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.**
   1. [**http://debunkingdenialism.com/2015/04/01/new-nature-methods-paper-argues-that-p-values-should-be-discarded/**](http://debunkingdenialism.com/2015/04/01/new-nature-methods-paper-argues-that-p-values-should-be-discarded/)**.**
   2. [**http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412**](http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412)
   3. [**http://www.nature.com/news/scientific-method-statistical-errors-1.14700**](http://www.nature.com/news/scientific-method-statistical-errors-1.14700)
   4. [**http://www.nature.com/nmeth/journal/v12/n3/full/nmeth.3288.html**](http://www.nature.com/nmeth/journal/v12/n3/full/nmeth.3288.html)

The comment in Nature made by Leek and Peng in *Pvalues are just the tip of the iceberg* 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. And I do think more rigorous statistical education can make a difference. 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 am, however, intrigued about their suggestions for an analogue of evidence-based medicine—‘evidence-based data analysis’. I think the true enemy here is the arbitrarily chosen threshold for significance. An easy solution is to 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.

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 into account the minimum significance level required for a traditional statistical analysis to conclude significance (AKA p-value). 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.

The high dimensionality of single-cell RNA-seq data, like other ‘omics’ datasets, share in the multiple comparisons problem (http://www.nature.com/nbt/journal/v27/n12/full/nbt1209-1135.html). Counters to this phenomenon arise from the idea of the adjusted p-value. The idea is to shring the p-value 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) (FDR; Benjamini & Hochberg, 1995; Genovese & Wassserman,

2002). (<http://www.stat.wisc.edu/~larget/Stat998/Fall2013/GelmanMultipleComparisons.pdf>) suggest that some Baysian approaches are more appropriate to counter multiple comparison artifacts in cases where classical null hypotheses of the form for some treatment effects are not likely to be true**.** However, (Efron & Tibshirani, 2002) disscuss 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.

As single-cell technology steadily advancing and massively parallel approaches generate ever greater sample sizes, ceteris paribus, statistical power of the investigations is expected to climb as well. Rather than throw out p-values they ought to be augmented in practice along with estimations of statistical power and the assumed effect size of interest. This of course would be the ideal, in practice these metrics are hard to estimate.

The counter argument in terms of single-cell resolution, however, is the innate heterogeneity of the samples—

In my research, I found the guidelines put forth by Robert Abelson in his book *Statistics as Principled Argument* to be rather compelling. He introduces

For instance, differential expression amongst genes in single-cells can be expressed as a fold change or a p-value.

# *Ching et al.* of the Cold Springs Harbor Laboratory in *Power analysis and sample size estimation for RNA-Seq differential expression* 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, EdgeR 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.

# Consider a clustering of single cells. Many works are interested in the heterogeneous populations of cells now that cellular resolution 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 ‘significant’ using a p-value. However, a p-values without the consideration of variance may not be appropriate when trying to characterize rare cell types. In such cases, a confidence interval would be

**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**](http://www.ncbi.nlm.nih.gov/pubmed/22323135)
2. [**http://www.nature.com/nmeth/journal/v9/n1/full/nmeth.1819.html**](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.**

**Observed Signal:**

Where is the biological signal, is the biological noise, is the technical noise, and is normally distributed error.

**Gene Abundance:**

**Biological Signal:**

**Noise:**

for most genes Accounting for technical noise in single-cell RNA-seq experiments). This is consistant with previous works done on bulk signal which is comparible because bulk signals average out biological variance.

This is a catch-all mixture model of both latent variables dependent functions and gene abundance dependent functions, . 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 , or even biological organizational levels latent variables: .

“the strength of technical noise of a given gene depends mainly on the gene's average read count” (Accounting for technical noise in single-cell RNA-seq experiments).

