scFEA: A graph neural network model to estimate cell-wise metabolic flux using single cell RNA-seq data

Norah Alghamdi[1+], Wennan Chang[1,2+], Pengtao Dang[1,2], Xiaoyu Lu[1], Changlin Wan[1,2], Zhi Huang[1,2], Jiashi Wang[1], Melissa Fishel[3], Sha Cao[1, 4]*, Chi Zhang[1,2]*

[1]Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics, [3]Department of Pediatrics, [4]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, 46202, USA.

[2]Department of Electrical and Computer Engineering, Purdue University, Indianapolis, IN, 46202, USA

*To whom correspondence should be addressed. +1 317-278-9625; Email: czhang87@iu.edu. Correspondence is also addressed to Sha Cao, Email: shacao@iu.edu.

[+]These authors have equal contribution to this work.

ABSTRACT

The metabolic heterogeneity, and metabolic interplay between cells and their microenvironment have been known as significant contributors to disease treatment resistance. Our understanding of the intra-tissue metabolic heterogeneity and cooperation phenomena among cell populations is unfortunately quite limited, without a mature single cell metabolomics technology. To mitigate this knowledge gap, we developed a novel computational method, namely scFEA (single cell Flux Estimation Analysis), to infer single cell fluxome from single cell RNA-sequencing (scRNA-seq) data. scFEA is empowered by a comprehensively reorganized human metabolic map as focused metabolic modules, a novel probabilistic model to leverage the flux balance constraints on scRNA-seq data, and a novel graph neural network based optimization solver. The intricate information cascade from transcriptome to metabolome was captured using multi-layer neural networks to fully capitulate the non-linear dependency between enzymatic gene expressions and reaction rates. We experimentally validated scFEA by generating an scRNA-seq dataset with matched metabolomics data on cells of perturbed oxygen and genetic conditions. Application of scFEA on this dataset demonstrated the consistency between predicted flux and metabolic imbalance with the observed variation of metabolites in the matched metabolomics data. We also applied scFEA on publicly available single cell melanoma and head and neck cancer datasets, and discovered different metabolic landscapes between cancer and stromal cells. The cell-wise fluxome predicted by scFEA empowers a series of downstream analysis including identification of metabolic modules or cell groups that share common metabolic variations, sensitivity evaluation of enzymes with regards to their impact on the whole metabolic flux, and inference of cell-tissue and cell-cell metabolic communications.

KEYWORDS

Single cell genomics, single cell metabolic flux estimation, graph neural network, scRNA-seq data, metabolic heterogeneity.

INTRODUCTION

Metabolic dysregulation is a hallmark of many disease types including cancer, diabetes, cardiovascular disease and Alzheimer's disease [1-7]. In cancer, the diseased cells are well understood to rewire their metabolism and energy production to support rapid proliferation, sustain viability, and promote acquired drug resistance [8-11]. Here, the diseased cells often react differently to the microenvironmental stress, resulting in an increased repertoire of possible cellular responses to compromise the efficacy of drug therapies, and synergistic cooperation among the cells that can ultimately enhance the survival of the entire population [12, 13]. The metabolome is an excellent indicator of phenotypic heterogeneity due to its high dynamics and plasticity [14]: one may expect to see a subset of cancerous cells, such as circulating tumor cells, that display abnormally high metabolic rates compared with many others with normal metabolism, and rare cells that successfully cope with microenvironmental stress, whereas the others die. Unfortunately, single cell metabolomics is still in its infancy, limited by its relatively low throughput and low sensitivity [14-20]. Hence, our understanding of metabolic dysregulation of human disease has been immensely limited by our technology to study the metabolic landscape at single-cell level and in the context of their tissue microenvironment [21-28].

Single cell RNA-Seq (scRNA-seq) data has been widely utilized to characterize cell type specific transcriptional states in a complex tissue. Large amount of scRNA-seq data are endowed with the potential to deliver information on a cell functioning state and its underlying phenotypic switches [29-38]. Realizing the strong connection between transcriptomic and metabolomic profiles, scRNA-Seq has found its application in portraying metabolic variations. Most of the existing studies examined single cell metabolic changes using the expression levels of key metabolic genes or pathways [29-36], without considering of constraints of metabolic network. On the contrary, the Flux Balance Analysis (FBA) describes the potential flux over the topological structure of a metabolic network, with a set of equations governing the mass balance at steady state. Studies coupling single cell transcriptomics data and the FBA steady-state framework have only recently emerged [37, 38]. It is noteworthy that these models are intended for modeling the whole tissue level fluxes, restricted to a small portion of the whole metabolic map, for cells of pre-defined groups. In other words, technology to integrate scRNA-seq data with whole metabolomic FBA constraints with single cell resolution is yet to be developed. Therefore, it is urgent to design advanced computational tools to empower a reliable estimation of cell-wise metabolic flux and states from scRNA-seq data by designing more appropriate and sophisticated model to translate single cell transcriptomes to single cell fluxomes [39, 40].

Computational challenges to estimate cell-wise metabolic flux arise from the following aspects: (1) multiple key factors determine cells' metabolic states, including exogeneous nutrients availability in the tissue microenvironment, leading to a disjunction of cell type specific markers and metabolic phenotypes, and making conventional single cell clustering methods inapplicable; (2) the whole metabolic network is of high complexity, hence a proper computational reorganization of the network is needed to reach a balance between resolution of metabolic state characterization and computational feasibility; (3) the intricate non-linear dependency between transcripts level and reaction rates calls for a more sophisticated model to fully capitulate the relationships; and (4) alternations on different enzymes of a metabolic pathway may result in common metabolic phenotypes, however, exactly which enzymes share such common effect to the metabolic flux change remains largely unknown.

In this study, we developed a novel computational method, namely **s**ingle-**c**ell **F**lux **E**stimation **A**nalysis (scFEA), to estimate the relative rate of metabolic flux at single cell resolution from scRNA-Seq data. Specially, scFEA is empowered by the following computational innovations that can effectively solve the above challenges: (i) a probabilistic model to leverage the flux balance constraint on varied metabolic fluxomes among a large number of single cells, (ii) a metabolic map reduction approach based on network topology and gene expression status, (iii) a multi-layer neural network model to capture the dependency of metabolic flux on the enzymatic gene expressions, and (iv) a novel graph neural network architecture and solution to maximize the overall flux balance of intermediate substrates through all cells. To experimentally validate scFEA, we generated an scRNA-seq data of a patient derived pancreatic cancer cells under four conditions of perturbed oxygen level and metabolic regulators, and matched tissue level metabolomics data and qRT-PCR profiles of key metabolic genes. We validated that the variations of metabolic flux predicted by scFEA are highly consistent with the observed metabolomic changes under different conditions. The scFEA predicted fluxome suggested the accumulation of glycolytic metabolites and depletion of TCA cycle metabolites, caused by suppression of the glycolysis pathway and TCA cycle pathways in both normoxia and hypoxia conditions. We also applied scFEA on scRNA-seq data collected from real cancer tumor microenvironment and quantified the level of metabolic shifts in cancer and stromal cells. Notably, the fluxome estimated by scFEA enables a series of downstream analysis including identification of cell or tissue level metabolic stress, sensitivity evaluation of enzymes to the metabolic flux, and inference of cell-tissue and cell-cell metabolic exchanges.

RESULTS

Problem formulation and analysis pipeline

scFEA consists of three major computational components, namely (1) network reorganization, (2) cell-wise metabolic flux estimation, and (3) downstream analyses including estimation of metabolic stress, perturbation of metabolic genes, and clustering of cells with different metabolic states. In this work, we mainly focus on solving cell-

wise metabolic flux and states for human cells. The input of scFEA is an scRNA-seq data, with cell labels and sets of to be analyzed metabolic reactions as optional information.

