Personalized Regression Enables Sample-Specific Pan-Cancer Analysis

Benjamin J. Lengerich, Bryon Aragam, Eric P. Xing
{blengeri, naragam, epxing}@cs.cmu.edu
@ben_lengerich, @itsrainingdata

Cancer is Complex

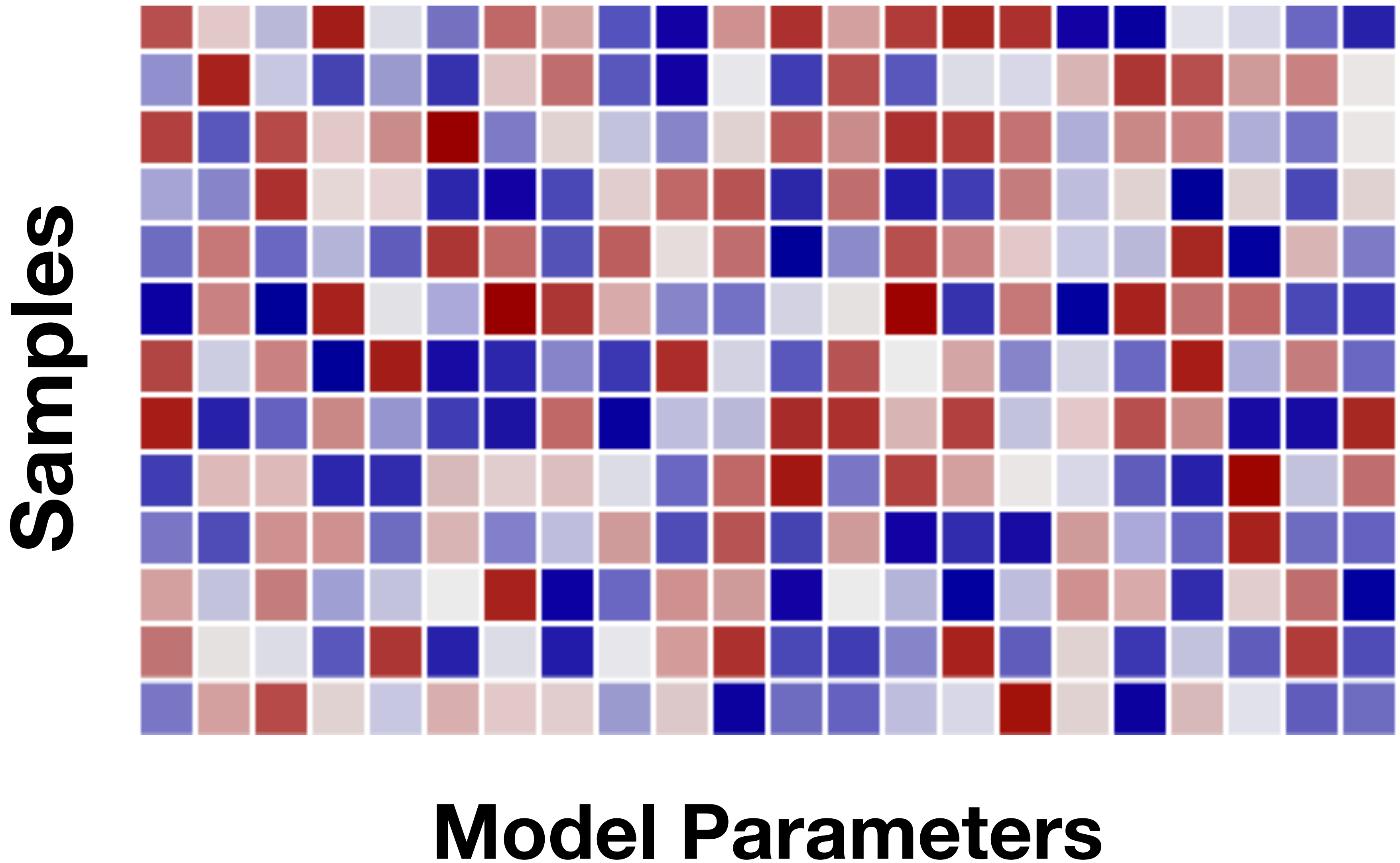
- Different mutations can cause similar phenotypes.
- There are many possible driver mutations.
- Do we need to build a *single* model that works for *all* cancers?
- Could we build a different model for each type of cancer?
  - But cancer “type” may not correspond to any single clinical covariate.

The Extreme: Sample-Specific Models

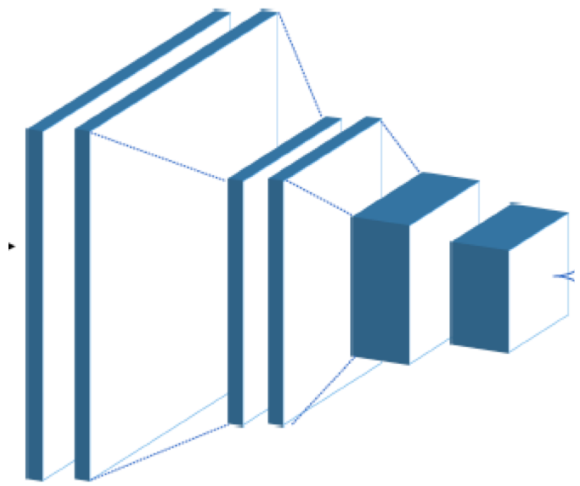
- What if we try to understand tumors one at a time?
- Could we use ***simple*** models that each work for a ***single*** patient?
    - Enable new types of questions to be asked: “How does this tumor’s model differ from the cohort’s?”

Our Goal

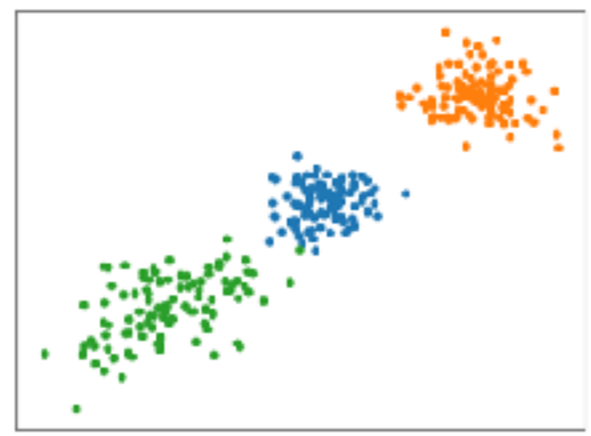
Sample-Specific, ***Pan-Cancer*** Models:



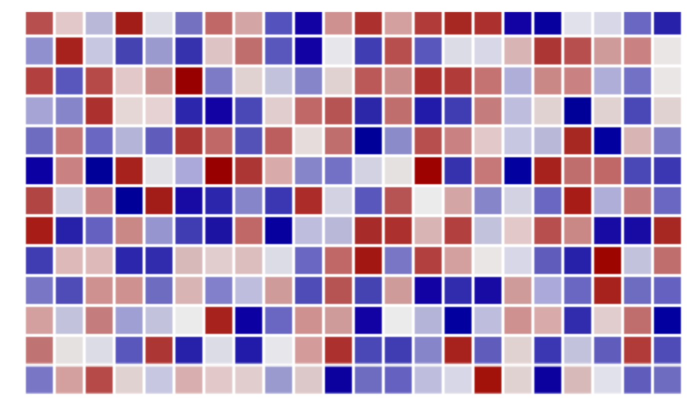
Why Sample-Specific Models?



