healthy human samples associated with 31 diseases that were analyzed on the Affymetrix Human Genome U133A array platform and submitted by Affymetrix. The associated information is listed in Table 1. The logarithm of the ratio of expression intensities from patients' samples to the mean of the corresponding healthy controls was calculated for each disease in turn to obtain log fold-change (logFC) values for disease-specific gene sets (Robinson and Oshlack, 2010). The obtained gene expression data were treated as a matrix of 262 × 22,283 dimensions in subsequent analyses (Supplementary Data S1).

Two different procedures were applied to normalize the experimental data (Park et al., 2003). For data with binary logarithm values, such as robust multiarray average (RMA) and GC-RMA, mean values of healthy samples were simply subtracted from those of diseased samples to yield the binary logarithm of the ratio. For antilogarithms, such as normalization by Affymetrix MicroArray Suite 5.0 (MAS5) algorithm, binary logarithms of values were calculated before subtraction.

Clustering of disease data and selection of target diseases using UMAP and k*-means methods*

UMAP was used to reduce the hyper dimensionality to two dimensions. UMAP analyzes topological data based on manifold theory, which is characterized by its ability to clearly separate clusters through dimensionality reduction (McInnes et al., 2018). In this study, the final parameters used were n_neighbors = 8, min_dist = 1, n_epochs = 500, metric = correlation, and set_op_mix_radio = 0.3, to allow a clearer classification of the different disorders to adjust the parameter values.

The resulting two-dimensional data were clustered using the *k*-means method that based on Euclidean distance to visualize target diseases among clearly classified groups. Since the basic idea of the *k*-means method is to select the centers of

Table 1. Datasets for Cases

GEO ID	*Disease*	*Tissue*
GSE475	Chronic obstructive pulmonary disease	Diaphragm muscle
GSE593	Uterine fibroid	Myometrial
GSE1297	Alzheimer's disease	Hippocampal
GSE128470	Dermatomyositis	Muscle
GSE1751	Huntington's disease	Blood
GSE1789	Down syndrome	Heart
GSE2712	Clear cell sarcoma of the kidney	Kidney
GSE3365	Crohn's disease	PBMC
GSE3365	Ulcerative colitis	PBMC
GSE5090	Polycystic ovary syndrome	Omental adipose tissue
GSE5667	Atopic dermatitis	Skin
GSE5808	Acute measles	Peripheral blood
GSE7429	Osteoporosis	Circulating B cell in blood
GSE9750	Cervical cancer	Cervical epithelium
GSE9877	Sickle cell disease	BOEC
GSE13785	Exercise-induced bronchoconstriction	Airways cell
GSE15568	Cystic fibrosis	Rectal mucosal epithelia
GSE25724	Type 2 diabetes	Islet
GSE47018	Polycythemia vera	$CD34^+$ cell
GSE55235	Rheumatoid arthritis	Synovial
GSE75415	Pediatric adrenocortical tumor	Adrenal gland
GSE110223	Colorectal cancer	Colon
GSE115810	Endometrial cancer	Endometrium
GSE124646	Breast cancer	Breast
GSE128470	Polymyositis	Muscle
GSE128470	Inclusion body myositis	Muscle
GSE6613	Parkinson's disease	Blood
GSE35487	IgA nephropathy	Kidney tubular epithelial cell
GSE41649	Allergic asthma	Bronchial
GSE43290	Meningioma	Meningeal
GSE55235	Osteoarthritis	Synovial

BOEC, blood outgrowth endonuclear cells; GEO, Gene Expression Omnibus; IgA, immunoglobulin A; PBMC, peripheral blood mononuclear cells.

k clusters based on a given *k*-value, and then assign the sample points to be classified to the clusters according to the nearest neighbor principle for calculation, selecting a *k*-value will have a decisive impact on data clustering. We chose the Silhouette method for the determination of *k*-value. The Silhouette value measures how similar a point is to its own cluster compared to others, so in which a high Silhouette value is desirable and indicates that the point is placed in the correct cluster (Khyati, 2019).

Disease groups in the same cluster obtained by the *k*-means method are expected to have similarities in cellular signaling and gene expression and are candidates for off-target drug effects.

