Comparing Models for Single-Cell RNA Sequencing Data

Final Project -- BIOS 6643 – FALL 2019

Lee Panter

Introduction

The discovery of DNA, RNA, and the development of genomic sequencing methods has proven to be a valuable tool of scientific research. However, the methods developed in the first-generation of technology relied on "bulk" sampling, which could only estimate population-average expression of RNA and DNA. In order to obtain a more complete understanding of how the cellular landscape functions, estimates of cell-to-cell variability would need to be estimated. Development of single-cell RNA sequencing (scRNAseq) technology has increased to satisfy this need, but a need for statistical analysis is still outstanding. Previous methods used to model bulk RNA sequencing data do not account for the correlated nature of scRNAseq data.

This paper will compare six different modeling approaches on an observational scRNAseq data set obtained from a Lupus Nephritis Study of 33 patients across the United States. Two RNA genes were selected to be the predictor-response pair to simplify the modeling process. The main goal of this report is to investigate the ways in which parameter estimates vary as the modeling methodology is altered. It is hoped that the results of this investigation are useful for the development of future models for scRNAseq data sets.

Data

A single-cell RNA sequencing (scRNAseq) expression profile is a matrix of count-values representing a "snapshot" of the magnitude of activity of genomic features of a single cell. ("Gene Expression Profiling - Wikipedia," n.d.) In its original form, the data being studied here had a data matrix that contained 9,560 single-cell observations clustered within 27 samples (5 control not included in data). Each observation contained the expression of 38,354 genetic features.

Single-Cell data is often unreliable, protocol dependent, and can often have batch effects. Data quality control (QC) filters out redundant measures, and dead cell observations. The Seurat guided tutorial (Satija and others 2018) was used to perform quality control, filtering out observations with:

- Percent Mitochondrial DNA $> 60\%$
- Genetic Features Expressed $< 1{,}000$
- Genetic Features Expressed $> 5{,}000$
- B-cells only

These quality control measures reduced the original data by 88%, leaving only 1,110 observations clustered within 15 samples. Two genes (MALAT1 and CD19) were then selected from the set of genetic features in the initial data to be studied due to a higher correlation. MALA1 has been consistently linked with cancer metastasis, cell migration, and cell regulation. ("MALAT1 Gene - Genecards | Malat1 Rna Gene," n.d.) CD19 encodes a cell surface molecule which regulates lymphocyte proliferation and differentiation. ("CD19 Gene - Genecards | Cd19 Protein | Cd19 Antibody," n.d.).

Histograms and a joint distribution scatter plot were constructed to visualize the distributions of the selected variables (Appendix: Fig1-Fig3). The presence of zeros in the data indicated that the distribution might be well suited for a zero-inflated mixture model. Specifically, since the response was count-valued, the histograms indicated that a zero-inflated Poisson Generalized Linear Model or Generalized Linear Mixed Model would be appropriate. Additionally, while normality was not expected, log-transformations were also applied (Appendix: fig4-fig6), and resulted in approximate normality of the response MALAT1 and a bimodal distribution of the predictor CD19.

Methods

The six different modeling methodologies explored in this investigation are: Linear Models with Fixed Effects (LMwFE), Linear Mixed Models with Random Effects (LMMwRE), Poisson Generalized Linear Models without

overdispersion (POI), Poisson quasi-likelihood Generalized Linear Models with over dispersion (POIql), Poisson Generalized Linear Mixed Models with over-dispersion (POILMM), and Zero-Inflated Poisson Generalized Linear Mixed Models (ZIP). Models are fit on non-transformed data.

We assume that repeated measure residual errors are independent, i.e for subject $i = 1, \ldots, 15$ and repeated measure $j = 1. \ldots, n_i$

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$$

In an attempt to gain insight into the processes governing cellular relationships with their host subjects, a subject parameter was also fit using either a fixed or a random effect.

$$Fixed\ Effect: \quad \beta_{0i}$$
$$Random\ Effect: \quad b_{0i}$$

where:

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

An intercept and a covariate parameter are included in all models, these parameter estimates will be the primary focus for comparing the modeling methods.

