doi:10.20944/preprints202011.0325.v1

Personalized beyond precision: designing unbiased gold standards to improve single-subject studies of personal genome dynamics from gene products

Samir Rachid Zaim[1,3], Colleen Kenost[1], Hao Helen Zhang[3,5], Yves A. Lussier[1-4,*]

1- Center for Biomedical Informatics & Biostatistics of the University of Arizona Health Sciences

2- Department of Medicine, College of Medicine Tucson

3- Graduate Interdisciplinary Program in Statistics

4- Arizona Cancer Center

5- Department of Mathematics

The University of Arizona, 1230 N. Cherry Ave, Tucson, AZ, 85721, USA

* Corresponding author

Email: yves@email.arizona.edu

Email of all authors:

SRZ: samirrachidzaim@arizona.edu

CK: ckenost@arizona.edu

HHZ: hzhang@math.arizona.edu

YAL: yves@arizona.edu

Abstract

Background: Developing patient-centric baseline standards that enable the detection of clinically significant outlier gene products on a genome-scale remains an unaddressed challenge required for advancing personalized medicine beyond the small pools of subjects implied by "precision medicine". This manuscript proposes a novel approach for reference standard development to evaluate the accuracy of single-subject analyses of metabolomes, proteomes, or transcriptomes. Since distributional assumptions of statistical testing may inadequately model genome dynamics of gene products, the so-called significant results of previous studies may artefactually conflate with real signals. Model confirmation biases escalate when studies use the same analytical methods in the discovery sets and reference standards, as corroboration of results leads to an evaluation of reproducibility confounded with replicated biases rather than a measure of accuracy. We hypothesized that developing method-agnostic reference standards using effect-size and expression-level filtering of results, obtained from multiple discovery methods that are distinct from the one evaluated, would maximize the evaluation of clinical-transcriptomic signals and minimize statistical artefactual biases. We developed and released an R package "***referenceNof1***" to facilitate the construction of robust reference standards.

Results: Since RNA-Seq data analysis methods often rely on binomial and negative binomial assumptions to non-parametric analyses, the differences create statistical noise and make the reference standards method dependent. In our experimental design, the accuracy of 30 distinct combinations of fold changes (**FC**) and expression levels (**EL**) were determined for five types of RNA analyses in two different datasets. This design was applied to two distinct datasets: breast cancer cell lines and a yeast study with isogenic biological replicates in two experimental conditions. In addition, the reference standard (**RS**) comprised all RNA analytical methods with the exception of the method testing accuracy. To mitigate for biased optimization of the RS parameters towards a specific analytical method, similarity between observed results of distinct analytical methods were calculated across all methods (Jaccard Concordance Index). The greatest differences were observed across diametric extremes. For example, filtering out differentially expressed genes (DEGs) using a fold change < 1.2 leads to a 50% increase in concordance between techniques when compared to results with FC > 1.2. Combining this FC cutoff with genes with mean expressions > 30 counts leads to a 65% increase in concordance in comparison to genes with expression levels < 30 counts and with FC < 1.2.

Conclusions: We have demonstrated that comparing accuracies of different single-subject analysis methods for clinical optimization requires a new evaluation framework. Reliable and robust reference standards, independent of the evaluated method, can be obtained under a limited number of parameter combinations: fold change (FC) ranges thresholds, expression level cutoffs, and exclusion of the tested method from the RS development process. When applying anticonservative reference standard frameworks (e.g., using the same method for RS development and for prediction), a majority of the concordant signal between prediction and Gold Standard (**GS**) cannot be confirmed by other methods, which we conclude as biased results. Statistical tests to determine DEGs from a single-subject study generate many biased results that require subsequent filtering for increasing their reliability. Conventional single-subject studies pertain to one or a few measures in one patient over time [1]and need a substantial conceptual framework extension in order to address the tens of thousands of measures in genome-wide analyses of gene products. The proposed ***referenceNof1*** framework addresses some of the inherent challenges in improving transcriptome scale single-subject analyses by providing a robust approach to constructing reference standards.

Github: https://github.com/SamirRachidZaim/referenceNof1

Keywords: single-subject studies, personalized medicine, precision medicine, reference standards, gold standards, biomarkers, open-source

1. Introduction

Reproducibility and accuracy of results are central issues in genome-wide Omics studies, from both biological and statistical standpoints. A 2016 survey by *Nature* [2] indicated that 70% of researchers attempted and failed to replicate studies conducted by other scientists, with more than half failing to replicate their own. While accuracy and reproducibility of an Omics signal in *multi-subject studies* can be assessed by comparing these subjects to comparable subjects in distinct datasets, evaluating the accuracy of a *single-subject study* (**SSS**) remains challenging. In principle, conventional statistics deriving dispersion parameters (e.g., variance) across samples can be applied to single-subject studies using multiple repeated measures in each compared condition (e.g., t-test) or many measures over time (e.g., time series)[3, 4] However this strategy is costly, wastes valuable clinical specimens, and is rate limiting. The foundation for single-subject studies (also referred to as “single-case designs”) [1, 5] dates back to the 1970s highlighting the challenges and issues associated with inferential statistics on cohorts of size N=1. Beyond the multiple repeated measures paradigm of conventional statistics, we and others have proposed new analytical methods designed to identify an effect size and statistical significance for a subject from an Omics sample per condition without replicates [3, 6-9]. A reference standard consisting of the genomes of other subjects is sufficient to qualify the frequency of a genetic variant or mutation in static DNA. However, when this strategy is applied to proteins or transcripts, it does not inform on the nature of the differences observed between an individual's gene products expression and that of a group. Are these differences attributable to a normal physiological adaption or to a pathological response to environmental factors unique to this individual (e.g., combination of medication)? Unfortunately, reference standards for evaluating a sample of a subject in one condition to another paired sample of the same subject in a second condition (e.g., pre- and during treatment) are rare as they require multiple repeated measurements in each condition for one subject. This manuscript addresses this problem and proposes a framework to improve upon the evaluation of software tools and algorithms for differential gene expression in one subject between two sampling conditions, in absence of replicate measures per condition. Such single-subject study Omics designs are more affordable and practical for clinical settings than repeated measures in one condition and generally provide a more interpretable effect size and p-value at a single subject than comparing an individual against a cohort. We contrast and compare this new evaluation framework to previous ones in terms of the accuracy of results beyond the previously proposed "naïve replication" and quantify the biases stemming from the anticonservative assumptions of previous evaluation frameworks.

When it comes to developing reference standards for large-scale biological data science studies, the so-called "*gold standard*" produced via biological validation is rate limiting and generally unfeasible for the entirety of predicted results at the Omics scale. Data scientists address this limitation with computational "*reference standards*" usually as a proxy for conventional biological gold standards. The most rigorous reference standards employ (i*) independent analytics* and (ii) *independent samples (datasets)* from predictions; however, these two conditions are not always feasible when evaluating novel analytical methods designed to analyze single-subject studies. Furthermore, most approaches generating reference standards from an Omics dataset rely overwhelmingly on p-values, despite the increasing recommendations from statistician scholars for effect-size informed approaches to address the limitations of null-hypothesis significance testing [10, 11]. We synthesize and incorporate these notions into a set of standard operating procedures for the development of reliable reference standards, as they build the foundation for evaluations upon which big data science reproducibility studies can be conducted.

Table 1. Current limitations with biased gold standards in transcriptomic gene expression in single-subject studies.

