Addressing the mean-variance relationship in spatially resolved transcriptomics data with *spoon*

Kinnary Shah[1], **Boyi Guo**[1], **and Stephanie C. Hicks**[1,2,3,4,*]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[2]Department of Biomedical Engineering, Johns Hopkins School of Medicine, Baltimore, MD, USA
[3]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA
[4]Malone Center for Engineering in Healthcare, Johns Hopkins University, MD, USA
[*]Corresponding author: `shicks19@jhu.edu`

Abstract

An important task in the analysis of spatially resolved transcriptomics data is to identify spatially variable genes (SVGs), or genes that vary in a 2D space. Current approaches rank SVGs based on either p-values or an effect size, such as the proportion of spatial variance. However, previous work in the analysis of RNA-sequencing identified a technical bias, referred to as the "mean-variance relationship", where highly expressed genes are more likely to have a higher variance. Here, we demonstrate the mean-variance relationship in spatial transcriptomics data. Furthermore, we propose *spoon*, a statistical framework using Empirical Bayes techniques to remove this bias, leading to more accurate prioritization of SVGs. We demonstrate the performance of *spoon* in both simulated and real spatial transcriptomics data. A software implementation of our method is available at `https://bioconductor.org/packages/spoon`.

Keywords: spatial transcriptomics, spatially variable gene, empirical Bayes, mean-variance bias, Gaussian process regression

1 Introduction

Advances in transcriptomics have led to profiling gene expression in a 2D space using spatially resolved transcriptomics (SRT) technologies [1]. These technologies have already led to novel biological insights across diverse application areas, including cancer [2], developmental biology [3, 4], and neurodegenerative disease [5, 6]. These emerging data types have also motivated new computational challenges, such as spatially-aware quality control to identify low-quality observations [7] and spatially-aware clustering to identify discrete spatial domains [8]. Another common data analysis task with these data is to perform feature selection by identifying a set of spatially variable genes (SVGs) [9–14]. The top SVGs are identified by ranking the genes based on some metric, such as p-values or an effect size like the proportion of spatial variance [10]. Accurately identifying SVGs is important because the top features are often used for downstream analyses, such as dimensionality reduction or unsupervised clustering [15–19].

Recently, Weber et al. [9] developed a computational method to identify SVGs based on a nearest-neighbor Gaussian process (NNGP) regression model [20]. In the paper, the authors identified an important relationship in SRT data. Specifically, they found a relationship between the estimated spatial variation and the overall expression, where genes that have higher overall expression are more likely to be more spatially variable. This phenomenon, known as the "mean-variance relationship", is a well-documented technical bias in genomics [21–29]. As previously shown in other sequencing-based technologies, the reason for this bias is due to the preprocessing and normalization steps that are often applied to raw gene expression counts, or the number of unique molecular identifiers (UMIs) mapping to each gene. Specifically, Weber et al. [9] used normalized $\log_2$-transformed gene expression as input to the NNGP model. These preprocessing techniques are widely used in bulk RNA-seq, scRNA-seq, and SRT data, because these transformations are assumed to enable the use of statistical models based on Gaussian distributions, rather than less tractable count-based distributions [10, 28, 30–32].

However, previous work in the analysis of bulk and scRNA-seq data has also shown that because counts have unequal variances (or larger counts have larger standard deviations compared to smaller counts [33]) (**Figure S1A**), applying these log-transformations is problematic as it can overcorrect (or large logcounts can have a smaller standard deviation than small logcounts) (**Figure S1B**). In these settings, it is important to account for the mean-variance relationship. Another way to think about the mean-variance relationship is to describe it as heteroskedasticity [34] in the context of using linear models. In contrast, homoskedasticity, in the case of profiling gene expression, would be if all genes in a sample had the same variance. When applying statistical models that assume homoskedasticity in the data, if we ignore the mean-variance relationship, our results would produce inefficient estimators or even incorrect results [22, 35, 36]. For example, in differential expression analysis, ignoring the mean-variance relationship can produce false positive differentially expressed genes [25].

To address this technical bias in SRT data, here we introduce the *spoon* framework, which was inspired by the limma-voom method [33] developed for bulk RNA-seq data. In this way, the name *spoon* incorporates the concepts of both "spatial" and its origin in RNA-seq. Using real and simulated SRT data, we show that *spoon* is able to correct for the mean-variance relationship leading to more accurately prioritizing SVGs. A software implementation of our method is available as an R/Bioconductor package (`https://bioconductor.org/packages/spoon`).

2 Materials and Methods

2.1 An overview of the *spoon* model and methodological framework

The *spoon* model was inspired by the limma-voom method [33], which estimates the mean-variance relationship to obtain precision weights for each observation to be used as input into a linear regression model to identify differentially expressed genes with bulk RNA-sequencing data [26]. In *spoon*, we use a similar idea. First, we use Empirical Bayes techniques to estimate observation- and gene-level weights. However, here we use a Gaussian process regression model, rather than a linear regression model, to model SRT data. Then, we leverage the Delta method to re-scale the data and covariates by these weights to address the heteroskedasticity in SRT data. Briefly, the Gaussian process (GP) regression model is specified as follows [20]:

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) \tag{1}$$

where $\boldsymbol{s}$ are the spatial locations, $y(\boldsymbol{s})$ is the response at a location, $\boldsymbol{x}(\boldsymbol{s})$ is a vector of explanatory variables, $w(\boldsymbol{s})$ is a function accounting for the spatial dependence, and $\epsilon(\boldsymbol{s}) \sim N(0, \tau^2)$ is noise. $\boldsymbol{\beta}$ is a fixed effect, while w and ϵ are random effects. $w(\boldsymbol{s})$ is modeled with a Gaussian process, $w(\boldsymbol{s}) \sim GP(\mu(\boldsymbol{s}), C(\boldsymbol{\theta}))$, where $\mu(\boldsymbol{s})$ is a mean function and $C(\boldsymbol{\theta})$ is a covariance function with parameters $\theta = (\sigma^2, \phi, ...)$ for the Matérn covariance function:

$$C(\boldsymbol{s_i}, \boldsymbol{s_j}|\boldsymbol{\theta} = (\sigma^2, \phi, \nu)) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(||\boldsymbol{s_i} - \boldsymbol{s_j}||\phi)^\nu K_\nu(||\boldsymbol{s_i} - \boldsymbol{s_j}||\phi); \phi > 0, \nu > 0$$

where σ^2 is the spatial component of variance, ϕ is the decay in spatial correlation, ν is the smoothness parameter, and K_ν is the Bessel function of the second kind with order ν. Because we fit these models on a per-gene basis with up to thousands of genes in a given dataset, we use a nearest-neighbor Gaussian process (NNGP) [37, 38] to reduce the computational running time and make *spoon* useful to practitioners. The key idea behind using NNGPs is that instead of conditioning on all of the points in the data, only a subset (a set of nearest neighbors) of the data are used for the conditioning. Conditioning on enough of the closest neighbors provides sufficient estimates of the needed information needed and improves storage and computational costs. Briefly, a NNGP is fit to the preprocessed expression values for each gene:

$$\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}, \tau^2)) \tag{2}$$

where the primary difference between a full GP model (**Equation 1**) and a NNGP (**Equation 2**) is that the NNGP covariance matrix, $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}, \tau^2)$, is a computationally fast approximation to the covariance matrix from a full GP model, $\boldsymbol{\Sigma}(\boldsymbol{\theta}, \tau^2) = \boldsymbol{C}(\boldsymbol{\theta}) + \tau^2\boldsymbol{I}$. In other words, $\tilde{\boldsymbol{\Sigma}}$ approximates the covariances from both from $w(\boldsymbol{s})$ and $\epsilon(\boldsymbol{s})$. For the kernel, $C(\boldsymbol{\theta}) = [C_{ij}(\boldsymbol{\theta})]$, we assume an exponential covariance function:

$$C_{ij}(\boldsymbol{\theta}) = \sigma^2 \exp\left(\frac{-||\boldsymbol{s_i} - \boldsymbol{s_j}||}{l}\right)$$

where $\boldsymbol{\theta} = (\sigma^2, l)$, and σ^2 is the spatial component of variance of interest. σ^2 is different from the nonspatial component of variance, τ^2, which is also referred to as the nugget. l is the lengthscale parameter, which sets how quickly the correlation decays with distance. $||\boldsymbol{s_i} - \boldsymbol{s_j}||$ is the Euclidean distance between spatial locations. To estimate the parameters in the NNGP model, we use the BRISC R package [20]. Using the estimated parameters, we calculate an effect size, the proportion of spatial variance ($\frac{\hat{\sigma}^2}{\hat{\sigma}^2+\hat{\tau}^2}$).

2.2 Calculating observation- and gene-level weights using Empirical Bayes techniques

Briefly, we calculate the average $\log_2$ expression values and the standard deviations of the residuals from fitting an NNGP model per gene using BRISC (**Figure 1A**). Then, we use splines to fit the gene-wise

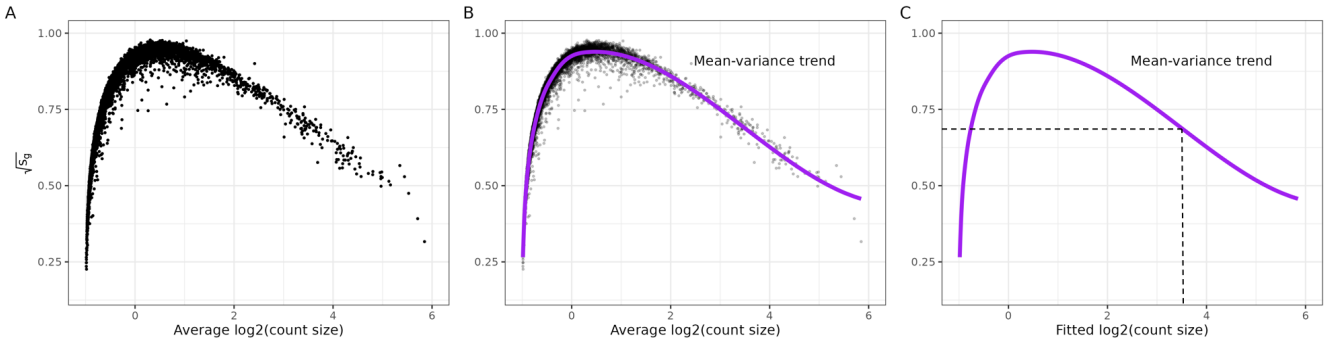

Figure 1: Calculating precision weights for individual observations. These data are from Invasive Ductal Carcinoma breast tissue analyzed with 10x Genomics Visium [39], hereafter referred to as "Ductal Breast". (**A-C**) The square root of the residual standard deviations estimated using nearest neighbor Gaussian processes ($\sqrt{s_g}$ defined in Equation 3) are plotted against average logcount ($\tilde{\boldsymbol{r}}$). (**B**) Same as A, except a spline curve (purple) is fitted to the data to estimate the gene-wise mean-variance relationship. (**C**) Using the fitted spline curve, each predicted count value ($\hat{\lambda}_{gi}$) is mapped to its corresponding square root standard deviation value using $\texttt{spl}(\hat{\lambda}_{gi})^{-4}$.

mean-variance relationship (**Figure 1B**). Finally, we use the fitted curve to estimate observation- and gene-level weights (**Figure 1C**). Next, we describe each of these steps in greater detail.

2.2.1 Fitting per-gene NNGP models using logCPM values

We start with a counts matrix, transposed so each row is a spot and each column is a gene. There are n spots and G genes in the counts matrix. The UMI counts can be indexed by r_{gi} for spots $i = 1$ to n and genes $g = 1$ to G. We define the total number of UMIs for sample i as $R_i = \sum_{g=1}^{G} r_{gi}$. Next, we transform r_{gi} to adjust for the total number of UMIs (R_i) by using logcounts per million (logCPM). We use a pseudocount of 0.5 to ensure we do not take the log of 0 and we add a pseudocount of 1 to the library size to make sure $0 < \frac{(r_{gi}+0.5)}{R_i+1} < 1$:

$$y_{gi} = \log_2\left(\frac{r_{gi}+0.5}{R_i+1} \times 10^6\right)$$

Using the normalized and $\log_2$-transformed data y_{gi}, we fit a NNGP model (**Equation 2**) per gene with a default of $\boldsymbol{X} = \mathbf{1}_{[N\times 1]}$, corresponding to including an intercept, with β_g representing the overall mean expression level for gene g. Using the observed data y_{gi} and the predicted value $\hat{\mu}_{gi} = x_i^T\hat{\beta}_g$, we can calculate the standard deviation of the residuals between y_{gi} and $\hat{\mu}_{gi}$:

$$s_g = \sqrt{\frac{\sum_{i=1}^{n}(y_{gi}-\hat{\mu}_{gi})^2}{n-1}} \tag{3}$$

The square root of s_g is what we use to represent the 'variance' in the mean-variance relationship (see y-axis in **Figure 1A-C**). This concept is used in limma-voom as well because the square root of the standard deviations is roughly symmetrically distributed.

2.2.2 Modeling the mean-variance relationship using $\sqrt{s_g}$ and $\tilde{\boldsymbol{r}}$

Next, we fit a nonparametric spline curve to model the mean-variance relationship in our data. Instead of using $\bar{y}_g$ directly to represent the 'mean' component, we convert $\bar{y}_g$ to average logcount using the geometric mean of library size, $\tilde{\boldsymbol{R}} = \exp\left(\sum_{i=1}^{n}\log(R_i)\right)$. We use the geometric mean to avoid integer overflow:

$$\tilde{\boldsymbol{r}} = \bar{y}_g + \log_2(\tilde{\boldsymbol{R}}) - \log_2(10^6) \tag{4}$$

Then, we use smoothing splines (specifically `smooth.spline()` in the base R `stats` package) to model the mean-variance relationship between $\sqrt{s_g}$ and $\tilde{\boldsymbol{r}}$. We use splines because we found they are a robust way to model the mean variance relationship seen across multiple datasets. We use the notation `spl()` to denote the fitted curve (**Figure 1B**), which represents an estimate of the mean-variance relationship.

2.2.3 Prediction modeling using fitted `spl()` curve

Similar to Equation 4, we convert the predicted value $\hat{\mu}_{gi}$ (on the logCPM scale) to a predicted count value:

$$\hat{\lambda}_{gi} = \hat{\mu}_{gi} + \log_2(R_i + 1) - \log_2(10^6) \tag{5}$$

The fitted counts values for each observation are used as input to predict the square root residual standard deviation values for each y_{gi} using the spline curve. **Figure 1C** shows an example of mapping an individual observation to a square-root standard deviation value using its fitted value from the BRISC models.

To avoid extrapolating beyond the range of the function, individual observations that have $\hat{\lambda}_{gi}$ more extreme than the range of $\tilde{\boldsymbol{r}}$ are constrained. If $\hat{\lambda}_{gi}$ is greater than $\max(\bar{\boldsymbol{y}})$, then the predicted square root residual standard deviation value for that observation is constrained to $\texttt{spl}(\max(\bar{\boldsymbol{y}}))$. If $\hat{\lambda}_{gi}$ is less than $\min(\bar{\boldsymbol{y}})$, then the predicted square root residual standard deviation value for that observation is constrained to $\texttt{spl}(\min(\bar{\boldsymbol{y}}))$. The final step is taking the inverse of the squared predicted standard deviation to compute the weight for each individual observation. The weight for each observation is defined as $w_{gi} = \texttt{spl}\,(\hat{\lambda}_{gi})^{-4}$, using the constrained values for observations outside of the range.

2.3 Correct for heteroskedasticity using observation- and gene-level precision weights

If the desired SVG detection method accepts observation- and gene-level weights, then the estimated weights w_{gi} (described in Section 2.2) can be used as input directly into the method. If the desired SVG detection method does not accept weights, then the Delta method is leveraged to rescale the data and covariates by the weights. These scaled data and covariates are used as inputs into the desired SVG detection function.

For example, the SVG detection tool called nearest neighbor SVGs (nnSVG) [9] uses a Gaussian process regression model and can have weights incorporated in the following way. We correct for the heteroskedasticity by adjusting with precision weights, w_{gi} for gene g at spatial location i. If $\boldsymbol{W}$ is a diagonal matrix where each diagonal element is w_{gi}, then we know:

$$\boldsymbol{W}\boldsymbol{y} \sim N(\boldsymbol{W}\boldsymbol{X}\beta, \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W})$$

where

$$\begin{aligned}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W} &= \boldsymbol{W}\boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{W} + \tau^2\boldsymbol{W}\boldsymbol{I}\boldsymbol{W} \\ &= \boldsymbol{W}\boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{W} + \tau^2\boldsymbol{W}\boldsymbol{W} \\ &= \boldsymbol{C}(\boldsymbol{\theta}') + \tau'^2\boldsymbol{I}\end{aligned}$$

and the new input data to nnSVG would be $\boldsymbol{W}\boldsymbol{y}$ and $\boldsymbol{W}\boldsymbol{X}$ where $\boldsymbol{X} = \begin{bmatrix} 1 \\ \boldsymbol{X}_{gi} \end{bmatrix}$.

