Application of augmented topic model to predicting biomarkers and therapeutic targets using multiple human disease-omics datasets

Satoshi Kozawa[1,2], Kyoji Urayama[1,2], Kengo Tejima[1,2], Hotaka Doi[1,2,3], Hirona Yokoyama[1,2,3], Yuhei Ueno[1,2,3], Thomas N. Sato[1,2,3,¶]

[1]The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan; [2]Karydo TherapeutiX, Inc., Kyoto, Japan; [3]V-iCliniX Laboratory, Nara Medical University, Nara, Japan

[¶]Corresponding Author

Thomas N. Sato, Ph.D.

[1]The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International (ATR)

2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan

TEL: +81-774-95-2311

Email: island1005@gmail.com

SUMMARY

Human diseases are multifactorial - hence it is important to characterize diseases on the basis of multiple disease-omics. However, the capability of the existing methods is largely limited to classifying diseases based on a single type or a few closely related omics data. Herein, we report a topic model framework that allows for characterizing diseases according to their multiple omics data. We also show that this method can be utilized to predict potential biomarkers and/or therapeutic targets. In this study, we illustrate a computational concept of this augmented topic model and demonstrate its prediction performance by a leave one-disease features out cross-validation scheme. Furthermore, we exploit this method together with human disease tissue/organ-transcriptome data and identify putative biomarkers and/or therapeutic targets across 79 diseases. In conclusion, this method and the prediction framework shown reported herein provide important tools for understanding complex human diseases and also facilitate diagnostic and/or therapeutic development.

INTRODUCTION

Computational methods for predicting potential biomarkers and/or therapeutic targets for human diseases are transforming medicine and drug development (Terkelsen et al., 2020; Zhang et al., 2021). One of such approaches is graph theory-based methods (Hwang et al., 2008; Leclercq et al., 2019; Wang et al., 2020). This builds a network of diseases on the basis of a single type of disease-related information, such as genetic influences on diseases. This allows for detecting the genetic biomarkers according to the genetic relatedness of the diseases. The same framework is used to build diseases on the basis of other information such as drug targets to predict novel therapeutic targets. Another approach is based on natural language processing models (Gyori et al., 2017; Haggerty and Purvis, 2017). This method exploits narrative descriptions of biological processes in numerous publications and other documents to build a model/models to search for "likely" biomarkers and/or therapeutic targets.

These approaches are common and, in some cases, effective in finding biomarkers and therapeutic targets. However, diseases are influenced by many factors, such as genes, proteins, cell-cell interactions, life-style, environments, microbiota in the human body, etc. Hence, it would be useful to have a method that identifies latent disease-disease association on the basis of multiple disease-omics data. Such an approach could facilitate the discovery of biomarkers and/or therapeutic targets across multiple disease related factors. For example, a close disease-disease relationship based on the genetics could be exploited to discover a life-style biomarker for the same diseases. A close microbiota relationship of the diseases could be translated to a related drug-targets.

Hence, we report a topic model framework to characterize disease-disease relationship based on multiple disease-omics and to discover biomarkers and therapeutic targets.

RESULTS

Augmented Latent Dirichlet Allocation (LDA) model

LDA model is a topic model for bag-of-words data (Blei et al., 2003). The conventional LDA model classifies sentences/documents as topics on the basis of their associated words. Diseases are influenced

by multiple-types of factors such as genes, microbiota, life-style, etc. This is equivalent to describing diseases by using words of genes, microbiota, life-style, etc.

To accommodate such multifactorial nature of human diseases, we augmented a conventional LDA model to classify diseases according to multiple disease-influencing factors (see METHODS) (Figure 1). In this paper, we classified diseases to three topics according to their associated single-nucleotide-polymorphisms (SNPs), drug-targets (i.e., therapeutic targets for diseases), microbiota, and/or human disease-associated tissue/organ transcriptome data using GWAS catalog (https://www.ebi.ac.uk/gwas/docs/file-downloads), SIDER (http://sideeffects.embl.de)/DrugBank (https://go.drugbank.com/), Disbiome (https://disbiome.ugent.be/home) (Janssens et al., 2018), NCBI GEO DataSets (https://www.ncbi.nlm.nih.gov/gds), respectively (see METHODS).

Performance of the augmented LDA

Our aim is to use this augmented LDA framework to predict disease-biomarkers and/or therapeutic targets. For this objective, we left out one type of disease-associated information and keeping the rest intact and evaluated the performance in predicting the missing information for each disease (Figure 2). We conducted this "leave one-disease features out cross-validation" scheme to evaluate two models (Figure 2, Supplementary table 1).

The first model consists of disease-associated SNPs and drug-targets features. In this model, there are 170 diseases that have both SNPs and the drug-targets information (Figure 2a, Supplementary table 1). We left out all SNPs associated with a disease and keeping all the SNPs associated with the 169 diseases and maintaining all the drug-targets information for all the 170 diseases. Then, we evaluated the performance in predicting the missing SNPs associated with its associated disease. We repeated this evaluation for all 170 diseases.

We also left out all drug-targets information associated with a disease and keeping all the drug-targets associated with the 169 diseases and keeping all the SNPs information for all the 170 diseases. Then, we evaluated the performance in predicting the missing drug-targets associated with the disease. We repeated this evaluation for all 170 diseases.

