UNIVERSITY OF CALIFORNIA SAN DIEGO

Expanding the applicability of gene set enrichment analysis with data-driven gene set refinement and adaptation for sparse data

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology
with a specialization in Biomedical Informatics

by

Alexander T. Wenzel

Committee in charge:

Professor Jill P. Mesirov, Chair
Professor Bing Ren, Co-Chair
Professor Hannah Carter
Professor J. Silvio Gutkind
Professor Pablo Tamayo

2024

The Dissertation of Alexander T. Wenzel is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

This dissertation is dedicated to my parents Susan and Tom. None of this work would have been possible without their lifelong dedication to my education and well-being, and their constant encouragement and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FILES

wenzel_ch3-supplemental-tables.xlsx

Table S3.1: Copy Number Segments called by CODEX

Table S3.2: List novel coding mutations not shared between all clones

Table S3.3: Single cell Pt content measured by Mass Cytometry

Table S3.4: Exome sequencing summary statistics

Table S3.5: Summary statistics of the single cell RNA sequencing

ACKNOWLEDGEMENTS

I would first like to thank my advisor Dr. Jill Mesirov. This work would not have been possible without her patience, guidance, and dedication to my training. Over my years in her lab, she has opened uncountable opportunities for me and fully supported me in pursuing them, and I owe my development as a scientist to her mentorship. I would also like to thank all the members of my thesis committee, Bing Ren, Hannah Carter, Silvio Gutkind, and Pablo Tamayo, for their invaluable input throughout this process, and every member of the Mesirov Lab from the beginning of my degree to the present, especially Owen Chapman, my fellow member of the 2017 cohort, for consistent scientific advice and friendship. I would also like to acknowledge my collaborators in addition to my committee, especially Dr. Jean Wang, whose insights into ERBB2 signaling helped me understand a key result in Chapter 1, and Dr. Olivier Harismendy, whose dedication and insight helped made Chapter 3 possible.

I want to thank all of my friends and colleagues in the BISB program, especially my fellow members of the 2017 cohort, for providing the critical support network that is needed to navigate graduate school. I want to thank my roommates Clarence, Michelle, Xiaomi, for creating a space that truly feels like home, and my (admittedly favorite) roommate Yuuki for accepting me as one of her humans (a rare honor) and making numerous days better with her enthusiastic front door greetings. I also want to thank my fellow academic workers in UAW 2865 and UAW 5810 for providing a network of support when it was needed most and for showing me a path to a better world, and my fellow workers and good friends in Graduate Workers United for Progress for showing me that the path never closes. Finally, I want to express my gratitude to my family, my brothers Jefferson and Elliott for keeping me grounded and making me laugh during even the most stressful times, and my parents Susan and Tom for

their unwavering support of my education and development as a person, this work would not have been possible without them.

Chapter 1, in full, is a reformatted presentation of the material currently being prepared for submission for publication as “Data driven refinement of expression signatures for enrichment analysis” by Alexander T. Wenzel, Farhoud Faraji, Kuniaki Sato, Kwat Medetgul-Ernar, Anthony Castanza, Romella Sagatelian, Gayathri Donepudi, Omar Halawa, Jean Y. J. Yang, J. Silvio Gutkind, Pablo Tamayo, and Jill P. Mesirov. The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, is a reformatted presentation of the material currently being prepared for submission for publication as “scGSEA: Adapting GSEA to single cell RNA-sequencing data” by Alexander T. Wenzel, John Jun, Tanja Eisemann, Robert Wechsler-Reya, Pablo Tamayo, and Jill P. Mesirov. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is a reformatted reprint of the material as it appears as “Single-cell characterization of step-wise acquisition of carboplatin resistance in ovarian cancer” by Alexander T. Wenzel, Devora Champa, Hrishi Venkatesh, Si Sun, Cheng-Yu Tsai, Jill P. Mesirov, Stephen B. Howell, and Olivier Harismendy. The dissertation author was one of the primary investigators and authors of this material.

VITA

2017 Bachelor of Science in Computer Science, Northwestern University

2024 Doctor of Philosophy in Bioinformatics and Systems Biology with a specialization in Biomedical Informatics, University of California San Diego

PUBLICATIONS *(optional)*

Wenzel, A. T*., Champa D*., Venkatesh, H., Sun, S., Tsai, C-Y., Mesirov, J.P., Bui, J.D., Howell, S.B., Harismendy, O. (2022). Single-cell characterization of step-wise acquisition of carboplatin resistance in ovarian cancer, npj Syst Biol Appl 8, 20

Yélamos, O., Merkel, E. A., Sholl, L. M., Zhang, B., Amin, S. M., Lee, C. Y., Guitart, G. E., Yang, J., **Wenzel, A. T**., Bunick, C. G., Yazdan, P., Choi, J., & Gerami, P. (2016). Nonoverlapping Clinical and Mutational Patterns in Melanomas from the Female Genital Tract and Atypical Genital Nevi. Journal of Investigative Dermatology, 136(9), 1858–1865.

Flores, M., **Wenzel, A.**, & Kuzmanovic, A. (2017). Oak: User-Targeted Web Performance. Proceedings - International Conference on Distributed Computing Systems.

Flores, M., **Wenzel, A.**, Chen, K., & Kuzmanovic, A. (2018). Fury Route: Leveraging CDNs to Remotely Measure Network Distance.

Flores, M., **Wenzel, A.**, & Kuzmanovic, A. (2016). Enabling router-assisted congestion control on the Internet. Proceedings - International Conference on Network Protocols, ICNP, 2016-Decem.

Park, J., Yang, J., **Wenzel, A. T.**, Ramachandran, A., Lee, W. J., Daniels, J. C., Kim, J., Martinez-Escala, E., Amankulor, N., Pro, B., Guitart, J., Mendillo, M. L., Savas, J. N., Boggon, T. J., & Choi, J. (2017). Genomic analysis of 220 CTCLs identifies a novel recurrent gain-of-function alteration in RLTPR (p.Q575E). Blood, 130(12), 1430–1440.

Haugh, A. M., Zhang, B., Quan, V. L., Garfield, E. M., Bubley, J. A., Kudalkar, E., Verzi, A. E., Walton, K., VandenBoom, T., Merkel, E. A., Lee, C. Y., Tan, T., Isales, M. C., Kong, B. Y., **Wenzel, A. T.**, Bunick, C. G., Choi, J., Sosman, J., & Gerami, P. (2018). Distinct Patterns of Acral Melanoma Based on Site and Relative Sun Exposure. Journal of Investigative Dermatology, 138(2), 384–393.

Zhou, X. A., Louissaint, A., **Wenzel, A.**, Yang, J., Martinez-Escala, M. E., Moy, A. P., Morgan, E. A., Paxton, C. N., Hong, B., Andersen, E. F., Guitart, J., Behdad, A., Cerroni,

L., Weinstock, D. M., & Choi, J. (2018). Genomic Analyses Identify Recurrent Alterations in Immune Evasion Genes in Diffuse Large B-Cell Lymphoma, Leg Type. Journal of Investigative Dermatology, 138(11), 2365–2376.

Mah, C. K., **Wenzel, A. T.**, Juarez, E. F., Tabor, T., Reich, M. M., & Mesirov, J. P. (2019). An accessible, interactive genepattern notebook for analysis and exploration of single-cell transcriptomic data. F1000Research, 7, 1–21.

Zhou, X. A., Yang, J., Ringbloom, K. G., Martinez-Escala, M. E., Stevenson, K. E., **Wenzel, A. T.**, Fantini, D., Martin, H. K., Moy, A. P., Morgan, E. A., Harkins, S., Paxton, C. N., Hong, B., Andersen, E. F., Guitart, J., Weinstock, D. M., Cerroni, L., Choi, J., & Louissaint, A. (2021). Genomic landscape of cutaneous follicular lymphomas reveals 2 subgroups with clinically predictive molecular features. Blood Advances, 5(3), 649–661.

Richelle, A., Kellman, B. P., **Wenzel, A. T.**, Chiang, A. W. T., Reagan, T., Gutierrez, J. M., Joshi, C., Li, S., Liu, J. K., Masson, H., Lee, J., Li, Z., Heirendt, L., Trefois, C., Juarez, E. F., Bath, T., Borland, D., Mesirov, J. P., Robasky, K., & Lewis, N. E. (2021). Model-based assessment of mammalian cell metabolic functionalities using omics data. Cell Reports Methods, 1(3), 100040.

Banerjee, S., Yoon, H., Ting, S., Tang, C.-M., Yebra, M., **Wenzel, A. T.**, Yeerna, H., Mesirov, J. P., Wechsler-Reya, R. J., Tamayo, P., & Sicklick, J. K. (2021). KIT low Cells Mediate Imatinib Resistance in Gastrointestinal Stromal Tumor . Molecular Cancer Therapeutics, 20(10), 2035–2048.

Wu, V. H., Yung, B. S., Faraji, F., Saddawi-Konefka, R., Zhiyong, W., **Wenzel, A. T.**, Song, M. J., Pagadala, M. S., Clubb L. M., Chiou, J., Sinha, S., Matic, M., Raimondi, F., Hoang, T. S., Berdeaux, R., Vignali, D. A. A., Iglesias-Bartolome, R., Carter H., Ruppin, E., Mesirov, J. P., Gutkind, J. S. (2023). The GPCR-Gas-PKA signaling axis promotes T cell dysfunction and cancer immunotherapy failure. Nature Immunology.

* These authors contributed equally

ABSTRACT OF THE DISSERTATION

Expanding the applicability of gene set enrichment analysis with data-driven gene set refinement and adaptation for sparse data

by

Alexander T. Wenzel

Doctor of Philosophy in Bioinformatics and Systems Biology
with a specialization in Biomedical Informatics

University of California San Diego, 2024

Professor Jill Mesirov, Chair
Professor Bing Ren, Co-Chair

The study of gene expression, the abundance of RNA transcripts in cells, has played a critical role in expanding knowledge of the molecular underpinnings of cancer and the ways in which treatments affect it. Because genes operate in interlinking pathways with nuanced and context-dependent behavior, deriving the phenotypic implication of changes in abundances of

individual genes can be challenging. For this reason, Gene Set Enrichment Analysis (GSEA) was developed. GSEA is a statistical method that quantifies the activation of pathways or processes as represented by *a priori* annotated gene sets and plays a critical role in constructing pathway-level understandings of cell states in diseases such as cancer. In the use of GSEA and the expansion of its companion database of gene sets, the Molecular Signatures Database (MSigDB), two challenges have emerged. First, some gene sets lack the properties of context-specificity and coordinate regulation, that is, not all their members be collectively more expressed in samples that have a specific phenotype to which the gene set should correspond. Second, the new technologies for measuring gene expression at single cell resolution yield data with different properties than the "bulk" gene expression data for which GSEA was initially developed. I will address the first challenge in Chapter 1, in which I propose a data-driven method for gene set refinement and show that this method yields new gene sets that are both more coordinately regulated and context-specific. I address the second challenge in Chapter 2, in which I present a characterization of the performance of the single sample version of GSEA (ssGSEA) in multiple datasets and propose adaptations which I will show produce enrichment scores that are more stable and certain in the single cell context. Finally, Chapter 3 is an application of gene set based pathway characterization to the problem of understanding the development of chemoresistance in ovarian cancer. Using a cell cycle-aware approach incorporating single sample GSEA, we propose a new framework for modeling the development of resistance to carboplatin. These analyses together lay the groundwork for improving the robustness of GSEA results, adapting it to emerging technologies for gene expression measurement, and applying it to address key challenges in the treatment of cancer.

INTRODUCTION

Following the advent of genomics technologies and the wealth of data it created, cancer is now understood as a heterogeneous collection of many diseases exhibiting tissue and context specific features that have a strong influence on disease progression and response to treatment (Tsimberidou et al., 2020). Therefore, the ability to model disease progression and response to treatment as a function of patients' distinct molecular characteristics if critical for making precision medicine a reality for more people with cancer.

The development of techniques to measure gene expression, the abundance of each RNA transcripts, has already enabled great strides towards building models of cell states in service of cancer precision medicine (Kim et al., 2016; Kim et al., 2017). Among the methods developed to analyze such data is Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005), a statistical method that quantifies the activate of pathways or processes which are represented by gene sets, *a priori* annotated set of genes whose members are all upregulated in cells expressing the phenotype to which the gene set is annotated to correspond. Because genes operate in interlinked pathways, often with tissue- or context-specific nuance, GSEA provides an intuitive notion of pathway- and process-level activity and has become a community-standard method for analyzing gene expression data.

As GSEA has continued to be developed alongside its companion database of gene sets, the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011), two key challenges have emerged. GSEA relies on gene sets that are *coordinately-regulated* and *context-specific*, meaning that their member genes should display collective transcriptional regulation in a clearly defined

biological context. However, some gene sets may lack these properties due to including genes that are related to the pathway but don't necessarily change in transcriptional abundance, or because the gene set encompasses multiple context-specific axes of a signaling cascade or some other process that may exhibit differing behavior and involve different genes in different cellular contexts. Additionally, GSEA was developed for "bulk" gene expression technologies, methods which measure abundance of genes from an admixture of sampled cells. The assumption of these bulk methods is that the sample is homogeneous, which, given recent observations of subclonal heterogeneity (Turajlic et al., 2019) and the dynamics of tumor-infiltrating immune cells(Zhang and Zhang, 2020), is often not the case in the study of cancer. Therefore, new technologies such as single cell RNA-sequencing (scRNA-seq), which can profile gene expression at the granularity of single cells, offer enormous potential for improving our understanding of the complex intra- and inter-cellular dynamics underlying cancer progression and response to treatment (Nofech-Mozes et al., 2023). However, this granularity comes at a cost of increased data sparsity since less RNA is available per-observation. Therefore, it is important to understand how GSEA performs in this new data with its underlying distribution and other properties that differ from bulk technologies.

In this work, I will address these challenges as well as apply gene set enrichment to questions in cancer biology. Here I propose a method for improving the gene sets used with GSEA by performing a data-driven *refinement* of existing signatures, deconvolving them based on patterns in space of their genes. This method uses large-scale multi-omics compendia to one or more coordinately regulated and context-specific signatures within the original set, yielding sets whose scores are more robust and whose enrichment provides a more granular and nuanced biological understanding of the data in question. Additionally, I will address the challenge of

sparsity and GSEA. Here I show that individual scRNA-seq cells are in fact too sparse in most cases for reliable enrichment scores to be computed. In order to adapt GSEA to this data, I propose aggregating cells based on cell type or some other dataset-specific meaningful category. I also benchmark various normalizations and methods of computing GSEA enrichment scores under a variety of conditions across multiple datasets to yield a combination of aggregation, normalization, and scoring metric which successfully adapts GSEA to the single cell context. Finally, I apply enrichment methods to the question of chemoresistance in ovarian cancer. Resistance to frontline therapy is a frequently occurring problem in ovarian cancer and a barrier to better patient outcomes (Siegel et al., 2013). Here, while analyzing data derived from a time course experiment of carboplatin resistance development, I develop a method of pairing GSEA with an awareness of each cell's position along a cell cycle continuum to develop a state model of transitions between plastic resistance states, which has the potential to inform treatment schedules for maximizing the efficacy of carboplatin and minimizing the development of chemoresistant cells. In summary, this work lays the groundwork for improving the robustness of GSEA scores and expanding its applicability into new gene expression technologies to answer key questions in cancer biology in service of enabling precision medicine and improving patient outcomes.

Chapter 1 Data driven refinement of expression signatures for enrichment analysis

1.1 Abstract

Gene set enrichment methods measure biological process or pathway activation in gene expression data by testing for coordinate up- or down-regulation of pathway members in a ranked list of genes. These methods rely on curated, annotated gene sets whose members' coordinate expression is an indicator of a process or phenotype. We developed the Molecular Signatures Database (MSigDB), a collection of expertly annotated gene sets for use with these methods. We have observed that some MSigDB gene sets can lack coordinate expression, especially those derived from canonical pathways. To address this challenge, we developed gene set refinement (GSR), a data-driven approach leveraging large-scale multi-omics compendia to extract context-specific coordinately regulated gene sets, deconvolve heterogeneity, and reveal multiple downstream signaling. We applied this method to address cancer biology questions, and demonstrated successful, targeted refinement of existing MSigDB gene sets.

1.2 Introduction

Gene set enrichment (GSE) methods allow researchers to quantify and assess the statistical significance of changes in gene expression at the process and pathway level, providing greater interpretability than examining the results of single-gene differential expression tests (Mootha et al., 2003; Subramanian et al., 2005; Hung et al., 2012). Most GSE methods either test for overrepresentation of gene set genes, either in the top or bottom of a ranked list of all genes , or in a smaller set of important genes, such as the significant results of a differential expression test (Thomas et al., 2022). However, in the latter case the up-regulation of a process-related set of genes may be subtle enough that none of its members are individually significantly upregulated and so the signal can be missed (Mootha et al., 2003). GSE methods rely on the existence of pre-defined, annotated gene sets to describe a variety of phenotypes, processes, and pathways. A gene set intended for use with GSE should have the property that its member genes are co-expressed in samples that have the phenotype that the gene set is annotated to represent. The Mesirov Lab developed Gene Set Enrichment Analysis (GSEA), a method that takes a gene expression dataset of samples that fall into two classes and tests for non-random distribution of gene set genes in a list of all genes ranked by their correlation with the sample class labels (Subramanian et al., 2005) . In addition, we also developed single sample GSEA (ssGSEA), which uses the expression of genes in an individual sample as their ranking rather than creating a ranking based on expression across samples (Barbie et al., 2009) .

Because GSE methods rely so strongly on high quality gene sets, the Mesirov Lab also developed the Molecular Signatures Database (MSigDB), a repository of gene sets organized into collections (www.msigdb.org) (Liberzon et al., 2011). MSigDB includes collections such as canonical pathway databases like Reactome (Croft et al., 2014) or WikiPathways (Martens et al.,

2021) (the “C2 “collection), ontologies such as Gene Ontology (Thomas et al., 2022) (the “C5” collection), and cell type signatures (the “C8” collection). Each MSigDB gene set is annotated with a description of the phenotype that the gene set represents and information on provenance such as the author, the data used to derive it, and the paper in which it first appeared. MSigDB contains both human- and mouse-native gene sets and provides tools for searching gene sets and checking for overlap between MSigDB gene sets and queried genes of interest. Our work to expand and enhance MSigDB has also included efforts to consolidate the existing sets and collections, which resulted in the Hallmark collection (Liberzon et al., 2015) . The Hallmark collection was created by clustering thousands of gene sets based on the degree of overlap in their member genes, identifying modules of gene sets representing a common biological theme, and using these clusters along with experimental data to create and validate 50 major “Hallmark” gene sets. Since their release, these Hallmark sets have been among the most frequently accessed gene sets in MSigDB.

While the Hallmarks significantly diminished redundancy among the existing collections and, in many cases, improved coordinate regulation, it was not as focused on the goal of context-specificity for the sets. By context-specificity we mean the degree to which genes are co-expressed in a specific tissue, cell type, phenotype, or disease type. Gene sets may lack context-specificity for two reasons. First, they may be annotated too broadly when in fact their coordinate regulation is most evident in samples from a more specific context. This occurs most commonly for experimentally derived gene sets, which are often the top n differentially expressed genes in a single dataset. Second, gene sets may contain multiple context-specific components but are not coordinately regulated as a whole. This is most common in ontology-derived gene sets that are created manually or are intended to encompass all genes related in

some way to a phenotype or pathway, which may represent a superset of the genes that are co-expressed.

Faced with this challenge, and the desire to increase the biological specificity of gene sets, we reasoned that we could take advantage of the substantial increase in the volume of publicly available multi-omics compendia (Weinstein et al., 2013; Tsherniak et al., 2017). We therefore developed *gene set refinement* (GSR), a data-driven approach to learn patterns from across multi-omics data from hundreds of samples to refine existing gene sets into one or more new, coordinately regulated gene sets. We found the method can both filter a single gene set to make it more sensitive and specific by removing genes that contribute noise and identify multiple sub-signatures within a gene set whose coordinate regulation indicates context-specific pathway activation. Both these improvements enable more robust and interpretable GSE results.

1.3 Results

Gene set refinement overview

The GSR method requires two inputs: a multi-omics compendium and a gene set to be refined. Appropriate multi-omics compendia consist of data for several hundred to a few thousand samples for which gene expression data and *direct phenotypic measurements* for the same samples are available (**Figure 1.1**). The refined gene sets are derived from these two types of data (Steps 1 and 2 below). By direct phenotypic measurements we mean characteristics of samples for which enrichment of a gene set should serve as a proxy. In multi-omics compendia these might include proteomics data, gene dependency screens, drug vulnerability screens, epigenetic data such as methylation, mutation status, and sample metadata such as disease type. These direct phenotype measurements are used to annotate the resulting refined gene sets (Step 3 below).

Step 1: Generate patterns with non-negative matrix factorization: We first define the matrix A as the subset of the full gene expression data matrix containing only the profiles of the genes in the input gene set. We then apply non-negative matrix factorization (NMF) to A. NMF decomposes an m x n matrix into two smaller matrices W (m x k) and H (k x n) where k is a user-chosen number of components and $A \approx WH$ (Lee and Seung, 1999; Lee and Seung, 2000). To avoid overfitting and to overcome NMF's local optimality (Vavasis et al., 2010), we define consensus NMF components using a down sampling and bootstrapping approach (see **1.5 Methods** for details). The method evaluates multiple values of k and we choose a final value by using the cophenetic coefficient (Sokal and Rohlf, 1962) as a heuristic to maximize pattern reproducibility and distinctness (Brunet et al. 2004) (see **1.5 Methods**).

Step 2: Define new refined gene sets from NMF components: We use each row or *component* of H from the decomposition of A to yield a new refined gene set as follows. First, we

compute the correlation of each gene's expression profile given in *A* with the row of *H*. Unless otherwise specified, all correlations are computed with the information coefficient (IC), an information theoretic correlation metric ranging from -1 to 1 that is more sensitive to non-linear relationships (Kinney et al. 2014, Joe 1989, Linfoot 1957) The genes with profiles whose IC with the component of at least 0.3 comprise that component's refined gene set. We use 0.3 as a cutoff based on benchmarking in prior work (Kim et al., 2016) . Because each component defines a new gene set, running GSR with a given parameter *k* will yield *k* possible new gene sets.

Step 3: Annotate gene sets using direct phenotype measurements: Each of the *k* resulting gene sets is characterized using its association with the direct phenotypic measurements in the multi-omics data. To do this, we first compute ssGSEA scores using each of the *k* refined gene sets in the test sample sets (see **1.5 Methods**). We then compute the IC between these ssGSEA vectors and the direct phenotypic measurements from the compendium. As we are correlating these measurements with the new gene sets' ssGSEA scores, we infer that a strong association between a gene set and a direct phenotypic measurement means that the new gene set serves as a good proxy for that phenotype when used with a gene set enrichment method.

Refinement of context specific signatures of ERBB2 signaling

We applied the GSR method to the *ERBB2* signaling pathway, specifically the gene set REACTOME_SIGNALING_BY_ERBB2 (Croft et al., 2014). *ERBB2*, the gene encoding the *HER2* protein, is a receptor tyrosine kinase which interacts with other RTKs including *ERBB3*, *ERBB4*, and *EGFR* to affect downstream changes in cell survival, proliferation, and differentiation (Hsu and Hung, 2016). As anti-*HER2* therapies have had success in cancers including some breast cancer subtypes, understanding context specific behaviors and

mechanisms of resistance is important for improving patient outcomes. In fact, recent research has highlighted the heterogeneity of *HER2*-driven cancers and the development of resistance to anti-*HER2* treatments (Pernas and Tolaney, 2020).

For the refinement of this gene set, we used the Cancer Dependency Map expression data and corresponding reverse-phase protein array (RPPA) data as the direct phenotype measurements (Tsherniak et al., 2017). GSR with $k = 3$, chosen based on the cophenetic coefficient (**Figure 1.2A**), produced three distinct gene sets. Refined Set 1 yielded a gene set consisting of *GRB7*, *ERBB3*, *ERBB2*, *EREG*, *ITPR3*, *SRC*, *PRKCD*, *BTC*, *TRIB3*, *EGFR*, and *ADCY6*. The enrichment scores with this gene set were most closely associated with the abundance of the *HER2* protein (IC=0.50, FDR q-value=0.0001) (**Figure 1.2B**), indicating that this signature is more sensitive and concise signature for *ERBB2* activation than the original. Refined Set 2 yielded a gene set consisting of *NRG1*, *PRKCA*, *EGFR*, *MAPKAP1*, *SHC1*, *AKT3*, *BAD*, and *HBEGF*. In contrast with Refined Set 1, this gene set's enrichment scores were most closely associated with the abundance of the *EGFR* protein (IC = 0.51, FDR q-value = 0.00005) (**Figure 1.2C**). Because *ERBB2* has context-dependent signaling activity depending on whether it homodimerizes or heterodimerizes with *EGFR* (Hsu and Hung 2016), these two sets, Refined Set 1 and Refined Set 2, emerging from different components of the refinement, represent two different context-specific patterns of *HER2* signaling within the data. As an external validation for these two gene sets, we compared their enrichment scores in TCGA breast cancer samples with the protein abundance of *HER2* and *EGFR* measured by the TCPA project (Weinstein et al., 2013; Li et al., 2013). Refined Set 1 was most associated with *HER2* protein abundance in patient tumors (IC = 0.48, FDR q-value = 0.00005) (**Figure 1.2E**) and Refined Set 2 was most

associated with *EGFR* abundance in patient tumors (IC= 0.26, FDR q-value = 0.00005) (**Figure 1.2F**).

Refined Set 3 consists of the genes *FYN*, *PCG1*, *PRKAR2B*, *ADCY1*, *NR4A1*, *AKT3*, *PRKACA*, *CAMK4*, *RAF1*, *MTOR*, *SOS1*, *THEM4*, *AKT2*, *GSK3A*, *PRKAR1A*, *RNF41*, and *UBA52*. This gene set includes *PRKACA* and other components of protein kinase A. Previous work by others points to a role for *PRKACA* in resistance of breast tumors to anti-*HER2* treatments (Moody et al., 2015). We therefore compared the enrichment scores for the three gene sets to the measured efficacy of various anti-*HER2* drugs measured in the same Cancer Dependency Map samples. The Refined Set 1 enrichment scores were the most anti-correlated with efficacy of Neratinib across DepMap cell lines (IC = -0.26, FDR = 0.0009) (**Figure 1.2D**); indicating the gene set derived from this component may represent a resistance or compensatory mechanism in *HER2*-driven tumors.

Refining *YAP* target gene set yields context-specific oncogenesis and tumor maintenance signatures

YAP, along with *TAZ*, mediates transcription downstream of the Hippo pathway (Zanconato et al., 2015). Nuclear *YAP*/*TAZ* bind with other proteins capable of binding to DNA, most frequently members of the *TEAD* family (Zanconato et al., 2015). *YAP* and *TAZ* mediate transcription in response to mechanotransduction and changes in the extracellular environment (Chang et al., 2018; Battilana et al., 2021) and are involved in many cancers (Zanconato et al., 2015; Battilana et al., 2021) . Because this pathway is involved in many different cancers across tissue types, we hypothesized that, given a list of *YAP* targets, GSR may identify context- and cell state-dependent signatures. As an input gene set, we used the validated *YAP* target genes

(n=379) identified in (Zanconato et al., 2015) . GSR was run using DepMap gene expression and refined sets from $k = 4$ were used (**Figure 1.3A**), yielding four gene sets (**Supplementary Table 1.1**). Two of the gene sets, Refined Set 1 and Refined Set 2, were associated with abundance of the *YAP* protein (**Figure 1.3B**). The other two sets were negatively correlated with *YAP* despite the input set consisting of validated *YAP* targets. We hypothesize that these components may represent non-*YAP*-mediated transcriptional states, so here we focus on refined sets 1 and 2. Refined gene sets were annotated by assessing their correlation with Reverse Phase Protein Array data also from DepMap. Refined Set 1 was associated with the abundance of receptor tyrosine kinases (RTKs) such as *EGFR*, *HER2*, and *HER3* (**Figure 1.3B**). RTKs are known to limit the Hippo pathway and promote *YAP*-mediated transcription via the translocation of *YAP* to the nucleus (Battilana et al., 2021; Zinatizadeh et al., 2021). Refined Set 1 may therefore represent *YAP*'s activity as an initiator of oncogenesis downstream from these genes. We next found that refined Set 2 was more associated with proteins involved in tumor maintenance and progression (**Figure 1.3B**). PAI-1 (SERPINE1), has been implicated in immunosuppression and cancer cell migration and is a marker of poor prognosis in multiple cancers (Chen et al., 2022; Li et al., 2023). Refined Set 2 is also associated with Caveolin-1 (*CAV1*) (**Figure 1.3B**), which has been reported to both negatively regulate *EGFR* as well as predict local tumor relapse and resistance to *EGFR* targeted therapy (Burgy et al., 2021). We therefore hypothesized that Refined Set 1 represents RTK/YAP-mediated oncogenesis, and that Refined Set 2 may represent *YAP*'s role in the epithelial-mesenchymal transition (EMT) and tumor maintenance.

We used two additional head and neck squamous cell carcinoma (HNSCC) datasets to validate the refined sets. The first is from a study of *YAP*-driven cancer initiation in which mice expressed constitutively active *YAP* and/or HPV oncoproteins. The second is a study of the effect

of *TEAD* inhibition in an established HNSCC cell line Ca133. In the *YAP* oncogenesis data, we observed that Refined Set 1 best separated *YAP*-active samples from non-*YAP* active samples (IC = 0.77, FDR = 0.0001) (**Figure 1.3C**). In contrast, Refined Set 2 best separated wildtype samples (those with intact *YAP*-mediated transcription) from *TEAD*-inhibited samples (IC = 0.70). Notably, while the original gene set scored well in both datasets (IC = 0.77, 0.71 respectively), the refined sets are better able to distinguish the context specific features of these two datasets, as Refined Set 2 scored significantly lower in the *YAP* oncogenesis dataset (IC = 0.46, FDR = 0.004) and Refined Set 1 scored lower in the *YAP TEAD*i dataset (IC = 0.58, FDR = 0.0027).

1.4 Discussion

We have developed a computational approach to refining gene sets for use with gene set enrichment methods. This method takes an existing gene set, and using NMF, deconvolves it into one or more refined sets, defined by NMF-derived patterns, which are more sensitive and specific indicators of a phenotype or of some contextually-specific pathway or process behavior. In the case of the Reactome *ERBB2* signaling gene set, GSR identified a decomposition into three major components. Two of these components represent two contexts of *HER2* signaling, whether homodimeric or heterodimeric with *EGFR* (Hsu and Hung, 2016). The third component defines a gene set that contains *PRKACA* and is anti-correlated with response to Neratinib, indicating that it may represent a compensatory or resistance mechanism to *HER2*-targeted therapy (Moody et al., 2015) . By subjecting a list of *YAP* targets to GSR, we identified two components that define two context-specific gene sets representing *YAP*-mediated oncogenesis by RTKs (Moody et al., 2015) and *YAP*-mediated tumor maintenance and progression (Burgy et al., 2021). Interestingly, while we focus here on finding signatures of YAP-mediated transcription, the other two sets did not correlate with YAP protein abundance. These may represent non-canonical or otherwise non-YAP-dependent regulation of Hippo pathway downstream targets which the field is still defining (Wang et al., 2022; Cho et al., 2021), and warrant further investigation.

The GSR method as described here has limitations which merit additional work. While the purpose of the method is to identify meaningful patterns in large-scale, multi-omics data for the purpose of generating refined gene sets, this does mean that the application of GSR is limited to domains of biological knowledge for which such large data exist. However, there is growing recognition of the importance of multi-omic sample characterization and data generation (Jung et

al., 2020; Santiago-Rodriguez and Hollister, 2021), and the utility of GSR will grow with the volume of available data. Additionally, while GSR can identify patterns without user intervention, a user must still supply a gene set that is comprehensive, meaning that it encompasses all genes necessary for generating the patterns that define the refined set. Using the entire transcriptome as an input to refinement is not feasible both due to excessive runtime and because NMF is unlikely to pathway level, context-specific signals within the space of all protein-coding genes. While we will continue to use existing MSigDB gene sets as the primary source of input sets for GSR, it may also be possible to leverage protein-protein interaction networks and other systems of organizing and annotating genes and their relationships to define input sets of generally related genes from which GSR can derive context-specific patterns and refined sets. In summary, we found GSR can computationally identify more sensitive and context-specific gene sets for use with gene set enrichment methods and other related applications, and, by adding to the MSigDB knowledge base, presents the opportunity to leverage the growing volume of multi-omic data to yield new biological insights from analyzing whole transcriptome data.

1.5 Methods

Data processing

DepMap Gene Expression:

The 23Q2 release "OmicsExpressionProteinCodingGenesTPMLogp1.csv" file was downloaded, and solid tumor samples were extracted. Genes with a sum of less than 5 TPM across all samples were removed. Each sample was subjected to a dense-rank normalization and scaling procedure,

$$10000 * \frac{R_i - \min(R_i)}{\max(R_i) - \min(R_i)}$$

where R_i is the result of applying the Python pandas.DataFrame.rank() function with the parameter method = 'dense' using Pandas version 1.2.3.

DepMap Reverse Phase Protein Array:

The file "CCLE_RPPA_20180123.csv" was retrieved from the DepMap Portal (release 23Q2). The data was used as originally processed (Ghandi et al., 2019). Only data from solid tumor samples were used.

DepMap PRISM Drug Repurposing:

The file "Repurposing_Public_23Q2_Extended_Primary_Data_Matrix.csv" was retrieved from the DepMap Poral (release 23Q2) and used as originally processed (Ghandi et al., 2019). Each entry in the matrix is the log fold change in cell viability following treatment in cell line *i* with drug *j*. The matrix was multiplied by -1 to interpret the data as a measure of drug efficacy. Only data from solid tumor samples were used.

The Cancer Proteome Atlas (TCPA):

Data from the The Cancer Proteome Atlas (Li et al., 2013) were retrieved from The Cancer Proteome Atlas Portal (https://tcpaportal.org/tcpa/download.html). Processed and batch-corrected data from "TCGA-BRCA-L4.zip" were used as provided in the porta (n = 891).

The Cancer Genome Atlas (TCGA):

TPM-normalized gene expression corresponding to samples available in TCPA were retrieved from the UCSC Xena portal (https://xena.ucsc.edu/). Ensembl gene IDs were converted to gene symbols using the org.Hs.eg.db R package (Carlson, 2023). Genes not present in the DepMap expression matrix were removed from the TCGA expression matrix.

Gene set refinement

Nonnegative Matrix Factorization (NMF):

NMF was run as implemented in scikit-learn version 1.0.2 (scikit-learn, 2024). In all uses of NMF, *init* was set to "random", *solver* was set to "mu" (multiplicative update), *beta_loss* was set to "kullback-leibler", and *max_iter* was set to 2000. For all examples, NMF was run for values of *k* from 2 to 9 (see *Choosing a value of k*).

Bootstrapping:

Because sample selection and NMF random initialization can influence the final refined gene set, a bootstrapping method was implemented in order to prevent overfitting and generate a consensus solution. The bootstrapping approach consists of an outer and inner loop. In each iteration of the outer loop, the gene expression samples are randomly split into two-thirds and one-third. The samples in the one-third set are set aside. The inner loop is then performed on the

two-thirds set. On each iteration of the inner loop, the two-thirds split is downsampled by 50%, and the remaining samples, comprising matrix *A*, are subjected to NMF, yielding matrices *W* and *H*. The pairwise correlations (using the IC) are then computed between every row of *A* and every row of *H*, which represent the gene expression and weight of the NMF component, respectively, in each sample. If *A* is an *m* x *n* matrix of m genes and n samples and NMF was run with inner dimension *k*, then the result of the entire bootstrapping process, consisting of *a* outer loop iterations and *b* inner loop iterations, is a gene *x* component matrix of dimension *m* genes and *a*b*k* components where an entry (*i*, *j*) represents the association between the *i*th gene and the *j*th NMF component across all NMF iterations. For all examples here, *a* = 10 and *b* = 50.

