*Bioinformatics*, 34, 2018, i178–i186
doi: 10.1093/bioinformatics/bty250
ISMB 2018

OXFORD

# Personalized regression enables sample-specific pan-cancer analysis

**Benjamin J. Lengerich[1], Bryon Aragam[2] and Eric P. Xing[1,2,3,*]**

[1]Computer Science Department and [2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA and [3]Petuum, Inc., Pittsburgh, PA 15222, USA

*To whom correspondence should be addressed.

## Abstract

**Motivation:** In many applications, inter-sample heterogeneity is crucial to understanding the complex biological processes under study. For example, in genomic analysis of cancers, each patient in a cohort may have a different driver mutation, making it difficult or impossible to identify causal mutations from an averaged view of the entire cohort. Unfortunately, many traditional methods for genomic analysis seek to estimate a single model which is shared by all samples in a population, ignoring this inter-sample heterogeneity entirely. In order to better understand patient heterogeneity, it is necessary to develop practical, personalized statistical models.

**Results:** To uncover this inter-sample heterogeneity, we propose a novel regularizer for achieving patient-specific personalized estimation. This regularizer operates by learning two latent distance metrics—one between personalized parameters and one between clinical covariates—and attempting to match the induced distances as closely as possible. Crucially, we do not assume these distance metrics are already known. Instead, we allow the data to dictate the structure of these latent distance metrics. Finally, we apply our method to learn patient-specific, interpretable models for a pan-cancer gene expression dataset containing samples from more than 30 distinct cancer types and find strong evidence of personalization effects between cancer types as well as between individuals. Our analysis uncovers sample-specific aberrations that are overlooked by population-level methods, suggesting a promising new path for precision analysis of complex diseases such as cancer.

**Availability and implementation:** Software for personalized linear and personalized logistic regression, along with code to reproduce experimental results, is freely available at github.com/blengerich/personalized_regression.

**Contact:** epxing@cs.cmu.edu

## 1 Introduction

A fundamental goal of pan-omic analysis, and a bottleneck for personalized medicine, is to understand the patterns of differentiation between individuals. With the advent of projects like The Cancer Genome Atlas (cancergenome.nih.giv) (TCGA) and the International Cancer Genome Consortium (ICGC) (dcc.icgc.org), genomic cancer data are generated at an unprecedented volume. We would like to use these data to understand patient-specific differences for personalized medicine, but many analysis pipelines discard sample heterogeneity in order to boost accuracy. Sample heterogeneity is particularly important for cancer, as cancer is increasingly appreciated as a complex disease in which many distinct underlying mutations may present with similar phenotypes (Fisher *et al.*, 2013); even within a single patient, there is increasing evidence of tumor mosaics composed of distinct cell lines (Marusyk *et al.*, 2012). This difficulty with complex diseases like cancer motivates us to find new ways of analyzing data at increasingly small granularities.

Toward this aim, the bioinformatics community has developed increasingly specific assays (Kumar-Sinha and Chinnaiyan, 2018). From targeted microarrays, to whole-genome RNA-Seq, to single cell RNA-Seq, the granularity of data collected by genomic assays has continued to be refined, to the point that we now possess data points representing the state of an individual cell at a single time point, unlocking the potential to study inter-patient, inter-tissue and inter-cell variability of complex diseases.

A classic approach to personalization is to assume that we have access to a large volume of multimodal data (e.g. clinical, genomic, proteomic, biometric, etc.) on each individual, which is used to build large predictive models. Given enough data per individual, clinical outcomes and decisions can be personalized (Kumar-Sinha and

Chinnaiyan, 2018; Pittman *et al.*, 2004), and recent work along these lines has leveraged a dizzying array of complex models including Gaussian processes (Alaa *et al.*, 2016), neural networks (Lopez-Martinez and Picard, 2017) and tree-based models (Moon *et al.*, 2007), just to name a few. Despite the successes of these methods, they are still limited to this 'one disease–one model' perspective, in which a single predictive model—often through model averaging—is built for a single cohort (e.g. corresponding to patients of a particular disease type). Furthermore, these complex models are often difficult to interpret and are not guaranteed to provide correct inference into the underlying biological drivers of disease.

Unfortunately, in many circumstances, we may only have access to a limited amount of measurements per individual (e.g. either for cost or privacy reasons). In this case, it is advantageous to leverage data from distinct but related cohorts in order to build personalized models for each individual. For example, in cancer applications we now have access to large datasets for commonly studied cancers such as breast and lung cancer through repositories such as TCGA. At the same time, less common cancers such as of the eye and lymph node, have much less data (Table 1). A true 'pan-cancer' study would combine all of this data, exploiting the similarities between different types of cancer to improve the accuracy of models for eye and lymph node cancer. That such similarities exist is well-established in the literature (e.g. Weinstein *et al.*, 2013). However, in the traditional 'one disease–one model' paradigm, data from other cancers play no role; while this makes sense for diseases which have a single root cause, the heterogeneity of complex diseases such as cancer renders these methods inadequate. Leveraging data from multiple cohorts while simultaneously obtaining distinct models for different diseases and different patients is a key challenge in personalized medicine.

Motivated by this new 'many disease–many model' paradigm, we propose a framework to estimate patient-specific models by learning patterns of differentation between samples. Instead of learning a single model for an entire cohort, our framework learns a unique model for each patient. The key is to leverage the fact that although each patient is expected to have a unique pattern of differentiation, these patterns are not independent of one another, and are expected to share substantial similarities. Leveraging this, we can 'borrow strength' from the entire cohort to learn a useful model that is specific to a given patient. To do this, we propose a novel *distance matching* regularizer and show how it can be applied to regression problems. Our main contributions are threefold:

1. A novel framework for personalized regression via distance matching;
2. We show that this framework can learn patient-specific models without prior knowledge of patient relatedness;
3. A TCGA pan-cancer study that illustrates the simultaneous similarities and differences between putative driver mutations of different cancers.

Although our main application will be to regression problems, our goal is *not* to simply predict an outcome, but instead to learn the underlying mechanisms that drive disease and lead to sample heterogeneity. By focusing our framework on learning patterns of differentiation, we can produce interpretable models of controllable granularity from patient-specific to pan-cancer.

## 2 Related work

Traditional models assume only one or a few statistical parameters for a given population. As a simple example, consider the case of linear regression with response $Y$ and predictors $X$. Then the typical model is $Y = X\beta + \epsilon$, where the parameter $\beta$ is shared across all individuals. More generally, mixture models allow for $K$ different parameters $\beta_k$, $k = 1 \ldots, K$. Another related generalization is multi-task learning (Maurer *et al.*, 2013). In both mixture modeling and multi-task learning, however, it is necessary that $K \ll N$ where $N$ is the total number of samples in the cohort. By contrast, we are interested in the case $K = N$, i.e. a single parameterization for each individual in the cohort. This is what is meant by *personalized* regression models.

