

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Less is More: Mitigating batch-effects in large scale RNA-Seq experiments by balancing experimental factors using a genetic algorithm

Permalink

<https://escholarship.org/uc/item/7343v3r1>

Author

Altieri, Mia

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Less is More: Mitigating batch-effects in large scale RNA-Seq experiments by balancing
experimental factors using a genetic algorithm

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Computer Science

by

Mia Gabrielle Altieri

Committee in charge:

Professor Nathan Lewis, Chair
Professor Debashis Sahoo, Co-Chair
Professor Melissa Gymrek

2021

Copyright

Mia Gabrielle Altieri, 2021

All rights reserved.

The Thesis of Mia Gabrielle Altieri is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

I would like the chance to dedicate this work to my friends, family, lab, and community that have supported me along the way. Without their continued support I would not have been able to complete this work. Most importantly I want to thank Alex Neuman, Matt Guzzo, Cameron Jones, my community at Lothlorien, and the housing commune Bug Farm; for their kindness and for reminding me to stay true and align myself with my long term goals. These colorful communities have brought much needed thoughtful discourse, supported me when I needed it the most, and reminded me to continue looking forward. Additionally I want to thank those at Lewis Labs, specifically Nathan Lewis and Austin Chiang. Without them this work would not have been possible, their guidance and patience proved to be invaluable. I feel both lucky and honored to have not just one but two amazing, once in a lifetime mentors. Finally I want to thank my parents Diane and Michael Altieri. Thank you for helping me follow my dreams and for supporting me while things were the most challenging, I owe you everything.

TABLE OF CONTENTS

Thesis Approval Page.....	iii
Dedication.....	iv
Table of Contents.....	v
Abstract of the Thesis.....	vii
Introduction.....	1
Chapter 1 Results.....	4
Chapter 2 Materials.....	7
Chapter 3 Methods.....	8
3.1 Simulating Data.....	8
3.2 Balancing Methods.....	11
3.3 The Genetic Algorithm.....	12
3.3.1 Prior Plate.....	12
3.3.2 Interplate.....	13
3.3.3 Initial Solutions.....	13
3.3.4 Fitness Score.....	14
3.3.5 Selection.....	15
3.3.6 Crossover.....	15
3.3.7 Mutation.....	15
3.5 Differential Expression Analysis (DEA) on Simulated Data.....	16
3.6 Simulating GSEA.....	16
Chapter 4 Conclusion.....	18
Chapter 5 Discussion.....	20
Figures.....	23

Figure 1: Covariate and Plate Balance Scores Before and After Library Balancing on Simulated Data.....	23
Figure 2: Gene Set Enrichment Analysis with and without Library and Covariate Balancing on Simulated Data.....	24
Figure 3: DE Analysis on Simulated Data.....	25
Figure 4: The Genetic Algorithm Process.....	27
Supplementary Figures.....	28
Supplementary Figure 1: Known Balanced Subset.....	28
Supplementary Figure 2: Covariate Distributions for 250 Unbalanced Samples.....	29
Supplementary Figure 3: Ideal Metrics.....	30
Supplementary Figure 4: LFC Patterns.....	31
Tables.....	32
Table 1. Overview of the samples used in the unbalanced library design of Simulated RNA-Seq experiments.....	32
Citations.....	34

ABSTRACT OF THE THESIS

Less is More: Mitigating batch-effects in large scale RNA-Seq experiments by balancing experimental factors using a genetic algorithm

by

Mia Gabrielle Altieri

Master of Science in
Computer Science

University of California San Diego, 2021

Professor Nathan Lewis, Chair
Professor Debashis Sahoo, Co-Chair

Randomization has been considered as the most important method to protect against bias and ensure the internal validity of clinical trial studies. Conducting randomization procedures could induce comparability with respect to known and unknown covariates, mitigate selection bias, and provide a basis for inference. However, randomization can't guarantee each covariate is balanced in large scale clinical samples. While the advent of next-generation sequencing (e.g., RNA-Seq) technologies allows us to measure global gene expression in a large number of samples with low cost, combining samples with imbalanced covariates in one RNA-Seq experiment can lead to the 'batch effect' problem. Specifically, the biological variation is confounded with unwanted variations from biased covariates. These unwanted variations must be effectively removed to eliminate batch effects that could significantly bias the biological conclusions. Unfortunately, they become indissociable and un-removable when examining samples with unbalanced experimental factors in the design process of a RNA-Seq experiment. Therefore, how to design a RNA-Seq experiment with fully balanced experimental factors to guarantee removable batch effects is an important task in the high-throughput RNA-Seq study era. In this study, we propose a genetic algorithm (GA)-based tool called BalanceIT to balance experimental factors prior to sequencing. BalanceIT identifies an optimal set of samples with balanced experimental factors to be used in the design of an RNA-Seq experiment. Using a panel of ~1000 simulated samples we demonstrate that our proposed GA-based tool is superior to the conventional randomization-based method in designing RNA-Seq experiments with samples of unbalanced experimental factors.

Introduction

With the advent of next-generation sequencing technologies, RNA-Seq now allows us to measure global gene expression in large number of samples with low cost (Fomina-Yadlin et al., 2015; Hsu et al., 2017; Vishwanathan et al., 2015; Wang et al., 2009; Yuk et al., 2014). However, combining samples from different batches in one RNA-Seq experiment can lead to the ‘batch effect’ problem (Akey et al., 2007; Sebastiani et al., 2012; Leek et al., 2010), in which biological variation is confounded with unwanted experimental variation from differences in sample processing procedures in different batches. These variations have to be effectively removed to eliminate the batch effects that could significantly bias the biological conclusions (Yang et al., 2008; Leek et al., 2010; Lambert & Black, 2012). Unfortunately, they become indissociable and un-removable when examining samples with unbalanced experimental factors in the design process of a RNA-Seq experiment. Therefore, how to design a RNA-Seq experiment with fully balanced experimental factors to guarantee removable batch effects has become an important task in the high-throughput RNA-Seq study era. The conventional randomization methods (Yang et al., 2008) that randomly sample two subsets of samples from a large dataset can’t guarantee all factors are balanced in the large-scale data set. GAs are powerful searching algorithms that are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection (Goldberg 1989 and Melanie 1996). In this study, we propose a genetic algorithm (GA)-based tool that can identify an optimal set of samples with balanced experimental factors that can be used in the design of an RNA-Seq experiment. This tool mitigates batch effects prior to sample sequencing by using an optimally balanced subset of samples. We refer to this pre-sequence balancing as library balancing (BL).

Post-sequencing batch effect mitigation has been proven to be an effective technique in adjusting for known and unknown covariates (Leek 2014; Müller et al., 2016). However this effectiveness has an upper limit depending on the severity of the batch effects (Zhou et al., 2019). Furthermore due to the nature of post-sequencing techniques, these techniques are at the mercy of their user since they cannot advise the user to organize samples in a way that would lead to less severe batch effects. Additionally some of these methods struggle to remove batch effects which are not orthogonal to one another due to the nature of their methods. More specifically, Combat is not well suited for removing multiple batch effects as Combat is designed to regress out one covariate at a time from the data, thus using combat iteratively for several covariates may result in errors. For example, if two covariates are non-orthogonal then estimating and removing the effect of the first covariate may interfere/hinder the estimation and removal of the second covariate as the first estimation may have been confounded with the second one. Similarly for SVA, estimating the effects of covariates may prove to be difficult if covariates are not orthogonal to one another (Lee et al., 2018). In either of these cases each may produce biased estimates and result in incorrect results. In this paper we explore the effect that these conventional post-sequencing methods have using both our GA-based approach and the randomized approach to illustrate the importance of quality library balancing.

Simulating genetic data has become a popular medium for validating bioinformatics methods (Chen et al., 2011; Engstrom et al., 2013; Li et al., 2013; Goldstein et al., 2016; Zhang et al., 2017). To illustrate the quantifiable differences between pre-sequencing and post-sequencing batch correction techniques we created a framework for simulating RNA-Seq data where the effect of covariates could be quantified and readily controlled. Using our framework we organized a panel of ~1000 simulated samples to compare the effects of

BalanceIT with other popular batch mitigation methods SVA and Combat. This application helped us to demonstrate that our proposed GA-based tool is superior to conventional randomization-based methods in designing RNA-Seq experiments with samples of balanced experimental factors.

Chapter 1 Results

Using our framework we simulated 1000 samples for 3 categorical covariates, 3 continuous covariates, 1 diagnosis, and a random plate design was used for testing library and covariate balancing metrics (**Table 1**). These covariates were simulated to have interactions between each other and thus not be orthogonal to each other. Prior to downstream analysis, balance scores were computed to illustrate the necessity of balancing samples prior to sequencing (**Figure 1**). After sample sequencing, we utilized the fact that our simulated RNA-Seq framework provides the user with the true differentially expressed genes (DEGs). These DEGs are used to perform Gene Set Enrichment Analysis (GSEA). Furthermore our framework provides ideal datasets that illustrate the effects of both perfect library and covariate balancing, serving as a baseline comparison metric.

