Gene Expression Deconvolution with Subpopulation Proportions

MSc Research Paper

Department of Computer Science

University of Toronto

Chris Cremer

Supervised by Dr. Quaid Morris

February 2016

I. ABSTRACT

Personalized cancer strategies are currently being hindered by intratumor heterogeneity. One source of heterogeneity, clonal evolution, can lead to genetically distinct subpopulations within a sample. Through the use of subclonal reconstruction methods, we can obtain estimates of the subpopulation proportions within a single sample. Here, I leverage these proportion estimates by incorporating it into the deconvolution of tumour gene expression data in order to estimate the subclone specific gene expression profiles. The deconvolution's effectiveness is demonstrated on simulated data and is analyzed on real gene expression data. I hope that future applications of this method to clinical data will improve our understanding of tumour evolution and help the development of improved treatments.

Table of Contents

II. INTRODUCTION

Tumors frequently display substantial intra-tumor heterogeneity [1]–[3]. This heterogeneity manifests in many phenotypic features, including cellular morphology, metabolism, motility, proliferation, metastatic potential, and gene expression. One model through which tumours develop their heterogeneity is clonal evolution [4]–[6]. This model can be regarded as a process of Darwinian evolution, where selection forces work on a population of cells with different heritable traits and after a large number of cell divisions, results in the emergence of multiple genetic mutants (subpopulations) that are better suited to thrive in their environment. Improvements in next-generation sequencing of tumor samples has led to the identification of somatic mutations at low allelic fractions, which has opened the way for new approaches to model the evolution of individual cancers [7]–[9].

Gene expression profiling has been shown to predict clinical outcomes in many cancer types [10]–[12]. The variability in gene expression caused by tumor heterogeneity interferes with the development and clinical use of gene signatures [13],[14]. There are numerous computational methods that aim to deconvolve gene expression profiles to reduce the effect of tumor heterogeneity [15]. In this report, I present a method to leverage tumor subpopulation frequency information to estimate subclone specific gene expression profiles, noting however, that the relationship between population frequencies and expression profiles is still unclear.

This report is structured as follows. First, I provide background on tumour gene expression deconvolution, subclonal reconstruction and the problems accompanied with combining the two. Next, I introduce a method that incorporates subclonal frequency information into the deconvolution of tumor samples so that we can estimate subclone specific gene expression profiles. In the Results section, I examine the performance of this method on

simulated heterogeneous tumor data, as well as real breast cancer gene expression data. Finally, I discuss the assumptions made by this model and the impact they can have on the results.

III. BACKGROUND

A. Gene Expression Deconvolution

The expression of most genes varies across different cell subsets, which implies that the measured abundance of any transcript is confounded by the composition of the sample. More precisely, the total measured abundance of a gene in a sample can be attributed to different factors, such as that due to the characteristic condition of a sample (e.g. type of cancer, etc.), that due to the individual variation or technical measurement variation, and that due to the average abundance of a gene as a function of the underlying cell subsets in a sample and their relative proportions [15]. The last factor, the sample heterogeneity, is the variation in gene expression that we try to capture by reporting differences in proportions of cell subsets.

Gene expression heterogeneity can be modelled by a system of linear equations such that each sample is composed of a convex combination of hidden 'pure' profiles. This linear system can be expressed in matrix notation as:

$$X = WZ$$

where X is the gene expressions of the observed samples, W is the proportion of each component in each sample, and Z is the expression profiles of the hidden components. Thus X is N samples by D genes, W is N samples by K components, and Z is K components by D genes. We can assume that D > N > K.

Given the gene expressions of a set of samples X, the challenge is to solve for W and Z. The constraints for this problem are that each element of W must be between zero and one and

each row must sum to one since they represent proportions. The elements of Z must be non-negative because they represent gene expressions, which cannot be below zero. A problem like this can be approached with non-negative matrix factorization (NMF) [16] but with the added constraint that each weight vector needs to sum to one. Therefore we are solving a convex optimization problem of this form:

$$min_{W,Z} \|X - WZ\|$$

$$s.t. W_{ik}, Z_{kj} \geq 0$$

$$1^T W_i = 1$$

for i=1…N, k=1…K, and j=1…D. The norm that is minimized is often chosen to be the L2 norm due to its simplicity and its efficient computational properties.

There are numerous computational tools available for the deconvolution of genomic data from heterogeneous samples. Present computational methodologies for extracting cell type-specific information differ by the type of input they require, the type of the output they offer, and the assumptions they make in the model [15]. For example, ISOpure [13] addresses the effects of normal tissue contamination in clinical tumor specimens by taking as input a panel of healthy tissue expression profiles in order to generate a purified cancer profile for each tumor sample, and an estimate of the proportion of RNA originating from cancerous cells. Similarly, CIBERSORT [17] takes as input the profiles of the hidden cell types in order to characterizes the cell composition of complex tissues from their gene expression profile. Lähdesmäki et. al. [14] developed a model to remove the effects of sample heterogeneity by taking as input accurate estimates of the mixing percentages of different cell types and outputs the estimates of the expression values of the pure (non-heterogeneous) cell samples. Similarly, DSection [18] is a probabilistic model for deconvolution which uses an estimate of the proportions as a prior and

determines the proportions and gene expression profiles by maximizing the likelihood of the data.

Other than computational approaches to purifying tumor profiles, there exists post-operative methods for sample purification, such as laser capture micro-dissection or cell sorting. These approaches require specialized equipment, are costly, delay the diagnostic cycle, and cannot always be used. [19]. Thus computational purification of tumor samples are important to avoid these problems.

B. Subpopulation Frequency Information

Through the successive iterations of expansion and selection over the lifetime of a single tumor, genetically diverse subclonal populations (subpopulations) of cells will evolve from a single pro-genitor population [20], [21]. The selective sweeps that cause subpopulation expansion will drive the various intratumor subpopulations to display differences in their frequency of driver and passenger simple somatic mutations (SSMs) [22]. Consequently, subpopulations are defined not only by the small number of oncogenic driver mutations but also by a larger number of passenger mutations acquired before the driver mutations. To infer the population structure of heterogeneous tumors, subclonal reconstruction algorithms use the measured variant allelic frequency (VAF) of their somatic mutations [9], [23], [24], [25]. Some subclonal reconstruction methods will combine the VAFs with the inferred copy number variations (CNVs) to identify genomic regions with an average ploidy that differs from normal [26], [27], [8]. Through the use of these subclonal reconstruction methods, we can obtain information about the genetically distinct populations within each sample. Part of the information we obtain from these methods includes an estimate on the number of subpopulations and their proportions within each sample.

We hypothesize that these cancerous subpopulations differ not only in their genetic mutations but also in their population gene expression profiles. Thus we can use these proportion values to push the gene expression deconvolution towards meaningful latent factors.

C. Combinatorial Optimization Problem

From the use of the subclonal reconstruction based on DNA sequencing data, we obtain, for each sample i, a set S_i of values that correspond to the proportions of the subpopulations within the sample. However one challenge that we are faced with is assigning the proportions to each of the hidden profiles. We assume that the number of possible hidden profiles is larger than the number of subpopulations within a single sample. In other words, we know the number of hidden profiles and their proportions within each sample but we don't know to which hidden profile of matrix Z each proportion refers to. The set S_i is padded with zeros so that its size is equal to the number of components K. Therefore we must find the permutation vector P_i of set S_i that minimizes the reconstruction error. Thus we are looking for the permutation vector that minimizes,

$$min_{P_i} \|Z^T P_i - X_i\|$$

This is a combinatorial optimization problem. The most naive method to solve this would be to perform an exhaustive search over all permutations of S_i. The running time would be K permute Q, $\frac{K!}{(K-Q)!}$, where K is the number of components and Q is the number of non-zero elements in the set S_i. This could be feasible for small values of K, however, it is unclear how many subpopulations actually exist within different types of cancers, and thus we cannot use the exhaustive search since it would limit our ability to model larger numbers of components.

To reduce the computational burden of combinatorial problems, there is occasionally

relaxations that can be made to the original problem so that it becomes a convex optimization problem. For instance, semidefinite programming (SDP) has wide applicability in combinatorial optimization. A number of NP−hard combinatorial optimization problems have convex relaxations that are SDPs [28]. Another example is decoding linear error correcting codes which have been shown to be solved by linear programming (LP) [29]. Therefore, it is possible that a relaxation can be made for our assignment problem so that we are not restricting the feasible region to be a member of the permutation set.