Choosing a value of k:

The choice of an optimal value of k is dependent on the expression space of a given gene set. Depending on the degree of context-specificity of complexity of the involved pathway, the different transcriptional axes may be adequately represented with a lower *k* (fewer patterns) or higher *k* (more patterns). To choose *k*, we use a heuristic that maximizes component stability and reproducibility. The gene *x* component matrix is subjected to NMF Consensus Clustering as described in (Brunet et al., 2004). The gene *x* component matrix is used instead of the *H* matrices because NMF results with different initializations are subject to random rotations (Paatero et al., 2002) and different samples are used in each iteration. Briefly, NMF Consensus Clustering finds a grouping of *k* clusters using a consensus-based approach across multiple random initializations of NMF. Following a single execution of NMF, any two samples *p* and *q* are grouped together in clustering *C* if they share the same maximum component in the *H* matrix:

$$C_p = C_q \iff \text{argmax}_{i \in [1,\text{nrow}(H)]} H_{i,p} = \text{argmax}_{i \in [1,\text{nrow}(H)]} H_{i,q}$$

NMF Consensus Clustering builds a square connectivity matrix where entry p, q contains the fraction of iterations in which sample p is clustered with sample q. The connectivity matrix is then hierarchically clustered and the cophenetic coefficient, a measure of similarity between dendrogram distances and Euclidean distances in the original data (Sokal and Rohlf, 1962), is calculated. This process is repeated for every value of k, yielding a cophenetic coefficient for each one. A high cophenetic coefficient indicates that the NMF patterns discovered at that value of k are distinct from one another and reproducible across sample bootstrapping and random NMF initializations. The results from the value of k with the highest cophenetic coefficient are used for the remainder of the method.

Defining consensus components and creating gene sets:

To define consensus patterns across bootstrapping iterations, k-means clustering is applied to the gene x component matrix. The number of clusters k is the same as the NMF k parameter. Each cluster, which consists of NMF components that have similar gene correlation profiles, defines a consensus component. Each of these consensus clusters is then used to generate a gene set. Each component consists of several vectors, each from a different NMF iteration, containing the correlations between each gene in the input gene set and the particular NMF component. The median of these associations is computed for each gene within each consensus component. Within each component, genes with a median association greater than 0.3, a threshold chosen based on benchmarking in previous work (Kim et al., 2016), are selected to comprise that consensus pattern's refined gene set.

Annotations with direct phenotype measurements

Following their creation, refined gene sets are annotated using the direct phenotype measurements. Enrichment scores are computed using ssGSEA for each of the refined gene sets in the *a* sets of samples that were held out in the two-thirds/one-third random split in each outer loop. Within each set, the IC between the vector of ssGSEA scores and every direct phenotypic measurement is computed, yielding a distribution of *a* associations between each gene set and each direct phenotype measurement. The medians of these scores are then computed and the top results for each gene set, sorted by median, are presented to the user.

Code availability

The Python source code for gene set refinement is available in GitHub at https://github.com/alex-wenzel/PheNMF and is installed as a Python package inside a Docker image available at https://hub.docker.com/repository/docker/atwenzel/gpnb-phenmf.

1.6 Figures

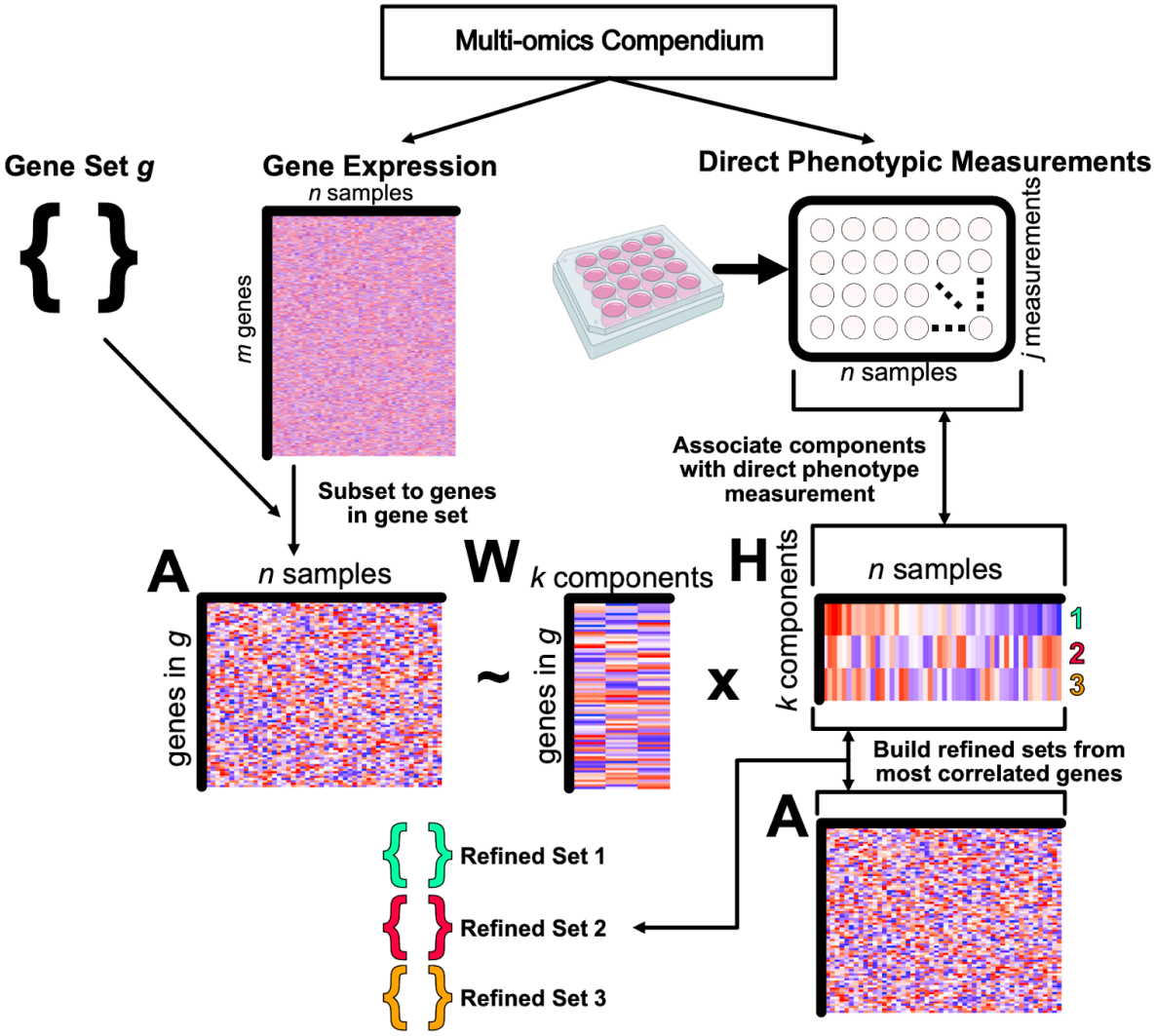


Figure 1.1: Outline of the Gene Set Refinement (GSR) method.

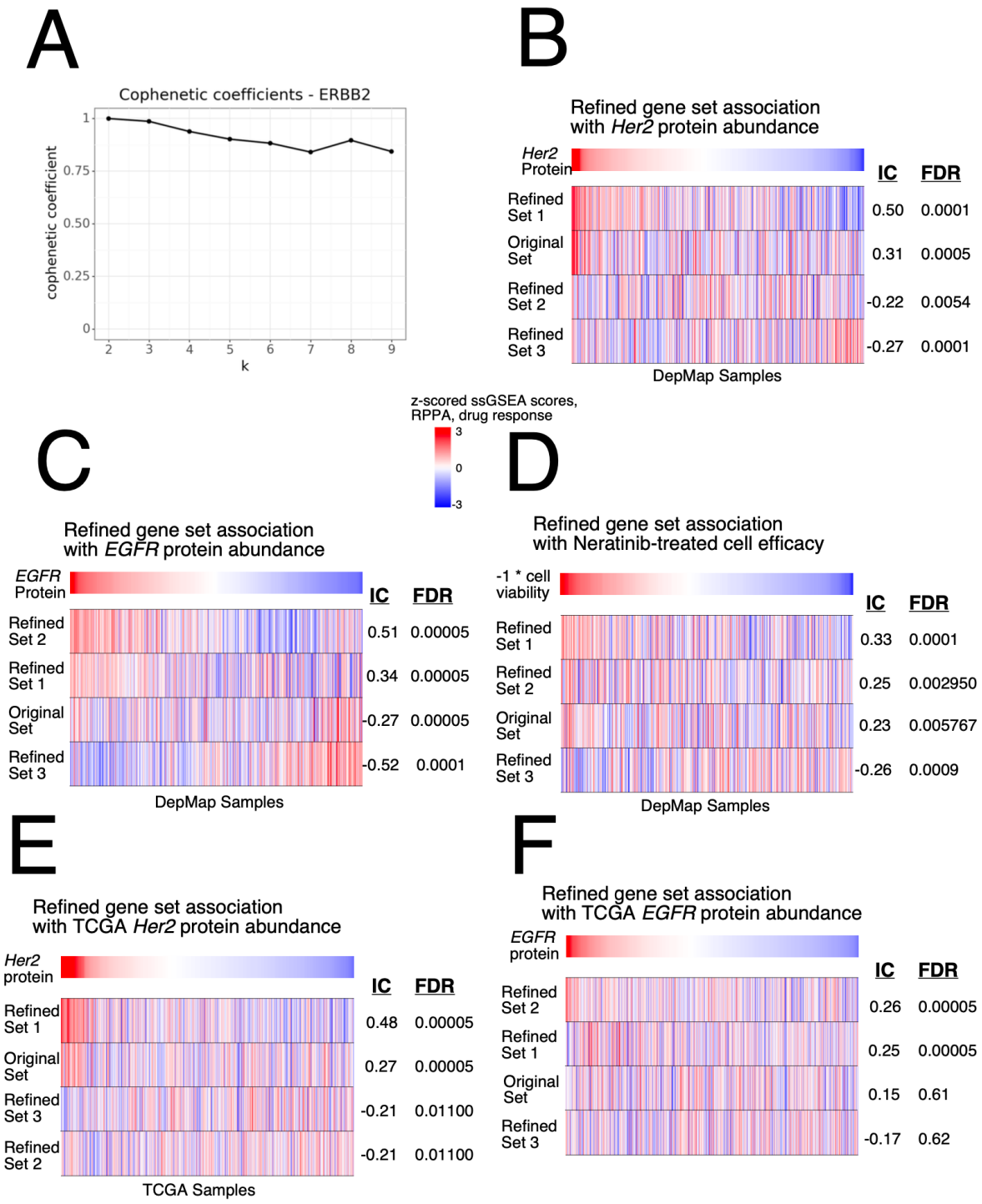


Figure 1.2: Refinement of an *ERBB2* signature

A) Cophenetic coefficients for lower dimensional patterns at $k = 2$ through $k = 9$. **B-F)** Associations between gene sets and direct phenotypic measurements. Gene set matrices are ssGSEA enrichment scores. IC: Information coefficient. FDR: Multiple-hypothesis corrected q-value from permutation testing. **B, E)** Association between gene sets and *HER2* protein in **(B)** DepMap and **(E)** TCPA . **C, F)** Association between gene sets and *EGFR* protein in **(C)** DepMap and **(F)** TCPA. **D)** Association between gene sets and Neratinib efficacy

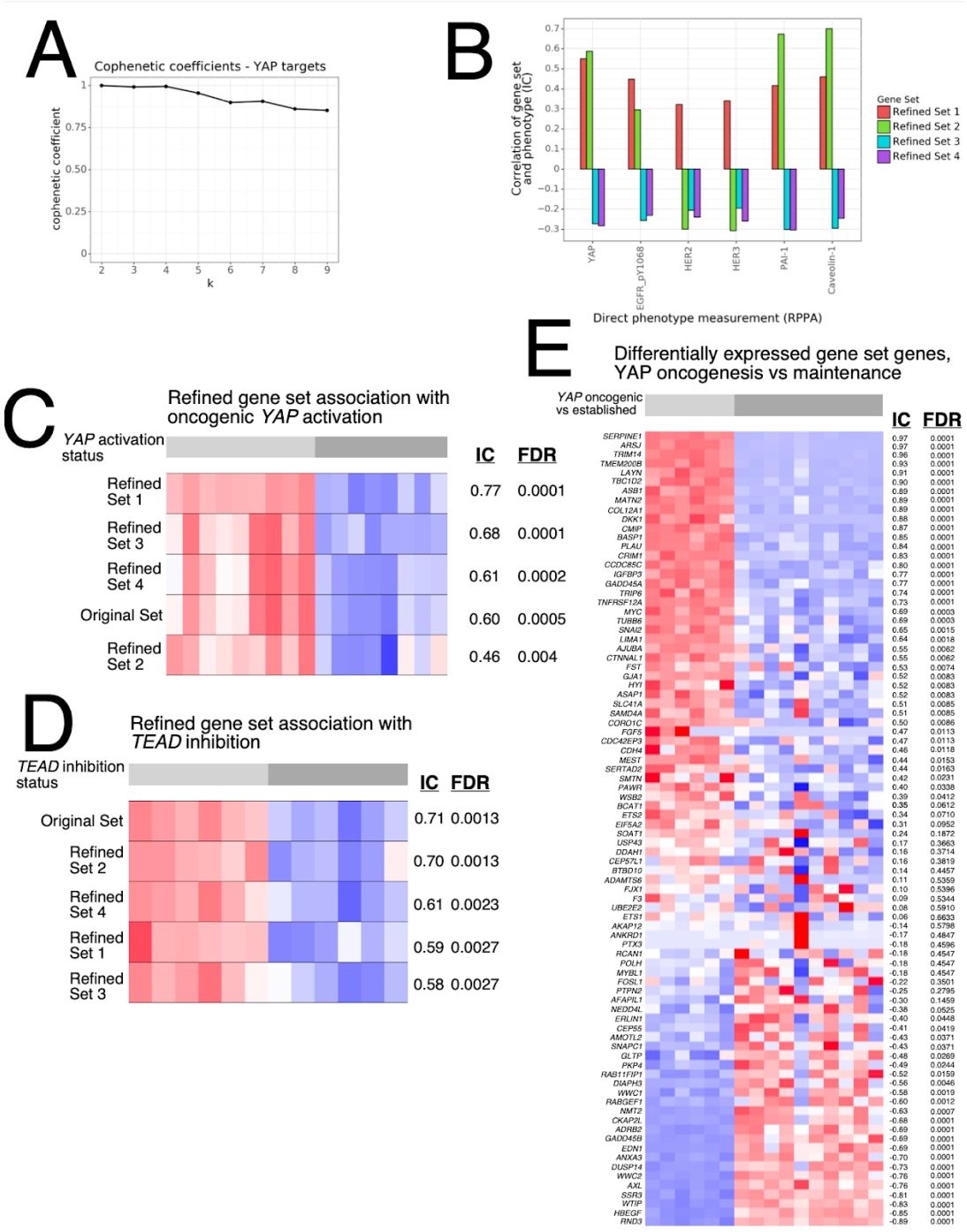


Figure 1.3: Refinement of a *YAP* signature

A) Cophenetic coefficients for refinement of *YAP* target genes. **B)** Median associations across bootstrapping iterations between gene set enrichment scores and direct phenotypic measurements. **C)** Enrichment of refined sets in mouse HNSCC tumors. Light gray: tumors expressing constitutively active *YAP*, dark gray: control samples. IC and FDR as in **Figure 2**. **D)** Enrichment of refined sets in Ca133 cell lines. Light gray: control samples, dark gray: *TEAD*i-treated samples. IC and FDR as in **Figure 2**. **E)** Expression of genes in Refined Set 1 and 2 in the *YAP*-active samples from **(D)** (light gray) and **(C)** (dark gray). IC and FDR as in **Figure 2**.

1.7 Supplementary Tables

Table S1.1: *ERBB2* refined gene sets.

Gene set name	**Gene set genes**
Refined Set 1	*FYN, PLCG1, PRKAR2B, ADCY1, NR4A1, AKT3, PRKACA, CAMK4, RAF1, MTOR, SOS1, THEM4, AKT2, GSK3A, PRKAR1A, RNF41, UBA52*
Refined Set 2	*GRB7, ERBB3, ERBB2, EREG, ITPR3, SRC, PRKCD, BTC, TRIB3, EGFR, ADCY6*
Refined Set 3	*NRG1, PRKCA, EGFR, MAPKAP1, SHC1, AKT3, BAD, HBEGF*

Table S1.2: *YAP* refined gene sets.

Gene set name	Gene set genes
Refined Set 1	*ANXA3, WWC1, TNFRSF12A, AJUBA, PAWR, F3, RAB11FIP1, TBC1D2, MYC, AMOTL2, CMIP, ETS2, RND3, PTPN2, MEST, C4BPB, GADD45B, CRIM1, CCDC85C, ADRB2, PKP4, PLAU, USP43, NEDD4L, EDN1, TRIM14, ARNTL2*
Refined Set 2	*AXL, CRIM1, SERPINE1, ETS1, COL12A1, GADD45B, NMT2, FOSL1, PTX3, LAYN, CDC42EP3, TNFRSF12A, AMOTL2, SNAPC1, TUBB6, RND3, ARSJ, CORO1C, ASAP1, SAMD4A, MYBL1, FGF5, GJA1, PLAU, DUSP14, WWC2, AJUBA, GADD45A, TBC1D2, FST, SNAI2, SOAT1, TRIP6, BASP1, FJX1, ASB1, DDAH1, LIMA1, ERLIN1, RABGEF1, CTNNAL1, RCAN1, WTIP, AFAP1L1, SLC41A1, DKK1, BCAT1, HYI, SMTN, IGFBP3, SSR3, ANKRD1, EIF5A2, TMEM200B, F3, ARNTL2, ADAMTS6, DIAPH3, GLTP, BTBD10, HBEGF, POLH, SERTAD2, AKAP12, CEP55, MATN2, ADRB2, WSB2, CDH4, CEP57L1, CKAP2L, UBE2E2*
Refined Set 3	*MAD2L1BP, FAM89A, PHACTR1, DPH5, SFXN4, URB2, CDC25A, NEIL2, MRPL24, C1QBP, TMEM201, PTRH2, TIMM8A, E2F3, NPM3, RRP1B, XPO5, TMEM209, CENPV, SRRD, TRMT1, SMC3, RPF2, UCK2, DARS2, BCS1L, GNL3, LZIC*

Table S1.2: YAP refined gene sets.

Gene set name	Gene set genes
Refined Set 4	TIMELESS, KIF18B, BUB1B, CDCA5, CDCA8, GINS1, KIF2C, NUP188, ATAD2, CDC6, KNTC1, RRM2, POLA2, MCM3, CENPF, CCNA2, DDX46, CDC25A, KIF14, EXOSC2, URB2, NUF2, SART3, LMNB2, SRSF2, ILF3, PSMC3IP, NUP85, KIF23, MCM7, PSRC1, TOP2A, XPO5, TROAP, ERCC6L, DTYMK, CENPL, SMC3, TDP1, POLE3, RRP1B, NASP, ZWILCH, SUV39H2, HNRNPM, CEP152, TTI1, POC1A, KIF20B, RCC1, POLR1A, C1orf109, CSE1L, TUBG1, PRPF4, TRA2B, CEP55, IQGAP3, DKC1, CKAP2L, TCP1, TK1, CDCA4, SRSF3, DEK, GEMIN4, THOC1, HAT1, RBL1, NUP50, TCOF1, MTHFD1, SEH1L, EFTUD2, KLHDC4, NAA25, NCL, SF3A3, NAA15, TMEM106C, CCDC137, SF3B3, ISG20L2, SET, UBE2G2, GPN3, MRPL9, GPATCH4, TOE1, SNRPF, TMEM201, DARS2, E2F3, TAF5L, UTP20, MTA2, ODC1, FARSA, DIS3L, PNPT1, PRMT5, METAP2, ZMYND19, CHUK, NUP93, CRY1, LSG1, KATNB1, LARP4B, HSPA14, CENPV, LARP4, PPRC1, ACAT2, TIMM8A, PKN3, C9orf40, RUVBL2, RAD18, C12orf65, TEAD4, CCDC15, PRMT1, TUBB, SRBD1, CDKN2, AIPNL, MUTYH, DDX21, GTPBP4, CEP57, NOP14, NAT10, SNRPD2, MIIP, CEP57L1, LSM3, CYC1, RBM22, BCS1, LUCK2, MTG1, DIAPH3, GNL3, TMEM209, ABHD10, MED30, DDX56, CCDC85C, NOC3L, DNAJA3, THOC6, BANF1, AIMP2, IRAK1, XPO6, ZCCHC8, HSPA9, UTP15, NEURL1B, CSPP1, B3GALNT2, MTPAP, NOL10, PKP4, TBCE, C1QBP, TRMT5, MTFP1, PCID2, ERLIN1, GNL2, RPUSD2, SPATA5L1, PHTF2, LZIC, NUDCD1, HAUS4, WDR74, MALSU1, POLH, USP36, TRMT1, OXNAD1, PPIF, NUP88, EID2, RTN4IP1, TSEN2, CUTC, NPM3, UBR4

1.8 Acknowledgements

Chapter 1, in full, is a reformatted presentation of the material currently being prepared for submission for publication as “Data driven refinement of expression signatures for enrichment analysis” by Alexander T. Wenzel, Farhoud Faraji, Kuniaki Sato, Kwat Medetgul-Ernar, Anthony Castanza, Romella Sagatelian, Gayathri Donepudi, Omar Halawa, Jean Y. J. Yang, J. Silvio Gutkind, Pablo Tamayo, and Jill P. Mesirov. The dissertation author was the primary investigator and author of this material.

Chapter 2 scGSEA: Adapting GSEA to sparse single cell data

2.1 Abstract

Gene Set Enrichment Analysis (GSEA) is a method for quantifying pathway and process activation in groups of samples, and its single sample version (ssGSEA) scores activation using mRNA abundance in a single sample. GSEA and ssGSEA were developed for "bulk" samples rather than individual cells technologies such as microarrays and bulk RNA-sequencing (RNA-seq) data. The growing use of single cell RNA-sequencing (scRNA-seq) raises the possibility of using ssGSEA to quantify pathway and process activation in individual cells. However, scRNA-seq data is much sparser than RNA-seq data. Here we show that ssGSEA as designed for bulk data is subject to some amount of score uncertainty and other technical issues when applied to individual cells from scRNA-seq data. We propose single cell GSEA (scGSEA), an adaptation of ssGSEA to scRNA-seq data based on a benchmarking of normalization methods and scoring metrics following the aggregation of "similar" cells, and we show how this method can be used to differentiate cell types based on gene set enrichment with greater certainty than ssGSEA applied to individual cells. We have made scGSEA available as a Python/R software package and a GenePattern module.

2.2 Introduction

The advent of scRNA-seq has enabled the profiling of gene expression in individual cells, a major advantage over past measurement of "bulk" samples which may mask inherent heterogeneity. scRNA-seq data can be used to quantify the abundance of fine grain cell types within a sample, giving rise to computational methods capable of labeling (Nofech-Mozes et al., 2023; Domínguez Conde et al., 2022) and modeling the interactions between these cell types (Armingol et al., 2021). Furthermore, with emerging spatial imaging technologies accompanying scRNA-seq, the expression patterns within individual cells can be interpreted in the context of their in situ arrangement and distribution, enabling a better understanding of healthy and diseased tissue function (Longo et al., 2021).

In the field of computational method development for scRNA-seq data, existing techniques for analyzing bulk data are often revisited to decide if they can be used "as-is", adapted to accommodate the differences between bulk and scRNA-seq data, or discarded in favor of new approaches (Chen et al., 2019). Among the most prominent differences between scRNA-seq, and a common cause of method incompatibility, is the sparsity of scRNA-seq data. The advantage of mapping RNA molecules back to their cell of origin comes with the drawback of substantially less material per observation (cell) than is available for a bulk RNA-seq observation (sampled admixture of cells). Therefore, computational methods applied to scRNA-seq data must accommodate a data distribution that is much sparser and includes fewer detected expressed genes per cell than bulk data (Chen et al., 2019).

Among the methods that may have applicability to scRNA-seq data is Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003, Subramanian et al., 2005). GSEA, as applied to bulk gene expression, first ranks all of the genes in the transcriptome by their differential

expression between two classes of samples. It then scores pre-existing gene sets by quantifying the degree to which gene set genes are non-uniformly distributed towards the top or bottom of the ranked list. While GSEA needs two groups of samples to compute a relative pathway enrichment score in one group compared to the other, single sample GSEA (ssGSEA) requires only a single sample and builds its ranked list based on the actual expression of genes in that sample (Barbie et al., 2009). When considering the problem of measuring pathway activity in single cell data, it appears attractive at first glance to apply ssGSEA to single cells, treating each one as a sample. In fact, we have already observed many users attempt this using ssGSEA as implemented in our GenePattern computational environment (Reich et al., 2006; Reich et al., 2017). We therefore sought to understand any issues with the application of ssGSEA to scRNA-seq data and what adaptations might be made to improve its performance with this type of data.

2.3 Results

Sparse data can cause instability in ssGSEA scores

We first assessed the performance of ssGSEA in six scRNA-seq datasets hosted on the CELLxGENE platform (Abdulla et al., 2023; Bhat-Nakshatri et al., 2021; MacParland et al., 2018; Zhang et al., 2021; Young et al., 2021; Fan et al., 2019; Solé-Boldo et al., 2019) (**Supplemental Table 2.1**) under varying conditions. All data were normalized according to the counts per 10,000 reads (CPTT) according to CELLxGENE specifications (**Table 2.1**). Datasets where a function other than the natural logarithm appeared to have been used in the CPTT computation were re-normalized. Each dataset was subset to only those genes shared among all six datasets in order to make enrichment scores comparable across datasets. Ten cells were randomly sampled from each dataset (**Figure 2.1B**). Gene sets for enrichment scoring were created by randomly sampling from the shared genes among all six datasets. Fifteen gene sets were created, five with 20 genes, five with 50 genes, and five with 100 genes (**Supplemental Table 2.2**). We used two different scoring metrics for evaluating gene set enrichment in a ranked list of genes. Both GSEA and ssGSEA compute a running sum along the ranked list, incrementing when a gene set member gene is encountered and decrementing otherwise. The weighted Kolmogorov-Smirnov (KS) statistic, most used in GSEA, is the maximum point, either positive or negative, along the running sum curve (Mootha et al., 2003; Subramanian et al., 2005). The area under the curve (AUC) statistic is the area under the running sum curve and is most used in ssGSEA (Barbie et al., 2009).

To assess the stability of enrichment scores under sparse conditions, we ranked genes based on their expression in each cell and permuted the order of tied genes in the ranked list. Genes are first grouped by their expression value, with all genes having the same expression

value placed in the same group. The order of these genes is then shuffled, and the permuted cell is formed by concatenating the shuffled groups. In this benchmarking, we performed 50 such permutations for each of the cells sampled from each dataset. The effect of these permutations is illustrated in the simplified example in **Figure 2.1A**, which demonstrates that it is possible to alter the enrichment score of a sparse sample (cell) with many genes having the same ranking (expression) value without changing the genes' expression or the gene set used. This is because the reordering of genes within groups of tied genes changes the peaks of the running sum curve, altering the AUC and, in some cases, the KS score. Because GSEA and ssGSEA break ties arbitrarily, both enrichment scores in **Figure 2.1A** are valid, resulting in uncertainty. We measure this uncertainty with the coefficient of variation (CoV), the standard deviation divided by the mean of the enrichment scores for all ranked list permutations, a higher CoV indicates more uncertainty in the enrichment score.

In general, the variance of the CoV itself increased with the number of genes expressed in the cell (**Figure 2.1C-E**), as more genes in the gene set being expressed offers more opportunities to perturb the enrichment score by moving gene set genes. Smaller gene sets were more likely to generate higher CoV values (**Figure 2.1C**), owing to the greater relative importance of moving one gene within a tied group of genes. AUC scores were more likely to have high AUC values than KS scores in cells with fewer genes expressed (**Figure 2.1D**), though we note that there are relatively few cells in our data with more than 3,000 genes expressed (**Figure 2.1B**). The KS score depends on the maximum point in the running sum statistic and having fewer gene set genes expressed offers fewer opportunities for perturbing the score. In the process of generating benchmarking, technical issues due to extreme sparsity emerged. If no members of a gene set are expressed at a level above 0, this will result in a division by zero

during the GSEA computation for either KS or AUC. For smaller gene sets, some enrichment scores could not be computed (**Figure 2.1F-2.1G**, **Supplementary Figures 2.1-2.10**). For sets where a score could be computed, scores tended to be relatively high (**Figure 2.1F-2.1G**), especially when using the KS scoring metric (**Figure 2.1G**). In a sparse sample, most of the signal for enrichment comes from the small fraction of strongly expressed genes in a cell. The reduced detection for moderately expressed genes results in more extreme enrichment scores driven by a small number of highly expressed genes. A major advantage of GSEA is the ability to detect collectively upregulated pathways even when no single gene is strongly differentially expressed (Mootha et al., 2005), and this advantage is reduced when computing enrichment on an extremely sparse single cell.

scGSEA: Adapting ssGSEA to high-sparsity conditions

As the sparsity of individual cells both causes uncertainty and prevents the computation of enrichment scores in some single cells with some gene sets, we concluded that aggregating multiple cells is preferable for using ssGSEA on scRNA-seq data. Here we describe our benchmarking of aggregation via averaging the expression of cells within each pre-annotated cell type in each CELLxGENE dataset. We benchmarked the performance of ssGSEA using four normalization methods (**Table 2.1**), as well as the two scoring metrics and 15 gene sets as described above. Each dataset was evaluated using raw counts and normalized with four normalization methods and the gene expression for cells within each annotated cell type were averaged. These aggregate average cells were subjected to the same ranked list permutation as described above.

Representative benchmarking results for one dataset are shown in **Figure 2.2A** and **Figure 2.2B** (for remaining datasets, see **Supplementary Figures 2.11-2.20**). For AUC scoring, we observed relatively little difference between non-normalized counts and normalized counts, indicating that sufficiently large groups of tied genes still exist to create uncertainty in the AUC score after normalization and aggregation (**Figure 2.2A**, **Supplementary Figures 2.11-2.20**). For KS scoring, most normalized scores had CoVs sufficiently small enough as to be considered effectively stable or constant (**Figure 2.2B**). Notably for some cell types, normalized data still resulted in relatively high CoV, usually for smaller gene sets (**Figure 2.2C**). For most cell types, the KS scores for unnormalized cells had uniformly higher CoV values than their normalized counterparts except for one cell type that contained 20,000 cells (**Figure 2.2D**). The distribution of counts for 20,000 averaged cells may begin to approach the distribution of a bulk RNA-seq sample depending on the transcriptional heterogeneity of the included cells, but a cluster or cell type that large in most contemporary datasets originating from a single experiment is uncommon. While both scoring metrics may yield high CoVs for some cell types and gene sets, all of the most stable scores were derived using the KS metric (**Figure 2.2E**). In fact, the combination of KS scoring with the centered-log-ratio (CLR) across cells (**Table 2.1**) resulted in the most stable scoring, with nearly all combinations of cell type and gene set yielding a CoV less than 10^{-5} (**Figure 2F**). Given these results, we selected KS as the scoring metric and the CLR applied across cells as the normalization method for scGSEA, as that combination yielded the most stable results.

scGSEA reduces uncertainty to separate cell types using expression signatures

As an application of scGSEA, we turned to the problem of separating cell types based on pathway enrichment scores. We obtained two datasets representing two immune cell populations in the Tabula Sapiens atlas (The Tabula Sapiens Consortium, 2022) as distributed in CELLxGENE (Abdulla et al., 2023) : spleen-resident B cells and bone marrow-resident plasma cells (**Figure 2.3A**). We selected two experimentally derived gene sets (Benson et al., 2012) from MSigDB (Liberzon et al., 2011) representing spleen-resident B cells and bone marrow-resident plasma cells respectively. We first used the currently release ssGSEA method in GenePattern (Reich et al. 2006), to score the two gene sets in each of the individual B cells and plasma cells (**Figure 2.3B**). In general, the B cell gene set scored higher in the B cells and the plasma cell gene set scored slightly higher in the plasma cells, but with a wide range of scores in both cases (**Figure 2.3B**). To understand the role of instability, we selected ten cells at random from each group and subjected them to the ranked list permutation approach described above. **Figure 2.3C** shows the range of scores obtained in each cell. Notably, the range of valid scores overlapped in many cases for pairs of cells from the two distinct populations, meaning that sparsity and the resulting uncertainty prevents the classification of individual cells as B cells or plasma cells based on cell type signature enrichment using ssGSEA.

We then used the scGSEA approach as described above, normalizing with the CLR across cells and creating an aggregate cell for each cell type. We then applied the same ranked list permutation approach to the two aggregate cells. Though we selected the KS score as the scoring metric based on scGSEA, we also evaluated the AUC as the scoring metric here. Though the uncertainty of AUC scores was greatly reduced (**Figure 2.3D**), the uncertainty was effectively eliminated with the KS score, as we had observed in the benchmarking (**Figure**

2.3E). Using this combination of normalization and KS scoring, we were able to separate the bone marrow plasma cell group from the spleen B cell group.

2.4 Discussion

To summarize, we propose scGSEA as an adaptation of ssGSEA to scRNA-seq data by applying an appropriate normalization across cells and using the KS score as opposed to the AUC score. We show that ssGSEA should not be applied to individual cells from scRNA-seq data, as the scores are either uncertain or cannot be computed due to the high sparsity of single cells. This work adds Gene Set Enrichment Analysis to the list of methods originally developed for bulk gene expression technologies, but which have utility when appropriately adapted to the scRNA-seq context. While the choice to aggregate cells is necessary to avoid uncertainty and cases where there is in fact no valid enrichment score for a given cell and gene set, this aggregation does not significantly reduce scRNA-seq's ability to yield more granular observations than are obtainable at the bulk level. Because the analysis of bulk data requires the often-oversimplifying assumption that a bulk sample is a homogeneous collection of cells, generating enrichment scores in different subtypes of cells is still a major advantage, even if all of the cells within each cell type must be averaged. Importantly, scGSEA does not require that cells be aggregated specifically by cell type. Any meaningful grouping of cells (e.g., metadata elements, clustering) is valid. One may even aggregate cells based on their similarity along a continuous axis, such as a pseudotemporal trajectory representing some process, as we demonstrated previously (Wenzel, Champa et al., 2022). In fact, where spatial imaging data is available, one may also aggregate cells that are physically co-located in tissue and exhibit similar expression patterns. Nevertheless, an enrichment method capable of operating in conditions of extreme sparsity by treating groups of tied genes the same, no matter how they are ordered within the group, could allow for more robust enrichment scoring in individual cells, allowing for the profiling of variance and heterogeneity of pathway activity within groups of cells, and we

leave such a method to potential future work. To conclude, scGSEA successfully adapts ssGSEA to the sparse scRNA-seq context by reducing the uncertainty of enrichment scoring and enabling the quantification of activated pathways and processes within groups of similar cells.

2.5 Figures

Figure 2.1: Ranked list permutations of individual cells.