Issue	Description
Statistical assumptions bias	When conditions of applicability (e.g., homoscedasticity assumptions) of the theoretical distribution of the underlying analytics are overlooked and unapplicable, prioritized results contain biases (false positives and false negatives) inherent to modeling inadequacies.
Analytical bias and systematic errors	Studies that use the same analytical method for the prediction calculation as for the reference standard construction incorrectly confirm systematic errors leading to *analytical biases*. For example, creating a reference standard with the same analytical method (isomorphic evaluation) as the one generating predictions can lead to "naive replication" of results comprising both true and false positives (biased systematic artefacts of a specific analytical method). *Isomorphic evaluations* in Omics analyses are anti-conservative by design.
Conflicting biomarker predictions in a single subject	Single-subject studies lack references by design: what happens when analytical method A and analytical method B disagree on a gene's significance? Is gene *x* really significant? There is a lack of accuracy framework for evaluating and resolving conflicting signal stemming from distinct DEG analytics in a single-subject analysis.
Dataset dependency biases	Reusing part of the reference standard data for generating predictions creates dependencies, an evaluation framework problem observed more frequently in statistical evaluation of isogenic data. [11, 12]

This manuscript focuses on improving the accuracy of single-subject studies evaluations, beyond "naïve reproducibility" of results and other biases described in **Table 1**. For example, in a given dataset, various tools may yield drastically different but still plausible algorithmic solutions depending on data distribution assumptions, resulting in technical noise that muddle the actual biological signal. In a prior study of 5 distinct RNA analysis methods in multiple isogenic datasets [12], we described a new method that combines the inconsistent signal between analytical methods that was previously unaddressed in the original study [13]. This inconsistency required methods such as DESeq [14]to impose a false discovery rate (**FDR**) cutoff of 0.001 to detect ~3,000 DEGs, while DEGseq [15] required a cutoff of FDR<3.6x10^{-12} for the same number of DEGs, with 2039 overlapping transcripts. Conversely, we also found that applying the same FDR cutoff (i.e., 0.001) resulted in methods producing various predictions. For instance, one method may produce ~3200 predictions, whereas another ~9000 with approximately 3000 overlapping transcripts, leaving ~6000 transcripts with a conflicting, unaddressed signal. Anticonservative isomorphic evaluations (**Table 1**) have been the conventional standard for evaluating DEG analytics in isogenic conditions (e.g., cell lines or inbred animal models), the closest datasets to single-subject studies [13, 16]. Such evaluations propose a naïve replication of results using the anticonservative assumption that the same DEG analytics can be employed to create the reference standard and the predictions. We have previously created a reference standard from multiple independent methods - without the method evaluated – and shown more conservative accuracy estimates. Specifically, we constructed an ensemble learner [12] to develop reference standards, where the ensemble approach resolves conflicting biomarker prediction, uses no statistical assumptions, and removes anti-conservative isomorphic evaluations. Of note, we avoid using the term gold standard which generally pertains to well-validated biological results and prefer the term reference standard for results constructed through high throughput analytical validations.

The framework is presented in **Figure 1**. We implicitly note that developing a reference standard and using an analytical method to develop a gene-level classifier are two different tasks and therefore separate processes. A reference standard's goal is to approximate the biological signal as best as possible to represent the "truth set", which requires addressing the issues highlighted in **Table 1**. For example, we must optimize the construction of reference standard conservative assumptions of agreement between analytics and avoid the anticonservative *isomorphic evaluations* (defined in **Table 1**). . Our prior study demonstrated that in situations comprising high technical noise, an ensemble learner maximizes the stability of a reference standard and the DEG predictions [12]. However, ensemble learners increase the "black-box" aspect of the data analysis and muddle its interpretability.

In principle, the reference standard should be independent from the predicted biological signal to evaluate an analytical method, requiring independent datasets for calculating and evaluating the prediction. We sought to improve evaluation of single-subject studies of Omics-scale gene products by generating unbiased reference standards. We focused on one framework of single-subject studies: those with two Omics-scale measures (one per condition) in one subject, which are designed to determined altered gene products using a subject as their own control. We hypothesized that these unbiased reference standard could be achieved with two methods: (i) using distinct analytical methods in the reference standard than the one being evaluated to avoid analytical biases, and (ii) selecting the most concordant results between multiple analytical methods as a reference standard according to ranges of fold change expression between two conditions and expression count cutoffs. We propose a framework, ***referenceNof1***, to resolve the challenges highlighted in **Table 1** among different modeling approaches. ***referenceNof1*** offers an alternate, yet related, evaluation framework for single-subject studies comprising explicit criteria for improving data quality and filtering out biases or noise to optimize the reference standard construction. We demonstrate the ***referenceNof1*** method accuracy with transcriptome simulations and historical transcriptome data (Section 2). . Section 3 discusses the implications and limitations of the current approaches, while Section 4 details the data and materials and formally introduces the ***referenceNof1*** algorithm. Section 5 concludes the study. The ***referenceNof1*** software is released as an R package.

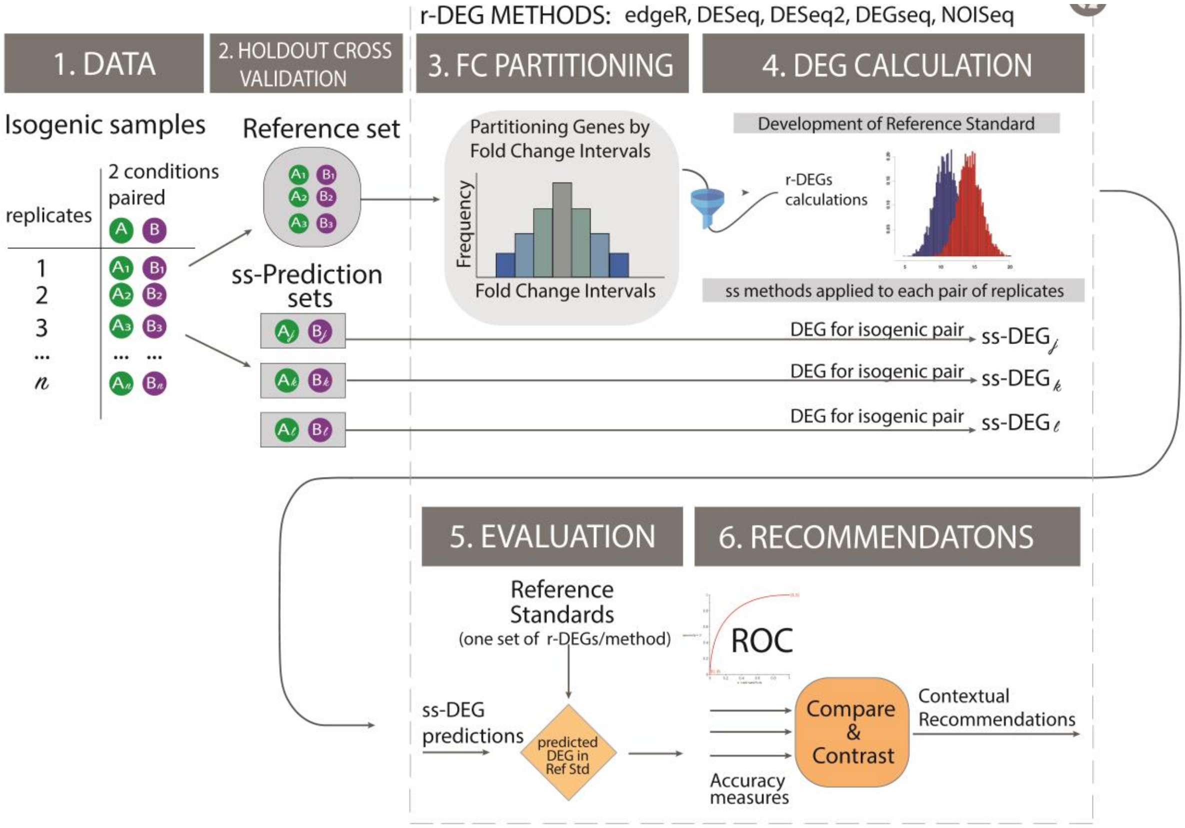

r-DEG METHODS: edgeR, DESeq, DESeq2, DEGseq, NOISeq
1. DATA
2. HOLDOUT CROSS VALIDATION
3. FC PARTITIONING
4. DEG CALCULATION
Isogenic samples
replicates
2 conditions paired
A B
1 2 3 ... n
A1 B1 A2 B2 A3 B3 ... ... An Bn
Reference set
A1 B1 A2 B2 A3 B3
ss-Prediction sets
Aj Bj
Ak Bk
Aℓ Bℓ
Partitioning Genes by Fold Change Intervals
Frequency
Fold Change Intervals
Development of Reference Standard
r-DEGs calculations
ss methods applied to each pair of replicates
DEG for isogenic pair ss-DEGj
DEG for isogenic pair ss-DEGk
DEG for isogenic pair ss-DEGℓ
5. EVALUATION
6. RECOMMENDATONS
Reference Standards
(one set of r-DEGs/method)
ROC
ss-DEG predictions
predicted DEG in Ref Std
Compare & Contrast
Contextual Recommendations
Accuracy measures

Figure 1. Reference Standard Construction Study Overview. In isogenic paired samples from historical cell lines datasets exposed to two conditions (condition A in green and B in purple), we first split the paired stimulus-control data into non-overlapping reference and prediction sets. In order to maximize biological interpretability and relevance, we then run all differentially expressed genes (DEG) calculations and organize the results by fold change regions and conduct all evaluations. We introduce the effect-size analysis into the study to mitigate noisy results (i.e., low p-values with negligible effect sizes) while maximizing biological interpretability.
Notation: ss = single-subject, ss-DEG*i* = differentially expressed genes in a single subject "*i*", conditions: A or B, A_k = gene product expression of gene "*k*" in condition A; B_k = gene product expression of gene "*k*" in condition B.