IV. METHODS AND MATERIALS

In this section, I describe the procedure that is used to deconvolve the gene expression samples using the proportion information. I explain how the combinatorial optimization problem is dealt with so that it becomes a convex optimization problem. I also describe the process used to simulate the data that will be used in the experiments.

A. Deconvolution Procedure

The goal of the deconvolution procedure is to retrieve the gene expression profiles of the hidden subpopulations and assign their proportions within each sample. More specifically, we are given matrix X, which is the gene expressions of the observed samples and a set S_i for each sample i which correspond to the proportions of the subpopulations within sample i. With this information, we are estimating the matrix W, which is the proportion of each component in each sample, and the matrix Z, which is the expression profiles of the hidden components. Note, however, that the set S_i is not ordered, meaning that we do not know which proportion belongs to which expression profile.

The first step of the deconvolution is to initialize the Z matrix with samples from X. The samples are selected in the following manner: a sample is randomly selected from X, then the sample that is most different (based on Euclidean distance) from the first sample is selected, then the sample that is most different from the average of the previous samples is selected and so on. This selection procedure is performed so that the Z matrix begins with samples that give a better representation of the space of gene expressions profiles, rather than samples that are similar to each other. In order to reduce the effect of poor initializations, the whole deconvolution is performed three times with different Z initializations and the deconvolution with the lowest final error is selected.

For the second step, we alternate between solving W then Z by minimizing the L2 norm while staying within the constraints. To be more specific, first, we solve for W using X and Z, then we solve for Z using X and W, and repeat. Note that each row of W must sum to one because it represents the percentage of each component found in that sample. To solve for each row W_i while having every row sum to one, I implemented a modified non-negative least-squares (NNLS) algorithm. I used a quadratic cone program solver to solve the following optimization problem and its constraints:

$$min_{W_i} \|Z^T W_i - X_i\|_2$$

$$s.t. W_{ik} \geq 0, \;\; 1^T W_i = 1$$

The first equation specifies that we are minimizing the L2-norm of the reconstruction error for each W_i. The constraints are that each element of W_i must be non-negative and each W_i must sum to one. Without the 'sum to one' constraint, this optimization would be equivalent to NNLS. Once we solve for matrix W, we use X and W to solve for each Z_i^T. In this case, we can use regular NNLS since each element represents gene expressions, which need to be non-negative

but don't need to sum to one. Thus the equation being optimized for each column of Z is:

$$min_{Z_i^T} \left\| WZ_i^T - X_i^T \right\|_2$$

$$s.t. Z_{ik} \geq 0$$

Once Z is optimized, we return to W, and repeat until convergence. If there are no prior proportions provided, the deconvolution stops here. In the Results section, I refer to this method as 'Without_Props'. This method would be similar to NMF but with the constraint that each row of W sums to one. If there is a set of prior values to fit, then we solve for W as before then fit the proportions to it. The prior proportions are fitted to the weight vectors using either the relaxation method or the exhaustive search. The relaxation method will be described in the next section. If the exhaustive search method is used, then we must try all permutations of the prior proportion vectors and select the one that minimized the error. After fitting, the Z matrix is optimized in the same manner as before. The method repeats the fitting of W followed by solving Z until some tolerance is reached. In the end, we have factorized the X matrix into two lower dimensional matrices W and Z and we have used the prior proportions to modify W, which in turn modifies Z. To summarize here are the steps used for the deconvolution:

Deconvolution with Proportions Procedure

1) Initialize Z with samples from X
2) Alternate between optimizing W and Z
3) Optimize W using modified NNLS
4) Assign prior proportions S_i to W_i, $\mathrm{i} = 1 \ldots \mathrm{N}$
5) Optimize Z using NNLS
6) While error after step 5) continues to decrease, repeat steps 3) to 5)

B. Combinatorial Optimization Relaxation

Let S_i be the set of prior proportions for sample i. For the exhaustive search, we need to test all the permutations of this set, which limits the number of components that can be modeled. Instead of evaluating all the permutations of S_i, we will solve for W using the modified NNLS algorithm then fit W_i to the nearest permutation vector. This is equivalent to projecting X_i onto Z^T, resulting in W_i then finding the nearest permutation vector to W_i. The feasible space is constrained by the non-negativity inequality and the affine equality of summing to one. The projection is equivalent to the modified NNLS, which is formulated as the following quadratic program (QP),

$$min_{W_i} \|Z^T W_i - X_i\|_2$$

$$s.t. W_{ik} \geq 0, \ 1^T W_i = 1$$

This projection step ignores the prior set of proportions. The next step is to fit W_i to the permutations of S_i so that it minimizes the same objective function. In other words, we are going to swap the values of W_i with those of S_i in such a way that minimizes $\|Z^T W_P - X_i\|_2$, where W_P is the new vector composed of only the values of S_i and zeroes so that the length of W_P is the same length as W_i. We don't need to concern ourselves with the constraints because the set S_i already satisfies them.

To fit S_i to W_i, first we sort the set S_i. Next, the largest value of W_i is replaced with the largest value of S_i, and the second largest value of W_i is replaced with the second largest value of S_i, and so on. I claim that this procedure will result in the optimal W_P vector, ie. $min \|Z^T W_P - X_i\|_2$. See Appendix A for a partial proof that this fitting procedure is optimal in this situation. I call this whole procedure a combinatorial relaxation because the projection step is a relaxation of

the combinatorial optimization problem and may not result in the same solution as the exhaustive search.

The purpose of doing the relaxation is to allow for the model to accommodate more components. The addition of components to the model is equivalent to assuming that there exists more possible subpopulations within a single cancer type. In order to claim that the combinatorial relaxation is sufficiently accurate to replace the exhaustive search, I need to define what is meant by 'sufficiently accurate'. To determine whether the relaxation performs as well as the exhaustive search, I will use Welch's t-test. This is a two-sided significance test where the null hypothesis is that the two methods have identical average performances. The test does not assume that the populations have equal variance. Welch's t-test defines the statistic t by the following formula [30]:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

where $\bar{X}$ is the sample mean, s^2 is the sample variance, and N is the sample size. The degrees of freedom v associated with this variance estimate is

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\nu_1} + \frac{s_2^4}{N_2^2\nu_2}}.$$

We will reject the null hypothesis if the p-value is less than 95%. In other words, if the p-value is greater than 95%, I will claim that the combinatorial relaxation is sufficiently similar to the exhaustive search.

C. Simulated Data

In order to test my deconvolution algorithm, I created simulated data so that we knew the ground truth and could evaluate the effectiveness of the model. In an effort to create samples that mimic the reality of heterogeneous gene expression tumor samples, I created artificial samples in-silico by mixing RNASeq data of breast invasive carcinomas from the Cancer Genome Atlas (TCGA). For each sample, a weight vector is made by random sampling from a uniform distribution between zero and one then normalized so that each weight vector sum to one. In order to make the weight vectors more realistic, I added a sparsity parameter. This parameter was set so that each weight vector has on average β non-zero values, where β is the 'sample heterogeneity' parameter. Thus, if K is the number of hidden profiles in the model, then the probability that any one of the hidden profiles is found in any specific sample is $\frac{\beta}{K}$. The average number of subpopulations measured in my real breast cancer data is 3.3, accordingly β will be set to 3 or 4 for the next experiments. The gene expression profiles of the components matrix Z were created by selecting random breast cancer samples. Finally, X is created by the dot product of W and Z plus the addition of noise. Thus we know the identity of the weights and profiles that went into making each sample of X.

Two types of noise were added to the data: sample noise and prior noise. Sample noise is supposed to mimic the individual sample variation and technical measurement variation. Sample noise was introduced to the gene expressions of each sample (X_{ij}) by adding values randomly sampled from the distribution $N(0, \sigma_j^2 * v)$, with σ_j^2 equal to the variance of gene j, v is the noise parameter, and $\sigma_j^2 * v$ is the standard deviation of the normal distribution. Prior noise is meant to represent inaccuracies in the prior estimates. For each prior proportion, random values sampled from the distribution $N(0, p^2)$, where p is the prior noise parameter and p^2 is the standard

deviation of the normal distribution. For either types of noise, the elements were thresholded at zero, meaning, if the value after the addition of noise is below zero, the value is returned to zero since gene expressions and proportions cannot be negative.

