

Candidacy Written Exam

Brian Arand

1. **Sort-of-basics:** Consider the venerable p-value. It is ubiquitously used by the bioinformatics community to determine the viability of their hypotheses results. However, it is running into a maelstrom of criticism as noted in the following commentaries found in the first two links. The other two in the list below are more informed manuscripts that dwell on this controversy.

You are asked to consider the criticism in the context of single-cell technologies. How will the large inherent dimensionality and relatively large samples (number of cells) will impact hypotheses testing in general for populations of cells? Will the use of CI, or confidence interval be of more value? Please carefully answer this question after perusing the commentaries and critiques. You are also welcome to peruse other material.

- a. <http://debunkingdenialism.com/2015/04/01/new-nature-methods-paper-argues-that-p-values-should-be-discarded/>.
- b. <http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412>
- c. <http://www.nature.com/news/scientific-method-statistical-errors-1.14700>
- d. <http://www.nature.com/nmeth/journal/v12/n3/full/nmeth.3288.html>

The comments to Nature made by Leek and Peng in *Pvalues are just the tip of the iceberg*[1] makes the point that p-values very often in practice are not the weakest link in the chain. They point out that drastic downstream changes in results can more easily be achieved by changing upstream experimental designs. In terms of many single-cell RNA-seq pipelines this is especially true. There are many sources of noise in RNA-seq data, many options for normalization, many options for alignment, and analysis. However, I still have to rebut. Sure, significance can be toggled by the specific choices made during data cleaning or the permutations of confounding factors for which one adjusts. Garbage in garbage out; I agree. But, the fact that there are potentially other causes to the increasingly arbitrarily results published in recent years is simply a red herring to the very specific dialogue surrounding the utility of the p-value. Just because there are loose and broken cogs in our machine doesn't mean we should ignore a cog that has been building up rust over the years. We need to dust off the p-value for single-cell RNA-seq data just as readily as we should standardize the handling of batch effects—for instance. A single faulty cog and we can be assured that our machine will produce garbage no matter the input.

With other points made by the co-authors, I can agree. (In spite of the slippery slop argument made when they claim, “deregulating statistical significance

opens the door to even more ways to game statistics.”). I don’t think deregulation is necessarily the best approach, however. The Leek and Peng are a bit alarmist with their claim that “people need to stop arguing about p-values, and prevent the [upstream factors in data analysis pipelines] from sinking science”. I think the true enemy here is the arbitrarily chosen threshold for significance, and the cultural phenomenon that has surrounded it. Investigators and their audiences are no longer interested in the whole statistical story. Potentially the phenomenon was born out of laziness, potentially ‘ignorance. A solution to the threshold problem is simple: report the p-value in results rather than just the dichotomous significance labeling. In previous work of my own, I report not only p-values but also categories of significance levels that may be of interest to the reader ($p < 0.01$, $p < 0.05$, $FDR < 0.01$, $FDR < 0.05$). I think that such a technique not only appeals to the classical standards but also forces the reader to do a little thinking for themselves, having to address the arbitrary nature of significance thresholds. This, I think, was how I took a personal effort toward solving the cultural problem. But, perhaps I need to do more in the future. Especially now that I’m entering a sub-discipline of widely unknown statically landscapes—single-cell RNA-seq. Two characteristics of interest in this field are the dimensionality and relative sample size of datasets.

The high dimensionality of single-cell RNA-seq data, like other ‘omics’ datasets, share in the multiple comparisons problem[2]. Counters to this phenomenon arise from the idea of the adjusted p-value. The idea is to shrink the alpha significance threshold so that a Type 1 error rate is achieved across the entire comparison set rather than just a given test. For instance, under the assumption of independence, the Bonferroni correction corrects the p-value by dividing it by the total number of comparisons being made. Realizing that independence is too strong of an assumption in many applications, and that Type 1 error control isn’t as much a concern when there are very few tests that meet a certain significance level, investigators proposed the false discovery rate (FDR)[3]. Gelman et al.[4] suggest that some Bayesian approaches are more appropriate to counter multiple comparison artifacts in cases where classical null hypotheses of the form $H_0^j: \tau^j = 0$ for some treatment effects τ are not likely to be true. However, a publication by Efron & Tibshirani[5] discuss how in many genetics testing environments, such as differential expression, this exact type of null hypothesis is prominent between genes. This brings us back to p-value approaches when considering multiple comparisons for single-cell RNA-seq.

Sample size is also of concern for *Ching et al.* of the Cold Springs Harbor Laboratory in *Power analysis and sample size estimation for RNA-Seq differential expression* [6] compared statistical power of various, popular RNA-seq differential expression pipelines. Using negative binomial parameters learned from public databases, a synthetic data was created and analyzed using DEseq[7], EdgeR[8] and other packages. They concluded that sample

size was more important than sequencing depth for improving statistical power. This is good news for single-cell RNA-seq—a technology that is often seen trading depth for increased sampling. (This is a trade off made when trying to fit in as many samples into a single RNA-sequencing lane to reduce costs. However, there's a risk of exhausting reagents with too many samples that can reduce sequencing depth of all samples.) This is assuming the same phenomenon is true for the technology, of course. To my knowledge, similar work has not been performed for single-cell RNA-seq datasets.