Classification of disease-specific gene expression

GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/) in NCBI GEO is a web-based interactive tool that can be used to compare multiple groups in the GEO series. $|\text{logFC}| > 1$ and *p*-value <0.05 are usually considered statistically significant (Zhang et al., 2020), so genes exhibiting significantly altered expression in a specific disease were identified on it based on the latter specifications.

Searching for small-molecule compounds

Using L1000CDS^2 (https://maayanlab.cloud/L1000CDS2/#/index), we identified compounds that reversed the changes in gene expression patterns for each disease while searching for drugs (Duan et al., 2016). Using the scatter plots of L1000 perturbation gene signatures and the results of chemical perturbations, genes directly affected by each identified drug were matched and organized.

GSEA and function prediction for gene sets

In preparation for GSEA, we collated raw genetic data for target diseases and inserted a ''Description'' (value = na) column for each disease to obtain three collated data tables. GSEA was performed using the ''Human_AFFY_HG_U133_MSigDB.v7.4.chip'' chip platform, and GO information was extracted for normalized *p*-value below 0.05 and false discovery rate (FDR) below 0.25 (Thomas et al., 2011). Finally, disease-specific expressed genes were classified using GO terms and their functions were compared.

Results

Dimensionality reduction with UMAP and clustering with k*-means*

For the datasets for cases and controls obtained from the NCBI GEO, by transforming gene expression data to analyze similarity between different diseases using clustering, and by comparing diseases classified in the same group to obtain a list of alternative drugs, we further processed the combined data and used UMAP and *k*-means methods for analysis.

UMAP was applied to the transformed expression data to reduce the hyper dimensionality of gene expression levels into two-dimension (Fig. 1a). Then, they were grouped by applying the *k*-means method to obtain similar changes in expression of gene sets, expecting they reflect similar changes in cytological conditions. The Silhouette Score reaches its global maximum at the optimal value (*k*-value = 19, Fig. 1b) as represented by the peak in the figure. The visualized groupings resulting from UMAP are shown in Figure 1c.

Target diseases were selected from the results of clustering by the *k*-means method, through which 31 diseases were clustered into several clearly classified groups. Figure 1c shows that inclusion body myositis (IBM), polymyositis (PM), and dermatomyositis (DM), which were classified into the same yellow cluster located on the upper right according to the results of the *k*-means method, shared similarity with each other; hence, they were chosen for the following analyses:

Target diseases and samples:

- IBM: GSM3676259-GSM3676284 (26 sets)
- PM: GSM3676317-GSM3676323 (7 sets)
- DM: GSM3676247-GSM3676258 (13 sets)
- Healthy: GSM3676285-GSM3676296 (12 sets)

Identification of disease-specific significant changes in gene expression by GEO2R

GEO2R was used to identify genes whose expression levels were specifically increased or decreased in the three