D. Preprocessing and Error Measurement

Prior to running the deconvolution, there are several preprocessing steps applied to the data. The purpose of these steps is mainly to speed up the computations. First, only the top 50% of genes with the highest mean expression are kept. Genes with low expressions are more susceptible to small measurement inaccuracies so this dimensionality reduction may reduce noise. Next, each gene is divided by its mean expression so that each gene's mean is scaled to one, while retaining its coefficient of variation. Finally, only the top D genes with highest variance are kept, where D is a parameter selected for each experiment. This step is meant to remove the features (genes) that do not provide much information (low variance) as well as speed up the computations.

Once the deconvolution is complete, we need a way to compare its prediction to the truth. The deconvolution will output a predicted weight matrix W_P as well as a predicted component (gene expression profiles) matrix Z_P. With the simulated data, we know the real proportions W_R and the real gene expressions Z_R. However, the order of the components in the prediction may not be the same as the real components. Consequently, I match each predicted component with a unique real component so that $\|Z_P - Z_R\|_2$ is minimized. This assignment problem is solved using the Hungarian method. The ordering of the components in W_P and Z_P are rearranged according to the Hungarian method's output. Since the deconvolution is an unsupervised problem, the performance of the deconvolution is based on the correctness of the predicted W and Z matrices. Specifically, the W error is the Frobenius norm of the difference between the

predicted weights and the real weights ($\|W_P - W_R\|_2$) and the Z error is the Frobenius norm of the difference between the predicted hidden profiles and the real hidden profiles ($\|Z_P - Z_R\|_2$).

V. RESULTS

The results section is structured as follows. First, I show that the combinatorial relaxation is sufficiently accurate to replace the exhaustive search, so that the relaxation can be used in the deconvolution model instead of the brute force exhaustive search. Second, I show how well the deconvolution performs with respect to different data variables, including sample noise, number of samples, number of genes, sample heterogeneity, and prior noise. Finally, I apply the deconvolution to real data and examine the assumptions made regarding the data.

A. Combinatorial Relaxation

In this first experiment, I examined the relationship between the amount of noise added to the data X and the performance of the methods. Performance is based on the method's ability to recover the latent matrices, so $\|W_P - W_R\|_2$ and $\|Z_P - Z_R\|_2$. The three methods being tested include the projection without the fitting of the proportions, the projection with the proportions (relaxation), and the exhaustive search. It's conceivable that the relaxation method may work when the problem is easy, but fail when there is a large amount of noise. The following are the settings of the parameters used in the experiment: 50 samples, 500 genes, 3 heterogeneity, 5 components, 0.5 sample noise, and 0.1 prior noise. I use 5 components because that is limit of the exhaustive search. For each experiment, I averaged over 10 iterations.

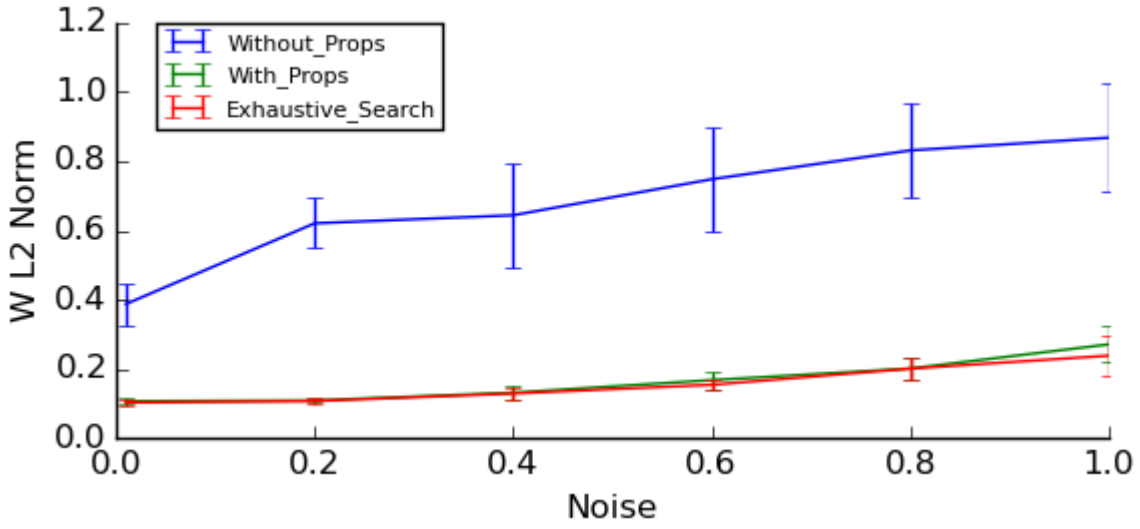

Fig. 1 Comparison of the deconvolution with and without subpopulation proportion (prop) information and the exhaustive search for selecting the weight vector for each sample. Sample noise was introduced by adding values randomly sampled from the distribution N(0,noise^2). 'W L2 Norm' is the Frobenius norm of the difference between the predicted weights and the real weights ($\|W_P - W_R\|_2$).

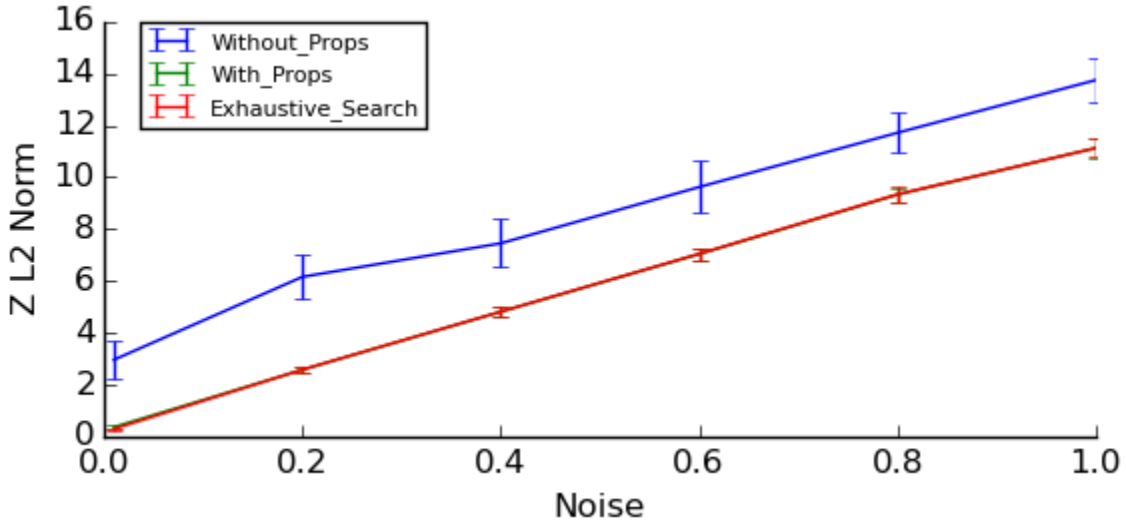

Fig. 2 Comparison of the deconvolution with and without subpopulation proportion (prop) information and the exhaustive search for selecting the weight vector for each sample. Sample noise was introduced by adding values randomly sampled from the distribution N(0,noise^2). 'Z L2 Norm' is the Frobenius norm of the difference between the predicted hidden profiles and the real hidden profiles ($\|Z_P - Z_R\|_2$). The green line is hidden behind the red.

Fig. 1 is a plot of the error in the prediction of weight matrix W using the various deconvolutions. As the noise increases, the error increases but what's important is that the deconvolution with proportions and the exhaustive search are nearly identical. The p-value of the 10 error samples for the two methods ('With_Props' vs 'Exhaustive Search') at noise =1 is

98.5%, meaning that the difference between the two is not significant. Fig. 2 shows that the expression profiles predicted by the relaxation and the brute force method are nearly identical (green line is underneath the red). These plots indicates that the relaxation appears to be a valid alternative to the exhaustive search at the current settings of the parameters. Replacing the brute force approach with the combinatorial relaxation approximation will allow for the model to accommodate more components. For the future experiments, the projection and fit method is used instead of the exhaustive search.

B. Deconvolution Performance

I test the performance of the deconvolution with and without the proportion information to examine the improvements the proportions bring to the accuracy of the deconvolution. Unless otherwise noted, the following are the default settings of the parameters used in the experiments: 200 samples, 1000 genes, 4 heterogeneity, 20 components, 0.5 sample noise, and 0.1 prior noise. See Methods section for a description of the parameters. Each experiment is averaged over 10 iterations and each iteration stopped converging when the change in error was less than 0.01.