2.4 Data

2.4.1 Real SRT data

Tissues from several regions of the human body analyzed with 10x Genomics Visium were used in the analyses. The datasets and preprocessing steps are further described below:

- Ductal Breast: Invasive Ductal Carcinoma breast tissue data are publicly available from the 10x Genomics website. It contains one tissue sample from one donor with Invasive Ductal Carcinoma [39]. After preprocessing, this dataset contains 12,321 genes and 4,898 spots.
- Lobular Breast: Invasive Lobular Carcinoma breast tissue data are publicly available from the 10x Genomics website. It contains one tissue sample from one donor with Invasive Lobular Carcinoma [40]. After preprocessing, this dataset contains 12,624 genes and 4,325 spots.
- Subtype Breast: Estrogen receptor positive (ER+) breast cancer tissue data are publicly available on Zenodo and contains several tissue samples of breast cancer tissue. Only sample CID4290 is used for this analysis [41]. After preprocessing, this dataset contains 12,325 genes and 2,419 spots.
- DLPFC: This dataset contains two pairs of spatial replicates of human postmortem dorsolateral prefrontal cortex (DLPFC) tissue from three neurotypical adult donors. Only tissue sample 151507 is used for this analysis [15]. After preprocessing, this dataset contains 7,343 genes and 4,221 spots.
- HPC: This dataset contains human postmortem hippocampus (HPC) tissue from several neurotypical adult donors. Each sample was broken up into four Visium slides due to the large size of the HPC. Only tissue sample V12D07_335, portion D1 is used for this analysis [16]. After preprocessing, this dataset contains 5,348 genes and 4,992 spots.
- LC: This dataset contains human postmortem locus coeruleus (LC) tissue from five neurotypical adult donors. Only tissue sample 2701 is used for this analysis [42]. After preprocessing, this dataset contains 1,331 genes and 2,809 spots.
- Ovarian: This dataset contains tissues collected during interval debulking surgery from eight high-grade serous ovarian carcinoma patients undergoing chemotherapy. Only one tissue sample from patient 2 is used for this analysis [43]. After preprocessing, this dataset contains 12,022 genes and 1,935 spots.

Preprocessing was performed as uniformly as possible across the datasets. For datasets that had an annotation for whether or not a spot was in the tissue, spots outside of the tissue were removed. For the Subtype Breast dataset, spots that were classified as artifacts were removed. `nnSVG::filter_genes()` was used to remove genes without enough data, specifically we kept genes with at least 2 counts in at least 0.2% of spots. For the LC dataset, we used a UMI filter instead of this function to remove genes with less than 80 total UMI counts summed across all spots. `scuttle::logNormCounts()` with default arguments was used to compute log-normalized expression values.

2.4.2 Simulated SRT data

To simulate the mean-variance relationship, we simulated raw gene expression counts following a Poisson distribution:

$$\boldsymbol{c}(\boldsymbol{s})|\lambda(\boldsymbol{s}) \sim Poisson(\lambda(\boldsymbol{s})); \lambda(\boldsymbol{s}) = \exp(\boldsymbol{\beta} + \boldsymbol{C}(\sigma^2))$$

where $\boldsymbol{s}$ are spatial locations, $\boldsymbol{\beta}$ is a vector of true mean expression per gene, σ^2 is the spatial component of variance, and $\boldsymbol{C}$ is the covariance function using a Matérn kernel with squared exponential distance. The σ^2 values and $\boldsymbol{\beta}$ values were randomly assigned from ranges of $[0.2, 1]$ and $[\ln(0.5), \ln(1)]$, respectively. We intentionally simulate σ^2 and β values so they are not correlated. In this way, we ensure we are simulating SVGs at all levels of mean expression. A fixed lengthscale parameter was chosen for all of the genes in a given simulation. Based on the estimated lengthscale distributions for four datasets, we chose to focus our simulations on smaller lengthscales because the majority of estimated lengthscales are between 0 to 0.15 (**Figure S2**). For reference, a scaled lengthscale value of 0.15 is interpreted as 15% of the maximum width or height of the tissue area on a standard Visium slide. We simulated 1000 genes in the following simulations.

In addition, we also considered the performance as a function of varying the lengthscale parameter l in $\theta = (\sigma^2, l)$. In the NNGP model, the lengthscale parameter sets how quickly the correlation decays with distance. In the nnSVG SVG detection method [9], a key innovation was using a flexible lengthscale parameter to fit the model for each gene. Genes within the same tissue can spatially vary with different ranges of sizes and patterns, so a flexible lengthscale parameter for each gene enables the discovery of distinct biological processes. For the primary simulation evaluation, a lengthscale of 100 was used. This corresponds to a scaled lengthscale value of roughly 0.02. For supplementary simulation evaluations, 50, 60, 100, and 500 lengthscales were used. These correspond to 0.010, 0.012, 0.020, and 0.100 of the maximum width or height of the tissue area on a standard Visium slide. The spatial coordinates from the example dataset `Visium_DLPFC()` in the `STexampleData` package were used. This dataset contains 4,992 spots. We used the subset of 968 spots with row and column coordinates between 20 to 65 as the spatial coordinates to reduce the amount of time to simulate data.

2.5 Methods to detect SVGs

For Moran's I [44], we ranked genes by the Moran's I value. For nnSVG [9], the genes were ranked within the method based on the estimated likelihood ratio test statistic values comparing the fitted model against a classical linear model, assuming the spatial component of variance is zero. For SpaGFT [45], the gene ranks were calculated within the method based on decreasing GFTscore, a measure of randomness of gene expression. For SPARK-X [11], adjusted combined p-values from multiple covariance matrices and kernels were used to rank genes. For SpatialDE2 [46], the genes were ranked by the negative of the fraction of spatial variance for each gene. All of the criteria were ranked using the `ties.method = ''first''` option.

1. Moran's I: `Rfast2::moranI()` [47] was used to compute Moran's I values, and the negative Moran's I value for each gene was ranked.

2. nnSVG: `nnSVG::nnSVG()` [9] was used, and the rank was calculated as part of the output of the function.

3. SpaGFT: `SpaGFT.detect_svg()` [45] was implemented in Python, and the rank was calculated as part of the output of the function.

4. SPARK-X: `SPARK::sparkx()` [11] was run with the option of a mixture of various kernels. The combined p-value from all the kernels for each gene was ranked.

5. SpatialDE2: `SpatialDE.fit()` [46] was implemented in Python to fit the model for each gene. The negative of the fraction of spatial variance for each gene was ranked.

An intercept-less covariate matrix is required to implement a weighted version of an SVG detection method. To the best of our knowledge, nnSVG is the only SVG detection tool with the option to include a covariate matrix without an intercept term. The weights from *spoon* have the potential to integrate with other methods based on the flexibility of their design.

2.6 Code Availability

spoon is freely available for use as an R package available from Bioconductor at `https://bioconductor.org/packages/spoon`. The code to reproduce the analyses in this paper is available on GitHub at `https://github.com/kinnaryshah/MeanVarBias`. We used *spoon* version 1.1.3 and R version 4.4.1 for the analyses in this manuscript.

3 Results

3.1 The mean-variance relationship exists in spatial transcriptomics data

We begin by systematically demonstrating the mean-variance relationship in SRT data. This finding builds upon the initial finding suggested in Weber et al. [9]. In contrast to investigating this bias in one tissue from one tissue section, here we explore this finding across multiple tissue sections from different regions in the human body, namely DLPFC, Ductal Breast cancer, HPC, LC, and Ovarian cancer. To visualize the mean-variance relationship, we plot the mean logcounts against different components (spatial and non-spatial components) of variance calculated using `nnSVG`. As seen in **Figure 2**, the mean-variance relationship is a concern in SRT data, specifically in the nonspatial component of variance, τ^2. Given τ^2 is used when calculating the proportion of spatial variance, this suggests the way genes are prioritized as spatially variable is dependent on the overall mean expression for the gene.

Next, we further investigated one of these tissues (DLPFC) to ask if the mean-variance relationship was due to differences in the spatial domains of the tissue. The six layers in the human neocortex are transcriptionally quite different from one another [15], so we wanted to show that the mean-variance relationship still exists when stratifying by layers. In order to control for differences in layer domains, the DLPFC data was first separated into Layers I-VI, and white matter and then the mean logcounts were plotted against the components of variance for each layer in the brain. However, we found that the mean-variance relationship was still observed within the different biological domains (**Figure S3**).

3.2 The mean-rank relationship exists in other SVG detection methods

Having established that the mean-variance relationship exists in SRT data across different tissues as measured by Gaussian processes in `nnSVG`, we next explored the mean-rank relationship as an extension of the mean-variance relationship. Other SVG detection methods do not separate out the total variance into spatial and nonspatial variance components, so we examine the mean-variance relationship using this proxy.

We examined the mean-rank relationship from several popular SVG detection methods on the DLPFC, Ovarian cancer, and Lobular Breast cancer datasets (**Figure 3**). The ranks were calculated for each SVG method (described in Section 2.5). We found that for almost every method, there is a clear relationship between the mean and the rank. Stated another way, the SVG detection methods that we evaluated rank and prioritize genes as SVGs, which is related to the overall mean expression. Because the overall mean expression is likely a technical artifact, we would expect that there should be genes that are highly ranked as SVGs within each mean-level decile. However, what we found is that the mean-

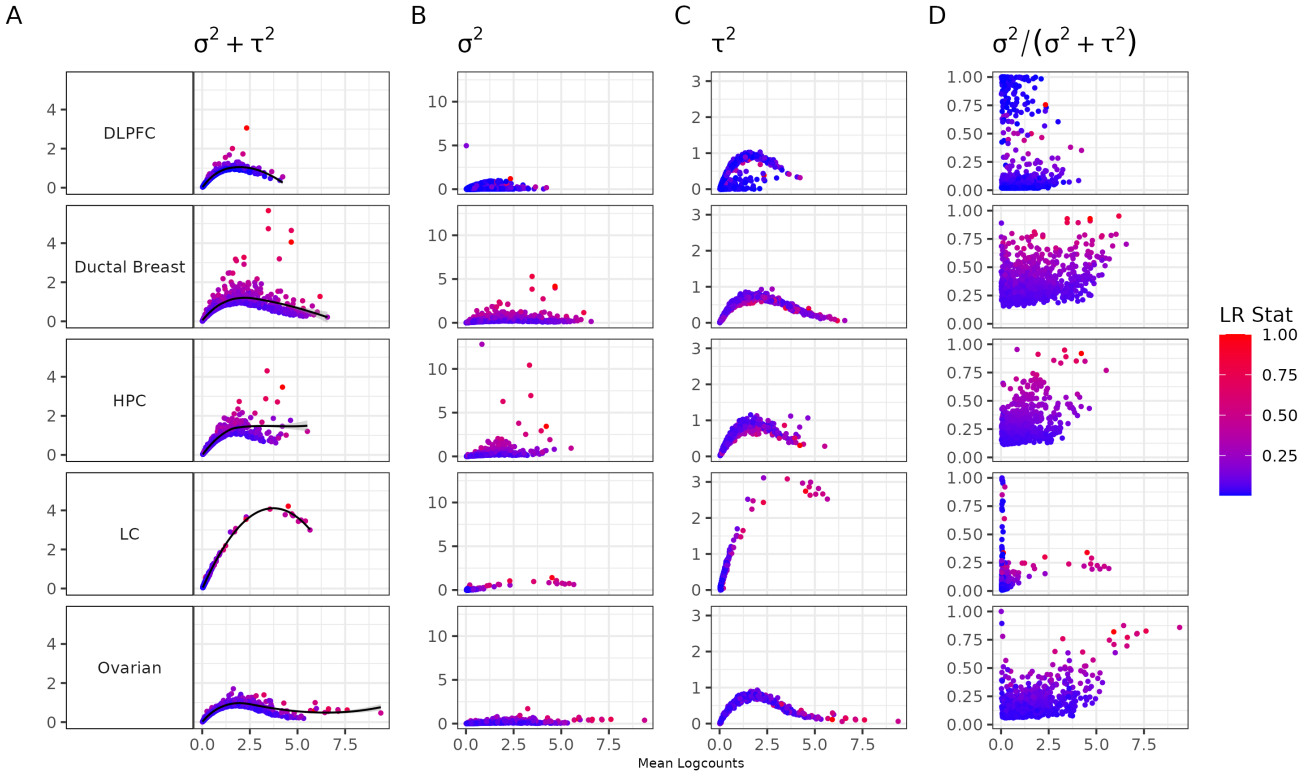

Figure 2: Mean-variance relationship exists in spatially resolved transcriptomics. Using data from different human tissues, in order from top to bottom: DLPFC [15], Ductal Breast cancer [39], HPC [16], LC [42], and Ovarian cancer [43], we quantified the mean-variance relationship. Each point is a gene colored by the likelihood ratio statistic for a test that compares the fitted model against a classical linear model for the spatial component of variance using a NNGP [9]. The likelihood ratio statistics (LR Stat) are scaled by the maximum likelihood ratio statistic for each dataset in order to have more uniform visualization. The x-axis is mean logcounts and the y-axes represent different components of variance, in order from left to right: total variance $\sigma^2 + \tau^2$, spatial variance σ^2, nonspatial variance τ^2, and proportion of spatial variance $\sigma^2/(\sigma^2 + \tau^2)$.

variance relationship biases genes towards the higher mean expression deciles. The extreme bias observed in SPARK-X is also noted in a recent benchmarking paper [48]. These are state of the art methods that perform well in recent benchmarking papers [14, 48, 49], yet they are sorely affected by the mean-variance bias.

3.3 Simulation: Weighted Spatially Variable Gene Evaluation

To address the mean-variance and mean-rank relationships, we began with simulation studies to evaluate the performance of *spoon* under different scenarios. Using simulated raw gene expression counts following a Poisson distribution (**Section 2.4.2**) with a fixed lengthscale (l=100), we ranked SVGs using nnSVG [9] without weights and with weights estimated via *spoon*. We found a strong mean-rank relationship using the unweighted SVGs (**Figure 4A**) compared to the weighted SVGs using *spoon* (**Figure 4B**). Stated differently, using observational- and gene-level weights, we can identify highly ranked SVGs even in lower deciles, demonstrating that *spoon* effectively addresses the mean-variance relationship.