$$\gamma_{ij} = \beta_0 + \beta_1\ CD19$$

Unless otherwise stated, the canonical link function will be used:

$$\mu_{ij} = g^{-1}(\eta_{ij}) = log(\eta_{ij})$$

Zero-Inflated models will use a fixed effect intensity model with only an intercept

$$R_{ij} \sim bernoulli(p_{ij}|\alpha_{0i}) \quad p(R_{ij} = 1) = \alpha_{0i}$$

And we will allow the intensity model parameters to vary across the subject parameter. This terminology allows the models to be written as:

Model	Subject as Fixed Effect-Predictor	Subject as Random Effect
LMwFE	$\beta_{0i} + \gamma_{ij} + \epsilon_{ij}$	NA
LMMwRE	NA	$b_{0i} + \gamma_{ij} + \epsilon_{ij}$
POI	$\mu_{ij} = g^{-1}(\eta_{ij} = \beta_{0i} + \gamma_{ij})$	NA
POIql	$\mu_{ij} = g^{-1}(\eta_{ij} = \beta_{0i} + \gamma_{ij})$	NA
POILMM	NA	$\mu_{ij} = g^{-1}(\eta_{ij} = b_{0i} + \gamma_{ij})$
ZIP	$R_{ij} \sim bernoulli(p_{ij}\|\alpha_{0i})$ $\mu_{ij} \mid (r_{ij} = 1) = g^{-1}(\eta_{ij} = \beta_{0i} + \gamma_{ij})$	$R_{ij} \sim bernoulli(p_{ij}\|\alpha_0)$ $\mu_{ij} \mid (r_{ij} = 1) = g^{-1}(\eta_{ij} = b_{0i} + \gamma_{ij})$

We note that there are four "NA" spots at which the model is either redundant (as in the case of LMwFE=LMMwRE with only fixed effects) or not possible to fit (cannot incorporate random effects). After these eliminations, there are a total of 7 models being estimated.

Results

Intercept

Subject Effect	LMwFE	LMMwRE	POI	POIql	POILMM	ZIP	Variable Labels
Fixed Effect	7.2786e3	X	8.821	8.957	X	8.958	Estimate
	2.4548e2		4.856e-4	3.007e-2		3.689e-4	Standard Error
	<2e-16		<2e-16	<2e-16		<2e-16	p-value
Random Effect	X	7.3324e3	X	X	8.8362	8.9402	Estimate
		7.6899e2			1.0163e-1	6.5229e-4	Standard Error
		<2e-16			<e-5	<e-4	p-value

Slope

Subject Effect	LMwFE	LMMwRE	POI	POIql	POILMM	ZIP	Variable Labels
Fixed Effect	2.562	X	3.246e-4	8.839e-5	X	8.559e-5	Estimate
	1.501		2.513e-6	1.913e-4		2.176e-6	Standard Error
	8.8131e-2		<2e-16	0.644		<2e-16	p-value
Random Effect	X	2.495	X	X	3.16e-4	2.9282e-4	Estimate
		1.491			1.6525e-4	2.1289e-6	Standard Error
		9.4193e-2			5.61e-2	<e-4	p-value

Discussion

Upon noting that:

$$6.7750 * 10^3 \approx e^{8.821} \leq \hat{\beta}_{0LMwFE}, \hat{\beta}_{0LMMwRE} \leq e^{8.958} \approx 7769$$

We see that changes in modeling strategy have little impact on the magnitude of the intercept estimate. In fact, all intercept estimates agreed in sign, and magnitude. The increase in p-value for intercept estimates from "POILMM-Random Subject" and "ZIP-Random Subject" are also associated with a corresponding drop in standard error and can therefore be attributed to improved model precision.

These statements are not true for covariate parameter estimates. While these parameters are comparable within similar model frameworks, the estimates agree on only sign across all models. This is not an unexpected result, since parameter estimates for slopes are heavily dependent on how observations are corelated, and this concept is approached differently in all of the models employed.