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 RNASeq transcript abundances via read counting would lead one to conclude that abundances assume a Poisson distribution. Poisson distributions are often used for running totals where the variance of the distribution increases as the total (mean) increases. However, the authors of \_\_\_ that introduce the RNASeq package DeSeq suggest that this classical approach 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 Poisson by decoupling the variance parameter from the mean parameter. This decision has been validated in a number of publications (Technical and biological variance structure in mRNA-Seq data: life in the real world,)

For this model we assume no batch affects, as this is ideally a concern of wet lab procedures. And these types of effects are generally trivial to correct with the inclusion of a batch dependent factor when calculating abundances. Other procedural errors that produce variability may be more difficult to counter, especially in the absence of spike-in controls and technical replicates. This is why we further assume best practices, which includes such measures.

In the study, (Validation of noise models for single-cell transcriptomics) investigators aimed to quantify both the technical and biological noise owing from the CEL-seq platform (CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification.) Two sets of cell samples were established. One set was sequenced individually, while the cells in the other were first pooled together and then separated out in order to unify biological variation—a negative control. Both sets of samples were sequenced and analyzed. Researchers found

**Bias + alignmentBias**

I would be remiss to fail to mention that the above signal-to-noise breakdown for a given gene was performed assuming independence. That is to say, that gene expression signal was considered to be independent of the biological signals and noise of other genes. This of course is not the case in reality. (In fact, this is arguably the whole point of my proposal). There exists a network of dependencies that contribute to the observed

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

Sources of biological noise:

Intrinsic biological noise ( [Elowitz et al. (2002)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3118044/#R27)):

* Low copy-number effects (wiki)
* Diffusive cellular dynamics (wiki)
* Noise propagation (wiki)
* Transcriptional bursting (https://www.ncbi.nlm.nih.gov/pubmed/16360033)

Extrinsic biological noise ( [Elowitz et al. (2002)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3118044/#R27)):

* Cellular age (wiki)
* Physical environment (wiki)
* Organelle distributions (wiki)
* Inheritance noise (wiki)

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
  + 2. Machine noise
    - Cluster generation
    - Potential batch effects
      * lane-to-lane variability
    - Molecular biology of sequencing
    - Amplification bias
      * GC content bias
      * Length bias (Jesse Dabney and Matthias Meyer)
  + Analysis Noise
    - Data quality Trimming
    - Alignment Error
    - Normalization artifacts
* Sampling Noise(Validation of noise models for single-cell)
* global cell-to-cell variation (Validation of noise models for single-cell)

1. **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 of cells and resulting stochastic interactions that happen at the molecular level within a cell. This remaining noise is coined intrinsic biological noise (Elowitz, Stochastic Gene expression in Single Cell).

Biological variation can be exaggerated by noise that affect rare molecules. Many biologically important macromolecules (DNA, transcription factors, etc.) only exist in single cells at very low copy numbers (P. Guptasarma, Bioessays). Biological noise, therefor, 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.

“To quantify technical noise, we eliminated biological variability of cellular mRNA abundance by pooling thousands of cells and then splitting them into single-cell equivalents (~20 pg) of RNA (Online Methods). These pool-and-split control samples”( Dominic Grün)

To my knowledge, at the level of single-cell RNA sequencing technologies intrinsic noise is only ever characterized mathematically by the inclusion of a latent variable in a given mixture model. This is most likely due to the fine resolution upon which intrinsic noise is defined.

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 (Quantitative single-cell RNA-seq with unique molecular identifiers). (UMI) tagging and massively parallelized single-cell RNA-seq technologies like DropSeq have improved on some aspects of error for single-cell. UMI counting can reduce bias due to PCR amplification (Shalek, A.K. et al. Nature 498, 236–240 (2013).;

15. Jaitin, D.A. et al. Science 343, 776–779 (2014))

PCR amplification biases are due, in part, to the preferencial amplification of transcripts exhibiting near 50% GC content. More extreme compositions are disproportionally under represented in the procedure. A number of studies have exposed this bias (Mamanova, L., Target-enrichment strategies for next-generation sequencing, and Sam, L.T., A comparison of single molecule and amplification based sequencing of cancer transcriptomes. PLoS One 6:e17305.). One study in particular by Aird et al. 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 in 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.

