

Comparing Models of Subject-Clustered Single-Cell Data

v1

Lee Panter

Abstract

Single-cell RNA sequencing (scRNA-seq) represents a revolutionary shift to the analytic approaches being used to decode the human transcriptome. Single-cell data has been used to: visualize cellular subpopulations with unsupervised clustering methods, test for differential expression rates across conditions using logistic and mixture modeling, and reconstruct spatio-temporal relationships using network analysis. While these successes demonstrate the utility and promise for single-cell methods, they do not demonstrate the practical need to generalize to single-cell data over multiple individuals. This paper looks to compare three different modeling strategies for RNA-seq expression estimation in data with individual-level clustering. The modeling approaches will be compared theoretically against Linear Regression Models, and analytically, motivated by data from a Lupus Nephritis study. It is hoped that this paper will present new approaches to modeling single-cell expression data, and will be useful not only for Statisticians, but also Geneticists and Microbiologists.

Introduction

18

Single-cell analysis has emerged as a leading methodology for transcriptome analytics. [1] 19
Single-cell data sets (i.e. data involving measurements with single-cell resolution) have demon- 20
strated their utility in research contexts for identifying rare subpopulations, characterizing 21
genes that are differentially expressed across conditions, and inferring spatio-temporal relation- 22
ships within the microbiome. [2] Additionally, advances in whole genome amplification and 23
cellular isolation techniques have made single-cell data sets more accessible, more informative, 24
and more diverse than ever before. [1] 25

Traditional methods for subpopulation exploration within single-cell data commonly involve 26
unsupervised clustering techniques including Principle Components Analysis (PCA) and 27
K-Nearest Neighbors (KNN). These methods have been shown to be effective in identifying 28
rare nerological cells within a homogeneous population. [3] Such clustering methods, and 29
additional (non-linear) methods such as the t-distributed stochastic neighborhood embedding 30
(t-SNE) are also useful for visualizing high-dimensional data and have been used to find 31
multi-dimensional boundary values for distinguishing heathly and cancerous bone marrow 32
samples. [4] While all these studies involve single-cell data that incorporates multiple subjects, 33
the modeling methodologies do not provide estimates for subject-factor effects. 34

Single-cell data has been used to target treatments by characterizing differential expression 35
across condition. Model-based Analysis of Single-cell Transcriptomics (MAST) has been used 36
to compare “primary human non-stimulated” and “cytokine-activated” mucosal-associated 37
invariant T-cells. [5] Additionally, Single-Cell Differential Expression (SCDE) was used to 38
compare 92 mouse embryonic fibroblasts to 92 embryonic stem cells. [6] Neither of these 39
studies included samples across multiple subjects, and the resulting models do not account 40
for possible correlation within subjects that might be present. 41

Network modeling approaches, in conjunction with single-cell data have provided the oppor- 42

tunity to learn about cellular heirarchies, spatial relationships, and temporal progressions 43
within the microbiome. Weighted Gene Co-Expression Network Analysis (WGCNA) has been 44
used to find delineations in both human and mouse embryonic transcriptome dynamics during 45
progression from oocyte to morula. [7] A similar analysis was performed using Single-cell 46
Clustering Using Bifurcation Analysis (SCUBA), and was verified using Reverse Transcription 47
Polymerase Chain Reaction (RT-PCR) data over the same single-cell measurements. [8] The 48
studies conducted using network modeling approaches target single-cell sources at multiple 49
time points, or distinct measures that could be compared using a pseudo-time mapping 50
Diversification of the single-cell data by incorporating multiple subjects is not considered or 51
addressed. 52

Down-stream analyses of single-cell data can be a very useful tool for transcriptome analytics. 53
Technological advances in cellular isolation and genetic material amplification will likely 54
lead to a rise in single-cell data prevalence, and a corresponding rise in the prevalence of 55
multiple-subject single-cell data sets. Therefore, there is a clear need to develope, test, and 56
integrate alternative methods that can accurately and precisely model single-cell data and 57
account for the correlation of repeated measures within subject samples. 58

This paper seeks to satisfy this need by suggesting three methods for modeling scRNA-seq 59
expression profiles that account for within-subject correlation differently. We provide a 60
motivating example consisting of scRNA-seq observations across multiple subjects with Lupus 61
Neprhitis. Modeling theory and comparisons will be provided in the context of this example 62
and the results of the various modeling approaches will be compared. We will discuss relevant 63
conclusions, implications, limitations and future research to illustrate our findings. 64

Description of Motivating Example

65

Throughout the course of this paper, references will be made to “The immune cell landscape
66
in kidneys with lupus nephritis patients” [9]. This paper references single-cell data collected
67
as part of a cross-sectional, case-control study of 24 Lupus Nephritis (LN) cases and ten
68
control (LD) subjects. Samples of kidney tissue and urine from LN subjects were taken from
69
ten clinical sites across the United States, LD subject samples were obtained at a single site
70
from a living kidney donor, after removal and prior to implantation in the recipient. No
71
LD urine samples were collected. Samples were cryogenically frozen and shipped to a central
72
processing facility where they were thawed, dissociated, and sorted into single-cell suspension
73
across 384-well plates using FlowJo 10.0.7, 11-color flow cytometry [10]. sc-RNA sequencing
74
was performed using a modified CEL-Seq2 method [11], followed by \sim 1 million paired-end
75
reads per cell. Data can be accessed through the ImmPort repository with accession code
76
SDY997.
77

The Seurat Guided Clustering Tutorial [12] was used to examine initial data and perform
78
quality control (QC) filtering. The Seurat package allows for easy classification of low-quality
79
observations by setting threshold values for:
80

1. the number of unique genes detected in each cell (*nFeature*), and
81
2. the percentage of reads that map to the mitochondrial genome (*perctMT*)
82

Item (1) can be useful for identifying empty or broken-cell measurements (indicated by
83
abnormally low gene detection numbers) as well as duplicate/multiplicate cells measures
84
(indicated by abnormally high gene detection numbers). Item (2) can help to identify dead
85
and/or broken cells since dead or dying cells will retain RNAs in mitochondria, but lose
86
cytoplasmic RNA [2]. Original QC measures employed by (Arazi A, Rao DA, Berthier CC,
87
et al.) and implemented using the Seurat package required:
88

1. $1,000 < nFeature < 5,000$
89

2. $perctMT \leq 25\%$

90

However, after inspecting the distribution of the $perctMT$ variable across subjects (Figure 1) 91
 (Figure 1)

92

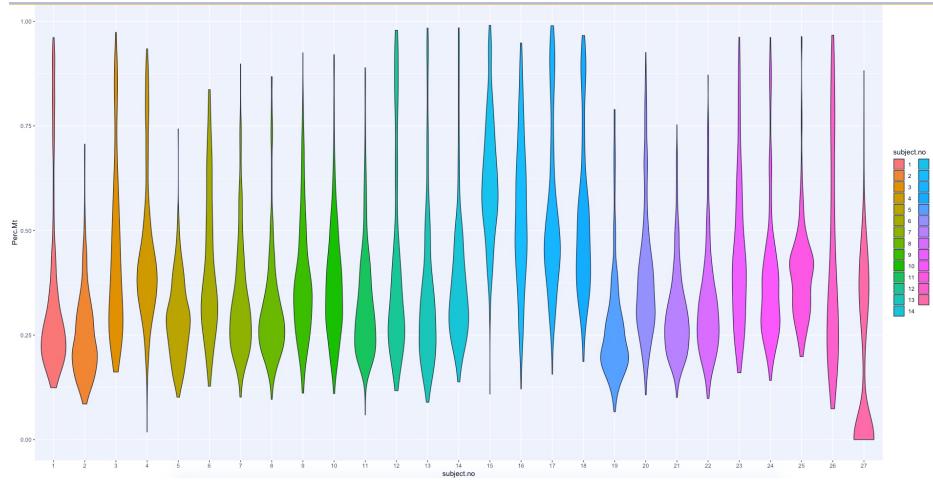


Figure 1:

and then observing the same distribution after the imposition of the QC measure $PerctMT \leq 25\%$ as employed by (Arazi A, Rao DA, Berthier CC, et al.) (Figure 2) 93
 94

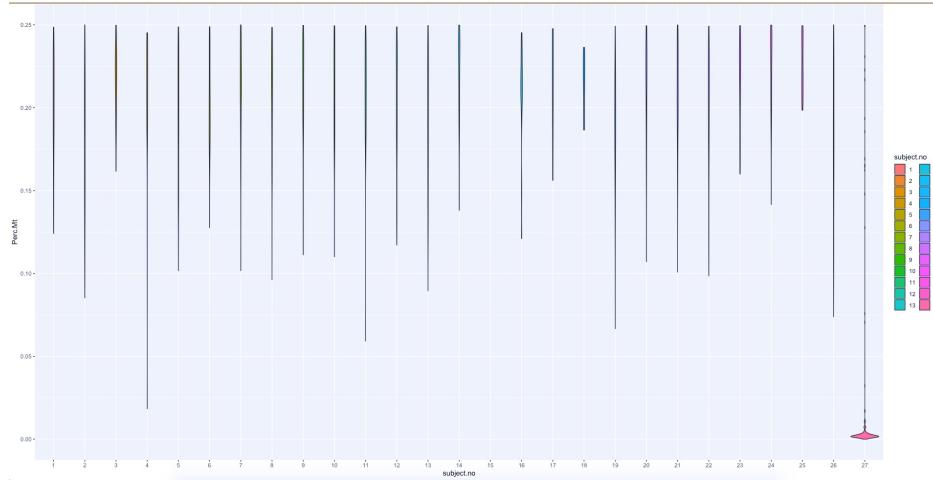


Figure 2:

a decision to increase the $perctMT$ threshold to 60% was made to preserve the inherent 95
 distribution structure across subjects (Figure 3) 96