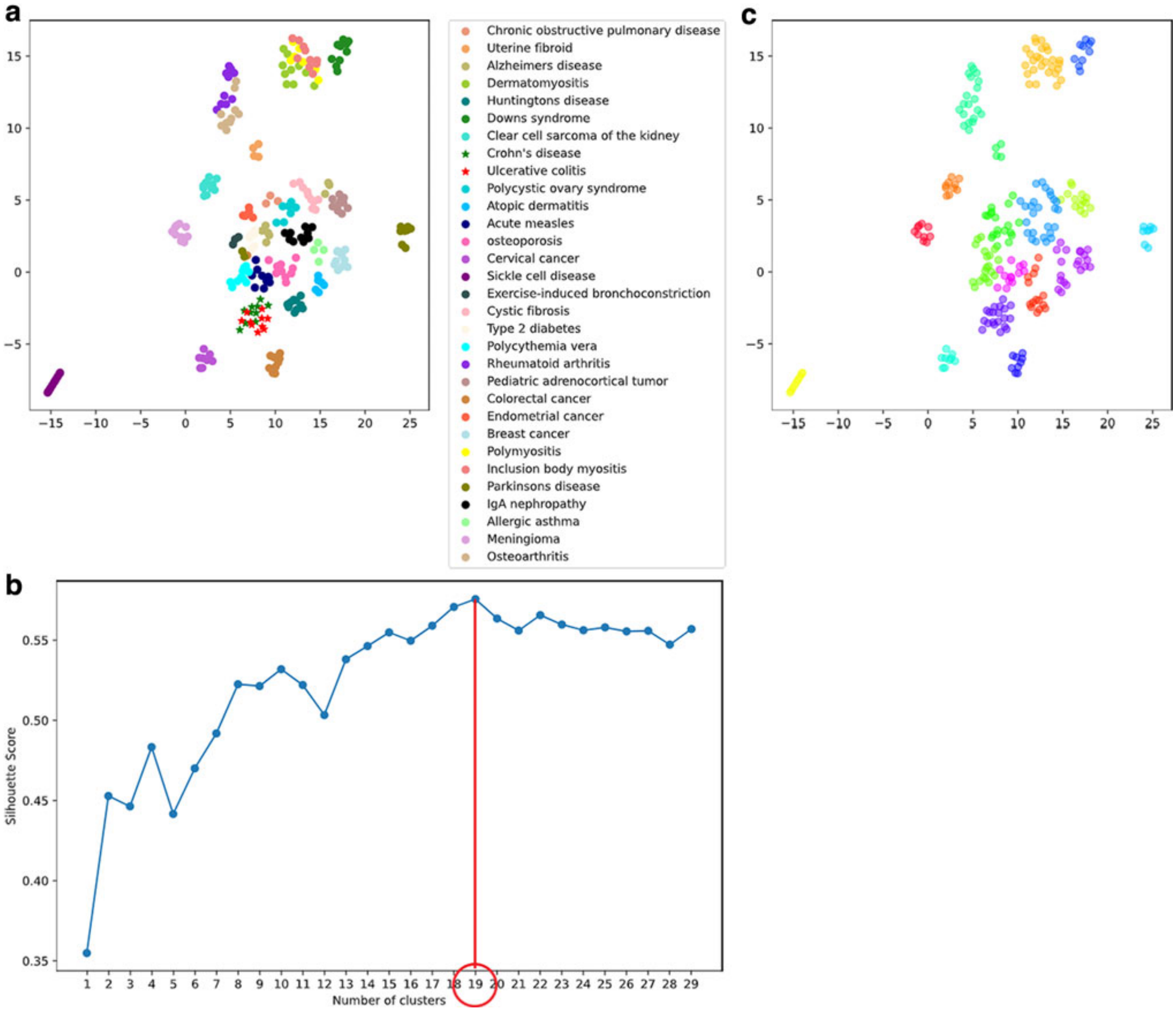

FIG. 1. Clustering of disease data. **(a)** UMAP analysis of disease-specific gene expression. *Different colors* represent different diseases, and positions reflect their degree of association. **(b)** Determination of *k*-value by Silhouette analysis. The peak of the results of Silhouette analysis provides the optimal *k*-value (19 in this case). The horizontal axis indicates the number of clusters (*k*-value) and the vertical axis indicates the degree of deviation of a cluster from its adjacent cluster at that time. The optimal *k*-value is reflected by the highest point. **(c)** Clustering by the *k*-means method. Clustering is shown by distinct colors and numbers were determined by Silhouette analysis. UMAP, Uniform Manifold Approximation and Projection for Dimension Reduction.