The following matrices of values represent the percent changes for the intercept and slope estimates as the model is changed:

Intercept

LMwFE	0.0000	-0.0008	0.0081	-0.0072	0.0064	-0.0073	-0.0053
LMMwRE	0.0008	0.0000	0.0089	-0.0064	0.0072	-0.0065	-0.0045
POI	-0.0081	-0.0090	0.0000	-0.0154	-0.0017	-0.0155	-0.0135
POIql	0.0072	0.0064	0.0152	0.0000	0.0135	-0.0001	0.0019
POILMM	-0.0064	-0.0072	0.0017	-0.0137	0.0000	-0.0138	-0.0118
ZIP Fixed	0.0073	0.0065	0.0153	0.0001	0.0136	0.0000	0.0020
ZIP Random	0.0053	0.0045	0.0133	-0.0019	0.0116	-0.0020	0.0000

The model intercept exhibits stability, while the covariate estimate does not.

SLOPE

LMwFE	0.0000	0.0262	0.9999	1.0000	0.9999	1.0000	0.9999
LMMwRE	-0.0269	0.0000	0.9999	1.0000	0.9999	1.0000	0.9999
POI	-7891.7911	-7685.3832	0.0000	0.7277	0.0265	0.7363	0.0979
POIql	-28984.1793	-28226.1750	-2.6724	0.0000	-2.5751	0.0317	-2.3128
POILMM	-8106.5949	-7894.5696	-0.0272	0.7203	0.0000	0.7291	0.0734
ZIP Fixed	-29932.4034	-29149.6017	-2.7925	-0.0327	-2.6920	0.0000	-2.4212
ZIP Random	-8748.4024	-8519.5929	-0.1085	0.6981	-0.0792	0.7077	0.0000

Limitations and Future Research

Initial quality control measures for the scRNAseq data found an extremely high presence of mitochondrial RNA (mRNA). Mitochondrial functionality is essential to cellular processeses, upon death such functions cease, and mitochondria degrade. As a result, mRNA content is used as a quality control measure to indicate progression of cellular death. The Seurat tutorial for single-cell analysis recommended filter parameters of %mRNA $<$ %5; however this threshold elimated all data when used in conjunction with the other recommended quality control measures. A final determination of %mRNA $<$ %60 was used to preserve as much of the clustering-structure as possible, with the trade-off of analyzing dead cells.

The results of this investigation are based upon mostly non-living biological samples. While interpretation of model parameters is certainly possible, the individual estimates are less meaningful considering the contex than the estimate trends that we are seeking.

The analysis conducted in this investigation is inadequate and inaccurate. The results are inaccurate because comparisons made in this investigation were not made carefully. Neither controll models or nested modeling techniques have been used to compare parameter changes. In order to fix this, the base model (in the case of this analysis-LMwFE) needs to incorperate BOTH fixed effects and random effects. Additionally, a parameter needs to be defined for both fixed and random effects (separate variables) which can be used to compare the effects of added variables. The results of this investigation are inadequate because the original goal of the investigation was to investigate models over single-cell data and isolate the specific effects that the specified modeling strategies have over parameter estimates. Since only a vauge conlcusion regarding slope and intercept was obtained, this investigation has fallen short of its objective.

Consequently, the future of research for this project is very broad in scope. Immediate considerations will be made to compare more simplistic modeling strategies over more similar models.

References

"CD19 Gene - Genecards | Cd19 Protein | Cd19 Antibody." n.d. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CD19&keywords=CD19.

"Gene Expression Profiling - Wikipedia." n.d. https://en.wikipedia.org/wiki/Gene_expression_profiling.

"MALAT1 Gene - Genecards | Malat1 Rna Gene." n.d. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MALAT1.

Satija, R, and others. 2018. "Seurat: Guided Clustering Tutorial." *Satija Lab Http://Satijalab. Org/Seurat/Pbmc3k_tutorial. Html*.