2. Results

2.1 Fold Change Region Analysis

We utilize a previously designed dataset comprising repeated biological assays and transcriptomes of MCF7 breast cancer cell lines in two conditions (expose to estrogen or deprived) to derive a reference standard in which we could evaluate analytical methods deriving DEGs from a mere 1 sample in each condition [3, 12, 17]. Concordances between conventional RNA-seq analytical methods for repeated measurements (Methods) applied to this MCF7 dataset were calculated and shown in **Table 2**. The calculated differentially expressed genes (**DEG**s) are filtered at different fold changes (**FC**). In low fold change regions, the agreement between analytical methods is at best 50%, providing low trust across the competing signal detected across analytical methods. That is, analytical methods are not validating each other's findings. However, as the FC effect size increases, the increasing FC regions demonstrate greater agreement across analytical methods, resulting in FC conditions enabling the evaluation of one method against another and thus devoid of isomorphic evaluation type of analytical biases. **Table 2** numerically illustrates this trend, noting that the Jaccard Indices (JIs) for the first pair of fold change regions is mostly near or at zero, except for a few combinations. This means that in absence of a framework or analytical method for resolving this conflict of signal, any study that chooses a given analytical method risks to detect a non-robust signal beyond naïve reproducibility. At higher fold change regions above 1.2, there is no one "best" fold change region, supporting the idea that filtering $FC \geq 1.3$ may provide the most reproducible results from method to method for this particular dataset.

2.2 Combining Fold Change and Low-expression Noise Reduction in Reference Standards

Combining gene expression levels (minimum expression cutoff; **Figs. 2** and **3**) and effect-size prefiltering results in a quasi-linear improvement in agreement as we imposed more rigid thresholding. The experimental design allowed us to examine the marginal effects of increasing the cutoffs for gene expression and effect size (see **Figure 2** for Breast Cancer and Figure 3 in Yeast).

In the breast cancer study, the reference standards were constructed using 4 replicates. The bottom right portion of concordances in **Figure 2** illustrates how all the analytical methods attained strong concordances due to the strictest thresholds (at least > 75% of all identified DEGs).

In the yeast study, smaller effect size thresholds were required for analytical methods to have complete agreement (see **Figure. 3**). The methods appear to agree and produce concordant reference standards after imposing moderate (at least on our scale of parameters) fold change and expression cutoff values. It suggests perhaps that one might benefit from developing a self-learning algorithm that finds the optimal cutoff values after conducting a grid search on the parameter space.

Table 2. Concordance between Analytic Methods of RNA sequencing According to Ranges of Gene Expression Fold Changes (FC) Between Two Conditions. The low concordance observed in most FC ranges illustrate the "*analytical bias of methods*" described in Table 1. Indeed, if the same method is used for prediction in one dataset and validation in a distinct dataset (*isomorphic evaluation*), the evaluation is considered anticonservative as it measures the reproducibility of true positive and false positive results (*analytical biases*) rather than a measure of accuracy. In addition, the table results illustrate the difficulty to create a conservative reference standard for which the analytical method would be independent from the predictive method (*heteromorphic evaluation*); there is no single method that would be the best choice "*a priori*" to evaluate a new method. **Legend**: Since the Jaccard Index is symmetric, for any two techniques, we present the Jaccard Indices for the $\binom{5}{2} = 10$ total possible pairwise combinations between the five analytical methods evaluated across the different fold change regions. DEGs were calculated using 5 repeated samples of MCF7 breast cancer cell lines exposed to estrogen and 5 unexposed samples. The high concordance for each pairwise comparison is bolded (JI>0.6) and shown in a larger font.

Analytical Method A	**Analytical method B**	**1<FC <1.1** **(~85 DEGs)**	**1.1<FC<1.2** **(~175 DEGs)**	**1.2<FC<1.3** **(~700 DEGs)**	**1.3<FC<1.5** **(~1100 DEGs)**	**1.5<FC<∞** **(~365 DEGs)**
NOISeq	edgeR	0.500	0.333	**0.885**	**0.819**	**0.631**
NOISeq	DESeq	0	0	**0.814**	**0.747**	0.586
NOISeq	DESeq2	0.005	0.002	0.311	0.436	0.372
NOISeq	DEGseq	0	0	0.355	0.569	**0.672**
edgeR	DESeq	0	0	**0.902**	**0.868**	**0.795**
edgeR	DESeq2	0.002	0	0.329	0.468	0.515
edgeR	DEGseq	0	0	0.387	0.558	**0.658**
DESeq	DESeq2	0	0.076	0.332	0.457	0.489
DESeq	DEGseq	0	0.285	0.415	0.555	**0.661**
DESeq2	DEGseq	0.005	0.135	0.452	**0.654**	0.450