Deep Learning
Mixed Effects



Mixtures



Sample-Specific
Varying-Coefficient

Universal Effects → **Personal Effects**

Complicated Effects ← **Simple Effects**

"Self-driving cars"

"This tumor is due to a mutation in gene TP53"

Why Pan-Cancer Models?

- Share information between ***rare*** and ***common*** cancer types
- Uncover molecular subtypes
- If we can handle clinical covariates well, tissue type can be simply treated as another covariate

Tissue	n	Tissue	n
Breast	1,092	Ovary	376
Lung	1,016	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		

Number of Samples by Tissue Type in TCGA[1]

1. cancergenome.nih.gov

Related Work

	Sample-Specific Models?	Unknown Covariate Effects?	General Framework?
Varying-Coefficient [1]	✓	✗	✓
Known Structure [2,3,4]	✓	✗	✗
Sample-Specific Network Estimation [5,6]	✓	✓	✗
Personalized Regression	✓	✓	✓

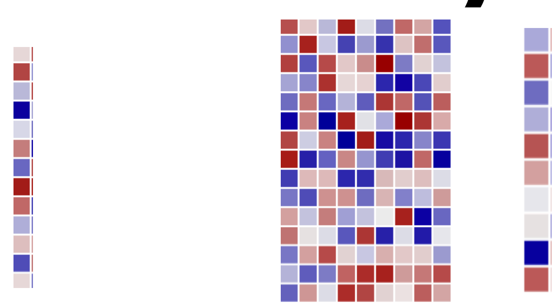
1. *Hastie and Tibshirani. Journal of the Royal Statistical Society 1993*
2. *Song et al. NIPS 2009, 3. Kolar et al. NIPS 2009, 4. Parikh et al. ISMB 2011*

5. Kuijjer et al. Arxiv 2015, 6. Liu et al. Nucleic Acids Research 2016

Personalized Regression

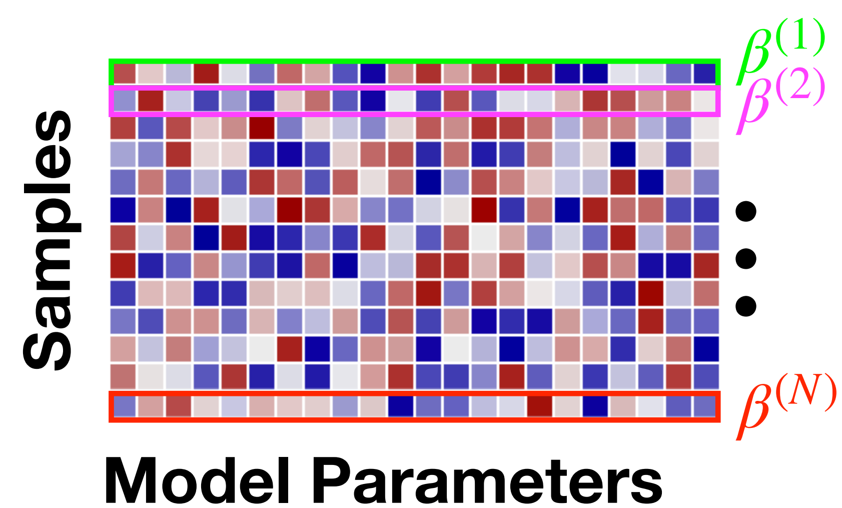
- From estimating a single model:

$$Y = X\beta^T + \epsilon$$



- To estimating sample-specific models:

$$Y^{(i)} = X^{(i)}\beta^{(i)T} + \epsilon^{(i)}$$



Overparameterized, but not hopeless!

Personalized Regression

- Define the sample-specific loss functional to be minimized:

$$\mathscr{L}(\beta; d_\beta, d_U) \propto \sum_{i=1}^{N} \mathscr{L}^{(i)}(\beta^{(i)}; d_\beta, d_U)$$

$$\mathscr{L}^{(i)}(\beta^{(i)}; d_\beta, d_U) \propto \boxed{\underbrace{f(X^{(i)}, Y^{(i)}, \beta^{(i)})}_{\textbf{Prediction Loss}} + \underbrace{\rho_\lambda^\beta(\beta^{(i)})}_{\textbf{Regularization}}} + \boxed{\underbrace{\varrho_\gamma^{(i)}(d_\beta, d_U)}_{\textbf{Distance-Matching}}}$$

Overparameterized, but not hopeless!

Distance Matching Regularization

- **Main idea:** Distance between sample **parameters** should be similar to distance between sample **covariates**.

- Define a regularization loss functional to be minimized:

$$\varrho_{\gamma}^{(i)}(d_\beta, d_U) = \gamma \sum_{j \neq i} \Big(\underbrace{d_\beta(\beta^{(i)}, \beta^{(j)})}_{\textbf{parameter distance}} - \underbrace{d_U(U^{(i)}, U^{(j)})}_{\textbf{covariate distance}} \Big)^2$$

Pairwise distances between all samples

Distance Metrics Can Be Learned From Data

- Define distance metrics as linear combinations of feature-wise distance metrics:

$$d_\beta(x, y) = \left[|x_1 - y_1|, \ldots, |x_P - y_P| \right] \phi_\beta^T$$

$$d_U(x, y) = \left[d_{U_1}(x_1, y_1), \ldots, d_{U_K}(x_K, y_K) \right] \phi_U^T$$

- After optimization, we can inspect the values in ϕ_β , ϕ_U to understand contributions to personalization.

- User must supply covariate-specific distance metrics.
  - Can use complicated covariate distance metrics.

When is Personalized Regression Useful?

- We are seeking a model for **inference**, not necessarily most accurate predictive model.
- We are seeking relatively simple **personalized** effects, not complex **universal** effects.
- We have covariate data which is informative of each sample.