We also explored the false discovery rate (FDR) (**Figure 4C**), true negative rate (TNR) (**Figure 4D**), and true positive rate (TPR) (**Figure 4E**). The red represents weighted `nnSVG` and the blue represents

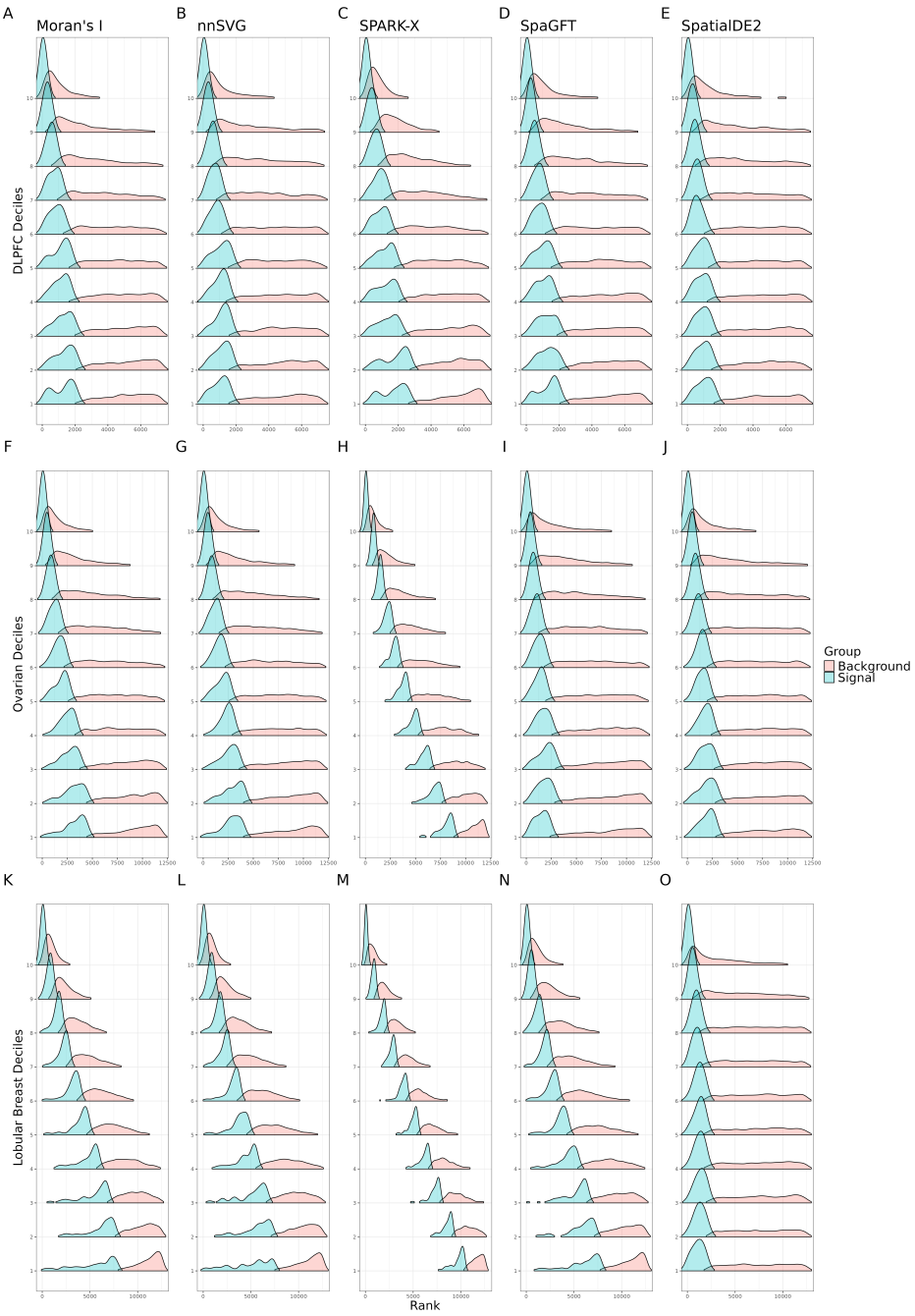

Figure 3: Mean-rank relationship exists in spatial transcriptomics data. Using three datasets, in order from top to bottom (DLPFC [15], Ovarian cancer [39], and Lobular Breast cancer [40]), we quantified the mean-rank relationship. The genes were binned into deciles based on mean logcounts. Decile 1 contains the lowest mean expression values. The x-axis represents the rank. Within each decile, the density of the top 10% ranks is plotted as the signal in blue, while the density of the remaining ranks is plotted as the background in orange. Each subfigure shows the mean-rank relationship that persists after applying each method, from left to right: Moran's I [47], nnSVG [9], SPARK-X [11], SpaGFT [45], and SpatialDE2 [46].

unweighted `nnSVG`. These plots represent the average of each respective rate over five iterations of the same simulation with unique random seeds. The FDR and TNR are similar between the unweighted and weighted methods, with a slight increase in performance observed in the unweighted method. The TPR,

however, is very similar for both methods. Finally, we considered other lengthscale values and found that the mean-variance relationship is improved for all values tested (**Figure S4**). We found that the weights from *spoon* improve the TPR for smaller lengthscale values, and there are diminishing returns regarding the convergence of the TPR for both the weighted method and unweighted methods at larger lengthscale values.

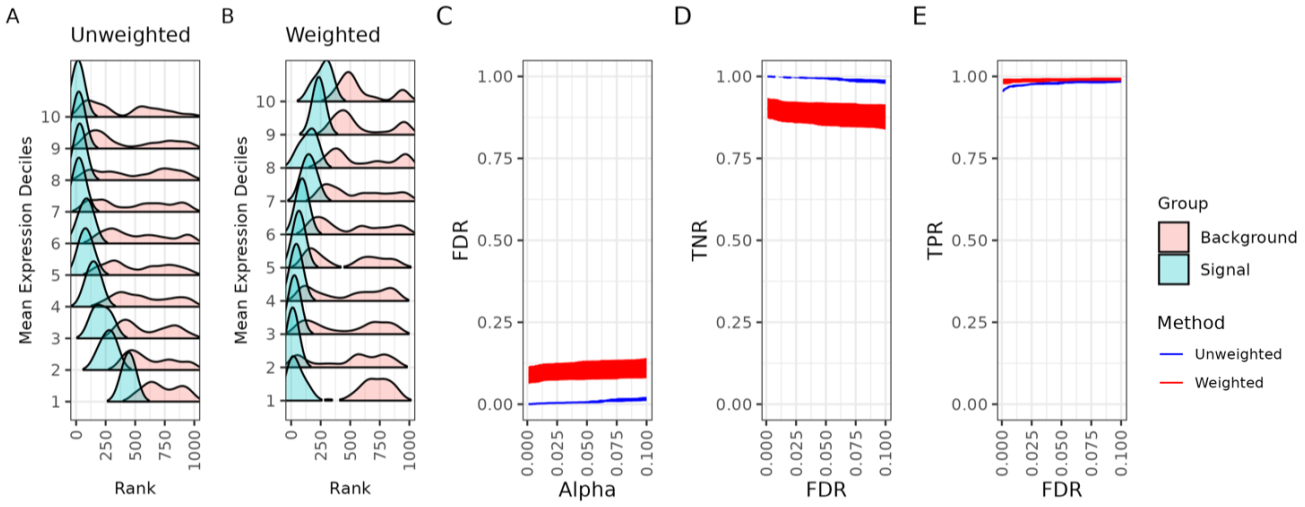

Figure 4: ***Spoon*** **removes the mean-variance relationship when detecting spatially variable genes.** This dataset consists of 1,000 simulated genes across 968 spots using a lengthscale of 100. Separately for unweighted and weighted methods, the genes were binned into deciles based on mean logcounts. Decile 1 contains the lowest mean expression values. Ridge plots for the (**A**) unweighted ranks and (**B**) weighted ranks are shown. Within each decile (y-axis), the density of the top 10% of ranks is plotted as the signal, while the density of the remaining ranks is plotted as the background. (**C**) False discovery rate (FDR) as a function of Type I error (α). As a function of FDR, we show the (**D**) true negative rate (TNR) and (**E**) true positive rate (TPR). The red represents weighted `nnSVG` and the blue represents unweighted `nnSVG`. These plots represent the average performance across five iterations of the same simulation, each with unique random seeds.

3.4 Real Data: Weighted Spatially Variable Gene Evaluation

Next, we evaluated the downstream impact of incorporating weights from *spoon* into SVG detection methods. Here, we aimed to demonstrate the impact of our method on recovering lowly expressed genes that become highly ranked in real biological datasets. We defined small mean gene expression genes as those with means less than the 25th percentile in the dataset. Within the set of small mean gene expression, we identified genes that were in the lowest 10% of ranks before weighting and then increased to the highest 10% of ranks after weighting. In the Ovarian cancer dataset, there are 7 genes that met this criterion. Out of these 7 genes, *TUFT1* and *DDX39B* are known to be implicated in ovarian cancer [50, 51]. These potentially important SVGs were ignored due to their low expression levels and our weighting algorithm can recapture them. Similar analyses were performed for the other three cancer datasets (**Figure 5**). The gene lists can be found in the supplemental materials.

Then, we explored the improvement in the small lengthscale set of genes. We defined small lengthscale genes as those with lengthscale values between 40 to 90. Within the set of small lengthscale genes, we found genes that were ranked higher after weighting. We also derived the "null distribution" — the underlying total SVGs for each dataset as a point of reference for the proportion of small lengthscale

genes that are ranked higher. We found that the differing proportions of small lengthscale genes that become higher ranked after weighting is appropriate based on the "null distribution" of the proportion of unweighted SVGs (**Figure S5**). Again, we related the higher-ranked small lengthscale genes to the cancer type of the dataset. In the Subtype Breast dataset, 59 small lengthscale genes were higher ranked after weighting, with 16 of these genes implicated in breast cancer. Full results are presented in **Figure 5** and gene lists are in supplemental materials.

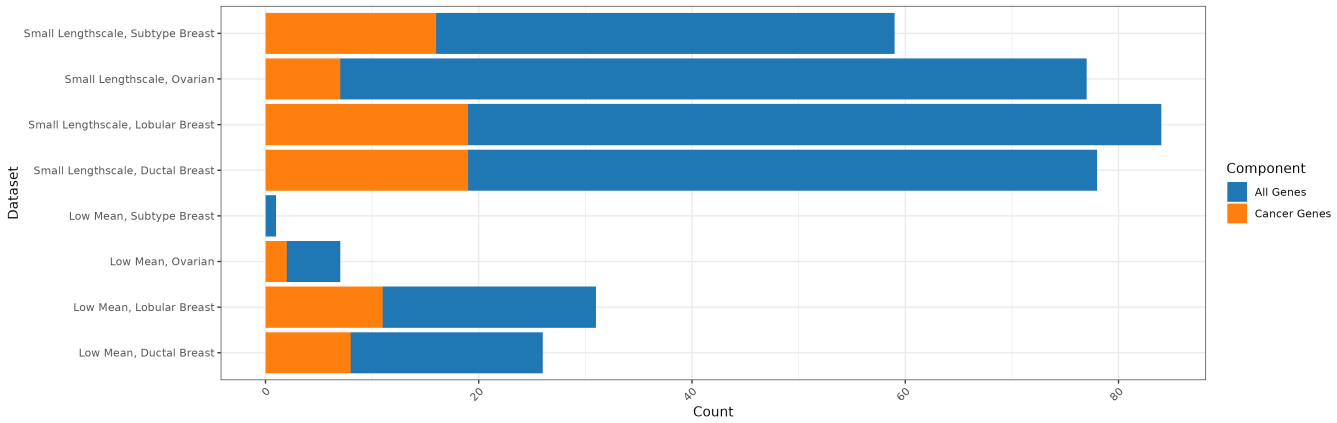

Figure 5: *Spoon* helps to detect SVGs associated with cancer that are lowly expressed. We used four datasets to evaluate the detection of cancer-related genes: Subtype Breast cancer [41], Ovarian cancer [43], Lobular Breast cancer [40], and Ductal Breast cancer [39]. Each bar contains the intersection of the set of genes of interest with genes within the set associated with cancer. For the first four rows, we defined low mean genes as those with means less than the 25th percentile in the dataset. Within the set of low mean genes, we found genes that were in the lowest 10% of ranks before weighting and then increased to the highest 10% of ranks after weighting. This is the set of genes of interest. The intersection in orange is the number of low mean and higher ranked genes that were found to be associated with the cancer of the dataset. For the last four rows, we defined small lengthscale genes as those with lengthscales between 40 to 90. Within the set of small lengthscale genes, we found genes that were ranked higher after weighting. This is the set of genes of interest. The intersection in orange shows the number of small lengthscale genes that were ranked higher and found to be associated with the cancer type of the dataset.

4 Discussion

In our work, we systematically demonstrate the mean-variance and the mean-rank relationships exist in spatially resolved transcriptomics data. Furthermore, we show this is not limited to just one SVG detection method. If researchers fail to adjust for this bias in spatial transcriptomics data, this can lead to false positives and inaccurate rankings of SVGs due to the violation of the homoskedasticity assumption. Here, we show that our method *spoon* is able to correct for this bias. Specifically, our approach uses Empirical Bayes techniques to generate weights for downstream analyses to remove the mean-variance relationship, leading to a more informative set of SVGs.

In a recent benchmark evaluation of SVG detection methods, the authors Chen et al. [48] noted a similar bias. NoVaTeST was recently proposed as a method to identify SVGs allowing noise variance to vary with spatial locations [52]. This method aims to identify genes that have location-dependent noise variance in SRT data, or genes that have statistically significant heteroskedasticity. This noise variation can be due to technical noise from the mean-variance relationship, variation due to sequencing processes, or underlying

biological differences, making it difficult to parse out the mean-variance relationship. Additionally, further analysis of the genes detected by NoVaTest showed that some genes are likely affected by the mean-variance relationship, and the authors suggest using a strong variance-stabilizing transformation.

We recognize there are limitations to our project and aim to address these in future work. Primarily, simulation studies for spatial transcriptomics data are difficult to design and execute due to numerical instability and limitations of parameterization. There is no clear consensus on the definition of an SVG, so we chose to simulate overall SVGs, defined in Yan et al. [53] as genes that exhibit non-random spatial patterns. To our knowledge, we are not aware of methods to simulate SVGs that include the mean-variance bias. In future work, we aim to refine spatial transcriptomics simulation study design to incorporate the mean-variance relationship and have more flexibility with various parameters, such as mean gene expression, degree of spatial variation, expression strength, and varying effect sizes in the same simulated dataset. We found that our method is most powerful for small lengthscale genes, and we hope to better understand medium and large lengthscale genes in future work as well.

In sum, we provide evidence for the mean-variance and mean-rank relationship in SRT data and show that our method *spoon* can mitigate these biases. We offer the software as an easily installable R/Bioconductor package that interfaces with `SpatialExperiment` to make this method broadly accessible to researchers.

5 Acknowledgments

Funding: Research reported in this publication was supported by the National Institute on Drug Abuse (NIDA) of the National Institutes of Health (NIH) under the award number R01DA053581, supported by the National Institute of Mental Health (NIMH) of the NIH under the award number R01MH126393, and also supported by National Cancer Institute (NCI) number R01CA237170. This project was also supported by CZF2019-002443 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. All funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflict of Interest: None declared.

Author Contributions:

- **Kinnary Shah**: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing - original draft; Writing - review & editing
- **Boyi Guo**: Investigation; Methodology; Software; Visualization; Writing - review & editing
- **Stephanie C. Hicks**: Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing - review & editing

Acknowledgements: We thank members of the Hansen-Hicks lab group and our collaborators at the Lieber Institute for Brain Development for their input and feedback on this project. We also thank maintainers of the Joint High Performance Computing Exchange (JHPCE) computing cluster at Johns Hopkins Bloomberg School of Public Health for computing resources.

References

[1] V. Marx. Method of the Year: spatially resolved transcriptomics. *Nature Methods*, 18(1):9–14, Jan. 2021. ISSN 1548-7091, 1548-7105. doi:10.1038/s41592-020-01033-y. URL https://www.nature.com/articles/s41592-020-01033-y.

[2] A. Deshpande, M. Loth, D. N. Sidiropoulos, S. Zhang, L. Yuan, A. T. Bell, Q. Zhu, W. J. Ho, C. Santa-Maria, D. M. Gilkes, S. R. Williams, C. R. Uytingco, J. Chew, A. Hartnett, Z. W. Bent, A. V. Favorov, A. S. Popel, M. Yarchoan, A. Kiemen, P.-H. Wu, K. Fujikura, D. Wirtz, L. D. Wood, L. Zheng, E. M. Jaffee, R. A. Anders, L. Danilova, G. Stein-O'Brien, L. T. Kagohara, and E. J. Fertig. Uncovering the spatial landscape of molecular interactions within the tumor microenvironment through latent spaces. *Cell Systems*, 14(4):285–301.e4, Apr. 2023. ISSN 24054712. doi:10.1016/j.cels.2023.03.004. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471223000807.

[3] A. Rao, D. Barkley, G. S. França, and I. Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, Aug. 2021. ISSN 0028-0836, 1476-4687. doi:10.1038/s41586-021-03634-9. URL https://www.nature.com/articles/s41586-021-03634-9.

[4] L. Garcia-Alonso, V. Lorenzi, C. I. Mazzeo, J. P. Alves-Lopes, K. Roberts, C. Sancho-Serra, J. Engelbert, M. Marečková, W. H. Gruhn, R. A. Botting, T. Li, B. Crespo, S. Van Dongen, V. Y. Kiselev, E. Prigmore, M. Herbert, A. Moffett, A. Chédotal, O. A. Bayraktar, A. Surani, M. Haniffa, and R. Vento-Tormo. Single-cell roadmap of human gonadal development. *Nature*, 607(7919):540–547, July 2022. ISSN 0028-0836, 1476-4687. doi:10.1038/s41586-022-04918-4. URL https://www.nature.com/articles/s41586-022-04918-4.

[5] K. S. Chen, M. H. Noureldein, D. M. Rigan, J. M. Hayes, M. G. Savelieff, and E. L. Feldman. Regional interneuron transcriptional changes reveal pathologic markers of disease progression in a mouse model of Alzheimer's disease, Nov. 2023. URL https://www.biorxiv.org/content/10.1101/2023.11.01.565165v1.

[6] Y. Vanrobaeys, Z. J. Peterson, E. N. Walsh, S. Chatterjee, L.-C. Lin, L. C. Lyons, T. Nickl-Jockschat, and T. Abel. Spatial transcriptomics reveals unique gene expression changes in different brain regions after sleep deprivation. *Nature Communications*, 14(1):7095, Nov. 2023. ISSN 2041-1723. doi:10.1038/s41467-023-42751-z. URL https://www.nature.com/articles/s41467-023-42751-z.

[7] M. Totty, S. C. Hicks, and B. Guo. SpotSweeper: spatially-aware quality control for spatial transcriptomics, June 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.06.06.597765.