The prediction performance was evaluated by calculating AUC scores for each disease (Figure 2a, see also METHODS). In predicting the missing drug-targets, 104 out of 170 diseases show the AUC scores over 0.7 and 81 show the AUC score over 0.8. In contrast, in predicting the missing SNPs, only four diseases show the AUC score over 0.7.

We also evaluated the performance of the augmented LDA model of the drug-targets/microbiota combination (Figure 2b). This combination has a total of 104 diseases (Figure 2b, Supplementary table 1). For this combination, 61 diseases show the AUC score over 0.7 for predicting the missing microbiota and 39 show the score over 0.8. In the prediction of the missing drug-targets, 48 diseases show the score over 0.7 and 24 are over 0.8.

Prediction of putative biomarkers and therapeutic-targets

Next, we applied this predication framework to the disease-associated tissue/organ transcriptome data (Figure 3). We combined the drug-targets, microbiota and the disease-associated tissue/organ transcriptome data. Out of the 104 diseases with the drug-targets and microbiota features, 25 are also tagged with the tissue/organ transcriptome features, hence 79 diseases miss the transcriptome information. These predicted tissue/organ-genes (transcriptome) could serve as biomarkers and/or therapeutic-targets for these diseases.

The analysis of the 79 diseases resulted in identifying a large number of tissue/organ-genes for each disease (Figure 3). In particular, some tissue/organ-genes show significantly higher counts (1,018 – 2,911 counts) than the rest (Figure 3, Supplementary table 3), indicating that they are more likely candidates for biomarkers and/or therapeutic-targets for the corresponding diseases.

DISCUSSION

In this paper, we report an augmented LDA framework that serves as another computational method to classify diseases and to predict their biomarkers and/or therapeutic-targets. In this method, diseases are tagged by multiple disease-omics features such as SNPs, microbiota, transcriptome, drug-targets, etc. This is an advantage over the other related methods where diseases are mostly described by a single type of omics features. Onset and progression of virtually all diseases are multifactorial, hence, our approach provides a useful tool to characterize diseases and also to identify candidate targets for diagnostic and therapeutic development.

The conventional topic model classify subjects to topics based on their associated single type of words/features (Blei et al., 2003). We augmented this framework to be able to characterize diseases based on multiple-types of disease-omics data, resulting in the augmented LDA scheme shown in Figure 1 (see also METHODS). This is on the assumption that each type of disease-omics features displays the same statistical distribution across different types of diseases (Figure 1, see also METHODS).

Using this augmented LDA model, we also devised a prediction framework that could be used to deduce new disease-associated features (Figure 2, see also METHODS). This is accomplished by predicting missing disease-omics features for diseases. To evaluate its performance, we conducted "leave one-disease features out cross-validation" scheme (Figure 2, see also METHODS). In predicting the missing microbiota and drug-targets, the 23% - 48% of the diseases show the AUC scores above 0.8, and the 46% - 61% exhibit the AUC scores above 0.7. In contrast, the AUC score for predicting the missing SNPs features is extremely low – only 2% of the diseases show the AUC scores above 0.7. This could be due to the far larger number of SNPs features associated with each disease as compared to the other features such as microbiota and drug-targets. Alternatively, this could also be due to the extensive overlaps of SNPs among many diseases, making it difficult to identify unique SNPs features for a disease.

To find potential biomarkers and/or therapeutic-targets, we conducted the augmented LDA based prediction scheme together with human disease-associated tissue/organ-transcriptome data (Figure 3, see also METHODS). This approach resulted in identifying a relatively small (1 – 20) number of tissue/organ-genes highly associated with each of the 79 diseases (Figure3, Supplementary table 3). We found that many are systemic origins (blood, immune cells) (Supplementary table 3). This is in part due to the biased tissue/organ origins of the transcriptome datasets that are available (see METHODS). Despite this biased result, such origins provide an opportunity to identify system/circulating biomarkers and/or therapeutic-targets which are easier targets for diagnostic and/or therapeutic development.

In conclusion, the augmented LDA model and the prediction framework reported in this paper, together with ever-growing disease-omics datasets, provide an opportunity to characterize complex nature of human diseases and also to facilitate mining latent biomarkers and/or therapeutic-targets.

METHODS

Augmented Latent Dirichlet Allocation (LDA) model

The conventional LDA model is augmented to classify diseases according to disease-associated SNPs, drug-targets, microbiota, transcriptome, etc. as follows (the framework is also schematically depicted in Figure 1):

Let $w_{di}^{(type)}$ be the i-th information of disease d acquired from (type) information, $z_{di}^{(type)}$ be the topic number of $w_{di}^{(type)}$, θ_d be the topic probability of disease d, and $\phi_k^{(type)}$be the word probability of topic k at (type) information. The type here is SNPs, drug-targets, microbiota, transcriptome, etc. The total number of topic K is a hand-tuned parameter and K was set to 3 in this paper. The generative model of the proposed topic models was defined as following,

$$p\left(w_{di}^{(type)}\middle|z_{di}^{(type)},\ \boldsymbol{\phi}^{(type)}\right) = Multinomial\left(w_{di}^{(type)}\middle|\boldsymbol{\phi}^{(type)}\right)$$

$$= \prod_{v=1}^{V^{(type)}} \prod_{k=1}^{K} \phi^{(type)}{}_{vk}^{\delta\left(z_{di}^{(type)}=k\right)\delta\left(w_{di}^{(type)}=v\right)}$$