**Table 1.** Number of samples by tissue in TCGA

| Tissue | $n$ | Tissue | $n$ |
|---|---|---|---|
| Breast | 1092 | Ovary | 376 |
| Lung | 1016 | Liver | 371 |
| Kidney | 885 | Cervix | 304 |
| Brain | 677 | Soft tissue | 259 |
| Colorectal | 623 | Adrenal gland | 258 |
| Uterus | 611 | Pancreas | 177 |
| Thyroid | 502 | Esophagus | 164 |
| Head and Neck | 501 | Bone marrow | 151 |
| Prostate | 495 | Eye | 80 |
| Skin | 468 | Lymph nodes | 48 |
| Bladder | 408 | Bile duct | 36 |
| Stomach | 380 | | |

As it has long been a goal of biologists to understand inter-sample variation, significant prior work has aimed to estimate model parameters that vary between samples. Unfortunately, prior work requires either (i) a small number of sub-groups relative to the number of samples (e.g. mixture models) (Roth *et al.*, 2014), or (ii) known patterns of variation (Kolar *et al.*, 2009; Parikh *et al.*, 2011; Song *et al.*, 2009), or (iii) significant domain knowledge to constrain the solutions (Xu *et al.*, 2015; Yamada *et al.*, 2016). Also closely related are random coefficient models, however, traditional random coefficient models do not allow for sample-specific sparsity patterns. In the presence of additional covariates $U$ (often, time or clinical variables), varying coefficient (VC) models have also been explored extensively (Fan and Zhang, 1999; Hastie and Tibshirani, 1993; Kolar *et al.*, 2009). In this VC framework, each regression parameter is modeled as a function of some external covariates $U$, i.e. $\beta = f(U)$. As with other models, VC models require significant domain knowledge in order to model a suitable relationship between $\beta$ and $U$.

The closest work in spirit to ours is arguably the recent work on sample-specific network estimation (Kuijjer *et al.*, 2015; Liu *et al.*, 2016). Although these papers also consider the problem of sample-specific estimation, they focus on the particular problem of network estimation, and hence are not directly comparable to the present work.

## 3 Model

We are interested in learning which features $X \in \mathbb{R}^P$ are relevant for predicting a phenotype $Y \in \mathbb{R}$ such as disease status. At the same time, we assume we have access to clinical covariates $U \in \Omega_1 \times \cdots \times \Omega_K$ for each individual, which are allowed to be arbitrary—unordered or ordered, categorical or continuous and even with missing values. Throughout, we let $N$ denote the total number of patients in the cohort and use superscripts to identify samples. Thus, $Y^{(i)}$, $X^{(i)}$ and $U^{(i)}$, denote the data for sample $i$ and $\beta^{(i)}$ is the personalized regression parameter for the $i$th sample.

### 3.1 Distance matching

To recover personalized model parameters $\beta^{(i)}$ without *a priori* knowledge of how samples are related, we assume that there are *unknown* distance (pseudo-)metrics $d_\beta$ and $d_U$ such that $d_\beta\left(\beta^{(i)}, \beta^{(j)}\right) \approx d_U\left(U^{(i)}, U^{(j)}\right)$. That is, similarity in parameters is related to similarity in covariates, however, the nature of this similarity is unobserved, unknown and may not correspond to usual notions of distance such as Euclidean distance. This is closely related to the notion of distance metric learning introduced by Xing *et al.* (2003). Existing work along these lines in the personalized estimation literature typically assumes that either (a) The metrics are Euclidean, or (b) The pairwise similarities are known (Xu *et al.*, 2015; Yamada *et al.*, 2016).

To learn these latent distance metrics, we model them as follows:

$$d_\beta(x, y) = \zeta\langle\phi_\beta, \left[d_{\beta_1}(x_1, y_1), \ldots, d_{\beta_P}(x_P, y_P)\right]\rangle, \tag{1a}$$

$$d_U(x, y) = \langle\phi_U, [d_{U_1}(x_1, y_1), \ldots, d_{U_K}(x_K, y_K)]\rangle, \tag{1b}$$

where $\langle\cdot,\cdot\rangle$ denotes the dot product of two vectors and $d_{\beta_p}$ $(p = 1, \ldots, P)$ are user-specified metrics between scalars and $d_{U_k}$ $(k = 1, \ldots, K)$ are user-specified metrics between covariates. Note that here we do not require these distance metrics to be differentiable. This allows for a wide variety of distance metrics, such as the discrete metric $d_{U_k}(x, y)$ that equals one if $x = y$ and is zero otherwise. This allows our framework to handle the realistic situation of categorical covariates without ordering. The parameters $\phi_\beta$ and $\phi_U$ represent unknown linear transformations of these 'simple' distances into more useful latent distance metrics given by (1a) and (1b) with scale $\zeta > 0$.

Define pairwise distance vectors for each $i, j$ by

$$\Delta_\beta^{(i,j)} = \left[d_{\beta_1}\left(\beta_1^{(i)}, \beta_1^{(j)}\right), \ldots, d_{\beta_P}\left(\beta_P^{(i)}, \beta_P^{(j)}\right)\right] \tag{2a}$$

$$\Delta_U^{(i,j)} = \left[d_{U_1}\left(U_1^{(i)}, U_1^{(j)}\right), \ldots, d_{U_K}\left(U_K^{(i)}, U_K^{(j)}\right)\right] \tag{2b}$$

Since the covariate values in $U$ are fixed, $\Delta_U^{(i,j)}$ is also fixed, whereas $\Delta_\beta^{(i,j)}$ is not fixed since the values of $\beta^{(i)}$ and $\beta^{(j)}$ will change during training. For simplicity, we take $d_{\beta_p}(x, y) = |x - y|$ $(p = 1, \ldots, P)$ in the remainder of this paper.

Now define the following *distance matching regularizer*:

$$\begin{aligned} \varrho_\gamma^{(i)}(d_\beta, d_U) &= \frac{\gamma}{2}\sum_{j\neq i}(d_\beta(\beta^{(i)}, \beta^{(j)}) - d_U(U^{(i)}, U^{(j)}))^2 \\ &= \frac{\gamma}{2}\sum_{j\neq i}(\zeta\langle\phi_\beta, \Delta_\beta^{(i,j)}\rangle - \langle\phi_U, \Delta_U^{(i,j)}\rangle)^2. \end{aligned} \tag{3}$$

This regularizer attempts to match the pairwise distances between covariate values to the pairwise distances in the learned regression parameters. Let $f$ be a loss function, e.g. least squares for regression or logistic loss for classification. Define a sample-specific objective by

$$\mathcal{L}^{(i)}\left(\beta^{(i)}; d_\beta, d_U\right) \propto f\left(X^{(i)}, Y^{(i)}, \beta^{(i)}\right) + \rho_\lambda^\beta\left(\beta^{(i)}\right) + \varrho_\gamma^{(i)}(d_\beta, d_U).$$

Summing these, we obtain the complete objective function:

$$\mathcal{L}(\boldsymbol{\beta}, \phi_\beta, \phi_U, \zeta) \propto \sum_{i=1}^N \mathcal{L}^{(i)}(\beta, d_\beta, d_U) + \psi_\alpha^\beta(d_\beta) + \psi_\upsilon^U(d_U).$$

where $\gamma$ trades off sensitivity to prediction of the response variable against sensitivity to sample distances, $f\left(X^{(i)}, Y^{(i)}, \beta^{(i)}\right)$ is the prediction loss for sample $i$, $\rho_\lambda^\beta : \mathbb{R}^P \to \mathbb{R}_{\geq 0}$ regularizes $\beta^{(i)}$ with strength set by $\lambda$, and $\psi_\alpha^\beta$, and $\psi_\upsilon^U$ regularize the distance functions $d_\beta$, $d_U$ with strengths set by $\alpha$, $\upsilon$, respectively.