Error vs Sample Noise

As expected, Fig. 3 and Fig. 4 show that as sample noise increases, the error in W and Z increase. The difference between the deconvolution with (With_Props) and without (Without_Props) the proportions stays nearly constant as noise increases for both the W and Z errors. Comparing Fig. 1 and Fig. 3, we see that the error more than doubles when increasing the number of componets from 5 to 20.

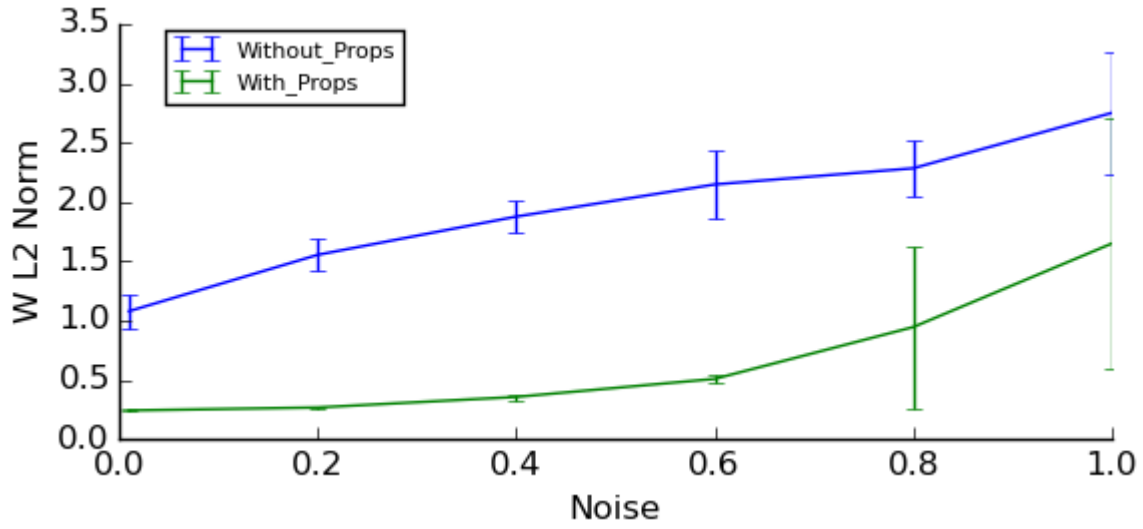

Fig. 3 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the amount of additive sample noise increases. 'W L2 Norm' is the Frobenius norm of the difference between the predicted weights and the real weights ($\|W_P - W_R\|_2$).

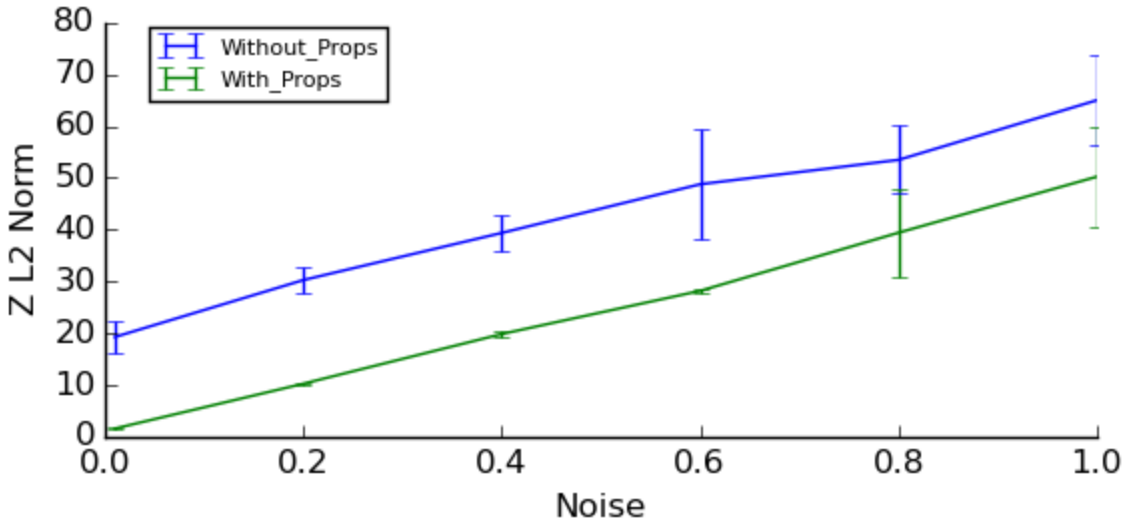

Fig. 4 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the amount of additive sample noise increases. 'Z L2 Norm' is the Frobenius norm of the difference between the predicted expression profiles and the real expression profiles ($\|Z_P - Z_R\|_2$).

Error vs Number of Samples

As the number of samples increases, the deconvolution gains more information regarding the identity of the hidden profiles. Consequently, the error in predicting the proportions and the expression profiles decreases with more samples as seen in Fig. 5 and Fig. 6. The deconvolution

with proportion information has lower error than without proportion information by a constant amount. Moreover, the improvement levels off at a certain point, in this case, at 200 samples. There are 20 hidden components in the simulated data, so the deconvolution seems to require that the fraction of hidden components to observed samples be less than 10% given the other settings of the parameters of the data.

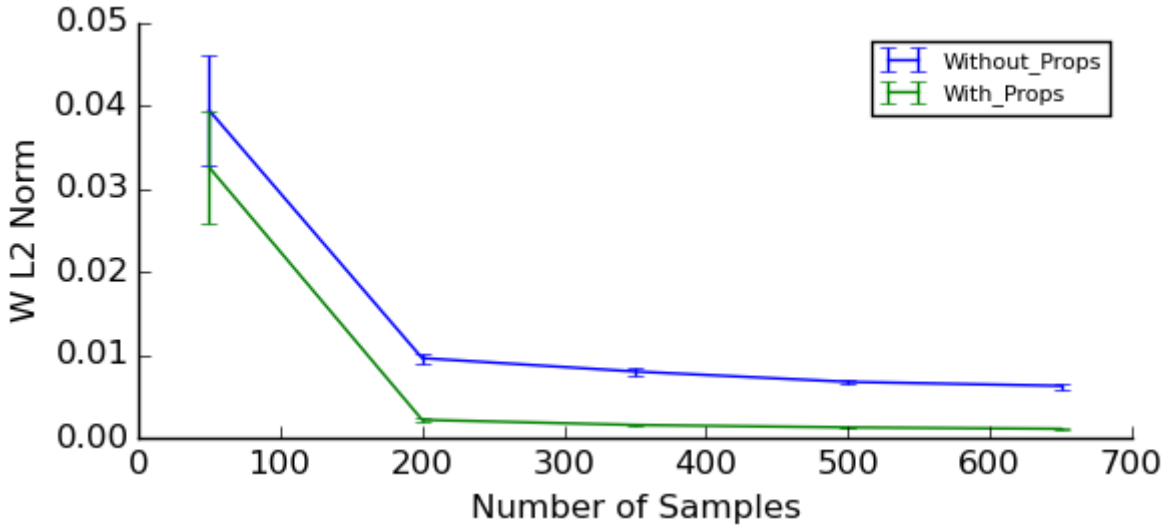

Fig. 5 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the number of samples increases. 'W L2 Norm' is the Frobenius norm of the difference between the predicted weights and the real weights divided by the number of samples, thus the y-axis is the error per sample ($\|W_P - W_R\|_2 / N$).