Experiments

TCGA Pan-Cancer Analysis

- Model: Logistic Regression with Lasso Regularization
- Task: Predict Case/Control Status
- Data:
    - 28 primary sites
    - 9663 samples (8944 case, 719 control)
    - 4123 RNA-Seq features
    - 14 clinical covariates

Tissue	n	Tissue	n
Breast	1,092	Ovary	376
Lung	1,016	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		

Number of Samples by Tissue Type in TCGA[1]

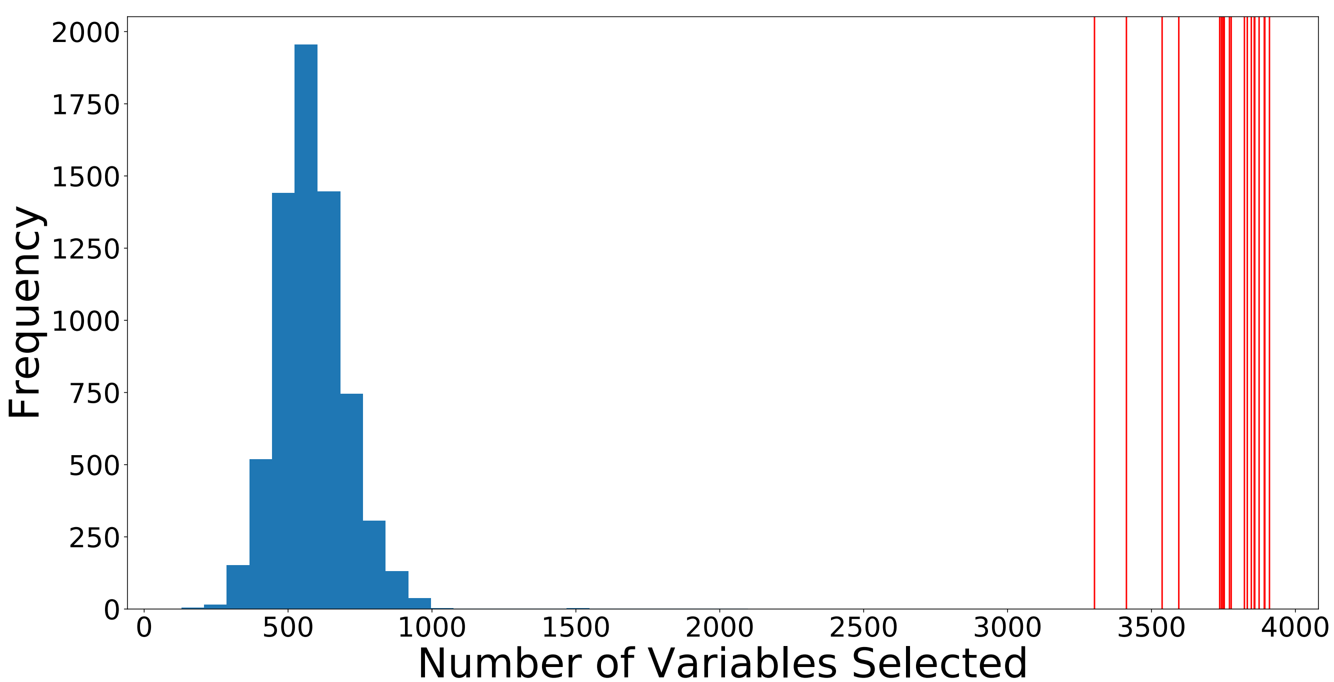
1. cancergenome.nih.gov

Clinical Covariates

- 14 Clinical Covariates:
  - **Tissue Features**: Disease Type, Primary Site, Days to Collection
  - **Sample Molecular Biomarkers**: Pct. Tumor Cells, Pct. Normal Cells, Pct. Tumor Nuclei, Pct. Lymphocyte Infiltration, Pct. Stromal Cells, Pct. Monocyte Infiltration, Pct. Neutrophil Infiltration
  - **Patient Demographic Features**: Age at Diagnosis, Year of Birth, Gender, Race
- Traditional methods expect these data encoded as one-hot vectors, which expands dimensionality 5X!

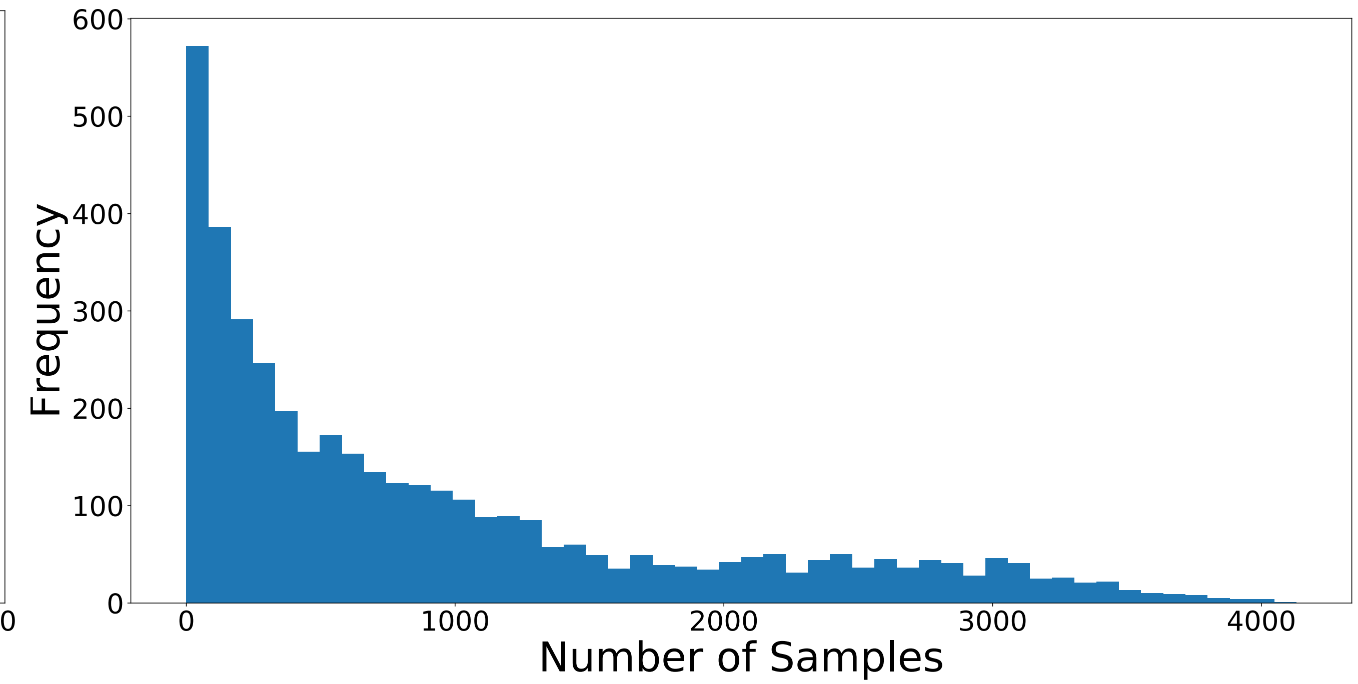
Personalized Models Are More Efficient with Variable Selection

Selects Fewer Genes Per Sample:



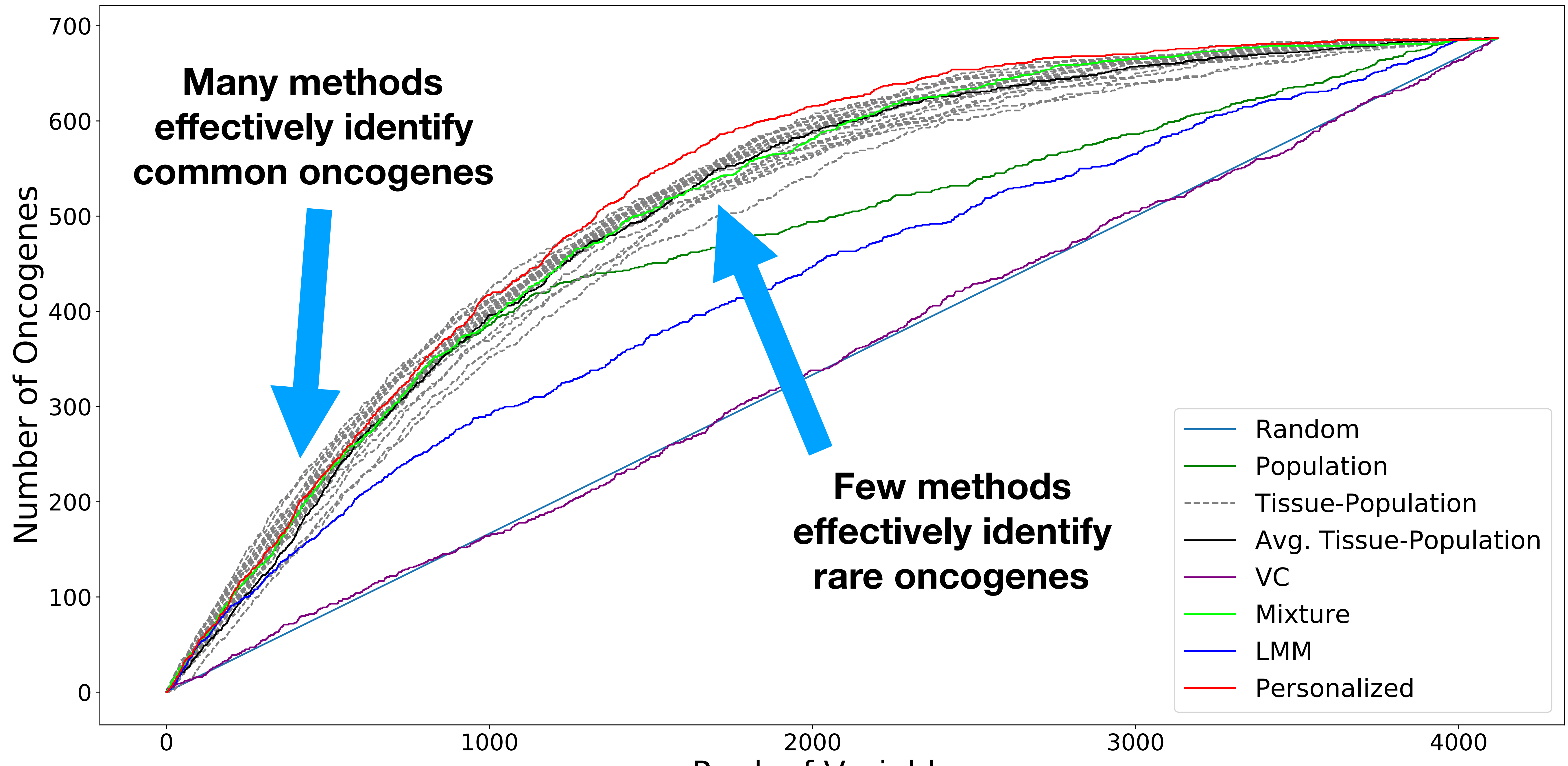
Red Lines Indicate Number of Variables Selected by Tissue-Specific Models

Uses each Gene in Fewer Samples:



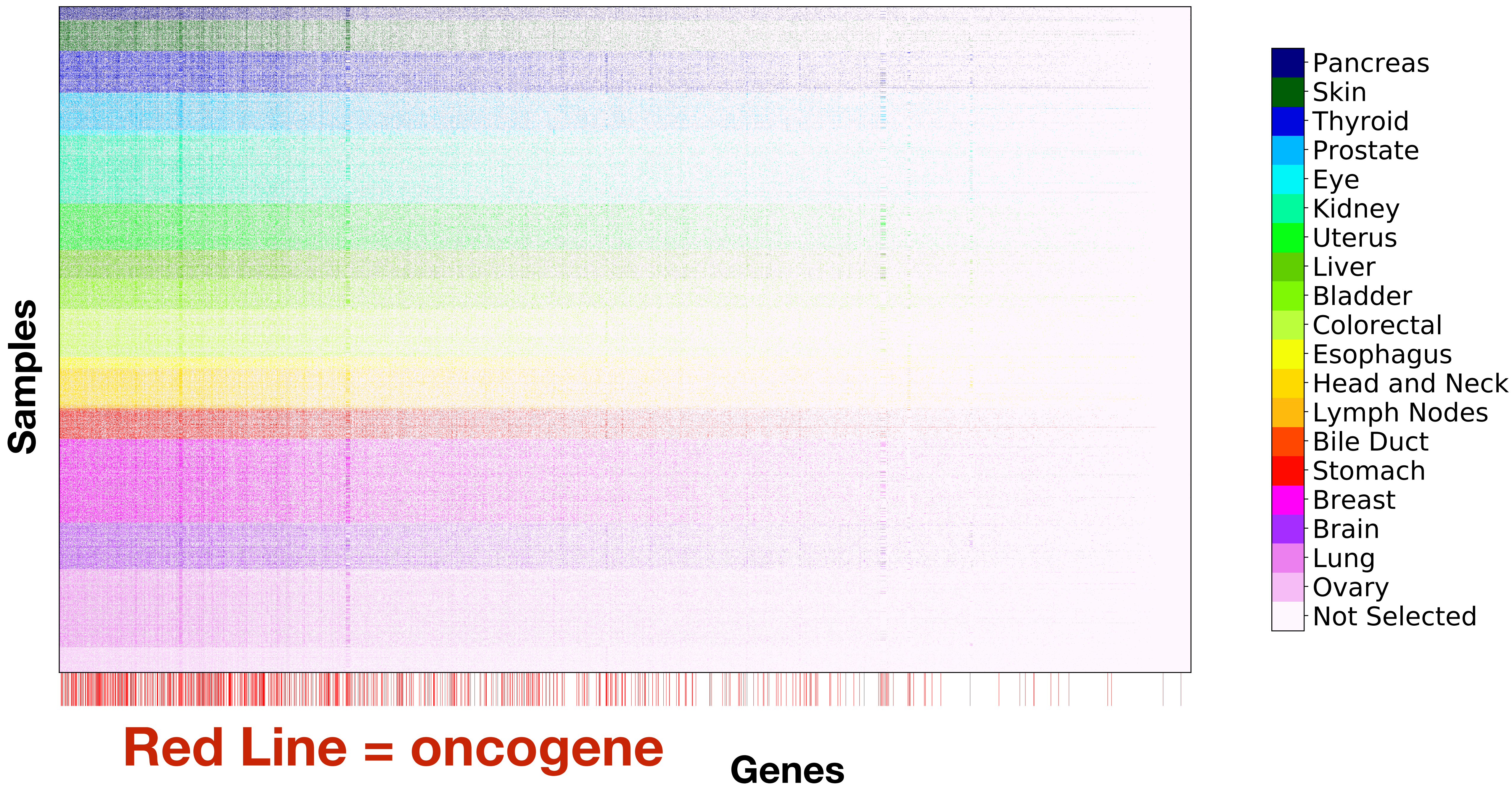
Most Genes are Selected for Fewer than 500 Samples

Personalized Regression Gives More Weight to Known Oncogenes [1]