### 3.2 Parametrization and initialization

Since the program (4) is nonconvex and the number of free parameters is large, some care must be taken to avoid degenerate solutions. We constrain the $\ell_1$ norm of both $\phi_\beta$ and $\phi_U$ to be equal to 1 and put all scaling into a single scale parameter $\zeta$. In addition, we require that each entry of $\phi_U$ and $\phi_\beta$ is non-negative, ensuring non-negative distances between samples. Placing appropriate priors on $\phi_\beta$ and $\phi_U$, we arrive at the final program we wish to optimize:

$$\begin{aligned} \min_{\beta,\phi_\beta,\phi_U,\zeta} \mathcal{L}(\boldsymbol{\beta}, \phi_\beta, \phi_U, \zeta) \quad &\text{such that} \quad \|\phi_\beta\|_1 = 1, \|\phi_U\|_1 = 1, \\ &\text{and} \quad \phi_\beta \geq 0,\ \phi_U \geq 0, \\ &\text{and} \quad \zeta \geq 0. \end{aligned} \tag{4}$$

where inequality here is interpreted component-wise.

After normalization, the model (4) has $(N + 1)P + K + 1$ free parameters to be learned from $N$ samples, which may seem significantly over-parameterized. Notwithstanding, although the technical details are beyond the scope of this short article, we can show that the distance matching regularizer (3) is able to constrain the personalized parameters $\widehat{\beta}^{(i)}$ so that they do not deviate too far from a *population* regression estimation $\widehat{\beta}^{\text{pop}}$, unless a substantial decrease in the loss can compensate for such deviations. Since (4) is a nonconvex program, proper initialization is crucial, and this gives us a practical strategy for initializing the personalized parameters: After solving for a population estimator $\widehat{\beta}^{\text{pop}}$, we initialize all $\widehat{\beta}^{(i)} = \widehat{\beta}^{\text{pop}}$. This initialization is important because the initial point is be a central point about which the personalized parameters are centered. As a result, our choice of regularizer allows for sample-specific personalization effects while preventing overfitting. This is a very desirable property for analysis of biological data: Suppose our data consists of microarray data from a diverse cohort of cancer patients. Each of these patients have experienced a series of mutations away from a healthy state; however, it is unlikely that they have experienced the same set of mutations. We would then like a personalized model to recover parameter values that are concentrated near a central model corresponding to a healthy state. This is precisely what distance matching does.

### 3.3 Missing values

When there is a missing value in the covariate data, we set the distance between this value and all others to zero. This underestimates the distance between samples, biasing the solution toward retaining a central population estimator rather than personalizing the models based on missing features.

### 3.4 Prediction

Although our main focus is on inference for a fixed sample cohort, given a new test point $X$, we can create a new model without rerunning the learning algorithm on the entire dataset by averaging the personalized parameters of the $K$ nearest models in the training set. This allows us to make predictions and inferences for new patients efficiently. Conveniently, since we have already learned a distance metric which we can use to accurately measure distance between samples, we can use this in the nearest neighbour search. Details are given in Algorithm 1. For linear regression, $p(x, \beta) = \langle x, \beta\rangle$. For logistic regression, $p(x, \beta) = \frac{\exp(\langle x,\beta\rangle)}{1+\exp(\langle x,\beta\rangle)}$.

**Algorithm 1** Inference Procedure

**Require:** Test point $(X^{(test)}, U^{(test)})$, predictive model $p(\cdot,\cdot)$, number of neighbors *N_neighbors*
$distances \leftarrow \{d_U(U^{(test)}, U^{(i)}) : i \in [1,\ldots,N_{train}]\}$
$neighbors \leftarrow sort(distances)[0 : N\_neighbors]$
$\beta^{(test)} \leftarrow mean(\{\beta^{(i)} : i \in neighbors\})$
**return** $p(X^{(test)}, \beta^{(test)})$

## 4 Optimization

We seek to minimize (4) by first estimating a traditional regression estimator such as the Lasso or OLS, and then gradually relaxing the personalized regression models away from this population model. For simplicitly, we describe the procedure as centered about a single population estimator, however, the method trivially extends to initialization about a mixture model. After setting hyperparameters $\gamma$, $\alpha$ and $\upsilon$ ($\lambda$ is dictated by the population estimator), we optimize by coordinate gradient descent with the following subgradients (note here that $Y^{(i)}$ is a scalar value, and $X^{(i)}$ and $\beta^{(i)}$ are both $p$-vectors):

$$\frac{\partial \mathcal{L}^{(i)}}{\partial \beta^{(i)}} = f'\left(Y^{(i)}, X^{(i)}, \beta^{(i)}\right) + \rho'\left(\beta^{(i)}\right) + \gamma \sum_{j\neq i}\left(\zeta\langle \phi_\beta, \Delta_\beta^{(i,j)}\rangle - \langle \phi_U, \Delta_U^{(i,j)}\rangle\right)\zeta \frac{\partial}{\partial \beta^{(i)}}\langle \phi_\beta, \Delta_\beta^{(i,j)}\rangle \tag{5a}$$

$$\frac{\partial \mathcal{L}^{(i)}}{\partial \phi_\beta} = \gamma \sum_{j\neq i}\left(\zeta\langle \phi_\beta, \Delta_\beta^{(i,j)}\rangle - \langle \phi_U, \Delta_U^{(i,j)}\rangle\right)\zeta\Delta_\beta^{(i,j)} + \psi'_\beta(\phi_\beta) \tag{5b}$$

$$\frac{\partial \mathcal{L}^{(i)}}{\partial \phi_U} = \gamma \sum_{j\neq i}\left(\zeta\langle \phi_\beta, \Delta_\beta^{(i,j)}\rangle - \langle \phi_U, \Delta_U^{(i,j)}\rangle\right)\left(-\Delta_U^{(i,j)}\right) + \psi'_U(\phi_U) \tag{5c}$$

$$\frac{\partial \mathcal{L}^{(i)}}{\partial \zeta} = \gamma \sum_{j\neq i}\left(\zeta\langle \phi_\beta, \Delta_\beta^{(i,j)}\rangle - \langle \phi_U, \Delta_U^{(i,j)}\rangle\right)\langle \phi_\beta, \Delta_\beta^{(i,j)}\rangle + \psi'_\zeta(\zeta) \tag{5d}$$

where $f'(\cdot,\cdot,\cdot)$, $\rho'(\cdot)$, $\psi'_\beta(\cdot)$ and $\psi'_U(\cdot)$ are subgradients of the predictive model $f(\cdot,\cdot,\cdot)$ and the regularizers $\rho_\lambda^\beta(\cdot)$, $\psi_\alpha^\beta(\cdot)$ and $\psi_\upsilon^U(\cdot)$, respectively.