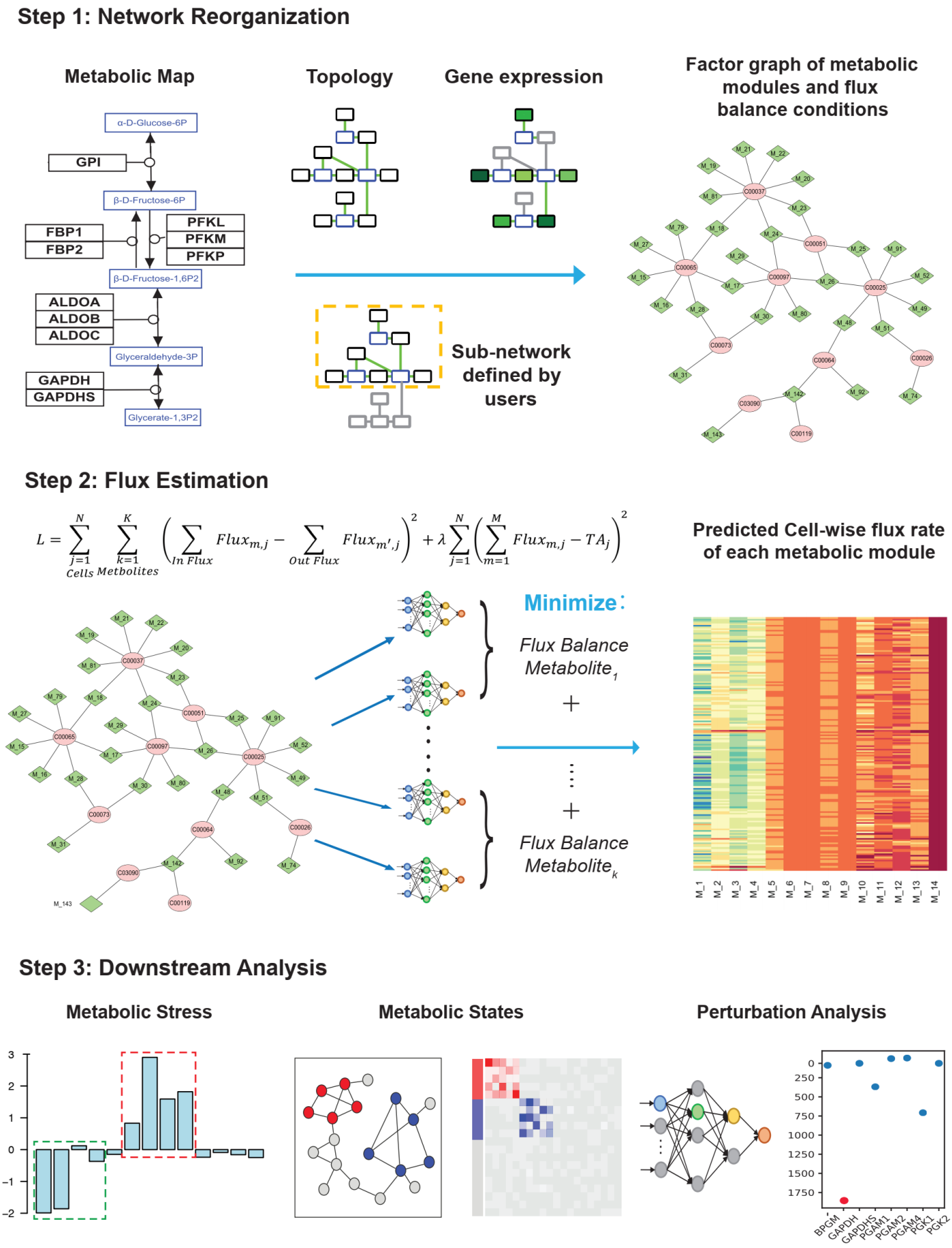

Fig 1. The computational framework of scFEA.

Figure 1 illustrated the detailed analysis framework of scFEA. In step (1), scFEA reorganizes the whole metabolic network into a factor graph composed by sets of metabolic modules and flux balance constraints, and the network reduction is achieved by considering the network topology, the expression status of metabolic genes, and optionally the customized network regions. This approach increases the robustness of flux estimation and reduces the computational complexity. In step (2), metabolic flux of each module will be modeled as a non-linear function of the expression levels of the enzymes in the module, and the non-linearity relationship is captured by a neural network with 2-4 layers. We assume that each cell is under metabolic steady state. To solve for the neural network parameter, scFEA introduced a flux balance constraint among the modules of all the single cells in the tissue based on a probabilistic model. Specially, scFEA optimizes an objective function that approximates the overall flux imbalance of the whole tissue, which assumes that the metabolic flux through all cells should reach such levels that would minimize the overall imbalance of the in-/out-flux of intermediate module substrates. In step (3), scFEA conducts downstream analysis to (i) detect the metabolites or pathways with high imbalance in certain cell group, (ii) assess the impact of metabolic genes on the current metabolic flux, and (iii) identify cell groups with distinct variations with regards to certain metabolic fluxes.

Reorganization of metabolic map

A metabolic network consists of reactions that fall under four major types, namely import, metabolism, biosynthesis, and export. For reactions in metabolisms, we collected the human metabolic pathways from KEGG database [41]; for import and export reactions, we collected the transporters from transporter classification database [42], for biosynthesis reactions, we collected the biosynthesis pathways from KEGG database and literatures (see details in Supplementary Methods). The collected metabolic map covers the metabolism, transport and biosynthesis of mono-/poly-saccharides, glycan, amino acids, fatty acids, and nucleic acids in human, including 727 genes of 541 enzymes, 1880 reactions, 8027 metabolites, and 116 transporter genes of 35 metabolites. Completes gene and reaction lists of the collected human metabolic map is given in Supplementary Table S1.

We first reorganize the metabolic network based on its topological structure. The compounds commonly involved in multiple reactions were excluded from further analysis, such as H2O, ATP, NADH, or other co-factors (Supplementary Table S1). Connected reactions will be coerced into a module if (1) the reactions could be linearized to have only one input and one output, and (2) the set of reactions has included all the reactions with in-flux or out-flux of all the intermediate compounds. **Figure 2C** illustrates several examples of how network motifs in the input metabolic network is transformed into metabolic modules. Specially, adjacent reactions without significant in-/out-flux other than the module input and output will be merged into one module. Such a network reduction approach will enable a more robust flux estimation by estimating the flux of one module instead of individual reactions, and a more efficient computation over a simplified network topological structure.

We reorganized the collected human metabolic map into a network of reaction modules, consisting 175 modules of 21 super module classes, 125 metabolites, out of which 84 are intermediate substrates, and 727 genes, as detailed in Table 1 and Supplementary Table S1. **Figure 2A** illustrates the functional group and complete topological structure of the collected metabolic modules and super modules. It is noteworthy the topology of the reorganized modules naturally forms a factor graph, in which each module and metabolite can be treated as a variable and factor node, respectively. **Figure 2B** shows the reorganized factor graph for human metabolic map, which is utilized in further flux estimation.

When reorganizing the metabolic map, scFEA also takes into consideration: (1) the user selected metabolic network and (2) the context specific expression levels of genes in the given dataset. For (2), the modules that are known a priori to carry little or no flux will be excluded from further analysis. Specifically, for a given scRNA-seq data, scFEA will first determine for all the genes whether they have an active expression state using our in-house Left Truncated Mixture Gaussian model [43] (see details in Methods). The default setting of scFEA considers a module is blocked, if the module becomes disconnected after removing the reactions whose associated genes do not have significantly non-zero expressions throughout all the cell,. The blocked modules will be removed before further analysis. On account of the common drop-out events in scRNA-Seq data, scFEA also enables a more conservative assumption to remove a module only if none of the genes involved in all reactions of this module has significantly active expressions. The genes, input and output metabolites, and topological structure of the filtered modules will be utilized for further flux estimation.

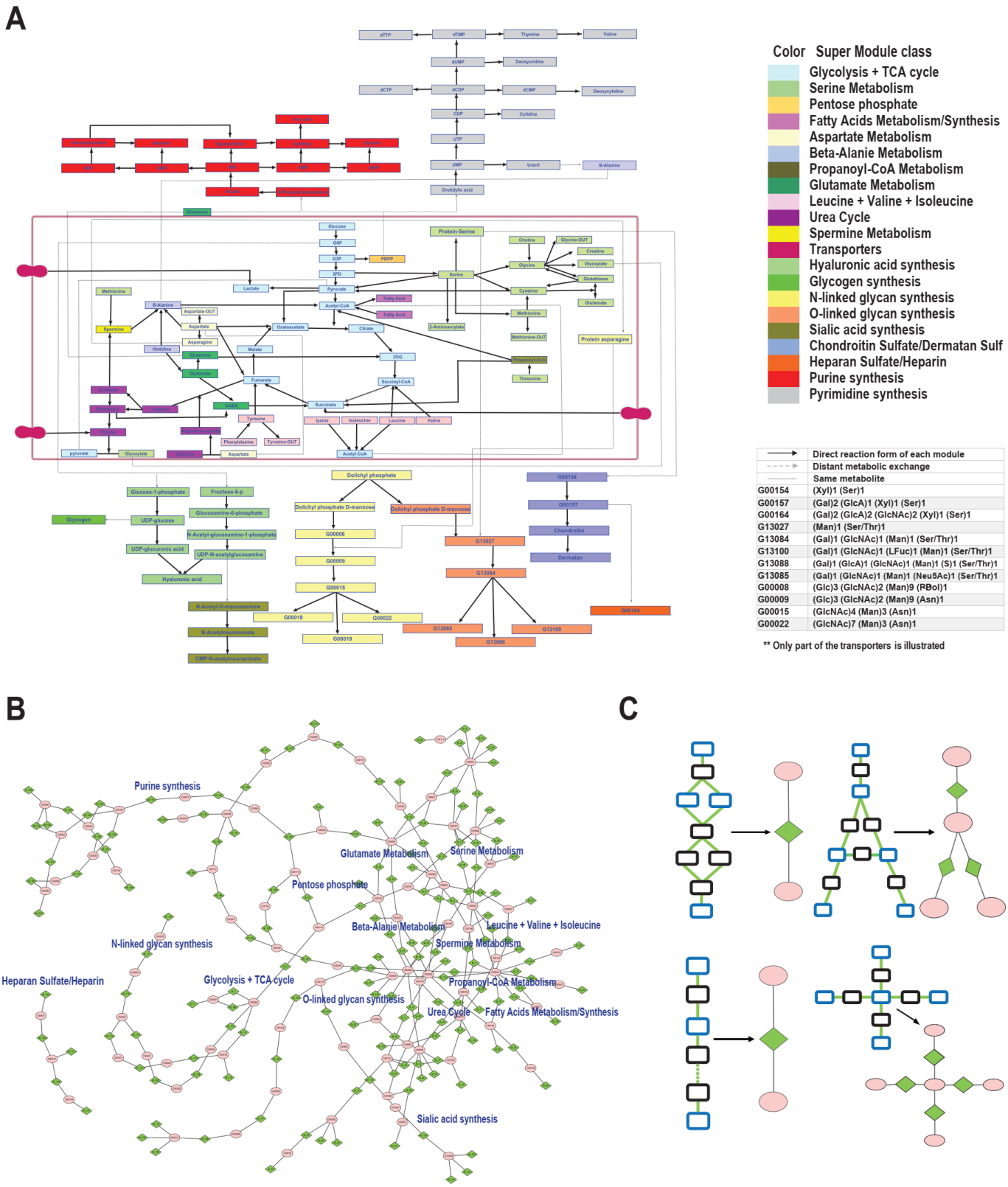

Figure 2. Reorganized human metabolic map. (A) Collected human metabolic modules and super module classes. (B) Factor graph representation of the reorganized human metabolic map, in which the modules and metabolites were colored by green and pink, respectively. (C) Examples of how the network motifs in the metabolic map are simplified into metabolic modules, where the reactions and metabolites are represented by black and blue rectangular, and modules and metabolites are colored by green and pink. Chain-like reactions can be directly simplified; a complicate module connected by multiple branches can be shrunk into one point linked with the multiple branches; and complicated intersections cannot be simplified.

