NAR Genomics and Bioinformatics, 2025, **7**, lqaf041
https://doi.org/10.1093/nargab/lqaf041
Computational Methods for Transcriptome Analysis

OXFORD

TRanscriptome ANalysis of StratifiEd CohorTs (TRANSECT) enables automated assessment of global gene regulation linked to disparate expression in user defined genes and gene sets

John Toubia[1,2,3,4,*,†], **Yasir Kusay**[1,2,3,4,†], **Muneeza Maqsood**[1,3,4], **Nicholas I. Warnock**[1,2,3,4], **David M. Lawrence**[1,2,3,4], **Cameron P. Bracken**[1,5], **Philip A. Gregory**[1,5], **Winnie L. Kan**[6], **Luke A. Selth**[5,7,8], **Simon J. Conn**[7], **Angel F. Lopez**[5,6], **Susan Branford**[1,3,4,5], **Hamish S. Scott**[1,3,4,5], **Chung Hoow Kok**[1,2,3,5,‡], **Gregory J. Goodall**[1,5,‡], **Andreas W. Schreiber**[1,9,10,‡]

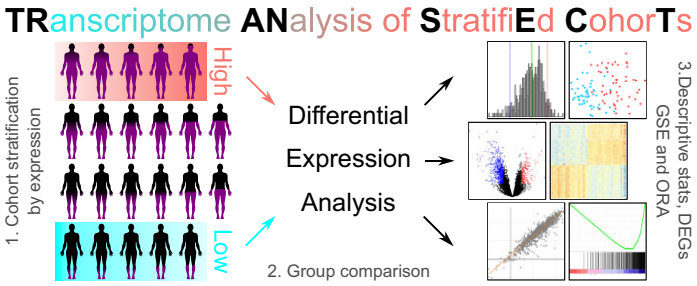
[1]Centre for Cancer Biology, University of South Australia and SA Pathology, Adelaide 5000, Australia
[2]Data and Bioinformatics Innovation, Department of Genetics and Molecular Pathology, SA Pathology, Adelaide 5000, Australia
[3]Department of Genetics and Molecular Pathology, Centre for Cancer Biology, SA Pathology, Adelaide 5000, Australia
[4]Clinical and Health Sciences, University of South Australia, Adelaide 5000, Australia
[5]Adelaide Medical School, The University of Adelaide, Adelaide 5000, Australia
[6]Cytokine Receptor Laboratory, Centre for Cancer Biology, SA Pathology and the University of South Australia, Adelaide 5000, Australia
[7]Flinders University, College of Medicine and Public Health, Flinders Health and Medical Research Institute, Adelaide 5042, South Australia
[8]Flinders University, College of Medicine and Public Health, Freemasons Centre for Male Health and Wellbeing, Adelaide 5042, Australia
[9]ACRF Genomics Facility, Centre for Cancer Biology, An alliance between SA Pathology and the University of South Australia, Adelaide 5000, Australia
[10]School of Biological Sciences, University of Adelaide, Adelaide 5000, Australia

[*]To whom correspondence should be addressed. Email: john.toubia@adelaide.edu.au
[†]The first two authors contributed equally to this work.
[‡]The last three authors contributed equally to this work.

Abstract

Publicly accessible expression data produced by large consortium projects like TCGA and GTEx are increasing in number and size at an unprecedented rate. Their utility cannot be underestimated given the diversity of valuable tools widely used to interrogate these data and the many discoveries of biological and clinical significance already garnered from these datasets. However, there remain undiscovered ways to mine these rich resources and a continuing need to provide researchers with easily accessible and user-friendly applications for complex or bespoke analyses. We introduce TRanscriptome ANalysis of StratifiEd CohorTs (TRANSECT), a bioinformatics application automating the stratification and subsequent differential expression analysis of cohort data to provide further insights into gene regulation. TRANSECT works by defining two groups within a cohort based on disparate expression of a gene or a gene set and subsequently compares the groups for differences in global expression. Akin to reverse genetics minus the inherent requirement of *in vitro* or *in vivo* perturbations, cell lines or model organisms and all the while working within natural physiological limits of expression, TRANSECT compiles information about global transcriptomic change and functional outcomes. TRANSECT is freely available as a command line application or online at https://transect.au.

Graphical abstract



Received: October 31, 2024. **Revised:** February 9, 2025. **Editorial Decision:** March 24, 2025. **Accepted:** March 27, 2025

Introduction

The appropriate and timely expression of mRNAs are crucial for the proper functioning of individual cells, tissues, and the greater organism. This process is exquisitely regulated throughout the entire life of a cell [1] and is a major determinant of the cellular phenotype and function [2]. As such, gene expression studies (the monitoring of mRNA abundance) are fundamental to our understanding and knowledge of the intricate mechanisms governing cellular life and processes. Knowing a gene's expression pattern—when, where, and to what degree it is expressed—is essential to comprehending the activity and biological functions of the protein it encodes. Furthermore, variations in the expression patterns of several genes together can offer insights into regulatory processes, as well as more general cellular activities and metabolic pathways.

There are many ways to measure mRNA abundance and detect changes in gene expression; however, RNA sequencing (RNA-Seq) [3] has evolved as arguably the most popular and powerful method for global transcriptomic analysis. Numerous steps are required to process raw RNA-Seq data into meaningful information from which biological insights can be gleaned. Over time, this processing has largely converged to encompass a standard set of bioinformatic steps. Despite this, differences in sequencing library preparation, reference genome builds and gene annotations used during data processing can introduce sufficient variation to make direct comparison or aggregation of results fraught with risk. In an effort to make these independently created and diverse studies amenable to large-scale integration, many efforts to apply consistent data processing to harmonize collections of RNA-seq data from different sources have been made [4–7].

Currently, there exists an unprecedented amount of publicly available RNA-seq expression profiling data both from targeted experiments stored in online repositories such as Gene Expression Omnibus [8] as well as collections of cohort data from groups of related and unrelated individuals accessed through repositories such as the Genotype-Tissue Expression database (GTEx) [9, 10] and the Cancer Genome Atlas (TCGA) [11]. These datasets, either independently or together with other forms of genomic and proteomic data, have led to many discoveries relating to human disease of biological and clinical significance [12–18]. Gene expression measurements in particular have been ascribed as the top-scoring most informative prognostic biomarkers for the average cancer type, accounting for 46% of all prognostic markers [19]. Thus, in-depth analysis of these resources has already and will undoubtedly continue to reveal novel insights into human health of biological and clinical relevance.

To this end, there currently exists an ever-growing number of effective analysis and visualization tools for the interrogation of gene expression data from public cohort sources. These include online applications such as cBioPortal [20, 21], Xena [22], FireBrowse (http://firebrowse.org/), GEPIA [23, 24], TCGAnalyzeR [25], and Web-TCGA [26] as well as standalone applications like TCGAbiolinks [27, 28], the newly released CRUX [29] and TCGAplot [30], among others. While these applications are exceptionally valuable to the research community, widely adopted and used, there remain unexplored angles to utilize and draw inference from these sources. Of particular interest, differential expression analyses (DEA) between groups are available in many of these applications, most commonly enabling normal-tumor comparisons [31, 32] or comparisons between different cancer subtypes. However, more bespoke DEA comparisons are challenging and remain in the domain only for the bioinformatically proficient.

To address this, we developed TRanscriptome ANalysis of StratifiEd CohorTs (TRANSECT), a standalone UNIX application and Web-accessible data mining and exploration tool to facilitate scientific inquiry using publicly available transcriptomic cohort data. The main function of TRANSECT is to stratify individuals within a large cohort into distinct groups (strata) based solely on gene expression. The strata are subsequently compared to each other to assess for global expression differences associated with distinct levels of expression for a particular gene or gene set. The partitioning of a cohort by a single gene's expression followed by a comparison between the resulting strata has been seen in literature [33]; however, at present, there does not exist a publicly available utility to easily accomplish this. Furthermore, the ability to combine expression profiles of multiple genes for subsequent partitioning and differential expression is highly complex. Having previously applied these types of analyses to studies of classes of cytokine receptor signaling responses [34] and to expression of individual genes in breast cancers stratified according to epithelial versus mesenchymal expression patterns [35], we have implemented the TRANSECT tool to facilitate such analyses for any chosen gene signature in a variety of cohort databases.

This type of analysis has only in recent times become feasible thanks to the large number of participants in cohort studies, and its power and utility will increase as these collections continue to grow in both size and diversity. Results from this type of analysis can be used to rapidly investigate transcriptome state within a natural physiological expression range and without economic burden. Despite this, it has been underutilized due to the lack of accessible tools to facilitate its implementation. TRANSECT addresses this deficit and can be applied to any of the growing number of public transcriptomic cohort studies. Additionally, TRANSECT can be applied to cohort data from any source including proteomic data, a combination of data types (multimodal), and even on non-biological datasets. In the age of big biological data, this tool will increasingly enhance our ability to query diverse datasets, driving hypothesis generation and testing toward advancements in scientific research and health outcomes.