Our tool BalanceIT aims to increase accuracy in downstream analysis by reducing collinearity between factors of interest and artifacts such as batch, biological factors, or other covariates. The results from our tool demonstrate its capability to decrease collinearity between the factor of interest and other factors through subsetting (**Figure 1**). Prior to subsetting, correlations between three of the six experimental factors and the diagnosis that were highly significant ($p\text{-value} < 0.05$), the sum of these correlations, also referred to as balance scores, was 0.464 meaning it was 8%~ balanced (maximum balance score achievable 6.0). The sum of the balance scores after the GA identified an optimal subset was 5.75, increasing the balance percentage to 96%~. Additionally, each balance score had a $p\text{-value} > 0.84$. Furthermore, the plate assignment from the GA increased the initial balance scores from the randomized design of 3.84 to 6.68. The GA demonstrated its capability to reduce collinearity between the biological

factor of interest and other unwanted factors and this decreased collinearity in turn improved the downstream results as illustrated in the GSEA and DE analysis. (**Figure 2 & Figure 3**)

As expected, for our dataset of collinear, non-orthogonal covariates the iterative use of Combat performs poorly. Specifically we see that Combat drastically reduces the set of TP DEGs from 359 to 58 from the fully unbalanced dataset ULUC to the dataset ULBC which is partially balanced using Combat; this in turn negatively effects Combats GSEA results reducing its overall sensitivity, specificity, and accuracy for gene sets of all sizes (**Figure 2**). The drastic loss in sensitivity (60%) makes Combat unsuitable for datasets with multiple collinear covariates. SVA's performance differs from Combat as it doesn't drastically decrease the quality of results, however we observe no substantial increase/decrease in GSEA results when using SVA, in fact the results are fairly comparable to ULUC. This in turn is observed in the DEA results where we see minimal increases/decreases in specificity, sensitivity, and accuracy.

In terms of library balancing we observe that our tool performs favorably on our dataset of collinear, non-orthogonal covariates. Specifically we see a significant reduction in the number of FP DEGs identified when no balancing is done from 390 to 152, thus leading to increased specificity and accuracy in the DE analysis, (20% and 14% increase respectively) (**Figure 3**). This in turn improves the GSEA results and we observe a substantial increase in sensitivity for gene sets of all sizes. As a comparison metric, we ran this analysis using the initial 750 balanced samples, and observed that the GA outperforms the initial set of 750 samples. We attribute this to the fact that the GA is a stochastic method in which more samples may be included if it increases the balance score. This trend can be observed in both Figure 1 and Supplementary Figure 1 where the resulting balance scores of the GA are higher than the initial 750 balanced samples. The GA outperforms the initial 750 balance samples most notably in sensitivity for the DE

analysis, where the GA retains more 44 TP DEGs than the initial 750 balanced samples. This performance difference can be further observed in the GSEA results where we see the GA outperforms the initial 750 balanced subset across all metrics on datasets of all sizes Figure 2.

Finally and most notably is the effect our tool has when used in tandem with Combat and SVA. Our dataset of collinear, non-orthogonal covariates proved to be an obstacle for both post-sequencing methods Combat and SVA. Despite the pre-balancing done by our tool we demonstrate that Combat is still unable to accurately remove batch effects in a dataset with collinear, non-orthogonal covariates. Yet, BalanceIT proved to be an effective measure for pre-balancing the data when using SVA and we observe an increased performance in SVA when BalanceIT is used. Specifically we observe an increase in accuracy from 76% to 90% and an increase in specificity from 81% to 100% when solely SVA is used and when SVA is used in conjunction with BalanceIT. This suggests that by in fact removing samples from an unbalanced panel of samples we may further improve our downstream results in RNA-Seq analysis, even if the samples have collinear, non-orthogonal covariates when using SVA. Further we observe a slightly boosted performance in GSEA sensitivity and accuracy results when GA-based method BalanceIT is applied to SVA.

Chapter 2 Materials

To evaluate balancing techniques a simulated dataset of 1000 samples and 2000 genes was generated. For the RNA-Seq count data, Five-hundred of the 2000 genes were simulated to be differentially expressed. For the sample metadata, diagnoses, categorical covariates, and continuous covariates were assigned for each sample (**Table 1**). Additionally samples were randomly assigned to plates. Of the 1000 samples generated 750 samples were generated such that their covariates were balanced with diagnosis (**Supplementary Figure 1**). Furthermore interactions were simulated between two pairs of covariates such that not all covariates were orthogonal to each other.

Chapter 3 Methods

3.1 Simulating Data

As aforementioned, a major quantifier for our tool BalanceIT was our simulated dataset. To quantify the effects of our tool it was requisite that we had the ability to both quantify and control batch effects as well as know the ground truth DEGs; thus we produced our own framework for simulating RNA-Seq counts. Simulated RNA-Seq datasets contained both the count data as well as a phenotypic/covariate matrix. We now outline our strategies for building our framework for our simulated datasets.

The simulated dataset is created in two steps, the first is the phenotypic/covariate matrix generation and the second being the RNA-Seq count matrix generation. A covariate matrix for 1000 samples was generated for three continuous covariates (a,b, and c), three categorical covariates (d,e, and f), and a single binary diagnosis. Seventy-five percent of the 1000 samples simulated were generated separately to be well balanced for the six covariates. For these 750 samples to be considered “balanced”, balance scores for each covariate with the diagnosis must be > 0.05 . Balance scores were computed using chi-squared tests for categorical covariates and linear models for non-categorical covariates. When simulating the covariate values for the balanced subset each covariate had a uniform probability of occurring regardless of the diagnosis, ensuring that there were no biases within the subset. The remaining 250 (25%) samples were simulated to be unbalanced, each covariate value had varying probabilities of occurring (**Supplementary Figure 2**). Furthermore, two pairs (a&d and c&e) of covariates were generated to have interactions between each other. To simulate interactions between covariates in the two pairs the covariate values were dependent on each other. For each pair there was one

binary categorical variable and one continuous variable, depending on the value of the categorical variable a specific portion of the possible values for the continuous variable was used. Additionally, a plate matrix was generated. The plate matrix, P, represents the various batches of the 1000 samples and is generated using a uniform random sampling of numbers 1 through 11 and assigning them to each sample.

After the phenotypic/covariate matrices are made, the diagnosis for each sample is used to generate raw RNA-seq counts for 2000 genes. The raw counts for each gene were generated using unique negative binomial distributions regardless of their diagnosis. For the DEGs, 500 genes were randomly selected to be DE, differential expression was simulated by first assigning a fold change based on the diagnosis and the direction of the log fold change (positive/negative). Magnitude (m) of the log fold change (LFC) were determined using a sampling of two random normal distributions, one with a mean of 0.75 and a standard deviation of 0.5 and another with a mean of -0.75 and a standard deviation of 0.5. Gene expression values of DEGs were changed according to the magnitude of their LFC using the formula $e^{(m)} \times \text{original expression}$.

After initialization, the covariate and plate matrix are scaled to ensure covariates and plates have a similar effect, each covariate and plate values were scaled to be within the range [0,0.5] by dividing each value by the maximum value for that covariate or for the maximum plate and then subtracting 0.5. Interactions between covariates were further simulated by combining covariates values together randomly. Resulting in 6 interactions: (a+d), (b+e), (c+f), (d+b), (e+c), and (f+a). All of the covariates and the plate values are multiplied by unique negative random binomial distributions, to ensure that covariates/plates affect different genes. To randomize the negative binomial distribution, while keeping them at the same magnitude, the same parameters

were used for each distribution (1 successful trial, .008 probability of success), and the distributions were taken for each factor.

Furthermore, to ensure that the covariates had the same effect on the dataset as the plate, multipliers were used for both the covariate and plate effects. Multipliers of 0.7 and 15 were used for the covariate and plate matrix respectively. To validate that these multipliers had a comparable effect for both the covariate and plate downstream analyses using sample distance, GSEA, and DEA were performed with datasets with perfectly balanced plates and perfectly balanced covariates. These downstream analyses illustrated that the effect of the plate and of the covariates were comparable with the aforementioned multipliers 15 and 0.7 (**Supplementary Figure 3**). While we required that the plate and covariate effect be comparable we allowed for the covariate effect to have a slightly stronger skew on the results as there are 6 covariates within our data and only one plate.

For additional random error, an error matrix was added to the RNA-Seq count data and is generated using the absolute value of a random normal distribution with a mean of 0 and a standard deviation equal to the maximum value of the raw RNA-Seq data divided by the number of genes present. This was done to ensure that the random error was sufficiently large for the number of genes. To ensure validity of the simulated RNA-Seq counts, in a separate analysis, an airway dataset was upsampled for the number of samples and downsampled for the number of genes to illustrate that our count data followed a standard distribution (**Supplementary Figure 4**).

Furthermore a necessary function of our framework was to have the capability to generate ideal scenarios in which datasets of RNA-Seq data could be without plate or covariate

effect to elucidate perfect library balancing and perfect batch effect correction. To enable this functionality we allow the user to choose whether to include plate or covariate effects in a dataset. Thereby creating datasets that could be considered ideally library balanced or ideally covariate balanced or both. These served as baseline comparisons for the real covariate balancing techniques (Combat and SVA) and the real library balancing techniques (BalanceIT and randomization). The ideal balancing scenarios were validated by demonstrating that sample distances become tightly associated with diagnosis as balancing techniques are applied (**Supplementary Figure 3**).

3.2 Balancing Methods

Two types of balancing were performed for this project: covariate balancing and library balancing. Covariate balancing, which refers to post sequencing balancing, was performed using two popular methods: Combat and Surrogate Variable Analysis (SVA). Combat adjusts for known batch effects in a data set using a normalized dataset and Empirical Bayes. Combat was provided with RNA-Seq counts that were normalized using a variance stabilizing transformation (VST). SVA removes unwanted artifacts by first estimating batch effects and then removing them by using an iteratively reweighted least squares approach.