A) An example of permuting the ranked list of genes in a cell and the resulting ssGSEA scores. Red: GSEA running sum for original list of genes sorted by expression. Blue: GSEA running sum for list of genes where order within genes with equivalent expression has been shuffled. Vertical lines: Gene set gene locations for the original ranked list (Red) and permuted ranked list (Blue). **B)** Distribution of the number of genes expressed in randomly sampled cells. **C-E)** Coefficient of variation for individual cells subjected to ranked list permutations (n = 50) as shown in **(A)**, colored by **(C)** gene set size, **(D)** scoring metric, and **(E)** dataset. **(F-G)** Distribution of enrichment scores in a single dataset "Aging skin". Each panel is a gene set, each row of panels is a gene set size. Each box represents a single cell and each point represents the enrichment score of a single ranked list permutation of that cell. Scores are separated into **(F)** AUC and **(G)** KS.

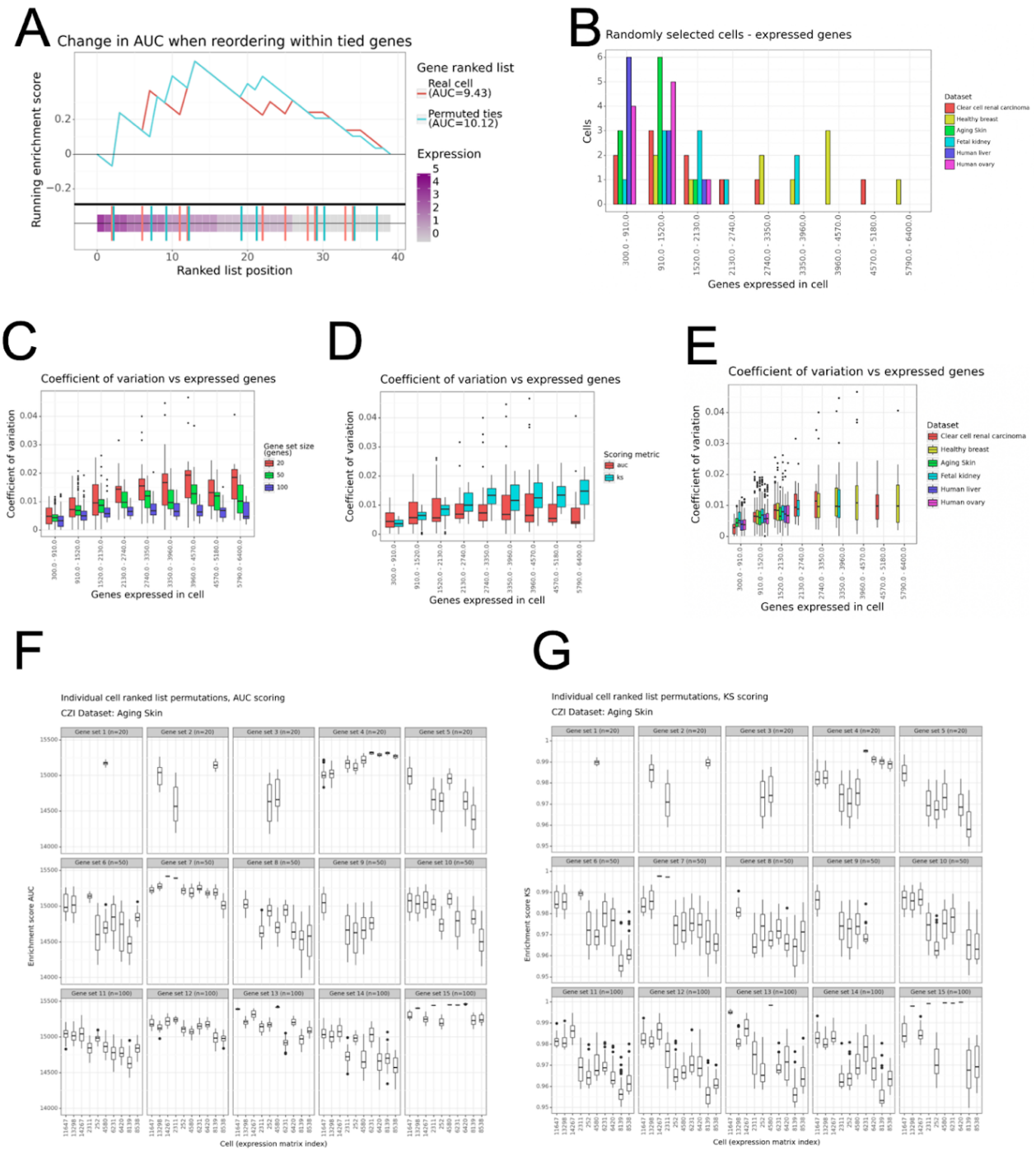

A
Change in AUC when reordering within tied genes
Running enrichment score
Ranked list position
Gene ranked list
Real cell (AUC=9.43)
Permuted ties (AUC=10.12)
Expression
B
Randomly selected cells - expressed genes
Cells
Genes expressed in cell
Dataset
Clear cell renal carcinoma
Healthy breast
Aging Skin
Fetal kidney
Human liver
Human ovary
C
Coefficient of variation vs expressed genes
Coefficient of variation
Genes expressed in cell
Gene set size (genes)
D
Coefficient of variation vs expressed genes
Scoring metric
auc
ks
E
Coefficient of variation vs expressed genes
F
Individual cell ranked list permutations, AUC scoring
CZI Dataset: Aging Skin
Enrichment score AUC
Cell (expression matrix index)
G
Individual cell ranked list permutations, KS scoring
CZI Dataset: Aging Skin
Enrichment score KS

Figure 2.2: Ranked list permutation of aggregate cells
A-B) Distribution of CoV values for each cell type in the "Aging skin" dataset. Each point represents the CoV of enrichment scoring for one gene set on permutations of a ranked list following normalization and averaging of cells in the cell type. Results are separated by **(A)** AUC and **(B)** KS scores. **C-D)** CoV of enrichment scores on ranked list permutations versus number of cells in the given cell type. Each point is the CoV of enrichment scores for a single cell type, gene set, normalization, and scoring metric. Points are colored by **(C)** gene set size, **(D)** normalization method, and **(E)** scoring metric. **F)** Fraction of CoV values below 1e-5 for each choice of normalization method and scoring metric.

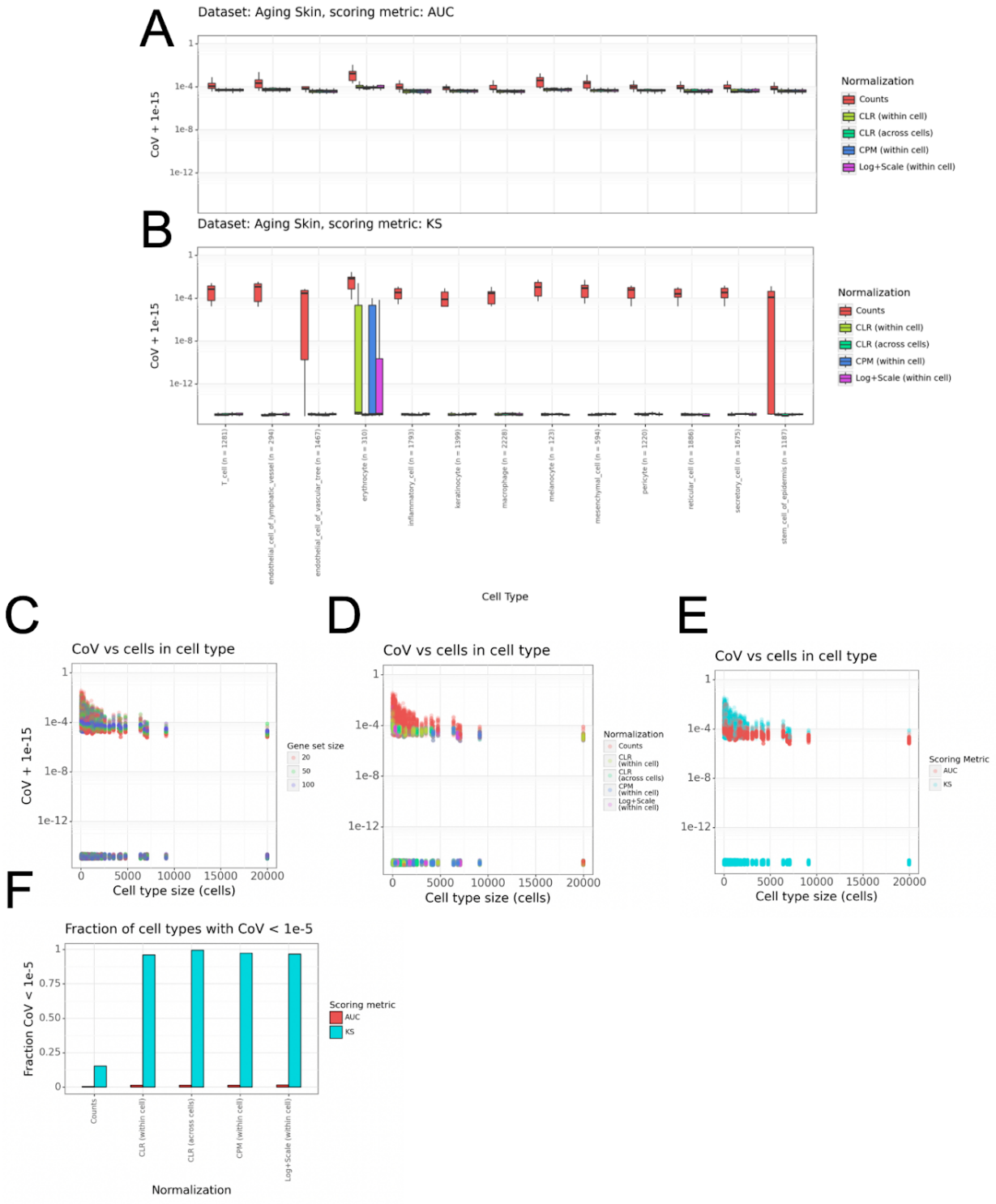
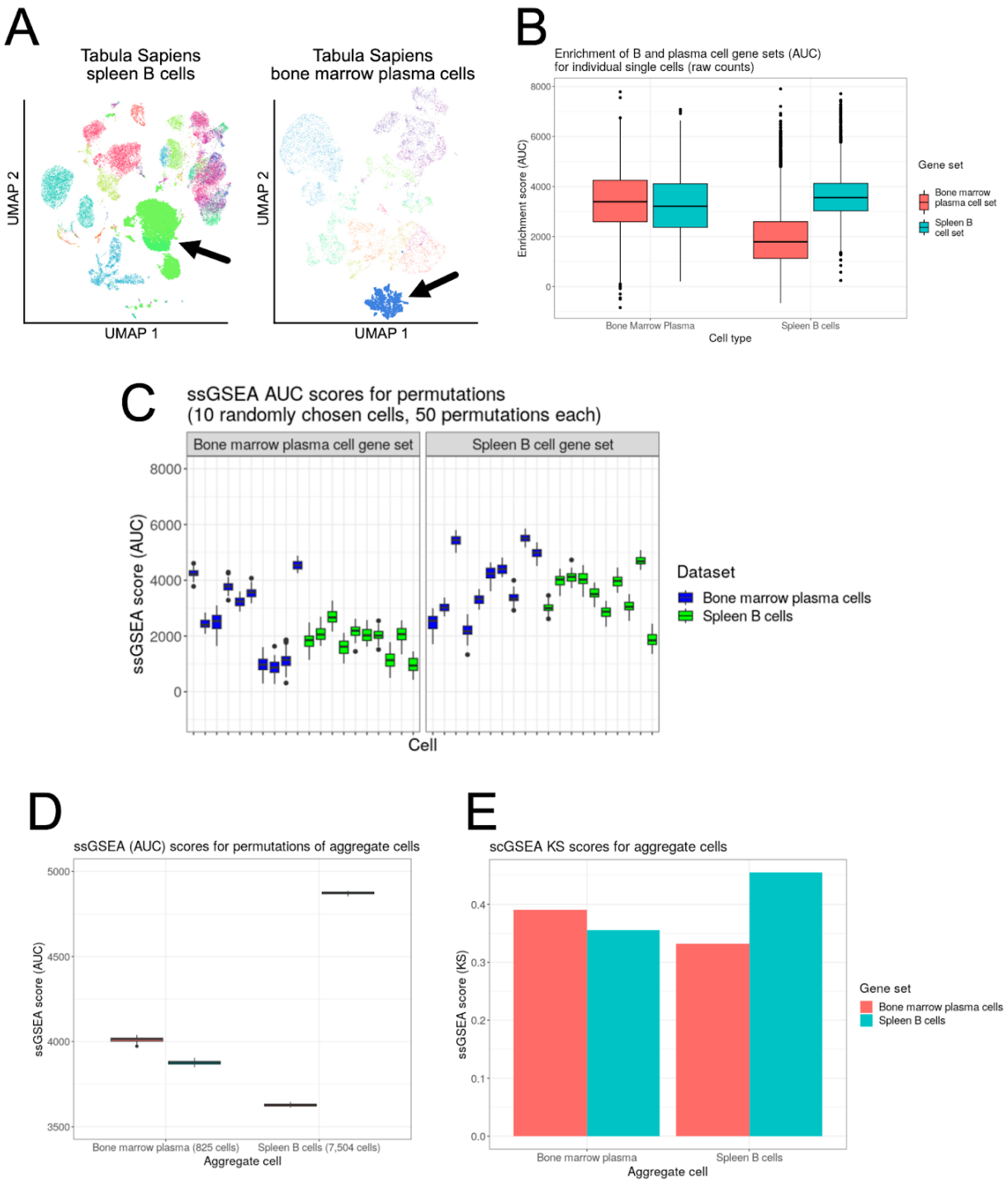

A
Dataset: Aging Skin, scoring metric: AUC
CoV + 1e-15
1
1e-4
1e-8
1e-12
Normalization
Counts
CLR (within cell)
CLR (across cells)
CPM (within cell)
Log+Scale (within cell)
B
Dataset: Aging Skin, scoring metric: KS
T_cell (n = 1281)
endothelial_cell_of_lymphatic_vessel (n = 294)
endothelial_cell_of_vascular_tree (n = 1467)
erythrocyte (n = 310)
inflammatory_cell (n = 1793)
keratinocyte (n = 1399)
macrophage (n = 2228)
melanocyte (n = 123)
mesenchymal_cell (n = 594)
pericyte (n = 1220)
reticular_cell (n = 1886)
secretory_cell (n = 1675)
stem_cell_of_epidermis (n = 1187)
Cell Type
C
CoV vs cells in cell type
Gene set size
20
50
100
0
5000
10000
15000
20000
Cell type size (cells)
D
CoV vs cells in cell type
E
CoV vs cells in cell type
Scoring Metric
AUC
KS
F
Fraction of cell types with CoV < 1e-5
Fraction CoV < 1e-5
0
0.25
0.50
0.75
1
Scoring metric
Counts
CLR (within cell)
CLR (across cells)
CPM (within cell)
Log+Scale (within cell)
Normalization

Figure 2.3: scGSEA separates B cells and plasma cells with more certainty

A) CELLxGENE-hosted datasets for Tabula Sapiens spleen cells (left) and bone marrow cells (right). UMAP plots are as rendered by the CELLxGENE data explorer. Arrows point to spleen B cells (lime green, left) and bone marrow plasma cells (blue, right). **B)** Enrichment scores for the spleen B cell gene set (teal) and bone marrow plasma cell gene set (teal) scores in the Tabula Sapiens bone marrow plasma cells (left) and spleen B cells (right). **C)** Scores for 10 randomly selected cells from the two Tabula Sapiens datasets following ranked list permutations (n = 50) applied to each cell. **D-E)** Scores for ranked list permutations (n = 50) following CLR normalization and averaging of cells within the spleen B cell group (teal) and bone marrow plasma cell group (salmon). Scores are split into **(D)** AUC and **(E)** KS. **(E)** is rendered as a bar plot because all scores were the same across permutations.


A
Tabula Sapiens
spleen B cells
Tabula Sapiens
bone marrow plasma cells
UMAP 2
UMAP 1
UMAP 2
UMAP 1
B
Enrichment of B and plasma cell gene sets (AUC)
for individual single cells (raw counts)
Enrichment score (AUC)
8000
6000
4000
2000
0
Bone Marrow Plasma
Spleen B cells
Cell type
Gene set
Bone marrow plasma cell set
Spleen B cell set
C
ssGSEA AUC scores for permutations
(10 randomly chosen cells, 50 permutations each)
Bone marrow plasma cell gene set
Spleen B cell gene set
ssGSEA score (AUC)
8000
6000
4000
2000
0
Cell
Dataset
Bone marrow plasma cells
Spleen B cells
D
ssGSEA (AUC) scores for permutations of aggregate cells
ssGSEA score (AUC)
5000
4500
4000
3500
Bone marrow plasma (825 cells)
Spleen B cells (7,504 cells)
Aggregate cell
E
scGSEA KS scores for aggregate cells
ssGSEA score (KS)
0.4
0.3
0.2
0.1
0.0
Bone marrow plasma
Spleen B cells
Aggregate cell
Gene set
Bone marrow plasma cells
Spleen B cells

2.6 Tables

Table 2.1: Normalization methods

Normalizations used in benchmarking analysis. For each formula, let *E* be a *m* x *n* scRNA-seq expression matrix of *m* genes and *n* cells. Each formula gives the normalized expression value for the raw counts for gene *g* and cell *c* from *E*.

Normalization	**Description**	**Formula**
Counts per ten thousand (CPTT)	The natural log of the number gene counts (with pseudocount) per 10,000 counts.	$ln\left(1e5 * \frac{E_{g,c}+1}{\sum_{j=0}^{m} E_{j,c}}\right)$
Centered log ratio (CLR) within cells	The natural log of the expression value divided by the geometric mean of the counts in the cell.	$\ln\left(\frac{E_{g,c}}{e^{\frac{\sum_{i=0}^{m}\ln(E_{j,c}+1)}{m}}}\right)$
Centered log ratio (CLR) across cells	The natural log of the expression value divided by the geometric mean of counts for the gene in all cells	$\ln\left(\frac{E_{g,c}}{e^{\frac{\sum_{i=0}^{n}\ln(E_{g,i}+1)}{n}}}\right)$
Counts per million (CPM)	The expression value divided by the sum of all counts in the cell, scaled by one million.	$\frac{E_{g,c}}{\sum_{j=0}^{m} E_{j,c}} * 1e^{6}$
Log+Scale	The natural log of the expression value divided by the sum of all counts with a pseudocount, scaled by 10,000.	$ln\left(\frac{E_{g,c}}{\sum_{j=0}^{m} E_{j,c}}+1\right) * 1e^{5}$

2.7 Supplementary Figures

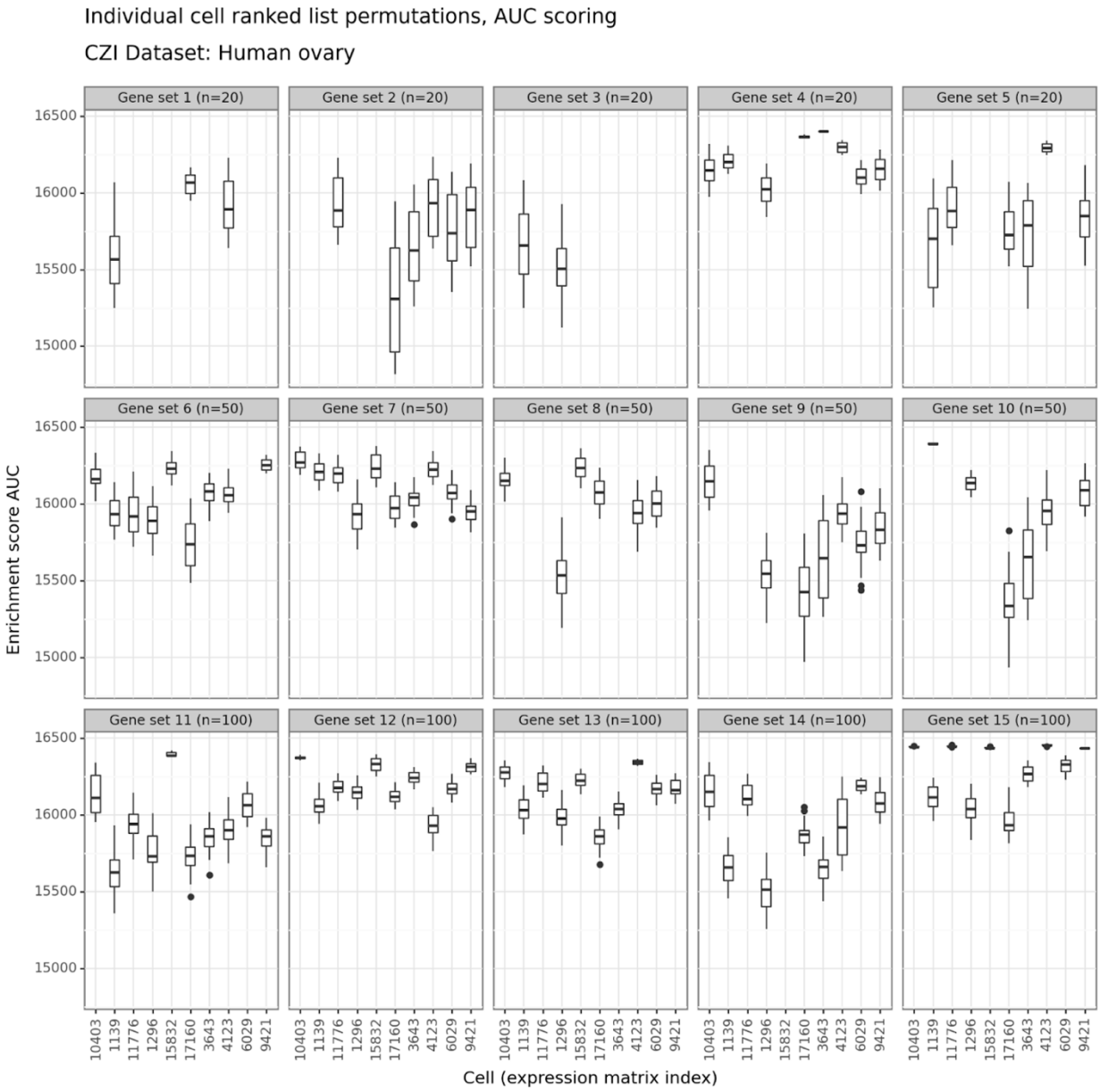


Figure S2.1: Individual cell permutations – Human ovary – AUC scores

Figure features as in **Figure 1F**

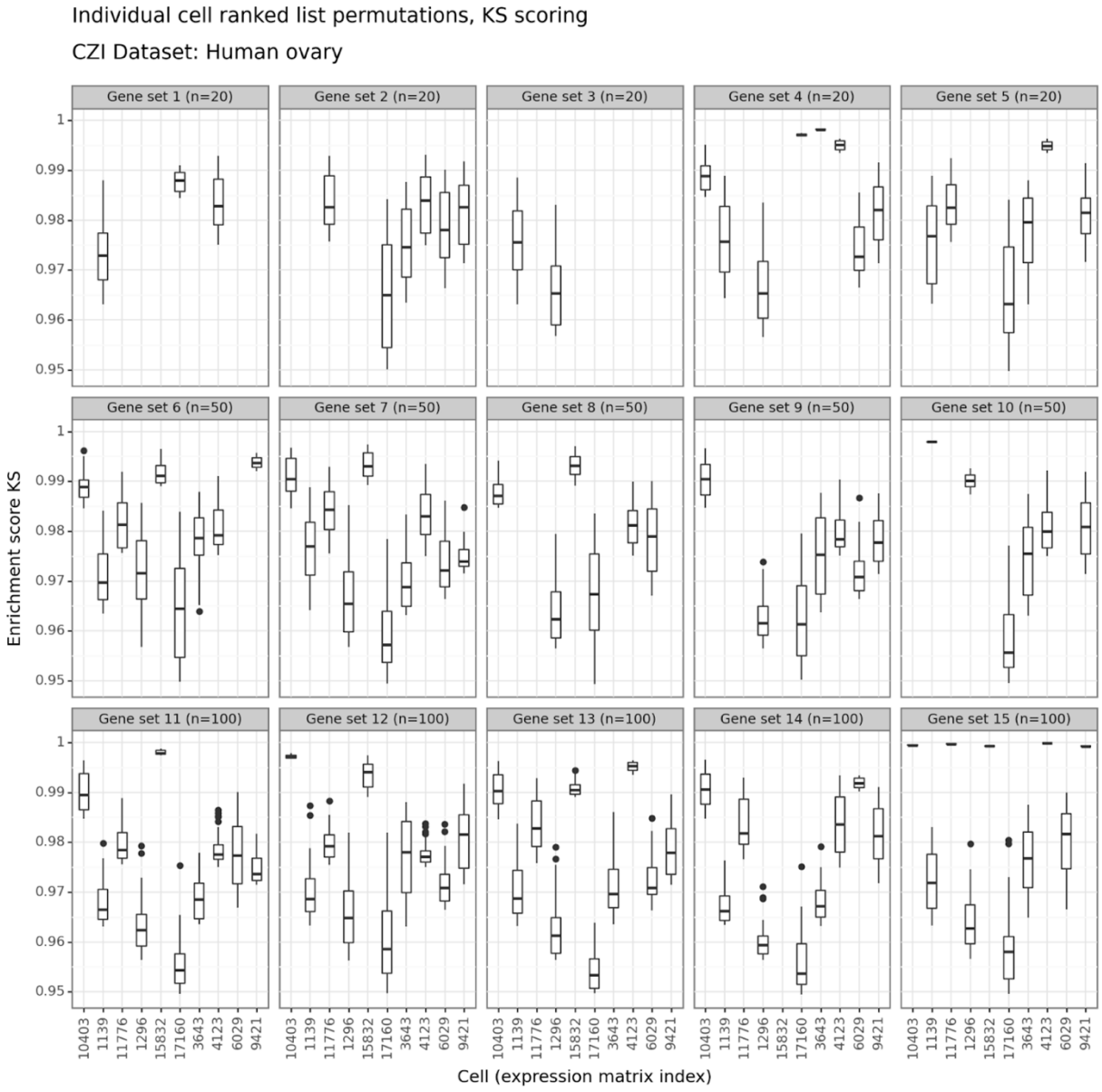


Figure S2.2: Individual cell permutations – Human ovary – KS scores

Figure features as in **Figure 1G**

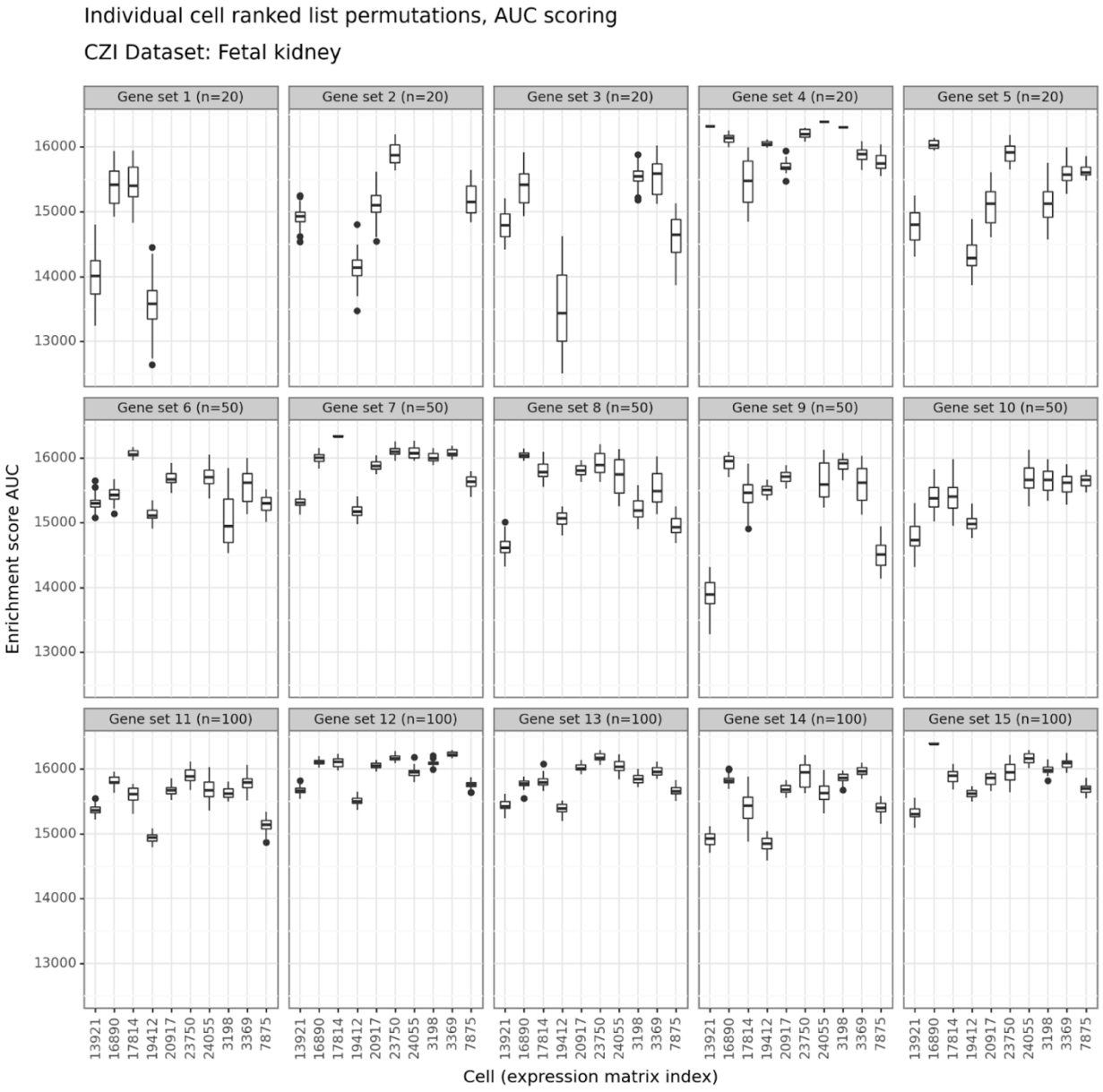


Figure S2.3: Individual cell permutations – Fetal kidney – AUC scores

Figure features as in **Figure 1F**

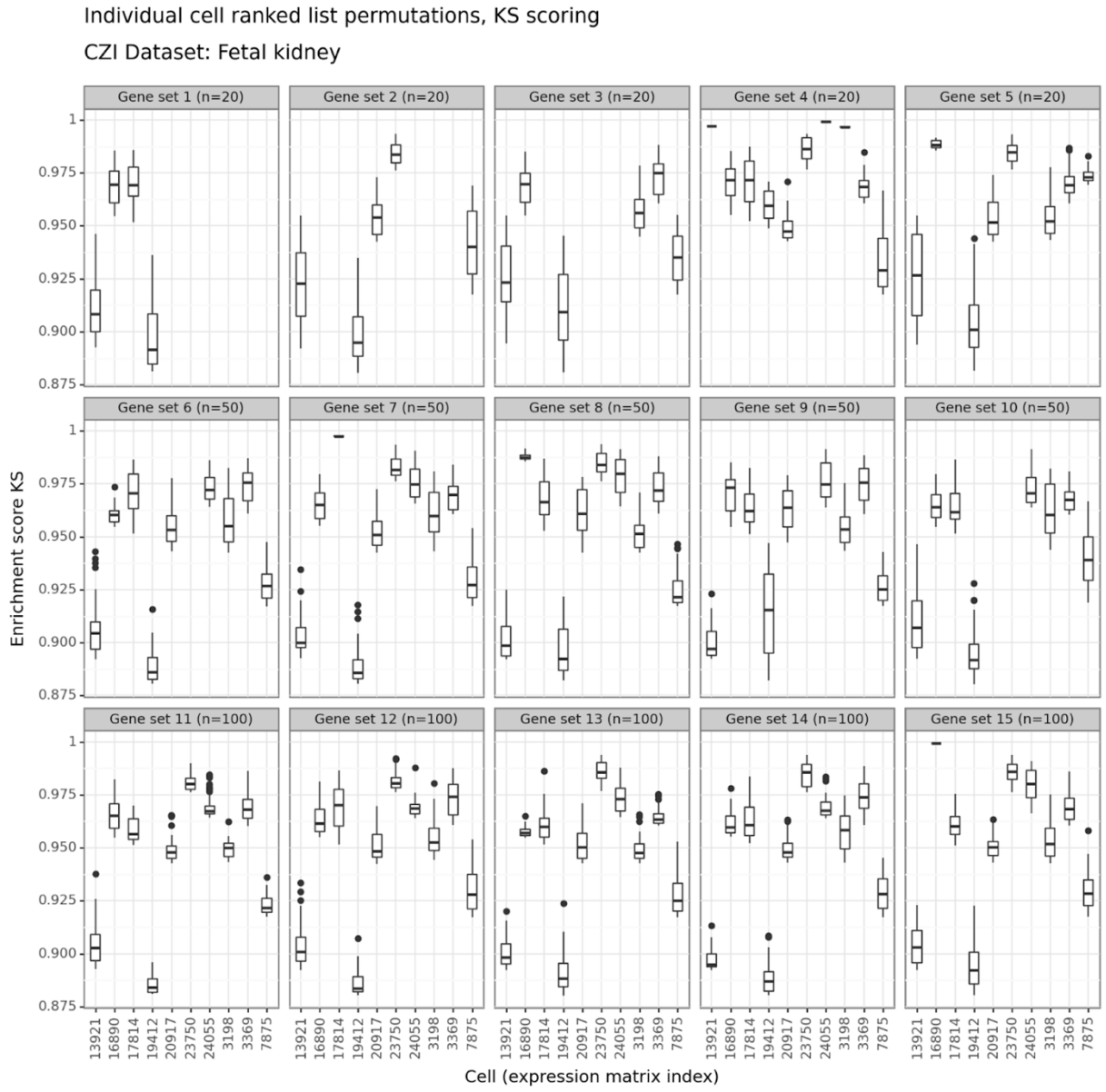


Figure S2.4: Individual cell permutations – Fetal kidney – KS scores

Figure features as in **Figure 1G**

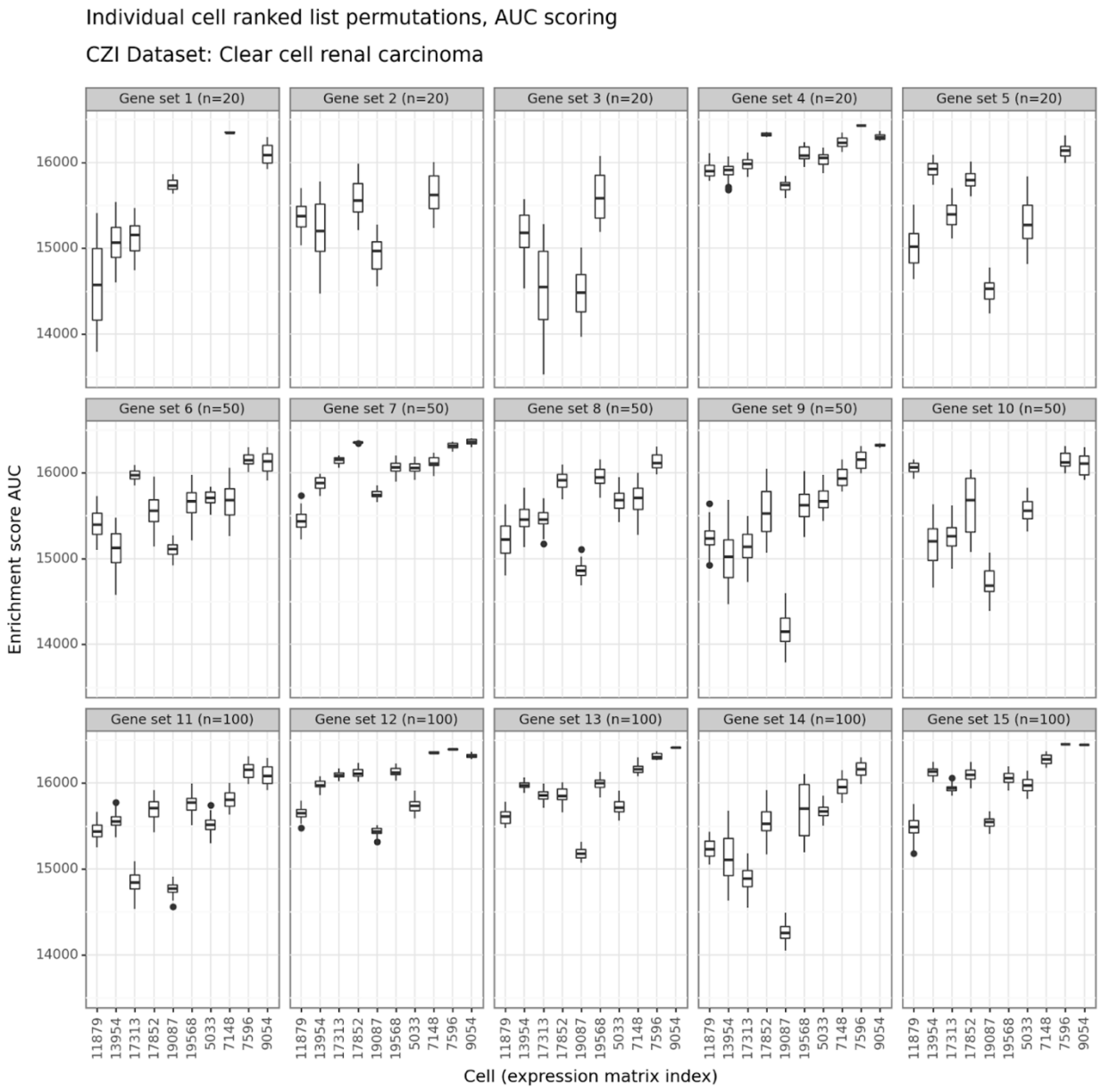


Figure S2.5: Individual cell permutations – Clear cell renal carcinoma – AUC scores