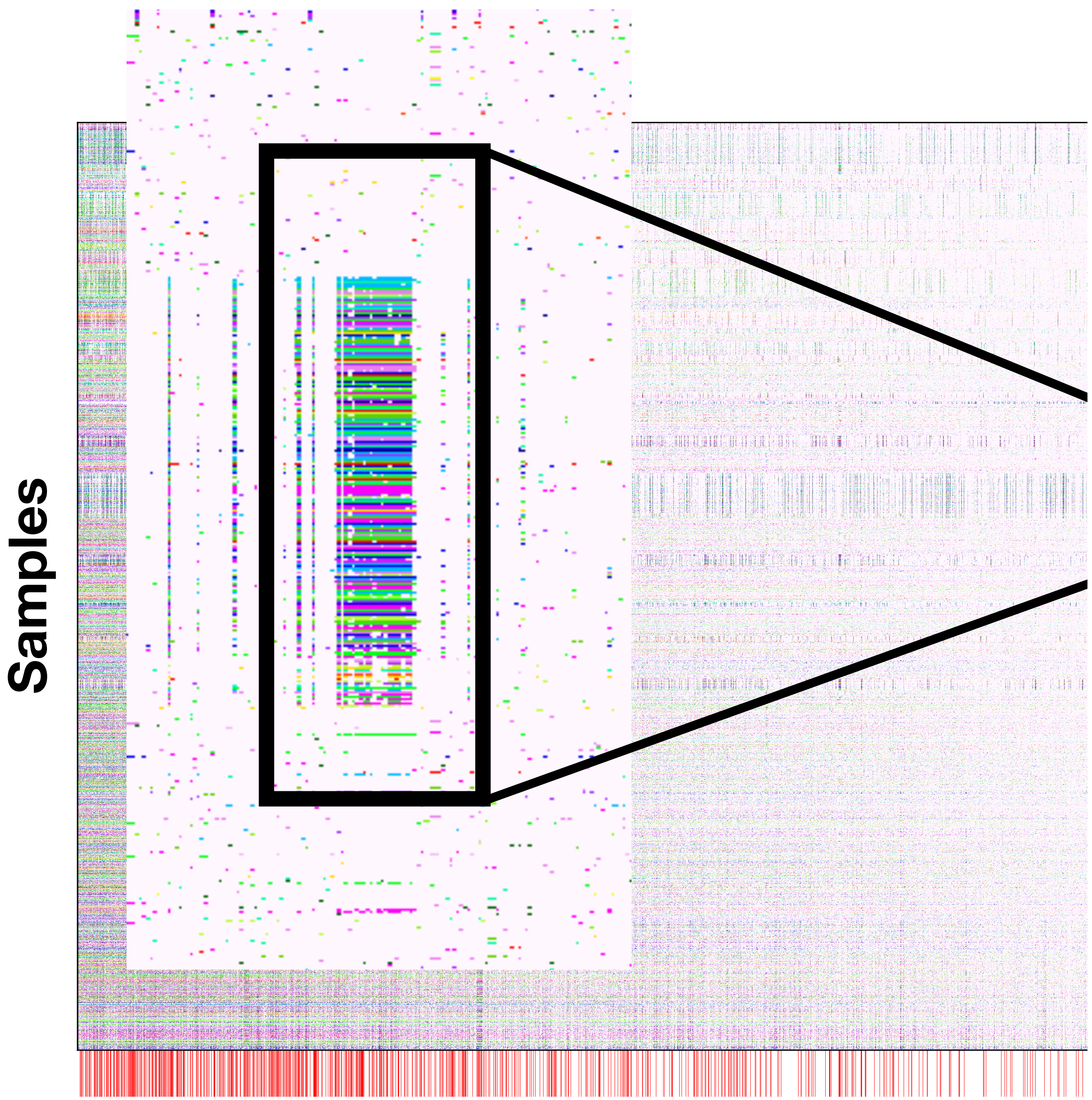
1. Oncogenes as annotated in COSMIC (Forbes et al. Nucleic Acids Research 2014)

Personalized Regression Produces Sample-Specific Pan-Cancer Models



Red Line = oncogene

Personalized Models Reveal Molecular Subtypes Which Span Tissues



- Over-represented for the GO biological process term “Modulation of Chemical Synaptic Transmission" (p <0.05FDR)
    - Includes genes ATP1A2, SLC6A4, ASIC1, GRM3, and SLC8A3, which code for **ion-transport processes**.
    - Ion-transport processes have long been seen in vivo as an important system in thyroid cancer [1] and in vitro from leukemic cells [2], but only recently as a functional marker across different cancer types [3].

1. *Filetti et al. European Journal of Endocrinology 1999*
2. *Morgan et al. Cancer Research 1986*
3. *Scafoglio et al. PNAS 2015*

Models Form
stinct Signatures

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 Reactions with Bulged Adenosine as Nucleophile	3.64e-7
		.00e-6

Extracellular Processes - Antigen

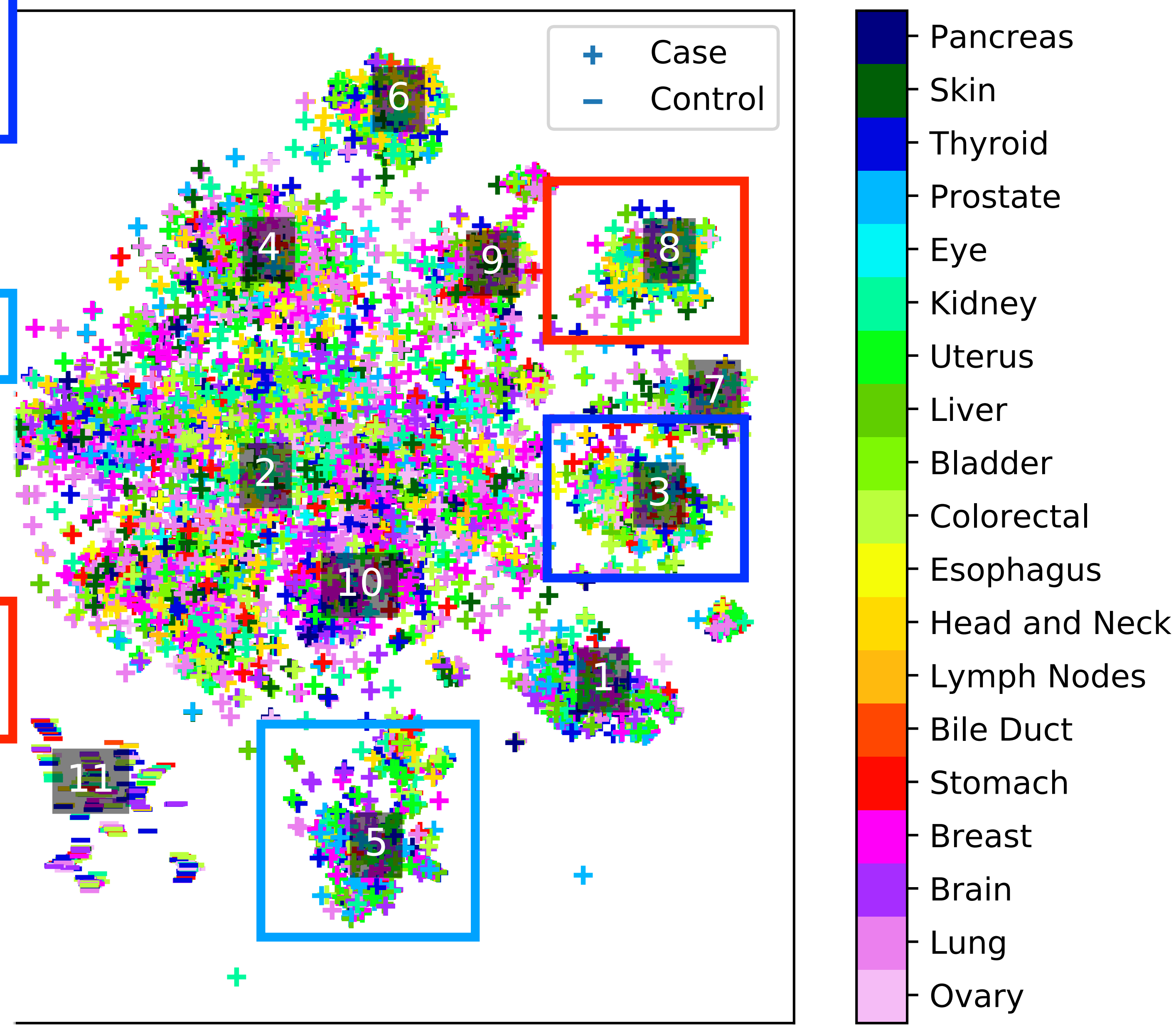
Cluster	Biological Process	p-value
3	Symbiont Process	1.4e-3
	Antigen Processing and Presentation of Peptide Antigen	1.06e-2
	Antigen Processing and Presentation of Exogenous Antigen	1.08e-2
4	DNA Metabolic Process	3.83e-8
	DNA repair	1.68e-6
		e-6