Appendix

Initial Variable Summary Plots

Figure 1: CD19 Histogram

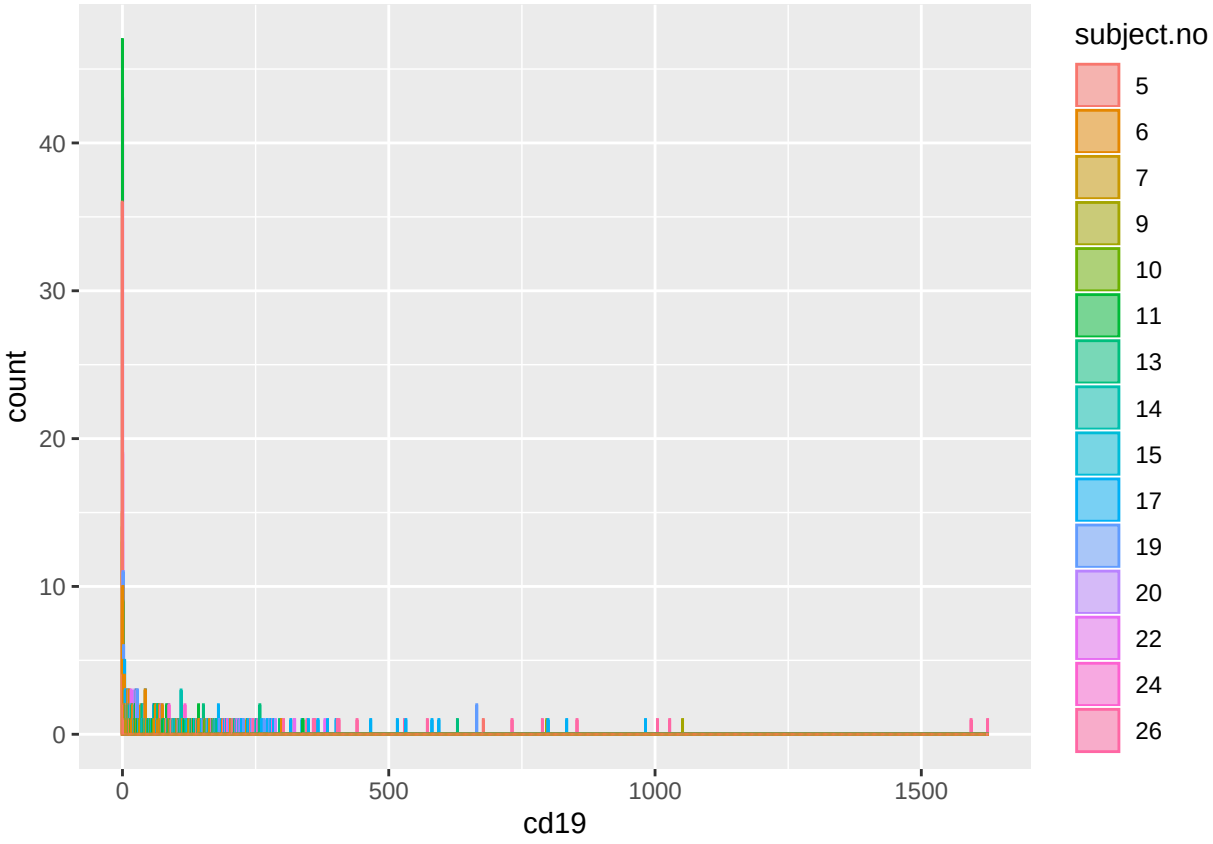