[8] Z. Yuan, F. Zhao, S. Lin, Y. Zhao, J. Yao, Y. Cui, X.-Y. Zhang, and Y. Zhao. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21(4):712–722, Apr. 2024. ISSN 1548-7091, 1548-7105. doi:10.1038/s41592-024-02215-8. URL https://www.nature.com/articles/s41592-024-02215-8.

[9] L. M. Weber, A. Saha, A. Datta, K. D. Hansen, and S. C. Hicks. nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes. *Nature Communications*, 14(1):4059, July 2023. ISSN 2041-1723. doi:10.1038/s41467-023-39748-z. URL https://www.nature.com/articles/s41467-023-39748-z.

[10] V. Svensson, S. A. Teichmann, and O. Stegle. SpatialDE: identification of spatially variable genes. *Nature Methods*, 15(5):343–346, May 2018. ISSN 1548-7091, 1548-7105. doi:10.1038/nmeth.4636. URL https://www.nature.com/articles/nmeth.4636.

[11] J. Zhu, S. Sun, and X. Zhou. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22(1):184, Dec. 2021. ISSN 1474-760X. doi:10.1186/s13059-021-02404-0. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02404-0.

[12] M. Hao, K. Hua, and X. Zhang. SOMDE: a scalable method for identifying spatially variable genes with self-organizing map. *Bioinformatics*, 37(23):4392–4398, Dec. 2021. ISSN 1367-4803, 1367-4811. doi:10.1093/bioinformatics/btab471. URL https://academic.oup.com/bioinformatics/article/37/23/4392/6308937.

[13] R. Dries, Q. Zhu, R. Dong, C.-H. L. Eng, H. Li, K. Liu, Y. Fu, T. Zhao, A. Sarkar, F. Bao, R. E. George, N. Pierson, L. Cai, and G.-C. Yuan. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22(1):78, Dec. 2021. ISSN 1474-760X. doi:10.1186/s13059-021-02286-2. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02286-2.

[14] Z. Li, Z. M.Patel, D. Song, G. Yan, J. J. Li, and L. Pinello. Benchmarking computational methods to identify spatially variable genes and peaks, Dec. 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.12.02.569717.

[15] K. R. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich, M. Tippani, J. Chew, Y. Yin, J. E. Kleinman, T. M. Hyde, N. Rao, S. C. Hicks, K. Martinowich, and A. E. Jaffe. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, Mar. 2021. ISSN 1546-1726. doi:10.1038/s41593-020-00787-0. URL https://www.nature.com/articles/s41593-020-00787-0.

[16] E. D. Nelson, M. Tippani, A. D. Ramnauth, H. R. Divecha, R. A. Miller, N. J. Eagles, E. A. Pattie, S. H. Kwon, S. V. Bach, U. M. Kaipa, J. Yao, J. E. Kleinman, L. Collado-Torres, S. Han, K. R. Maynard, T. M. Hyde, K. Martinowich, S. C. Page, and S. C. Hicks. An integrated single-nucleus and spatial transcriptomics atlas reveals the molecular landscape of the human hippocampus, Apr. 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.04.26.590643.

[17] B. L. Walker, Z. Cang, H. Ren, E. Bourgain-Chang, and Q. Nie. Deciphering tissue structure and function using spatial transcriptomics. *Communications Biology*, 5(1):1–10, Mar. 2022. ISSN 2399-3642. doi:10.1038/s42003-022-03175-5. URL https://www.nature.com/articles/s42003-022-03175-5.

[18] Y. Wang, S. Ma, and W. L. Ruzzo. Spatial modeling of prostate cancer metabolic gene expression reveals extensive heterogeneity and selective vulnerabilities. *Scientific Reports*, 10:3490, Feb. 2020. ISSN 2045-2322. doi:10.1038/s41598-020-60384-w. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7044328/.

[19] J. F. Navarro, D. L. Croteau, A. Jurek, Z. Andrusivova, B. Yang, Y. Wang, B. Ogedegbe, T. Riaz, M. Stoen, C. Desler, L. J. Rasmussen, T. Tonjum, M.-C. Galas, J. Lundeberg, and V. A. Bohr. Spatial Transcriptomics Reveals Genes Associated with Dysregulated Mitochondrial Functions and Stress Signaling in Alzheimer Disease. *iScience*, 23(10), Oct. 2020. ISSN

2589-0042. doi:10.1016/j.isci.2020.101556. URL https://www.cell.com/iscience/abstract/S2589-0042(20)30748-3.

[20] A. Saha and A. Datta. BRISC: bootstrap for rapid inference on spatial covariances: Rapid bootstrap for spatial covariances. *Stat*, 7(1):e184, 2018. ISSN 20491573. doi:10.1002/sta4.184. URL https://onlinelibrary.wiley.com/doi/10.1002/sta4.184.

[21] N. Eling, A. C. Richard, S. Richardson, J. C. Marioni, and C. A. Vallejos. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Systems*, 7(3):284–294.e12, Sept. 2018. ISSN 2405-4712. doi:10.1016/j.cels.2018.06.011. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6167088/.

[22] C. Ahlmann-Eltze and W. Huber. Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, 20(5):665–672, May 2023. ISSN 1548-7105. doi:10.1038/s41592-023-01814-1. URL https://www.nature.com/articles/s41592-023-01814-1.

[23] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095, Nov. 2013. ISSN 1548-7105. doi:10.1038/nmeth.2645. URL https://www.nature.com/articles/nmeth.2645.

[24] V. Antolović, A. Miermont, A. M. Corrigan, and J. R. Chubb. Generation of Single-Cell Transcript Variability by Repression. *Current Biology*, 27(12):1811, June 2017. doi:10.1016/j.cub.2017.05.028. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5483230/.

[25] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. ISSN 1474-760X. doi:10.1186/s13059-014-0550-8.

[26] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, Apr. 2015. ISSN 1362-4962. doi:10.1093/nar/gkv007.

[27] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan. 2010. ISSN 1367-4811, 1367-4803. doi:10.1093/bioinformatics/btp616. URL https://academic.oup.com/bioinformatics/article/26/1/139/182458.

[28] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):295, Dec. 2019. ISSN 1474-760X. doi:10.1186/s13059-019-1861-6. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1861-6.

[29] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, and R. Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021. ISSN 00928674. doi:10.1016/j.cell.2021.04.048. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867421005833.

[30] C. Hafemeister and R. Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, Dec. 2019. ISSN 1474-760X. doi:10.1186/s13059-019-1874-1. URL https://doi.org/10.1186/s13059-019-1874-1.

[31] D. Edsgärd, P. Johnsson, and R. Sandberg. Identification of spatial expression trends in single-cell gene expression data. *Nature Methods*, 15(5):339–342, May 2018. ISSN 1548-7105. doi:10.1038/nmeth.4634.

[32] A. S. Booeshaghi and L. Pachter. Normalization of single-cell RNA-seq counts by $\log(x + 1)$ or $\log(1 + x)$. *Bioinformatics*, 37(15):2223–2224, Mar. 2021. ISSN 1367-4803. doi:10.1093/bioinformatics/btab085. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7989636/.

[33] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29, 2014. ISSN 1465-6906. doi:10.1186/gb-2014-15-2-r29. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29.

[34] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, Feb. 2015. ISSN 1546-1696. doi:10.1038/nbt.3102.

[35] K. Yang, J. Tu, and T. Chen. Homoscedasticity: an overlooked critical assumption for linear regression. *General Psychiatry*, 32(5):e100148, Oct. 2019. ISSN 2517-729X. doi:10.1136/gpsych-2019-100148. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6802968/.

[36] S. Sun, J. Zhu, and X. Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, 17(2):193–200, Feb. 2020. ISSN 1548-7105. doi:10.1038/s41592-019-0701-7. URL https://www.nature.com/articles/s41592-019-0701-7.

[37] A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association*, 111(514):800–812, Apr. 2016. ISSN 0162-1459, 1537-274X. doi:10.1080/01621459.2015.1044091. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2015.1044091.

[38] A. O. Finley, A. Datta, B. C. Cook, D. C. Morton, H. E. Andersen, and S. Banerjee. Efficient algorithms for Bayesian Nearest Neighbor Gaussian Processes. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 28(2):401–414, 2019. ISSN 1061-8600. doi:10.1080/10618600.2018.1537924.

[39] 10x Genomics. Human Breast Cancer: Visium Fresh Frozen, Whole Transcriptome, July 2022. URL https://www.10xgenomics.com/resources/datasets/human-breast-cancer-visium-fresh-frozen-whole-transcriptome-1-standard.

[40] 10x Genomics. Human Breast Cancer: Whole Transcriptome Analysis, Oct. 2020. URL https://www.10xgenomics.com/datasets/human-breast-cancer-whole-transcriptome-analysis-1-standard-1-2-0.

[41] S. Z. Wu, G. Al-Eryani, D. Roden, S. Junankar, K. Harvey, A. Andersson, A. Thennavan, C. Wang, J. Torpy, N. Bartonicek, T. Wang, L. Larsson, D. Kaczorowski, N. I. Weisenfeld, C. R. Uytingco, J. G. Chew, Z. W. Bent, C.-L. Chan, V. Gnanasambandapillai, C.-A. Dutertre, L. Gluch, M. N. Hui, J. Beith, A. Parker, E. Robbins, D. Segara, C. Cooper, C. Mak, B. Chan, S. Warrier, F. Ginhoux, E. Millar, J. E. Powell, S. R. Williams, X. S. Liu, S. O'Toole, E. Lim, J. Lundeberg, C. M. Perou, and A. Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, Sept. 2021. ISSN 1061-4036. doi:10.1038/s41588-021-00911-1. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9044823/.

[42] L. M. Weber, H. R. Divecha, M. N. Tran, S. H. Kwon, A. Spangler, K. D. Montgomery, M. Tippani, R. Bharadwaj, J. E. Kleinman, S. C. Page, T. M. Hyde, L. Collado-Torres, K. R. Maynard, K. Martinowich, and S. C. Hicks. The gene expression landscape of the human locus coeruleus revealed by single-nucleus and spatially-resolved transcriptomics. *eLife*, 12, Feb. 2023. doi:10.7554/eLife.84628. URL https://elifesciences.org/reviewed-preprints/84628.

[43] E. Denisenko, L. de Kock, A. Tan, A. B. Beasley, M. Beilin, M. E. Jones, R. Hou, D. O. Muiri, S. Bilic, G. R. K. A. Mohan, S. Salfinger, S. Fox, K. P. W. Hmon, Y. Yeow, Y. Kim, R. John, T. S. Gilderman, E. Killingbeck, E. S. Gray, P. A. Cohen, Y. Yu, and A. R. R. Forrest. Spatial transcriptomics reveals discrete tumour microenvironments and autocrine loops within ovarian cancer subclones. *Nature Communications*, 15(1):2860, Apr. 2024. ISSN 2041-1723. doi:10.1038/s41467-024-47271-y. URL https://www.nature.com/articles/s41467-024-47271-y.

[44] P. A. P. Moran. NOTES ON CONTINUOUS STOCHASTIC PHENOMENA. *Biometrika*, 37(1-2): 17–23, 1950. ISSN 0006-3444, 1464-3510. doi:10.1093/biomet/37.1-2.17. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/37.1-2.17.

[45] Y. Chang, J. Liu, A. Ma, Z. Li, B. Liu, and Q. Ma. SpaGFT is a graph Fourier transform for tissue module identification from spatially resolved transcriptomics, Dec. 2022. URL https://www.biorxiv.org/content/10.1101/2022.12.10.519929v1.

[46] I. Kats, R. Vento-Tormo, and O. Stegle. SpatialDE2: Fast and localized variance component analysis of spatial transcriptomics, Nov. 2021. URL https://www.biorxiv.org/content/10.1101/2021.10.27.466045v2.

[47] M. Tsagris and M. Papadakis. Taking R to its limits: 70+ tips, Mar. 2018. URL https://peerj.com/preprints/26605v1.

[48] C. Chen, H. J. Kim, and P. Yang. Evaluating spatially variable gene detection methods for spatial transcriptomics data. *Genome Biology*, 25(1):18, Jan. 2024. ISSN 1474-760X. doi:10.1186/s13059-023-03145-y. URL https://doi.org/10.1186/s13059-023-03145-y.

[49] X. Chen, Q. Ran, J. Tang, Z. Chen, S. Huang, X. Shi, and R. Xi. Benchmarking algorithms for spatially variable gene identification in spatial transcriptomics, July 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.07.04.602147.

[50] M. Opławski, A. Srednicka, E. Niewiadomska, D. Boron, P. Januszyk, and B. O. Grabarek. Clinical and molecular evaluation of patients with ovarian cancer in the context of drug resistance to chemotherapy. *Frontiers in Oncology*, 12:954008, Aug. 2022. ISSN 2234-943X. doi:10.3389/fonc.2022.954008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9389532/.

[51] Z. Xu, X. Li, H. Li, C. Nie, W. Liu, S. Li, Z. Liu, W. Wang, and J. Wang. Suppression of DDX39B sensitizes ovarian cancer cells to DNA-damaging chemotherapeutic agents via destabilizing BRCA1 mRNA. *Oncogene*, 39(47):7051–7062, Nov. 2020. ISSN 1476-5594. doi:10.1038/s41388-020-01482-x. URL `https://www.nature.com/articles/s41388-020-01482-x`.

[52] M. A. Abrar, M. Kaykobad, M. S. Rahman, and M. A. H. Samee. NoVaTeST: identifying genes with location-dependent noise variance in spatial transcriptomics data. *Bioinformatics*, 39(6):btad372, June 2023. ISSN 1367-4811. doi:10.1093/bioinformatics/btad372. URL `https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btad372/7191774`.

[53] G. Yan, S. H. Hua, and J. J. Li. Categorization of 33 computational methods to detect spatially variable genes from spatially resolved transcriptomics data, 2024. URL `https://arxiv.org/abs/2405.18779`.

[54] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, Dec. 2017. ISSN 1474-760X. doi:10.1186/s13059-017-1305-0. URL `http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1305-0`.

[55] R. L. Milne, B. Burwinkel, K. Michailidou, J.-I. Arias-Perez, M. P. Zamora, P. Menéndez-Rodríguez, D. Hardisson, M. Mendiola, A. González-Neira, G. Pita, M. R. Alonso, J. Dennis, Q. Wang, M. K. Bolla, A. Swerdlow, A. Ashworth, N. Orr, M. Schoemaker, Y.-D. Ko, H. Brauch, U. Hamann, I. L. Andrulis, J. A. Knight, G. Glendon, S. Tchatchou, K. Matsuo, H. Ito, H. Iwata, K. Tajima, J. Li, J. S. Brand, H. Brenner, A. K. Dieffenbach, V. Arndt, C. Stegmaier, D. Lambrechts, G. Peuteman, M.-R. Christiaens, A. Smeets, A. Jakubowska, J. Lubinski, K. Jaworska-Bieniek, K. Durda, M. Hartman, M. Hui, W. Yen Lim, C. Wan Chan, F. Marme, R. Yang, P. Bugert, A. Lindblom, S. Margolin, M. García-Closas, S. J. Chanock, J. Lissowska, J. D. Figueroa, S. E. Bojesen, B. G. Nordestgaard, H. Flyger, M. J. Hooning, M. Kriege, A. M. van den Ouweland, L. B. Koppert, O. Fletcher, N. Johnson, I. dos Santos-Silva, J. Peto, W. Zheng, S. Deming-Halverson, M. J. Shrubsole, J. Long, J. Chang-Claude, A. Rudolph, P. Seibold, D. Flesch-Janys, R. Winqvist, K. Pylkäs, A. Jukkola-Vuorinen, M. Grip, A. Cox, S. S. Cross, M. W. Reed, M. K. Schmidt, A. Broeks, S. Cornelissen, L. Braaf, D. Kang, J.-Y. Choi, S. K. Park, D.-Y. Noh, J. Simard, M. Dumont, M. S. Goldberg, F. Labrèche, P. A. Fasching, A. Hein, A. B. Ekici, M. W. Beckmann, P. Radice, P. Peterlongo, J. Azzollini, M. Barile, E. Sawyer, I. Tomlinson, M. Kerin, N. Miller, J. L. Hopper, D. F. Schmidt, E. Makalic, M. C. Southey, S. Hwang Teo, C. Har Yip, K. Sivanandan, W.-T. Tay, C.-Y. Shen, C.-N. Hsiung, J.-C. Yu, M.-F. Hou, P. Guénel, T. Truong, M. Sanchez, C. Mulot, W. Blot, Q. Cai, H. Nevanlinna, T. A. Muranen, K. Aittomäki, C. Blomqvist, A. H. Wu, C.-C. Tseng, D. Van Den Berg, D. O. Stram, N. Bogdanova, T. Dörk, K. Muir, A. Lophatananon, S. Stewart-Brown, P. Siriwanarangsan, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, X.-O. Shu, W. Lu, Y.-T. Gao, B. Zhang, F. J. Couch, A. E. Toland, D. Yannoukakos, S. Sangrajrang, J. McKay, X. Wang, J. E. Olson, C. Vachon, K. Purrington, G. Severi, L. Baglietto, C. A. Haiman, B. E. Henderson, F. Schumacher, L. Le Marchand, P. Devilee, R. A. Tollenaar, C. Seynaeve, K. Czene, M. Eriksson, K. Humphreys, H. Darabi, S. Ahmed, M. Shah, P. D. Pharoah, P. Hall, G. G. Giles, J. Benítez, A. M. Dunning, G. Chenevix-Trench, D. F. Easton, A. Berchuck, R. A. Eeles, A. A. A. Olama, Z. Kote-Jarai, S. Benlloch, A. Antoniou, L. McGuffog, K. Offit, A. Lee, E. Dicks, C. Luccarini, D. C. Tessier, F. Bacot, D. Vincent, S. LaBoissière, F. Robidoux, S. F. Nielsen, J. M. Cunningham, S. A. Windebank, C. A. Hilker, J. Meyer, M. Angelakos, J. Maskiell, E. van der Schoot, E. Rutgers, S. Verhoef, F. Hogervorst, P. Boonyawongviroj, P. Siriwanarungsan, M. Schrauder, M. Rübner, S. Oeser, S. Landrith, E. Williams, E. Ryder-Mills, K. Sargus, N. McInerney, G. Colleran, A. Rowan, A. Jones, C. Sohn,

