



HOKKAIDO UNIVERSITY

Title	Disease Similarity Analysis Based on Disease-Specific Transcriptomic Alterations : Development of an Automated Framework and Application to Drug Repositioning
Author(s)	CONG, Yi
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第16989号
Issue Date	2026-03-25
DOI	https://doi.org/10.14943/doctoral.k16989
Doc URL	https://hdl.handle.net/2115/99833
Type	doctoral thesis
File Information	Cong_Yi.pdf, 全文



Disease Similarity Analysis Based on Disease-Specific Transcriptomic Alterations: Development of an Automated Framework and Application to Drug Repositioning

疾患特異的転写発現変動に基づく疾患類似性解析:自動化フレームワークの構築と
ドラッグリポジショニングへの応用

北海道大学大学院情報科学院
情報科学専攻 生体情報工学コース
情報生物学研究室

博士 3 年

CONG YI (ソウイ)

1. Introduction.....	5
2. Materials	12
2.1 Overview of Transcriptomic Datasets.....	12
2.1.1 Multi-disease bulk transcriptomic datasets	12
2.1.2 Curated public single-cell transcriptomic datasets	13
2.2 Definition of Disease and Control Samples	14
3. Methods	15
3.1 Conceptual Framework for Disease Similarity-Based DR	15
3.2 Overview of the disint Pipeline.....	16
3.3 Normalization with Housekeeping Genes.....	18
3.4 Disease-Specific log ₂ FC Computation.....	19
3.5 Embedding and Hyperparameter Optimization	20
3.6 Determination of the Number of Clusters.....	21
3.7 K-Means Clustering and Output Generation.....	22
3.8 Implementation and Reproducibility	22
3.9 Prototype Drug Repositioning Workflow (Demonstration)	22
4. Results.....	24
4.1 Overview of datasets used for evaluation	24

4.2 Normalization Effects and Batch Alignment	25
4.3 Disease-Specific Expression Shifts.....	27
4.4 Embedding and Clustering of Disease Profiles.....	29
4.4.1 Multi-Disease Transcriptomic Dataset (Bulk RNA-seq)	30
4.4.2 Single-Cell Dataset from Industrial Collaboration.....	34
4.5 Selection and in-depth analysis of the optimal disease similarity structure.....	36
4.5.1 Quantitative criteria for selecting the optimal configuration	36
4.5.2 Visualization and cluster characteristics of the selected embedding.....	38
4.5.3 Evaluation of tissue-of-origin effects (UMAP colored by tissue).....	41
4.5.4 High-dimensional validation of cluster coherence.....	43
4.6 Application to Drug Repositioning	46
5. Discussion.....	51
5.1 Limitations of classical batch-correction approaches under transcript-level quantification	51
5.2 Effectiveness and limitations of housekeeping-gene-based normalization.....	53
5.3 Interaction between normalization strategy and distance metrics.....	54
5.4 Internal clustering evaluation and practical considerations	55
5.5 Implications for downstream Drug Repositioning analysis	55
5.6 Biological interpretation of partially coherent clusters.....	58

5.7 Broader implications and future directions	62
Data and Code Availability	63
Acknowledgements.....	63
Funding statement.....	64
References.....	65
Appendix A.....	70
A.1 Spearman Correlation Validation Script.....	70
A.2 Shared DE Direction Validation Script	74
A.3 DrugBank Target Extraction and Signaling Module Annotation Script.....	76
Supplementary Figures	81
Supplementary Tables	88

1. Introduction

Drug repositioning (DR) is a strategy for identifying new therapeutic indications for existing drugs whose safety and pharmacokinetic profiles have already been established in humans.

Compared with de novo drug development, which typically requires 6-9 years and costs 2-3 billion dollars, DR can substantially reduce development time and cost by enabling rapid transition to preclinical testing and clinical trials [1, 2]. As a result, DR has attracted increasing attention as an efficient approach for therapeutic innovation in both infectious diseases and complex chronic disorders. Indeed, approximately 30% of newly approved drugs and vaccines by the U.S. Food and Drug Administration have originated from repurposing strategies [3], highlighting its growing importance in modern drug development.

The success of DR relies heavily on the availability of systematic and reproducible analytical frameworks. However, the optimization of efficient DR pipelines remains challenging, particularly in the absence of complete drug-target interaction networks [4, 5]. Traditional discovery strategies based on serendipity or phenotypic screening are inherently limited in scalability and reproducibility, making them insufficient for comprehensive exploration of disease-drug relationships [6, 7]. Given that most diseases arise from complex interactions among multiple genetic variants [8], systematic computational approaches are required to enable large-scale and unbiased DR analyses.

With the rapid development of high-throughput transcriptomic technologies, large volumes of

gene expression data derived from microarray and RNA-sequencing experiments have become publicly available [1]. Signature-based DR approaches leverage disease-specific gene expression profiles to infer potential therapeutic compounds capable of reversing disease-associated transcriptional alterations [9]. These methods provide a practical means of uncovering previously unknown disease mechanisms and off-target drug effects by exploiting transcriptomic signatures extracted from public genomic repositories.

In parallel, advances in artificial intelligence (AI) and machine learning (ML) have further expanded the analytical capacity of computational DR. Unlike earlier paradigms such as rule-based artificial intelligence or expert systems, modern ML algorithms can infer complex patterns directly from large-scale biological data without explicit knowledge encoding [10]. The integration of ML with transcriptomic analyses has enabled the development of diverse data-driven repurposing strategies, including approaches that integrate drug properties, molecular targets, and induced gene expression profiles to predict therapeutic potential [11, 12]. Nevertheless, many of these methods remain drug-centric, rely on predefined drug sets, or lack the flexibility to explore disease-specific molecular heterogeneity.

More recent approaches have attempted to incorporate disease-specific molecular information into DR pipelines. For example, DRIAD integrates drug-induced transcriptomic perturbations with disease-associated molecular changes to evaluate the association between drug effects and disease

progression [13]. While such methods demonstrate improved interpretability and reduced overfitting, their applicability is often constrained by disease specificity, reliance on predefined drug lists, or limited generalizability across disease contexts.

A major technical challenge underlying these approaches is the high dimensionality and heterogeneity of transcriptomic data. Clustering methods such as K-means offer scalability but can suffer from poor performance in high-dimensional spaces, particularly when applied directly to gene expression data [14]. Dimensionality reduction is therefore essential for extracting latent disease structures. Although principal component analysis (PCA) is commonly used due to its simplicity, it is limited in capturing nonlinear relationships. Uniform manifold approximation and projection (UMAP) has emerged as a powerful alternative that better preserves local and global data structures in low-dimensional embeddings, making it suitable for large-scale transcriptomic analyses [14, 15]. However, UMAP alone does not define explicit cluster boundaries, necessitating complementary clustering and evaluation strategies.

Recent studies have emphasized the importance of reproducible transcriptomic data integration and clustering across heterogeneous datasets [14, 16, 17]. Differences in experimental platforms, normalization procedures, and parameter selection can introduce batch effects and obscure biologically meaningful disease relationships [18–20]. To address these issues, horizontal integration strategies and ratio-based normalization methods using reference genes have been proposed to

enhance robustness and comparability across studies [20]. Moreover, objective evaluation metrics such as Silhouette analysis provide a principled basis for determining cluster validity and stability.

In this study, we focus on the development of a reproducible and general-purpose computational pipeline for disease similarity analysis based on disease-specific transcriptomic alterations. We developed an open-source Python framework, *disint* (disease integration and clustering toolkit), which standardizes key analytical steps including housekeeping gene (HKG)-based normalization, disease-specific $\log_2$ fold-change ($\log_2\text{FC}$) computation, PCA-UMAP embedding with automated hyperparameter optimization, and K-means clustering with quantitative cluster quality assessment. By organizing diseases within a shared transcriptomic space, this framework enables systematic visualization and comparative analysis of disease relationships across heterogeneous datasets. Importantly, the resulting disease similarity structures also serve as a conceptual basis for downstream translational analyses, including transcriptome-guided DR, by highlighting groups of diseases that share common molecular abnormalities.

Parts of the results presented in this thesis have been previously published in two peer-reviewed research articles [17, 21]. This thesis builds upon these published works and provides a unified and extended presentation of the disease similarity analysis framework and its application to DR. Some figures and textual descriptions are reused from these publications with appropriate copyright permissions.

Positioning of the Present Study and Its Novelty

The novelty of this study lies in its analytical perspective, in which diseases are explicitly treated as the fundamental units of analysis, and disease–disease similarity is characterized at the molecular level through the integration of dimensionality reduction and clustering of disease-specific transcriptomic alterations.

To clarify the methodological positioning of this analytical framework, a systematic literature survey was conducted in PubMed using the keywords ("Uniform Manifold Approximation" OR "UMAP") AND "k-means". The search was performed on January 19, 2026, and retrieved 11 relevant studies published up to June 3, 2022, including one review article and the present study. These studies span a wide range of application domains and are summarized in **Table 1**.

Table 1. Studies applying dimensionality reduction and clustering prior to the initial publication of the present study (June 2022).

Study context	Example application	Studies
Viral genomics	SARS-CoV-2 mutation pattern analysis	Hozumi et al.[22]
Chemical space analysis	Odorant and compound fingerprint classification	Rugard et al.[23]
Mass spectrometry–based profiling	LC-MS and ion imaging data analysis	Tian et al.[24]; Zhang et al.[25]

Methodological benchmarking	Evaluation of clustering and dimensionality reduction methods	Dalmaijer et al.[26]
Immunological profiling	Immune landscape analysis (review)	Sharma et al.[27]
Variant analysis	Functional classification of genetic variants	Bennett et al.[28]
Cell-type analysis	Cell-type identification and prediction	Cheng et al.[29]
Single-cell analysis	Subgroup discovery within cell populations	Wang et al.[30]
Neuroimaging (fMRI)	UMAP used as a comparative baseline	Spencer and
dFC analysis	method	Goodfellow[31]
Disease-level transcriptomics	Cross-disease similarity analysis	Cong et al.[17]

All studies listed in this table were identified under the predefined search strategy described in the main text.

Abbreviations: LC-MS, liquid chromatography coupled with mass spectrometry; fMRI, functional magnetic resonance imaging; dFC, dynamic functional connectivity.

While these studies demonstrate the broad applicability of dimensionality reduction and clustering techniques, their primary analytical objectives focus on structure discovery within a single system, such as cells, molecular entities, or subpopulations within a specific disease context. A systematic framework that treats diseases themselves as analytical entities and enables transverse comparison across multiple diseases based on transcriptomic alterations had not been established.

In contrast, the present study performs disease-level dimensionality reduction and clustering based on disease-specific transcriptional changes, thereby enabling molecular-level similarity analysis

across multiple diseases. The initial outcome of this analytical concept was first reported in June 2022. Subsequently, from late 2024 onward, studies incorporating partially overlapping analytical ideas have increased rapidly. However, to the best of our knowledge, an automated and reproducible pipeline that implements disease-level similarity analysis through the integration of dimensionality reduction and clustering has not yet been reported.

Building upon this conceptual foundation, the present doctoral research establishes a standardized and reproducible analytical framework for disease-level similarity analysis, in which diseases are explicitly treated as the fundamental units of analysis. By integrating dimensionality reduction and clustering at the disease level, this framework enables systematic and scalable comparison of molecular alterations across heterogeneous disease contexts.

Furthermore, this study explores the applicability of disease similarity analysis as a foundation for transcriptome-guided DR. By extending the disease similarity framework, shared molecular signatures among related disease groups can be leveraged to identify candidate compounds with potential therapeutic relevance. At present, only a limited number of studies have attempted to integrate disease-level similarity analysis with DR within a unified and reproducible computational framework. Accordingly, this work is positioned as a methodological contribution that provides both a reproducible foundation for cross-disease transcriptomic analysis and a basis for downstream translational applications.

2. Materials

2.1 Overview of Transcriptomic Datasets

To evaluate the general applicability of the proposed framework, we analyzed multiple publicly available transcriptomic datasets encompassing diverse diseases, experimental platforms, and biological resolutions. All datasets used in this study were obtained from established public repositories and processed in accordance with standard community practices.

2.1.1 Multi-disease bulk transcriptomic datasets

The primary dataset collection consisted of bulk transcriptomic profiles from multiple disease conditions. In total, 28 independent datasets covering 34 diseases and 386 samples were analyzed, including 255 disease samples and 131 corresponding healthy controls. All bulk datasets were obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO). A complete list of accession numbers and sample identifiers is provided in the public code repository accompanying this dissertation.

Raw sequencing data were retrieved in SRA format using the SRA Toolkit and converted to FASTQ files. Transcript-level quantification was performed using Salmon with the *Homo sapiens* GRCh38 reference transcriptome (Ensembl release 110). The resulting transcript-level expression matrices were aggregated into a unified matrix comprising 194,182 transcript entries across all

samples. Subsequent normalization and transformation procedures were applied uniformly across datasets to ensure comparability.

These bulk transcriptomic datasets served as a representative public resource for disease-level transcriptomic analysis across heterogeneous conditions.

2.1.2 Curated public single-cell transcriptomic datasets

To examine the applicability of the framework across different biological resolutions, we additionally analyzed a curated collection of publicly available single-cell RNA sequencing (scRNA-seq) datasets. This dataset collection was provided through an industrial collaboration and was originally compiled from major public repositories, including NCBI GEO, ArrayExpress, the Human Cell Atlas Data Portal, and the Broad Institute Single Cell Portal.

The curated dataset comprised 8 integrated scRNA-seq datasets spanning 12 disease conditions and 10 major cell types, encompassing 15,844 genes across 8,064 samples. Due to data usage agreements, the specific dataset identifiers are not disclosed in this dissertation.

All datasets in this collection were preprocessed and curated prior to analysis. In this dataset, cell-type identity constituted the dominant source of biological variation, providing a complementary setting to bulk disease-level transcriptomic analysis.

2.2 Definition of Disease and Control Samples

For all transcriptomic datasets analyzed in this study, disease and control samples were defined according to the original experimental annotations provided by the respective data sources. Healthy control samples were included for each dataset and were matched to disease samples within the same experimental context whenever available.

Disease-specific transcriptomic alterations were quantified by comparing disease samples against their corresponding control samples within each dataset. This dataset-level definition ensured that disease-associated expression changes were evaluated independently of cross-study technical variation.

3. Methods

This section describes the design and implementation of the disint pipeline used in this study. The pipeline integrates multiple analytical steps commonly employed in transcriptomic data analysis, including normalization, dimensionality reduction, clustering, and evaluation, into a unified and reproducible workflow.

Each component of disint is implemented as a modular unit and can be executed independently or as part of the full pipeline, allowing flexible application to diverse transcriptomic datasets. Unless otherwise stated, all analyses were performed using consistent preprocessing strategies and parameter settings to ensure comparability across experiments.

3.1 Conceptual Framework for Disease Similarity-Based DR

The methodological framework presented in this dissertation builds upon and extends our previous work on transcriptome-based disease analysis and DR.

In our earlier study [17], we proposed a two-stage computational framework that combined disease-specific transcriptomic profiling with unsupervised structure discovery for DR analysis. That work demonstrated the feasibility of representing diseases as differential gene expression signatures and of exploiting transcriptomic similarity for systematic downstream analysis. All analytical procedures were explicitly defined and reproducible at the method level; however, the workflow was implemented as a sequence of individual analytical steps rather than as a unified and automated

pipeline.

During the subsequent doctoral research period, this conceptual framework was generalized and formalized into an integrated, reproducible, and automated computational framework for multi-study transcriptomic integration and disease similarity analysis. This effort resulted in the development of *disint*, an open-source Python framework designed to standardize and orchestrate key analytical stages, including normalization, disease-specific representation, dimensionality reduction, clustering, and evaluation.

Accordingly, the core contribution of this dissertation lies in the establishment of a pipeline-level computational abstraction that enables consistent execution, parameter control, and reproducibility across heterogeneous transcriptomic datasets. By integrating method-oriented analytical workflows into a unified computational framework, this work provides methodological support for robust disease similarity analysis and establishes a reliable foundation for downstream applications such as DR.

The following sections describe the design principles and implementation of the *disint* pipeline in detail, with a particular focus on its methodological components for transcriptomic integration, low-dimensional embedding, and unsupervised clustering.

3.2 Overview of the *disint* Pipeline

The *disint* framework was implemented as an open-source Python package to provide a

standardized and automated workflow for large-scale transcriptomic analyses (**Figure 1**). The pipeline integrates multiple analytical stages, including HKG-based normalization, disease-specific log₂FC computation, dimensionality reduction, clustering, and quantitative evaluation.

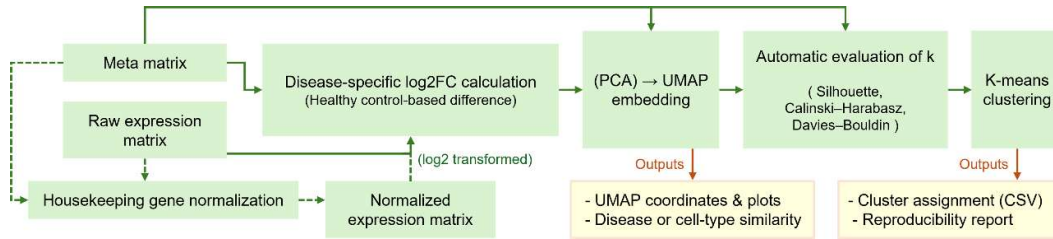


Figure 1. Overview of the disint framework for transcriptomic data integration and analysis

The schematic illustrates the modular workflow implemented in disint, including normalization, disease-specific log₂ fold-change computation, dimensionality reduction, and clustering. Solid arrows indicate the core analysis steps executed in the full pipeline, whereas dashed arrows denote optional or downstream procedures that can be enabled depending on the analysis objective.

Each analytical component is implemented as an independent yet interoperable module, allowing flexible execution either as part of the full pipeline or as standalone procedures. The input to disint consists of an expression matrix (genes × samples), accepting any count- or abundance-based values, together with an optional metadata table describing sample-level attributes such as disease, condition, or cell type. The pipeline outputs a comprehensive set of results, including normalized and log₂FC-transformed matrices, two-dimensional UMAP embeddings, automatically determined

optimal cluster numbers, cluster assignment tables, and evaluation visualizations. All outputs are saved in standard formats (CSV, JSON, PNG) to ensure reproducibility and downstream compatibility.

Two operational modes are supported. In the embedding-only mode, disint performs PCA-UMAP dimensionality reduction to visualize global similarity structures without subsequent clustering, which is particularly suitable for exploratory analyses. In the full-pipeline mode, embedding is followed by automatic cluster-number determination and K-means clustering, producing discrete cluster assignments and quantitative evaluation metrics. This dual-mode design allows disint to function both as a visualization tool and as a complete integration and clustering framework adaptable to diverse transcriptomic contexts.

3.3 Normalization with Housekeeping Genes

To minimize cross-dataset variability and establish a unified scale for expression quantification, disint adopts a HKG-based normalization strategy.

A panel of 20 candidate HKGs was predefined within the package based on the Housekeeping and Reference Transcript Atlas (HRT Atlas v1.0), which systematically identified stable human reference transcripts through large-scale RNA-seq meta-analysis across 52 tissues [32]. These genes exhibit consistent expression across tissues and experimental conditions, providing a robust baseline for cross-dataset normalization.

For each dataset group, `disint` computes three complementary stability indices for each candidate HKG: the maximum-to-minimum fold change (MFC), the median absolute deviation (MAD), and the coefficient of variation (CV). Optional trimming can be applied to exclude extreme sample values prior to index computation, reducing the influence of technical or biological outliers. Each index is normalized across genes and aggregated to generate a composite stability score reflecting overall expression consistency.

The three HKGs with the highest stability scores are automatically selected as reference genes. For each sample, the geometric mean of these reference genes is computed and used as an internal scaling factor. All expression values are rescaled so that the geometric mean of the selected HKGs matches a unified reference value shared across all dataset groups. This procedure equalizes baseline expression magnitudes while preserving relative gene-to-gene relationships and reducing inter-batch bias.

The resulting normalized expression matrix is used directly for subsequent analyses, including $\log_2\text{FC}$ computation, dimensionality reduction, and clustering.

3.4 Disease-Specific $\log_2\text{FC}$ Computation

Following normalization, `disint` computes per-sample disease-specific $\log_2\text{FC}$ profiles relative to matched healthy controls. This transformation converts absolute expression values into relative expression signatures that enable robust comparison across heterogeneous datasets.

For each disease D , all samples annotated with D in the metadata are identified. Each disease sample is associated with a subtype label stored in the metadata. Control samples are filtered to retain only those whose subtype tokens overlap with the disease samples, ensuring biologically appropriate matching. The mean expression of matched controls is calculated for each gene, and $\log_2\text{FC}$ values are computed as:

$$\log_2\text{FC}(g, s) = \log_2(x_{g,s} + \varepsilon) - \log_2(\bar{x}_{g,HC} + \varepsilon)$$

where $x_{g,s}$ denotes the normalized expression of gene g in sample s , $\bar{x}_{g,HC}$ denotes the mean expression in matched controls, and $\varepsilon = 10^{-5}$ is a pseudocount to ensure numerical stability.

The resulting $\log_2\text{FC}$ matrix captures disease-specific transcriptional shifts while minimizing inter-group bias. Each column represents an individual disease sample or cell-type-specific condition, enabling similarity estimation at both disease and subtype resolution.

3.5 Embedding and Hyperparameter Optimization

To identify intrinsic similarity structures among disease-level transcriptomic profiles, `disint` applies a two-step dimensionality reduction strategy consisting of PCA followed by UMAP. PCA is first used to reduce noise and computational complexity, retaining by default the top 50 components capturing the majority of variance. These components are subsequently embedded into two dimensions using UMAP.

Key UMAP hyperparameters, including the number of nearest neighbors, minimum inter-point

distance, and distance metric, are automatically optimized using Bayesian optimization implemented via Optuna [33]. The optimization process evaluates candidate embeddings using silhouette-based criteria and adapts to both supervised and unsupervised settings depending on the availability of sample-level biological labels.

The search space spans biologically relevant parameter ranges and multiple distance metrics (Euclidean, Manhattan, Canberra, Cosine). Independent optimization is performed for each metric, allowing complementary similarity structures to be evaluated rather than enforcing a single representation.

3.6 Determination of the Number of Clusters

Following UMAP embedding, `disint` determines the optimal number of clusters in a data-driven manner. A range of candidate cluster numbers is evaluated using three internal validation indices: the Silhouette coefficient [34], the Calinski-Harabasz index [35], and the Davies-Bouldin index [36].

When biological group labels are available, the search range for the number of clusters is defined based on the number of unique groups. In the absence of such labels, a predefined exploratory range is used. All validation indices are rescaled to a common range prior to integration, and a consensus criterion is applied to select the final cluster number.

3.7 K-Means Clustering and Output Generation

Using the selected cluster number, disint performs final clustering with the K-means algorithm implemented in scikit-learn, employing multiple initializations and a fixed random seed to ensure reproducibility. Cluster assignments are generated for each sample and visualized on the optimized embedding.

The pipeline outputs cluster membership tables, embedding coordinates, evaluation metrics, and structured metadata describing the analysis configuration. These outputs are stored in standard formats to facilitate reproducibility and integration with downstream analytical workflows.

3.8 Implementation and Reproducibility

The disint framework was implemented in Python (versions 3.9-3.11) and packaged using a standard setuptools-based structure. Core dependencies include NumPy, pandas, scikit-learn, umap-learn, matplotlib, Optuna, and related scientific libraries. All analyses reported in this dissertation were conducted using validated package versions to ensure reproducibility.

3.9 Prototype Drug Repositioning Workflow (Demonstration)

To demonstrate the downstream applicability of disint, we implemented a prototype DR analysis that builds upon the conceptual framework established in our previously published work [17].

Diseases assigned to the same cluster were analyzed to identify consensus transcriptional signatures,

which were subsequently queried against the L1000CDS² database to identify compounds predicted to reverse disease-associated gene expression patterns [37].

This demonstration illustrates how disint-based disease similarity analysis can serve as a systematic and reproducible foundation for translational extensions, including DR.

4. Results

This section demonstrates the performance and applicability of the disint framework.

We first assessed the reproducibility, cluster quality, and metric-dependent behavior of embeddings across multiple disease-level transcriptomic datasets. We then applied the same optimized configuration to an independent single-cell dataset collection to evaluate cross-platform generalizability. Finally, we illustrate how the pipeline outputs can be extended toward biological interpretation and prototype DR analysis.

4.1 Overview of datasets used for evaluation

The performance and versatility of the proposed framework were evaluated using two transcriptomic dataset collections described in Sections 2.1.1-2.1.2, representing distinct experimental modalities and biological resolutions.

First, a multi-disease bulk transcriptomic dataset was used to assess normalization behavior, embedding reproducibility, and clustering performance at the disease level. Second, a curated collection of public single-cell RNA-seq datasets obtained through industrial collaboration was analyzed to examine cross-platform and cross-resolution generalizability.

Unless otherwise noted, all results reported in this chapter are based on these two dataset collections, with disease and control samples defined as described in Section 2.2.

4.2 Normalization Effects and Batch Alignment

We first evaluated the impact of the HKG-based normalization strategy implemented in `disint` (Section 3.3) using the multi-disease bulk transcriptomic datasets. For each dataset, candidate housekeeping transcripts were ranked by cross-dataset stability, and the top-ranked genes were used to derive an internal scaling reference for sample-wise normalization.

To visualize the impact of this normalization step, we compared gene-level expression patterns before and after normalization using the 50 most variable transcripts across all samples (**Figure 2**). In the raw expression matrix, some samples exhibited extremely high or low overall intensity, producing vertical stripe-like patterns with a compressed dynamic range for subsets of genes. After HKG-based normalization, these global intensity differences were substantially reduced: highly saturated columns were compressed, while suppressed columns were lifted toward a shared dynamic range. Importantly, this adjustment equalized baseline expression magnitudes without forcing all samples to become identical.

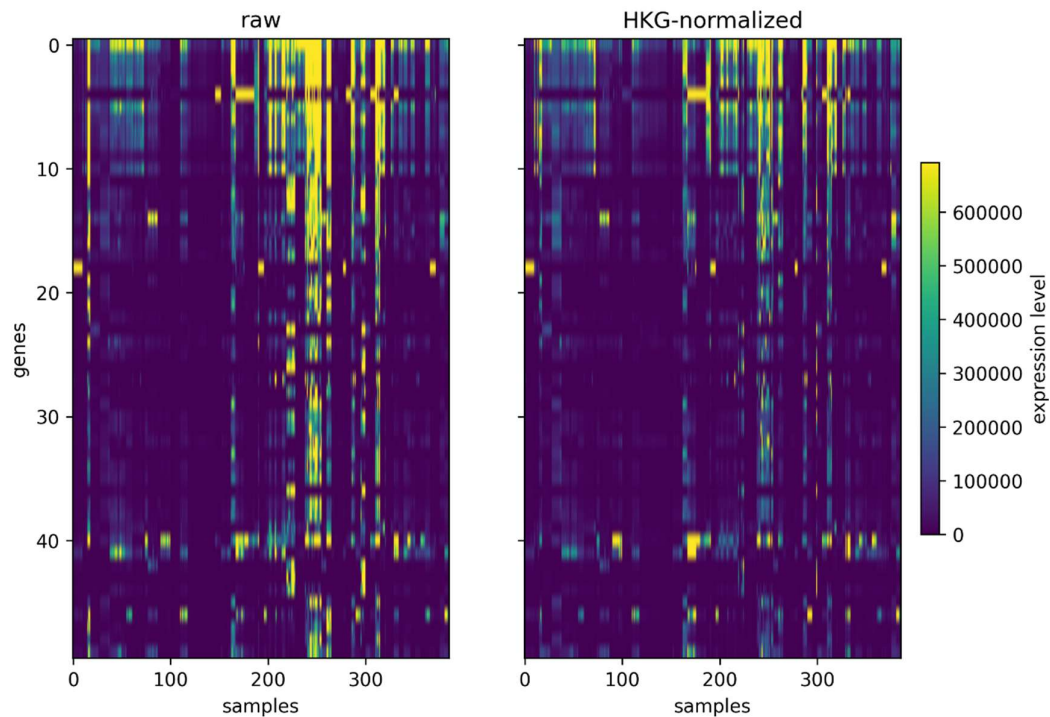


Figure 2. Effect of HKG-based normalization on gene-level expression patterns.