$$p\left(z_{di}^{(type)}\middle|\boldsymbol{\theta}_d\right) = Multinomial\left(z_{di}^{(type)}\middle|\boldsymbol{\theta}_d\right) = \prod_{k=1}^{K} \theta_{dk}^{\delta\left(z_{di}^{(type)}=k\right)}$$

$$p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) = Dirchlet(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1}$$

$$p\left(\boldsymbol{\phi}_k^{(type)}\middle|\boldsymbol{\beta}^{(type)}\right) = Dirichlet\left(\boldsymbol{\phi}_k^{(type)}\middle|\boldsymbol{\beta}^{(\boldsymbol{type})}\right) = \frac{\Gamma\left(\sum_v \beta_v^{(type)}\right)}{\prod_v^{V^{(type)}} \Gamma\left(\beta_v^{(type)}\right)} \prod_{v=1}^{V^{(type)}} \phi^{(type)}{}_{kv}^{\beta_v^{(type)}-1}$$

where, $\boldsymbol{\alpha} \in \mathbb{R}^{+K}$ and $\boldsymbol{\beta}^{(type)} \in \mathbb{R}^{+V^{(type)}}$ are the hyperparameters and we set these to a vector that elements have 0.1, v is the index word, $V^{(type)}$ is the total kind of word at a (type) data source, $\Gamma(\cdot)$ is the Gamma function and $\delta(\cdot)$ is the Kronecker delta function. Based on the model, we estimated the variables $z_{di}^{(type)}, \boldsymbol{\theta}_d\ and\ \boldsymbol{\phi}_k^{(type)}$, and also estimated a part of $w_{di}^{(type)}$ that not observed at some (type) data sources by using Gibbs sampling method. The conditional distributions were calculated based on Gibbs sampling method's manner as following,

$$p\left(z_{di}^{(type)} = k \middle| w_{di}^{(type)} = v, \boldsymbol{\theta}_d, \boldsymbol{\phi}^{(type)}\right) = \frac{\theta_{dk}\phi_{kv}^{(type)}}{\sum_{k'} \theta_{dk'}\phi_{k'v}^{(type)}}$$

$$p\left(\boldsymbol{\theta}_d \middle| \boldsymbol{z}^{(type)}, \boldsymbol{\alpha}\right) = Dirichlet(\boldsymbol{\theta}_d | \boldsymbol{\alpha}') \propto \prod_{k=1}^{K} \theta_{dk}{}^{\alpha_k + \sum_t \sum_i \delta\left(z_{di}^{(type)} = k\right) - 1}$$

$$p\left(\boldsymbol{\phi}_k^{(type)} \middle| \boldsymbol{z}^{(type)}, \boldsymbol{w}^{(type)}, \boldsymbol{\beta}^{(type)}\right) = Dirichlet\left(\boldsymbol{\phi}_k^{(type)} \middle| \boldsymbol{\beta}^{(type)\prime}\right)$$

$$\propto \prod_{v} \phi^{(type)}{}_{kv}^{\beta_k^{(type)} + \sum_d \sum_i \delta\left(z_{di}^{(type)} = k\right) \delta\left(w_{di}^{(type)} = v\right) - 1}$$

$$p\left(w_{di}^{(type)} = v \middle| z_{di}^{(type)} = k, \boldsymbol{\phi}^{(type)}\right) = \phi_{kv}^{(type)}$$

From this conditional distribution, we repeated to sample $z_{di}^{(type)}, \boldsymbol{\theta}_d$ *and* $\boldsymbol{\phi}_k^{(type)}$ and $w_{di}^{(type)}$, until they all converge. In this paper, the number of iterations was set to 5,000, that was decided by manually decision. The initial value of $\boldsymbol{\theta}_d$ *and* $\boldsymbol{\phi}_k^{(type)}$ were set to uniform distribution, $w_{di}^{(type)}$ was set to word from random sampled from $\boldsymbol{\phi}_k^{(type)}$. The total number of $w_{di}^{(type)}$ at disease d was decided by sampling from binominal distribution. The binominal distribution's parameters were estimated by maximum likelihood estimation method by using only observed data. The estimated variables $z_{di}^{(type)}, \boldsymbol{\theta}_d$ *and* $\boldsymbol{\phi}_k^{(type)}$ could be predicted by averaging sampled variables.

SNPs and Microbiota datasets

The disease-associated SNPs and the disease-associated microbiota datasets are from "All associations" (v1.0.2) from GWAS Catalog (https://www.ebi.ac.uk/gwas/docs/file-downloads) and Disbiome (version on 11th of November 2020, https://disbiome.ugent.be/home) (Janssens et al., 2018).

Drug-targets dataset

The disease-associated drug-targets dataset is built as follows: First, we downloaded therapeutic-targets (i.e., diseases) information for drugs from SIDER (version 4.1) (http://sideeffects.embl.de). Next, we searched for drug-targets from DrugBank (https://go.drugbank.com/). This process was performed by web-scraping on Python (version 3.8.1).