We can also look at an example of sample sizes' role in single-cell RNA-seq statistical analysis. Consider a clustering of single cells. Many works are interested in the heterogeneous populations of cells now that cellular resolution sequencing is available. A subset of those works are specifically interested in the rare cell subtypes that are no longer masked by the averaging effects of bulk cell analysis. Typically, when comparing clusters, a set of differentially expressed genes can be found and statistically categorized as either 'significant' or 'not' using a p-value and alpha significance level. However, Harsey et al [9] points out the fickle nature of this setup upon repeated testing, especially when applied to a sparsely sampled population. It may therefore be inappropriate to apply traditional differential analysis techniques to clusters of rare cell types. This poses the practical problem in single-cell RNA-seq cluster analysis where, it might be acceptable to apply this significance test between some clusters but not others due to differences in expected reproducibility.

So what about providing more information, in the form of a confidence interval (CI)? I'll conclude that, at least, this much should be provided in addition to the p-value. By defining the span of values your test statistic is expected to achieve upon repetition, the CI can give an intuition for the robustness of a certain significance result. The confidence interval definitely conveys more analytics data to an audience. However, implicit in this confidence interval selection is yet another arbitrary threshold. Typically a 95% CI is used, but there's no restriction here. Moreover, the span of a CI is dependent upon sample size. Effect size on the other hand, is not. It reports just how 'big' of a difference is being observed independent of the sample size.

In cases similar to differential expression when an effect size like Cohen's d can be estimated, I think that this is the ultimate. In lieu of that, I think, a visualization of the distributions in question can convey the same story (with less rigor, yes, but the same story). So in the case of differential expression analysis for single-cell data, report the violin plots along side your p-values, and confidence intervals, and effect sizes. (Why isn't this a standard practice already?)

So why not throw out the p-value all together? Well, yes, it does have a historical precedence and is a standard source for cross study comparison. However, my primary claim is that p-values can be useful in practice. They provide a convenient way to sort findings and draw attention to a subset of

potentially interesting ones when reporting many findings at a time. The p-value can also be used as a net to filter out the vast majority of the truly-insignificant results in a multiple comparisons context. After such filtering, other analytic metrics can be consulted.

Another suggestion that might be useful for single-cell RNA-seq data as well as many other bioinformatics disciplines is, in the case in which a p-value calculation is an intermediate step, generalize downstream analysis wherever possible to take the p-value into account. For instance when constructing gene networks, instead of choosing an arbitrary threshold of significance for the inclusion of edges, is it possible to broaden the inclusion of edges and then weight all included edges by a achieved p-value? Results (like cluster or hub detection) drawn from topologies that change in weight distribution may be more readily replicable than topologies that differ by the toggling of entire edges.

So in conclusion, just because a finding is statistically significant doesn't mean that the finding is practically significant. But!...just because the p-value isn't statistically valuable, doesn't mean that the p-value isn't practically valuable. In practice, we as bioinformaticians need to broaden our statistical vernacular. And single-cell RNA-seq analytics—in its sprawling, statistically diverse landscape—will challenge us and our ability to tell the whole story.

2. A-survey-of-sorts: Many argue that single-cell methods are here to stay as stated in the following publications:

1. <http://www.ncbi.nlm.nih.gov/pubmed/22323135>
2. <http://www.nature.com/nmeth/journal/v9/n1/full/nmeth.1819.html>

There is much work on the actual acquisition of the single cell measurements through appropriate micro-fluidics and chemistry. However, there is a paucity of work and surveys on techniques of analysis. Now the questions -

1. Given your understanding of acquisition technologies, please provide a systematic and mathematical formal description replete with symbols and detailed formulation of available signal and confounding noise.

For sample gene g from sample i :

Observation: $O_{ig} = B_{ig} + L_{ig} + T_{ig} + \epsilon$
 $\sim \text{NegativeBinomial}(p_{ig}, s_{ig})$

Where B_{ig} is the biological signal, L_{ig} is the biological bias and noise, T_{ig} is the technical bias and noise, and ϵ is normally distributed error.

Gene Abundance: $Y_{ig} = B_{ig} + L_{ig} = R_i \cdot \pi_{ig} \sim \text{Poisson}(\mu_{ig})$
Where R_i is the total genetic abundance, and π_{ig} is the proportion attributed to gene g . Considerations of over-dispersion have left most models to conclude $\mu_{ig} \sim \text{Gamma}$

Technical Noise: $T_{ig} \sim \text{Gamma}(k_{ig}, \theta_{ig})$

It has been suggested that $k_{ig} \cdot \theta_{ig}$ (the mean technical noise) is functionally dependent upon transcript abundance Y_{ig} [10, 11]. The largest source of technical bias and variation PCR amplification was also suggested to fit this model [12].

Mixture Models: $B_{ig} = \sum F_j(f_j) + \epsilon = \sum_{h \neq g} G_{Y_{ih}}(Y_{ih}) + \epsilon$

This is a catch-all mixture model of both latent variables dependent functions $F_j(f_j)$ and gene abundance dependent functions, $G_{Y_{ih}}(Y_{ih})$. This is my way of identifying that there are many ways to break down biological signal in to functional components. By the individual contributions of other transcripts, or by the contributions of functional components like $B_{ig} = \text{cell cycle} + \text{differentiation} + \epsilon$, or even biological organizational levels latent variables:

$B_{ig} = \text{organism} + \text{tissue} + \text{cell type} + \text{cell profile} + \epsilon$. I

will admit that even this is not a comprehensive model. As discussed in Shalek et al. [13], some transcript expression, such as inflammatory genes and antiviral genes, are highly dependent upon signaling from the extracellular environment—not just intracellular transcriptomics. A computational model for such interactions has yet to be proposed.