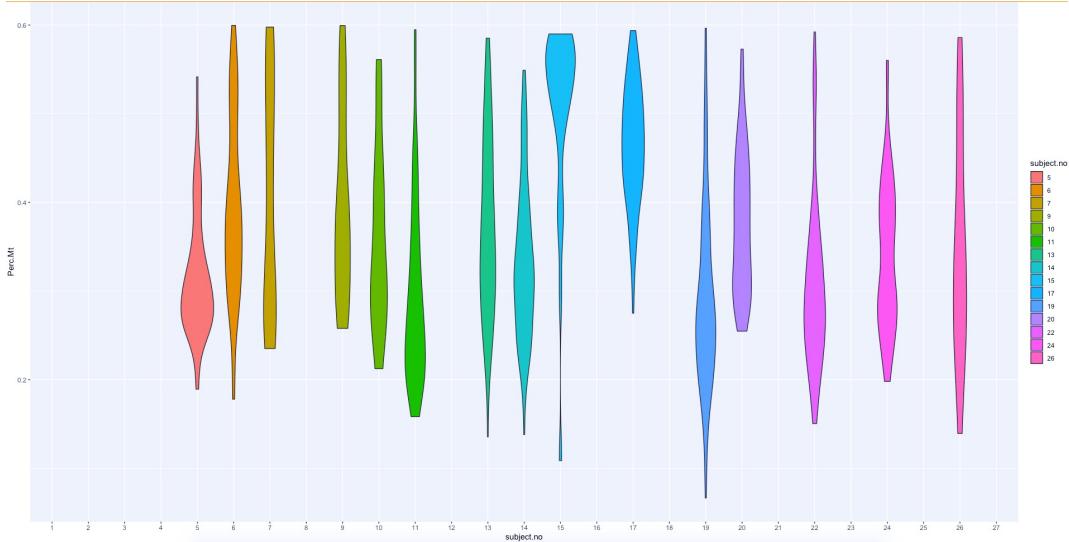


Figure 3:

After application of quality control filters, we are left with 1,110 scRNA-seq observations 97 of 38,354 genetic variables distributed across 15 subjects (originally 27 subjects in data, 12 98 removed due to poor quality data). There are a minimum of 21 and a maximum of 127 99 observations per subject, with a mean of 74 and a median of 79 observations per subject. 100

In order to simplify the analysis and make more significant insights into model comparisons, 101 we have chosen two pairs of variables from the 38,354 genetic markers to model in a predictor- 102 response relationship. The variables chosen indicated higher values of correlation than 103 arbitrary pairings, and could be associate with important outcomes of interest (e.g. cancer 104 treatment research in the case of MALAT1 [13], or observed limb malformations in the 105 case of FBLN1 variation [14]). An attempt was also made to choose predictor-pairings of 106 interest. The CD19 marker (paired with MALAT1) is a transmembrane protein, encoded 107 by the CD19 gene. Since the FlowJo cytometry measurement contain CD19 readings, the 108 relationship between a proteomic, a transcriptomic predictor, and an outcome of interest 109 could be modeled simultaneously. CD34, the predictor which we will link with FBLN1 is 110 also a transmembrane protein encoded by a gene, and similarly interesting. 111

Single-cell RNA sequencing data is represented as non-negative integer count data. The 112

observations are specific to much finer genomes (i.e. single-cell resolution), so while the overall scope of gene expression is the same as a traditional bulk experiment, individual observations have a biologically inflated zero-component. Additionally, there are technical zero-inflation components that are associated with protocol variations. Together, these factors contribute to heavily right-skewed variables (Figure 4)

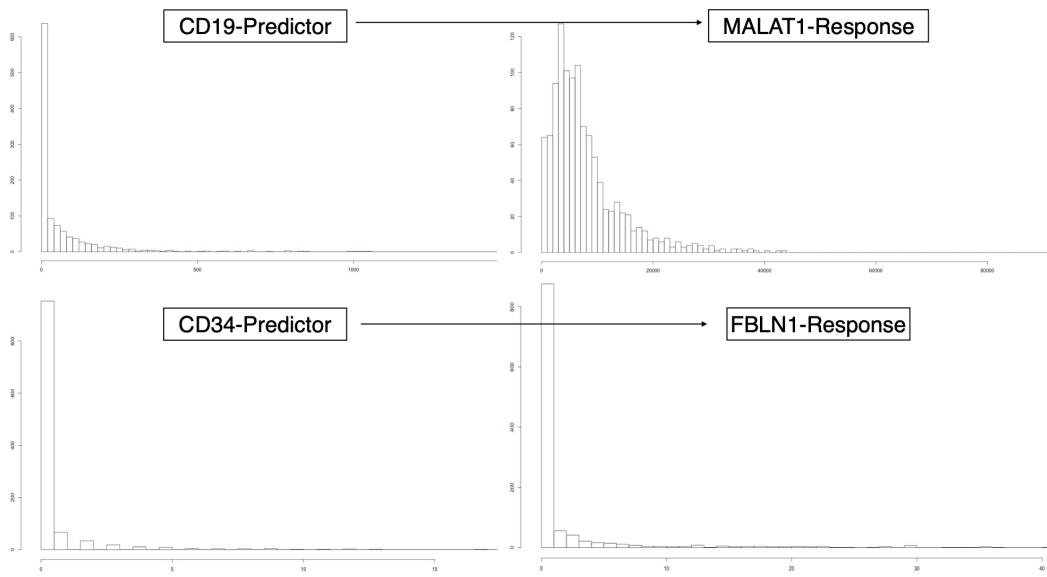


Figure 4:

We therefore perform a log-transform on the predictors and response variables, in an attempt to achieve approximate normality. Figure 4 displays the log transformations given by:

$$X \mapsto \log(X + 1)$$

We can see that the log-transformed response for the MALAT1 ~ CD19 pairing has resulted in an acceptably normal distribution for the response. Since the methods we will employ in this paper do not make the assumption of error normality it should suffice to have approximately normal response variables (i.e. predictor variable distributions may be reasonably non-normal provided that appropriate residual analysis is performed).

Conversely, the log-transformed response for the FBLN1 ~ CD34 pairing is not inherently

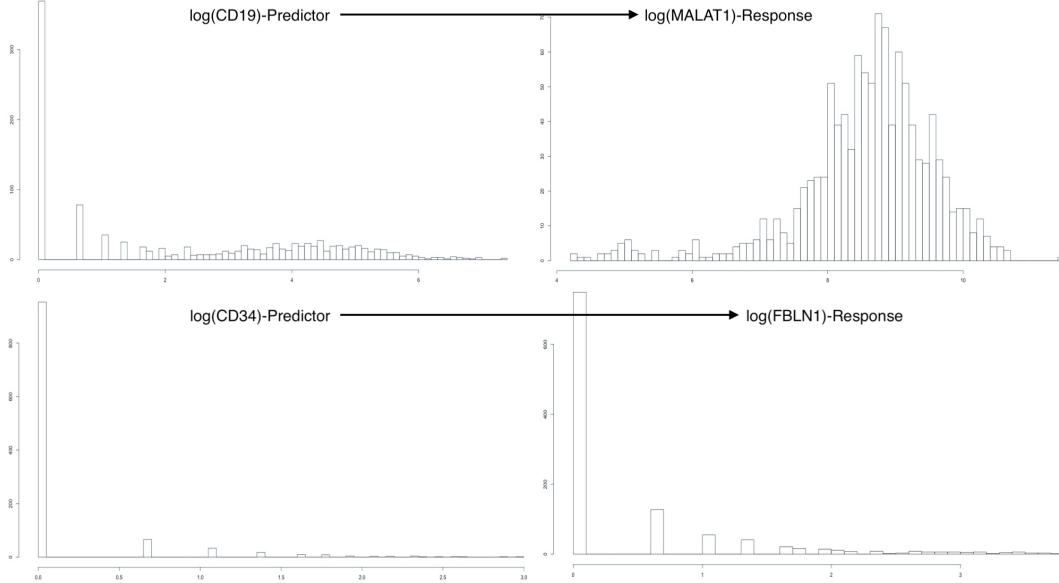


Figure 5:

better than the un-transformed response. We can clearly see the heavy influence of zero-
 inflation in these variables. Even though this is an alarming result, we will continue to model
 the FBLN ~ CD34 relationship, with the hope that the resulting model residuals will display
 conditionally zero expectation.

Model Parameters

The data we acquired from (Arazi A, Rao DA, Berthier CC, et al) allows us to fully define
 variables and parameters and outline each model clearly.

We define our outcome(s) of interest to be one of the following transformed variables:

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

where

$$Y_1 = \text{MALAT1} \quad \text{and} \quad Y_2 = \text{FBLN1}$$

We aslo define the predictor attached to R_k

135

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

where

136

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

Let a single response be designated as: R_{hij} . The index $i = 1, \dots, 15$ represents the subject 137 from which the observation originated, and the index $j = 1, \dots, n_i$ represents the repeated 138 observation number within subject-i. We note that $n_i \in \{21, 22, 23, \dots, 127\}$ in the context 139 of the Lupus Nephritis Data. We present the theoretical model frameworks here as “Less 140 Than Full Rank” (LTFR) representations. The Full-Rank model results presented create 141 full-rank model and design matrices by droping the first level in all factors, and using this as 142 the referrence level. 143

Linear Regression

144

We begin the process of comparing models for scRNA-seq expression profiles in Lupus 145 Nephritis subject-clustered tissue data, by describing three linear regression models, with 146 parameters estimated using Least Squares optimizations. 147

It should be noted that these methods make the assumption that observations are independent, 148 but account for some observational correlation with the use of subject specific intercept and 149 slope terms. 150

Ultimately, all these methods assume an identical error structure across all observations of 151 the form: 152

$$\epsilon_{hij} \sim N(0, \sigma_\epsilon^2 * I_{1110})$$

where we are assuming that σ^2 is a fixed variance parameter for all subjects and I_{1110} is the 153
 1110 X 1110 identity matrix. 154

Simple Linear Regression (Model 0) 155

Using the notation we defined above, we can write the first model as: 156

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

which is equivalent to: 157

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

Fixed-Effect Subject-Intercept (Model 1) 158

Adding a subject-specific intercept term, allows us to account for within-subject correlation 159
 by uniformly shifting the fitted values specific to a subject. This model may be written as: 160

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

where the term: 161

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

Fixed-Effect Subject-Slope (Model 2) 162

We may further account for within-subject correlation by adding a term which will ensure 163
 that individual subjects' relationships with the covariate of interest is accounted for. This will 164

help to reduce within-subject variation across the predictor space, and will be more noticeable ¹⁶⁵
for stronger, subject-specific interactions with covariates. ¹⁶⁶