Materials and methods

Implementation

Command line application

TRANSECT is an application that is freely available and is licensed under the MIT license. TRANSECT was developed to run solely on Ubuntu Linux as a command line application; however, it can also be implemented on other operating systems such as MacOS. The application requires substantial resources both in terms of the physical hardware required to retrieve, store, and compute as well as the many vitally important third-party packages required to run the analysis (Supplementary Table S1). Differential expression, graphical representation, and the subsequent functional annotation operations of TRANSECT are implemented through edgeR [36], Glimma [37], GSEA [38, 39], and WebGestalt [40]. There are two main operations of the application, *Prepare* and *Analyse* (Fig. 1A). *Prepare* is a process that retrieves expression data

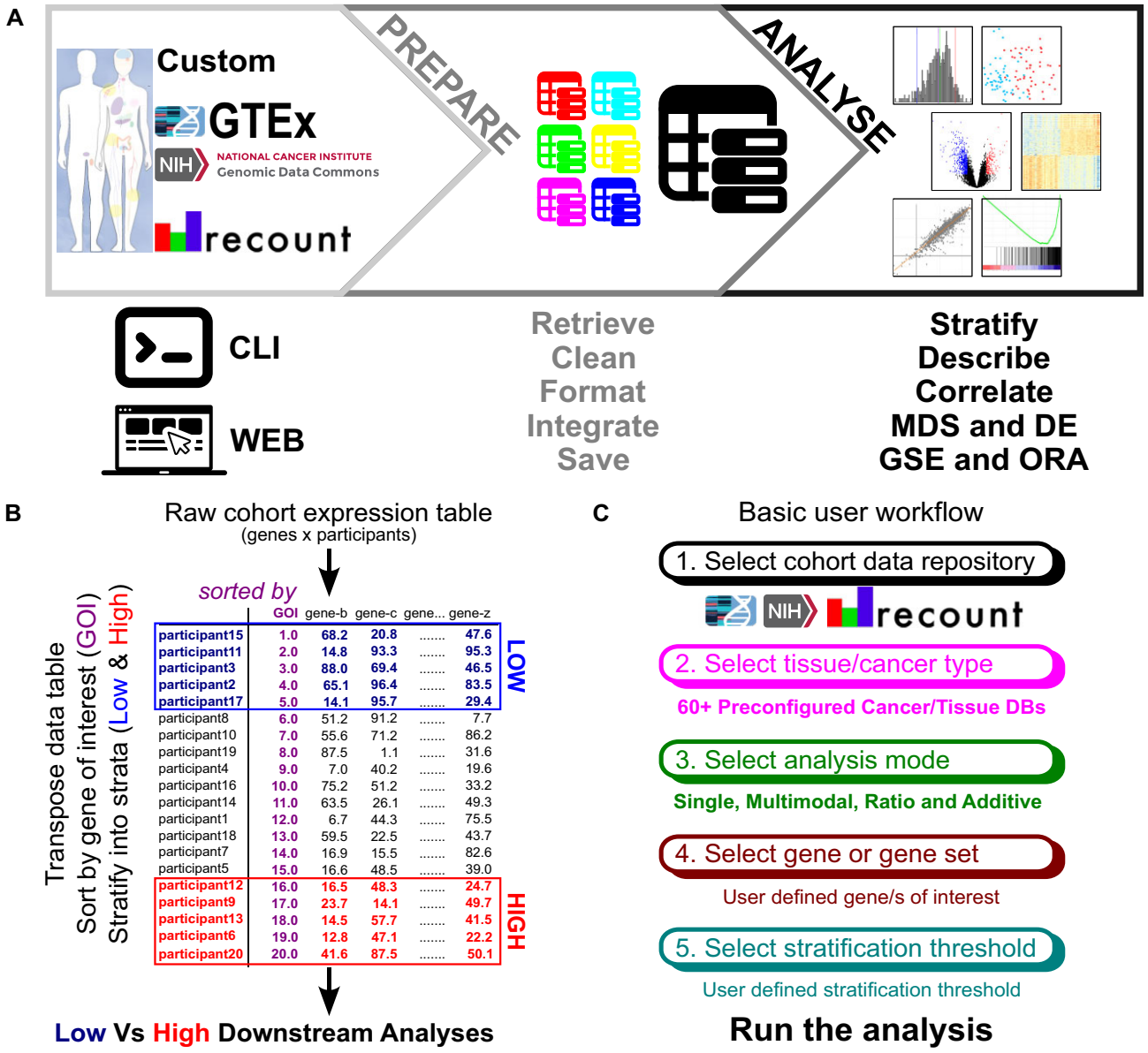


Figure 1. Schematic of the TRANSECT workflow and stratification process. (**A**) Schema depicting TRANSECTs' primary cohort data sources, the main operations: Prepare and Analyze and, the functionality of each. (**B**) A simplified example of the stratification process by a single gene of interest (GOI) and the partitioning into distinct strata (low and high), based solely on gene expression. (**C**) The basic required user input and workflow of TRANSECT for both the web and command line applications. The user first selects the cohort repository followed by the tissue or cancer type within. Subsequently, the user chooses between analysis modes and selects the appropriate gene(s) of interest to match. Finally, the user chooses the percentile threshold for inclusion into low and high stratum.

from online repositories and prepares them (if required) for analysis. *Analyse* is a process that, using the prepared public data, conducts the stratification and differential expression analysis and outputs the results in tables and figures. TRANSECT is a multi-language application built in R, Python, and Bash, which is accessible at https://github.com/twobeers75/TRANSECT. Installation instructions as well as documentation for its use are also available.

Web-based application

To assist researchers lacking time or the local resources necessary for this tool, we have implemented a web-based instance that provides a convenient input and output interface to TRANSECT. This website is freely available at https://transect.au where users can submit their analyses, requiring no login nor configuration. The website is hosted on the Australian Research Data Commons (ARDC) Nectar cloud and has been configured for use with the cohort datasets utilized for this study as well as others. We employed NGINX (v1.18.0) [41] as our webserver and our web application was implemented via the Django web framework (v5.0.1) (https://www.djangoproject.com/). The front-end is written with HTML5, CSS, and Javascript with additional libraries including Bootstrap (v5.0.2), jQuery (v3.7.1), Datatables (v1.3.10), Plotly (v1.53.0) and Popper (v2.92.0). We also maintain a PostgreSQL database (v14.12) that keeps a record of all gene names, all available analysis databases and all recently submitted analyses, which can be queried via Django. When a user submits a job, the application checks for an identical pre-existing request and returns it. Otherwise, a new entry is

created and run asynchronously on the server with the help of the Celery task queue (v5.3.6) (https://docs.celeryq.dev/en/stable/index.html). The code for the web application is available (https://github.com/SACGF/transect_web) with an extensive Wiki for those wishing to delve more into the webtool.

Data and processing

TRANSECT uses publicly available transcriptomic data from large cohort datasets. The current implementation of TRANSECT retrieves expression measurements in the form of raw and TPM normalized counts from the GTEx [9, 10] and TCGA [11] studies as well as retrieving the same data, uniformly processed, through the RECOUNT3 [7] project. Users are advised to work with the RECOUNT3 data for most cases. However, the use of the original GTEx and Genomic Data Commons (GDC) TCGA data is available to those wanting to explore these resources to make comparisons to the harmonized data or for other purposes. Simple statistics about cohort size and disk space requirements for the storage of the resultant database are shown in Supplementary Table S2. A limitation of the application pertains to the size of the cohort, which is required to be large in participant numbers to achieve adequate stratification. The web-accessible version of TRANSECT is restricted to datasets having over 100 participants. The databases are stored in a human readable tab-delimited file format with cohort participant IDs in columns and gene IDs in rows as depicted in Supplementary Table S3. Custom built databases for any suitable cohort dataset (public or private) can be generated in the same manner for use by the application (basic coding skills required, see the TRANSECT online manual for examples and detailed instructions https://transect.au/analysis/manual). A depiction of the workflow and settings used by TRANSECT are shown in Supplementary Table S4.

Functionality

The basic premise of TRANSECT is to stratify individuals from large cohort transcriptomic data into defined strata (plural for stratum) based on single gene expression or multiple gene expression sets. The stratified participant strata are subsequently compared one to the other in order to assess global expression changes and functional differences (Fig. 1B). TRANSECT requires five parameters to be set at a minimum (Fig. 1C). Both gene(s) and number of members within each stratum (stratification threshold) are user-defined and customizable. Below we provide recommendations for the appropriate selection of *n* members per group.

Single analysis

Single gene stratification is simply the division of individuals within a cohort population into distinct strata based solely on the expression of a single gene. In the current version of TRANSECT, individuals with expression levels at or near both ends of the physiological limits for the gene of interest are grouped separately and subsequently compared (Fig. 1B).