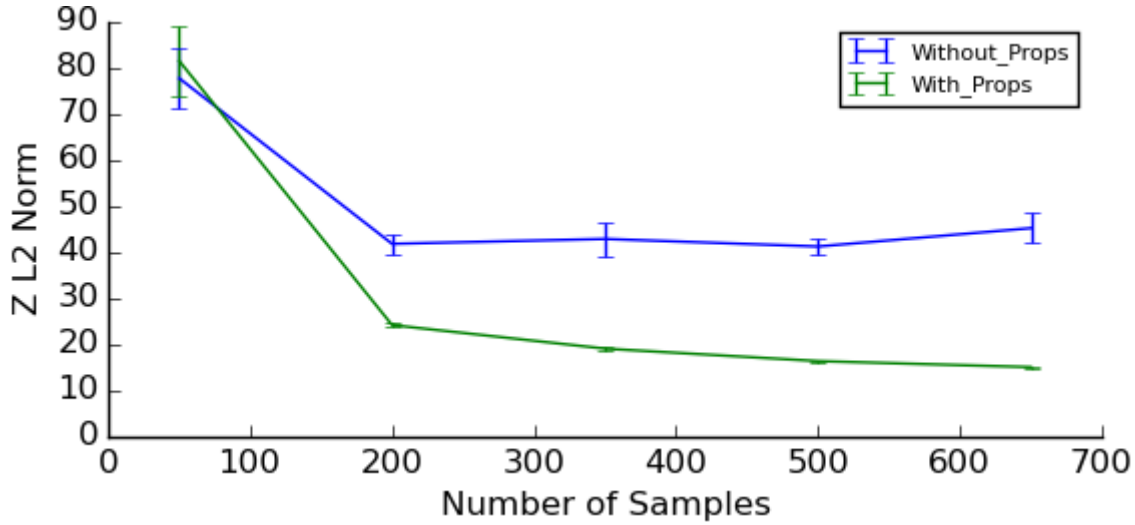

Fig. 6 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the number of samples increase. 'Z L2 Norm' is the Frobenius norm of the difference between the predicted expression profiles and the real expression ($\|Z_P - Z_R\|_2$).

Error vs Features

The number of features corresponds to the number of genes kept after preprocessing of the data. Fig. 8 demonstrates that with less information (<400 genes), the subpopulation propositions significantly help the deconvolution identify the hidden gene expression profiles. With excessive information (>400 genes), the benefit that the proportions provide is less significant. The improvement in accuracy of the predicted weights that the proportion information provides is relatively constant with the increase in genes, shown by Fig. 7.

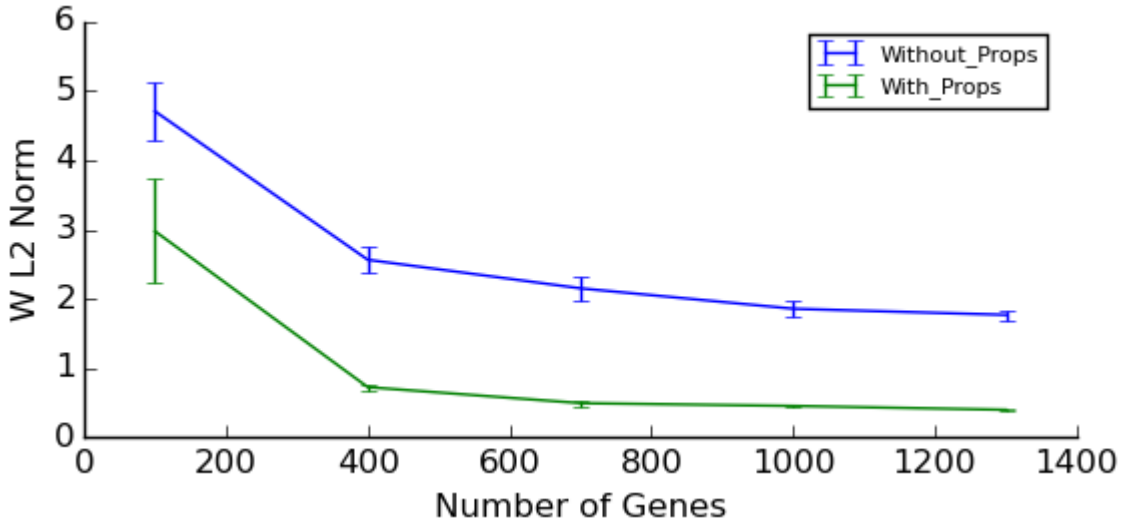

Fig. 7 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the number of features (genes) increase. 'W L2 Norm' is the Frobenius norm of the difference between the predicted weights and the real weights ($\|W_P - W_R\|_2$).

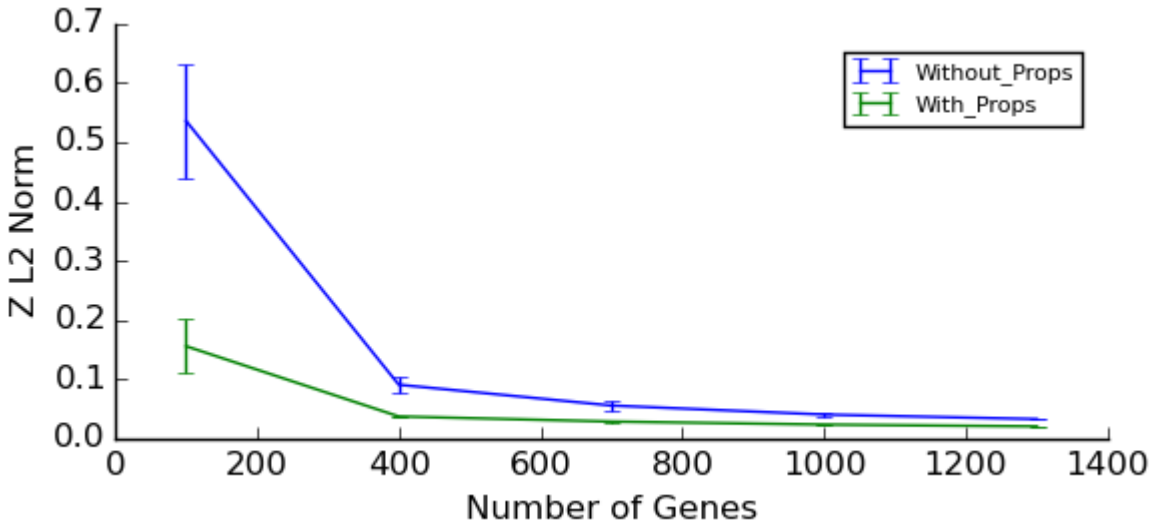

Fig. 8 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the number of samples increase. 'Z L2 Norm' is the Frobenius norm of the difference between the predicted expression profiles and the real expression divided by the number of genes ($\|Z_P - Z_R\|_2 \,/\, D$).

Error vs Prior Noise

Prior noise is noise added to the proportions, thus the deconvolution without proportion information is unaffected by the prior noise. Fig. 9 indicates that when the prior noise surpasses 0.3, then the predicted weights of the deconvolution without proportions is more accurate than the deconvolution with the noisy proportions. The prediction of the hidden gene expression profiles (Z) is slightly more robust to noise in the proportions compared to W, since the deconvolution without proportions only surpasses the deconvolution with proportions near prior noise of 0.45 (Fig. 10).

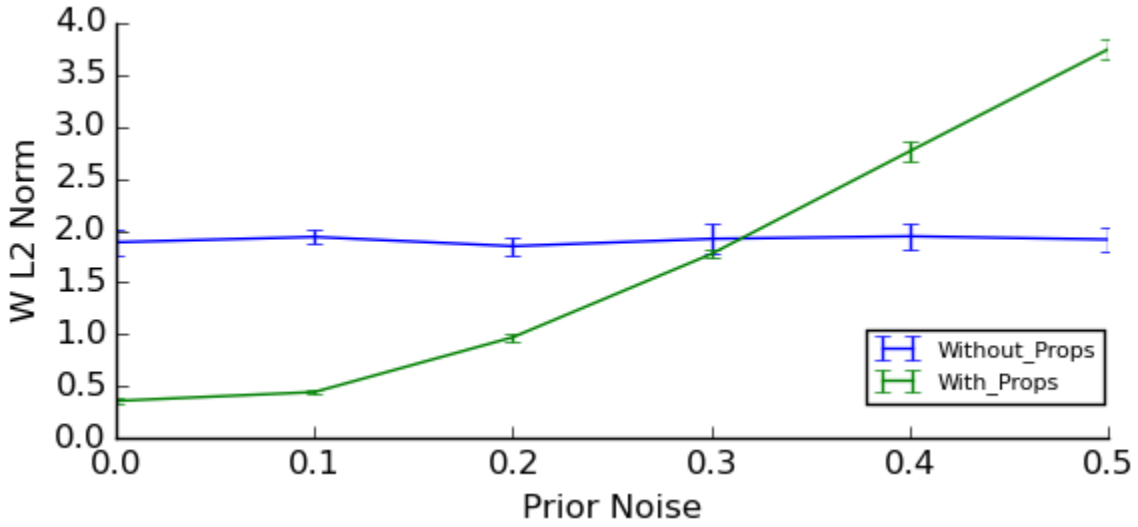

Fig. 9 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the amount of additive prior noise increases. 'W L2 Norm' is the Frobenius norm of the difference between the predicted weights and the real weights ($\|W_P - W_R\|_2$).

