IndepthPathway: an integrated tool for in-depth pathway enrichment analysis based on bulk and single cell sequencing data

Letian Deng[1,2], Sanghoon Lee[1,2], Kai Wang[4], Yuehua Zhu[1,3], Yue Wang[1,2], Maureen A. Sartor[4], Xiaosong Wang[1,2#]

[1]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, 15232, U.S.A.

[2]Department of Pathology, University of Pittsburgh, Pittsburgh, PA, 15232, U.S.A.

[3]School of Medicine, Tsinghua University, Beijing, China, 100084.

[4]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, 48109, U.S.A.

[#]**Corresponding Author**:

Xiaosong Wang, M.D., Ph.D.
Associate Professor of Pathology
UPMC Hillman Cancer Center
5117 Centre Avenue,
Pittsburgh, PA 15232
Phone: 412-623-1587
Email: xiaosongw@pitt.edu

ABSTRACT

Single-cell sequencing (SCS) enables exploring the pathways and processes of cells and cell populations, however, there is a paucity of pathway enrichment methods designed to tolerate the high noise and low gene coverage of this technology. When gene expression data are noisy and signals are sparse, testing pathway enrichment based on the genes measured may not yield statistically significant results which is particularly problematic when detecting the pathways enriched in less abundant cells that are vulnerable to disturbances. In this project, we developed a Weighted Concept Signature Enrichment Analysis (WCSEA) algorithm specialized for pathway enrichment analysis from single cell transcriptomics (scRNA-seq). WCSEA took a broader approach for assessing the functional relations of pathway gene sets to differentially expressed genes, and leverage the cumulative signature of molecular concepts characteristic of the highly differentially expressed genes, which we termed as the universal concept signature, to tolerate the high noise and low coverage of this technology. We then incorporated WCSEA into a R package called "IndepthPathway" for biologists to broadly leverage this method for pathway analysis based on bulk and single cell sequencing data. Through simulating the technical variability and dropouts in gene expression characteristic of scRNA-seq, WCSEA yielded overall low deviations in pathway enrichment results. Leveraging its unique strength, IndepthPathway will enhance the application of bulk and single cell sequencing technologies to explore the cellular pathway mechanisms at more precise resolution.

BACKGROUND

The identification of causal pathways underlying physiological process or disease initiation, progression, or therapeutic resistance based on proteomic, genomic, and transcriptomic datasets are major challenges of genomics research, particularly for single cell sequencing-based studies. Pathway Enrichment (PE) methods can be classified chronologically into three generations [1]: 1) over-representation analysis (ORA) such as Fisher's exact test, 2) functional class scoring such as Gene Set Enrichment Analysis (GSEA)[2], and 3) network-topology based such as network enrichment analysis (NEA) [3]. Based on their dependency on gene weights (i.e., levels of differential expressions), PE methods can be classified into unweighted or weighted methods. The unweighted methods (i.e., ORA or NEA) are commonly used for analysis of genetic traits extracted from genomic analysis. For RNA-seq analysis, these methods can be used to identify the pathways within the user-defined DE genes based on thresholding. The weighted methods (i.e., GSEA) are commonly used for analysis of continuous expression data. Incorporating the degrees of differential gene expressions allows for detection of different magnitudes of pathway alterations.

It is notable that PE methods have distinct functionality from pathway activity inference, pathway and gene set over-dispersion analysis, regulon activity inference, or functional gene set inference. PE methods aim to test the enrichment of a wide array of pathways in differentially expressed genes thus are commonly used for pathway discovery. Whereas pathway activity inference tools assess the activity of handful pathways in each sample or single cell. Popular pathway activity inference tools such as PROGENy require consensus signatures obtained from publicly available perturbation experiments, which further limit the scope of the pathways to be assessed [4]. Pathway and gene set over-dispersion analysis (i.e., PAGODA) is used to cluster cells into transcriptional subpopulations [5]. Regulon inference (i.e., AUCel) are designed for testing the activity of the large footprint regulons for gene regulatory network reconstruction [6], thus are not dedicated for pathways that have much smaller footprints. Functional gene set inference depends on PE methods for testing gene set enrichments based on matrix factorizations, which we will discuss further in the results section.

While there are a handful pathway activity inference methods for single cell sequencing (SCS) data, to date

there is a paucity of PE methods tailored for SCS data despite its popularity for pathway discovery in biomedical research. Compared to bulk-sequencing, SCS enables exploring pathways and processes of individual cells. One of the most important analyses to achieve this is to identify the pathways that exhibit significant differences among different cell states or cell types. Due to low genome coverage and high amplification bias, bioinformatics tools developed for bulk sequencing may not work well with SCS data [7, 8]. The conspicuous challenges of SCS data analysis include dissecting single cell expression variability, identifying rare cell populations, and understanding spatio-temporal transitions [9, 10]. In the case of PE analysis, the problem is less self-evident due to lack of golden standard and explicit modeling methods to assess the impact of SCS technical variability on the pathway interpretations. Further, it is even more challenging to detect the pathways enriched in less abundant cells which is more vulnerable to disturbances.