Composite analysis—additive mode

Composite analyses use information from multiple genes simultaneously to divide individuals within a cohort population into distinct strata. Consider the case where a cellular process requires the cooperation of multiple genes to function in a particular manner. In this scenario, we expect all cooperative genes in this process to be expressed at or near their respective physiological extreme to achieve the required outcome. It may be the case that this conceived cellular function is perturbed or modified in the absence of one or more of these components. The additive mode of TRANSECT uses expression information from multiple genes (2–5 genes in the current implementation) to rank individuals expressing each of the component genes at near to physiological extreme and separate them into low and high strata. This is achieved by computing the average of rank positions, a simple form of rank aggregation [42, 43], across all component genes for each participant and using the metric to position each individual within the cohort in order. Once this is achieved, individuals with extreme high average rank positions can be grouped and compared to individuals with extreme low rank positions.

Composite analysis—ratio mode

In the same manner as the additive mode described above, the ratio mode also considers information from multiple genes simultaneously to partition individuals within a cohort population into distinct strata. Consider a cellular process where one gene acts antagonistically with another gene. In this case we assume gene A's maximum effect is achieved when its competitive element, gene B is conversely expressed (ie. gene A high: gene B low), and both at their respective extremes. This can also extend to non-competitive partners that require a fine-tuned balance or "ratio" of expression which, when perturbed, can lead to alternate outcomes for the cell. To identify individuals in a cohort that comply with the above, TRANSECT calculates a simple log-ratio statistic between two genes for each patient. Extremely low ratio scores will demarcate participants where geneA >>> geneB and vice versa. Again, individuals at both extremes are grouped and compared.

Multimodal analysis

In select large cohort studies there exist measurements derived from multiple omics for the same individual at the same or similar timepoints. For example, the TCGA study consists of RNA (mRNA, miRNA) and DNA (methylation, mutation, and copy number) data in addition to the associated global proteomics data generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [44]. The integrated analysis of multi-omics data is commonplace, conducted widely and facilitates great advances in disease research [45]. TRANSECT has the facility to survey changes in one omics data type based on the stratification of individuals using matched data from another omics. For portrayal purposes, consider assessing the consequence of elevated levels of a miRNA on the global expression of all other miRNAs. This may not be a misplaced task. However, it is well known that most functioning miRNA primarily exert their influence on mRNA targets. Therefore, it may be better placed to assess the effect of the miRNA on global expression of all mRNAs. As in the use cases above, individuals at each extreme are grouped and compared (in case study 4 below, we depict this facility using TCGA miRNA and mRNA data).

Results

To introduce and describe the functionality of TRANSECT, we first describe the outcomes of repeated random stratification trials to interrogate baseline results and to select appropriate parameters for later investigations. Subsequently, us-

ing three different GTEx tissue (Prostate, Blood, and Breast) and TCGA cancer types (PRAD, LAML, and BRCA) as case study examples, we conduct tests using well studied gene(s) with recognized effects on cellular outcomes. For each case, we use a different stratification strategy and explore the descriptive statistics of the resulting strata and the outputs of differential expression and enrichment analyses that are provided by TRANSECT. Finally, we comment on the conformity of TRANSECT results compared to the documented effects of the gene(s) in question.

Random stratification—determination of optimal numbers for analysis

In order to explore the effect of stratum size on analysis outcomes, we first performed random cohort stratification rather than expression-based stratification. To accomplish this, we randomly selected and assigned cohort members into one of two non-overlapping groups (random sampling without replacement) to achieve n individuals per group, where $n = 5$–50 (in increments of 5), 75 or 100. We repeated the random allocation for each trial one hundred consecutive times (a total of 1 200 random trials) and assessed the resulting number and level of significant differentially expressed genes (DEGs). To minimize sex differences from unintentional allocation of males and females into alternate strata (confounding factor), we used the RECOUNT3 TCGA BRCA (*Breast invasive carcinoma*) dataset refined to only include females. We additionally filtered out non-diseased matched samples for the same reason (*original* $n = 1227$, *filtered* $n = 1122$).

When n was low (<20) the median number of significant DEGs (FDR < 0.05) was also low, however we did observe a number of individual random trials with excessively large numbers of significant DEGs (863 and 1015 for trials n5 and n10, respectively) (Supplementary Fig. S1A). Further investigation revealed that this was caused by known substructure present within the cohort. Specifically, the cohort contains subtypes of breast cancer with different transcriptional profiles. For small n, it is not unlikely for this substructure to filter through to the two extreme strata purely by chance. However, as n increased in size (>20), this phenomenon became statistically less likely, and we found that the maximum number of significant DEGs first stabilized and then became smaller as n was further increased (Supplementary Fig. S1A and associated summary table). Conversely, the median number of significant DEGs continued growing as n increased beyond 20 (Supplementary Fig. S1A and associated summary table) and the minimum FDR returned decreased (Supplementary Fig. S1B), which is likely due to increased statistical power. In short, these trials illustrated the complexities of extracting truly random samples from a cohort with substructure (known and unknown) and the effects of the interplay of cohort size and stratum size. Notionally, selecting a large stratum size overcomes issues caused by cohort substructures, but excessively large stratum size relative to total cohort size can be detrimental as well. Ideally, members in each stratum will share like attributes; however, this becomes less achievable as strata size approaches total cohort size (Supplementary Fig. S1C).

These results suggest that for the cohort considered here, choosing n between 15 and 30 reduces the chance of returning large numbers of DEGs derived from known or unknown cohort substructures whilst maintaining shared attributes within each stratum and aligns well with previously derived power analysis results that suggest 33 samples per group to achieve 80% power [46]. However, the trials also revealed that regardless of n, with moderate FDR cutoffs some false positive DEGs are to be expected. As ultimately no significant DEGs should be returned in a random trial, we proceed by employing a stricter FDR cutoff (FDR < 1e-05) than the commonly adopted FDR < 0.05.

Case study 1: ZEB1 stratified single gene analysis, TCGA PRAD cohort

In this case study, we investigated the effect on the transcriptome associated with extreme variations of expression of a single gene ZEB1, a transcription factor known to promote tumor invasion and metastasis by inducing epithelial-mesenchymal transition (EMT) [47, 48]. For this example, we used harmonized data derived from the RECOUNT3 project for TCGA PRAD (Prostate adenocarcinoma, $n = 554$). We sought to test whether separation of the PRAD cohort into low and high strata based solely on the participants expression of ZEB1 and the subsequent comparison of these strata (e.g. the TRANSECT analysis), would return DEGs known to be regulated by ZEB1 and whether the resulting collection of DEGs matched those from experimentally derived ZEB1 and EMT studies.

Fig. 2A shows the distribution of ZEB1 expression for all individuals in the TCGA PRAD cohort demarcating the thresholds for inclusion into low and high expression strata (blue and red lines, respectively) as well as the average and median expression level for ZEB1 across the cohort (purple and green lines, respectively). Individuals with expression levels beyond the low and high thresholds were grouped ($n = 25$ and 26; average log2 TPM expression of 0 and 4 for low and high strata, respectively) (Fig. 2B), and differential expression analysis between the two strata was performed. The resulting multidimensional scaling (MDS) plot shows clear separation of individuals across dimension 1 based on low and high stratum allocation (Dim1 accounts for ∼30% of the variance between all the individuals) (Fig. 2C). In total, 4318 genes were found to be significantly differentially expressed (FC > 2, FDR < 1e-05), 1959 down- and 2359 upregulated (Fig. 2D), most of which exhibiting a consistent trend between groups (Supplementary main figure data). Known EMT marker genes were found to be amongst the most highly altered between the two groups, including ZEB2, VIM, FN1, and TGFB1 in addition to many extracellular matrix and collagen related genes. Gene set enrichment analysis (GSEA) of the DEA results using the Molecular Signature Database (MSigDB) Hallmark gene sets showed a strong match to the Hallmark EMT collection (NES = 2.73, FDR q-value < 0.0001) (Fig. 2E and F). Concordantly, WebGestalt over-representation analysis (ORA) using the top 500 upregulated genes showed significant matches to GO annotations related to ZEB1 functions and EMT, including extracellular structure organization, cell-substrate adhesion and collagen binding [49, 50] (Fig. 2G).