Figure 2: MALAT1 Histogram

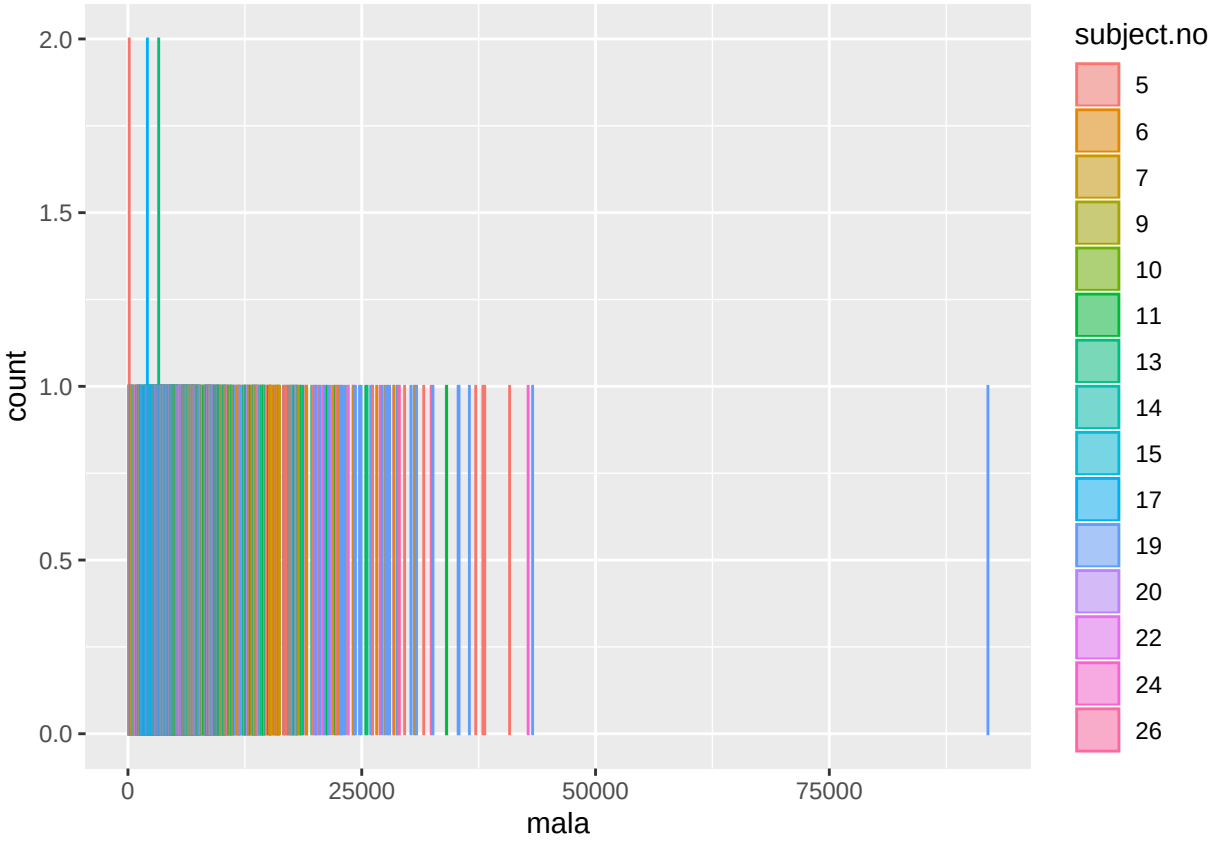
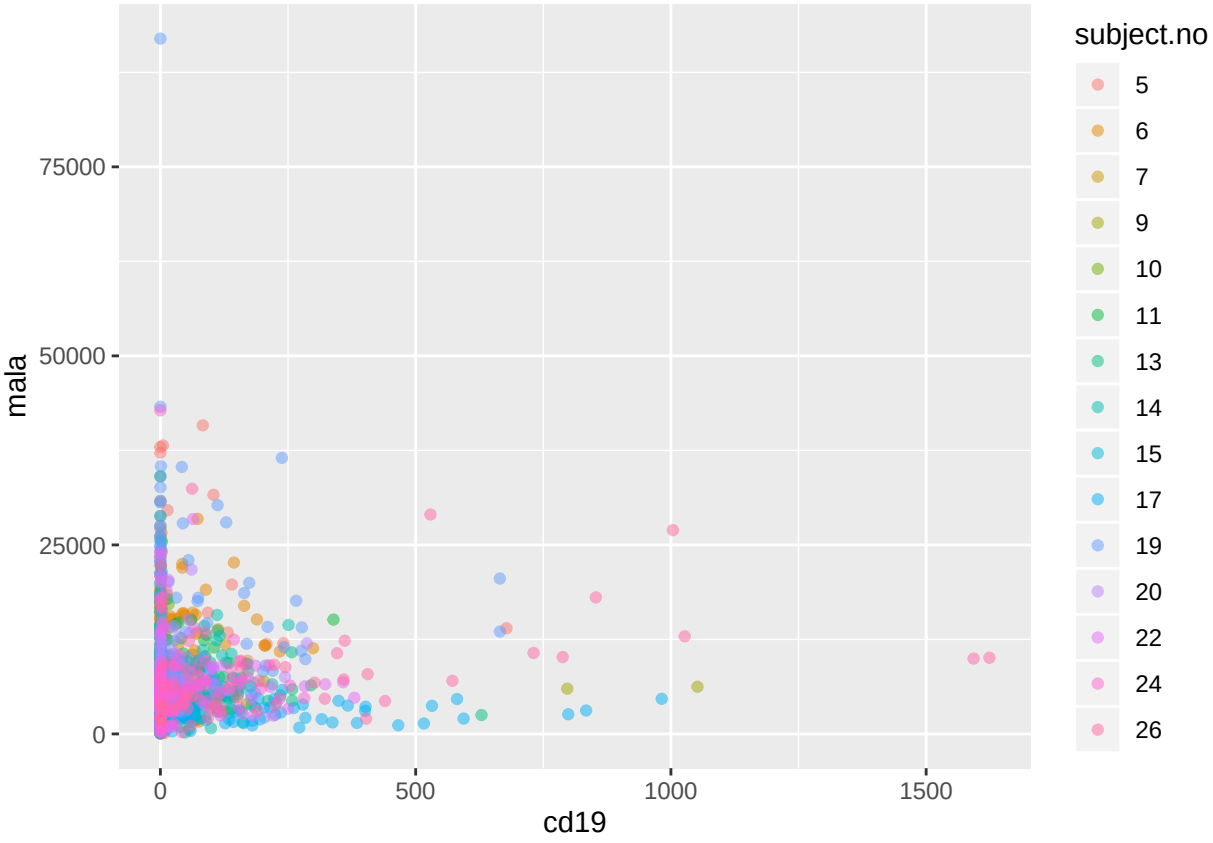