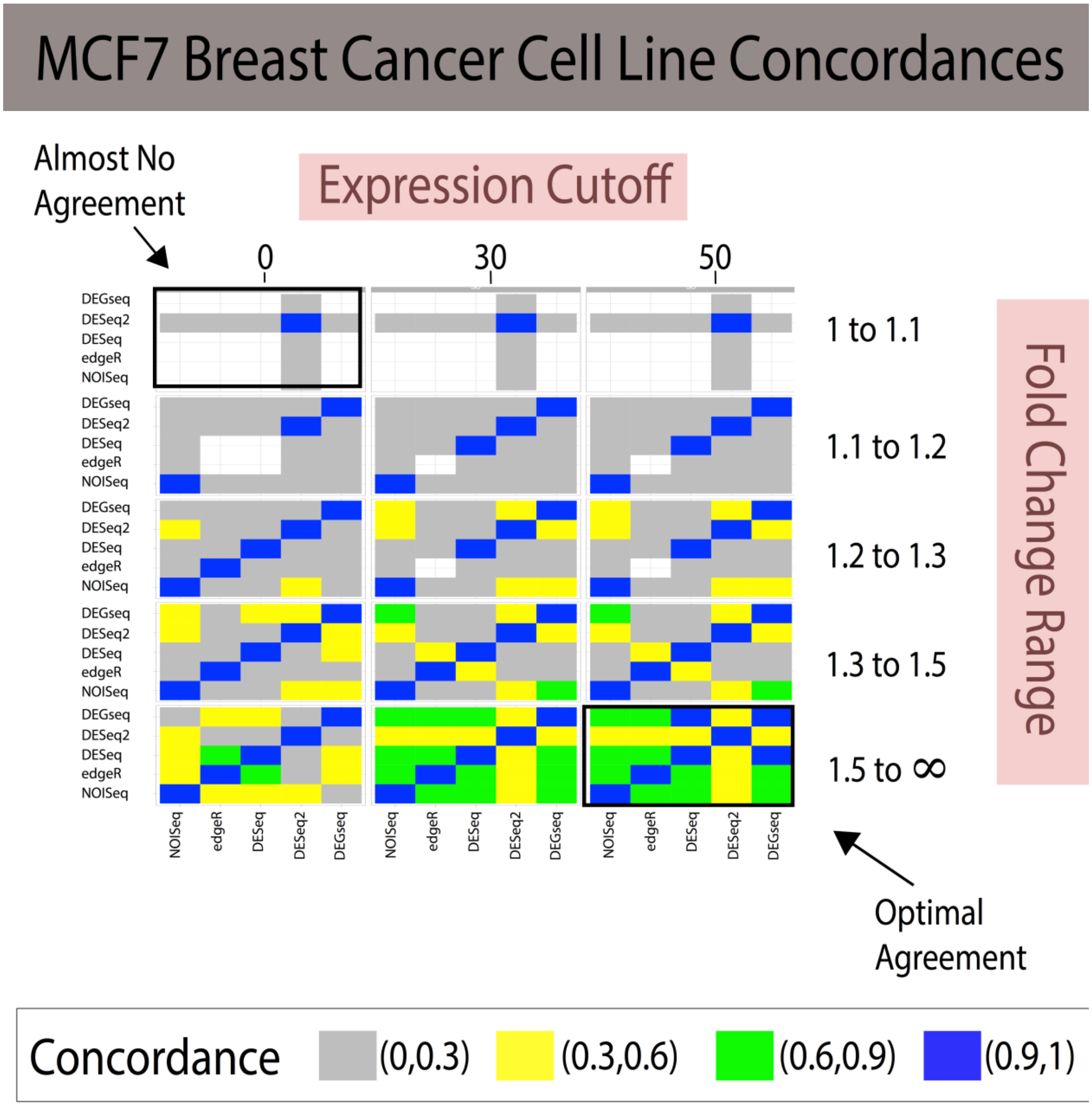


Figure 2. Combining fold change and expression level filtering leads to robust, method-agnostic reference standards for single-subject studies in breast cancer. The grid of heatmaps illustrates that as low-expression genes with small fold changes across individuals are filtered out, the reference standards constructed agree increasingly more with one another. In the bottom right, most methods attain a 100% concordance with one another, providing a reliable gold standard. Expression Cutoffs are applied to genes whose average counts across samples fall under the threshold. White cells indicate that no predictions were made, and therefore the Jaccard Index cannot be calculated. Note: when the JI cannot be calculated due to the lack of transcripts, the color of the rectangle is white; in addition, FC ranges are symmetric 1/1.1 to 1 and 1 to 1.1, 1/1.2 to 1 and 1 to 1.2, etc.

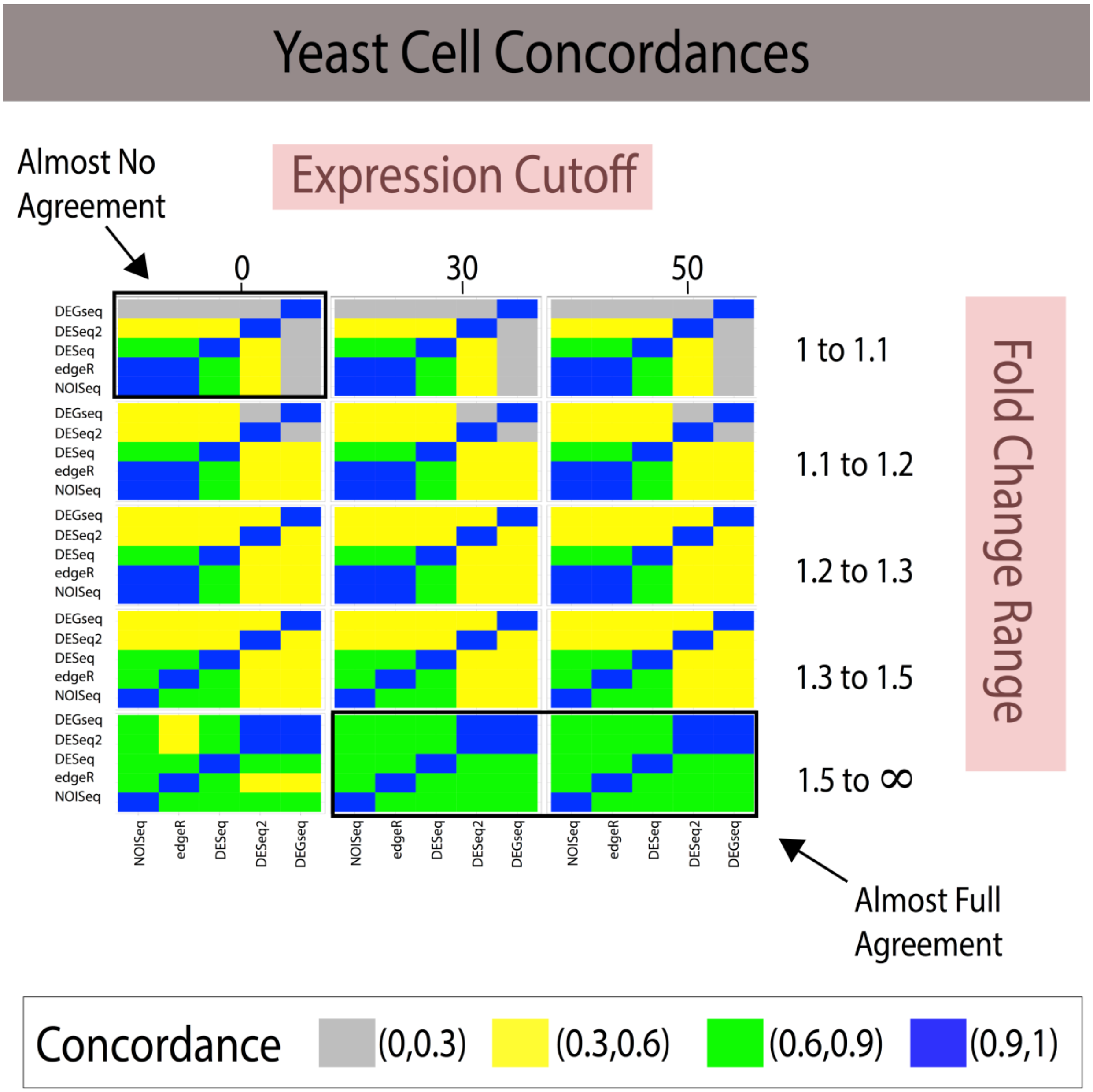


Figure 3. Combining fold change and expression level filtering leads to robust, method-agnostic reference standards for single-subject studies in yeast. The grid of heatmaps illustrates that as low-expression genes with small fold changes across individuals are filtered out, the reference standards constructed provide higher concordance with one another. Given the larger number of replicates in yeast, it may be that less rigid filters are required to produce reliable, concordant reference standards. Expression Cutoffs in average counts across samples. White cells indicate that no predictions were made, and therefore the Jaccard Index cannot be calculated. Note: when the JI cannot be calculated due to the lack of transcripts, the color of the rectangle is white; in addition, FC ranges are symmetrical 1/1.1 to 1 and 1-1.1, 1/1.2-1 and 1-1.2, etc.

2.3 DESeq example analysis with robust reference standards

As shown in **Figure 1**, we have calculated single-subject DEGs (ss-DEGs) from two samples (one in each condition, without replicate) using MCF breast cancer cells exposed to estrogens or not and evaluated them against either a reference standard constructed from the intersection of DESeq2, edgeR, NOISeq, and DEGseq (**Table 3 , top row**) or using the optimized reference standard using the proposed ***referenceNof1*** method (**Table 3, bottom row**). This experimental design using isogenic cells lines with replicates in two conditions enables validation of single-subject transcriptome analysis methods. We have previously documented that some DEGs methods designed for comparisons requiring repeated measures of isogenic samples or many heterogeneric samples for each condition claim to be applicable to single-subject analyses of transcriptome methods, however none had documented their validation [3]. Subsequently, Schurch et al. [16] and Liu et al. [13] validated NOISeq, DEGseq, DESeq2, and EdgeR using replicates of isogenic samples in two conditions conducted in MCF7 and yeast data, respectively; however, their evaluations were conducted using anticonservative designs generating analytical biases due to isomorphic evaluations (**Table 1**) as shown in our recent study [12]. Here, we conduct an evaluation that addresses these previous limitations. The results in **Table 3** indicate a substantial increase in accuracy in single-subject studies using single-subject DESeq taking one sample in each of the two MCF7 cell lines conditions without replication (average recall and precision shown as we obtained three measures of each for every reference standard). As shown in **Figure 1**, we took three distinct pairs of estrogen exposed and unexposed samples and calculated for each pair the ss-DEGs using DESeq as an exemplar method being applied to a clinical sample. Prior studies using DESeq in single-subject studies indicate a conservative prediction approach, producing few but highly precise DEG calls [12, 17, 18]. These results indicate a consistent operational characteristic as well as an improved region of algorithm accuracy.