For the most part, single-cell RNA-seq is not unlike conventional RNA-seq technology when it comes to a breakdown of its signal—except for the fact that tissue samples average out the biological variation found in individual cells. Meaning intuition behind biological noise must come specifically from single-cell RNA-seq investigations. For everything else, we can look primarily to the literature for conventional bulk RNA-seq error models for insight. Let's start our break down with available biological signal.

A classical statistical view of the calculation of RNA-seq transcript abundances via read counting would lead one to conclude that abundances, Y_{ig} , assume Poisson distributions. Poisson distributions are often used for estimating sampling totals where the variance (sampling noise) of the distribution increases as the total (mean) increases. Imagine a process atop a transcriptome, where we chose a location at random to select a read. Each transcript has a probability of being selected that is theoretically proportional to transcript abundance. It so happens that upon repeated sampling the number of reads selected for a given transcript form approximately poisson distribution about the true number of reads present in the sample. This theoretical view is complicated slightly by the existence of transcripts with different abundances. However, the DeSeq2[7] authors suggest that this classical view of biological signal is oversimplified. They claim that, in practice, transcript abundances exhibit more variance than that accounted for by the Poisson distribution. A quick perusing of the potential sources of error in a typical single-cell RNA-Seq pipeline listed in section 2 of this question gives an intuition behind this over-dispersion phenomenon. The authors, decide on a more general negative-binomial distribution which can account for over-dispersion of a Poissonian biological signal by decoupling the variance parameter from the mean parameter.

2. Please list all the sources of noise, outliers, and confounding (and possibly latent) factors.

Sources of biological noise:

Intrinsic biological noise[14]:

- Diffusive molecular dynamics within a cell[15]
- Low copy-number effects of transcription factors or regulatory molecules
- Transcriptional bursting [16]

Extrinsic biological noise[14]:

- Cellular age as discovered in *Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise* [17]
- Environment factors include extracellular signaling pathways
- Organelle distributions, copy number, or other structural phenotypes [18]
- Inheritance noise [19]

Sources of technical noise:

- Process Noise
 - Library preparation
 - Potential batch effects
 - Different human factors
 - Different times
 - Different kits
 - Pipetting errors
 - Molecular variability in cDNA generation
 - Machine noise
 - Potential batch effects
 - lane-to-lane variability
 - Molecular biology of sequencing
 - Single-molecule capture efficiency [20]
 - Amplification bias
 - GC content bias
 - Length bias [21]
 - Analysis Noise
 - Data quality Trimming
 - Alignment Error
 - Normalization artifacts
- Sampling Noise [22]
- global cell-to-cell variation [22]

3. What are the essential and underlying reasons for the sources of noise and how are they characterized?

Consider a pair of isogenic cells. For a given gene, expression levels may vary in spite of the fact that the cells are genetically identical. This variation, in total, can be defined as the biological noise in the system. Biological noise can be broken down further by considering the factors that are globally defined for cells—for instance the temperature of the environment in which the cell resides; copy number and locations of organelles; densities, states, and locations of regulatory molecules within the cells etc... Variance in gene expression linked to variance in these global properties is categorized as extrinsic biological noise. However, if all global cellular properties were held constant between our two aforementioned isogenic cells, there would still be some biological noise in our system. This is due to the molecular Brownian motion in cells and resulting stochastic interactions that happen at the molecular level within a cell. This remaining noise is coined intrinsic biological noise[14].

Biological variation can be exaggerated by noise that affects rare molecules. Many biologically important macromolecules (DNA, transcription factors, etc.) only exist in single cells at very low copy numbers [23]. Biological noise, therefore, can disproportionately affect the downstream transcripts' abundances of these very sparsely represented particles. For instance, consider again a pair of isogenic cells. If a stochastic process (either intrinsic or extrinsic noise) interferes with the reactions of a particular rare molecule in one cell but not a neighboring cell's the two cells' transcriptomic profile may vary drastically.

Similar introduction of noise due to the stochastic interactions of rare molecules is found among the technical sources of variation. Single-cell RNA-seq is known to preferentially detect highly expressed transcripts. Many transcripts only exist in small abundances, however. So this, paired with the fact that single-cells. One of the largest drawbacks of single-cell RNA-seq technology to date is the skew that arises from preferential Single-cell RNA-seq like conventional, bulk RNA-seq technology suffers from many sources of noise. Single-cell RNA-seq, however, further suffers from the exaggerating effect low quantities of genetic material. Recent developments in unique molecule identifiers (UMI) tagging [24] and massively parallelized single-cell RNA-seq technologies like DropSeq [25] have improved on some aspects of error for single-cell. UMI counting can reduce bias due to PCR amplification [13].

PCR amplification biases are due, in part, to the preferential amplification of transcripts exhibiting near 50% GC content. More extreme compositions are disproportionately under represented in the procedure. A number of studies have exposed this bias[26, 27]. One study in particular by Aird et al. [28] showed this by dissecting various stages of the Illumina sequencing library generation pipeline and quantified transcripts at the various library stages with qPCR. The investigators, in conclusion, listed a number of alterations that could be made to the library generation pipeline to minimize the effects of GC content and overall PCR amplification biases. One of the suggestions for instance was a specific temperature ramping regimen between the various cycles of the amplification process. This could be shown to attenuate the bias but not eliminate it. Thereby, this places GC content squarely in the realm of informatics analysis to be dealt with fully. Sufficient use of technical replicates and EECR spike in controls can provide vital information to help isolate technological noise.