Library balancing is performed prior to sequencing and aims to minimize the impact of covariates on RNA-Seq counts by minimizing the correlation between experimental factors and the biological factor of interest. Our proposed tool BalanceIT utilizes a GA in order to minimize these effects. The GA minimizes batch effects in two steps, first by reducing the collinearity

between covariates and the diagnosis, and then by reducing collinearity between plates and covariates and the diagnosis.

3.3 The Genetic Algorithm:

Our proposed tool hinges on the capability of GAs to optimize solutions. GAs work by iteratively finding the best solution through repeatedly computing fitness scores, mixing samples, and adding mutations (**Figure 5**). In the scope of this work the GA aims to reduce collinearity so that covariates and plates do not drive the RNA-Seq count data. The Genetic Algorithm reduces unwanted collinearity by working in two steps. The first step aims to reduce collinearity between covariates and the diagnosis through finding the optimal subset, such that the optimal subset increases the balance scores for each covariate. Balance scores are computed using chi-squared and anova tests using covariate values and the diagnosis. After the GA has obtained an optimal subset it assigns plate values to each sample such that balance scores for each plate are maximized.

3.3.1 Prior Plate

Since GAs work by iteratively finding the best solution, it is crucial that solutions are well defined. The Prior Plate GA identifies a subset of samples to sequence. For a subset of samples to sequence to be considered optimal, it must increase the sum of the balance scores for each covariates and retain the most number of samples possible. The Prior Plate GA takes in a predetermined number of sample solutions, each solution takes the form of a list of zeros and

ones, corresponding to the number of samples. A one in the solution indicates a sample being included and a zero corresponds to a sample being removed. Using the included samples, balance scores are computed and used to determine the fitness score in conjunction with the number of samples included. (See 3.4.3 - 3.4.7 for full details).

3.3.2 Interplate

Interplate balancing using the GA is similar to that of balancing the covariates with the diagnosis, except instead of identifying the optimal subset, it determines the optimal plate assignments and needs not to discard samples. Using a modified version of the Prior Plate GA, samples are assigned plates such that it minimizes covariate and diagnosis correlation with any of the plates. (See 3.4.3 - 3.3.7 for full details).

3.3.3 Initial Solutions

Initial solutions are generated randomly unless provided by the user. Each initial solution contains a valid potential solution for either Interplate balancing or Prior Plate balancing. For Prior Plate balancing a solution thus takes the form of $\{S_1, S_2, \dots, S_n\}$ where $S \in \{0,1\}$, indicating whether or not a sample is included. For interplate balancing, solutions take the same form but $S \in \{1, \dots, \text{max_number_of_plates}\}$, as no samples are discarded in interplate balancing. Each sample S_i in the individual is annotated with the known experimental factors $\{F_1, F_2, \dots, F_n\}$ where n is equal to the number of experimental factors. **(Figure 5 i)**

3.3.4 Fitness Score

To assess whether the distributions of experimental factors in a solution is favorable, each experimental factor is evaluated by performing the ANOVA (Analysis of Variance) test for continuous factors and by the Chi-square (2) test for the categorical factors of the solution. The p-value of these statistical tests are considered as the ‘balance’ score of the experimental factor. For Prior Plate balancing the experimental factors are the covariates of interest and for Interplate balancing, the plates are used.

Experimental factors are handled differently for Prior Plate Balancing than they are for Interplate Balancing. The degree that an experimental factor is confounded with diagnosis affects that factors balance score and consequently all factors balance scores do not increase uniformly. To facilitate even distributions of balance scores for a solution, each experimental factor has a weight $W = \{W_1, \dots, W_k\}$ and scalar $V = \{V_1, \dots, V_k\}$ associated with it, where k is the length of experimental factors to be balanced with the diagnosis. The weights and scalars for each experimental factor are set prior to the start of the GA and are held constant through all solutions and through all iterations. The relation to a balance score of an experimental factor and its corresponding weight and scalar takes the form of $V_i * (\text{balance score}_i)^{W_i}$.

Additionally, since the Prior Plate balancing aims to retain as many samples as possible, size awards are given to solutions depending on the number of samples retained. To motivate the GA to generate an optimal solution of a substantial size an award is given to each solution with respect to the number of samples included. To influence the GA to generate a solution within an acceptable range the sigmoid function is used to push solutions into an acceptable range. Prior to the start of the GA the parameters for the sigmoid function: slope, offset, and size weight are set.

3.3.5 Selection

The population is then sorted by descending fitness values, and the top 80% of solutions are kept to be used in the next generation. The top 80% of the population (800 solutions) are considered as candidates to select two individuals as parents (N_i and N_j) for further crossover and mutation to generate a new individual of next-generation population. New individuals are generated until there are a full set of samples. **(Figure 5 iv)**

3.4.6 Crossover

Crossovers occur between two parents at a rate of 10%, in the case that a crossover does not occur, the child takes the father's solution. On the occasion that a crossover occurs, a single crossover point on both parents chromosome strings $\{P_1, P_2\}$ is selected, and all data beyond that point in either chromosome string is swapped between the two parent organisms. The resulting chromosome string is the child (N'). **(Figure 5 v)**

3.4.7 Mutation

Mutation used to maintain diversity in individuals, mutation occurs in a child solution after crossover. In the child solution 10% of the samples are mutated. **(Figure 5 vi)** For Prior Plate balancing, mutations correspond with including or removing a sample. Mutations occur by randomly changing the value of a sample flag in the child from 0 to 1 or vice versa. To influence growth in the number of samples selected by the GA, mutations were biased to change sample

flags to 1 by 5% . In Interplate balancing, mutation corresponds to changing plate assignments for a given sample and needs not to consider the growth of the sample set.

3.5 Differential Expression Analysis (DEA) on Simulated Data

Validating BalanceIT, required utilizing known downstream methods of analysis. DEA is a common method of downstream analysis for many modern day RNA-Seq studies. Differentially expressed genes (DEGs) were determined using DESeq, for a gene to be a DEG it must have an FDR < 0.05 . All combinations of balancing methods (covariate and balancing) were used on the unbalanced dataset, DESeq was used to identify predicted DEGs and non-DEGs. The 500 known DE genes provided by our simulated RNA-Seq framework served as a metric for comparing balancing methods, using the known DE and non-DE genes confusion matrices were generated from the predicted DE and non-DE. The confusion matrices were used to compute accuracy, sensitivity, and specificity; and used to further compare balancing methods.

3.6 Simulating GSEA

GSEA is a common down RNA-Seq analysis, thus motivating us to simulate GSEA as a form of measuring the performance of our tool with respect to covariate and library balancing. We simulated GSEA using a variety of gene sets and determined whether a gene set was enriched using the true DEGs provided by our simulated RNA-Seq framework. One-hundred gene sets were simulated containing 100, 200, 300, 400, and 500 genes, totaling for 500 total gene sets. For each gene set, g_i genes were selected randomly without replacement. Using the

true DE genes and the genes in the gene set, the hypergeometric distribution was used to identify which gene sets were enriched and which ones were not enriched. Using the same strategy the DE genes from the datasets ULUC, ULBC, BLUC, and BLBC were used to predict which gene sets were enriched and which were not. Using the predicted and real enrichment results, confusion matrices were generated for each gene set size (100, 200, 300, 400, and 500). Using these confusion matrices sensitivity, specificity, and accuracy were calculated. This analysis was then repeated 10,000 times, resulting in 50,000 confusion matrices (10,000 confusion matrices for each size gene set: 100, 200, 300, 400, and 500). The sensitivity, specificity, and accuracy metrics from these results were used to determine the enrichment quality of each dataset: ULUC, ULBC, BLUC, and BLBC.

Chapter 4 Conclusion

There have been many techniques pioneered for the downstream removal of batch effects after sequencing. However when a dataset is fraught with many batch effects, particularly significant batch effects, or non-orthogonal covariates, these downstream removal techniques may fail to eliminate artifacts. Randomization, a popular method for mitigating batch effects prior to sequencing, may not always properly mitigate these affects. In fact, we demonstrate that it may be best for a researcher to discard samples to obtain a clear biological signal. If a researcher can identify problematic samples and remove them, the dataset may provide a clear biological signal and improve the effects of modern downstream removal of batch effects. We created a tool for identifying problematic samples using a genetic algorithm to optimize balance scores and number of samples, and demonstrated its effectiveness in mitigating batch effects and improving pre-existing balancing techniques such as SVA and Combat.

Furthermore, we demonstrated that iterative methods for batch effect removal, such as Combat, can dramatically fail when batch effects are collinear and non orthogonal and further that our tool BalanceIT may not aid these iterative methods. Additionally, we illustrated that the performance of SVA may be further enhanced when the GA-based tool BalanceIT is applied. Interestingly enough, without post-sequencing balancing tools such as Combat and SVA BalanceIT still proves to improve GSEA and DEA results. More specifically BalanceIT improves DEA results with respect to accuracy and specificity and GSEA results with respect to sensitivity and accuracy. Interestingly enough BalanceIT outperformed our initial subsample of 750 balanced samples due to the stochastic nature of the GA approach which enabled it to include more samples that would increase its balance score.