Using the same approach, we repeated the above test replacing the TCGA PRAD cohort with the GTEx Prostate cohort, again from the RECOUNT3 project. The distribution of ZEB1 expression in the GTEx cohort matched more closely with the TCGA PRAD normal group (Supplementary Fig. S2A). The DEA and associated downstream analyses using

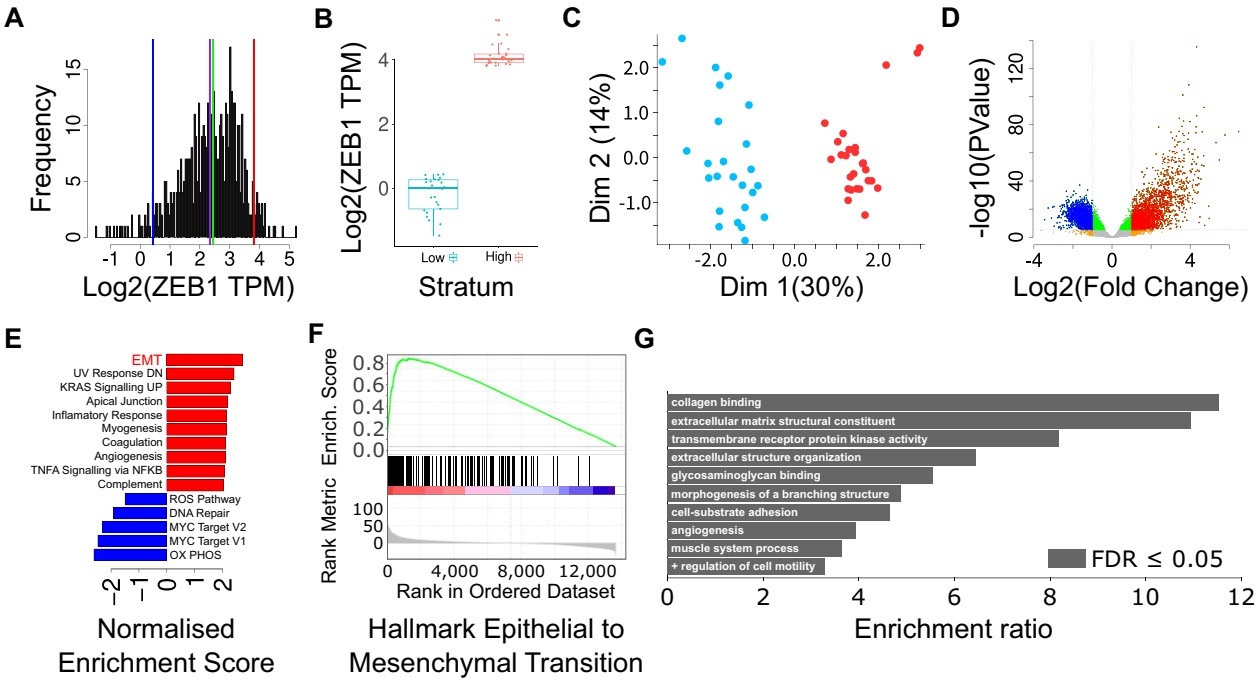


Figure 2. Case study 1: Select TRANSECT outputs from the ZEB1 single gene analysis using the RECOUNT3 TCGA-PRAD cohort dataset. (**A**) Frequency distribution histogram of ZEB1 expression in TCGA PRAD with vertical lines demarcating low stratum threshold in blue, high stratum threshold in red, average and median expression in purple and green, respectively. (**B**) Boxplot of ZEB1 expression solely for participants stratified into low and high strata. (**C**) MDS plot of ZEB1 low stratum members in blue and high members in red. (**D**) Volcano plot (fold change of expression versus significance) of the DEA results with significant upregulated genes in red and downregulated in blue. (**E**) Summary barplot of GSEA normalized enrichment scores against MSigDB Hallmark collections. (**F**) Select GSEA barcode plot of the top positive hit to the Hallmark EMT collection. (**G**) WebGestalt ORA enrichment ratio summary barplot against GO (Biological Process and Molecular Function) functional databases enriched with the top 500 upregulated DEGs.

the GTEx Prostate data returned strikingly similar results compared to those returned by the TCGA PRAD analysis (Supplementary Fig. S2B-G), exhibiting TRANSECTs consistency when used on different cohorts.

Case study 2: IL3RA CSF2RB composite ratio analysis, TCGA LAML cohort

In case study 2, we assessed the effect on the transcriptome given the ratio of two cytokine-mediated receptor subunit genes, IL3RA and CSF2RB. Expression of these genes on the cell surface constitute the heterodimeric IL-3 receptor (IL3R) that, when dimerized by the IL3 cytokine, triggers a cascade of cellular events governing cell fate [34]. The overexpression of IL3RA in AML blasts and leukemia stem cells has been demonstrated to be associated with poor patient survival [51, 52]. More recently, we showed that the ratio of IL3RA to CSF2RB (e.g. discordant expression of IL3RA and CSF2RB) drives distinct transcriptional and phosphosignaling programs in AML patient samples leading to poor patient survival [34]. For this example, we again used data derived from the RECOUNT3 project, only this time for TCGA LAML (Acute Myeloid Leukemia, $n = 178$). Here, we tested TRANSECT's ability to derive the significance of altered cytokine receptor expression in leukemia using the relatively small LAML cohort.

IL3RA and CSF2RB show no clear correlation with respect to gene expression ($R = -0.04$, FDR $= 0.6$) and possess very similar average expression levels across the entire cohort (average log2 expression of 5.9 and 5.6 for IL3RA and CSF2RB respectively). Fig. 3A shows the expression of both IL3RA and CSF2RB (y-axis) for each patient (x-axis) ranked from low (left) to high (right) IL3RA:CSF2RB ratio. The plot clearly shows that the average expression ratio between these two genes (5.9:5.6) is not consistently representative of the ratio of expression between these genes independently for each individual in this population. Again, blue and red dashed vertical lines on the plot demarcate the thresholds for inclusion into low and high ratio strata, respectively. Extracting and grouping separately only those individuals with ratio levels beyond the low and high thresholds ($n = 21$ and 22, respectively) (Fig. 3B), we next performed a differential expression analysis between the two strata. The resulting MDS plot does not show a distinct separation of strata as in case study 1 (Dim1 accounts for only 15% of the variance between all the individuals) (Fig. 3C). Concordantly, the DEA returned only 32 significant DEGs (FC > 2, FDR $<$ 1e-05), 20 down and 12 upregulated (Fig. 3D). GSEA analysis of the DEA results using the MSigDB Curated gene sets showed a strong match to the JAATINEN_HEMATOPOIETIC_STEM_CELL_DN collection (NES=-2.73, FDR q-value < 0.0001) (Fig. 3E and F and Supplementary Fig. S3A) as well as matches to many other haematopoietic stem cell related collections. WebGestalt ORA using only downregulated genes with a relaxed FDR threshold to attain adequate numbers for ORA ($n = 397$, FC > 2, FDR < 0.05) showed significant matches to expected GO annotations including inflammatory and immune response and leukocyte activation (Fig. 3G).

This case study is analogous to the analysis we performed using the same approach on a cohort of 672 primary speci-

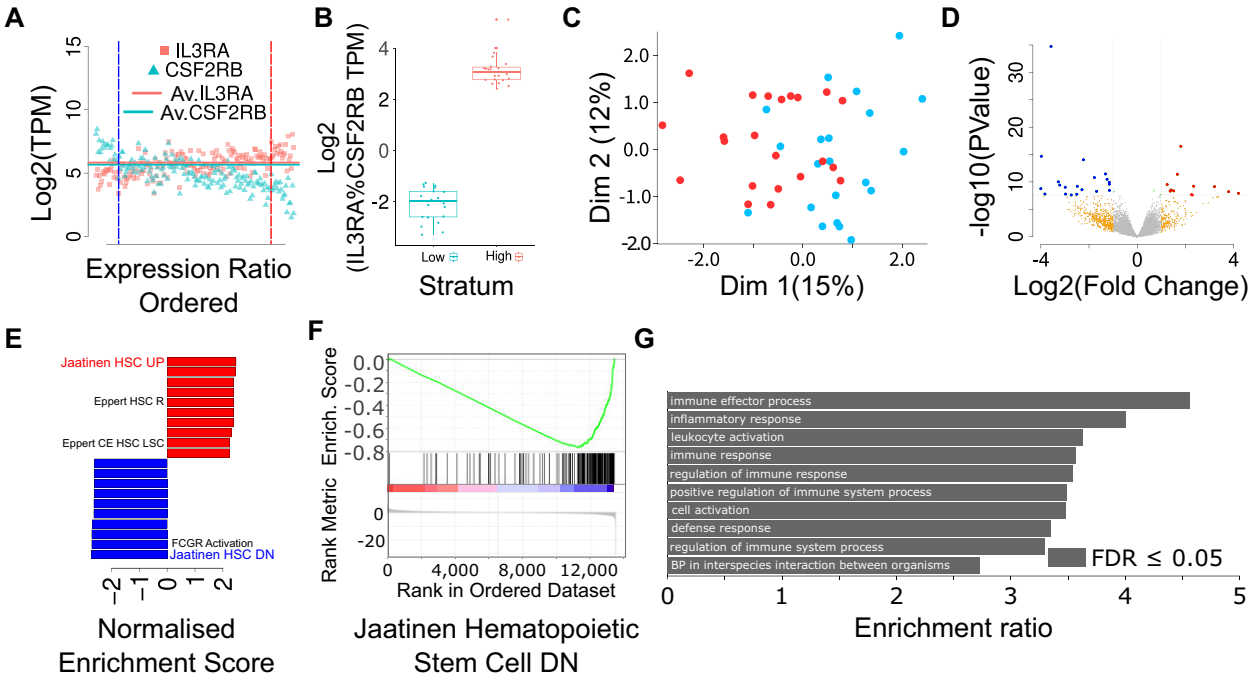


Figure 3. Case study 2 - Select TRANSECT outputs from the IL3RA CSF2RB composite ratio analysis using the RECOUNT3 TCGA-LAML cohort dataset. (**A**) Log2 expression scatterplot of IL3RA (magenta square) and CSF2RB (cyan triangle) for each participant (*y*-axis) ordered by their IL3RA:CSF2RB ratio (low to high, *x*-axis) with vertical lines demarcating low stratum threshold in blue and high stratum threshold in red. In addition, horizontal magenta and cyan lines show average IL3RA and CSF2RB expression, respectively, across all cohort participants. (**B**) Boxplot of IL3RA:CSF2RB expression ratios solely for participants stratified into low and high strata. (**C**) MDS plot of IL3RA:CSF2RB low stratum members in blue and high members in red. (**D**) Volcano plot (fold change of expression versus significance) of the DEA results with significant upregulated genes in red and downregulated in blue. (**E**) Summary barplot of GSEA normalized enrichment scores against MSigDB C2-Curated collections. (**F**) Select GSEA barcode plot of the top negative hit to the C2 Jaatineen Hematopoietic Stem Cell collection. (**G**) WebGestalt ORA enrichment ratio summary barplot against GO (Biological Process and Molecular Function) and Pathways (Kegg and Reactome) functional databases enriched with the 397 most downregulated DEGs.