This model may be written as: ¹⁶⁷

$$R_{hij} = \beta_0 + \beta_{1i} (\text{subject}_i) + [\beta_{2i} (\text{subject}_i) * P_{hij}] + \beta_3 P_{hij} + \epsilon_{hij}$$

where we are using the same definitions of (subject_i), R_{hij} , and P_{hij} as in Models 0 and 1. ¹⁶⁸

Motivated Results-Linear Regression ¹⁶⁹

The Linear Regression models described in the previous section(s) were each fit in R to the ¹⁷⁰
data from the motivating example (Arazi A, Rao DA, Berthier CC, et al) [9]. Estimates ¹⁷¹
for the main intercept (β_0), and the main-effect slope (β_1 in Model 0, β_2 in Model 1, β_3 in ¹⁷²
Model 2) along with the estimated standard errors, test-statistics, and p-values are displayed ¹⁷³
in (table 1) and (table 2) for the MALAT1 ~ CD19 relationship and in (table 3) and (table ¹⁷⁴
4) for the FBLN1 ~ CD34 relationship. ¹⁷⁵

Intercept Estimates (MALAT1 ~ CD19)

176

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 0	8.4618	4.568e-2	1.8526e+2	< 2e-16
Model 1	7.5486	1.2261e-1	6.1564e+1	< 2e-16
Model 2	6.9793	1.3896e-1	5.0226e+1	< 2e-16

Table 1

178

Main-Effect Slope (MALAT1 ~ CD19)

179

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 0	4.918e-2	1.455e-2	3.381	7.47e-4
Model 1	4.833e-2	1.381e-2	3.500	4.84e-4
Model 2	5.143e-1	6.017e-2	8.546	< 2e-16

Table 2

180

Intercept Estimates (FBLN1 ~ CD34)

182

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 0	3.510e-1	2.45e-2	1.43e+1	< 2e-16
Model 1	2.7572	6.52e-2	4.23e+1	< 2e-16
Model 2	2.7973	7.68e-2	3.643e+1	< 2e-16

Table 3

183

184

Main-effect Slope Estimates (FBLN1 ~ CD34)

185

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 0	7.884e-1	4.92e-2	1.6e+1	< 2e-16
Model 1	1.306e-1	3.42e-2	3.82	1.4e-4
Model 2	8.38e-2	5.89e-2	1.42	1.5492e-1

Table 4

186

We note that in each of the relationships being modeled, we see a decrease in the standard error associated with the main-effect slope as we incorporate subject-specific information regarding the intercept (i.e. when we compare model 0 to model 1). Conversely, we see an increase in the standard error associated with the main-effect slope as we incorporate subject-specific information about the slope.

188

189

190

191

192

Model diagnostics including plots of model fit, residual vs fitted value plots, and quantile-quantile residual distribution plots are included as part of the appendix material. *A discussion of modeling assumptions will also be included (wondering if this needs to be addressed)*

193

194

195

Linear Mixed Models

196

The next category of approaches to modeling scRNA-seq expression profiles in Lupus Nephritis, subject-clustered, data we will describe is two distinct Linear Mixed Models. These modeling methods account for subject-clustering differently than the previously discussed Linear Regression models. Linear Mixed Models do not necessarily assume observational independence, and they can even systematically account for correlation among repeated measures within a subject. Additionally, if we can rationally assume that the responses shown in Figure 3 have a multivariate distribution, the model parameters can be easily estimated using Maximum Likelihood Estimation techniques [15].

197

198

199

200

201

202

203

204

Linear Mixed Model with Random Intercept (Model 3)

205

Model 1 accounted for subject-clustering by assuming that observations within a subject were 206 uniformly influenced by the nested nature (observations within subjects) of the sampling 207 method. However, this assumption may not always be reasonable, as we could imagine that 208 responses within each subject exhibit random variation that is also related to nested sampling 209 methods. 210

A Linear Mixed Effects Model that includes a Random Intercept accounts for observational 211 correlation due to subject-clustering by assuming that observations within a subject are a 212 consequence of the nested nature of the sampling method, and therefore a consequence of 213 an additive (covariate-independent), subject-specific, effect; AND due to subject-specific 214 random variation in response measurement associated with measurement instability for 215 THAT subject. 216

This model may be written as: 217

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + b_{0i} (\text{subject}_i) + \epsilon_{hij}$$

where 218

$$b_{0i} \sim N(0, \sigma_b^2)$$

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

and we assume that b_{0i} and ϵ_{hij} are independent. 220

We note that both random-components can be assumed to have a mean of zero as non-zero 221 components are inherently deterministic and can be integrated into intercept terms. 222

Linear Mixed Effect Model with Random Slope (Model 4)

223

Model 2 implemented a Fixed Effect slope in an attempt to reconcile the effects of observational
224 clustering inadequately accounted for by the subject-specific Fixed Effect Intercept in Model 1.
225 However, in light of the information surrounding the development of Model 3, it is incumbent
226 for us to develop an analogous correction for Model 2. Such a correction would allow us to
227 account for observational correlation due to subeject-clustering as sourced from:
228

- additive, effects due to subject-clustered nested sampling methods
229
- subject-specific random variation associated with measurement instability
230
- covariate-dependent, subject-specific effects
231
- covariate-dependent, subject-specific random variation associated with measurement
232
instability
233

We write this model as:
234

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + b_{0i} (\text{subject}_i) + [b_{1i} (\text{subject}_i) P_{hij}] + \epsilon_{hij}$$

where
235

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

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

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

Motivated Results-Linear Mixed Models

236

The tables displayed (5 - 8) are analogous to Tables (1 - 4) for Model 3 and Model 4.

237

Intercept Estimates (MALAT1 ~ CD19)

238

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 3	8.4137	1.1825e-1	7.1151e+1	< 2e-16
Model 4	8.3972	1.3957e-1	6.0166e+1	<2e-16

Table 5

240

Main-Effect Slope (MALAT1 ~ CD19)

241

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 3	4.920e-2	1.374e-2	3.579	3.6e-4
Model 4	5.938e-2	3.538e-2	1.678	1.19e-1

Table 6

243

Intercept Estimates (FBLN1 ~ CD34)

244

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 3	6.53e-1	2.22e-1	2.94	1.1e-2
Model 4	6.491e-1	2.223e-1	2.92	1.1e-2

Table 7

246

Main-effect Slope Estimates (FBLN1 ~ CD34)

247

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 3	1.35e-1	3.42e-2	3.95	8.4e-5
Model 4	1.705e-1	7.29e-2	2.34	6.7e-2

Table 8

248

Again, we note that in each of the relationships being modeled, we see an increase in 250
the standard error associated with the main-effect slope as we incorporate subject-specific 251
information about the slope. 252

Model diagnostics including plots of model fit, residual vs fitted value plots, and quantile - 253
quantile residual distribution plots are included as part of the appendix material. *A discussion* 254
*of modeling assumptions will also be included (**wondering if this needs to be addressed**)* 255

Generalized Estimating Equations (Model 5)

256

Our final method for modeling scRNA-seq expression profiles is Generalized Estimating 257
Equations (GEE). Dissimilar to each of the methods previously described, GEE regression 258
coefficient esitimates are obtained using methodologies that allow for non-continuous responses. 259
GEE extrapolates on the techniques used for estimating Generalized Linear Models by 260
incorporating the effects of observational correlation and clustering. 261

The choice of GEE to model scRNA-seq expression profiles in subject-clustered data is a 262
natural continuation to the progression of methods chosen for comparison in this paper 263
becuase it represents a third, distinct approach to accomodating observational clustering 264
across subjects. 265

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

266

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

where:

267

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

represents the relationship between the expected value of the response μ_i (not necessarily 268 assumed to be related to a distribution) and the linear predictor η_i , 269

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

is the first derivative matrix, and

270

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

271

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

We note that $\phi_j(t_{ij})$ and $\nu(\mu_{hij})$ are hyperparameters defined so that we may know the 272 variance as a function of the mean and a scale parameter, i.e: 273

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

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

274

1. Generalized Linear Modeling methods employing Maximum Likelihood Estimation are 275 used to obtain intial estimates for β

276

- | | |
|---|------------|
| 2. Estimates for β used to compute hyper-parameters | 277 |
| 3. New estimates for hyper-parameters and working covariance matrix (\mathbf{V}_{hi}) used to obtain new estimates for β by solving (1) | 278
279 |
| 4. Repeat Steps 2 & 3 until algorithm converges | 280 |

The GEE algorithm has a quality which makes it very appealing for many applications with observational clustering. Specifically, the 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.

The stability of the GEE algorithm is in-part due to the effects that it estimates. Whereas each of the previous methods (Model 0 notwithstanding) had subject-specific interpretations, the GEE algorithm provides marginal parameter estimates. These values do not represent any specific subject, but rather the population-average.

According to (Fitzmaurice GM, Laird NM, Ware JH (2012)) [15] we also need to ensure that any responses modeled in the GEE process are stationary, i.e:

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

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

$$\begin{aligned} E[Y_{hij}|X_{hij}] &= E[Y_{hij}|X_{hij'}] \\ \forall j \in \{1, \dots, n_i\} \quad j &\neq j' \end{aligned}$$

Since the use of the scRNA-seq data would not violate the GEE assumptions, we proceed with the description of the model that we will fit.

