Communications for Statistical Applications and Methods
2019, Vol. 26, No. 4, 411–430

https://doi.org/10.29220/CSAM.2019.26.4.411
Print ISSN 2287-7843 / Online ISSN 2383-4757

Iterative integrated imputation for missing data and pathway models with applications to breast cancer subtypes

Henry Linder[a], Yuping Zhang[1,a]

[a]Department of Statistics, University of Connecticut, USA

Abstract

Tumor development is driven by complex combinations of biological elements. Recent advances suggest that molecularly distinct subtypes of breast cancers may respond differently to pathway-targeted therapies. Thus, it is important to dissect pathway disturbances by integrating multiple molecular profiles, such as genetic, genomic and epigenomic data. However, missing data are often present in the -omic profiles of interest. Motivated by genomic data integration and imputation, we present a new statistical framework for pathway significance analysis. Specifically, we develop a new strategy for imputation of missing data in large-scale genomic studies, which adapts low-rank, structured matrix completion. Our iterative strategy enables us to impute missing data in complex configurations across multiple data platforms. In turn, we perform large-scale pathway analysis integrating gene expression, copy number, and methylation data. The advantages of the proposed statistical framework are demonstrated through simulations and real applications to breast cancer subtypes. We demonstrate superior power to identify pathway disturbances, compared with other imputation strategies. We also identify differential pathway activity across different breast tumor subtypes.

Keywords: imputation, multi-platform data integration, tumor subtypes, pathway analysis

1. Introduction

Large-scale genomic studies offer opportunities for comprehensive analysis of complex diseases at the molecular level. These studies produce rich datasets measured across hundreds of patients, permitting comparisons not only between healthy and diseased genomic profiles, but also to contrast different subtypes of a single umbrella disease. Integrated datasets observed on -omics data types beyond gene expression allow for an unprecedented level of detail.

To match the complexity and novelty of these datasets, statistical methods have been developed to model complex biological processes and to integrate multiple data types. Recent work has gone beyond simpler methods for statistical analysis, such as gene-set enrichment analysis (GSEA) (Subramanian *et al.*, 2005). Recent work emphasizes application of system-level models of biological processes. Pathway analysis incorporates knowledge of biological network structure into statistical models of expression. Pathway models, such as the NetGSA model (Shojaie and Michailidis, 2009) exploit results from graph theory to account for co-expression due to pass-through effects from features with differential activity to those that are not. The availability of -omics data platforms has also spurred research into integrative genomics. In their clinical study, Danielsen *et al.* (2015) integrated

[1] Corresponding author: Department of Statistics, University of Connecticut, Storrs, CT 06269, USA.
E-mail: yuping.zhang@uconn.edu

Published 31 July 2019 / journal homepage: http://csam.or.kr

the results of multiple separate analyses by data type in an ad-hoc, manual fashion. This approach is typical of the literature.

Systematic statistical models provide a solid theoretical foundation for integrative analysis. Much work has been done to identify common signals across genomic features and -omics data types. The iCluster model of Shen *et al.* (2009) applies PCA-style dimension reduction for cluster analysis. PARADIGM (Vaske *et al.*, 2010) uses network information to integrate -omics data via a latent factor model. DIG (Zhang *et al.*, 2017b) is a statistical framework for estimating networks integrating multiple -omics data types and biological conditions.

These methods are based on complete data that exhibit no missingness. In practice, datasets collected across dozens of clinical locations on thousands of genomic features contain substantial missing data. This limits the viability of many statistical methods, a problem that will only worsen as integrative studies become more common. Some work has begun to look at imputation of integrative datasets. Fryett *et al.* (2018) reviewed imputation methods for transcriptome data, highlighting in particular the FUSION model (Gusev *et al.*, 2016). That method integrates genotype and expression, followed by downstream analysis to identify phenotypic drivers. Several imputation schemes were also considered, including a variant of nearest neighbors imputation, linear predictors, and the Bayesian linear mixed models introduced by Zhou *et al.* (2013). Gamazon *et al.* (2015) employed a penalized regression framework to obtain similar linear predictions.

Schulz *et al.* (2017) performed integrative analysis of methylation and expression data in the human brain. For expression imputation, they use the IMPUTE model. IMPUTE (Howie *et al.*, 2009) uses hidden Markov models for simulation-based imputation of missing data. Chudasama *et al.* (2018) also applied IMPUTE for imputation, and performed an integrative analysis of expression and transcriptome observations in cancer.

Work on imputation for genomic pathway analysis is in the early stages. Some authors simply introduce methodological adjustments to work around missing data. Zhao *et al.* (2017) scored pathways and corrected for a deterministic impact of smaller sample sizes due to missing data on the pathway rank. Likewise, Köksal *et al.* (2018) appealed to parsimony by assuming that missing values have an insignificant effect on the overall analysis.

Other work uses predictive models for imputation: PAIRUP-MS (Hsu *et al.*, 2019) imputes metabolic data for pathway analysis using a linear regression model to predict unobserved biological signals. The Perseus -omics software (Tyanova *et al.*, 2016) uses simulation-based methods to obtain imputed values suitable for downstream analysis. Outside of the genomics literature, Krause *et al.* (2018) imputed longitudinal data for social networks analysis via simulation-based methods.

Dimension-reduction is also applied for imputation. Some methods based on the singular value decomposition expression imputation (Troyanskaya *et al.*, 2001). Mazumder *et al.* (2010) iteratively imputed missing data with a soft-thresholding algorithm. Tsuchiya *et al.* (2017) used reduced-rank methods to impute values in unequally-spaced gene expression time series.

In this paper, we address the widespread missingness in data from large-scale, multi-platform genomic studies. We propose a procedure for iterative integrated imputation (I3) that adapts the low-rank, structured matrix completion (SMC) method (Cai *et al.*, 2016). Whereas the original SMC method imputes rectangular sub-matrices for downstream statistical analysis of a complete dataset, our imputation is applicable to datasets with arbitrary configurations of missing data. After imputation, we then integrate gene expression (E), methylation (M) and copy number variations (C) to perform multi-modal network-based gene-set significance analysis based on the EMC-NetGSA model (Zhang *et al.*, 2017a).

Our approach enables analysis of a broader range of pathways than is possible without imputa-

tion. This allows analysis of biological processes using all data available, not just data that is well-formatted. In turn, this provides the opportunity to examine the organic systems that underlie complex diseases in finer detail and greater depth. We demonstrate both these perspectives with our data analysis: large-scale discovery among many pathways and granular analysis of specific pathways.

The paper proceeds as follows. We introduce a motivating breast cancer dataset from The Cancer Genome Atlas. We give details of our iterative imputation procedure for multi-platform -omics data, and give an overview of the EMC-NetGSA pathway model. We demonstrate our method's improved statistical power to detect pathway disturbances via a simulation study. Finally, we apply our strategy to analyze pathway disturbance across several breast tumor subtypes.

2. Methods

We constructed a multi-platform breast cancer dataset from data published by The Cancer Genome Atlas (TCGA) (Tomczak *et al.*, 2015). This long-running study collects observations of cancerous tissue in more than 30 cancers, on multiple -omics data platforms. For each of the project's cancer types, we acquired multi-platform -omics measurements of gene expression, copy number variation, and methylation for all tumor tissue samples, as well as control tissue samples collected from corresponding matched, healthy tissue.

The end goal is to perform integrative pathway analysis on all available samples, so we downloaded the NCI Pathway Interaction Database (PID) (Schaefer *et al.*, 2008), which contains 212 genomic signaling pathways defined across 2,393 genes. These pathways specify functional relationships between genes, and correspond to $q = 6,973$ -omics features in the TCGA dataset, after the processing steps discussed below in Section 3.2.

We format the -omics data as a $q \times N$ matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, where N is the total number of subjects. We denote by q_E the number of genes with observed mRNA expression in the dataset; q_C the number of genes with copy number observed; and q_M the methylation features. In principle, a full dataset for the NCI PID would contain $q_E = q_C = q_M = 2,393$ features, for a total of 7,179 features. However, in practice, we have $q_E \neq q_C \neq q_M$.

The data exhibit several patterns of missingness, particularly entire -omics platforms missing in all samples collected at specific sites, and individual -omics features with missing values that vary according to subject, typically missing because of data quality problems.

Structured matrix completion (SMC) addresses the first of these scenarios. The method builds on results for recovery of missing-at-random elements of low-rank matrices, for example, Candès and Tao (2010). The authors consider matrix $\mathbf{X}$ that is approximately rank r, in the sense that its r^{th}-largest singular value is much larger than the next; contains continuous elements; and the matrix of deviations $(\mathbf{X} - \mathbf{X}_r)$ is well-conditioned. Here, $\mathbf{X}_r$ is the rank-r singular value decomposition (SVD) approximation to $\mathbf{X}$ constructed using the first r singular dimensions.