Disease-associated tissue/organ transcriptome dataset

Human disease-associated bulk tissue/organ transcriptome data were automatically collected from NCBI GEO DataSets (https://www.ncbi.nlm.nih.gov/gds) by using the following keywords. The keywords were "RNA-seq"[All Fields] AND "Homo sapiens"[organism] NOT "scRNA"[All Fields] NOT "Single cell"[All Fields] NOT "culture"[All Fields] NOT "cultured"[All Fields] NOT "cell line" [All Fields] NOT "treatment"[All Fields] NOT "single-nucleus"[All Fields] NOT "microRNA"[All Fields] NOT "iPSC"[All Fields] NOT "organoids"[All Fields] AND "expression profiling by high throughput sequencing"[DataSet Type] AND "SIDER's indication"[All Fields]. Only SIDER's indication input-term was automatically changed in these keywords. The SIDER's indication list is available from the download page of SIDER 4.1 (http://sideeffects.embl.de). After the first automated selection of GEO datasets, we manually selected the GEO datasets that contain both human disease-associated bulk tissue/organ transcriptome data and control or healthy data in the same project, resulting in 50 GEO datasets useful for our analyses. The selected GEO datasets are listed in Supplementary table 2. Next, we prepared raw count data for each selected GEO dataset. In 17 GEO datasets (GSE107638, GSE112057, GSE112523, GSE114007, GSE117970, GSE121212, GSE124439, GSE131031, GSE131525, GSE133624, GSE134056, GSE138082, GSE138614, GSE146889, GSE42546, GSE80336, GSE80655), the raw count data were downloaded from NCBI GEO DataSets websites, whereas in the other 33 GEO datasets, the raw count data were produced from FASTQ data as follows. Briefly, FASTQ data were downloaded from European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) via FTP downloading or the fast file-downloading software, Aspera Connect (https://www.ibm.com/aspera/connect/; version 3.11.1.58). The downloaded FASTQ data were mapped onto human genome (UCSC hg38) with STAR (version 2.7.8a) to produce BAM data. The produced BAM data were converted into raw count data by the human gene GTF file (UCSC hg38). This data conversion was performed by the software of Subread (version 2.0.1), where we used `featureCounts` command with the option of `-Q 10` and, only in the case that data was paired-end, additionally with the options, `-p` and `-B`. The processes of STAR and Subread were performed on Linux of CentOS6 or CentOS7. If necessary, the raw count data were combined into one raw count data

by GSM ID, because some data were deposited as splitted FASTQ data. For our analyses, we extracted the differentially expressing genes (DEGs) with DESeq2 (version 1.30.1) (Love et al., 2014) on free software R (version 4.0.3). DESeq2 analysis was conducted for the respective GSE project. First, we classified transcriptome data into the control group from healthy tissue/normal tissue parts and the disease group from lesions by each tissue. This classification was performed on sample information describing in NCBI GEO DataSets. Then, we performed DESeq2 analysis for each disease in each GSE project. To select DEGs, we typically filtered genes by adjusted p value below 0.0001. However, if we could not have any DEGs in filtering by adjusted p value below 0.0001, the threshold of adjusted p value was incremented from 0.0001, 0.001, 0.01, to 0.05 until we could have at least one DEG. In addition, tissue/organ names were manually unified to the rough class.

UMLS IDs assignment to the disease names

To combine disease names acquired from multiple data sources, we add UMLS IDs to the disease names. The UMLS ID annotation was performed by `UMLS_AUI.extract_terminology ("ICD10")` function from python library PyMedTermino (version 0.3.3) (Lamy et al., 2015). Prior to performing this function, `'s` was replaced by a blank space in disease names. Following assigning the UMLS ID annotations, the UMLS IDs were combined by string '|' if these UMLS ID have the same disease names. For the disease names that this UMLS ID annotation method failed, the actual disease names in the datasets were used and only those with the exact matching names were combined.

AUC score calculations

The prediction performance of the augmented LDA model was evaluated. In the SNPs/drug-targets model, a total of 170 diseases are tagged with both SNPs and drug-targets features (i.e., "type" in the model description of the augmented LDA above). The SNPs or the drug-targets features were left-out from one disease at a time and the missing features were predicted. The prediction performance was evaluated by calculating the AUC score for each type of features (i.e., SNPs, drug-targets, microbiota) for each disease. The same evaluation was conducted on the microbiota/drug-targets model (104 diseases with both features). The AUC scores were calculated as follows: Each feature was ranked

according to the sampling frequency with 1 as the highest and 0 as the lowest. This ranking score for each feature is compared to the score 1 for the known (observed) features and the score 0 for all the other features (i.e., unobserved features). The AUC scores are calculated for each type of features by comparing the score of all the sampled features for each type. Hence, for each type of features, the higher AUC score indicates the higher prediction performance for the disease.

ACKNOWLEDGEMENTS

We thank R. Takahashi, R. Kitaura, K. Sugisaka for administrative assistance. We are also grateful to the members of Sato lab at ATR and Karydo TherapeutiX, Inc. and for advice and discussion throughout the course of this work. This work was supported in part by JST ERATO Grant Number JPMJER1303(T.N.S), Nakatani Foundation (T.N.S) and AMED under Grant Number JP21he2102002 (T.N.S).

AUTHOR CONTRIBUTIONS

T.N.S. conceived the idea of the project, designed the study and supervised the overall research project. S.K. designed and performed the augmented LDA analyses. K.U., K.T., H.D., H.Y., Y.U. establishing the datasets and analyzing human transcriptome data. T.N.S., S.K., K.U., K.T. wrote the manuscript.