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

**EdgeR:**

Where is the estimated abundance for gene g in sample i. is the total number of reads in the generated library. is the proportion of the total abundance attributed to gene g in group j to which sample i belongs. is the dispersion parameter for gene g. The major shortcoming of this model is that, although no work has been completed on the matter, probably cannot account for the extreme variance in gene expressions found in single-cell RNA-seq data. EdgeR’s choice of dispersion estimation has been criticized for it’s sensitivity to outliers in conventional RNA-seq data (). And since, EdgeR was developed with bulk samples in mind. It is unlikely that the module will be robust to they additional noisy effects found in single-cell RNA-seq data—like transcriptional bursting or low-copy number artifacts.

**DeSeq2:**

[**https://cofactorgenomics.com/process-noise-rna-seq/**](https://cofactorgenomics.com/process-noise-rna-seq/)

[**http://www.hubrecht.eu/research/vanoudenaarden/publications/NMETH2014b.pdf**](http://www.hubrecht.eu/research/vanoudenaarden/publications/NMETH2014b.pdf)

**scLVM:**

http://www.nature.com/nbt/journal/v33/n2/images/nbt.3102-M2.gif

Accounting for technical noise in single-cell RNA-seq experiments

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 fit 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. 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.

1. **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. Less work has been done on the quantification of biological noise. Presumably, the literature is so imbalance because conventional bulk transcriptomics does not exhibit the degrees of biological variance seen in single-cells. And many of the techniques being used are holdover from conventional RNA-seq. 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 error. The most effective techniques to counter technical sources of variation 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 (GLM gamma spike in model).

Other questions under investigation, seek to deconvolute abundances as a mixture of signals (scLVM).

1. **You are welcome to consult any manuscript to answer this question including these two:**
2. **Quantitative assessment of single-cell RNA-sequencing methods, *Nature Methods* 11, 41–46 (2014) doi:10.1038/nmeth.2694** [**http://www.nature.com/nmeth/journal/v11/n1/full/nmeth.2694.html**](http://www.nature.com/nmeth/journal/v11/n1/full/nmeth.2694.html)
3. **Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells *Nature Biotechnology* 33, 155–160 (2015), doi:10.1038/nbt.3102** [**http://www.nature.com/nbt/journal/v33/n2/full/nbt.3102.html**](http://www.nature.com/nbt/journal/v33/n2/full/nbt.3102.html)

Notes**:**

From Quantitative assessment of single-cell RNA-sequencing methods we find:

relatively low sequencing depth (~2million raw reads)

3’ end bias: coverage drops with distance from 3’ (Ramskold et al)

GC bias

Length bias

Emulsion Methods: DropSeq, InDrop, HiSCL

Single-cell microfluidics C1 recapitulates bulk RNASeq in amalgam.

Genes with low expression tend to have high variability.

From scLVM:

From DeSeq:

From DropSeq:

“DropSeq retains transcripts with high GC content at lower rates.”

Error:

Technical (pcr bias+ gc bias + indicator function for threshold of missing values)+ Biological

Signal:

Organism + Tissue(environment) + cell type + cell state

**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**](http://www.ncbi.nlm.nih.gov/pubmed/23088656)
2. [**http://www.ncbi.nlm.nih.gov/pubmed/26078586**](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 \_\_\_\_ refers to measurements made from cancer ***tissue*** samples—not individual malignant cells. However, there may still be something to say about this analysis methodology in terms of single-cell RNA-seq technology. Let’s first formalize hypervariability as used in the aforementioned publication. The statistic for determining hypervariable gene sets is as follows:

Where is the standard deviation of expression across tumor samples for gene *g,* and is the analogous measurement for normal tissue. The authors selected all genes with 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 > 0.92.

Does hypervariability have merit in the context of single-cell rna-seq? Of course it does. 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 that standard deviation of the microarray values themselves. 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.

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 approach could be incorporated into my own project.