Table 1: Super metabolic module information

SM ID	Super Module class	#Modules	#Genes
1	Glycolysis + TCA cycle	14	83
2	Serine Metabolism	18	114
3	Pentose phosphate	1	28
4	Fatty Acids Metabolism/Synthesis	2	81
5	Aspartate Metabolism	5	35
6	Beta-Alanine Metabolism	5	48
7	Propionyl-CoA Metabolism	2	25
8	Glutamate Metabolism	5	13
9	Leucine + Valine + Isoleucine	8	99
10	Urea Cycle	8	30
11	Spermine Metabolism	2	7
12	Transporters	35	80
13	Hyaluronic acid synthesis	9	26
14	Glycogen synthesis	1	4
15	Glycosaminoglycan synthesis	3	14
16	N-linked glycan synthesis	13	88
17	O-linked glycan synthesis	5	17
18	Sialic acid synthesis	4	12
19	Glycan synthesis	1	5
20	Purine synthesis	17	67
21	Pyrimidine synthesis	17	49

Mathematical consideration and formulation of metabolic flux in individual cells

We developed a novel graph neural network architecture to model cell-wise metabolic flux of each module by using their gene expression levels in each individual cell. For a clear model setup, we formulate the metabolic network as a factor graph, where each module represents a variable and each compound a factor node carrying a likelihood function describing the flux balance among modules (**Figure 2B**). The computational consideration is that the flux shift of a metabolic module generally impacts its neighboring modules, which can be characterized by aggregating the expression variations of the genes in its neighborhood over the metabolic network. We also hypothesize the metabolic flux through all cells should minimize the overall imbalance of the in-/out-flux of intermediate substrates, i.e. maintaining the maximal flux balance of intermediate substrates on the whole tissue level.

We denote $FG(C^{1\times K}, RM^{1\times M}, E = \{E_{C\to R}, E_{R\to C}\})$ as the factor graph, where $C^{1\times K} = \{C_k, k = 1, \dots, K\}$ is the set of K compounds, $RM^{1\times M} = \{R_m, m = 1, \dots, M\}$ is the set of M metabolic modules, $E_{C\to R}$ and $E_{R\to C}$ represent direct edges from module R_m to compound C_k and from compound C_k to module R_m, respectively. For the k-th compound C_k, we define the set of reactions consuming and producing C_k as $F_{in}^{C_k} = \{R_m | (R_m \to C_k) \in E_{C\to R}\}$ and $F_{out}^{C_k} = \{R_m | (C_k \to R_m) \in E_{R\to C}\}$. For a scRNA-seq data set with N samples, we denote $Flux_{m,j}$ as the flux of the mth module in the sample $j, j = 1 \dots N$, and $F_j = \{Flux_{1,j}, \dots, Flux_{M,j}\}$ as the whole set of the reaction fluxes. Our computational hypothesis for the whole tissue system is that flux imbalance of the intermediate metabolites at the whole tissue level should be minimized. Noting the tissue level flux balance can be reflected as the total flux balance of all measured cells, the likelihood function of the tissue level flux can be written as:

$$\phi(C, F) = \prod_{j=1}^{N} \prod_{k=1}^{K} \phi\left(C_{k,j} | F_j\right) \phi\left(F_j\right)$$

, where $\phi\left(C_{k,j} | F_j\right) = \phi\left(C_{k,j} | F_{in}^{C_k}, F_{out}^{C_k}\right) \propto e^{-\frac{\beta\left(\sum_{m\in F_{in}^{C_k}} Flux_{m,j} - \sum_{m\prime\in F_{out}^{C_k}} Flux_{m\prime,j}\right)^2}{2}}$ and $\phi\left(F_j\right) \propto e^{-\frac{\gamma\left(\sum_{m=1}^{M} Flux_{m,j} - TA_j\right)^2}{2}}$, β and γ are hyperparameters, and TA_j is a surrogate for total metabolic activity level of cell j, which can be assigned as a constant or total expression of metabolic genes in j. Here we introduce $\phi\left(F_j\right)$ and TA_j to avoid a trivial solution of $Flux_{m,j} \equiv 0$.

scFEA models the flux of reach reaction, namely $Flux_{m,j}$, as a nonlinear function of the expression changes of the genes associated with the module. This hypothesis can be supported by many existing studies that reveal the high

explainability of transcriptomic levels for the protein level of enzymes [24, 44, 45]. Denote $\boldsymbol{G^m} = \{G_1^m, \dots, G_{i_m}^m\}$ as the genes associated with the reactions in R_m, and $\boldsymbol{G_j^m} = \{G_{i_1,j}^m, \dots, G_{i_m,j}^m\}$ as their expressions in sample j, where i_m stands for the number of genes in R_m. We model $Flux_{m,j} = f_{nn}^m(\boldsymbol{G_j^m}|\,\boldsymbol{\theta_m})$ as a multi-layer fully connected neural network with the input $\boldsymbol{G_j^m}$, where $\boldsymbol{\theta_m}$ denotes the parameters of the neural network (**Figure 3**). Then the $\boldsymbol{\theta_m}$ and cell-wise flux $Flux_{m,j}$ can be solved by minimizing the following loss function L, where $\lambda \sim \frac{\gamma}{\beta}$ serves as a hyperparameter:

$$\mathrm{L} = -\log(\phi(C,F)) = \sum_{j=1}^{N}\sum_{k=1}^{K}\left(\sum_{m\in F_{in}^{C_k}} Flux_{m,j} - \sum_{m'\in F_{out}^{C_k}} Flux_{m',j}\right)^2 + \lambda\sum_{j=1}^{N}\left(\sum_{m=1}^{M} Flux_{m,j} - TA_j\right)^2$$

It is noteworthy that the above formulation defines a new graph neural network architecture for flux estimation over a factor graph, where each variable is defined as a neural network of biologically meaningful attributes, i.e. the genes participating in each metabolic module, and the information aggregation between adjacent variables is constrained by the balance of chemical mass of the in- and out- flux of each intermediate metabolites. Noted, the number of intermediate constraints (K) and large sample size (N) of scRNA-seq data ensures the identifiability of $\boldsymbol{\theta_m}$ for the f_{nn}^m at a certain complexity level (see details in Methods).

The challenges to minimize the objective function include the following: (1) the flux of each module affects the balance of its input and output and multiple modules are involved in the balance of one intermediate substrate, hence perturbing one single flux at each step may not converge, and on the other hand (2) the direction for simultaneously updating a large group of fluxes cannot be theoretically derived. The two challenges prohibit a direct utilization of back propagation or gradient descending methods. We developed an effective optimization strategy for L by adopting the idea of information transfer in belief propagation, which has been commonly utilized in analyzing cyclic networks such as Markov random field [46]. Specifically, L is minimized by iteratively minimizing the flux balance of each intermediate metabolite C_k and the weighted sum of the flux balance of the Hop-2 neighbors of C_k in the factor graph, as the L_k^* defined below:

$$L_k^* = \sum_{j=1}^{N}\left(\sum_{m\in F_{in}^{C_k}} Flux_{m,j} - \sum_{m'\in F_{out}^{C_k}} Flux_{m',j}\right)^2 + \sum_{k'} W_{k'}\sum_{j=1}^{N}\left(\sum_{m\in F_{in}^{C_{k'}}} Flux_{m,j} - \sum_{m'\in F_{out}^{C_{k'}}} Flux_{m',j}\right)^2$$

, where $C_{k'}$ are the Hop-2 neighbors of C_k, $W_{k'}$ is proportional to the current total imbalance of all the Hop-2 neighbors of $C_{k'}$ except for C_k itself (see more details in Methods). Such a regional perturbation strategy over the whole graph can effectively leverage the search of global minimum and computational feasibility.

The output of scFEA includes f_{nn}^m, $\boldsymbol{\theta_m}$ for each module and predicted cell-wise metabolic flux $Flux_{m,j}$.

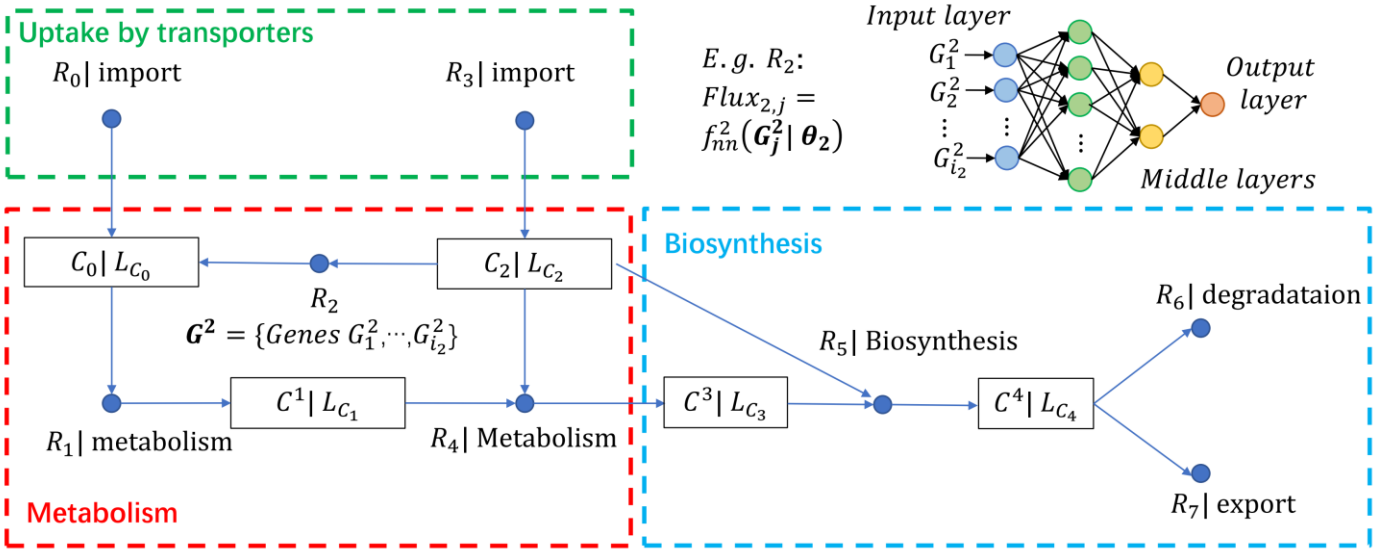

Figure 3. A toy model of the factor graph of metabolic modules, flux balance conditions, and the flux model for the module R_2 (top-right). In the factor graph, each metabolite (C) corresponds to one flux balance condition and serves as a factor, and each reaction (R) is a variable. Import and export/degradation reactions are considered as having no input or output substrates.