4. List various quantitative models of signal that have been reported and what are their deficiencies.

EdgeR[8]:

$$Y_{gi} \sim NB(M_i p_{gj}, \phi_g)$$

$$\phi_g \sim Gamma()$$

DeSeq2[7]:

$$K_{ij} \sim NB(s_{ij} q_{ij}, a_i)$$

$$a_i = \frac{Var(K_{ij}) - s_{ij} q_{ij}}{(s_{ij} q_{ij})^2}$$

Both EdgeR and DeSeq2 are popular models that originate from conventional RNA-seq transcriptomics. Despite their origin, these two packages are still used widely in single-cell analysis. However, their utility is limited to differential expression or relative abundance between groups of samples. EdgeR's choice of dispersion estimation ϕ_g has been criticized for its sensitivity to outliers in conventional RNA-seq data. So, EdgeR, in particular, may not be a wise choice for analyzing single-cell RNA-seq data which often times contain extreme variation and outliers. Similarly, recent studies suggest that DeSeq2 may not be as sensitive or specific as those developed with single-cell RNA-seq sources of variation in mind [29].

scLVM (single-cell Linear Variable Model)[30]:

$$y_g \sim \mathcal{N}\left(\mu_g \mathbf{1}, \sum_{h=1}^H \sigma_{gh}^2 \Sigma_h + v_g^2 \mathbf{I} + \delta_g^2 \mathbf{I}\right)$$

This method requires *a priori* information regarding about functional gene sets. But, if this information is available scLVM is able to infer biological variance contributions from a latent variable that represent effects of those functional groups. Such a decomposition provides investigators the ability to remove inferred signals from single-cell expression data. Signals such as cell-cycle that may confound downstream analysis. The major drawback of this besides the need for gene set membership information, is the assumption of linear Gaussian noise for each of the latent variables. This model is restrictive and likely cannot capture the true components of latent variance in some cases. Inappropriate applications like this will lead to bias in the remaining biological variance.

Spike-in GRM (Gamma Regression Model)[12]:

$$y \sim \text{Gamma}(\mu(y), \varphi)$$

Bo Ding et al. have proposed an excitingly simple and effective way to counter technical sources of variation in single-cell RNA-seq experiments with spike-in controls. They train a gamma regression model to the calculated RNA-seq read abundances (FPKM, RPKM, TMM) versus the known transcript concentration for all spike-in control genes. This helps to counter for technical biases felt dependent upon of how highly expressed a given gene is. The gamma distribution was selected for it's wide range of distribution shapes rather than some biologically driven theory regarding the true underlying distribution of technical variation.

A drawback of this normalization is, standalone, spike-in controls only represent a subset of genes in the transcriptome being sampled. A number of works have commented on the functional relationship between technical noise and transcript length, GC content, and other gene-dependent characteristics. Exhaustive representation of genes is thereby preferable for increased regression accuracy. A universal model would incorporate not only spike-in controls to help attenuate technical noise, but also the consideration of split-pool control samples that would represent all genes being measured. This techniques described thoroughly in Marinov et al.[20]. Furthermore, this type of regression seems to work well for bias due to gene average expression. What about an extension of the regression to correct for, say GC content, biases?

SCDE (single-cell differential expression)[29]:

$$\left\{ \begin{array}{ll} r_1 \approx \text{Poisson}(\lambda_0) & \text{Dropout in } c_1 \\ \left\{ \begin{array}{l} r_1 \approx \text{NB}(r_2) \\ r_2 \approx \text{NB}(r_1) \end{array} \right. & \text{Amplified} \\ r_2 \approx \text{Poisson}(\lambda_0) & \text{Dropout in } c_2 \end{array} \right.$$

Kharchenko et al. propose a Bayesian model that mixes two distinct events and their corresponding distributional contributions to differential expression of a gene between two sets of samples. The paper is concerned with the number of so-called 'dropout' events in single-cell data. An expression value is said to drop out if, by means of biological noise, fails to amplify during sequencing. The expression contributions of a dropout data point are modeled as a Poisson distribution. The authors mention that a constant of 0 expression could have been used in the mixing instead, however the Poisson helps account for background noise still observed in missing (close to

missing) values. A negative binomial distribution is used to model properly amplified expression values.

This model is correct to focus on the prevalent problem of missing data points in single-cell RNA-seq data. The technology owes its' erroneously sparse nature to the extreme PCR amplification biases that come with such small starting genetic material. However, this model doesn't differentiate dropout error due to biologically relevant variance versus technical variance. Perhaps the inclusion of spike-in control regression could aid in that determination.

5. What are the main questions that are sought with the quantitative models? How well have they been answered.

The main questions surrounding quantitative models of expression currently being investigated are those seeking to model technical noise. To this aim, technical noise has been investigated and modeled well. A consensus of models has predominantly been reached for technical noise and sources of technical bias. The most effective techniques to counter technical sources of error have been the inclusion of spike-in controls of known concentrations in RNA samples. This gives rise to a subset of technological noise filtering techniques that estimate abundances by reversing a regression performed observed spike-in quantities [12]. Other works pose questions regarding differential expression of genes across sample groups or clusters. Monocle [31], for instance, is a comprehensive single-cell RNA-seq algorithm that clusters samples, orders samples, and using a dynamic model presumes differential gene status between samples. EdgeR, DeSeq2, and SCDE also weigh in on the differential expression efforts

Biological sources of variation oriented questions require much more investigation when it comes to mathematical modeling. Some questions regarding biological sources of variation for single-cell RNA-seq technology have been posed, however. scLVM attempts to use user-defined *a priori* knowledge about functional gene sets to decompose biological variance into components. There of course are some shortcomings with this methodology as mentioned above, and more work needs to be done in both the biology and analytics.