The three-part specification includes:

1. The link function and linear predictor	296
2. Variance function	297
3. A working covariance matrix	298

The link function and linear predictor are chosen so that the resulting model estimates will 299
be comparable to preceding estimates for intercept and slope. Therefore, we will use the 300
identity link function in conjunction with the linear predictor: 301

$$g(x) = x$$

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

which implies we will be assuming the general modeling structure: 303

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

In the absence of further information, we will assume a variance function of the form: 304

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

and we will be using a working covariance matrix structure for repeated measures that 305
corresponds to the assumption of independence of observations within a subject. 306

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

$$\text{for } j, k \in \{1, \dots, n_i\}$$

Motivated Results - Generalized Estimating Equations

308

Intercept Estimates (MALAT1 ~ CD19)

309

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 5	8.4618	1.842e-1	2.1098e+3**	<2e-16**

Table 9

310

Note: ** These are Wald test of a single parameter (not t-tests)

311

Main-Effect Slope (MALAT1 ~ CD19)

313

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 5	4.92e-2	3.97e-1	1.53**	2.2e-1**

Table 10

314

315

Note: ** These are Wald test of a single parameter (not t-tests)

316

Intercept Estimates (FBLN1 ~ CD34)

317

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 5	3.51e-1	1.25e-1	7.85**	5.09e-3**

Table 11

318

319

Note: ** These are Wald test of a single parameter (not t-tests)

320

Main-effect Slope Estimates (FBLN1 ~ CD34)

321

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 5	7.88e-1	2.2e-1	1.281e+1**	3.4e-4**

Table 12

322

323

Note: ** These are Wald test of a single parameter (not t-tests)

324

Appendix

325

Linear Regression Model Diagnostic Plots

326

This section will contain plots of:

327

- Model vs Original data 328
- Fitted vs Original data 329
- Quantile - Quantile distributions of the model residuals. 330