The update to $\beta^{(i)}$ is dependent on the distance metric chosen for parameter values. For $d_{\beta_m}(x,y) = |x-y|$ , (5a) becomes

$$\frac{\partial \mathcal{L}^{(i)}}{\partial \beta^{(i)}} = f'\left(Y^{(i)}, X^{(i)}, \beta^{(i)}\right) + \rho'\left(\beta^{(i)}\right) + \gamma \sum_{j\neq i}\left(\zeta\langle \phi_\beta, \Delta_\beta^{(i,j)}\rangle - \langle \phi_U, \Delta_U^{(i,j)}\rangle\right)\mathrm{sgn}\left(\beta^{(i)} - \beta^{(j)}\right)\phi_\beta$$

where the sgn($\cdot$) function is applied element-wise. Finally, to ensure that each coordinate of $\phi$ is non-negative (i.e. distances cannot be negative), we project the updated value of $\phi$ into the non-negative reals. This is summarized in Algorithm 2. Each iteration of the naïve optimization procedure has computational time complexity in $O(N^2PK)$ where $P$ is the number of features, $K$ is the number of covariates and $N$ is the number of samples. This can be reduced to $O(NPK)$ by defining a constant-size set of neighbors for each sample and only calculating pairwise distances within neighborhoods, as illustrated in Algorithm 2. The use of these neighborhoods naturally extends this procedure to optimization of personalized models centered around mixture models.

**Algorithm 2** Optimization

**Require:** step size $a$, intializations $\beta^{\mathrm{pop}}$, $\phi^0_{beta}$, $\phi^0_U$, $\zeta^0$, covariate distances $\Delta_U$, training data $X$, $Y$, hyperparameters $\gamma$, $\alpha$, $\upsilon$
$\beta^i \leftarrow \beta^{\mathrm{pop}} \quad \forall i \in [1,\ldots,N_{train}]$
$\phi_\beta \leftarrow \phi_\beta^0$
$\phi_U \leftarrow \phi_U^0$
$\zeta \leftarrow \zeta^0$
**while** not converged **do**
  $d\phi_\beta \leftarrow \psi'_\beta(\phi_\beta)$
  $d\phi_U \leftarrow \psi'_U(\phi_U)$
  **for** $i \in \{1,\ldots,N\}$ **do**
    $d\beta[i] \leftarrow f'(Y^{(i)}, X^{(i)}, \beta^{(i)}) + \rho'(\beta^{(i)})$
    **for** $j \in$neighbors[i] **do**
      $g \leftarrow \gamma(\zeta\langle\phi_\beta, \delta_{beta}^{(i,j)}\rangle - \langle\phi_U, \Delta_U^{(i,j)}\rangle)$
      $d\beta[i] \leftarrow d\beta[i] + g\langle\phi_\beta, sign(\beta^{(i)} - \beta^{(j)})\rangle$
      $d\phi_\beta \leftarrow d\phi_\beta + g(\beta^{(i)} - \beta^{(j)})$
      $d\phi_U \leftarrow d\phi_U - g\Delta_U^{(i,j)}$
      $d\zeta \leftarrow g\langle\phi_{beta}, \delta_\beta^{(i,j)}\rangle$
    **end for**
  **end for**
  **for** $i \in \{1,\ldots,N\}$ **do**
    $\beta^{(i)} \leftarrow \beta^{(i)} - a * d\beta[i]$
  **end for**
  $\phi_\beta \leftarrow softmax(\phi_\beta - ad\phi_\beta)$
  $\phi_U \leftarrow softmax(\phi_U - ad\phi_U)$
  $\zeta \leftarrow max(0, \zeta - ad\zeta)$
**end while**
**return** $\beta, \phi_\beta, \phi_U, \zeta$

### 4.1 Linear regression

As an example application, let us instantiate the model (4) for personalized linear regression with Lasso regularization by

$$f\left(Y^{(i)}, X^{(i)}, \beta^{(i)}\right) = \frac{1}{2}\left(Y^{(i)} - \langle X^{(i)}, \beta^{(i)}\rangle\right)^2$$

$$f'\left(Y^{(i)}, X^{(i)}, \beta^{(i)}\right) = -\left(Y^{(i)} - \langle X^{(i)}, \beta^{(i)}\rangle\right)X^{(i)}$$

$$\rho_\lambda^\beta\left(\beta^{(i)}\right) = \lambda\|\beta^{(i)}\|_1.$$

### 4.2 Logistic regression

Similarly, we instantiate personalized logistic regression with response variables $Y^{(i)} \in \{0,1\}$ and Lasso regularization by

$$f\left(Y^{(i)}, X^{(i)}, \beta^{(i)}\right) = \log\left(1 + \exp\left(\langle X^{(i)}, \beta^{(i)}\rangle\right)\right) - Y^{(i)}X^{(i)}\beta^{(i)}$$

$$f'\left(Y^{(i)}, X^{(i)}, \beta^{(i)}\right) = \left(\frac{\exp\left(\langle X^{(i)}, \beta^{(i)}\rangle\right)}{1 + \exp\left(\langle X^{(i)}, \beta^{(i)}\rangle\right)} - Y^{(i)}\right)X^{(i)}$$

$$\rho_\lambda^\beta\left(\beta^{(i)}\right) = \lambda\|\beta^{(i)}\|_1.$$

## 5 Simulation study

To test the performance of personalized regression, we measure the recovery of personalized parameters on simulated data. For fixed $X \in \mathbb{R}^{N\times P}$, we generate sample-specific effect size vectors $\beta^{(i)} \sim$ *Unif* $(0,1)$ and sample $Y^{(i)} \in \{0,1\}$ according to a logistic regression model. The covariates $U^{(i)}$ are generated by projecting $\beta^{(i)}$ into $K < P$