Finally, we tested our methods on a simulated clinical RNA-Seq dataset which we generated using a framework we built in house. This framework allowed us to properly control and quantify the effects of the covariates and of the plates. Our results from these datasets demonstrated that the GA method for library balancing strengthens the biological signal and reduces the number of false positives, thereby increasing DE results and GSEA accuracy, going directly against the notion that more samples is better, illustrating that in some occasions less is more.

Chapter 5 Discussion

Batch effect mitigation has been proven essential in related works (Leek JT. 2014, Müller et al.; 2016), and there exist several methods for mitigating batch effects after samples have been sequenced i.e. Combat and SVA. The most popular methods for post-sequencing batch removal in 2021 are Combat and SVA. Combat, a supervised batch effect removal technique, is designed to remove a single covariate, or multiple orthogonal covariates using Empirical Bayes. Due to the stringent nature of the covariates needing to be orthogonal, Combat may not be effective in removing multiple collinear covariates. This can be a significant drawback as it is not uncommon for covariates to be dependent on one another. In this work we demonstrate Combat's inability to remove multiple collinear covariates and effectively show that using Combat with collinear covariates may worsen the quality of the resulting RNA-Seq counts.

SVA in contrast is an unsupervised method which identifies and estimates covariates to be removed using methods defined by the user (i.e. regression). SVA estimates covariates by finding a set of orthogonal vectors that span the same linear space as a single experimental factor across all genes (Varma, 2020). As in Combat, SVA may fail if collinear covariates are within the data as they may be difficult to estimate. However one of the main benefits of SVA is that it can identify covariates that are unknown to the user providing a hidden utility that is unavailable to supervised methods. Despite the challenges of the non-orthogonal covariates presented in our work, SVA proved to be successful in improving overall GSEA and DEA results. Yet, our work demonstrates that SVA performs best when used with a library balanced dataset.

Similarly, the package DESeq2, which identifies differentially expressed genes, has been designed for batch effect removal through a user-specified design. Covariates provided by the

user are included in the DESeq model to inform the value for a fitted mean for each gene on each sample. More specifically, for each gene on each sample there is a covariate dependent parameter which measures the variance which is included when calculating the fitted mean. While this mathematical modeling of covariates is intuitive and follows traditional approaches to modeling covariates, it simply does not remove batch effects as effectively as other downstream methods such as SVA. This fact is demonstrated in our work where we compare DESeq on its own (ULUC) with DESeq in conjunction with SVA (ULBC); in which we illustrate SVAs improving effect on DESeq2's performance.

While there exist a variety of methods for post sequencing correction of batch effects, the primary method for library balancing is randomization. Randomization has been proven to mitigate batch effects to a degree in regards to plate and sample sequencing. However, randomization makes no attempt to mitigate covariates since randomization only controls the plate design. Thus if you have a sample design in which a covariate (i.e. sex) is heavily correlated with your diagnosis randomized plate design cannot mitigate this affect. Hence motivating the need for a library balancing technique that handles more than the plate effect.

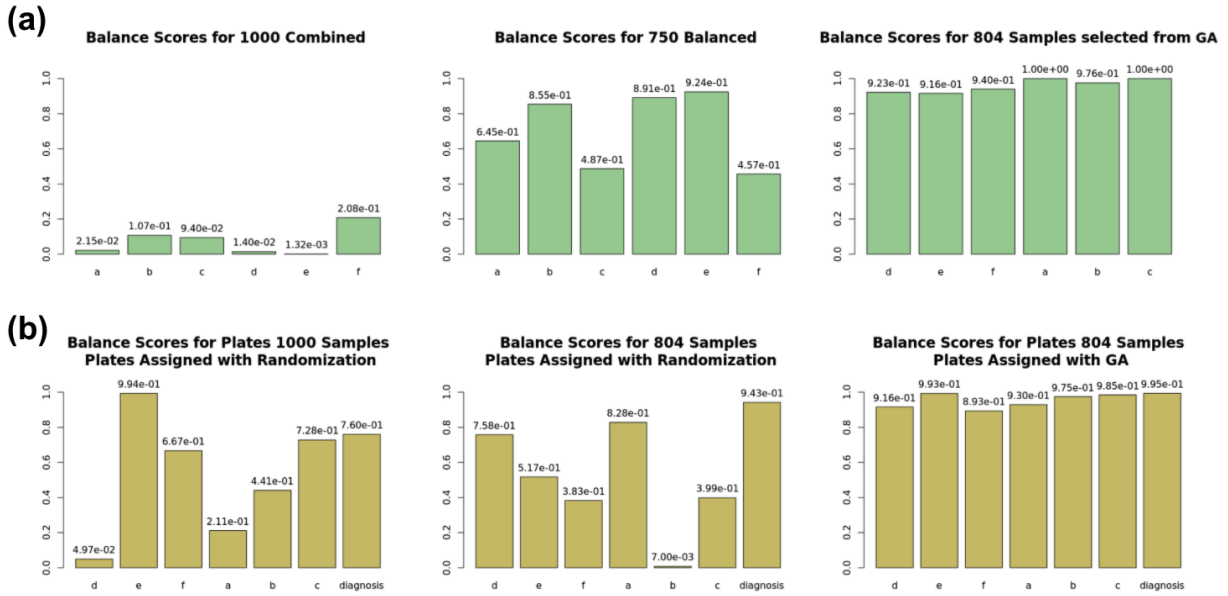
Our tool offers to balance both covariates and the plate effect via a GA-based method. In using a stochastic optimization scheme the GA-based tool identifies a subset of samples in which covariates are optimally balanced with the diagnosis and plates are balanced for all covariates and the diagnosis of interest. What we demonstrate in this work is that in intelligently removing 20% of samples and assigning plates using the GA optimization scheme we achieve better accuracy in both GSEA and DEA; and further remove correlations pre-existing in the data.

In this paper we illustrated that our GA-based method BalanceIT can effectively mitigate batch effects when used prior to sequencing, when samples are fraught with co-linear, non-orthogonal covariates. However we have not fully explored the utility of this tool when re-sequencing is not an option, i.e. when samples along with their RNA-Seq counts are provided for the user. Exploring whether the GA may be used post-sequencing can help to further studies in which combining samples would result in indistinguishable batch effects. To illustrate the effectiveness of BalanceIT on already sequenced RNA-Seq samples we suggest to use data that is publicly available from The Cancer Genome Atlas (TCGA), as it is a comprehensive dataset with many covariates, cancer types, tissues, and is collected from a variety of studies.

Finally, in this paper we explore a dataset with two highly collinear, non-orthogonal covariate pairs. Finding the threshold of collinearity in which the GA becomes comparable to standard post-sequencing batch effect mitigation methods would prove useful to illustrate when the GA should be used and when it may not be used. Exploring this threshold would provide a good utility to future researchers.

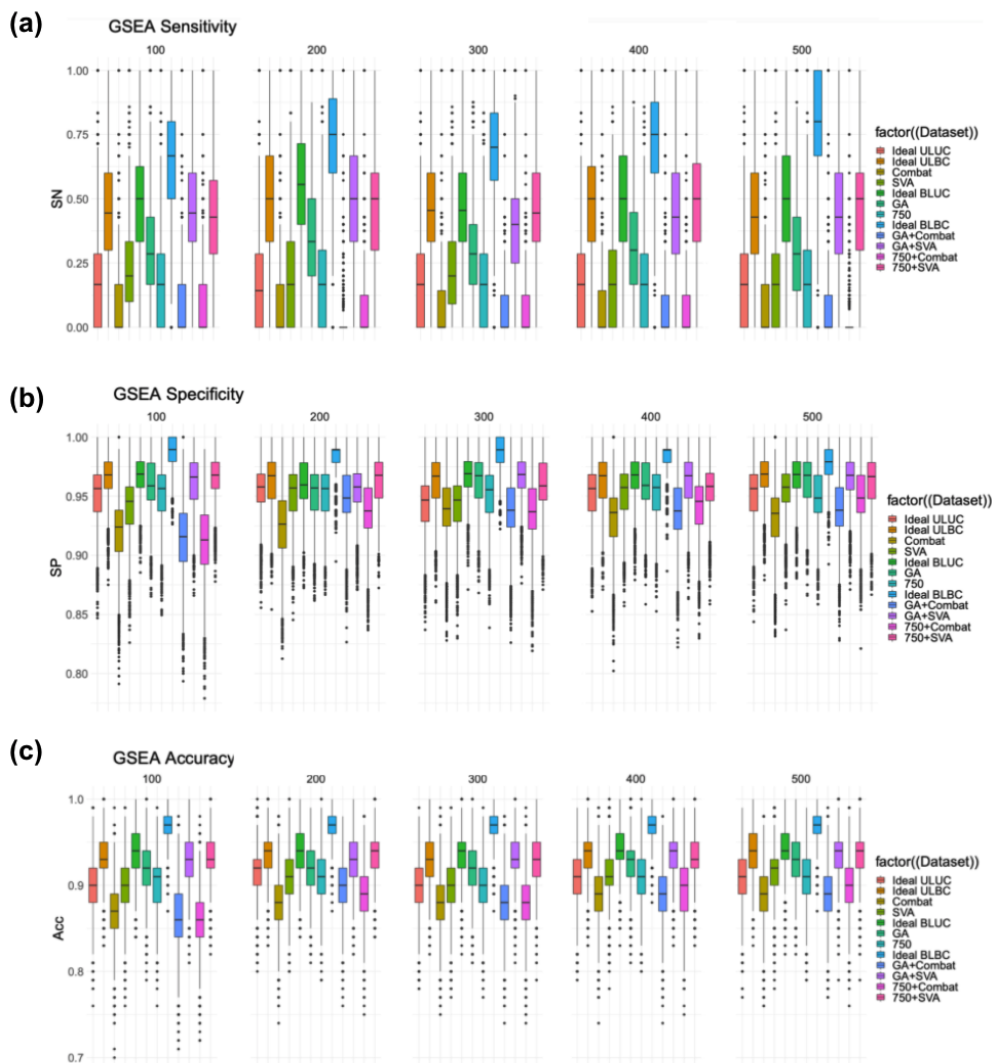
FIGURES:

Figure 1: Covariate and Plate Balance Scores Before and After Library Balancing on Simulated Data



a) Covariate Balancing Balance scores of covariates before and after balancing using the GA. Continuous covariates balance scores were computed using anova test with the covariate value and the diagnosis. Categorical covariates balance scores were computed using a chi-squared test with the covariate value and the diagnosis. **b) Plate Balancing** Balance scores for the covariates and diagnosis with the plate assignment for random assignment with the full sample set, random assignment with subsetting sample data, and GA assignment with subsetting sample set. Balance scores were computed using anova tests for continuous variables and chi-square tests for categorical variables.