Model 0

331

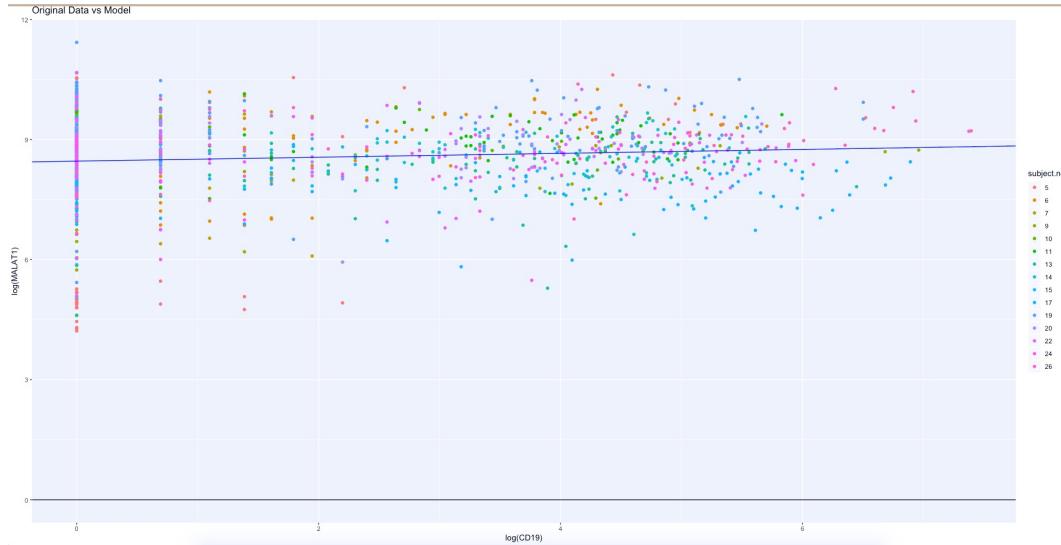


Figure 6: MALAT1 – CD19: Model v Original Data

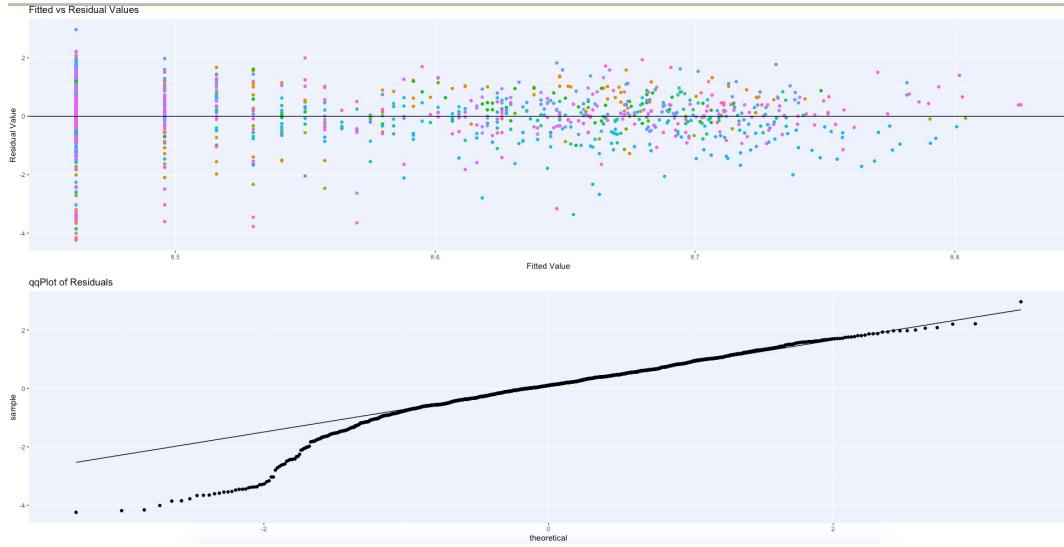


Figure 7: MALAT1 – CD19: Diagnostic Plots

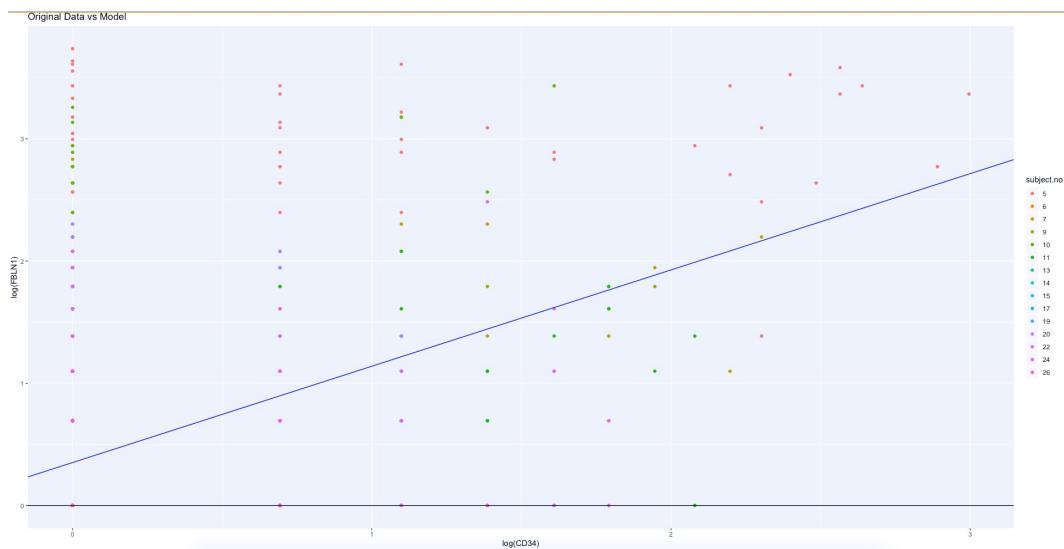


Figure 8: FBLN1 – CD34: Model v Original Data

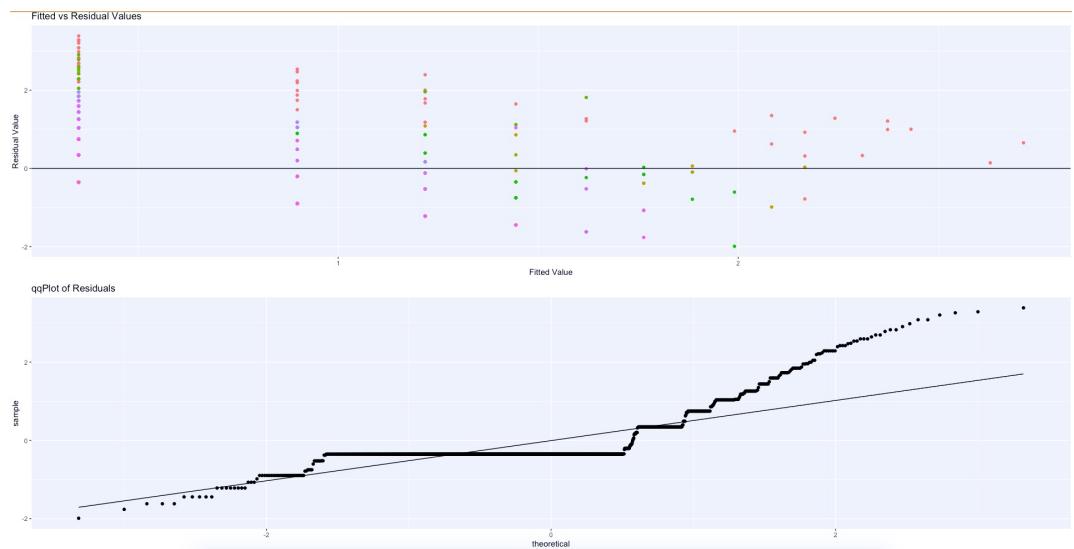


Figure 9: FBLN1 – CD34: Diagnostic Plots

Model 1

332

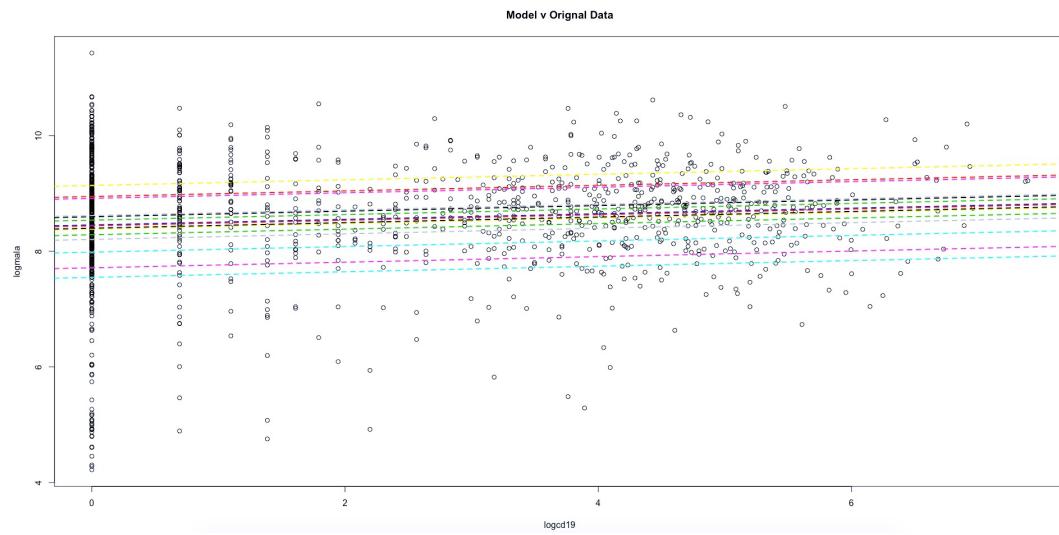


Figure 10: MALAT1 – CD19: Model v Original Data

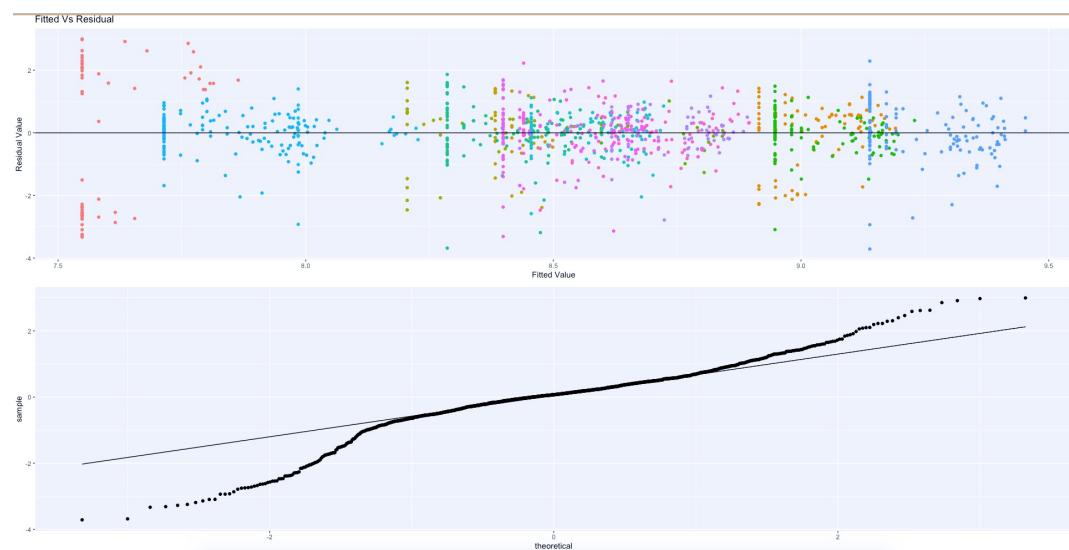


Figure 11: MALAT1 – CD19: Diagnostic Plots

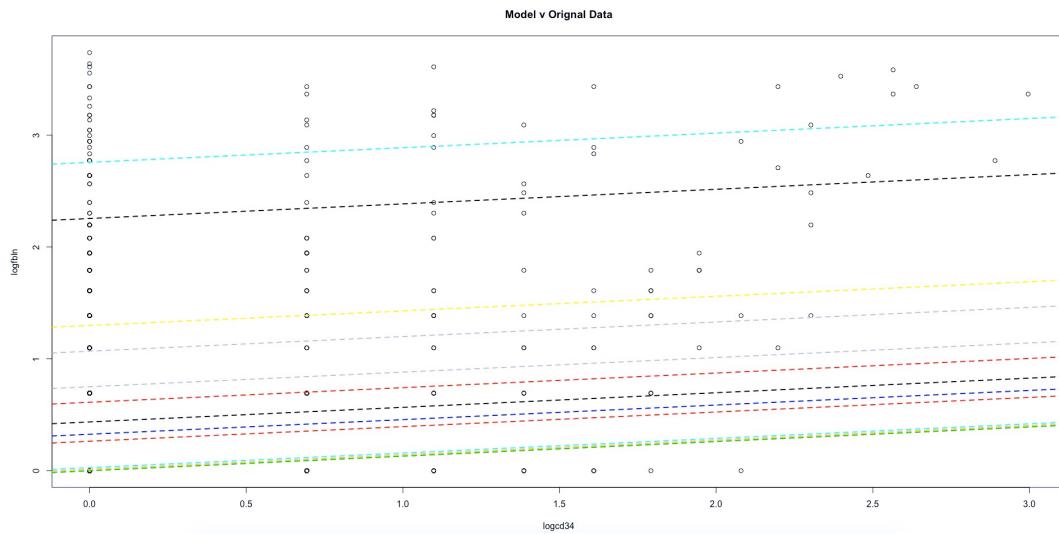


Figure 12: FBLN1 – CD34: Model v Original Data

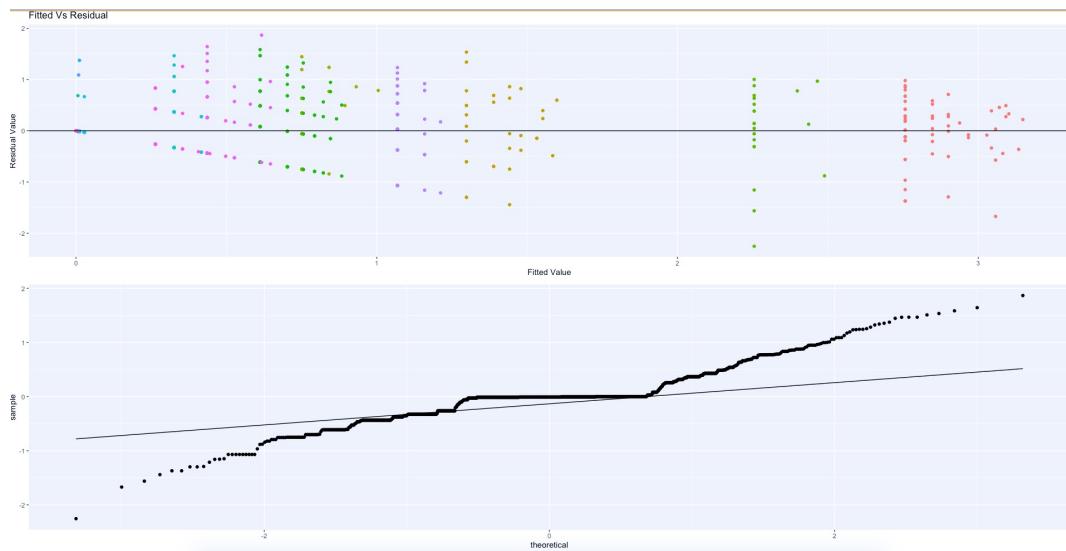


Figure 13: FBLN1 – CD34: Diagnostic Plots

Model 2

333

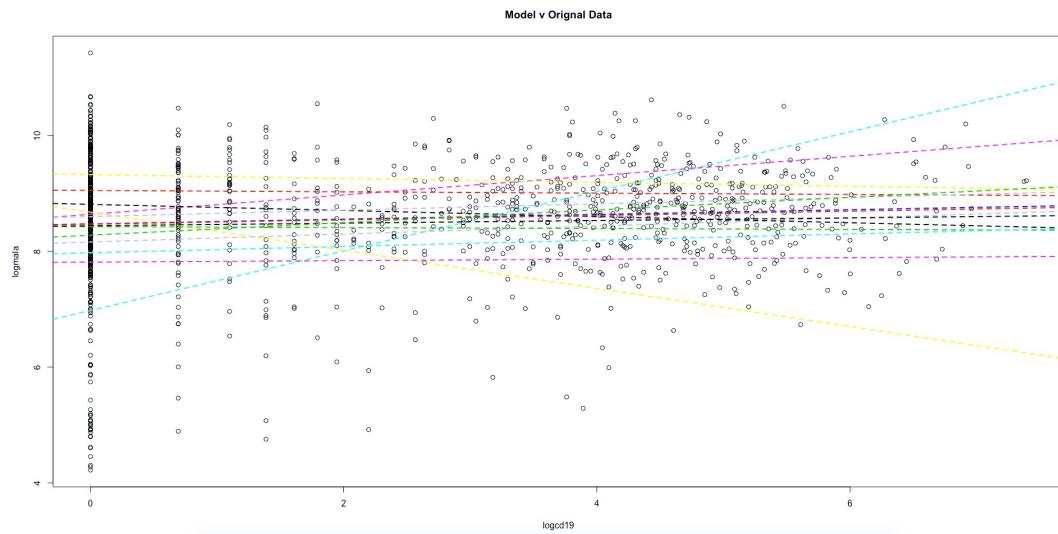


Figure 14: MALAT1 – CD19: Model v Original Data

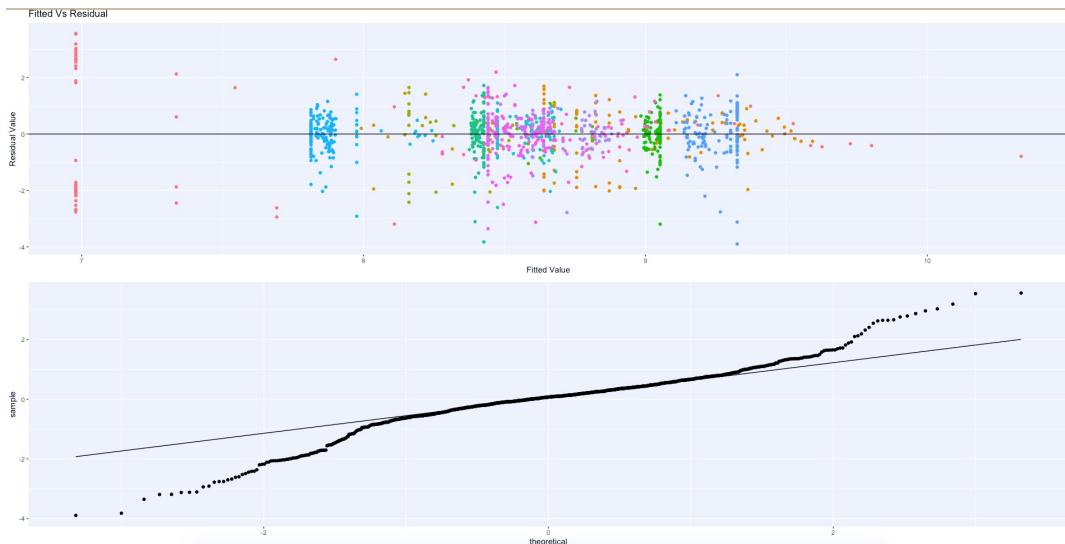


Figure 15: MALAT1 – CD19: Diagnostic Plots

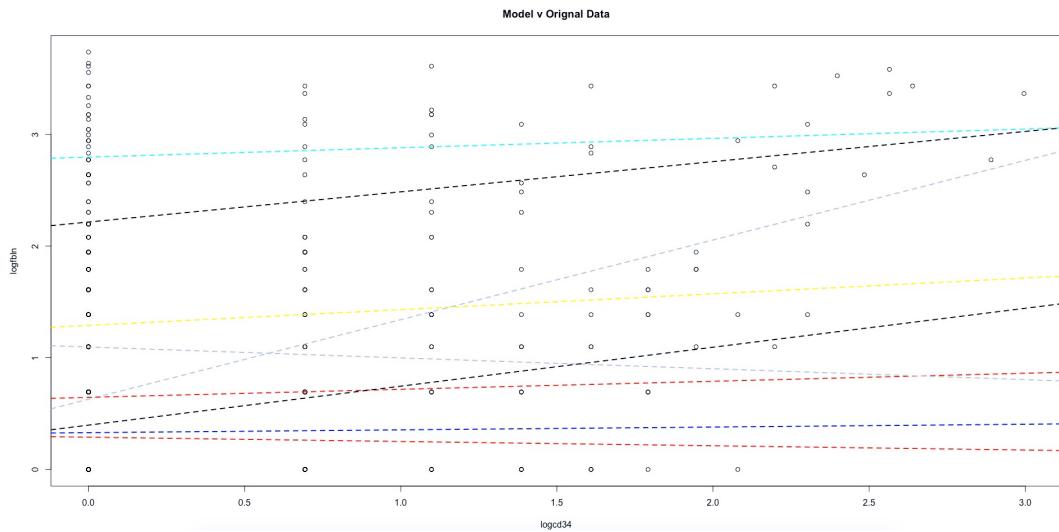


Figure 16: FBLN1 – CD34: Model v Original Data

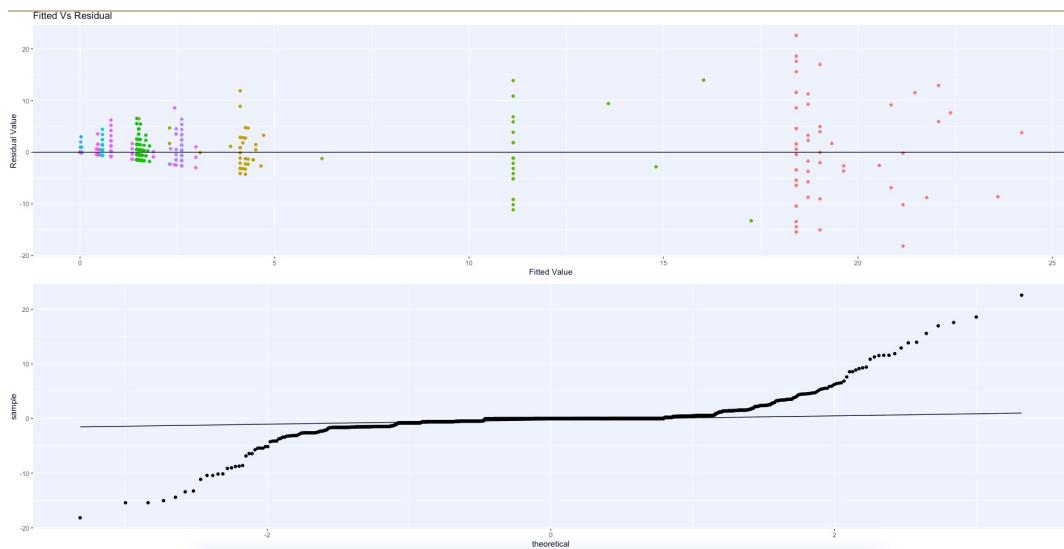


Figure 17: FBLN1 – CD34: Diagnostic Plots

Linear Mixed Model Diagnostic Plots

334

Model 3

335

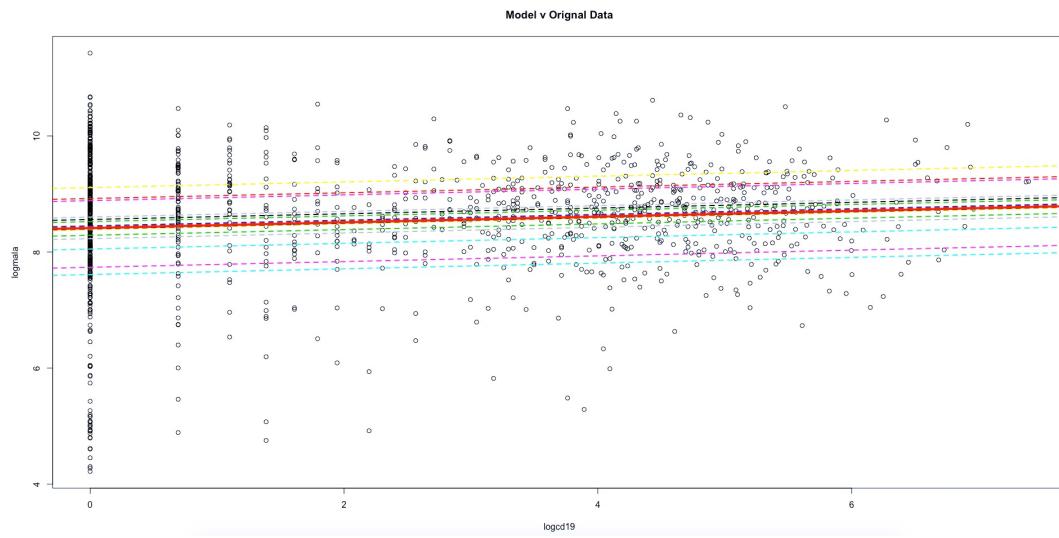


Figure 18: MALAT1 – CD19: Model v Original Data

