The impact of intratumor single-cell heterogeneity on breast cancer prognosis prediction

Fraunfelder, M.*, Chen, Y.*, Al-Mahroos, M.*,
*These authors contributed equally to this work

Abstract

The management of breast cancer by physicians requires a balanced dance between patient-specific treatment regimens and prediction of tumor stage trajectory. Clinics need better tools to predict the prognosis of a given breast cancer; however, given the inherent heterogeneity of the disease, training clinically significant prognosis prediction models has proven difficult. Here we identify specific measures of breast cancer intratumor heterogeneity markers and discuss the utility of heterogeneity as a predictor of prognosis. Ultimately, we demonstrate multiple measures of heterogeneity across a patient cohort and set up future studies to train and test a prognosis prediction model using our unique, heterogeneity-based feature inputs.

Introduction

Breast cancer affects a large population of individuals in the United States and the world. Breast cancer is a highly heterogenous disease that is clinically classified into three primary subtypes based on the dominant expression of a given receptor in the characteristic neoplastic epithelial cells: progesterone receptor (PR), estrogen receptor (ER), and human epidermal growth factor receptor-2 (HER2)[1,2,3]. After the explosion of microarray methods that improved our ability to analyze the tumor microenvironment, the PAM50 method identified five additional "intrinsic" subtypes based on a given tumor's transcriptomic and proteomic profiles: Luminal A, Luminal B, HER2+ enriched, normal-like, and basal-like[2,3].

Given the high variation in the clinical presentation of breast cancer, one of the challenges in the treatment of breast cancer is understanding individual tumor responses to a variety of treatment options, including chemotherapy and immunotherapy, and subsequently predicting the prognosis of a patient based on the clinical markers of their tumor[1,2,3]. Clinical treatment decisions must consider both predicted treatment response, treatment impact on quality of life, and how these two values intersect with predicted patient prognosis. Thus, there is a need for prognosis prediction models to evolve as our understanding of breast cancer does.

Despite the recent progress in microarray methods and transcriptomic analysis, building prediction prognosis models requires an understanding of the wide range of factors, or "features", that affect tumor progression and severity; however, we currently lack a robust understanding of all possible biologically significant features due to the high spread and variability in these features throughout the patient population[1,3]. Genetic mutations, protein expression profiles, secreted cytokines, infiltrating immune cell populations, and even stromal cell identities can vary across patients and even within individual tumors, and the specific role of intratumor heterogeneity (IH) in tumor treatment response and prognosis is currently unclear[1,3,4,5,6,8,9]. We hypothesize that by quantifying IH across multiple features of single-cell RNA sequencing (scRNAseq) data, we might generate unique feature inputs that can further classify tumor subtype and be used to train a novel prediction model that considers transcriptomics, neoplastic cell characteristics, and clinical subtype alongside heterogeneity of the tumor micro-environment.

To test our hypothesis, we identified three features of scRNAseq data to wholistically characterize a single human breast tumor. Our chosen analysis methods – scSubtype, inferCNV, and gene module analysis – are adapted from (Wu et al, 2021), where scSubtype provides single

cell classification, inferCNV provides genetic mutation classification, and gene module analysis groups cells by common transcriptomic by profiles[8]. We calculated the coefficient of variation (CV) across each of these three features to represent a given patient's IH for that feature, with the intent to subsequently assign each patient a feature input and a heterogeneity input in a prognosis prediction model[9]. In this study we establish three unique measures of IH that we hope can be used to train a multinomial logistic regression model against patient prognosis data and reveal any links between patient responses to therapy and the inherent variation within the tumor microenvironment.

Methods

Data

We used the 26 individuals from the Wu et al study to build our feature analysis methods[8]. Datasets can be downloaded from the following link: GSE176078. Additional training data for our prediction model was compiled from GSE161529[7]. Non-cancerous control data from GSE161892 was used to train the model against control samples[10].

Features

ScSubtype. ScSubtype code was adapted from the Wu et al method, original code can be found in the linked repository: https://github.com/Swarbricklab-code/BrCa_cell_atlas[8]. In short, our pipeline can be summarized as follows (Figure 1):

1. Sparse matrix files are integrated and converted into an RDS object for Seurat to read.
2. A Seurat object is created and scaled.
3. The scSubtype gene signatures to compare the data against are loaded into the environment.

4. Mean scSubtype scores based on the comparison to the gene signatures for each cell in the dataset are calculated.
5. A series of normalization and processing steps.
6. Highest scSubtype call is calculated and the output is two separate text files, scores.txt and highestcalls.txt.

Genomic Instability. Genomic instability scores of neoplastic epithelial cells were calculated using inferCNV. Stromal and immune cells were used as the non-neoplastic reference group. Cells were annotated using the Garnett method, neoplastic cells were identified using inferCNV (v0.99.7), and genomic instability scores and CV scores were calculated R (Figure 2). The pipeline is describe here in short:

1. Cell types were annotated using the Garnett method to group cells based on immune, stromal, or epithelial type identities. Immune and stromal cell reads represented the reference genome due to their relatively stable genetic content.
2. Utilizing a 100-gene sliding window, all cell reads were run through inferCNV. To enhance the reliability of the analysis, genes exhibiting a mean count of less than 0.1 across the entire cell population were excluded from the dataset. Subsequently, the CNV signal was denoised using an adaptive threshold, set at 1.3 standard deviations from the mean, to minimize background noise while preserving biologically relevant signals.
3. Genomic instability scores were generated by inferCNV, and the CV for each individual was calculated using R. Detailed results, including the distribution of genomic instability scores, can be found in the Supplementary file: GIS.

ITTH & Gene Module Analysis. Transcriptomic profiles of scRNAseq data were classified using a modified gene module analysis[8]. The code for our modified method is original, however

the original method can be found here: https://github.com/Swarbricklab-code/BrCa_cell_atlas[8].

In short, our pipeline progresses as follows (Figure 3):

1. First, we performed single-cell RNA sequencing data analysis using Seurat v4.3.0 and clustered cells at five distinct resolutions (0.4, 0.8, 1.2, 1.6, and 2.0) for each patient. Next, the Seurat's findAllMarkers function was applied to determine the top 200 differentially expressed genes (DEGs) for each cluster in every patient.
2. We refined the identified DEGs using the following criteria: DEG clusters with fewer than six genes or originating from a cell cluster with fewer than six cells were excluded.
3. We employed the Jaccard index to further filter DEG clusters by computing the Jaccard similarity matrix for all gene signatures identified within each patient and subsequently removing pairs with the fewest genes from those with a Jaccard index greater than 0.75.
4. We identified 2,234 gene signatures after DEG filtering was complete. To uncover robust gene modules among these signatures, we employed consensus clustering based on the Jaccard similarity matrix using spherical k-means, as implemented in the Cola v2.4.0 package.
    a. We initially attempted to cluster gene signatures with k equal to 7 but observed that cells were predominantly assigned to only three modules. Consequently, we evaluated k values ranging from 2 to 10 and determined the optimal k using the elbow method.
    b. Additionally, we tested multiple top value methods and partition methods in combination with multiple k-values to optimize consensus clustering analysis. The cola report for parameter optimization is provided in the supplementary file: cola report.