3. Use-of-hypervariability: Very large variance or hypervariability in expression can be used in differential studies as demonstrated in the following two manuscripts:

1. <http://www.ncbi.nlm.nih.gov/pubmed/23088656>
2. <http://www.ncbi.nlm.nih.gov/pubmed/26078586>

The authors essentially propose the use of anti-profiles. The premise of this method is that tumor phenotypes are best characterized by extreme variability.

Some of your own proposed methods do not take this extreme variability into account. For instance, consider the boolean quantification techniques you wish to borrow and use.

Evaluate the notion of hyper-variability in the context of single-cell measurements. If you believe that this notion has no merit what-so-ever, please provide careful reasoning to support your stand. Otherwise, propose ways to include hypervariability in your methods. Also, how will you extend this idea to co-expression networks. Could you one discuss extreme hypervariability of “functional groups” and “networks”. What are the possible perils and useful outcomes of such an approach?

Note upfront that the ‘hypervariability’ coined by Bravo et al. refers to measurements made from cancer *tissue* samples—not individual malignant cells. Keeping this in mind, there may still be something to say about this analysis methodology in terms of single-cell RNA-seq technology. First, let’s formalize hypervariability as it is presented in the aforementioned publications. A gene is said to be hypervariable in a set of samples compared to some control set of samples, if the ratio of that gene’s standard deviation between those two sets achieves some threshold value. Specifically, the statistic for determining hypervariable gene sets is as follows:

$$r_g = \log_2 \left(\frac{S_{gc}}{S_{gn}} \right)$$

Where S_{gc} is the standard deviation of expression across tumor samples for gene g , and S_{gn} is the analogous measurement for normal tissue. The authors selected all genes with r_g values greater than 1 to be in the so-called ‘anti-profile’ for a given tumor. Then a sample classifier statistic was calculated as the total number of anti-profile genes with expressions beyond 5 standard deviations of the median normal expression value for that gene. The authors then learn a cutoff for this classifier statistic that maximizes the area under a ROC curve when applied to a test set. This classification methodology achieved an impressive area under curve score of > 0.92 .

I believe the notion of hypervariability does indeed have merit in the context of single-cell rna-seq. Knowing what we know about the variance in single-cell data, it is likely that the standard deviation of the single-cell gene expression values underlying the microarray measurements in Bravo et al. were larger than the standard deviation of the microarray values themselves. So, it's no stretch of the imagination to picture this exact same methodology being applied to single-cell RNA-seq data successfully. Different anti-profile genes (no longer obscured by averaging effects) may, however, result. So, yes I think there is merit in applying hypervariability analysis to single-cell transcriptomics.

Before attempt to expand my proposed methodology, we must understand something further. The entire notion of hypervariability, as evident by the equation above, relies on the comparison of variance between two groups. Bravo et al. was only able to identify samples with extreme expression values, by comparing values to the spread of a group to which that sample did not belong. The set of bulk cancer tissue samples' expression distributions were fatter than those for normal samples. The authors exploited this fact to derive a metric for a classifier. Perhaps, a similar group comparison could be incorporated into our proposal.

For inspiration, a quick review of the literature reveals a predominance of work concerned with the change in gene variance due to pathology. From this, I could imagine a version of our gene implication methodology based on hypervariability in single-cell pathogenic datasets. Currently we propose to capture implications of the form: "If gene A is highly expressed, then gene B is highly expressed". What about the implications between hypervariant genes? If gene A is hyperexpressed, can we say anything regarding the hyperexpression of gene B? Here we consider a gene to be hyperexpressed in a sample, if its expression value lays some number, k , of standard deviations from the median of a group of comparison samples.

The next couple of steps are analogous to our proposed gene expression based implication methodology. We can compile a contingency table of hyperexpression between every pair of genes and test for sparse table entries. This will produce a number of implication classes similar to those introduced in our proposed methodology. Implication classes in hand, we can then construct a network of hyperexpression implication between genes for a subpopulation of samples with respect to some control subpopulation of samples.

With the central workflow established we dive into the details and utility of this construction. I can imagine two possible implementations of hyperexpression discretization. I will call one 'within-group' hypervariability and another 'between group' hypervariability. The original paper by Bravo et al. is an example of the latter, because they use the normal range of expression from one group of samples to categorized samples in another group. The within-group approach follows as a categorization of samples using the median and standard deviation

of all samples within that group. This second approach may pose some implicit obstacles, but investigation may still be fruitful.

Let's look at implications that can be made using within-group hypervariability. Single-cell expression is known to be widely variable. Do expressions vary enough to make meaningful implications when the control and comparison groups are derived from the same populations? Quite possibly, Shalek et al. [13] in an investigation of immune response gene pathways using single-cell RNA-seq technology reported instances of extreme variability. Approximately, 1000 fold-change ranges in gene expression were observed in the more extreme instances.

Figure 3A illustrates the major difference in the discretization step between the original proposed methodology and the new discretization step when considering within-group hypervariability. Since both the control and comparison groups will have the same mean values. We essentially label samples as either being 'in' – within a rectangle signifying non-extreme expression values—or 'out'—those exhibiting extreme expression values in a given joint distribution. This is opposed to the original proposal in which fuzzy labels were 'high' and 'low'.