In our previous study, we sought to improve the reproducibility of PE analysis based on experimentally defined gene sets through leveraging the vast variety of knowledge-based gene sets including gene ontologies, pathways, interactions, and domains etc., to help inform complex functional relations, which we collectively termed as "molecular concepts". Based on this idea, we developed a concept signature enrichment analysis (CSEA) for deep assessing the functional relations between pathway gene sets and a target gene list using a compendia of molecular concepts [11]. Here "a target gene list" is defined as a gene list of interest cataloged by comparative genomic analysis, and a "pathway gene set" is defined as a set of genes known to function in a certain pathway (**Fig.1a**). The concept signatures of a target gene list are collectively termed as a Universal Concept Signature (uniConSig). CSEA deeply interprets the function of a target gene list via computing their overrepresentations in a wide array of molecular concepts, which were then used as weights to compute a cumulative genome-wide uniConSig score that represent the functional relevance of human genes underlying this target gene list (**Fig. 1d**). Then the UniConSig-sorted genome will be used for testing pathway enrichment through weighted Kolmogorov–Smirnov (WKS) tests (**Fig. 1e**). If a pathway gene set is functionally like the target gene list, the uniConSig scores will be high for that pathway genes, which will result in a high normalized enrichment score (NES) of that pathway. CSEA represents a new class of PE method that is distinct from all earlier generations with much improved depth in assessing gene set functional relations. A major advantage of

CSEA over the approach based on interactome network topology is that CSEA is grounded on the framework of the vast knowledge databases, and thus can comprehensively assess most, if not all, functional aspects when computing the functional relations [11].

In this study, we speculate that CSEA could be particularly suited for addressing the characteristics of SCS data. The rationale is that CSEA may tolerate the high noise and stochasticity of SCS data via leveraging the stable concept signatures of the target gene list to improve the reproducibility and scientific rigor of PE results. While the expressions of individual genes may be noisy, coordinated upregulation of genes within a gene set could provide a more stable signature. A similar principle has been leveraged by PAGODA to cluster cells into transcriptional subpopulations [5]. In this project, we sought to develop a purpose-built PE package called "IndepthPathway" for deep pathway enrichment analysis from bulk and single cell sequencing data that took a broader approach for assessing gene set relations and leverage the universal concept signature of the target gene list to tolerate the high noise and low gene coverage of this technology. Our results show that IndepthPathway achieved better reproducibility under technical noise simulations of single cell sequencing data compared to other popular PE methods. Empowered by its strength in deep functional interpretation, IndepthPathway will have broad applications on inferring the pathway alterations from bulk and SCS data.

METHODS

Compiling molecular concept databases and calculating weighted uniConsig scores in a target gene list

The uniConSig score calculation and deep functional interpretation for a target gene list rely on the precompiled human molecular concept databases. To generate a comprehensive knowledge base, we compiled 45,522 molecular concepts from the Molecular Signatures Database (MSigDB) C2CP pathways, or hallmark (v.7.0) [12], NCBI EntrezGene interactome [13]database, and conserved domain database[13]. To get a target gene list, we used single-cell transcriptomic data of hematopoietic stem cells (HSCs) [Yang et al. 2017 NAR] and ran Single Cell Differential Expression (SCDE) that applies Bayesian approach for single-cell DE analysis [14].

The identified up- or downregulated gene lists in quiescent HSCs compared to active HSCs were used as the target gene list for CSEA analysis, and the statistical values are used as weights for WCSEA analysis.

The basic algorithm to calculate uniConSig score for an experimental gene list in CSEA analysis is explained in our previous study[15]. In WCSEA analysis, the calculation of uniConSig scores from a weighted gene list is based on the enrichment of each molecular concepts associated with Gene *x* in the weighted gene list calculated using weighted K-S test. The resulting normalized enrichment score (NES) score for each concept is used as a weight for calculating the genome-wide uniConSig score.

i) $NES = {ES}/{mean(ES_{random})}$

To reduce the inflation effect on the uniConSig scores by the degree of overlaps between the molecular concepts, we penalized the scores based on the similarity matrix of molecular concepts. We generated the similarity matrix by computing the Otsuka-Ochiai coefficient between molecular concepts. Compared to the Jaccard Index used in our previous study [11], the Otsuka-Ochiai coefficient can better assess the similarity between a small molecular concept and a large molecular concept. We then introduced a penalization factor (ε) calculated for each molecular concept *i* based on the cumulated similarities of concept *i* with other molecular concepts associated with $Gene_x$. The Effective Weight (EW) is then calculated as:

ii) $EW_i = \frac{\omega_i}{\varepsilon_i}$

Finally, the uniConSig score of the given $Gene_x$ is calculated as following:

iii) $uniConSig_{|Gene_x} = \frac{\sum_{i=1}^{n} EW_i}{EFN_i} = \frac{\sum_{i=1}^{n} {\omega_i}/{\varepsilon_i}}{\sqrt{\sum_{i=1}^{n} {1}/{\varepsilon_i}}}$

The cumulative genome-wide uniConSig scores reflect the functional relevance of the human genes underlying the highly weighted target genes. The normalized enrichment scores of pathway gene sets were then calculated by the weighted step-up of the random walk K-S test using uniConSig sorted human gene list to assess the functional associations of these pathways with the original weighted gene list.

Disambiguation of the crosstalk effects between gene signatures

Many gene signatures or pathways have varying degrees of overlap each other due to shared gene families. We performed disambiguation steps to correct the crosstalk effects between the enriched pathways of the WCSEA output. The significance of the disambiguation process was determined by p-value < 0.01. We then assessed the functional association between the WCSEA output pathways using our CSEA method as we described in our previous study[15], and scaled the WCSEA scores in the range of -1 and +1. The pathway network was visualized using the igraph R package (ver. 1.2.4.2).

Simulations of multiple faceted variability of scRNA-seq and dropouts in gene expression.