Predominantly the work that has been done regarding variance of genes and gene networks has been concerned with the change in variance due to pathology. I could imagine a version of our gene implication methodology based on this idea in hypervariability in pathogenic datasets. Currently we 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? We consider a gene to be hyperexpressed in a sample, if its expression value lies some number, *k,* of standard deviations from the median of a group of comparison samples.

We can then compile a contingency table of hyperexpression between every pair of genes and test for sparse quadrants. This will produce a number of implication classes similar to those introduced in our proposed methodology. From these implication classes we can then construct a network of hyperexpression implication between genes for a subpopulation of samples with respect to some control subpopulation of samples. I can imagine two possible implimentations of this hypervariability approach. 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 they vary enough to make meaningful implications when the control and comparison groups are derived from the same populations? If so, then perhaps extreme variability amongst genes in single cells isn’t completely a result of biological noise. I currently hypothesize that such an investigation would produce very few implications. But, methodologically it could work. 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’.

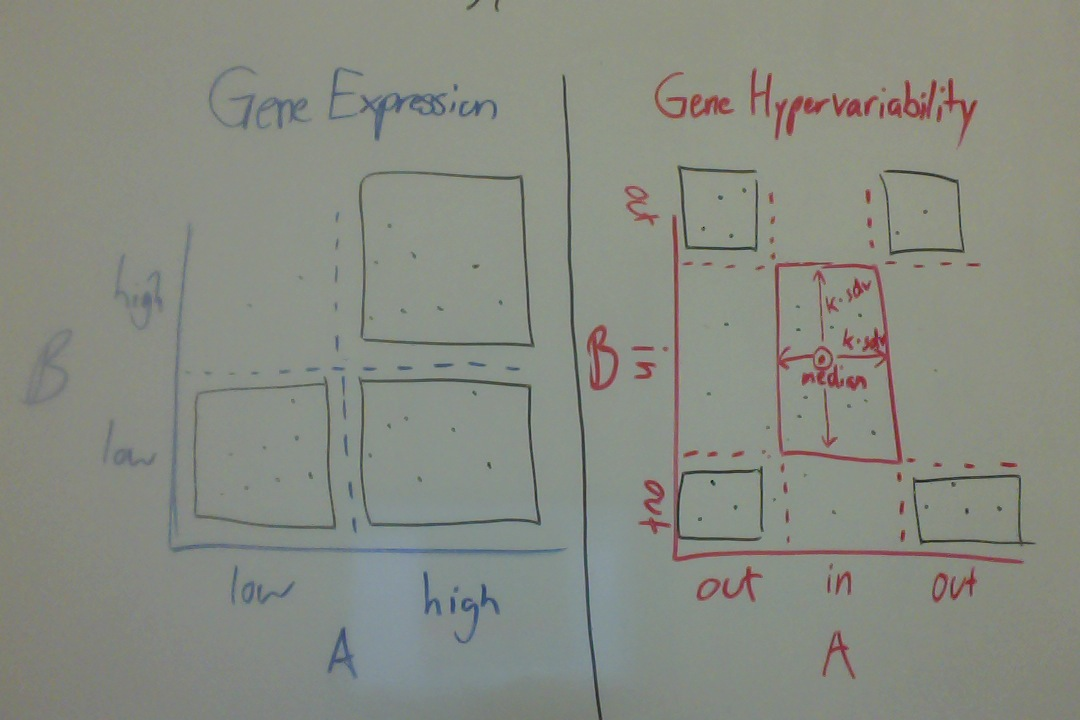


Figure 3A| Comparison of ‘high’ vs ‘low’ discretization of gene expression values in original proposal (Blue) versus the discretization of gene hypervariability (Red).

Hypervariability may only be a characteristic of pathogenic samples though. So, perhaps the between-group hypervariability is a safer/more appropriate implementation. It follows that our control group should, like in Bravo et al., be a set of normal samples, while an implication network is constructed for a pathogenic sample set. Figure 3B illustrates what can be expected when discretizing between-group hypervariability versus within-group hypervariability. This shift in median and ‘in’ rectangle is possible due to the decoupling of the control and comparison groups’ distributions.