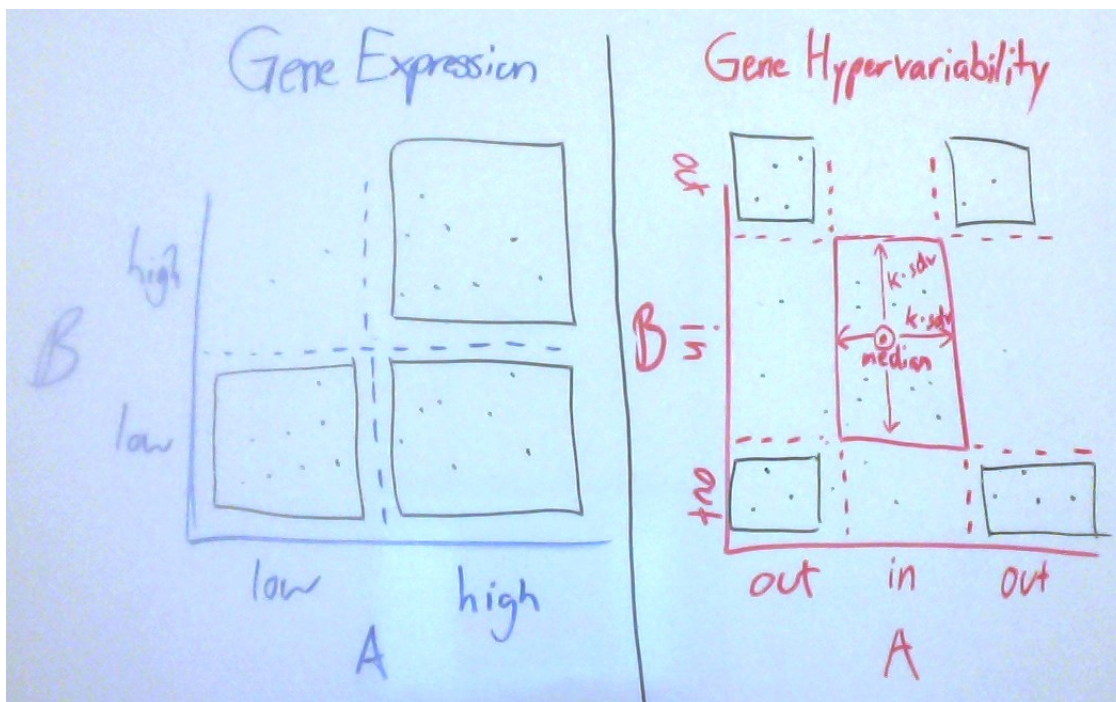


Figure 3A| Comparison of 'high' vs 'low' discretization of gene expression values in original proposal (Blue) versus the discretization of gene hypervariability (Red).

Let's return to the consideration of a pathological sample versus a normal sample of single cells when considering hypervariability between groups. Figure 3B illustrates what can be expected when discretizing between-group hypervariability versus within-group hypervariability. Note that the shift in median and 'in-rectangle' seen in Figure 3B is possible due to the decoupling of the control and comparison groups' distributions. For both approaches, network

construction follows with the same steps mentioned above. This concludes our discussion of the details of both the within-group and between-group hypervariability pipelines.

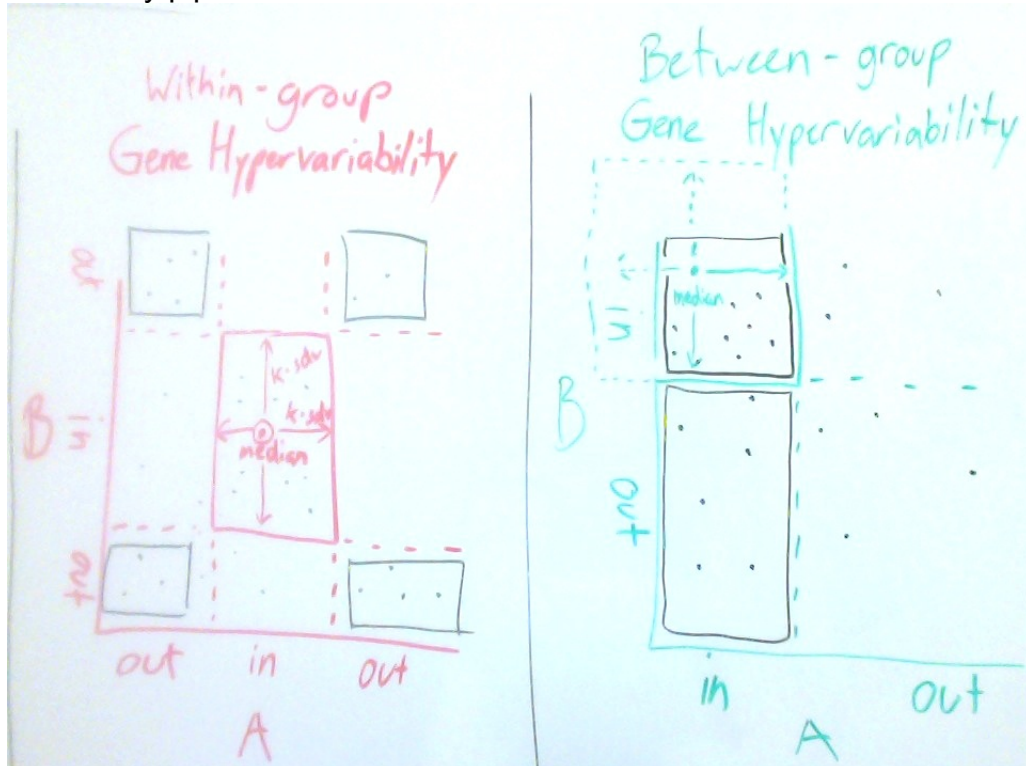



Figure 3B| Comparison of within-group hypervariability (Red) and between-group hypervariability (Green). Note that there is no guarantee that the between-group hypervariability's 'in' rectangle will be centered in the joint distribution.