mens from 562 unique AML patients from the BeatAML cohort [53]. Compared to the TCGA LAML cohort which has only 178 participants, the BeatAML cohort is much larger, providing better stratification of individuals (Supplementary Fig. S3B and Fig. 1A in [34]). Remarkably, our findings with the TCGA LAML cohort are consistent with our published data showing that AML patients with high IL3RA and low CSF2RB expression (i.e. high IL3RA:CSF2RB ratio) have enrichment of stemness programs. In addition, the same analysis using the GTEX Whole Blood cohort (Supplementary Fig. S3C–F) returned similar results albeit with a clearly skewed expression ratio profile compared to the TCGA LAML cohort (Supplementary Fig. S3C). These results demonstrate the robustness of TRANSECT and significance of the differential gene expression results between the high vs low IL3RA:CSF2RB ratio strata across various patient cohorts.

Case study 3: ESR1 PGR ERBB2 composite additive analysis, TCGA BRCA cohort

In case study 3, using the harmonized RECOUNT3 data from TCGA BRCA (Breast invasive carcinoma, $n = 1227$), we explored the effect of three important BRCA related genes. In breast cancer, the absence of the estrogen (ER) and progesterone (PR) receptors (mRNAs: ESR1, PGR respectively) together with low or absent HER2 protein (mRNA: ERRB2) are classified as triple-negative breast cancer (TNBC). TNBC differs from other classifications of breast cancer in that they tend to be more aggressive having poorer prognosis and survival rates in part due to having fewer treatment options [54]. For this example, we employed multiple TRANSECT modes starting with the additive mode to interrogate TNBC gene changes. Following this, we employed a ratio test solely on ESR1 and PGR to survey the changes in the transcriptome in the presence of only one of these receptors.

Fig. 4A shows the distribution of expression across all individuals in the cohort for all three TNBC receptor genes ESR1, PGR, and ERRB2. The average expression and overall distribution differ between the three receptors with ERBB2 possessing the most diverged and tightest distribution in addition to having a small but observable group of individuals with significantly high expression (likely to be HER2-amplified breast cancer individuals). ESR1 and PGR appear to be bimodal, showing a weak but visible secondary distribution around log2(TPM) of zero likely attributable to log transforming very low expression measurements (variations and noise in the RNA-Seq and counting algorithms in the absence of the expression of a gene). As shown in the plot (Fig. 4A), individuals who rank low for expression of all three receptors are marked in cyan (one point per individual on each of the three distributions). Conversely, individuals ranking high for elevated expression of all three genes are marked in magenta. Interestingly, most of the individuals in the cohort with very high levels of ERBB2 do not rank highly for expression of ESR1 and PGR suggesting a different, possibly opposing, pattern of expression for these genes within this group (Fig. 4A). Grouping separately those individuals ranked in low expressing (cyan, considered TNBC) and high

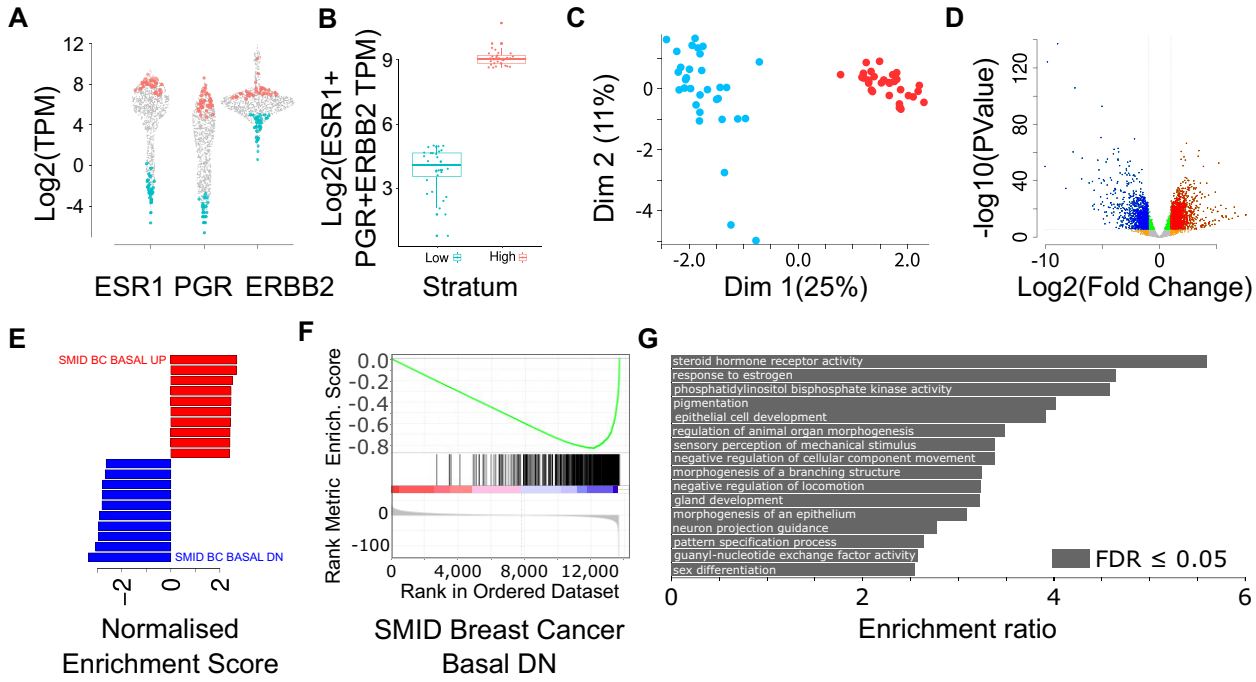


Figure 4. Case study 3 - Select TRANSECT outputs from the ESR1 PGR ERBB2 composite additive analysis using the RECOUNT3 TCGA-BRCA cohort dataset. (**A**) Violin plot showing the distribution of expression for ESR1, PGR and ERBB2 in TCGA BRCA with individual participants classified into low or high strata for all three genes shown in cyan and magenta, respectively. (**B**) Boxplot of the cumulative expression for all three genes solely for participants stratified into low and high strata. (**C**) MDS plot of ESR1 + PGR + ERBB2 low stratum members in blue and high members in red. (**D**) Volcano plot (fold change of expression versus significance) of the DEA results with significant upregulated genes in red and downregulated in blue. (**E**) Summary barplot of GSEA normalized enrichment scores against MSigDB C2-Curated collections. (**F**) Select GSEA barcode plot of the top negative hit to the C2 SMID Breast Cancer Basal DN collection. (**G**) WebGestalt ORA enrichment ratio summary barplot against GO (Biological Process and Molecular Function) functional databases enriched with the top 500 upregulated DEGs.

expressing (magenta, considered non-TNBC) strata (n = 32 and 34, respectively) (Fig. 4B), we conducted a differential expression analysis between these two groups. As we were most interested in the absence of expression for these genes, we swapped the direction of the comparison from the default to assess differences in the low compared to high stratum. The resulting MDS plot shows a clear separation of individuals across dimension 1 based on low and high stratum allocation (Dim1 accounts for near 25% of the variance between all the individuals) (Fig. 4C). In total, 2916 genes were found to be significantly differentially expressed (FC > 2, FDR < 1e-05), 1090 down and 1826 upregulated (Fig. 4D), many of which show consistent trends between groups (Supplementary main figure data). GSEA analysis of the DEA results using the MSigDB Curated gene sets returned best hits to the SMID_BREAST_CANCER_BASAL collections in both positive and negative directions (NES = 2.69 and -3.36 for pos and neg, respectively, FDR q-value < 0.0001) (Fig. 4E and F). This is consistent with the known overlap between the gene signatures of TNBC and basal-like breast cancers [55]. Concordantly, WebGestalt ORA using the top 500 downregulated genes showed significant matches to expected GO annotations including steroid hormone receptor activity and response to estrogen (Fig. 4G).