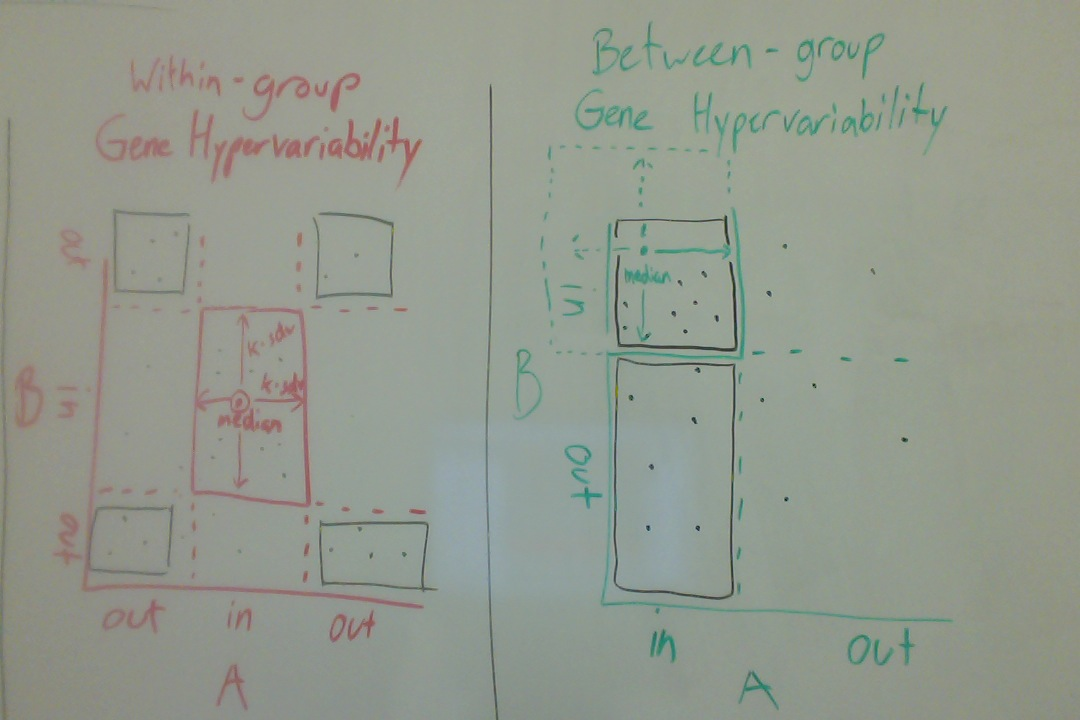


Figure 3B| Comparison of within-group hypervariability (Red) and between-group hypervariability (Green). Note that there is no guarantee that the between-group hypervariability’s ‘in’ rectangle will be centered in the joint distribution.

Furthermore, it may also be interesting to compare the gene expression implication network constructed by our original methodology to that constructed by hypervariability considerations. In a pathogenic context, can all the hypervariable implications be explained by gene expression implications? If not, are any of the genes involved in these unexplained relationships of biological interest given the pathogen? Are there any hypervariable clusters in our network? Another way to address the concern of not witnessing enough hypervariable samples is simple by adjusting our standard deviation threshold k.

Some works such as (Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease) have gathered that variance of gene expression is not randomly distributed about gene signaling pathways. Rather low-variance genes are significantly more connected in gene regulatory networks and high-variance genes are significantly less connected. Perhaps simply overlaying variance in our proposed network visualization can further aid investigators in hypothesis generation.

And when we incorporate Trapnell et al.’s pseudotemporal ordering techniques—not only could we investigate the dynamics of Boolean implication—but also the relative changes in expression through a dynamic biological process.

We could even discuss hypervariability in the context of groups of functional genes. 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 help identify the causal biological mechanism underlying extreme trends in variance.