A robust unsupervised clustering approach for high-dimensional biological imaging data reveals shared drug-induced morphological signatures

Shaine Chenxin Bao[1,9], Dalia Mizikovsky[1,9], Kathleen Pishas[2,9], Karla J Cowley[3], Qiongyi Zhao[1], Evanny Marinovic[2], Mark Carey[4,5,6], Ian Campbell[2], Kaylene J Simpson[3,7,8], Dane Cheasley[2,#], Nathan J. Palpant[1,#,*]

[1] Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia

[2]Peter MacCallum Cancer Centre, Melbourne Australia

[3]Victorian Centre for Functional Genomics, Peter MacCallum Cancer Centre, Melbourne, Australia

[4]Division of Gynecology Oncology, Department of Obstetrics and Gynecology, University of British Columbia, Vancouver, British Columbia, V5Z 1M9, Canada

[5]Department of Clinical Research, BC Cancer, Vancouver, British Columbia, V5Z 4E6, Canada.

[6]Division of Gynecologic Oncology, Diamond Health Centre, Vancouver, British Columbia, Canada.

[7]Department of Biochemistry and Pharmacology, University of Melbourne, Parkville, Australia

[8]Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Australia

[9] Co-first authors

[#]Co-last authors

*Corresponding Author:

Associate Professor Nathan Palpant BSc PhD
Institute for Molecular Bioscience
The University of Queensland
St. Lucia, QLD 4072, Australia
Tel: +61 0439241069
Email: n.palpant@imb.uq.edu.au

Abstract

High-throughput analysis methods have emerged as central technologies to accelerate discovery by accelerating generation of large-scale data. Analysis and interpretation of these datasets remains challenging due to limitations in computational approaches for dimensionality reduction. Here, we present UnTANGLeD, a versatile computational pipeline that prioritises biologically robust and meaningful information to guide actionable strategies from input screening data which we demonstrate using results from image-based drug screening. By providing a robust framework for analysing high dimensional biological data, UnTANGLeD offers a powerful tool for analysis of theoretically any data type from any screening platform.

Advancements in sequencing technologies have enabled the generation of biobanks, cell-atlases and high-throughput perturbation platforms. Traditional analysis techniques, which often rely on direct comparison between predefined groups, are inadequate for parsing these complex, high dimensional datasets. Instead, unsupervised clustering can be applied to identify groups with shared changes across the characterized conditions, without requiring established ground truth. However, high-throughput datasets often suffer from increased noise, sparsity and multi-collinearity between features, which makes output from unsupervised clustering challenging to evaluate and interpret[1].

UnTANGLeD (Unsupervised Trait Analysis of Networks from Gene Level Data), an ensemble clustering approach, was first developed to identify conserved gene programs from genome wide association study (GWAS) data linking genes to over 1,000 diverse traits and diseases[2] **(Figure 1A)**. UnTANGLeD employs iterative clustering to construct a consensus matrix that captures robust relationships across varying resolutions. This approach, combined with silhouette score as a robustness metric to optimize cluster number, enables UnTANGLeD to extract biologically relevant clusters from large, noisy datasets by focusing on reproducible associations rather than spurious correlations. Here, we extend UnTANGLeD to high-content imaging screens, applied to the phenotypic response of three ovarian cancer cell lines with unique genetic profiles (AOCS-2, VOA-6406, SLC-58), and one control ovarian cell line (IOSE-523) to a diverse selection of 5596 FDA approved drugs screened at 3 different concentrations (10, 1, 0.1uM final) in 384 well format **(Figure 1B**). Cells were stained with DAPI and phalloidin (F-actin) and imaged using a CellInsight CX7-LZR high content microscope, at 20X magnification and 9 fields per well. By analysing quantitative information extracted from these images, drugs causing shared morphological changes in cells can be grouped together for downstream investigation.

Each dataset, comprising all drug conditions (defined as compound with dosage) and fields imaged, underwent cytotoxic compound filtering based on cell count thresholds, feature extraction with principal component analysis, and iterative shared nearest neighbour clustering across increasing cluster numbers. This enabled the calculation of the consensus matrix, an extremely sparse matrix that captures the frequency at which any two drugs cluster together as the granularity of clustering increases (**Figure 1C**). We use the silhouette score to determine the optimal number of clusters. This score measures how well each object fits within its assigned cluster compared to other clusters with value ranging from –1 (least robust) to 1 (most robust). Applied to the consensus matrix, it directly assesses the consistency of drug sets clustering together compared to their likelihood of clustering with other sets. A higher silhouette score indicates more robust and distinct clusters, while a plateau in this score indicates that further splitting does not improve distinction between clusters. The score plateaued at approximately 150 clusters across all four datasets, suggesting that this clustering solution optimally captures the inherent structure of the consensus matrix (**Figure 1D**).