Figure features as in **Figure 1F**

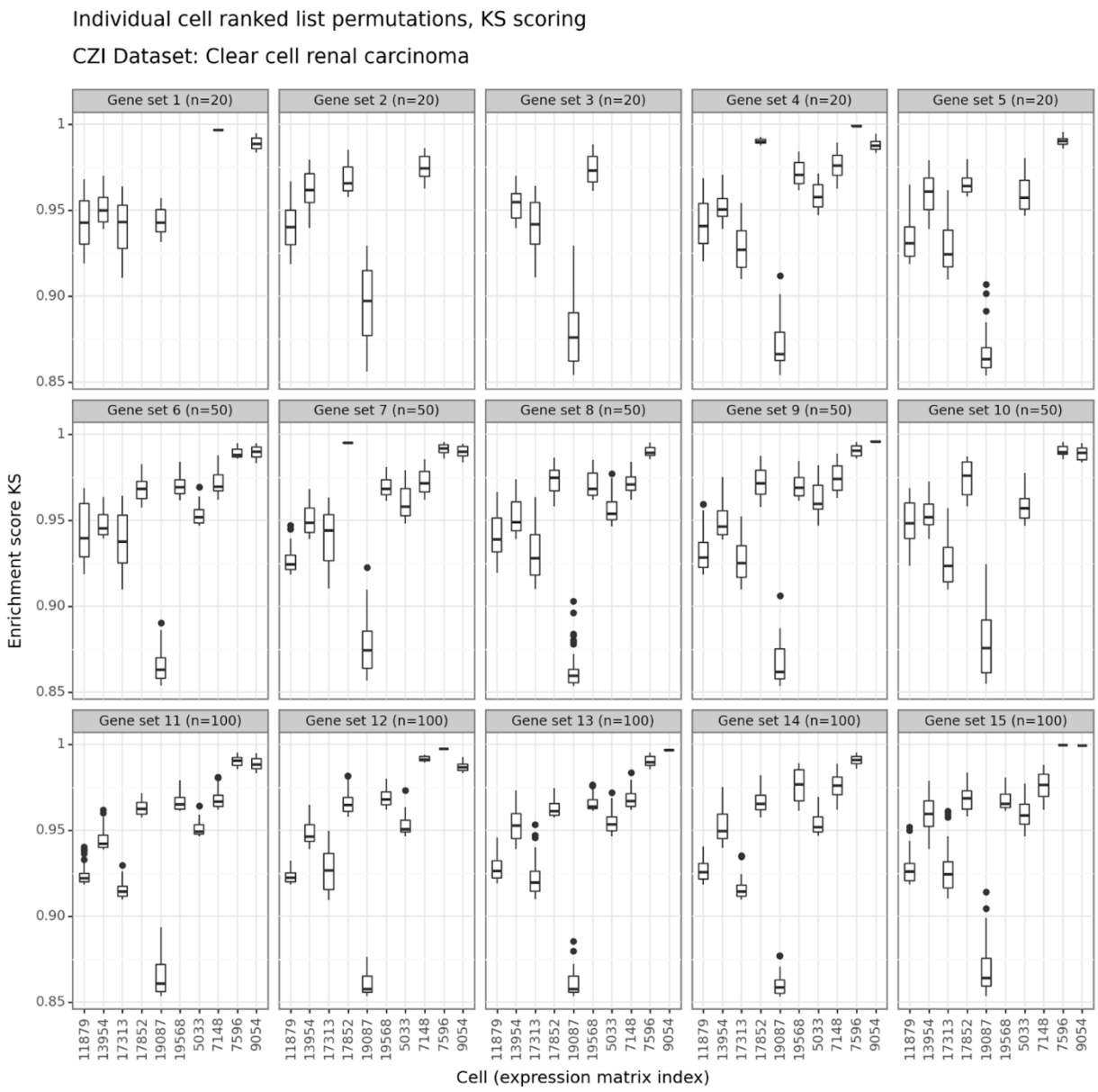


Figure S2.6: Individual cell permutations – Clear cell renal carcinoma – KS scores

Figure features as in **Figure 1G**

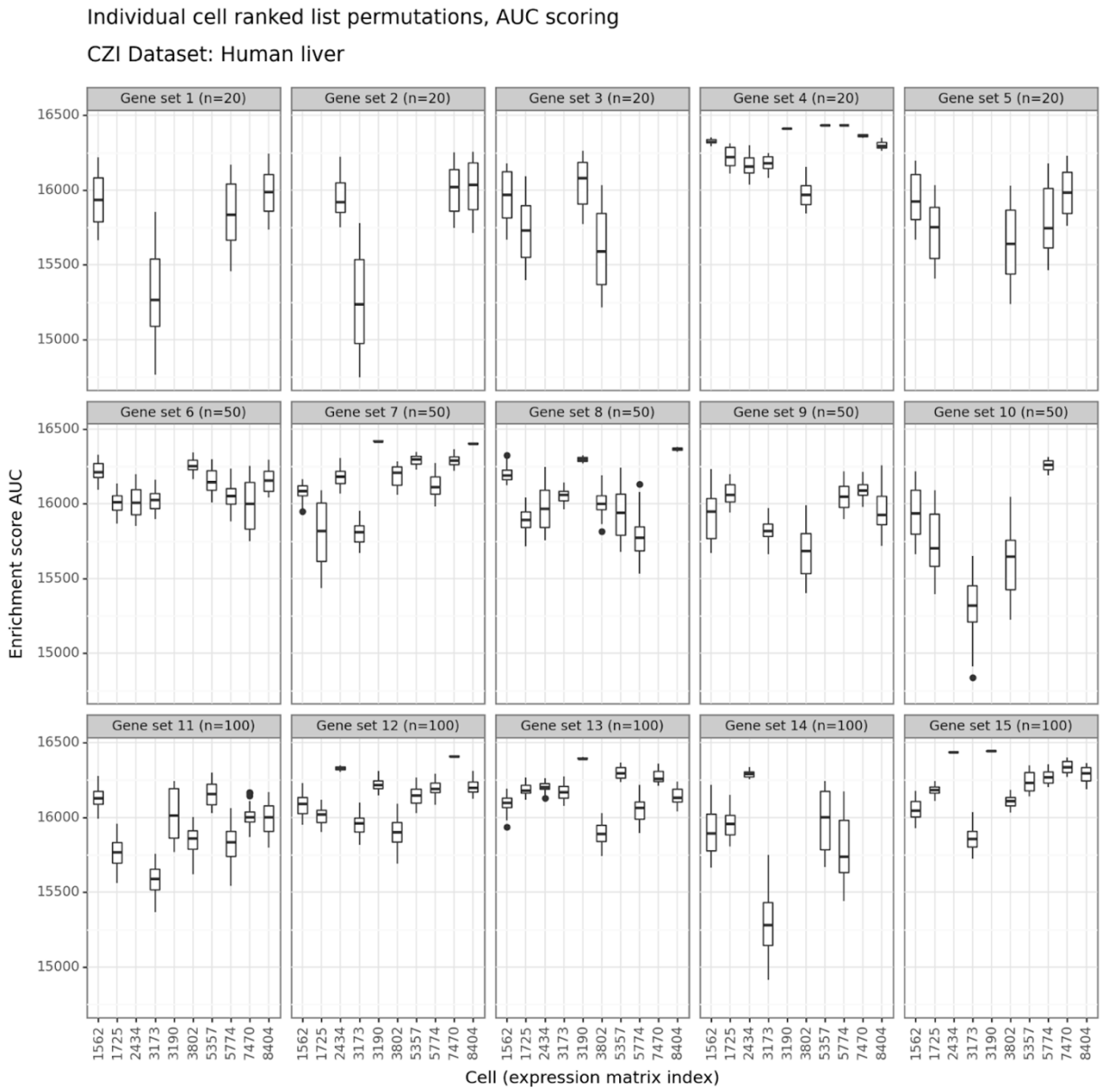


Figure S2.7: Individual cell permutations – Human liver – AUC scores

Figure features as in **Figure 1F**

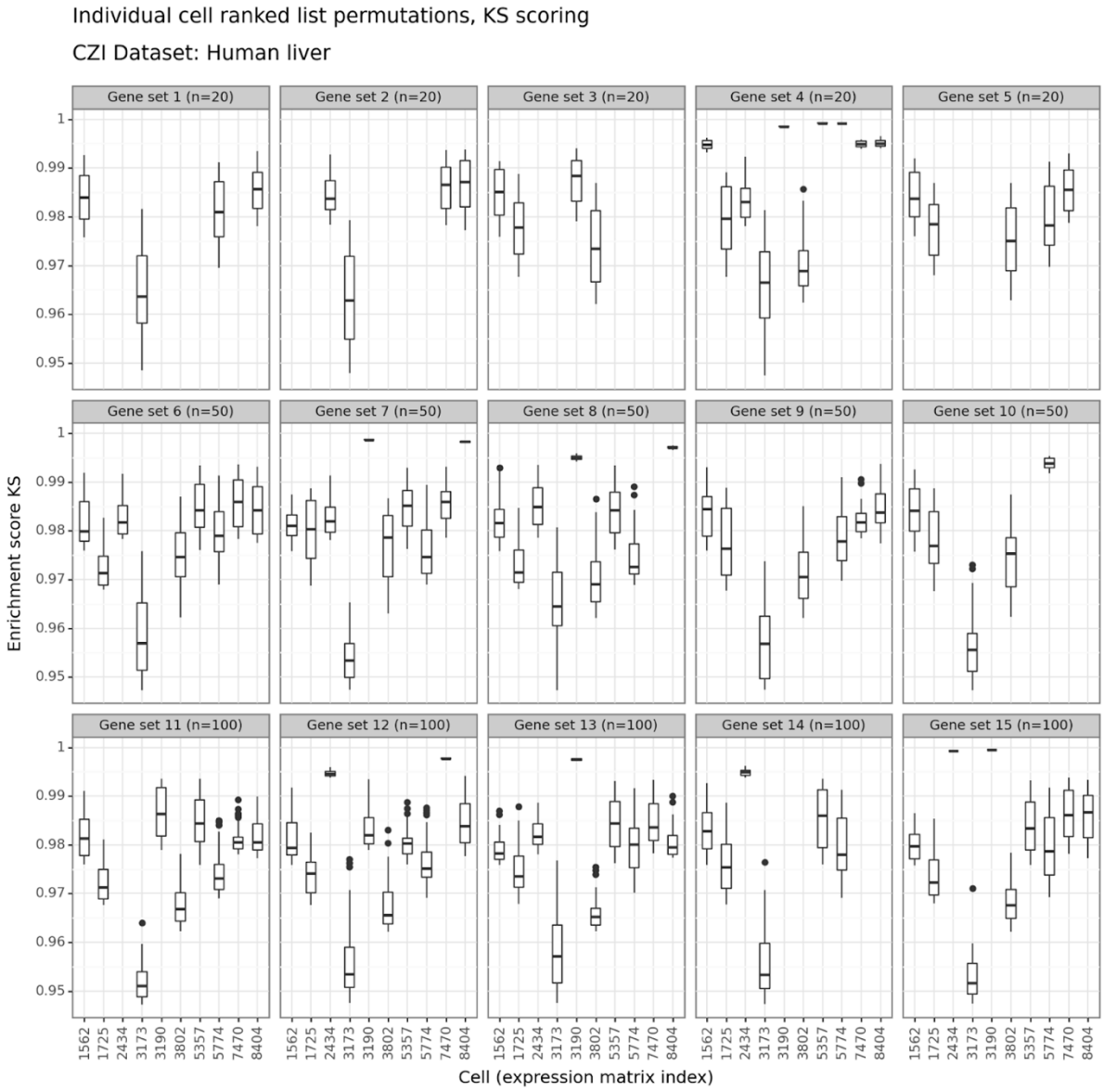


Figure S2.8: Individual cell permutations – Human liver – KS scores

Figure features as in **Figure 1G**

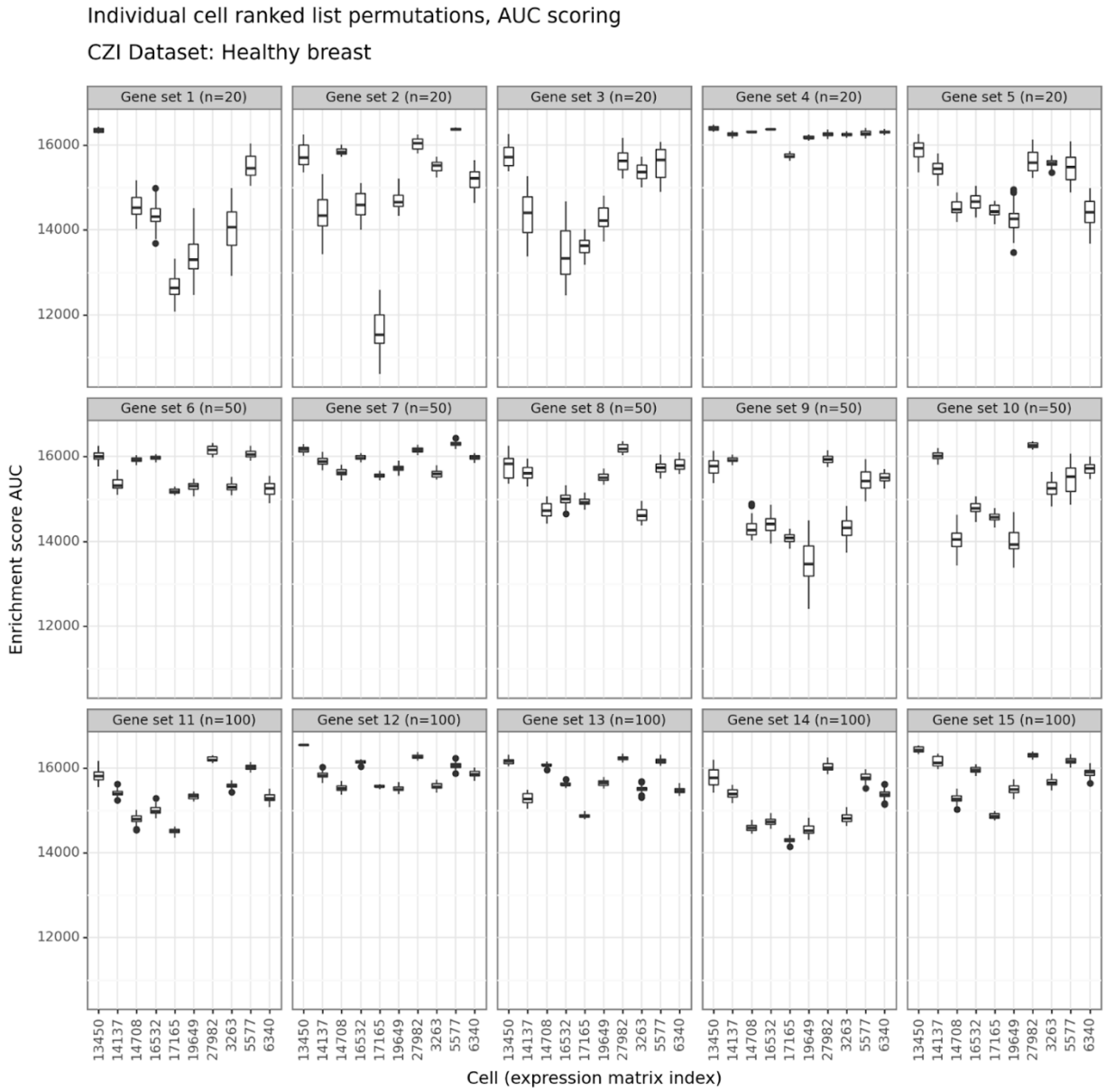


Figure S2.9: Individual cell permutations – Healthy breast – AUC scores

Figure features as in **Figure 1F**

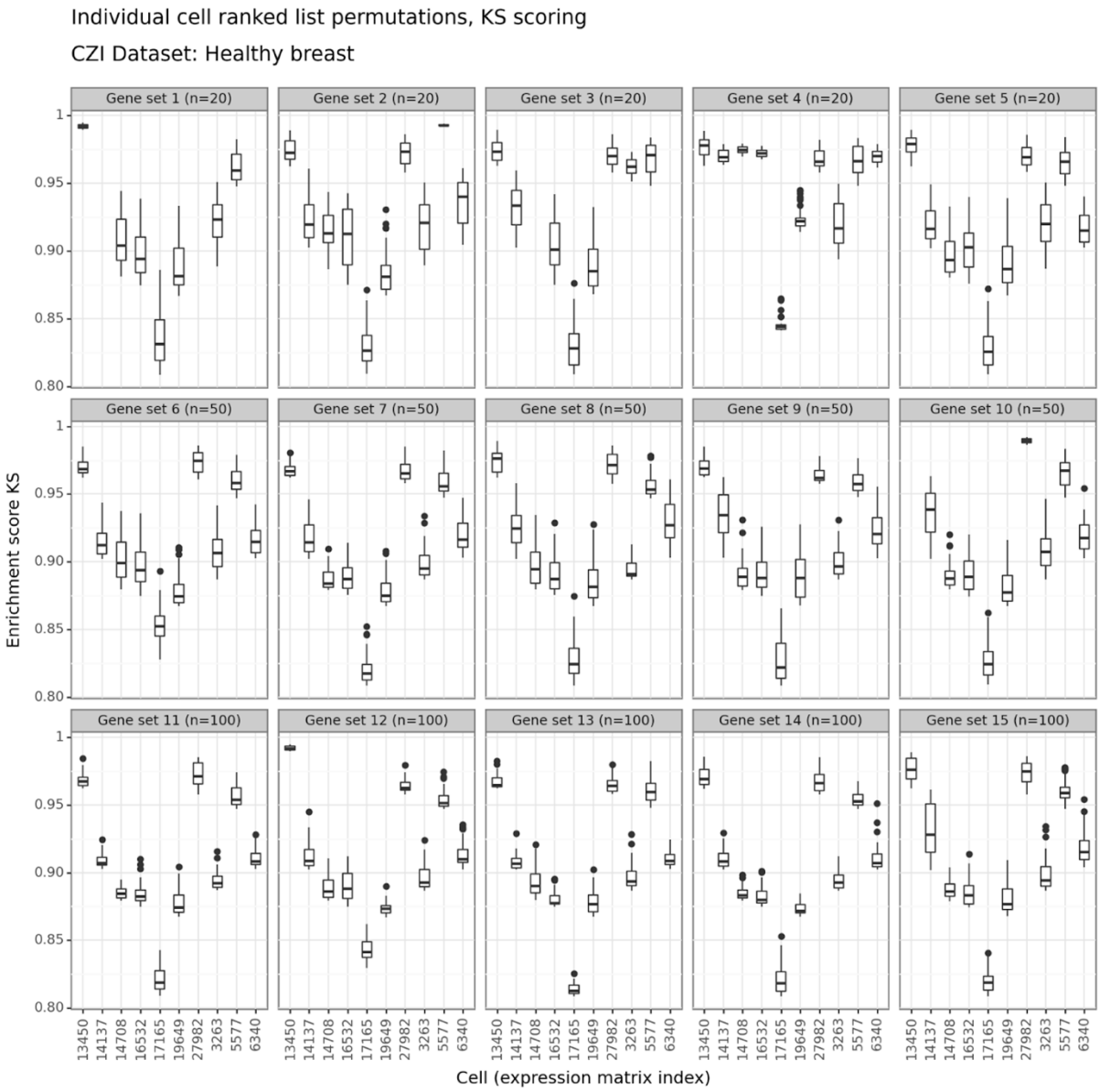


Figure S2.10: Individual cell permutations – Healthy breast – KS scores

Figure features as in **Figure 1G**

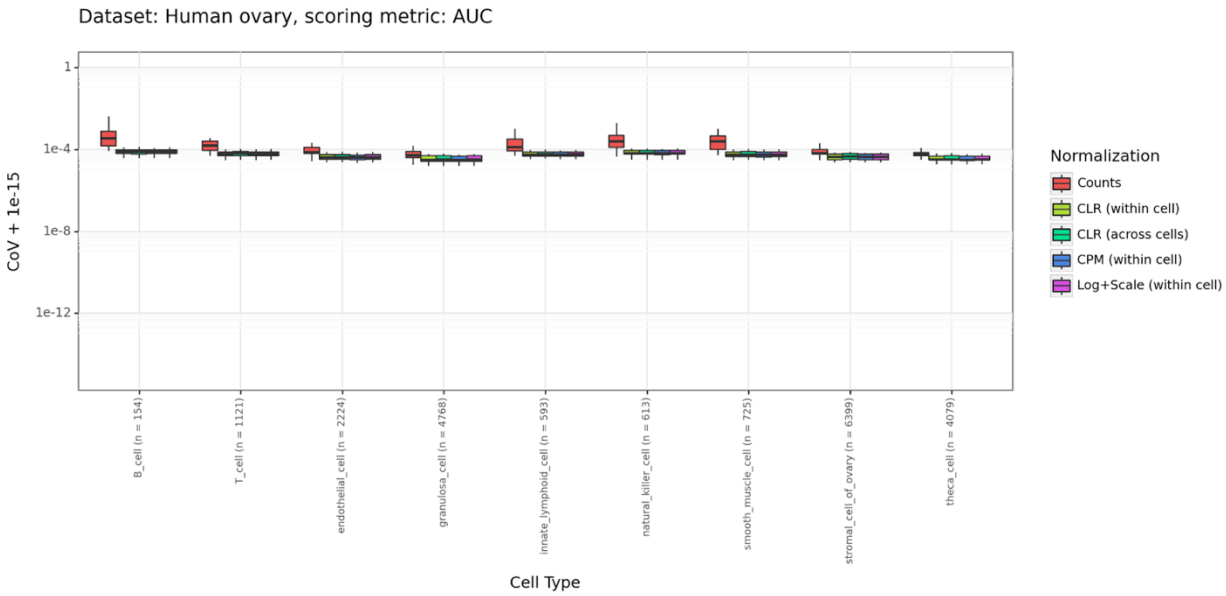


Figure S2.11: Aggregate cell permutations – Human ovary – AUC scores

Figure features as in **Figure 2A**

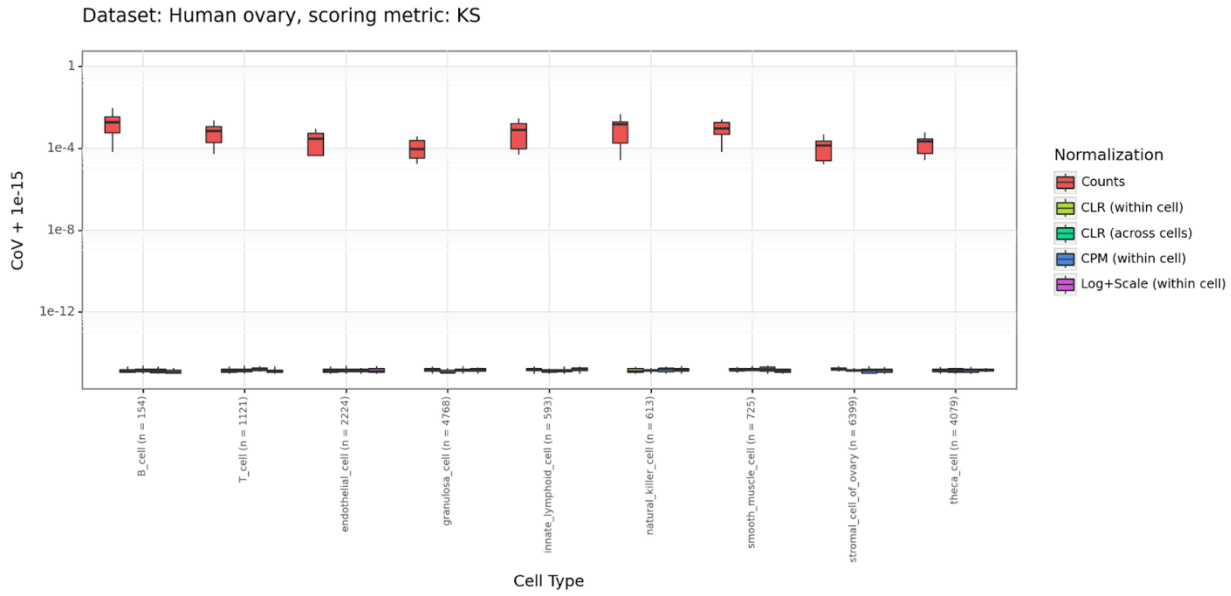


Figure S2.12: Aggregate cell permutations – Human ovary – KS scores

Figure features as in **Figure 2B**

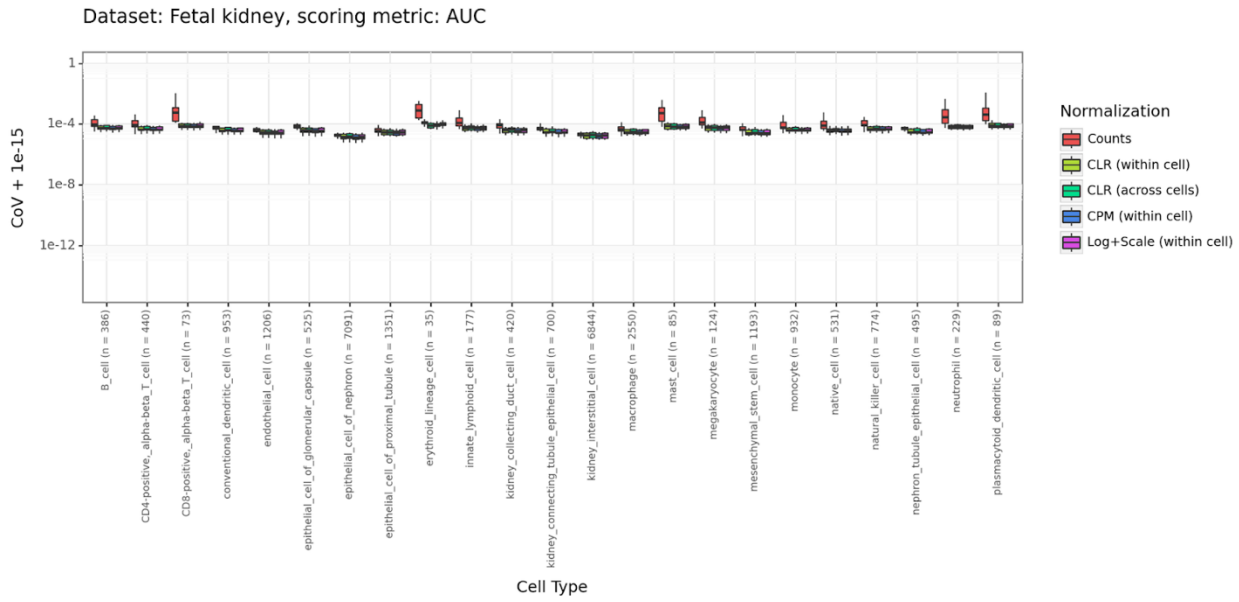


Figure S2.13: Aggregate cell permutations – Fetal kidney – AUC scores

Figure features as in **Figure 2A**

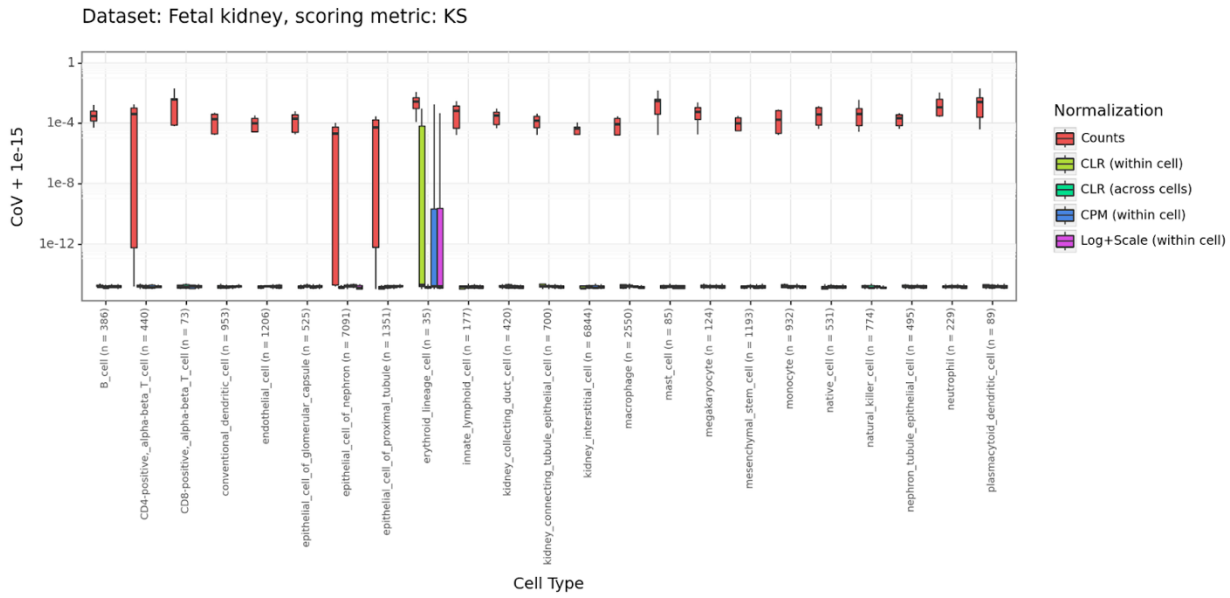


Figure S2.14: Aggregate cell permutations – Fetal kidney – KS scores

Figure features as in **Figure 2B**

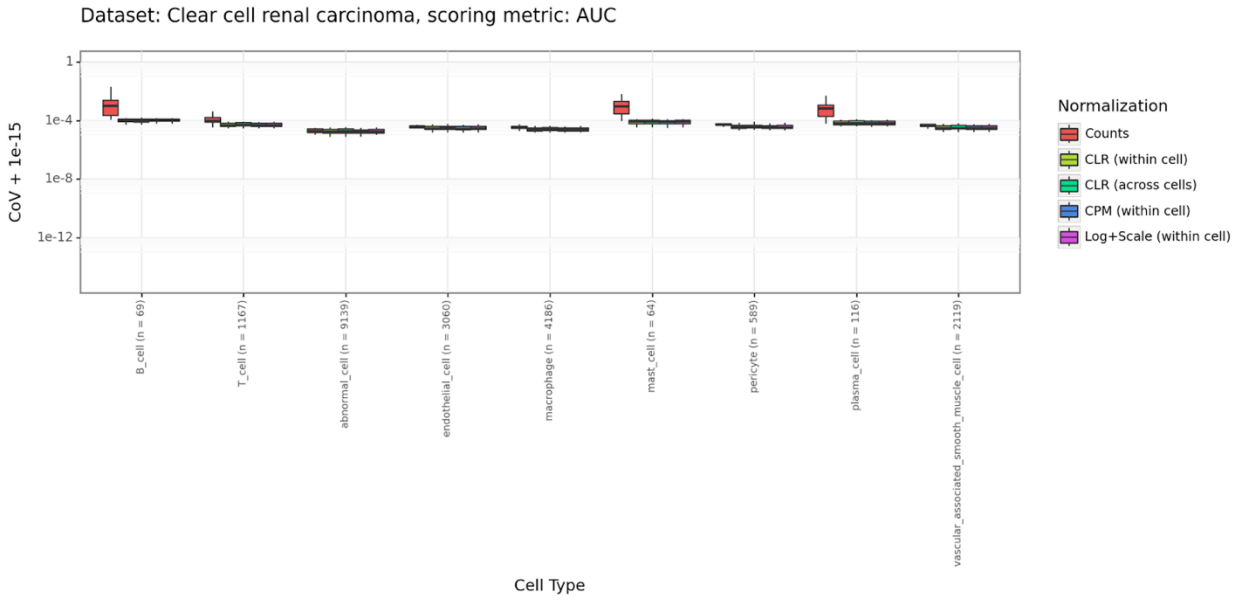


Figure S2.15: Aggregate cell permutations – Clear cell renal carcinoma – AUC scores

Figure features as in **Figure 2A**

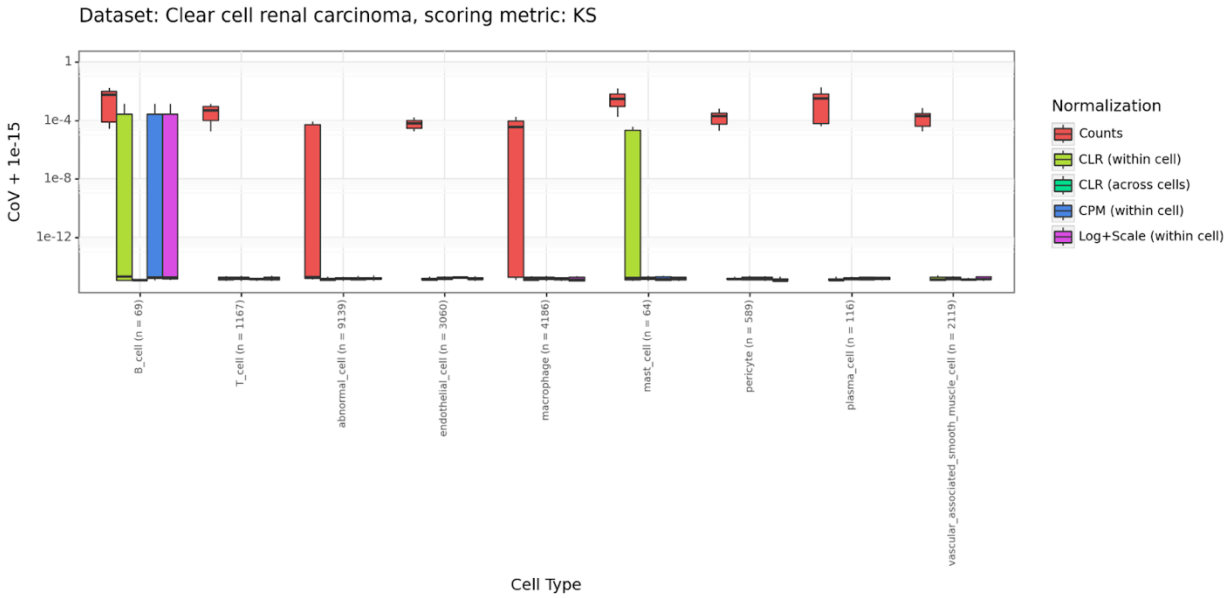


Figure S2.16: Aggregate cell permutations – Clear cell renal carcinoma – KS scores