DECLARATION OF INTERESTS

T.N.S., S.K., K.U., K.T., H.D., H.Y., Y.U. are employees of Karydo TherapeutiX, Inc.

Correspondence and requests for materials should be addressed to T.N.S. (island1005@gmail.com).

REFERENCES

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research *3*, 993-1022.

Gyori, B.M., Bachman, J.A., Subramanian, K., Muhlich, J.L., Galescu, L., and Sorger, P.K. (2017). From word models to executable models of signaling networks using automated assembly. Molecular systems biology *13*, 954.

Haggerty, R.A., and Purvis, J.E. (2017). Natural language processing: put your model where your mouth is. Molecular systems biology *13*, 958.

Hwang, T., Sicotte, H., Tian, Z., Wu, B., Kocher, J.P., Wigle, D.A., Kumar, V., and Kuang, R. (2008). Robust and efficient identification of biomarkers by classifying features on graphs. Bioinformatics *24*, 2023-2029.

Janssens, Y., Nielandt, J., Bronselaer, A., Debunne, N., Verbeke, F., Wynendaele, E., Van Immerseel, F., Vandewynckel, Y.P., De Tre, G., and De Spiegeleer, B. (2018). Disbiome database: linking the microbiome to disease. BMC Microbiol *18*, 50.

Lamy, J.B., Venot, A., and Duclos, C. (2015). PyMedTermino: an open-source generic API for advanced terminology services. Stud Health Technol Inform *210*, 924-928.

Leclercq, M., Vittrant, B., Martin-Magniette, M.L., Scott Boyer, M.P., Perin, O., Bergeron, A., Fradet, Y., and Droit, A. (2019). Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. Front Genet *10*, 452.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology *15*, 550.

Terkelsen, T., Krogh, A., and Papaleo, E. (2020). CAncer bioMarker Prediction Pipeline (CAMPP)-A standardized framework for the analysis of quantitative biological data. PLoS computational biology *16*, e1007665.

Wang, Z., He, Z., Shah, M., Zhang, T., Fan, D., and Zhang, W. (2020). Network-based multi-task learning models for biomarker selection and cancer outcome prediction. Bioinformatics *36*, 1814-1822.

Zhang, X., Jonassen, I., and Goksoyr, A. (2021). Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data. In Bioinformatics, I.N. Helder, ed. (Brisbane (AU)).

FIGURE LEGENDS

Figure 1. The concept of the augmented topic model based on latent Dirichlet allocation (LDA).

The concept is graphically shown. The empty circles are random variables. The grey circles represent observed/input data. The black dots are hyperparameters. See METHODS for details.

Figure 2. The "leave one-disease features out cross-validation".

The input data are (a) SNPs/drug-targets and (b) microbiota/drug-targets. The prediction performance is evaluated and shown as AUC scores. The diseases are shown as disease ID numbers (see Supplementary table 1 for the correspondence between the ID numbers and the diseases).

Figure 3. Prediction of the tissue/organ-transcriptome features for 79 diseases.

The results for each of the 79 diseases are shown as histogram of the likeliness of the identified tissue/organ-transcriptome features. Each histogram represents the result of each disease which is indicated as disease ID number at the top. The likeliness is represented as the frequency of the sampling of each feature (sampling counts) on the vertical axis. Each tissue/organ-transcriptome feature is represented on the horizonal axis. The overall total number of sampling is indicated at the top-right corner in each histogram (i.e., each disease). The top 20 most frequently predicated features (tissue/organ-gene) for each disease are summarized in Supplementary table 3.

SUPPLEMENTARY TABLES

Supplementary table 1. Disease information.

The disease information is summarized in this table. The UMLS IDs, the presence (1) or the absence (0) of the data (SNPs, drug-targets, microbiota, human disease-associated tissue/organ-transcriptome) are indicated. The human disease-associated tissue/organ-transcriptome is labeled as "Human transcriptome" in the table. The origin of the human disease-associated tissue/organ-transcriptome data for each disease is indicated as GSE numbers.

Supplementary table 2. The human disease-associated tissue/organ-transcriptome datasets.

The source of the human disease-associated tissue/organ-transcriptome datasets for the 25 diseases analyzed in this study are summarized. The diseases are identified as UMLS ID or the actual disease names. The dataset sources are identified as GSE numbers. The number of datasets for the diseases samples ("Disease_num") and the matched controls (Ctrl_num") are indicated. Tissue/organ origins are indicated ("Tissue_origin")

Supplementary table 3. The top 20 most frequently predicated tissue/organ-gene features for each disease.

The frequency for each feature ("Tissue-gene") is indicated as "count" in the table. Each disease is indicated as disease ID and the UMLS ID (or disease names in the case of no available UMLS ID).