Agglomerative clustering using Ward's minimum variance was performed on the consensus matrix, yielding 150 drug clusters. UnTANGLeD then stratified these clusters based on their silhouette score and internal drug correlation across imaging features (**Figure 1E, Extended Figure 1-3).** Clusters scoring highly on both metrics were deemed as top-ranked clusters, indicating robust groupings with shared morphological signatures. They showed stronger enrichment for cancer-relevant drug pathways such as DNA damage and MAPK/ERK pathways (**Figure 1F**), and exhibited more consistent morphological changes in treated cells, contrasting with inconsistent morphologies in poorly ranked clusters (**Extended Figure 4)**. Notably, the drugs in top-ranked Cluster 125 from cell line VOA-6406 were enriched for cytoskeleton-relevant pathways, and accordingly induced consistent cytoskeletal changes characterized by increased cell area, irregular extensions and enhanced cytoplasmic granularities compared to DMSO vehicle controls (**Figure 1G**). Lastly, to evaluate how well the clustering relationships identified in one cell line are preserved in others, we employed three measures of cluster conservation and found that top ranked clusters (123, 124, 125) are more conserved across all three ovarian cancer cell lines, demonstrating UnTANGLeD's ability to identify robust, reproducible drug-

response signatures across diverse genetic profiles of ovarian carcinoma cells (**Figure 1E, Extended Figure 1-3, 5**).

We further validated our approach using the Connectivity Map (CMap)[3], an independent dataset characterising changes in gene expression in response to drug perturbations. UnTANGLeD was applied to CMap to yield 175 drug clusters. Next, we assessed if the drug clusters identified using imaging data were conserved with the drug clusters identified using gene expression data. Indeed, drug sets between the two modalities were significantly conserved, with stronger conservation observed in highly ranked clusters **(Figure 2A)**. This alignment between morphological and transcriptomic clusters provides orthogonal validation of our method, suggesting that the drug-induced morphological changes captured by our imaging-based approach are concordant with underlying transcriptomic alterations.

Finally, we compared UnTANGLeD to traditional cluster methods, k-means and hierarchical clustering, on their ability to define and evaluate cluster quality and biological relevance. While attempting to identify optimal cluster numbers using silhouette scores, we found that unlike UnTANGLeD, both traditional methods' scores rapidly fell and plateaued near 0 (**Figure 2B**). This phenomenon occurs because the matrix is dense, and therefore the distinction between clusters quickly diminishes as the data is split further. For fair comparison, we set the number of clusters to 150 for all methods, based on the optimal number determined by UnTANGLeD's silhouette score analysis. We found that UnTANGLeD-defined clusters were more conserved across ovarian cancer lines and on average had more enrichment for biological pathways **(Figure 2C, D)**. Furthermore, UnTANGLeD's cluster quality metrics (correlation and silhouette score) also better correlated with increased conservation and enrichment compared to traditional methods. **(Extended Figure 6)**. Notably, while UnTANGLeD and k-means had comparable distributions of cluster size, hierarchical clustering had many clusters with very few drugs, potentially indicating random groupings (**Figure 2E**).

Collectively, UnTANGLeD was able to parse large-scale morphology data to reveal robust, meaningful drug groupings characterized by strong pathway enrichment, high conservation across cell lines, and alignment with orthogonal gene expression data. Building upon the consensus clustering approach of Monti et al.[4], UnTANGLeD aggregates clustering across increasing levels of granularity from a single clustering algorithm, in contrast to other methods like SC3[5], ConsensusClusterPlus[6], and Cluster Of Clusters Analysis (COCA)[7], which typically aggregate multiple algorithms or use subsampling techniques to improve clustering robustness. UnTANGLeD's strength lies in its sparse consensus matrix that emphasizes robust, recurring relationships, and its use of silhouette score as a cluster quality metric to this consensus matrix rather than raw data. The consensus matrix, built from multiple clustering iterations, inherently reduces the impact of noise and sparsity. By evaluating cluster quality on this consensus matrix, UnTANGLeD effectively minimizes the influence of spurious associations common in high-dimensional datasets and prioritise meaningful relationships, enabling it to outperform conventional clustering methods. While we've demonstrated UnTANGLeD's effectiveness on imaging and gene expression data, its approach is fundamentally agnostic to data type, with an R package available for application upon request. In conclusion, UnTANGLeD prioritises biologically meaningful clusters and extracts interpretable insights from highly complex, noisy datasets, positioning it as a versatile tool for analysis of high-dimensional data with applications spanning drug discovery, target identification, disease subtyping, and biomarker selection.