Further, suppose $\mathbf{X}$ is a block matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{pmatrix}, \tag{2.1}$$

where $\mathbf{X}_{22}$ is entirely missing. We approximate $\mathbf{X}_{22}$ using the SVD. If $\mathbf{X}$ were exactly rank r, one could exactly recover $\mathbf{X}_{22}$ with the component $\mathbf{X}_{21}\mathbf{X}_{11}^{-}\mathbf{X}_{12}$ of the Schur complement of $\mathbf{X}$, where $\mathbf{X}_{11}^{-}$ is the Moore-Penrose pseudo-inverse. For known r but $\mathbf{X}$ *approximately* rank r, we may recover $\mathbf{X}_{22}$ using the rank-r SVD approximation to $\mathbf{X}_{11}, \mathbf{X}_{12}, \mathbf{X}_{21}$. Finally, we choose $\hat{r}$ as the largest value of r for which the SVD approximation to $\mathbf{X}_{11}$ is non-singular, and the approximation to $\mathbf{X}_{22}$ is well-conditioned.

This block approach works when the missing data in $\mathbf{X}$ has clean margins, i.e., the same features are either all missing or all observed in all samples. However, this is not typically true of real-world datasets, in which some subjects are naturally missing a small number of non-overlapping features. A direct solution is to form $\mathbf{X}_{22}$ as be the minimal covering submatrix for all missing values. When the number of observed values in $\mathbf{X}_{22}$ is small, we may justify discarding small portions of data prior to imputation, especially when the entire dataset at large is of interest, as opposed to individual features.

In real-world data matrices, however, the missingness in $\mathbf{X}_{22}$ is sparse. In our TCGA dataset, for example, less than 5% of its elements are missing. The dimensions of $\mathbf{X}_{22}$ must extend to cover all features with missing values, even those that are observed in most samples. In the pathway analysis framework, $\mathbf{X}$ is a composite matrix of data generated from multiple signaling pathways. It is thus important to maximize the amount of information we use when imputing each given feature, so that we do not discard information of importance to a specific subset of pathway features of interest.

Also of critical importance is the assumption that $\mathbf{X}_{12}$ is non-degenerate, i.e., has a positive (nonzero) number of rows. This assumption requires the presence of at least *some* features observed in all samples, which is not generally guaranteed in -omics studies. We demonstrate a toy example, below, that illustrates this case.

To address these limitations of SMC, we perform an iterative imputation independently across each sample. In the notation of (2.1), and without loss of generality, $\mathbf{X}_1$ is $q \times N_1$, $N_1 < N$, and contains no missing values; and $\mathbf{X}_2$ is $q \times N_2$. Here, N_1 is the number of samples with complete observations for all q genomic features, while N_2 is the number of samples with any missing values, $N_1 + N_2 = N$. Suppose further that $q_1 \times N_2$ matrix $\mathbf{X}_{12}$ contains no missing values, and $\mathbf{X}_{22}$ is the $q_2 \times N_2$ minimal covering submatrix for all missing values in $\mathbf{X}$, $q_1 + q_2 = q$. $q_1 \geq 0$ is the number of features with no missing values, and may equal to zero, in which case every row of matrix $\mathbf{X}_2$ contains at least one missing value, so that $\mathbf{X}_2 \equiv \mathbf{X}_{22}$.

We may then consider individual columns $\mathbf{x}_i$ of $\mathbf{X}_2$, $i = N_1 + 1, \ldots, N$, each of which is a sample with complete observations on at least q_1 features. Of the elements of $\mathbf{x}_i$, we denote by $q_i \leq q_2$ the number of missing features.

If most of the values of $\mathbf{X}_{22}$ were missing, we might apply SMC directly, replacing the imputed values with observed elements of $\mathbf{X}_{22}$, where available. However, $\mathbf{X}_{22}$ is often quite dense: in our data, more than 95% of the entries $\mathbf{X}_2$ are observed, so imputing all elements of $\mathbf{X}_{22}$ with SMC risks losing possibly considerable structural information in its many nonzero elements.

We apply SMC iteratively and independently to each of the samples in columns of $\mathbf{X}_2$. Sample $\mathbf{x}_i$ is a vector of q elements, $i = N_1 + 1, \ldots, N$. By invariance of the singular values of $\mathbf{X}$ under row and column permutations, we may suppose without loss of generality that the first $(q - q_i)$ elements of $\mathbf{x}_i$ are entirely observed, so that the matrix $\begin{pmatrix}\mathbf{X}_1 & \mathbf{x}_i\end{pmatrix}$ is entirely observed except for a missing $q_i \times 1$ submatrix in its lower-right corner.

Turning our attention to the remaining columns of $\mathbf{X}_2$, we identify all $0 \leq k_i \leq N_2 - 1$ other $\mathbf{x}_{i'}$ with complete observed values for the $(q - q_i)$ features observed in $\mathbf{x}_i$. Denoting the index set for these vectors by $\{\omega_{in}\}_{n=1}^{k_i}$, we form the matrix

$$\chi_i = \begin{pmatrix}\mathbf{X}_1 & \mathbf{x}_i & \mathbf{x}_{\omega_{i1}} & \cdots & \mathbf{x}_{\omega_{ik_i}}\end{pmatrix} \tag{2.2}$$

χ_i is $q \times (N_1 + k_i + 1)$ matrix, and we treat the entire lower-right $q_i \times (k_i + 1)$ submatrix as missing. The first column of that submatrix, which corresponds to the missing elements in $\mathbf{x}_i$, is entirely missing. Subsequent columns come from $\mathbf{X}_2$ and are missing a subset, possibly improper, of the same missing values in $\mathbf{x}_i$. We impute this corner matrix using SMC and use the first column to form imputed vector $\hat{\mathbf{x}}_i$.

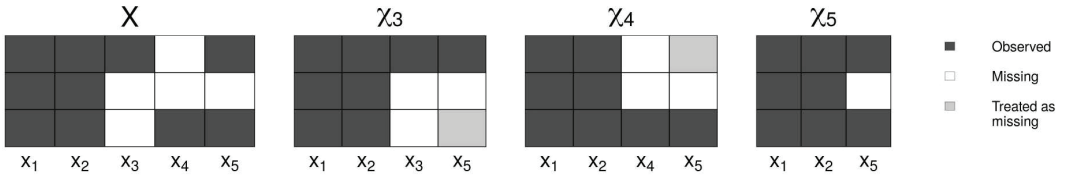

Figure 1: *Iterative integrated imputation with a toy example.* $\mathbf{X}$ *is a data matrix of 5 observation vectors with missing values that do not form a rectangle. Iteratively for each column with missing elements* $\mathbf{x}_i$ *(*$i = 3, 4, 5$*), we apply structured matrix completion to the matrix* $\boldsymbol{\chi}_i$*, consisting of all complete data vectors;* $\mathbf{x}_i$*; and any columns with missing values that are a subset of those missing from* $\mathbf{x}_i$*. Imputation is performed on the entire rectangular submatrix of all features missing in* $\mathbf{x}_i$*, with any observed elements not from* $\mathbf{x}_i$ *treated as missing. So, imputing* $\mathbf{x}_3$ *and* $\mathbf{x}_4$ *uses* $\mathbf{x}_5$*, whereas* $\mathbf{x}_5$ *is imputed using only the complete data vectors.*

We repeat this procedure for all $i = 1, \ldots, N_2$, at the end of which, we have an imputed matrix $\widehat{\mathbf{X}}_2 = \begin{pmatrix} \widehat{\mathbf{x}}_1 & \cdots & \widehat{\mathbf{x}}_N \end{pmatrix}$ which contains no missing values. In turn, this yields an imputed matrix $\widehat{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_1 & \widehat{\mathbf{X}}_2 \end{pmatrix}$.

As a toy example to demonstrate the procedure, we consider a matrix with elements that equal to 1, or are missing (denoted " - "):

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 & \mathbf{x}_5 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & - & 1 \\ 1 & 1 & - & - & - \\ 1 & 1 & - & 1 & 1 \end{pmatrix}. \tag{2.3}$$

We observe that, as originally proposed, SMC is unable to directly impute any of the missing values in $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$, because none of the rows of $\mathbf{X}$ is completely observed.

Figure 1 displays the configuration of $\mathbf{X}$, as well as the corresponding $\boldsymbol{\chi}_i$, $i = 3, 4, 5$. $\mathbf{x}_5$ contains a single missing value, which is a subset of the missing elements in $\mathbf{x}_3$, whereas $\mathbf{x}_4$ is missing that element. So, $k_3 = 1$, and we impute the missing values in $\mathbf{x}_3$, using the element in $\mathbf{x}_5$ corresponding to the observed value in $\mathbf{x}_3$; but we ignore the observed element in $\mathbf{x}_5$ that is missing in $\mathbf{x}_3$ and $\mathbf{x}_4$, so $k_5 = 0$. Implicitly, we impute this value of $\mathbf{x}_5$, as well, but we discard this value, and keep only the imputed values for $\mathbf{x}_3$. A symmetric argument applies to $\mathbf{x}_4$.

On the other hand, because both $\mathbf{x}_3$ and $\mathbf{x}_4$ are missing elements that are observed in $\mathbf{x}_5$, we cannot use either of these in imputing the missing value in $\mathbf{x}_5$.