To benchmark analyzing differential expression and identifying enriched pathways in scRNA-seq data, we simulated scRNA-seq data using **Sy**nthetic model of **m**ultiple variability factors for **Sim**ulation (SymSim)[16]. We simulated technical variations by noise intrinsic to the process of transcription, extrinsic indication of different cell states, and low sensitivity in measurement. In addition, we simulated dropouts in gene expressions by randomly deleting 20% or 50% of the expressed genes. We then compared the performance of CSEA, WCSEA and other PE discovery tools, such as ORA[17], NEA[3], and GSEA[2], based on the variations in pathway ranks in the simulated scRNA-seq data with or without the technical variations.

RESULTS

Development of a weighted CSEA method to enhance its utility and reproducibility in detecting pathway alterations from single cell transcriptomics data.

Next, we developed a Weighted CSEA (WCSEA) method to enhance its functionality and reproducibility in analysis of single cell transcriptomics. We propose WCSEA to achieve deep functional assessment of the weighted gene list (based on differential expression statistics) by performing Weighted KS tests to quantitate the enrichment of each molecular concept associated with a given Gene x in the highly weighted genes. The resulting NES score for each concept is used as weight for calculating the uniConSig score of Gene x (**Fig. 1b**).

These weights are then penalized by the degree of overlaps between the Gene x concepts, computed based on their Ochiai's similarity indexes (**Fig. 1c**). The resulting uniConSig scores reflect the functional relevance of human genes underlying the highly weighted genes in the list (i.e., highly overexpressed genes) (**Fig. 1d**). Then pathway enrichment analysis will be performed on this genome-wide weighted gene list based on uniConSig scores (**Fig. 1e**). WCSEA leverages the stable concept signatures of highly weighted genes to create a buffer zone for tolerating the high noise and stochasticity of scRNA-seq data.

Development of IndepthPathway package for biologists to broadly leverage this technology to enhance pathway discovery from experimentally defined gene sets and SCS data.

Next, we implemented CSEA and WCSEA in a R package called "IndepthPathway" to widely disseminate this technology. This purpose-built tool allows users to provide a weighted or unweighted defined by comparative analysis of bulk or single cell sequencing data for WCSEA or CSEA analysis respectively. Users can perform differential expression (DE) analysis between the comparing sample groups or cell groups using the bulk or single cell sequencing DE methods, such as DESeq2 (Love et al. 2014 Genome Biology) and Limma (Ritchie et al. 2015 NAR), and the resulting statistical values can be used to generate a weighted gene list for WCSEA analysis. Users can also use their pre-defined weighted gene list from DE analysis or target gene list from comparative analysis of scRNA-seq or scDNA-seq. This will allow user to implement their preferred DE methods to identify differentially expressed genes. Over- or under-expressed pathways should be analyzed separately. Users can also implement unsupervised Matrix factorizations (MF) techniques to reveal the low-dimensional components (LDC) for functional pathway inference [18]. MF is not a PE method, but instead, it relies on PE analysis to explore the pathways enriched in an LDC using its amplitude matrix which is known as functional pathway inference. Users can select popular MF methods such as UMAP [19] or t-SNE[20], and perform WCSEA analysis for selected LDC to identify the key pathways underlying that component.

For deep functional interpretations, we have precompiled the human concept datasets from the molecular signatures database (MSigdB) [2], NCBI EntrezGene interactome database [13], and conserved domain database [13]. Users can use our pre-compiled human concept datasets or provide their own concept dataset. We expect inclusion of the more available concepts will substantially "thicken" the "buffering zone" via increasing the

number of signature concepts, thus will further improve the stability of WCSEA under technical noise. For pathway enrichment analysis, the pathway gene sets for curated canonical pathways and hallmark pathways have been precompiled from MSigDb [2] and provided to users. Users may select the pre-compiled pathway gene sets or provide their own pathway gene sets in specific format. In addition, we have integrated downstream analysis modules to improve its usefulness. Following pathway enrichment analysis, we have implemented a module for users to disambiguate and identify independent pathway modules via correcting the crosstalk effects between pathways, as previously reported [21]. The pathway enrichment results include the pathways ranked by NES scores, together with the p-values and FDR q-values. We also implemented a module to allow users to generate a pathway network showing the functional relations between significant pathways which are computed based on the CSEA method. The IndepthPathway package has been released to Github and we have provided a detailed manual with examples.

IndepthPathway yielded reproducible pathway results under simulations of multiple faceted variability of scRNA-seq and dropouts in gene expression.

Next, we applied the IndepthPathway tool to examine the pathways enriched in active vs quiescent populations of hematopoietic stem cells (HSC) based on a publishes single cell sequencing dataset [22]. Differential expressions (DE) are analyzed using the Single Cell Differential Expression (SCDE) package that applies Bayesian approach for single-cell DE analysis [14]. The resulting statistical values are used as weights for WCSEA analysis. Following WCSEA analysis, the top 30 upregulated and top 30 downregulated pathways are disambiguated based on the method discussed in the **Methods** section, and plotted into a network space based on the functional associations of these pathways computed through CSEA analysis (**Fig. 2a**). Our results show that the WCSEA methods can comprehensively detect the pathways known to characterize active HSC: 1) Upregulation of E2F targets, DNA replication, and G2M checkpoint pathways, indicating increased cell cycle activities [22, 23]; 2) Upregulation of MYC targets known to control HSC self-renewal and differentiation [24]; 3) Upregulation of spliceosome, which is consistent with the increased total RNA amount in active HSCs [25]; 4) Downregulation of pathway required for the quiescent HSCs to respond to interferon α/β stimulation [26, 27]; 5) upregulation of various DNA repair pathways. It is reported that DNA damages accumulated during

quiescent state will be repaired upon entry into active state [28]. In contrast, the commonly used Gene Set Enrichment Analysis (GSEA) [2]. appears to enrich the most significant pathways but lack the depth to interpret complex pathway alternations (**Fig. 2b**).