Method validation on a scRNA-seq data with perturbed metabolic conditions and matched metabolomics data

To validate the cell-wise flux estimated by scFEA, we generated an scRNA-seq dataset consisting of 162 patient-

derived pancreatic cancer cells (Pa03c cell) under two crossed experimental conditions: APEX-1 knockdown (APEX-1 KD) or control, and under hypoxia or normoxia conditions (see detailed experimental procedure and data processing in Methods). Metabolomics profiling of 14 metabolites, namely glucose, glucose-1 phosphate, glucose-6 phosphate, pyruvate, and lactate in the glycolysis pathway, citrate, 2-oxoglutarate, succinate, fumarate, malate in the TCA cycle, and amino acids glutamate, glutamine, serine, and ornithine were collected on bulk wildtype Pa03c cells and APEX-1 inhibition cells under the normoxia conditions, each with three replicates (Supplementary Table S2). We utilized the Smart-seq2-fluidigm protocol for single cell RNA sequencing for saturated gene detection of each single cell, to enable a more accurate modeling of metabolic flux. *APEX1* is a multifunctional protein that interacts with multiple transcriptional factors (TFs) to regulate cellular responses to hypoxia and oxidative stress [47]. Our previous studies identified significant roles of *APEX1* in the regulation of Pa03c cells' response to metabolic environment changes [48, 49].

To the best of our knowledge, scFEA is the first computational tool to estimate metabolic flux at single cell level. Without baseline methods for comparisons, we validate scFEA by examining the consistency between the metabolic flux variation predicted by scFEA and experimental observations. We identified 126 up- and 443 down- regulated genes in APEX-1 KD vs Control under the normoxia condition, and 260 up- and 1496 down- regulated genes under hypoxia condition. Pathway enrichment analysis showed that the TCA cycle pathway (normoxia: p=0.003, hypoxia: p=1.12e-07) and oxidative phosphorylation (normoxia: p=3.17e-4, hypoxia: p=1.77e-08,) are significantly enriched by down regulated genes, under both normoxia and hypoxia conditions. This suggests that the knock down of APEX-1 may lead to inhibited cellular aerobic respiration. In addition, genes regulated by *HIF1A* (hypoxia-inducible factor 1-alpha), including glycolysis and TCA cycle genes, were observed to be up- and down-regulated respectively, in comparing the hypoxia vs normoxia conditions in the control Pa03c cells. This is consistent to the common knowledge of hypoxia response. Our of the 14 metabolites, we have seen an increase in glucose, glucose-1 phosphate, glucose-6 phosphate, and lactate, and a decrease in 2-oxoglutarate, succinate, fumarate, and malate in APEX1-KD vs control cells under the normoxia condition. In summary, analysis of the single cell gene expression and bulk cell metabolomic data revealed that knockdown of APEX1 affects the cells' glucose metabolism and inhibits the cells' TCA cycle pathway, under both normoxia and hypoxia condition. **Figure 4A** illustrates the variation of genes and metabolites involved in glycolysis, pentose phosphorylation, TCA cycle, glutaminolysis and aspartate metabolism pathways in APEX1-KD vs control under normoxia condition. Complete list of differentially expressed genes and pathway enrichment results were provided in Supplementary Table S3.

Consistency between the scFEA predicted flux variation and the metabolomics data. We applied scFEA to the aforementioned scRNA-seq data of the four conditions. We first focus on the normoxia conditions where matched single cell expression and metabolomics data are available. scFEA predicted decreased metabolic flux for the modules in glycolysis and TCA cycle in APEX1-KD vs control, i.e. glucose → glyceraldhyde-3P (G3P) → pyruvate → citrate → succinate → oxaloacetate (OAA), where particularly, the reactions in the downstream of the reaction chain is more suppressed in APEX1-KD (**Figure 4B**). We examined the correlation between the predicted flux change with the observed metabolomic change of intermediate metabolites in glycolysis and TCA cycle pathways, and observed a Pearson Correlation Coefficient (PCC) of 0.86 (p=0.006) (**Figure 4B**), suggesting the high consistency between predicted flux variation with the metabolic changes. Using metabolomics data, we observed increase of production for glucose, G1P, G3P and lactate, and decrease of production for 2OG, succinate, fumarate, and malate in APEX1-KD vs control (**Figure 4C**). We also correlated the metabolomic change with the averaged expression change of the enzymes catalyzing the reactions involving the metabolomics. However, no significant correlation was observed (PCC=-0.03, p=0.943, **Figure 4C**), suggesting that single cell gene expression itself, without considering the constraints from the intricate metabolic network as in scFEA, doesn't produce a good estimate of single cell metabolic landscape. scFEA leveraged the non-linear relationships between gene expression and enzymatic reaction rate, and the flux balance constraints of the metabolites, and hence its predicted metabolic flux is more consistent to the true metabolomics changes.

High consistency of the predicted metabolic stress with experimentally observed metabolomic changes. scFEA predicted in and out flux for each metabolite allows us to investigate the cell-wise metabolic stress, which was defined as the

imbalance of the in-/out-fluxes of each intermediate metabolites in each cell. **Figure 4D** shows that the G1P, G6P and lactate were accumulating while 2OG, succinate, succinyl-CoA, and fumarate were depleted in APEX1-KD vs control. A PCC of 0.67 (p=0.024) was observed between the predicted metabolic stress and the true metabolic change, demonstrating a high accuracy of the predicted metabolic stress level. Detailed predicted and observed metabolic imbalance were provided in Supplementary Table S2. **Figure 4E** shows the predicted cell-wise fluxome of the glycolysis and TCA cycle modules for cells of the four conditions. Here each column represents the flux between two metabolites, shown on the x-axis, for all the cells, shown on the y-axis. For two neighboring fluxes, the product of the reaction on the left is the substrate of the reaction on the right, and in a perfectly balanced flux condition, the two neighboring fluxes should be equal. We observed, in general, higher flux of the glycolytic modules than the TCA cycle modules, with the largest average flux gap seen on Pyruvate → Acetyl-CoA and Acetyl-CoA → Citrate. In addition, the flux of the downstream reactions (citrate → 2OG → succinyl-CoA → succinate) of the TCA cycle is lower than the upstream reactions (succinate → fumarate → malate → OAA). A possible explanation for the leaky metabolic flux is that some of the intermediate substrates flow to other branches, majorly for biosynthesis of amino acids. Among the four conditions, we identified that the hypoxia control group has the highest flux rate of glycolysis and TCA cycle modules. Clearly, the inhibition of APE1-X significantly decreased the metabolism rate of glucose. Combined with the accumulations of glycolytic substrates and depletions of TCA cycle substrates identified by the metabolic stress and metabolomics data analysis, our speculate that the knock-down of APE1-X may directly impact the downstream part of glycolysis, the whole TCA cycle and further oxidative phosphorylation, leading to accumulation of G1P and G6P as a result of the blockage. Up regulation of glucose transporters was also observed in APE1-X KD vs control, further suggesting the accumulation of glycolytic substrates.

Perturbation analysis of flux deterministic genes. We also conducted a perturbation analysis to tease out the key genes with high impact on each metabolic module (see details in Methods). The following genes were identified to have the highest impact on metabolic flux: HK1 and HK2 (Glucose→G6P), ALDOA and GPI (G6P→G3P), GAPDH and PGK1 (G3P → 3PD, ENO1, PGAM1, and PKM (3PD → Pyruvate), PDHA2 (Pyruvate → Acetyl-Coa), LDHA (Pyruvate→Lactate), ACLY (Acetyl-CoA+OAA→Citrate), IDH1 (Citrate→2OG), DLD and OGDH (2OG→Succinyl-CoA), SUCLG1 (Succinyl-CoA → Succinate), SDHB (Succinate → Fumarate), FH (Fumarate → Malate), MDH1 (Malate→OAA). A qRT-PCR experiment was conducted to confirm the down regulation of the above key metabolic genes (see details in Methods). We also conducted a module level perturbation analysis by increasing or decreasing the expression of genes in a certain module (see details in Methods). Non-surprisingly, a decrease of expression on genes of the downstream part of glycolysis pathway in the control cells will lower the flux of the TCA cycle, causing the accumulation of glycolytic intermediate substrates and depletion of TCA cycle metabolites, which is consistent to our experimental observations.