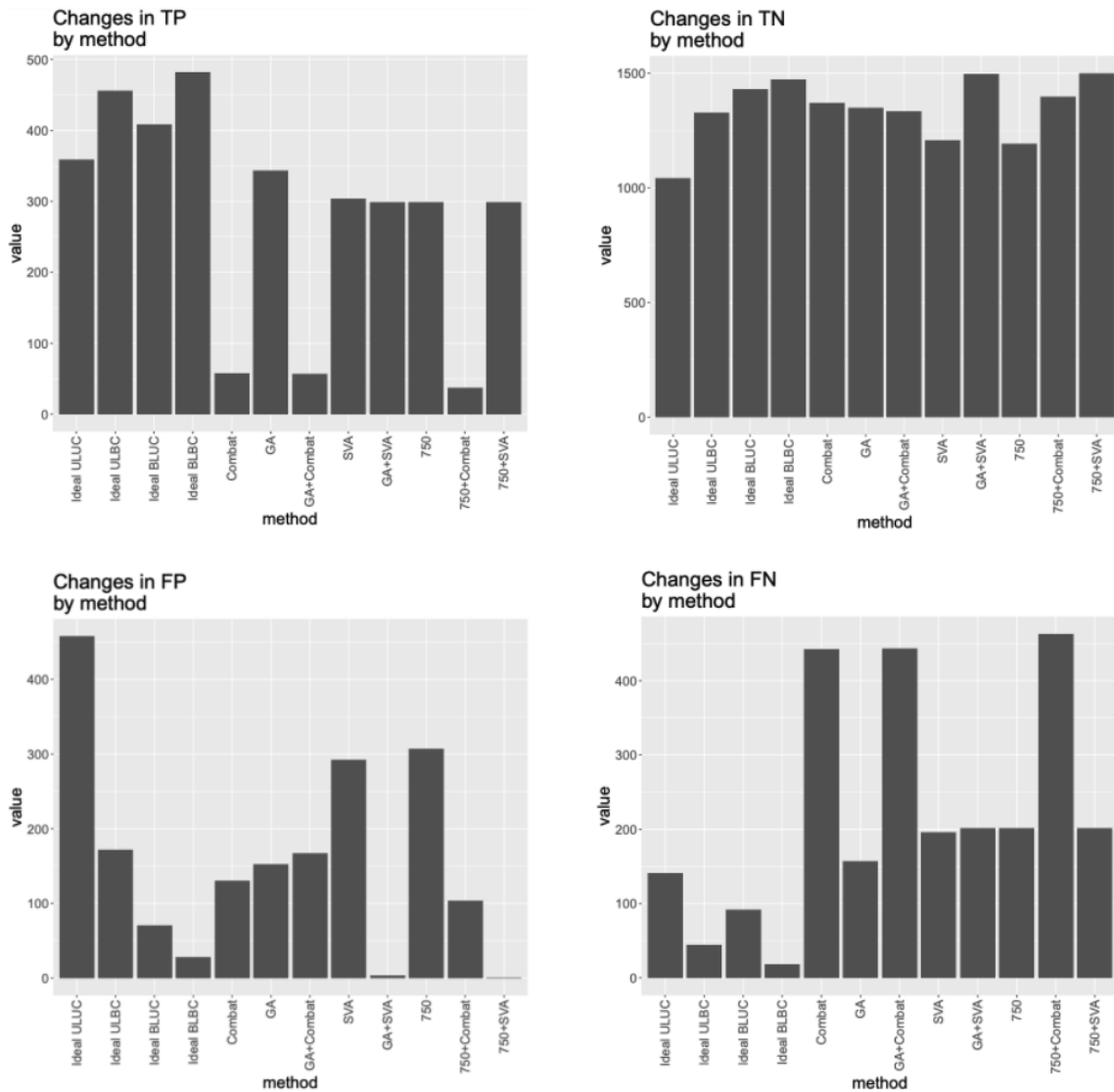
Figure 2: Gene Set Enrichment Analysis with and without Library and Covariate Balancing on Simulated Data.



a-c) Sensitivity, Specificity, and Accuracy for GSEA. Using the original unbalanced dataset (ULUC) as a base, three metrics were tested: applying only covariate balancing (ULBC) with either SVA or Combat, applying only library balancing (BLUC) with either the GA or with the original 750 balanced, and applying both library and covariate balancing (BLBC). These results are displayed in tandem with the ideal dataset, demonstrating ULUC when it is entirely free of either covariate or library effects, or both. GSEA was performed using 5 million randomly generated gene sets. Over 10,000 iterations, 100 gene sets were generated for each respective size gene set 100, 200, 300, 400, and 500; totalling for 500 total gene sets per iteration. Using the true DE genes from the simulated RNA-Seq count data the 500 total gene sets were labeled as enriched or unenriched, the predicted DE genes from each method were then used to predict whether each gene set is enriched or unenriched. The results of the true and predicted gene sets were used to compute Sensitivity, Specificity, and Accuracy metrics for covariate and library balancing.

Figure 3: DE Analysis on Simulated Data

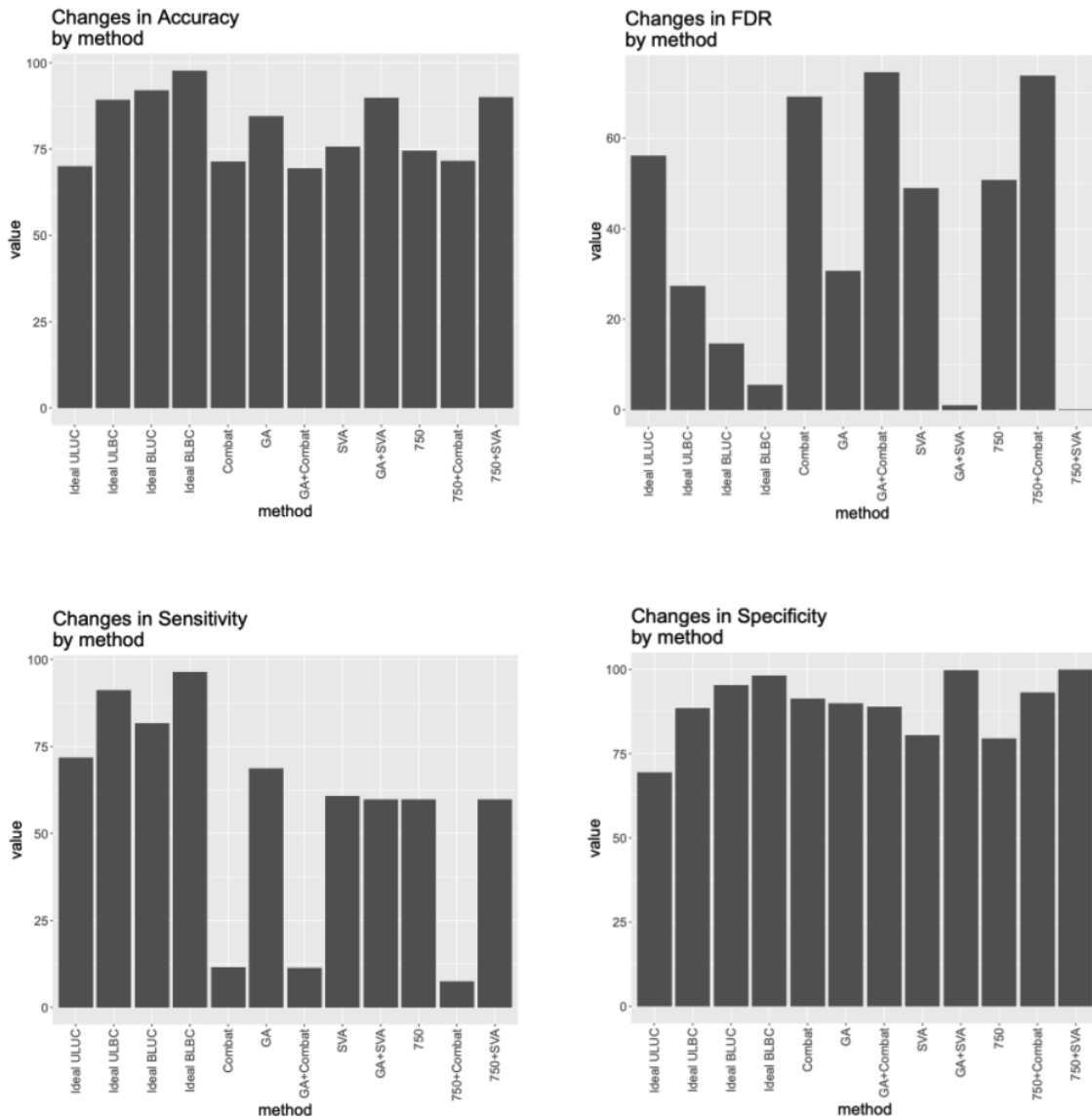
(a)



a) Categorizing DE Genes, TP, TN, FP, FN. Numbers from confusion matrices using the known DE genes and non-DE genes and the predicted DE genes and non-DE genes by balancing method. For post-sequencing balancing (covariate balancing) SVA and Combat are used, for pre-sequencing balancing (library balancing) the GA and the original 750 samples are used. **b) Changes in DEA Results.** Scores for accuracy, FDR, sensitivity, and specificity are computed for each method using confusion matrices from DE gene classification.