Acknowledgements

This work has been supported by grant funding from the NHMRC (MRFCDDM000033 to NP and 2007625 to NP) and the National Heart Foundation of Australia (106721 to NP). The Victorian Centre for Functional Genomics RRID:SCR_025582 (K.J.S.) is funded by the Australian Cancer Research Foundation (ACRF), Phenomics Australia, through funding from the Australian Government’s National Collaborative Research Infrastructure Strategy (NCRIS) program, the Peter MacCallum Cancer Centre Foundation and the University

of Melbourne Collaborative Research Infrastructure Program. We thank Jennii Luu and Robert Vary from the Victorian Centre for Functional Genomics for their screening support. We thank Compounds Australia for the provision of compounds and logistical support. We also thank Sophie Shen for her assistance with the schematic in Figure 1.

Code Availability: The code used in this study is available upon request.

Disclosures

No conflicts of interest are reported for any of the authors.

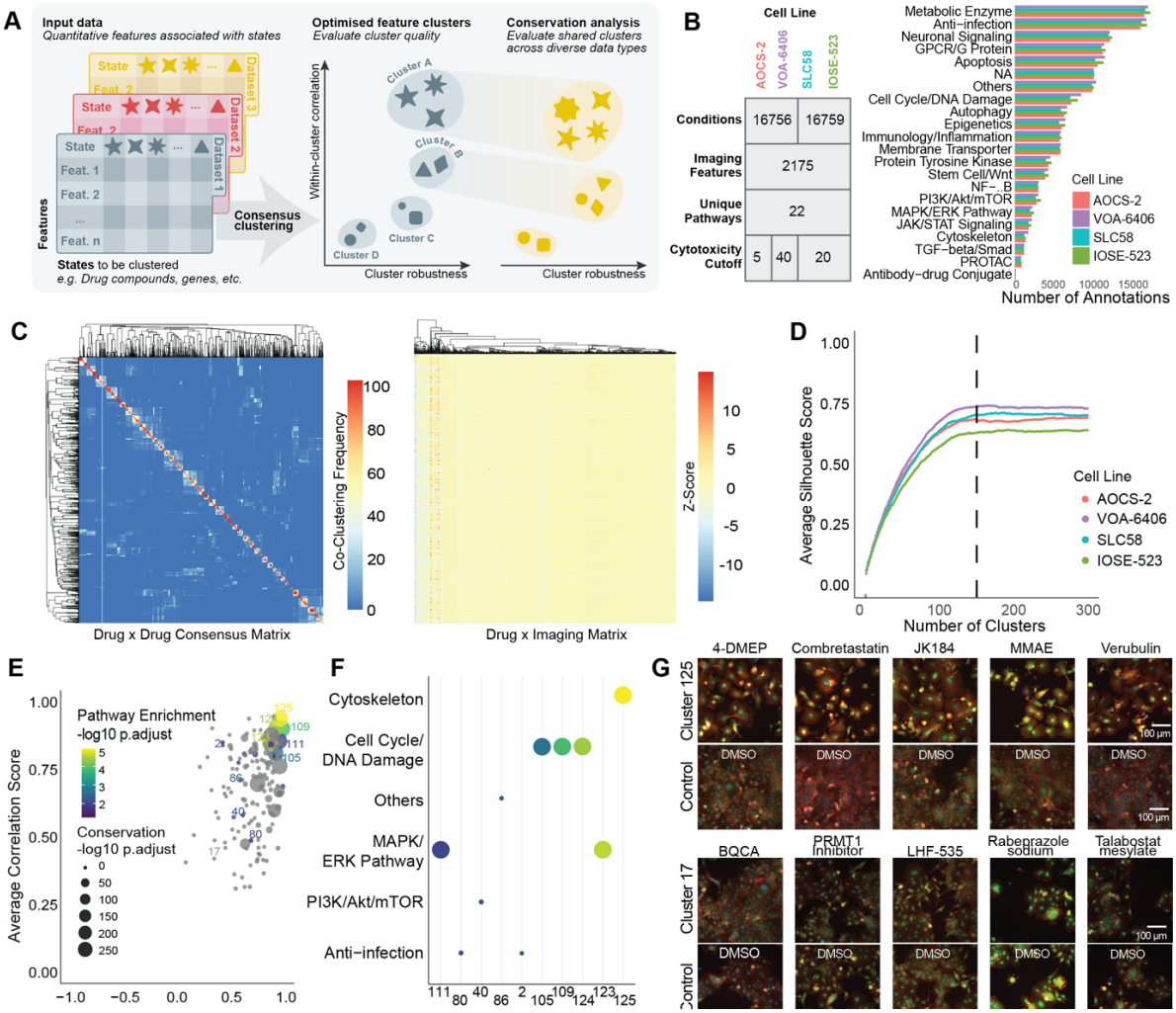

Figure 1: Unsupervised clustering reveals biologically meaningful drug clusters based on cellular morphology. (A) Schematic of computational pipeline. (**B**) Four cell lines (3 ovarian cancer: AOCS-2, VOA-6406, SLC58; 1 normal: IOSE-523) were treated with 5596 FDA approved compound library for 72 hours. The table shows the number of conditions (compound name followed by concentration), number of imaging features initially extracted, unique pathways, and cell count threshold for cytotoxicity cutoff for each