Heatmaps showing gene-level expression profiles before (left) and after (right) HKG-based normalization, visualized using the 50 most variable transcripts across all samples. Normalization reduces global intensity differences between samples while preserving relative expression patterns.

This normalization effect was not merely cosmetic. When propagated into downstream analyses, two improvements became apparent. First, disease-specific $\log_2\text{FC}$ profiles computed relative to matched healthy controls (Section 3.4) became more comparable across datasets, reducing spurious disease signals driven by dataset-specific baseline shifts. Second, subsequent PCA-UMAP embeddings and K-means clustering (Sections 3.5-3.7) preserved disease-level structure while

reducing separation attributable purely to technical batch effects.

Together, these results indicate that HKG-based scaling improves cross-dataset alignment while retaining biologically meaningful disease-specific variation. Rather than aggressively eliminating variation through batch correction, disint adopts a lightweight, sample-wise normalization strategy that enhances comparability across studies without erasing signals later exploited for embedding, cluster definition, and downstream interpretation.

4.3 Disease-Specific Expression Shifts

To quantitatively assess disease-associated transcriptomic alterations, we computed disease-specific $\log_2\text{FC}$ profiles by comparing each disease group with its matched healthy controls within the same dataset. $\log_2\text{FC}$ captures both the direction and magnitude of transcriptional dysregulation and constitutes a core feature for subsequent disease similarity analysis and downstream DR applications.

Because the interpretability of $\log_2\text{FC}$ patterns is sensitive to baseline expression intensity, $\log_2\text{FC}$ matrices were computed both before and after HKG-based normalization to evaluate how normalization influences the detectability of disease-specific signals. The complete $\log_2\text{FC}$ matrices computed before and after normalization are provided in the public code repository associated with this dissertation.

To visualize the global impact of normalization, we constructed group-mean $\log_2\text{FC}$ matrices for

all disease groups and summarized them as a heatmap (**Figure 3**). For each dataset group, $\log_2FC$ values were averaged across samples and displayed for the top 50 genes showing the largest absolute variation. Prior to normalization, many genes exhibited $\log_2FC$ values close to zero across multiple disease groups, indicating that disease-associated transcriptional changes were difficult to discern due to dataset-specific baseline and amplitude differences. After HKG-based normalization, $\log_2FC$ distributions became more balanced across groups, and clearer positive and negative deviations emerged, reflecting enhanced detectability of disease-related up- and down-regulation patterns.

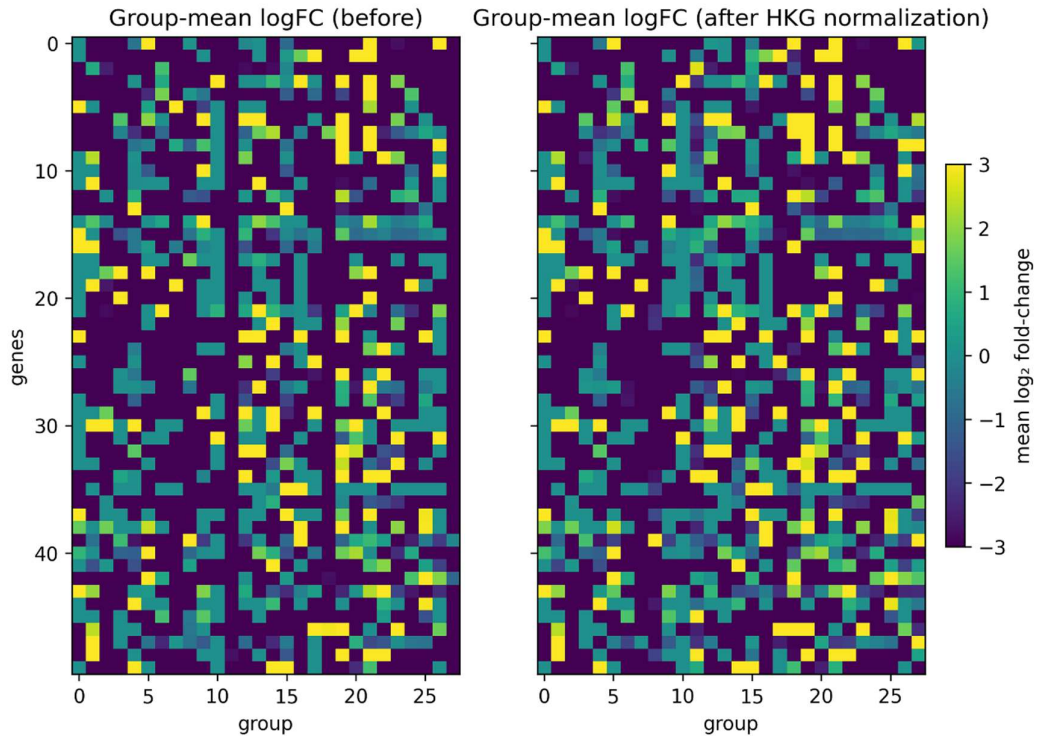


Figure 3. Group-mean disease-specific $\log_2FC$ profiles before and after normalization.

Heatmap representation of group-mean $\log_2FC$ values computed for each disease group relative to matched healthy controls. The left panel shows $\log_2FC$ profiles derived from raw expression values, whereas the right panel shows

log₂FC profiles after HKG-based normalization. Normalization enhances the contrast of disease-associated

transcriptional alterations by reducing baseline intensity bias across datasets.

Importantly, this change reflects an improvement in the interpretability of disease-specific transcriptional shifts rather than artificial signal amplification. By reducing baseline intensity bias and aligning expression magnitudes across datasets, HKG-based scaling increased the dynamic range of log₂FC values while preserving relative gene-to-gene expression relationships. As a result, disease-associated transcriptional trends that were previously obscured became consistently observable across all dataset groups.

4.4 Embedding and Clustering of Disease Profiles

This section evaluates the dimensionality-reduction and clustering performance of disint using the modules described in Sections 3.5-3.7. Depending on the analysis objective, these modules can be applied either independently—for example, embedding alone for exploratory visualization—or in an integrated full-pipeline mode that sequentially performs PCA-UMAP optimization, and clustering.

The framework supports both supervised and unsupervised configurations, allowing flexible adaptation to diverse transcriptomic data structures, including bulk, pseudo-bulk, and single-cell datasets. All outputs, including optimized UMAP coordinates, cluster assignments, k-selection profiles, and visualizations, are generated automatically in a standardized format to ensure

reproducibility and consistency across analyses.

4.4.1 Multi-Disease Transcriptomic Dataset (Bulk RNA-seq)

We first applied disint to the multi-disease bulk transcriptomic dataset described in Section 2.1.1 to assess how HKG-based normalization influences disease-level embeddings and clustering. Both pre- and post-normalized log₂FC matrices were analyzed to enable direct comparison. Each matrix (genes × samples) was transposed into a sample-wise representation and processed through UMAP dimensionality reduction followed by automatic K-means clustering.

The optimized UMAP parameters and the corresponding selected numbers of clusters (k_{best}) for all four distance metrics—Euclidean, Manhattan, Canberra, and Cosine—are summarized in **Table 2**. Representative results obtained using the Manhattan distance metric are shown in **Figure 4**. Results for the remaining distance metrics are provided in **Supplementary Figures S1-3**.

Table 2. Optimized UMAP parameters and selected numbers of clusters before and after normalization across four distance metrics.

after_norm				before_norm			
metric	n_neighbors	min_dist	k _{best}	metric	n_neighbors	min_dist	k _{best}
euclidean	10	0.0135235	26	euclidean	9	0.0713402	28
manhattan	9	0.1046117	18	manhattan	10	0.1216534	27
canberra	13	0.1309211	25	canberra	6	0.2072781	21

cosine	12	0.0134146	32	cosine	98	0.0290928	25
--------	----	-----------	----	--------	----	-----------	----

Summary of UMAP hyperparameters optimized by Optuna and the corresponding selected numbers of clusters (k_{best})

for four distance metrics (Euclidean, Manhattan, Canberra, and Cosine). Results are shown separately for embeddings

derived from HKG-normalized and unnormalized log₂FC matrices.

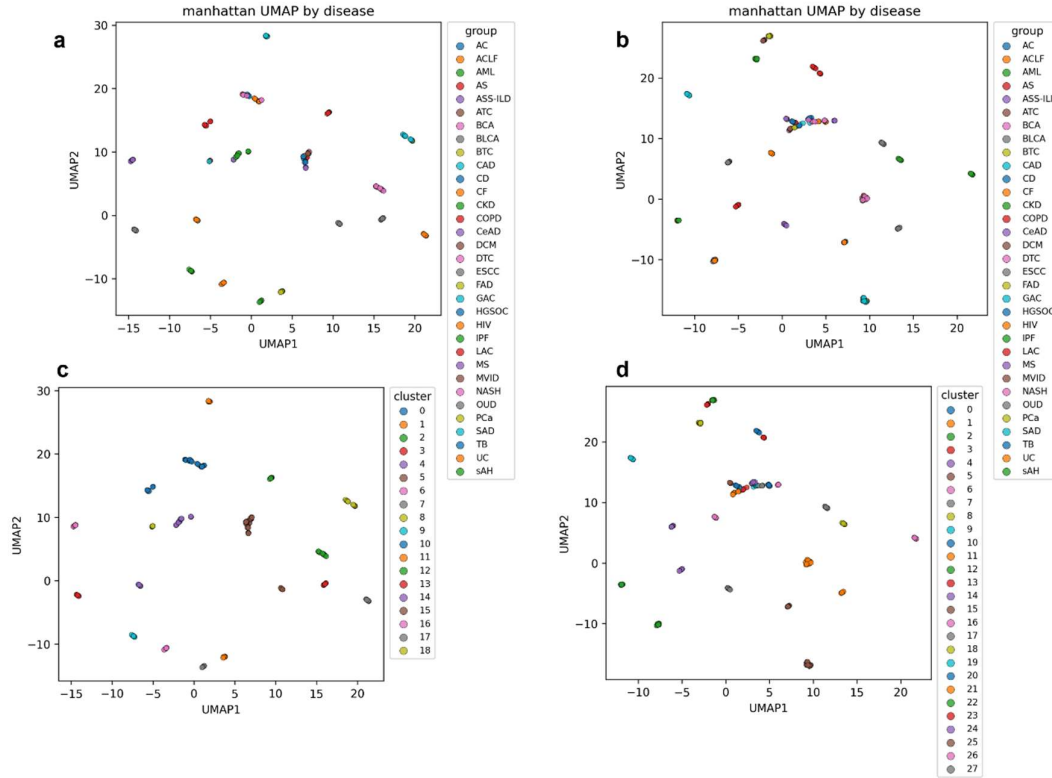


Figure 4. Effect of HKG-based normalization on disease-level embeddings and clustering (Manhattan distance).

Comparison of UMAP embeddings and K-means clustering results computed using the Manhattan distance metric before and after HKG-based normalization. Panels (a) and (b) show UMAP embeddings derived from normalized and raw log₂FC matrices, respectively, while panels (c) and (d) show the corresponding K-means clustering results. After normalization, diseases sharing similar transcriptional perturbation patterns form more compact and well-separated clusters, whereas raw embeddings exhibit blurred boundaries and partial overlap. These results indicate that HKG-

based scaling enhances biological separability by reducing inter-cohort amplitude bias.

In the normalized embedding (**Figure 4a**), diseases exhibiting similar transcriptional perturbation patterns formed compact and well-separated clusters. In contrast, the raw embedding (**Figure 4b**) displayed blurred boundaries and partial overlap among disease groups. These trends were consistently reflected in the clustering results (Figures 4c and 4d), where samples with highly similar log₂FC profiles were assigned to the same clusters, indicating shared molecular alteration patterns.

Selection of the optimal number of clusters was driven primarily by the Silhouette coefficient, which exhibited the most stable behavior across distance metrics after normalization. For the Manhattan metric, representative *k*-selection profiles are shown in **Figure 5**, with all evaluation indices min-max normalized to the [0, 1] range for comparability. Corresponding results for the other metrics are shown in **Supplementary Figures S4-6**. Complete cluster assignments and UMAP coordinates are provided in **Supplementary Tables S1 and S2**.

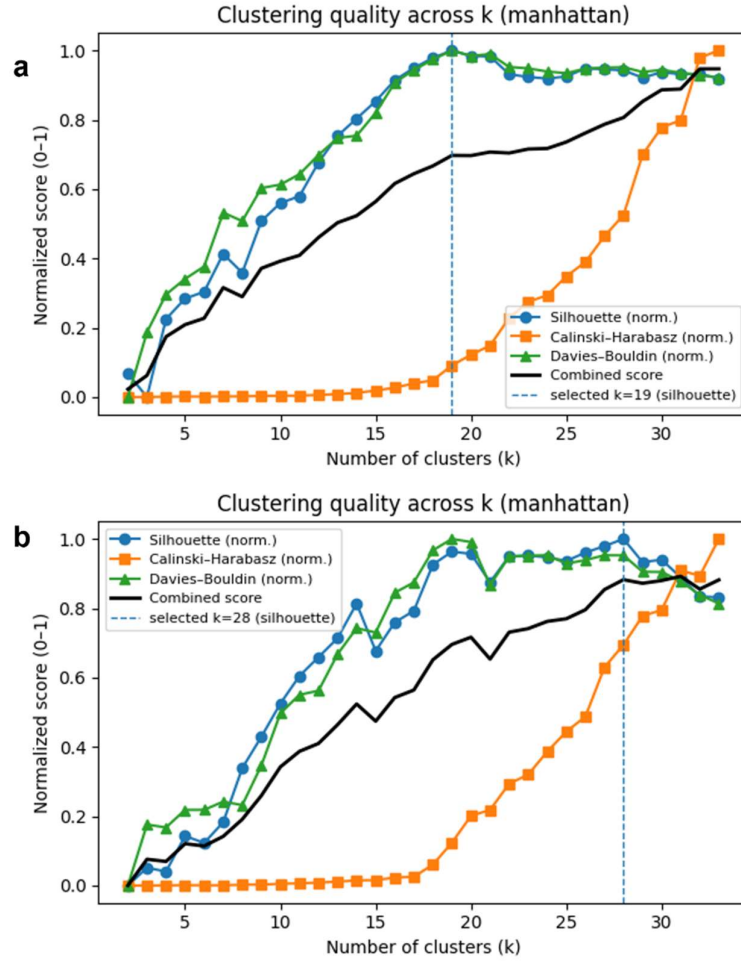


Figure 5. Determination of the optimal number of clusters (k) before and after HKG-based normalization using the Manhattan distance metric.

Selection curves for the optimal number of clusters (k) based on internal validation indices using the Manhattan distance metric. Panels (a) and (b) correspond to the HKG-normalized and raw $\log_2FC$ matrices, respectively. The Silhouette coefficient, CH index, and DB index were each min-max normalized to the range $[0, 1]$ to allow direct comparison. In this dataset, the CH index exhibited unstable behavior across k values; therefore, the Silhouette criterion was adopted as the primary decision rule. The vertical dashed line indicates the final number of clusters (k_{best}) selected by disint.

Asthma (AS) served as a representative case for assessing inter-dataset harmonization. AS samples originated from two independent datasets and appeared as separate clusters in the raw embedding, reflecting dataset-specific effects. After HKG-based normalization, these samples converged into a single coherent cluster under both the Euclidean and Manhattan metrics. This convergence indicates that HKG-based scaling effectively mitigates batch-level discrepancies while preserving disease-specific transcriptional structure.

4.4.2 Single-Cell Dataset from Industrial Collaboration

To further evaluate the generalizability of disint across data modalities, we applied the framework to an independent single-cell transcriptomic dataset provided through industrial collaboration. In this analysis, only the embedding modules were used, as the objective was to assess manifold structure rather than clustering performance.

UMAP embeddings were generated under four distance metrics (Euclidean, Manhattan, Canberra, and Cosine), with hyperparameters optimized using the same Optuna-based strategy described in Section 3.5. For visualization, embeddings were annotated by both disease condition and cell type.

Representative results for the Canberra and Manhattan metrics are shown in **Figure 6**, while embeddings based on Euclidean and Cosine distances are provided in **Supplementary Figure S7**. In disease-labeled embeddings, prefix numbers denote dataset identifiers and suffixes correspond to individual patients.

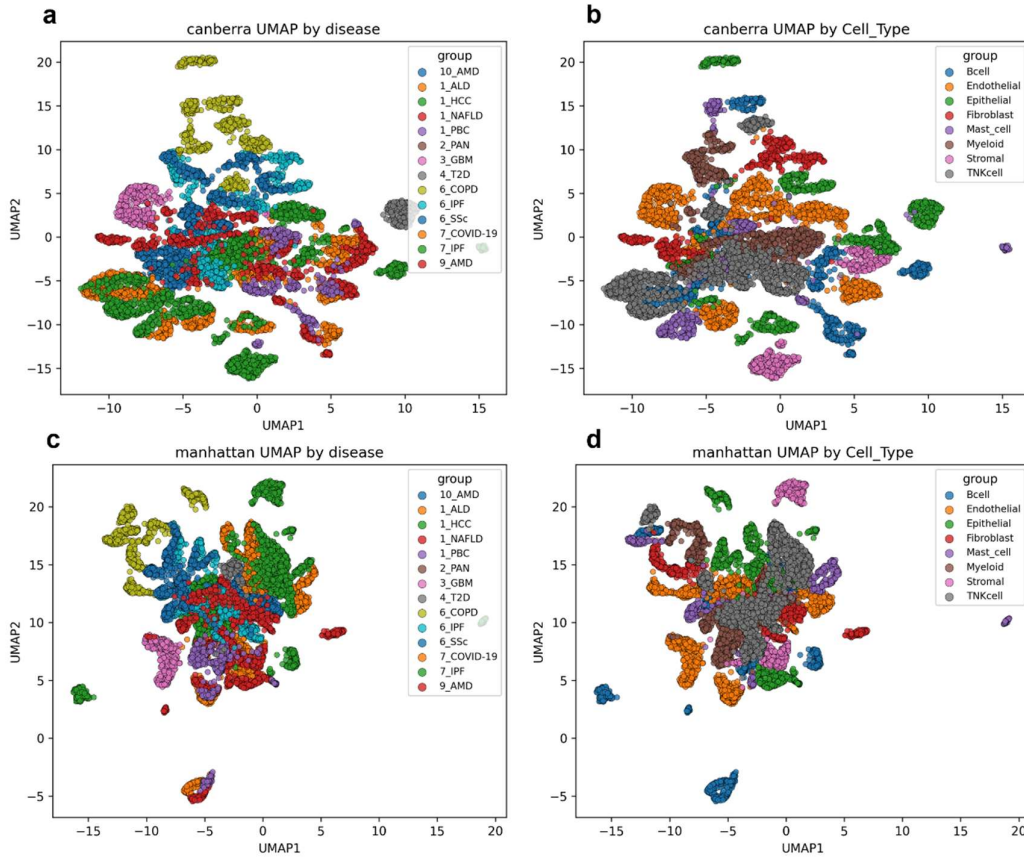


Figure 6. UMAP embeddings of the industrial single-cell cohort under different distance metrics and annotation schemes.

Panels (a) and (c) are colored by disease condition, and panels (b) and (d) by cell type. Prefix numbers in disease labels indicate dataset identifiers, while suffix numbers correspond to individual patients with different severity levels.

Diseases occupying multiple spatial regions typically reflected differences in dominant cell-type composition, whereas cell-type annotations formed distinct clusters within disease contexts. This reciprocal structure indicates that disint captures both inter-disease and inter-cell-type variation in a

coherent manner. Consistent embedding behavior across all four distance metrics highlights the robustness and versatility of the framework for large-scale, multi-label transcriptomic datasets.

4.5 Selection and in-depth analysis of the optimal disease similarity structure

Based on the systematic evaluation presented in Section 4.4.1, we selected the optimal embedding and clustering configuration according to predefined quantitative criteria.

The following analyses focus on an in-depth biological interpretation of this selected disease similarity structure, which also serves as the basis for the downstream DR analysis presented in Section 4.6.

4.5.1 Quantitative criteria for selecting the optimal configuration

To identify a representative disease similarity structure for in-depth biological interpretation, we first compiled all clustering results obtained in Section 4.4.1 into two comprehensive summary tables, corresponding to analyses performed before and after HKG normalization (**Supplementary Tables S3 and S4**). Each table summarizes the cluster assignments and clustering-quality characteristics for all diseases under the four evaluated distance metrics.

Based on these compiled results, we then quantitatively evaluated disease-level clustering coherence using a composite scoring scheme, which enabled systematic comparison across distance metrics and normalization conditions. For each distance metric, coherence scores were computed for all disease groups and averaged to yield an overall performance score, which served as the primary

criterion for selecting the optimal configuration.

For a given disease, let $f_{\max}$ denote the maximum fraction of its samples assigned to a single cluster, and let k denote the number of clusters to which samples from that disease are distributed.

To penalize fragmentation across multiple clusters, a fragmentation term was defined as $f_{\text{frag}} = \frac{1}{k}$,

such that diseases concentrated within fewer clusters receive higher scores.

In addition, we considered the internal composition of the dominant cluster by defining $m_{\max}$ as the maximum number of distinct diseases co-occurring within that cluster. Based on this measure, two bounded bonus terms were introduced to capture biologically interpretable clustering patterns. A consistency bonus of 0.15 was assigned when a disease was split across multiple clusters but remained consistently associated with a small and compositionally stable set of diseases ($k > 1$ and $m_{\max} \leq 2$). Conversely, a complete-retention bonus of 0.10 was assigned when all samples of a disease were retained within a single cluster ($k = 1$ and $f_{\max} = 1$).

The final disease-level coherence score was computed as

$$S = 0.7 f_{\max} + 0.2 f_{\text{frag}} + b_{\text{consistency}} + b_{\text{retention}},$$

where the weighting scheme emphasizes dominant-cluster retention while moderately penalizing excessive fragmentation, and selectively rewards structured multi-cluster patterns or complete sample retention.

For each distance metric, coherence scores were computed for all disease groups and averaged to

obtain a single global clustering score under each normalization condition. The resulting average scores are summarized in **Table 3**, which enables direct comparison of clustering performance across distance metrics. These average scores were used as the primary criterion for selecting the optimal embedding and clustering configuration for subsequent analyses.

Table 3. Average disease-level coherence scores for each distance metric before and after HKG normalization.

Distance metric	Before HKG	After HKG
Euclidean	0.830025	0.918945946
Manhattan	0.681904762	0.852055556
Canberra	0.751794872	0.745
Cosine	0.845555556	0.787756757

4.5.2 Visualization and cluster characteristics of the selected embedding

Based on the quantitative evaluation described above, we focused subsequent analyses on the embedding obtained using Euclidean distance after HKG normalization. This configuration achieved the highest average coherence score and was therefore selected as the representative disease similarity structure.

Figure 7 presents the global PCA-UMAP embedding derived from the disease-level $\log_2FC$ matrix using Euclidean distance after HKG normalization, together with magnified views of representative local neighborhoods. In this embedding, each point corresponds to a single disease,

and distances between points reflect the similarity of disease-induced transcriptomic alteration patterns rather than absolute expression levels. Consequently, spatial proximity in this space indicates shared molecular perturbations across diseases.

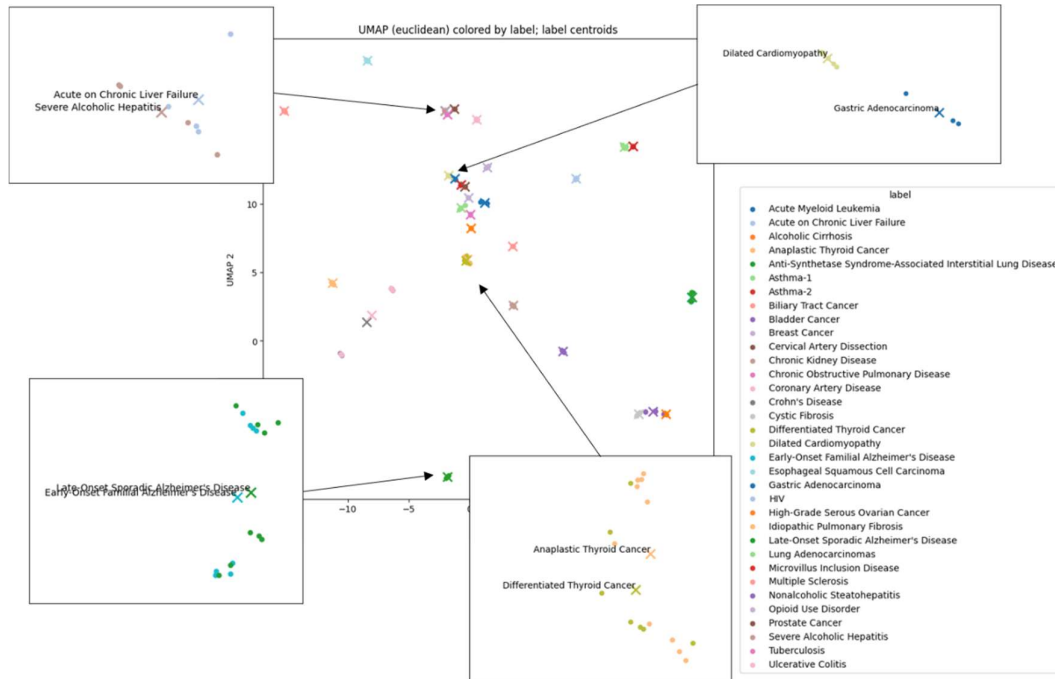


Figure 7. Disease-level PCA-UMAP embedding based on transcriptomic alteration profiles

Global PCA-UMAP embedding of diseases based on disease-specific $\log_2FC$ profiles after HKG normalization using Euclidean distance. Each point represents a single disease, positioned according to similarity in transcriptomic alteration patterns. The central panel shows the complete two-dimensional embedding, while the surrounding panels present magnified views of representative local neighborhoods.

Inspection of the enlarged regions reveals several biologically interpretable disease groupings. For

example, anaplastic thyroid cancer (ATC) and differentiated thyroid cancer (DTC) are positioned in close proximity, consistent with their shared tissue origin and partially overlapping oncogenic transcriptional programs. Similarly, early-onset familial Alzheimer's disease (FAD) and late-onset sporadic Alzheimer's disease (SAD) form a compact local cluster, reflecting common neurodegenerative molecular signatures despite differences in genetic background and disease onset.

Additional examples include the close spatial association between acute on chronic liver failure (ACLF) and severe alcoholic hepatitis (sAH), as well as between dilated cardiomyopathy (DCM) and gastric adenocarcinoma (GAC). Importantly, these disease groupings are not imposed a priori but emerge naturally from similarity in disease-specific transcriptomic alteration patterns captured by the low-dimensional embedding.

To further formalize these spatial relationships, K-means clustering was applied to the UMAP coordinates using the previously determined optimal cluster number ($k = 29$). The resulting cluster assignments are shown in **Figure 8**. Diseases assigned to the same cluster share highly similar gene expression change profiles, as defined by their proximity in the embedded space. Importantly, the disease groupings highlighted in the PCA-UMAP visualization are consistently recovered by K-means clustering, indicating concordance between the continuous embedding and discrete cluster structure.