Extracellular Processes - Membrane

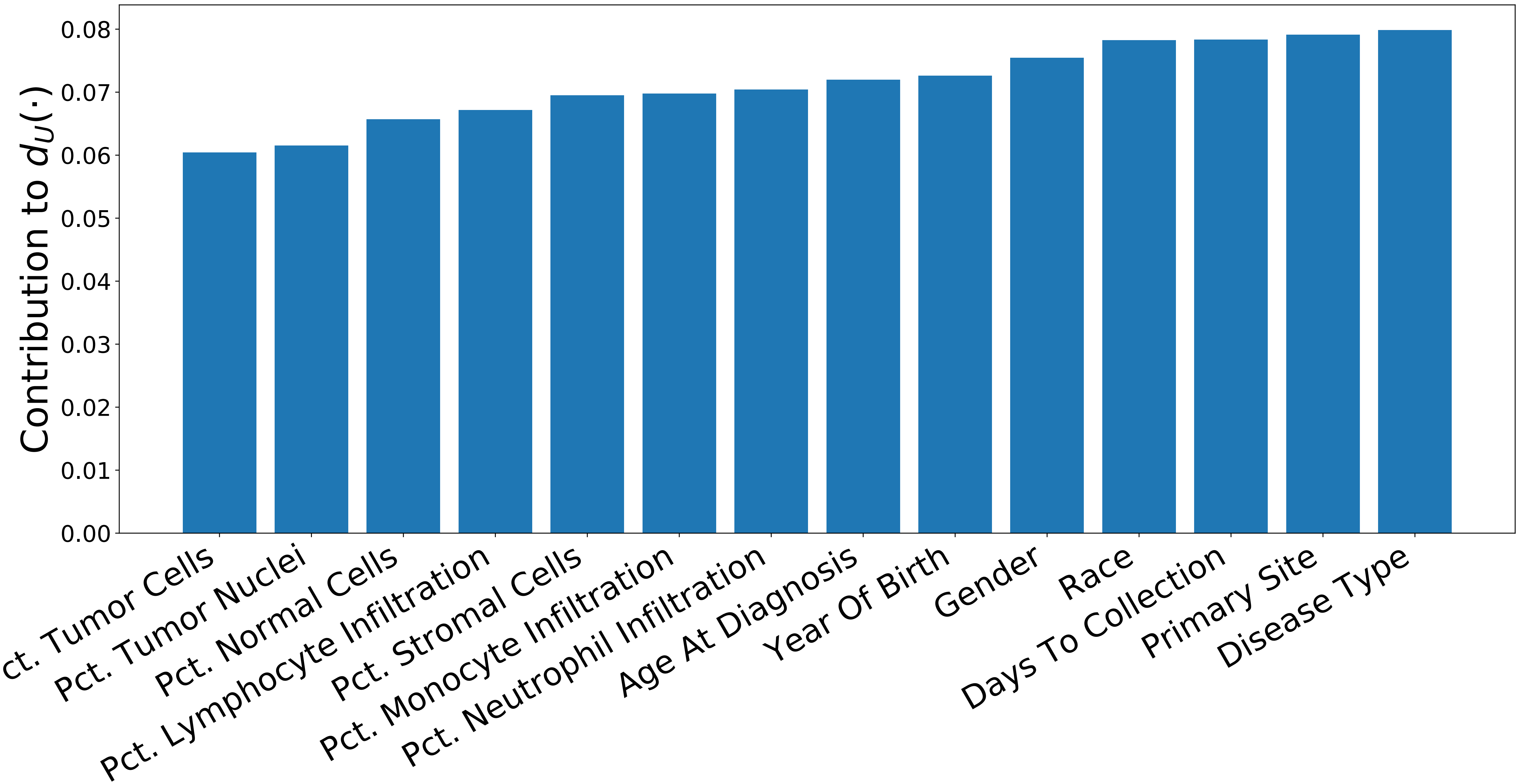
Cluster	Biological Process	p-value
5	Plasma Membrane Bounded Cell Projection Morphogenesis	1.45e-2
	Neuron Projection Development	3.02e-2
6	mRNA Catabolic Process	8.78e-4
	Gene Expression	6.02e-4
	Macromolecule Biosynthetic Process	3.32e-2
7	None	N/A

Cellular Metabolism

Cluster	Biological Process	p-value
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



Personalized Regression Learns Clinical Distance Metrics



Conclusions

- Sample-specific models can give us a new perspective.
    - Unlock bottom-up in addition to traditional top-down analyses.
- ***Personalized Regression*** with ***Distance-Matching Regularization*** effectively learns sample-specific models.
- ***Personalized Regression*** reveals patterns in pan-cancer transcriptomic data that are overlooked by traditional analyses.