5. Using our training dataset[8], consensus clustering identified 6 gene modules within the neoplastic cell population, and 10 gene modules within the entire cell population. We computed gene-module signature scores for each cell using AUCell v1.20.2 and assigned cells to gene modules based on the highest scaled area under the curve (AUC) score for each module.
6. Lastly, to test multiple quantifications of gene module IH, we calculated the CV, Gini-Simpson index, and entropy. Detailed information regarding the implementation and specific results can be found in the supplementary file: GMA.

Measures of heterogeneity

Coefficient of variation (CV) represents the extent of spread in relation to the mean of the population, and thus is a suitable measure of heterogeneity across given cell clusters within a single patient[9]. CV is calculated as the standard deviation divided by the mean, the quotient of which is then multiplied by 100%. All CV calculations were done in R. Other measures of heterogeneity, such as entropy and Gini-Simpson's method were also tested during study development and can be found in supplementary file: GMA.

Multinomial Logistic Regression

Multinomial logistic regression (MLR) was chosen as a classifier because of its adaptability to multiple feature inputs and a range of prediction criteria outputs, as well as its resistance to data overfitting[11,12]. A sample of MLR code was written for the purpose of this study can be found in in the following repository: IntratumorHeterogeneity. The statistics report can be found at: StatisicalAnalysis. Sample data from this model can be found in supplementary folder: MLPdata.

Results

ScSubtype intratumor heterogeneity

To incorporate the impact of tumor intrinsic subtype on breast cancer prognosis and determine the effect of intratumor intrinsic subtype heterogeneity (IISH), scSubtype – a single cell intrinsic subtyping algorithm designed as an improvement upon the preceding PAM50 method – was used to type all scRNAseq datasets[2,3,8]. After normalizing and annotating the reads from a single patient sample, the scSubtype pipeline assigns four "scSubtype Scores" (S-scores) between -1 and 1 to each cell (Figure 4). The S-score represents the similarity or dissimilarity of any given cell's transcriptomic profile to that of four primary intrinsic subtypes: Luminal A (LumA), Luminal B (LumB), HER2-enriched (HER2-E), and basal-like (Basal). Here, an S-score of -1 represents negligible similarities between the cell's transcriptomic profile and the given subtype, and a score of +1 represents perfect alignment of said cell's transcriptomic profile and the given subtype.

After S-scores are determined, the scSubtype pipeline produces a Highest Call (HC) table that summarizes the 'dominant' subtype of each cell (Figure 5). For each individual, we use the average HC to determine the intrinsic subtype (IS) and calculate the scSubtype coefficient of variation (S-CV). The IS represents a 'standard' feature input, and the S-CV represents a measure of IISH. In Figure 6, IS for the 26 training samples is plotted, demonstrating subtype heterogeneity across tumors, whereas Figure 7 represents the variations in IISH across the 26 patients.

Ultimately the measurements of intrinsic subtype and S-CV serve as feature inputs to represent both the impact of tumor subtype and IISH on breast cancer prognosis.

Genetic instability intratumor heterogeneity

Genetic instability of cancerous tumors can be measured by multiple metrics, however a common quantification is copy number variation (CNV). CNV is a measure of the number of changes to the chromosomes of a given cell – these changes can be loss of all or part of chromosome loci, gain of additional DNA segments, and significant insertions or deletions within the genome[8]. inferCNV is a method commonly used to characterize the CNVs apparent within scRNAseq datasets by comparing 100-200 sliding scale of chromosomal loci within non-cancerous reference cells from the subject to potentially cancerous 'experimental' cells (Figure 8). We utilized inferCNV to identify neoplastic cells within each patient's dataset and determine the genetic instability score (GIS) of each cell within a given individual (Figure 9). Subsequently, we calculated the GIS-CV for each individual. Figure 10 represents the GIS-CVs across all confirmed neoplastic cells, and Figure 11 illustrates the GIS-CVs across all cell types.

The average GIS of each patient characterizes the genetic stability of each tumor, while the GIS-CV of each individual represents the intratumor neoplastic heterogeneity (ITNH).

Intratumor transcriptional heterogeneity

The comparison of single-cell transcriptomic profiles is a common method to characterize tumor function and prognosis across patients; however, as our understanding of intratumor micro-environments grows, the importance of intratumor transcriptional heterogeneity (ITTH) in tumor response to treatment and risk of metastasis has become apparent[2,3,4,7,8,10]. To characterize single-cell transcriptomic profiles, we utilized a modified version of the gene module analysis method[8]. In short, cells were clustered by expression similarities across 100-200 genes and 10 gene modules were identified based on transcriptomic

profile clustering for all cell types (T-cells, Cancer Epithelial, Myeloid, Endothelial, CAFs, PVL, Normal Epithelial, Plasma blasts, B-cells), 6 gene modules were identified based on the transcriptomic profile clustering of neoplastic cells (Cancer Epithelial, CAFs), and each cell within an individual was thus assigned a gene module classification (Figure 12). Gene module coefficient of variability (GM-CV) was then determined for each individual. Figure 13 represents the GM-CV across all confirmed neoplastic cells, and Figure 14 illustrates the GM-CV across all cell types.

This method established individual scores for each gene module to represent transcriptional characterization features, and allowed us to generate a measure of ITTH through the calculation of GM-CV.

Multinomial Logistic Regression Training with and without heterogeneity features

There are a wide range of prediction models and classifiers that have been used to characterize the relationship between single-cell transcriptomics and breast cancer prognosis. For our dataset, we chose to train a multinomial logistic regression classifier because of its adaptability to multiple feature inputs and its ability to output a range of prediction criteria[11,12].

The intention after completing the scRNAseq analyses is to train two MLR prediction models (Figure 15). For each individual in both models, the following feature inputs will be included: average scSubtype highest call, average genomic instability score, and 10 gene module scores. For Model 2, the experimental MLR, additional measures of heterogeneity will be inputted for each individual: S-CV, GIS-CV, and GM-CV. Models are to be trained against patient prognosis data and against a control dataset to prevent false-positive predictions. Additional information about our current work on the MLR code can be found here: MLPcode.

Discussion

Breast cancer is a highly heterogenous disease with micro-environment variation across patients and within individual tumors. Intratumor heterogeneity (IH) within a given patient has significant impacts on tumor response to treatment, and prognosis prediction tools are becoming increasingly valuable in both the management and research of breast cancer therapy. High resolution transcriptomic analysis provides the opportunity to comprehensively characterize multiple measures of IH across single cells within a single tumor. Here we seek to establish the significance of intratumor genetic, transcriptomic, and subtype heterogeneity in breast cancer prognosis prediction using a multinomial logistic regression model (MLR).

scRNAseq data was run through three pipelines to gather each measure of tumor classification. Intrinsic cell subtypes were determined using the novel scSubtype pipeline, genetic instability scores were calculated using inferCNV, and transcriptomic profiles were characterized using a modified gene module analysis. Heterogeneity of each characteristic was measured by calculating the coefficient of variation (CV) across all the cells within the individual. An MLR model was chosen for its adaptability to the desired range of prognosis outputs. Input features include patient intrinsic subtypes, average genetic instability scores, 10 gene module scores, and the CV values for each characteristic.