Currently I foresee sample size being a potential obstacle for implication analysis. Hypervariability for within-group comparisons will, by definition, only exhibit a fraction of 'out' discretized samples relative to the total sample size. This problem can be addressed potentially by adjusting the parameter k , which slides the hypervariability threshold.

Informative biological hypotheses could potentially be drawn from hypervariability implication networks. We could discuss hypervariability in the context of groups of functional genes in general. Again in the context of some pathology, one might expect to be able to mine gene modules from Boolean networks derived from the between-group hypervariability methodology outlined above. An ontology enrichment of such groups may identify causal biological mechanisms underlying extreme trends in variance. Aside from our implication network construction, ontology enrichment of hypervariant functional groups could also be investigated in a standard gene co-expression network. Assuming a control group of samples with which to compare still exists. We can label genes in a co-expression network as either hypervariable or not by making use of a count of hyperexpressed samples per gene. We can then select modules with statistically significant number of hyperexpressed genes, and perform enrichment.

References

- [1] L. JT and P. RD, *Statistics: P values are just the tip of the iceberg.*, Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, USA., pp. --.
- [2] W. S. Noble, "How does multiple testing correction work?," *Nat Biotech*, vol. 27, no. 12, pp. 1135-1137, #dec# 2009.
- [3] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289-300, 1995.
- [4] A. Gelman, J. Hill and M. Yajima, *Why we (usually) don't have to worry about multiple comparisons* , 2008.
- [5] E. B and T. R, *Empirical bayes methods and false discovery rates for microarrays.*, Department of Statistics and Division of Biostatistics, Stanford University, Stanford, California 94305, USA. FAU - Tibshirani, Robert, pp. --.
- [6] T. Ching, S. Huang and L. X. Garmire, "Power analysis and sample size estimation for RNA-Seq differential expression," *RNA*, 2014.
- [7] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, p. R106, 2010.
- [8] M. D. Robinson, D. J. McCarthy and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139-140, #oct# 2009.
- [9] L. G. Halsey, D. Curran-Everett, S. L. Vowler and G. B. Drummond, "The fickle P value generates irreproducible results," *Nat Meth*, vol. 12, no. 3, pp. 179-185, #mar# 2015.
- [10] P. Brennecke, S. Anders, J. K. Kim, A. A. Kolodziejczyk, 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," *Nat Meth*, vol. 10, no. 11, pp. 1093-1095, #nov# 2013.
- [11] A. Oberg, B. Bot, D. Grill, G. Poland and T. Therneau, "Technical and biological variance structure in mRNA-Seq data: life in the real world," *BMC Genomics*, vol. 13, no. 1, p. 304, 2012.
- [12] B. Ding, L. Zheng, Y. Zhu, N. Li, H. Jia, R. Ai, A. Wildberg and W. Wang, "Normalization and noise reduction for single cell RNA-seq experiments," *Bioinformatics*, 2015.
- [13] A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaubomme, N. Yosef, S. Schwartz, B. Fowler, S. Weaver, J. Wang, X. Wang, R. Ding, R. Raychowdhury, N. Friedman, N. Hacohen, H. Park, A. P. May and A. Regev, "Single-cell RNA-seq reveals dynamic paracrine control of cellular variation," *Nature*, vol. 510, no. 7505, pp. 363-369, #jun# 2014.
- [14] E. MB, E. D. Levine AJ FAU Siggia, P. S. Siggia ED FAU Swain and S. PS, *Stochastic gene expression in a single cell.*, Laboratory of Cancer Biology, Center for Studies in Physics and Biology, Rockefeller University, New York, NY 10021, USA. elowitm@rockefeller.edu FAU - Levine, Arnold J, pp. --.
- [15] M. Morelli, R. Allen and P. ReinÂ tenÂ Wolde, "Effects of Macromolecular Crowding on Genetic Networks," *Biophysical Journal*, vol. 101, no. 12, pp. 2882-2891, 2011.
- [16] G. I, S. M. Paulsson J FAU Zawilski, E. C. Zawilski SM FAU Cox and C. EC, *Real-time kinetics of gene activity in individual bacteria.*, Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA. igolding@princeton.edu FAU - Paulsson, Johan, pp. --.
- [17] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi and J. S. Weissman, "Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise," *Nature*, vol. 441, no. 7095, pp. 840-846, #jun# 2006.

