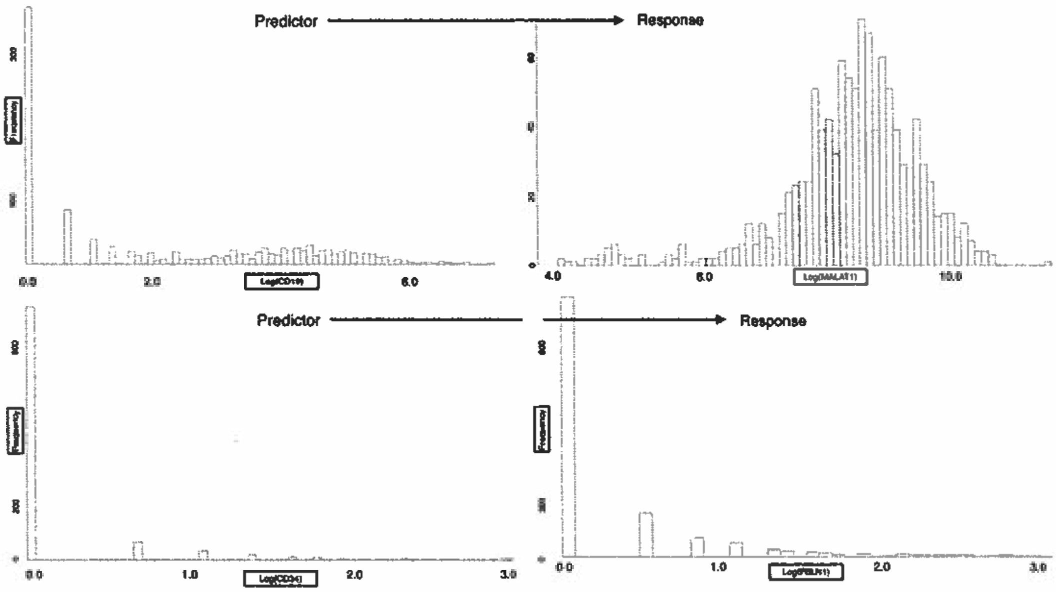

Figure 5:

We see that the log-transformed response MALAT1 is approximately normal distribution. Conversely, the log-transformed response FBLN1 is not inherently better than the untransformed response. We can clearly see the heavy influence of zero-inflation in these variables as is apparent from the dominance of the "zero-bins" in Figure (5).

Regardless, we model each outcome under the assumption that: compensating for observational correlation will sufficiently account for non-normality of the responses. This may not generally the case, and additional transformations or modeling methodologies may be needed to improve model error distributions. However, for the purpose of comparing the previously mentioned models on subject-correlated single-cell data, we will proceed with this assumption and verify ridual homoscedasticity, normality and independence using fitted vs residual plots and quantile-quantile plots.

Subtitle?

Model Descriptions

What is meant by transformed? and is this necessary? don't we just need a variable?

We define our outcome(s) of interest to be one ~~of the following transformed variables~~ as taken from Arazi, Rao, Berthier, et al. Let a single observation be designated as: R_{hij}. The

Lee, please describe the models in more general terms. No need to have the index h

index $h = \{1, 2\}$ represents the model pairing number ($CD19 \sim MALAT1$ is pairing #1 $CD34 \sim FBLN1$ is pairing #2) $i \in \{5, 6, \ldots, 26\}$ represents the subject (~~name of subject by~~ number) from which ~~the observation originated~~, and the index $j = 1, \ldots, n_i$ represents the single-cell observation within subject-i. We note that $n_i \in \{21, 22, 23, \ldots, 127\}$ in the context of the Lupus Data.

why not start at 1?

as the # of single-cell obs. in each subject

please be more general

We perform the transformations:

why?

$$R_h = \log(Y_h^* + 1)$$

where

$$Y_1^* = \text{MALAT1} - 67 \quad \text{and} \quad Y_2^* = \text{FBLN1}$$

and

$$Y_h = Y_h^* - min(Y_h^*)$$

We aslo define the predictor attached to R_h as:

?

$$P_h = \log(X_h + 1) \quad \text{for} \quad h = 1, 2$$

where

$$X_1 = \text{CD19} \quad \text{and} \quad Y_2 = \text{CD34}$$

?