A. Schneeweiß, P. Bugert, N. Álvarez, J. Lacey, S. Wang, H. Ma, Y. Lu, D. Deapen, R. Pinder, E. Lee, F. Schumacher, P. Horn-Ross, P. Reynolds, D. Nelson, H. Ziegler, S. Wolf, V. Hermann, W.-Y. Lo, C. Justenhoven, C. Baisch, H.-P. Fischer, T. Brüning, B. Pesch, S. Rabstein, A. Lotz, V. Harth, T. Heikkinen, I. Erkkilä, K. Aaltonen, K. von Smitten, N. Antonenkova, P. Hillemanns, H. Christiansen, E. Myöhänen, H. Kemiläinen, H. Thorne, E. Niedermayr, D. Bowtell, G. Chenevix-Trench, A. deFazio, D. Gertig, A. Green, P. Webb, A. Green, P. Parsons, N. Hayward, P. Webb, D. Whiteman, A. Fung, J. Yashiki, G. Peuteman, D. Smeets, T. V. Brussel, K. Corthouts, N. Obi, J. Heinz, S. Behrens, U. Eilber, M. Celik, T. Olchers, S. Manoukian, B. Peissel, G. Scuvera, D. Zaffaroni, B. Bonanni, I. Feroce, A. Maniscalco, A. Rossi, L. Bernard, M. Tranchant, M.-F. Valois, A. Turgeon, L. Heguy, P. Sze Yee, P. Kang, K. I. Nee, S. Mariapun, Y. Sook-Yee, D. Lee, T. Y. Ching, N. A. M. Taib, M. Otsukka, K. Mononen, T. Selander, N. Weerasooriya, O. staff, E. Krol-Warmerdam, J. Molenaar, J. Blom, L. Brinton, N. Szeszenia-Dabrowska, B. Peplonska, W. Zatonski, P. Chao, M. Stagner, P. Bos, J. Blom, E. Crepin, A. Nieuwlaat, A. Heemskerk, S. Higham, S. Cross, H. Cramp, D. Connley, S. Balasubramanian, I. Brock, C. Luccarini, D. Conroy, C. Baynes, and K. Chua. Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium. *Human Molecular Genetics*, 23(22):6096–6111, Nov. 2014. ISSN 0964-6906. doi:10.1093/hmg/ddu311. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4204770/`.

[56] S. Jusino, Y. Rivera-Rivera, C. Chardón-Colón, A. J. Ruiz-Justiz, J. Vélez-Velázquez, A. Isidro, M. E. Cruz-Robles, M. Bonilla-Claudio, G. N. Armaiz-Pena, and H. I. Saavedra. E2F3 drives the epithelial-to-mesenchymal transition, cell invasion, and metastasis in breast cancer. *Experimental Biology and Medicine*, 246(19):2057–2071, Oct. 2021. ISSN 1535-3702, 1535-3699. doi:10.1177/15353702211035693. URL `http://journals.sagepub.com/doi/10.1177/15353702211035693`.

[57] Y. Jin, M. Zhai, R. Cao, H. Yu, C. Wu, and Y. Liu. Silencing MFHAS1 Induces Pyroptosis via the JNK-activated NF-KB/Caspase1/ GSDMD Signal Axis in Breast Cancer. *Current Pharmaceutical Design*, 29(42):3408–3420, 2023. ISSN 1873-4286. doi:10.2174/0113816128268130231026054649.

[58] W. Yang, B. Han, Y. Chen, and F. Geng. SAAL1, a novel oncogene, is associated with prognosis and immunotherapy in multiple types of cancer. *Aging (Albany NY)*, 14(15):6316, Aug. 2022. doi:10.18632/aging.204224. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9417231/`.

[59] X. Jiao, S. D. Hooper, T. Djureinovic, C. Larsson, F. Wärnberg, C. Tellgren-Roth, J. Botling, and T. Sjöblom. Gene rearrangements in hormone receptor negative breast cancers revealed by mate pair sequencing. *BMC Genomics*, 14:165, Mar. 2013. ISSN 1471-2164. doi:10.1186/1471-2164-14-165. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3600027/`.

[60] L. Li, X. Li, L. Qi, P. Rychahou, N. Jafari, and C. Huang. The role of talin2 in breast cancer tumorigenesis and metastasis. *Oncotarget*, 8(63):106876, Dec. 2017. doi:10.18632/oncotarget.22449. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5739781/`.

[61] Q. Yin, C. J. Wyatt, T. Han, K. S. Smalley, and L. Wan. ITCH as a potential therapeutic target in human cancers. *Seminars in cancer biology*, 67(Pt 2):117–130, Dec. 2020. ISSN 1044-579X. doi:10.1016/j.semcancer.2020.03.003. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7724637/`.

[62] Y. You, Y. Ma, Q. Wang, Z. Ye, Y. Deng, and F. Bai. Attenuated ZHX3 expression serves as a potential biomarker that predicts poor clinical outcomes in breast cancer patients. *Cancer Manage-*

ment and Research, 11:1199–1210, Feb. 2019. ISSN 1179-1322. doi:10.2147/CMAR.S184340. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6368119/.

[63] Y. Wei, D. Zhang, H. Shi, H. Qian, H. Chen, Q. Zeng, F. Jin, Y. Ye, Z. Ou, M. Guo, B. Guo, and T. Chen. PDK1 promotes breast cancer progression by enhancing the stability and transcriptional activity of HIF-1alpha. *Genes & Diseases*, 11(4):101041, July 2024. ISSN 2352-3042. doi:10.1016/j.gendis.2023.06.013. URL https://www.sciencedirect.com/science/article/pii/S2352304223003112.

[64] D. Zhu, Z. Zhao, G. Cui, S. Chang, L. Hu, Y. X. See, M. G. L. Lim, D. Guo, X. Chen, B. Poudel, P. Robson, Y. Luo, and E. Cheung. Single-Cell Transcriptome Analysis Reveals Estrogen Signaling Coordinately Augments One-Carbon, Polyamine, and Purine Synthesis in Breast Cancer. *Cell Reports*, 25(8):2285–2298.e4, Nov. 2018. ISSN 22111247. doi:10.1016/j.celrep.2018.10.093. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124718317145.

[65] S. Ma, N. Ren, and Q. Huang. rs10514231 Leads to Breast Cancer Predisposition by Altering ATP6AP1L Gene Expression. *Cancers*, 13(15):3752, July 2021. ISSN 2072-6694. doi:10.3390/cancers13153752. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8345087/.

[66] A. Shinde, N. Chandak, J. Singh, M. Roy, M. Mane, X. Tang, H. Vasiyani, F. Currim, D. Gohel, S. Shukla, S. Goyani, M. V. Saranga, D. N. Brindley, and R. Singh. TNF-alpha induced NF-KB mediated LYRM7 expression modulates the tumor growth and metastatic ability in breast cancer. *Free Radical Biology and Medicine*, 211:158–170, Feb. 2024. ISSN 0891-5849. doi:10.1016/j.freeradbiomed.2023.12.018. URL https://www.sciencedirect.com/science/article/pii/S0891584923011747.

[67] H. Wu, X. Guo, Y. Jiao, Z. Wu, and Q. Lv. TRIM35 ubiquitination regulates the expression of PKM2 tetramer and dimer and affects the malignant behaviour of breast cancer by regulating the Warburg effect. *International Journal of Oncology*, 61(6):144, Oct. 2022. ISSN 1019-6439. doi:10.3892/ijo.2022.5434. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9581112/.

[68] T.-T. Zhao, F. Jin, J.-G. Li, Y.-Y. Xu, H.-T. Dong, Q. Liu, P. Xing, G.-L. Zhu, H. Xu, S.-C. Yin, and Z.-F. Miao. TRIM32 promotes proliferation and confers chemoresistance to breast cancer cells through activation of the NF-KB pathway. *Journal of Cancer*, 9(8):1349–1356, Apr. 2018. ISSN 1837-9664. doi:10.7150/jca.22390. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5929078/.

[69] Y.-H. Ko, M. Domingo-Vidal, M. Roche, Z. Lin, D. Whitaker-Menezes, E. Seifert, C. Capparelli, M. Tuluc, R. C. Birbe, P. Tassone, J. M. Curry, A. Navarro-Sabate, A. Manzano, R. Bartrons, J. Caro, and U. Martinez-Outschoorn. TP53-inducible Glycolysis and Apoptosis Regulator (TIGAR) Metabolically Reprograms Carcinoma and Stromal Cells in Breast Cancer. *The Journal of Biological Chemistry*, 291(51):26291–26303, Dec. 2016. ISSN 0021-9258. doi:10.1074/jbc.M116.740209. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5159492/.

[70] V. Raghavan and D. B. Manasa. Identification and Analysis of Disease Target Network of Human MicroRNA and Predicting Promising Leads for ZNF439, a Potential Target for Breast Cancer. *International Journal of Bioscience, Biochemistry and Bioinformatics*, pages 358–362, 2012. ISSN 20103638. doi:10.7763/IJBBB.2012.V2.132. URL http://www.ijbbb.org/show-33-382-1.html.

[71] Y. Salem, N. Yacov, O. Propheta-Meiran, E. Breitbart, and I. Mendel. Newly characterized motile sperm domain-containing protein 2 promotes human breast cancer metastasis. *International Journal of Cancer*, 144(1):125–135, Jan. 2019. ISSN 1097-0215. doi:10.1002/ijc.31665.

[72] E. J. Suh, M. H. Kabir, U.-B. Kang, J. W. Lee, J. Yu, D.-Y. Noh, and C. Lee. Comparative profiling of plasma proteome from breast cancer patients reveals thrombospondin-1 and BRWD3 as serological biomarkers. *Experimental & Molecular Medicine*, 44(1):36–44, Jan. 2012. ISSN 1226-3613. doi:10.3858/emm.2012.44.1.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3277896/.

[73] L. He, J. Yang, Y. Hao, X. Yang, X. Shi, D. Zhang, D. Zhao, W. Yan, X. Bie, L. Chen, G. Chen, S. Zhao, X. Liu, H. Zheng, and K. Zhang. DDX20: A Multifunctional Complex Protein. *Molecules*, 28(20):7198, Oct. 2023. ISSN 1420-3049. doi:10.3390/molecules28207198. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10608988/.

[74] U.-H. Park, M.-R. Kang, E.-J. Kim, Y.-S. Kwon, W. Hur, S. K. Yoon, B.-J. Song, J. H. Park, J.-T. Hwang, J.-C. Jeong, and S.-J. Um. ASXL2 promotes proliferation of breast cancer cells by linking ERalpha to histone methylation. *Oncogene*, 35(28):3742–3752, July 2016. ISSN 1476-5594. doi:10.1038/onc.2015.443. URL https://www.nature.com/articles/onc2015443.

[75] D. Li, Y. Li, X. Wu, Q. Li, J. Yu, J. Gen, and X.-L. Zhang. Knockdown of Mgat5 Inhibits Breast Cancer Cell Growth with Activation of CD4+ T Cells and Macrophages. *The Journal of Immunology*, 180(5):3158–3165, Mar. 2008. ISSN 0022-1767, 1550-6606. doi:10.4049/jimmunol.180.5.3158. URL https://journals.aai.org/jimmunol/article/180/5/3158/78739/Knockdown-of-Mgat5-Inhibits-Breast-Cancer-Cell.

[76] P. Wang, Q. Zhang, H. Zhang, J. Shao, H. Zhang, and Z. Wang. Molecular and clinical characterization of ICOS expression in breast cancer through large-scale transcriptome data. *PLOS ONE*, 18(12):e0293469, Dec. 2023. ISSN 1932-6203. doi:10.1371/journal.pone.0293469. URL https://dx.plos.org/10.1371/journal.pone.0293469.

[77] I. X. Perez-Añorve, C. H. Gonzalez-De la Rosa, E. Soto-Reyes, F. O. Beltran-Anaya, O. Del Moral-Hernandez, M. Salgado-Albarran, O. Angeles-Zaragoza, J. A. Gonzalez-Barrios, D. A. Landero-Huerta, M. Chavez-Saldaña, A. Garcia-Carranca, N. Villegas-Sepulveda, and E. Arechaga-Ocampo. New insights into radioresistance in breast cancer identify a dual function of miR-122 as a tumor suppressor and oncomiR. *Molecular Oncology*, 13(5):1249–1267, May 2019. ISSN 1574-7891. doi:10.1002/1878-0261.12483. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6487688/.

[78] C. Tian, S. Zhou, and C. Yi. High NUP43 expression might independently predict poor overall survival in luminal A and in HER2+ breast cancer. *Future Oncology (London, England)*, 14(15): 1431–1442, June 2018. ISSN 1744-8301. doi:10.2217/fon-2017-0690.

[79] V. Quereda, S. Bayle, F. Vena, S. M. Frydman, A. Monastyrskyi, W. R. Roush, and D. R. Duckett. Therapeutic Targeting of CDK12/CDK13 in Triple-Negative Breast Cancer. *Cancer Cell*, 36(5): 545–558.e7, Nov. 2019. ISSN 15356108. doi:10.1016/j.ccell.2019.09.004. URL https://linkinghub.elsevier.com/retrieve/pii/S1535610819304246.

[80] S. Mohammed Zaidh, K. B. Aher, G. B. Bhavar, N. Irfan, H. N. Ahmed, and Y. Ismail. Genes adaptability and NOL6 protein inhibition studies of fabricated flavan-3-ols lead skeleton intended to treat breast carcinoma. *International Journal of Biological Macromolecules*, 258:127661, Feb.

2024. ISSN 0141-8130. doi:10.1016/j.ijbiomac.2023.127661. URL https://www.sciencedirect.com/science/article/pii/S0141813023045592.

[81] B. Arko-Boham, B. A. Owusu, N. A. Aryee, R. M. Blay, E. D. A. Owusu, E. A. Tagoe, A. R. Adams, R. K. Gyasi, N. A. Adu-Aryee, and S. Mahmood. Prospecting for Breast Cancer Blood Biomarkers: Death-Associated Protein Kinase 1 (DAPK1) as a Potential Candidate. *Disease Markers*, 2020: 6848703, May 2020. ISSN 0278-0240. doi:10.1155/2020/6848703. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7267859/.

[82] N. Huang, P. Li, X. Sun, L. Tong, X. Dong, X. Zhang, J. Duan, X. Sheng, and H. Xin. TRIM21 mediates the synergistic effect of Olaparib and Sorafenib by degrading BRCA1 through ubiquitination in TNBC. *npj Breast Cancer*, 9(1):1–11, Oct. 2023. ISSN 2374-4677. doi:10.1038/s41523-023-00588-1. URL https://www.nature.com/articles/s41523-023-00588-1.

[83] W. Li, S. Hu, Z. Han, and X. Jiang. YY1-Induced Transcriptional Activation of FAM111B Contributes to the Malignancy of Breast Cancer. *Clinical Breast Cancer*, 22(4):e417–e425, June 2022. ISSN 1526-8209. doi:10.1016/j.clbc.2021.10.008. URL https://www.sciencedirect.com/science/article/pii/S1526820921002986.

[84] X. Chen, H. Peng, Z. Zhang, C. Yang, Y. Liu, Y. Chen, F. Yu, S. Wu, and L. Cao. SPDYC serves as a prognostic biomarker related to lipid metabolism and the immune microenvironment in breast cancer. *Immunologic Research*, June 2024. ISSN 0257-277X, 1559-0755. doi:10.1007/s12026-024-09505-5. URL https://link.springer.com/10.1007/s12026-024-09505-5.