- [18] I. G. a. G. B. a. N. R. P. d. a. E. T. a. I. F. J. a. J. N. S. Johnston, "Mitochondrial Variability as a Source of Extrinsic Cellular Noise," *PLoS Comput Biol*, vol. 8, no. 3, p. e1002416, 03 2012.
- [19] D. Huh and J. Paulsson, "Random partitioning of molecules at cell division," *Proceedings of the National Academy of Sciences*, vol. 108, no. 36, pp. 15004-15009, 2011.
- [20] G. K. Marinov, B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers and B. J. Wold, "From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing," *Genome Research*, vol. 24, no. 3, pp. 496-510, 2014.
- [21] D. J and M. M, *Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries.*, Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. jesse_dabney@eva.mpg.de FAU - Meyer, Matthias, pp. --.
- [22] D. Grun, L. Kester and A. van Oudenaarden, "Validation of noise models for single-cell transcriptomics," *Nat Meth*, vol. 11, no. 6, pp. 637-640, #jun# 2014.
- [23] G. P, *Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of Escherichia coli?*, Centre for Cellular and Molecular Biology, Hyderabad, India., pp. --.
- [24] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lonnerberg and S. Linnarsson, "Quantitative single-cell RNA-seq with unique molecular identifiers," *Nat Meth*, vol. 11, no. 2, pp. 163-166, #feb# 2014.
- [25] E. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. Bialas, N. Kamitaki, E. Martersteck, J. Trombetta, D. Weitz, J. Sanes, A. Shalek, A. Regev and S. McCarroll, "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets," *Cell*, vol. 161, no. 5, pp. 1202-1214.
- [26] L. Mamanova, A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure and D. J. Turner, "Target-enrichment strategies for next-generation sequencing," *Nat Meth*, vol. 7, no. 2, pp. 111-118, #feb# 2010.
- [27] L. T. a. L. D. a. R. T. a. C. X. a. T. J. a. M. P. M. a. R. D. a. C. A. M. a. K.-S. C. a. M. C. A. Sam, "A Comparison of Single Molecule and Amplification Based Sequencing of Cancer Transcriptomes," *PLoS ONE*, vol. 6, no. 3, p. e17305, 03 2011.
- [28] D. Aird, M. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. Jaffe, C. Nusbaum and A. Gnirke, "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries," *Genome Biology*, vol. 12, no. 2, p. R18, 2011.
- [29] P. V. Kharchenko, L. Silberstein and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nat Meth*, vol. 11, no. 7, pp. 740-742, #jul# 2014.
- [30] 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," *Nat Biotech*, vol. 33, no. 2, pp. 155-160, #feb# 2015.
- [31] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nat Biotech*, vol. 32, no. 4, pp. 381-386, #apr# 2014.
- [32] L. JT and P. RD, *Statistics: P values are just the tip of the iceberg.*, Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, USA., pp. --.
- [33] A. R. Wu, N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke and S. R. Quake, "Quantitative assessment of single-cell RNA-sequencing methods," *Nat Meth*, vol. 11, no. 1, pp. 41-46, #jan# 2014.
- [34] D. W and B. HC, *Gene Expression Signatures Based on Variability can Robustly Predict Tumor Progression and Prognosis.*, Center for Bioinformatics and Computational Biology, Department of Computer Science and UMIACS, University of Maryland, College Park, MD, USA., pp. --.
- [35] N. R, *Scientific method: statistical errors.*, Gallaudet University in Washington DC., pp. --.
- [36] E. MB, E. D. Levine AJ FAU Siggia, P. S. Siggia ED FAU Swain and S. PS, *Stochastic gene expression*

in a single cell., Laboratory of Cancer Biology, Center for Studies in Physics and Biology, Rockefeller University, New York, NY 10021, USA. elowitm@rockefeller.edu FAU - Levine, Arnold J, pp. --.

- [37] M. I. Love, W. Huber and S. Anders, "Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2," *bioRxiv*, 2014.
- [38] M. Komorowski, J. MiÅ™kisz and M. Stumpf, "Decomposing Noise in Biochemical Signaling Systems Highlights the Role of Protein Degradation," *Biophysical Journal*, vol. 104, no. 8, pp. 1783-1793, 2013.
- [39] L. JT and P. RD, *Statistics: P values are just the tip of the iceberg.*, Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, USA., pp. --.
- [40] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay and I. Amit, "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types," *Science*, vol. 343, no. 6172, pp. 776-779, 2014.
- [41] B. HC, M. Pihur V FAU McCall, R. A. McCall M FAU Irizarry, J. T. Irizarry RA FAU Leek and L. JT, *Gene expression anti-profiles as a basis for accurate universal cancer signatures.*, Department of Computer Science, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA. hcorrada@umiacs.umd.edu FAU - Pihur, VasyI, pp. --.
- [42] W. Dinalankara and H. C. Bravo, "Gene Expression Signatures Based on Variability can Robustly Predict Tumor Progression and Prognosis," *Cancer Informatics*, vol. 14, pp. 71-81, #mar# 2015.
- [43] D. Di Carlo, H. Tse and D. Gossett, "Introduction: Why Analyze Single Cells?," in *Methods in Molecular Biology*, vol. 853, S. LindstrÅ™m and H. Andersson-Svahn, Eds., Humana Press, 2012, pp. 1-10--.
- [44] N. de Souza, "Single-cell methods," *Nat Meth*, vol. 9, no. 1, pp. 35-35, #jan# 2012.
- [45] J. Dabney, M. Meyer and S. PÅäbo, "Ancient DNA Damage," *Cold Spring Harbor Perspectives in Biology*, 2013.
- [46] H. Corrada Bravo, V. Pihur, M. McCall, R. A. Irizarry and J. T. Leek, "Gene expression anti-profiles as a basis for accurate universal cancer signatures," *BMC Bioinformatics*, vol. 13, pp. 272-272, #oct# 2012.
- [47] C. A. Athale and H. Chaudhari, "Population length variability and nucleoid numbers in Escherichia coli," *Bioinformatics*, vol. 27, no. 21, pp. 2944-2948, 2011.
- [48] S. AK, X. Satija R FAU Adiconis, R. S. Adiconis X FAU Gertner, J. T. Gertner RS FAU Gaublomme, R. Gaublomme JT FAU Raychowdhury, S. Raychowdhury R FAU Schwartz, N. Schwartz S FAU Yosef, C. Yosef N FAU Malboeuf, D. Malboeuf C FAU Lu, J. J. Lu D FAU Trombetta, D. Trombetta JJ FAU Gennert, A. Gennert D FAU Gnirke, A. Gnirke A FAU Goren, N. Goren A FAU Hacohen, J. Z. Hacohen N FAU Levin, H. Levin JZ FAU Park, A. Park H FAU Regev and R. A, *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.*, Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA. FAU - Satija, Rahul, pp. --.