We present the theoretical model frameworks here as "Less Than Full Rank" (LTFR) representations. The Full-Rank model results presented in the *Results* section to follow are created by droping the first level in all factors and using this as the referrence level. ?

Linear Regression

We begin the model framework definitions by describing two l Linear r Regression models, with f Fixed e Effect parameters estimated using maximum likelihood optimization. It should be noted that ~~these methods~~ linear regression make the assumption that observations are independent, ~~and should therefore be used for comparison to modeling methods to come. However, the~~ l linear regression models ~~we present here~~ can account for some ~~observational correlation~~ structure with the use of a subject specific intercept term as we will see in the second model.

Ultimately, all the methods defined in this section assume an identical error structure across all observations of the form:

$$\epsilon_{hij} \sim N\left(0, \sigma^2_\epsilon \cancel{*} \boxed{I_{1110}}\right) \quad ?$$

where we are assuming that σ^2 is ~~a~~ the common variance parameter for all subjects and I_{1110} is the 1110 X 1110 identity matrix. ?

Simple Linear Regression (Model 0)

Using the notation ~~we~~ defined above, we write the first model as:

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + \epsilon_{hij}$$

remove h from all models

which is equivalent to:

$$\log(Y_{hij}) = \beta_0 + \beta_1 \log(X_{hij}) + \epsilon_{hij}$$

We note that this model does not account for ~~observational correlation~~ any structure in the observations, and ~~instead~~ provides an estimation estimates for population-averaged relationships, ~~namely~~:

- ~~What is~~ T the estimated average (across all observations, across all subjects) value of R_{hij} when $P_{hij} = 0$ (intercept)
- On average (across all observations, across all subjects) ~~what is~~ the ~~average~~ rate of change in R_{hij} per unit increase in P_{hij} (slope)

Fixed-Effect Subject-Specific Intercept (Model 1)

Adding a subject-specific intercept term allows us to account for within-subject correlation by uniformly shifting the fitted values specific to a subject. This model ~~may be~~ is written as:

the mean of the fitted values!

$$R_{hij} = \beta_0 + \beta_{1i}(subject_i) + \beta_2 P_{hij} + \epsilon_{hij}$$

where we define the term:

~~How~~ what do you use to define the total # of subjects ?

$$\beta_{1i}\left(subject_i\right) = \begin{cases} \beta_{1i} & \text{if} \quad subject_i = i \\ 0 & \text{if} \quad subject_i \neq i \end{cases}$$

This model ~~provide the added~~ adds (N-1) estimated parameter $\hat{\beta}_{1i}$ which ~~tells us a uniform estimated~~ is the average deviation for each ~~subjects'~~ subject ~~fitted response~~ from the global estimated mean ~~provided by~~ n Model 0 (Simple Linear Regression).

Linear Mixed Effects Models

change all to lower case

l m e m

The next category of modeling approaches we describe is Linear Mixed Effect Models with Random Effects. Specifically, we describe two distinct Linear Mixed Effect Models that account for subject-correlation in a different manner than the previously discussed Linear Regression models. Linear Mixed Efffects Models do not neccessarily ~~assume~~ require the assumption of independence of observations. Correlation structures such as AR(1), spatial power, or unstructured can be

? write out

we don't need the responses to be normally distributed, just the random error, ε

used ~~to estimate parameters determining correlation amongst observations within a~~ subject and between observations ~~(across subjects)~~. Additionally, if we can ~~rationally~~ assume that the responses shown in Figure 3 have a multivariate normal distribution, the model parameters can be easily estimated using Maximum Likelihood Estimation techniques [9].

such as REML?

Linear Mixed Effects Model with Random Intercept (Model 2)

Model 1 (Linear Regression with Fixed Effect Intercept) accounts for subject correlation by assuming that observations within a subject are uniformly correlated ~~influenced by~~ the nested nature of the sampling ~~method~~ (i.e. observations are sampled so that they are identically correlate within each subject). However, this assumption may not always be reasonable, as we could imagine that responses within each subject also exhibit random variation that is related to nested sampling methods.