Table 3. Single-subject DEGs predictions evaluated by conventional methods and *refereneNof1*. In order to simulate transcriptomic data from a single patient, single-subject DEGs were calculated by the DESeq method from two samples (MCF7 cell vs MCF7 exposed to estrogen) and evaluated against two reference standards derived from Y samples in each condition (2Y samples total). The two reference standards constructed in this exemplary study illustrate the increases in accuracy provided by the proposed ***referenceNof1*** method to increase the agreement between DEGs methods used for as a reference and mitigate analytical biases from isomorphic evaluations. The optimal region identified by the ***referenceNof1*** *algorithm* resulted in a DESeq prediction set with a substantially higher precision with a slightly higher recall.

		Predictions of ss-DEGs calculate by DESeq	
Reference Standard Construction	**References** True DEGs (*False= remaining transcripts*)	**Average Precision**	**Average Recall**
Intersection of DEGs between methods*	522 (*16,625*)	0.57	0.08
Majority vote of DEGs between methods*	1424 (*15,723*)	0.77	0.04
referenceNof1 applied to intersection of DEGs* between methods	165 (*16,982*)	0.70	0.12
referenceNof1 applied to majority vote of DEGs* between methods	406 (*16741*)	0.85	0.06

* calculated by DESeq2, EdgeR, NOISeq, DEGseq NOT using DESeq to avoid analytical biases

3 Discussion, Limitations, and Future Studies

We and others have proposed that while identifying altered DNA is possible using reference standards derived from populations, determining altered gene products (e.g., transcriptomes, proteomes and metabolomes) are better determined in isogenic conditions[3]. Indeed, identical twins sharing the same DNA but living in diametrically different environments (polar vs artic) and having different sites, sleeping and exercise regimen may have totally adapted and normal, yet quite distinct, gene products in their cells. This motivated us and others to design analytical methods to determine personalized differentially expressed genes (DEGs) from two samples without replicate, each taken in a different condition. These approaches have been applied to (i) comparing a cancer transcriptome to a paired control tissue [6], and (ii) comparing peripheral blood mononucleocytes of a single subject either taken in two conditions (e.g., before and during therapy to predict response[19], separated in two petri dishes with one experimentally exposed to a virus vs a control to determine ulterior hospitalizations in pediatric asthmatic subjects [19, 20], etc.). While these single-subject study designs are economical and more informative than a single measure of the transcriptome, they remain difficult to evaluate. We have previously shown that these single-subject DEGs analytics designed for two paired samples can be better evaluated by using previous cell lines datasets comprising multiple replicates in each of the two conditions. However, most conventional methods generated a conflicting discordant reference standard, motivating the current study.

The task of building a robust and reproducible reference standard should not be confounded with identifying/predicting DEGs for a gene-expression classifier. Since gold standards on an Omics scale are truly only available in simulation studies, we propose to call standards derived from Omics-scale analyses of biologic datasets reference standards. Therefore, a novel framework is required to address the difficulties with generating a reliable reference standard as highlighted in **Table 1**. The proposed framework provides an alternative reference standard construction that is robust against violations of statistical assumptions, resolves competing signal across analytical methods, and produces accuracy beyond naïve reproducibility.

The modus operandi of developing a reference standard consisted of using a technique to build a reference standard (i.e, DEGseq) and measuring it against itself via some criterion (i.e., AUC). This provides naïve reproducibility that does not generalize across methods. Our proposed framework, ***referenceNof1***, generates an unbiased and robust reference standard using concordance between heterogenic methods. We note how filtering out noise (i.e., genes with small fold change) can have a drastic effect for improving on how to build reference standards as well as provides a framework that quantifies these differences and guidance on how to construct the refence standard for a biomarker analysis independent of the used technique. In addition, filtering out genes with low expression [21-23] improves the power and removes the noise in bioinformatics, therefore ensuring that uniform cutoffs across techniques improves their concordance, thus increasing their reliability.

Constructing methods-agnostic reference standards will only enable the community to continue improving the state of reproducibility in bioinformatics data analysis. Clearly, the solution range is dataset specific as shown by the different concordances of the same methods applied to two different datasets (**Figures 2** and **3**). This study suggests that a “stratification data analysis model” could be applied to determine the optimal gene expression cutoff required to obtain the desired concordance for each range of fold changes and the minimal fold change required to include results in the reference standard. For example, if a concordance of JI>75% for a simple majority of DEG methods is considered sufficient generating a reference standard, then **Figure 2** shows that an expression>30 and FC>1.3 meet these criteria. If the criteria were reduced to JI>50% and simple majority vote, then the reference standard comprises the DEGs discovered at the union of [expression cutoff>0 and 1.3<FC<1.5] and [expression cutoff>30 and FC>1.5]. Reference standards built for a single subject are isogenic by design. Since most of the publicly available human transcriptome datasets comprise of measures for multiple subjects (heterogenic), we conducted the validation of our proposed methods using multiple measures in two experimental isogenic conditions in cell lines as proxy for a single subject study.

As shown in **Figures 2** and **3**, stricter cutoffs resulted in higher concordances in the breast cancer cell line dataset, whereas in the yeast dataset, the concordances followed more of a parabola, with initial increases and then a decrease in concordances. This suggests that there is not necessarily a universal effect-size “cutoff” after which methods agree, but rather data-driven fold change (FC) regions in which the agreements are maximized. Since these FC regions have to be identified on each dataset, we posit that a self-learning algorithm can be implemented to identify high-agreement regions. Preliminary work in ***referenceNof1*** has been extended in this direction with future studies focused on fine-tuning and refining the approach to having a self-learning algorithm optimize the operating space to better guarantee biological validity for the statistical results. Currently, the search is designed to identify the minimal set of operating parameters to attain a certain Jaccard Index concordance using the intersection of techniques. Future work will expand to include a “majority-voting” rule and generalize the voting scheme to allow greater flexibility.

Furthermore, as shown in [24], biological signal in heterogenic subjects can be unstable and expressed inconsistently across subjects, suggesting that alternatively it might be more effective to conduct pathway-level analyses and train pathway-level classifiers. Therefore, another direction to extend this strategy is to improve and refine the state of pathway-level reference standards by incorporating ontologies like Gene Ontology [25] and other network-analysis tools into the process of building a robust and reproducible reference standard.

Finally, the evaluation framework, ***referenceNof1***, can be extended in future studies to single-subject studies of biomolecular pathway dynamics. We and others have previously demonstrated the utility of single-subject analytics of gene products at Omic scale. For example, we have shown how to compare treated vs untreated peripheral blood mononucleocytes (PBMCs) using single-subject transcriptome analyses designs [26], as well as contrasting experimentally stimulated vs unstimulated PBMCs of a subject *ex vivo* using rhinovirus to predict hospitalization in asthmatic subjects [27], or comparing cancer vs adjacent control tissue of a subject pr predicting response to therapy[6].

4 Methods & Materials

The study design is illustrated in **Figure 1**, and the following subsections detail the datasets and materials used throughout the study. To improve the state of the art in building reference standards, the study is designed around isogenic datasets to address the issues highlighted in **Table 1**.

4.1 Datasets

In this study, two different isogenic datasets with biological replicates were used to evaluate the construction of reference standards, shown in **Table 4.**

Table 4. The two isogenic datasets include a single individual's gene expression dataset with 7 biological replicates while the second dataset has 48 wild-type and mutant biological replicates.