Next, we sought to examine the performance of IndepthPathway under the simulated technical variability of scRNA-seq data. This can be achieved through "SymSim" that explicitly models the multiple faceted variability of single cell RNA-seq (scRNA-seq) experiments [16]. In addition, we simulated dropouts in gene expressions by randomly deleting 20% or 50% of expressed genes and assessed the changes in pathway ranks from the original results. We then tested the performance of IndepthPathway on pathway discovery from scRNA-seq with or without simulated technical variabilities. We compared IndepthPathway with the representative PE tools of each generations including ORA, NEA, and GSEA [2, 3, 17] (**Fig. 3**). As discussed above, PE methods have distinct functionality from pathway activity inference, regulon activity inference, or functional gene set inference, all of which are not considered as classic PE methods, thus are excluded from our following comparisons. Pathways are usually interpreted based on their enrichment ranks, thus the rank variations for the top-30 pathways identified by these methods compared to the original PE results are used for benchmarking.

As expected, the simulations resulted in high variations in the pathway ranks from representative PE methods (**Fig. 3**). This suggests that PE tools developed for bulk RNA-seq may yield inconsistent results under technical variability of scRNA-seq. When gene expression data are noisy and signals are sparse, testing pathway enrichment based on the genes measured may not yield statistically significant results. Tolerating dropouts in gene expression will require deep interpreting the functions of the differently expressed genes in the comparing pathways, instead of simply mapping the gene names in the pathways. Indeed, our result showed that CSEA and WCSEA produced overall lowest variations in pathway ranks with highest density around zero compared to other PE methods under simulations of both technical noises and gene expression dropouts, and WCSEA showed further improved reproducibility than CSEA under 50% expression dropouts. Together, this result supports the advantage of IndepthPathway on reproducibility under high technical noise and dropouts in gene expression.

DISCUSSION

Bioinformatics tools for bulk-sequencing may not work well with SCS data due to the low genome coverage and high amplification bias. In case of pathway enrichment (PE) analysis, the problem is less self-evident and suitability of existing PE methods on SCS data are hard to assess due to the lack of golden standard. When gene expression data are noisy and signals are sparse, testing pathway enrichment (PE) based on the genes measured may not yield statistically significant results. Indeed, our simulation analyses revealed that popular PE methods yielded highly inconsistent results under the simulated multiple faceted variability of single cell RNA-seq. Here we introduce a weighted CSEA analysis for deep assessing the pathways enriched in a weighted gene list such as differentially expressions through exploiting comprehensive sets of molecular concepts (i.e. gene ontologies, pathways, interactions, and domains, etc.) to compute the concept signatures shared by the highly weighted genes, then testing the pathway enrichments over all genes sorted by their functional similarity to the highly weighted genes based on their associations with these concept signatures. WCSEA takes a broader approach for assessing gene set relations and leverages universal concept signature to tolerate the stochasticity of the data. The utilization of the universal concept signature for PE analysis will be particularly useful for scRNA-seq data as technical noise and dropouts in gene expressions are less likely to affect the universal concept signature. This creates a “buffer zone” to tolerate the technical noise of SCS data and improves reproducibility through leveraging the stable concept signatures of the highly weighted genes for PE analysis.

Through analysis of a single cell transcriptomic (scRNA-seq) dataset comparing active and quiescent hematopoietic stem cells (HSCs), we demonstrate the outstanding performance of WCSEA on detecting complex pathway alterations characteristic of quiescent HSCs. Further, under simulated technical variability of scRNA-seq, WCSEA yielded lowest deviations in PE results compared to other commonly used PE methods, suggesting that WCSEA can endure the stochasticity of data, and thus improve the reproducibility and scientific rigor of the PE analysis.

Merging a wide array of concept datasets carries inherent concept redundancy, as different gene set sources follow different nomenclatures, and different categories of concepts possess different levels of inherent redundancy. Our uniConSig algorithm possesses a unique ability to compute the redundancy within molecular concept data and sieve only the unique parts of signature concepts for further calculation via an innovative mathematical scheme to penalize the partially overlapping concepts (**Fig. 1b**). This allows for comprehensively assessing the function relations through merging a wide array of molecular concept datasets. More important, leveraging the stability of the universal concept signature (the cumulative gene set signatures) instead of individual genes to rank the pathways will allow this method to tolerate the high noise and intrinsic stochasticity of individual gene expressions in the scRNA-seq data. Furthermore, the new design of the Weighted CSEA (WCSEA) algorithm will empower PE analysis based on a weighted gene list to provide more power to detect different magnitudes of pathway alterations reflected by the levels of differential expressions and better tolerance of the technical noise of scRNA-seq. Purpose-designed PE methods such as IndepthPathway that can better address the characteristics of scRNA-seq data will be of utmost importance to ensure the accuracy and reproducibility of the PE results. The integration of the up and downstream analysis modules such as differential gene expression, pathway disambiguation, functional associations of pathways, and heatmap and network visualization of the top enriched pathways will greatly enhance the functionalities of IndepthPathway.

Code availability

The R modules for iGenSig modeling are available through: https://github.com/wangxlab/indepthPathway

Acknowledgements

This study was supported by P30 CA047904 CCSG Dev Funds (X-S. W.), NIH grant 1R01CA181368 (X-S.W.), and 1R01CA183976 (X-S.W.). This study is also in part supported by 1R21CA237964 (X-S.W.), P50 CA097190 Head and Neck SPORE DRP (X-S. W.), PA breast cancer coalition (X-S.W.), Commonwealth of PA Tobacco Phase 15 Formula Fund (X-S. W.), the Shear Family Foundation, and the Hillman Foundation.