Figure features as in **Figure 2B**

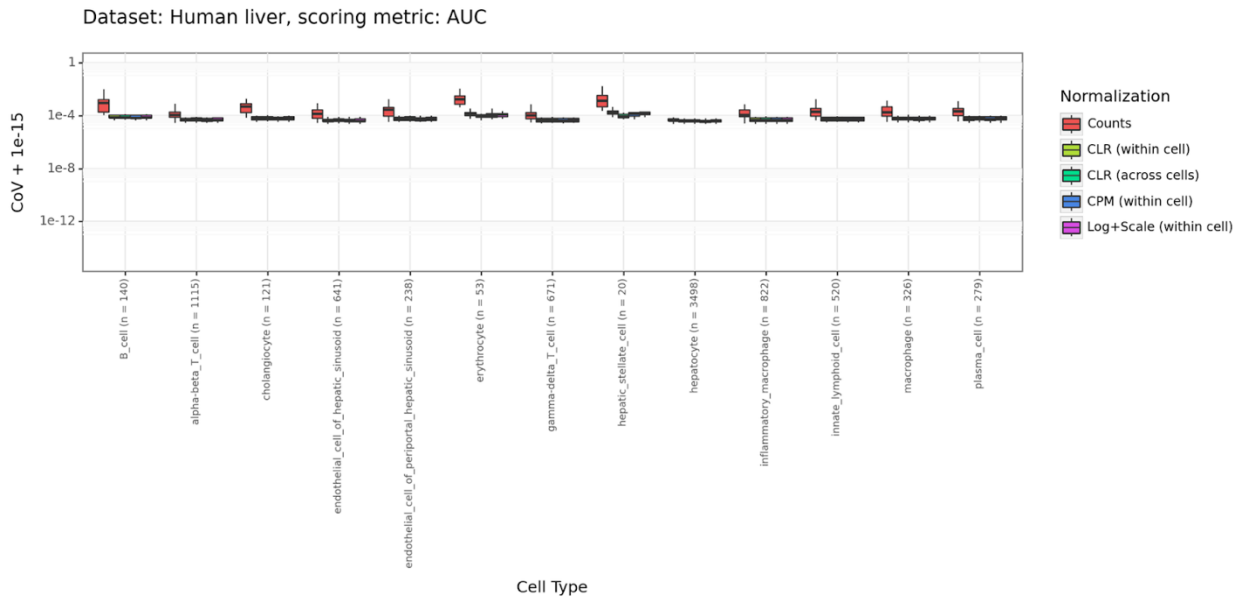


Figure S2.17: Aggregate cell permutations – Human liver – AUC scores

Figure features as in **Figure 2A**

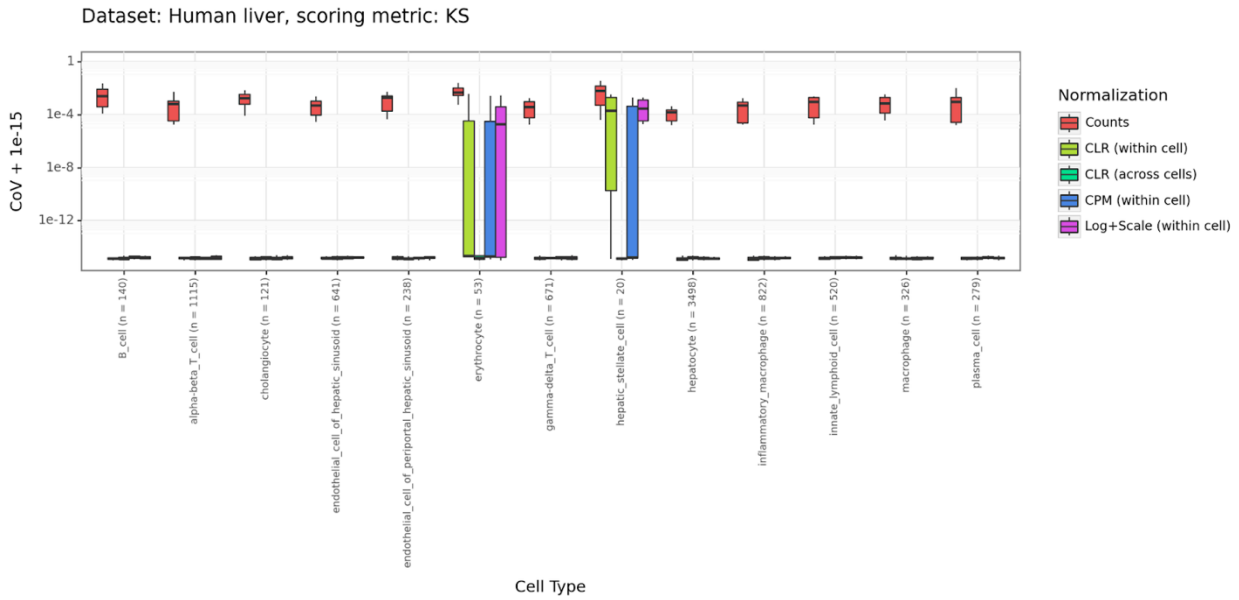


Figure S2.18: Aggregate cell permutations – Human liver – KS scores

Figure features as in **Figure 2B**

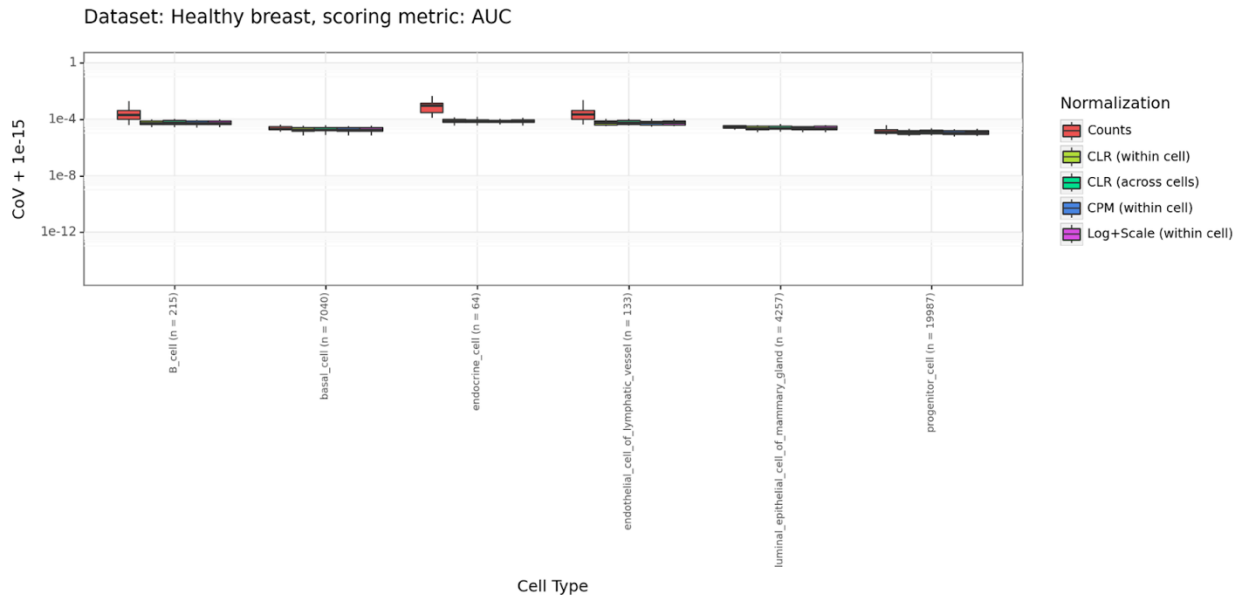


Figure S2.19: Aggregate cell permutations – Healthy breast – AUC scores

Figure features as in **Figure 2A**

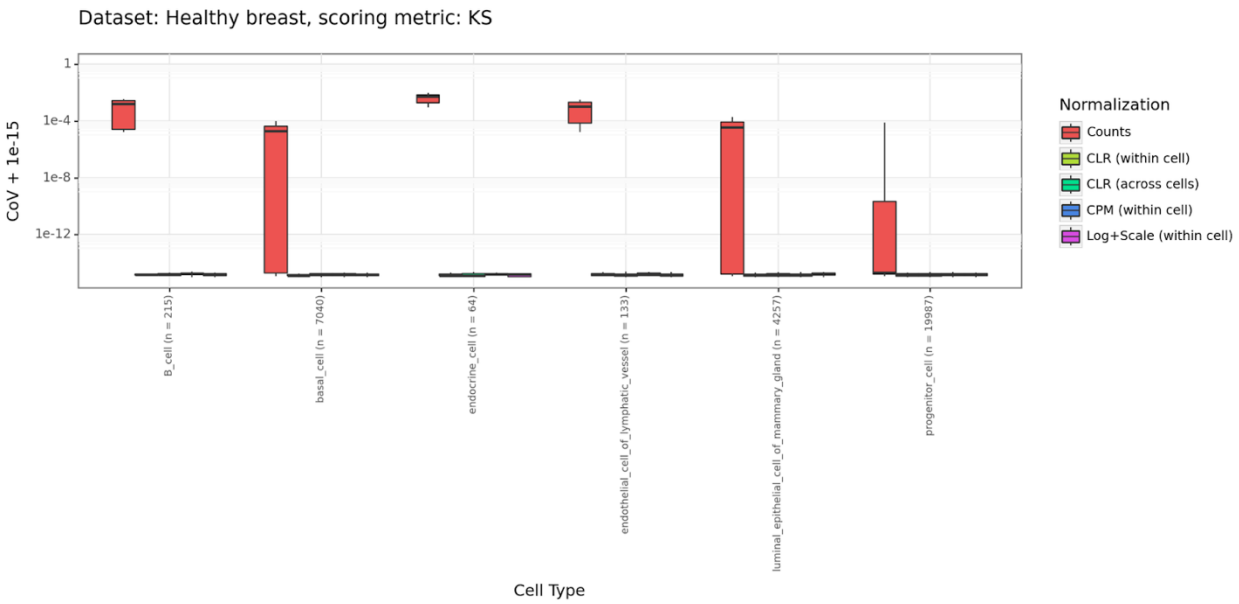


Figure S2.20: Aggregate cell permutations – Healthy breast – KS scores

Figure features as in **Figure 2B**

2.8 Supplementary Tables

Table S2.1: Datasets used in individual and aggregate cell benchmarking.

First three columns contain metadata from the CZI CELLxGENE data portal as provided by the authors of the corresponding studies. The fourth column "Brief names" were assigned for the purpose of this work.

CELLxGENE Dataset ID	CELLxGENE URL	CELLxGENE Full title	Brief name
793e90d3-62ab-44c4-98f5-6a090f321ba4	https://cellxgene.cziscience.com/collections/c353707f-09a4-4f12-92a0-cb741e57e5f0	Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming	Aging Skin
e62ae178-63f8-4025-91b2-33b15b48abe3	https://cellxgene.cziscience.com/collections/2902f08c-f83c-470e-a541-e463e25e5058	Single-cell reconstruction of follicular remodeling in the human adult ovary	Human ovary
837599ec-6580-42b4-86b7-2fb9ad232734	https://cellxgene.cziscience.com/collections/13d1c580-4b17-4b2e-85c4-75b36917413f	Single cell derived mRNA signals across human kidney tumors	Fetal kidney
231d025d-6b31-40da-aa38-cf618d53b544	https://cellxgene.cziscience.com/collections/1df8c90d-d299-4b2e-a54d-a5a80f36e780	ccRCC - Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response	Clear cell renal carcinoma
91a16284-28c5-41f8-a8c8-b283d2e5b6ea	https://cellxgene.cziscience.com/collections/bd5230f4-cd76-4d35-9ee5-89b3e7475659	Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations	Human liver
64f14a2b-d754-4bc9-b496-b26f05ebfe4e	https://cellxgene.cziscience.com/collections/c9706a92-0e5f-46c1-96d8-20e42467f287	A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells	Healthy breast

Table S2.2: Randomly created gene sets used in benchmarking

Gene set name	Gene set genes
gset_0	ENSG00000180316, ENSG00000158955, ENSG00000182583, ENSG00000139131, ENSG00000158716, ENSG00000223800, ENSG00000259124, ENSG00000205929, ENSG00000229924, ENSG00000227260, ENSG00000265100, ENSG00000225206, ENSG00000256671, ENSG00000143570, ENSG00000144843, ENSG00000155087, ENSG00000236790, ENSG00000156453, ENSG00000260442, ENSG00000255026
gset_1	ENSG00000171960, ENSG00000101191, ENSG00000165816, ENSG00000258107, ENSG00000249000, ENSG00000139620, ENSG00000107862, ENSG00000162753, ENSG00000184897, ENSG00000171492, ENSG00000176040, ENSG00000260162, ENSG00000261717, ENSG00000155827, ENSG00000183134, ENSG00000164458, ENSG00000237441, ENSG00000259129, ENSG00000233256, ENSG00000147592
gset_2	ENSG00000095932, ENSG00000226386, ENSG00000214360, ENSG00000073578, ENSG00000231987, ENSG00000155903, ENSG00000157077, ENSG00000135972, ENSG00000265907, ENSG00000167371, ENSG00000066248, ENSG00000185899, ENSG00000105671, ENSG00000249236, ENSG00000237419, ENSG00000268129, ENSG00000186297, ENSG00000088320, ENSG00000272456, ENSG00000250626
gset_3	ENSG00000185347, ENSG00000261129, ENSG00000204084, ENSG00000254344, ENSG00000108100, ENSG00000241370, ENSG00000268854, ENSG00000133112, ENSG00000104131, ENSG00000205186, ENSG00000237001, ENSG00000221909, ENSG00000227518, ENSG00000117691, ENSG00000083828, ENSG00000182095, ENSG00000267004, ENSG00000160867, ENSG00000186860, ENSG00000248393
gset_4	ENSG00000004848, ENSG00000063761, ENSG00000257746, ENSG00000188037, ENSG00000081320, ENSG00000273413, ENSG00000108961, ENSG00000141002, ENSG00000272861, ENSG00000100418, ENSG00000250855, ENSG00000234380, ENSG00000136939, ENSG00000249494, ENSG00000135899, ENSG00000170579, ENSG00000188803, ENSG00000112667, ENSG00000171862, ENSG00000143479

Table S2.2: Randomly created gene sets used in benchmarking

Gene set name	Gene set genes
gset_5	ENSG00000113739, ENSG00000137824, ENSG00000261360, ENSG00000266947, ENSG00000213585, ENSG00000119778, ENSG00000232298, ENSG00000068079, ENSG00000231914, ENSG00000260456, ENSG00000232555, ENSG00000266696, ENSG00000121274, ENSG00000011485, ENSG00000214614, ENSG00000171970, ENSG00000164815, ENSG00000231046, ENSG00000130640, ENSG00000267360, ENSG00000169413, ENSG00000136270, ENSG00000187033, ENSG00000101251, ENSG00000164323, ENSG00000273442, ENSG00000186001, ENSG00000115053, ENSG00000248262, ENSG00000067445, ENSG00000248243, ENSG00000138207, ENSG00000100138, ENSG00000162494, ENSG00000106128, ENSG00000260642, ENSG00000167447, ENSG00000105085, ENSG00000269307, ENSG00000104218, ENSG00000256084, ENSG00000197991, ENSG00000273211, ENSG00000256137, ENSG00000264575, ENSG00000267219, ENSG00000159182, ENSG00000129932, ENSG00000223569, ENSG00000136100
gset_6	ENSG00000188385, ENSG00000196562, ENSG00000040275, ENSG00000103067, ENSG00000163431, ENSG00000124508, ENSG00000256894, ENSG00000111832, ENSG00000198947, ENSG00000198673, ENSG00000105254, ENSG00000168509, ENSG00000248809, ENSG00000247775, ENSG00000225867, ENSG00000185347, ENSG00000163220, ENSG00000172840, ENSG00000166780, ENSG00000260416, ENSG00000182776, ENSG00000260430, ENSG00000226438, ENSG00000074660, ENSG00000224347, ENSG00000180957, ENSG00000001497, ENSG00000137726, ENSG00000236869, ENSG00000255462, ENSG00000108107, ENSG00000163637, ENSG00000221938, ENSG00000085721, ENSG00000197498, ENSG00000026103, ENSG00000165660, ENSG00000225697, ENSG00000197641, ENSG00000086015, ENSG00000233848, ENSG00000130695, ENSG00000139914, ENSG00000255845, ENSG00000148824, ENSG00000266696, ENSG00000160216, ENSG00000149679, ENSG00000160131, ENSG00000221869
gset_7	ENSG00000135454, ENSG00000234244, ENSG00000227252, ENSG00000249388, ENSG00000253715, ENSG00000197172, ENSG00000187010, ENSG00000137824, ENSG00000250313, ENSG00000164920, ENSG00000257365, ENSG00000187566, ENSG00000233397, ENSG00000106069, ENSG00000167711, ENSG00000236914, ENSG00000268205, ENSG00000127337, ENSG00000134852, ENSG00000141696, ENSG00000244953, ENSG00000154359, ENSG00000240498, ENSG00000108559, ENSG00000183035, ENSG00000100307, ENSG00000135547, ENSG00000270249, ENSG00000134716, ENSG00000131669, ENSG00000196459, ENSG00000259635, ENSG00000260701, ENSG00000253955, ENSG00000257341, ENSG00000249693, ENSG00000256417, ENSG00000138778, ENSG00000148942, ENSG00000272384, ENSG00000260185, ENSG00000231851, ENSG00000143252, ENSG00000154723, ENSG00000170961, ENSG00000261090, ENSG00000044090, ENSG00000146910, ENSG00000232645, ENSG00000115084

Table S2.2: Randomly created gene sets used in benchmarking

Gene set name	Gene set genes
gset_8	ENSG00000080603, ENSG00000113924, ENSG00000172375, ENSG00000112273, ENSG00000100605, ENSG00000197993, ENSG00000272502, ENSG00000233214, ENSG00000267224, ENSG00000248559, ENSG00000141639, ENSG00000267284, ENSG00000269243, ENSG00000235168, ENSG00000111536, ENSG00000235431, ENSG00000184432, ENSG00000265643, ENSG00000178821, ENSG00000257155, ENSG00000262791, ENSG00000133030, ENSG00000169062, ENSG00000116478, ENSG00000267152, ENSG00000183621, ENSG00000128536, ENSG00000158480, ENSG00000249392, ENSG00000258749, ENSG00000260278, ENSG00000223572, ENSG00000142166, ENSG00000165724, ENSG00000261334, ENSG00000139793, ENSG00000254968, ENSG00000150977, ENSG00000261749, ENSG00000198003, ENSG00000111863, ENSG00000146830, ENSG00000197061, ENSG00000234390, ENSG00000272139, ENSG00000180543, ENSG00000125107, ENSG00000129596, ENSG00000167633, ENSG00000143363
gset_9	ENSG00000224916, ENSG00000227507, ENSG00000253682, ENSG00000046653, ENSG00000177000, ENSG00000205155, ENSG00000169340, ENSG00000205208, ENSG00000268120, ENSG00000041880, ENSG00000140564, ENSG00000113810, ENSG00000187627, ENSG00000249096, ENSG00000254580, ENSG00000111684, ENSG00000258345, ENSG00000170473, ENSG00000139656, ENSG00000119913, ENSG00000250098, ENSG00000206105, ENSG00000135378, ENSG00000249637, ENSG00000106524, ENSG00000255733, ENSG00000197283, ENSG00000162825, ENSG00000232252, ENSG00000226476, ENSG00000116774, ENSG00000254006, ENSG00000152192, ENSG00000240012, ENSG00000236032, ENSG00000130377, ENSG00000185933, ENSG00000134627, ENSG00000186326, ENSG00000233365, ENSG00000127920, ENSG00000231212, ENSG00000182824, ENSG00000249988, ENSG00000234306, ENSG00000125820, ENSG00000260570, ENSG00000164236, ENSG00000188725, ENSG00000163933

Table S2.2: Randomly created gene sets used in benchmarking

Gene set name	Gene set genes
gset_10	ENSG00000077044, ENSG00000183484, ENSG00000261286, ENSG00000163564, ENSG00000250131, ENSG00000214285, ENSG00000131966, ENSG00000219438, ENSG00000261433, ENSG00000235601, ENSG00000158850, ENSG00000253932, ENSG00000110436, ENSG00000253792, ENSG00000125450, ENSG00000227269, ENSG00000101347, ENSG00000006210, ENSG00000180008, ENSG00000269343, ENSG00000225187, ENSG00000250765, ENSG00000224405, ENSG00000176402, ENSG00000250298, ENSG00000225785, ENSG00000106701, ENSG00000253708, ENSG00000236740, ENSG00000151726, ENSG00000261367, ENSG00000171951, ENSG00000254369, ENSG00000176984, ENSG00000181222, ENSG00000236914, ENSG00000119707, ENSG00000170876, ENSG00000125149, ENSG00000150051, ENSG00000179630, ENSG00000148926, ENSG00000255302, ENSG00000239453, ENSG00000174718, ENSG00000171435, ENSG00000075643, ENSG00000223631, ENSG00000169813, ENSG00000254587, ENSG00000260930, ENSG00000168303, ENSG00000260854, ENSG00000273419, ENSG00000182993, ENSG00000225807, ENSG00000099721, ENSG00000144283, ENSG00000267909, ENSG00000230159, ENSG00000165732, ENSG00000249236, ENSG00000168772, ENSG00000253344, ENSG00000265799, ENSG00000243896, ENSG00000203485, ENSG00000223725, ENSG00000177463, ENSG00000121390, ENSG00000248813, ENSG00000008196, ENSG00000233538, ENSG00000243903, ENSG00000198542, ENSG00000231238, ENSG00000170248, ENSG00000151366, ENSG00000253653, ENSG00000225612, ENSG00000246662, ENSG00000100412, ENSG00000259376, ENSG00000055332, ENSG00000171617, ENSG00000131115, ENSG00000258804, ENSG00000135828, ENSG00000174132, ENSG00000173705, ENSG00000231873, ENSG00000267325, ENSG00000182156, ENSG00000223601, ENSG00000162881, ENSG00000250328, ENSG00000261648, ENSG00000172534, ENSG00000086991, ENSG00000055955

Table S2.2: Randomly created gene sets used in benchmarking

Gene set name	Gene set genes
gset_11	ENSG00000132329, ENSG00000185009, ENSG00000221989, ENSG00000269053, ENSG00000230433, ENSG00000224101, ENSG00000128346, ENSG00000005022, ENSG00000183066, ENSG00000234068, ENSG00000253673, ENSG00000227297, ENSG00000111727, ENSG00000273210, ENSG00000250614, ENSG00000005884, ENSG00000099377, ENSG00000063169, ENSG00000226149, ENSG00000258935, ENSG00000261766, ENSG00000172340, ENSG00000233760, ENSG00000184220, ENSG00000261307, ENSG00000165323, ENSG00000237101, ENSG00000151923, ENSG00000101417, ENSG00000159263, ENSG00000162458, ENSG00000235497, ENSG00000189280, ENSG00000254039, ENSG00000229483, ENSG00000187944, ENSG00000116922, ENSG00000198521, ENSG00000204421, ENSG00000135945, ENSG00000126882, ENSG00000254163, ENSG00000250067, ENSG00000253309, ENSG00000228799, ENSG00000229307, ENSG00000156642, ENSG00000169174, ENSG00000154263, ENSG00000269416, ENSG00000125611, ENSG00000166917, ENSG00000224361, ENSG00000169548, ENSG00000039600, ENSG00000205238, ENSG00000231662, ENSG00000182776, ENSG00000127989, ENSG00000103066, ENSG00000189056, ENSG00000263684, ENSG00000234184, ENSG00000206559, ENSG00000139797, ENSG00000145817, ENSG00000165983, ENSG00000174738, ENSG00000101336, ENSG00000231424, ENSG00000081923, ENSG00000115109, ENSG00000172724, ENSG00000259351, ENSG00000154027, ENSG00000205669, ENSG00000127995, ENSG00000088992, ENSG00000253821, ENSG00000148300, ENSG00000149452, ENSG00000230172, ENSG00000268804, ENSG00000235819, ENSG00000257986, ENSG00000162244, ENSG00000205642, ENSG00000171863, ENSG00000215784, ENSG00000244558, ENSG00000221855, ENSG00000127561, ENSG00000157653, ENSG00000128654, ENSG00000175198, ENSG00000088832, ENSG00000258325, ENSG00000102384, ENSG00000166090, ENSG00000117600

Table S2.2: Randomly created gene sets used in benchmarking

Gene set name	Gene set genes
gset_12	ENSG00000256072, ENSG00000111886, ENSG00000173825, ENSG00000227475, ENSG00000183908, ENSG00000058673, ENSG00000178803, ENSG00000197381, ENSG00000255944, ENSG00000247993, ENSG00000228072, ENSG00000166343, ENSG00000248397, ENSG00000126947, ENSG00000152760, ENSG00000121853, ENSG00000264595, ENSG00000124172, ENSG00000173890, ENSG00000185532, ENSG00000062096, ENSG00000169862, ENSG00000015676, ENSG00000242221, ENSG00000183035, ENSG00000105404, ENSG00000257474, ENSG00000087266, ENSG00000236437, ENSG00000239900, ENSG00000114054, ENSG00000125170, ENSG00000108950, ENSG00000236104, ENSG00000120327, ENSG00000087111, ENSG00000258444, ENSG00000251417, ENSG00000012660, ENSG00000166106, ENSG00000130255, ENSG00000088002, ENSG00000116005, ENSG00000160801, ENSG00000249846, ENSG00000162755, ENSG00000172059, ENSG00000114745, ENSG00000260658, ENSG00000184897, ENSG00000254528, ENSG00000155719, ENSG00000135773, ENSG00000132603, ENSG00000107672, ENSG00000130182, ENSG00000255686, ENSG00000259383, ENSG00000146666, ENSG00000224857, ENSG00000240006, ENSG00000185761, ENSG00000226741, ENSG00000250056, ENSG00000229275, ENSG00000205279, ENSG00000164023, ENSG00000231143, ENSG00000176840, ENSG00000069020, ENSG00000173451, ENSG00000164853, ENSG00000138472, ENSG00000005448, ENSG00000224832, ENSG00000138623, ENSG00000116984, ENSG00000214575, ENSG00000228351, ENSG00000246695, ENSG00000075914, ENSG00000197993, ENSG00000198169, ENSG00000268173, ENSG00000224505, ENSG00000145721, ENSG00000251218, ENSG00000271011, ENSG00000149571, ENSG00000229928, ENSG00000102174, ENSG00000214940, ENSG00000266933, ENSG00000162761, ENSG00000223430, ENSG00000138074, ENSG00000170310, ENSG00000253479, ENSG00000140043, ENSG00000231420

Table S2.2: Randomly created gene sets used in benchmarking

Gene set name	Gene set genes
gset_13	ENSG00000145794, ENSG00000171189, ENSG00000163714, ENSG00000232057, ENSG00000189180, ENSG00000267644, ENSG00000147121, ENSG00000136802, ENSG00000132510, ENSG00000140157, ENSG00000174485, ENSG00000232613, ENSG00000163295, ENSG00000183246, ENSG00000266202, ENSG00000131653, ENSG00000272760, ENSG00000250375, ENSG00000130054, ENSG00000234943, ENSG00000272338, ENSG00000155962, ENSG00000250822, ENSG00000124092, ENSG00000196498, ENSG00000175334, ENSG00000223960, ENSG00000078900, ENSG00000119333, ENSG00000250091, ENSG00000164080, ENSG00000114520, ENSG00000228421, ENSG00000214872, ENSG00000272647, ENSG00000198129, ENSG00000198685, ENSG00000129028, ENSG00000144366, ENSG00000171401, ENSG00000224596, ENSG00000186924, ENSG00000236543, ENSG00000237741, ENSG00000153266, ENSG00000113070, ENSG00000168306, ENSG00000261431, ENSG00000114656, ENSG00000148677, ENSG00000214432, ENSG00000088038, ENSG00000231226, ENSG00000260141, ENSG00000251573, ENSG00000253474, ENSG00000186509, ENSG00000171130, ENSG00000237781, ENSG00000170871, ENSG00000250081, ENSG00000125686, ENSG00000204677, ENSG00000234396, ENSG00000263159, ENSG00000164284, ENSG00000155307, ENSG00000149196, ENSG00000100387, ENSG00000249662, ENSG00000224903, ENSG00000140057, ENSG00000149532, ENSG00000227482, ENSG00000163626, ENSG00000261121, ENSG00000173068, ENSG00000138615, ENSG00000052344, ENSG00000255054, ENSG00000250719, ENSG00000141646, ENSG00000166869, ENSG00000156968, ENSG00000167081, ENSG00000229278, ENSG00000166454, ENSG00000123810, ENSG00000273011, ENSG00000213015, ENSG00000260362, ENSG00000231597, ENSG00000091106, ENSG00000256969, ENSG00000217555, ENSG00000213420, ENSG00000214940, ENSG00000247708, ENSG00000059573, ENSG00000135318

Table S2.2: Randomly created gene sets used in benchmarking

Gene set name	Gene set genes
gset_14	ENSG00000124532, ENSG00000248235, ENSG00000125246, ENSG00000174292, ENSG00000183423, ENSG00000081760, ENSG00000175065, ENSG00000146828, ENSG00000226067, ENSG00000272259, ENSG00000257135, ENSG00000225243, ENSG00000262358, ENSG00000204278, ENSG00000143811, ENSG00000148343, ENSG00000186910, ENSG00000268810, ENSG00000227851, ENSG00000268754, ENSG00000187754, ENSG00000102934, ENSG00000101190, ENSG00000267726, ENSG00000110025, ENSG00000262133, ENSG00000233651, ENSG00000226516, ENSG00000259529, ENSG00000052795, ENSG00000231128, ENSG00000260470, ENSG00000260303, ENSG00000138115, ENSG00000229671, ENSG00000153086, ENSG00000221923, ENSG00000267204, ENSG00000188859, ENSG00000185668, ENSG00000168496, ENSG00000100162, ENSG00000248508, ENSG00000273149, ENSG00000233234, ENSG00000261448, ENSG00000163629, ENSG00000167526, ENSG00000198001, ENSG00000007372, ENSG00000254136, ENSG00000227059, ENSG00000272502, ENSG00000188828, ENSG00000107036, ENSG00000259523, ENSG00000236668, ENSG00000230552, ENSG00000214970, ENSG00000155886, ENSG00000112837, ENSG00000257634, ENSG00000269901, ENSG00000255548, ENSG00000234906, ENSG00000164076, ENSG00000158806, ENSG00000204228, ENSG00000261227, ENSG00000122952, ENSG00000185352, ENSG00000260088, ENSG00000272431, ENSG00000269292, ENSG00000232767, ENSG00000237837, ENSG00000070729, ENSG00000249924, ENSG00000260759, ENSG00000155897, ENSG00000105971, ENSG00000162408, ENSG00000242082, ENSG00000214413, ENSG00000171819, ENSG00000258580, ENSG00000223991, ENSG00000205863, ENSG00000164841, ENSG00000230894, ENSG00000137947, ENSG00000255372, ENSG00000198874, ENSG00000248973, ENSG00000134538, ENSG00000263688, ENSG00000258853, ENSG00000250049, ENSG00000178700, ENSG00000257835

2.9 Acknowledgements

Chapter 2, in full, is a reformatted presentation of the material currently being prepared for submission for publication as “scGSEA: Adapting GSEA to single cell RNA-sequencing data” by Alexander T. Wenzel, John Jun, Tanja Eisemann, Robert Wechsler-Reya, Pablo Tamayo, and Jill P. Mesirov. The dissertation author was the primary investigator and author of this material.

Chapter 3 Single-cell characterization of step-wise acquisition of carboplatin resistance in ovarian cancer

3.1 Abstract

The molecular underpinnings of acquired resistance to carboplatin are poorly understood and often inconsistent between in vitro modeling studies. After sequential treatment cycles, multiple isogenic clones reached similar levels of resistance, but significant transcriptional heterogeneity. Gene-expression based virtual synchronization of 26,772 single cells from 2 treatment steps and 4 resistant clones was used to evaluate the activity of Hallmark gene sets in proliferative (P) and quiescent (Q) phases. Two behaviors were associated with resistance: (1) broad repression in the P phase observed in all clones in early resistant steps and (2) prevalent induction in Q phase observed in the late treatment step of one clone. Furthermore, the induction of IFNα response in P phase or Wnt-signaling in Q phase were observed in distinct resistant clones. These observations suggest a model of resistance hysteresis, where functional alterations of the P and Q phase states affect the dynamics of the successive transitions between drug exposure and recovery and prompts for a precise monitoring of single-cell states to develop more effective schedules for, or combination of, chemotherapy treatments.

3.2 Introduction

Patients diagnosed with high grade serous ovarian cancer are generally treated initially with either cisplatin or carboplatin (CBDCA) in combination with paclitaxel. While 65–75% of patients respond to the primary treatment (Ozols et al. 2003), resistance emerges frequently during therapy and this is a major obstacle to cure (Siegel et al., 2013). Unlike targeted agents where high levels of resistance are common, repeated treatment of sensitive cells with clinically relevant levels of exposure to cisplatin or CBDCA produces only low-level resistance, typically in the range of 1.5–3-fold, a level sufficient to account for clinical failure of treatment in vivo (Andrews et al., 1990).

The mechanisms underlying acquired resistance to platinum-containing drugs have been the subject of intense study ever since their discovery. Acquired resistance has been attributed to changes in many types of cellular functions including import and export of the drug, enhanced detoxification and DNA adduct repair, inactivation of the mismatch repair checkpoint, and repression of apoptotic signaling (Galluzi et al., 2012). Findings from single genes or transcriptome-wide studies of bulk cell populations can usually be validated through overexpression or knock-out, but these studies have failed to disclose any actionable gene or set of genes that are consistently altered across different cell types or experiments and that would point toward the need for widely useful approaches for preventing or overcoming the development of resistance in patients.

Apart from BRCA1/2 mutation reversion, which occurs in up to 26% of mutated patients (Sakai et al., 2008) , the acquisition of CBDCA resistance is believed to be epigenetically mediated (Brown et al., 2014). Recent advances in the study of resistance to kinase inhibitors has revealed the existence of “persister” cells in lung cancer cell lines that are present at low

prevalence and can resist treatment through epigenetic mediated mechanisms (Sharma et al., 2010). Previously, single cell tracing had shown that, within a cell population, the immediate response to genotoxic treatment can vary extensively from cell to cell giving rise to considerable heterogeneity within the surviving population (Cohen et al., 2008; Gascoigne and Taylor, 2008). More recently, similar observations were made in vemurafenib-treated melanoma cells where cells in a transient resistant state are present in the population prior to drug exposure and display increased levels of expression of resistance genes (Shaffer et al., 2017). Importantly, the relevance of these recent models to the acquisition of resistance to platinum-containing drugs has not been established for either in vitro or in vivo models nor have the concepts been validated in clinical studies.

Here, we present a comprehensive phenotypic and molecular characterization of a set of ovarian cancer clones derived from a single cell and selected in parallel for acquired resistance to CBDCA. We show that the resistance is unlikely to be due to genetic mutations, copy number changes, or differences in CBDCA uptake. Resistance was associated with significant changes in proliferation rate and the capacity to form colonies and organoids, and there is substantial heterogeneity between clones. Transcriptome profiling showed a common association of CBDCA resistance with slow proliferation and high interferon signaling but also demonstrated marked heterogeneity between resistant clones. Importantly, single-cell transcriptome analysis allowed us to characterize the resistant states at single-cell resolution, eliminating the confounding effect of cell cycle variation and revealing functional changes specific to proliferative and quiescent phases.