An appealing property of our iterative imputation procedure arises when the missing data has a block structure, as in the original SMC setting. In that case, iterative imputation produces identical imputed values to those from SMC. Moreover, as demonstrated in our toy example, we may apply our procedure to matrices for which $q_1 = 0$, that is, no features are fully observed across all subjects. In contrast, SMC cannot be used to impute that type of matrix.

After imputation of the full dataset, we turn our attention to downstream analysis of the dataset. Between the q_E genes in $\mathbf{X}$ are a known set of genomic signaling pathways. These pathways specify directed functional relationships between sets of genes, where the relationships represent known biological interactions between genes. In particular, for a specific pathway we consider a set of p genes, $p \leq q_E$. We define the graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is a set of p vertices (genes), and $\mathcal{E}$ is a set of directed edges between the elements of $\mathcal{V}$ (functional relationships).

We represent $\mathcal{G}$ by the $p \times p$ graph adjacency matrix $\mathbf{A}^\star = \{\alpha_{jk}\}_{j=1,\ldots,p;\ k=1,\ldots,p}$, the elements of which indicate conditional dependence of gene j on gene k, that is, $\alpha_{jk} = \mathbb{1}(\exists$ directed edge from gene k to gene $j)$. For a given pathway pathway, we observe a vector $\mathbf{y}_{i1}$ of p gene expression values,

$i = 1, \ldots, N$. $\mathbf{y}_{i1}$ corresponds to the column vector $\mathbf{x}_i$ of $\mathbf{X}$, with its elements comprising a subset of the elements of $\mathbf{x}_i$.

For each element in $\mathbf{y}_{i1}$, we also observe copy number and methylation beta values for the corresponding genes. Denote the vectors of copy number and methylation by $\mathbf{y}_{i2}$ and $\mathbf{y}_{i3}$, respectively. We use the EMC-NetGSA to integrate methylation and copy number into the graph topology. We augment $\mathcal{G}$ by adding $2p$ vertices to $\mathcal{V}$, one each for copy number and methylation for every gene. We also add $2p$ edges to $\mathcal{E}$, with directed from the copy number or methylation vertex to the corresponding gene expression vertex. Therefore, the final EMC-NetGSA graph consists of two separate layers of network topology:

1. The primary, inter-gene signaling pathway network, giving edges between elements of $\mathbf{y}_{i1}$;
2. A secondary integration network of edges between elements of $\mathbf{y}_{i2}$ or $\mathbf{y}_{i3}$, and the corresponding elements in $\mathbf{y}_{i1}$.

Denote the full observation vector for sample i by $\mathbf{y}_i = (\mathbf{y}_{i1}', \mathbf{y}_{i2}', \mathbf{y}_{i3}')'$. Without loss of generality, we assume $\mathbf{y}_{i2}$ and $\mathbf{y}_{i3}$ each contain p elements. As discussed above, some genes' copy number or methylation values are missing across all subjects, and so cannot be imputed. In these cases, the number of copy number (methylation) observations in $\mathbf{y}_{i2}$ ($\mathbf{y}_{i3}$) will be less than p. We resolve this by removing the non-expression vertex from $\mathcal{V}$ and the corresponding integrated edge from $\mathcal{E}$.

The EMC-NetGSA integration is implemented in the adjacency matrix by concatenating identity matrices along the right-hand margin and zero matrices along the bottom:

$$\mathbf{A}^{\star}_{\text{EMC}} = \begin{pmatrix} \mathbf{A}^{\star}_{\text{E}} & \mathbf{I}_{p\times p} & \mathbf{I}_{p\times p} \\ \mathbf{O}_{2p\times p} & \mathbf{O}_{2p\times p} & \mathbf{O}_{2p\times p} \end{pmatrix}, \tag{2.4}$$

where $\mathbf{O}_{2p\times p}$ is a $2p \times p$ matrix of zeros. Here, the identity matrices reflect the directed edges from copy number and methylation vertices to expression, and the zero matrices reflect the lack of any network structure within or between the additional data types, or a directed relationship from genes to -omics features.

Each pair of vertices in $\{(j,k) \mid \alpha_{jk} \neq 0\}$ specifies a directed edge from y_{ik} to y_{ij}, $i = 1, \ldots, N$. This is a conditional dependence relation of vertex j on vertex k, given the effects of the other $(3p-2)$ network features. In the context of Gaussian graphical models, conditional dependence of random variables X_j and X_k, conditional on a set $\mathcal{Z}$ of additional random variables of interest, is formalized as the partial correlation ρ_{jk} with respect to $\mathcal{Z}$. That is, $\rho_{jk} = \text{corr}(X_{j\setminus\mathcal{Z}}, X_{k\setminus\mathcal{Z}})$, where $X_{j\setminus\mathcal{Z}} = X_j - \mathcal{P}_{\mathcal{Z}} X_j$ is the orthogonal complement of X_j with respect to $\mathcal{Z}$, and $\mathcal{P}_{\mathcal{Z}}$ is a projection onto $\mathcal{Z}$ (Krämer *et al.*, 2009). Intuitively, ρ_{jk} represents the association between -omics features j and k, controlling for each of their associations with other features in the pathway. We estimate ρ_{jk} by the sample partial correlation r_{jk}, which is obtained by regressing each of X_j and X_k separately on $\mathcal{Z}$, and then calculating Pearson's correlation coefficient between the vectors of residuals. Using the partial correlations and $\mathbf{A}^{\star}$, we construct the weighted adjacency matrix $\mathbf{A}$ with elements $a_{jk} = r_{jk}\alpha_{jk}$, $j, k = 1, \ldots, p$.

To build a statistical model for the data $\{\mathbf{y}_i\}_{i=1}^N$, we consider a transformation of $\mathbf{A}$ introduced by Shojaie and Michailidis (2009). The influence matrix $\mathbf{\Lambda}$ captures the cumulative network effect of each gene on the expression of all others. In the case of directed acyclic graphs (DAGs), the authors derive the identity $\mathbf{\Lambda} = (\mathbf{I}_{3p} - \mathbf{A})^{-1}$. Shojaie and Michailidis (2010) extended this formula to apply to general, non-DAG graphs.

NetGSA uses $\mathbf{\Lambda}$ to structure the mean in a mixed-effects model with unknown regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, with $\mathbb{E}\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\beta}$. We may interpret $\boldsymbol{\beta}$ as the network-adjusted expression coefficients for the $3p$ -omics features, and we note $\mathbf{\Lambda}$ also structures the covariance of $\mathbf{y}_i$.

NetGSA offers a significance testing framework to compare two populations, control (healthy) and treatment (disease). Denote the population label for sample i as $c_i \in \{\mathrm{C}, \mathrm{T}\}$. The control and treatment populations are modeled using separate adjacency matrices $\mathbf{A}_\mathrm{C}$, $\mathbf{A}_\mathrm{T}$, corresponding to influence matrices $\mathbf{\Lambda}_\mathrm{C}$, $\mathbf{\Lambda}_\mathrm{T}$, and parameterized with population-specific pathway-adjusted expression parameters $\boldsymbol{\beta}_\mathrm{C}, \boldsymbol{\beta}_\mathrm{T}$, respectively. The statistical model is

$$\begin{aligned} \mathbf{y}_i &= \mathbf{\Lambda}_{c_i}\boldsymbol{\beta}_{c_i} + \mathbf{\Lambda}_{c_i}\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N, \\ \boldsymbol{\gamma}_i &\sim N_{3p}\left(\mathbf{0}_{3p}, \sigma^2_\gamma \mathbf{I}_{3p}\right), \\ \boldsymbol{\epsilon}_i &\sim N_{3p}\left(\mathbf{0}_{3p}, \sigma^2_\epsilon \mathbf{I}_{3p}\right). \end{aligned} \tag{2.5}$$

To test differential activity in subsets of the pathway's genomic features, represented by elements of $\boldsymbol{\beta}_\mathrm{C}$ and $\boldsymbol{\beta}_\mathrm{T}$, we specify features of interest through an indicator vector $\mathbf{b}$. We use the NetGSA network contrast $\boldsymbol{\ell} = (-\mathbf{b} \cdot \mathbf{b}\mathbf{\Lambda}_\mathrm{C}, \mathbf{b} \cdot \mathbf{b}\mathbf{\Lambda}_\mathrm{T})$ to compute the test statistic $T \propto \ell\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathrm{C}'}, \boldsymbol{\beta}_{\mathrm{T}'})'$. T follows a Student's t distribution with degrees of freedom estimated by the Satterthwaite approximation.

3. Results

Applied to large-scale datasets, iterative imputation produces a complete matrix $\widehat{\mathbf{X}}$, using which we perform downstream analysis of pathway disturbance. We explore the power of our two-stage imputation and pathway analysis using a simulation. Then, we apply the method to breast cancer data to analyze pathway disturbances by tumor subtype.

3.1. Simulations

We follow a similar simulation design to that of Zhang *et al.*, (2017a). For the treatment group, we construct a five-level binary tree. The control group network derives from the treatment network, with all edges in the left branch (including the root) removed. Correlation between expression vertices is set to 0.8 in the tree's top third (two levels); the middle third (third level) has association 0.5; and the lower third (last level) has association 0.2. We also add vertices for methylation and copy number, integrated via directed edges to the expression vertex. We set correlation of expression with copy number to 0.5, and correlation with methylation to −0.25.

