Emery and Ayu Variational Autoencoder Project

**HVG detection**

See https://www.ncbi.nlm.nih.gov/pubmed/29481632

How do other programs determine the most variable genes?

Normalization has a great impact on HVG detection. Because heteroskedacity is observed in expression data, variance cannot be directly used as an indicator of HVGs. Most existing methods use the relationship between variance (or something closely related to variance) and the mean as an indicator. A model of the relationship is fitted and statistical test for high biological variability is then performed.

**Seurat:**

After controlling for the relationship between mean expression and variability (using log CPM) they calculate a mean and dispersion measure (variance/mean) for each gene across all cells. Put genes into 20 bins based on average expression. Within each bin, z-normalize the dispersion measure of all genes within the bin to identify genes whose expression values were highly variable. They used a z-score cutoff of 2 to identify 160 significantly variable genes, after excluding genes in QC. This strategy decreases the results’ dependency on the mean.

**Scran**

Similar to DESeq for normalization, in which a specialized scaling factor calculation algorithm is used for normalization, but with multiple scaling factors calculated for multiple pools of cells. A more accurate estimate of the scaling factor is used for each cell by using systems of linear equations. Variance calculated with log transformed expression matrix. Then uses LOESS with the mean-variance relationship of log-transformed expression values. After obtaining a LOESS fit, it estimates the technical variance in each gene. Then subtracts inferred technical component from each variance to obtain biological variation.

**scLVM**

Uses DESeq’s method for normalization and then detects HVG using two different approaches. Either log-mean to log CV2 wi a log linear fit (scLVM log). LogVar algorithm calculates variance with a logarithmic transformed expression matrix (similar to in scran). The second method uses a LOESS method (locally weighted scatterplot smoothing) with the mean variance relationship after a log transformation.

**BASiCS**

Uses variance and a Bayesian hierarchical model. Seems to require spike-ins. Firs fits spike-ins to a hierarchical model and then uses all genes in the dataset to extend the hierarchical model to account for biological cell-to-cell variability. Decomposes variance into biological background, technical variability, and biological variability. Then use biological variability to test for HVGs.

**scVEGs**

normalizes data using the relative expression method, which then is multiplied by the mean of the samples total counts. uses coefficient of variation (CV) to model variations in genes. Model mean versus CV relationship by assuming a negative binomial distribution, which is then fitted with a modified local regression. After getting the parameters from the fitted model, p-values are obtained by assuming the normal distribution in the difference between each gene and the model curve.

**Brennecke**

Uses DESeq’s method for normalization. Uses coefficient of variation, squared to analyze biological variation. Technical noise is modeled by fitting a generalized linear model onto the mean versus the CV2 plot. Using this model the coefficient of biological variation is obtained and a chi-squared distribution is used to test for high variance. They use the chi-square distribution because the extraordinarily high variance indicates a high degree of uncertainty of variation in a gene with low expression. You can also filter genes with high uncertainty using a CV2 threshold, which allows for better control of false positive rates.

Vast differences are seem between methods. Each tool performs optimally in different situations. Brennecke has good stable performance, great sensitivity and specificity, though scran and Seurat performed optimally with some data sets.

BASiCS performed best with simulated data, but did not do well with real data. Also BASiCS uses spike-ins, which many, or even most projects do not use. So I do not think we should use a method like BASiCS uses.

BASiCS and Brennecke tend to call the most HVgS, WHEREAS SCvegS AND Seurat tend to call far fewer HVGs. ScLVMs performance seems to be very dependent on the expression mean.

Scran tended to have good performancein clustering, stable number of detected HVGs, good independency from the mean, and good run time. Something to consider is that HVG analysis requires large sample sizes in order to be reproducible. Many methods show increasing rediscovery rates as the sample size increases. Higher sequencing depth also seems to be able to improve reproducibility. Utilization of overlaps between tools seemed to really help these measures. One issue with the review is that the stringency of assigning P-values between each tool is different, which can result in different numbers of detected genes when different tools are used. Also each tool used a slightly different normalization method and so it is hard to truly compare which method is best in terms of simply selected HVGs.