Auxiliary to the main functions of TRANSECT, one can use the findings from an TRANSECT analysis to probe matching clinical observations. For example, using the available clinical classifications for ER+, PR+ and HER+ per participant in the cohort used here, we can visualize their corresponding mRNA expression levels for each of these genes (Supplementary Fig. S4A). We can also ask where these classifications lie, given the expression of each gene for each participant ordered independently from low to high. Using ESR1 as an example, we can easily and counterintuitively see that the individual in the cohort with the lowest gene expression measurement for ESR1 has a classification of ER+ (Supplementary Fig. S4B). Of more interest here, we can assess the concordance between the TRANSECT stratification and the clinical classification for each participant. This shows overwhelming concordance in the low stratified groups for each gene (average 86% concordance) and the same for ER and PR in the high stratified group (average 92% concordance), but with mixed results for ERBB2 (28% concordance) (Supplementary Fig. S4C). A better way to assess concordance is to build a complete collection of TNBC-IDs based on each participants clinical classifications (ER-, PR- and HER-). Using the ranks derived from TRANSECT converted into z-scores, we can then run a custom pre-ranked gene-set enrichment analysis through GSEA. As expected, there was a strong negative match to the TNBC-ID collection (NES = -3.86, FDR < 0.0001) (Supplementary Fig. S4D). This type of analysis (using TRANSECT ranking together with clinical observations) is, however, not a current feature of TRANSECT due to the array of different clinical observations between different cohort data but can be manually implemented.

As an assessment of the specificity of TRANSECT, we queried the same cohort above for changes in the transcriptome when the estrogen receptor gene (ESR1)

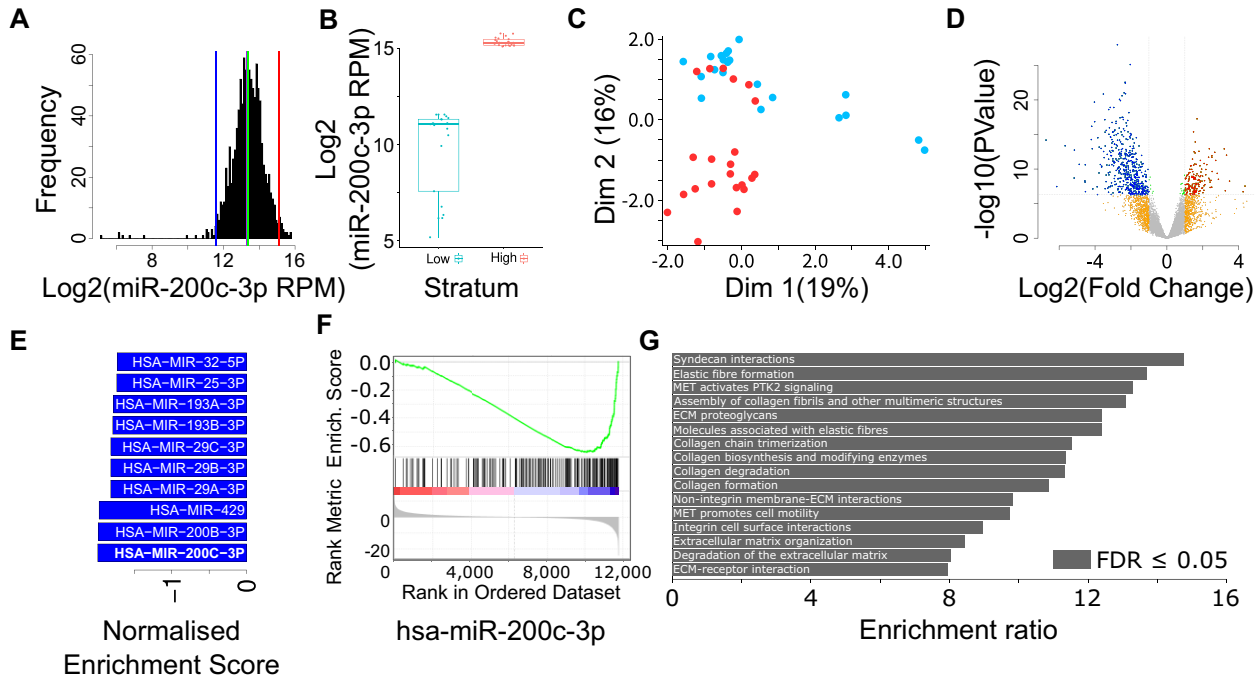


Figure 5. Case study 4 - Select TRANSECT outputs from the hsa-miR-200c-3p multimodal analysis using the RECOUNT3 TCGA-BRCA cohort dataset. (**A**) Frequency distribution histogram of miR-200c-3p expression in TCGA BRCA with vertical lines demarcating low stratum threshold in blue, high stratum threshold in red, average and median expression in purple and green, respectively. (**B**) Boxplot of miR-200c-3p expression solely for participants stratified into low and high strata. (**C**) MDS plot of miR-200c-3p low stratum members in blue and high members in red. (**D**) Volcano plot (fold change of expression versus significance) of the DEA results with significant upregulated genes in red and downregulated in blue. (**E**) Summary barplot of GSEA normalized enrichment scores against TargetScan Human (conserved site context++ score, 8mers) collections. Note: no significant matches in the positive direction. (**F**) Select GSEA barcode plot of the top negative hit to the hsa-miR-200c-3p collection. (**G**) WebGestalt ORA enrichment ratio summary barplot against pathway (KEGG and Reactome) functional databases enriched with the 476 significantly downregulated DEGs.

is expressed in the absence of the PGR. Upon applying the ratio mode of TRANSECT, substantially different results are returned culminating in strong GSEA best matches to SMID_BREAST_CANCER_LUMINAL_B collections (Supplementary Fig. S5A–G). These data are consistent with subgroups of luminal B profile breast cancers expressing ER but not PR [56].

Case study 4: hsa-miR-200c-3p multimodal analysis, TCGA BRCA cohort

In this final case study, we used RECOUNT3 data from TCGA BRCA to explore the effect on the transcriptome given discrete expression of a miRNA. To expand on this, we use measurements from one omics type (smRNA-seq) to stratify participants within the cohort into low and high miRNA expression strata. We then use this stratification to explore the resulting changes when the strata are compared in a completely different omics (mRNA-seq). This can be achieved for most omics when matched data exists for participants across the omics of interest. In this example, we explore the stratification of participants based on the expression of a single miRNA, hsa-miR-200c-3p (single analysis), a key regulatory repressor of EMT and known interactor of the mRNA ZEB1 [57–59]. We sought to test TRANSECT's ability to infer the known influence of a miRNA as measured in participants using one omics assay, on the global transcriptome as measured using a separate assay.

Fig. 5A shows the distribution of hsa-miR-200c-3p expression for all individuals in the TCGA BRCA cohort demarcating the boundaries for inclusion into low and high strata (blue and red lines, respectively) as well as showing the average and median expression levels (purple and green lines, respectively). We extracted and grouped separately the individuals with expression levels beyond the low and high thresholds (n = 21 and 22; average log2 reads per million (RPM) expression of ≈ 11 and ≈ 15 for low and high strata, respectively) (Fig. 5B). Next, we used this stratification to perform a differential expression analysis on matched participants in the mRNA-seq transcriptomics data. The resulting MDS plot shows the separation of individuals mixed across two dimensions based on low and high stratum allocation (Dim1 accounts for near 19% of the variance between all the individuals and Dim2 for 16%) (Fig. 5C). In total, 631 genes were found to be significantly differentially expressed (FC > 2, FDR < 1e-05), 476 down and 155 upregulated (Fig. 5D), most of which showing consistent trends between groups (Supplementary main figure data). As expected, known miR-200 target genes, ZEB1 and ZEB2, were found to be amongst the most highly downregulated genes. GSEA analysis of the DEA results using a custom collection of gene sets (TargetScanHuman - conserved site context++ score, 8mers) showed strong matches to miR-200 family members, miR-200b, miR-200c and miR429 (NES = −1.99, −1.98, −1.97, respectively and FDR q-value < 0.001 for all) (Fig. 5E and F), and no significant positive matches. WebGestalt ORA using the downregulated genes showed significant matches to EMT associated pathways including Syndecan interactions, multiple Collagen and Extracellular matrix pathways, and Focal adhesion (Fig. 5G).

While it is known that miR-200 exerts influence on the cell beyond most other miRs, this case study nevertheless demonstrates the use of TRANSECT to uncover effects on global mRNA expression in a multi-modal manner. Like analyses with alternate omics where available can also be conducted. Additionally, multi-modal analyses can be run in a bi-directional manner. For example, here we could have stratified participants by their mRNA ZEB1 expression and used the resulting stratification results to conduct a DE analysis on the miRNA data.

Limitations

Critical evaluation of analytical tools is necessary to understand their strengths and weaknesses, and to fully appreciate and interpret their outputs accurately. The case studies above provide compelling evidence of TRANSECT's usefulness given an appropriate scientific enquiry. However, it comes with inherent limitations. As mentioned previously, TRANSECT requires "big data" with large numbers of participants in the cohort datasets to adequately achieve appropriate stratification and grouping. Presently, we require more than 100 participants within a cohort to achieve reasonable separation between low and high groups (when selecting $\approx$ 30 members in each). Ideally, individual members of each stratum derived from the stratification process will share highly similar attributes or characteristics (here, gene expression levels). Cohort datasets with low participant numbers are unlikely to possess the required random sampling of a population to achieve defined strata containing members with shared similar attributes or characteristics and may force the allocation of members with different characteristics into the same stratum. As participant numbers grow, the power to stratify "like" members into disparate groups will also increase. Simultaneously, adequately large cohorts will provide the facility to extract and compare strata other than low and high and conduct investigations using a greater number of genes in more complex scenarios than the ratio and additive modes described here.