Detecting groups of metabolic modules with similar variations and cells with distinct metabolic states. We also applied scFEA to a larger metabolic map, with 11 metabolic super modules and transporters, and then examined the high-level organization of the modules. Based on only the metabolic network connectivity, **Figure 4F** illustrated classes of metabolic modules derived using a spectral clustering method (see Methods), in which glycolysis, TCA cycle and acetyl-coA related modules, serine metabolism, urea cycle, and other amino acids metabolism form distinct module classes. To examine the high level structure based on the flow of flux, we further conducted a clustering analysis of the metabolic modules by considering both the network connectivity and flux similarity. The distance between two modules R_i and R_j is defined as $\alpha d(R_i, R_j) + (1-\alpha) d^F(R_i, R_j)$, where $d(R_i, R_j)$ is their normalized spectral distance, and $d^F(R_i, R_j)$ is their normalized similarity in estimated flux through different cells. Here $\alpha = 0.3$ is used in the analysis. **Figure 4G** shows the metabolic module clusters by integrating topological and flux distance information. Three distinct clusters were identified, including (1) glycolysis and fatty acids metabolism, (2) TCA cycle and pyruvate metabolism, and (3) metabolism of amino acids and other metabolites, which correspond to the modules with (i) decreased flux and accumulated substrates, (ii) decreased flux and depleted substrates, and (iii) unchanged flux and metabolites, supported by scFEA prediction and metabolomic observations, respectively. This observation further validated the rationale of scFEA predicted fluxomes.

We also conducted cell clustering based on the metabolomic modules with varied flux (Methods). Non-surprisingly, the cells clusters were aligned with experimental conditions, forming four group of cells with high, intermediate, low, and extremely low metabolic rates (Supplementary Table S2).

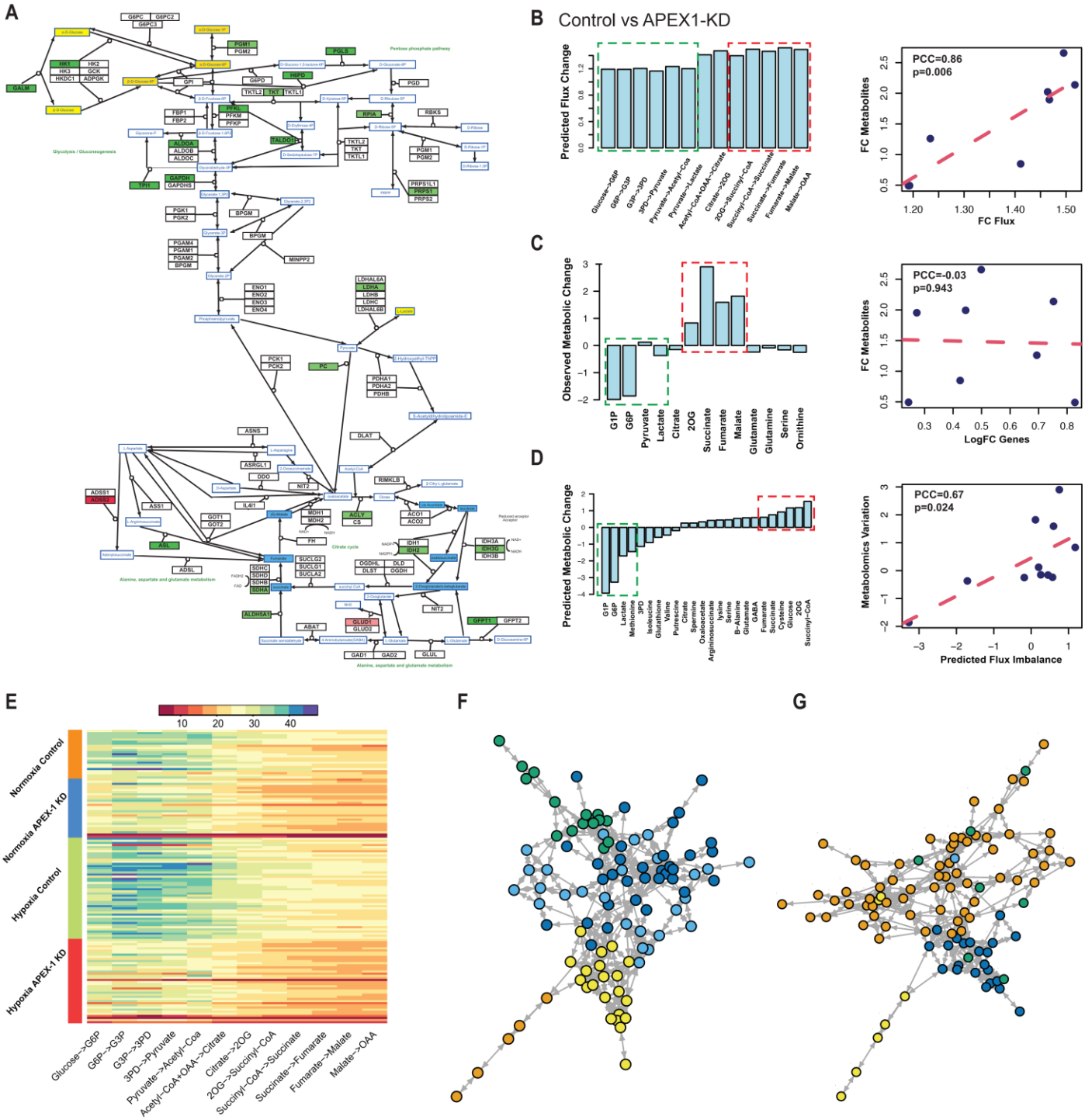

Figure 4. Application of scFEA on matched scRNA-seq and metabolomics data of Pa03C cells. (A) Gene expression and metabolomic variations of the glycolysis, pentose phosphate, TCA cycle, glutamine, and aspartate metabolic pathways in APEX1-KD vs control under normoxia condition. Genes/metabolites were shown in rectangular boxes with black/blue borders, up/down regulated genes were colored in red/green, increased and decreased metabolites were colored in yellow/blue, respectively. The darker color suggests a higher variation. (B) Predicted flux change (left, *x*-axis: metabolic module, *y*-axis: predicted flux change) in control vs APEX1-KD, and correlation between predicted flux change and observed metabolite change (right, *x*-axis: fold change of predicted flux, *y*-axis: fold change of observed metabolite abundance, each data point is one metabolite). (C) Observed metabolite change (left, *x*-axis: metabolites, y-axis: observed abundance difference) in control vs APEX1-KD, and correlation between expression change of the genes of each reaction and observed metabolomics change (right, *x*-axis: averaged fold change of the expression of the genes involved in each reaction, y-axis: fold change of observed metabolote abundance, each data point is one metabolite).

(D) Predicted metabolic stress (left, *x*-axis: metabolites, y-axis: predicted abundance change) in control vs APEX1-KD and correlation between predicted flux imbalance and metabolite variation (right, *x*-axis: predicted imbalance of the in-/out- flux of each metabolite, *y*-axis: difference of observed metabolomic abundance, each data point is one metabolite). In (B-D) all comparisons were made by comparing control vs APEX1-KD under normoxia. The green and red dash-blocks represents the accumulated (green) and depleted (red) metabolites in control vs APEX1-KD. (E) Profile of the predicted fluxome of 13 glycolytic and TCA cycle modules (x-axis: metabolic modules, y-axis: experimental conditions). (F) Clusters of metabolic modules inferred by using the network connectivity structure only. (G) Clusters of metabolic modules inferred by using the network topological structure (weight of 0.3) combined with predicted fluxome (weight of 0.7).

Application on scRNA-seq data of tumor microenvironment revealed distinct metabolic stress, exchange and varied metabolic states in cancer and stromal cells.

We also applied scFEA on two publicly available scRNA-seq datasets collected from the microenvironment of melanoma (GSE72056) and head and neck cancer (GSE103322) (see detailed data information in Methods). All the metabolomic modules across the whole metabolomic network was used. Due to the lack of matched metabolomic information, we focused on demonstrating the capability of scFEA in inferring metabolic stress, exchange of metabolic flows and metabolic modules and cells with distinct variations, for both cancer and stromal cells.

In both data sets, we identified that the malignant cells have the highest metabolic stresses, i.e. the total imbalance of intermediate substrates, followed by fibroblast and endothelial cells, and then immune cells. Specially, the malignant cells have the highest metabolic rates in most metabolic reactions comparing to other cell types in both melanoma and head and neck cancer, especially for the glucose and amino acids metabolic modules. On average, the estimated flux of TCA cycle and lactate production account for 43.4% and 52.5% of the glycolysis flux in head and neck cancer and 65.3% and 46.1% of the glycolysis flux (with additional carbon flow from other metabolites such as glutaminolysis) in melanoma, respectively, while the ratio of lactate production is much lower in other cell types. Our observation clearly suggested the Warburg effect and metabolic shift in cancer cells, which is consistent to our previously findings of high lactate production in melanoma [50]. Similar to the pancreatic cancer cell line data, we identified that both cancer and stromal cells in both cancer types tend to have depleted glucose, G1P and G6P. In addition, cancer cells tend to suffer from a high depletion level of acetyl-coA. On the other hand, TCA cycle intermediates and amino acids tend to be accumulated in cancer cells. These observations are consistent to the quantitative metabolomics data collected from solid cancer [51].

Interestingly, we noticed that the direction of imbalance for most intermediate metabolites seem to be the same throughout different cell types, though the imbalance level is much lower in stromal cells comparing to cancer cells. A possible explanation is that these cells were collected in a focused region of the same microenvironment, subject to similar microenvironmental stresses, such as hypoxia and altered pH level, which causes a similar impact on the metabolic landscape of cells of different types. This suggests that to better see the metabolomic heterogeneity, it is better to use spatial scRNA-Seq data, where the cells are more scattered away. Cell clusters with distinct metabolic states were identified in both data sets. Cancer cells and fibroblast cells form more similar metabolic characteristics comparing to immune cells while fibroblast cells show distinct fluxome profile of biosynthesis.

The predicted cell type specific fluxome and imbalance level of metabolites were given in Supplementary Table S4. In this study, we majorly focused on validating the computational concept and model of scFEA. Detailed analysis procedure and a comprehensive discussion of the cell-wise fluxome of the two data sets were available in the Github link: https://github.com/changwn/scFEA.