Dataset	Samples	Genome size (# genes)	Access to Data
MCF7 Breast Cancer [13].	7	~ 20,000	GEO: GSE51403
Yeast [16]	48	~ 7,000	Github: bartongroup/profDGE48

The first is an MCF7 breast cancer dataset that contains replicated gene expression data in isogenic conditions with 7 human biological replicates of MCF7 cells which were either treated with 10 nM 17β-estradiol (E2) or cultured as unstimulated controls [13]. The data contains replicates at various read depths with all analyses being conducted using the 30M read replicates, which are available open source online under the Gene Expression Omnibus repository [28] under the "GSE51403" GEO tag. Normalized and preprocessed data were downloaded on January 21, 2018. The data was used as obtained with no additional pre-processing steps to conduct the reproducibility analyses (preprocessing details and correction details can be found in the original publication) [13], and we randomly selected four biological replicates ("565-

doi:10.20944/preprints202011.0325.v1

576","564-572","566-570","562-574") to construct the reference standard, while the remaining three ("563-577","568-575","569-571") were used to evaluate how well the DEG methods recapture the signal under different reference standard settings.

The second dataset [16] is a yeast study with biological replicates comprised of 48 wild-type (BY4741 strain, WT) or **Δ**snf2 mutant biological yeast replicates (*Saccharomyces cerevisiae)*. These include a total of 7,126 measured genes, and we replicated followed the author's data preprocessing framework and conducted our studies using their suggested 42 WT and 44 **Δ**snf2 'clean' replicates. The preprocessed and normalized data were downloaded as prepared by the original authors from their GitHub[1] repository.

4.2 Software Environment

All analyses in this study were conducted in the R programming language, using R 3.5.0[29] on a 2017 MacBook Pro under the macOS High Sierra (10.13.6) OS system.

All code and analyses are freely available at

http://www.lussiergroup.org/publications/EffectSize_ReferenceStandard

4.3 Differential Expression Software Tools

To evaluate the robustness and reproducibility of various differential gene expression analyses techniques, we evaluated 5 different cohort-based (cb) RNA-seq tools, found in **Table 1** [15, 30-33]. Since the study was designed to evaluate the robustness of reference standards, we omitted the single-subject analytics included in our prior study (i.e., Mixture Models [7] and iDEG [18]), as these methods are designed to predict DEGs in isogenic settings, two conditions without replicates (**TWCR**) design, rather than produce reference standards in replicated settings. Since this study required techniques with p-value-based decision-making, we omitted GFOLD [34] and similar ranking-based methods in order to establish consistent cutoffs in our experimental design. **Table 5** provides the individual parameter settings for each method.

[1] Data downloaded from on August 27th, 2018. https://github.com/bartongroup/profDGE48

Table 5. Replicated DEG Methods for Single-Subject Studies and their previous validations.

Method	Distribution assumptions	Experimental design & parameter settings	P-value
edgeR[30]	Negative Binomial	"genetically identical model organisms"	✓
DESeq[31]	Negative Binomial	*estimateDispersions* function, the method parameter is set to 'per-condition'	✓
DESeq2[32]	Negative Binomial	Default parameters	✓
DEGseq[15]	Binomial	Default parameters	✓
NOISeq[2] [33]	Non-parametric	Noiseqbio set to default parameters	✓

4.4 Building Effect-size-informed Reference Standards

We hypothesize that low effect sizes (i.e., low-fold change) introduces statistical noise into the reference standard construction in isogenic conditions, which can introduce biases when using only p-value informed DEGs. Therefore, in order to test this hypothesis, we first construct a reference standard for each method using all the data, and then degrade the dataset by filtering out genes with effect sizes, in an increasing fashion and evaluate the strength of the agreement across them. Thus, if we use fold change (FC) as a proxy for effect size (as calculated by equation 1),

$$\text{Fold Change of gene } k = A_k/B_k \qquad \textbf{Equation (1)}$$

where A_k and B_k are the expression of gene product k in condition A and B **(Figure 1).** The experimental design was comprised of constructing the reference standard across different levels of fold change (FC) and evaluating the concordance as a consequence of the effect size filter. Since we do not distinguish between up and down regulated genes, for down-regulated genes, we take their reciprocal, ($FC^{-1} = 1/FC$) when we filter for FC thresholds.

[2] NOISeq-Bio was used to construct the reference standard, while NOISeq-sim was used in the single-subject prediction sets.

4.5 Low Expression Pre-filtering

We also hypothesize that low gene expression introduces statistical instability in the task of identifying differentially expressed genes. A common preprocessing approach in differential gene expression is prefiltering genes [21-23] with low expression as this may increase the power of the subsequent statistical test, and we extend this work into precision medicine by examining the effects of gene pre-filtering in constructing reference standards in isogenic conditions.

4.6 Experimental Design

To evaluate the power of combining prefiltering genes based on their effect size (fold change) and their expression level, we considered an array of possible low-expression cutoffs and fold-change regions (see **Table 6**) and selected genes in these windows.

Table 6. Parameter Settings in Experimental Design.

Parameter	**Values**
Fold change window	[1 -1.1], [1.1-1.2], [1.2-1.3], [1.3-1.5], [1.5 - ∞]
Low expression cutoff	0, 5, 10, 20, 30,50

We then used these selected genes to construct reference standards for all the methods presented in **Table 5** and evaluated their concordances. To evaluate the results of the different experimental runs, heatmaps were used to visually compare their concordances across different experimental runs (see **Figures 2 and 3**) and the Jaccard Index to numerically evaluate the agreement between them, where the Jaccard Index is given by **Equation 2**.

$$\text{Jaccard Index (JI)} = \frac{|M \cap N|}{|M \cup N|} \qquad \textbf{Equation (2)}$$

In our study, these metrics translate to the similarity and dissimilarity between the DEG calls between method A and method B (i.e., between edgeR and DESeq), which quantifies the biological signal reproducibility between analytical approaches in our reference standard construction study.

4.7 Optimization of a Reference Standard using maximum Jaccard Index Concordance

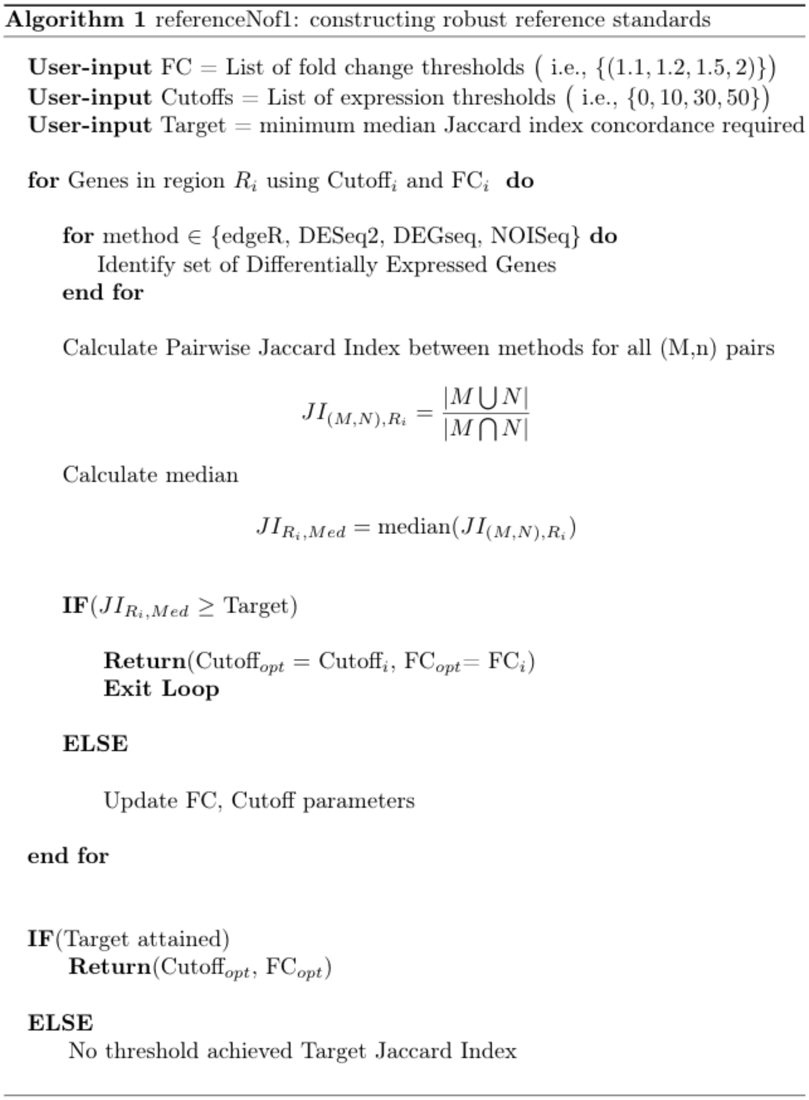
The parameters in the experimental design were constructed to identify regions in which the Jaccard Index was maximized. In this grid search, each set of parameter combinations results in a Concordance Matrix of Jaccard distances (**Table 7**).