Figure 1

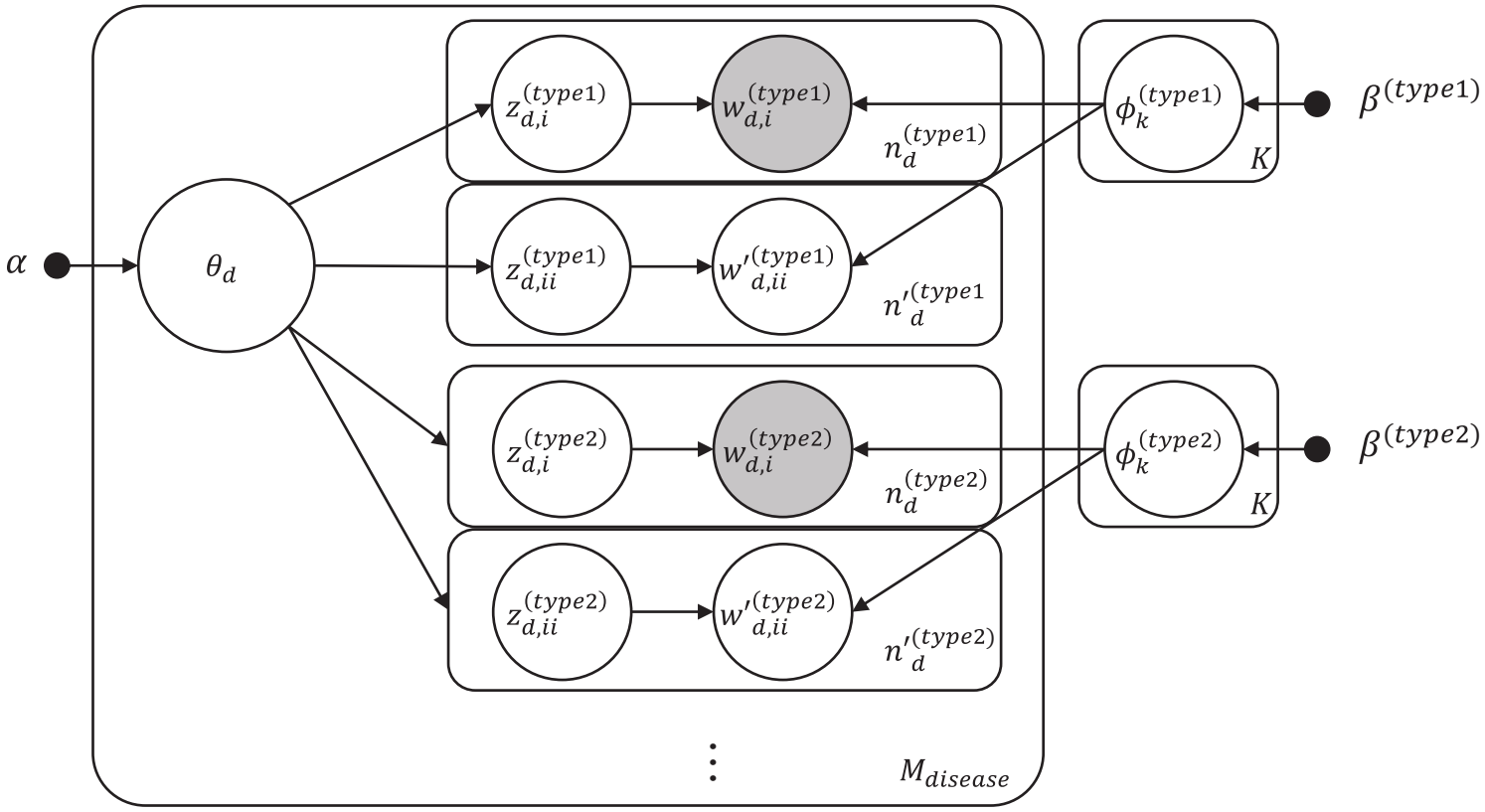
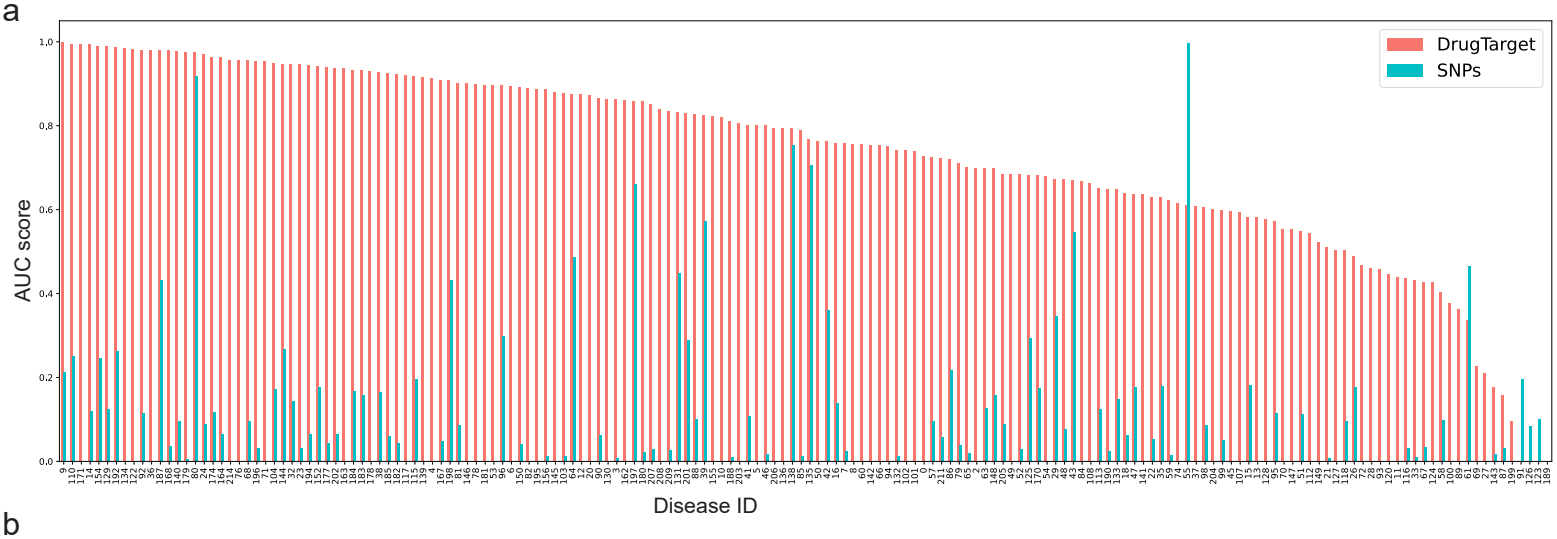