DISCUSSION

Despite a plethora of knowledge we have gained on metabolic dysregulation for many disease types, there are still major gaps in our understanding of the integrated behavior and metabolic heterogeneity of cells in the context of their microenvironment. Essentially, the metabolic behavior can vary dramatically from cell to cell due to their high plasticity, driven by the need to cope with various dynamic metabolic requirements. Considering single cell metabolomics technique is still in its infancy, large amount of transcriptomics data obtained by scRNA-seq has proven to be endowed with the potential to deliver information on a cell functioning state and its underlying phenotypic switches. Hence,

advanced computational tools are in pressing need to empower reliable prediction of cell-wise metabolic flux and states from scRNA-seq data. In this study, we developed a novel computational concept and method to predict metabolic flux at single cell resolution from single cell transcriptomics data, and the ultimate goal is to accurately construct and portray a compendium of metabolic states for different cell types and tissue contexts, and their relevance to various disease phenotypes.

The scFEA model has the following advantages: (1) the model characterizes true biological flux by leveraging the metabolic networks, and it is generally applicable as it requires only the input of scRNA-seq data; (2) the number of constraints, i.e. the number of flux balance conditions multiplied by the single cell number, is larger than the number of parameters, avoiding possible overfitting; and (3) The neural network based flux estimation can well handle the non-linear dependency between enzymatic gene expression and reaction rates. The scFEA model can also be extended to estimate activity level of functional modules in a general biological network such as signaling pathways. The expression level of a signaling path reflects its capacity and the signaling molecules can be viewed as intermediates.

The neural network based optimization framework of scFEA could enable a seamless integration of metabolomics data, kinetic parameters, spatial information, or other prior knowledge of metabolic imbalance, to better characterize cell and tissue level metabolic shifts of the target system. Specifically, metabolomics data, kinetic parameters or other prior knowledge can be utilized to better design the first layer of the neural network in modeling the flux of each module. Spatial information can be utilized to preselect group of cells for training spatially dependent model. A potential future direction is to implement the current flux estimation analysis in spatial transcriptomics to infer (1) metabolic shifts specific to spatial patterns and (2) metabolic exchange between adjacent cells. This application to spatial transcriptomics data will be particularly interesting for cancer tissue, to reveal how the metabolism and macromolecule biosynthesis in stromal cells such as cancer associated fibroblast affect the adjacent cancer cells.

scFEA seeks for a constrained optimization of flux balance, where each flux was modeled as a non-linear function of the relevant enzymatic gene expression levels. The flux of each module is currently constrained to be similar to the cell-wise total metabolic activity, TA_j, to avoid trivia solution. However, our analysis suggested one TA_j for each cell may lead to similar metabolic flux distribution for different cells. Although our current setting has been validated by our matched scRNA-seq and metabolomics data, applications on publicly available cancer data suggested a similar metabolic imbalance trend among different cell types. We speculate that setting TA_{mj} for each super module m in cell j may increase the flexibility of cell specific metabolic imbalance, but at the price of possible over-fitting. A more sensitive approach is to train a specific model for each pre-defined cell group. The biological rationale is that the neural network parameters contains the information of "kinetic parameters" that link gene expression with metabolic rate, which differ among distant cell types, or cells under different conditions. Hence it is rationale to assume cell type specific parameters.

Overall, scFEA can efficiently delineate the sophisticated metabolic flux and imbalance specific to certain cell groups. We anticipate it has the potential to decipher metabolomic heterogeneity, and teasing out the metabolomic susceptibility to certain drugs, and ultimately warrant novel mechanistic and therapeutic insights of a diseased biological system at an unprecedented resolution.

METHODS

Collection and reorganization of human metabolic map

We consider the human metabolic network as composed of different reaction types including metabolism, transport (including uptake and export), and biosynthesis. As detailed in Results, the reorganized network consists of 21 super module classes of 175 modules. All reactions related to metabolism were collected from the Kyoto Encyclopedia of Genes and Genomes database (KEGG) (61). In total, 11 metabolism related super modules were manually summarized, which is comprised of glycolysis, TCA cycle, pentose phosphate, fatty acids metabolism and synthesis, metabolism of amino acids namely serine, aspartate, beta-alanine, glutamate, leucine/valine/isoleucine and urea cycle, propionyl-CoA and spermidine metabolism [52]. The 11 metabolism super modules contain 1388 reactions, 317 enzymes, which corresponds to 3508 genes.

Transporters enable the trafficking of molecules in and out of cell membranes. We collected the human transporter proteins, their corresponding genes and metabolite substrates from the Transporter Classification Database [53, 54]. In total, 116 transporter genes, and 35 related metabolites were collected.

An essential part of metabolic map is the biosynthesis pathways. KEGG database and literature [7, 55-64] are the main information sources used for building biosynthesis modules. We collected 69 biosynthesis modules forming 9 super modules, namely biosynthesis of hyaluronic acid, glycogen, glycosaminoglycan, N-linked glycan, O-linked glycan, Sialic acid, Glycan, Purine and Pyrimidine. Overall, the biosynthesis modules include 142 enzymes catalyzing 280 reactions.

More details of the collection and reorganization of human metabolic map were provided in Supplementary Methods.

Selecting genes of significant expression.

We applied our inhouse method, LTMG, to determine the expression status of each genes in each single cell. LTMG considers the multi-modality of the expression profile of each gene throughout all the single cells, by assuming that the gene's expression follows a mixture of suppressed state and activated states, as represented by the following likelihood function [49].

$$\prod_{j=1}^{N}\left(\sum_{i=1}^{K} a_i p_i(x_j|u_i,\sigma_i) + a_{K+1}p_{K+1}(x_j|u_{K+1},\sigma_{K+1})\right)$$

, where $x_j, j = 1 \dots N$ are the expression profile of gene x in N cells, the index $1 \dots K$ are the K active expression states and $K+1$ is the suppressed expression state, a_i is the proportion of each state with $a_1 + \cdots + a_{K+1} = 1$, $a_{1\dots K} > 0$ and $a_{K+1} \geq 0$, p_i, u_i, and σ_i are the pdf, mean and standard deviation of each expression state. Specifically, LTMG considers the distribution of each mixing component, p_i, as a left truncated Gaussian distribution, to account for the noise of drop out events. In this work, LTMG was used to fit to each gene's expression and a gene was determined to have significant expression if $\sum_{i=1}^{K} a_i \geq 0.1$, i.e. the gene has active expression states in at least 10% cells.

Pre-filtering of active modules based on gene expression.

Each metabolic module contains an input, an output, and a number of enzymes catalyzing the reactions. A reaction is considered as disconnected if none of the genes encoding its catalyzing enzymes is significantly expressed. A metabolic module is considered as blocked if there is no connected path from the input to the output. Considering the common drop-out events in scRNA-Seq data, especially for the drop-seq data, we adopted a conservative approach to pre-trim the metabolic modules: essentially, a module will be removed from further analysis if none of the genes involved in all reactions of this module has significantly active expressions.

scFEA model setup and a belief propagation based solution of the flux model

<u>*Model Setup.*</u> We developed a novel optimization strategy to minimize L similar to the idea of belief propagation [65]. Specifically, the flux balance of each metabolite C_k, $L_K \triangleq \sum_{j=1}^{N}\left(\sum_{m\in F_{in}^{C_k}} Flux_{m,j} - \sum_{m'\in F_{out}^{C_k}} Flux_{m',j}\right)^2$, will be iteratively optimized, by taking into account all the Hop-2 neighbors in the factor graph (metabolites), denoted as $Ne(C_k)$, and Hop-4 neighbors (metabolites), i.e., $Ne^2(C_k) \coloneqq \{C_{k'} | C_{k'} \in Ne\big(Ne(C_k)\big)\backslash C_k\}$. Specifically, for a more efficient optimization, we adopt the idea of belief propagation by minimizing a reweighted flux imbalance: $L_K^* \triangleq L_K + \sum_{C_{k'}\in Ne^2(C_k)} W_{k'}L_{k'}$ at each iteration, where $W_{k'}$ is a weight value in $(0,1]$ representing the reliability of the current flux balance of $C_{k'}$. We set $W_{k'} = exp\left(-\frac{\sum_{C_{k''}\in Ne(Ne(C_{k'}))\backslash\{C_{k'},C_k\}} L_{k''}}{|Ne^2(C_{k'})\backslash\{C_{k'},C_k\}|}\right)$ as an exponential function of the negative averaged imbalance level of 2-hop neighbors (metabolite) of $C_{k'}$ excluding C_k, with higher $W_{k'}$ denoting lower imbalance of the metabolites. The underlying idea is that the more reliable the current flux is estimated for the modules involving $C_{k'}$, which is reflected by the averaged imbalance level of its 2-hop neighbors, a higher weight $W_{k'}$ should be given to $C_{k'}$, such that when minimizing L_K, a disruption of the flux balance of $C_{k'}$ will be more heavily penalized.

<u>*Neural network*</u> <u>*model setup.*</u> For each module, a neural network is used to represent the non-linear dependency between gene expressions and reaction rates. Each neural network has a_1 hidden layers each with a_2 hidden nodes,

and one output node. In this study, we took $a_1 = 3$ and $a_2 = 8$. A Hyperbolic Tangent activation function, $Tanhshrink(x) = x - \tanh(x)$, is used.

Clustering analysis of cells with distinct metabolic states

scFEA adopts an attributed graph clustering approach to identify the group of cells and metabolic modules forming a distinct metabolic state. Three clustering approaches were provided to the results of scFEA for different tasks, namely clustering of (1) metabolic modules, (2) cells share a common state on the overall metabolic map, and (3) cells share a common state on selected metabolic modules.