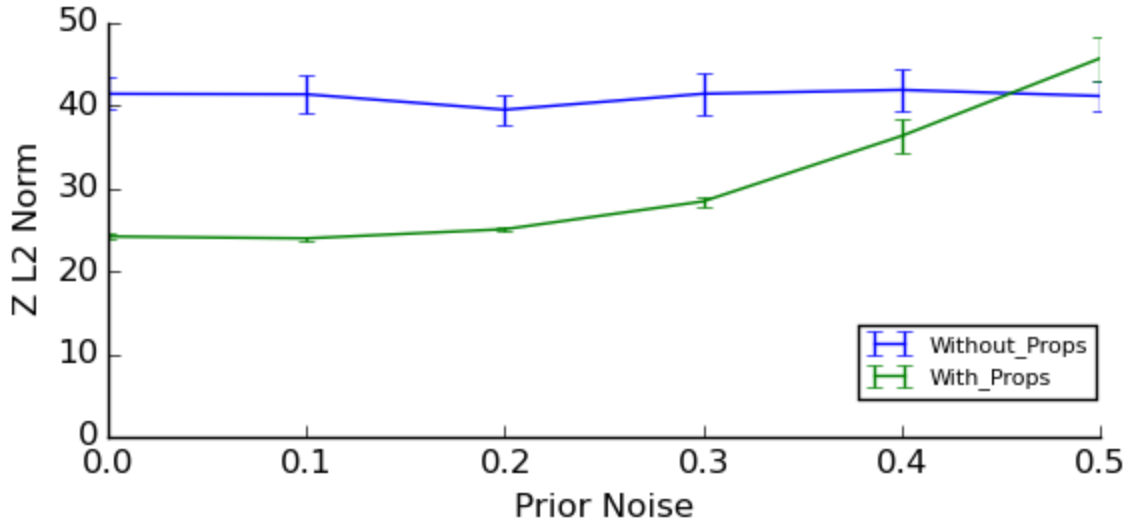

Fig. 10 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the amount of additive prior noise increases. 'Z L2 Norm' is the Frobenius norm of the difference between the predicted expression profiles and the real expression profiles ($\|Z_P - Z_R\|_2$).

Error vs Sample Heterogeneity

In this experiment, sample heterogeneity is the average number of hidden profiles contained in each sample. The results show that deconvolution with and without the proportions behave differently with changes in heterogeneity. More specifically, the deconvolution without

proportions improves its W matrix prediction with increasing heterogeneity, whereas the deconvolution with proportions makes a worse prediction of W as the heterogeneity increases (Fig. 11). Regarding the gene expression profiles, heterogeneity doesn't seem to affect the deconvolution when the proportions are provided (Fig. 12). The prediction of Z is also constant without the proportions when the heterogeneity is greater than 4.

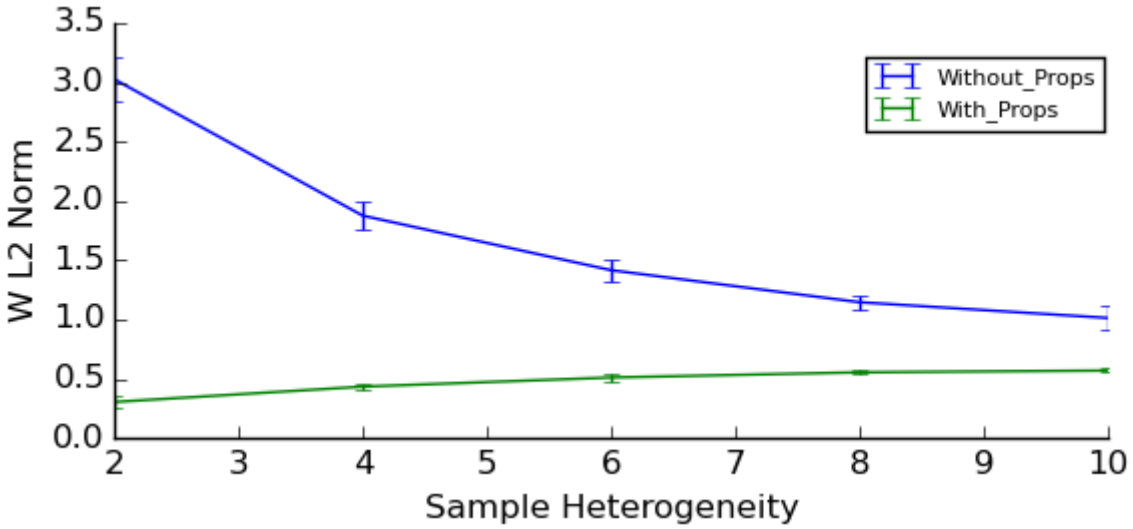

Fig. 11 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the sample heterogeneity increases. 'W L2 Norm' is the Frobenius norm of the difference between the predicted weights and the real weights ($\|W_P - W_R\|_2$).

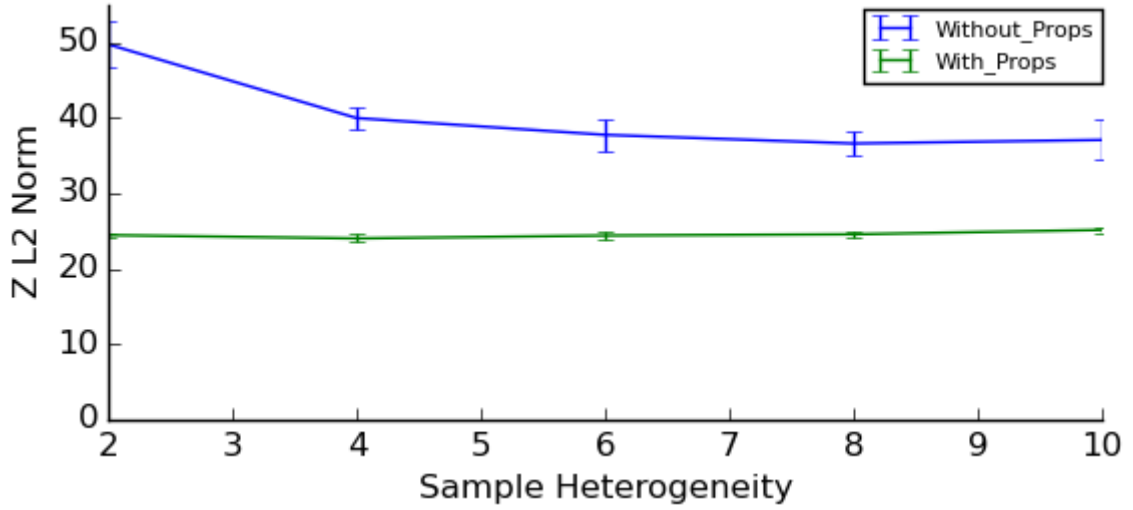

Fig. 12 Comparison of the deconvolution with and without subpopulation proportion (prop) information as the sample heterogeneity increases. 'Z L2 Norm' is the Frobenius norm of the difference between the predicted expression profiles and the real expression profiles ($\|Z_P - Z_R\|_2$).

C. Real Data Experiment

This experiment will examine whether subpopulations share the same gene expressions or whether subpopulations gene expressions differ within an individual. For this experiment, I used 52 RNASeq breast cancer samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project along with the predicted cellular prevalence of their subpopulations. The subpopulation information came from running PhyloWGS [8] on the samples. The average number of subpopulations per sample is 3.3. The cellular prevalence values are converted to proportions by simply subtracting the sum of the child prevalences from their parents. While I test the assumption that subpopulations share gene expressions, I also introduce a new assumption, which is that the PhyloWGS output is accurate. In the previous section, we showed that the deconvolution with proportions is superior to the deconvolution without the proportions in nearly all scenarios. The proportions only decrease the performance of the deconvolution when they are inaccurate (Fig. 9 and Fig. 10). Thus in this experiment, we are assuming that the proportions provided by the phylogenetic reconstruction are accurate.

In this experiment, the deconvolution of the real data is done in two scenarios: 1) using all the predicted proportions, and 2) using only the cellularity proportion and one minus the cellularity proportion. The cellularity proportion is the percent of the tumor that is non-cancerous. Thus the idea behind the experiment is that if subpopulations do have different gene expressions then scenario 1) should result in lower reconstruction error. Reconstruction error is the Frobenius norm of the difference between the predicted sample expressions and the actual sample expressions ($\|\mathrm{X} - \mathrm{WZ}\|_2$). However, if subpopulations don't have different gene

expressions (share the same expressions) then we would expect the model with only two proportions for each sample to have lower error.