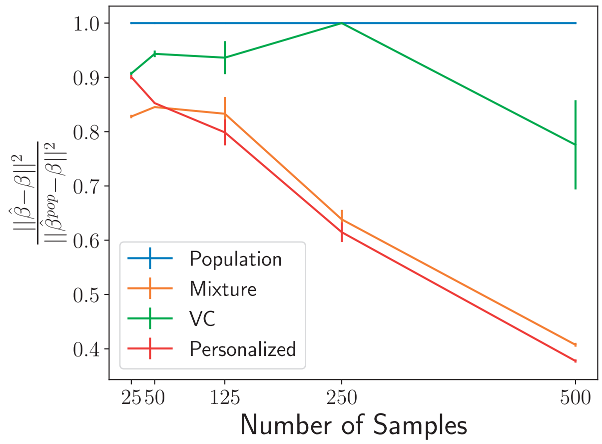


**Fig. 1.** Recovery of regression parameters for the simulated data described in Section 5. Values indicate the mean error of the personalized parameter matrix normalized to the performance of the population estimator and averaged over 20 data generation processes, with error bars to denote the variance. The personalized model struggles at extremely low sample sizes but quickly surpasses the performance of the baseline models

dimensions by multi-dimensional scaling. This produces covariates that are related to the personalized regression coefficients in a highly nonlinear, nonparametric manner. Recovery of the ground truth effect size vectors for fixed $P = 10$ variables and $K = 3$ covariates is depicted in Figure 1. In general, the personalized model outperforms other baselines except when the sample size is small, which is to be expected.

## 6 Sample-specific pan-cancer analysis

Here, we investigate the potential for personalized cancer analysis. We use gene expression (RNA-Seq) quantification data from The Cancer Genome Atlas (TCGA). This dataset compiles data from 37 projects spanning 36 disease types in 28 primary sites. After pruning for missing values, this dataset contains 9663 profiles for 8944 case and 719 matched control samples; we divide this set into 75% training data and 25% testing. While this full dataset is sizable, previous analyses have been hampered by the small number of samples for each particular cancer sub-type (e.g. there are only 36 cases present in the bile duct cancer dataset). Because our framework of personalized regression allows models to share information across diverse settings, we are able to jointly analyze the cancer subtypes while still recovering subtype-specific characteristics. The number of samples available from each dataset was shown in Table 1.

We subsample genes based on annotations in the COSMIC Catalogue of Somatic Mutations in Cancer (Forbes *et al.*, 2015), so that there is exactly one putatively causal gene for each 5 non-annotated genes. This resulting in $P = 4123$ features when an intercept term is added. We train each logistic regression model to predict the case/control status of each sample with $\ell_1$ regularization to perform variable selection in order to study which genes are relevant for classification. Our baseline models include: $\ell_1$-regularized logistic regression model trained on all pan-cancer data ('Population'), $\ell_1$-regularized logistic regression model trained on each primary tissue type ('Tissue-Population'), $\ell_1$-regularized mixture model with the number of clusters equal to the number of tissue types in the pan-cancer dataset ('Mixture'), a logistic regression model with parameters that follow a linear varying coefficients model ('VC'), and the mixed model recently proposed by Hayeck *et al.* (2015).

**Table 2.** Classification errors

| Model | Train error (%) | Test error (%) |
|---|---|---|
| Population | 6.9 | 6.8 |
| Tissue-population | 6.5 | 6.8 |
| Mixture | 6.7 | 6.8 |
| VC | 7.5 | 8.7 |
| LMM | 7.0 | 7.1 |
| Personalized | **6.3** | **6.7** |

Bold indicates the best performing model.

In addition to the RNA-seq data, we used the following 14 covariates: disease type, primary tumor site, age of the patient at diagnosis, year of birth of the patient, the number of days to sample collection, gender of the patient, race of the patient, percent of neutrophil infiltration, percent monocyte infiltration, percent normal cells, percent tumor nuclei, percent lymphocyte infiltration, percent stromal cells and percent tumor cells in the sample. These covariates span a range of different types, including both continuous and discrete values; for continuous-valued covariates, we use the $\ell_1$ distance function, for discrete-valued covariates, we use the discrete distance metric. For the VC model, unordered discrete covariates such as primary tissue must be converted into one-hot vectors. This procedure increases the number of covariate features to 64, underscoring the benefit of our model's ability to directly use the 14 unordered, discrete covariates without modification.

To predict case/control status of each sample, we implemented the personalized logistic regression model with Lasso regularization described in Section 4.2. We selected $\lambda$ in the population estimator by 10-fold cross-validation on the training set. This value of $\lambda$ is held fixed between the population estimator and the personalized estimator. Next, we set $\gamma$ so that the loss due to the distance matching regularizer is similar in magnitude to the prediction loss. Finally, we set $\upsilon$ and $\alpha$ so that the loss due to distance metric regularization is one order of magnitude smaller than the logistic classification loss. This heuristic represents our uncertainty in the form of personalization for cancer; we prefer to rely on the data than to set a rigid form of personalization. Empricially, we observe robustness in the solutions up to an order of magnitude change in these hyperparameters. By inspecting the variables (mRNA transcripts) selected by this method, we find that personalized regression identifies (i) individualized genetic aberrations, (ii) interpretable patterns of differentiation and (iii) patient sub-typing that is more meaningful than clustering based on covariate data.

### 6.1 Predictive accuracy

To verify accuracy of the model, we first examine the classification loss of the case/control status target. Although our main goal is to study the selection of important genes for this task, the overall classification error is a convenient benchmark for sanity checking the learned models. Training and testing error values are shown in Table 2 with testing error rates calculated using $n\_neighbors = 3$, as described in Algorithm 1. For the Tissue-Population model, we report the sample-weighted mean performance of the tissue-specific models. We see that the predictive accuracy for both training and testing sets is meaningfully improved by this method of personalization. We expect a low training error by virtue of the large number of parameters in the personalized models; the low testing error indicates that personalized patterns of differentiation are generalizable throughout the patient cohort and that the learned distant metrics are effective at finding related samples at test-time.

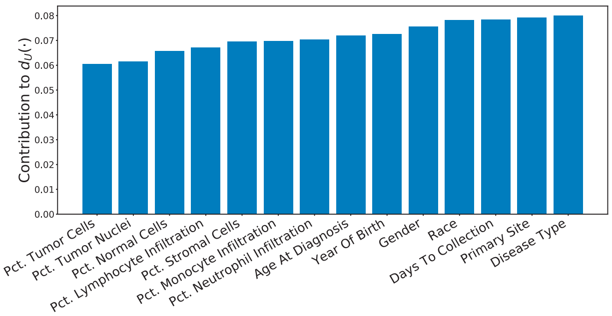


**Fig. 2.** Contribution of each covariate to the learned personalization distance in the pan-cancer dataset. We see that, as expected, this method learns to upweight differences in disease type and primary site, along with other demographic features

## 6.2 Personalization effects

We also examine the learned distance metrics for contributions to personalization by each covariate. The linear form of the distance metric makes interpretation of $\phi_U$ straightforward by inspection of the loadings (Fig. 2). As expected, the disease and primary tissue site of the sample have the heaviest influence on personalization, confirming our intuition that the variation between cell types is highest in cells of distinct differentiations. Next in importance to $\phi_U$ are demographic and clinical features, which may be interpreted as a coarse-grained view of the patient's SNPs. Important molecular markers of cancer subtype appear to be (a) percent of neutrophil infiltration, (b) percent monocyte infiltration and (c) percent stromal cells, confirming clinical findings these phenotypic characteristics as indicative of molecular subtypes, especially in breast cancers (Dennison *et al.*, 2016; Livasy *et al.*, 2006; Isella *et al.*, 2015).