3.3 Results

While resistance acquisition can be well recapitulated in vitro through repeated drug exposure, variation between cell lines, methodologies and the lack of selection replicates compromise the identification of widely shared molecular changes associated with resistance to platinum-containing drugs. We chose an experimental design that would allow us to determine whether genetically identical clones undergo the same molecular changes during acquisition of resistance. Specifically, we isolated a single cell from a non-clonal population of human ovarian cancer CAOV3 cells, and grew them to a small population (the parental clone) from which 12 clones were isolated (**Figure 3.1A**). Four of these clonal populations were grown continuously in the absence of CBDCA ("S clones": S01–S04). The remaining 8 ("R clones": R06, R07, R14-R19) were each individually subjected to 4 cycles of exposure to CBDCA at which point they tolerated 5 μM drug (subsequently referred to as step 5) and averaged 1.7-fold resistance relative to the parental clone (**Figure 3.1B**). Four of these eight clones were then treated with additional cycles of CBDCA at gradually increasing concentrations until they tolerated 15 μM CBDCA (referred to as step 15) and averaged 7.8-fold resistance relative to the parental clone. Subsequent passages of the resistant clonal populations in the absence of CBDCA for 63 doublings did not result in loss of resistance (**Supplementary Figure 3.1**), indicating that the phenotype was stable within the number of passages used in the study.

Phenotypic and molecular characterization of isogenic resistant clones

We next compared the phenotypic and molecular characteristics of 4 S and 4 step 15 R clones to identify features associated with acquired resistance. As shown in **Figure 3.1C-E**, the growth of the R clones was slower, they formed fewer large colonies in 2D culture, and in low-attachment plates they formed a higher proportion of small spheres. Importantly, the reduced proliferation can partly explain the resistant phenotype (**Supplementary Figure 3.2**) as

previously proposed (Shah and Schwartz, 2001; Kondoh et al., 2010). However, given the large differences in proliferation between R clones despite their similar level of resistance, it is likely that other processes contribute to the reduced cytotoxicity. Cells from all clones had similar distribution of CBDCA content after 1 h exposure (**Supplementary Figure 3.3**), suggesting that unlike other models (Abada and Howell, 2010), reduced drug uptake was not a major contributor to resistance in the studied clones. Exome sequencing of 4 S clones and 8 R clones (step 5) was used to identify copy number alterations (**Supplementary Table 3.1**). Two clones (S01 and R06) were affected by copy number gains (4 and 10 Mbp) and 7 clones (S01–03, R06, R16, R18, R19) had copy number losses (0.2–12 M bp). When copy number changes occurred, they were small in magnitude (less than 1.5-fold) and none of them affected multiple R clones. We also identified a median of 39 coding mutations per clone affecting a total of 74 genes. Neither total mutation burden nor gene-specific mutational burden was significantly associated with the resistant phenotype, albeit with limited statistical power (**Supplementary Table 3.2**). Interestingly, *CAAP1* p.T103P, was identified in all 8 R clones and 1 S clone, and while the mutation is predicted to be deleterious (CADD score = 2214), a role for *CAAP1* in apoptosis signaling has not been validated (Aslam et al., 2019; Zhang et al., 2011). Thus, these clonal populations derived from a single cell showed small genetic differences, unlikely to be the result of the treatment, but due to the small number of clones studied as well as the lack of functional information for most variants, a genetic cause to the resistance cannot be fully ruled out.

In order to identify molecular processes associated with resistance, we measured the expression level of all genes in the 4 S and the 4 R clones at step 15 using RNA-seq. We identified 186 genes that were differentially expressed between S and R clones (**Figure 3.2A**). An enrichment analysis of Hallmark (Liberzon et al., 2015) and Reactome (Fabregat et al., 2018)

gene sets from MSigDB (Liberzon et al., 2011), revealed that resistance was associated with a global repression of proliferation and translation and the activation of genes involved with interferon and *KRAS* signaling, and epithelial-to-mesenchymal transition (EMT) (**Supplementary Figure 3.4**). All of these are processes previously reported to be involved in chemo-resistance or response to genotoxic injury (Yang et al., 2006; Pavan et al., 2013; Brzostek-Racine et al., 2011). An unsupervised analysis revealed that, while the S clones had similar transcriptional profiles, the profiles of the R clones were highly heterogeneous (**Figure 3.2B**). Processes related to cellular proliferation (*E2F* targets) were repressed in all R clones, and interferon and *KRAS* signaling were induced in all R clones (**Figure 3.2C**). In contrast, a clone-specific analysis revealed that cell cycle and nucleotide excision repair were induced at higher level in R06, EMT in R14, oxidative phosphorylation in R16. These processes are not mutually exclusive and were dysregulated to different degrees in the various resistant clones. Importantly, the regulation of these resistance-associated processes is not specific to CAOV3. In order to validate these findings in additional ovarian cancer cell lines, we re-analyzed gene expression profiles associated with acquired cisplatin resistance in 8 ovarian cancer cell lines (Marchion et al., 2011). We could confirm the greater heterogeneity of resistant clones compared to untreated cells and showed that genes involved in *IFNα* or *KRAS* signaling, or EMT were frequently activated while those related to cell cycle and proliferation were repressed at equivalent time point and treatment regimen (cycle 6, schedule C, **Supplementary Figure 3.5**). This analysis, therefore, confirms that the functional changes observed in our system are consistent in cell lines and models of platinum resistance. Interestingly, we observed that ruxolitinib, a strong inhibitor of *IFNα* signaling via the *JAK*/*STAT* pathway, could re-sensitize the CAOV3 cells to CBDCA (**Figure 3.2D**) suggesting that the *JAK*/*STAT* pathway activation is required for the resistance in

these cells. Similarly, inhibition of *KRAS* signaling via silencing of *MEK* or inhibition of *ERK1/2* has been shown to increase platinum sensitivity in other ovarian cancer cell lines (Pénzvált et al., 2014; Persons et al., 1999; Cui et al., 2000). Beyond common mechanisms of resistance highlighted or confirmed by this analysis, our observations suggest important differences between clones prompting a more detailed investigation of the dynamic changes in the processes mediating treatment resistance.

Characterization of chemo-resistance at single cell resolution

Suspecting that phenotypic heterogeneity within the cell population may be underlying the differences in the acquisition of drug resistance, we measured the expression level of individual genes in 26,772 single cells from the original CAOV3 cell population, the parental clone derived from a single cell in this population, two S clones and two time points for each of the R clones (step 5 and step 15). The cells were classified according to their cell cycle signatures and, in agreement with the slower proliferation of the resistant clones, step 5 resistant cells were more likely to be quiescent than the untreated cells (61% vs 21% in G0- **Figure 3.3A**), while step 15 cells resumed proliferation (~30% in G0) with the exception of R16 (57% in G0).

The expression profiles were used to group all cells into 6 expression clusters, A through F (**Figure 3.3B, C**). Cells from step 15 R clones were distributed across 5 different clusters, whereas cells from step 5 R clones clustered together, indicating that additional treatment cycles increase transcriptional heterogeneity (**Figure 3.3D, E**). Interestingly, cells from R14 step 15 cells likely split into two sub-populations which could not be reliably distinguished by chromosomal copy number analysis (**Supplementary Figure 3.6A**) suggesting their differences in expression are not genetically driven. Similarly, step 15 clones display increased aneuploidy compared to their step 5 clones, suggesting that genetic drift may contribute to their increased

heterogeneity (**Supplementary Figure 3.6B, C**). The single-cell analysis further revealed that a small fraction (230/7814, 2.9%) of untreated cells were not in cluster A and were primarily in cluster B (133/230, 57%). The converse was not true with fewer than 0.2 % of the treated cells found in cluster A. From this observation, one could speculate that untreated cells could exist in a pre-resistant state prior to exposure to the drug, but isolation or tracking of these cells would be required to demonstrate this property. Cluster B had the largest fraction of cells in G0 indicating that quiescence characterizes both pre-resistant cells, step 5 and R16 step 15 cells (**Figure 3.3E**). However, differences in cell-cycle alone is unlikely to explain the variation in cellular states as cells in all phases of the cell cycle exist in all clusters. This observation prompted us to more carefully account for cell cycle differences before characterizing the functional changes associated with the different resistant states.

Variation of expression-based activity of biological processes along the cell-cycle

The contribution of cell cycle to single-cell gene expression can be mathematically subtracted to study the source of the residual variation and its association with the resistant phenotype. However, this correction method assumes the independence of gene expression measurements. Nonetheless, gene expression is tightly coordinated and gene-based cell-cycle correction may mask important intrinsic differences as cells progress through the cell cycle. Thus, to make the correction, we chose instead to use the expression of 603 genes associated with cell cycle to order the cells along a linear pseudotime (pt) trajectory from mitosis (G2M, M; $pt < 0.20$) to growth and replication (MG1, G1S, S; $0.20 \leq pt < 0.69$) and quiescence (G0; $pt \geq 0.69$) (**Figure 3.4A, B**). Consistent with the cell cycle phase analysis, the distribution of cells along the trajectory varied between clusters with the majority of cells in cluster A (referred to as

A cells) in proliferation (92% with pt < 0.69), B and C cells mostly in quiescence (79% and 74% pt ≥ 0.69, respectively), while D, E, and F cells, corresponded to step 15 treatment of R06 and R14 clones, resuming proliferation (19%, 15% and 15% with pt ≥ 0.69, respectively). The resulting virtual synchronization allowed us to identify cells from different clones that are in similar phases of the cell cycle, which facilitates the interpretation of functional differences associated with resistance.

The changes in activity, or enrichment score (ES), of the Hallmark gene sets can be measured along the pseudo-time using the smoothed geometric average of the gene expression (see **3.5 Methods** — **Figure 3.4C**). In untreated cells (Cluster A), the activity varied most strongly between proliferation (referred to as P; pt < 0.69) and quiescence (referred to as Q; pt ≥ 0.69). In particular, and as expected from their use in the trajectory inference, processes associated with cell cycle progression had the highest activity in P and shut down in Q (G2M checkpoint, *E2F* target, *MTOR* signaling). The activity of other processes were either unchanged along the cell cycle (*TGFβ* signaling, *P53* pathway) or were induced as cells progressed from P to Q (Hedgehog signaling, KRAS_DN). Such time-resolved functional changes can also help interpret some of the result observed in bulk RNA-seq as the differences proportion of cells in P and Q phases between R and S clones would result in apparent downregulation of process high in P and low in Q (e.g., *E2F* targets), or upregulation or processes low in P and high in Q (e.g., *IFN* alpha response). Hence the overall variation in gene expression - with and without dependency to changes in cell-cycle distribution - justify the necessity to distinguish between cells in P and Q to identify and properly interpret intrinsic functional differences associated with acquired resistance.

Drug resistance associates with distinct patterns of activity changes between proliferation and quiescence

As most clones and gene set activity showed the strongest differences between P and Q phases, we next summarized the observations across these two phases (**Figure 3.4D**, and **Supplementary Figure 3.7**). The median activity level of cluster B, C, and D was the lowest, irrespective of the phase. In contrast, the median activity in cluster E and F had levels similar to cluster A and even slightly higher in the Q phase. Across all gene sets activities, we observed three correlation patterns relative to cluster A untreated cells (**Figure 3.4E**): (1) a P-like pattern observed in P phase of cluster E and F, highly correlated to P phase of cluster A, (2) a Q-like pattern observed in Q phases of cluster E and F, correlated to Q phase of cluster A, (3) an anti-P pattern observed in both P and Q phases of cluster B, C, and D, with activity levels with strong negative correlations in comparison to P phase of cluster A.

The cells following the anti-P pattern all belonged to cluster B, C, and D, the clusters with the most repressed activity in both P and Q phases. The gene sets that are the most active in the proliferating cells from cluster A are also the most repressed in these cells, irrespective of the cell cycle phase, suggesting the cells adopt an expression state the most functionally distant from untreated proliferative cells and more closely resembling untreated quiescent cells. The resistant phenotype in these cells could therefore be due to the maintenance of a quiescence-like state, even in cells that are proliferating. In contrast, cells from cluster E and F in P and Q phases followed the P-like and Q-like patterns respectively, suggesting they resemble more closely cluster A cells. The level of correlation with the Q phase is however weaker, suggesting the difference may contribute to the resistance. Importantly, the patterns observed are unlikely due to genetic changes or associated with resistant steps as one clone switched from Anti-P (R14 step 5) to P-like and Q-like (R14 step 15) during the course of the treatment. Furthermore, the patterns

are independent from differences in proliferation rate, or fraction of cells in Q, as cells from slow proliferation (cluster B) or faster proliferation (cluster D) are both following the anti-P pattern.

The activity of some gene sets did not follow the patterns described above and their analysis may help determine which process are essential or dispensable to the resistant phenotype in specific clusters (**Supplementary Figure 3.7**). Increased *IFNα* response signaling in both P and Q phases of cluster C suggests that this process has been activated for clone R18 between step 5 and step15 of the treatment. Cells from cluster D showed induction of Wnt-signaling in both P and Q phases and UV response in P phase to levels similar to untreated cells, suggesting these processes may not contribute to the resistant phenotype as their activity was restored during subsequent treatment of that clone. Compared to Q phase untreated cells, Q phase cells in both cluster E and F have more active inflammatory response (E and F), *IFNα* (E only), Wnt signaling (F only) suggesting that these processes may contribute to the resistance by altering the expression state of cells in Q phase.

3.4 Discussion

As shown from the multiple replicates, treatment steps, and gene sets analyzed, the acquisition of resistance to CBDCA is a highly heterogeneous process with the repression of proliferation, and transition to a quiescent state as a common underlying factor. As such, the variation of activity of biological process along the cell cycle, and in particular between proliferation and quiescence can confound the functional annotation of the resistant phenotype. In order to determine whether the changes in expression of specific processes mediate the resistant phenotype, we used single-cell expression profiling to virtually synchronize cells and compare the activity of biological processes along the cell cycle. This analysis revealed that processes may be altered differently between P and Q phases during the acquisition of resistance and that different clones and treatment steps may follow different patterns.

In contrast to previous studies, the use of isogenic, single-cell derived cell lines allowed the generation and comparison of multiple matching resistant and sensitive samples. This rigorous experimental design, applied to the CAOV3 cell line, substantially reduced experimental noise. Additionally, CAOV3 has been unambiguously characterized as a serous ovarian cancer cell line in contrast to other models of platinum resistance such as A2780 (Beaufort et al., 2014). Most experimental studies of resistance choose a continuous exposure to the drug, selecting for cells in a drug persister state which expand to a drug-tolerant persister state. In contrast, in our study we chose to mimic chemotherapy cycles, giving an opportunity for the cells to recover between treatment steps through multiple cycles. Carrying out such multi-clone, multi-step experiments offers significant challenges in the laboratory, since clones rapidly acquire variable proliferation and treatment recovery dynamics. While the source of this diversity is not completely understood, it is likely rooted in the differences in Lamarckian

induction of adaptive response observed in other systems (Su et al., 2017). Variability between ovarian cancer cell models exists and while our observations were replicated in public studies with similar experimental design (**Supplementary Figure 3.5**), their validity in additional HGSOC models or even organoid systems remains to be established. Importantly, our study addressed the possibility that the resistant phenotype is genetically acquired. The analysis was however limited to mutations and copy number profiling and, starting from isogenic clones, the design assumed the convergent evolution of several resistant clones potentially altered in the same gene or locus. The analysis was however limited by the small number of clones studied and the reduced statistical power to conduct multi-genic or genetic burden tests. Hence, while our study does not completely rule out the possibility of a genetically driven phenotype in our model, the comprehensive remodeling of the cell cycle and associated processes observed and discussed below gave us an opportunity to explore non-genetic causes.

Interestingly, sequential treatment cycles such as the one used in our study may allow multiple rounds of adaptation to two different types of transitions to occur: from drug-free to treatment (drug on) and reciprocally (drug-off), providing, therefore, multiple chances for such adaptation – and associated heterogeneity – to occur. There is a strong association between cell cycle transition and drug exposure transition, as cells exposed to the drug will activate G1 checkpoint and may enter into quiescence, until repair can be completed and drug removed. Reciprocally, drug removal eventually leads to re-entry in cell cycle. As a consequence processes active in P phase are more likely to impact the drug-on transition whereas processes active in Q phase would be more likely to impact the drug-off transition. Specifically, processes accelerating P to Q transition or slowing Q to P transition are both likely to increase resistance by keeping cells in a low proliferative state. Such a hysteresis resistance model (**Supplementary Figure 3.8**)

is compatible with the following observations: processes in cells from cluster B had activity levels similar to, or lower than, untreated cells in Q phase, suggesting that cells in P phase are primed for a fast P to Q transition in this earlier treatment step. Alternatively, processes in cells from cluster E and F in Q phase have higher activity levels than untreated cells in Q phase suggesting that these processes slow down Q to P transition in support of resistance.

Cellular hysteresis models similar to the one discussed here, have been used previously to describe antibiotic resistance in bacteria (Roemhild et al., 2018) or transition between epithelial and mesenchymal states in cancer cells (Celià-Terrassa et al., 2018), but are not commonly used to model cancer drug resistance. The model accounts for the treatment memory effect that has been observed (Bell and Gilan, 2020), proposing that transition between states is not symmetric and can be mediated by distinct processes. Single-cell RNA-sequencing (scRNA-seq) and virtual synchronization allowed us, for the first time, to independently study cells undergoing drug exposure in each state thereby offering an observation window on the processes already primed before a transition occurs. However, the treatment time course (step 5 and step 15) and observation window (post recovery) used in our experiments lacks sufficient resolution to fully validate the model on individual treatment cycles and to precisely follow how the activity of different processes changes as a function of drug concentration, exposure or recovery duration, or number of cycles. The collection of multiple time points, within hours before and after each transition is likely to provide much clearer information. Importantly, the use of single-cell assays such as scRNA-seq or RNA-FISH is critical to capture events shortly after treatment where only few cells remain and to alleviate the need to expand them and introduce undesirable variation.

Based on our observations, it is unlikely that a single signaling or Hallmark process exclusively contributes to the acquisition of CBDCA resistance. The bulk analysis identified

IFNα response signaling as an induced process shared by multiple clones. Activation of interferon signaling is triggered by the DNA damage response (DDR) (Nakad and Schumacher, 2016) and was previously observed in response to genotoxic stress. Expression of *IRF1*, a main effector of interferon signaling, is induced by cisplatin and may limit this drug's effectiveness (Pavan et al., 2013). The process could be successfully blocked by pre-exposure to ruxolitinib which restored sensitivity. Hence, consistent with the findings of the genetic profiling, this observation suggested that resistance is unlikely to be inherited. The virtual synchronization showed little variation in *IFNα* response signaling along the cell cycle, which suggests that despite a strong shift of resistant clones into quiescence, the bulk analysis may have measured intrinsic changes independent of cell cycle. Interestingly, the level of induction of *IFNα* in both P and Q phase was correlated with the response to ruxolitinib with R14 and R18 clones being the most sensitive. *IFNα* response was, however, not induced in cluster B corresponding to step 5 treatment, perhaps suggesting that the sensitizing effect of ruxolitinib, although visible in untreated cells, may not apply to earlier treatment steps. Furthermore, it is not clear whether the inhibition of *JAK*/*STAT* signaling would restore sensitivity by accelerating the Q to P transition or slowing down P to Q. Studies in tissue regeneration have shown that inhibition of *JAK*/*STAT* promotes stem cell expansion suggesting that inhibition of *JAK*/*STAT* could wake up quiescent cells (Price et al., 2014). Such considerations on the direction of the effect are important for the design of combination therapy to prevent or reverse resistance and more precisely determine the treatment schedule. One can envision a combination of platinum drugs with treatments slowing down P to Q transition to prevent resistance development. Reciprocally, treatment accelerating Q to P transition, targeting processes active in Q phase, would increase the benefit of drug holidays and accelerate the re-sensitization of the tumor.

Beyond *IFNα* response signaling, it is possible that many other processes are impacting the dynamics of the transition at every treatment cycle and such redundancy may explain the heterogeneity observed between clones or after multiple cycles. However, cellular processes are constrained by the underlying regulatory circuitry and strongly inter-dependent. Multiple research efforts are underway to map these dependencies and reduce the complexity to fewer comprehensible dimensions (Kim et al., 2016; Tsherniak et al., 2017). The cellular hysteresis model of transition however adds a dynamic dimension ignored so far. The addition of single-cell assays and virtual synchronization to systematic, large scale assays will likely help generalize the phenomenon to multiple stimuli and understand which processes are acting in concert along each transition direction. It is likely that some of them may be key regulators of one direction without affecting the other. Such precise mapping would be important to determine which processes, or combination of processes, can be targeted to more effectively prevent the acquisition of resistance or more rapidly and durably re-sensitize cells. Similarly, the analysis of the epigenetic changes associated with the transitions are likely to capture underlying regulatory mechanisms supporting the hysteretic memory of change in growth conditions. Subsequent analysis at single-cell resolution would be needed to distinguish regulatory elements primed or poised for each transition direction.

Importantly, while our demonstration relies on in vitro observations, these processes are at play in vivo, in patients, where multiple stimuli from the micro-environmental niche, in addition to the treatment itself, may impact the cells' decision to enter or exit a proliferative state and it is likely that the hysteresis paradigm applies to any transition between states, which in oncology, often boils down to proliferation, quiescence or cell-death.

3.5 Methods

Generation of CBDCA resistant clones

CAOV3 cells and all sublines were grown in RPMI 1640 containing 5% fetal bovine serum and 1X penicillin/streptomycin. The clonally derived cell lines were always plated at 40,000 cells per well in 6 well plates and allowed to attach overnight before adding the drug. A selection cycle consisted of exposure to the drug for 7 days following which cells were allowed to recover in drug-free medium for ~2 weeks until they resumed growth and reached confluence. CBDCA sensitivity was determined from concentration-survival curves using ≥5 concentrations; viability was determined with the Cell Counting Kit 8 (Dojindo Molecular Technologies, Rockville, MD) or Crystal Violet reagent after 96 h of drug exposure.

2D and 3D growth assays

CAOV3 cells were seeded at 200 cells/well in a 6-well plate in replicates of 3 or 6 and allowed to form colonies for 9 days after which they were stained with Crystal Violet. Colonies were counted microscopically. The capacity to grow in 3 dimensions was tested by seeding 20,000 cells/well in ultra-low attachment 6 well plates (Corning Ref 3471) in stem cell medium (1:1 DMEM:F12 plus L-glutamine, 15 mM HEPES, 100 U/mL penicillin, 100 µg/mL streptomycin, 1% knockout serum replacement, 0.4% bovine serum albumin, and 0.1% insulin-transferrin-selenium (Corning, Corning, NY). The stem cell medium was further supplemented with human recombinant epidermal growth factor (20 ng/mL) and human recombinant basic fibroblast growth factor (10 ng/mL). The medium in each well was refreshed every 3 days by adding 500 µL/well of fresh stem cell media supplemented with the growth factors. Spheres were counted under a microscope and subclassified as either tight spheres or organoids after 7 and 14 days of culture.

Ruxolitinib treatment

Validation of JAK inhibition

The parental clones (6 wells seeded at 105 cells/well) were pre-treated with 5 μM of Ruxolitinib (50 mM in DMSO – LC laboratories) or vehicle. The cells were then treated with 1000 units of human Interferon-β for 24 h. RNA was then isolated using TRIzol (ThermoFisher), quantified using Nanodrop and 1 μg was converted to cDNA (High Capacity cDNA Reverse Transcription Kit - Applied Biosystems). The quantitative real-time PCR amplification −95 °C (10 min) and 34 cycles of 95 °C (1 min), 60 °C (30 s), 72 °C (1 min) – final: 72 °C (5 min) - was carried out in duplicate using the following pairs of primers: ISG15_FWD (GAGAGGCAGCGAACTCATCT), ISG15_REV (CTTCAGCTCTGACACCGACA) and 18S_FWD (CGCCGCTAGAGGTGAAATTCT) and 18S_REV (CGAACCTCCGACTTTCGTTCT). Expression data were normalized to the geometric mean of housekeeping gene 18 s to control the variability in expression levels and were analyzed using the 2-ΔΔCT method.

Dose-response assay

The growth rate of each clone was first established by seeding in 500, 1000, or 2000 cell per well and monitoring for 96 h (3 wells per time point, per sample). The density allowing exponential growth at 96 h was chosen to seed the cells in triplicate (96-well plates): Parent and R06: 2000 cells/well, R14: 2500 cells/well, R16: 3000 cells/well, and R18: 3000 cells/well. The cells were pre-treated with 5 μM Ruxolitinib or vehicle. Two additional wells were seeded in parallel to correct for growth rate differences (with and without Ruxolitinib) at the time of CBDCA treatment (time 0). The remaining wells were treated at time 0 with increasing CBDCA dose for 96 h, following which the cells were detached using 0.25% Tryspin-EDTA, diluted 1:2 in trypan blue and viable cells counted using a Hemocytometer. The Growth Rate and Growth-

Rate corrected half-maximal inhibitory concentration (GR50) were calculated following Hafner et al.36 and the GRmetrics bioconductor package.

Mass cytometry

Cells were incubated with 15 μM CBDCA for 1 h at 37 oC, washed and then exposed to a 1:500 dilution of Cell-ID Intercalator 103Rh for 15 min at 37 oC to mark the dead cells in the population. Analysis of ~2 × 105 cells from each sample was carried out on a Fluidigm Helios mass cytometer using EQ Four Element Calibration Beads for normalization. The results are presented in **Supplementary Table 3**.

Exome sequencing and analysis

The sequencing libraries were prepared and captured using SureSelect Human All Exon V4 kit (Agilent Technologies) following the manufacturer's instructions. The sequencing was performed using the Illumina HiSeq 2000 system, generating 100 bp paired-end reads. All raw 100 bp paired-end reads were aligned to the human genome reference sequence (hg19) using BWA (Li and Durbin, 2009) and further jointly realigned around indels sites using GATK's (McKenna et al., 2010) IndelRealigner. Duplicate reads were removed using Picard Tools MarkDuplicates (GitHub repository, 2024). **Supplementary Table 3.4** presents the summary statistics of the sequencing. The variants were called using Freebayes (Garrison and Marth, 2012) and filtered for high quality (QUAL/AO > 10). We annotated the variants with ANNOVAR (Wang et al., 2010), removed non-coding and synonymous variants, variants in dbSNP147, or shared between all samples, leading to a total of 93 variants across all 8 samples (**Supplementary Table 3.2**). The copy number changes were called independently on each chromosome using CODEX (Jiang et al., 2015) with default settings (**Supplementary Table 3.1**), limited to the expected exonic target from the SureSelect capture kits and expecting

fractional copy number from aneuploidy. Segments smaller than 100 kb, supported by less than 3 exons, or with copy number between 1.5 and 2.5 were excluded.

RNA sequencing and analysis

RNA was extracted using Qiagen RNAEasy and the libraries were prepared from 1 µg of RNA using TruSeq following the manufacturer instructions (shear time modified to 5 min). The libraries were sequenced on HiSeq 4000 (paired end 100 nt reads) and analyzed using BCBio-nextgen 1.0.1 (GitHub Repository, 2024). RNA-Seq default pipeline which included adapter removal with cutadapt v1.12 (Martin et al., 2011), read splice aware alignment with Bowtie2/Tophat suite v2.22.8 (Kim et al., 2013; Langmead and Salzberg, 2012), for quality control, and isoform expression level estimation using sailfish 0.10.1 (Patro et al., 2014). The differential expression was determined using DESeq2 (Love et al., 2014). We performed Gene Set Enrichment Analysis (Subramanian et al., 2005) implemented in the liger R package using gene sets from MSigDB (Liberzon et al., 2011).

Single cell RNA-sequencing

Data generation

We used the 10x Chromium (10x Genomics v2 reagents) to isolate ~2000 single cells from each sample following the manufacturer's instructions. Briefly, the cells/GEM droplets emulsion was formed using the 10x Chromium controller. The reverse transcription and template switching steps added both a cell-specific barcode and unique molecular identifier to each cDNA. The emulsion was then broken up and the GEM cleaned up. The single-strand cDNA was fragmented enzymatically and subjected to library preparation, including clean-up, end-repair, adapter ligation and enrichment PCR to add a sample-specific index. The libraries were quantified using Agilent Tape-station, and pooled for sequencing on the Illumina HiSeq 4000 for

single index paired-end sequencing (28 + 98nt reads). The resulting sequencing reads were separated using bcl2fastq and analyzed using the Cell Ranger v2 pipeline count, combining reads from different sequencing runs.

Data analysis

The barcode/cell matrices from different samples were further aggregated using Cell Ranger aggregate normalizing to total number of reads (**Supplementary Table 3.5**). The aggregated count matrices were processed with Seurat version 3.1.1 (Stuart et al., 2019). Genes expressed in fewer than three cells were removed. Additionally, cells were removed if fewer than 200 genes or more than 3700 genes were detected, or if mitochondrial genes made up more than 10% of their transcriptome. The remaining data was log normalized and scaled using the NormalizeData and ScaleData functions in Seurat with default parameter settings. To identify clusters, Seurat FindVariableGenes and RunPCA functions were run with default parameters and followed by FindNeighbors with the top 50 principal components and a resolution of 0.2 to identify the cell clusters. To assign cells to cell cycle phases, the Seurat CellCycleScoring function was modified to accept more than the 2 cell cycle gene sets. Using this function, each cell was scored for six cell cycle related gene sets derived from Xue et al. (G0.1 gene set excluded) (Xue et al., 2020) and was assigned the cell-cycle phase corresponding to the gene set with the maximum score.

Pseudo-time gene set enrichment analysis

The expression of 603 genes distributed across six cell cycle gene sets (Xue et al., 2020) was used to organize cells along a linear pseudo-time trajectory using SCORPIUS version 1.0.2 (Cannoodt et al., 2016), following the default workflow from the tutorial (GitHub Repository, 2024). Cells from each cluster were grouped in smoothing pseudo-time windows (interval = 0.2,

increment = 0.01 as a fraction of the pseudo-time range). The gene expression of all cells in a window was summarized into a virtual cell gene expression using a method derived from Baran et al. (Baran et al., 2019). In brief, the geometric mean of each gene's expression was calculated, divided by the median across all windows and clusters, and log transformed. The resulting gene expression profile of each virtual cell was then used to calculate the enrichment score of each gene set in the MSigDB's Hallmark collection (Liberzon et al., 2015) using single sample GSEA (ssGSEA) (Barbie et al., 2009) as implemented in the GenePattern (Reich et al., 2006) module. The enrichment score values were then scaled between 0 and 1 scaling the value in a given bin and cluster to the full range of values across all bins and clusters (**Figure 3.4C**). These scaled enrichment scores were further used to calculate the median in each P and Q phase for each cluster (**Figure 3.4D**).

3.6 Acknowledgements

Chapter 3, in full, is a reformatted reprint of the material as it appears as “Single-cell characterization of step-wise acquisition of carboplatin resistance in ovarian cancer” by Alexander T. Wenzel, Devora Champa, Hrishi Venkatesh, Si Sun, Cheng-Yu Tsai, Jill P. Mesirov, Stephen B. Howell, and Olivier Harismendy. The dissertation author was one of the primary investigators and authors of this material.

We thank Drs. Pablo Tamayo and Xin Lian Zhang for helpful discussions. We thank Mr. Gerald Manorek for technical assistance as well as Dr. Lyn Hedrick and Erik Ehinger for assistance with the Mass Cytometry assays. The work was supported by grants from the National Cancer Institute (R21CA177519, U24CA220341, U24CA194107, U24CA248457), from the Department of Defense (OC140179), grants from The Hartwell Foundation and The Immunotherapy Foundation, and an award from Pedal the Cause and a grant from the UC San Diego Academic Senate. A.T.W. was supported by an NLM T-15 training grant (T15LM011271) and a Training Fellowship from UC San Diego Cancer Cell Map Initiative (U54CA209891).

3.7 Author Contributions

D.C., S.S., C.-Y.T., H.V. conducted the experiments; D.C., O.H., A.T.W., J.P.M. analyzed the data; A.T.W., D.C., J.D.B., S.B.H., J.P.M., and O.H. interpreted the data and S.B.H., O.H. wrote the manuscript. A.T.W. and D.C. contributed equally. All authors approved the manuscript.

3.8 Figures

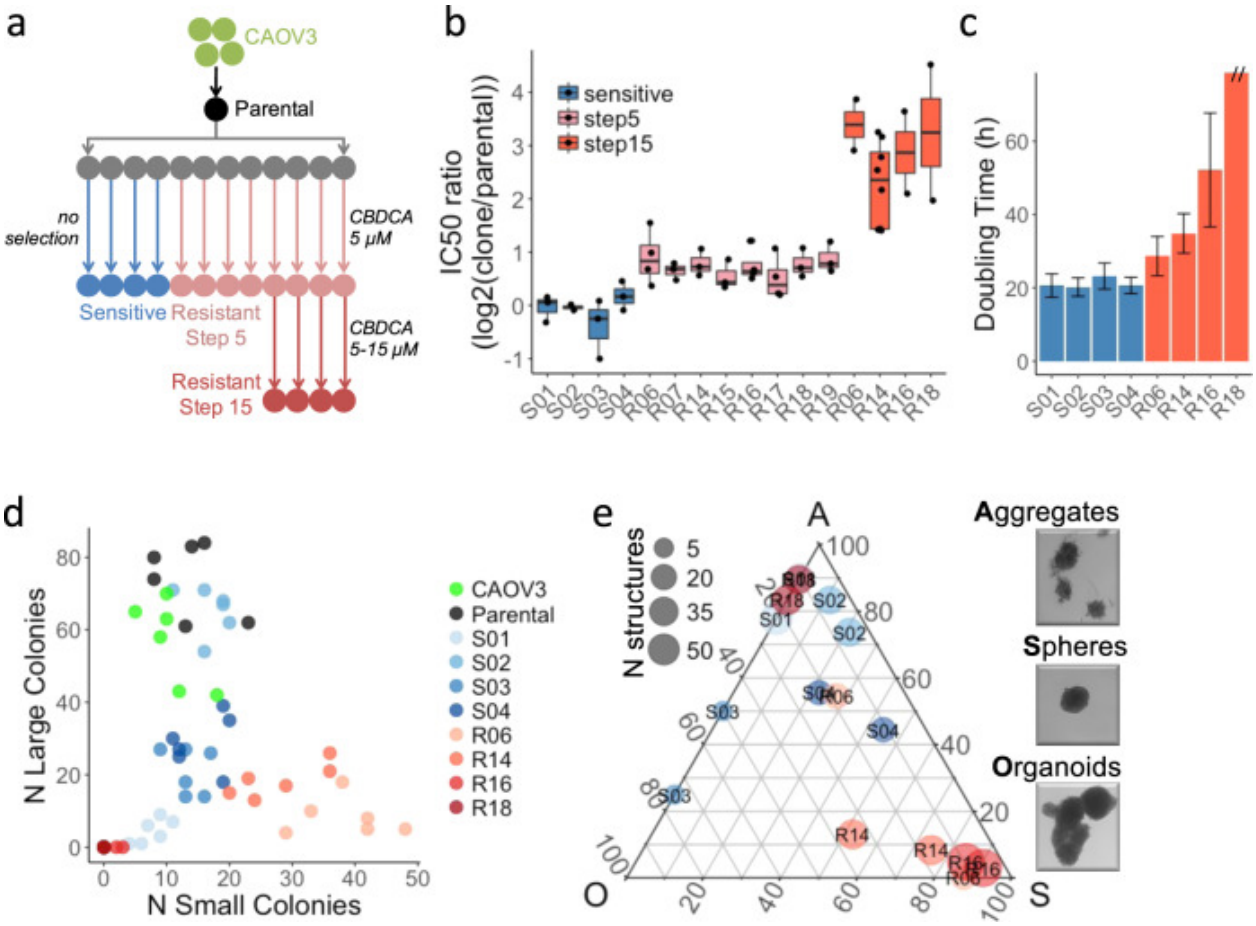


Figure 3.1: Phenotypic characterization of the resistant clones.