Our data demonstrate the variety of IH profiles across patients, suggesting that heterogeneity could be a useful tool to expand our classifications of breast cancer. After training and testing the MLR model we hope to see a significant relationship between intratumor heterogeneity measurements and patient prognosis, and it is possible that the the heterogeneity-based prediction model will demonstrate a low agreement with the control prediction model. Together we hope these data might suggest an important role for IH in patient prognosis, and we

hope to lend support to the notion that IH could account for the high variation in patient response to treatment observed in clinics[1,3,4,5,6,8,9]. Once we complete our prediction model, further study and a more robust training dataset will be required to comprehensively characterize the significance of IH in prognosis prediction.

There are multiple shortcomings to the design, code, and methods utilized within this study. Beyond standard code optimization concerns and the issues with the completion of our study, a primary concern is the selection of coefficient of variation (CV) as our measure of heterogeneity. Although CV is a superior representation of abstract variable spread compared to standard deviation and was selected based on its use in other cancer heterogeneity studies, CV has multiple shortcomings and may not be the optimal measure of breast cancer IH[9]. Other measures of variation, including entropy and Gini-Simpson analysis, may be more appropriate for our chosen feature sets. Further study that compares multiple types of heterogeneity-based prognosis prediction may be valuable in characterizing the most appropriate methods to classify IH.

Additionally, we must acknowledge that there may be more appropriate prediction models for our data than the multinomial logistic regression. Future study should consider additional classifier algorithms to examine the significance of IH in breast cancer, such as a convoluted neural network or a more complex logistic regression model. In choosing our prediction model, we recognize the importance of controlling for the biases not only within one's dataset and feature analyses, but within the framework of the chosen model as well.

In this study we establish a starting point to investigate the impact of genetic, transcriptional, and subtype intratumor heterogeneity on breast cancer prognosis. To address this question in the future, we intend to optimize a robust training pipeline for our prediction model

and rigorously test multiple measures of heterogeneity. Ultimately, our study represents one of the first attempts to characterize the impact of intratumor micro-environment heterogeneity on breast cancer prognosis prediction.

Author contributions

Mikayla Fraunfelder

Study design and manuscript authorship, scSubtype code, Garnett annotation method and some inferCNV code.

Mustafa Al-Mahroos

Manuscript authorship, scSubtype feature analysis, background literature, pipeline organization, training dataset collection and curation of Wu et al paper.

Yibo Chen

Manuscript authorship, gene module analysis code, inferCNV code, gene module feature analysis, inferCNV feature analysis, experimental dataset collection.

Data & Code Availability

All relevant files for both code and data can be found in our repository [https://github.com/1boChen/intratumor_heterogeneity/tree/main].

References

1. Luond, F., Tiede, S., Christofori, G. (2021). Breast cancer as an example of tumour heterogeneity and tumor cell plasticity during malignant progression. *British Journal of Cancer, 125*(2021), pp. 164-175.

2. Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. (2009). *J Clin Oncol. 27*(8), pp. 1160-1167.

3. Barry, W. T., Kernagis, D. N., Dressman, H. K., Griffis, R. J., Hunter, JV. D., Olson, J. A., Marks, J. R., Ginsburg, G. S., Marcom, P. K., Nevins, J. r., Geradts, J., Datto, M. B. (2010). Intratumor heterogeneity and precision of microarray-based predictors of breast cancer biology and clinical outcome. *J Clin Oncol., 28*(13), pp. 2198-2206.

4. Wang, F., Dohogne, Z., Yang, J. Liu, Y., Soibam, B. (2018). Predictors of breast cancer cell types and their prognostic power in breast cancer patients. *BMC Genomics, 19*(137).

5. Holm, J., Eriksson, L., Ploner, A., Eriksson, M., Rantalainen, M., Li, J., Hall, P., Czene, K. (2017). Assessment of Breast Cancer Risk Factors Reveals Subtype Heterogeneity. *Cancer Res, 77*(13), pp. 3708-3717.

6. Liu, J., Dang, H., Wang, X. W. (2018). The significance of intertumor and intratumor heterogeneity in liver cancer. *Experimental and Molecular Medicine, 50*(e416).

7. Chen, Y., Pal, B., Lindeman, G. J., Visvader, J. E., Smyth, G. K. (2022). R code and downstream analysis objects for the scRNA-seq atlas of normal and tumorigenic human breast tissue. *Scientific Data, 9*(2022), Article no. 96.

8. Wu, S. Z., Al-Eryani, G., Roden, D. L., Junakar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J. R., Bartonicek, N., Wang, T., Larsson, L., Kaczorowski, D., Weisenfeld, N. I., Uytingco, C. R., Chew, J. G., Bent, Z. W., Chan, C. L., Gnanasambandapillai, V., Dutertre, C. A., Gluch, L., Hui, M. N., Beith, J., Parker, A., Robbins, E., Segara, D., Cooper, C., Mak, C., Chan, B., Warrier, S., Ginhoux, F., Millar, E., Powell, J. E., Williams, S. R., Liu, X. S., O'Toole, S., Lim, E., Lundeberg, J., Perou, C. M., Swarbrick, A. (2021) A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics, 53*(2021), pp. 1334-1347.

9. Nguyen, A., Yoshida, M., Goodarzi, H., Tavazoie, S. F. (2016). Highly variable cancer subpopulations that exhibit enhanced transcriptome variability and metastatic fitness. *Nat Commun. 2016*(7): 11246.

10. Pal, B., Chen, Y., Vaillant, F., Capaldo, B. D., Joyce, R., Song, X., Bryant, V. L., Penington, J. S., Stefano, L., D., Ribera, N. T., Wilcox, S., Mann, G. B., kConFab., Papenfuss, A. T., Lindeman, G. J., Smyth, G. K., Visvader, J. E. (2021). A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *Embo J., 40*(11), e107333.

11. Ranganathan, P., Pramesh, C. S., Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic Regression. *Perspect Clin Res., 8*(3), pp. 148-151.

12. Konopinksi, M. K. (2020). Shannon diversity index: a call to replace the original Shannon's formula with unbiased estimator in the population genetics studies. *PeerJ. 2020*(8), e9391.

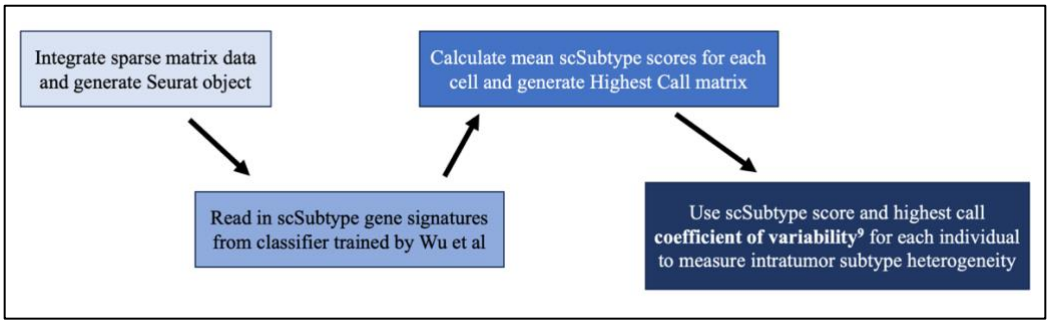

Figure 1: scSubtype pipeline. Schematic summarizing the scSubtype feature analysis method. Method adapted from Wu et al, 2021[8].