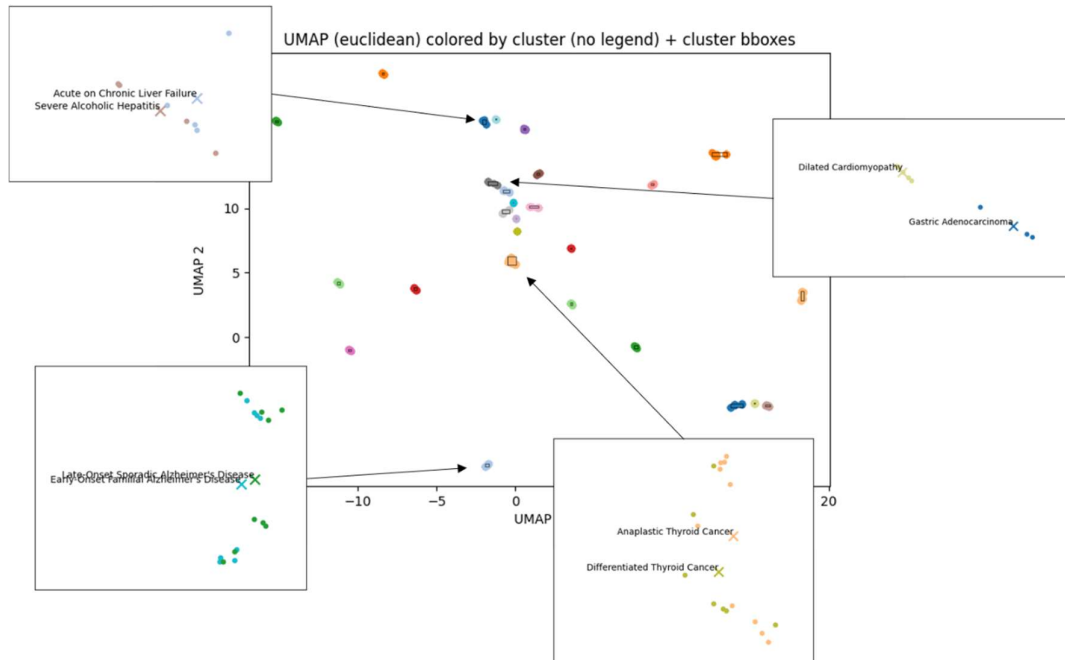


Figure 8. K-means clustering of diseases in the selected PCA-UMAP embedding

K-means clustering results applied to the PCA-UMAP embedding shown in **Figure 7** using the selected configuration (Euclidean distance after HKG normalization, $k = 29$). Diseases are projected onto the same two-dimensional UMAP coordinates, with colors indicating cluster assignments

Clusters containing multiple related diseases were subsequently treated as unified analytical units in downstream analyses. In particular, such disease clusters serve as the basis for transcriptome-guided DR, under the rationale that diseases sharing coherent molecular alteration patterns may also respond to overlapping therapeutic interventions.

4.5.3 Evaluation of tissue-of-origin effects (UMAP colored by tissue)

To examine whether the disease similarity structure identified in this study is primarily driven by

tissue-of-origin effects, the selected PCA-UMAP embedding was further visualized with diseases colored according to their tissue of origin (**Figure 9**).

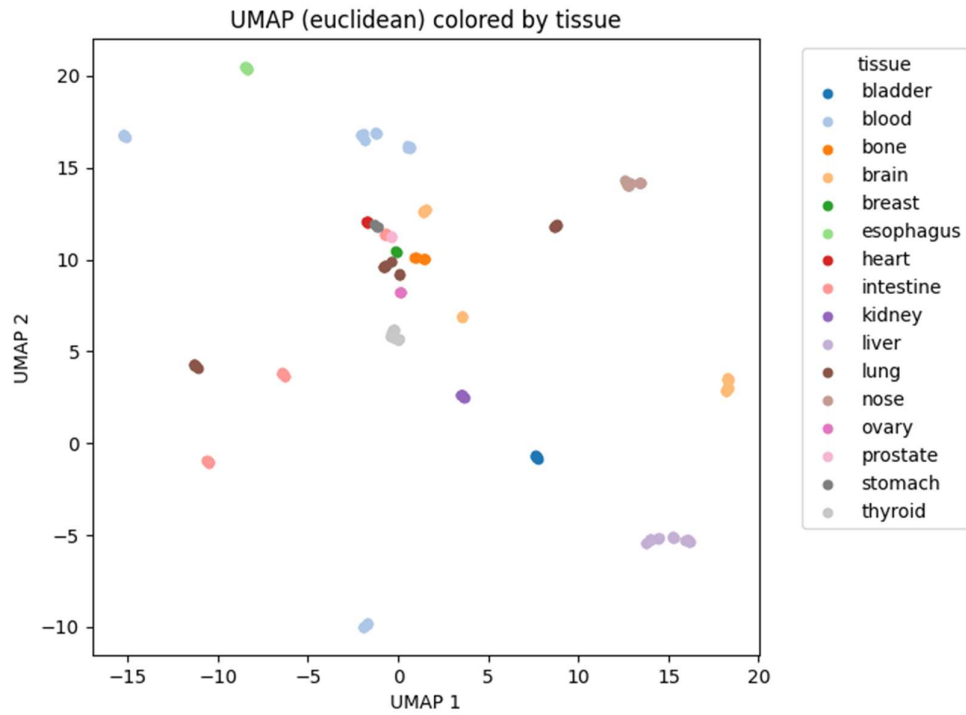


Figure 9. UMAP embedding of diseases colored by tissue of origin.

Diseases are projected onto the selected PCA-UMAP space (Euclidean distance after HKG normalization) and colored according to their tissue of origin.

As shown in **Figure 9**, diseases originating from the same tissue do not form perfectly segregated or isolated clusters. No extreme separation was observed in which all diseases from a given tissue collapsed into a single compact group. This observation indicates that the disease similarity structure

captured by the proposed framework cannot be explained solely by tissue-of-origin information.

Instead, the spatial proximity between diseases in the embedding space reflects similarities in disease-specific transcriptomic alterations.

At the same time, partial proximity among diseases derived from the same tissue was observed in several regions of the embedding. This tendency is likely attributable to the fact that the datasets analyzed in this study predominantly consist of samples derived from affected tissues, such that biologically related diseases may share both molecular characteristics and tissue context.

Taken together, these results suggest that the disease similarity structure obtained in this analysis is not determined exclusively by tissue-of-origin effects, while also not being entirely independent of tissue background. Rather, the embedding captures disease-level molecular similarities that extend beyond tissue identity, while remaining biologically consistent with underlying tissue-related influences.

4.5.4 High-dimensional validation of cluster coherence

Although UMAP preserves local neighborhood structure, it remains a nonlinear projection method. Therefore, to examine whether the observed low-dimensional proximity reflects genuine high-dimensional transcriptional similarity rather than projection artifacts, we conducted validation analyses directly in the original $\sim 194,182$ -dimensional $\log_2\text{FC}$ space.

(1) Rank-based similarity in the original $\log_2\text{FC}$ space

For each disease d , a disease-level mean $\log_2\text{FC}$ vector was computed by averaging all samples annotated with the same disease label:

$$\mathbf{v}_d = \frac{1}{m_d} \sum_{i=1}^{m_d} \mathbf{x}_i$$

where m_d is the number of samples belonging to disease d .

For each disease pair (d_i, d_j) , similarity was quantified using Spearman rank correlation:

$$\rho_{ij} = \text{Spearman}(\mathbf{v}_{d_i}, \mathbf{v}_{d_j})$$

Within-cluster similarities were computed for predefined cluster pairs (ATC-DTC, FAD-SAD, ACLF-sAH, DCM-GAC). Between-cluster similarity was defined as the mean correlation across all disease pairs that do not belong to the same predefined cluster. Random similarity was estimated by sampling 1000 random disease pairs.

Within-cluster mean Spearman similarities were 0.858 (ATC-DTC), 0.839 (FAD-SAD), 0.694 (ACLF-sAH), and 0.246 (DCM-GAC). In contrast, the mean between-cluster similarity was 0.068, and the mean random similarity was 0.071.

(2) Directional concordance of differentially expressed genes

To further evaluate biological consistency, we quantified directional concordance of differentially expressed genes using a predefined absolute $\log_2\text{FC}$ threshold of 0.5.

For a disease pair (d_i, d_j) , genes were selected if:

$$|v_{d_i, g}| > 0.5 \text{ or } |v_{d_j, g}| > 0.5$$

Let S denote this selected gene set. For each gene $g \in S$, directional concordance was defined as either: $(v_{d_i,g} > 0 \wedge v_{d_j,g} > 0)$ or $(v_{d_i,g} < 0 \wedge v_{d_j,g} < 0)$, The shared-direction ratio was computed as:

$$R_{ij} = \frac{N_{\text{up-up}} + N_{\text{down-down}}}{|S|}$$

where $N_{\text{up-up}}$ and $N_{\text{down-down}}$ denote the number of genes with concordant upregulation and concordant downregulation, respectively.

Within-cluster shared-direction ratios were 0.783 (ATC-DTC; 159,247 genes selected), 0.766 (FAD-SAD; 152,237 genes), 0.700 (ACLF-sAH; 115,297 genes), and 0.594 (DCM-GAC; 161,735 genes). In contrast, the mean between-cluster shared-direction ratio was 0.409, and the mean random shared-direction ratio was 0.408.

Taken together, both validation analyses consistently support the robustness of the UMAP-derived clustering. The three clusters ATC-DTC, FAD-SAD, and ACLK-sAH exhibited markedly elevated coherence in both metrics: rank-based similarity in the original $\log_2\text{FC}$ space and shared-direction ratios substantially exceeding background levels. In particular, their Spearman correlations ranged from 0.694 to 0.858, compared to a between-cluster mean of 0.068 and a random mean of 0.071. Similarly, their shared-direction ratios ranged from 0.700 to 0.783, clearly above the between-cluster and random baselines (~ 0.409).

Although the DCM-GAC pair showed comparatively lower coherence (Spearman correlation

0.246; shared-direction ratio 0.594), both metrics remained distinctly elevated relative to between-cluster and random controls. This indicates that the observed proximity in UMAP space does not arise from projection artifacts, but reflects genuine high-dimensional transcriptional similarity. The comparatively moderate similarity observed for DCM-GAC suggests a weaker but non-random form of shared transcriptional response, potentially associated with higher-order stress or inflammatory modules emerging during disease progression. The biological interpretation of this pattern is discussed in Section 5.5.

The Python scripts used for the high-dimensional validation analyses described above are provided in the Appendix of this dissertation for reproducibility.

4.6 Application to Drug Repositioning

As a proof-of-concept demonstration of downstream applicability, we applied disint-derived disease signatures to a transcriptome-guided DR analysis focusing on two liver-associated diseases: sAH and ACLF. These diseases were selected because they formed well-defined and distinct clusters in HKG-normalized embeddings, indicating coherent disease-level transcriptional signatures suitable for downstream exploratory analysis.

For each disease, differential expression analysis was performed between disease and matched healthy control samples using the limma framework. Significantly up- and down-regulated genes were visualized using volcano plots (**Figure 10**), highlighting characteristic expression patterns

associated with each condition.

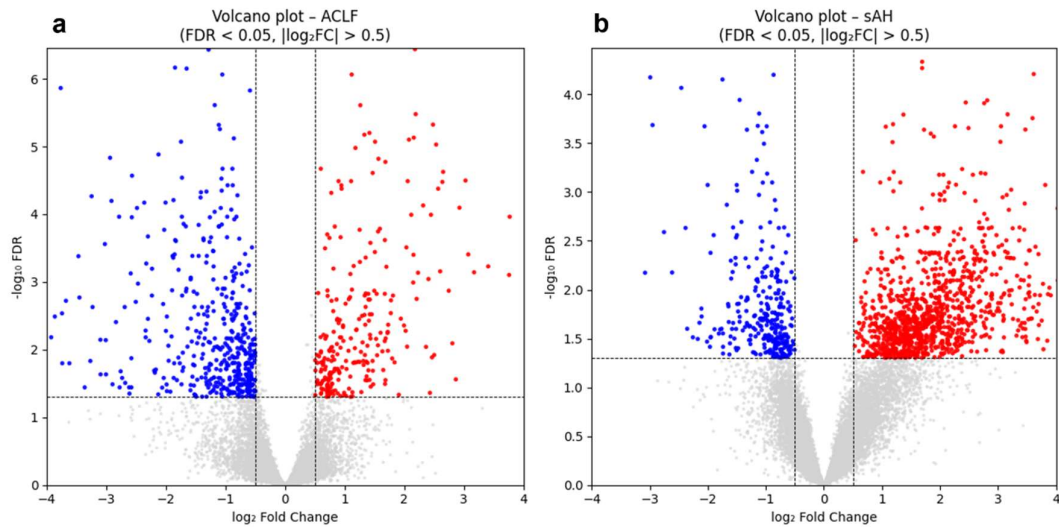


Figure 10. Volcano plots illustrating significantly up- and down-regulated genes identified in (a) ACLF and (b) sAH.

Each dot represents a transcript, with red and blue indicating up- and down-regulated genes, respectively, and gray indicating non-significant changes.

Using thresholds of $|\log_2\text{FC}| > 1$ and $\text{padj} < 0.05$, ACLF and sAH exhibited 169 and 1000 up-regulated genes and 247 and 136 down-regulated genes, respectively. Among these, 92 up-regulated and 51 down-regulated genes were shared between the two diseases, suggesting partially overlapping transcriptional perturbations.

The resulting disease-specific gene signatures were subsequently submitted to the L1000CDS² platform to identify compounds predicted to reverse observed expression change. The ranked

compound lists are summarized in **Table 4**. Several top-ranked compounds were associated with reported hepatoprotective or anti-inflammatory functions, suggesting qualitative consistency between the inferred transcriptional signatures and relevant biological contexts.

Table 4. Top-ranked compounds predicted by L1000CDS² to reverse the transcriptional signatures of ACLF and sAH.

ACLF				sAH			
Rank	score	Perturbation	Cell-line	Rank	score	Perturbation	Cell-line
1	0.0623	NNC 55-0396	HA1E	1	0.0238	trichostatin A	A549
2	0.0623	SB 218078	SKB	2	0.0238	curcumin	PC3
3	0.0584	RESERPINE	HA1E	3	0.0238	BRD-A36630025	PC3
4	0.0584	BRD-A34205397	SKB	4	0.0238	KUC104495	A375
5	0.0545	Rimcazole	HA1E	5	0.0221	trichostatin A	ASC
6	0.0545	AMSACRINE	HA1E	6	0.0221	trichostatin A	ASC
7	0.0545	PERHEXILINE	HA1E	7	0.0204	topotecan hcl	HT29
8	0.0545	BRD-K04853698	MCF7	8	0.0204	PF-431396	PC3
9	0.0545	NCGC00185684-	HT29	9	0.0187	trichostatin A	HA1E
10	0.0545	BRD-K10846167	A375	10	0.0187	vorinostat	HA1E
11	0.0545	KUC104495	A375	11	0.0187	trichostatin A	A375
12	0.0545	JNK-9L	A375	12	0.0187	cercosporin	A549
13	0.0545	NVP-TAE684	HA1E	13	0.0187	BRD-K15563106	A375
14	0.0545	BMS-387032	HA1E	14	0.0187	GF-109203X	HT29
15	0.0545	NVP-TAE684	HA1E	15	0.0187	YM-155	PC3
16	0.0506	Fluorofamide	HA1E	16	0.0187	MLN2238	SKLU1
17	0.0506	trichostatin A	A375	17	0.0187	BRD-K84203638	A549
18	0.0506	MK-2206	A375	18	0.0187	trichostatin A	PC3
19	0.0506	Ingenol 3		19	0.0187	F3055	A375
20	0.0506	SU11652	HT29	20	0.0187	trichostatin A	ASC
21	0.0506	trichostatin A	MCF7	21	0.0187	trichostatin A	ASC
22	0.0506	BRD-K04853698	VCAP	22	0.0187	Camptothecine (S	

23	0.0506	BRD-K37798499	A375	23	0.0187	I-BET151	HEPG2
24	0.0506	trichostatin A	PC3	24	0.0187	WZ-4-145	HME1
25	0.0506	BRD-A43155244	HA1E	25	0.0187	BI-2536	HS578T
26	0.0506	BRD-A80502530	A549	26	0.017	trichostatin A	HA1E
27	0.0506	BRD-A94756469	A549	27	0.017	TENIPOSIDE	HA1E
28	0.0506	trichostatin A	PC3	28	0.017	vorinostat	A549
29	0.0506	nilotinib	A375	29	0.017	MLN4924	A375
30	0.0506	LDN-193189	A549	30	0.017	528116.cdx	A549
31	0.0506	WZ-3105	A549	31	0.017	vorinostat	HA1E
32	0.0506	saracatinib	A375	32	0.017	TG101348	HA1E
33	0.0506	AZD4547	A375	33	0.017	HDAC6 inhibitor	HCT116
34	0.0506	PD-0325901	A375	34	0.017	vorinostat	VCAP
35	0.0467	Vinblastine sulfate	HA1E	35	0.017	BRD-K84203638	A375
36	0.0467	JAK3 Inhibitor VI	PC3	36	0.017	NCGC00167398-	A375
37	0.0467	wiskostatin	HA1E	37	0.017	trichostatin A	ASC
38	0.0467	U-46619	HA1E	38	0.017	HY-10518	MCF7
39	0.0467	NNC 26-9100	HA1E	39	0.017	S1003	PC3
40	0.0467	trichostatin A	HA1E	40	0.017	trichostatin A	ASC
41	0.0467	AMSACRINE	VCAP	41	0.017	trichostatin A	VCAP
42	0.0467	Etoposide	A375	42	0.017	PF-3758309	PC3
43	0.0467	16-	A375	43	0.017	GSK-690693	MDAM
44	0.0467	GW405833	A375	44	0.017	linifanib	A549
45	0.0467	curcumin	A375	45	0.017	entinostat	A375
46	0.0467	4-	A375	46	0.0153	JAK3 Inhibitor VI	HCC51
47	0.0467	BRD-K17140735	A375	47	0.0153	trichostatin A	PC3
48	0.0467	BRD-K17140735	A549	48	0.0153	Scriptaid	HA1E
49	0.0467	phorbol-12-	HA1E	49	0.0153	WAY 170523	HA1E
50	0.0467	trichostatin A	HA1E	50	0.0153	CCT 018159	HCC51

Compounds highlighted in orange are shared between ACLF and sAH.

Compounds shared between ACLF and sAH reflect similarities in transcriptional alterations

between the two diseases and illustrate how disint outputs can be used to explore cross-disease

patterns in compound prioritization. In contrast, disease-specific compounds may represent condition-dependent transcriptional responses and provide additional hypotheses for targeted DR.

Taken together, this analysis illustrates how disint-derived disease clusters can be coupled with external perturbation databases to generate reproducible, transcriptome-based hypotheses for DR.

The present analysis is intended as a demonstration of downstream applicability, and further biological and clinical validation is required to assess the therapeutic relevance of individual compounds.

5. Discussion

The results presented in this study highlight several methodological considerations that are particularly relevant for disease-level integration of heterogeneous transcriptomic datasets. Rather than aiming to maximize performance under idealized assumptions, the disint framework was designed to operate robustly under realistic data constraints commonly encountered in large-scale public repositories, where datasets differ substantially in experimental design, sequencing depth, and preprocessing history.

5.1 Limitations of classical batch-correction approaches under transcript-level quantification

In this study, we first assessed the removal of batch effects from a multi-disease, multi-dataset expression matrix using several representative integration approaches, including ComBat, BBKNN, and scVI. To systematically evaluate the raw matrix and each integration result, we applied the scIB framework proposed by Luecken et al., which combines batch-removal and biological-conservation metrics into a composite score [18]. None of these methods yielded satisfactory performance on the present dataset (**Table 5**).

Table 5. scIB-based evaluation of batch correction and integration methods

Method	batch score	bio score	final score
raw	0.592299	0.435621	0.498292

bbknn	0.597094	0.516994	0.549034
combat	0.604608	0.280187	0.409955
scvi	0.638721	0.493506	0.551592

$$\text{final_score} = 0.4 \times \text{batch_score} + 0.6 \times \text{bio_score}$$

A likely explanation lies in the upstream data-processing constraints. Several datasets did not provide true raw count matrices, and to minimize computational bias introduced by heterogeneous quantification pipelines, all samples were uniformly quantified using Salmon. Unlike genome-aligned count generation, Salmon performs quasi-mapping of sequencing reads directly to transcript sequences. As a consequence, the resulting NumReads matrix contains 194,182 transcript-level entries, a substantial proportion of which exhibit extremely low or zero counts because many annotated transcripts are unexpressed or only weakly expressed in the sampled tissues.

Although this uniform quantification strategy eliminates methodological discrepancies across datasets, it may also obscure dataset-specific technical structure that batch-correction tools rely on to estimate batch effects. Under such conditions, explicit batch modeling becomes unstable, and aggressive integration may distort genuine biological variation rather than remove technical noise. This observation suggests that the failure of classical integration pipelines in this setting reflects a mismatch between data characteristics and modeling assumptions, rather than shortcomings of the methods themselves.

5.2 Effectiveness and limitations of housekeeping-gene-based normalization

In contrast to explicit batch-correction approaches, normalization based on HKGs demonstrated markedly improved performance across the majority of datasets. As shown in Sections 4.2 and 4.3, HKG normalization enhanced the clarity of patient-control log₂FC patterns, and these improvements were preserved through downstream UMAP embedding and clustering.

By rescaling samples using a small set of highly stable reference genes, this approach reduces large-scale amplitude differences between datasets without enforcing a global transformation that homogenizes all samples. As a result, disease-specific transcriptional signatures become more comparable across studies while preserving relative gene-to-gene relationships. This lightweight normalization strategy proved particularly effective for disease-level integration, where global intensity shifts across cohorts can otherwise dominate similarity structure.

Nevertheless, HKG-based normalization is not universally optimal. Diseases characterized by global transcriptional dysregulation, metabolic reprogramming, or widespread cellular stress may partially violate the assumption of HKG stability. In such cases, ratio-based normalization using a small set of reference genes may attenuate disease-specific signals or introduce bias. Importantly, disint mitigates this risk by adopting a data-driven HKG selection procedure, prioritizing genes with the highest observed stability within each dataset rather than relying on a fixed reference list.

5.3 Interaction between normalization strategy and distance metrics

The observed interaction between normalization strategy and distance metric underscores the importance of considering preprocessing and similarity definition jointly. Using the quantitative disease-level clustering coherence framework described in Section 4.5.1, we observed that under HKG normalization, Euclidean and Manhattan distances exhibited the most substantial improvements in clustering coherence, reflecting their sensitivity to magnitude-based transcriptomic variation. Quantitatively, these metrics showed clear gains in per-disease clustering quality compared with unnormalized data.

In contrast, Canberra distance showed minimal change, likely because its intrinsic normalization already down weights large absolute differences. Cosine distance, which emphasizes vector orientation rather than magnitude, exhibited reduced coherence after normalization. Removing global intensity differences in this case may amplify subtle cohort-specific fluctuations, leading to fragmented cluster structures with limited biological interpretability.

These results suggest that magnitude-aware distance metrics such as Euclidean and Manhattan represent more robust default choices for cross-dataset disease-level transcriptomic analysis within this framework. More broadly, they highlight that preprocessing choices cannot be evaluated independently of downstream similarity definitions.

5.4 Internal clustering evaluation and practical considerations

To determine the optimal number of clusters, disint integrates three complementary internal validation indices: Silhouette, CH, and DB. While this multi-criteria strategy provides flexibility across diverse datasets, certain indices may become uninformative in practice. In the present study, the CH index increased monotonically with the number of clusters and did not provide meaningful guidance.

This behavior is consistent with datasets containing non-convex cluster structures, variable densities, or diseases composed of multiple molecular subtypes. In such settings, reliance on a single evaluation metric can lead to misleading conclusions. Allowing users to prioritize more robust criteria, such as the Silhouette coefficient, provides a practical compromise between automation and interpretability. Future extensions of disint may incorporate adaptive weighting schemes or alternative indices tailored to complex disease architectures.

5.5 Implications for downstream Drug Repositioning analysis

In the downstream DR analysis, disease clusters derived from disint were coupled with reverse-signature matching using the LINCS L1000 reference database. Diseases grouped within the same cluster exhibited highly similar log₂FC patterns, enabling the identification of compounds predicted to antagonize shared transcriptional perturbations. This cluster-level strategy provides a natural extension from disease similarity analysis to systematic therapeutic hypothesis generation, without

relying on disease-specific or manually curated drug sets.

At the current stage, these results should be interpreted as proof-of-concept rather than validated therapeutic predictions. The biological efficacy of individual compounds was not experimentally assessed, and no in vitro or in vivo validation was performed. Moreover, due to instability of the external L1000CDS² API at the time of analysis, the DR module was not incorporated into the core disint package and is instead provided as an optional post-processing workflow.

Nevertheless, to assess whether the extracted DR candidates reside within a realistic drug development space, we further examined their annotation status in DrugBank and associated clinical trial phases. Focusing on the ACLF and sAH disease cluster, a total of 21 known compounds (9 for ACLF and 12 for sAH) were identified as candidate drugs registered in DrugBank (**Table 6**).

Table 6. Clinical development status of DrugBank-annotated DR candidates for ACLF and sAH

	Compound name	DrugBank ID	Clinical Trials				
			Phase 0	Phase 1	Phase 2	Phase 3	Phase 4
ACLF	BRD-K37798499	DB00773	5	377	938	399	27
	nilotinib	DB04868	3	47	91	35	17
	PD-0325901	DB07101	0	22	8	0	0
	vorinostat	DB02546	1	184	143	7	0
sAH	MLN2238	DB09570	1	81	90	16	7
	F3055	DB03496	0	40	32	0	0
	TG101348	DB12500	0	17	9	3	0
	S1003 / linifanib	DB06080	0	13	11	1	0

Compound names correspond to perturbagen annotations provided by the LINCS L1000 dataset; Clinical trial phase information was retrieved from DrugBank (accessed January 19, 2026).

Importantly, a substantial fraction of these compounds have progressed to clinical evaluation, spanning Phase 1 to Phase 3 trials, with several agents reaching Phase 4. This observation indicates that the cluster-derived DR candidates are not dominated by experimental or implausible molecules, but instead include compounds whose safety profiles and pharmacological properties have already been investigated in human studies.

It should be emphasized that the presence of clinically evaluated compounds does not imply therapeutic efficacy for ACLF or sAH. However, the enrichment of clinically staged drugs among the DR candidates provides indirect support for the translational relevance of the proposed disease similarity-driven framework. Specifically, it suggests that transcriptome-based disease clustering can guide DR hypothesis generation toward regions of the drug development landscape that are consistent with existing pharmacological knowledge, rather than arbitrary or biologically disconnected chemical space.

Taken together, these findings position disint as a hypothesis-generating framework that bridges disease-level transcriptomic similarity analysis and downstream drug repositioning. While further experimental validation is required, the integration of disease clustering with reverse-signature matching offers a reproducible and conceptually coherent basis for prioritizing candidate therapeutics in a cross-disease context.

5.6 Biological interpretation of partially coherent clusters

Drawing upon the representative clustering configuration presented in Section 4.5 (HKG normalization with Euclidean distance), we further explored the biological interpretation of selected disease proximities. Several clusters correspond to intuitively related disease entities. For instance, ATC and DTC share a common tissue origin and partially overlapping oncogenic transcriptional programs. Likewise, FAD and SAD represent distinct genetic contexts of Alzheimer's disease but converge at the level of neurodegenerative molecular signatures. The ACLF-sAH pairing, while less immediately intuitive than the thyroid or Alzheimer's disease examples, nevertheless demonstrates recognizable biological commonality, as both conditions involve acute inflammatory injury superimposed on chronic liver dysfunction, accompanied by extensive immune activation and stress-related transcriptional responses.