TRANSECT is most effective when the targets of interest are transcription factors or other regulators that impact significant change to the transcriptome. Differences in the expression of targets that assert changes in other omics without affecting the transcriptome may still be useful if the appropriate multi-omics data are available. In addition, TRANSECT provides no measure of association or causation; this requires follow-up laboratory or *in silico* analysis.

The heterogeneity of cell types within bulk tissue samples that are present in these cohort datasets can lead to misleading observations if not carefully considered. There may also exist distinctions between different omics cohort datasets negating the fine tuning and parameter selection chosen in this study. In these cases, as well as for custom investigations, further evaluation and fine tuning of parameters may be required.

Discussion

TRANSECT is an innovative application providing researchers the ability to determine and examine global expression changes between distinct groups of individuals from large cohorts without requiring extensive computational programming or bioinformatics skills. To our knowledge, no application currently exists to easily achieve stratification of cohorts by user defined gene or gene set expression followed by differential expression analysis and functional annotation. TRANSECT addresses this and can be customized to incorporate any cohort data set, is not restricted by omics type and can even be used across omics where available. The program is freely available online as an interactive web application and can also be retrieved in standalone command line form for local installations. The current online version is configured with over 60 cohort datasets from the GTEx and GDC repositories as well as the uniformly processed data for GTEx and TCGA provided by the RECOUNT3 project. The standalone version includes the necessary scripts to retrieve and prepare select public cohort data and can be configured for use with in-house and private datasets.

Additional functionality and improvements to TRANSECT will enhance its utility and ease of use. Currently, features in production and testing for upcoming versions include: (i) Separation of cohort datasets by known substructures into distinct groups for targeted investigations. (ii) Implementation of advanced edgeR differential-expression tests using generalized linear models (GLMs) alongside the currently employed exact test allowing for the integration of fold-change thresholds into significance tests. GLM functionality will also enable the incorporation of covariates, such as cancer subtypes, into the model eliminating the need to split datasets when participant numbers are low. Future enhancements under development include: (i) Automated export of stratified participant IDs to bioinformatics algorithms such as survival analyses and (ii) Integration with external bioinformatics tools through web links to applications like GEPIA and UCSC Xena, enabling seamless exploratory options for users investigating specific genes or gene sets.

TRANSECT was designed and developed to be an effective mining and exploration tool, enabling users the ability to effortlessly query large sets of clinically derived data using known genes of interest. The use cases for TRANSECT are wide and varied, however the application was primarily designed to aid in the extrapolation of *in vitro* observations towards elucidating possible outcomes in a clinical setting using clinical data. Candidate genes or gene set selection is currently not provided; however, there currently exists a number of publicly available and widely used bioinformatics tools that can be employed for this purpose [60] including the R packages WGCNA [61] and GWENA [62].

TRANSECT enables researchers to query complex multi-gene expression scenarios that are not easily achieved through *in vitro* or *in vivo* experimentation and releases any economic burden associated with these types of projects. TRANSECT can also be used in conjunction with laboratory-based experimentation for hypothesis testing and validation or solely for hypothesis generation before further validation. The output from a TRANSECT analysis can also be used for additional investigations, for example, stratum participants can be used in survival analyses and clinical data association can be examined. The case studies here demonstrate simple yet powerful examples of the usefulness and robustness of TRANSECT, describe a set of parameters for general practice and outline the limitations of the application that need to be considered.

With ongoing global efforts to collect and collate large cohort data on a multitude of different disease and non-disease individuals with progressively larger participant numbers and clinical data content, TRANSECT will prove to be an increasingly powerful instrument toward a better understanding of

gene regulation and the perturbations that can occur in a diseased state. Moreover, with ongoing refinement and user-feedback, TRANSECT has the potential to evolve into an integral bioinformatics tool for experimental researchers into the future.

Acknowledgements

This research was supported by using the ARDC Nectar Research Cloud and by Internode. The ARDC Nectar Research Cloud is a collaborative Australian research platform supported by the NCRIS-funded Australian Research Data Commons (ARDC). The compute infrastructure used for testing and benchmarking was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide. We like to thank Dr Stephen Love for his generous and ongoing support towards the online instance of TRANSECT.

Author contributions: J.T.: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Visualization, Supervision, and Project administration. Y.K.: Software, Validation, Visualization, and Supervision. M.M.: Software and Visualization. N.I.W., C.P.B., P.A.G., W.L.K., L.A.S., S.J.C.: Methodology. D.M.L.: Software. A.F.L., S.B., H.S.S.: Supervision. C.H.K., G.J.G., A.W.S.: Methodology and Supervision. All authors reviewed and approved the final version of the manuscript.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

All authors declare no conflict of interest to disclose.

Funding

This work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC) to P.A.G. and G.J.G. (1128479, 1164669), to A.F.L. and W.L.K. (2021560), to G.J.G. (1089167, 1118170), and to H.S.S. (1164601); the National Breast Cancer Foundation to P.A.G. (IIRS18147 and IIRS0098) and to G.J.G. (IN-16-072); the Hospital Research Foundation to P.A.G. (S-13-DTFA-2021 and C-PJ-16-Breast-2020); the Tour de Cure to P.A.G. (RSP-419-2024); the Leukaemia Foundation of Australia and the RAH Research Foundation to H.S.S.; the Cancer Council SA, Beat Cancer and Therapeutic Innovation Australia to H.S.S. and A.F.L.; and from the Australian Cancer Research Foundation (ACRF) to G.J.G., A.F.L., and H.S.S.

P.A.G. is supported by a Principal Cancer Research Fellowship awarded by Cancer Council's Beat Cancer project on behalf of its donors, the state Government through the Department of Health, and the Australian Government through the Medical Research Future Fund.

L.A.S. is supported by a Principal Cancer Research Fellowship (PRF2919) awarded by Cancer Council's Beat Cancer project on behalf of its donors, the State Government through the Department of Health and the Australian Government through the Medical Research Future Fund

S.J.C. is supported by a NHMRC Leadership Investigator grant and research fellowship (GNT1198014) and Tour de Cure senior research grant (RSP-089-2020).

Data availability

The datasets used in this article were derived from sources in the public domain:

1. RECOUNT3, https://rna.recount.bio/
2. GTEx, https://gtexportal.org/home/
3. GDC Data Portal, https://portal.gdc.cancer.gov/

Reproducibility

The TRANSECT commands and parameters used to produce the outputs in the case studies can be retrieved by downloading and installing the standalone command line application from https://github.com/twobeers75/TRANSECT or https://doi.org/10.5281/zenodo.15083663. A wrapper script named "SCA_recreate_manuscript_data.sh" within the "bin/Accessory/" directory of TRANSECT can be modified to suit the user's setup, and run on any compatible operating system to reproduce the case study results.

References

1. Bissell MJ, Weaver VM, Lelievre SA *et al.* Tissue structure, nuclear organization, and gene expression in normal and malignant breast. *Cancer Res* 1999;**59**(7 Suppl):1757–1763s.
2. Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. *Nature* 2000;**405**:827–36. https://doi.org/10.1038/35015701
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63. https://doi.org/10.1038/nrg2484
4. Papatheodorou I, Fonseca NA, Keays M *et al.* Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res* 2018;**46**:D246–51. https://doi.org/10.1093/nar/gkx1158
5. Vivian J, Rao AA, Nothaft FA *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nat Biotechnol* 2017;**35**:314–6. https://doi.org/10.1038/nbt.3772
6. Wang Q, Armenia J, Zhang C *et al.* Unifying cancer and normal RNA sequencing data from different sources. *Sci Data* 2018;**5**:180061. https://doi.org/10.1038/sdata.2018.61
7. Wilks C, Zheng SC, Chen FY *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol* 2021;**22**:323. https://doi.org/10.1186/s13059-021-02533-6
8. Barrett T, Wilhite SE, Ledoux P *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2012;**41**:D991–5.https://doi.org/10.1093/nar/gks1193
9. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;**369**:1318–30. https://doi.org/10.1126/science.aaz1776
10. Lonsdale J, Thomas J, Salvatore M *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;**45**:580–5.
11. The Cancer Genome Atlas Research, Weinstein N, Collisson JN, Mills EA *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
12. Akbani R, Akdemir KC, Aksoy BA *et al.* Genomic classification of cutaneous melanoma. *Cell* 2015;**161**:1681–96.https://doi.org/10.1016/j.cell.2015.05.044
13. Brennan CW, Verhaak RGW, McKenna A *et al.* The somatic genomic landscape of glioblastoma. *Cell* 2013;**155**:462–77.https://doi.org/10.1016/j.cell.2013.09.034