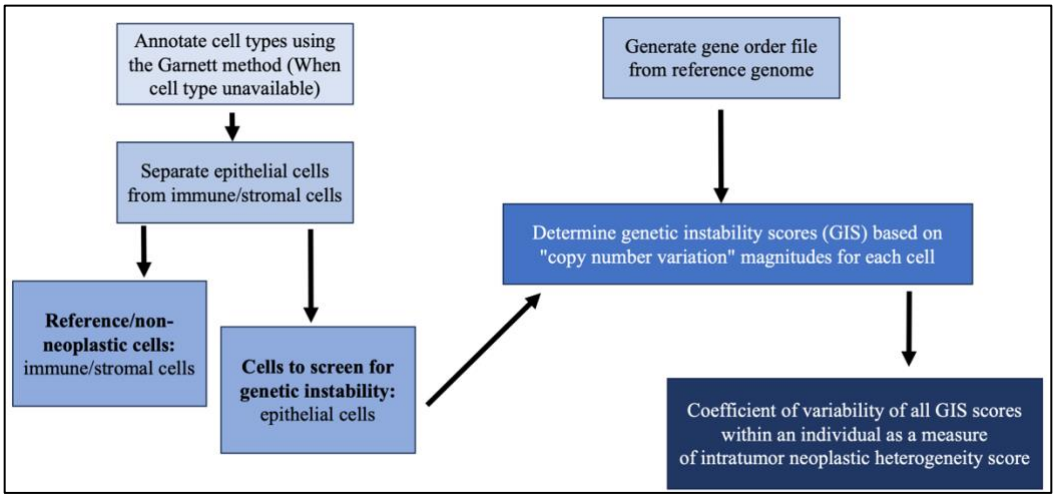

Figure 2: inferCNV pipeline. Schematic summarizing the inferCNV feature analysis method. Method adapted from Wu et al, 2021[8].

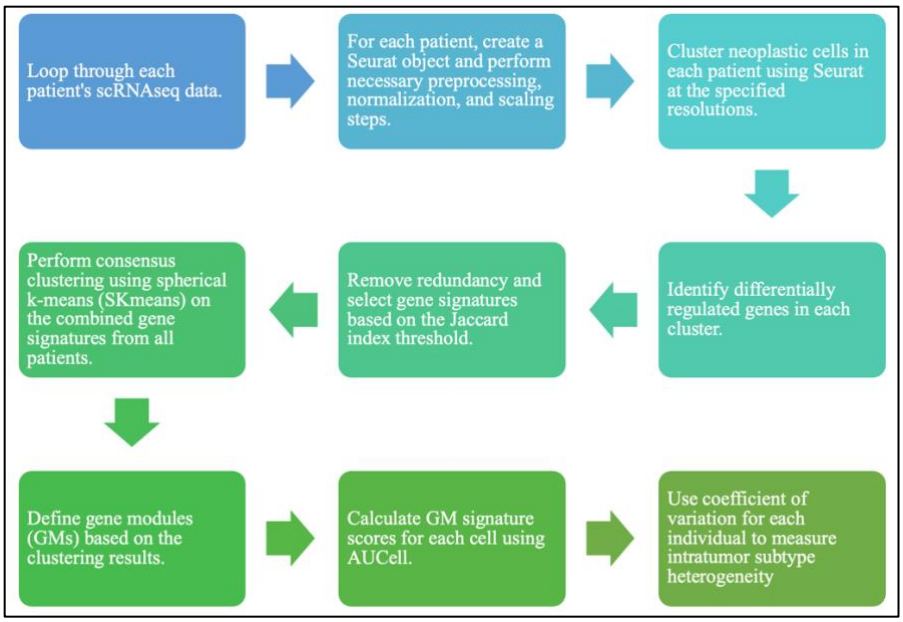

Figure 3: Gene module analysis pipeline. Schematic summarizing the gene module analysis method used to classify cell transcriptomic profiles. Method adapted from Wu et al, 2021[8].

"Basal_SC"	"Her2E_SC"	"LumA_SC"	"LumB_SC"	"SCSubtypeCall"
"CID3586_AAGACCTCAGCATGAG"	0.174360875868884	0.137992837693497	-0.0354952392757317	0.520428438972256
"LumB_SC"				
"CID3586_AAGGTTCGTAGTACCT"	-0.185088780289713	-0.103936202178561	-0.13593264046078	0.0168210895342234
"LumB_SC"				
"CID3586_ACCAGTAGTTGTGGCC"	-0.102430744814834	-0.126338944484228	-0.212838425935857	-0.147064693401202
"Basal_SC"				
"CID3586_ACCCACTAGATGTCGG"	-0.111999694799964	-0.163415543433994	-0.207691010206432	0.0080816473925833
"LumB_SC"				
"CID3586_ACTGATGGTCAACTGT"	-0.1737717322002	-0.0431016603714389	-0.164468831158179	-0.116568529626923
"Her2E_SC"				
"CID3586_ACTTGTTAGGGAAACA"	-0.0803713399394437	-0.0411808639477596	-0.125243523649613	-0.0289874203991664
"LumB_SC"				
"CID3586_AGCAGCCTCCCTCTTT"	-0.0692859918113872	-0.0116658142600138	-0.250359653482033	-0.165127330480541
"Her2E_SC"				
"CID3586_AGCTTGATCGGCGCTA"	-0.105870117404699	0.0476807215345729	-0.204554401727728	-0.0967442066619456
"Her2E_SC"				
"CID3586_ATCATCTAGGGATACC"	-0.139840647186428	-0.115594605967278	0.0591404297663801	-0.0609850515162906
"LumA_SC"				
"CID3586_ATGGGAGAGGAGCGAG"	-0.0270124556882305	-0.0557460951509101	0.0551237220554522	0.0148764022782442
"LumA_SC"				
"CID3586_ATGTGTGAGTCAATAG"	-0.1829891436712	-0.0467935320046146	-0.158440042772702	-0.0637253202332304
"Her2E_SC"				
"CID3586_ATTATCCGTCCTAGCG"	-0.105457151267233	0.0376096925416164	-0.201584733020721	-0.0408509185485331
"Her2E_SC"				
"CID3586_ATTGGACTCCTATTCA"	-0.137478110686426	0.00680550347668075	-0.130111443269417	-0.107210321063591
"Her2E_SC"				
"CID3586_CAACTAGGTGTGAAAT"	-0.113633037933326	-0.0587518604425362	-0.130114579648624	-0.101576460969693
"Her2E_SC"				
"CID3586_CAAGGCCAGTGTTTGC"	-0.134090127874958	-0.0273734191834624	-0.202370444045953	

Figure 4: scSubtype Scores output. Example table of scSubtype scores. For each cell, a score between -1 and +1 was calculated across four intrinsic subtypes: Basal, HER2-E, LumA, and LumB.