In contrast, the DCM-GAC cluster represents a less intuitive association. Gene-level inspection did not reveal substantial overlap in reported disease-associated genes between DCM and GAC, and disease-specific Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps (hsa05414 for DCM and hsa05226 for GAC) do not indicate a shared primary etiological mechanism. However, analysis at the level of higher-order signaling modules suggested convergence in inflammatory and stress-response pathways, including NF- κ B-related signaling, cytokine-mediated signaling, PI3K-Akt survival signaling, and TGF- β -associated remodeling responses. These pathways are not

necessarily core etiological triggers in both diseases but may function as regulatory hubs during disease progression and cellular stress adaptation.

To further examine whether this convergence extends to therapeutically targetable mechanisms, we analyzed DrugBank-annotated mechanisms of action for candidate compounds associated with DCM and GAC. Detailed drug–target associations, including specific compound names and annotated protein targets, are summarized in **Table 7** and provided in full in **Supplementary Table S5**. Although the downstream drug repositioning analysis was conducted independently from the UMAP embedding, the target profiles provide an orthogonal perspective on shared regulatory axes.

Table 7. Summary of DrugBank target classes for DCM- and GAC-associated candidate compounds

candidate_source	drug_id	drug_name	n_rows	n_unique_ta	is_kinase	is_mapk_fa	is_map2k	is_map3k	is_pi3k	is_akt	is_cdk	is_jak_stat	is_tgf_beta	is_hdac
DCM_only	DB00694	Daunorubicin	3	2	0	0	0	0	0	0	0	0	0	0
DCM_only	DB08073	(2S)-1-(1H-INDOL-3-YL)-3-{[5-(3-METHYL-1H-INDAZOL-5-YL)PYRIDIN-3-YL]OXY}PR OPAN-2-AMINE	5	5	5	0	0	0	0	1	0	0	0	0
DCM_only	DB11805	Saracatinib	2	2	2	0	0	0	0	0	0	0	0	0
DCM_only	DB12686	PHA-793887	3	3	3	0	0	0	0	0	3	0	0	0

GAC_only	DB01254	Dasatinib	23	23	13	2	1	1	0	0	0	1	0	0
GAC_only	DB04865	Omacetaxine mepesuccinate	2	2	0	0	0	0	0	0	0	0	0	0
GAC_only	DB06061	AZD-8330	1	1	1	1	1	1	0	0	0	0	0	0
GAC_only	DB06999	PLX-4720	2	2	1	0	0	0	0	0	0	0	0	0
GAC_only	DB07101	Mirdametinib	2	2	2	2	2	0	0	0	0	0	0	0
GAC_only	DB07159	Tamatinib	2	2	2	0	0	0	0	0	0	0	0	0
GAC_only	DB08597	Dorsomorphin	5	5	3	0	0	0	0	0	0	0	1	0
GAC_only	DB11663	Pictilisib	1	1	1	0	0	0	1	0	0	0	0	0
GAC_only	DB11689	Selumetinib	2	2	2	2	2	0	0	0	0	0	0	0
GAC_only	DB12429	CI-1040	15	15	15	7	2	2	0	1	0	0	0	0
GAC_only	DB12847	Pyroxamide	2	1	0	0	0	0	0	0	0	0	0	2
GAC_only	DB12904	ZSTK-474	1	1	1	0	0	0	1	0	0	0	0	0
overlap	DB03496	Alvocidib	12	12	8	0	0	0	0	0	8	0	0	0
overlap	DB08142	AT-7519	2	2	2	0	0	0	0	0	2	0	0	0

n_rows indicates the number of target entries annotated in DrugBank for each compound.

n_unique_target_genes represents the number of unique target genes associated with the compound.

Binary class indicators (e.g., is_kinase, is_mapk_family, is_pi3k, is_cdk, etc.) represent the count of targets belonging to predefined mechanistic categories based on DrugBank target names (accessed February 12, 2026). These categories were defined by keyword-based classification of annotated target proteins and are intended to provide a coarse-grained overview of signaling module engagement rather than formal pathway enrichment.

Two compounds shared between DCM and GAC (Alvocidib and AT-7519) predominantly target cyclin-dependent kinases (CDKs), indicating convergence at the level of cell-cycle regulatory control. Disease-specific candidates further extended this pattern: DCM-associated compounds showed enrichment for CDK and AKT-related targets, whereas GAC-associated compounds prominently targeted MAPK cascade components, receptor tyrosine kinases, and PI3K signaling

elements. These kinase-centered profiles are consistent with the higher-order stress and survival signaling modules identified above.

Importantly, this pattern does not imply that DCM and GAC share a primary disease mechanism. Rather, it suggests that distinct pathological processes may converge on common regulatory circuits—particularly proliferation control and stress-adaptive signaling—during disease progression. Notably, these inflammatory and stress-response signaling modules often function as regulatory hubs in both disease progression and pharmacological modulation. Such convergence at the level of signaling hubs may provide a mechanistic basis for the moderate but reproducible high-dimensional similarity observed for the DCM-GAC pair relative to inter-cluster and random controls.

Taken together, these findings indicate that disease proximity in the embedding space does not necessarily reflect shared etiological origin. Instead, certain clusters may capture higher-order regulatory similarities, representing transcriptional modules recurrently activated across distinct pathological contexts and potentially susceptible to overlapping classes of therapeutic intervention. In this sense, potential drug repositioning across such clusters would not depend on shared primary etiology, but rather on shared regulatory vulnerabilities at the level of signaling architecture.

The Python scripts used for the high-dimensional validation and DrugBank target analyses are provided in Appendix A.3.

5.7 Broader implications and future directions

Taken together, this study demonstrates that robust disease-level integration of heterogeneous transcriptomic datasets can be achieved without explicit batch modeling, provided that normalization and similarity definitions are carefully aligned with data characteristics. By elevating method-oriented workflows to a pipeline-level abstraction, disint enables consistent execution, parameter control, and reproducibility across diverse datasets.

Future work will focus on extending the framework toward a fully integrated end-to-end pipeline that includes systematic validation of downstream applications such as DR. Incorporating additional normalization strategies, adaptive distance selection, and more robust evaluation criteria may further enhance applicability across broader disease contexts. Ultimately, such developments may support scalable, transcriptome-guided hypothesis generation for translational research.

Data and Code Availability

The full source code of the disint framework is openly available at:

<https://github.com/Soui816/disint/>

The package can be installed via pip: `pip install disint`

The repository provides the complete implementation, documentation, and example workflows required to reproduce all analyses presented in this dissertation. The software is released under the CC BY-NC-SA 4.0 license; commercial use requires explicit permission from the author.

Large supplementary data matrices, including the complete log₂FC matrices before and after HKG-based normalization and a full dataset manifest, are publicly available at:

<https://github.com/Soui816/disint-supplementary-data>

Drug target annotation was performed using DrugBank (version 5.1), accessed under an academic license. Due to licensing restrictions, the full DrugBank database is not redistributed. Researchers may obtain access directly from DrugBank (<https://go.drugbank.com/>) subject to their terms and conditions.

Acknowledgements

The author expresses her sincere gratitude to Professor Toshinori Endo for his supervision, insightful guidance, and continuous support throughout the doctoral program.

The author also sincerely thanks Associate Professor Naoki Osada, Associate Professor Shingo Matsumoto, and Professor Masaharu Yoshioka for their valuable comments, constructive suggestions, and thoughtful evaluation of this dissertation.

The author appreciates the members of the Information Biology Laboratory for maintaining a stimulating and supportive research environment.

The author thanks Professor Vural Özdemir for his editorial guidance and valuable academic advice during the publication process.

The author acknowledges NB Health Laboratory Co. Ltd. for providing access to curated single-cell datasets and for their support during this research collaboration.

The author also acknowledges DrugBank for providing access to its database under an academic license.

The author also sincerely appreciates the quiet support and understanding of her family and friends throughout this journey.

Finally, the author is deeply grateful for the gentle companionship of her rabbits, Gen and Muru, throughout the course of this study, and thanks Linlin for kindly taking care of them during this period.

Funding statement

This work was supported by JST SPRING, Grant Number JPMJSP2119.

References

1. Rapicavoli RV, Alaimo S, Ferro A, Pulvirenti A (2022) Computational Methods for Drug Repurposing. pp 119–141
2. Jourdan J-P, Bureau R, Rochais C, Dallemagne P (2020) Drug repositioning: a brief overview. *Journal of Pharmacy and Pharmacology* 72:. <https://doi.org/10.1111/jphp.13273>
3. Kwon O-S, Kim W, Cha H-J, Lee H (2019) In silico drug repositioning: from large-scale transcriptome data to therapeutics. *Arch Pharm Res* 42:879–889. <https://doi.org/10.1007/s12272-019-01176-3>
4. Jin G, Wong STC (2014) Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discov Today* 19:637–644. <https://doi.org/10.1016/j.drudis.2013.11.005>
5. Zeng X, Zhu S, Lu W, et al (2020) Target identification among known drugs by deep learning from heterogeneous networks. *Chem Sci* 11:. <https://doi.org/10.1039/C9SC04336E>
6. Ciallella JR, Reaume AG (2017) In vivo phenotypic screening: clinical proof of concept for a drug repositioning approach. *Drug Discov Today Technol* 23:45–52. <https://doi.org/10.1016/j.ddtec.2017.04.001>
7. Roundtable on Translating Genomic-Based Research for Health, Board on Health Sciences Policy; Institute of Medicine (2014) Drug Repurposing and Repositioning. National

Academies Press, Washington, D.C.

8. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6:95–108. <https://doi.org/10.1038/nrg1521>
9. 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.
<https://doi.org/10.1038/ncomms16022>
10. Garvey C (2018) Interview with Colin Garvey, Rensselaer Polytechnic Institute. *Artificial Intelligence and Systems Medicine Convergence. OMICS* 22:.
<https://doi.org/10.1089/omi.2017.0218>
11. Napolitano F, Zhao Y, Moreira VM, et al (2013) Drug repositioning: a machine-learning approach through data integration. *J Cheminform* 5:.
<https://doi.org/10.1186/1758-2946-5-30>
12. Farghali H, Kutinová Canová N, Arora M (2021) The potential applications of artificial intelligence in drug discovery and development. *Physiol Res* S715–S722.
<https://doi.org/10.33549/physiolres.934765>
13. Rodriguez S, Hug C, Todorov P, et al (2021) Machine learning identifies candidates for drug repurposing in Alzheimer’s disease. *Nat Commun* 12:.
<https://doi.org/10.1038/s41467-021-21330-0>
14. Hozumi Y, Wang R, Yin C, Wei G-W (2021) UMAP-assisted K-means clustering of large-

- scale SARS-CoV-2 mutation datasets. *Comput Biol Med* 131:104264.
- <https://doi.org/10.1016/j.combiomed.2021.104264>
15. McInnes L, Healy J, Melville J (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
 16. Cong Y, Endo T (2022) Multi-Omics and Artificial Intelligence-Guided Drug Repositioning: Prospects, Challenges, and Lessons Learned from COVID-19. *OMICS* 26:361–371.
 - <https://doi.org/10.1089/omi.2022.0068>
 17. Cong Y, Shintani M, Imanari F, et al (2022) A New Approach to Drug Repurposing with Two-Stage Prediction, Machine Learning, and Unsupervised Clustering of Gene Expression. *OMICS* 26:339–347. <https://doi.org/10.1089/omi.2022.0026>
 18. Luecken MD, Büttner M, Chaichoompu K, et al (2022) Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 19:41–50. <https://doi.org/10.1038/s41592-021-01336-8>
 19. Jiang W, Ye W, Tan X, Bao Y-J (2025) Network-based multi-omics integrative analysis methods in drug discovery: a systematic review. *BioData Min* 18:27.
 - <https://doi.org/10.1186/s13040-025-00442-z>
 20. Zheng Y, Liu Y, Yang J, et al (2024) Multi-omics data integration using ratio-based quantitative profiling with Quartet reference materials. *Nat Biotechnol* 42:1133–1149.

<https://doi.org/10.1038/s41587-023-01934-1>

21. Cong Y, Osada N, Endo T (2026) Reproducible and Multi-Study Transcriptomic Integration with *disint* , Disease Integration and Clustering Toolkit, and Application to Drug Repositioning. OMICS. <https://doi.org/10.1177/15578100251413812>
22. Hozumi Y, Wang R, Yin C, Wei G-W (2021) UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. Comput Biol Med 131:104264.
<https://doi.org/10.1016/j.combiomed.2021.104264>
23. Rugard M, Jaylet T, Taboureau O, et al (2021) Smell compounds classification using UMAP to increase knowledge of odors and molecular structures linkages. PLoS One 16:e0252486.
<https://doi.org/10.1371/journal.pone.0252486>
24. Tian M, Lin Z, Wang X, et al (2021) Pure Ion Chromatograms Combined with Advanced Machine Learning Methods Improve Accuracy of Discriminant Models in LC–MS-Based Untargeted Metabolomics. Molecules 26:2715. <https://doi.org/10.3390/molecules26092715>
25. Zhang W, Claesen M, Moerman T, et al (2021) Spatially aware clustering of ion images in mass spectrometry imaging data using deep learning. Anal Bioanal Chem 413:2803–2819.
<https://doi.org/10.1007/s00216-021-03179-w>
26. Dalmaijer ES, Nord CL, Astle DE (2022) Statistical power for cluster analysis. BMC Bioinformatics 23:205. <https://doi.org/10.1186/s12859-022-04675-1>

27. Sharma S, Quinn D, Melenhorst JJ, Pruteanu-Malinici I (2021) High-Dimensional Immune Monitoring for Chimeric Antigen Receptor T Cell Therapies. *Curr Hematol Malig Rep* 16:112–116. <https://doi.org/10.1007/s11899-020-00602-4>
28. Bennett JS, Gordon DM, Majumdar U, et al (2022) Use of machine learning to classify high-risk variants of uncertain significance in lamin A/C cardiac disease. *Heart Rhythm* 19:676–685. <https://doi.org/10.1016/j.hrthm.2021.12.019>
29. Cheng L, Karkhanis P, Gokbag B, et al (2022) DGCyTOF: Deep learning with graphic cluster visualization to predict cell types of single cell mass cytometry data. *PLoS Comput Biol* 18:e1008885. <https://doi.org/10.1371/journal.pcbi.1008885>
30. Wang H-Y, Zhao J, Zheng C-H (2021) SUSCC: Secondary Construction of Feature Space based on UMAP for Rapid and Accurate Clustering Large-scale Single Cell RNA-seq Data. *Interdiscip Sci* 13:83–90. <https://doi.org/10.1007/s12539-020-00411-6>
31. Spencer APC, Goodfellow M (2022) Using deep clustering to improve fMRI dynamic functional connectivity analysis. *Neuroimage* 257:119288. <https://doi.org/10.1016/j.neuroimage.2022.119288>
32. Hounkpe BW, Chenou F, de Lima F, De Paula EV (2021) HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res* 49:D947–D955.

<https://doi.org/10.1093/nar/gkaa609>

33. Akiba T, Sano S, Yanase T, et al (2019) Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, New York, NY, USA, pp 2623–2631
34. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
35. Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Theory Methods 3:1–27. <https://doi.org/10.1080/03610927408827101>
36. Davies DL, Bouldin DW (1979) A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell PAMI-1:224–227. <https://doi.org/10.1109/TPAMI.1979.4766909>
37. Duan Q, Reid SP, Clark NR, et al (2016) L1000CDS2: LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl 2:. <https://doi.org/10.1038/npjsba.2016.15>

Appendix A

A.1 Spearman Correlation Validation Script

```
import pandas as pd
import numpy as np
```

```

import random
from itertools import combinations

CLUSTERS = {
    "ATC_DTC": ["ATC", "DTC"],
    "FAD_SAD": ["FAD", "SAD"],
    "ACLF_sAH": ["ACLF", "sAH"],
    "DCM_GAC": ["DCM", "GAC"],
}

LOGFC_THRESHOLD = 0.5
RANDOM_ITER = 1000
SEED = 0

logfc = pd.read_csv("E:/disint/logFC_HKG_normalized_matrix.csv", index_col=0)
meta = pd.read_csv("E:/disint/meta.csv")

sample2disease = dict(zip(meta["sample"], meta["disease"]))
logfc = logfc.loc[:, logfc.columns.isin(sample2disease)]

disease_vectors = {}
for disease in meta["disease"].unique():
    samples = meta.loc[meta["disease"] == disease, "sample"].tolist()
    samples = [s for s in samples if s in logfc.columns]
    if len(samples) == 0:
        continue
    disease_vectors[disease] = logfc[samples].mean(axis=1)

all_diseases = list(disease_vectors.keys())

def shared_de_direction(v1, v2, threshold=0.5):
    mask = (v1.abs() > threshold) | (v2.abs() > threshold)
    v1s = v1[mask]
    v2s = v2[mask]
    if len(v1s) == 0:
        return np.nan, 0, 0, 0

```

```

    upup = ((v1s > 0) & (v2s > 0)).sum()
    dndn = ((v1s < 0) & (v2s < 0)).sum()
    opp = len(v1s) - (upup + dndn)
    ratio = (upup + dndn) / len(v1s)
    return ratio, len(v1s), int(upup), int(dndn)

within_rows = []
for cname, ds in CLUSTERS.items():
    d1, d2 = ds
    if (d1 not in disease_vectors) or (d2 not in disease_vectors):
        within_rows.append({
            "pair_type": "within",
            "cluster": cname,
            "d1": d1,
            "d2": d2,
            "threshold": LOGFC_THRESHOLD,
            "shared_direction_ratio": np.nan,
            "n_selected_genes": 0,
            "n_upup": 0,
            "n_dndn": 0,
        })
        continue

    ratio, nsel, upup, dndn = shared_de_direction(
        disease_vectors[d1], disease_vectors[d2], LOGFC_THRESHOLD
    )
    within_rows.append({
        "pair_type": "within",
        "cluster": cname,
        "d1": d1,
        "d2": d2,
        "threshold": LOGFC_THRESHOLD,
        "shared_direction_ratio": ratio,
        "n_selected_genes": nsel,
        "n_upup": upup,
        "n_dndn": dndn,
    })

```

```

    })

within_df = pd.DataFrame(within_rows)

def in_same_defined_cluster(a, b):
    return any((a in v) and (b in v) for v in CLUSTERS.values())

between_vals = []
between_rows = []

for d1, d2 in combinations(all_diseases, 2):
    if in_same_defined_cluster(d1, d2):
        continue
    ratio, nsel, upup, dndn = shared_de_direction(
        disease_vectors[d1], disease_vectors[d2], LOGFC_THRESHOLD
    )
    between_vals.append(ratio)
    between_rows.append({
        "pair_type": "between",
        "cluster": "",
        "d1": d1,
        "d2": d2,
        "threshold": LOGFC_THRESHOLD,
        "shared_direction_ratio": ratio,
        "n_selected_genes": nsel,
        "n_upup": upup,
        "n_dndn": dndn,
    })

between_df = pd.DataFrame(between_rows)
between_mean = float(np.nanmean(between_vals)) if len(between_vals) else np.nan

random.seed(SEED)
rand_vals = []
for _ in range(RANDOM_ITER):
    d1, d2 = random.sample(all_diseases, 2)

```

```

    ratio, _, _, _ = shared_de_direction(
        disease_vectors[d1], disease_vectors[d2], LOGFC_THRESHOLD
    )
    rand_vals.append(ratio)

random_mean = float(np.nanmean(rand_vals)) if len(rand_vals) else np.nan

print("=== Within-cluster shared DE direction ===")
print(within_df[["cluster", "d1", "d2", "shared_direction_ratio",
    "n_selected_genes", "n_upup", "n_dndn"]])
print()

print("Mean between-cluster shared_direction_ratio:", between_mean)
print("Mean random shared_direction_ratio:", random_mean)
print()

```

A.2 Shared DE Direction Validation Script

```

import pandas as pd
import numpy as np
from scipy.stats import spearmanr
from itertools import combinations
import random

CLUSTERS = {
    "ATC_DTC": ["ATC", "DTC"],
    "FAD_SAD": ["FAD", "SAD"],
    "ACLF_sAH": ["ACLF", "sAH"],
    "DCM_GAC": ["DCM", "GAC"],
}

logfc = pd.read_csv("E:/disint/logFC_HKG_normalized_matrix.csv", index_col=0)
meta = pd.read_csv("E:/disint/meta.csv")

sample2disease = dict(zip(meta["sample"], meta["disease"]))
logfc = logfc.loc[:, logfc.columns.isin(sample2disease)]

```

```

# disease-level mean logFC vector
disease_vectors = {}

for disease in meta["disease"].unique():
    samples = meta.loc[meta["disease"] == disease, "sample"]
    samples = [s for s in samples if s in logfc.columns]
    disease_vectors[disease] = logfc[samples].mean(axis=1)

def spearman_similarity(v1, v2):
    return spearmanr(v1, v2).correlation

cluster_similarities = {}

for cluster_name, diseases in CLUSTERS.items():
    sims = []
    for d1, d2 in combinations(diseases, 2):
        sim = spearman_similarity(
            disease_vectors[d1],
            disease_vectors[d2]
        )
        sims.append(sim)
    cluster_similarities[cluster_name] = sims

all_diseases = list(disease_vectors.keys())

between_sims = []

for d1, d2 in combinations(all_diseases, 2):
    if any(d1 in v and d2 in v for v in CLUSTERS.values()):
        continue
    sim = spearman_similarity(
        disease_vectors[d1],
        disease_vectors[d2]
    )
    between_sims.append(sim)

```

```

random_sims = []

for _ in range(1000):
    d1, d2 = random.sample(all_diseases, 2)
    sim = spearman_similarity(
        disease_vectors[d1],
        disease_vectors[d2]
    )
    random_sims.append(sim)

summary = []

for cluster, sims in cluster_similarities.items():
    summary.append({
        "cluster": cluster,
        "mean_within_similarity": np.mean(sims),
        "n_pairs": len(sims)
    })

summary_df = pd.DataFrame(summary)

print(summary_df)
print("Mean between-cluster similarity:", np.mean(between_sims))
print("Mean random similarity:", np.mean(random_sims))

```

A.3 DrugBank Target Extraction and Signaling Module Annotation Script

```

import xml.etree.ElementTree as ET
import pandas as pd

XML_PATH = r"C:/Users/ivy15/Downloads/full database.xml"
OUT_DETAIL = r"E:/disint/DR_candidate_targets_detail.csv"
OUT_SUMMARY = r"E:/disint/DR_candidate_targets_summary.csv"

# 非共通(DR 候補薬剤)
DCM_IDS = ["DB00694", "DB11805", "DB08073", "DB12686"]

```

```

GAC_IDS =
["DB12429", "DB11689", "DB11663", "DB04865", "DB01254", "DB07101", "DB07159", "DB08597", "D
B12847", "DB06999", "DB06061", "DB12904"]

# 共通
OVERLAP_IDS = ["DB08142", "DB03496"]

CANDIDATES = set(DCM_IDS + GAC_IDS + OVERLAP_IDS)

source_map = {dbid: "DCM_only" for dbid in DCM_IDS}
source_map.update({dbid: "GAC_only" for dbid in GAC_IDS})
source_map.update({dbid: "shared_overlap" for dbid in OVERLAP_IDS})

NS = "{http://www.drugbank.ca}"

def classify_target(target_name: str) -> dict:
    t = (target_name or "").lower()
    return {
        "is_kinase": int("kinase" in t),
        "is_mapk_family": int("mitogen-activated protein kinase" in t or
t.startswith("mapk")),
        "is_map2k": int("mitogen-activated protein kinase kinase" in t or
t.startswith("map2k")),
        "is_map3k": int("mitogen-activated protein kinase kinase kinase" in t or
t.startswith("map3k")),
        "is_pi3k": int("phosphatidylinositol" in t or t.startswith("pik3")),
        "is_akt": int("rac-alpha serine/threonine-protein kinase" in t or
t.startswith("akt")),
        "is_cdk": int("cyclin-dependent kinase" in t or t.startswith("cdk")),
        "is_jak_stat": int("janus kinase" in t or "signal transducer and activator
of transcription" in t or t.startswith("stat")),
        "is_tgf_beta": int("activin receptor" in t or "tgf" in t),
        "is_hdac": int("histone deacetylase" in t or t.startswith("hdac")),
    }

rows = []

```

```

ctx = ET.iterparse(XML_PATH, events=("end",))
for _, drug in ctx:
    if drug.tag != f"{NS}drug":
        continue

    primary_id = None
    for did in drug.findall(f"{NS}drugbank-id"):
        if did.attrib.get("primary") == "true":
            primary_id = (did.text or "").strip()
            break

    if not primary_id or primary_id not in CANDIDATES:
        drug.clear()
        continue

    drug_name = (drug.findtext(f"{NS}name") or "").strip()
    cand_src = source_map.get(primary_id, "")

    targets = drug.find(f"{NS}targets")

    if targets is None:
        rows.append({
            "drug_id": primary_id,
            "drug_name": drug_name,
            "candidate_source": cand_src,
            "target_gene": "",
            "target_name": "",
            "action": "",
            **{k: 0 for k in classify_target("").keys()}
        })
        drug.clear()
        continue

    any_target = False
    for t in targets.findall(f"{NS}target"):

```

```

any_target = True

target_name = (t.findtext(f"{NS}name") or "").strip()

gene = ""
poly = t.find(f"{NS}polypeptide")
if poly is not None:
    gene = (poly.findtext(f"{NS}gene-name") or "").strip()

act_list = []
acts = t.find(f"{NS}actions")
if acts is not None:
    for a in acts.findall(f"{NS}action"):
        if a.text:
            act_list.append(a.text.strip())
action = ";".join(act_list)

flags = classify_target(target_name)

rows.append({
    "drug_id": primary_id,
    "drug_name": drug_name,
    "candidate_source": cand_src,
    "target_gene": gene,
    "target_name": target_name,
    "action": action,
    **flags
})

if not any_target:
    rows.append({
        "drug_id": primary_id,
        "drug_name": drug_name,
        "candidate_source": cand_src,
        "target_gene": "",
        "target_name": "",
    })