But the random var is random so doesn't need to be modeled

awkward & repetitive

A Linear Mixed Effects Model that includes a Random Intercept accounts for subject-level observational correlation by inducing individual-specific levels of random variation into all observations specific to each subject. Such a model may be written as:

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + b_{0i}\,(subject_i) + \epsilon_{hij}$$

where

$$b_{0i} \sim N\left(0, \sigma_b^2\right) \quad \text{for} \quad i \in \{5, 6, \ldots, 26\}$$

make more general

$$\epsilon_{hij} \sim N\left(0, \sigma_\epsilon^2 I_{n_i}\right)$$

does this need to be indexed by n_i?

and we assume that b_{0i} and ϵ_{hij} are independent.

~~We note that~~ B both random-components ~~can be~~ are assumed to have a mean of zero ~~as non-zero components are inherently deterministic and can be integrated into intercept terms.~~

Confusing → a random slope allows the slope estimate to differ for each subject

Linear Mixed Effect Model with Random Slope (Model 3)

A further accounting for the effects of subject-level observational clustering may be made by extrapolating on Model 2 (Linear Mixed Effects Model with Random Intercept) with the addition of a random intercept.

accounts additional structure for

The incoperation of a Fixed Effect subject-specific slope ~~would account~~ for subject-level observational correlation by assuming that the relationship between predictor and response are uniformly influenced ~~across observations~~. Implying that, in addition to the average response devation from the estimated average response, there is also an average ~~uniform~~ shift in how each subjects' response changes with respect to a unit shift in the predictor. Again, this assumption may not be reasonable, as we may expect variation in how responses within a subject deviate from the estimated ~~average change~~ in response over the predictor space.

effects

allowing

awkward. rewrite simply

We will therefore incorperated a Random Slope into the format of the Random Intercept model (Model 2) to attempt to reconcile these effects. This will allow for us to account for observational correlation due to subject-level sampling as sourced from: ??

- subject-specific random variation associated with measurement instability
- predictor-dependent, subject-specific random variation associated with measurement instability

We write this model as:

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + b_{0i}\,(subject_i) + \left[b_{1i}\,(subject_i)\,P_{hij}\right] + \epsilon_{hij}$$

where

$$\mathbf{b} = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\mathbf{0}, \mathbf{G}\right)$$

$$G = \begin{bmatrix} \sigma_b^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}$$

are these really the same?

$$\epsilon_{hij} \sim N\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}\right)$$

Generalized Estimating Equations (Model 4)

so do the others if we treat them as generalized

~~Our final method for modeling scRNA-seq expression profiles is Generalized Estimating Equations (GEE)~~. ~~Dissimilar to each of the methods previously described,~~ GEE regression esitimates are obtained using methodologies that allow for non-continuous responses. GEE also extrapolates on the techniques used for modleing non-normal responses by incorperating the effects of observational correlation.

I would remove this section as LME & OLS regression can be easily extended to generalized models

GEE estimates are computed by solving the estimating equation(s):

? ? General please

$$0 = U(\beta) = \sum_{i=1}^{15} \left\{ \mathbf{D}_{hi}^T \mathbf{V}_{hi}^{-1} \left(\mathbf{y}_{hi} - \mu_{hi}\right) \right\} \qquad (1)$$

define all terms please

where:

- instead, the main difference is marginal (GEE) vs conditional (LME) controlling of the groups.

$$\mu_{hi} = \mu_{hi}(\beta) = E\left[\mathbf{Y}_{hi}\right] = \eta_{hi}$$

represents the relationship between the expected value of the response μ_i (not necessarily assumed to be a distribution) and the linear predictor η_i,

★ DISCUSS HOW TO INTERPRET LME (conditional) & GEE (marginal)

$$\mathbf{D}_{hi} = \begin{bmatrix} \frac{\partial\mu_{hi1}}{\beta_1} & \frac{\partial\mu_{hi1}}{\beta_2} & \cdots & \frac{\partial\mu_{hi1}}{\beta_p} \\ \frac{\partial\mu_{hi2}}{\beta_1} & \frac{\partial\mu_{hi2}}{\beta_2} & \cdots & \frac{\partial\mu_{hi2}}{\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\mu_{hin_i}}{\beta_1} & \frac{\partial\mu_{hin_i}}{\beta_2} & \cdots & \frac{\partial\mu_{hin_i}}{\beta_p} \end{bmatrix}$$