This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. We especially thank Fangping Mu for his kind assistance with the research computing. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE) supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

Competing Interests Statement

The authors declare no competing interests.

REFERENCES

1. Garcia-Campos MA, Espinal-Enriquez J, Hernandez-Lemus E. Pathway Analysis: State of the Art, Front Physiol 2015;6:383.
2. 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 U S A 2005;102:15545-15550.
3. Jeggari A, Alekseenko Z, Petrov I et al. EviNet: a web platform for network enrichment analysis with flexible definition of gene sets, Nucleic Acids Res 2018;46:W163-W170.
4. Schubert M, Klinger B, Klunemann M et al. Perturbation-response genes reveal signaling footprints in cancer gene expression, Nat Commun 2018;9:20.
5. Fan J, Salathia N, Liu R et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis, Nat Methods 2016;13:241-244.
6. Aibar S, Gonzalez-Blas CB, Moerman T et al. SCENIC: single-cell regulatory network inference and clustering, Nat Methods 2017;14:1083-1086.
7. Poirion OB, Zhu X, Ching T et al. Single-Cell Transcriptomics Bioinformatics and Computational Challenges, Front Genet 2016;7:163.
8. Ning L, Liu G, Li G et al. Current challenges in the bioinformatics of single cell genomics, Front Oncol 2014;4:7.
9. Yuan GC, Cai L, Elowitz M et al. Challenges and emerging directions in single-cell analysis, Genome Biol 2017;18:84.
10. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics, Nat Rev Genet 2015;16:133-145.
11. Chi X, Sartor MA, Lee S et al. Universal concept signature analysis: genome-wide quantification of new biological and pathological functions of genes and pathways, Brief Bioinform 2019.
12. Zhao D, Lu X, Wang G et al. Synthetic essentiality of chromatin remodelling factor CHD1 in PTEN-deficient cancer, Nature 2017;542:484-488.
13. Brown GR, Hem V, Katz KS et al. Gene: a gene-centered information resource at NCBI, Nucleic Acids Res 2015;43:D36-42.
14. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis, Nat Methods 2014;11:740-742.
15. Chi X, Sartor MA, Lee S et al. Universal concept signature analysis: genome-wide quantification of new biological and pathological functions of genes and pathways, Brief Bioinform 2020;21:1717-1732.

16. Zhang X, Xu C, Yosef N. Simulating multiple faceted variability in single cell RNA sequencing, Nat Commun 2019;10:2611.
17. Yu G, Wang LG, Han Y et al. clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS 2012;16:284-287.
18. Stein-O'Brien GL, Arora R, Culhane AC et al. Enter the Matrix: Factorization Uncovers Knowledge from Omics, Trends Genet 2018;34:790-805.
19. Becht E, McInnes L, Healy J et al. Dimensionality reduction for visualizing single-cell data using UMAP, Nat Biotechnol 2018.
20. Linderman GC, Rachh M, Hoskins JG et al. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data, Nat Methods 2019;16:243-245.
21. Donato M, Xu Z, Tomoiaga A et al. Analysis and correction of crosstalk effects in pathway analysis, Genome Res 2013;23:1885-1893.
22. Yang J, Tanaka Y, Seay M et al. Single cell transcriptomics reveals unanticipated features of early hematopoietic precursors, Nucleic Acids Res 2017;45:1281-1296.
23. Kwon JS, Everetts NJ, Wang X et al. Controlling Depth of Cellular Quiescence by an Rb-E2F Network Switch, Cell Rep 2017;20:3223-3235.
24. Wilson A, Murphy MJ, Oskarsson T et al. c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation, Genes Dev 2004;18:2747-2763.
25. Huttmann A, Liu SL, Boyd AW et al. Functional heterogeneity within rhodamine123(lo) Hoechst33342(lo/sp) primitive hemopoietic stem cells revealed by pyronin Y, Exp Hematol 2001;29:1109-1116.
26. Robb L. Cytokine receptors and hematopoietic differentiation, Oncogene 2007;26:6715-6723.
27. Essers MA, Offner S, Blanco-Bose WE et al. IFNalpha activates dormant haematopoietic stem cells in vivo, Nature 2009;458:904-908.
28. Beerman I, Seita J, Inlay MA et al. Quiescent hematopoietic stem cells accumulate DNA damage during aging that is repaired upon entry into cell cycle, Cell Stem Cell 2014;15:37-50.

FIGURE LEGENDS

Figure 1. Schematic of the unweighted and weighted CSEA methods for PE analysis based on bulk or scRNA-seq data. a) For CSEA analysis, a target gene list will be generated from genomic datasets (i.e., differentially expressed or mutated genes). CSEA achieves deep functional assessment of the target gene list by performing similarity tests to quantitate the enrichment of each molecular concept associated with a given Gene x in the target gene list. The resulting similarity index for each concept will be used for calculating uniConSig scores. **b**) For WCSEA analysis, the weighted gene list can be created based on differential expression statistics comparing different cell groups, or from the amplitude matrix of a low dimensional component identified by matrix factorization methods. UniConSig analysis achieves deep functional assessment of the weighted gene list (based on differential expression statistics) by performing Weighted KS tests to quantitate the enrichment of each molecular concept associated with a given Gene x in the highly weighted genes. The resulting NES score for each concept is used as weight for calculating the uniConSig score of Gene x, which will be normalized to the range of [0,1]. **c**) These weights generated by similarity tests or WKS tests will be penalized by the degree of overlaps between the Gene x concepts, computed using their Ochiai's similarity indexes based on molecular concept database. **d**) The uniConSig scores will be computed using the indicated formula to reflect the functional relevance of human genes underlying the target gene list (for CSEA analysis) or highly weighted genes (for WCSEA analysis). **e**) The enrichment of pathways in the pathway datasets (i.e., MSigdb C2CP canonical pathways, or hallmark pathways) in this uniConSig-sorted gene list will then be assessed using weighted K-S tests. The resulting normalized enrichment scores can be used as a quantitative measure of the functional associations between these pathways underlying the target gene list (CSEA) or weighted gene list (WCSEA).