```

```

        "action": "",
        **{k: 0 for k in classify_target("").keys()}
    })

    drug.clear()

df = pd.DataFrame(rows)
df.to_csv(OUT_DETAIL, index=False, encoding="utf-8-sig")

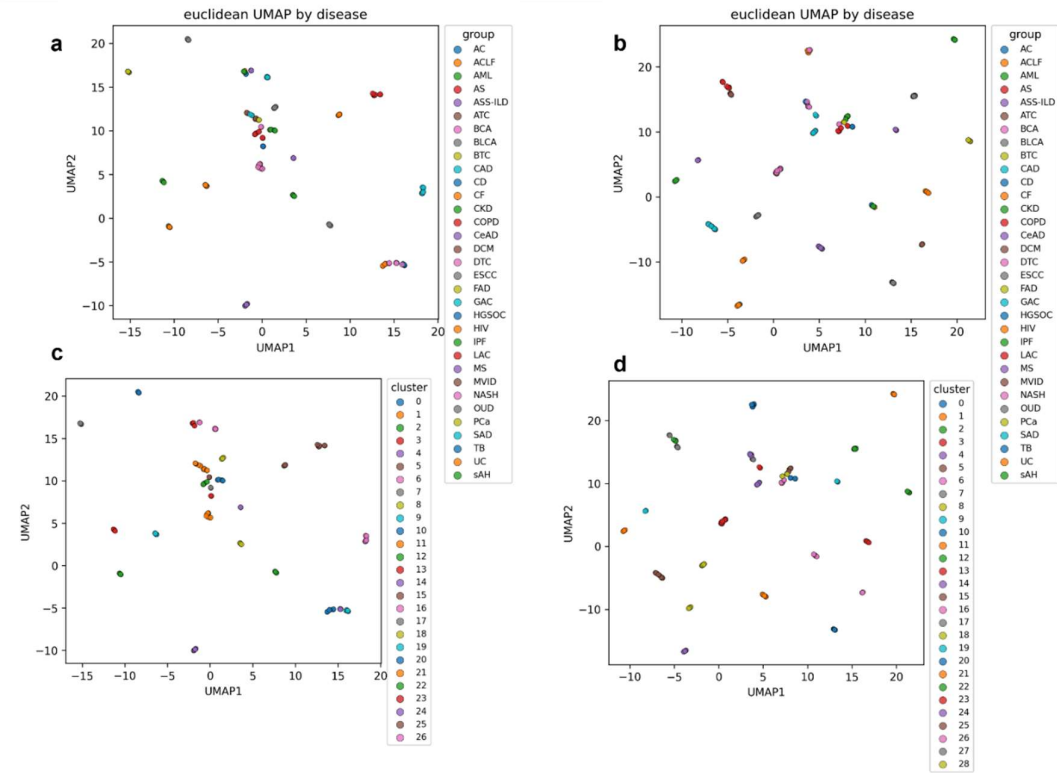
flag_cols = [c for c in df.columns if c.startswith("is_")]
summary = (
    df.groupby(["candidate_source", "drug_id", "drug_name"], as_index=False)
        .agg(
            n_rows=("drug_id", "size"),
            n_unique_target_genes=("target_gene", lambda s: s.replace("",
pd.NA).dropna().nunique()),
            **{c: (c, "sum") for c in flag_cols}
        )
        .sort_values(["candidate_source", "drug_id"])
    )

summary.to_csv(OUT_SUMMARY, index=False, encoding="utf-8-sig")

print("Saved detail :", OUT_DETAIL)
print("Saved summary:", OUT_SUMMARY)

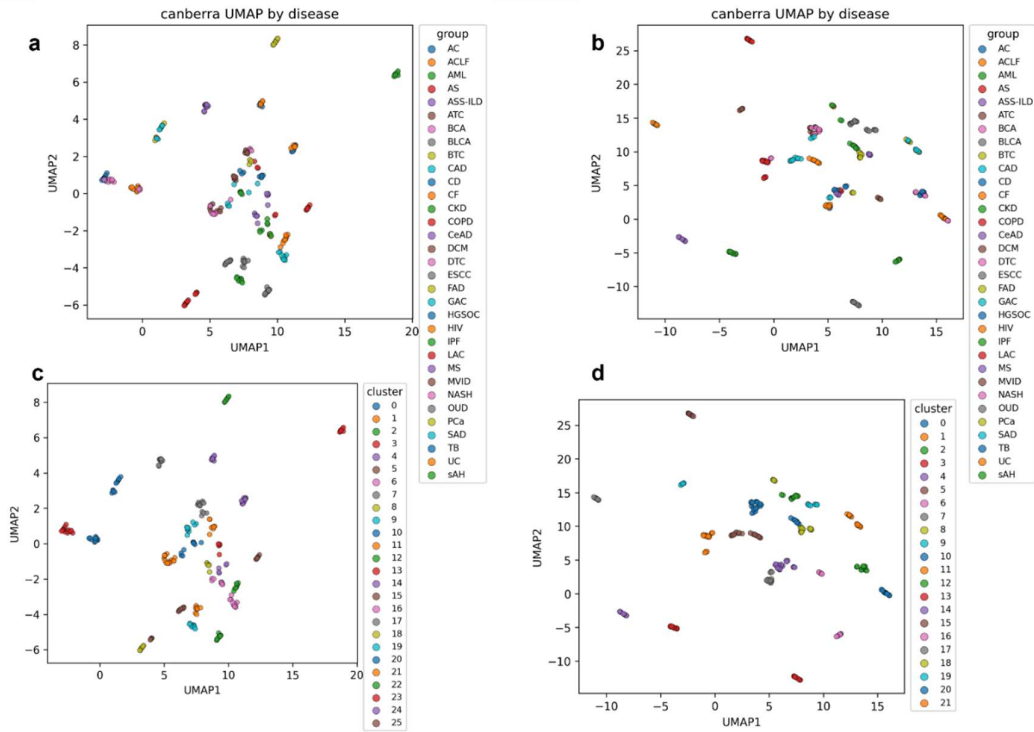
```

Supplementary Figures



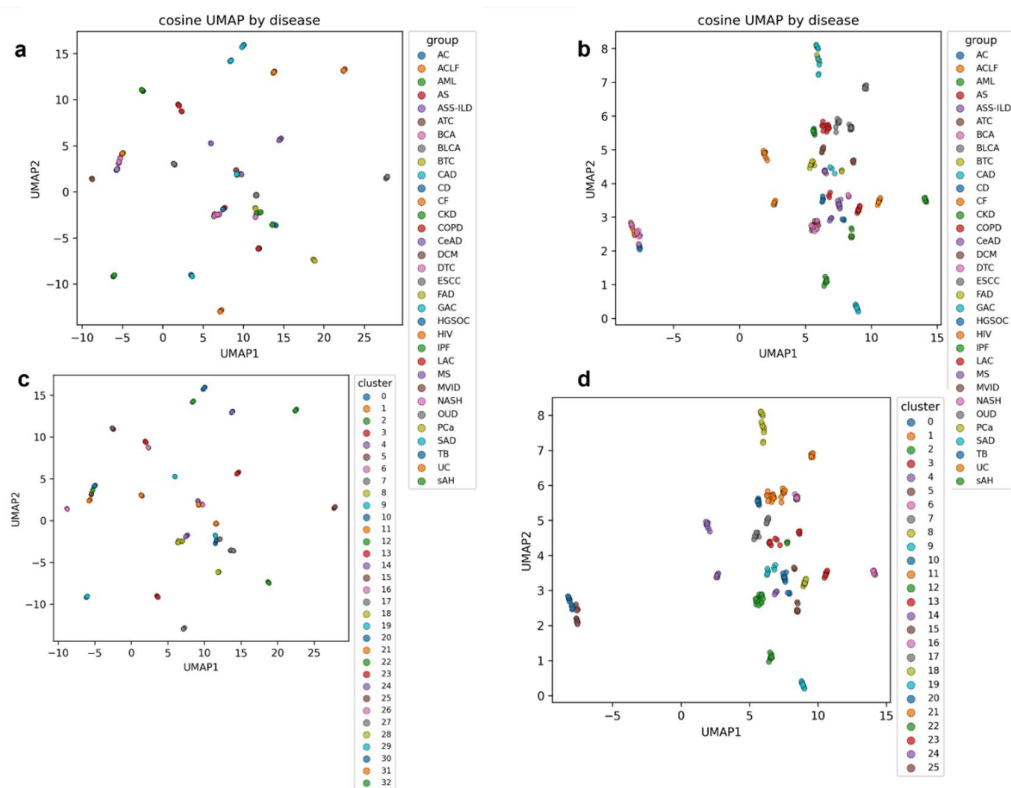
Supplementary Figure S1. Effects of HKG normalization on disease-level embeddings and clustering (Euclidean distance)

UMAP embeddings and K-means clustering results generated using the Euclidean distance metric are shown before and after HKG-based normalization.



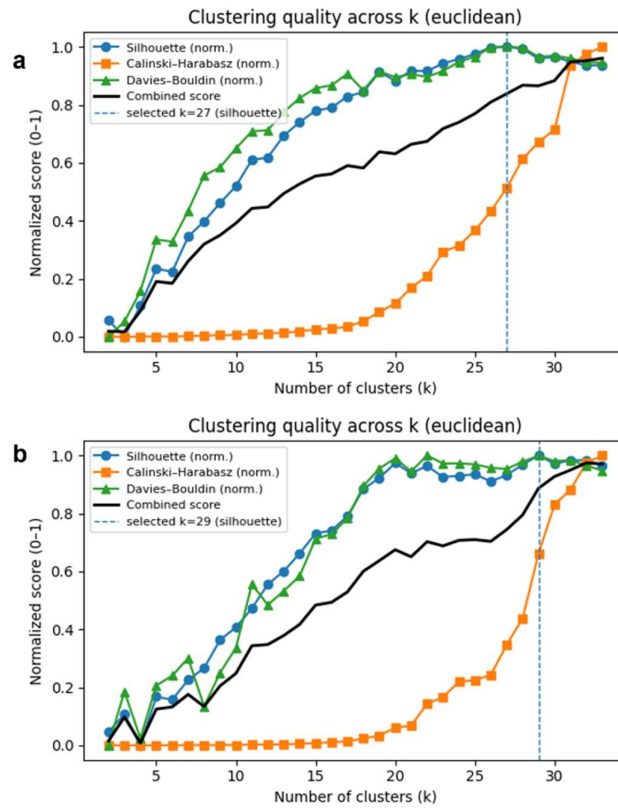
Supplementary Figure S2. Effects of HKG normalization on disease-level embeddings and clustering (Canberra distance)

UMAP embeddings and K-means clustering results generated using the Canberra distance metric are shown before and after HKG-based normalization.



Supplementary Figure S3. Effects of HKG normalization on disease-level embeddings and clustering (Cosine distance)

UMAP embeddings and K-means clustering results generated using the Cosine distance metric are shown before and after HKG-based normalization.

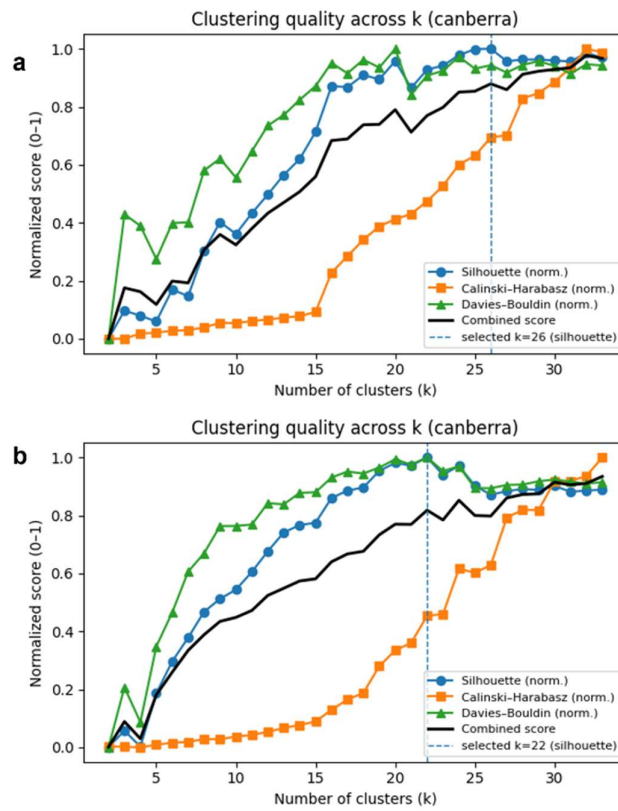


Supplementary Figure S4. Determination of the optimal number of clusters (k) before and after

HKG normalization (Euclidean distance)

Cluster-number selection curves based on internal validation indices for the Euclidean distance metric, shown for

HKG-normalized (a) and raw (b) $\log_2$ FC matrices.

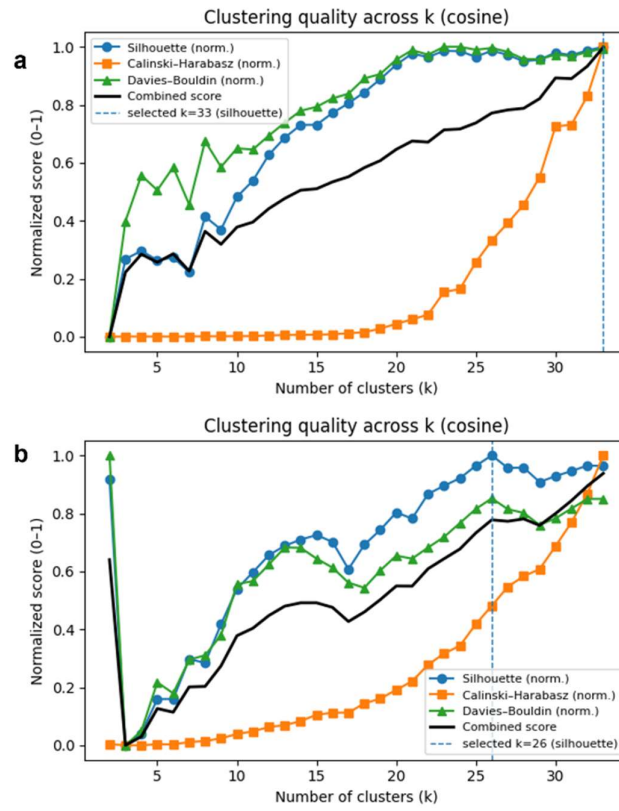


Supplementary Figure S5. Determination of the optimal number of clusters (k) before and after

HKG normalization (Canberra distance)

Cluster-number selection curves based on internal validation indices for the Canberra distance metric, shown for

HKG-normalized (a) and raw (b) $\log_2$ FC matrices.

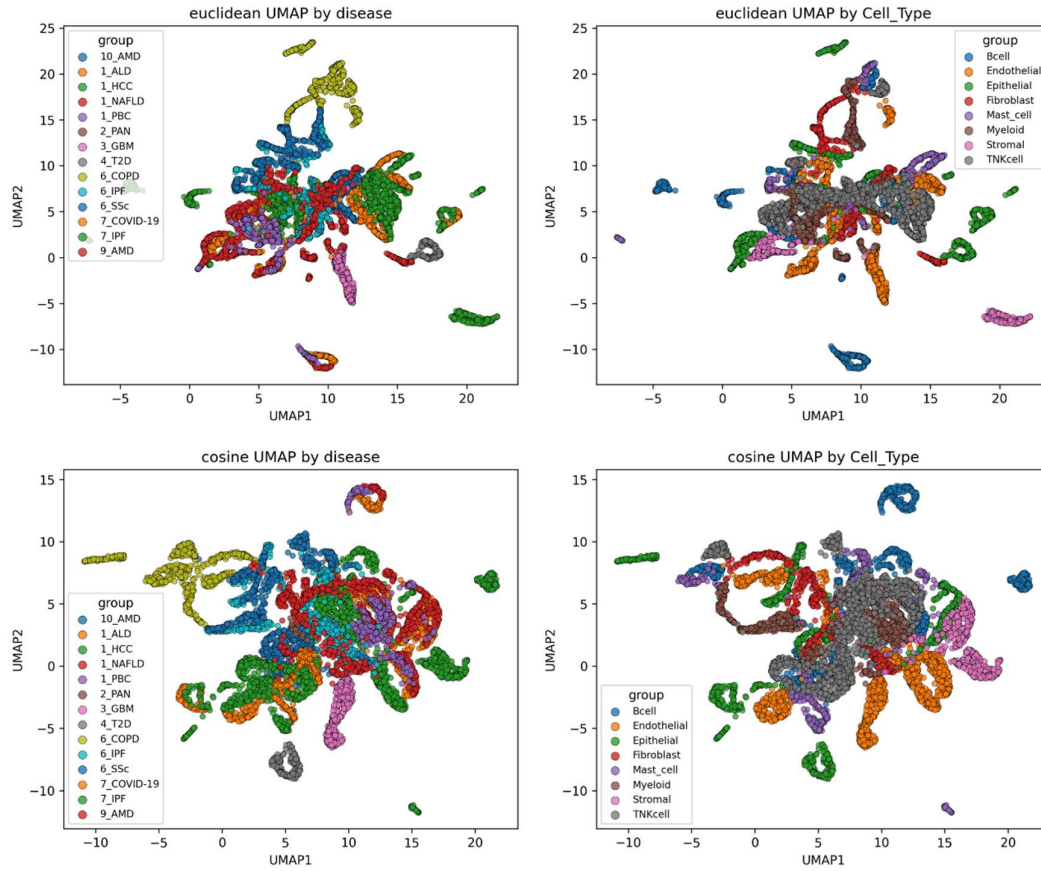


Supplementary Figure S6. Determination of the optimal number of clusters (k) before and after

HKG normalization (Cosine distance)

Cluster-number selection curves based on internal validation indices for the Cosine distance metric, shown for HKG-

normalized (a) and raw (b) $\log_2$ FC matrices.



Supplementary Figure 7. UMAP embeddings of the industrial single-cell cohort based on

Euclidean and Cosine distances

Supplementary Tables

Supplementary Table S1. Cluster Assignments and UMAP Coordinates Before HKG Normalization

	euclidean			manhattan			canberra			cosine		
sample	UMAP1	UMAP2	cluster	UMAP1	UMAP2	cluster	UMAP1	UMAP2	cluster	UMAP1	UMAP2	cluster
SRR25836907	4.3113	9.816983	4	-10.7642	17.294	9	1.506456	8.620649	15	8.791637	0.409315	9
SRR25836909	4.300675	9.808733	4	-10.8414	17.37213	9	1.506856	8.695667	15	8.798976	0.373159	9
SRR25836910	4.415505	9.93888	4	-10.672	17.20291	9	1.813659	8.901665	15	8.809542	0.410416	9
SRR25836912	4.378691	9.900311	4	-10.8364	17.36716	9	1.67341	8.748926	15	8.973298	0.219593	9
SRR25836913	4.591143	10.19536	4	-10.6715	17.2026	9	2.50052	8.962561	15	8.935583	0.262939	9
SRR25836914	4.478565	10.02986	4	-10.5891	17.12012	9	1.955663	9.129435	15	8.844835	0.34657	9
SRR25836915	4.55966	10.17298	4	-10.591	17.12215	9	2.186542	8.975081	15	8.835316	0.35192	9
SRR25836916	4.527327	10.11334	4	-10.6172	17.14822	9	2.20443	9.070947	15	8.924244	0.297795	9
SRR25836917	4.294703	9.802946	4	-10.8912	17.42092	9	1.637023	8.717138	15	9.000001	0.201938	9
SRR25836920	4.354407	9.874446	4	-10.8674	17.39811	9	1.763116	8.781299	15	8.881897	0.307606	9
SRR26261421	-8.21492	5.681773	9	5.94982	12.93795	26	8.723407	9.713672	8	6.566368	4.296259	23
SRR26261422	-8.21293	5.686073	9	6.002093	12.99867	26	8.891794	9.507098	8	6.47596	4.404175	23
SRR26261423	-8.20352	5.693234	9	5.935912	12.93479	26	8.803955	9.500851	8	6.447192	4.341343	23
SRR26261424	-8.2172	5.679189	9	5.915858	12.93474	26	8.870931	9.618343	8	6.414802	4.377408	23
SRR26261425	-8.28435	5.612579	9	6.039414	12.9582	26	8.799745	9.667605	8	6.469192	4.370638	23
SRR14150310	8.092552	10.91579	20	1.520138	12.62881	20	6.222753	4.144233	4	6.872257	3.735768	19
SRR14150311	8.076291	10.91046	20	1.554557	12.61234	20	6.102806	4.171899	4	6.762442	3.649862	19
SRR14150312	8.04114	10.92954	20	1.558409	12.40879	20	6.101464	4.354567	4	6.819326	3.593469	19
SRR16230213	-10.6944	2.491788	21	21.6227	4.165889	6	-3.52288	-5.11181	3	14.16599	3.447999	6
SRR16230214	-10.7188	2.467096	21	21.65552	4.133048	6	-3.51406	-5.17851	3	14.04691	3.566873	6
SRR16230215	-10.7095	2.476531	21	21.75695	4.031626	6	-3.50983	-5.15819	3	14.13426	3.478628	6
SRR16230216	-10.7006	2.485512	21	21.68944	4.09918	6	-3.54899	-5.0961	3	14.09247	3.520996	6
SRR16230217	-10.6113	2.574976	21	21.74189	4.04663	6	-4.11489	-4.85112	3	14.1063	3.517488	6
SRR16230218	-10.6374	2.548811	21	21.76822	4.020336	6	-4.09875	-4.79047	3	14.03823	3.568292	6
SRR16230219	-10.7322	2.453857	21	21.54027	4.248356	6	-3.72862	-5.1222	3	14.08398	3.529781	6
SRR16230220	-10.7798	2.40511	21	21.66931	4.11934	6	-3.80921	-4.952	3	14.0582	3.583174	6
SRR16230221	-10.5381	2.648233	21	21.54474	4.243888	6	-3.92588	-4.92463	3	14.06398	3.482341	6
SRR16230222	-10.6555	2.530914	21	21.56177	4.226837	6	-4.01115	-4.79729	3	14.00882	3.567985	6

SRR8325947	7.1013	10.13083	26	-5.04013	-0.98669	14	-2.22574	26.55735	5	9.010014	3.251884	18
SRR8325948	7.133336	10.11987	26	-5.14008	-1.0865	14	-2.43037	26.76202	5	8.96484	3.175332	18
SRR8325949	7.186527	10.18134	26	-5.07358	-1.02	14	-2.27	26.60149	5	9.087754	3.262141	18
SRR8325950	7.099037	10.13665	26	-5.09359	-1.03988	14	-2.48152	26.81335	5	9.110682	3.334604	18
SRR8325951	7.09417	10.06715	26	-4.95976	-0.90646	14	-2.41971	26.75106	5	8.924498	3.124026	18
SRR8325952	7.103951	10.21222	26	-5.11338	-1.05962	14	-2.29981	26.63142	5	8.993824	3.209746	18
SRR8325953	7.09975	10.09271	26	-5.07014	-1.01646	14	-2.35194	26.68336	5	8.975564	3.17249	18
SRR8325954	7.034469	10.20607	26	-5.07174	-1.01824	14	-2.15942	26.49063	5	9.034928	3.196812	18
SRR8325955	7.353812	10.58022	26	-5.28502	-1.22626	14	-2.06303	26.39089	5	9.016718	3.221755	18
SRR8325956	7.362859	10.61478	26	-5.32204	-1.26271	14	-1.99676	26.32656	5	9.131099	3.328383	18
SRR21619385	-6.38	-4.96321	5	9.564403	-16.8006	5	13.36262	10.00303	11	5.852967	8.059834	8
SRR21619383	-6.46553	-4.83873	5	9.698184	-16.8042	5	13.11015	10.32324	11	5.876123	8.041863	8
SRR21619384	-6.41783	-4.98213	5	9.659163	-16.9243	5	13.41603	9.943102	11	5.840539	8.064692	8
SRR21619382	-6.33537	-5.01612	5	9.656217	-16.7688	5	13.13547	10.12797	11	5.891517	8.030063	8
SRR21619410	-6.9875	-4.27824	25	9.319427	-16.3832	5	12.18051	11.86277	11	6.067627	7.580731	8
SRR21619411	-7.03334	-4.23476	25	9.374606	-16.4793	5	12.56248	11.41476	11	5.926405	7.752401	8
SRR21619412	-7.09076	-4.17674	25	9.248771	-16.3886	5	12.46826	11.56778	11	5.858358	7.810741	8
SRR21619413	-7.0641	-4.20395	25	9.219936	-16.4278	5	12.35355	11.63024	11	5.977531	7.664879	8
SRR21619393	-6.74344	-4.52377	25	9.228921	-16.7951	5	3.563734	12.09726	10	5.989721	7.251364	8
SRR21619392	-6.66752	-4.59676	5	9.278886	-16.8184	5	3.409523	11.96788	10	5.98368	7.211577	8
SRR21619391	-7.12545	-4.14203	25	9.271345	-16.2724	5	12.38016	11.66803	11	6.045746	7.656817	8
SRR21619390	-7.05587	-4.21047	25	9.335245	-16.2783	5	12.38628	11.72678	11	5.941683	7.700506	8
SRR21619381	-6.37729	-4.94455	5	9.536888	-16.8604	5	13.06693	10.36642	11	5.809763	8.112489	8
SRR21619380	-6.43028	-4.91835	5	9.553002	-17.0135	5	13.3617	9.990555	11	5.965419	7.983481	8
SRR21619379	-6.38376	-4.96013	5	9.587299	-16.8537	5	13.28435	10.13015	11	5.927849	8.035265	8
SRR21619378	-6.47997	-4.84048	5	9.453437	-16.8966	5	13.13992	10.28756	11	5.821132	8.104075	8
SRR21619414	-6.81388	-4.45329	25	9.379786	-16.7969	5	3.733558	12.23623	10	6.033709	7.518215	8
SRR25654709	16.13536	-7.33873	16	0.890748	11.55191	21	9.735839	3.072267	16	8.609942	4.639623	13
SRR25654705	16.23186	-7.24235	16	0.934265	11.58469	21	9.590134	3.220202	16	8.587521	4.676421	13
SRR25654706	16.14133	-7.33319	16	0.888924	11.54955	21	9.75968	3.041442	16	8.609329	4.639835	13
SRR25654707	16.1653	-7.31275	16	0.978609	11.63392	21	9.590207	3.209138	16	8.654092	4.698196	13
SRR25654708	16.1659	-7.30863	16	0.997864	11.67504	21	9.80162	3.009154	16	8.636195	4.610655	13
SRR25654710	16.15904	-7.31562	16	0.908707	11.61583	21	9.852665	2.951996	16	8.574265	4.674224	13
SRR21609197	-1.75303	-2.87346	18	11.4561	9.251464	17	7.627071	14.49151	2	8.48744	5.610899	16
SRR21609198	-1.8087	-2.92887	18	11.57336	9.087981	17	7.021047	14.20872	2	8.400465	5.597454	16