A) Schematic representation of the workflow to generate CBDCA resistant clones from CAOV3. **B)** Changes in IC50 of S clones (unselected) or R clones (8 at step 5 and 4 at step 15). Each IC50 is calculated from dose-response curves of 6 replicates and experiments repeated twice or more (dots). Boxes represent the top and bottom quartiles of the distribution and whiskers are extended to 1.5 time the interquartile range. **C)** Doubling time measured over a 48 h time course—y axis cut for R18 (>100 h). **D)** Counting of colonies formed in a period of 9 days after seeding 200 cells per well. Experimental replicates (N = 6) are shown. **E)** Fraction of organoids (O), spheres (S) and cell aggregates (A) observed after 14 days growth in low adherence 3D culture model. For each sample (N = 8) and replicates (N = 2), the total number (point size) and relative abundance (Gibbs triangle coordinates) of each type of structure are indicated.

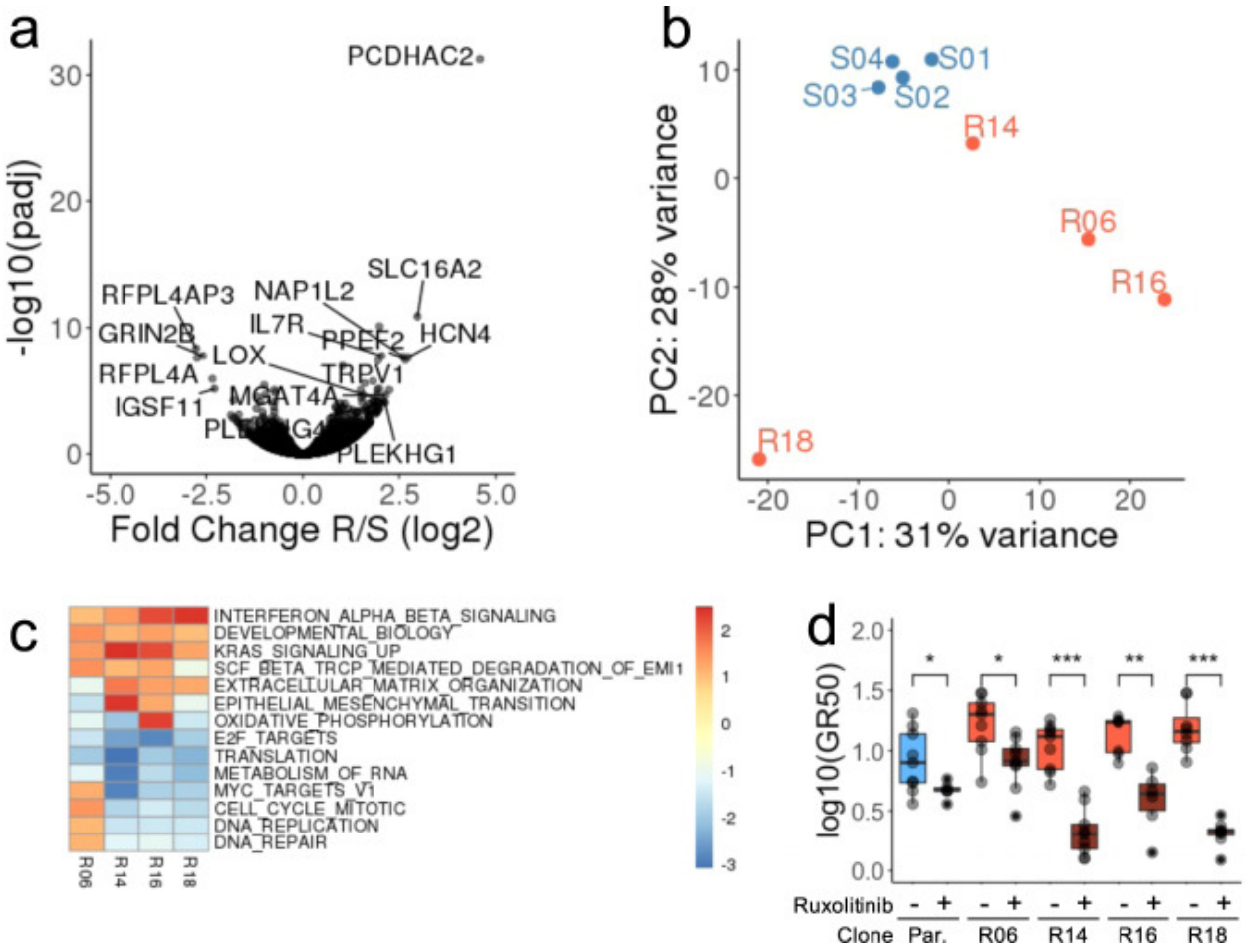


Figure 3.2: Expression profiling of the derived clones.

A) Volcano plot indicating the fold change (y axis) and significance (x axis) of the genes differentially expressed between S and R clones. **B)** First two principal components derived from the expression profiles of each clone. **C)** Most significantly up or down-regulated gene sets (Hallmark and Reactome from MSigDB) in individual R clones compared to all S clones. Significant gene sets ($q < 0.005$) enriched (score > 1.5) or depleted (score < -2) in at least one clone are reported. Color gradient indicates enrichment score. **D)** Treatment with 5 µM ruxolitinib (Rux) significantly decreases the growth-rate corrected half maximal inhibitory concentration (GR50) in both Parental and R clones. The results of 3 dose-response experiments, 3 replicates per experiment are presented. Significance was measured using Wilcoxon Test (* < 0.05, ** < 0.01, *** < 0.001). Boxes represent the top and bottom quartiles of the distribution and whiskers are extended to 1.5 time the interquartile range.

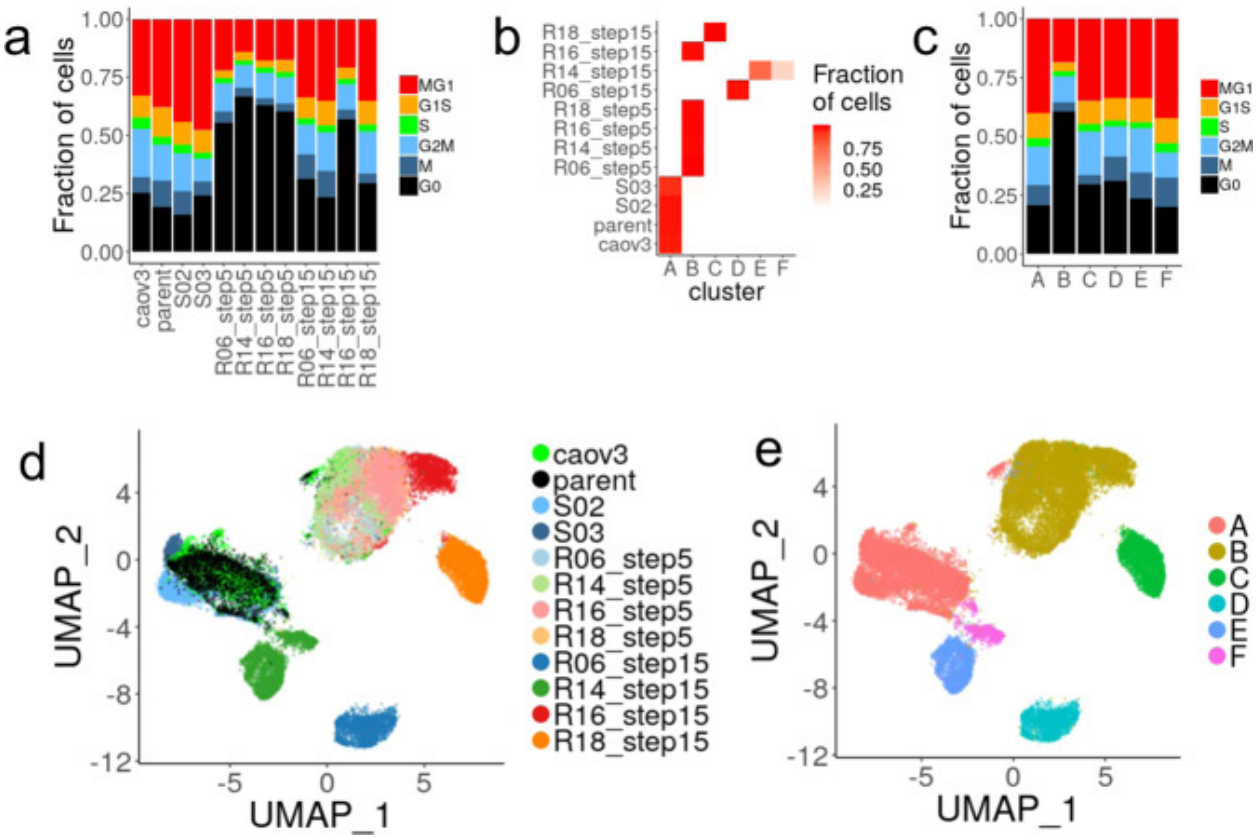


Figure 3.3: Evolution of expression states in all clones.

A) Distribution of cells in three phases of the cell cycles estimated from the expression signatures. **B)** Distribution of the cells from each clone and treatment group across the 6 clusters. **C)** Distribution of cells in three phases of the cell cycles estimated from the expression signatures. **D, E)** Uniform Manifold Approximation and Projection (UMAP) of cells from the aggregated analysis based on the first 2 principal components. Cells are colored according to their sample of origin **(D)** or Louvain cluster **(E)**.

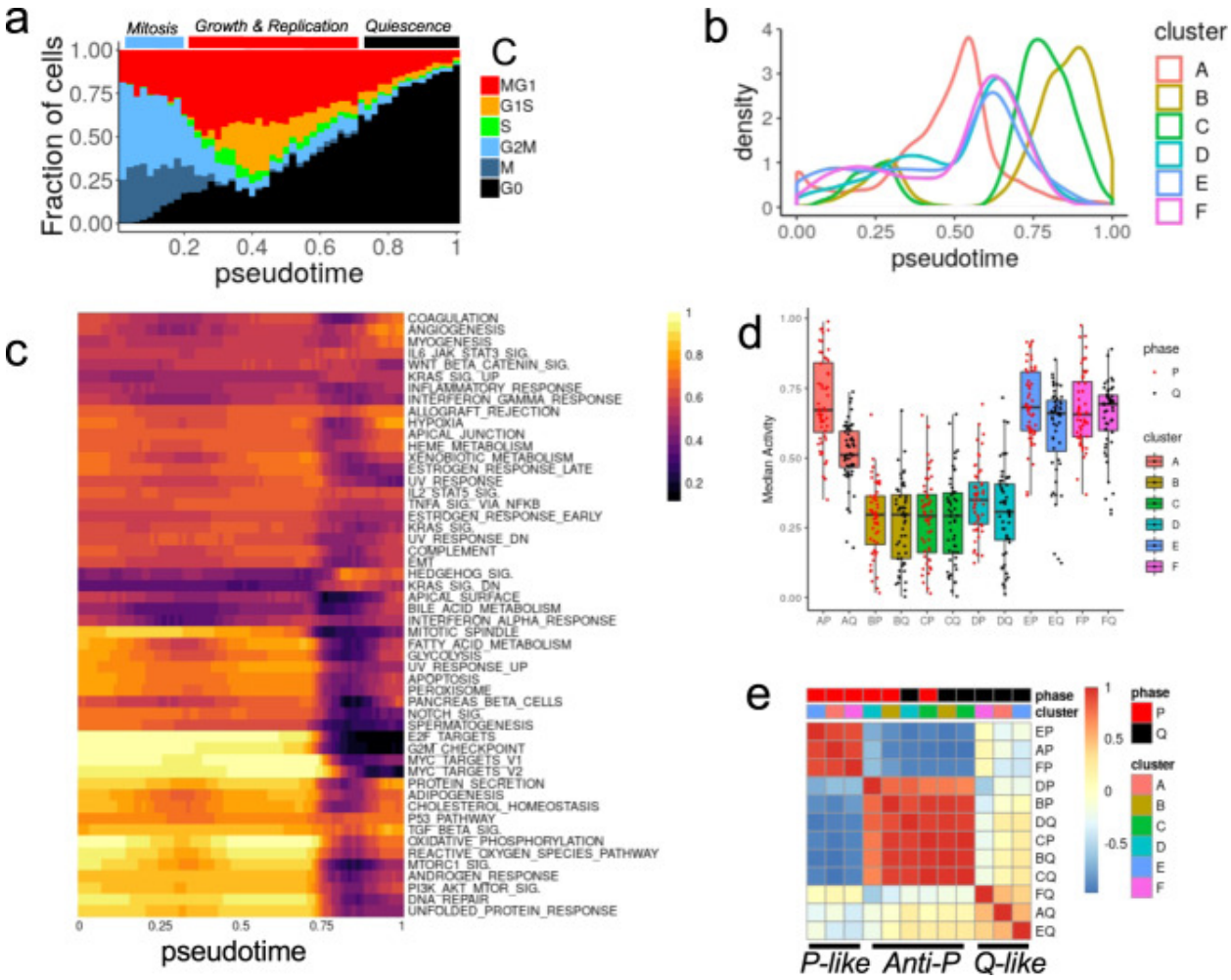


Figure 3.4: Functional analyses of single-cell resistant states after in silico synchronization.

A)Fraction of cells in each inferred cell cycle phase as a function of the pseudo-time (x axis bins). Three aggregated cell cycle phases are indicated above the plot and determined by bins containing more than 50% of the cells in G2/M or M (mitosis, blue bar, pt < 0.2), in G1S or MG1 (growth and proliferation, red bar, 0.2 ≤ pt < 0.69) or in G0 (quiescence, black bar pt ≥ 0.69). **B)** The distribution of cells (kernel density – y axis) along the pseudo-time trajectory (x axis) is represented for each expression-based clustering. **C)** Scaled enrichment score (ES) of the hallmark gene sets (clustered rows) observed in cluster A cells as a function of pseudo-time (columns). **D)** Median scaled enrichment score of all hallmark gene sets across cells from P (red points) and Q (black points) phases for each of the 6 clusters. Boxes represent the top and bottom quartiles of the distribution and whiskers are extended to 1.5 time the interquartile range. **E)** Correlation of hallmark gene sets median scaled enrichment score for all clusters and proliferation phases.

3.9 Supplementary Figures

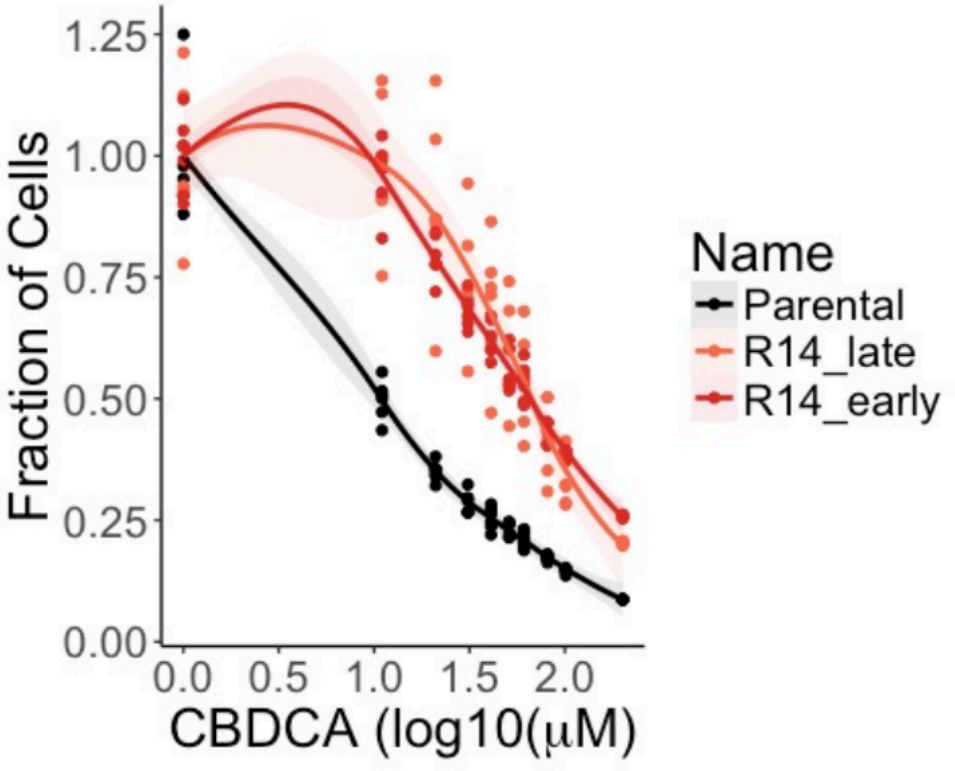


Figure S3.1: Carboplatin dose response curve

Dose Response comparing the drug sensitivity of R14 clones after 36 (R14_early) and 99 (R14_late) doublings after step 15 selection.

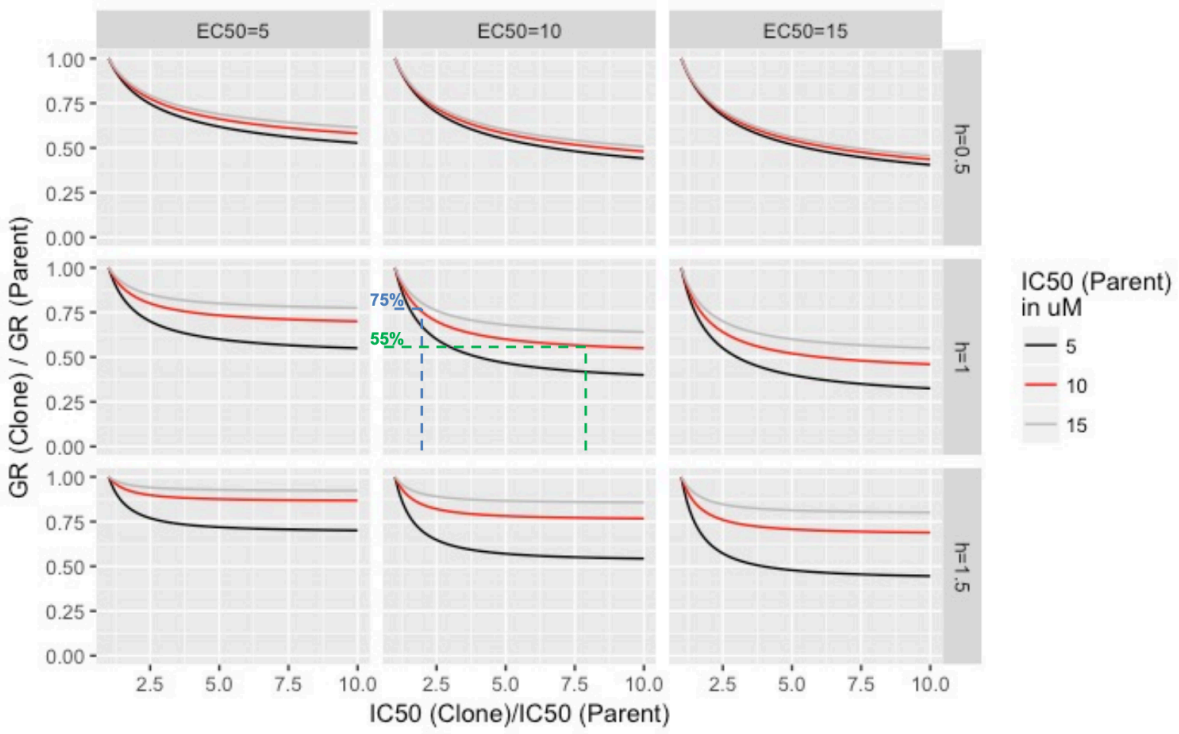


Figure S3.2: Simulated impact of growth rate on resistance.

The reduction in growth rate (y axis) is associated with an increase in IC50 (x-axis). The simulations were performed according to Hafner et al (formula IC(c,t) located at http://www.grcalculator.org/grtutorial/). The IC50 of the parental clone (colored lines), hill coefficient (grid rows) and EC50 (grid columns) were tested around the experimental ones observed (IC50=10 µM, h=1, EC50=10 µM, middle panel of the grid, red line). The level of resistance observed in step 5 (blue lines) and step 15 (green lines) is indicated.

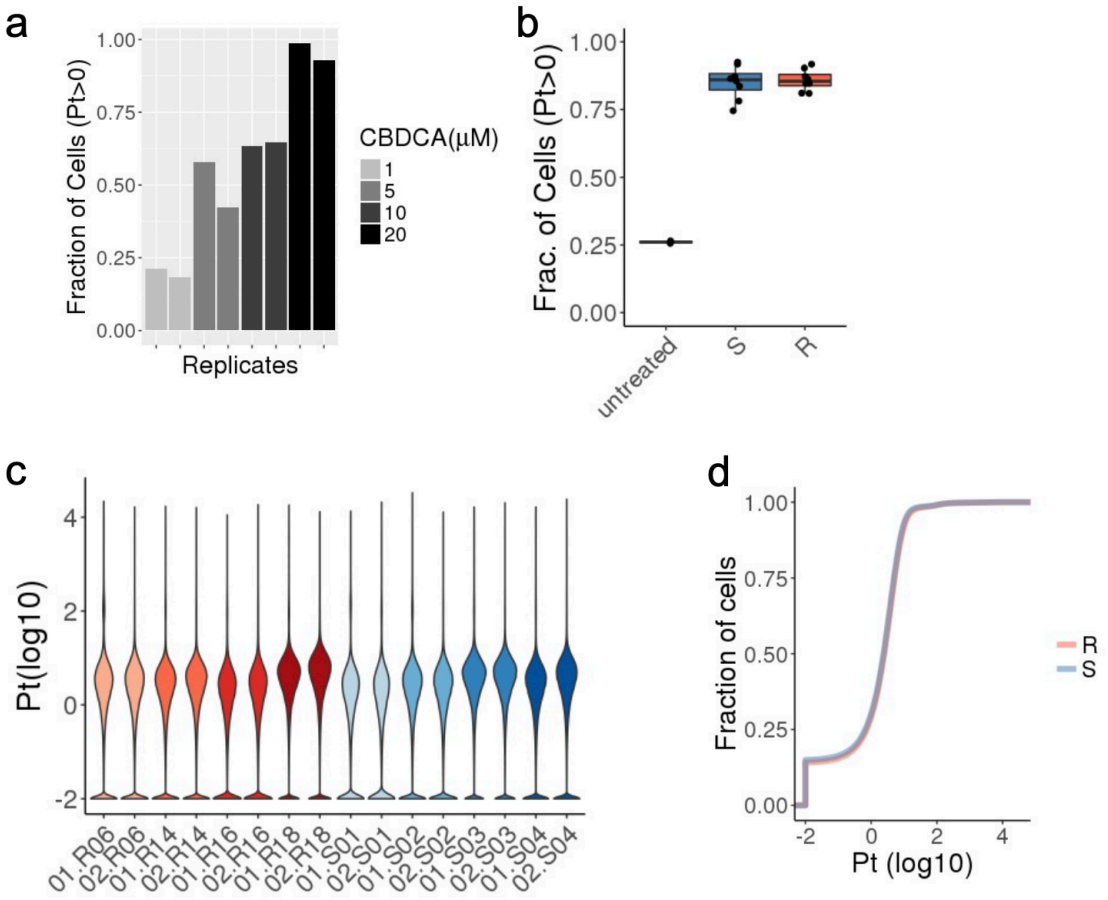


Figure S3.3: Platinum uptake measured by mass cytometry.

A) The fraction of cells forms the parental clone with detectable Pt content was measured after 1 h treatment with increasing dose of carboplatin (x axis, grey color gradient). **B)** Fraction of cells with detectable levels of Pt. Each clone (S and step 15 R clones) and replicate are represented. Boxes represent the top and bottom quartiles of the distribution and whiskers are extended to 1.5 time the interquartile range. **C)** Violin plot showing the distribution of Pt content across cells from all clones and replicates. **D)** Cumulative distribution of Pt content between cells from S clones and step 5 R clones and replicates. The Pt negative cells were assigned at 0.01 Pt content for graphical representation **(C, D)**.

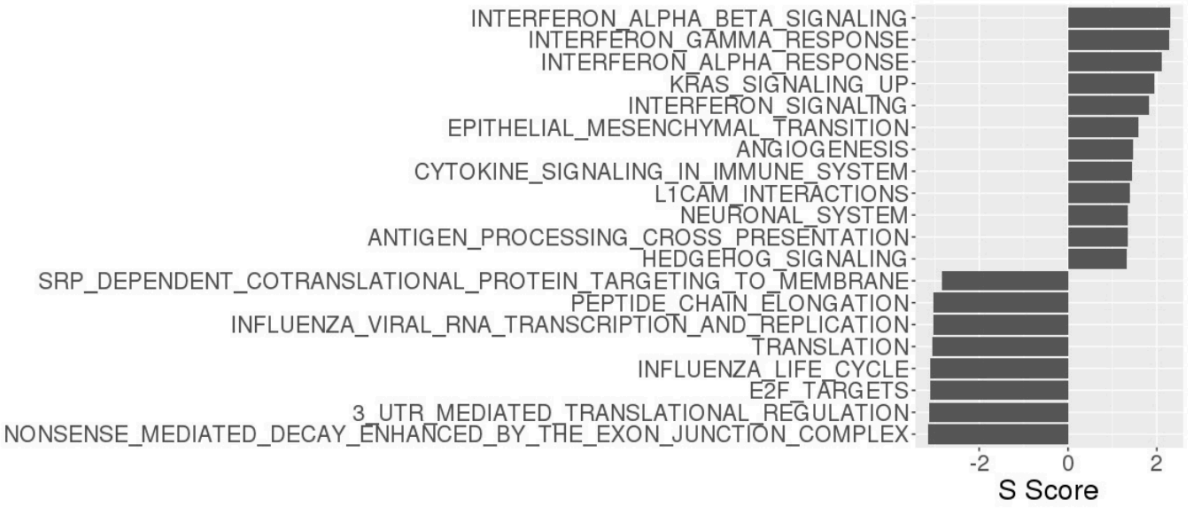
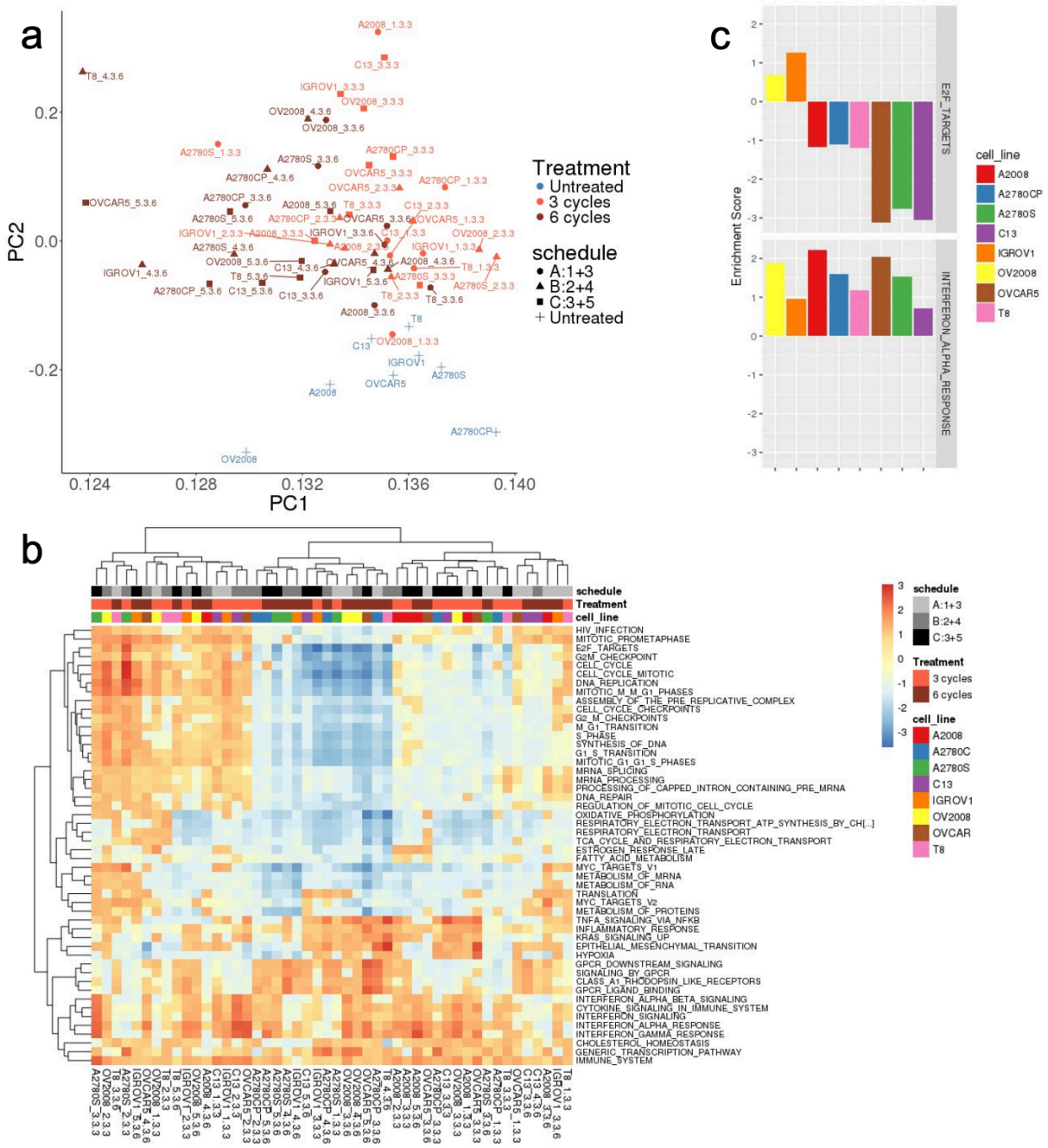


Figure S3.4: Most affected pathways in the resistant clones.

Most significantly up or down-regulated processes in CBDCA resistant clones after step 15 selection (q.value<0.05). All Hallmark (N=50) and Reactome (N=674) gene sets from MSigDB were included in the analysis

Figure S3.5: Signatures of acquired cisplatin resistance in 8 ovarian cancer cell lines

A) Principal component analysis of 56 samples from 8 cell lines. Resistant lines were derived from each cell line using 3 treatment schedules of 6 cycles each, increasing cisplatin concentration after 3 cycles. Cells were analyzed after 3 and 6 treatment cycles. Cell line specific variance was corrected for batch effect. See Methods and Marchion et al for details **B)** Fold change in gene set expression levels compared to all untreated samples. In each treated sample and the set of 8 untreated samples was used to calculate the enrichment score of all Hallmarks and Reactome gene sets from MSigDB. Significant gene sets more than 4 fold induced or repressed in at least one sample are represented. The color scale represents the enrichment score (log scale). **C)** Enrichment score for two representative gene sets for samples after 6 cycles of treatment schedule C (3 μM then 5 μM)


a
PC2
PC1
0.2
0.0
-0.2
0.124
0.128
0.132
0.136
0.140
Treatment
Untreated
3 cycles
6 cycles
schedule
A:1+3
B:2+4
C:3+5
Untreated
c
E2F_TARGETS
INTERFERON_ALPHA_RESPONSE
Enrichment Score
cell_line
A2008
A2780CP
A2780S
C13
IGROV1
OV2008
OVCAR5
T8
b
schedule
Treatment
cell_line
HIV_INFECTION
MITOTIC_PROMETAPHASE
E2F_TARGETS
G2M_CHECKPOINT
CELL_CYCLE
CELL_CYCLE_MITOTIC
DNA_REPLICATION
MITOTIC_M_M_G1_PHASES
ASSEMBLY_OF_THE_PRE_REPLICATIVE_COMPLEX
CELL_CYCLE_CHECKPOINTS
G2_M_CHECKPOINTS
M_G1_TRANSITION
S_PHASE
SYNTHESIS_OF_DNA
G1_S_TRANSITION
MITOTIC_G1_G1_S_PHASES
MRNA_SPLICING
MRNA_PROCESSING
PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA
DNA_REPAIR
REGULATION_OF_MITOTIC_CELL_CYCLE
OXIDATIVE_PHOSPHORYLATION
RESPIRATORY_ELECTRON_TRANSPORT_ATP_SYNTHESIS_BY_CH[...]
RESPIRATORY_ELECTRON_TRANSPORT
TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT
ESTROGEN_RESPONSE_LATE
FATTY_ACID_METABOLISM
MYC_TARGETS_V1
METABOLISM_OF_MRNA
METABOLISM_OF_RNA
TRANSLATION
MYC_TARGETS_V2
METABOLISM_OF_PROTEINS
TNFA_SIGNALING_VIA_NFKB
INFLAMMATORY_RESPONSE
KRAS_SIGNALING_UP
EPITHELIAL_MESENCHYMAL_TRANSITION
HYPOXIA
GPCR_DOWNSTREAM_SIGNALING
SIGNALING_BY_GPCR
CLASS_A1_RHODOPSIN_LIKE_RECEPTORS
GPCR_LIGAND_BINDING
INTERFERON_ALPHA_BETA_SIGNALING
CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM
INTERFERON_SIGNALING
INTERFERON_ALPHA_RESPONSE
INTERFERON_GAMMA_RESPONSE
CHOLESTEROL_HOMEOSTASIS
GENERIC_TRANSCRIPTION_PATHWAY
IMMUNE_SYSTEM
schedule
A:1+3
B:2+4
C:3+5
Treatment
3 cycles
6 cycles
cell_line
A2008
A2780C
A2780S
C13
IGROV1
OV2008
OVCAR
T8

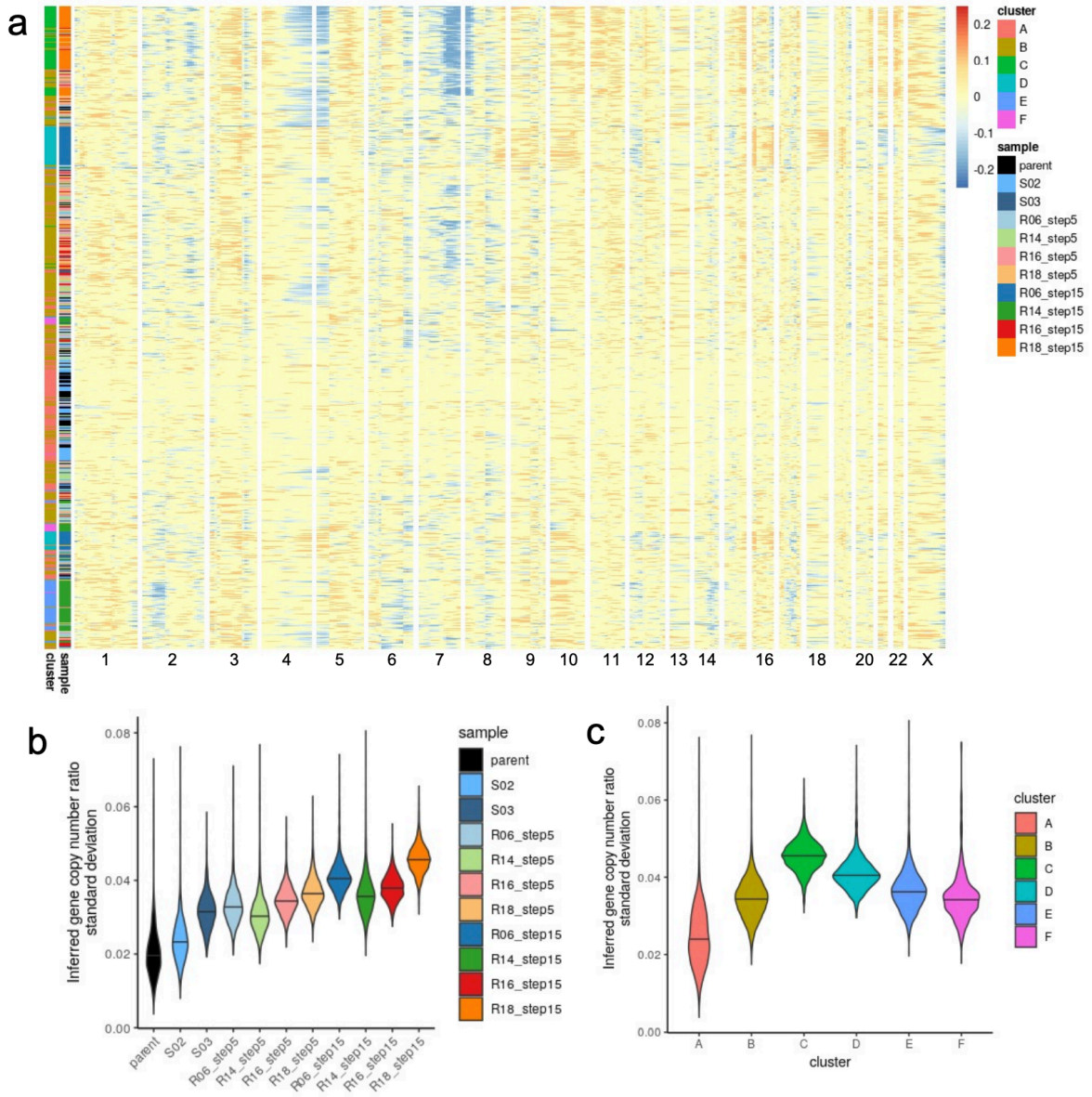


Figure S3.6: Chromosome Copy Number Profile inferred from scRNA-seq expression.