Figure 2. Pathways characteristic of active HSCs revealed by WCSEA or GSEA based on a single cell sequencing dataset. Pathway association networks showing the top upregulated and downregulated pathways in active HSCs compared to quiescent HSCs detected by WCSEA (**a**) or GSEA (**b**) analyiss. Top 30

upregulated and top 30 downregulated pathways based on WCSEA or GSEA analysis are disambiguated, and the resulting non-ambiguous pathways are plotted in the network. The node size correlates with the levels of enrichments of the respective pathways, and the edge thickness correlates with the levels of associations between the pathways computed based on CSEA method.

Figure 3. The reproducibility of WCSEA under simulations of scRNA-seq technical variability and dropouts in gene expression. Noise simulations are performed using SymSim with two sets of parameters representing different noise levels. The dropouts are simulated via randomly deleting 20% or 50% expressed genes. Five permutations are done for each simulation. Read counts for 177 HSC cells (77 quiescent HSCs and 100 active HSCs) are used as true counts before simulations. SCDE is used for DE analysis, and the results are used for PE analysis. The z scores of signed p-values are used for weighted methods (WCSEA, GSEA). Top 300 downregulated genes in quiescent HSCs are used for unweighted methods (CSEA, NEA, ORA). Variations in pathway ranks for top-30 downregulated pathways in quiescent HSCs are used for benchmarking with negative variations indicating reduced significance of the pathways.