cell line. Bar plot displays the distribution of annotated biological pathways across all compounds. (**C**) Heatmaps comparing a subset of the sparse drug-by-drug consensus matrix (left) to the dense, raw drug-by-imaging matrix (right). The consensus matrix represents drug co-clustering frequency across 100 iterations of Seurat clustering at resolutions 0.2 to 20 with an interval of 0.2. (**D**) Average silhouette score across 2-300 clusters for all cell lines. ~150 clusters were selected for all cell lines based on the average silhouette score plateau point (indicated by dashed line). (**E**) Scatter plot of clusters in cell line VOA-6406 stratified by average silhouette score and correlation score. (**F**) Pathway enrichment for the top 10 clusters with the highest Enrichment -log10 p-adjust for cell line VOA-6406. Enrichment calculated using hypergeometric test. **(E-F)** Size indicates conservation significance (Conservation -log10 p-adjust) across cell lines while colour represents pathway enrichment significance (Enrichment -log10 p-adjust). (**G**) Representative images of cells (1 field per well) from cell line VOA-6406 treated with selected drug conditions from top-ranked Cluster 125 (4-DEMP 1μM, Combretastatin 0.1μM, JK184 10μM, MMAE 1μM, Verubulin 10μM) and bottom-ranked Cluster 17 (BQCA 10μM, PRMT1 inhibitor 0.1μM, LHF-535 1μM, Rabeprazole sodium 10μM, Talabostat mesylate 1μM). Images shown with DMSO control and 100μm scale bar. Cells were stained with DAPI, CellMask (plasma membrane) and phalloidin/rhodamine (F-actin satin).