"Cell"	
"1"	"LumB_SC"
"2"	"LumB_SC"
"3"	"Basal_SC"
"4"	"LumB_SC"
"5"	"Her2E_SC"
"6"	"LumB_SC"
"7"	"Her2E_SC"
"8"	"Her2E_SC"
"9"	"LumA_SC"
"10"	"LumA_SC"
"11"	"Her2E_SC"
"12"	"Her2E_SC"
"13"	"Her2E_SC"
"14"	"Her2E_SC"
"15"	"LumB_SC"
"16"	"LumB_SC"
"17"	"LumB_SC"
"18"	"LumB_SC"
"19"	"Basal_SC"
"20"	"LumB_SC"
"21"	"LumB_SC"

Figure 5: scSubtype highest calls. Example table of scSubtype highest calls outputs. For each cell, the scSubtype call score was calculated to determine the dominant intrinsic subtype. S-CV scores for each individual were calculated based on scSubtype highest call data.

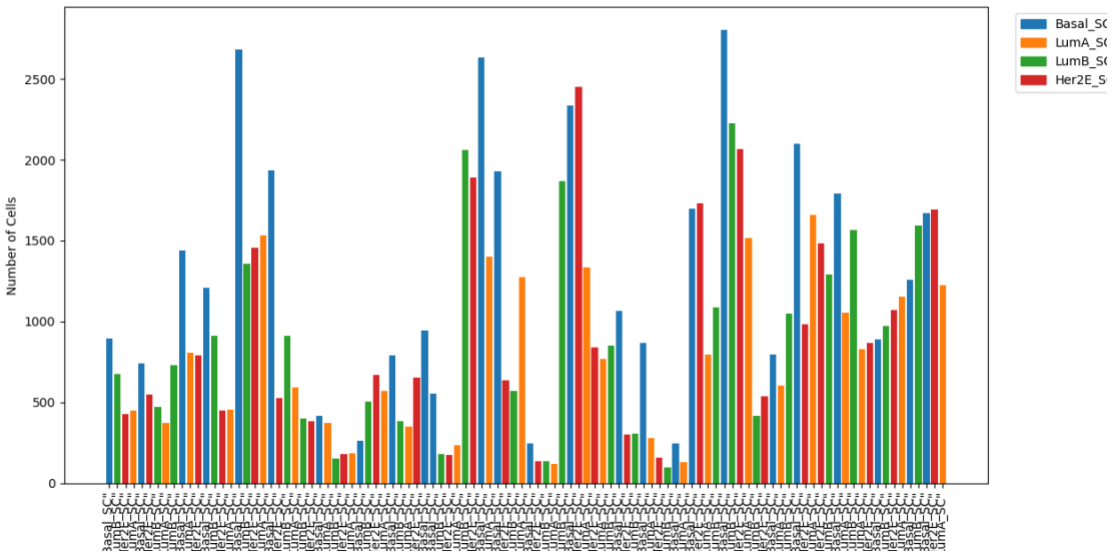

Figure 6: Visualization of scSubtype highest calls. Illustration of dominant subtype calls across patients. The y-axis represents cell count, and the x-axis is clustered by cell type. Each color represents one of four intrinsic subtypes. These data demonstrate the subtype heterogeneity inherent across cell types and within our sample set of ~100,000 cells.

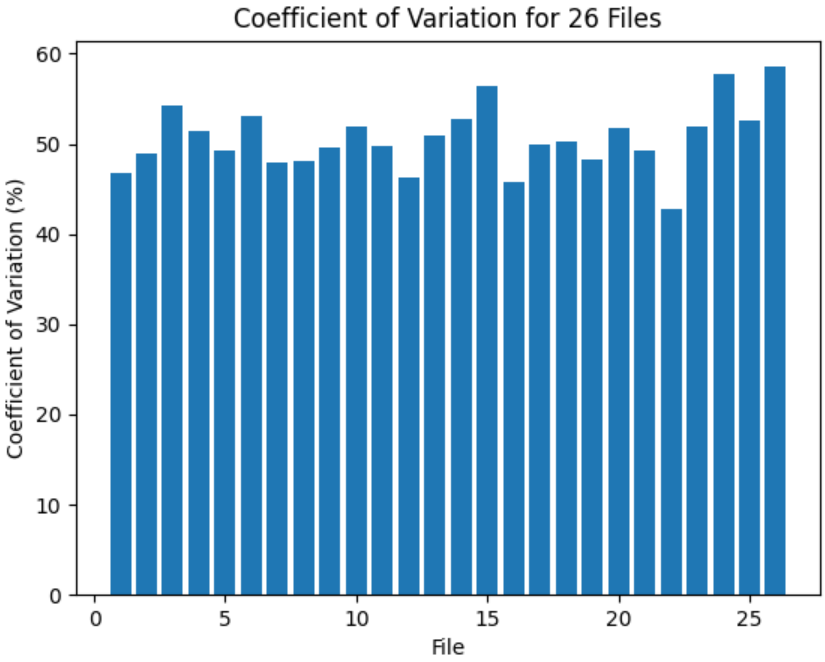

Figure 7: Visualization of S-CV scores. Plot of S-CV across all 26 patients in our training sample set. These data demonstrate the magnitude of IH within each patient tumor, as well as spread of S-CV within our patient population.