Figure 3: DE Analysis on Simulated Data, Continued

(b)



a) Categorizing DE Genes, TP, TN, FP, FN. Numbers from confusion matrices using the known DE genes and non-DE genes and the predicted DE genes and non-DE genes by balancing method. For post-sequencing balancing (covariate balancing) SVA and Combat are used, for pre-sequencing balancing (library balancing) the GA and the original 750 samples are used. **b) Changes in DEA Results.** Scores for accuracy, FDR, sensitivity, and specificity are computed for each method using confusion matrices from DE gene classification.

Figure 4: The Genetic Algorithm Process

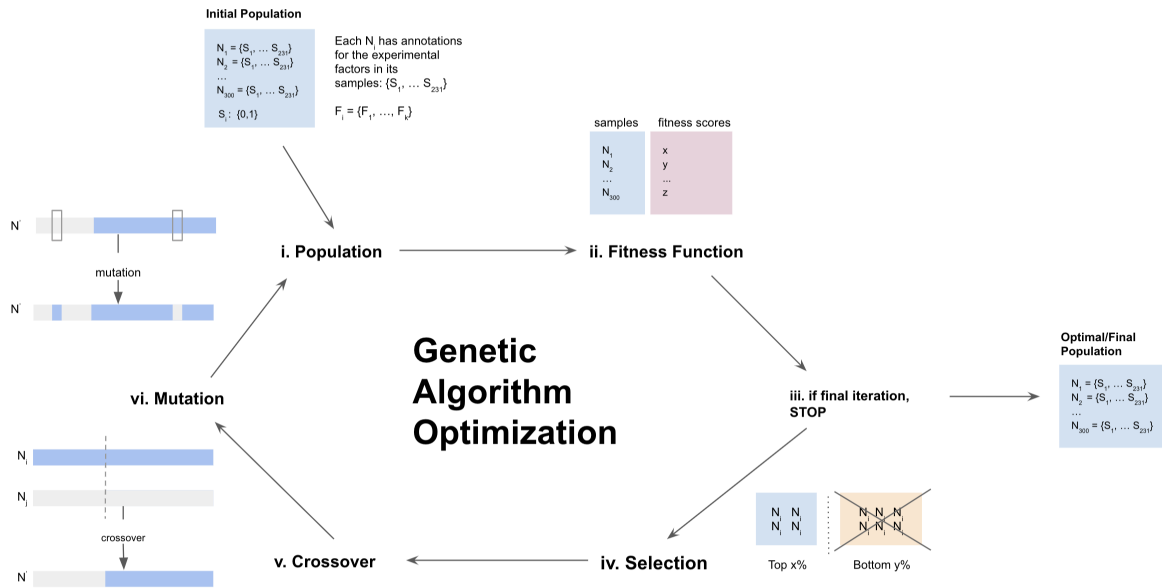
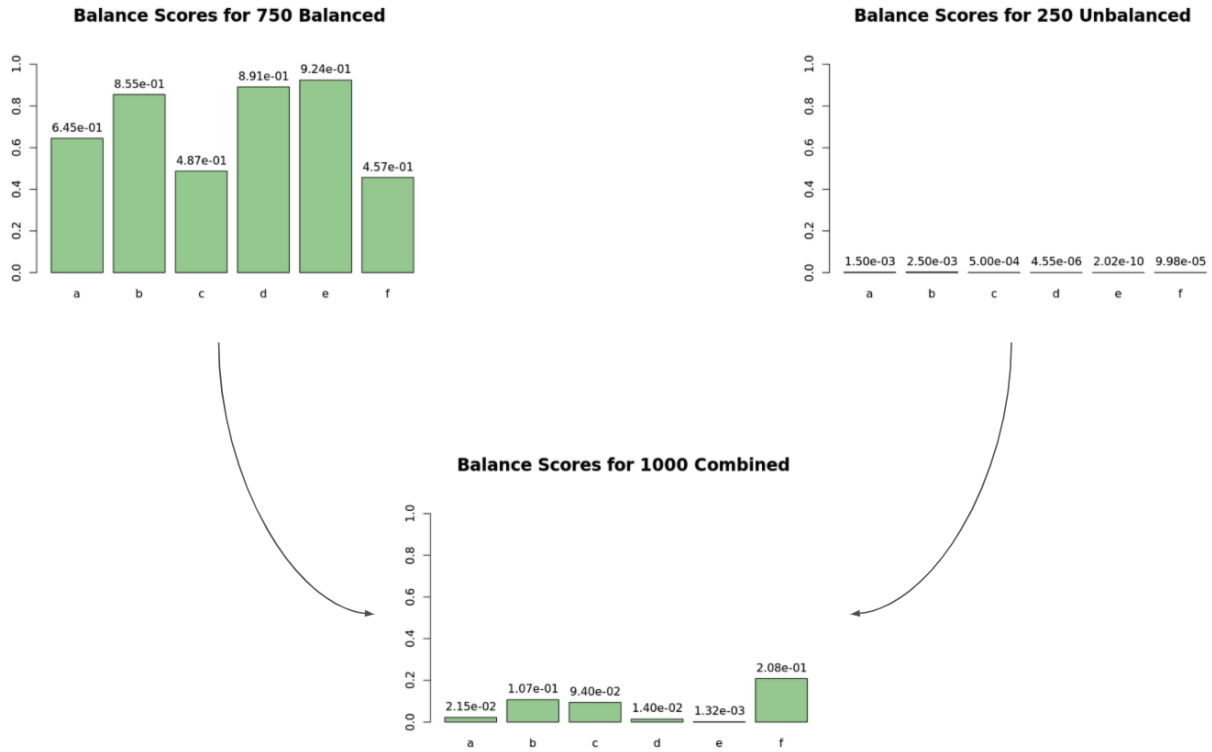


Figure 4 The GA-based optimization approach consists of six steps: (i) initial population—the initial population termed as $N = \{N_1, N_2, \dots, N_{300}\}$; (ii) fitness function—the function to assess how balance of the experimental factors in an individual solution N_i . For each N_i , each experimental factor is evaluated by performing the *ANOVA test* for continuous factors and by the *Chi-square test* for the categorical factors. The p-value of statistical test is considered as the ‘balance score’ of the testing experimental factor; (iii) stop criteria—the optimization process will be terminated after a predetermined number of iteration; (iv) selection— top 30% of individuals served as candidates for further crossover and mutation to generate a new individual of the next-generation population; (v) crossover— the child individual is generated by swapping two parent chromosomes with a single crossover point; and (vi) mutation— a mutation occurs by randomly changing the value of a chromosome in the child individual to maintain population diversity.

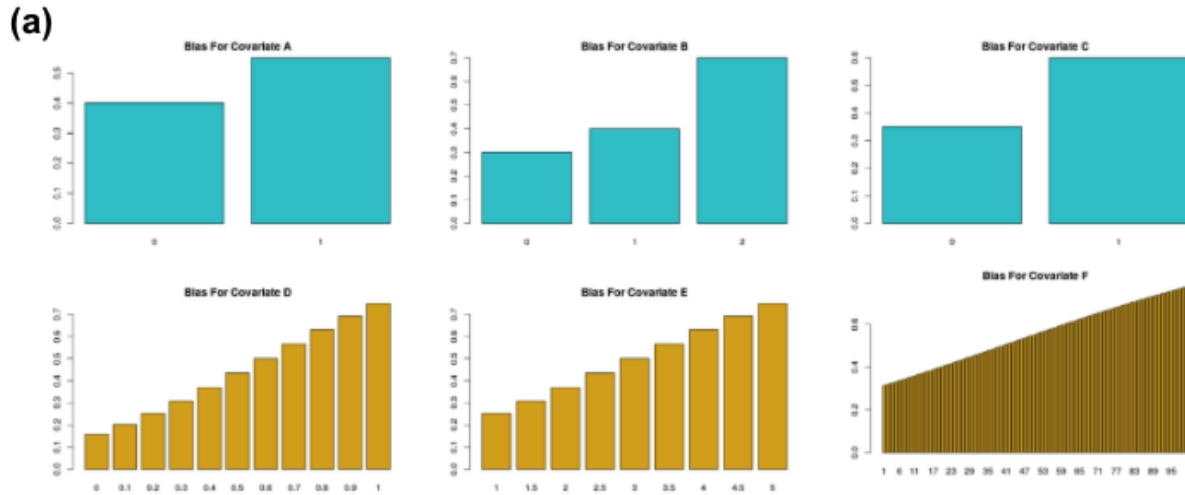
SUPPLEMENTARY FIGURES:

Supplementary Figure 1: Known Balanced Subset



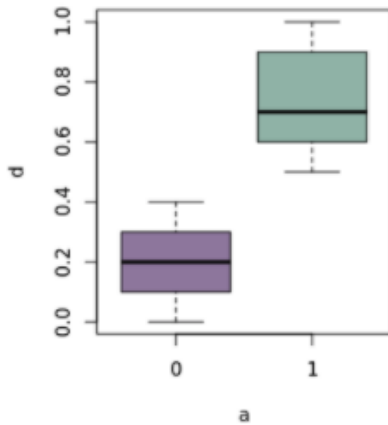
Known Balanced Subset. The 1000 samples were divided up into a balanced subset of 750 samples and an unbalanced subset of 250 samples. The two of these combined produced a final unbalanced dataset of 1000 samples. Continuous covariates (d-f) balance score was computed by fitting a model with the covariate value and the diagnosis. Categorical covariates (a-c) balance score was computed using a chi-squared test with the covariate value and the diagnosis.