## 6.3 Accurate recovery of personalized parameters

Personalized regression selects variables on a sample-specific level. Such fine-grained analytic power, unobscured by cohort averaging, enables more accurate recovery of important features than is possible by population-scale models. As a result, the number of variables selected for each sample-specific model is much lower than the number of variables selected by the population estimator (Fig. 3, top). In addition, the number of samples for which each variable is selected follow a long-tailed distribution in which a few genes are selected for many samples, but many genes are selected for a few samples (Fig. 3, bottom). The set of common gene selections represents well-studied oncogenes that are common to many types of cancer while the infrequently selected genes may correspond to less common oncogenes.

To investigate this possibility of many infrequently selected oncogenes, we further examine the oncogene distribution by rank of variable. Ranks are calculated by ordering the sums of the magnitudes of each coefficient along the sample axis (for population models, this is simply the magnitude of the coefficient associated with that variable). In this way, the rank captures both the number of samples for which the variable was selected and the magnitude of the implied effect size. As shown in Figure 4, the overlap between selected genetic markers and the annotations in COSMIC (Forbes *et al.*, 2015) is improved by the process of personalization. We see that the highly ranked oncogenes are efficiently selected by nearly all methods, but the performance of the baseline models lags as the rank diminishes. In particular, although the Tissue-Population models that are learned independently using only samples from a given tissue tend to select highly ranked genes that are also annotated in

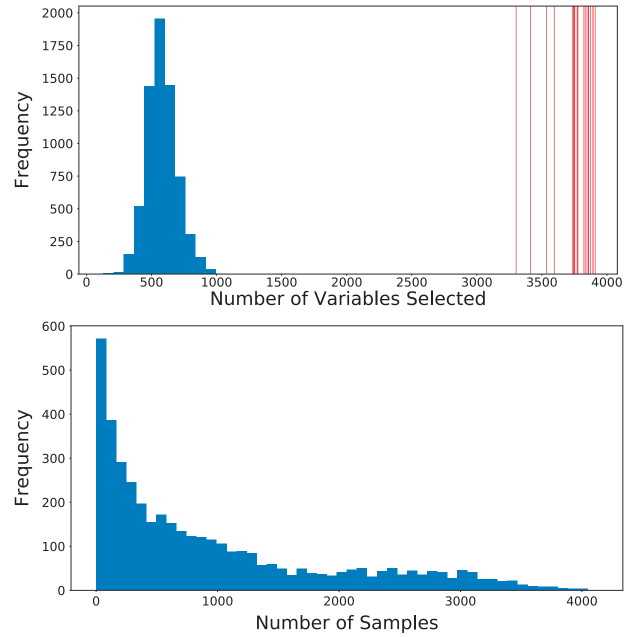


**Fig. 3.** The sample-specific variable selection of personalized regression results in models with fewer selected variables than those selected by population-level models. (Top) Histogram of the number of variables selected for each patient by personalized regression. Vertical red lines indicate the number of variables selected by the Tissue-Population model trained on a single cancer type. Personalized models achieve similar or improved predictive performance with fewer selected genes. (Bottom) Histogram of the number of samples for which each gene is selected

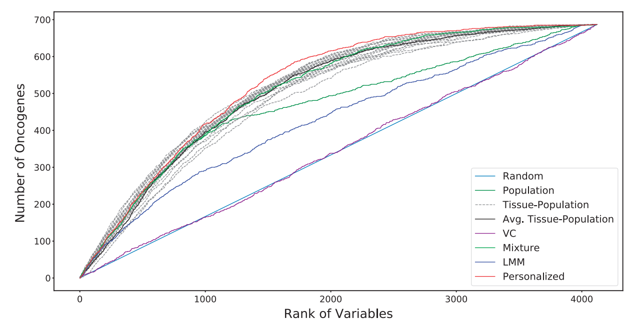


**Fig. 4.** Overlap of selected variables with annotated oncogenes (best viewed in color). Results for each tissue-specific model are displayed in dashed gray lines, with the sample-weighted mean displayed in a solid black line. We see that the personalized models select oncogenes at higher ranks than do the baseline methods, especially for the long tail of low rank oncogenes

COSMIC, the performance in the long-tail of infrequently selected genes is less competitive compared to the personalized model. This confirms the intuition that personalization is the most useful in this latter regime.

To test whether this increase in oncogene selection is due to novel identification of genetic processes, we perform enrichment analysis of the ranked lists of genes. Reported in Table 3 are the most significant Gene Ontology (GO) terms from a ranked enrichment test using Panther 13.1 (Mi *et al.*, 2017) on the Panther GO-SLIM Biological Process dataset (Mi and Thomas, 2009) with a cutoff of $P < 0.05$ for the Bonferroni-corrected $P$-values. The genes selected by personalized models are enriched with similar GO terms compared to the baseline models, which is expected since the gene ontology is largely comprised of well-studied annotations from large

**Table 3.** Enrichment analysis of complete variable rankings

| Model | Biological process | *P*-value |
|---|---|---|
| Population | mRNA processing | 2.06e-8 |
| | DNA metabolic process | 3.18e-6 |
| | Organelle organization | 3.86e-2 |
| Tissue-Population | mRNA processing | 3.09e-9 |
| | Metabolic process | 3.26e-5 |
| | Transcription, DNA-dependent | 9.61e-5 |
| | DNA metabolic process | 5.9e-3 |
| Mixture | mRNA processing | 1.45e-8 |
| | DNA Metabolic process | 1.96e-5 |
| | Transcription, DNA-dependent | 2.62e-4 |
| | Organelle organization | 7.32e-3 |
| VC | None | NA |
| LMM | DNA metabolic process | 2.02e-2 |
| Personalized | mRNA processing | 5.83e-6 |
| | Metabolic process | 1.1e-3 |
| | DNA metabolic process | 3.15e-2 |

cohorts as opposed to harder to detect personalized effects. This validates our hypothesis that the improved performance of variable selection is not due to identification of a single group of genes, but rather is due to the identification of many sample-specific effects.