A) clustered heatmap of the copy number ratio observed in individual cells (rows) annotated based on their sample and cluster assignment. Genomic interval bins (columns – 10 Mbp windows) are group by chromosome (grouped columns). Values represent the log2 of the average copy number. The COAV3 cells were used as reference group. **(B, C)** Distribution of the standard deviation in gene copy number ratio for all cells in each sample **(B)** or expression-based cluster **(C)**

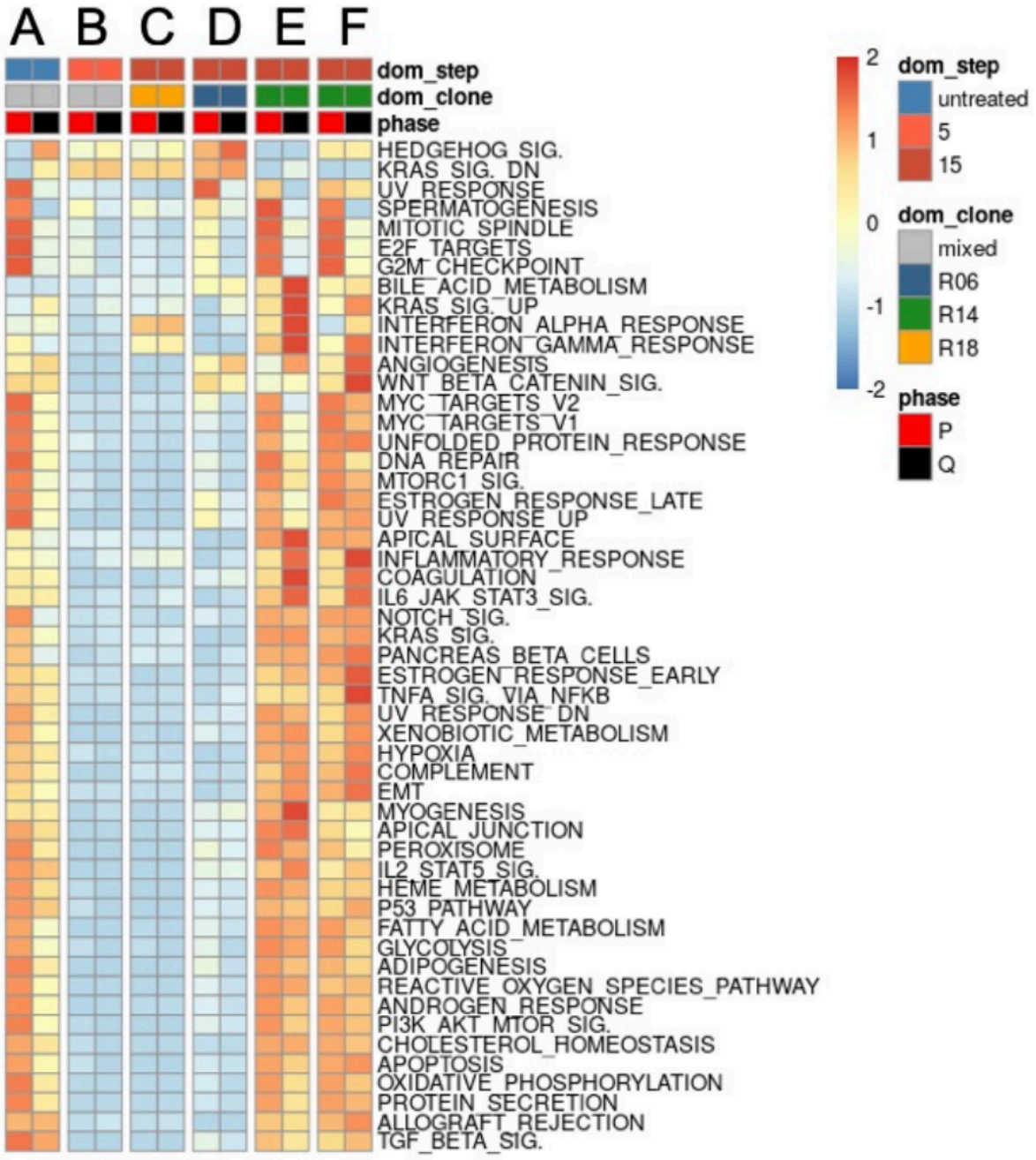


Figure S3.7: Clustered heatmap

Clustered heatmap representing the median gene set enrichment scores (zscore - color-scale) for hallmark gene sets (rows) for all clusters in proliferative (P) or quiescence (Q) phases. Dominant treatment step (dom_step) and clone (dom_clone) in each cluster are indicated.

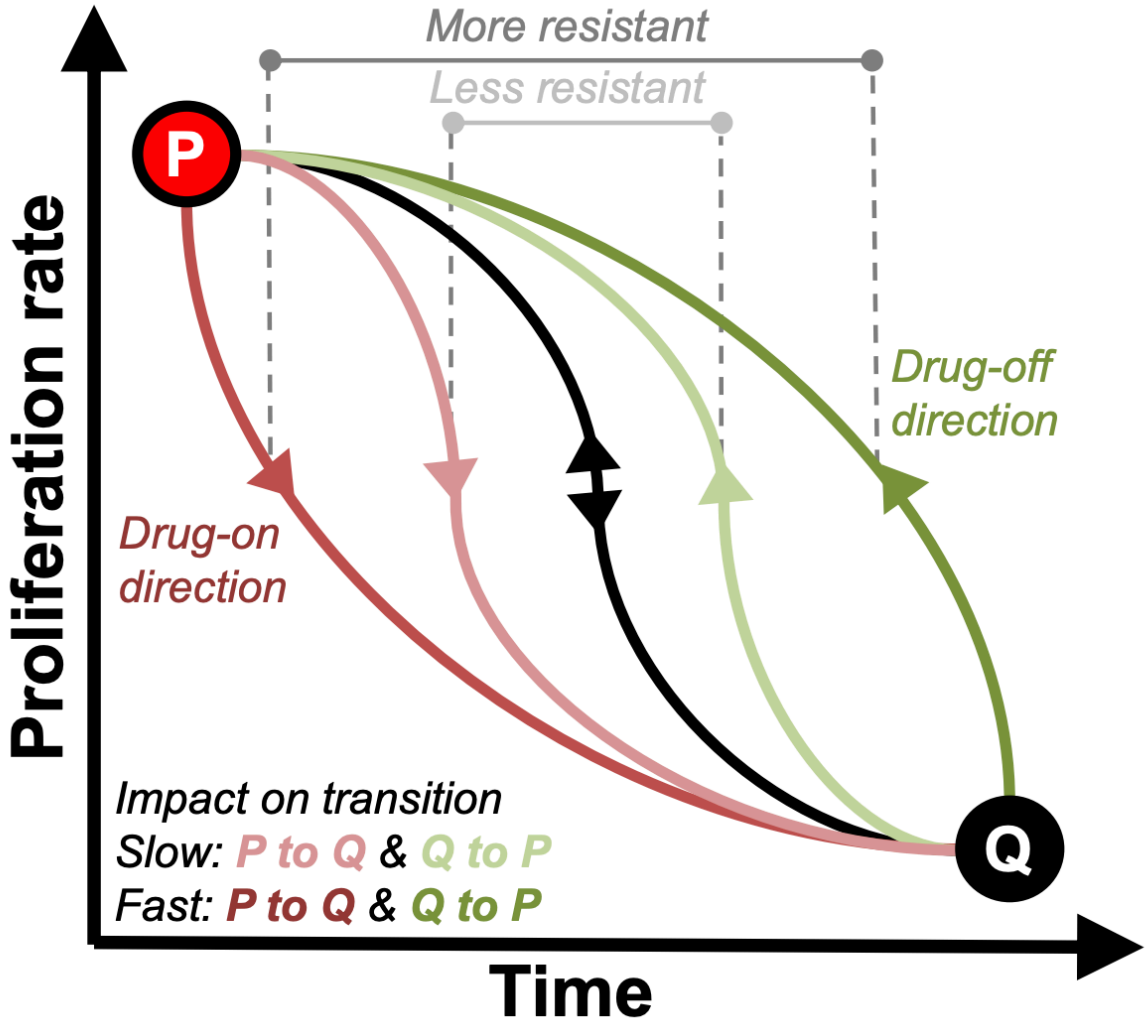


Figure S3.8: A hysteresis model of resistance

Transition between proliferation (P) and quiescence (Q) occurs at each treatment cycle following directional transitions on the hysteresis diagram: drug-on (red, downward paths) and drug-off (green upward paths) transitions. The functional state of the cells in P and Q, different after each cycle, impacts the effectiveness of the transition: light colors and low time-wise slope: slow path, dark color and steep time-wise slope: fast path.

References

1. Tsimberidou AM, Fountzilas E, Nikanjam M, Kurzrock R. Review of precision cancer medicine: Evolution of the treatment paradigm. Cancer Treatment Reviews. 2020 Jun 1;86:102019.

2. Kim JW, Botvinnik OB, Abudayyeh O, Birger C, Rosenbluh J, Shrestha Y, Abazeed ME, Hammerman PS, DiCara D, Konieczkowski DJ, Johannessen CM, Liberzon A, Alizad-Rahvar AR, Alexe G, Aguirre A, Ghandi M, Greulich H, Vazquez F, Weir BA, Van Allen EM, Tsherniak A, Shao DD, Zack TI, Noble M, Getz G, Beroukhim R, Garraway LA, Ardakani M, Romualdi C, Sales G, Barbie DA, Boehm JS, Hahn WC, Mesirov JP, Tamayo P. Characterizing genomic alterations in cancer by complementary functional associations. Nature Biotechnology. 2016 Apr 18;34(5):539–46.

3. Kim JW, Abudayyeh OO, Yeerna H, Yeang CH, Stewart M, Jenkins RW, Kitajima S, Konieczkowski DJ, Medetgul-Ernar K, Cavazos T, Mah C, Ting S, Van Allen EM, Cohen O, Mcdermott J, Damato E, Aguirre AJ, Liang J, Liberzon A, Alexe G, Doench J, Ghandi M, Vazquez F, Weir BA, Tsherniak A, Subramanian A, Meneses-Cime K, Park J, Clemons P, Garraway LA, Thomas D, Boehm JS, Barbie DA, Hahn WC, Mesirov JP, Tamayo P. Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States. Cell Systems. 2017 Aug;5(2):105-118.e9.

4. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics. 2003 Jul;34(3):267–73.

5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545–50.

6. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011 Jun 15;27(12):1739–40.

7. Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. Nat Rev Genet. 2019 Jul;20(7):404–16.

8. Zhang Y, Zhang Z. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. Cell Mol Immunol. 2020 Aug;17(8):807–21.

9. Nofech-Mozes I, Soave D, Awadalla P, Abelson S. Pan-cancer classification of single cells in the tumour microenvironment. Nat Commun. 2023 Mar 23;14(1):1615.

10. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA Cancer J Clin. 2013 Jan;63(1):11–30.

11. Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform. 2012 May;13(3):281–91.

12. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci. 2022 Jan;31(1):8–22.

13. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C, Fröhling S, Chan EM, Sos ML, Michel K, Mermel C, Silver SJ, Weir BA, Reiling JH, Sheng Q, Gupta PB, Wadlow RC, Le H, Hoersch S, Wittner BS, Ramaswamy S, Livingston DM, Sabatini DM, Meyerson M, Thomas RK, Lander ES, Mesirov JP, Root DE, Gilliland DG, Jacks T, Hahn WC. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009 Nov 5;462(7269):108–12.

14. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014 Jan 1;42(D1):D472–7.

15. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, A Miller R, Digles D, Lopes EN, Ehrhart F, Dupuis LJ, Winckers LA, Coort SL, Willighagen EL, Evelo CT, Pico AR, Kutmon M. WikiPathways: connecting communities. Nucleic Acids Res. 2021 Jan 8;49(D1):D613–21.

16. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems. 2015 Dec;1(6):417–25.

17. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20.

18. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J, Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn WC. Defining a Cancer Dependency Map. Cell. 2017 Jul;170(3):564-576.e16.

19. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999 Oct;401(6755):788–91.

20. Lee DD, Seung HS. Algorithms for Non-negative Matrix Factorization. :7.

21. Vavasis SA. On the Complexity of Nonnegative Matrix Factorization. SIAM J Optim. 2010 Jan;20(3):1364–77.

22. Sokal RR, Rohlf FJ. The Comparison of Dendrograms by Objective Methods. Taxon. 1962;11(2):33–40.

23. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences. 2004 Mar 23;101(12):4164–9.

24. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences. 2014 Mar 4;111(9):3354–9.

25. Joe H. Relative Entropy Measures of Multivariate Dependence. Journal of the American Statistical Association. 1989 Mar 1;84(405):157–64.

26. Linfoot EH. An informational measure of correlation. Information and Control. 1957 Sep 1;1(1):85–9.

27. Hsu JL, Hung MC. The role of HER2, EGFR, and other receptor tyrosine kinases in breast cancer. Cancer Metastasis Rev. 2016 Dec;35(4):575–88.

28. Pernas S, Tolaney SM. Targeting HER2 heterogeneity in early-stage breast cancer. Current Opinion in Oncology. 2020 Nov;32(6):545–54.

29. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, Yang JY, Broom BM, Verhaak RGW, Kane DW, Wakefield C, Weinstein JN, Mills GB, Liang H. TCPA: a resource for cancer functional proteomics data. Nat Methods. 2013 Nov;10(11):1046–7.

30. Moody SE, Schinzel AC, Singh S, Izzo F, Strickland MR, Luo L, Thomas SR, Boehm JS, Kim SY, Wang ZC, Hahn WC. PRKACA mediates resistance to HER2-targeted therapy in breast cancer cells and restores anti-apoptotic signaling. Oncogene. 2015 Apr;34(16):2061–71.

31. Zanconato F, Forcato M, Battilana G, Azzolin L, Quaranta E, Bodega B, Rosato A, Bicciato S, Cordenonsi M, Piccolo S. Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. Nat Cell Biol. 2015 Sep;17(9):1218–27.

32. Chang L, Azzolin L, Di Biagio D, Zanconato F, Battilana G, Lucon Xiccato R, Aragona M, Giulitti S, Panciera T, Gandin A, Sigismondo G, Krijgsveld J, Fassan M, Brusatin G, Cordenonsi M, Piccolo S. The SWI/SNF complex is a mechanoregulated inhibitor of YAP and TAZ. Nature. 2018 Nov;563(7730):265–9.

33. Battilana G, Zanconato F, Piccolo S. Mechanisms of YAP/TAZ transcriptional control. Cell Stress. 2021 Nov;5(11):167–72.

34. Zinatizadeh MR, Miri SR, Zarandi PK, Chalbatani GM, Rapôso C, Mirzaei HR, Akbari ME, Mahmoodzadeh H. The Hippo Tumor Suppressor Pathway (YAP/TAZ/TEAD/MST/LATS) and EGFR-RAS-RAF-MEK in cancer metastasis. Genes Dis. 2021 Jan;8(1):48–60.

35. Chen S, Li Y, Zhu Y, Fei J, Song L, Sun G, Guo L, Li X. SERPINE1 Overexpression Promotes Malignant Progression and Poor Prognosis of Gastric Cancer. J Oncol. 2022;2022:2647825.

36. Li L, Li F, Xu Z, Li L, Hu H, Li Y, Yu S, Wang M, Gao L. Identification and validation of SERPINE1 as a prognostic and immunological biomarker in pan-cancer and in ccRCC. Front Pharmacol. 2023;14:1213891.

37. Burgy M, Jehl A, Conrad O, Foppolo S, Bruban V, Etienne-Selloum N, Jung AC, Masson M, Macabre C, Ledrappier S, Burckel H, Mura C, Noël G, Borel C, Fasquelle F, Onea MA, Chenard MP, Thiéry A, Dontenwill M, Martin S. Cav1/EREG/YAP Axis in the Treatment Resistance of Cav1-Expressing Head and Neck Squamous Cell Carcinoma. Cancers (Basel). 2021 Jun 18;13(12):3038.

38. Wang H, Zhang S, Zhang Y, Jia J, Wang J, Liu X, Zhang J, Song X, Ribback S, Cigliano A, Evert M, Liang B, Wu H, Calvisi DF, Zeng Y, Chen X. TAZ is indispensable for c-MYC-induced hepatocarcinogenesis. J Hepatol. 2022 Jan;76(1):123–34.

39. Cho YS, Jiang J. Hippo-Independent Regulation of Yki/Yap/Taz: A Non-canonical View. Frontiers in Cell and Developmental Biology [Internet]. 2021 [cited 2024 Jan 11];9. Available from: https://www.frontiersin.org/articles/10.3389/fcell.2021.658481

40. Jung S, del Sol A. Multiomics data integration unveils core transcriptional regulatory networks governing cell-type identity. npj Syst Biol Appl. 2020 Aug 24;6(1):1–4.

41. Santiago-Rodriguez TM, Hollister EB. Multi 'omic data integration: A review of concepts, considerations, and approaches. Seminars in Perinatology. 2021 Oct 1;45(6):151456.

42. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, Barretina J, Gelfand ET, Bielski CM, Li H, Hu K, Andreev-Drakhlin AY, Kim J, Hess JM, Haas BJ, Aguet F, Weir BA, Rothberg MV, Paolella BR, Lawrence MS, Akbani R, Lu Y, Tiv HL, Gokhale PC, de Weck A, Mansour AA, Oh C, Shih J, Hadi K, Rosen Y, Bistline J, Venkatesan K, Reddy A, Sonkin D, Liu M, Lehar J, Korn JM, Porter DA, Jones MD, Golji J, Caponigro G, Taylor JE, Dunning CM, Creech AL, Warren AC, McFarland JM, Zamanighomi M, Kauffmann A, Stransky N, Imielinski M, Maruvka YE, Cherniack AD, Tsherniak A, Vazquez F, Jaffe JD, Lane AA, Weinstock DM, Johannessen CM, Morrissey MP, Stegmeier F, Schlegel R, Hahn WC, Getz G, Mills GB, Boehm JS, Golub TR, Garraway LA, Sellers WR. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019 May;569(7757):503–8.

43. Bioconductor [Internet]. [cited 2024 Jan 21]. org.Hs.eg.db. Available from: http://bioconductor.org/packages/org.Hs.eg.db/

44. scikit-learn/scikit-learn at 7e1e6d09bcc2eaeba98f7e737aac2ac782f0e5f1 [Internet]. [cited 2024 Jan 11]. Available from: https://github.com/scikit-learn/scikit-learn/tree/7e1e6d09bcc2eaeba98f7e737aac2ac782f0e5f1

45. Paatero P, Hopke PK, Song XH, Ramadan Z. Understanding and controlling rotations in factor analytic models. Chemometrics and Intelligent Laboratory Systems. 2002 Jan 28;60(1):253–64.

46. Domínguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, Howlett SK, Suchanek O, Polanski K, King HW, Mamanova L, Huang N, Szabo PA, Richardson L, Bolt L, Fasouli ES, Mahbubani KT, Prete M, Tuck L, Richoz N, Tuong ZK, Campos L, Mousa HS, Needham EJ, Pritchard S, Li T, Elmentaite R, Park J, Rahmani E, Chen D, Menon DK, Bayraktar OA, James LK, Meyer KB, Yosef N, Clatworthy MR, Sims PA, Farber DL, Saeb-Parsy K, Jones JL, Teichmann SA. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022 May 13;376(6594):eabl5197.

47. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. Nat Rev Genet. 2021 Feb;22(2):71–88.

48. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 2021 Oct;22(10):627–44.

49. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Frontiers in Genetics [Internet]. 2019 [cited 2024 Jan 21];10. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2019.00317

50. Reich M, Tabor T, Liefeld T, Thorvaldsdóttir H, Hill B, Tamayo P, Mesirov JP. The GenePattern Notebook Environment. Cell Systems. 2017 Aug;5(2):149-151.e1.

51. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nature Genetics. 2006 May;38(5):500.

52. Program CSCB, Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, Bezzi E, Cakir B, Chaffer J, Chambers S, Cherry JM, Chi T, Chien J, Dorman L, Garcia-Nieto P, Gloria N, Hastie M, Hegeman D, Hilton J, Huang T, Infeld A, Istrate AM, Jelic I, Katsuya K, Kim YJ, Liang K, Lin M, Lombardo M, Marshall B, Martin B, McDade F, Megill C, Patel N, Predeus A, Raymor B, Robatmili B, Rogers D, Rutherford E, Sadgat D, Shin A, Small C, Smith T, Sridharan P, Tarashansky A, Tavares N, Thomas H, Tolopko A, Urisko M, Yan J, Yeretssian G, Zamanian J, Mani A, Cool J, Carr A. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data [Internet]. bioRxiv; 2023 [cited 2024 Jan 21]. p. 2023.10.30.563174. Available from: https://www.biorxiv.org/content/10.1101/2023.10.30.563174v1

53. Solé-Boldo L, Raddatz G, Schütz S, Mallm JP, Rippe K, Lonsdorf AS, Rodríguez-Paredes M, Lyko F. Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. Commun Biol. 2020 Apr 23;3(1):1–12.

54. Fan X, Bialecka M, Moustakas I, Lam E, Torrens-Juaneda V, Borggreven NV, Trouw L, Louwe LA, Pilgram GSK, Mei H, van der Westerlaken L, Chuva de Sousa Lopes SM. Single-cell reconstruction of follicular remodeling in the human adult ovary. Nat Commun. 2019 Jul 18;10(1):3164.

55. Young MD, Mitchell TJ, Custers L, Margaritis T, Morales-Rodriguez F, Kwakwa K, Khabirova E, Kildisiute G, Oliver TRW, de Krijger RR, van den Heuvel-Eibrink MM, Comitani F, Piapi A, Bugallo-Blanco E, Thevanesan C, Burke C, Prigmore E, Ambridge K,

Roberts K, Braga FAV, Coorens THH, Del Valle I, Wilbrey-Clark A, Mamanova L, Stewart GD, Gnanapragasam VJ, Rampling D, Sebire N, Coleman N, Hook L, Warren A, Haniffa M, Kool M, Pfister SM, Achermann JC, He X, Barker RA, Shlien A, Bayraktar OA, Teichmann SA, Holstege FC, Meyer KB, Drost J, Straathof K, Behjati S. Single cell derived mRNA signals across human kidney tumors. Nat Commun. 2021 Jun 23;12(1):3896.

56. Zhang Y, Narayanan SP, Mannan R, Raskind G, Wang X, Vats P, Su F, Hosseini N, Cao X, Kumar-Sinha C, Ellison SJ, Giordano TJ, Morgan TM, Pitchiaya S, Alva A, Mehra R, Cieslik M, Dhanasekaran SM, Chinnaiyan AM. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. Proceedings of the National Academy of Sciences. 2021 Jun 15;118(24):e2103240118.

57. MacParland SA, Liu JC, Ma XZ, Innes BT, Bartczak AM, Gage BK, Manuel J, Khuu N, Echeverri J, Linares I, Gupta R, Cheng ML, Liu LY, Camat D, Chung SW, Seliga RK, Shao Z, Lee E, Ogawa S, Ogawa M, Wilson MD, Fish JE, Selzner M, Ghanekar A, Grant D, Greig P, Sapisochin G, Selzner N, Winegarden N, Adeyi O, Keller G, Bader GD, McGilvray ID. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun. 2018 Oct 22;9(1):4383.

58. Bhat-Nakshatri P, Gao H, Sheng L, McGuire PC, Xuei X, Wan J, Liu Y, Althouse SK, Colter A, Sandusky G, Storniolo AM, Nakshatri H. A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. Cell Reports Medicine. 2021 Mar 16;2(3):100219.

59. THE TABULA SAPIENS CONSORTIUM. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science. 2022 May 13;376(6594):eabl4896.

60. Benson MJ, Aijö T, Chang X, Gagnon J, Pape UJ, Anantharaman V, Aravind L, Pursiheimo JP, Oberdoerffer S, Liu XS, Lahesmaa R, Lähdesmäki H, Rao A. Heterogeneous nuclear ribonucleoprotein L-like (hnRNPLL) and elongation factor, RNA polymerase II, 2 (ELL2) are regulators of mRNA processing in plasma cells. Proc Natl Acad Sci U S A. 2012 Oct 2;109(40):16252–7.

61. Wenzel AT, Champa D, Venkatesh H, Sun S, Tsai CY, Mesirov JP, Bui JD, Howell SB, Harismendy O. Single-cell characterization of step-wise acquisition of carboplatin resistance in ovarian cancer. npj Syst Biol Appl. 2022 Jun 17;8(1):1–9.

62. Ozols RF, Bundy BN, Greer BE, Fowler JM, Clarke-Pearson D, Burger RA, Mannel RS, DeGeest K, Hartenbach EM, Baergen R, Gynecologic Oncology Group. Phase III trial of carboplatin and paclitaxel compared with cisplatin and paclitaxel in patients with optimally resected stage III ovarian cancer: a Gynecologic Oncology Group study. J Clin Oncol. 2003 Sep 1;21(17):3194–200.

63. Andrews PA, Jones JA, Varki NM, Howell SB. Rapid emergence of acquired cis-diamminedichloroplatinum(II) resistance in an in vivo model of human ovarian carcinoma. Cancer Commun. 1990;2(2):93–100.

64. Galluzzi L, Senovilla L, Vitale I, Michels J, Martins I, Kepp O, Castedo M, Kroemer G. Molecular mechanisms of cisplatin resistance. Oncogene. 2012 Apr 12;31(15):1869–83.

65. Sakai W, Swisher EM, Karlan BY, Agarwal MK, Higgins J, Friedman C, Villegas E, Jacquemont C, Farrugia DJ, Couch FJ, Urban N, Taniguchi T. Secondary mutations as a mechanism of cisplatin resistance in BRCA2-mutated cancers. Nature. 2008 Feb 28;451(7182):1116–20.

66. Brown R, Curry E, Magnani L, Wilhelm-Benartzi CS, Borley J. Poised epigenetic states and acquired drug resistance in cancer. Nat Rev Cancer. 2014 Nov;14(11):747–53.

67. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S, McDermott U, Azizian N, Zou L, Fischbach MA, Wong KK, Brandstetter K, Wittner B, Ramaswamy S, Classon M, Settleman J. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. Cell. 2010 Apr 2;141(1):69–80.

68. Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, Sigal A, Milo R, Cohen-Saidon C, Liron Y, Kam Z, Cohen L, Danon T, Perzov N, Alon U. Dynamic proteomics of individual cancer cells in response to a drug. Science. 2008 Dec 5;322(5907):1511–6.

69. Gascoigne KE, Taylor SS. Cancer cells display profound intra- and interline variation following prolonged exposure to antimitotic drugs. Cancer Cell. 2008 Aug 12;14(2):111–22.

70. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, Beqiri M, Sproesser K, Brafford PA, Xiao M, Eggan E, Anastopoulos IN, Vargas-Garcia CA, Singh A, Nathanson KL, Herlyn M, Raj A. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017 Jun 15;546(7658):431–5.

71. Shah MA, Schwartz GK. Cell cycle-mediated drug resistance: an emerging concept in cancer therapy. Clin Cancer Res. 2001 Aug;7(8):2168–81.

72. Kondoh E, Mori S, Yamaguchi K, Baba T, Matsumura N, Cory Barnett J, Whitaker RS, Konishi I, Fujii S, Berchuck A, Murphy SK. Targeting slow-proliferating ovarian cancer cells. Int J Cancer. 2010 May 15;126(10):2448–56.

73. Abada P, Howell SB. Regulation of Cisplatin cytotoxicity by cu influx transporters. Met Based Drugs. 2010;2010:317581.

74. Aslam MA, Alemdehy MF, Pritchard CEJ, Song JY, Muhaimin FI, Wijdeven RH, Huijbers IJ, Neefjes J, Jacobs H. Towards an understanding of C9orf82 protein/CAAP1 function. PLoS One. 2019;14(1):e0210526.

75. Zhang Y, Johansson E, Miller ML, Jänicke RU, Ferguson DJ, Plas D, Meller J, Anderson MW. Identification of a conserved anti-apoptotic protein that modulates the mitochondrial apoptosis pathway. PLoS One. 2011;6(9):e25284.

76. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018 Jan 4;46(D1):D649–55.

77. Yang AD, Fan F, Camp ER, van Buren G, Liu W, Somcio R, Gray MJ, Cheng H, Hoff PM, Ellis LM. Chronic oxaliplatin resistance induces epithelial-to-mesenchymal transition in colorectal cancer cell lines. Clin Cancer Res. 2006 Jul 15;12(14 Pt 1):4147–53.

78. Pavan S, Olivero M, Corà D, Di Renzo MF. IRF-1 expression is induced by cisplatin in ovarian cancer cells and limits drug effectiveness. Eur J Cancer. 2013 Mar;49(4):964–73.

79. Brzostek-Racine S, Gordon C, Van Scoy S, Reich NC. The DNA damage response induces IFN. J Immunol. 2011 Nov 15;187(10):5336–45.

80. Marchion DC, Cottrill HM, Xiong Y, Chen N, Bicaku E, Fulp WJ, Bansal N, Chon HS, Stickles XB, Kamath SG, Hakam A, Li L, Su D, Moreno C, Judson PL, Berchuck A, Wenham RM, Apte SM, Gonzalez-Bosquet J, Bloom GC, Eschrich SA, Sebti S, Chen DT, Lancaster JM. BAD phosphorylation determines ovarian cancer chemosensitivity and patient survival. Clin Cancer Res. 2011 Oct 1;17(19):6356–66.

81. Pénzváltó Z, Lánczky A, Lénárt J, Meggyesházi N, Krenács T, Szoboszlai N, Denkert C, Pete I, Győrffy B. MEK1 is associated with carboplatin resistance and is a prognostic biomarker in epithelial ovarian cancer. BMC Cancer. 2014 Nov 18;14:837.

82. Persons DL, Yazlovitskaya EM, Cui W, Pelling JC. Cisplatin-induced activation of mitogen-activated protein kinases in ovarian carcinoma cells: inhibition of extracellular signal-regulated kinase activity increases sensitivity to cisplatin. Clin Cancer Res. 1999 May;5(5):1007–14.

83. Cui W, Yazlovitskaya EM, Mayo MS, Pelling JC, Persons DL. Cisplatin-induced response of c-jun N-terminal kinase 1 and extracellular signal--regulated protein kinases 1 and 2 in a series of cisplatin-resistant ovarian carcinoma cell lines. Mol Carcinog. 2000 Dec;29(4):219–28.

84. Beaufort CM, Helmijr JCA, Piskorz AM, Hoogstraat M, Ruigrok-Ritstier K, Besselink N, Murtaza M, van IJcken WFJ, Heine AAJ, Smid M, Koudijs MJ, Brenton JD, Berns EMJJ, Helleman J. Ovarian cancer cell line panel (OCCP): clinical importance of in vitro morphological subtypes. PLoS One. 2014;9(9):e103988.

85. Su Y, Wei W, Robert L, Xue M, Tsoi J, Garcia-Diaz A, Homet Moreno B, Kim J, Ng RH, Lee JW, Koya RC, Comin-Anduix B, Graeber TG, Ribas A, Heath JR. Single-cell analysis resolves the cell state transition and signaling dynamics associated with melanoma drug-induced resistance. Proc Natl Acad Sci U S A. 2017 Dec 26;114(52):13679–84.

86. Roemhild R, Gokhale CS, Dirksen P, Blake C, Rosenstiel P, Traulsen A, Andersson DI, Schulenburg H. Cellular hysteresis as a principle to maximize the efficacy of antibiotic therapy. Proc Natl Acad Sci U S A. 2018 Sep 25;115(39):9767–72.

87. Celià-Terrassa T, Bastian C, Liu DD, Ell B, Aiello NM, Wei Y, Zamalloa J, Blanco AM, Hang X, Kunisky D, Li W, Williams ED, Rabitz H, Kang Y. Hysteresis control of epithelial-mesenchymal transition dynamics conveys a distinct program with enhanced metastatic ability. Nat Commun. 2018 Nov 27;9(1):5005.

88. Bell CC, Gilan O. Principles and mechanisms of non-genetic resistance in cancer. Br J Cancer. 2020 Feb;122(4):465–72.

89. Nakad R, Schumacher B. DNA Damage Response and Immune Defense: Links and Mechanisms. Front Genet. 2016;7:147.

90. Price FD, von Maltzahn J, Bentzinger CF, Dumont NA, Yin H, Chang NC, Wilson DH, Frenette J, Rudnicki MA. Inhibition of JAK-STAT signaling stimulates adult satellite cell function. Nat Med. 2014 Oct;20(10):1174–81.

91. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

92. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

93. broadinstitute/picard [Internet]. Broad Institute; 2024 [cited 2024 Jan 21]. Available from: https://github.com/broadinstitute/picard

94. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing.

95. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164.

96. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Res. 2015 Mar 31;43(6):e39.

97. bcbio/bcbio-nextgen [Internet]. Blue Collar Bioinformatics; 2024 [cited 2024 Jan 21]. Available from: https://github.com/bcbio/bcbio-nextgen

98. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2;17(1):10–2.

99. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014 May;32(5):462–4.

100. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

101. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun 13;177(7):1888-1902.e21.

102. Xue JY, Zhao Y, Aronowitz J, Mai TT, Vides A, Qeriqi B, Kim D, Li C, de Stanchina E, Mazutis L, Risso D, Lito P. Rapid non-uniform adaptation to conformation-specific KRAS(G12C) inhibition. Nature. 2020 Jan;577(7790):421–5.

103. Cannoodt R, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M, Lambrecht B, Preter KD, Saeys Y. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development [Internet]. bioRxiv; 2016 [cited 2024 Jan 21]. p. 079509. Available from: https://www.biorxiv.org/content/10.1101/079509v2

104. Cannoodt R. rcannood/SCORPIUS [Internet]. 2023 [cited 2024 Jan 21]. Available from: https://github.com/rcannood/SCORPIUS

105. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, Meir Z, Hoichman M, Lifshitz A, Tanay A. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 2019 Oct 11;20(1):206.