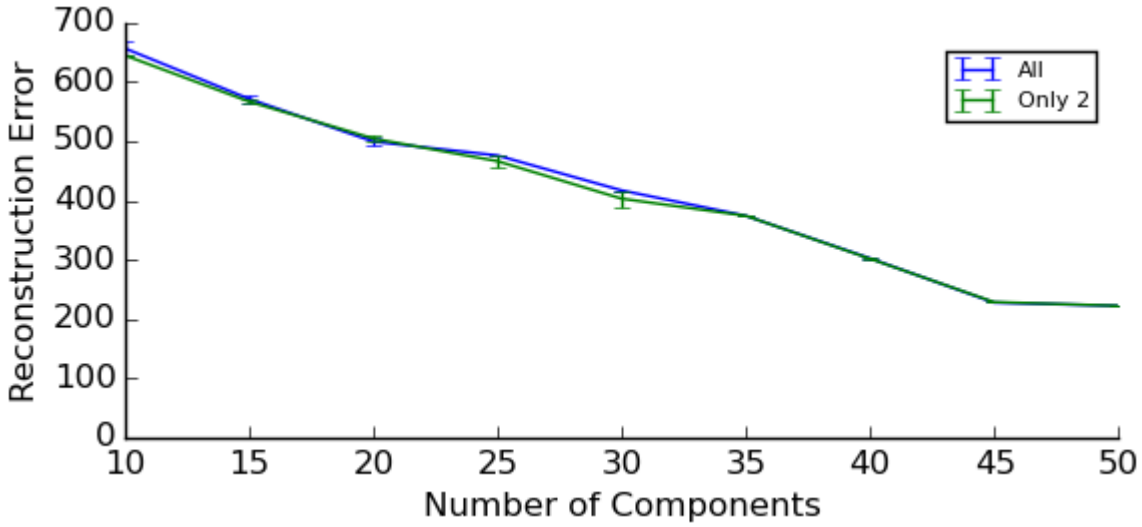

Fig. 13 Comparison of the reconstruction error of the deconvolution method using all the predicted subpopulation proportions versus the deconvolution restricted to only two proportions. Both deconvolutions are compared over an increasing number of model components. Reconstruction error is the Frobenius norm of the difference between the predicted sample expressions and the actual sample expressions ($\|X - WZ\|_2$).

Fig. 13 is a plot of the reconstruction error of these two scenarios with varying number of components in the model. The two scenarios end up having nearly identical error. There are many possible explanations for this result. Here are two possible explanations. It could be that the proportions are incorrect, leading to both models being equally wrong. Another explanation is that there are too many hidden profiles to model. As seen in Fig. 5 and Fig. 6, without sufficient information (samples), the model will perform very poorly. Thus it is very possible that there are too many subpopulations and that 52 samples didn't provide enough information to capture them all.

VI. Discussion

A. Assumptions

For the deconvolution method, there are two major assumptions:

1) Different tumour subpopulations have different gene expressions.
2) Samples from different patients share common subpopulations.

The simulated data reflects these assumptions since each sample is composed of a set of distinct tumor samples. However, the assumptions may not be appropriate for real tumor gene expression data. The reason for assuming 2) is so that we can gain information from a large set of samples to improve the deconvolution of any individual sample. It's unclear whether samples from different patients share common subpopulations, however, we do know that there exists common subtypes within a cancer type [32] [33], thus it may be the same case for subpopulations. The problem is that even if assumption 2) holds, two different subpopulations within a sample may have more similar gene expression profiles than a single subpopulation coming from two different samples. Therefore we must assume 1), meaning that we assume that the distinct genetic mutations of the subpopulations affect their gene expression levels. The reality is likely somewhere in between: 1) the genetic differences of subpopulations affect the expressions of some genes and 2) some subpopulations are common among different samples whereas some are unique.

Another assumption that is made by the model is that the mixture of gene expressions is linear in all genes. Although this assumption may not hold for some genes, it is expected that a linear model can, to some extent, capture nearly linear responses with sufficient accuracy [33],[18].

B. Proportions Act as a Regularization

When we compare the projection with and without the fitting of the prior values, the main benefit that the fitting has is that it acts as a type of regularization. This is most notably seen in Fig. 11. The increase in heterogeneity increases the complexity of the problem since each sample is a mixture of more components. Accordingly we see that the method with the proportions

increases in error with an increase in heterogeneity. In contrast, the method without the proportions benefits from the increase in heterogeneity since there are fewer non-zero proportions. Even in a sparse setting, the method without the proportions spreads out the weights to more components, leading to higher error. The prior proportions provides not only the magnitude of the weights, but also the number of non-zero weights. In the cancer deconvolution setting, the weight vectors are sparse, so having the prior values constrains the solution to be sparse as well. In other applications, the L1 weight regularization is used to push models towards more sparse solutions. In this scenario, the L1 regularization would not provide any benefit because the proportions are constrained to sum to one, thus the L1 norm of the weights is equal for all feasible solutions. Another alternative could be to add L0 regularization, which penalizes based on the number of non-zero elements, thus increasing sparsity. The L0 optimization is an NP-hard problem and therefore would need to be approximated. This approach was not explored in this report but it could lead to improved predictions for the deconvolution without proportion information.

VII. FUTURE DIRECTIONS

One of the challenges of studying clonal heterogeneity is the availability of representative biopsies. Bulk tumor samples provide an average picture, but the problem is that large numbers of cells pooled together will quench the signal from minor subpopulations [3]. On the other hand, single cell sequencing provides a powerful approach for resolving clonal substructure, reconstructing phylogenetic lineages, and identifying unique tumor cell-specific gene expression profiles [34][35]. However, the resolution is so small that most subpopulations will be missed. This can be dealt with by analyzing larger numbers of individual cells, but scaling up the analysis will inevitably drive up the cost and labor required, making these studies impractical.

Accordingly, the future direction of tumor devolution could incorporate both bulk and single cell sequencing. The single cell data could be easily integrated into the model by merging their gene expression profiles with into the Z matrix. Their gene expression profiles would represent pure subpopulations. This data would help reduce the uncertainty in the deconvolution of bulk tumour samples.

VIII. CONCLUSION

Intratumor heterogeneity currently interferes with the development of personalized cancer treatments. With the advent of tumour evolution modelling, we now have access to estimates of the proportions of the subpopulations within a sample. In this report, I hypothesized that these cancerous subpopulations differ not only in their genetic mutations but also in their population gene expression profiles. To this end, I showed how proportion estimates can be incorporated into the deconvolution of tumour samples. In addition, I addressed the combinatorial optimization problem of matching the proportions to the components by introducing a relaxation to the problem that sufficiently approximates the correct solution. Using simulated data, I demonstrate that the deconvolution benefits from the incorporation of the subpopulation proportions but only when the proportions are precise. I also investigate the deconvolution of real tumour data which reveals that the assumptions made by the model are not perfect. In all, through the combination of different sources of information and techniques, we become closer to alleviating the problems caused by tumour heterogeneity. Further research into this area will improve our understanding of tumour evolution and help the development of improved treatments.

IX. REFERENCES

[1] I. J. Fidler and I. R. Hart, "Biological diversity in metastatic neoplasms: origins and implications.," *Science*, vol. 217, no. 4564, pp. 998–1003, Sep. 1982.

[2] G. H. Heppner, "Tumor heterogeneity.," *Cancer Res.*, vol. 44, no. 6, pp. 2259–2265, Jun. 1984.

[3] A. Marusyk and K. Polyak, "Tumor heterogeneity: causes and consequences," *Biochim Biophys Acta*, vol. 1805, no. 1, pp. 1–28, 2011.

[4] P. C. Nowell, "The clonal evolution of tumor cell populations.," *Science*, vol. 194, no. 4260, pp. 23–28, Oct. 1976.

[5] C. Swanton, "Intratumor heterogeneity: evolution through space and time.," *Cancer Res.*, vol. 72, no. 19, pp. 4875–4882, Oct. 2012.

[6] L. M. F. Merlo, J. W. Pepper, B. J. Reid, and C. C. Maley, "Cancer as an evolutionary and ecological process.," *Nat. Rev. Cancer*, vol. 6, no. 12, pp. 924–935, Dec. 2006.