SRR21609201	-1.9092	-3.0296	18	11.49724	9.18583	17	7.596253	14.53901	2	8.428193	5.607429	16
SRR21609202	-1.81704	-2.93755	18	11.3905	9.315748	17	7.129468	14.25575	2	8.412102	5.64012	16
SRR21609204	-1.87304	-2.99326	18	11.60329	9.100105	17	7.045571	14.10607	2	8.488632	5.691278	16
SRR21609205	-1.84787	-2.96752	18	11.63645	9.065275	17	7.070689	14.0418	2	8.316455	5.751312	16
SRR21609203	-1.67165	-2.79217	18	11.53152	9.171709	17	7.237227	14.32882	2	8.421783	5.689792	16
SRR21609200	-1.6439	-2.76449	18	11.33654	9.374154	17	7.300438	14.37356	2	8.452949	5.667422	16
SRR21609199	-1.63059	-2.75122	18	11.33829	9.371378	17	7.431941	14.63342	2	8.475224	5.626292	16
SRR21609206	-1.73688	-2.85728	18	11.43376	9.27247	17	7.467004	14.48367	2	8.430279	5.646571	16
SRR19732846	4.665117	12.4786	23	2.072406	12.32563	13	5.178675	3.309361	17	7.207944	4.295267	23
SRR19732849	4.649097	12.48749	23	2.386578	12.53898	13	5.091054	3.16815	17	6.999484	4.443515	23
SRR19732852	4.542741	12.66273	23	3.166023	12.61579	19	5.236589	3.204989	17	6.863988	4.489938	23
SRR23226403	7.1525	11.18226	28	0.806401	11.36151	21	-0.28606	9.031417	1	8.274597	3.609121	25
SRR23226404	7.147884	11.17249	28	0.812449	11.37557	21	-0.29158	9.013201	1	8.287644	3.626553	25
SRR23226405	7.142164	11.17807	28	0.797185	11.34798	21	-0.22974	9.060672	1	8.241664	3.664197	25
SRR13571325	16.82896	0.671003	13	-1.29278	7.665036	16	3.688833	8.762155	15	10.64225	3.538991	3
SRR13571327	16.66857	0.831321	13	-1.18371	7.555198	16	3.901171	8.597477	15	10.62184	3.517656	3
SRR13571331	16.84394	0.656033	13	-1.21042	7.582357	16	3.513418	8.86539	15	10.56457	3.461264	3
SRR13571339	16.68461	0.815333	13	-1.19847	7.570008	16	4.039536	8.461938	15	10.58253	3.475211	3
SRR13571349	16.56448	0.935277	13	-1.20132	7.574441	16	4.185818	8.300369	15	10.56246	3.514366	3
SRR13571321	16.80447	0.695364	13	-1.1288	7.501056	16	3.273269	9.021604	15	10.51853	3.4211	3
SRR13571347	16.64597	0.853858	13	-1.31658	7.688341	16	4.115266	8.386552	15	10.54668	3.447481	3
SRR13571355	16.75771	0.742228	13	-1.10514	7.477191	16	3.683029	8.790817	15	10.47501	3.364159	3
SRR13571357	16.69696	0.802977	13	-1.15645	7.527911	16	4.016034	8.486517	15	10.55289	3.437336	3
SRR13571359	16.88935	0.61061	13	-1.29272	7.664739	16	3.67811	8.783434	15	10.67855	3.574337	3
SRR15021920	7.673073	11.50053	28	1.439021	11.89405	21	7.210246	3.950529	4	7.72243	4.374474	22
SRR15021921	7.651555	11.54055	28	1.526183	11.9157	13	7.262701	3.965819	4	7.773593	4.392214	22
SRR15021922	7.69559	11.49776	28	1.430656	11.81967	21	7.340519	3.931818	4	7.735868	4.345703	22
SRR25606246	-4.90065	16.81995	22	3.733603	21.65396	10	-0.98277	8.600983	1	6.701106	5.689332	21
SRR25606247	-4.93366	16.85446	22	3.69738	21.6868	10	-0.74867	8.611488	1	6.745411	5.696446	21
SRR25606248	-4.87275	16.78686	22	3.74377	21.64418	10	-0.83462	8.706407	1	6.734025	5.634461	21
SRR25606249	-4.85272	16.86405	22	3.539492	21.84271	10	-0.72009	8.477146	1	6.691928	5.747414	21
SRR25606250	-4.88983	16.80767	22	3.668496	21.72229	10	-1.03474	8.698287	1	6.783411	5.644045	21
SRR25606252	-4.95835	16.89831	22	3.555139	21.82425	10	-0.51978	8.503635	1	6.730303	5.723432	21
SRR25606253	-4.81578	16.73623	22	3.725961	21.66257	10	-1.07023	8.741175	1	6.638434	5.594345	21
SRR25606254	-4.98294	16.88877	22	3.563046	21.80443	10	-0.55095	8.398031	1	6.544418	5.703319	21

SRR25606255	-4.9116	16.77925	22	3.82346	21.55823	10	-1.06116	8.632369	1	6.751888	5.640791	21
SRR25606256	-5.1243	17.03116	22	3.46481	21.91273	10	-0.62099	8.526364	1	6.597598	5.529162	21
SRR15334651	-5.54961	17.69173	27	4.387103	20.69529	23	-0.82929	6.191422	21	6.26436	5.724485	21
SRR15334652	-5.53452	17.67529	27	4.309315	20.77581	23	-0.78362	6.223565	21	6.348416	5.553076	21
SRR15334653	-5.58424	17.72723	27	4.356941	20.73425	23	-0.85025	6.230995	21	6.364466	5.718784	21
SRR15334654	-5.52029	17.66342	27	4.275644	20.80974	23	-0.74455	6.270931	21	6.263777	5.706028	21
SRR15334655	-5.51499	17.66021	27	4.32389	20.77082	23	-0.93672	6.100945	21	6.317136	5.839685	21
SRR18520163	8.042152	12.361	15	13.36861	6.622191	8	11.51586	-6.02523	6	6.488107	1.042429	12
SRR18520164	8.0208	12.36409	15	13.51852	6.47236	8	11.41791	-6.1221	6	6.533444	1.09081	12
SRR18520165	7.942638	12.26555	15	13.36738	6.62343	8	11.45942	-6.08169	6	6.634466	1.087494	12
SRR18520166	7.999743	12.32035	15	13.34999	6.641028	8	11.41732	-6.12388	6	6.557974	1.064038	12
SRR18520167	8.064693	12.43252	15	13.57774	6.412423	8	11.6188	-5.92171	6	6.580273	1.169007	12
SRR18520168	8.108701	12.46915	15	13.57845	6.412273	8	11.54939	-5.99113	6	6.593188	1.183242	12
SRR18520169	8.016234	12.37985	15	13.44517	6.545573	8	11.515	-6.02601	6	6.59178	1.164223	12
SRR18520170	8.052612	12.41052	15	13.53974	6.450616	8	11.50281	-6.03823	6	6.599604	1.107123	12
SRR18520171	7.867174	11.94103	15	13.29036	6.700757	8	11.25622	-6.28498	6	6.493172	1.238685	12
SRR18520172	7.853389	11.93971	15	13.4304	6.560551	8	11.1964	-6.3448	6	6.407889	0.965113	12
SRR9888136	10.90134	-1.45666	6	-11.9345	-3.51061	2	5.44364	16.86659	18	8.469753	2.396657	5
SRR9888137	10.80912	-1.36328	6	-11.7976	-3.48062	2	6.248739	14.64244	2	8.472996	2.677459	5
SRR9888138	10.94237	-1.49781	6	-11.8844	-3.50527	2	5.338821	16.94608	18	8.488342	2.445927	5
SRR9888139	10.87095	-1.42508	6	-11.8038	-3.4879	2	6.132355	14.70848	2	8.515707	2.6285	5
SRR9888140	10.98387	-1.53907	6	-11.9771	-3.59526	2	5.526484	16.72701	18	8.523278	2.400845	5
SRR9888141	10.95097	-1.50617	6	-12.0077	-3.50289	2	5.367354	16.91218	18	8.458813	2.450525	5
SRR9888142	10.98725	-1.54255	6	-11.8894	-3.56076	2	5.480955	16.80094	18	8.453305	2.402401	5
SRR9888143	11.03144	-1.58686	6	-11.9161	-3.53435	2	5.407686	16.86669	18	8.532381	2.436315	5
SRR15254306	3.973815	22.5309	0	4.852144	12.97612	0	16.03233	-0.18208	0	-8.21623	2.815939	0
SRR15254307	3.970946	22.66266	0	4.896878	13.01644	0	16.09098	-0.24075	0	-8.13799	2.742919	0
SRR15254308	3.885216	22.57908	0	4.804953	12.9859	0	16.0355	-0.18699	0	-8.19781	2.790038	0
SRR15254309	3.754337	13.93054	17	3.471651	12.85648	19	13.93666	3.532002	12	-7.54991	2.458106	15
SRR15254310	3.784122	13.93002	17	3.433618	12.77948	19	13.89491	3.502798	12	-7.58051	2.486676	15
SRR15254311	3.836332	13.8735	17	3.480427	12.77331	19	13.99256	3.444009	12	-7.61104	2.426274	15
SRR15254312	3.847647	13.85811	17	3.811157	12.7872	27	13.0687	4.001149	12	-7.6741	2.60498	15
SRR15254313	3.797323	13.93265	17	3.571286	12.82447	19	13.07362	3.992695	12	-7.7962	2.514584	0
SRR15254314	3.895409	13.82068	17	3.666712	12.81674	19	13.10258	4.042237	12	-7.79027	2.570426	0
SRR15254315	3.674667	14.6539	24	3.022765	13.08277	24	13.08801	4.056338	12	-7.64618	2.206996	15

SRR15254333	3.760077	22.47574	0	4.870298	12.87102	0	15.66223	0.149671	0	-8.19743	2.751341	0
SRR15254334	3.719811	22.53382	0	4.837397	12.85972	0	15.66836	0.241493	0	-7.9995	2.566641	0
SRR15254335	3.79549	22.44532	0	4.950125	12.75114	0	15.59127	0.226914	0	-8.13539	2.681018	0
SRR15254336	3.789721	22.20528	0	4.195997	12.8235	27	15.37834	0.588603	0	-7.95785	2.464388	0
SRR15254337	3.809028	22.189	0	4.171502	12.85732	27	15.34443	0.624241	0	-7.98713	2.589451	0
SRR15254338	3.804798	22.19117	0	4.127726	12.8051	27	15.46584	0.484643	0	-7.93485	2.492147	0
SRR15254339	3.748309	22.49252	0	5.022922	12.75796	0	15.77294	0.092678	0	-8.18174	2.781911	0
SRR15254340	3.721228	22.5465	0	5.006771	12.77572	0	15.81751	0.050141	0	-8.21101	2.804647	0
SRR15254341	3.70047	22.54267	0	4.903265	12.92759	0	15.85779	0.00709	0	-8.26238	2.845534	0
SRR15254325	3.506955	14.64211	24	3.130189	13.35571	24	13.79365	4.066649	12	-7.64476	2.201111	15
SRR15254326	3.496417	14.70828	24	3.066461	13.25876	24	13.82414	4.034943	12	-7.60326	2.139624	15
SRR15254327	3.594061	14.53561	24	3.067393	13.26085	24	13.54145	3.574554	12	-7.58068	2.150608	15
SRR15254328	3.580062	14.54467	24	2.949394	13.17749	24	13.48075	3.616912	12	-7.57892	2.124473	15
SRR15254329	3.676094	14.45171	24	2.997198	13.23819	24	13.4752	3.544348	12	-7.54543	2.084489	15
SRR15254330	3.634126	14.45165	24	3.209333	13.38669	24	13.69807	3.901898	12	-7.65001	2.195061	15
SRR15254331	3.559729	14.56294	24	3.2427	13.27674	24	13.76685	4.053564	12	-7.5946	2.134087	15
SRR15254332	3.637112	14.43221	24	3.350773	13.42182	24	13.64489	4.055524	12	-7.59756	2.083689	15
SRR15254324	3.498984	14.68644	24	3.089791	13.39572	24	13.67762	4.130749	12	-7.5644	2.041313	15
SRR15835822	-4.7311	15.93518	7	-2.02524	26.30696	3	-3.06677	16.24196	9	6.237861	4.918256	7
SRR15835823	-4.7106	15.92282	7	-2.02051	26.31198	3	-3.14882	16.15976	9	6.278913	5.030919	7
SRR15835826	-4.59016	15.70508	7	-2.11144	26.21233	3	-2.90276	16.40688	9	6.337438	5.013525	7
SRR15835827	-4.67946	15.79573	7	-2.12176	26.2018	3	-2.91812	16.39157	9	6.368664	5.083634	7
SRR15835828	-4.66717	15.78189	7	-2.15542	26.16634	3	-2.871	16.43862	9	6.296739	5.031629	7
SRR15835829	-4.64742	15.76127	7	-2.17538	26.1496	3	-2.9731	16.33666	9	6.244098	4.934022	7
SRR25223472	13.09587	-13.2774	10	13.40747	-4.70802	1	9.334145	13.23761	19	7.299642	5.674815	1
SRR25223473	13.06558	-13.2474	10	13.35965	-4.75539	1	9.230927	13.27929	19	7.433689	5.831975	1
SRR25223476	13.01919	-13.2009	10	13.40438	-4.70883	1	9.402174	13.23589	19	7.524444	5.794036	1
SRR25223478	12.90631	-13.0882	10	13.30738	-4.80655	1	8.617538	13.19414	19	7.485162	5.850149	1
SRR25223479	12.95853	-13.1404	10	13.18046	-4.89002	1	8.55358	13.33767	19	7.431445	5.78901	1
SRR25223480	12.98831	-13.1702	10	13.18911	-4.92294	1	8.662694	13.17109	19	7.505229	5.81869	1
SRR25223481	12.98553	-13.1673	10	13.29646	-4.81567	1	8.616125	13.21378	19	7.527825	5.816856	1
SRR25223483	12.98708	-13.1688	10	13.39584	-4.71823	1	9.273884	13.25017	19	7.427717	5.917853	1
SRR25223484	13.09096	-13.2725	10	13.44023	-4.6741	1	9.419276	13.19464	19	7.319099	5.522723	1
SRR25223485	13.02524	-13.2071	10	13.24356	-4.87026	1	8.801828	13.08671	19	7.573599	5.864323	1
SRR16093329	8.572838	10.81988	20	1.240583	12.74271	20	5.823517	4.2215	4	6.339496	3.538869	19

SRR16093333	8.610179	10.78018	20	1.125539	12.86869	20	5.562275	4.28574	4	6.329792	3.556808	19
SRR16093337	8.58252	10.81029	20	1.262433	12.73955	20	5.865755	4.311416	4	6.257401	3.449151	19
SRR16093341	8.597617	10.79667	20	1.115542	12.87348	20	5.677236	4.326865	4	6.312393	3.488091	19
SRR16093345	8.569006	10.82033	20	1.204798	12.80178	20	5.581661	4.412432	4	6.30645	3.62119	19
SRR21227478	21.44404	8.563091	12	-1.31248	26.88261	22	7.951139	9.878182	8	5.671398	4.534706	17
SRR21227479	21.35298	8.654337	12	-1.41531	26.96534	22	8.129116	9.588871	8	5.567113	4.660011	17
SRR21227482	21.47056	8.536699	12	-1.35037	27.01312	22	7.852454	9.903205	8	5.467397	4.552805	17
SRR21227483	21.336	8.671375	12	-1.59199	26.93178	22	7.917089	9.260684	8	5.486003	4.619196	17
SRR21227484	21.40399	8.603411	12	-1.52394	26.90684	22	7.939187	9.298229	8	5.460661	4.614574	17
SRR21227485	21.24285	8.764605	12	-1.58284	26.95094	22	8.009263	9.367305	8	5.336874	4.459248	17
SRR21227486	21.41238	8.594934	12	-1.46025	26.97784	22	8.005139	9.684529	8	5.417712	4.55278	17
SRR21227490	21.26059	8.74688	12	-1.44693	26.86544	22	7.924409	9.094453	8	5.372822	4.547512	17
SRR21227491	21.22734	8.780247	12	-1.50589	26.90816	22	7.976782	9.136118	8	5.498351	4.660681	17
SRR17173441	4.969776	-7.6418	1	0.394104	-4.25834	7	-8.21592	-3.1872	14	7.534615	3.334345	10
SRR17173442	4.948761	-7.6208	1	0.420784	-4.28502	7	-8.19999	-3.20375	14	7.556494	3.310877	10
SRR17173443	4.902956	-7.57499	1	0.389821	-4.25409	7	-8.26706	-3.13559	14	7.594473	3.247148	10
SRR17173444	4.988381	-7.66044	1	0.315172	-4.18001	7	-8.15284	-3.25057	14	7.606216	3.316677	10
SRR17173445	5.288113	-7.9604	1	0.192016	-4.05613	7	-8.69087	-2.70064	14	7.582378	3.387077	10
SRR17173446	5.202945	-7.87525	1	0.228277	-4.09238	7	-8.74402	-2.64646	14	7.664836	3.528662	10
SRR17173447	5.300179	-7.97252	1	0.288535	-4.15266	7	-8.70921	-2.68113	14	7.486408	3.395819	10
SRR17173448	5.265086	-7.93749	1	0.215786	-4.07998	7	-8.7636	-2.6284	14	7.44052	3.489704	10
SRR17173449	5.152352	-7.82462	1	0.423763	-4.28795	7	-8.40154	-2.99878	14	7.470771	3.388635	10
SRR17173450	5.143446	-7.8157	1	0.508387	-4.37247	7	-8.43598	-2.96224	14	7.542111	3.384253	10
SRR15235946	13.39991	10.26531	19	0.569284	13.1876	25	5.839203	3.75537	4	6.980972	2.977198	24
SRR15235947	13.32749	10.33778	19	0.53009	13.22681	25	5.908519	3.706353	4	6.938027	2.962264	24
SRR15235948	13.34644	10.31882	19	0.547704	13.20833	25	5.957546	3.581319	4	6.885516	2.946222	24
SRR15235949	13.32994	10.33552	19	0.467621	13.28556	25	5.980311	3.675903	4	6.952191	2.97419	24
SRR18050576	13.26872	10.39654	19	0.439177	13.30618	25	5.770713	3.800289	4	6.837607	2.891173	24
SRR16207089	10.71818	-1.27413	6	1.878814	12.09112	13	6.709923	4.91547	4	7.903795	2.904935	10
SRR16207090	10.69721	-1.25165	6	1.97705	12.12155	13	6.580639	4.904693	4	7.912956	2.936701	10
SRR16207092	10.67164	-1.2272	6	1.983082	12.15422	13	6.600247	4.77146	4	7.787107	2.938612	10
SRR23955781	-3.75951	-16.5487	14	-7.72042	-10.212	12	5.067438	2.047196	17	1.907131	4.917967	4
SRR23955789	-3.9388	-16.728	14	-7.64508	-9.94421	12	4.76648	1.96627	17	1.954911	4.937429	4
SRR23955806	-3.34218	-9.78129	8	7.055155	-7.18789	15	-10.9841	14.18021	7	2.659143	3.453356	14
SRR23955805	-3.37253	-9.8117	8	7.15488	-7.08964	15	-11.0254	14.22122	7	2.626746	3.395312	14

SRR23955813	-3.30417	-9.74328	8	7.262506	-6.98179	15	-10.9551	14.15095	7	2.638877	3.428322	14
SRR23955816	-3.69522	-16.4843	14	-7.91495	-10.308	12	5.194617	1.876921	17	1.837025	4.89121	4
SRR23955822	-3.8141	-16.6034	14	-7.68302	-10.0611	12	5.133552	1.634763	17	1.86957	4.89342	4
SRR23955825	-3.74918	-16.5384	14	-7.82566	-9.99851	12	5.192513	1.92623	17	1.886345	4.870108	4
SRR23955830	-3.1789	-9.61807	8	7.162685	-7.08181	15	-10.8664	14.06207	7	2.647689	3.375976	14
SRR23955837	-3.24488	-9.68416	8	7.116128	-7.12866	15	-11.0592	14.25527	7	2.663936	3.453991	14
SRR23955775	-3.94058	-16.7297	14	-7.62763	-9.98845	12	4.882772	2.148351	17	2.093127	4.686366	4
SRR23955776	-3.8045	-16.5936	14	-7.82123	-10.214	12	5.185094	2.234712	17	2.003763	4.770388	4
SRR23955778	-3.82802	-16.6172	14	-7.7231	-10.0784	12	5.09288	1.773568	17	1.874115	4.917404	4
SRR23955794	-3.82808	-16.6174	14	-7.64222	-10.1617	12	4.910828	1.810938	17	1.810465	4.970501	4
SRR23955795	-3.93815	-16.7273	14	-7.66581	-10.0424	12	4.780634	2.033615	17	1.90844	4.830856	4
SRR23955797	-3.29609	-9.73545	8	7.147946	-7.09672	15	-11.171	14.36701	7	2.562358	3.371492	14
SRR23955800	-3.15627	-9.59543	8	7.114368	-7.13035	15	-10.7397	13.93537	7	2.682852	3.47504	14
SRR23955803	-3.2917	-9.73092	8	7.093866	-7.1509	15	-10.7246	13.92031	7	2.705047	3.496586	14
SRR23955814	-3.38122	-9.82055	8	7.107874	-7.13681	15	-10.8325	14.02818	7	2.600952	3.433909	14
SRR23955820	-3.80742	-16.5974	14	-7.71497	-10.1557	12	5.089465	1.975087	17	1.873152	4.907518	4
SRR16208969	19.75728	24.14617	11	-3.08479	23.1765	18	7.746674	10.38931	20	5.699621	5.442085	20
SRR16208975	19.60988	24.29366	11	-2.96854	23.05622	18	7.008456	11.2405	20	5.665463	5.483382	20
SRR16208977	19.7725	24.13105	11	-3.11912	23.22866	18	7.569223	10.69335	20	5.626238	5.603169	20
SRR16208978	19.74891	24.15467	11	-2.84894	23.21088	18	7.455959	10.8244	20	5.640158	5.632399	20
SRR16208979	19.63716	24.26646	11	-2.96785	23.04811	18	7.269738	11.0283	20	5.676983	5.549341	20
SRR16208980	19.64949	24.25415	11	-3.01571	23.08254	18	7.394112	10.8928	20	5.598925	5.578031	20
SRR16208981	19.70575	24.19782	11	-2.97659	23.10443	18	7.590565	10.54337	20	5.604053	5.505495	20
SRR16208982	19.64008	24.26358	11	-2.89849	22.99338	18	7.372336	10.91741	20	5.58023	5.563184	20
SRR16208983	19.75608	24.14756	11	-2.89452	23.19949	18	7.540106	10.58828	20	5.57069	5.56366	20
SRR17736802	15.34162	15.5631	2	-6.00678	6.220913	4	7.776203	-12.7284	13	9.579387	6.91387	11
SRR17736803	15.2609	15.51519	2	-6.19859	6.029937	4	7.405829	-12.3565	13	9.575831	6.894815	11
SRR17736804	15.26887	15.55478	2	-6.1215	6.107069	4	7.288284	-12.2387	13	9.507695	6.807768	11
SRR17736805	15.43913	15.61163	2	-6.08017	6.147718	4	7.848115	-12.7993	13	9.622806	6.80498	11
SRR17736806	15.36165	15.58939	2	-6.03094	6.197148	4	7.793677	-12.7449	13	9.489136	6.893055	11
SRR17736807	15.41935	15.45941	2	-6.14084	6.087613	4	7.642861	-12.594	13	9.532393	6.856345	11
SRR17736808	15.20217	15.45388	2	-6.17138	6.057318	4	7.274833	-12.2254	13	9.525824	6.86845	11
SRR17736809	15.24452	15.49646	2	-6.2218	6.00706	4	7.320283	-12.2712	13	9.611351	6.923341	11
SRR17736810	15.24838	15.57285	2	-6.2288	5.999823	4	7.442292	-12.3925	13	9.499694	6.811144	11
SRR17736811	15.30016	15.49328	2	-6.19065	6.039328	4	7.335969	-12.2866	13	9.514324	6.825947	11

SRR17593559	0.334732	3.579196	3	9.170665	-0.15128	11	3.398829	13.48302	10	5.466286	2.696128	2
SRR17593560	0.40308	3.7095	3	9.131592	0.1914	11	3.364514	13.009	10	5.690191	2.893668	2
SRR17593561	0.756064	4.260757	3	9.465399	0.414202	11	3.852704	12.78407	10	5.724018	2.69895	2
SRR17593562	0.293754	3.724008	3	9.320704	-0.15672	11	3.790184	13.69504	10	5.586194	2.784097	2
SRR17593563	0.345268	3.933583	3	9.450348	-0.10507	11	3.850339	13.64673	10	5.561266	2.768097	2
SRR17593564	0.721753	4.376581	3	9.314738	0.55595	11	3.626742	12.5565	10	5.891824	2.882407	2
SRR17593565	0.763021	4.274097	3	9.349255	0.529886	11	3.573201	12.76475	10	5.908533	2.832623	2
SRR17593566	0.24263	3.689147	3	9.223443	-0.21495	11	3.308518	13.55004	10	5.473203	2.60357	2
SRR17593567	0.750564	4.398139	3	9.295081	0.599124	11	3.558384	12.57568	10	5.972706	2.759115	2
SRR17593568	0.288993	3.66749	3	9.185253	-0.26542	11	3.356457	13.62174	10	5.505105	2.754249	2
SRR17593570	0.669582	4.270456	3	9.589853	0.210672	11	4.173019	13.17753	10	5.790797	2.63678	2
SRR17593571	0.631211	4.245461	3	9.656546	0.240247	11	4.252444	13.19595	10	5.735822	2.715179	2
SRR17593572	0.735637	4.308431	3	9.617016	0.311655	11	4.167586	13.00592	10	5.868514	2.649814	2
SRR17593573	0.298857	3.918277	3	9.580644	-0.04909	11	3.922075	13.56373	10	5.523251	2.706767	2
SRR17593575	0.65025	4.160249	3	9.423165	0.448774	11	3.773843	12.68498	10	5.878316	2.897413	2
SRR17593576	0.347128	3.748337	3	9.217938	-0.11738	11	3.46756	13.57194	10	5.437727	2.760962	2
SRR17593577	0.404546	4.052458	3	9.729597	0.11854	11	4.170084	13.36961	10	5.664636	2.583132	2

Supplementary Table S2. Cluster Assignments and UMAP Coordinates After HKG Normalization

	euclidean			manhattan			canberra			cosine		
sample	UMAP1	UMAP2	cluster	UMAP1	UMAP2	cluster	UMAP1	UMAP2	cluster	UMAP1	UMAP2	cluster
SRR25836907	0.637468	16.09754	16	1.812011	28.32023	11	10.56671	-3.55331	16	3.646353	-9.1546	3
SRR25836909	0.569635	16.17548	16	1.822158	28.24457	11	10.66812	-3.30176	16	3.656344	-9.16454	3
SRR25836910	0.642734	16.14616	16	1.769085	28.27068	11	10.5193	-3.40438	16	3.591561	-9.09933	3
SRR25836912	0.578792	16.16627	16	1.794553	28.34366	11	10.41931	-3.48769	16	3.559168	-9.06692	3
SRR25836913	0.545616	16.08319	16	1.911551	28.30527	11	10.20532	-3.15161	16	3.531273	-9.03731	3
SRR25836914	0.627137	16.11354	16	1.876049	28.278	11	10.42077	-3.45217	16	3.499784	-9.00687	3
SRR25836915	0.569957	16.15528	16	1.962659	28.24541	11	10.11142	-3.15665	16	3.477268	-8.98258	3
SRR25836916	0.538278	16.19414	16	1.872193	28.30437	11	10.31513	-3.32563	16	3.557966	-9.0655	3
SRR25836917	0.653691	16.08367	16	1.765292	28.36298	11	10.55273	-3.58638	16	3.627386	-9.13561	3
SRR25836920	0.610013	16.13116	16	1.77963	28.37953	11	10.46904	-3.54361	16	3.650557	-9.15883	3
SRR26261421	-1.24134	16.89409	26	-2.18979	8.785837	4	8.528756	-1.59374	18	9.75337	1.868585	26
SRR26261422	-1.25441	16.91106	26	-2.16247	8.822983	4	8.328183	-1.06576	18	9.646386	1.963957	26
SRR26261423	-1.24595	16.90555	26	-2.16361	8.820834	4	8.350031	-1.16288	18	9.72878	1.900869	26
SRR26261424	-1.22312	16.88638	26	-2.13544	8.805499	4	8.50085	-1.19011	18	9.728704	1.88694	26
SRR26261425	-1.22162	16.90371	26	-2.17323	8.802882	4	8.457793	-1.22637	18	9.760627	1.870827	26
SRR14150310	0.080122	9.180163	17	6.677066	8.985872	5	8.553106	1.38079	1	7.677742	-1.72311	24
SRR14150311	0.063705	9.207833	17	6.724001	9.078658	5	8.522732	1.402566	1	7.635713	-1.75284	24
SRR14150312	0.052805	9.234087	17	6.808362	9.151969	5	8.29582	1.740899	17	7.70041	-1.7174	24
SRR16230213	-11.3168	4.267149	13	1.214291	-13.4614	17	7.296725	-4.58825	19	-6.05568	-9.09022	9
SRR16230214	-11.3384	4.288833	13	1.108009	-13.5676	17	7.234733	-4.6201	19	-6.08648	-9.12102	9
SRR16230215	-11.2561	4.207628	13	1.248301	-13.4275	17	7.30891	-4.59113	19	-6.02238	-9.05694	9
SRR16230216	-11.3204	4.270855	13	1.181794	-13.494	17	7.32913	-4.62275	19	-6.13982	-9.17438	9
SRR16230217	-11.2455	4.195958	13	1.102022	-13.5738	17	7.125141	-4.65622	19	-6.04698	-9.08146	9
SRR16230218	-11.2085	4.158992	13	1.130723	-13.5451	17	7.144435	-4.69791	19	-6.14904	-9.18351	9
SRR16230219	-11.2846	4.235073	13	1.124739	-13.5512	17	7.359715	-4.72006	19	-6.05111	-9.08563	9
SRR16230220	-11.3343	4.284649	13	1.247473	-13.4285	17	7.416275	-4.80392	19	-6.13075	-9.16528	9
SRR16230221	-11.1792	4.129596	13	1.074878	-13.6008	17	7.005525	-4.48627	19	-6.07218	-9.10667	9
SRR16230222	-11.1383	4.088893	13	0.984594	-13.6909	17	6.963308	-4.55011	19	-5.99227	-9.02678	9
SRR8325947	-0.76822	9.642004	22	9.438238	16.2086	12	12.28591	-0.74651	15	11.91784	-6.10553	18
SRR8325948	-0.7768	9.633431	22	9.414689	16.18725	12	12.14766	-0.87577	15	11.93796	-6.15276	18
SRR8325949	-0.72387	9.677396	22	9.412829	16.18555	12	12.2354	-0.8107	15	11.96031	-6.12816	18