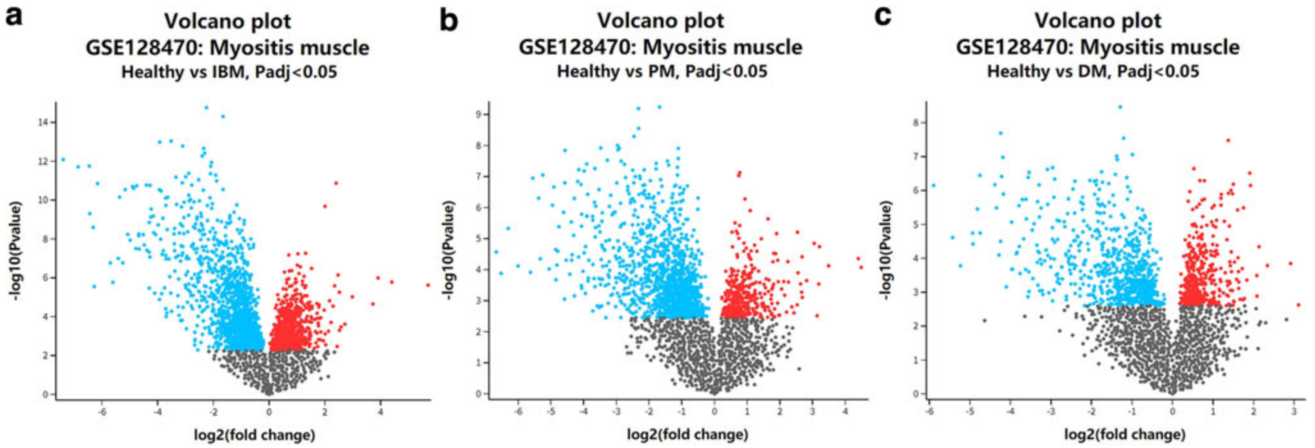

FIG. 2. Discovery of differentially expressed genes. **(a)** The differentially expressed genes for IBM. **(b)** The differentially expressed genes for PM. **(c)** The differentially expressed genes for DM. The vertical axis reflects the intentionality of the statistic (–log10[*p*-value]) and the horizontal axis reflects the magnitude of the change (log2[fold-change]). *Red and blue colors* indicate positive and negative directions of significant expression changes, and *gray* indicates changes below the significance level. DM, dermatomyositis; IBM, inclusion body myositis; PM, polymyositis.

target diseases. Figure 2 shows the significance levels and *p*-values on the ordinate versus the fold change on the abscissa. Genes satisfying the conditions | logFC | > 1 and *p*-value <0.05 (Supplementary Tables S1 and S2) were used for further analysis to identify common gene expression patterns (Supplementary Tables S3–S5).

Exploration of small-molecule compounds that recover gene expression patterns altered by disease

Using data for genes with variable expression, we searched for small-molecule compounds that reversed the expression pattern changes for each disease using L1000CDS^2. The obtained compounds corresponding to each disease are listed in Supplementary Table S6.

When these compounds were compared according to the L1000CDS^2 results, common drugs were identified (Fig. 3, Table 2 and Supplementary Table S7). We found 10 drugs that corresponded to IBM and PM, 10 drugs that corresponded to PM and DM, and 14 drugs that corresponded to all three diseases.

These drugs are small-molecule inhibitors and other interfering agents that reverse the changes in gene-specific expression patterns. Genes regulated by each drug were collated, the number of times the drug appeared was calculated in a sum set of results for drugs that appeared more than once, and areas where two diseases had something in common were sorted (Supplementary Tables S8 and S9). The following L1000 perturbation gene signatures scatter plots were generated:

IBM: https://maayanlab.cloud/clustergrammer/l1000cds2/61082faed99ec600506a634e

PM: https://maayanlab.cloud/clustergrammer/l1000cds2/612306c0d99ec600506a6a88

DM: https://maayanlab.cloud/clustergrammer/l1000cds2/6123071fd99ec600506a6a8a

Biological processes significantly affected by the diseases

The three target disease datasets were formatted and GSEA was performed to compare the functions of the expressed genes. The GO results for the top 20 functions with Nominal *p*-value <0.05 and FDR <0.25 are shown in Supplementary Table S10. Genes obtained from GSEA were classified by function using GO terms (Supplementary Table S11).

Discussion

In this study, we took advantage of the large-scale identification and integration of different levels of information and biological insights that ML offers. The efficiency and accuracy of drug candidate calculations were superior to those of previous studies, effectively improving the likelihood of successful drug repurposing, since all drugs were derived from agents effective against other diseases that clustered together in the same group.

To examine the efficacy of dimensionality reduction by UMAP, we directly clustered the 262 × 22,283 dimensions gene expression data using the *k*-means method as shown in Supplementary Table S12. The result failed to show effective

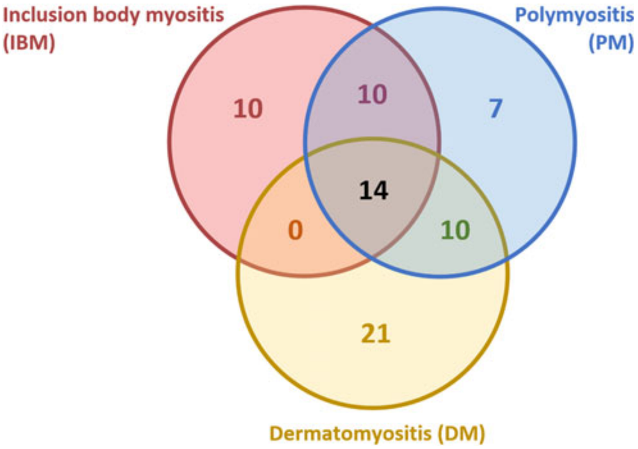

FIG. 3. Number of small-molecule compounds for each disease. Numbers in overlapping regions are compounds shared between the disorders.