Clustering of metabolic modules. Denote the adjacency matrix of the context specific metabolic map as $A^{M\times M}$ and predicted metabolic flux as $Flux^{M\times N}$, where $Flux_{m,j}$ represents the predicted flux rate of the module m in cell j, a two-stage spectral clustering was applied to cluster the metabolic modules based on $A^{M\times M}$ and predicted $Flux^{M\times N}$. It is noteworthy here the $Flux^{M\times N}$ is usually much denser than the input scRNA-seq data since the metabolic modules without significant expression were excluded before the analysis. Specifically, denote $A^{F,M\times M}$ as the Euclidean distance of the M modules in $Flux^{M\times N}$, and $D^{M\times M}$ and $D^{F,M\times M}$ as the two diagonal matrices, in which $D_{ii} = \sum_{j=1}^{M} A_{ij}$ and $D_{ii}^{F} = \sum_{j=1}^{M} A_{ij}^{F}$. The normalized graph Laplacian matrices for the network topology and attributes similarity were defined as $L = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ and $L^F = I - D^{F^{-\frac{1}{2}}}A^F D^{F^{-\frac{1}{2}}}$. The normalized graph Laplacian matrices scale the topology and attributes similarity into the same scale. Denote $d(R_i, R_j)$ and $d^F(R_i, R_j)$ as the Euclidean distance between the metabolic modules R_i and R_j of the smallest P_1 eigenvectors of L and the smallest P_2 eigenvectors of L^F, the modules were clusters by the *K-mean* method with using the following distance

$$\alpha d(R_i, R_j) + (1-\alpha) d^F(R_i, R_j)$$

, here α, P_1 and P_2, and the number of clusters are hyperparameters. Our empirical analysis suggested a default setting as $\alpha = 0.3$, which assigns a higher weight to the similarity of the predict flux; $P_1 = \max\{3, floor(\frac{\#SM}{2})\}$, where $\#SM$ is the number super-modules in the current metabolic map; and $P_2 = \max\{3, floor(\frac{\#M}{17})\}$, where $\#M$ is the number of non-zero modules in the current metabolic map. The number of clusters should be pre-given by users, which depends on the number of cells, cell types, and metabolic modules.

Clustering of cells. For a given metabolic map or a predefined group of metabolic modules, such an identified module cluster, scFEA conducts cell clustering analysis by using the spectral cl
ustering approach based on the L^F and d^F as defined above.

Analysis of cell group specific metabolic stress and metabolic exchanges among cell groups.

The cell-wise metabolic flux estimated by scFEA enables the analysis of metabolic stress. For a pre-defined cell group such as cells of the same type, the total imbalance of each compound will be computed and ranked. One-way t-test was applied to test if the imbalance is significantly different to 0. The metabolic exchange among different cell groups from one tissue sample were identified as the metabolites with different sign of metabolic imbalance in different cell groups, such as accumulation and depletion, or exporting or importing. Tissue level metabolic stress is computed as the total imbalance throughout multiple cells.

Perturbation analysis

scFEA encodes a perturbation analysis to evaluate the impact of the change of each gene on the whole metabolic map. The perturbation analysis includes three components: (1) the direct impact of each gene G_i^m to the flux module m can be directly computed by its derivative $\frac{df_m^m}{dG_i^m}$ for all the modules containing G_i^m; (2) the impact of the flux change of one module A on other modules and flux balance of metabolites can be computed as the difference of flux of other modules estimated by scFEA while fixing the flux of module A at different values; (3) the impact of each gene's expression to the flux of distant modules and the flux balance was evaluated by integrating the approach of (1) and (2),

i.e. first computing the flux change of the modules containing the gene and then evaluating the change of other modules and flux balance of other metabolites.

Patient-derived cell line models of pancreatic cancer

Pa03C cells were obtained from Dr. Anirban Maitra's lab at The Johns Hopkins University[66]. All cells were maintained at 37°C in 5% CO2 and grown in DMEM (Invitrogen; Carlsbad, CA) with 10% Serum (Hyclone; Logan, UT). Cell line identity was confirmed by DNA fingerprint analysis (IDEXX BioResearch, Columbia, MO) for species and baseline short-tandem repeat analysis testing in February 2017. All cell lines were 100% human and a nine-marker short tandem repeat analysis is on file. They were also confirmed to be mycoplasma free.

ScRNA-seq experiment

Cells were transfected with either Scrambled (SCR) (5′ CCAUGAGGUCAGCAUGGUCUG 3′, 5′ GACCAUGCUGACCUCAUGGAA 3′) or siAPE1 (5′ GUCUGGUACGACUGGAGUACC 3′, 5′ UACUCCAGUCGUACCAGACCU 3′ siRNA). Briefly, 1×10^5 cells are plated per well of a 6-well plate and allowed to attach overnight. The next day, Lipofectamine RNAiMAX reagent (Invitrogen, Carlsbad, CA) was used to transfect in the APE1 and SCR siRNA at 20 nM following the manufacturer's indicated protocol. Opti-MEM, siRNA, and Lipofectamine was left on the cells for 16 h and then regular DMEM media with 10% Serum was added.

Three days post-transfection, SCR/siAPE1 cells were collected and loaded into 96-well microfluidic C1 Fluidigm array (Fluidigm, South San Francisco, CA, USA). All chambers were visually assessed and any chamber containing dead or multiple cells was excluded. The SMARTer system (Clontech, Mountain View, CA) was used to generate cDNA from captured single cells. The dscDNA quantity and quality was assessed using an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) with the High Sensitivity DNA Chip. The Purdue Genomics Facility prepared libraries using a Nextera kit (Illumina, San Diego, CA). Unstrained 2x100 bp reads were sequenced using the HiSeq2500 on rapid run mode in one lane.

ScRNA-seq data processing and analysis

FastQC was applied to evaluate the quality of the single cell RNA sequencing data. Counts were called for each cell sample by using STAR alignment pipeline against human GRCh38 reference genome. Cells with less than 250 or more than 10000 non-zero expressed genes were excluded from the analysis. Cells with more than 15% counts mapped to the mitochondrial genome were excluded as low quality cells, resulting 40 APEX-1 KD and 48 Control cells under hypoxia condition and 27 APEX-1 KD and 46 Control cells under normoxia condition for further analysis.

We utilized our in-house developed left truncated mixture Gaussian model to identify differentially expressed genes [49]. Pathway enrichment analysis of the genes in the identified bi-clusters are computed using hypergeometric test against the 1329 canonical pathway in MsigDB database [67], with $p<0.001$ as a significance cutoff.

qRT-PCR

qRT-PCR was used to measure the mRNA expression levels of the various genes identified from the scRNA-seq analysis. Following transfection, total RNA was extracted from cells using the Qiagen RNeasy Mini kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. First-strand cDNA was obtained from RNA using random hexamers and MultiScribe reverse transcriptase (Applied Biosystems, Foster City, CA). Quantitative PCR was performed using SYBR Green Real Time PCR master mix (Applied Biosystems, Foster City, CA) in a CFX96 Real Time detection system (Bio-Rad, Hercules, CA). The relative quantitative mRNA level was determined using the comparative Ct method using ribosomal protein L6 (RPL6) as the reference gene. Experiments were performed in triplicate for each sample. Statistical analysis performed using the $2^{-\Delta\Delta CT}$ method and analysis of covariance (ANCOVA) models, as previously published [68].

Metabolomics experiment and data analysis

We utilized the *MitoPlates Assay* from BiOLOG to measure relative abundance of 14 metabolites in central metabolic pathways, namely glucose, glucose-1 phosphate, glucose-6 phosphate, pyruvate, and lactate in the glycolysis

pathway, citrate, 2-oxoglutarate, succinate, fumarate, malate in the TCA cycle, and amino acids glutamate, glutamine, serine, and ornithine. See details in [69]. Metabolomics profiling of the 14 metabolites were collected from three replicates of bulk cell samples of wildtype Pa03c cells and APEX-1 inhibition under the normoxia condition.

ScRNA-seq data of head and neck cancer microenvironment

We collected melanoma and head and neck cancer scRNA-seq data from Gene Expression Omnibus (GEO) database, with accession ID GSE72056 and GSE103322. Basic QC for SC using the Seurat default parameter to filter out cells with high expressions of MT-coding genes. The cell type label and sample information provided in the original work were directly utilized. The GSE72056 data is collected on human melanoma tissues. The original paper provided cell classification and annotations including B cells, cancer-associated fibroblast (CAF) cells, endothelial cells, macrophage cells, malignant cells, NK cells, T cells, and unknown cells. The GSE103322 data is collected on head and neck cancer tissues. The original paper provided cell classification and annotations including B cells, dendritic cells, endothelial cells, fibroblast cells, macrophage cells, malignant cells, mast cells, myocyte cells, and T cells. Notably, as indicated by the original work, malignant cells have high intertumoral heterogeneity.

DATA ACCESS

The single cell sequencing data with matched metabolomic data collected on the Pa03C cells is being submitted to Gene Expression Omnibus, and is currently accessible via https://github.com/changwn/scFEA.

FUNDING SUPPORTS

This work was supported by R01 award #1R01GM131399- 01, NSF IIS (N0.1850360).