Supplementary Figure 2: Covariate Distributions for 250 Unbalanced Samples

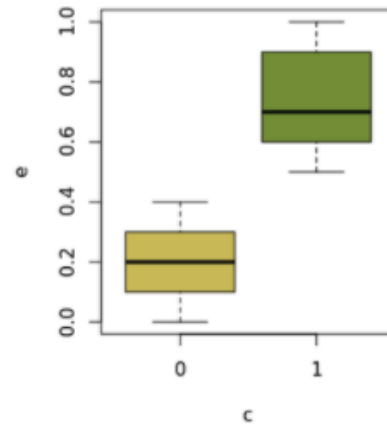


(b)

Covariate Distribution A & D
p-val: 2.2e-299

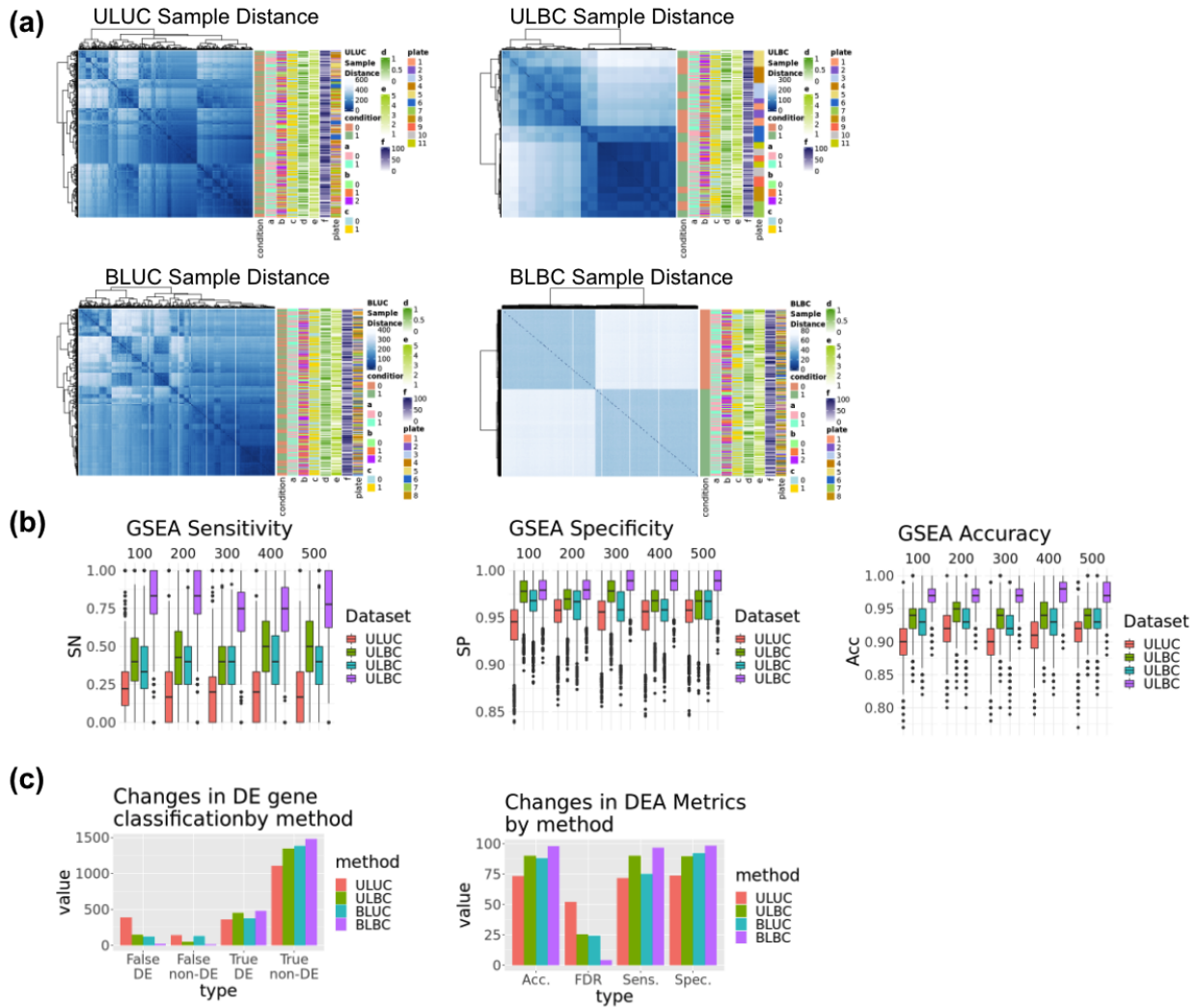


Covariate Distribution E & C
p-val: 1.08e-311



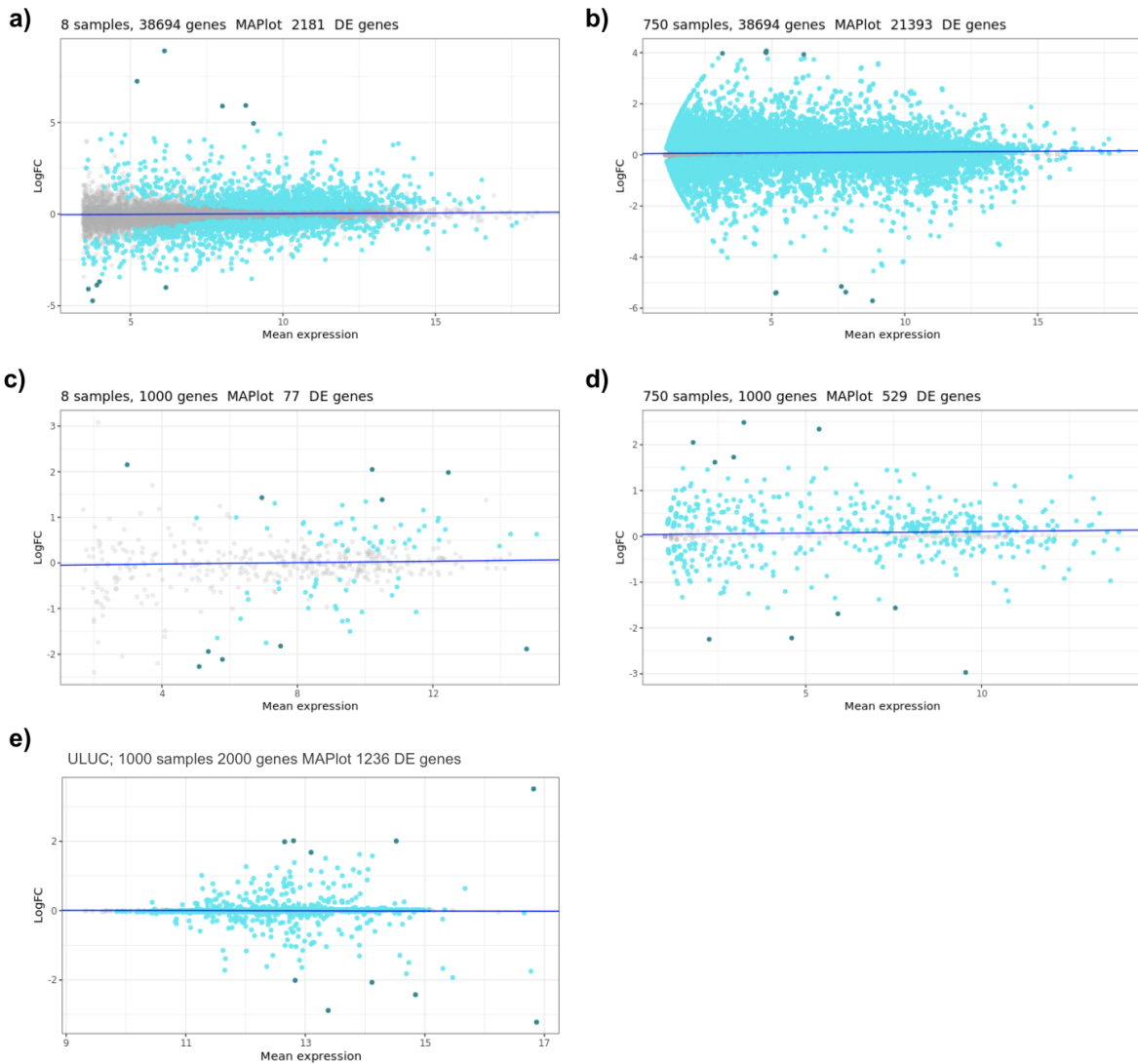
a) Covariate Distributions of Covariates based on Diagnosis. Covariate distributions without the posterior probability of interactions. Covariates were assigned to the 250 unbalanced samples according to the above biased distributions. For diagnosis 0, the above distributions were used, for diagnosis 1 covariates were assigned using the original distribution, dist , in the formula $\text{abs}(\text{dist} - 1)$. **b) Covariate Distributions for Covariates based on Interactions.** Covariate distributions for two interacting covariate pairs after covariates were simulated, p values computed using a linear model fitting a relationship between the two covariates.