14. Jiang F, Wu C, Wang M *et al.* Identification of novel cell glycolysis related gene signature predicting survival in patients with breast cancer. *Sci Rep* 2021;**11**:3986.https://doi.org/10.1038/s41598-021-83628-9
15. Liu X, Miao Y, Liu C *et al.* Identification of multiple novel susceptibility genes associated with autoimmune thyroid disease. *Front Immunol* 2023;**14**:1161311.https://doi.org/10.3389/fimmu.2023.1161311
16. The Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;**487**:330–7.
17. Verhaak RGW, Hoadley KA, Purdom E *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 2010;**17**:98–110.https://doi.org/10.1016/j.ccr.2009.12.020
18. Wang L, Xue Y, Wang X *et al.* DEPDC1 is a potential therapeutic target in lung adenocarcinoma. *Nano Today* 2024;**56**:102249.https://doi.org/10.1016/j.nantod.2024.102249
19. Smith JC, Sheltzer JM. Genome-wide identification and analysis of prognostic features in human cancers. *Cell Rep* 2022;**38**:110569.
20. Cerami E, Gao J, Dogrusoz U *et al.* The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**:401–4.https://doi.org/10.1158/2159-8290.CD-12-0095
21. Gao J, Aksoy BA, Dogrusoz U *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1. https://doi.org/10.1126/scisignal.2004088
22. Goldman MJ, Craft B, Hastie M *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* 2020;**38**:675–8.
23. Tang Z, Kang B, Li C *et al.* GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res* 2019;**47**:W556–60.
24. Tang Z, Li C, Kang B *et al.* GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res* 2017;**45**:W98–W102.
25. Zengin T, Masud BA, Önal-Süzek T. TCGAnalyzeR: an online pan-cancer tool for integrative visualization of molecular and clinical data of cancer patients for cohort and associated gene discovery. *Cancers* 2024;**16**:345.https://doi.org/10.3390/cancers16020345
26. Deng M, Brägelmann J, Schultze JL *et al.* Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. *BMC Bioinf* 2016;**17**:72.https://doi.org/10.1186/s12859-016-0917-9
27. Colaprico A, Silva TC, Olsen C *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016;**44**:e71. https://doi.org/10.1093/nar/gkv1507
28. Mounir M, Lucchetta M, Silva TC *et al.* New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput Biol* 2019;**15**:e1006701. https://doi.org/10.1371/journal.pcbi.1006701
29. El-Kamand S, Quinn JMW, Sareen H *et al.* CRUX, a platform for visualising, exploring and analysing cancer genome cohort data. *NAR Genom Bioinform* 2024;**6**:lqae003. https://doi.org/10.1093/nargab/lqae003
30. Liao C, Wang X. TCGAplot: an R package for integrative pan-cancer analysis and visualization of TCGA multi-omics data. *BMC Bioinf* 2023;**24**:483. https://doi.org/10.1186/s12859-023-05615-3
31. Bartha A, Gyorffy B. TNMplot.com: a web tool for the comparison of gene expression in normal, tumor and metastatic tissues. *Int J Mol Sci* 2021;**22**:2622. https://doi.org/10.3390/ijms22052622
32. Chen HM, MacDonald JA. Network analysis of TCGA and GTEx gene expression datasets for identification of trait-associated biomarkers in human cancer. *STAR Protoc* 2022;**3**:101168. https://doi.org/10.1016/j.xpro.2022.101168
33. Dvinge H, Git A, Gräf S *et al.* The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* 2013;**497**:378–82. https://doi.org/10.1038/nature12108
34. Kan WL, Dhagat U, Kaufmann KB *et al.* Distinct assemblies of heterodimeric cytokine receptors govern stemness programs in leukemia. *Cancer Discov* 2023;**13**:1922–47. https://doi.org/10.1158/2159-8290.CD-22-1396
35. Liu D, Dredge BK, Bert AG *et al.* ESRP1 controls biogenesis and function of a large abundant multiexon circRNA. *Nucleic Acids Res* 2024;**52**:1387–403. https://doi.org/10.1093/nar/gkad1138
36. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40. https://doi.org/10.1093/bioinformatics/btp616
37. Su S, Law CW, Ah-Cann C *et al.* Glimma: interactive graphics for gene expression analysis. *Bioinformatics* 2017;**33**:2050–2. https://doi.org/10.1093/bioinformatics/btx094
38. Mootha VK, Lindgren CM, Eriksson KF *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;**34**:267–73. https://doi.org/10.1038/ng1180
39. Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50. https://doi.org/10.1073/pnas.0506580102
40. Liao Y, Wang J, Jaehnig EJ *et al.* WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* 2019;**47**:W199–205. https://doi.org/10.1093/nar/gkz401
41. Reese WD. Nginx: the high-performance web server and reverse proxy. *Linux Journal* 2008;**2008**:2.
42. Pihur V, Datta S, Datta S. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinf* 2009;**10**:62. https://doi.org/10.1186/1471-2105-10-62
43. Kolde R, Laur S, Adler P *et al.* Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012;**28**:573–80. https://doi.org/10.1093/bioinformatics/btr709
44. Ellis MJ, Gillette M, Carr SA *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI clinical proteomic tumor analysis consortium. *Cancer Discov* 2013;**3**:1108–12. https://doi.org/10.1158/2159-8290.CD-13-0219
45. Chen C, Wang J, Pan D *et al.* Applications of multi-omics analysis in human diseases. *MedComm* 2023;**4**:e315. https://doi.org/10.1002/mco2.315
46. Yu L, Fernandez S, Brock G. Power analysis for RNA-Seq differential expression studies. *BMC Bioinf* 2017;**18**:234. https://doi.org/10.1186/s12859-017-1648-2
47. Thiery JP. Epithelial-mesenchymal transitions in development and pathologies. *Curr Opin Cell Biol* 2003;**15**:740–6. https://doi.org/10.1016/j.ceb.2003.10.006
48. Aigner K, Dampier B, Descovich L *et al.* The transcription factor ZEB1 (deltaEF1) promotes tumour cell dedifferentiation by repressing master regulators of epithelial polarity. *Oncogene* 2007;**26**:6979–88. https://doi.org/10.1038/sj.onc.1210508
49. Drapela S, Bouchal J, Jolly MK *et al.* ZEB1: a critical regulator of cell plasticity, DNA damage response, and therapy resistance. *Front Mol Biosci* 2020;**7**:36. https://doi.org/10.3389/fmolb.2020.00036
50. Krebs AM, Mitschke J, Lasierra Losada M *et al.* The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. *Nat Cell Biol* 2017;**19**:518–29. https://doi.org/10.1038/ncb3513
51. Testa U, Riccioni R, Militi S *et al.* Elevated expression of IL-3Ralpha in acute myelogenous leukemia is associated with enhanced blast proliferation, increased cellularity, and poor prognosis. *Blood* 2002;**100**:2980–8. https://doi.org/10.1182/blood-2002-03-0852

52. Jordan CT, Upchurch D, Szilvassy SJ *et al.* The interleukin-3 receptor alpha chain is a unique marker for human acute myelogenous leukemia stem cells. *Leukemia* 2000;**14**:1777–84. https://doi.org/10.1038/sj.leu.2401903
53. Tyner JW, Tognon CE, Bottomly D *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* 2018;**562**:526–31. https://doi.org/10.1038/s41586-018-0623-z
54. Aysola K, Desai A, Welch C *et al.* Triple negative breast cancer - an overview. *Hereditary Genet* 2013;**2013**(Suppl 2):001. https://doi.org/10.4172/2161-1041.S2-001
55. Yin L, Duan JJ, Bian XW *et al.* Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Res* 2020;**22**:61. https://doi.org/10.1186/s13058-020-01296-5
56. Cancello G, Maisonneuve P, Rotmensz N *et al.* Progesterone receptor loss identifies Luminal B breast cancer subgroups at higher risk of relapse. *Ann Oncol* 2013;**24**:661–8. https://doi.org/10.1093/annonc/mds430
57. Bracken CP, Gregory PA, Kolesnikoff N *et al.* A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition. *Cancer Res* 2008;**68**:7846–54. https://doi.org/10.1158/0008-5472.CAN-08-1942
58. Gregory PA, Bert AG, Paterson EL *et al.* The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol* 2008;**10**:593–601. https://doi.org/10.1038/ncb1722
59. Wellner U, Schubert J, Burk UC *et al.* The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. *Nat Cell Biol* 2009;**11**:1487–95. https://doi.org/10.1038/ncb1998
60. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat Commun* 2018;**9**:1090. https://doi.org/10.1038/s41467-018-03424-4
61. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf* 2008;**9**:559. https://doi.org/10.1186/1471-2105-9-559
62. Lemoine GG, Scott-Boyer MP, Ambroise B *et al.* GWENA: gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. *BMC Bioinf* 2021;**22**:267. https://doi.org/10.1186/s12859-021-04179-4

Received: October 31, 2024. **Revised:** February 9, 2025. **Editorial Decision:** March 24, 2025. **Accepted:** March 27, 2025