Table 7. Example of building a Jaccard Index matrix to optimize the Reference Standard Optimization. An example Jaccard Index matrix where once all techniques have produced a list of DEG calls, then the pairwise Jaccard Index (**Equation 3**) can be calculated as illustrated in Algorithm 1, resulting in a Jaccard Index matrix below. Then, for each set of parameter combinations (one expression cutoff value, one fold-change region), a Jaccard concordance matrix is constructed, and summarized using the median value. Then all medians are compared to identify the best parameter configuration to identify the region for constructing the optimal, most robust reference standard.

	NOISeq	edgeR	DESeq	DESeq2	DEGseq
NOISeq	1	0.8	0.71	0.2	0.17
edgeR	0.8	1	0.57	0.21	0.2
DESeq	0.71	0.57	1	0.25	0.29
DESeq2	0.2	0.21	0.25	1	0.15
DEGseq	0.17	0.2	0.29	0.15	1

From this Jaccard matrix, each method’s median JI can be calculated (i.e., NOISeq’s median JI is 0.71 while DEGseq’s is 0.21). Using this information, one can summarize the agreement either by method to see which methods agree more with one another and which differ. To construct the most robust reference standard, one needs to identify the optimal parameter combination (fold change and expression-level thresholds) to maximize the Jaccard Index across all reference standards. Therefore, we constructed an R package, ***referenceNof1***, to do exactly this, enabling bioinformaticians to then construct the optimal reference standard. If a user inputs a vector of effect size windows, a vector of minimum value expression cutoffs, and a desired level of concordance, the ***referenceNof1*** calculates all pair-wise Jaccard similarity indices for all parameter combinations and identifies the minimum parameter combination that attains the desired pairwise concordance across all techniques. This algorithm is formalized in **Algorithm 1**:

doi:10.20944/preprints202011.0325.v1



```
Algorithm 1 referenceNof1: constructing robust reference standards

User-input FC = List of fold change thresholds ( i.e., {(1.1, 1.2, 1.5, 2)})
User-input Cutoffs = List of expression thresholds ( i.e., {0, 10, 30, 50})
User-input Target = minimum median Jaccard index concordance required

for Genes in region $R_i$ using Cutoff$_i$ and FC$_i$ do

    for method ∈ {edgeR, DESeq2, DEGseq, NOISeq} do
        Identify set of Differentially Expressed Genes
    end for

    Calculate Pairwise Jaccard Index between methods for all (M,n) pairs

        $JI_{(M,N),R_i} = \frac{|M \bigcup N|}{|M \bigcap N|}$

    Calculate median

        $JI_{R_i,Med} = \text{median}(JI_{(M,N),R_i})$

    IF($JI_{R_i,Med} \geq$ Target)

        Return(Cutoff$_{opt}$ = Cutoff$_i$, FC$_{opt}$= FC$_i$)
        Exit Loop

    ELSE

        Update FC, Cutoff parameters

end for

IF(Target attained)
    Return(Cutoff$_{opt}$, FC$_{opt}$)

ELSE
    No threshold achieved Target Jaccard Index
```

Algorithm 1. The *referenceNof1* algorithm pseudocode to construct an optimized and unbiased reference standard. The *referenceNof1* algorithm requires a user to input the FC and expression cutoff filters for it to then identify the optimal region for producing the reference standard. For each pair of FC-region and expression cutoff combination, it calculates each method's list of differentially expressed genes (DEGs), and then for each DEG list it calculates the Jaccard Index (JI) as a set-theoretic pairwise similarity measure. After calculating all pairwise Jaccard indices, it calculates the median JI for each region. If a parameter combination attains the desired median Jaccard Index, an early stopping rule is implemented and the optimal parameter combination is returned. Otherwise, it continues the search until the target JI is attained or the search through the parameter space is complete.

4.8 Comparing the proposed Reference Standard Optimization with a single heteromorphic one

To illustrate the benefits of creating more robust reference standards, we conducted an exemplar study using DESeq. As shown in **Figure 1**, the analysis consisted of constructing a reference standard using the intersection of all DEG calls by edgeR, NOISeq, DEGseq, and DESeq2 using four MCF7 samples in two conditions (8 samples), while the prediction of single-subject DEGs(ss-DEGs) was conducted on independent sample pairs three times using DESeq (3 independent pairs). Note that the prediction method is not part of the reference standard construction to mitigate for analytic biases (**Table 1**). The analysis consists of constructing the reference standard first on the entire set of MCF7 gene product counts and second on the optimal region of concordance as identified by the Jaccard Indices. Then, using this robust reference standard identified by ***referenceNof1***, an exemplar analysis in a hold-out pair of single-subject paired-transcriptomes, DESeq is used to identify altered genes and the results are compared against the reference standard. This process is repeated three times, once on each of the three hold-out sets (ss-DEGs). The average results across all three hold-out sets are presented in **Table 3**. For comparison, the results are shown for an equivalent DESeq analysis on the full, unfiltered hold-out sets.

4.9 Code availability

All the code used to carry out this study is readily available in GitHub under the SamirRachidZaim/referenceNof1_study repository.

4.10 referenceNof1 R package

The code *SamirRachidZaim/referenceNof1_study* was re-packaged into a reproducible and shareable R-package format, available for installation on GitHub under the following repository: *SamirRachidZaim/referenceNof1.*

5 Conclusions

Reproducibility and accuracy are not only central to Omics studies but to precision medicine. Improving existing techniques and frameworks in single-subject studies allows us to separate clinically-relevant biomarkers from statistical artefacts. Transforming these initiatives into open-source software greatly enables reproducibility and furthers the space of open precision medicine. Prior studies [3] illustrate how the unique challenges of single-subject analyses of transcriptomes in absence of replicates remains challenging. However, we posit that an improvement in evaluation methods, as proposed here, provides the rigorous framework for assessing objectively ulterior proposed improvements. In addition, pathway-level single-subject studies of transcriptomes have been shown more accurate than gene product level ones [3], suggesting potential future pathway-level applications of the methods we proposed. This manuscript highlights four types of biases (**Table 1**) that confound results in both conventional analyses and the clinical translation of single-subject studies. The proposed ***referenceNof1***, complementary to [12], follows a suite of recent work [12, 17, 24] in which we seek to address these challenges, resulting in a new framework for creating robust reference standards. We proposed, tested, and developed an open-source software using a single strategy that reduces two additional biases: (i) statistical distribution bias and (ii) systematic bias from isomorphic evaluations (using the same analysis in the prediction and validation sets). Despite the specific challenges posed in single-subject studies, these advances create new opportunities to combine single-subject and conventional cohort studies. In essence, this manuscript continues recent work and addresses existing knowledge gaps and challenges in the single-subject domain to bring our tools, technology, and analyses closer to delivering the promise made by precision medicine: “the right treatment, for the right patient, at the right time.”

Author Contributions: SRZ and YAL conceived the analyses. SRZ conducted the statistical analyses in R; SRZ, HHZ, and YAL reviewed and interpreted results. SRZ, CK and YAL conceived and created the figures and tables. All authors wrote and revised the manuscript.