Figure 3: MALAT1 vs CD19 Scatter Plot



Log Transformed Variables

Figure 4: CD19 Histogram

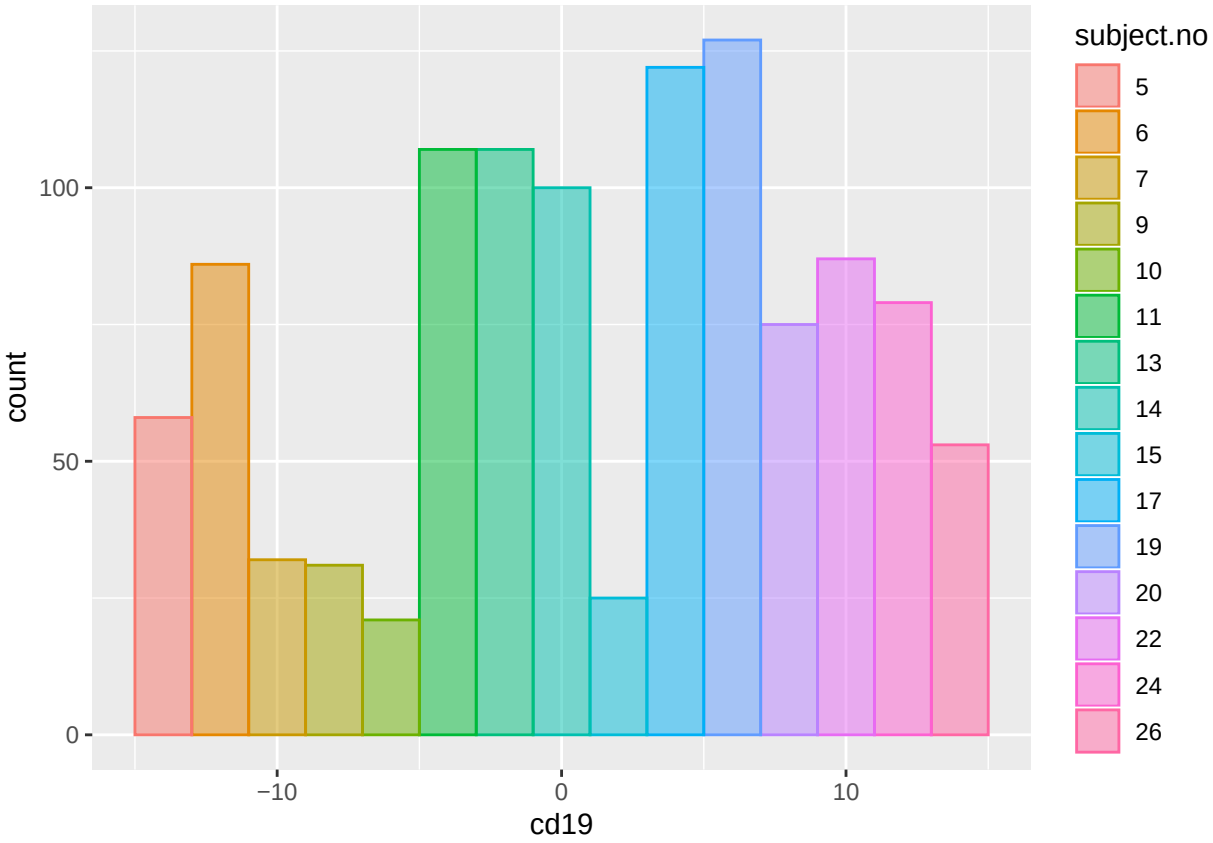