Table 2. Breakdown of Common Drugs

	IBM	*PM*	*DM*
PX12	○	○	○
Salermide	○	○	○
NCGC00185684-02	○	○	○
NVP-TAE684	○	○	○
AMSACRINE	○	○	○
Ingenol 3, 20-dibenzoate	○	○	○
THIOTHIXENE	○	○	○
BRD-K32896438	○	○	○
BRD-A34205397	○	○	○
Wortmannin	○	○	○
SB 218078	○	○	○
RESERPINE	○	○	○
CHR 2797	○	○	○
NCGC00188536-01	○	○	○
AT-7519	○	○	—
PMA	○	○	—
CGP-60474	○	○	—
BMS-387032	○	○	—
Mitoxantrone	○	○	—
Alvocidib	○	○	—
NSC 3852	○	○	—
Gemcitabine	○	○	—
NTNCB hydrochloride	○	○	—
Vorinostat	○	○	—
PERHEXILINE MALEATE	—	○	○
BRD-K57080016	—	○	○
Ro 28-1675?	—	○	○
BRD-K25737009	—	○	○
7878890	—	○	○
WZ-4-145	—	○	○
BRD-K19181733	—	○	○
Wiskostatin	—	○	○
BRD-A90643929	—	○	○
Cyclosporine	—	○	○

Dark and light yellow shades represent compounds shared by three and only two diseases, respectively.

DM, dermatomyositis; IBM, inclusion body myositis; PM, polymyositis; PMA, phorbol-12-myristate-13-acetate.

clustering as suggested by Hozumi et al. (2021), although some of the groups appeared to be clustered appropriately as Osteoarthritis were clustered into the same group (no. 7).

In addition, for the *k*-value calculation in the *k*-means method, we first considered using the elbow method, but due to the complexity of the data, a clear inflection point could not be determined. Therefore, we instead used the Silhouette method, which has a wider range of applicability. It combines both cohesion and separation to evaluate the impact on the clustering results produced by different algorithms, or different ways of running the algorithm, based on the same original data. This yielded 20 small-molecule compounds for only one disease, for which applicability to other diseases was unconfirmed, 10 with unconfirmed efficacy only for DM, and 10 with unconfirmed efficacy only for IBM.

Of the 421 genes directly affected by the 10 compounds that were validated for IBM and PM using L1000CDS^2 but not validated for DM, 224 were common to the disease-specific gene expression data for DM obtained using GEO2R. Of the 407 genes directly affected by the 10 compounds whose efficacy was unconfirmed only for IBM, 307 were common to the disease-specific genes for IBM obtained using GEO2R.

Next, we obtained BRD-IDs and PubChem Names for 20 small-molecule compounds with unconfirmed efficacy for only one of the three diseases using side effect prediction based on the L1000 data (http://maayanlab.net/SEP-L1000/index.html). PubChem names were obtained and their efficacy as drugs was investigated in several databases (Supplementary Table S13). Of the 10 compounds that were only validated for IBM and PM, AT-7519, CGP-60474, BMS-387032, and alvocidib inhibit cyclin-dependent kinases; mitoxantrone inhibits DNA cleavage by topoisomerase II; dFdCTP, a conversion product of gemcitabine, competes with deoxycytidine triphosphate (dCTP) to inhibit DNA replication (Chabner et al., 2011); SB-218078 inhibits checkpoint kinase 1 (Chk1); and CHR-2797 converted into a pharmacologically active acid product named CHR-79888 inside cells that inhibits the M1 family aminopeptidases; all of these compounds ultimately induce apoptosis.