Figure 1

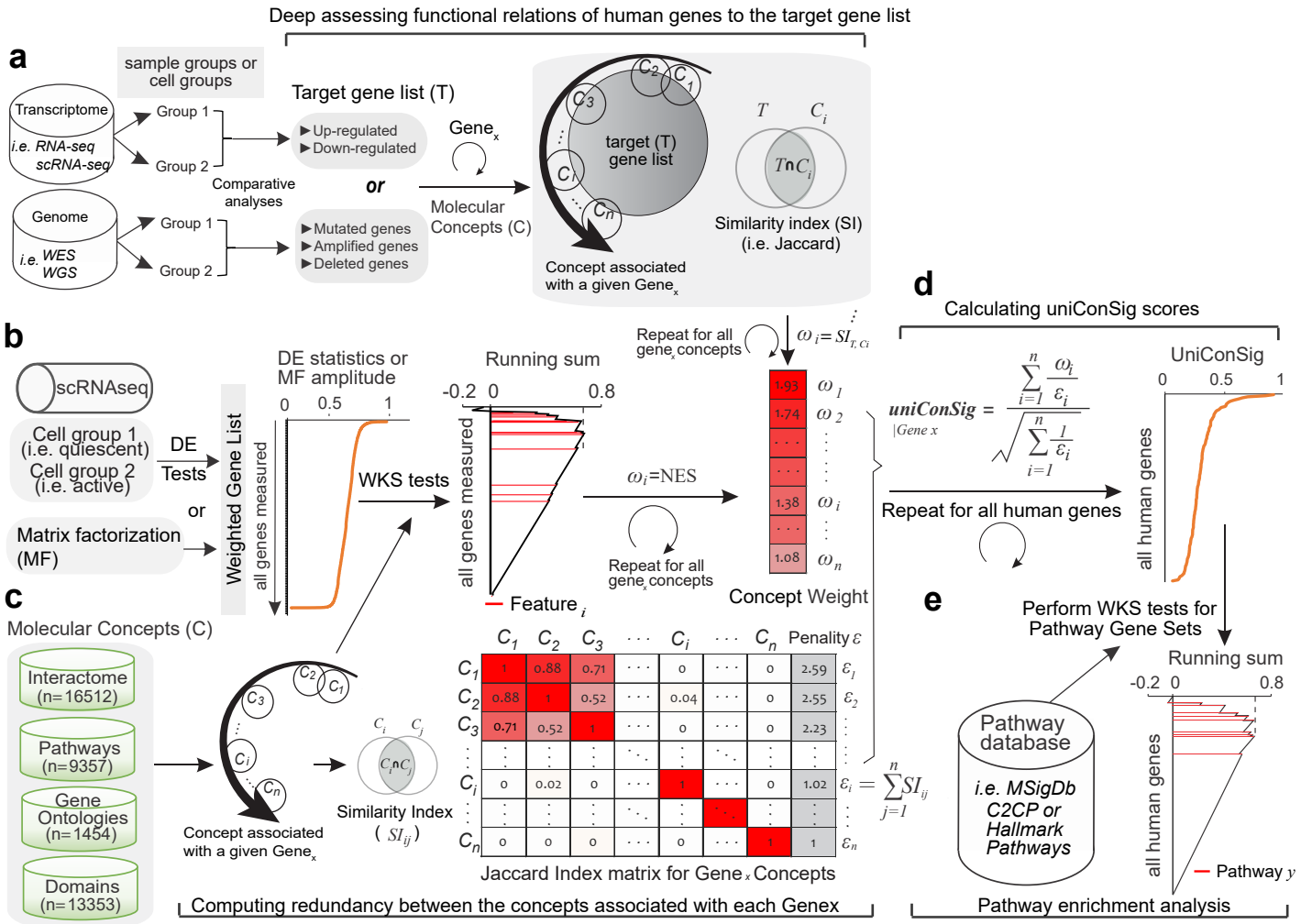

Deep assessing functional relations of human genes to the target gene list
a
sample groups or cell groups
Transcriptome
i.e. RNA-seq scRNA-seq
Group 1
Group 2
Comparative analyses
Genome
i.e. WES WGS
Group 1
Group 2
Target gene list (T)
▶Up-regulated
▶Down-regulated
or
▶Mutated genes
▶Amplified genes
▶Deleted genes
Gene$_x$
Molecular Concepts (C)
target (T) gene list
Concept associated with a given Gene$_x$
Similarity index (SI) (i.e. Jaccard)
b
scRNAseq
Cell group 1 (i.e. quiescent)
Cell group 2 (i.e. active)
DE Tests
or
Matrix factorization (MF)
Weighted Gene List
DE statistics or MF amplitude
all genes measured
WKS tests
Running sum
all genes measured
Feature $_i$
Repeat for all gene$_x$ concepts
ω_i=NES
Repeat for all gene$_x$ concepts
Concept Weight
c
Molecular Concepts (C)
Interactome (n=16512)
Pathways (n=9357)
Gene Ontologies (n=1454)
Domains (n=13353)
Concept associated with a given Gene$_x$
Similarity Index (SI_{ij})
Penality ε
Jaccard Index matrix for Gene$_x$ Concepts
Computing redundancy between the concepts associated with each Genex
d
Calculating uniConSig scores
uniConSig
UniConSig
all human genes
Repeat for all human genes
e
Perform WKS tests for Pathway Gene Sets
Pathway database
i.e. MSigDb C2CP or Hallmark Pathways
Running sum
all human genes
Pathway y
Pathway enrichment analysis

Figure 2

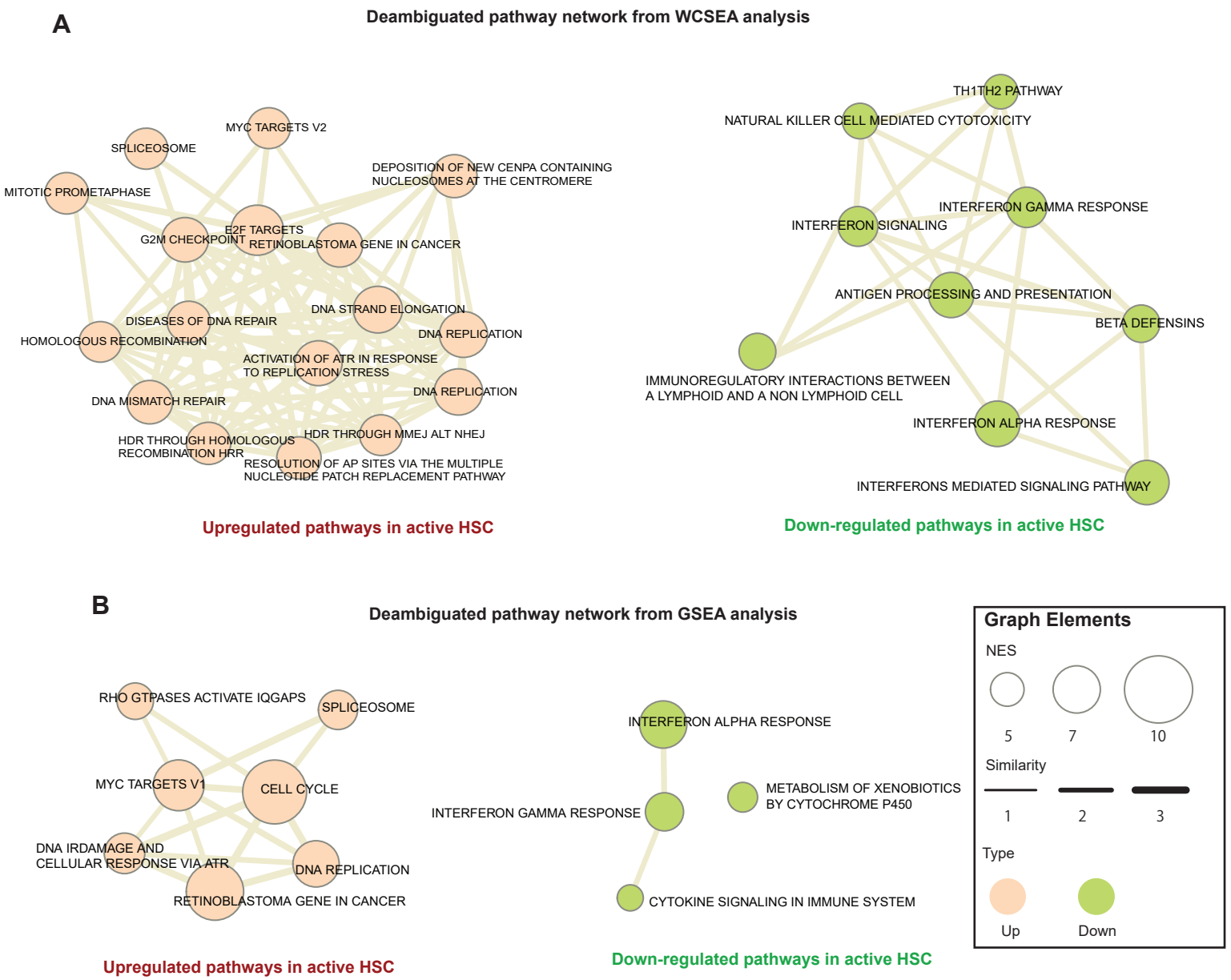
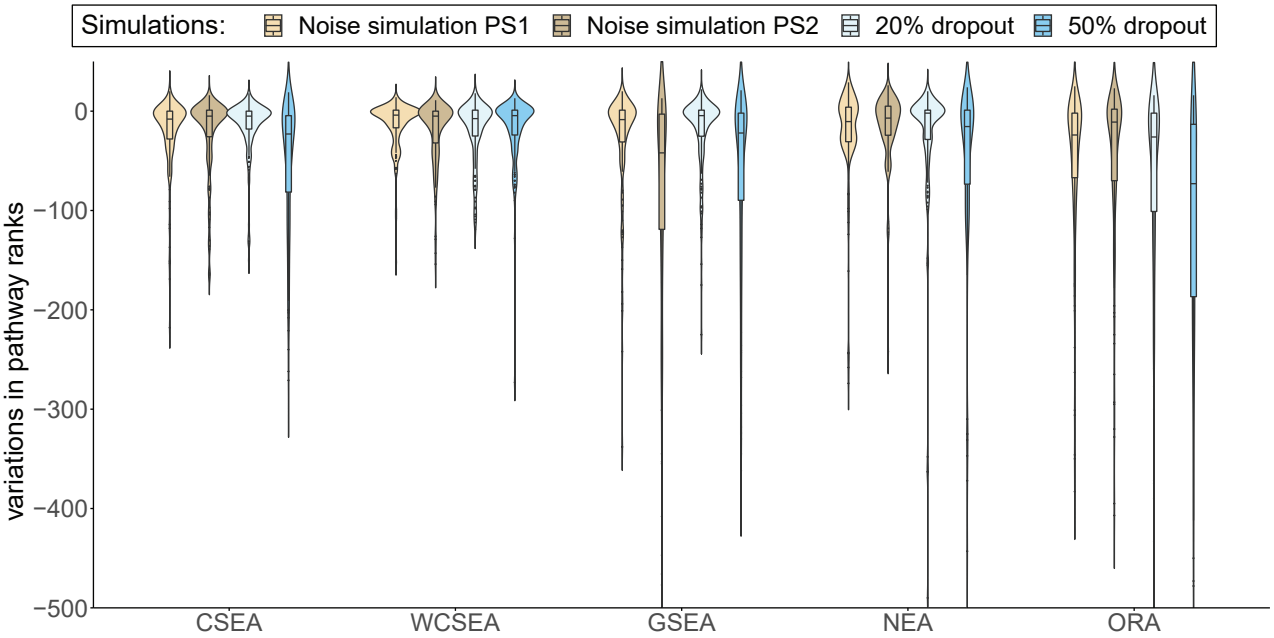