– which term is the correlation structure?

is the first derivative matrix, and

$$\mathbf{V}_{hi} = \mathbf{A}_{hi}^{\frac{1}{2}} Corr(\mathbf{Y}_{\mathbf{hi}}) \mathbf{A}_{hi}^{\frac{1}{2}}$$

– define all terms

$$\mathbf{A}_{hi} = \underset{n_i}{diag}\left\{\phi_j\left(t_{ij}\right)\nu\left(\mu_{hij}\right)\right\}$$

We note that $\phi_j\left(t_{ij}\right)$ and $\nu\left(\mu_{hij}\right)$ are hyperparameters defined so that we may know the variance as a function of the mean and a scale parameter, i.e:

$$Var\left(Y_{hij}\right) = \phi_j\left(t_{ij}\right)\nu\left(\mu_{hij}\right)$$

The GEE algorithm is iterative and used the following steps to converge ~~at~~ to an estimate:

1. Generalized Linear Modeling methods employing Maximum Likelihood Estimation are used to obtain intial estimates for β
2. Estimates for β used to compute hyper-parameters
3. New estimates for hyper-parameters and working covariance matrix $(\mathbf{V}_{hi})$ used to obtain new estimates for β by solving (1)
4. Repeat Steps 2 & 3 until algorithm converges

~~The~~ GEE algorithm has a quality which ~~makes it very appealing for many applications~~ with observational clustering. ~~Specifically~~, T the GEE algorithm is robust to misspecification of the observational correlation structure. That is, the estimates $\hat{\beta}_{GEE}$ are consistent with β irrespective of the estimates for within-subject correlation.

It is not clear to me what you mean here which implies we ~~will be~~ assumi~~ng~~ e the general modeling structure:

$$E\left[Y_{hij}\right] = \mu_{hij} = \eta_{hij} = \beta_0 + \beta_1 P_{hij}$$

we will assume a variance function of the form:

$$Var\left(Y_{hij}\right) = \phi$$

→ is ϕ a matrix?

avoid ~~will~~ be -ing verbs

and we ~~will be~~ use using a working covariance matrix structure for repeated measures that correspond~~s~~ to the assumption of independence of observations within a subject.

$$\left[Corr\left(Y_{hij}, Y_{hik}\right)\right]_{jk} = \begin{cases} 1 & \text{if} \quad j = k \\ 0 & \text{if} \quad j \neq k \end{cases}$$

$$for \quad j, k \in \{1, \ldots, n_i\}$$

Results

Table 8 and table 9 display parameter value estimates, standard errors, test statistics, and p-values for the main-effect slope term estimated by all five modeling approaches:

The GEE algorithm is also very stable, in-part due to the fact that the effect(s) ~~that it~~ estimates are population-averaged. Each of the previous methods (Model 0 withstanding) ~~had~~ have subject-specific interpretations, but the GEE algorithm provides marginal parameter estimates. These values do not represent any specific subject, but rather the population-average. averaging over all subjects

According to Fitzmaurice, Laird, and Ware [9] ~~we also need to ensure that any~~ responses modeled in the GEE process ~~are~~ need to be stationary, i.e:

$$E\left[Y_{hij}|\mathbf{X}_{hi}\right] = E\left[Y_{hij}|X_{hi1}, \ldots, X_{hin_i}\right] = E\left[Y_{hij}|X_{hij}\right]$$

The scRNA-seq data ~~has been~~ is assumed to be independent within-subject, therefore we have:

$$E\left[Y_{hij}|X_{hij}\right] = E\left[Y_{hij}|X_{hij'}\right]$$

$$\forall j \in \{1, \ldots, n_i\} \quad j \neq j'$$

as needed.

The three-part specification of the GEE framework includes:

1. The link function and linear predictor
2. Variance function
3. A working covariance matrix

The link function and linear predictor are chosen so that the resulting model estimates ~~will be~~ are comparable to preceeding estimates for intercept and slope. Therefore, we will use the identity link function:

$$g(x) = x$$

in conjunction with the linear predictor:

$$g(\mu_{hij}) = \eta_{hij} = \beta_0 + \beta_1 P_{hij}$$