We generate integrated data vectors $\mathbf{y}_i$, $i = 1, \ldots, N$, from the NetGSA model in Equation 2.5, where $N = N_\mathrm{T} + N_\mathrm{C}$. N_T (N_C) is the number of treatment (control) samples. We set $\sigma^2_\gamma = 5$, $\sigma^2_\epsilon = 0.5$, and $N_\mathrm{C} = 50$ and $N_\mathrm{T} = 150$. We consider four gene sets:

1. All genes in the network;
2. Top one-third levels of the tree;
3. First two-thirds levels of the tree;
4. The last level of the tree.

Denote by $\boldsymbol{\beta}_{\mathrm{C}1}, \boldsymbol{\beta}_{\mathrm{C}2}, \boldsymbol{\beta}_{\mathrm{C}3}$, the mean vectors of gene expression, copy number, and methylation, respectively, in the control population. The treatment population is defined analogously. We simulate two scenarios for the network-adjusted mean coefficient $\boldsymbol{\beta}$:

1. $\boldsymbol{\beta}_{\mathrm{T}1} = \boldsymbol{\beta}_{\mathrm{C}1} = \boldsymbol{\beta}_{\mathrm{T}2} = \boldsymbol{\beta}_{\mathrm{C}2} = \boldsymbol{\beta}_{\mathrm{T}3} = \boldsymbol{\beta}_{\mathrm{C}3} = \mathbf{0}$;
2. $\boldsymbol{\beta}_{\mathrm{T}1} = 0.25\mathbf{1}, \beta_{\mathrm{T}2} = 1\mathbf{1}, \beta_{\mathrm{T}3} = 0.5\mathbf{1}$ for top two-thirds levels, otherwise same as first scenario.

The TCGA dataset, introduced in Section 2 and analyzed below, displays missingness of two types. The first is block-wise missing data in samples collected at specific research sites for entire classes of measurements. This type of missing data motivated the original SMC method, and as noted previously, our iterative imputation is identical to SMC in the special case that the missingness in $\mathbf{X}$ forms a block.

The second type of missingness varies the specific features that are missing within any single sample. In this case, the SMC procedure cannot simultaneously impute all missing values *and* use all observed data points. We note that the missingness does not alter the low-rank structure of the matrix. Therefore, we designed our simulation to mimic this variety of missingness, namely, where the features with missing values vary by sample, but we also include block-wise missingness, as well. This allows us to compare the novel aspect of our iterative imputation with the performance of the original formulation of SMC.

Within each simulation replicate, we generated a full data matrix, $\mathbf{X}$. We randomly selected subsets of features and subjects to exhibit missing data, giving a data matrix with missing values. Additionally, we removed a full rectangular submatrix, which reflects the composition of missing data that is sometimes observed. We removed all samples with missing observations to form $\tilde{\mathbf{X}}$. Finally, we iteratively imputed the missing values to obtain $\widehat{\mathbf{X}}$. For comparison, we also imputed the matrix using SMC on the covering submatrix $\mathbf{X}_{22}$, as well as the K-nearest neighbors (KNN) method taking median among $K = 10$ nearest neighbors, as implemented in `bnstruct` (Franzin *et al.*, 2017).

We performed 1,000 simulation replicates for each mean scenario. We performed NetGSA on all five data matrices for both mean scenarios, and we calculated the power for each test by checking the significance of the Benjamini-Hochberg (BH) adjusted p-value (Benjamini and Hochberg, 1995), at the $\alpha = 0.05$ level. The power is then the proportion of simulations in which the null hypothesis of no difference is rejected.

Our combination of mean scenarios and gene sets provides simulations in which all genes in a gene set are differential; some (but not all) genes are differential; and none of the genes are differential. Moreover, our simulation scenario includes missing data that is characteristic of the real data, with a composition of data that is missing in block form, as well as data that is missing on a by-feature basis within individual samples. This simulation setting more accurately characterizes the real-world missing data, but this type of by-feature missingness was not considered by the authors of SMC.

The left panel of Figure 2 gives boxplots of $-\log_{10}(p)$ for the p-values in the first mean scenario, in which the expression coefficients are equal across the two populations. The power of the pathway analysis that uses our imputed matrix $\widehat{\mathbf{X}}$ is comparable to that of the true data. Our procedure corresponds to a minor increase in the type I error rate over the true data. In comparison, direct application of SMC to the minimal covering submatrix for the missing values results in a higher false positive rate.

The right panel of Figure 2 displays simulation significance results from the second mean scenario, in which the top 2/3 levels of nodes in the binary tree are differentially expressed. Our iterative integrated imputation procedure exhibits power comparable to that of the true data, and consistently dominates use of the truncated data, $\tilde{\mathbf{X}}$. Application of SMC to a rectangular submatrix is similar, but with a higher rate of false positives. In turn, this results in power that is higher than the true data, because of the inflated propensity to reject the null hypothesis.

KNN is comparable to our iterative imputation, but with weaker power. This relative performance of KNN to our method is robust to a range of K, both large and small. A benefit of our method is that the imputed values do not change the outcome of inference, as compared with the true data. The KNN imputation uses real observed values from other observations that are "close." In contrast, the

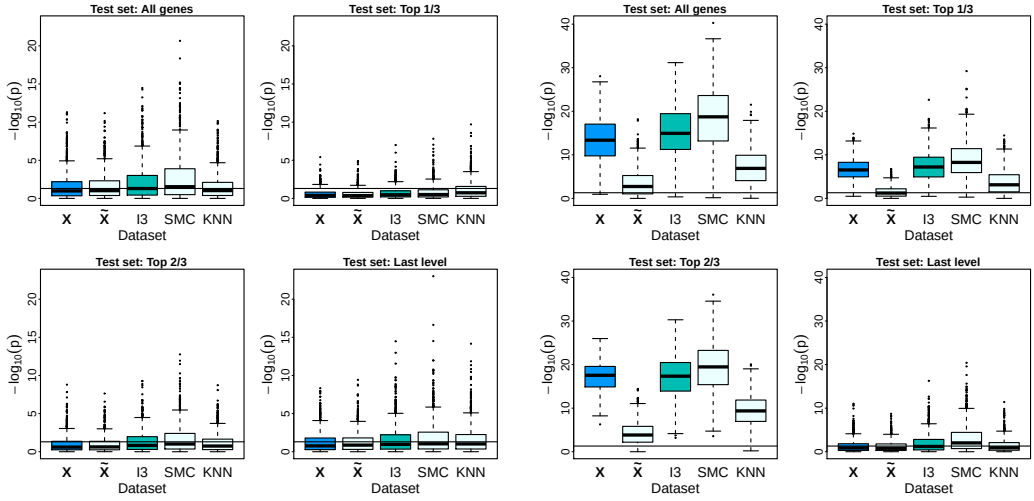

Figure 2: *Boxplots of* $-\log_{10}(p)$*-values for simulation study tests of pathway disturbance. Network structure is a binary tree.* $\mathbf{X}$ *(shown in blue) represents EMC-NetGSA applied to the complete data matrix, with no missing values.* $\tilde{\mathbf{X}}$ *is the matrix with all samples with any missing values dropped. I3 (in green) refers to our integrated, iteratively imputed matrix. SMC uses imputed values obtained by treating the entire minimal covering submatrix as missing, and KNN is* $K = 10$*-nearest neighbors imputation. Horizontal line gives* $-\log_{10}(0.05)$. Scenario 1 (left): *No features differential.* Scenario 2 (right): *Top 2/3 features differential.*

SVD-based methods construct a (linear) functional model for the data, and can thus directly predict the missing values.

The direct application of SMC to the minimal covering submatrix of the missing data in $\mathbf{X}$ is similar to our iterative method, but with a higher false positive rate. The elevated false discovery rate reflects that SMC accentuates statistical noise, thereby reinforcing and strengthening spurious deviations due to statistical variation. SMC propagates dominant low-rank structure in $\mathbf{X}$. But, discarding the information contained in samples with missing values results in an over-emphasis of features that are not reflective of the overall data matrix. This occurs even when imputing across a data matrix for which the mean structure is the same across all subjects, as in the first mean scenario.

Our imputation leverages the imputation using linear dependence due to the underlying SMC method, while also maximizing the available information about each sample. This provides a more nuanced and complete picture of the linear structure of the network data. For pathway analysis, it is particularly important that any imputed values be locally accurate. The features in the TCGA data matrix $\mathbf{X}$, discussed below, encompass hundreds of different pathways, so the matrix-wide contamination we observed in the SMC simulations poses a serious obstacle to valid statistical inference.