[85] P. Zhang, Y. Yang, K. Qian, L. Li, C. Zhang, X. Fu, X. Zhang, H. Chen, Q. Liu, S. Cao, and J. Cui. A novel tumor suppressor ZBTB1 regulates tamoxifen resistance and aerobic glycolysis through suppressing HER2 expression in breast cancer. *Journal of Biological Chemistry*, 295(41):14140–14152, Oct. 2020. ISSN 00219258. doi:10.1074/jbc.RA119.010759. URL https://linkinghub.elsevier.com/retrieve/pii/S0021925817498097.

[86] X. Wang, Y. Zheng, and Y. Wang. PEAK1 promotes invasion and metastasis and confers drug resistance in breast cancer. *Clinical and Experimental Medicine*, 22(3):393–402, Aug. 2022. ISSN 1591-9528. doi:10.1007/s10238-021-00761-5. URL https://link.springer.com/10.1007/s10238-021-00761-5.

[87] Y. Peng, X. Liu, X. Liu, X. Cheng, L. Xia, L. Qin, S. Guan, Y. Wang, X. Wu, J. Wu, D. Yan, J. Liu, Y. Zhang, L. Sun, J. Liang, and Y. Shang. RCCD1 promotes breast carcinogenesis through regulating hypoxia-associated mitochondrial homeostasis. *Oncogene*, 42(50):3684–3697, Dec. 2023. ISSN 1476-5594. doi:10.1038/s41388-023-02877-2. URL https://www.nature.com/articles/s41388-023-02877-2.

[88] Y. Niu, Z. Lin, A. Wan, H. Chen, H. Liang, L. Sun, Y. Wang, X. Li, X.-f. Xiong, B. Wei, X. Wu, and G. Wan. RNA N6-methyladenosine demethylase FTO promotes breast tumor progression through inhibiting BNIP3. *Molecular Cancer*, 18(1):46, Dec. 2019. ISSN 1476-4598. doi:10.1186/s12943-019-1004-4. URL https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-019-1004-4.

[89] P. Liang, J. Zhang, Y. Wu, S. Zheng, Z. Xu, S. Yang, J. Wang, S. Ma, L. Xiao, T. Hu, W. Jiang, C. Huang, Q. Xing, M. Kundu, and B. Wang. An ULK1/2-PXN mechanotransduction pathway suppresses breast cancer cell migration. *EMBO reports*, 24(11):e56850, Nov. 2023. ISSN

1469-221X, 1469-3178. doi:10.15252/embr.202356850. URL https://www.embopress.org/doi/10.15252/embr.202356850.

[90] Y. Lu, Y. Xiao, J. Yang, H. Su, X. Zhang, F. Su, B. Tian, D. Zhao, X. Ling, and T. Zhang. TRIM65 Promotes Malignant Cell Behaviors in Triple-Negative Breast Cancer by Impairing the Stability of LATS1 Protein. *Oxidative Medicine and Cellular Longevity*, 2022:1–16, Aug. 2022. ISSN 1942-0994, 1942-0900. doi:10.1155/2022/4374978. URL https://www.hindawi.com/journals/omcl/2022/4374978/.

[91] X. Man, Q. Li, B. Wang, H. Zhang, S. Zhang, and Z. Li. DNMT3A and DNMT3B in Breast Tumorigenesis and Potential Therapy. *Frontiers in Cell and Developmental Biology*, 10:916725, May 2022. ISSN 2296-634X. doi:10.3389/fcell.2022.916725. URL https://www.frontiersin.org/articles/10.3389/fcell.2022.916725/full.

[92] N. J. Camp, M. Parry, S. Knight, R. Abo, G. Elliott, S. H. Rigas, S. P. Balasubramanian, M. W. R. Reed, H. McBurney, A. Latif, W. G. Newman, L. A. Cannon-Albright, D. G. Evans, and A. Cox. Fine-Mapping *CASP8* Risk Variants in Breast Cancer. *Cancer Epidemiology, Biomarkers & Prevention*, 21(1):176–181, Jan. 2012. ISSN 1055-9965, 1538-7755. doi:10.1158/1055-9965.EPI-11-0845. URL https://aacrjournals.org/cebp/article/21/1/176/157359/Fine-Mapping-CASP8-Risk-Variants-in-Breast.

[93] J. Xu, S. M. Su, X. Zhang, U. I. Chan, R. Adhav, X. Shu, J. Liu, J. Li, L. Mo, Y. Wang, T. An, J. H. Lei, K. Miao, C.-X. Deng, and X. Xu. ATP11B inhibits breast cancer metastasis in a mouse model by suppressing externalization of nonapoptotic phosphatidylserine. *Journal of Clinical Investigation*, 132(5):e149473, Mar. 2022. ISSN 1558-8238. doi:10.1172/JCI149473. URL https://www.jci.org/articles/view/149473.

[94] S. Kim, K. Kim, J. Ryu, T. Ryu, J. H. Lim, J. Oh, J. Min, C. Jung, R. Hamamoto, M. Son, D. Kim, and H. Cho. The novel prognostic marker, EHMT2, is involved in cell proliferation via HSPD1 regulation in breast cancer. *International Journal of Oncology*, Oct. 2018. ISSN 1019-6439, 1791-2423. doi:10.3892/ijo.2018.4608. URL http://www.spandidos-publications.com/10.3892/ijo.2018.4608.

[95] A. K. Pullikuth, E. D. Routh, K. D. Zimmerman, J. Chifman, J. W. Chou, M. H. Soike, G. Jin, J. Su, Q. Song, M. A. Black, C. Print, D. Bedognetti, M. Howard-McNatt, S. S. O'Neill, A. Thomas, C. D. Langefeld, A. B. Sigalov, Y. Lu, and L. D. Miller. Bulk and Single-Cell Profiling of Breast Tumors Identifies TREM-1 as a Dominant Immune Suppressive Marker Associated With Poor Outcomes. *Frontiers in Oncology*, 11:734959, Dec. 2021. ISSN 2234-943X. doi:10.3389/fonc.2021.734959. URL https://www.frontiersin.org/articles/10.3389/fonc.2021.734959/full.

[96] L. Pei, Y. Li, H. Gu, S. Wang, W. Wu, S. Fan, X. Shi, and X. Si. Identification of SMC2 and SMC4 as prognostic markers in breast cancer through bioinformatics analysis. *Clinical and Translational Oncology*, May 2024. ISSN 1699-3055. doi:10.1007/s12094-024-03521-5. URL https://link.springer.com/10.1007/s12094-024-03521-5.

[97] M. Viera, G. W. C. Yip, H.-M. Shen, G. H. Baeg, and B. H. Bay. Targeting CD82/KAI1 for Precision Therapeutics in Surmounting Metastatic Potential in Breast Cancer. *Cancers*, 13(17):4486, Sept. 2021. ISSN 2072-6694. doi:10.3390/cancers13174486. URL https://www.mdpi.com/2072-6694/13/17/4486.

[98] D. Samanta, T. Y.-T. Huang, R. Shah, Y. Yang, F. Pan, and G. L. Semenza. BIRC2 Expression Impairs Anti-Cancer Immunity and Immunotherapy Efficacy. *Cell Reports*, 32(8):108073, Aug. 2020. ISSN 22111247. doi:10.1016/j.celrep.2020.108073. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124720310585.

[99] E. Hervouet, A. Claude-Taupin, T. Gauthier, V. Perez, A. Fraichard, P. Adami, G. Despouy, F. Monnien, M.-P. Algros, M. Jouvenot, R. Delage-Mourroux, and M. Boyer-Guittaut. The autophagy GABARAPL1 gene is epigenetically regulated in breast cancer models. *BMC Cancer*, 15(1):729, Dec. 2015. ISSN 1471-2407. doi:10.1186/s12885-015-1761-4. URL http://bmccancer.biomedcentral.com/articles/10.1186/s12885-015-1761-4.

[100] M. Siouda, A. D. Dujardin, L. Barbollat-Boutrand, M. A. Mendoza-Parra, B. Gibert, M. Ouzounova, J. Bouaoud, L. Tonon, M. Robert, J.-P. Foy, V. Lavergne, S. N. Manie, A. Viari, A. Puisieux, G. Ichim, H. Gronemeyer, P. Saintigny, and P. Mulligan. CDYL2 Epigenetically Regulates MIR124 to Control NF-KB/STAT3-Dependent Breast Cancer Cell Plasticity. *iScience*, 23(6):101141, June 2020. ISSN 25890042. doi:10.1016/j.isci.2020.101141. URL https://linkinghub.elsevier.com/retrieve/pii/S2589004220303266.

[101] M.-H. Kang, K. J. Jeong, W. Y. Kim, H. J. Lee, G. Gong, N. Suh, B. Győrffy, S. Kim, S.-Y. Jeong, G. B. Mills, and Y.-Y. Park. Musashi RNA-binding protein 2 regulates estrogen receptor 1 function in breast cancer. *Oncogene*, 36(12):1745–1752, Mar. 2017. ISSN 0950-9232, 1476-5594. doi:10.1038/onc.2016.327. URL https://www.nature.com/articles/onc2016327.

[102] W.-F. Hu, K. L. Krieger, D. Lagundžin, X. Li, R. S. Cheung, T. Taniguchi, K. R. Johnson, T. Bessho, A. N. A. Monteiro, and N. T. Woods. CTDP1 regulates breast cancer survival and DNA repair through BRCT-specific interactions with FANCI. *Cell Death Discovery*, 5(1):105, June 2019. ISSN 2058-7716. doi:10.1038/s41420-019-0185-3. URL https://www.nature.com/articles/s41420-019-0185-3.

[103] J. Zhang, Y. Li, J.-G. Wang, J.-Y. Feng, G.-D. Huang, and C.-G. Luo. Dihydroartemisinin Affects STAT3/DDA1 Signaling Pathway and Reverses Breast Cancer Resistance to Cisplatin. *The American Journal of Chinese Medicine*, 51(02):445–459, Jan. 2023. ISSN 0192-415X, 1793-6853. doi:10.1142/S0192415X23500234. URL https://www.worldscientific.com/doi/10.1142/S0192415X23500234.

[104] Y. Peng, H. Li, Y. Fu, S. Guo, C. Qu, Y. Zhang, B. Zong, and S. Liu. JAM2 predicts a good prognosis and inhibits invasion and migration by suppressing EMT pathway in breast cancer. *International Immunopharmacology*, 103:108430, Feb. 2022. ISSN 15675769. doi:10.1016/j.intimp.2021.108430. URL https://linkinghub.elsevier.com/retrieve/pii/S1567576921010663.

[105] S. Zhang, X. Liu, W. Chen, K. Zhang, Q. Wu, and Y. Wei. Targeting TAF1 with BAY-299 induces antitumor immunity in triple-negative breast cancer. *Biochemical and Biophysical Research Communications*, 665:55–63, July 2023. ISSN 0006291X. doi:10.1016/j.bbrc.2023.04.100. URL https://linkinghub.elsevier.com/retrieve/pii/S0006291X23005314.

[106] C. N. Johnstone, A. D. Pattison, P. F. Harrison, D. R. Powell, P. Lock, M. Ernst, R. L. Anderson, and T. H. Beilharz. FGF13 promotes metastasis of triple-negative breast cancer. *International Journal of Cancer*, 147(1):230–243, July 2020. ISSN 0020-7136, 1097-0215. doi:10.1002/ijc.32874. URL https://onlinelibrary.wiley.com/doi/10.1002/ijc.32874.

[107] Z. Chen, N. Cui, J.-s. Zhao, J.-f. Wu, F. Ma, C. Li, and X.-y. Liu. Expressions of ZNF436, beta-catenin, EGFR, and CMTM5 in breast cancer and their clinical significances. *European Journal of Histochemistry*, 65(1), Jan. 2021. ISSN 2038-8306, 1121-760X. doi:10.4081/ejh.2021.3173. URL https://www.ejh.it/index.php/ejh/article/view/3173.

[108] M. Ngubo, F. Moradi, C. Y. Ito, and W. L. Stanford. Tissue-Specific Tumour Suppressor and Oncogenic Activities of the Polycomb-like Protein MTF2. *Genes*, 14(10):1879, Sept. 2023. ISSN 2073-4425. doi:10.3390/genes14101879. URL https://www.mdpi.com/2073-4425/14/10/1879.

[109] L. Jin, C. Luo, X. Wu, M. Li, S. Wu, and Y. Feng. LncRNA-HAGLR motivates triple negative breast cancer progression by regulation of WNT2 via sponging miR-335-3p. *Aging*, 13(15):19306–19316, Aug. 2021. ISSN 1945-4589. doi:10.18632/aging.203272. URL https://www.aging-us.com/lookup/doi/10.18632/aging.203272.

[110] H.-J. Han, J. Russo, Y. Kohwi, and T. Kohwi-Shigematsu. SATB1 reprogrammes gene expression to promote breast tumour growth and metastasis. *Nature*, 452(7184):187–193, Mar. 2008. ISSN 0028-0836, 1476-4687. doi:10.1038/nature06781. URL https://www.nature.com/articles/nature06781.

[111] N. Erin, A. Podnos, G. Tanriover, O. Duymus, E. Cote, I. Khatri, and R. M. Gorczynski. Bidirectional effect of CD200 on breast cancer development and metastasis, with ultimate outcome determined by tumor aggressiveness and a cancer-induced inflammatory response. *Oncogene*, 34(29):3860–3870, July 2015. ISSN 0950-9232, 1476-5594. doi:10.1038/onc.2014.317. URL https://www.nature.com/articles/onc2014317.

[112] P. Tan, Y. Ye, L. He, J. Xie, J. Jing, G. Ma, H. Pan, L. Han, W. Han, and Y. Zhou. TRIM59 promotes breast cancer motility by suppressing p62-selective autophagic degradation of PDCD10. *PLOS Biology*, 16(11):e3000051, Nov. 2018. ISSN 1545-7885. doi:10.1371/journal.pbio.3000051. URL https://dx.plos.org/10.1371/journal.pbio.3000051.

[113] M. Mondal, D. Conole, J. Nautiyal, and E. W. Tate. UCHL1 as a novel target in breast cancer: emerging insights from cell and chemical biology. *British Journal of Cancer*, 126(1):24–33, Jan. 2022. ISSN 0007-0920, 1532-1827. doi:10.1038/s41416-021-01516-5. URL https://www.nature.com/articles/s41416-021-01516-5.

[114] D. Chen, Y. Sun, Y. Wei, P. Zhang, A. H. Rezaeian, J. Teruya-Feldstein, S. Gupta, H. Liang, H.-K. Lin, M.-C. Hung, and L. Ma. LIFR is a breast cancer metastasis suppressor upstream of the Hippo-YAP pathway and a prognostic marker. *Nature Medicine*, 18(10):1511–1517, Oct. 2012. ISSN 1078-8956, 1546-170X. doi:10.1038/nm.2940. URL https://www.nature.com/articles/nm.2940.

[115] I. A. Ivanova, J. F. Vermeulen, C. Ercan, J. M. Houthuijzen, F. A. Saig, E. J. Vlug, E. Van Der Wall, P. J. Van Diest, M. Vooijs, and P. W. B. Derksen. FER kinase promotes breast cancer metastasis by regulating alpha6- and beta1-integrin-dependent cell adhesion and anoikis resistance. *Oncogene*, 32(50):5582–5592, Dec. 2013. ISSN 0950-9232, 1476-5594. doi:10.1038/onc.2013.277. URL https://www.nature.com/articles/onc2013277.

[116] W. Yang, J. Li, M. Zhang, H. Yu, Y. Zhuang, L. Zhao, L. Ren, J. Gong, H. Bi, L. Zeng, Y. Xue, J. Yang, Y. Zhao, S. Wang, S. Gao, Z. Fu, D. Li, J. Zhang, T. Wang, M. Shan, B. Tang, and X. Li. Elevated expression of the rhythm gene NFIL3 promotes the progression of TNBC by activating NF-KB signaling through suppression of NFKBIA transcription. *Journal of Experimental & Clinical*

Cancer Research, 41(1):67, Dec. 2022. ISSN 1756-9966. doi:10.1186/s13046-022-02260-1. URL https://jeccr.biomedcentral.com/articles/10.1186/s13046-022-02260-1.

[117] D. B. Shropshire, F. M. Acosta, K. Fang, J. Benavides, L.-Z. Sun, V. X. Jin, and J. X. Jiang. Association of adenosine signaling gene signature with estrogen receptor-positive breast and prostate cancer bone metastasis. *Frontiers in Medicine*, 9:965429, Sept. 2022. ISSN 2296-858X. doi:10.3389/fmed.2022.965429. URL https://www.frontiersin.org/articles/10.3389/fmed.2022.965429/full.