Future Work

- Biological Questions - Sample-Specific Processes?
- More complex personalized models
- Personalized Regression for Single-Cell Data, Election Modeling, Stock Prediction

Thank You

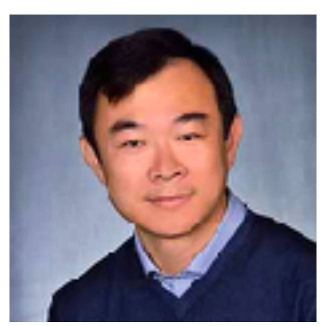
Code available at:

github.com/blengerich/personalized_regression

Collaborators:

- Bryon Aragam

- Eric P. Xing



- Contact: {blengeri, epxing}@cs.cmu.edu

Travel to ISMB generously supported by ISCB

Research supported by NIH

The Gory Details

Personalized Regression: Optimization

- Define pairwise distance vectors by:

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

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

- Construction of the covariate distance tensor can be amortized

Avoiding Degenerate Solutions

- Add priors to distance metrics
- From:

$$\varrho_{\gamma}^{(i)}(d_{\beta}, d_{U}) = \gamma \sum_{j \neq i} \Big(\underbrace{d_{\beta}(\beta^{(i)}, \beta^{(j)})}_{\text{parameter distance}} - \underbrace{d_{U}(U^{(i)}, U^{(j)})}_{\text{covariate distance}} \Big)^2$$

- To:

$$\varrho_{\gamma}^{(i)}(d_{\beta}, d_{U}) = \gamma \sum_{j \neq i} \Big(\underbrace{d_{\beta}(\beta^{(i)}, \beta^{(j)})}_{\text{parameter distance}} - \underbrace{d_{U}(U^{(i)}, U^{(j)})}_{\text{covariate distance}} \Big)^2 + \psi_{\alpha}(d_{\beta}) + \psi_{\upsilon}(d_{U})$$

Avoiding Degenerate Solutions

- Add priors to distance metrics

$$\varrho_{\gamma}^{(i)}(d_\beta, d_U) = \gamma \sum_{j \neq i} \Big(\underbrace{d_\beta(\beta^{(i)}, \beta^{(j)})}_{\text{parameter distance}} - \underbrace{d_U(U^{(i)}, U^{(j)})}_{\text{covariate distance}} \Big)^2 + \psi_\alpha(d_\beta) + \psi_\upsilon(d_U)$$

- where

$$\psi_\alpha(d_\beta) = \alpha || \phi_\beta - \phi_{beta}^0 ||^2$$

$$\psi_\upsilon(d_U) = \upsilon || \phi_U - \phi_U^0 ||^2$$

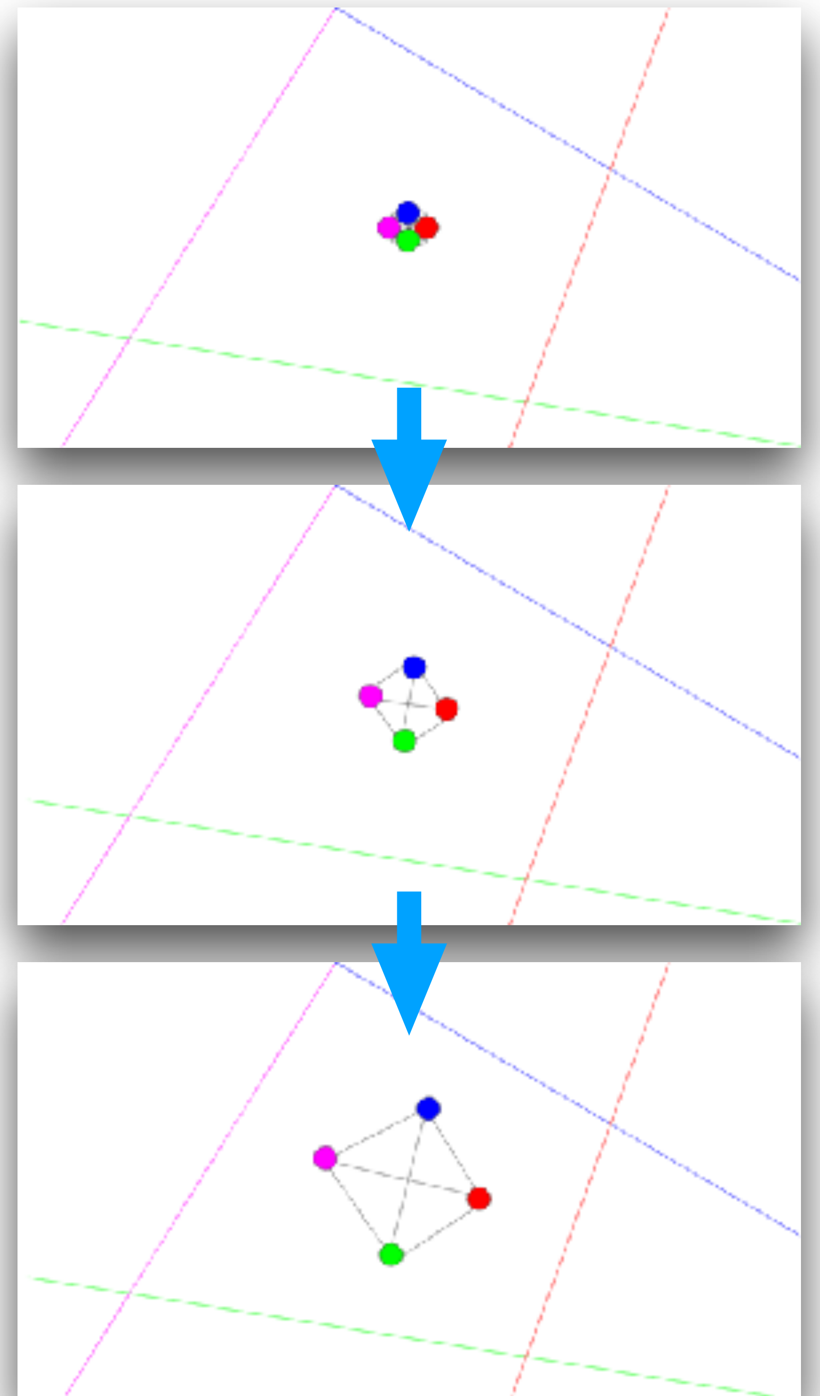
- and we project loadings into the non-negative reals.

Personalized Regression

- Initialize at population solution

- Allow each personalized model to “fine-tune” away from the central population solution (block coordinate descent)

- **Distance-matching regularization** ensures the personalized models respect covariate structure



Inference Procedure

- Conveniently, we have already learned distance metrics to use for predictions.
- On test data, we identify the closest neighbors and use their sample-specific models.