SRR8325950	-0.75036	9.655676	22	9.529046	16.30157	12	12.40718	-0.625	15	11.94709	-6.10244	18
SRR8325951	-0.7046	9.708675	22	9.511692	16.28404	12	12.16138	-0.86513	15	11.92886	-6.11934	18
SRR8325952	-0.79922	9.60979	22	9.420514	16.19323	12	12.29063	-0.74295	15	11.85483	-6.19529	18
SRR8325953	-0.82975	9.583513	22	9.445258	16.21788	12	12.25527	-0.77707	15	11.94959	-6.14733	18
SRR8325954	-0.79906	9.617329	22	9.433876	16.20647	12	12.22377	-0.82004	15	11.90828	-6.13176	18
SRR8325955	-0.39528	9.881922	22	9.280161	16.05553	12	9.816703	-1.18222	24	11.88548	-6.08923	18
SRR8325956	-0.35147	9.909968	22	9.387483	16.16151	12	9.875906	-1.12365	24	11.83364	-6.16154	18
SRR21619385	18.27614	3.482633	6	19.62556	11.84472	8	1.417334	3.584678	10	8.397773	14.21201	22
SRR21619383	18.28423	3.471016	6	19.72063	11.75268	8	1.434913	3.606151	10	8.416972	14.2311	22
SRR21619384	18.25445	3.531898	6	19.64349	11.82874	8	1.474997	3.669228	10	8.368177	14.18238	22
SRR21619382	18.29633	3.459175	6	19.66519	11.80674	8	1.614258	3.789091	10	8.448842	14.26301	22
SRR21619410	18.22287	2.9065	6	18.79702	12.5337	8	1.040467	2.962973	10	9.995771	15.89132	30
SRR21619411	18.17244	2.856779	6	18.83624	12.49665	8	1.017497	3.038077	10	10.01821	15.9147	30
SRR21619412	18.1725	2.872928	6	18.73614	12.59605	8	1.019797	2.996491	10	10.05436	15.9516	30
SRR21619413	18.21623	2.862158	6	18.80475	12.52743	8	0.938933	2.867244	10	10.03599	15.93256	30
SRR21619393	18.31274	3.006286	6	18.65813	12.66859	8	6.332615	-0.65303	20	9.829545	15.71436	10
SRR21619392	18.27683	3.03538	6	18.5401	12.77615	8	6.309485	-0.52762	20	9.862208	15.74463	10
SRR21619391	18.18188	2.856531	6	18.75424	12.57768	8	1.153846	2.932258	10	10.01812	15.91518	30
SRR21619390	18.21651	2.899299	6	18.72697	12.60278	8	1.060967	2.966323	10	9.962137	15.85781	30
SRR21619381	18.32014	3.450461	6	19.50028	11.96659	8	1.313718	3.471631	10	8.405896	14.22034	22
SRR21619380	18.30059	3.483675	6	19.60641	11.86251	8	1.479589	3.665086	10	8.481994	14.29652	22
SRR21619379	18.23303	3.565309	6	19.53562	11.93421	8	1.484063	3.657326	10	8.338136	14.1524	22
SRR21619378	18.36099	3.493832	6	19.43134	12.03223	8	1.267674	3.43354	10	8.431659	14.24609	22
SRR21619414	18.30362	3.021566	6	18.8442	12.49423	8	6.434876	-0.67579	20	9.809333	15.69484	10
SRR25654709	-0.74203	11.33296	1	6.952032	9.75372	5	7.781986	2.152915	17	-8.80665	1.442443	16
SRR25654705	-0.73094	11.40288	1	6.996987	9.754171	5	7.599364	2.20378	17	-8.78513	1.42087	16
SRR25654706	-0.67621	11.36552	1	6.997608	9.740746	5	7.670281	2.281662	17	-8.77246	1.408324	16
SRR25654707	-0.6926	11.42441	1	6.92189	9.823341	5	7.666697	2.341178	17	-8.76271	1.398739	16
SRR25654708	-0.76392	11.35316	1	6.97634	9.821857	5	7.756031	2.250938	17	-8.79996	1.435787	16
SRR25654710	-0.70234	11.41225	1	6.884069	9.761442	5	7.772229	2.1789	17	-8.72762	1.363365	16
SRR21609197	-8.35605	20.37713	10	10.84896	-1.38925	15	7.424493	-3.92144	21	1.336235	3.084619	11
SRR21609198	-8.42563	20.44669	10	10.67179	-1.21301	15	7.477462	-3.49797	21	1.42406	2.997963	11
SRR21609201	-8.43465	20.45586	10	10.81876	-1.35788	15	7.575133	-3.70294	21	1.396876	3.024321	11
SRR21609202	-8.37406	20.39586	10	10.70507	-1.24621	15	7.440462	-3.61917	21	1.40959	3.01173	11
SRR21609204	-8.4831	20.50428	10	10.72373	-1.2646	15	7.602212	-3.64919	21	1.505557	2.916554	11

SRR21609205	-8.4166	20.43777	10	10.64354	-1.18464	15	7.556542	-3.60795	21	1.417226	3.005342	11
SRR21609203	-8.44694	20.46814	10	10.69216	-1.2333	15	7.553494	-3.62818	21	1.367957	3.052975	11
SRR21609200	-8.36995	20.39118	10	10.7491	-1.2904	15	7.579747	-3.68393	21	1.363047	3.057795	11
SRR21609199	-8.48014	20.50133	10	10.60221	-1.14302	15	7.821588	-3.60858	21	1.435731	2.986032	11
SRR21609206	-8.36443	20.3856	10	10.84266	-1.38335	15	7.535103	-3.98658	21	1.370512	3.050485	11
SRR19732846	-1.13609	11.76761	21	6.470648	9.100002	5	8.536029	0.533097	1	9.191383	1.824324	31
SRR19732849	-1.15996	11.77905	21	6.392434	8.95553	5	7.890655	0.066733	20	9.167458	1.915628	31
SRR19732852	-1.37135	11.89896	21	-5.14121	8.520024	18	6.785324	0.493417	9	9.170181	1.948288	31
SRR23226403	-0.10042	10.439	25	7.056944	10.02484	5	7.984636	2.41371	17	11.49454	-2.73346	20
SRR23226404	-0.09709	10.44115	25	7.072964	10.02266	5	7.987478	2.325709	17	11.50063	-2.72402	20
SRR23226405	-0.12579	10.45504	25	7.06685	9.999729	5	8.084921	2.300368	17	11.47487	-2.73773	20
SRR13571325	8.754931	11.84422	15	21.07278	-2.92645	7	10.48631	-2.54353	22	7.23895	-12.8497	17
SRR13571327	8.744152	11.83355	15	21.2743	-3.12795	7	10.61661	-2.4228	22	7.153835	-12.9347	17
SRR13571331	8.757384	11.84655	15	21.14416	-2.99784	7	10.42466	-2.67157	22	7.195213	-12.8935	17
SRR13571339	8.807775	11.89742	15	21.39028	-3.24412	7	10.68319	-2.29656	22	7.161216	-12.9277	17
SRR13571349	8.677598	11.76819	15	21.34812	-3.20206	7	10.7682	-2.20695	22	7.107397	-12.9812	17
SRR13571321	8.767246	11.85642	15	21.03333	-2.8871	7	10.21596	-2.88317	16	7.298301	-12.7903	17
SRR13571347	8.676387	11.76656	15	21.25753	-3.11136	7	10.65006	-2.3553	22	7.127858	-12.9609	17
SRR13571355	8.759578	11.84886	15	21.12157	-2.97514	7	10.60271	-2.46517	22	7.109374	-12.9795	17
SRR13571357	8.699534	11.78913	15	21.30952	-3.16326	7	10.62825	-2.24071	22	7.153054	-12.9355	17
SRR13571359	8.801703	11.89103	15	21.12668	-2.98034	7	10.55596	-2.5019	22	7.094762	-12.994	17
SRR15021920	-0.39728	11.25211	1	6.475349	9.418926	5	7.964628	1.821363	17	11.52162	-1.83835	29
SRR15021921	-0.37638	11.22844	1	6.422394	9.316325	5	7.929108	1.578754	17	11.45774	-1.728	29
SRR15021922	-0.35732	11.24651	1	6.421695	9.378596	5	8.060658	1.740685	17	11.50059	-1.80932	29
SRR25606246	12.75564	14.13886	5	-5.61163	14.2194	10	3.289673	-5.85232	8	1.908805	9.462292	23
SRR25606247	12.68845	14.12817	5	-5.5775	14.26994	10	3.275057	-5.86589	8	1.991458	9.377451	23
SRR25606248	12.7086	14.04259	5	-5.70131	14.31055	10	3.240089	-5.89066	8	1.878335	9.49018	23
SRR25606249	12.74908	14.11294	5	-5.56574	14.21972	10	3.176271	-5.93677	8	2.026829	9.359171	23
SRR25606250	12.76769	14.07148	5	-5.63422	14.15181	10	3.146663	-6.0205	8	1.889793	9.481037	23
SRR25606252	12.81095	14.06163	5	-5.62478	14.2683	10	3.209173	-5.92299	8	1.964226	9.396893	23
SRR25606253	12.75791	14.0741	5	-5.6325	14.20265	10	3.296921	-5.84274	8	1.940072	9.428735	23
SRR25606254	12.63765	14.18378	5	-5.60699	14.25543	10	3.396812	-5.74836	8	2.009627	9.34809	23
SRR25606255	12.8472	14.15261	5	-5.50373	14.12355	10	3.094923	-6.02574	8	1.880255	9.48827	23
SRR25606256	12.54798	14.29579	5	-5.70353	14.32392	10	3.347238	-5.79708	8	2.002518	9.352263	23
SRR15334651	13.41284	14.16198	5	-4.99995	14.84546	10	4.043541	-5.32628	25	2.342507	8.727723	6

SRR15334652	13.43622	14.1729	5	-5.01506	14.82815	10	4.014353	-5.32441	25	2.301291	8.76505	6
SRR15334653	13.40025	14.15099	5	-5.02967	14.80961	10	3.977538	-5.37533	25	2.341349	8.723169	6
SRR15334654	13.42586	14.17525	5	-5.01878	14.82562	10	3.982668	-5.37502	25	2.309441	8.736601	6
SRR15334655	13.41126	14.18498	5	-5.03922	14.80649	10	3.914452	-5.42567	25	2.334527	8.7298	6
SRR18520163	3.556009	2.607624	8	-7.28705	-8.83333	9	18.82503	6.472229	3	-2.54093	11.01014	15
SRR18520164	3.552758	2.610938	8	-7.36629	-8.75395	9	18.73563	6.420202	3	-2.5377	11.00693	15
SRR18520165	3.542678	2.621013	8	-7.29881	-8.82134	9	18.74202	6.425176	3	-2.4591	10.92836	15
SRR18520166	3.50635	2.657334	8	-7.24183	-8.87851	9	18.66815	6.33212	3	-2.52439	10.99353	15
SRR18520167	3.619898	2.543913	8	-7.40898	-8.71141	9	18.86431	6.517338	3	-2.61139	11.08049	15
SRR18520168	3.642518	2.521232	8	-7.4599	-8.6605	9	18.93074	6.370324	3	-2.44408	10.91327	15
SRR18520169	3.498315	2.665395	8	-7.43798	-8.6826	9	18.7226	6.36402	3	-2.50917	10.97842	15
SRR18520170	3.50471	2.658924	8	-7.61396	-8.50639	9	18.92545	6.582908	3	-2.51944	10.98861	15
SRR18520171	3.562659	2.601507	8	-7.35556	-8.76469	9	18.65254	6.313722	3	-2.57633	11.0455	15
SRR18520172	3.659265	2.504484	8	-7.40183	-8.71841	9	18.75918	6.400988	3	-2.63181	11.10102	15
SRR9888136	-1.97076	16.70064	3	-1.87534	9.24353	4	9.474889	-2.16296	6	13.62229	-3.52913	7
SRR9888137	-2.0859	16.79267	3	-1.75097	9.343451	4	8.716196	-1.9691	6	13.53315	-3.5843	7
SRR9888138	-2.00562	16.74407	3	-1.9036	9.184374	4	9.540354	-2.28393	6	13.67095	-3.52355	7
SRR9888139	-2.08751	16.79508	3	-1.69194	9.436118	4	8.706652	-2.07744	6	13.55483	-3.59113	7
SRR9888140	-2.02904	16.76555	3	-1.8147	9.287869	4	9.450339	-2.22973	6	13.60364	-3.5954	7
SRR9888141	-1.99315	16.73164	3	-1.89488	9.211009	4	9.422419	-2.17359	6	13.6241	-3.57567	7
SRR9888142	-1.99566	16.73878	3	-1.75598	9.396833	4	9.447186	-2.1919	6	13.62777	-3.52673	7
SRR9888143	-1.95548	16.8641	3	-1.89925	9.205341	4	9.513372	-2.29192	6	13.65469	-3.54084	7
SRR15254306	14.42824	-5.18649	0	1.214707	18.17996	0	-0.18289	0.196449	0	-5.27192	3.674983	32
SRR15254307	14.44381	-5.15696	0	1.252149	18.18744	0	-0.16266	0.264429	0	-5.27017	3.606445	32
SRR15254308	14.45975	-5.12298	0	1.265471	18.20411	0	-0.26878	0.23195	0	-5.27709	3.621337	32
SRR15254309	15.2907	-5.12135	24	-1.0667	19.13785	0	-2.42538	0.780026	13	-5.4763	3.112664	25
SRR15254310	15.28715	-5.12018	24	-1.09448	19.13058	0	-2.51896	0.62144	13	-5.47343	3.114584	25
SRR15254311	15.25154	-5.09201	24	-1.01328	19.08586	0	-2.3867	0.710306	13	-5.42929	3.178314	25
SRR15254312	15.22773	-5.08527	24	-0.97582	19.04339	0	-2.12164	0.620832	13	-5.4219	3.207871	25
SRR15254313	15.22029	-5.11084	24	-0.96758	19.02337	0	-2.2472	0.701148	13	-5.45908	3.166507	25
SRR15254314	15.26431	-5.14234	24	-0.96624	19.01024	0	-2.23746	0.758927	13	-5.44533	3.268119	25
SRR15254315	15.90896	-5.3106	19	-0.55093	18.8857	0	-2.80935	0.746993	13	-5.6896	2.45437	21
SRR15254333	13.96339	-5.25568	0	0.903008	18.01582	0	-0.3965	0.163549	0	-5.0439	4.133886	0
SRR15254334	13.96524	-5.24216	0	0.859054	18.03865	0	-0.40825	0.10249	0	-5.0786	4.099321	0
SRR15254335	13.97391	-5.22394	0	1.042574	18.06551	0	-0.36546	0.14864	0	-5.05432	4.127941	0

SRR15254336	13.72803	-5.44412	0	0.536715	18.33134	0	-0.70977	0.2541	0	-4.98825	4.175907	0
SRR15254337	13.73398	-5.43926	0	0.378831	18.47471	0	-0.84311	0.320061	0	-4.91916	4.243823	0
SRR15254338	13.73162	-5.43697	0	0.466746	18.4008	0	-0.75961	0.347216	0	-4.98044	4.186297	0
SRR15254339	14.0157	-5.19225	0	0.922393	17.99791	0	-0.33051	0.407661	0	-5.02178	4.152823	0
SRR15254340	13.98842	-5.23001	0	0.946438	17.92615	0	-0.28122	0.32574	0	-5.0515	4.129317	0
SRR15254341	13.97591	-5.25087	0	1.001138	18.00647	0	-0.34582	0.31058	0	-5.00672	4.172944	0
SRR15254325	16.10933	-5.30543	19	-0.33954	18.89396	0	-2.72924	0.676193	13	-5.76048	2.368534	21
SRR15254326	16.1284	-5.33662	19	-0.36307	18.92556	0	-2.8213	0.806239	13	-5.76421	2.365747	21
SRR15254327	16.0743	-5.31079	19	-0.31293	18.84008	0	-2.63727	1.092436	13	-5.76457	2.360428	21
SRR15254328	16.16013	-5.36918	19	-0.38997	18.88075	0	-2.73173	0.964239	13	-5.7154	2.412143	21
SRR15254329	16.19464	-5.38375	19	-0.45921	18.93589	0	-2.81308	0.853486	13	-5.65119	2.474787	21
SRR15254330	16.14499	-5.34975	19	-0.42284	18.97171	0	-2.86741	0.80575	13	-5.65765	2.469174	21
SRR15254331	16.01535	-5.22166	19	-0.52483	19.1409	0	-2.8969	0.774756	13	-5.62271	2.508457	21
SRR15254332	16.1052	-5.31247	19	-0.44487	18.9793	0	-2.90614	0.766105	13	-5.75847	2.372785	21
SRR15254324	16.12567	-5.31281	19	-0.26205	18.80072	0	-2.97816	0.717426	13	-5.78169	2.345574	21
SRR15835822	-1.67945	12.01538	21	-5.03727	8.627539	18	6.934812	0.708576	9	9.119469	2.317412	14
SRR15835823	-1.69355	12.02785	21	-5.03046	8.634398	18	6.811591	0.780338	9	9.088136	2.370492	14
SRR15835826	-1.73913	12.0756	21	-5.06121	8.604692	18	6.809488	0.835867	9	9.09483	2.351589	14
SRR15835827	-1.71813	12.05301	21	-5.10864	8.557354	18	6.910833	0.886399	9	9.099861	2.354806	14
SRR15835828	-1.73806	12.07027	21	-5.08043	8.58299	18	6.743074	0.94395	9	9.072486	2.352847	14
SRR15835829	-1.74794	12.07829	21	-5.01471	8.651657	18	6.799202	0.910994	9	9.090601	2.387657	14
SRR25223472	7.669239	-0.74358	12	-14.0796	-2.43292	3	6.255251	-3.72191	5	11.52139	-0.35259	1
SRR25223473	7.716647	-0.79104	12	-14.1342	-2.37845	3	6.225082	-3.72936	5	11.55413	-0.40517	1
SRR25223476	7.661234	-0.73562	12	-14.1554	-2.35731	3	6.152584	-3.77442	5	11.61443	-0.39448	1
SRR25223478	7.686163	-0.76108	12	-14.3121	-2.20052	3	6.473179	-3.65091	5	11.61013	-0.3344	1
SRR25223479	7.674119	-0.74853	12	-14.3493	-2.16277	3	6.533161	-3.5771	5	11.57201	-0.3069	1
SRR25223480	7.586679	-0.6611	12	-14.1098	-2.40293	3	6.478398	-3.59549	5	11.58619	-0.33869	1
SRR25223481	7.63932	-0.71378	12	-14.2561	-2.25652	3	6.483409	-3.62659	5	11.59339	-0.33525	1
SRR25223483	7.783178	-0.8576	12	-14.2367	-2.27594	3	6.109246	-3.83313	5	11.63754	-0.38083	1
SRR25223484	7.707359	-0.78174	12	-14.0876	-2.42509	3	6.16993	-3.76676	5	11.53377	-0.40926	1
SRR25223485	7.637456	-0.71195	12	-14.2639	-2.24891	3	6.302444	-3.66548	5	11.58436	-0.34063	1
SRR16093329	0.113018	8.213095	23	6.61147	8.533154	5	8.933238	0.989141	1	7.483215	-1.87482	24
SRR16093333	0.11273	8.219997	23	6.602382	8.420519	5	8.682083	0.918529	1	7.503598	-1.85746	24
SRR16093337	0.125791	8.235179	23	6.560963	8.47361	5	8.834085	0.974745	1	7.419766	-1.93478	24
SRR16093341	0.126591	8.216094	23	6.528762	8.366819	5	8.80757	0.854512	1	7.478918	-1.88153	24

SRR16093345	0.09524	8.236283	23	6.65837	8.403739	5	8.853895	0.943304	1	7.44726	-1.9117	24
SRR21227478	-15.2047	16.75205	7	3.798462	-11.9255	1	9.88528	8.19749	12	18.78633	-7.44592	12
SRR21227479	-15.2499	16.79716	7	3.657292	-12.0713	1	9.983598	8.307985	12	18.74927	-7.40889	12
SRR21227482	-15.2285	16.77584	7	3.800244	-11.9951	1	9.803969	8.116013	12	18.82315	-7.4828	12
SRR21227483	-15.1227	16.66988	7	3.737008	-11.9663	1	9.826625	8.148629	12	18.75526	-7.41541	12
SRR21227484	-15.2185	16.7658	7	3.671358	-12.056	1	9.688387	8.003021	12	18.77259	-7.43219	12
SRR21227485	-15.1868	16.73407	7	3.622974	-12.0038	1	9.998163	8.330504	12	18.65974	-7.31934	12
SRR21227486	-15.159	16.70629	7	3.766765	-12.0066	1	10.01914	8.344889	12	18.69852	-7.35809	12
SRR21227490	-15.209	16.75623	7	3.706069	-12.0403	1	9.840044	8.155994	12	18.82264	-7.48235	12
SRR21227491	-15.2648	16.81203	7	3.569234	-12.1458	1	9.718343	8.03124	12	18.81109	-7.47092	12
SRR17173441	-1.9178	-10.0063	4	-14.6842	8.603637	6	4.747981	4.643874	7	14.46729	5.625091	13
SRR17173442	-1.91226	-10.0008	4	-14.7043	8.583533	6	4.630002	4.771717	7	14.50233	5.66021	13
SRR17173443	-1.91532	-10.0039	4	-14.7229	8.564938	6	4.687711	4.73744	7	14.45277	5.610543	13
SRR17173444	-1.9053	-9.9939	4	-14.75	8.537885	6	4.870629	4.706231	7	14.4661	5.623652	13
SRR17173445	-1.70465	-9.79365	4	-14.494	8.794047	6	4.770478	4.670307	7	14.63172	5.789424	13
SRR17173446	-1.71805	-9.80707	4	-14.4804	8.807701	6	4.698091	4.792667	7	14.62641	5.784002	13
SRR17173447	-1.6978	-9.78683	4	-14.5696	8.71843	6	4.719481	4.806682	7	14.66079	5.818419	13
SRR17173448	-1.74211	-9.83109	4	-14.4676	8.820507	6	4.782838	4.786739	7	14.5884	5.745925	13
SRR17173449	-1.81882	-9.90767	4	-14.5279	8.760148	6	4.584483	4.402066	7	14.43982	5.594685	13
SRR17173450	-1.78288	-9.87174	4	-14.5023	8.785704	6	4.564156	4.437564	7	14.42484	5.581862	13
SRR15235946	3.544336	6.913195	14	6.642362	7.591155	5	9.317066	-0.06886	23	5.953834	5.273391	19
SRR15235947	3.575821	6.882396	14	6.616561	7.538298	5	9.254237	-0.1019	23	5.975318	5.25064	19
SRR15235948	3.565091	6.893329	14	6.68569	7.545242	5	9.309664	-0.01195	23	5.977237	5.248244	19
SRR15235949	3.547173	6.910427	14	6.665017	7.55795	5	9.208066	0.001004	23	5.975352	5.250588	19
SRR18050576	3.554002	6.904496	14	6.638898	7.467313	5	9.287005	-0.61172	23	5.975233	5.250767	19
SRR16207089	-1.85226	16.52478	3	6.358807	9.243958	5	7.490559	1.14103	9	13.96381	-3.62236	7
SRR16207090	-1.84472	16.50944	3	6.307185	9.268138	5	7.36618	1.071309	9	13.99966	-3.64591	7
SRR16207092	-1.83224	16.48907	3	6.342004	9.283494	5	7.230716	1.226624	9	13.95287	-3.57671	7
SRR23955781	-6.3019	3.682804	9	-3.5388	-10.7707	16	11.12246	2.358302	14	22.54866	13.28355	2
SRR23955789	-6.40473	3.786176	9	-3.38312	-10.6152	16	11.32995	2.609954	14	22.38913	13.12398	2
SRR23955806	-10.5481	-0.97963	2	-6.58166	-0.79498	14	8.725196	4.827238	4	13.82205	13.03301	4
SRR23955805	-10.4927	-1.03556	2	-6.54057	-0.836	14	8.739986	4.864262	4	13.84561	13.06161	4
SRR23955813	-10.4742	-1.05396	2	-6.55403	-0.82257	14	8.684788	4.747002	4	13.84067	13.0565	4
SRR23955816	-6.29584	3.676723	9	-3.47458	-10.7066	16	11.22123	2.372685	14	22.5035	13.23837	2
SRR23955822	-6.4246	3.805931	9	-3.43084	-10.6631	16	11.06941	2.260061	14	22.51069	13.24556	2