[118] A. M. Marrufo, S. O. Mathew, P. Chaudhary, J. D. Malaer, J. K. Vishwanatha, and P. A. Mathew. Blocking LLT1 (CLEC2D, OCIL)-NKRP1A (CD161) interaction enhances natural killer cell-mediated lysis of triple-negative breast cancer cells. *American Journal of Cancer Research*, 8 (6):1050–1063, June 2018. ISSN 2156-6976. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6048397/.

[119] Y. Tai, A. Chow, S. Han, C. Coker, W. Ma, Y. Gu, V. Estrada Navarro, M. Kandpal, H. Hibshoosh, K. Kalinsky, K. Manova-Todorova, A. Safonov, E. M. Walsh, M. Robson, L. Norton, R. Baer, T. Merghoub, A. K. Biswas, and S. Acharyya. FLT1 activation in cancer cells promotes PARP-inhibitor resistance in breast cancer. *EMBO Molecular Medicine*, 16(8):1957–1980, July 2024. ISSN 1757-4684. doi:10.1038/s44321-024-00094-2. URL https://www.embopress.org/doi/full/10.1038/s44321-024-00094-2.

[120] Y.-J. Lee, S.-R. Ho, J. D. Graves, Y. Xiao, S. Huang, and W.-C. Lin. CGRRF1, a growth suppressor, regulates EGFR ubiquitination in breast cancer. *Breast Cancer Research*, 21(1):134, Dec. 2019. ISSN 1465-542X. doi:10.1186/s13058-019-1212-2. URL https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-019-1212-2.

[121] R. C. Ekyalongo, T. Mukohara, Y. Funakoshi, H. Tomioka, Y. Kataoka, Y. Shimono, N. Chayahara, M. Toyoda, N. Kiyota, and H. Minami. TYRO3 as a potential therapeutic target in breast cancer. *Anticancer Research*, 34(7):3337–3345, July 2014. ISSN 1791-7530.

[122] Y. Cheng, L. Lin, X. Li, A. Lu, C. Hou, Q. Wu, X. Hu, Z. Zhou, Z. Chen, and F. Tang. ADAM10 is involved in the oncogenic process and chemo-resistance of triple-negative breast cancer via regulating Notch1 signaling pathway, CD44 and PrPc. *Cancer Cell International*, 21(1):32, Jan. 2021. ISSN 1475-2867. doi:10.1186/s12935-020-01727-5. URL https://cancerci.biomedcentral.com/articles/10.1186/s12935-020-01727-5.

[123] Y. Wei, H. Huang, Z. Qiu, H. Li, J. Tan, G. Ren, and X. Wang. NLRP1 Overexpression Is Correlated with the Tumorigenesis and Proliferation of Human Breast Tumor. *BioMed Research International*, 2017:1–9, 2017. ISSN 2314-6133, 2314-6141. doi:10.1155/2017/4938473. URL https://www.hindawi.com/journals/bmri/2017/4938473/.

[124] Z. Huang, J. Yang, W. Qiu, J. Huang, Z. Chen, Y. Han, and C. Ye. HAUS5 Is A Potential Prognostic Biomarker With Functional Significance in Breast Cancer. *Frontiers in Oncology*, 12:829777, Feb. 2022. ISSN 2234-943X. doi:10.3389/fonc.2022.829777. URL https://www.frontiersin.org/articles/10.3389/fonc.2022.829777/full.

[125] D. Tong, J. Zhang, X. Wang, Q. Li, L. Y. Liu, J. Yang, B. Guo, L. Ni, L. Zhao, and C. Huang. MeCP2 facilitates breast cancer growth via promoting ubiquitination-mediated P53 degradation by inhibiting

RPL5/RPL11 transcription. *Oncogenesis*, 9(5):56, June 2020. ISSN 2157-9024. doi:10.1038/s41389-020-0239-7. URL https://www.nature.com/articles/s41389-020-0239-7.

[126] M. Spears, G. J. Rabiasz, D. Scott, E. Ntougkos, S. Fegan, E. P. Miller, J. F. Smyth, and G. C. Sellar. The function of tumor suppressor genes in ovarian cancer: the role of LSAMP. *Cancer Res*, 66 (8_Supplement)(587), Apr. 2006. ISSN 1538-7445.

[127] Y. Shi, S. Gao, Y. Zheng, M. Yao, and F. Ruan. LncRNA CASC15 Functions As An Unfavorable Predictor Of Ovarian Cancer Prognosis And Inhibits Tumor Progression Through Regulation Of miR-221/ARID1A Axis. *OncoTargets and therapy*, 12:8725–8736, Oct. 2019. ISSN 1178-6930. doi:10.2147/OTT.S219900. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6815787/.

[128] A.-r. Liu, Y.-n. Liu, S.-x. Shen, L.-r. Yan, Z. Lv, H.-x. Ding, A. Wang, Y. Yuan, and Q. Xu. Comprehensive Analysis and Validation of Solute Carrier Family 25 (SLC25) and Its Correlation with Immune Infiltration in Pan-Cancer. *BioMed Research International*, 2022:1–23, Oct. 2022. ISSN 2314-6141, 2314-6133. doi:10.1155/2022/4009354. URL https://www.hindawi.com/journals/bmri/2022/4009354/.

[129] J. Wang, D. C. Dean, F. J. Hornicek, H. Shi, and Z. Duan. Cyclin-dependent kinase 9 (CDK9) is a novel prognostic marker and therapeutic target in ovarian cancer. *The FASEB Journal*, 33 (5):5990–6000, May 2019. ISSN 0892-6638, 1530-6860. doi:10.1096/fj.201801789RR. URL https://onlinelibrary.wiley.com/doi/10.1096/fj.201801789RR.

[130] T. Gruosso, C. Garnier, S. Abelanet, Y. Kieffer, V. Lemesre, D. Bellanger, I. Bieche, E. Marangoni, X. Sastre-Garau, V. Mieulet, and F. Mechta-Grigoriou. MAP3K8/TPL-2/COT is a potential predictive marker for MEK inhibitor treatment in high-grade serous ovarian carcinomas. *Nature Communications*, 6(1):8583, Oct. 2015. ISSN 2041-1723. doi:10.1038/ncomms9583. URL https://www.nature.com/articles/ncomms9583.

[131] S. Wang, Y. Xia, P. Huang, C. Xu, Y. Qian, T. Fang, and Q. Gao. Suppression of GCH1 Sensitizes Ovarian Cancer and Breast Cancer to PARP Inhibitor. *Journal of Oncology*, 2023:1–16, Feb. 2023. ISSN 1687-8469, 1687-8450. doi:10.1155/2023/1453739. URL https://www.hindawi.com/journals/jo/2023/1453739/.

[132] Y. Xie, H. Chen, P. Shen, Q. Shen, and Y. Luo. PCYT2-Mediated Regulation of Phospholipid Metabolism Enhances Metastasis in Epithelial Ovarian Cancer via the AKT/mTOR and HIPPO Signaling Pathways. *Journal of Biological Regulators and Homeostatic Agents*, 38(2):1351–1364, Feb. 2024. ISSN 0393-974X. doi:10.23812/j.biol.regul.homeost.agents.20243802.108. URL https://www.biolifesas.org/EN/10.23812/j.biol.regul.homeost.agents.20243802.108.

Supplementary Materials

Addressing the mean-variance relationship in spatially resolved transcriptomics data with *spoon*

Kinnary Shah, Boyi Guo, Stephanie C. Hicks*

*Correspondence to `shicks19@jhu.edu`

Contents

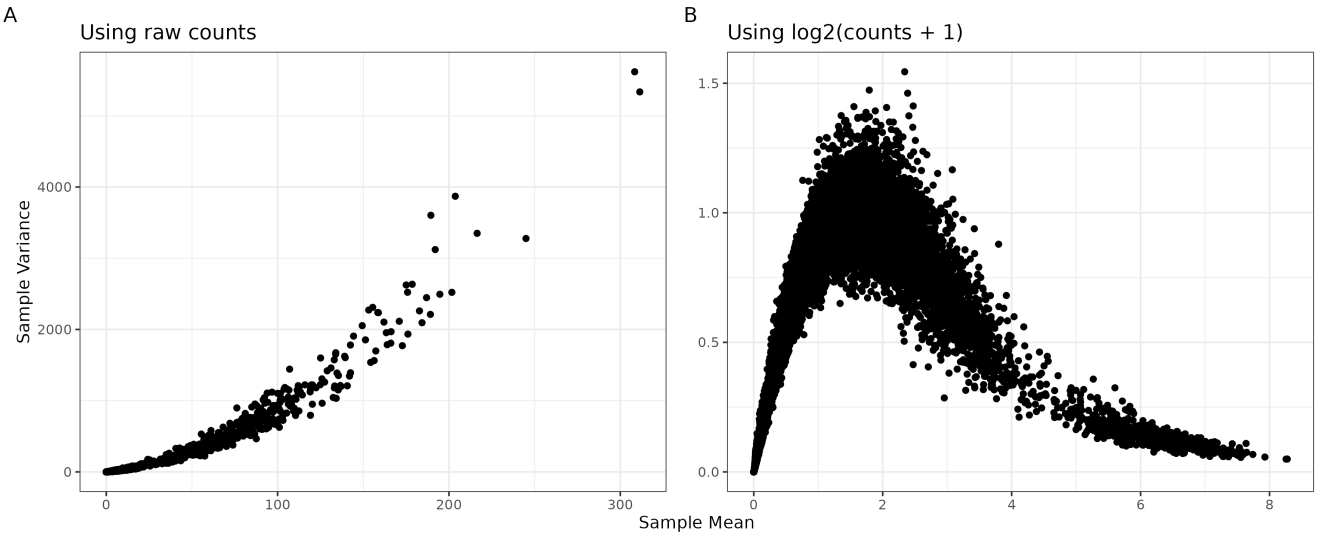

Figure S1: Visualizing the mean-variance relationship on different scales. The mean-variance relationship exists with or without a log-transformation. Gene expression counts were simulated using the `splatter` R/Bioconductor package [54] for G=10,000 genes and N=100 observations (or cells) under a Gamma-Poisson model. Each point represents one gene. Both representations illustrate the mean-variance relationship where the x-axis is the sample mean and y-axis is the sample variance using either (**A**) the raw counts or (**B**) the $\log_2$-transformed counts with a pseudocount of 1 (or $\log_2(\text{counts}+1)$). Here, the log-transformation overcorrects for the mean-variance relationship for the larger counts.

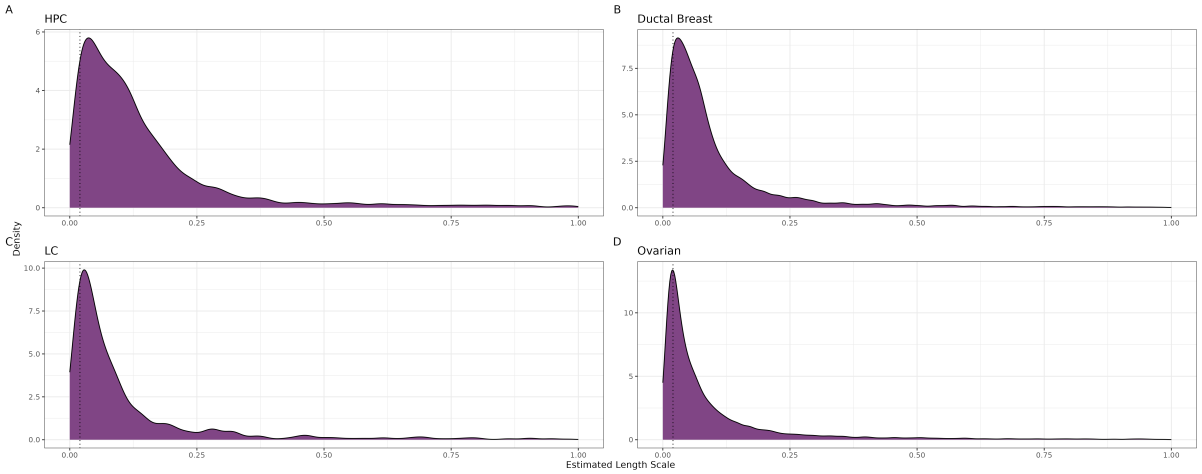

Figure S2: Real data estimated lengthscale distributions using `nnSVG`. This figure shows the estimated lengthscale distributions for four real datasets (**A**) HPC [16], (**B**) Ductal Breast cancer [39], (**C**) LC [42], and (**D**) Ovarian cancer [43]. For each dataset, `nnSVG` was used to calculate the estimated lengthscale value for each gene and the distribution of values between 0 and 1 is plotted. The dotted line highlights the lengthscale value used in the primary simulation evaluations.

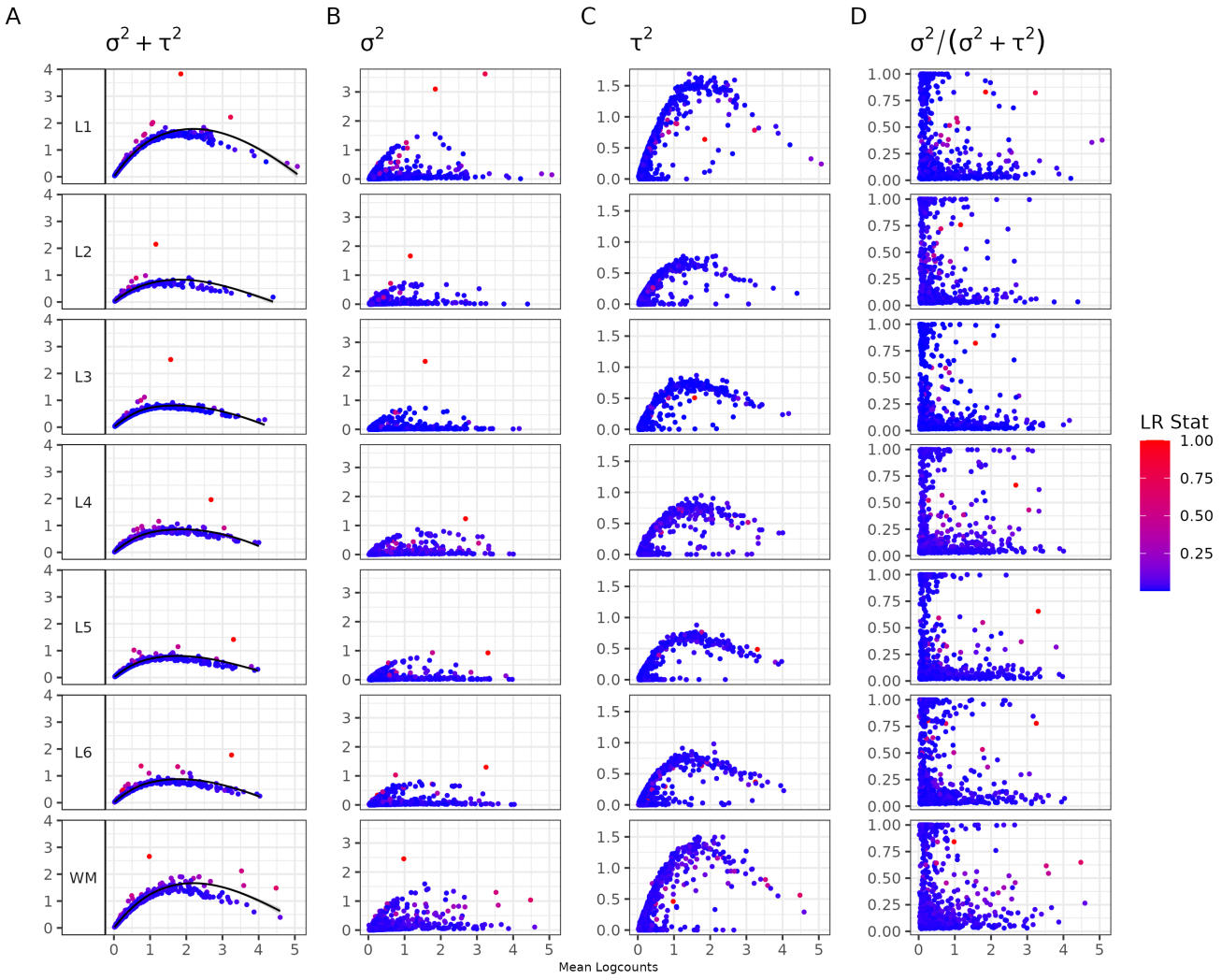

Figure S3: Mean-variance relationship after conditioning out biological variance measured by Gaussian process. Each row is a cortical layer from the DLPFC dataset, in order from top to bottom: Layers I-VI, white matter (WM). Each point is a gene colored by the likelihood ratio statistic (LR Stat) for a test comparing the fitted model against a classical linear model for the spatial component of variance. The likelihood ratio statistics are scaled by the maximum likelihood ratio statistic for each layer in order to have more uniform visualization. The x-axis represents mean logcounts and the y-axes represent different components of variance, in order from left to right: total variance $\sigma^2 + \tau^2$, spatial variance σ^2, nonspatial variance τ^2, and proportion of spatial variance $\sigma^2/(\sigma^2 + \tau^2)$.