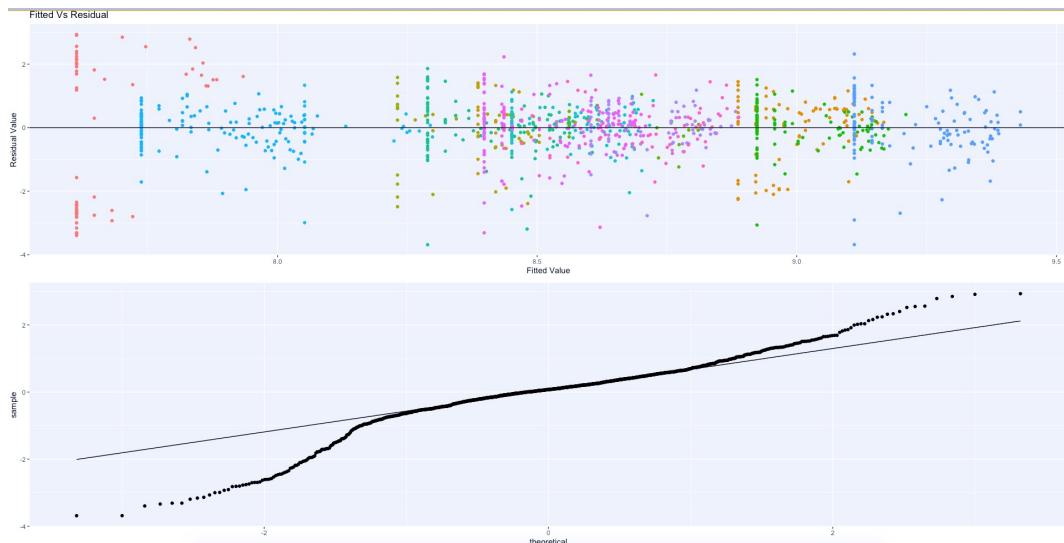


Figure 19: MALAT1 – CD19: Diagnostic Plots

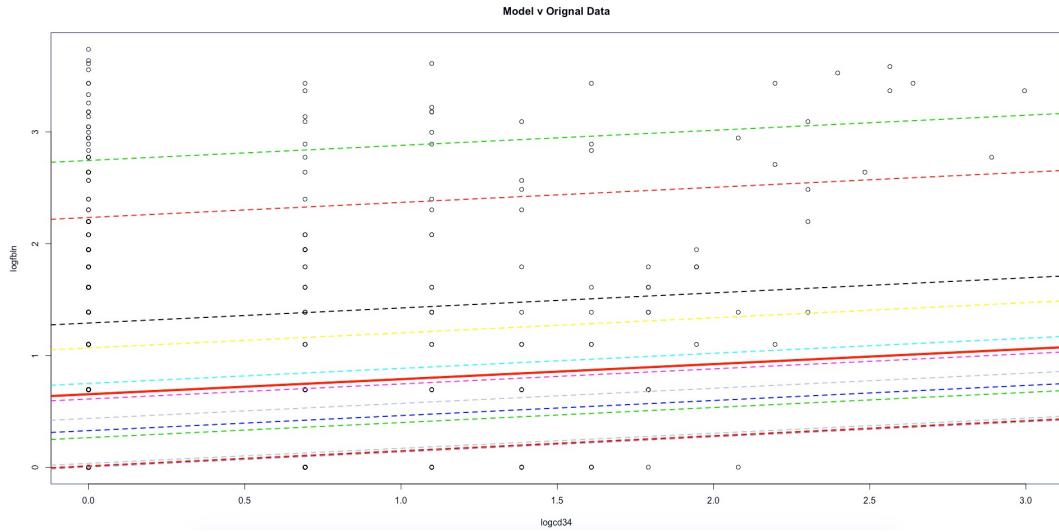


Figure 20: FBLN1 – CD34: Model v Original Data

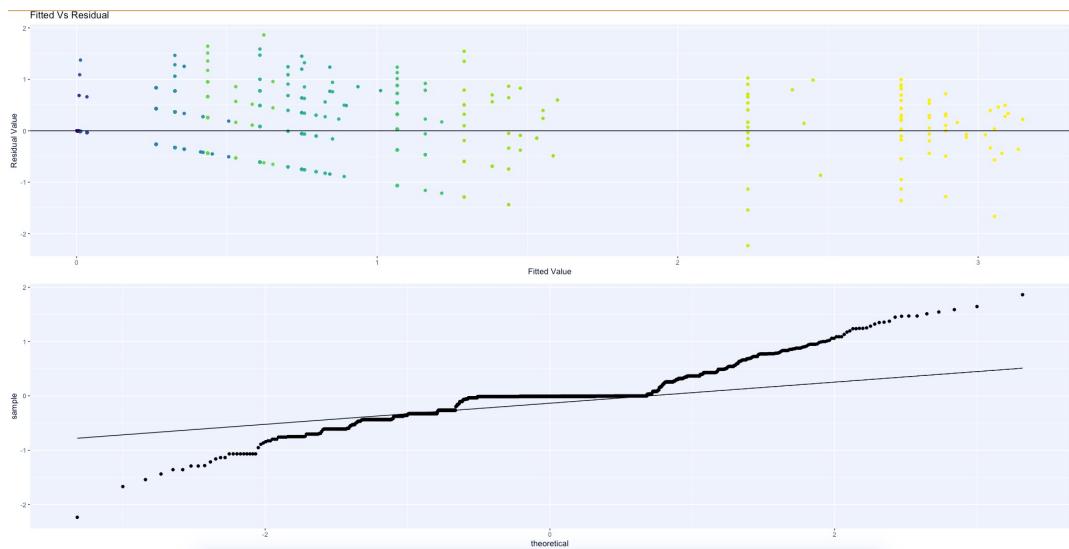


Figure 21: FBLN1 – CD34: Diagnostic Plots

Model 4

336

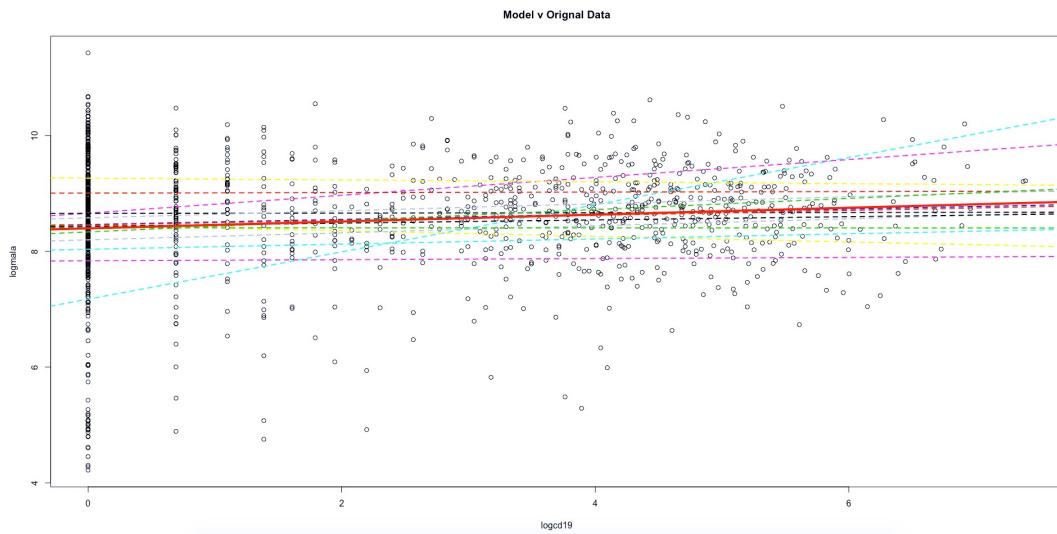


Figure 22: MALAT1 – CD19: Model v Original Data

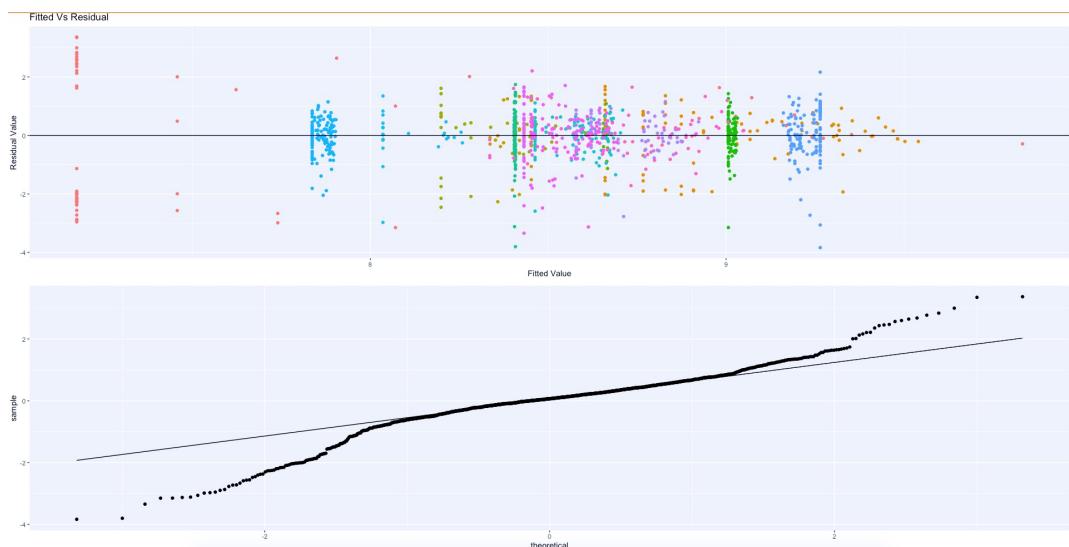


Figure 23: MALAT1 – CD19: Diagnostic Plots

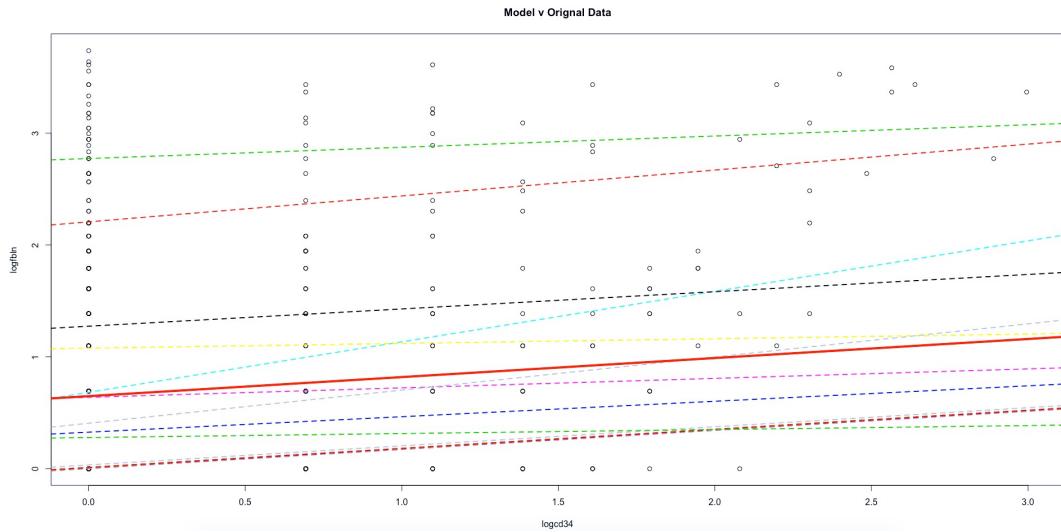


Figure 24: FBLN1 – CD34: Model v Original Data

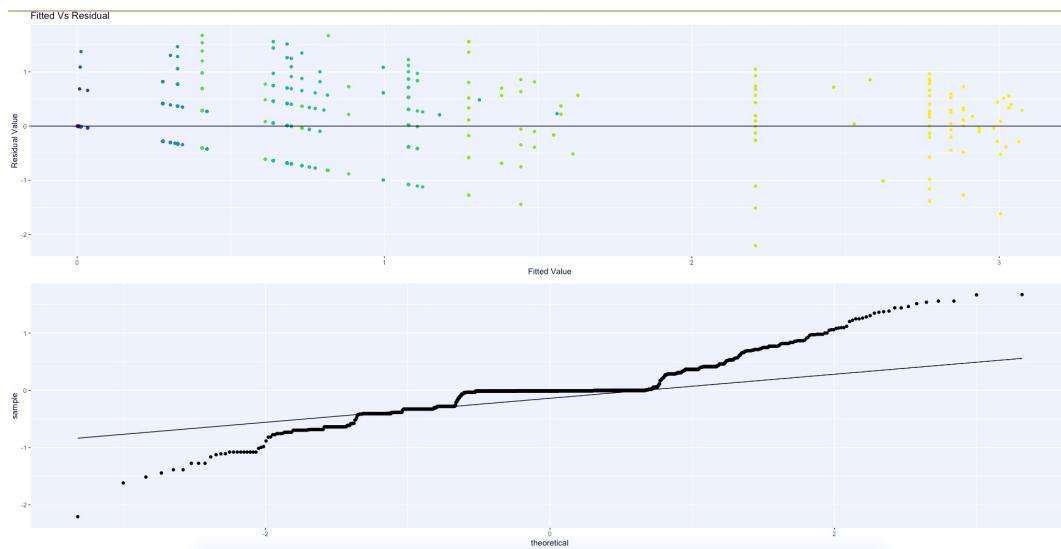


Figure 25: FBLN1 – CD34: Diagnostic Plots

Generalized Estimating Equations

337

Model 5

338

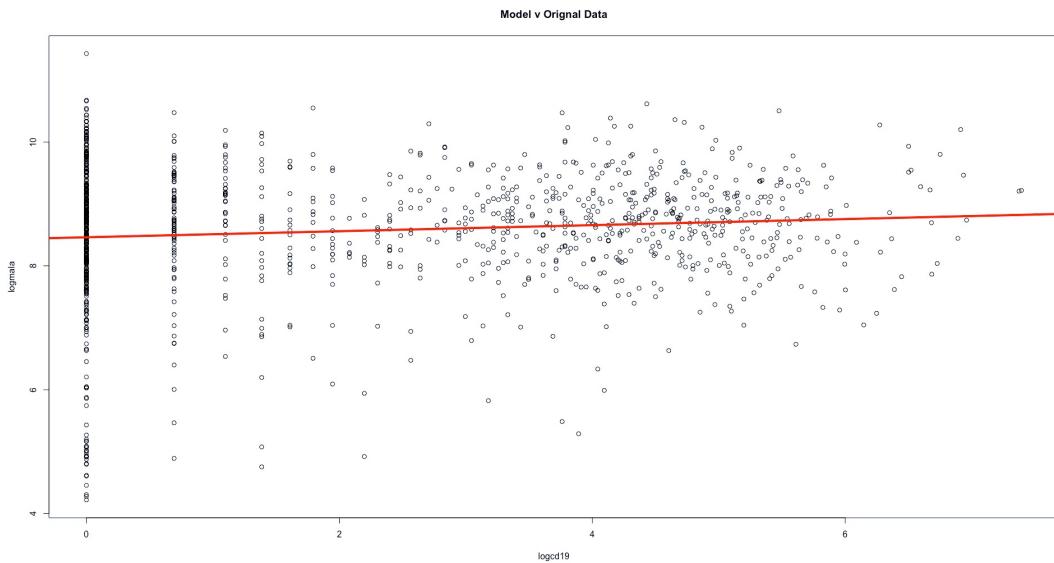


Figure 26: MALAT1 – CD19: Model v Original Data

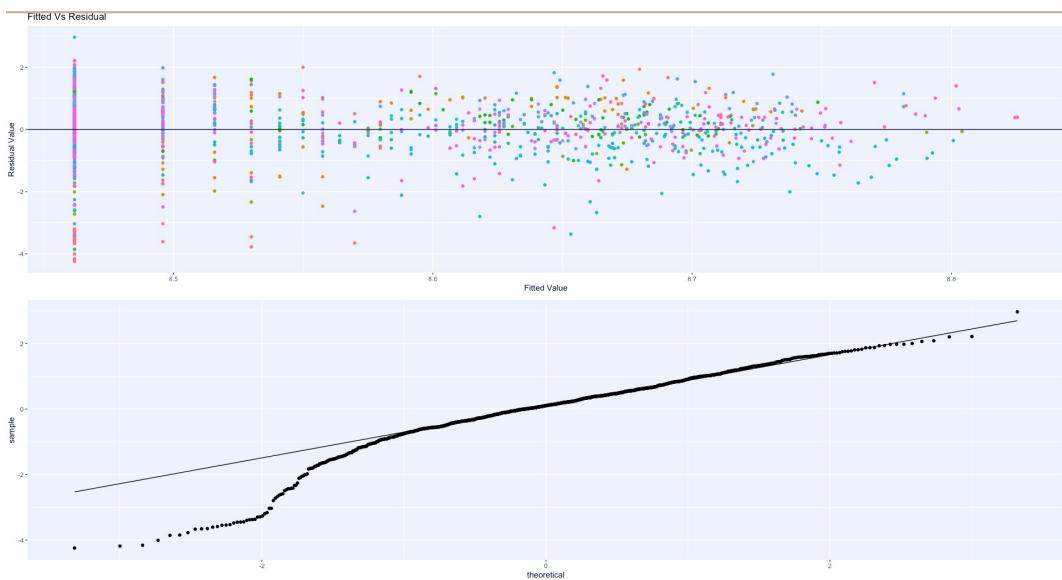


Figure 27: MALAT1 – CD19: Diagnostic Plots

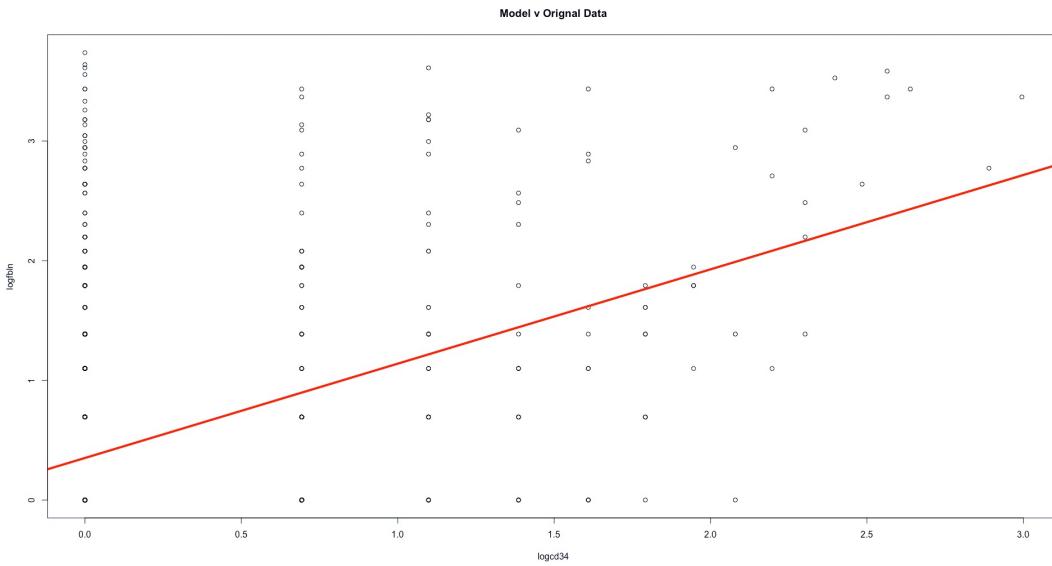


Figure 28: FBLN1 – CD34: Model v Original Data

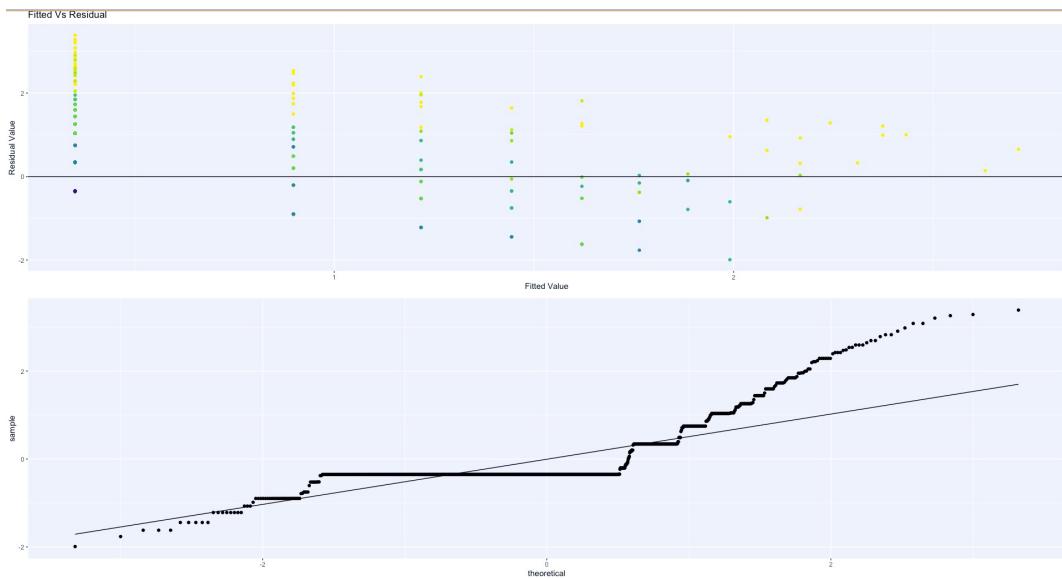


Figure 29: FBLN1 – CD34: Diagnostic Plots

Code and Data

339

All code for the above analysis was written and evaluated in RStudio Version 1.2.1335, and
340
is available for download at the following GitHub repository:

341