These results are robust to variations in the combination of block-wise and at-random missingness. The balance between these two types may vary by dataset: for instance, the original SMC method was designed for block-missing data from the TCGA study, but our data analysis indicates that at-random missing values are also widespread within individual subjects. As the missingness progresses from at-random to block-wise, we find that the performance of SMC improves relative to the iterative imputation. Equivalent power is achieved when the missing data is fully rectangular, which reflects that our method gives the same result as SMC for this edge case. This power is shown in Figure 3. However, when the data composes both block-wise and at-random missingness, our method offers

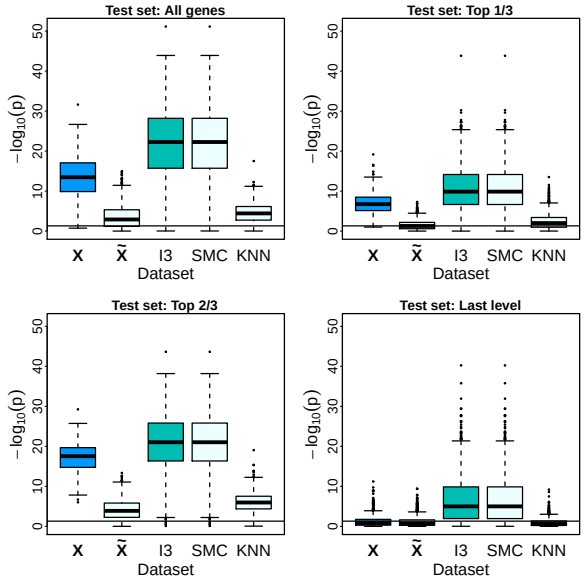

Figure 3: *Simulation results for missing data in a rectangular submatrix, using the second mean scenario in which the top 2/3 of binary tree nodes are differentially expressed. In this special case, our iterative imputation method (I3, in green) has identical power to that of the basic SMC method.*

improved control of false discoveries, without a cost in terms of power.

Finally, we note that real-world data matrices, as in the TCGA dataset we analyze, the features in $\mathbf{X}$ may be drawn from many different pathways, some of which may exhibit disturbance in tumor samples, but others of which may have equal means in the sample populations. This is reflected in the fourth test set in the second simulation scenario. The genes of interest in the test set are not differentially expressed, although other features in the data matrix do feature different means. Our iterative method identifies the differential structure in the regions of the data matrix, without a corresponding increase in false positives. If the minimal covering submatrix for the missing values contains all features that are differentially expressed, in the samples with the pathway disturbance, the original SMC method will fail to reveal any pathway disturbance. On the other hand, our method will use the partial information available in each sample, thereby retaining the differential structure in the imputed values it returns.

3.2. Data analysis

To demonstrate the superiority of the imputed data in pathway analysis, compared with the truncated data, we performed an analysis of pathway disturbances in breast tumor subtypes.

We downloaded the TCGA data from the NCI Genomic Data Commons (GDC) (Grossman *et al.*, 2016), and using the TCGA-Assembler software, v2.0.0 (Zhu *et al.*, 2014; Wei *et al.*, 2017).

Our dataset is similar to that of the original SMC paper. We acquired level 3 TCGA data for all breast cancer tumors and matched control samples, measured on three -omics platforms: (1) gene expression (RNASeqV2); (2) copy number variation (CNV, germline CNVs omitted); (3) methylation (HumanMethylation450 BeadChip). We used normalized RNASeq read counts, and applied a $\log_2$ transformation. We averaged gene-level CNV by DNA region, and we averaged methylation beta

values across CpG sites for each gene.

Our TCGA dataset contained 838 tissue samples. After downloading the NCI Pathway Interaction Database (PID) using the `graphite` software (Sales *et al.*, 2018), which spans 2,393 genes across 212 pathways, the matrix $\mathbf{X}$ had $q = 6,973$ rows (genomic features), observed on $N = 838$ tissue samples.

It might be possible to impute the tumor and healthy sample populations as separate matrices. However, a major limitation of the TCGA datasets is imbalanced sample sizes, relative to the large number of features. This could result in sensitivity of the imputed values in the control population, and may be unnecessary. In principle, healthy and tumorous tissues should share many structural characteristics, namely, all non-malignant signaling activity. Therefore, we impute both populations jointly.

Of the 838 samples, $N_2 = 719$ samples contained missing values (681 tumor, 38 healthy), spread across $q_2 = 171$ features. 157 features measured copy number, and the other 14 measured methylation. The values k_i, giving the number of additional columns of $\mathbf{X}_2$ used in forming χ_i, were generally quite high: only 6 columns used fewer than 50 other samples for imputation, or less than 1% of the total number of columns we imputed. Our imputation procedure uses a substantial amount of the non-missing information contained in $\mathbf{X}_{22}$ when forming χ_i, despite our separate imputation of each column of $\mathbf{X}_2$.

Several distinct breast tumor subtypes are defined in terms of the immunohistochemistry characteristics of specific genes (Dai *et al.*, 2015). These subtypes are strongly associated with tumor grade, clinical outcome, and overall subtype prevalence. It is therefore critical to understand the distinctions between separate tumor types, which may operate under substantially different biological mechanisms.

A key assumption of the SMC imputation method is that the data matrix is approximately low rank. Conceptually, the data satisfy the low-rank assumption: within each population, conditional on the subject-level random effects, the mean matrix is simply noisy observations of a rank 1 matrix. We verified empirically the low-rank structure of the submatrix $\mathbf{X}_1$, i.e., all rows with no missing values.

To determine the breast tumor intrinsic subtype, we followed the classification given in Dai *et al.* (2015). These subtypes are defined according to immunohistochemistry (IHC) status of several specific genes: ER, PR, HER2, and KI67. The TCGA dataset contains the IHC status for ER, PR, and HER2, but not KI67, so we employ four tumor subtypes: luminal A / HER2- luminal B; HER2+ luminal B; HER2 over-expression; and basal, or triple negative. The specific immunohistochemical criteria used is given in Table 1. We categorized tumors described as "normal-like" as luminal A / HER2- luminal B, since they cannot be distinguished on the basis of ER, PR, and HER2 IHC status.

Although we were unable to distinguish tumors on the basis of KI67, this reflected practical considerations that apply beyond the current study. Indeed, Cheang *et al.* (2009) note that "Ki67 is not included in routine clinical decision-making because of a lack of clarity regarding how Ki67 measurements should influence clinical decisions." In the BRCA dataset, 479 samples had IHC status available for ER, PR, and HER. Frequency and prevalence of the subtypes is given in Table 1, and clearly matches that in Dai *et al.* (2015).

We applied EMC-NetGSA to the raw and imputed datasets for 176 of the 212 NCI pathways. In the remaining 36 pathways, our dataset was missing expression observations across all subjects for at least one gene, so we were unable to analyze these pathways. We applied the BH adjustment to the pathway-level p-values.

We ranked the pathways by the variance between the $-\log_{10}(p)$-values in the 4 subtypes. We considered the five pathways that exhibited the highest variability, which reflects major differences in the disturbance of these pathways between subtypes. These most-variable pathways were (1) EPHB

Table 1: Breast tumor subtype definitions and prevalence in data from The Cancer Genome Atlas (TCGA). IHC status refers to immunohistochemical status for the three specified genes. Prevalence measures a proportion.

Subtype	**Luminal A / HER2- luminal B**	**HER2+ Luminal B**	**HER2 over-expression**	**Basal**
IHC status	ER+ or PR+, HER2-	ER+ or PR+, HER2+	ER-, PR-, HER2+	ER-, PR-, HER2-
Tumor subtype population characteristics from Dai *et al.* (2015)				
Prevalence	0.62	0.14	0.11	0.12
Tumor grade	1, 2, 3	2, 3	2, 3	3
Outcome	Good, Intermediate	Intermediate, Poor	Poor	Poor
TCGA breast cancer tumor sample population				
Count	307	74	17	81
Prevalence	0.64	0.15	0.04	0.17

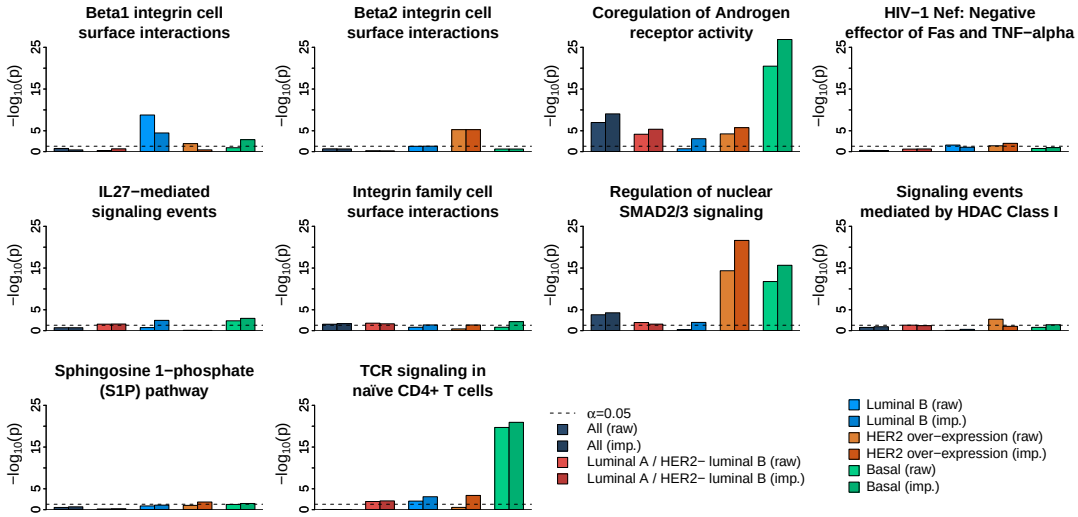

Figure 4: $-\log_{10}(p)$-*values for integrative tests of network disturbance in genetic signaling pathways. Shown are pathways that changed significance after imputation in at least one sample population. Separate differential analyses were performed for each of four breast cancer subtypes, as well as a composite population of all tumors.*

forward signaling, (2) FOXA2 and FOXA3 transcription factor networks, (3) ErbB2/ErbB3 signaling events, (4) Validated nuclear estrogen receptor alpha network, and (5) E2F transcription factor network. The nuclear estrogen receptor network contains known oncogenes and plays an established role in breast cancer (Sommer and Fuqua, 2001), as does EPHB (Pasquale, 2010), (Kaenel *et al.*, 2012). But while the role of E2F is established in cancers such as retinal cancer (Nevins, 2001), its possible role in breast cancer has only been explored recently (Johnson *et al.*, 2016). This is also true of ErbB2/ErbB3 (Ma *et al.*, 2014), and while FOXA1 is implicated in breast cancer (Bochkis *et al.*, 2012), this is not true of FOXA2 and FOXA3.