Wortmannin is a cell-permeable Phosphatidylinositol 3-kinase (PI3K) inhibitor. Phorbol-12-myristate-13-acetate is a potent tumor promoter that is often used in biomedical research to activate the signal transduction enzyme protein kinase C, and is linked to gene transcription, cell proliferation, differentiation, programmed cell death, and immune pathways.

Of the 10 compounds that were only validated for PM and DM, BRD-K19181733 is a receptor antagonist that inhibits and attenuates dopaminergic effects. Perhexiline maleate is a potent inhibitor of carnitine palmitoyltransferase 1. Selumetinib and 7878890 are highly selective methyl ethyl ketone inhibitors. WZ-4-145 is a selective inhibitor of epidermal growth factor receptor tyrosine kinase activity, which inhibits cell growth and proliferation. Wiskostatin and BRD-A90643929 inhibit protein synthesis and function. RO-28-1675 is a potent allosteric glucokinase activator. CD-1530 markedly enhances catalytic activity, increases glucose metabolism, and lowers blood glucose levels. The mechanism of action of cyclosporin-a is not clear, but it is thought to bind to cytophilin and inhibit calcineurin.

When we contrasted the functions of genes that were specifically altered in one of the target diseases identified by GSEA with the effects of drugs that were found to be effective only in the other two diseases according to L1000CDS^2, many overlaps were found, and most were related to immune responses. Since IBM and DM are both autoimmune inflammatory muscle diseases (Nishino, 2020), it is likely that the reactions inhibited by the 10 validated compounds need to be suppressed for treatment.

The two-stage prediction approach to drug repurposing presented here offers innovation to inform future drug discovery and clinical trials in a variety of human diseases. We predict that drugs shown to be effective for only two diseases may also be effective for other diseases, IBM and DM, for which no efficacy was reported previously.

It should be noted, however, that the clustering of gene expression might reflect shared tissue of origin instead of disease mechanism in common. The fact that body myositis (IBM), PM, and DM clustered together could be due to tissue origin (muscle) rather than common disease etiology or mechanism, or both. Nonetheless, we believe that our proposed method would be useful for drug repurposing because the method focuses of the genes with altered expression under the condition of disorders, which is canceled by the treatment. Further study shall give more concrete perspectives.

Authors' Contributions

Y.C. designed the study, performed data analysis, and wrote the article. M.S. contributed the conceptual design. F.I. contributed writing the article. T.E. and N.O. contributed to the research question and edited the article. All authors have made a significant intellectual contribution, read, and approved the article.

Acknowledgments

The authors thank the anonymous reviewers as well as the editor for their valuable and constructive suggestions to improve the article.

Author Disclosure Statement

The authors declare they have no conflicting financial interests.

Funding Information

No funding was received for this article.

Supplementary Material

Supplementary Data SD1
Supplementary Table S1
Supplementary Table S2
Supplementary Table S3
Supplementary Table S4
Supplementary Table S5
Supplementary Table S6
Supplementary Table S7
Supplementary Table S8
Supplementary Table S9
Supplementary Table S10
Supplementary Table S11
Supplementary Table S12
Supplementary Table S13

References

Chabner BA, Amrein PC, Druker BJ, et al. (2011). Antineoplastic agents. In: *Goodman & Gilman's The Pharmacological Basis of Therapeutics*, 12th ed. Brunton LL, and Parker KL, eds. New York: McGraw-Hill Education Press, 1315–1403.

Chen B, Ma L, Paik H, et al. (2017). Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nat Commun 8, 16022.

Ciallella JR, and Reaume AG. (2017). *In vivo* phenotypic screening: Clinical proof of concept for a drug repositioning approach. Drug Discov Today Technol 23, 45–52.

Duan Q, Reid SP, Clark NR, et al. (2016). L1000CDS2: LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl 2, 16015.

Farghali H, Kutinová CN, and Arora M. (2021). The potential applications of artificial intelligence in drug discovery and development. Physiol Res 70(Suppl. 4), S715–S722.

Garvey C. (2018). Interview with colin garvey, rensselaer polytechnic institute. Artificial intelligence and systems medicine convergence. OMICS J Integr Biol 22. DOI: 10.1089/omi.2017.0218.

Hirschhorn JN, and Daly MJ. (2005). Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6, 95–108.