REFERENCES

1. Kochanek, K.D., et al., Deaths: final data for 2017. 2019.
2. Matsuzawa, Y., Therapy insight: adipocytokines in metabolic syndrome and related cardiovascular disease. Nature clinical practice Cardiovascular medicine, 2006. 3(1): p. 35-42.
3. Mattson, M.P. and S.L. Chan, Dysregulation of cellular calcium homeostasis in Alzheimer's disease. Journal of Molecular Neuroscience, 2001. 17(2): p. 205-224.
4. Dunn, L., et al., Dysregulation of glucose metabolism is an early event in sporadic Parkinson's disease. Neurobiology of aging, 2014. 35(5): p. 1111-1115.
5. Hirschey, M.D., et al. Dysregulated metabolism contributes to oncogenesis. in Seminars in cancer biology. 2015. Elsevier.
6. Rask, E., et al., Tissue-specific dysregulation of cortisol metabolism in human obesity. The Journal of clinical endocrinology & metabolism, 2001. 86(3): p. 1418-1421.
7. Sun, H., et al., Metabolic reprogramming in cancer is induced to increase proton production. Cancer Research, 2020. 80(5): p. 1143-1155.
8. Thompson, C., et al. How do cancer cells acquire the fuel needed to support cell growth? in Cold Spring Harbor symposia on quantitative biology. 2005. Cold Spring Harbor Laboratory Press.
9. DeBerardinis, R.J., et al., The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. Cell metabolism, 2008. 7(1): p. 11-20.
10. Hanahan, D. and R.A. Weinberg, Hallmarks of cancer: the next generation. cell, 2011. 144(5): p. 646-674.
11. Ward, P.S. and C.B. Thompson, Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer cell, 2012. 21(3): p. 297-308.
12. Lidstrom, M.E. and M.C. Konopka, The role of physiological heterogeneity in microbial population behavior. Nature chemical biology, 2010. 6(10): p. 705-712.
13. Bishop, A.L., et al., Phenotypic heterogeneity can enhance rare - cell survival in 'stress - sensitive' yeast populations. Molecular microbiology, 2007. 63(2): p. 507-520.
14. Zenobi, R., Single-cell metabolomics: analytical and biological perspectives. Science, 2013. 342(6163): p. 1243259.
15. Ahl, P.J., et al., Met-Flow, a strategy for single-cell metabolic analysis highlights dynamic changes in immune subpopulations. Communications Biology, 2020. 3(1): p. 305.
16. Ali, A., et al., Single-cell metabolomics by mass spectrometry: Advances, challenges, and future applications. TrAC Trends in Analytical Chemistry, 2019.
17. Duncan, K.D., J. Fyrestam, and I. Lanekoff, Advances in mass spectrometry based single-cell metabolomics. Analyst, 2019. 144(3): p. 782-793.
18. Emara, S., et al., Single-cell metabolomics, in Metabolomics: from fundamentals to clinical applications. 2017, Springer. p. 323-343.
19. Fessenden, M., Metabolomics: Small molecules, single cells. Nature, 2016. 540(7631): p. 153-155.
20. Zampieri, M., et al., Frontiers of high-throughput metabolomics. Curr Opin Chem Biol, 2017. 36: p. 15-23.
21. Kim, J. and R.J. DeBerardinis, Mechanisms and Implications of Metabolic Heterogeneity in Cancer. Cell Metab, 2019. 30(3): p. 434-446.
22. Robertson-Tessi, M., et al., Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. Cancer Res, 2015. 75(8): p. 1567-79.
23. Dunham, I., et al., An integrated encyclopedia of DNA elements in the human genome. Nature, 2012. 489(7414): p. 57-74.
24. Roadmap Epigenomics, C., et al., Integrative analysis of 111 reference human epigenomes. Nature, 2015. 518(7539): p. 317-30.
25. Feinberg, A.P., Phenotypic plasticity and the epigenetics of human disease. Nature, 2007. 447(7143): p. 433-440.
26. Harris, R.A., et al., Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nature Biotechnology, 2010. 28(10): p. 1097-U194.

27. Heintzman, N.D., et al., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics, 2007. 39(3): p. 311-318.
28. Jaenisch, R. and A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nature Genetics, 2003. 33: p. 245-254.
29. Honkoop, H., et al., Single-cell analysis uncovers that metabolic reprogramming by ErbB2 signaling is essential for cardiomyocyte proliferation in the regenerating heart. Elife, 2019. 8: p. e50163.
30. Xiao, Z., Z. Dai, and J.W. Locasale, Metabolic landscape of the tumor microenvironment at single cell resolution. Nature communications, 2019. 10(1): p. 1-12.
31. Vasdekis, A.E. and G. Stephanopoulos, Review of methods to probe single cell metabolism and bioenergetics. Metabolic engineering, 2015. 27: p. 115-135.
32. Saurty-Seerunghen, M.S., et al., Capture at the single cell level of metabolic modules distinguishing aggressive and indolent glioblastoma cells. Acta Neuropathologica Communications, 2019. 7(1): p. 1-16.
33. Xiao, Z., J.W. Locasale, and Z. Dai, Metabolism in the tumor microenvironment: insights from single-cell analysis. Oncoimmunology, 2020. 9(1): p. 1726556.
34. Evers, T.M., et al., Deciphering metabolic heterogeneity by single-cell analysis. 2019, ACS Publications.
35. Levine, L.S., et al., Single-cell metabolic dynamics of early activated CD8 T cells during the primary immune response to infection. bioRxiv, 2020: p. 2020.01.21.911545.
36. Rohlenova, K., et al., Single-Cell RNA Sequencing Maps Endothelial Metabolic Plasticity in Pathological Angiogenesis. Cell Metabolism, 2020. 31(4): p. 862-877. e14.
37. Zhang, Y., et al., Modeling metabolic variation with single-cell expression data. bioRxiv, 2020.
38. Damiani, C., et al., Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. PLoS computational biology, 2019. 15(2): p. e1006733.
39. Damiani, C., et al., Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. PLoS computational biology, 2019. 15(2): p. e1006733-e1006733.
40. Evers, T.M.J., et al., Deciphering Metabolic Heterogeneity by Single-Cell Analysis. Analytical Chemistry, 2019. 91(21): p. 13314-13323.
41. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000. 28(1): p. 27-30.
42. Saier, M.H., Jr., C.V. Tran, and R.D. Barabote, TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res, 2006. 34(Database issue): p. D181-6.
43. Wan, C., et al., LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. Nucleic Acids Res, 2019.
44. Schnell, S., Validity of the Michaelis-Menten equation--steady-state or reactant stationary assumption: that is the question. FEBS J, 2014. 281(2): p. 464-72.
45. Liu, Y., A. Beyer, and R. Aebersold, On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell, 2016. 165(3): p. 535-50.
46. Lan, X., et al. Efficient belief propagation with learned higher-order markov random fields. in European conference on computer vision. 2006. Springer.
47. Kelley, M.R., M.M. Georgiadis, and M.L. Fishel, APE1/Ref-1 role in redox signaling: translational applications of targeting the redox function of the DNA repair/redox protein APE1/Ref-1. Curr Mol Pharmacol, 2012. 5(1): p. 36-53.
48. Shah, F., et al., APE1/Ref-1 knockdown in pancreatic ductal adenocarcinoma–characterizing gene expression changes and identifying novel pathways using single-cell RNA sequencing. 2017. 11(12): p. 1711-1732.
49. Wan, C., et al., LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. Nucleic acids research, 2019. 47(18): p. e111-e111.
50. Xu, K., et al., A comparative study of gene-expression data of basal cell carcinoma and melanoma reveals new insights about the two cancers. PloS one, 2012. 7(1): p. e30750.
51. Hirayama, A., et al., Quantitative Metabolome Profiling of Colon and Stomach Cancer Microenvironment by Capillary Electrophoresis Time-of-Flight Mass Spectrometry. Cancer Research, 2009. 69(11): p. 4918-4925.

52. Cao, S., et al., Competition between DNA Methylation, Nucleotide Synthesis, and Antioxidation in Cancer versus Normal Tissues. Cancer Res, 2017. 77(15): p. 4185-4195.
53. Bhutia, Y.D., et al., SLC transporters as a novel class of tumour suppressors: identity, function and molecular mechanisms. The Biochemical journal, 2016. 473(9): p. 1113-1124.
54. Lin, L., et al., SLC transporters as therapeutic targets: emerging opportunities. Nature reviews. Drug discovery, 2015. 14(8): p. 543-560.
55. DeAngelis, P.L., J. Liu, and R.J. Linhardt, Chemoenzymatic synthesis of glycosaminoglycans: Re-creating, re-modeling and re-designing nature's longest or most complex carbohydrate chains. Glycobiology, 2013. 23(7): p. 764-777.
56. Gao, C. and K.J. Edgar, Efficient Synthesis of Glycosaminoglycan Analogs. Biomacromolecules, 2019. 20(2): p. 608-617.
57. Krasnova, L. and C.-H. Wong, Understanding the Chemistry and Biology of Glycosylation with Glycan Synthesis. 2016. 85(1): p. 599-630.
58. Lv, X., et al., Synthesis of Sialic Acids, Their Derivatives, and Analogs by Using a Whole-Cell Catalyst. Chemistry (Weinheim an der Bergstrasse, Germany), 2017. 23(60): p. 15143-15149.
59. Moffatt, B.A. and H. Ashihara, Purine and pyrimidine nucleotide synthesis and metabolism. The arabidopsis book, 2002. 1: p. e0018-e0018.
60. Zulueta, M.M., et al., Synthesis of glycosaminoglycans. 2016. p. 235-261.
61. Sun, H., et al., Metabolic Reprogramming in Cancer Is Induced to Increase Proton Production. Cancer Res, 2020. 80(5): p. 1143-1155.
62. Sun, H., et al., Fenton reactions drive nucleotide and ATP syntheses in cancer. J Mol Cell Biol, 2018. 10(5): p. 448-459.
63. Zhang, C., et al., Cancer may be a pathway to cell survival under persistent hypoxia and elevated ROS: a model for solid-cancer initiation and early development. Int J Cancer, 2015. 136(9): p. 2001-11.
64. Zhang, C., et al., Elucidation of drivers of high-level production of lactates throughout a cancer development. J Mol Cell Biol, 2015. 7(3): p. 267-79.
65. Yedidia, J.S., W.T. Freeman, and Y. Weiss. Generalized belief propagation. in Advances in neural information processing systems. 2001.
66. Jones, S., et al., Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. 2008.
67. Liberzon, A., et al., Molecular signatures database (MSigDB) 3.0. 2011. 27(12): p. 1739-1740.
68. Fishel, M.L., et al., Apurinic/apyrimidinic endonuclease/redox factor-1 (APE1/Ref-1) redox function negatively regulates NRF2. 2015. 290(5): p. 3057-3068.
69. Mitochondrial Function Assays with MitoPlates. 2020; https://www.biolog.com/products-portfolio-overview/mitochondrial-function-assays/].