https://github.com/leepanter/MSproject_RBC.git

342

Additionally, a link to all necessary and reference data files (including original data) are contained in the following Google Drive:

https://drive.google.com/open?id=1gjHaMJG0Y_kPYWj5bIE4gRJU5z9R2Wqb

343

344

345

References

346

1. Macaulay IC, Voet T (2014) Single cell genomics: Advances and future perspectives. *PLoS genetics* 10: e1004126. 347
348
2. Bacher R, Kendziora C (2016) Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology* 17: 63. 349
350
3. Ståhlberg A, Andersson D, Aurelius J, et al. (2010) Defining cell populations with single-cell gene expression profiling: Correlations and identification of astrocyte subpopulations. 351
352
Nucleic acids research 39: e24–e24. 353
4. Amir E-aD, Davis KL, Tadmor MD, et al. (2013) ViSNE enables visualization of 354
high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology* 31: 545. 355
356
5. Wang Z, Gerstein M, Snyder M (2009) RNA-seq: A revolutionary tool for transcriptomics. 357
Nature reviews genetics 10: 57. 358
6. Kharchenko PV, Silberstein L, Scadden DT (2014) Bayesian approach to single-cell 359
differential expression analysis. *Nature methods* 11: 740. 360
7. Xue Z, Huang K, Cai C, et al. (2013) Genetic programs in human and mouse early 361