We identified the pathways for which the test for overall pathway disturbance changed significance between the raw and imputed datasets in at least one subtype. This resulted in ten pathways with different statistical conclusions between the raw and imputed datasets. Figure 4 shows the $-\log_{10}(p)$-values for the raw and imputed datasets in these pathways, across the different subtypes.

Changes in significance are in some cases minimal, and likely represent insubstantial statistical variation—for example, significance in any subtype is questionable in the HIV-1 and HDAC1-

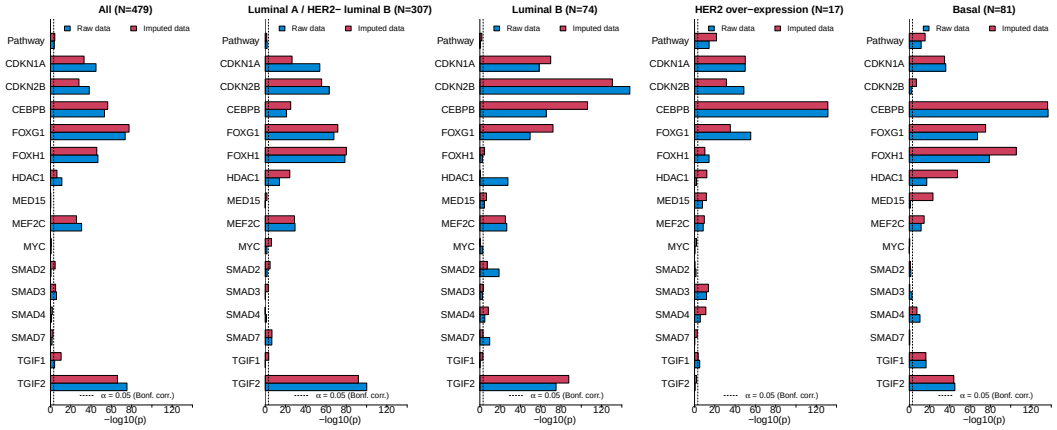

Figure 5: $-\log_{10}(p)$-*values for significance tests of integrated genomics data for the entire Smad2 / Smad3 signaling pathway and selected genes, by subtype. Gene-level significance tests considered integrated triplets of pathway nodes, namely, expression, copy number, and methylation observations for each gene.*

mediated signaling events. Likewise, β_1 integrin cell surface interactions appear significant only in luminal B subtypes, and β_2 integrin cell surface interactions is significant in HER2 over-expression subtypes. In these cases, the significance is strong in both the raw and imputed datasets. Moreover, we verified the robustness of the pathways shown in Figure 4 with SMC on the entirety of $\mathbf{X}_{11}$, as well as KNN imputation, and in both cases, the same pathways changed significance, with similar conclusions.

Of the pathways we tested, regulation of Smad2 and Smad3 signaling was identified as significantly disturbed across all subtypes, although only marginally so for luminal subtypes; and the pathway contains ESR1, an oncogene implicated in breast cancer. Smad2 and Smad3 play a well-established role in transcription growth factor β (TGF-β) (Brown *et al.*, 2007), and we considered this regulatory pathway for further analysis.

Our data contained a total of $N = 479$ breast tumor samples with valid IHC observations. The Smad2 / Smad3 signaling pathway consists of 68 genes, comprising 200 -omics features after integration of copy number and methylation data. We performed NetGSA on the raw and imputed datasets, which contained 371 and 543 samples, respectively. Of the 172 samples with imputation, 119 were luminal A / HER2- luminal B subtype; 23 were luminal B; 6 were HER2 over-expression; and 24 were basal. We conducted significance tests for the full pathway, as well as each integrated triplet corresponding to one gene: three nodes, one for each integrated platform.

The network topology for the pathway consists of 423 directed edges between the expression nodes, as well as two directed edges for each expression node to integrate copy number and methylation, where available. A graphical representation of the network may be accessed online, as discussed below.

Figure 5 plots p-values for all breast cancer samples, as well as individual subtypes. Comparison is given between the raw and imputed datasets. Shown are the results of testing the entire pathway, as well as selected individual genes, with Bonferroni testing for multiple correction. Gene-level significance tests considered gene-level trios of pathway vertices, namely, those for expression, copy number, and methylation.

We note that, tested across all tumors, the pathway is only marginally significant. In contrast,

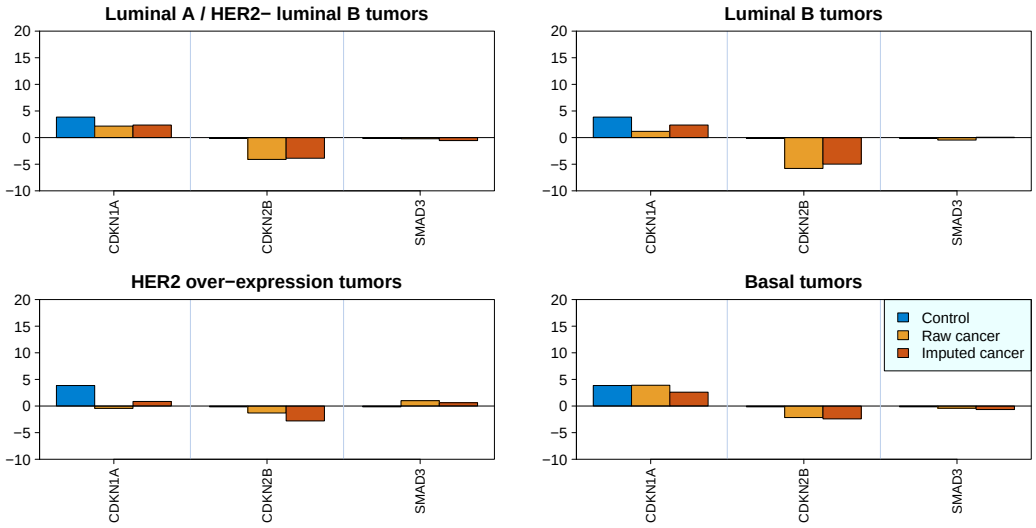

Figure 6: *Coefficient estimates for expression in the Smad2 / Smad3 signaling pathway, by subtype. Selected genes shown only. Coefficients were estimated separately by population (cancerous tumor and healthy control). Within the tumor population, separate estimates were obtained using the raw data with missing data removed, and using the imputed data.*

analyzed at the subtype level, a clean partition is clear: Smad2 and Smad3 regulation is disturbed in HER2 over-expression and basal subtypes, with strong statistical significance, but not disturbed significantly in either luminal A or B subtypes. These results are consistent between the raw and imputed datasets, but exhibit stronger significance in the imputed dataset.

Considered more granularly, TG-interacting factor 1 (TGIF1) displays only marginal significance in HER2 over-expression subtypes, consistent with the low significance of TGIF2. TGIF2 is strongly significant in the other three subtypes, but in the raw data, TGIF1 is strongly significant only in basal subtypes. However, our imputed method identifies these as significantly disturbed in luminal subtypes.

Our imputation method also reveals disturbance in Smad3 in luminal A / HER2- luminal B subtypes, and reinforces the lack of significance in basal tumors. The effect of Smad2 is reduced in luminal B tumors but increased in luminal A.

Across all subtypes, MED15 is not identified as disturbed, but within each subtype it is. This suggests a biological difference between the role this gene places in different subtypes. Figure 6 shows gene expression coefficient estimates $\widehat{\boldsymbol{\beta}}$ from the NetGSA analysis, and MED15 demonstrates down-regulation in HER2 over-expressed tumors, compared with up-regulation in basal tumors. Luminal B and basal tumors have comparable, mid-level expression in CEBPB, small in comparison with control; in HER2 over-expressed tumors, this value is under-expressed, whereas in luminal A tumors it is over-expressed.