α
θ_d
$z_{d,i}^{(type1)}$
$w_{d,i}^{(type1)}$
$n_d^{(type1)}$
$z_{d,ii}^{(type1)}$
$w'^{(type1)}_{d,ii}$
$n'^{(type1}_d$
$z_{d,i}^{(type2)}$
$w_{d,i}^{(type2)}$
$n_d^{(type2)}$
$z_{d,ii}^{(type2)}$
$w'^{(type2)}_{d,ii}$
$n'^{(type2)}_d$
$M_{disease}$
$\phi_k^{(type1)}$
K
$\beta^{(type1)}$
$\phi_k^{(type2)}$
K
$\beta^{(type2)}$

Figure 2

a


DrugTarget
SNPs
AUC score
Disease ID

b

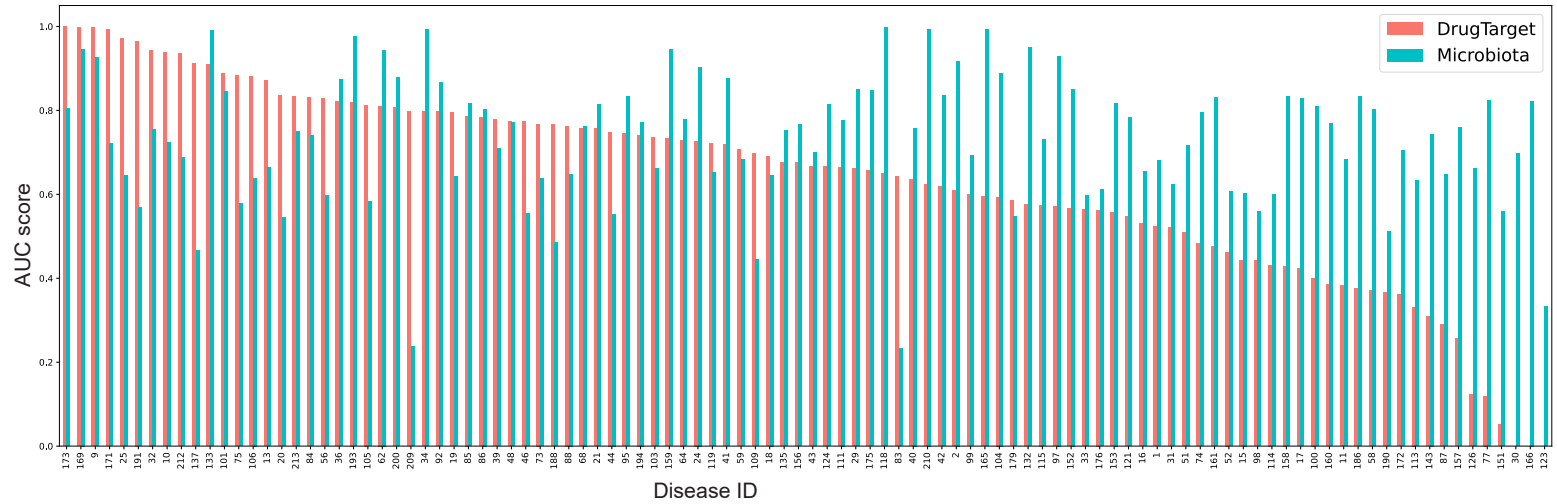
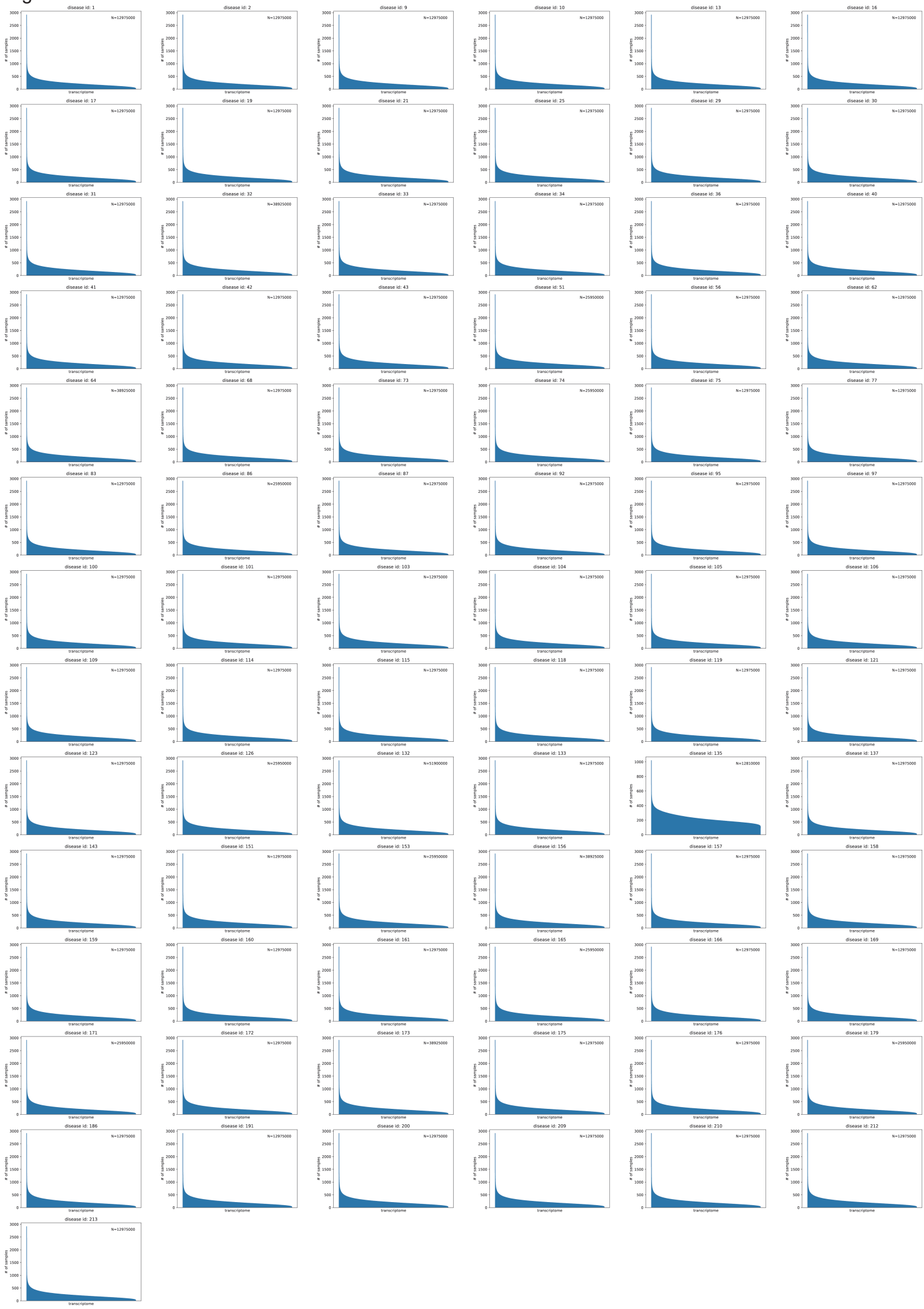