## 6.4 Discovery of molecular subtypes

The pattern of selection of genes is of particular interest for clinical application. As seen in Figure 5, there are a number of common oncogenes that are repeatedly selected throughout many cancer types, including FOXA1, HOXC13 and FCGR2B. This set combines with a sparse selection of a number of oncogenes specific to each cancer type. These cancer types span surface-level characteristics such as tissue type. Interestingly, we also see a small set of rarely selected oncogenes that are consistently selected for a cluster of about 300 patients (outlined in Fig. 5). This set of oncogenes is highly over-represented for the GO biological process term ‘Modulation of Chemical Synaptic Transmission’ (Bonferroni corrected *P*-values of 2.32e-2), which includes genes ATP1A2, SLC6A4, ASIC1,

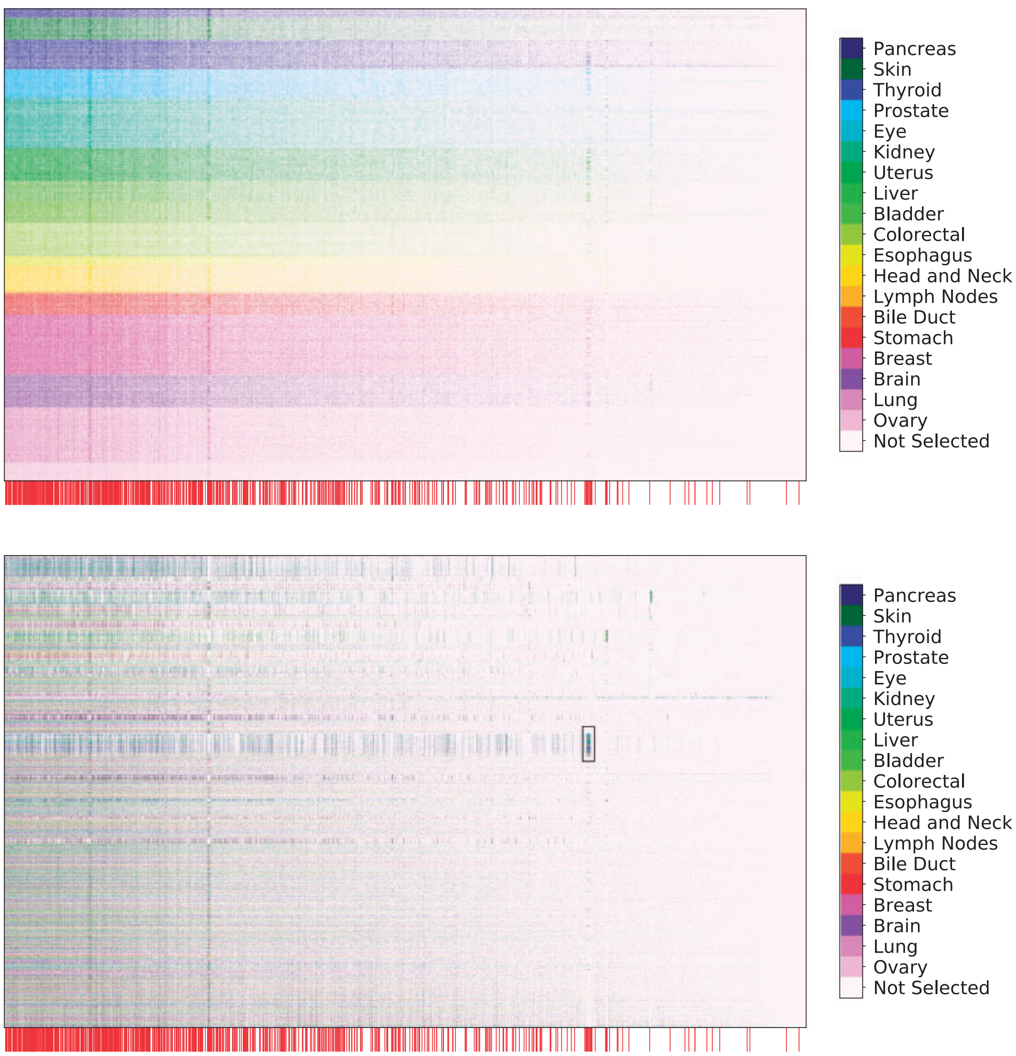


**Fig. 5.** Selection of genetic markers as predictive of case/control status from a pan-cancer dataset. The horizontal axis denotes genes while the vertical axis indexes samples. Selected variables in each row are colored by the primary tumor site of the sample, with unselected variables colored white. We observe consistent selection of a number of common oncogenes throughout all cancer types along with the sparse selection of a small number of oncogenes specific to each cancer type. Genes annotated as oncogenes in the COSMIC census are marked by a red line along the horizontal axis (zoom in for more detail as these lines may be difficult to differentiate on some screens). (Top) Rows ordered by primary tissue site, (Bottom) Rows clustered according to personalized variable selection. The boxed region is analyzed in Section 6.4

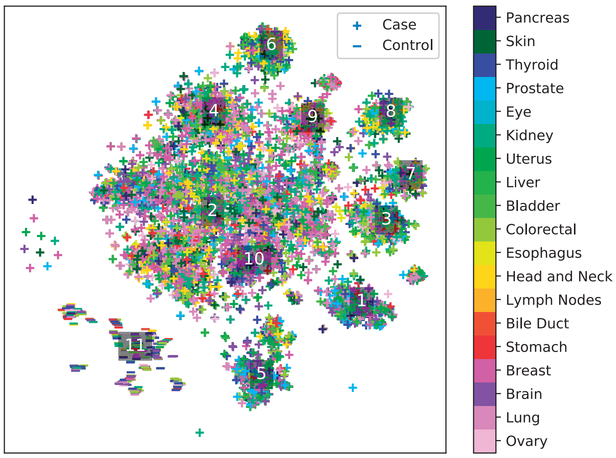


**Fig. 6.** tSNE projection of personalized regression parameters learned from a pan-cancer dataset. Each point represents a single sample with color indicating primary tumor site and marker type indicating case/control status of the patient. Labelled points indicate the centroids of clusters analyzed in Table 4

GRM3 and SLC8A3. These genes code for ion-transport processes, which have long been seen in vivo as an important system in thyroid cancer (Filetti *et al.*, 1999) and in vitro from leukemic cells (Morgan *et al.*, 1986), but only recently been appreciated as a functional marker across many different cancer types (Scafoglio *et al.*, 2015).

Figure 6 depicts a tSNE projection of the learned effect vector for each sample, colored by the primary tumor site. While the samples appear to form clusters, and the case samples are separated from the control samples by a large margin, again these clusters do not appear to correspond to any individual covariate. This complexity of personalization underscores the need for learned distant metrics to capture relationships corresponding to molecular characterization of tumors.

To identify molecular subtypes, we cluster the parameter embeddings using the HDBSCAN algorithm and perform an enrichment analysis of each cluster's variable selection in an analogous manner to the procedure described in Section 6.3. The top 3 over-enriched leaf terms from the GO biological process dataset are shown in Table 4. We see that the different clusters of models correspond to different biological processes. For instance, cluster 3 is enriched for several terms associated with extracellular interactions, while cluster 2 emphasizes terms associated with nucleotide modification via splicing and repair. These results suggest that the clusters discovered by personalized regression may correspond to clinically meaningful molecular subtypes.