Figure 5: MALAT1 Histogram

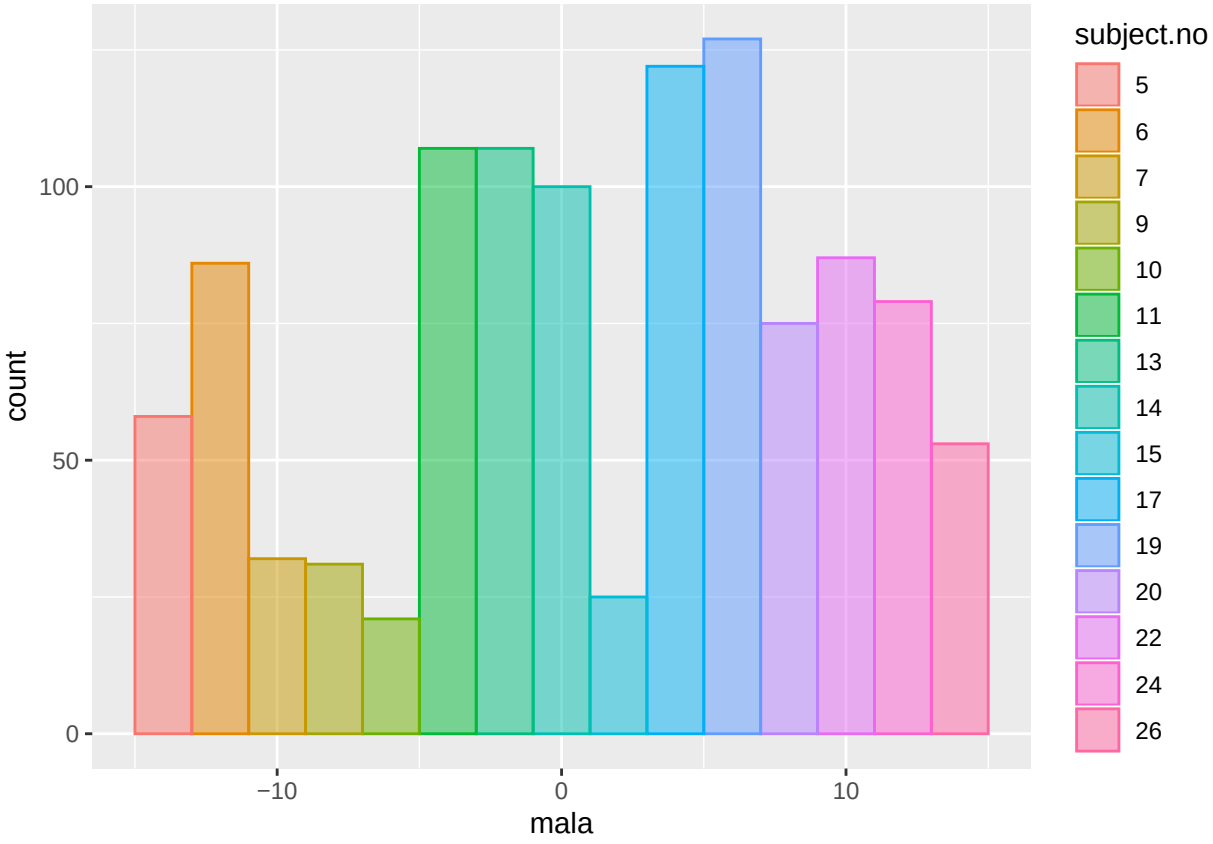


Figure 6: MALAT1 vs CD19 Scatter Plot