[7] N. Niknafs, V. Beleva-Guthrie, D. Q. Naiman, and R. Karchin, "SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing.," *PLoS Comput. Biol.*, vol. 11, no. 10, p. e1004416, Oct. 2015.

[8] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris, "Reconstructing subclonal composition and evolution from whole genome sequencing of tumors," *Genome Biol.*, pp. 0–29, 2014.

[9] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris, "Inferring clonal evolution of tumors from single nucleotide somatic mutations.," *BMC Bioinformatics*, vol. 15, p. 35, 2014.

[10] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning, and A. L. Borresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 19, pp. 10869–10874, Sep. 2001.

[11] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer.," *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002.

[12] G. M. Dancik and D. Theodorescu, "Robust prognostic gene expression signatures in bladder cancer and lung adenocarcinoma depend on cell cycle related genes.," *PLoS One*, vol. 9, no. 1, p. e85249, 2014.

[13] G. Quon, S. Haider, A. G. Deshwar, A. Cui, P. C. Boutros, and Q. Morris, "Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction.," *Genome Med.*, vol. 5, no. 3, p. 29, 2013.

[14] H. Lahdesmaki, L. Shmulevich, V. Dunmire, O. Yli-Harja, and W. Zhang, "In silico microdissection of microarray data from heterogeneous cell populations.," *BMC Bioinformatics*, vol. 6, p. 54, 2005.

[15] S. S. Shen-Orr and R. Gaujoux, "Computational deconvolution: extracting cell type-specific information from heterogeneous samples," *Curr. Opin. Immunol.*, vol. 25, no. 5, pp. 571–578, 2013.

[16] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. October 1999, pp. 788–91, 1999.

[17] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. a Alizadeh, "Robust enumeration of cell subsets from tissue expression profiles," *Nat. Methods*, vol. 12, no. 5, pp. 1–10, 2015.

[18] T. Erkkilä, S. Lehmusvaara, P. Ruusuvuori, T. Visakorpi, I. Shmulevich, and H. Lähdesmäki, "Probabilistic analysis of gene expression measurements from heterogeneous tissues," *Bioinformatics*, vol. 26, no. 20, pp. 2571–2577, 2010.

[19] B. W. Okaty, K. Sugino, and S. B. Nelson, "A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain," *PLoS One*, vol. 6, no. 1, pp. 1–10, 2011.

[20] M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, and C. Swanton, "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing," *N. Engl. J. Med.*, vol. 366, no. 10, pp. 883–892, 2012.

[21] A. E. O. Hughes, V. Magrini, R. Demeter, C. a. Miller, R. Fulton, L. L. Fulton, W. C. Eades, K. Elliott, S. Heath, P. Westervelt, L. Ding, D. F. Conrad, B. S. White, J. Shao, D. C. Link, J. F. DiPersio, E. R. Mardis, R. K. Wilson, T. J. Ley, M. J. Walter, and T. a. Graubert, "Clonal Architecture of Secondary Acute Myeloid Leukemia Defined by Single-Cell Sequencing," *PLoS Genet.*, vol. 10, no. 7, p. e1004462, 2014.

[22] C. A. Klein, "Selection and adaptation during metastatic cancer progression," *Nature*, vol. 501, no. 7467, pp. 365–372, Sep. 2013.

[23] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger, "TrAp: a tree approach for fingerprinting subclonal tumor composition," *Nucleic Acids Res.*, vol. 41, no. 17, pp. e165–e165, 2013.

[24] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah, "PyClone: statistical inference of clonal population structure in cancer," *Nat. Methods*, vol. 11, no. 4, pp. 396–398, 2014.

[25] N. Andor, J. V. Harness, S. Müller, H. W. Mewes, and C. Petritsch, "Expands: Expanding ploidy and allele frequency on nested subpopulations," *Bioinformatics*, vol. 30, no. 1, pp. 50–60, 2014.

[26] L. Oesper, A. Mahmoody, and B. J. Raphael, "Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7821 LNBI, no. 7, pp. 171–172, 2013.

[27] M. Chen, M. Gunel, and H. Zhao, "SomatiCA: Identifying, Characterizing and Quantifying Somatic Copy Number Aberrations from Cancer Genome Sequencing Data.," *PLoS One*, vol. 8, no. 11, p. e78143, 2013.

[28] M. X. Goemans, "Semidefinite Programming and Combinatorial Optimization," *Doc. Math.*, vol. Extra Volu, pp. 657–666, 1998.

[29] J. Feldman, M. J. Wainwright, and D. R. Karger, "Using linear programming to decode binary linear codes," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 954–972, 2005.

[30] B. L. WELCH, "The Generalization of 'Student's' problem when several different population varlances are involved," *Biometrika*, vol. 34, no. 1–2, pp. 28–35, 1947.

[31] "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[32] "Comprehensive molecular characterization of urothelial bladder carcinoma," *Nature*, vol. 507, no. 7492, pp. 315–322, Mar. 2014.

[33] M. Hoffmann, D. Pohlers, D. Koczan, H.-J. Thiesen, S. Wolfl, and R. W. Kinne, "Robust computational reconstitution - a new method for the comparative analysis of gene expression in tissues and isolated cell fractions.," *BMC Bioinformatics*, vol. 7, p. 369, 2006.

[34] N. E. Navin, "The first five years of single-cell cancer genomics and beyond.," *Genome Res.*, vol. 25, no. 10, pp. 1499–1507, Oct. 2015.

[35] K.-T. Kim, H. W. Lee, H.-O. Lee, S. C. Kim, Y. J. Seo, W. Chung, H. H. Eum, D.-H. Nam, J. Kim, K. M. Joo, and W.-Y. Park, "Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells.," *Genome Biol.*, vol. 16, p. 127, 2015.

X. APPENDIX A

A. Proof of Fitting Optimality

After the projection of X_i onto Z^T, we end up with the vector W_i. The next step is to fit the set of values S_i to W_i. To fit S_i to W_i, first we sort the set S_i. Next, the largest value of W_i is replaced with the largest value of S_i, and the second largest value of W_i is replaced with the second largest value of S_i, and so on. The optimal ordering of W_P is defined as the permutation of S_i (with zeros added so that the length of W_P is equal to the number of components) that minimizes $\|Z^T W_P - X_i\|_2$. Thus we are searching for the ordering of W_P that is closest to W_i, ie. $\|W_P - W_R\|_2$.

I claim that the fitting method of swapping the values in order of magnitude obtains the optimal ordering. This may be seem obvious, nonetheless, I will prove it in order to remove any doubt. In order to prove that it is optimal, I will show that any other ordering is sub-optimal. Let

W_i have values i and j where $i \geq j$ and let S_i have values k and l where $k \geq l$. Consequently, using my method of matching, i will be matched with k and j will be matched with g. The alternative ordering is that i is matched with l and j is matched with k. I need to show that the first ordering is always better than or equal to the second. 'Better' in this case means less error. This can be defined as follows,

$$\|i - k\| + \|j - l\| \leq \|i - l\| + \|j - k\|.$$

For our application we use the L2-norm, however this proof does apply to all norms. There are two general cases which we need to consider. The first case is where the second ordering has a term that is greater than both terms of the first ordering. The second case is where the each term of the second ordering is individually greater that one of the terms of the first ordering. To demonstrate the first case, let us assume that $k \geq i \geq l \geq j$. Then the proof is as follows,

$$\begin{aligned} \|i - k\| + \|j - l\| &\leq \|j - k\| \\ &\leq \|j - k\| + \|i - l\|. \end{aligned}$$

The first inequality stems from the assumption $k \geq i \geq l \geq j$. The second inequality comes from the fact that all norms are positive. To demonstrate the second case, let us assume that $k \geq i \geq j \geq l$. Then the proof is as follows,

$$\begin{aligned} \|i - k\| + \|j - l\| &\leq \|j - k\| + \|j - l\| \\ &\leq \|j - k\| + \|i - l\|. \end{aligned}$$

The first inequality is explained by the assumption $k \geq i \geq j$ so that $\|i - k\| \leq \|j - k\|$ and the second inequality is explained by $i \geq j \geq l$ so that $\|j - l\| \leq \|i - l\|$. All other orderings of $i, j, k,$ and l follow very similar proofs as the ones above. Either the second ordering has one norm that is greater than both norms of the first ordering or each norm of the second

ordering is individually greater than one of the norms of the first ordering. Thus, though it may have been obvious to some, I have shown that swapping the values in order of magnitude obtains the optimal ordering.