Funding: This work was supported in part by The University of Arizona Health Sciences Center for Biomedical Informatics and Biostatistics, the BIO5 Institute, and the NIH (U01AI122275).

5.1 Conflicts of Interest: The authors have no conflict of interest.

8 Mathematical Notation

Notation & Variable	Variable Description	Equation
A_k and B_k	Gene product expression of gene *k* in conditions *A* and *B*	Figure 1
FC	Fold change	Equation 1
Jaccard Index (JI) = $\frac{\|M \cap N\|}{\|M \cup N\|}$	Jaccard index is the ratio of significant gene products in common between results derived from analytical methods *M* and *N* divided by the union of these sets	Equation 2
$M \cap N$	Intersection of sets *M* and *N*	Equation 2
$M \cup N$	Union of *M* and *N*	Equation 2
$\|M\|$	Cardinality or size of set *M*	Equation 2
Ri	Region i, portion of the transcriptome resulting from the filters and cutoffs selected in ***referenceNof1***	Algorithm 1
$JI_{(M,N),Ri}$	The Jaccard index between analytics methods M and N for genes in region *Ri*	Algorithm 1

9. Acronyms and Abbreviations

Abbreviation	Name
DEGs	Differentially expressed genes
DGE	Differential gene expression
SS	single-subject
EL	Ensemble learner
FC	Fold change
FCR	Fold change region
JI	Jaccard Index
NHST	Null Hypothesis Significance Testing

References

1. Kratochwill TR, Hitchcock J, Horner R, Levin JR, Odom S, Rindskopf D, Shadish W: **Single-case designs technical documentation.** *What works clearinghouse* 2010.
2. Baker M: **1,500 scientists lift the lid on reproducibility.** *Nature News* 2016, **533:**452.

3. Vitali F, Li Q, Schissler AG, Berghout J, Kenost C, Lussier YA: **Developing a 'personalome'for precision medicine: emerging methods that compute interpretable effect sizes from single-subject transcriptomes.** *Briefings in bioinformatics* 2017.
4. Lim S, Lee S, Jung I, Rhee S, Kim S: **Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data.** *Briefings in bioinformatics* 2020, **21:**36-46.
5. Kratochwill TR, Brody GH: **Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification.** *Behavior Modification* 1978, **2:**291-307.
6. Schissler AG, Gardeux V, Li Q, Achour I, Li H, Piegorsch WW, Lussier YA: **Dynamic changes of RNA-sequencing expression for precision medicine: N-of-1-pathways Mahalanobis distance within pathways of single subjects predicts breast cancer survival.** *Bioinformatics* 2015, **31:**i293-i302.
7. Li Q, Schissler AG, Gardeux V, Achour I, Kenost C, Berghout J, Li H, Zhang HH, Lussier YA: **N-of-1-pathways MixEnrich: advancing precision medicine via single-subject analysis in discovering dynamic changes of transcriptomes.** *BMC Medical Genomics* 2017, **10:**27.
8. Gardeux V, Achour I, Li J, Maienschein-Cline M, Li H, Pesce L, Parinandi G, Bahroos N, Winn R, Foster I: **'N-of-1-pathways' unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine.** *Journal of the American Medical Informatics Association* 2014, **21:**1015-1025.
9. Li Q, Zaim SR, Aberasturi D, Berghout J, Li H, Vitali F, Kenost C, Zhang HH, Lussier YA: **Interpretation of Omics dynamics in a single subject using local estimates of dispersion between two transcriptomes.** *bioRxiv* 2019**:**405332.
10. McShane BB, Gal D, Gelman A, Robert C, Tackett JL: **Abandon statistical significance.** *The American Statistician* 2019, **73:**235-245.
11. Wasserstein RL, Schirm AL, Lazar NA: **Moving to a world beyond "p< 0.05".** Taylor & Francis; 2019.
12. Zaim SR, Kenost C, Berghout J, Vitali F, Zhang HH, Lussier YA: **Evaluating single-subject study methods for personal transcriptomic interpretations to advance precision medicine.** *BMC medical genomics* 2019, **12:**96.
13. Liu Y, Zhou J, White KP: **RNA-seq differential expression studies: more sequence or more replication?** *Bioinformatics* 2014, **30:**301-304.
14. Anders S, Huber W: **Differential expression of RNA-Seq data at the gene level–the DESeq package.** *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)* 2012, **10:**f1000research.
15. Wang L, Feng Z, Wang X, Wang X, Zhang X: **Degseq: an R Package for Identifying Differentially Expressed Genes From Rna-Seq Data.** *Bioinformatics* 2009, **26:**136-138.
16. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, et al: **How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?** *RNA* 2016, **22:**839-851.
17. Li Q, Zaim SR, Aberasturi D, Berghout J, Li H, Vitali F, Kenost C, Zhang HH, Lussier YA: **Interpretation of 'Omics dynamics in a single subject using local estimates of**

dispersion between two transcriptomes. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association; 2019: 582.

18. Li Q, Zaim SR, Aberasturi D, Berghout J, Li H, Vitali F, Kenost C, Zhang HH, Lussier YA: **iDEG: a single-subject method utilizing local estimates of dispersion to impute differential expression between two transcriptomes.** *bioRxiv* 2018:405332.
19. Gardeux V, Bosco A, Li J, Halonen MJ, Jackson D, Martinez FD, Lussier YA: **Towards a PBMC "virogram assay" for precision medicine: Concordance between ex vivo and in vivo viral infection transcriptomes.** *Journal of biomedical informatics* 2015, **55:**94-103.
20. Gardeux V, Berghout J, Achour I, Schissler AG, Li Q, Kenost C, Li J, Shang Y, Bosco A, Saner D, others: **A genome-by-environment interaction classifier for precision medicine: personal transcriptome response to rhinovirus identifies children prone to asthma exacerbations.** *Journal of the American Medical Informatics Association* 2017, **24:**1116-1126.
21. Sha Y, Phan JH, Wang MD: **Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data.** In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2015: 6461-6464.
22. Hackstadt AJ, Hess AM: **Filtering for increased power for microarray data analysis.** *BMC bioinformatics* 2009, **10:**11.
23. Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power for high-throughput experiments.** *Proceedings of the National Academy of Sciences* 2010, **107:**9546-9551.
24. Zaim SR, Li Q, Schissler AG, Lussier YA: **Emergence of pathway-level composite biomarkers from converging gene set signals of heterogeneous transcriptomic responses.** In *Pac Symp Biocomput*. World Scientific; 2018: 484-495.
25. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: **Gene ontology: tool for the unification of biology.** *Nature genetics* 2000, **25:**25.
26. Li Q, Schissler AG, Gardeux V, Berghout J, Achour I, Kenost C, Li H, Zhang HH, Lussier YA: **kMEn: Analyzing noisy and bidirectional transcriptional pathway responses in single subjects.** *Journal of biomedical informatics* 2017, **66:**32-41.
27. Gardeux V, Berghout J, Achour I, Schissler AG, Li Q, Kenost C, Li J, Shang Y, Bosco A, Saner D: **A genome-by-environment interaction classifier for precision medicine: personal transcriptome response to rhinovirus identifies children prone to asthma exacerbations.** *Journal of the American Medical Informatics Association* 2017, **24:**1116-1126.
28. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic acids research* 2002, **30:**207-210.
29. Team RC: **R: A language and environment for statistical computing.** 2013.
30. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.** *Bioinformatics* 2009, **26:**139-140.
31. Anders S, Huber W: **Differential expression of RNA-Seq data at the gene level–the DESeq package.** *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)* 2012.

doi:10.20944/preprints202011.0325.v1

32. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome biology* 2014, **15:**550.

33. Tarazona S, García F, Ferrer A, Dopazo J, Conesa A: **NOIseq: a RNA-seq differential expression method robust for sequencing depth biases.** *EMBnet journal* 2011, **17:**18-19.

34. Feng J, Meyer CA, Wang Q, Liu JS, Shirley Liu X, Zhang Y: **GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data.** *Bioinformatics* 2012, **28:**2782-2788.