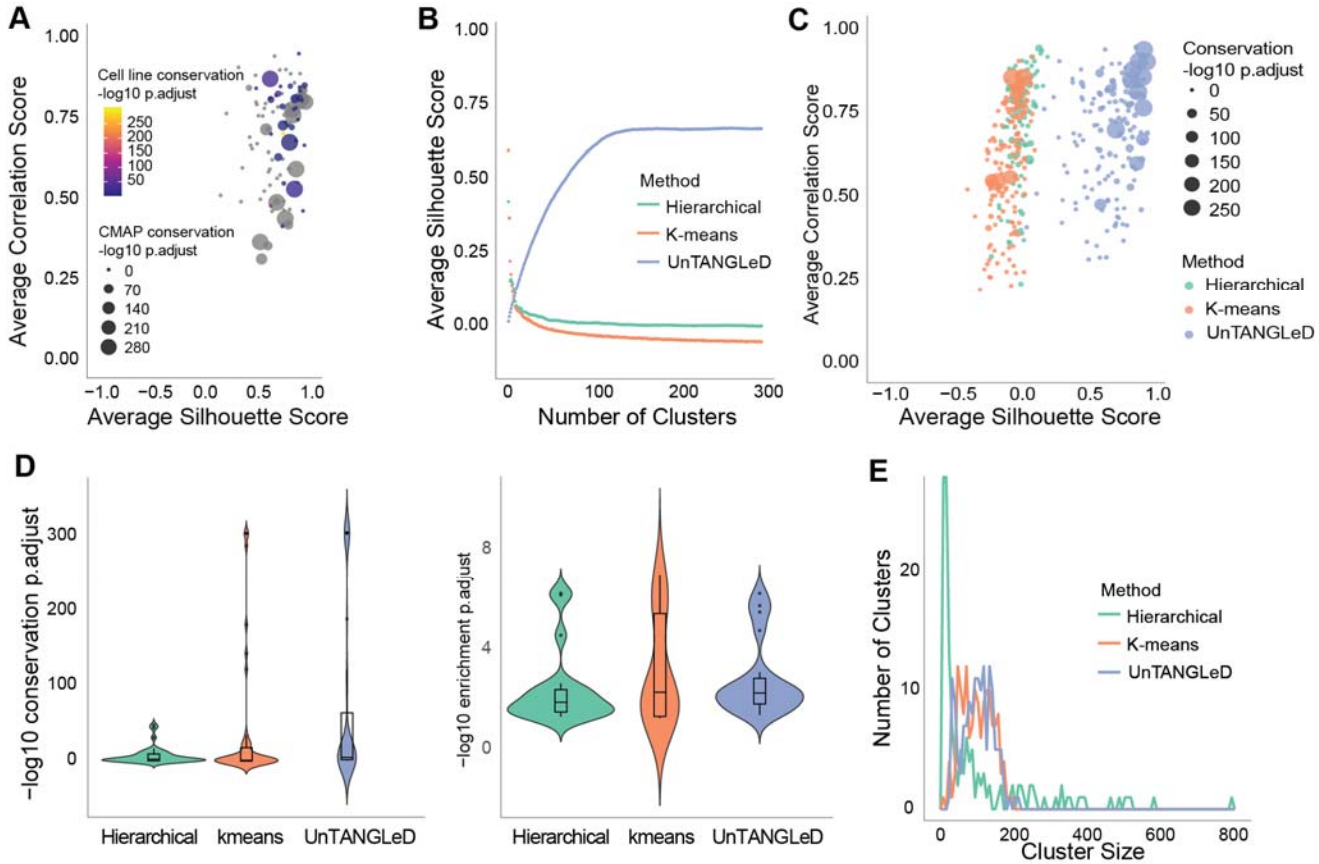

Figure 2: UnTANGLeD outperforms traditional clustering methods in identifying biologically relevant drug groups with orthogonal validation from CMap. (A) Scatter plot of UnTANGLeD clusters stratified by average silhouette score and correlation score. Size indicates conservation significance of UnTANGLeD drug clusters with those derived from Connectivity Map (CMap) using gene expression, limited to clusters containing overlapping drugs only. **(B)** Comparison of average silhouette scores across cluster numbers for UnTANGLeD, k-means, and hierarchical clustering methods. **(C)** Scatter plot comparing cluster quality metrics (average silhouette score vs. average correlation score) for UnTANGLeD, k-means, and hierarchical clustering in cell line VOA-6406. Dot size indicates conservation significance across cell lines (-log10 p-adjust). **(D)** Violin plots comparing conservation (left) and pathway enrichment (right) for significant clusters with adjusted p-value < 0.05 across the three clustering methods in cell line VOA-6406. **(E)** Distribution of cluster sizes for UnTANGLeD, k-means, and hierarchical clustering in cell line VOA-6406.

References

1 Donoho, D. L. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture 1(2000), 32*. (2000).

2 Mizikovsky, D., Sanchez, M. N., Nefzger, C. M., Partida, G. C. & Palpant, N. J. Organization of gene programs revealed by unsupervised analysis of diverse gene-trait associations. *Nucleic Acids Res* **50** (2022). https://doi.org:10.1093/nar/gkac413

3 Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452 e1417 (2017). https://doi.org:10.1016/j.cell.2017.10.049

4 Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* **52**, 91-118 (2003). https://doi.org:Doi 10.1023/A:1023949509487

5 Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486 (2017). https://doi.org:10.1038/nmeth.4236

6 Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010). https://doi.org:10.1093/bioinformatics/btq170

7 Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012). https://doi.org:10.1038/nature11412