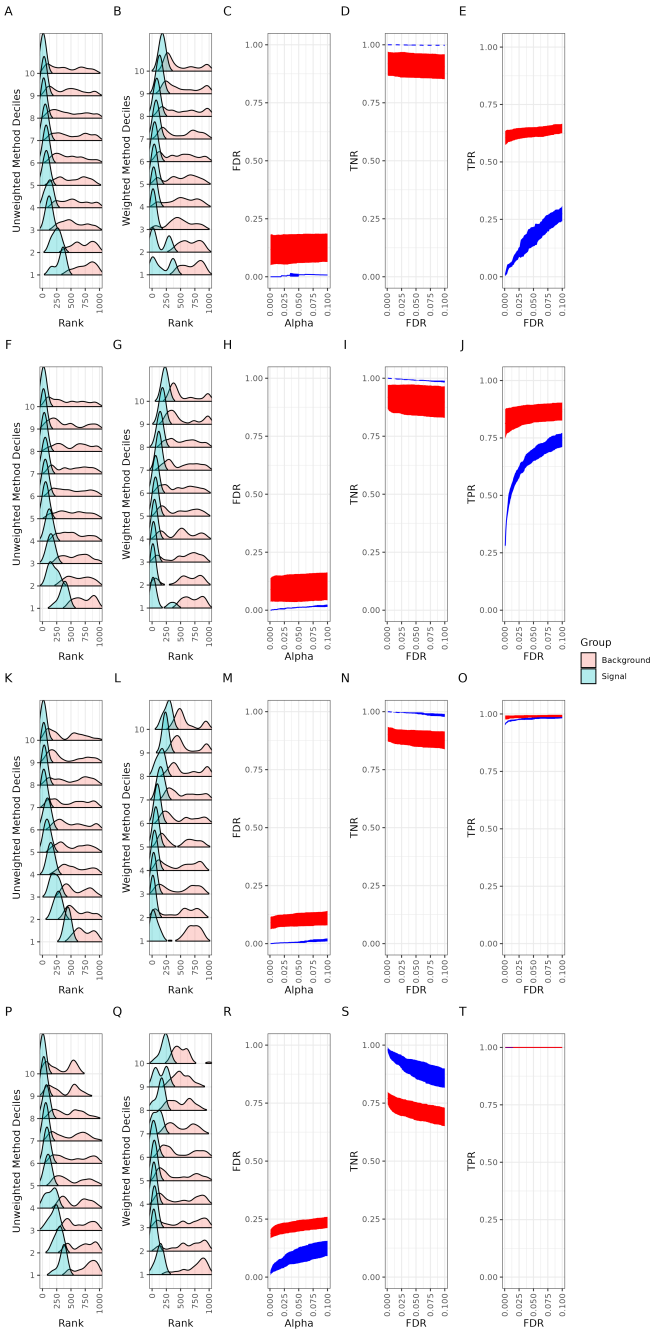

Figure S4: Removing the mean-variance relationship with expanded lengthscale metrics. This dataset contains 1,000 simulated genes across 968 spots. Each row represents a simulation setting with unique lengthscales, in order from top to bottom: 50, 60, 100, 500. Separately for unweighted and weighted methods, the genes were binned into deciles based on mean logcounts. Decile 1 is the lowest mean expression values. The first column of plots is unweighted ranks and the second column of plots is weighted ranks. Within each decile, the density of the top 10% ranks is plotted as the signal and the density of the remaining ranks is plotted as the background. The final three columns show the false discovery rate (FDR), true negative rate (TNR), and true positive rate (TPR). The red represents weighted nnSVG and the blue represents unweighted nnSVG. These plots represent the average of each respective rate over five iterations of the same simulation with unique random seeds.

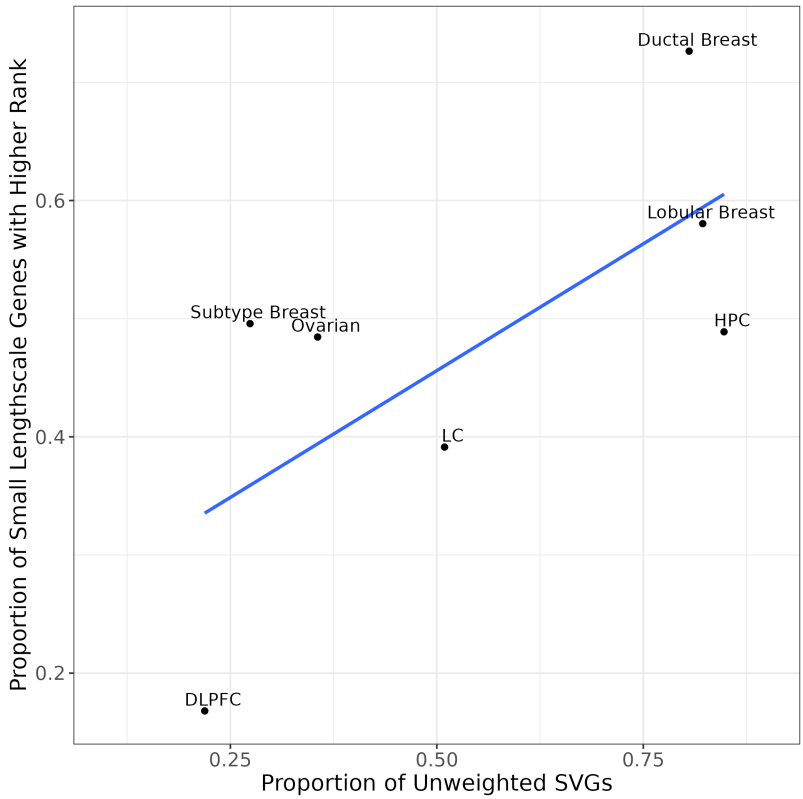

Figure S5: Ranking small lengthscale genes after weighting. Each point is a unique real dataset analyzed with 10x Genomics Visium. The x-axis is the proportion of SVGs from running unweighted `nnSVG` on the dataset. The y-axis is the proportion of genes with a small lengthscale (40-90) that are higher ranked in weighted `nnSVG` compared to unweighted `nnSVG`.

Table S1: Low Mean, Ductal Breast Genes Relation to Cancer. This table shows all genes with means less than the 25th percentile in the Ductal Breast [39] dataset which were in the lowest 10% of ranks before weighting and then increased to the highest 10% of ranks after weighting. The second column indicates if the gene is related to Breast cancer, with the corresponding reference in the third column.

Gene	Cancer-related?	References
TMEM39B		
ETAA1		
ATXN7	TRUE	[55]
BBS7		
MFSD8		
ETFDH		
E2F3	TRUE	[56]
FIG4		
TSPYL4		
METTL2B		
MFHAS1	TRUE	[57]
MAK16		
GKAP1		
SAAL1	TRUE	[58]
DDX10	TRUE	[59]
B3GLCT		
GNB5		
TLN2	TRUE	[60]
VPS33B		
ASB7		
MOSMO		
MED26		
ZNF227		
ITCH	TRUE	[61]
ZHX3	TRUE	[62]
ZBTB21		

Table S2: Low Mean, Subtype Breast Genes Relation to Cancer. This table shows all genes with means less than the 25th percentile in the Subtype Breast [41] dataset which were in the lowest 10% of ranks before weighting and then increased to the highest 10% of ranks after weighting. The second column indicates if the gene is related to Breast cancer, with the corresponding reference in the third column.

Gene	Cancer-related?	References
RELA-DT		

Table S3: Low Mean, Lobular Breast Genes Relation to Cancer. This table shows all genes with means less than the 25th percentile in the Lobular Breast [40] dataset which were in the lowest 10% of ranks before weighting and then increased to the highest 10% of ranks after weighting. The second column indicates if the gene is related to Breast cancer, with the corresponding reference in the third column.

Gene	Cancer-related?	References
TMEM51		
PDK1	TRUE	[63]
QTRT2		
OSBPL11		
TMEM44		
CPLX1		
PPAT	TRUE	[64]
ATP6AP1L	TRUE	[65]
LYRM7	TRUE	[66]
MSH5		
ZBTB24		
SHPRH		
TRIM35	TRUE	[67]
NCALD		
SNX30		
TRIM32	TRUE	[68]
FRAT1	TRUE	
RAB11FIP2		
TIGAR	TRUE	[69]
INTS13		
DLEU2		
ATXN1L		
SNHG30		
SEPTIN4		
ZNF439	TRUE	[70]
ZNF181		
TMEM191B		
ZNF74		
MOSPD2	TRUE	[71]
FAAH2		
BRWD3	TRUE	[72]

Table S4: Low Mean, Ovarian Genes Relation to Cancer. This table shows all genes with means less than the 25th percentile in the Ovarian [43] dataset which were in the lowest 10% of ranks before weighting and then increased to the highest 10% of ranks after weighting. The second column indicates if the gene is related to Ovarian cancer, with the corresponding reference in the third column.

Gene	Cancer-related?	References
TUFT1	TRUE	[50]
EHHADH		
DDX39B	TRUE	[51]
NAV2		
AC026471.4		
SMYD4		
HEXIM2		

Table S5: Small Lengthscale, Ductal Breast Genes Relation to Cancer. This table shows all genes with lengthscale values between 40 to 90 in the Ductal Breast [39] dataset which were ranked higher after weighting. The second column indicates if the gene is related to Breast cancer, with the corresponding reference in the third column.

Gene	Cancer-related?	References
S100PBP		
MED8		
DDX20	TRUE	[73]
GPR89A		
ASXL2	TRUE	[74]
PARTICL		
IMP4		
MGAT5	TRUE	[75]
ICOS	TRUE	[76]
AC112220.2		
PHF7		
ATXN7		
DNAJC13		
GAPT		
AC008608.2		
AC106795.2		
SAPCD1		
TNFRSF21	TRUE	[77]
CRYBG1		
MOXD1		
NUP43	TRUE	[78]
CDK13	TRUE	[79]
ZKSCAN1		
CLDN15		
NOM1		
NOL6	TRUE	[80]
ZCCHC7		
DAPK1	TRUE	[81]
MFSD14B		
CIZ1		
BORCS7		
TRIM21	TRUE	[82]
APIP		
FAM111B	TRUE	[83]
SPDYC	TRUE	[84]
AP5B1		
SPTBN2		
P2RY6		
CWC15		
CLEC4A		
AC087239.1		
ETFBKMT		
ZC3H10		
CRYL1		
IFT88		
MIS18BP1		
ZBTB1	TRUE	[85]
ERG28		
PEAK1	TRUE	[86]
RCCD1	TRUE	[87]
TEDC2		
NDE1		
FTO	TRUE	[88]
B3GNT9		
PARD6A		
SNAI3-AS1		
ALOX15		
ULK2	TRUE	[89]
TANC2		
TRIM65	TRUE	[90]
UBE2O		
ZNF532		
ZNF557		
MAP2K7		
ZNF490		
ZNF429		
C19orf47		
IRF2BP1		
AC010331.1		
DNMT3B	TRUE	[91]
PPP1R3D		
SEC14L2		
PMM1		
CRELD2		
ZMAT1		
NKRF		
MMGT1		

Table S6: Small Lengthscale, Subtype Breast Genes Relation to Cancer. This table shows all genes with lengthscale values between 40 to 90 in the Subtype Breast [41] dataset which were ranked higher after weighting. The second column indicates if the gene is related to Breast cancer, with the corresponding reference in the third column.

Gene	Cancer-related?	References
MTFR1L		
PRPF38A		
TBCE		
KIAA1841		
TSN		
CASP8	TRUE	[92]
AC022007.1		
HEMK1		
TRMT10C		
PLD1		
ATP11B	TRUE	[93]
YEATS2		
E2F3	TRUE	[56]
EHMT2	TRUE	[94]
PRRT1		
TREM1	TRUE	[95]
MSC		
SMC2	TRUE	[96]
CIZ1		
GTF3C4		
BRD3OS		
SEC31B		
CD82	TRUE	[97]
BIRC2	TRUE	[98]
UPK2		
HMBS		
GABARAPL1	TRUE	[99]
STAT2		
NUAK1		
RHOF		
SLC7A1		
EFCAB11		
ZNF770		
HOMER2		
POLR2C		
CDYL2	TRUE	[100]
MBTPS1		
CTU2		
DRG2		
BLMH		
RASL10B		
NBR2		
MSI2	TRUE	[101]
QRICH2		
ME2		
CTDP1	TRUE	[102]
DDA1	TRUE	[103]
SLC25A42		
AC006504.5		
STRN4		
TTYH1		
ZNF544		
JAM2	TRUE	[104]
AC245060.5		
NOL12		
GTPBP1		
TAF1	TRUE	[105]
FGF13	TRUE	[106]
LINC00893		

Table S7: Small Lengthscale, Lobular Breast Genes Relation to Cancer. This table shows all genes with lengthscale values between 40 to 90 in the Lobular Breast [40] dataset which were ranked higher after weighting. The second column indicates if the gene is related to Breast cancer, with the corresponding reference in the third column.

Gene	Cancer-related?	References
FBXO2		
IFFO2		
ZNF436	TRUE	[107]
EXTL1		
MTF2	TRUE	[108]
PHTF1		
HAGLR	TRUE	[109]
ORC2		
SATB1	TRUE	[110]
C3orf38		
CMSS1		
CD200	TRUE	[111]
TRIM59	TRUE	[112]
ETV5		
UCHL1	TRUE	[113]
INTS12		
LIFR	TRUE	[114]
ZBED3		
FER	TRUE	[115]
DCP2		
MIR3936HG		
SH3RF2		
SLC36A1		
BNIP1		
HIST1H4J		
NHSL1		
C7orf25		
HUS1		
ZSCAN21		
AASS		
AC004918.1		
FANCG		
NFIL3	TRUE	[116]
AKNA		
PROSER2		
THNSL1		
ENTPD1	TRUE	[117]
CWF19L1		
ARMH3		
ROM1		
AP000438.1		
FDXACB1		
CLEC2D	TRUE	[118]
RASSF3		
PRDM4		
FLT1	TRUE	[119]
BIVM		
ZNF219		
NYNRIN		
CGRRF1	TRUE	[120]
SGPP1		
FAM71D		
GABRB3		
CHST14		
TYRO3	TRUE	[121]
ADAM10	TRUE	[122]
GRAMD2A		
ASPHD1		
PLEKHG4		
PDPR		
MLYCD		
ZC3H18		
CENPBD1		
NLRP1	TRUE	[123]
PIGW		
CEP95		
ABCA5		
MEX3C		
ZNF77		
ZGLP1		
HAUS5	TRUE	[124]
ZNF574		
ZNF628		
ZNF579		
ELMO2		
DIP2A		
PLA2G3		
KCTD17		
APOBEC3C		
TCF20		
JADE3		
TRO		
MECP2	TRUE	[125]

Table S8: Small Lengthscale, Ovarian Genes Relation to Cancer. This table shows all genes with lengthscale values between 40 to 90 in the Ovarian [43] dataset which were ranked higher after weighting. The second column indicates if the gene is related to Ovarian cancer, with the corresponding reference in the third column.

Gene	Cancer-related?	References
HYI		
ECHDC2		
RAVER2		
WDR3		
PAQR6		
DESI2		
STRN		
ACYP2		
PAIP2B		
LIMD1		
PRICKLE2		
BBX		
LSAMP	TRUE	[126]
TMEM39A		
DTX3L		
B3GNT5		
IQCG		
LIN54		
PDE5A		
LARP1B		
HPF1		
DCP2		
WDR55		
CASC15	TRUE	[127]
SLC18B1		
ICA1		
PSPH		
RCC1L		
COG5		
WASL		
KLHDC10		
ZNF775		
INSIG1		
SLC25A6	TRUE	[128]
CXorf36		
HNRNPH2		
PRPS1		
SGK3		
AF117829.1		
ZC3H3		
NPR2		
SCAI		
ZBTB34		
CDK9	TRUE	[129]
LRRC56		
INTS5		
HIKESHI		
CASP1		
SIDT2		
MAP3K8	TRUE	[130]
ERCC6		
FAM149B1		
NOC3L		
LRRC27		
LRRC23		
RDH5		
ARHGEF25		
TXNRD1		
MVK		
B3GLCT		
GCH1	TRUE	[131]
C16orf87		
ZNF319		
SLC12A4		
AC008105.3		
PCYT2	TRUE	[132]
TRAPPC8		
SDCBP2		
BSG		
ZBTB7A		
ZNF443		
ASF1B		
ZNF135		
ATP6V1E1		
ZNRF3		
NOL12		
SREBF2		