Supplementary Figure 3: Ideal Metrics



Ideal Metrics for the four ideal datasets: Unbalanced Library and Unbalanced Covariates (ULUC), Unbalanced Library and Balanced Covariates (ULBC), Balanced Library and Unbalanced Covariates (BLUC), and Balanced Library and Balanced Covariates (BLBC). BLUC and BLBC use the 750 samples that are known to be balanced for all the covariates. In these four datasets balancing is Ideal, meaning that covariate/plate effect is fully removed. **a) Sample Distances for Ideal Metrics.** Heat Maps visualizing sample distances for the four balancing scenarios. Sample distances are computed using the euclidean distances between samples using their VST normalized RNA-Seq counts. **(b) GSEA for Ideal Metrics** Specificity, sensitivity, and accuracy scores for GSEA for the four datasets. **(c) DE Analysis for Ideal Metrics**

Supplementary Figure 4: LFC Patterns



a) Default Size. LFC by mean value for a dataset of airway scaled counts. **b) Increased Sample Size.** Using the original dataset of 8 scaled airway samples, the dataset was randomly upsampled to reach a sample size of 750 samples. **c) Decreased Number of Genes.** One thousand genes were randomly selected from the dataset in a) using a uniform distribution, the remaining genes were removed from the gene pool. **d) Decreased Number of Genes and Increased Sample Size.** Using the methods in b) and c) the original dataset in a) is modified to have 750 samples and 1000 genes. **e) Unbalanced Simulated MA plot.** 1000 samples with 2000 genes

TABLES:

Table 1. Overview of the samples used in the unbalanced library design of Simulated RNA-Seq experiments

Diagnosis		f			
0	491	1	8	51	15
1	509	2	2	52	10
a		3	7	53	12
0	491	4	7	54	15
1	509	5	9	55	11
b		6	10	56	11
0	338	7	8	57	10
1	335	8	5	58	11
2	327	9	10	59	7
c		10	16	60	11
0	508	11	7	61	9
1	492	12	12	62	14
d		13	6	63	14
1	91	14	11	64	16
2	103	15	11	65	10
3	99	16	15	66	10
4	98	17	7	67	7
5	100	18	7	68	12
6	105	19	11	69	4
7	63	20	9	70	12
8	98	21	19	71	9
9	79	22	11	72	3
10	92	23	14	73	12
11	72	24	9	74	13
e		25	7	75	9
1	108	26	10	76	15
2	125	27	3	77	9
3	154	28	8	78	15

Table 1. Overview of the samples used in the unbalanced library design of Simulated RNA-Seq experiments, Continued

e		f			
4	121	29	8	79	10
5	96	30	10	80	10
6	94	31	10	81	8
7	97	32	7	82	14
8	97	33	10	83	11
9	108	34	8	84	10
plates		35	18	85	9
1	95	36	11	86	5
2	81	37	11	87	9
3	96	38	8	88	9
4	103	39	4	89	4
5	105	40	8	90	12
6	99	41	10	91	10
7	109	42	11	92	9
8	95	43	12	93	12
9	108	44	7	94	12
10	109	45	12	95	17
		46	16	96	7
		47	4	97	6
		48	11	98	11
		49	14	99	9
		50	7	100	14

Citations

Akey, J., Biswas, S., Leek, J. Storey J. On the design and analysis of gene expression studies in human populations. *Nat Genet* 39, 807–808 (2007). <https://doi.org/10.1038/ng0707-807>

Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C(2011) Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS ONE* 6(2): e17238. <https://doi.org/10.1371/journal.pone.0017238>

Engström, P., Steijger, T., Sipos, B. Grant, G., Kahles, A., The RGASP Consortium, Räscht, G., Goldman, N., Hubbard, T., Harrow, J., Guigó, R., Bertone, P. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* 10, 1185–1191 (2013). <https://doi.org/10.1038/nmeth.2722>

Fomina-Yadlin D, Mujacic M, Maggiora K, Quesnell G, Saleem R, McGrew JT. 2015. Transcriptome analysis of a CHO cell line expressing a recombinant therapeutic protein treated with inducers of protein expression. *J Biotechnol* 212:106-15.

Fomina-Yadlin, D., Mujacic, M., Maggiora, K. Quesnell, G., Saleem, R., McGrew, J. Transcriptome analysis of a CHO cell line expressing a recombinant therapeutic protein treated with inducers of protein expression, *Journal of Biotechnology*, Volume 212, 2015, Pages 106-115, ISSN 0168-1656, <https://doi.org/10.1016/j.jbiotec.2015.08.025>.

Goldberg DE. Genetic Algorithm in Search, Optimization and Machine Learning. New York: Addison Wesley, 1989. ISBN:0201157675

Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, Gentleman R. (2016) Prediction and Quantification of Splice Events from RNA-Seq Data. *PLoS ONE* 11(5): e0156132. <https://doi.org/10.1371/journal.pone.0156132>

Hsu HH, Araki M, Mochizuki M, Hori Y, Murata M, Kahar P, Yoshida T, Hasunuma T, Kondo A. 2017. A Systematic Approach to Time-series Metabolite Profiling and RNA-seq Analysis of Chinese Hamster Ovary Cell Culture. *Sci Rep* 7:43518.

Lambert CG, Black LJ. Learning from our GWAS mistakes: from experimental design to scientific method. *Biostatistics*. 2012 Apr;13(2):195-203.

Lee, D., Cheng, A., Lawlor, N., Bolisetty, M., Ucar, D. Detection of correlated hidden factors from single cell transcriptomes using Iteratively Adjusted-SVA (IA-SVA). *Sci Rep* 8, 17040 (2018). <https://doi.org/10.1038/s41598-018-35365-9>

Leek JT, (2014). svaseq: removing batch effects and other unwanted noise from sequencing data *Nucleic Acids Research*, Volume 42, Issue 21, 1 December 2014, Page e161, <https://doi.org/10.1093/nar/gku864>

Leek, J., Johnson, E., Parker, H., Jaffe, A., Storey, J. The sva package for removing batch effects and other unwanted variation in high-throughput experiments *Bioinformatics*, Volume 28, Issue 6, 15 March 2012, Pages 882–883, <https://doi.org/10.1093/bioinformatics/bts034>

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K and Irizarry RA. (2010). Tackling the widespread and critical impact of batch effects in high throughput data. *Nature reviews. Genetics*, 11, 733–739.

Li, B., Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323 (2011). <https://doi.org/10.1186/1471-2105-12-323>

Melanie M. (1996). An introduction to genetic algorithms. Cambridge, MA: *MIT Press*. ISBN: 9780585030944.

Müller C, Schillert A, Röthemeier C, Trégouët D, Proust C, Binder H, Pfeiffer N, Beutel M, Lackner K, Schnabel R, Tired L, Wild P, Blankenberg S, Zeller T, Ziegler A, (2016) Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data <https://doi.org/10.1371/journal.pone.0156594>

Sebastiani P, Solovieff N, Dewan AT, Walsh KM, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wilk JB, Myers RH, Steinberg MH, Montano M, Baldwin CT, Hoh J, Perls TT. Genetic signatures of exceptional longevity in humans. *PLoS One*. 2012;7(1): e29848.

Varma S. 2020 Blind estimation and correction of microarray batch effect. *PLoS One* <https://doi.org/10.1371/journal.pone.0231446>

Vishwanathan N, Yongky A, Johnson KC, Fu HY, Jacob NM, Le H, Yusufi FNK, Lee DY, Hu WS. 2015. Global insights into the Chinese Hamster and CHO cell transcriptomes. *Biotechnology and Bioengineering* 112(5):965-976.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10(1):57-63.

Yang H, Harrington CA, Vartanian K, Coldren CD, Hall R, and Churchill GA. (2008). Randomization in Laboratory procedure is key to obtaining reproducible microarray results. *PLoS ONE*, 3, e3724.

Yuk IH, Zhang JD, Ebeling M, Berrera M, Gomez N, Werz S, Meiringer C, Shao Z, Swanberg JC, Lee KH and others. 2014. Effects of copper on CHO cells: insights from gene expression analyses. *Biotechnol Prog* 30(2):429-42.

Zhang, C., Zhang, B., Lin, LL. Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* 18, 583 (2017). <https://doi.org/10.1186/s12864-017-4002-1>

Longjian Zhou, Andrew Chi-Hau Sue, Wilson Wen Bin Goh, Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects?, *Journal of*

Genetics and Genomics, Volume 46, Issue 9, 2019, Pages 433-443, ISSN 1673-8527,
<https://doi.org/10.1016/j.jgg.2019.08.002>.