HER2 over-expressed tumors exhibit expression in MYC in the opposite direction of the other subtypes, while the imputation actually causes a change in the sign of HDAC1 in luminal B tumors. Recent work found inconclusive results on the role of this gene in cancer tumors (Tang *et al.*, 2015). The sign switch is only in luminal B tumors, whereas in other tumors the difference is absent, or at least more muted.

Also of note is the HER2 over-expression subtype. The sample size for this subtype is already

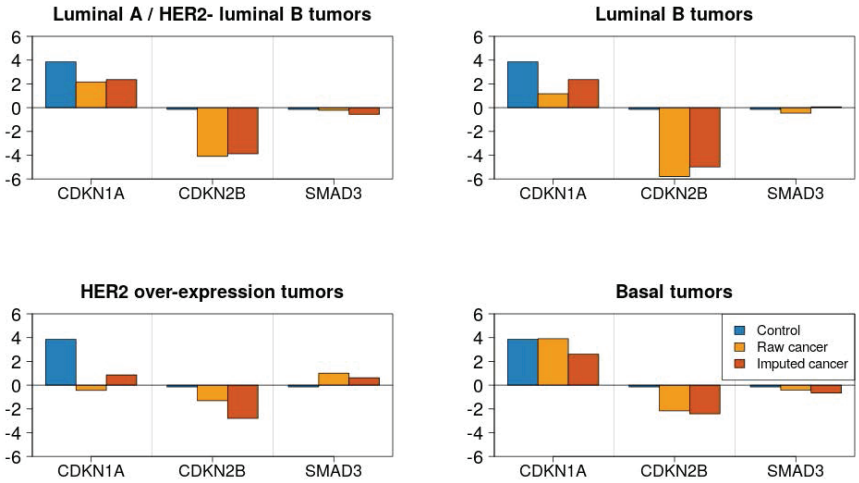

Figure 7: *Expression coefficient estimates for a selected subset of Smad2 / Smad3 signaling pathway genes, by subtype. Coefficients were estimated separately by population (cancerous tumor and healthy control). Within the tumor population, separate estimates were obtained using the raw data with missing data removed, and using the imputed data.*

small, at 17, but with the raw data, only 11 samples are available for analysis of HER2 over-expression. We observed sign changes between the raw and imputed data among several genes in this subtype, including the gene CDKN1A, FOXH1, and MEF2C. The latter two genes are only marginally significant, especially compared with the very large p-values in other subtypes. But CDKN1A is strongly significant, and is known to relate to Smad signaling.

Smad3 is known to inhibit tumor cell growth (Zelivianski *et al.*, 2010), but phosphorylation by CDK prevents this tumor-suppressive behavior (Liu, 2011). Counteracting this effect, inhibiting CDK promotes transcription by Smad3, with the effect of suppressing cancer cell growth. This is speculated to reflect a non-canonical interaction between CDK and Smad3 that promotes tumor growth (Tarasewicz *et al.*, 2014).

Addressing a lack of bio-marker targets for basal (triple-negative) breast tumors, recent research on this subtype has considered targeted suppression of CDK phosphorylation of Smad3. In particular, CDK has been observed to mediate a malignant interaction between Pin1 and Smad3 that increases proliferation of cancer cells in aggressive breast tumors (Thomas *et al.*, 2017). Targeted inhibition of CDK2 and CDK4 was prevented the Pin1-Smad3 interaction in basal subtypes, confirming the finding in (Tarasewicz *et al.*, 2014) that suppression of CDK2/4 leads to increased Smad3 activity in basal subtypes.

Figure 7 shows expression coefficient estimates for the four tumor subtypes, for CDKN1A, CDKN 2B, and Smad3. CDKN1A inhibits CDK2 phosphorylation of Smad3, while CDKN2B inhibits CDK4 (Driver *et al.*, 2008). In the healthy control tissue, CDKN1A expression is strong, whereas CDKN2B and Smad3 are not expressed at all.

We observe negative expression in Smad3 in the basal subtype, and the same in luminal A / HER2- luminal B tumors. This compares with slight positive expression of Smad3 in the other two subtypes. All four subtypes display positive expression for CDKN1A, which inhibits CDK2, however, the luminal and HER2 over-expression subtypes all display negative expression for CDKN2B, with

the strongest effect in luminal B tumors. This corresponds to small, positive expression in Smad3.

For comparison, we also analyzed the pathway using NetGSA on only expression data; with integration of only one of methylation and copy number at a time; and GSEA of expression (Wu *et al.*, 2010). The GSEA analysis identified only differential expression in HER2+ subtypes. The other three NetGSA methods were strongly significant in HER2+ and basal subtypes, as well as across all tumor types. However, while all four methods found marginal to no significance in luminal subtypes, the EC-NetGSA exhibited stronger significance. This supports the premise that Smad2 and Smad3 are central to transcription.

Taken as a whole, the results of our data analysis give strong support to the value of subtype analysis in complex diseases such as cancer. Among breast tumors, we identify Smad2 and Smad3 signaling pathway disturbance in basal and HER2 over-expression subtypes, with weaker disturbance in luminal subtypes. We identify separate roles of Smad3 in basal and HER2 over-expressed subtypes, suggesting the gene-level mechanism for the pathway disturbance may differ in the two subtypes.

Given the multivariate nature of our data, as well as the cross-sectional comparisons performed in our inference, even straightforward analyses generate large volumes of statistical output that requires detailed attention. Our discussion above of Smad2 and Smad3 signaling demonstrates this point. However, considered across five sub-populations compared with normal tissue, separate raw and imputed datasets, 176 pathways, and many hypothesis tests within each pathway, it is prohibitively impractical to generate these results "on-demand" for individual pathways, or to manually explore thousands of procedurally-generated plots.

To address this problem and assist in our data analysis, we implemented an interactive web application using the Shiny software package in R (Chang *et al.*, 2018). This framework generates responsive, modern web content form within R, with a focus on data-driven and statistical applications and requiring no direct understanding of HTML.

The application aggregates pathway-level significance test results for all 176 pathways, for discovery at the level of pathways and subtypes. We also implement an interface to explore individual pathways. We provide a dynamic visualization of the network topology and adjacency matrix partial correlations, built with `igraph` (Csardi and Nepusz, 2006) and `visnetwork` (Almende *et al.*, 2018) packages. We also display p-values and test statistics for NetGSA tests of the integrated gene-level subsets, and a comparison view of the expression coefficient estimates $\boldsymbol{\beta}$.

The application is publicly available online at

`https://zhang-lab.shinyapps.io/pathway-analysis-missing-data/`

4. Conclusion

In this paper, we presented a procedure to impute missing data in integrative genomics. We applied a matrix completion method in an iterative fashion, imputing non-rectangular missing data in a way that uses all available information in a genomic sample. Our approach permits application of a theoretically-appealing model for imputation to real-world data. We applied pathway analysis to the imputed data, and demonstrated through simulation the increased power of our strategy, with a smaller rise in false discoveries than the original SMC method. We demonstrated that the I3 method improves the precision of the basic SMC method, while still leveraging the linear dependencies that make SMC attractive in the first place.

The results of our data analysis demonstrate the power to stabilize coefficient estimates of subtypes, despite small sample sizes. Furthermore, they reinforce the importance of subtype analysis of tumors, rather than across entire populations. Among the pathways in the NCI Pathway Interaction

Database, we found many possible pathway disturbances. Ranked by variability between the subtypes, we identified the pathways (1) EPHB forward signaling, (2) FOXA2 and FOXA3 transcription factor networks, (3) ErbB2/ErbB3 signaling events, (4) Validated nuclear estrogen receptor alpha network, and (5) E2F transcription factor network, as warranting further study. Some of these pathways play known roles in breast cancer, which supports evidence of disturbance in other pathways not previously known. The FOXA2 and FOXA3 transcription factor networks and the ErbB2/ErbB3 signaling events pathways, in particular, display pathway disturbances that have been relatively unexplored in breast cancer.

Our method identifies differential disturbances of the Smad2 and Smad3 signaling pathway across subtypes, including disturbance in basal and HER2 over-expressed tumors, but lesser disturbance in luminal subtypes. Moreover, our findings suggest different mechanisms by which Smad3 operates in basal and HER2 over-expressed tumors. We provide access to the results of the paper in a web application available for public use.

Based on the results of the simulation study and the data analysis, we conclude that iterative integrated imputation (I3) successfully strikes a balance between local accuracy and global structural constraints when imputing missing data. We demonstrate the bias introduced by the assumption that missing data occurs in blocks. We propose a procedure based on SMC that includes the original block-missing SMC as a special case, while flexibly handling non-rectangular missing data. Our method imputes values that permit analysis of datasets with missing data, without providing imputed values that distort the output statistical analysis.

Acknowledgement

We sincerely thank the Editor, Associate Editor and anonymous reviewers for their careful review and constructive comments. YZ acknowledges University of Connecticut Faculty Research Excellence Program Award.

References

Almende BV, Thieurmel B, and Robert T (2018). visNetwork: Network Visualization using vis.js Library, R package version 2.0.4, https://CRAN.R-project.org/package=visNetwork

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B (Methodological)*, **57**, 289–300.