Hozumi Y, Wang R, Yin C, et al. (2021). UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. Comput Biol Med 131, 104264.

Jin G, and Wong ST. (2014). Toward better drug repositioning: Prioritizing and integrating existing methods into efficient pipelines. Drug Discov Today 19, 637–644.

Jourdan J-P, Bureau R, Rochais C, et al. (2020). Drug repositioning: A brief overview. J Pharm Pharmacol 72, 1145–1151.

Khyati M. (2019). Analytics Vidhya. How to Determine the Optimal K for K-Means? Pune, Maharashtra, 2019. https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb. Accessed July 17, 2019.

Kwon OS, Kim W, Cha HJ, et al. (2019). In silico drug repositioning: From large-scale transcriptome data to therapeutics. Arch Pharm Res 42, 879–889.

McInnes L, Healy J, and Melville J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. DOI: 10.48550/arXiv.1802.03426.

Napolitano F, Zhao Y, Moreira VM, et al. (2013). Drug repositioning: A machine-learning approach through data integration. J Cheminform 5, 30.

Nishino I. Classification of Autoimmune Myositis. Neurodiem, 2020. https://www.neurodiem.jp/news/classification-of-autoimmune-myositis-5VZWDib2pYfHBLISriomq8. Accessed February 19, 2020.

Park T, Yi S-G, Kang S-H, et al. (2003). Evaluation of normalization methods for microarray data. BMC Bioinform 4, 44.

Rapicavoli RV, Alaimo S, Ferro A, et al. (2022). Computational methods for drug repurposing. Adv Exp Med Biol 1361, 119–141.

Robinson MD, and Oshlack A. (2010). Scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11, R25.

Rodriguez S, Hug C, Todorov P, et al. (2021). Machine learning identifies candidates for drug repurposing in Alzheimer's disease. Nat Commun 12, 1033.

Roundtable on Translating Genomic-Based Research for Health, Board on Health Sciences Policy, Institute of Medicine. (2014). *Drug Repurposing and Repositioning*. Washington (DC): National Academies Press, 2014.

Thomas MA, Yang L, Carter BJ, et al. (2011). Gene set enrichment analysis of microarray data from Pimephales promelas (Rafinesque), a non-mammalian model organism. BMC Genom 12, 66.

Turanli B, Grøtli M, Boren J, et al. (2018). Drug repositioning for effective prostate cancer treatment. Front Physiol 9, 500.

Zeng X, Zhu S, Lu W, et al. (2020). Target identification among known drugs by deep learning from heterogeneous networks. Chem Sci 11,1775–1797.

Zhang C, Berndt-Paetz M, and Neuhaus J. (2020). Identification of key biomarkers in bladder cancer: Evidence from a bioinformatics analysis. Diagnostics 10, 66.

Address correspondence to:
Toshinori Endo, PhD
Laboratory of Information Biology
Information Science and Technology
Hokkaido University
Sapporo 060-0814
Japan

E-mail: endo@ist.hokudai.ac.jp

Abbreviations Used

AD	=	Alzheimer’s disease
AI	=	artificial intelligence
BOEC	=	blood outgrowth endothelial cells
Chk1	=	checkpoint kinase 1
CPT1	=	carnitine palmitoyltransferase 1
dCTP	=	deoxycytidine triphosphate
DM	=	dermatomyositis
DRIAD	=	drug repurposing in Alzheimer’s disease
FDA	=	U.S. Food and Drug Administration
FDR	=	false discovery rate
GEO	=	Gene Expression Omnibus
GO	=	Gene Ontology
GOFAI	=	good old fashioned artificial intelligence
GSEA	=	gene set enrichment analysis
IBM	=	inclusion body myositis
IgA	=	immunoglobulin A
logFC	=	log fold-change
MAS5	=	Affymetrix MicroArray Suite 5.0
ML	=	machine learning
NCBI	=	National Center for Biotechnology Information
PBMC	=	peripheral blood mononuclear cells
PCA	=	principal component analysis
PI3K	=	permeable Phosphatidylinositol 3-kinase
PM	=	polymyositis
PMA	=	phorbol-12-myristate-13-acetate
RMA	=	robust multiarray average
UMAP	=	Uniform Manifold Approximation and Projection for Dimension Reduction