A
Deambiguated pathway network from WCSEA analysis
MYC TARGETS V2
SPLICEOSOME
MITOTIC PROMETAPHASE
DEPOSITION OF NEW CENPA CONTAINING NUCLEOSOMES AT THE CENTROMERE
E2F TARGETS
G2M CHECKPOINT
RETINOBLASTOMA GENE IN CANCER
DNA STRAND ELONGATION
DISEASES OF DNA REPAIR
HOMOLOGOUS RECOMBINATION
DNA REPLICATION
ACTIVATION OF ATR IN RESPONSE TO REPLICATION STRESS
DNA REPLICATION
DNA MISMATCH REPAIR
HDR THROUGH HOMOLOGOUS RECOMBINATION HRR
HDR THROUGH MMEJ ALT NHEJ
RESOLUTION OF AP SITES VIA THE MULTIPLE NUCLEOTIDE PATCH REPLACEMENT PATHWAY
Upregulated pathways in active HSC
TH1TH2 PATHWAY
NATURAL KILLER CELL MEDIATED CYTOTOXICITY
INTERFERON GAMMA RESPONSE
INTERFERON SIGNALING
ANTIGEN PROCESSING AND PRESENTATION
BETA DEFENSINS
IMMUNOREGULATORY INTERACTIONS BETWEEN A LYMPHOID AND A NON LYMPHOID CELL
INTERFERON ALPHA RESPONSE
INTERFERONS MEDIATED SIGNALING PATHWAY
Down-regulated pathways in active HSC
B
Deambiguated pathway network from GSEA analysis
RHO GTPASES ACTIVATE IQGAPS
SPLICEOSOME
MYC TARGETS V1
CELL CYCLE
DNA IRDAMAGE AND CELLULAR RESPONSE VIA ATR
DNA REPLICATION
RETINOBLASTOMA GENE IN CANCER
Upregulated pathways in active HSC
INTERFERON ALPHA RESPONSE
METABOLISM OF XENOBIOTICS BY CYTOCHROME P450
INTERFERON GAMMA RESPONSE
CYTOKINE SIGNALING IN IMMUNE SYSTEM
Down-regulated pathways in active HSC
Graph Elements
NES
5
7
10
Similarity
1
2
3
Type
Up
Down

Figure 3


Simulations:
Noise simulation PS1
Noise simulation PS2
20% dropout
50% dropout
variations in pathway ranks
0
−100
−200
−300
−400
−500
CSEA
WCSEA
GSEA
NEA
ORA