Bochkis IM, Schug J, Diana ZY, Kurinna S, Stratton SA, Barton MC, and Kaestner KH (2012). Genome-wide location analysis reveals distinct transcriptional circuitry by paralogous regulators Foxa1 and Foxa2, *PLoS Genetics*, **8**, e1002770.

Brown KA, Pietenpol JA, and Moses HL (2007). A tale of two proteins: Differential roles and regulation of Smad2 and Smad3 in TGF-beta signaling, *Journal of Cellular Biochemistry*, **101**, 9–33.

Cai T, Cai TT, and Zhang A (2016). Structured matrix completion with applications to genomic data integration, *Journal of the American Statistical Association*, **111**, 621–633.

Candès EJ and Tao T (2010). The power of convex relaxation: Near-optimal matrix completion, *IEEE Transactions on Information Theory*, **56**, 2053–2080.

Chang W, Cheng J, Allaire JJ, Xie Y, and McPherson J (2018). shiny: Web Application Framework for R, R package version 1.2.0, https://CRAN.R-project.org/package=shiny

Cheang MCU, Chia SK, Voduc D, *et al.* (2009). Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer, *JNCI: Journal of the National Cancer Institute*, **101**, 736–750.

Chudasama P, Mughal SS, Sanders MA, *et al.* (2018). Integrative genomic and transcriptomic analysis of leiomyosarcoma, *Nature Communications*, **9**, 144.

Csardi G and Nepusz T (2006). The igraph software package for complex network research, *InterJournal*, Complex Systems, 1695.

Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, and Shi B (2015). Breast cancer intrinsic subtype classification, clinical use and future trends, *American Journal of Cancer Research*, **5**, 2929.

Danielsen SA, Eide PW, Nesbakken A, Guren T, Leithe E, and Lothe RA (2015). Portrait of the PI3K/AKT pathway in colorectal cancer, Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, **1855**, 104–121.

Driver KE, Song H, Lesueur F, *et al.* (2008). Association of single-nucleotide polymorphisms in the cell cycle genes with breast cancer in the British population, *Carcinogenesis*, **29**, 333–341.

Franzin A, Sambo F, and di Camillo B (2017). bnstruct: an R package for Bayesian Network structure learning in the presence of missing data, *Bioinformatics*, **33**, 1250–1252.

Fryett JJ, Inshaw J, Morris AP, and Cordell HJ (2018). Comparison of methods for transcriptome imputation through application to two common complex diseases, *European Journal of Human Genetics*, **26**, 1658–1667.

Gamazon ER, Wheeler HE, Shah KP, *et al.* (2015). A gene-based association method for mapping traits using reference transcriptome data, *Nature Genetics*, **47**, 1091.

Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, and Staudt LM (2016). Toward a shared vision for cancer genomic data, *New England Journal of Medicine*, **375**, 1109–1112.

Gusev A, Ko A, Shi H, *et al.* (2016). Integrative approaches for large-scale transcriptome-wide association studies, *Nature Genetics*, **48**, 245.

Howie BN, Donnelly P, and Marchini J (2009). *PLoS Genetics*, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, **5**, e1000529.

Hsu YHH, Churchhouse C, Pers TH, *et al.* (2019). PAIRUP-MS: Pathway analysis and imputation to relate unknowns in profiles from mass spectrometry-based metabolite data, *PLoS Computational Biology*, **15**, e1006734.

Johnson J, Thijssen B, McDermott U, Garnett M, Wessels LFA, and Bernards R (2016). Targeting the RB-E2F pathway in breast cancer, *Oncogene*, **35**, 4829.

Kaenel P, Mosimann M, and Andres AC (2012). The multifaceted roles of Eph/ephrin signaling in breast cancer, *Cell Adhesion & Migration*, **6**, 138–147.

Köksal AS, Beck K, Cronin DR, *et al.* (2018). Synthesizing signaling pathways from temporal phosphoproteomic data, *Cell Reports*, **24**, 3607–3618.

Krause RW, Huisman M, and Snijders TA (2018). Multiple imputation for longitudinal network data, *Italian Journal of Applied Statistics*, **30**, 33–58.

Krämer N, Schäfer J, and Boulesteix AL (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models, *BMC Bioinformatics*, **10**, 384.

Liu F (2011). Inhibition of Smad3 activity by cyclin D-CDK4 and cyclin E-CDK2 in breast cancer cells, *Cell Cycle*, **10**, 190–191.

Ma J, Lyu H, Huang J, and Liu B (2014). Targeting of erbB3 receptor to overcome resistance in cancer treatment, *Molecular Cancer*, **13**, 105.

Mazumder R, Hastie T, and Tibshirani R (2010). Spectral regularization algorithms for learning large incomplete matrices, *Journal of Machine Learning Research*, **11**, 2287–2322.

Nevins JR (2001). The Rb/E2F pathway and cancer, *Human Molecular Genetics*, **10**, 699–703.

Pasquale EB (2010). Eph receptors and ephrins in cancer: bidirectional signalling and beyond, *Nature*

Reviews Cancer, **10**, 165.

Sales G, Calura E, and Romualdi C (2018). *graphite: GRAPH Interaction from pathway Topological Environment*, R package version 1.26.1.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, and Buetow KH (2008). PID: the pathway interaction database, *Nucleic Acids Research*, **37**, D674–D679.

Schulz H, Ruppert AK, Herms S, *et al.* (2017). Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus, *Nature Communications*, **8**, 1511.

Shen R, Olshen AB, and Ladanyi M (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis, *Bioinformatics*, **25**, 2906–2912.

Shojaie A and Michailidis G (2009). Analysis of gene sets based on the underlying regulatory network, *Journal of Computational Biology*, **16**, 407–426.

Shojaie A and Michailidis G (2010). Network enrichment analysis in complex experiments, *Statistical Applications in Genetics and Molecular Biology*, **9**, 22.

Sommer S and Fuqua SA (2001). Estrogen receptor and breast cancer, *Seminars in Cancer Biology*, **11**, 339–352.

Subramanian A, Tamayo P, Mootha VK, *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. In Proceedings of the National Academy of Sciences, **102**, 15545–15550.

Tang YN, Ding WQ, Guo XJ, Yuan XW, Wang DM, and Song JG (2015). Epigenetic regulation of Smad2 and Smad3 by profilin-2 promotes lung cancer growth and metastasis, *Nature Communications*, **6**, 8230.

Tarasewicz E, Rivas L, Hamdan R, *et al.* (2014). Inhibition of CDK-mediated phosphorylation of Smad3 results in decreased oncogenesis in triple negative breast cancer cells, *Cell Cycle*, **13**, 3191–3201.

Thomas AL, Lind H, Hong A, *et al.* (2017). Inhibition of CDK-mediated Smad3 phosphorylation reduces the Pin1-Smad3 interaction and aggressiveness of triple negative breast cancer cells, *Cell Cycle*, **16**, 1453–1464.

Tomczak K, Czerwińska P, and Wiznerowicz M (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, *Contemporary Oncology*, **19**, A68.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, and Altman RB (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520–525.

Tsuchiya T, Fujii M, Matsuda N, Kunida K, Uda S, Kubota H, Konishi K, and Kuroda S (2017). System identification of signaling dependent gene expression with different time-scale data, *PLoS Computational Biology*, **13**, e1005913.

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, and Cox J (2016). The Perseus computational platform for comprehensive analysis of (prote) omics data, *Nature Methods*, **13**, 731.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, and Stuart JM (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, *Bioinformatics*, **26**, i237–i245.

Wei L, Jin Z, Yang S, Xu Y, Zhu Y, and Ji Y (2017). TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data, *Bioinformatics*, **34**, 1615–1617.

Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, and Smyth GK (2010). ROAST: rotation gene set tests for complex microarray experiments, *Bioinformatics*, **26**, 2176–2182.

Zelivianski S, Cooley A, Kall R, and Jeruss JS (2010). Cyclin-dependent kinase 4–mediated phosphorylation inhibits Smad3 activity in cyclin D–overexpressing breast cancer Cells, *Molecular Cancer Research*, **8**, 1375–1387.

Zhang Y, Linder MH, Shojaie A, Ouyang Z, Shen R, Baggerly KA, Baladandayuthapani V, and Zhao H (2017a). Dissecting pathway disturbances using network topology and multi-platform genomics data, *Statistics in Biosciences*, **10**, 1–21.

Zhang Y, Ouyang Z, and Zhao H (2017b). A statistical framework for data integration through graphical models with application to cancer genomics, *The Annals of Applied Statistics*, **11**, 161–184.

Zhao Y, Hoang TH, Joshi P, Hong SH, Giardina C, and Shin DG (2017). A route-based pathway analysis framework integrating mutation information and gene expression data, *Methods*, **124**, 3–12.

Zhou X, Carbonetto P, and Stephens M (2013). Polygenic modeling with Bayesian sparse linear mixed models, *PLoS Genetics*, **9**, e1003264.

Zhu Y, Qiu P, and Ji Y (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data, *Nature Methods*, **11**, 599.

Received April 30, 2019; Revised June 10, 2019; Accepted June 16, 2019