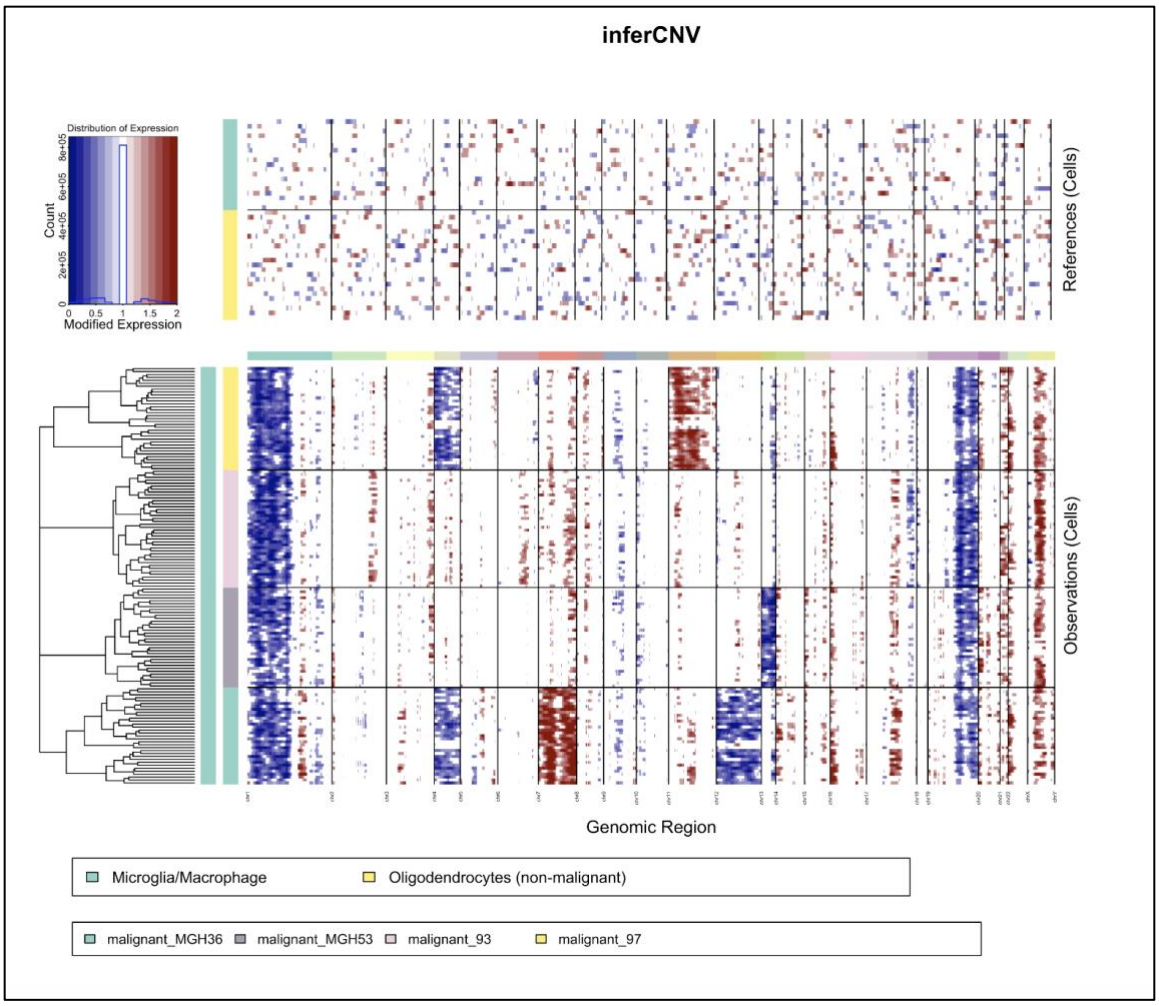

Figure 8: Example inferCNV analysis. Example schematic of inferCNV analysis from method authors: https://github.com/broadinstitute/inferCNV/wiki. Reference cells (top) are screened from modified expression levels across 100-200 different chromosomal loci. Potentially neoplastic cells (bottom) are compared to the reference genome and reference cells. Deviations from reference genome, i.e. copy number variations (CNVs), are denoted as insertions (red) or deletions (blue) based on expected read length for a given gene. For both plots, the x-axis represents gene loci across 23 unique chromosomes and the y-axis represents cells. The magnitude of CNVs in a potentially neoplastic cell, and thus the genetic instability score, is determined from this plot.

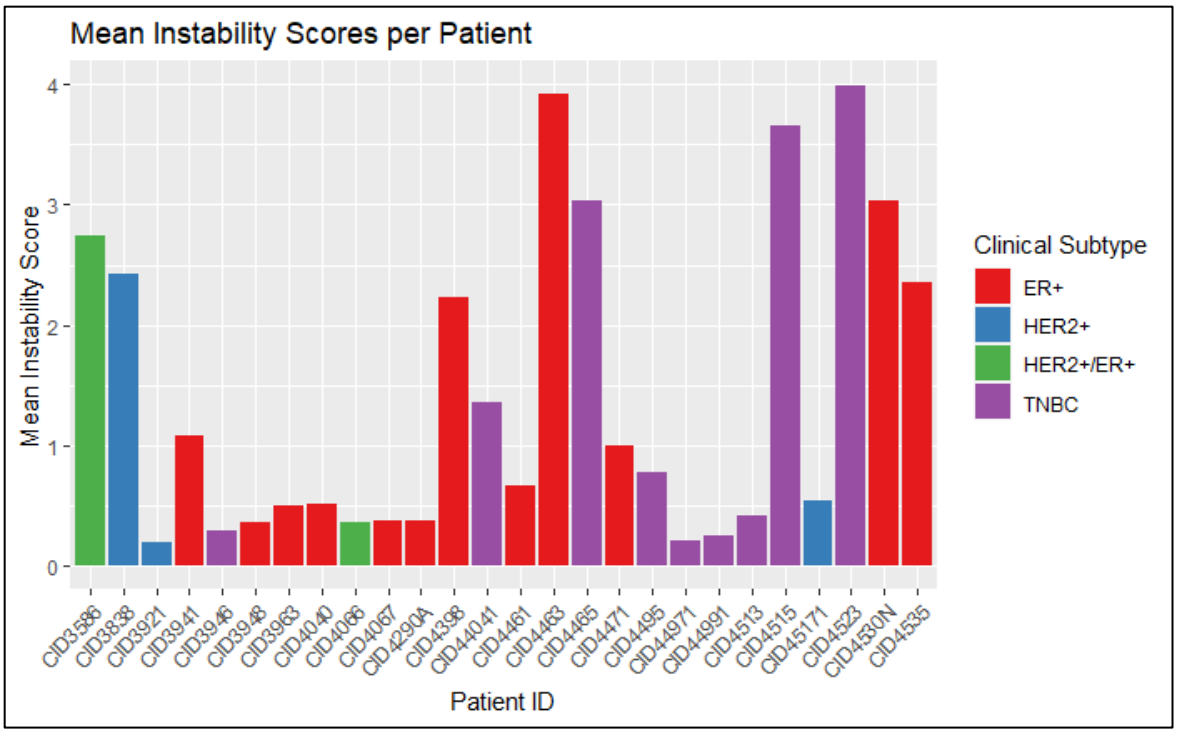

Figure 9: Patient genetic instability scores. Plot of average GIS scores calculated by inferCNV across all patients in our training sample set. These data illustrate the genetic inter-tumor heterogeneity inherent in our patient population.

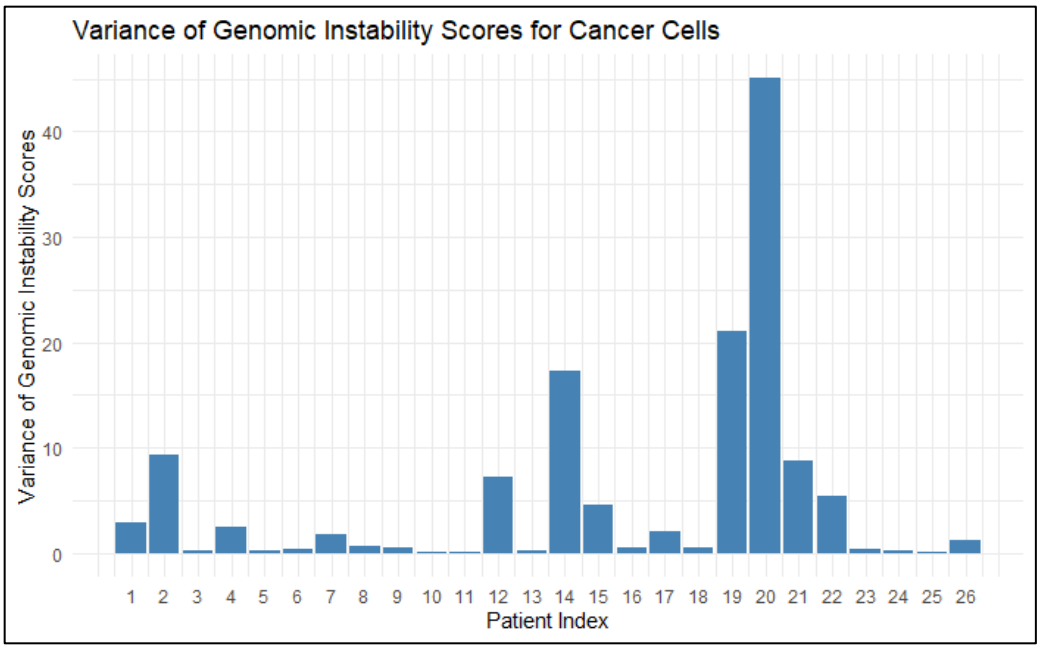

Figure 10: Visualization of GIS-CV across confirmed neoplastic cells. Plot of neoplastic GIS-CV scores across our patient sample set. These data illustrate the variations in ITNH both within and across patients.