**Table 4.** Enrichment analysis of tumor clusters

| Cluster | Biological process | *P*-value |
|---|---|---|
| 1 | Symbiont process | 2.62e-3 |
| | Regulation of cellular catabolic process | 1.96e-2 |
| | Protein modification process | 3.43e-2 |
| 2 | DNA repair | 3.21e-12 |
| | RNA splicing, via transesterification | 3.64e-7 |
| | Reactions with bulged adenosine as nucleophile | |
| | DNA replication | 1.00e-6 |
| 3 | Symbiont process | 1.4e-3 |
| | Antigen processing and presentation of peptide antigen | 1.06e-2 |
| | Antigen processing and presentation of exogenous antigen | 1.08e-2 |
| 4 | DNA metabolic process | 3.83e-8 |
| | DNA repair | 1.68e-6 |
| | Regulation of cellular macromolecule biosynthetic process | 5.06e-6 |
| 5 | Plasma membrane bounded cell projection morphogenesis | 1.45e-2 |
| | Neuron projection development | 3.02e-2 |
| 6 | mRNA catabolic process | 8.78e-4 |
| | Gene expression | 6.02e-4 |
| | Macromolecule biosynthetic process | 3.32e-2 |
| 7 | None | N/A |
| 8 | Generation of precursor metabolites and energy | 4.75e-5 |
| | Oxidation-reduction process | 4.52e-5 |
| | Citrate metabolic process | 9.84e-3 |
| 9 | DNA metabolic process | 3.96e-10 |
| | Cellular response to DNA damage stimulus | 5.57e-9 |
| | Protein complex subunit organization | 1.41e-4 |
| 10 | DNA metabolic process | 7.15e-8 |
| | ncRNA metabolic process | 1.33e-4 |
| | Chromatin organization | 8.27e-4 |
| 11 | Negative regulation of phosphorylation | 3.74e-2 |
| | Hematopoietic or lymphoid organ development | 4.46e-2 |

## 7 Conclusions and future work

In this work, we have presented a framework for estimating sample-specific regression models via the introduction of a novel regularizer that matches distance in covariate values to distance in regression parameters. We have demonstrated the effectiveness of this paradigm for sample-specific tumor analysis by gene selection on a pan-cancer dataset. Much work remains to be done in the application of this method to cancer analysis. We are particularly interested in the potential to uncover novel molecular subtypes that correspond to shared mutational patterns of tumors, especially for analysis of the long tail of understudied genetic factors. In addition, we would like to apply this paradigm of sample-specific estimation to more complicated models. With the increasing number of biological assays for precise granularity buoyed by the rising tide of genomic data availability, we anticipate sample-specific modeling to continue to increase in importance and relevance to the bioinformatics community.

## Acknowledgements

We thank Maruan Al-Shedivat, Avinava Dubey and Michael Kleyman for insightful discussion, and anonymous reviewers for constructive criticism.

## Funding

This work was supported by NIH R01 GM114311-02.

*Conflict of Interest*: none declared.

## References

Alaa,A.M. *et al.* (2016) Personalized risk scoring for critical care prognosis using mixtures of gaussian processes. *arXiv preprint arXiv: 1610.08853*.

Dennison,J.B. *et al.* (2016) High intratumoral stromal content defines reactive breast cancer as a low-risk breast cancer subtype. *Clinical Cancer Res.*, **22**, 5068–5078.

Fan,J. and Zhang,W. (1999) Statistical estimation in varying coefficient models. *Ann. Stat.*, **27**, 1491–1518.

Filetti,S. *et al.* (1999) Sodium/iodide symporter: a key transport system in thyroid cancer cell metabolism. *Eur. J. Endocrinol.*, **141**, 443–457.

Fisher,R. *et al.* (2013) Cancer heterogeneity: implications for targeted therapeutics. *Br. J. Cancer*, **108**, 479.

Forbes,S.A. *et al.* (2015) Cosmic: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.

Hastie,T. and Tibshirani,R. (1993) Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**, 757–796.

Hayeck,T.J. *et al.* (2015) Mixed model with correction for case-control ascertainment increases association power. *Am. J. Hum. Genet.*, **96**, 720–730.

Isella,C. *et al.* (2015) Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.*, **47**, 312.

Kolar,M. *et al.* (2009) Sparsistent learning of varying-coefficient models with structural changes. In: *Advances in Neural Information Processing Systems*, pp. 1006–1014.

Kuijjer,M.L. *et al.* (2015) Estimating sample-specific regulatory networks. *arXiv preprint arXiv: 1505.06440*.

Kumar-Sinha,C. and Chinnaiyan,A.M. (2018) Precision oncology in the age of integrative genomics. *Nat. Biotechnol.*, **36**, 46.

Liu,X. *et al.* (2016) Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res.*, **44**, e164–e164.

Livasy,C.A. *et al.* (2006) Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. *Modern Pathol.*, **19**, 264.

Lopez-Martinez,D. and Picard,R. (2017) Multi-task neural networks for personalized pain recognition from physiological signals. *arXiv preprint arXiv: 1708.08755*.

Marusyk,A. *et al.* (2012) Intra-tumour heterogeneity: a looking glass for cancer? *Nature Rev. Cancer*, **12**, 323.

Maurer,A. *et al.* (2013) Sparse coding for multitask and transfer learning. In: *ICML (2)*, pp. 343–351.

Mi,H. and Thomas,P. (2009) *PANTHER Pathway: An Ontology-Based Pathway Database Coupled with Data Analysis Tools*. Humana Press, Totowa, NJ, pp. 123–140.

Mi,H. *et al.* (2017) Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189.

Moon,H. *et al.* (2007) Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artif. Intell. Med.*, **41**, 197–207.

Morgan,K. *et al.* (1986) Release of a sodium transport inhibitor (inhibitin) from cultured human cancer cells. *Cancer Res.*, **46**, 6095–6100.

Parikh,A.P. *et al.* (2011) Treegl: reverse engineering tree-evolving gene networks underlying developing biological lineages. *Bioinformatics*, **27**, i196–i204.

Pittman,J. *et al.* (2004) Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl. Acad. Sci. USA*, **101**, 8431–8436.

Roth,A. *et al.* (2014) Pyclone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396.

Scafoglio,C. *et al.* (2015) Functional expression of sodium-glucose transporters in cancer. *Proc. Natl. Acad. Sci. USA*, **112**, E4111–E4119.

Song,L. *et al.* (2009) Time-varying dynamic Bayesian networks. In: *Advances in Neural Information Processing Systems*, pp. 1732–1740.

Weinstein,J.N. *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113.

Xing,E.P. *et al.* (2003) Distance metric learning with application to clustering with side-information. In: *Advances in Neural Information Processing Systems*, pp. 521–528.

Xu,J. *et al.* (2015) Formula: factorized multi-task learning for task discovery in personalized medical models. In: *Proceedings of the 2015 International Conference on Data Mining*. SIAM.

Yamada,M. *et al.* (2016) Localized lasso for high-dimensional regression. *STAT*, **1050**, 20.