Algorithm 1 Inference Procedure

Require: Test point $(X^{(test)}, U^{(test)})$, predictive model $p(\cdot,\cdot)$, number of nearest neighbors m

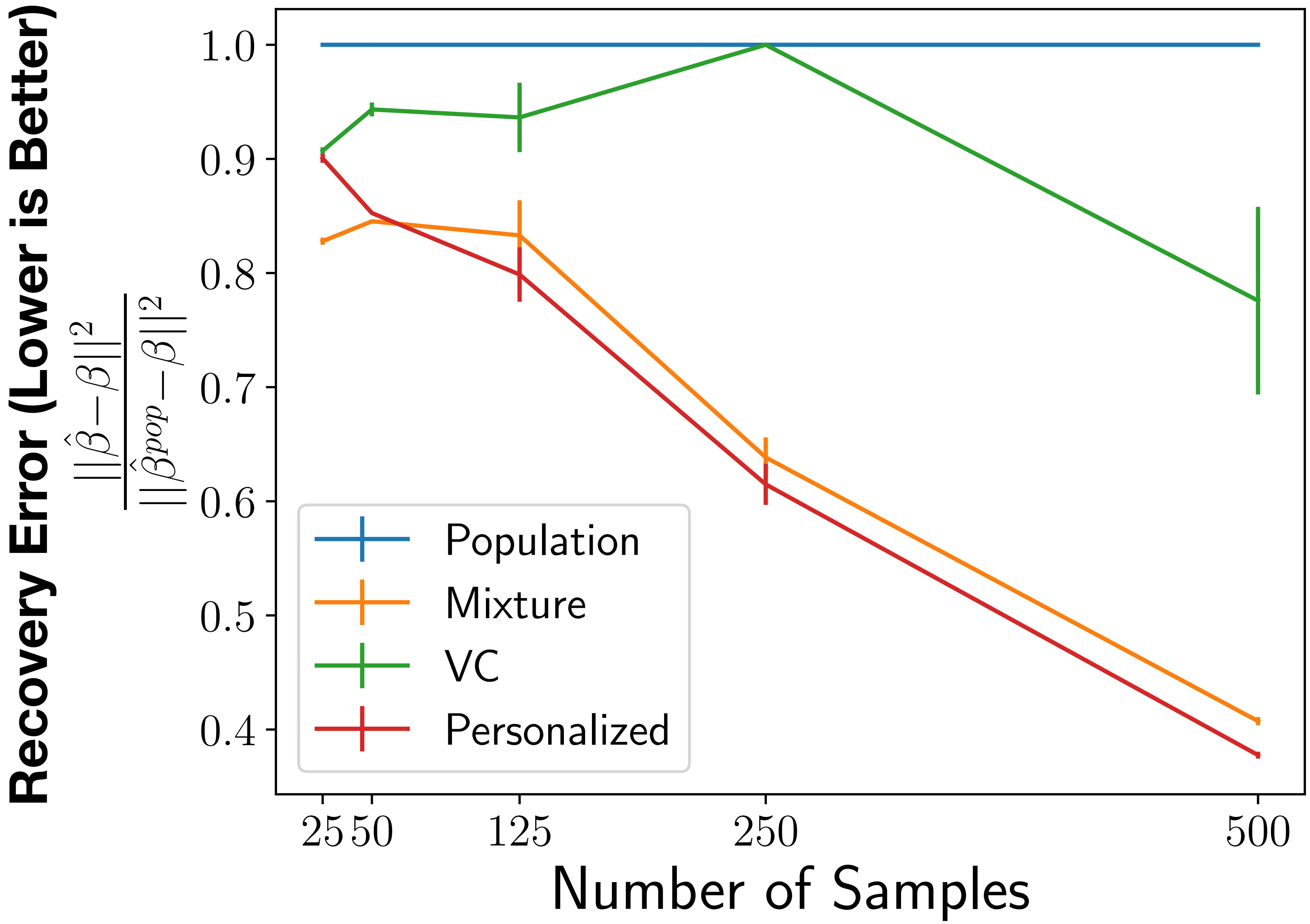
$distances \leftarrow \{d_U(U^{(test)}, U^{(i)}) : i \in [1, \ldots, N_{train}]\}$

$neighbors \leftarrow \text{sort}(distances)[0:m]$

$\beta^{(test)} \leftarrow \text{mean}(\{\beta^{(i)} : i \in neighbors\})$

return $p(X^{(test)}, \beta^{(test)})$

Simulation Results



- At moderate sample sizes, personalized regression recovers parameters well.
- At low sample size, cannot learn distance metrics.

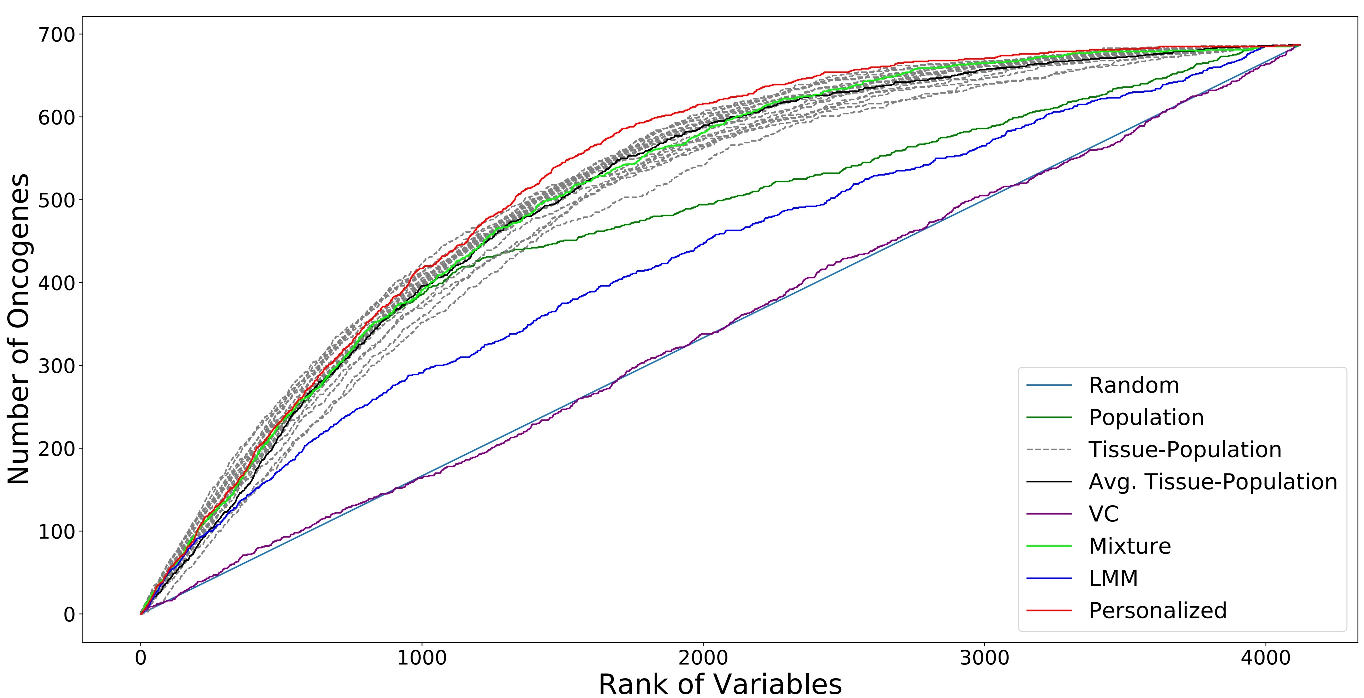
Personalized Regression Fine-Tunes Accuracy

- Here, personalized regression overfits the data but is still better than competing methods.
- Better clinical distance metrics and hyperparameter tuning will likely alleviate overfitting.

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**

Personalized Regression Does Not Merely Identify More Enriched Gene Sets

Enrichment Analysis of Complete 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

Instead, it identifies a variety of sample-specific patterns which do not fit into a small number of mixtures