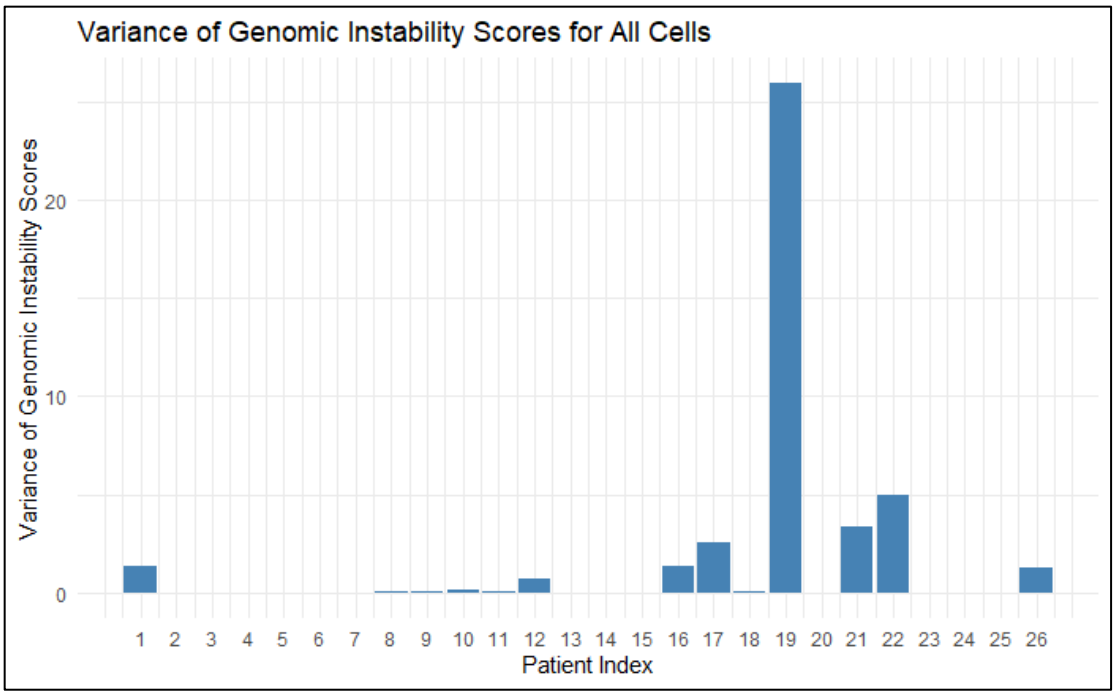

Figure 11: Visualization of GIS-CV across all cells. Plot of all GIS-CV scores across our patient sample set. These data illustrate the variations in ITNH both within and across patients.

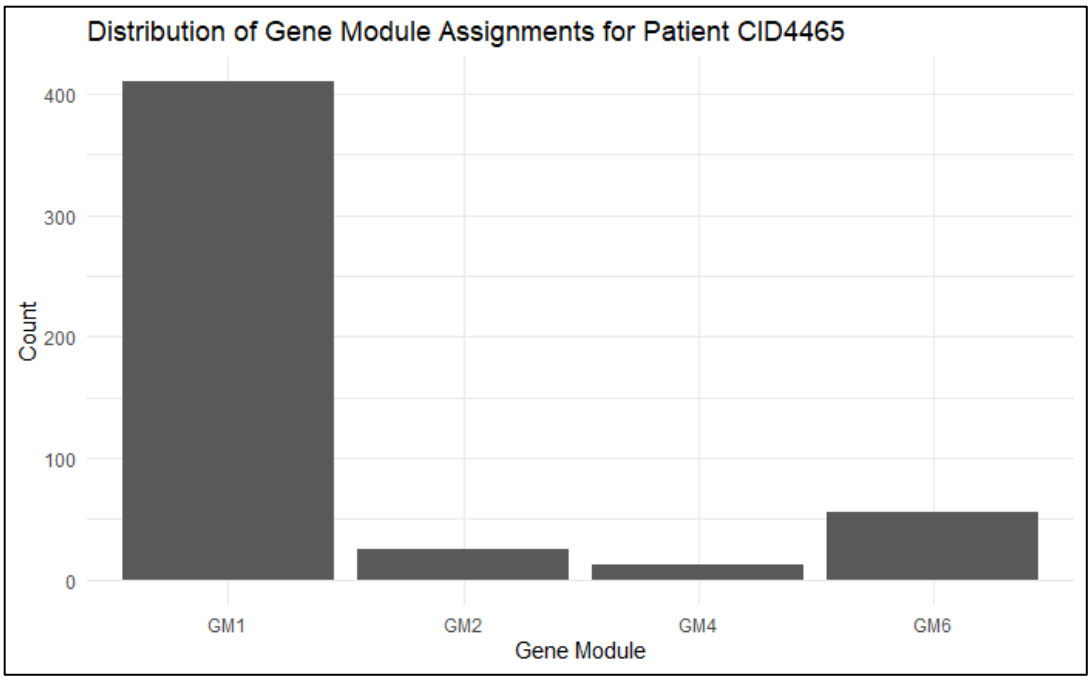

Figure 12: Gene module analysis cluster plot. Distribution of gene module assignments for a single individual. These data represent ITTH within a single individual, as well as a dominant transcriptomic profile for that individual.