DrugTarget
Microbiota
AUC score
Disease ID

Figure 3


disease id: 1
N=12975000
disease id: 2
N=12975000
disease id: 9
N=12975000
disease id: 10
N=12975000
disease id: 13
N=12975000
disease id: 16
N=12975000
disease id: 17
N=12975000
disease id: 19
N=12975000
disease id: 21
N=12975000
disease id: 25
N=12975000
disease id: 29
N=12975000
disease id: 30
N=12975000
disease id: 31
N=12975000
disease id: 32
N=38925000
disease id: 33
N=12975000
disease id: 34
N=12975000
disease id: 36
N=12975000
disease id: 40
N=12975000
disease id: 41
N=12975000
disease id: 42
N=12975000
disease id: 43
N=12975000
disease id: 51
N=25950000
disease id: 56
N=12975000
disease id: 62
N=12975000
disease id: 64
N=38925000
disease id: 68
N=12975000
disease id: 73
N=12975000
disease id: 74
N=25950000
disease id: 75
N=12975000
disease id: 77
N=12975000
disease id: 83
N=12975000
disease id: 86
N=25950000
disease id: 87
N=12975000
disease id: 92
N=12975000
disease id: 95
N=12975000
disease id: 97
N=12975000
disease id: 100
N=12975000
disease id: 101
N=12975000
disease id: 103
N=12975000
disease id: 104
N=12975000
disease id: 105
N=12975000
disease id: 106
N=12975000
disease id: 109
N=12975000
disease id: 114
N=12975000
disease id: 115
N=12975000
disease id: 118
N=12975000
disease id: 119
N=12975000
disease id: 121
N=12975000
disease id: 123
N=12975000
disease id: 126
N=25950000
disease id: 132
N=51900000
disease id: 133
N=12975000
disease id: 135
N=12810000
disease id: 137
N=12975000
disease id: 143
N=12975000
disease id: 151
N=12975000
disease id: 153
N=25950000
disease id: 156
N=38925000
disease id: 157
N=12975000
disease id: 158
N=12975000
disease id: 159
N=12975000
disease id: 160
N=12975000
disease id: 161
N=12975000
disease id: 165
N=25950000
disease id: 166
N=12975000
disease id: 169
N=12975000
disease id: 171
N=25950000
disease id: 172
N=12975000
disease id: 173
N=38925000
disease id: 175
N=12975000
disease id: 176
N=12975000
disease id: 179
N=25950000
disease id: 186
N=12975000
disease id: 191
N=12975000
disease id: 200
N=12975000
disease id: 209
N=12975000
disease id: 210
N=12975000
disease id: 212
N=12975000
disease id: 213
N=12975000
# of samples
transcriptome