embryos revealed by single-cell rna sequencing. *Nature* 500: 593. 362
8. Marco E, Karp RL, Guo G, et al. (2014) Bifurcation analysis of single-cell gene expression 363
data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences* 111: 364
E5643–E5650. 365
9. Arazi A, Rao DA, Berthier CC, et al. (2018) The immune cell landscape in kidneys of 366
lupus nephritis patients. *bioRxiv* 363051. 367
10. FlowJo X V10. 0.7 r2 flowjo. *LLC* <https://www.flowjo.com>. 368

11. Hashimshony T, Senderovich N, Avital G, et al. (2016) CEL-seq2: Sensitive highly- 369
multiplexed single-cell rna-seq. *Genome biology* 17: 77. 370
12. Satija R, others (2018) Seurat: Guided clustering tutorial. *Satija Lab* http://satijalab.org/seurat/pbmc3k_tutorial.html. 371
372
13. Gutschner T, Hä默le M, Diederichs S (2013) MALAT1—a paradigm for long noncoding 373
rna function in cancer. *Journal of molecular medicine* 91: 791–801. 374
14. Debeer P, Schoenmakers E, Twal W, et al. (2002) The fibulin-1 gene (fbln1) is disrupted 375
in at (12; 22) associated with a complex type of synpolydactyly. *Journal of medical genetics* 376
39: 98–104. 377
15. Fitzmaurice GM, Laird NM, Ware JH (2012) Applied longitudinal analysis, John Wiley 378
& Sons. 379