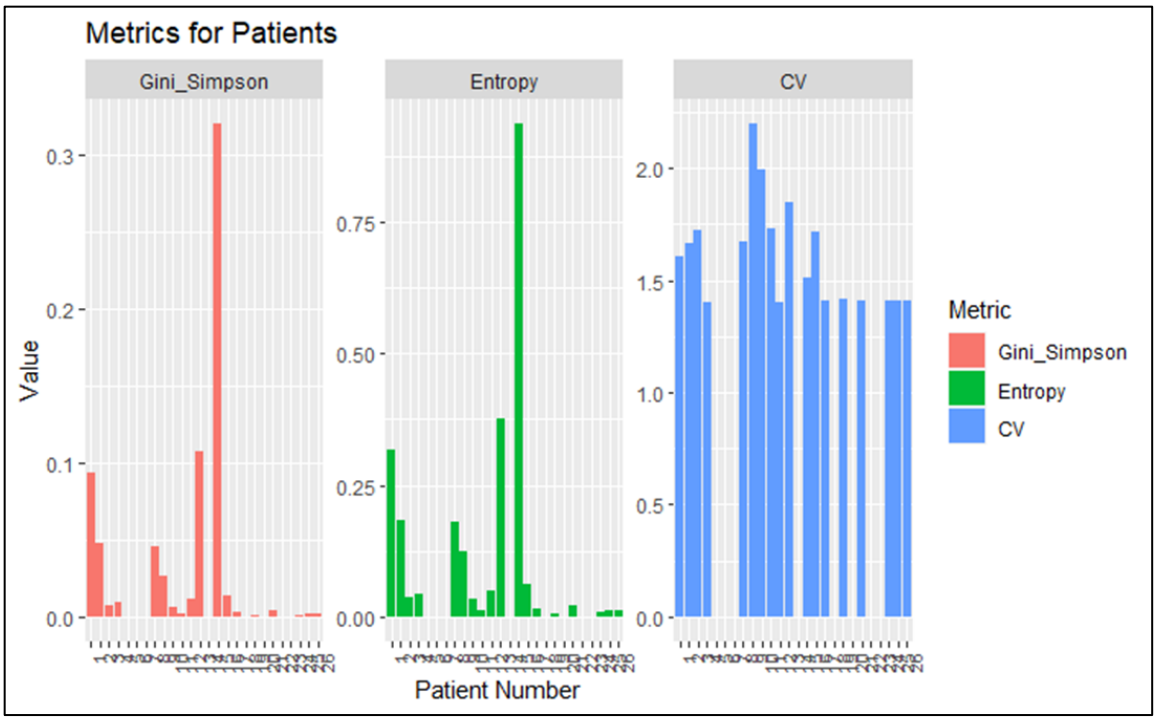

Figure 13: GM-CV scores across neoplastic cells. Plot of neoplastic GM-CV scores, in addition to Gini Simpson and Entropy measures, across our patient sample set. These data illustrate the variations in ITTH both within and across patients.

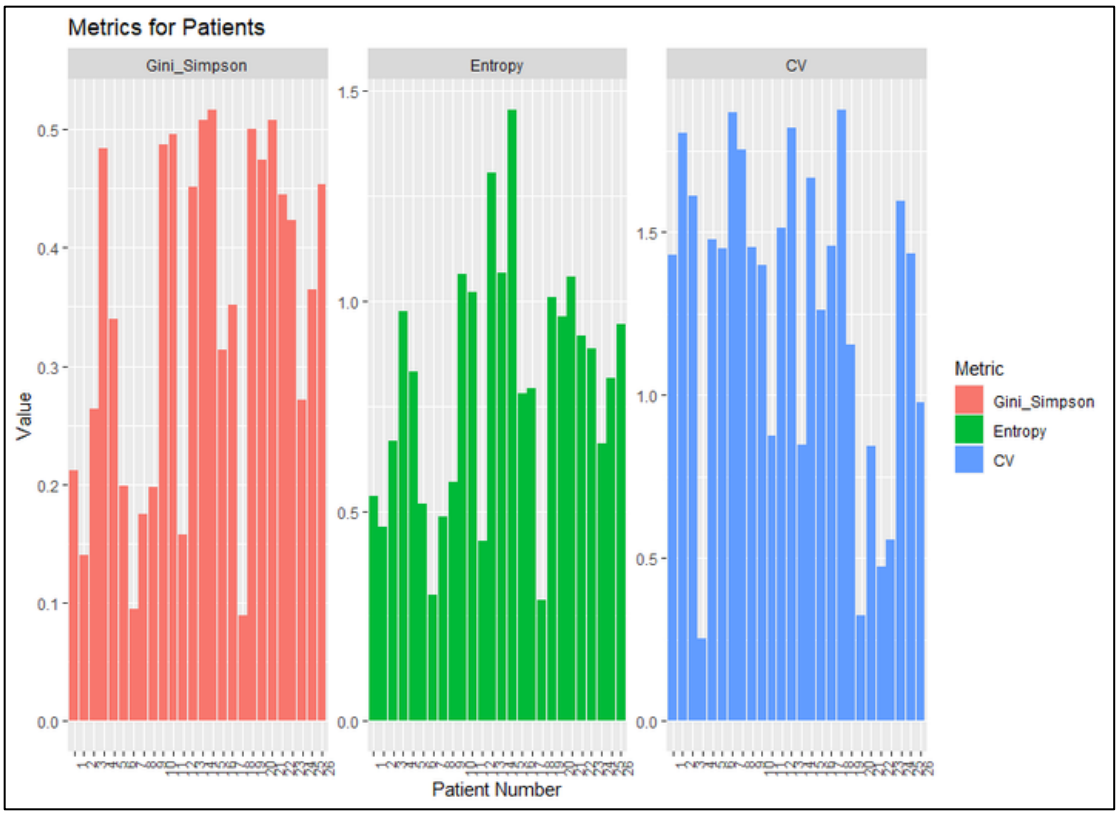

Figure 14: GM-CV scores across all cells. Plot of all GM-CV scores, in addition to Gini Simpson and Entropy measures, across our patient sample set. These data illustrate the variations in ITTH both within and across patients.

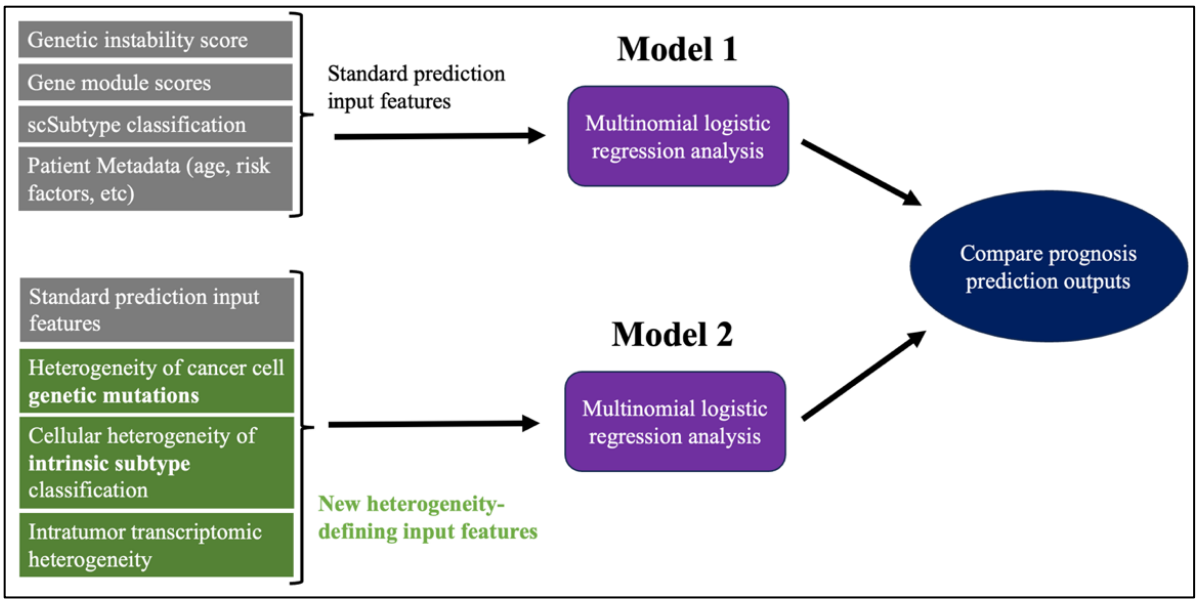

Figure 15: Multinomial logistic regression prediction model. Schematic of our intended experimental design to measure the biological significance of IH in the context of prognosis prediction.