SRR23955825	-6.42486	3.806151	9	-3.59822	-10.8308	16	11.13814	2.275592	14	22.57947	13.3144	2
SRR23955830	-10.5617	-0.96662	2	-6.71756	-0.65905	14	8.797682	4.856172	4	13.77466	12.99085	4
SRR23955837	-10.6292	-0.89993	2	-6.66121	-0.7155	14	8.86536	4.685067	4	13.81548	13.03168	4
SRR23955775	-6.39044	3.773273	9	-3.44215	-10.6734	16	11.3417	2.559803	14	22.37053	13.10545	2
SRR23955776	-6.30014	3.680966	9	-3.36989	-10.6012	16	11.1742	2.462322	14	22.59184	13.32672	2
SRR23955778	-6.36787	3.749158	9	-3.65822	-10.8903	16	11.19935	2.595298	14	22.44686	13.18172	2
SRR23955794	-6.46446	3.845758	9	-3.4	-10.6316	16	11.00796	2.442489	14	22.48603	13.2209	2
SRR23955795	-6.43974	3.821139	9	-3.33022	-10.5615	16	11.29507	2.557022	14	22.37058	13.10546	2
SRR23955797	-10.5306	-0.9978	2	-6.58377	-0.79298	14	8.781322	4.764106	4	13.80709	13.02365	4
SRR23955800	-10.5911	-0.93743	2	-6.7705	-0.60611	14	8.893433	4.956779	4	13.67546	12.89225	4
SRR23955803	-10.5243	-1.00407	2	-6.78638	-0.59028	14	8.885008	4.989218	4	13.7364	12.95342	4
SRR23955814	-10.5064	-1.02215	2	-6.70183	-0.67474	14	8.76199	4.82075	4	13.771	12.98676	4
SRR23955820	-6.43487	3.815887	9	-3.38384	-10.6154	16	11.29212	2.515456	14	22.45105	13.1859	2
SRR16208969	0.877268	10.10999	20	-1.65898	9.707197	4	8.854502	-1.97345	6	11.58317	-2.38642	20
SRR16208975	1.396259	10.07228	20	-0.34452	10.07628	4	7.245726	0.093267	20	12.13141	-2.20339	27
SRR16208977	1.459616	10.01247	20	-0.35898	10.07832	4	7.278282	0.099408	20	12.09637	-2.24123	27
SRR16208978	0.939412	10.14045	20	-1.58702	9.776353	4	9.27148	-1.65151	24	11.62891	-2.3805	20
SRR16208979	1.055304	10.15219	20	-1.51049	9.796628	4	9.245602	-1.37436	24	11.69288	-2.34671	20
SRR16208980	0.952221	10.14222	20	-1.53552	9.8224	4	9.25557	-1.60606	24	11.60573	-2.30424	20
SRR16208981	1.460554	10.01096	20	-0.37604	10.0632	4	7.389841	-0.03824	20	12.09903	-2.23766	27
SRR16208982	1.420917	10.04177	20	-0.37708	10.102	4	7.297705	0.041699	20	12.09615	-2.2438	27
SRR16208983	1.470981	10.03894	20	-0.39263	10.1073	4	7.2935	0.033947	20	12.15581	-2.17869	27
SRR17736802	1.529948	12.75285	18	16.05868	-0.4619	13	9.074331	-5.45772	2	27.70635	1.49834	5
SRR17736803	1.451994	12.67083	18	15.86629	-0.65628	13	9.288689	-5.27422	2	27.80065	1.592739	5
SRR17736804	1.4359	12.65475	18	15.91852	-0.60416	13	9.400973	-5.16676	2	27.84098	1.63305	5
SRR17736805	1.528296	12.74968	18	16.11961	-0.4001	13	9.134615	-5.403	2	27.78385	1.57591	5
SRR17736806	1.531316	12.75431	18	16.13835	-0.38253	13	9.227694	-5.27558	2	27.6601	1.451988	5
SRR17736807	1.367978	12.58539	18	15.97908	-0.5432	13	9.092935	-5.38747	2	27.78516	1.577193	5
SRR17736808	1.382092	12.59931	18	15.82886	-0.69395	13	9.238647	-5.03623	2	27.82352	1.615541	5
SRR17736809	1.428055	12.64655	18	15.86383	-0.65906	13	9.308284	-5.18755	2	27.76318	1.555202	5
SRR17736810	1.364759	12.58317	18	15.96659	-0.55647	13	9.37792	-5.20445	2	27.82004	1.612113	5
SRR17736811	1.455284	12.67403	18	16.02588	-0.49673	13	9.347688	-5.20789	2	27.8752	1.667212	5
SRR17593559	-0.23992	6.200316	11	15.21721	4.607532	2	5.069645	-0.63607	11	6.351726	-2.64116	28
SRR17593560	-0.20927	6.1286	11	15.38308	4.467568	2	5.198878	-0.56657	11	6.48185	-2.59691	28
SRR17593561	-0.1978	5.745361	11	15.83705	4.180102	2	5.549867	-0.92797	11	6.878805	-2.49109	8

SRR17593562	-0.26348	6.177592	11	15.38339	4.454138	2	5.001977	-0.80698	11	6.43268	-2.56482	28
SRR17593563	-0.3856	5.997238	11	15.4072	4.467914	2	5.074184	-0.95748	11	6.397688	-2.3919	28
SRR17593564	-0.07354	5.694269	11	16.04015	3.970902	2	5.821395	-0.78237	11	6.994854	-2.42059	8
SRR17593565	-0.03658	5.658595	11	15.91294	4.105467	2	5.715765	-0.9549	11	6.95742	-2.44083	8
SRR17593566	-0.23022	6.218947	11	15.21647	4.573615	2	5.019648	-0.67679	11	6.346985	-2.62507	28
SRR17593567	-0.00301	5.628704	11	16.033	3.978087	2	5.75765	-0.96966	11	6.967976	-2.43245	8
SRR17593568	-0.25998	6.197285	11	15.21921	4.607998	2	5.041412	-0.64352	11	6.408865	-2.56359	28
SRR17593570	-0.29799	5.750539	11	15.82051	4.170051	2	5.35812	-1.01114	11	6.850234	-2.48831	8
SRR17593571	-0.22975	5.728715	11	15.83074	4.224083	2	5.440757	-1.05213	11	6.825265	-2.50936	8
SRR17593572	-0.2456	5.735733	11	15.81167	4.193029	2	5.3622	-1.10485	11	6.837284	-2.49173	8
SRR17593573	-0.40778	6.033335	11	15.40228	4.474672	2	5.155973	-1.08292	11	6.488333	-2.48606	28
SRR17593575	0.035735	5.684929	11	16.15873	3.850916	2	6.525689	-0.33197	20	7.026688	-2.42675	8
SRR17593576	-0.29914	6.18805	11	15.24346	4.620036	2	5.034178	-0.65544	11	6.371019	-2.64943	28
SRR17593577	-0.45344	5.842089	11	15.69799	4.313965	2	5.266088	-1.10585	11	6.77712	-2.47004	8

Supplementary Table S3. Cluster assignments and clustering-quality scores before HKG normalization

Euclidean						Manhattan					
cluster	disease	n_in_th is_clust er	group_t otal	frac_of _group	clusteri ng- quality	cluster	disease	n_in_th is_clust er	group_t otal	frac_of _group	clusteri ng- quality
0	CF	9	9	1	1	0	CF	6	9	0.66666	0.1
0	NASH	3	10	0.3		0.487	0	NASH	3	10	0.3
1	ASS-	10	10	1	1	1	BLCA	10	10	1	0.9
2	ODU	10	10	1	1	2	ACLF	4	4	1	0.9
3	ATC	10	10	1	1	2	sAH	4	4	1	0.9
3	DTC	7	7	1	1	3	DCM	6	6	1	0.9
4	CAD	10	10	1	1	4	ODU	10	10	1	0.9
5	FAD	4	8	0.5	0.6	5	FAD	8	8	1	0.9
5	SAD	5	9	0.55555	0.25	5	SAD	9	9	1	0.9
6	ACLF	4	4	1	1	6	IPF	10	10	1	0.9
6	TB	3	3	1	1	7	ASS-	10	10	1	0.9
6	sAH	4	4	1	1	8	CKD	10	10	1	0.9
7	DCM	6	6	1	1	9	CAD	10	10	1	0.9
8	CD	5	10	0.5	0.6	10	AS	10	15	0.66666	0.25
8	UC	4	10	0.4	0.67	11	ATC	10	10	1	0.9
9	CeAD	5	5	1	1	11	DTC	7	7	1	0.9
10	BLCA	10	10	1	1	12	CD	5	10	0.5	0.6
11	AML	9	9	1	1	12	UC	6	10	0.6	0.67
12	BTC	9	9	1	1	13	GAC	2	3	0.66666	0.1
13	HIV	10	10	1	1	13	PCa	1	3	0.33333	0.1
14	CD	5	10	0.5	0.6	13	TB	3	3	1	0.9
14	UC	6	10	0.6	0.67	14	LAC	10	10	1	0.9
15	CKD	10	10	1	1	15	CD	5	10	0.5	0.6
16	MVID	6	6	1	1	15	UC	4	10	0.4	0.67
17	NASH	6	10	0.6	0.487	16	HIV	10	10	1	0.9
18	ESCC	10	10	1	1	17	ESCC	10	10	1	0.9
19	MS	5	5	1	1	18	AML	9	9	1	0.9
20	COPD	3	3	1	1	19	GAC	1	3	0.33333	0.1
20	HGSO	5	5	1	1	19	NASH	5	10	0.5	0.4

21	IPF	10	10	1	1	20	COPD	3	3	1	0.9
22	AS	10	15	0.66666	0.25	20	HGSO	5	5	1	0.9
23	GAC	3	3	1	1	21	BCA	3	3	1	0.9
24	AC	9	9	1	1	21	MVID	6	6	1	0.9
24	NASH	1	10	0.1	0.487	21	PCa	2	3	0.66666	0.1
25	FAD	4	8	0.5	0.6	22	BTC	9	9	1	0.9
25	SAD	4	9	0.44444	0.25	23	AS	5	15	0.33333	0.25
26	LAC	10	10	1	1	24	AC	9	9	1	0.9
27	AS	5	15	0.33333	0.25	24	NASH	1	10	0.1	0.4
28	BCA	3	3	1	1	25	MS	5	5	1	0.9
28	PCa	3	3	1	1	26	CeAD	5	5	1	0.9
						27	CF	3	9	0.33333	0.1
						27	NASH	1	10	0.1	0.4
Canberra					Cosine						
cluster	disease	n_in_th is_clust er	group_t otal	frac_of _group	clusteri ng- quality	cluster	disease	n_in_th is_clust er	group_t otal	frac_of _group	clusteri ng- quality
0	CF	9	9	1	0.9	0	CF	9	9	1	0.9
0	NASH	3	10	0.3	0.59	0	NASH	5	10	0.5	0.45
1	AS	10	15	0.66666	0.25	1	BLCA	10	10	1	0.9
1	BCA	3	3	1	0.9	2	ATC	10	10	1	0.9
2	ESCC	10	10	1	0.9	2	DTC	7	7	1	0.9
2	sAH	2	4	0.5	0.6	3	HIV	10	10	1	0.9
3	IPF	10	10	1	0.9	4	CD	5	10	0.5	0.6
4	COPD	3	3	1	0.9	4	UC	6	10	0.6	0.67
4	HGSO	5	5	1	0.9	5	ACLF	4	4	1	0.9
4	MS	5	5	1	0.9	5	sAH	4	4	1	0.9
4	PCa	3	3	1	0.9	6	IPF	10	10	1	0.9
4	TB	3	3	1	0.9	7	DCM	6	6	1	0.9
5	LAC	10	10	1	0.9	8	FAD	8	8	1	0.9
6	CKD	10	10	1	0.9	8	SAD	9	9	1	0.9
7	CD	5	10	0.5	0.45	9	CAD	10	10	1	0.9
7	UC	4	10	0.4	0.52	10	ASS-	10	10	1	0.9
8	BTC	9	9	1	0.9	10	TB	3	3	1	0.9
8	CeAD	5	5	1	0.9	11	OULD	10	10	1	0.9

9	DCM	6	6	1	0.9	12	CKD	10	10	1	0.9
10	ATC	10	10	1	0.9	13	MVID	6	6	1	0.9
10	DTC	7	7	1	0.9	14	CD	5	10	0.5	0.6
10	SAD	3	9	0.33333	0.1	14	UC	4	10	0.4	0.67
11	FAD	8	8	1	0.9	15	AC	9	9	1	0.9
11	SAD	6	9	0.66666	0.1	15	NASH	5	10	0.5	0.45
12	AC	9	9	1	0.9	16	ESCC	10	10	1	0.9
12	NASH	7	10	0.7	0.59	17	BTC	9	9	1	0.9
13	ODU	10	10	1	0.9	18	LAC	10	10	1	0.9
14	ASS-	10	10	1	0.9	19	COPD	3	3	1	0.9
15	CAD	10	10	1	0.9	19	HGSO	5	5	1	0.9
15	HIV	10	10	1	0.9	20	AML	9	9	1	0.9
16	MVID	6	6	1	0.9	21	AS	15	15	1	0.9
17	CD	5	10	0.5	0.45	22	PCa	3	3	1	0.9
17	GAC	3	3	1	0.9	23	CeAD	5	5	1	0.9
17	UC	6	10	0.6	0.52	23	GAC	3	3	1	0.9
18	ACLF	4	4	1	0.9	24	MS	5	5	1	0.9
18	sAH	2	4	0.5	0.6	25	BCA	3	3	1	0.9
19	BLCA	10	10	1	0.9						
20	AML	9	9	1	0.9						
21	AS	5	15	0.33333	0.25						

Color-coded cluster assignments for all diseases before HKG normalization, together with the corresponding clustering-quality score for each disease–cluster mapping.

Colors indicate qualitative coherence categories used throughout this study: Red — all samples of a disease assigned to a single cluster with complete retention; Green — a disease split across multiple clusters but consistently co-occurring with a small and compositionally stable set of diseases; Yellow — clusters with minor impurity, in which most samples remain together but a small number of unrelated diseases co-occur; Blue — fragmented diseases distributed across multiple unrelated clusters; Gray — a single mixed cluster containing multiple diseases without a clear internal structure.

Supplementary Table S4. Cluster assignments and clustering-quality scores after HKG normalization

Euclidean						Manhattan						
cluster	disease	n_in_th is_clust er	group_t otal	frac_of _group	clusteri ng- quality	cluster	disease	n_in_th is_clust er	group_t otal	frac_of _group	clusteri ng- quality	
0	CF	9	9	1	1 0.487	0	AC	9	9	1	0.9	
0	NASH	3	10	0.3		0	CF	9	9	1	0.9	
1	MVID	6	6	1		0	NASH	10	10	1	0.9	
1	PCa	3	3	1	1	1	BTC	9	9	1	0.9	
2	CD	5	10	0.5	0.6	2	ATC	10	10	1	0.9	
2	UC	4	10	0.4	0.67	2	DTC	7	7	1	0.9	
3	ACLF	4	4	1	1	3	BLCA	10	10	1	0.9	
3	TB	3	3	1	1	4	ACLF	4	4	1	0.9	
3	sAH	4	4	1	1	4	AML	9	9	1	0.9	
4	ASS-	10	10	1	1	4	CeAD	5	5	1	0.9	
5	AS	15	15	1	1	4	sAH	4	4	1	0.9	
6	FAD	8	8	1	1	5	BCA	3	3	1	0.9	
6	SAD	9	9	1	1	5	COPD	3	3	1	0.9	
7	BTC	9	9	1	1	5	GAC	2	3	0.66666	0.567	
8	CKD	10	10	1	1	5	HGSO	5	5	1		0.9
9	CD	5	10	0.5	0.6	5	MS	5	5	1		0.9
9	UC	6	10	0.6	0.67	5	MVID	6	6	1	0.9	
10	ESCC	10	10	1	1	5	PCa	3	3	1	0.9	
11	ATC	10	10	1	1	5	TB	3	3	1	0.9	
11	DTC	7	7	1	1	6	ASS-	10	10	1	0.9	
12	BLCA	10	10	1	1	7	HIV	10	10	1	0.9	
13	IPF	10	10	1	1	8	FAD	8	8	1	0.9	
14	MS	5	5	1	1	8	SAD	9	9	1	0.9	
15	HIV	10	10	1	1	9	CKD	10	10	1	0.9	
16	CAD	10	10	1	1	10	AS	15	15	1	0.9	
17	COPD	3	3	1	1	11	CAD	10	10	1	0.9	
18	ODU	10	10	1	1	12	LAC	10	10	1	0.9	
19	AC	9	9	1	1	13	ODU	10	10	1	0.9	
19	NASH	1	10	0.1	0.487	14	CD	5	10	0.5	0.6	

20	AML	9	9	1	1	14	UC	4	10	0.4	0.67
21	DCM	6	6	1	1	15	ESCC	10	10	1	0.9
21	GAC	3	3	1	1	16	CD	5	10	0.5	0.6
22	LAC	10	10	1	1	16	UC	6	10	0.6	0.67
23	HGSO	5	5	1	1	17	IPF	10	10	1	0.9
24	NASH	6	10	0.6	0.487	18	DCM	6	6	1	0.9
25	BCA	3	3	1	1	18	GAC	1	3	0.333333	0.567
26	CeAD	5	5	1	1						
Canberra						Cosine					
cluster	disease	n_in_th is_clust er	group_t otal	frac_of _group	clusteri ng- quality	cluster	disease	n_in_th is_clust er	group_t otal	frac_of _group	clusteri ng- quality
0	CF	9	9	1	0.9	0	CF	9	9	1	0.9
0	NASH	3	10	0.3	0.59	1	BLCA	10	10	1	0.9
1	COPD	2	3	0.66666	0.567	2	CD	5	10	0.5	0.6
1	GAC	1	3	0.33333	0.3	2	UC	6	10	0.6	0.67
1	HGSO	5	5	1	0.9	3	CAD	10	10	1	0.9
2	ODU	10	10	1	0.9	4	CD	5	10	0.5	0.6
3	CKD	10	10	1	0.9	4	UC	4	10	0.4	0.67
4	CD	5	10	0.5	0.6	5	ODU	10	10	1	0.9
4	UC	4	10	0.4	0.67	6	AS	5	15	0.333333	0.717
5	BLCA	10	10	1	0.9	7	ACLF	4	4	1	0.9
6	ACLF	4	4	1	0.9	7	TB	3	3	1	0.9
6	AML	1	9	0.11111	0.456	7	sAH	4	4	1	0.9
6	sAH	4	4	1	0.9	8	ATC	4	10	0.4	0.67
7	ASS-	10	10	1	0.9	8	DTC	5	7	0.71428	0.75
8	AS	10	15	0.66666	0.717	9	IPF	10	10	1	0.9
9	DCM	6	6	1	0.9	10	SAD	3	9	0.333333	0.528
9	GAC	1	3	0.33333	0.3	11	ESCC	10	10	1	0.9
9	TB	3	3	1	0.9	12	BTC	9	9	1	0.9
10	FAD	8	8	1	0.9	13	ASS-	10	10	1	0.9
10	SAD	6	9	0.66666	0.567	14	DCM	6	6	1	0.9
11	ATC	10	10	1	0.9	15	CKD	10	10	1	0.9
11	DTC	6	7	0.85714	0.7	16	MVID	6	6	1	0.9
12	BTC	9	9	1	0.9	17	HIV	10	10	1	0.9

13	AC	9	9	1	0.9	18	LAC	10	10	1	0.9
13	NASH	7	10	0.7	0.59	19	MS	5	5	1	0.9
14	CD	5	10	0.5	0.6	20	AML	4	9	0.44444	0.639
14	UC	6	10	0.6	0.67	20	BCA	3	3	1	0.9
15	LAC	8	10	0.8	0.81	21	AC	9	9	1	0.9
16	CAD	10	10	1	0.9	21	NASH	1	10	0.1	0.637
16	HIV	1	10	0.1	0.88	22	FAD	4	8	0.5	0.6
17	BCA	3	3	1	0.9	22	SAD	4	9	0.44444	0.528
17	COPD	1	3	0.33333	0.567	23	AS	10	15	0.66666	0.717
17	MVID	6	6	1	0.9	24	COPD	3	3	1	0.9
17	PCa	3	3	1	0.9	24	HGSO	5	5	1	0.9
18	CeAD	5	5	1	0.9	25	NASH	6	10	0.6	0.637
19	IPF	10	10	1	0.9	26	CeAD	5	5	1	0.9
20	AML	5	9	0.55555	0.456	27	AML	5	9	0.55555	0.639
20	DTC	1	7	0.14285	0.7	28	ATC	6	10	0.6	0.67
20	GAC	1	3	0.33333	0.3	28	DTC	2	7	0.28571	0.75
20	SAD	3	9	0.33333	0.567	29	PCa	3	3	1	0.9
21	ESCC	10	10	1	0.9	30	FAD	4	8	0.5	0.6
22	HIV	9	10	0.9	0.88	30	SAD	2	9	0.22222	0.528
23	MS	5	5	1	0.9	31	GAC	3	3	1	0.9
24	AML	3	9	0.33333	0.456	32	NASH	3	10	0.3	0.637
24	LAC	2	10	0.2	0.81						
25	AS	5	15	0.33333	0.717						

Color-coded cluster assignments for all diseases after HKG normalization, together with the corresponding

clustering-quality score for each disease–cluster mapping.

The color scheme and qualitative coherence categories are identical to those described in **Supplementary Table S3**.

Supplementary Table S5. Detailed DrugBank-annotated drug–target associations for DCM- and GAC-associated candidate compounds

drug_id	drug_name	candidate_source	target_gene	target_name	action
DB00694	Daunorubicin	DCM_only		DNA	intercalation
			<i>TOP2A</i>	DNA topoisomerase 2-alpha	inhibitor
			<i>TOP2B</i>	DNA topoisomerase 2-beta	inhibitor
DB01254	Dasatinib	GAC_only	<i>ABL1</i>	Tyrosine-protein kinase ABL1	inhibitor;multitarget
			<i>SRC</i>	Proto-oncogene tyrosine-protein kinase Src	inhibitor;multitarget
			<i>EPHA2</i>	Ephrin type-A receptor 2	antagonist
			<i>LCK</i>	Tyrosine-protein kinase Lck	inhibitor;multitarget
			<i>YES1</i>	Tyrosine-protein kinase Yes	inhibitor
			<i>KIT</i>	Mast/stem cell growth factor receptor Kit	antagonist
			<i>PDGFRB</i>	Platelet-derived growth factor receptor beta	antagonist
			<i>FYN</i>	Tyrosine-protein kinase Fyn	inhibitor;multitarget
			<i>BCR</i>	Breakpoint cluster region protein	inhibitor
			<i>STAT5B</i>	Signal transducer and activator of transcription 5B	inhibitor
			<i>ABL2</i>	Tyrosine-protein kinase ABL2	multitarget
			<i>BTK</i>	Tyrosine-protein kinase BTK	inhibitor
			<i>NR4A3</i>	Nuclear receptor subfamily 4 group A member 3	inhibitor
			<i>CSK</i>	Tyrosine-protein kinase CSK	inhibitor
			<i>EPHA5</i>	Ephrin type-A receptor 5	inhibitor
			<i>EPHB4</i>	Ephrin type-B receptor 4	inhibitor
			<i>FGR</i>	Tyrosine-protein kinase Fgr	inhibitor
			<i>FRK</i>	Tyrosine-protein kinase FRK	binder
			<i>HSPA8</i>	Heat shock cognate 71 kDa protein	binder
			<i>LYN</i>	Tyrosine-protein kinase Lyn	inhibitor
			<i>MAP3K20</i>	Mitogen-activated protein kinase kinase kinase 20	binder
			<i>MAPK14</i>	Mitogen-activated protein kinase 14	binder
			<i>PPAT</i>	Amidophosphoribosyltransferase	binder
DB03496	Alvocidib	shared_overlap	<i>PYGM</i>	Glycogen phosphorylase, muscle form	inhibitor
			<i>CDK9</i>	Cyclin-dependent kinase 9	inhibitor

			<i>PYGB</i>	Glycogen phosphorylase, brain form	
			<i>PYGL</i>	Glycogen phosphorylase, liver form	
			<i>CDK7</i>	Cyclin-dependent kinase 7	
			<i>CDK5</i>	Cyclin-dependent kinase 5	
			<i>CDK1</i>	Cyclin-dependent kinase 1	
			<i>CDK6</i>	Cyclin-dependent kinase 6	
			<i>EGFR</i>	Epidermal growth factor receptor	
			<i>CDK4</i>	Cyclin-dependent kinase 4	
			<i>CDK8</i>	Cyclin-dependent kinase 8	
			<i>CDK2</i>	Cyclin-dependent kinase 2	inhibitor
DB04865	Omacetaxine mepesuccinate	GAC_only	<i>rpl2</i>	Large ribosomal subunit protein uL2	antagonist
			<i>RPL3</i>	Large ribosomal subunit protein uL3	antagonist
DB06061	AZD-8330	GAC_only	<i>MAP3K1</i>	Mitogen-activated protein kinase kinase kinase 1	
DB06999	PLX-4720	GAC_only	<i>BRAF</i>	Serine/threonine-protein kinase B-raf	inhibitor
			<i>KDR</i>	Vascular endothelial growth factor receptor 2	inhibitor
DB07101	Mirdametinib	GAC_only	<i>MAP2K1</i>	Dual specificity mitogen-activated protein kinase kinase 1	inhibitor
			<i>MAP2K2</i>	Dual specificity mitogen-activated protein kinase kinase 2	
DB07159	Tamatinib	GAC_only	<i>SYK</i>	Tyrosine-protein kinase SYK	inhibitor
			<i>RET</i>	Proto-oncogene tyrosine-protein kinase receptor Ret	inhibitor
DB08073	(2S)-1-(1H-INDOL-3-YL)-3-{{[5-(3-METHYL-1H-INDAZOL-5-YL)PYRIDIN-3-YL]OXY}PROPAN-2-AMINE	DCM_only	<i>AKT1</i>	RAC-alpha serine/threonine-protein kinase	inhibitor
			<i>AKT2</i>	RAC-beta serine/threonine-protein kinase	
			<i>GSK3B</i>	Glycogen synthase kinase-3 beta	
			<i>PRKACA</i>	cAMP-dependent protein kinase catalytic subunit alpha	
			<i>PKIA</i>	cAMP-dependent protein kinase inhibitor alpha	

DB08142	AT-7519	shared_overlap	<i>CDK2</i> <i>CDK1</i>	Cyclin-dependent kinase 2 Cyclin-dependent kinase 1	inhibitor
DB08597	Dorsomorphin	GAC_only	<i>PRKAG1</i> <i>PRKAB1</i> <i>PRKAA1</i> <i>ACVR1</i> <i>FKBP1A</i>	5'-AMP-activated protein kinase subunit gamma-1 5'-AMP-activated protein kinase subunit beta-1 5'-AMP-activated protein kinase catalytic subunit alpha-1 Activin receptor type-1 Peptidyl-prolyl cis-trans isomerase FKBP1A	inhibitor inhibitor inhibitor
DB11663	Pictilisib	GAC_only	<i>PIK3CG</i>	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit gamma isoform	inhibitor
DB11689	Selumetinib	GAC_only	<i>MAP2K1</i> <i>MAP2K2</i>	Dual specificity mitogen-activated protein kinase kinase 1 Dual specificity mitogen-activated protein kinase kinase 2	inhibitor inhibitor
DB11805	Saracatinib	DCM_only	<i>ABL1</i> <i>SRC</i>	Tyrosine-protein kinase ABL1 Proto-oncogene tyrosine-protein kinase Src	modulator modulator
DB12429	CI-1040	GAC_only	<i>AKT1</i> <i>GSK3B</i> <i>MAPK12</i> <i>CHEK1</i> <i>LCK</i> <i>MAPK14</i> <i>MAPK11</i> <i>PRKCA</i> <i>MAPK8</i> <i>ROCK1</i> <i>RPS6KB1</i> <i>MAPK1</i> <i>SGK1</i> <i>MAP3K2</i> <i>MAP3K1</i>	RAC-alpha serine/threonine-protein kinase Glycogen synthase kinase-3 beta Mitogen-activated protein kinase 12 Serine/threonine-protein kinase Chk1 Tyrosine-protein kinase Lck Mitogen-activated protein kinase 14 Mitogen-activated protein kinase 11 Protein kinase C alpha type Mitogen-activated protein kinase 8 Rho-associated protein kinase 1 Ribosomal protein S6 kinase beta-1 Mitogen-activated protein kinase 1 Serine/threonine-protein kinase Sgk1 Mitogen-activated protein kinase kinase kinase 2 Mitogen-activated protein kinase kinase kinase 1	inhibitor inhibitor inhibitor inhibitor inhibitor inhibitor inhibitor inhibitor inhibitor inhibitor inhibitor inhibitor inhibitor inhibitor
DB12686	PHA-793887	DCM_only	<i>CDK1</i> <i>CDK4</i> <i>CDK2</i>	Cyclin-dependent kinase 1 Cyclin-dependent kinase 4 Cyclin-dependent kinase 2	inhibitor inhibitor inhibitor
DB12847	Pyroxamide	GAC_only	<i>HDAC1</i>	Histone deacetylase 1	inhibitor

			<i>HDAC1</i>	Histone deacetylase	inhibitor
DB12904	ZSTK-474	GAC_only	<i>PIK3CG</i>	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit gamma isoform	modulator

This table lists all DrugBank-annotated target proteins associated with candidate compounds identified for DCM and GAC. For each compound, the corresponding target gene symbol, annotated protein name, and reported mechanism of action (e.g., inhibitor, antagonist, modulator) are provided.

DrugBank data were accessed under an academic license (version 5.1).