BST283: Final Project

Alex Munoz, Emily Simons, Margaux Hujoel

May 8, 2017

1 Introduction

Intratumor genetic variation is an established risk factor for decreased survival and increased likelihood of treatment failure (Mroz et al. 2015; Landau et al. 2013). One approach to assessing heterogeneity is to sample a tumor in multiple regions and evaluate for the presence of distinct subclones based on differences in somatic single nucleotide variants (SNVs). We present here a method for clustering shared versus private sSNVs and inferring a phylogenetic tree.

2 Background

2.1 Clinical basis of tumor phylogenetics

Intratumor genetic variation affects treatment response and recurrence, but is not currently part of clinical management. Clinical management of cancer is comprised of two components: treatment (resection, medications, and/or radiation) and surveillance (via imaging and/or biomarkers) to determine treatment response and recurrence. Both cancer treatment and surveillance decisions pivot on pathologic stage. The fundamental purpose of pathologic stage is to capture the most important prognostic factors of a cancer type in order to facilitate cost-benefit decision making involved in selecting therapies and surveillance frequencies that have significant adverse effects, such as risk of secondary malignancy associated with traditional cytotoxic chemotherapy. Stage has been conventionally determined from tumor depth, nodal involvement and presence of metastasis, which for most tumors are derived from histopathologic analysis of tissue and imaging studies (e.g. FDG PET-CT.) Conventional pathologic stage, however, often does not adequately predict response nor durability of response to therapy. Genomics offers a new dimension- the molecular anatomy of a tumor – that is augmenting the power of pathologic stage to tailor treatment and surveillance approaches. Many genetic characteristics have been incorporated into treatment algorithms. For example, stage IIA (T3N0M0) colon cancer is treated exclusively with surgery if microsatellite instable, but adjuvant chemotherapy can be beneficial for microsatellite stable stage IIA tumors (see Figure 1 below.) Intratumor genetic heterogeneity, however, has long been recognized as a critical factor that could explain why some tumor cells develop resistance mechanisms (Samuel et al. 2013), but is not yet a feature of genetic profiling practices in clinical oncology.

The degree of intratumor genetic heterogeneity is an important marker of survival and likelihood of developing resistance mechanisms. Evidence for this comes from numerous cancer types. In metastatic renal-cell carcinoma, samples from different regions of the same tumor revealed distinct driver mutations. It follows that intratumor genetic heterogeneity would be associated with heterogeneous protein function, facilitating tumor adaptation and therapeutic failure through selection (Gerlinger et al. 2012). In head and neck cancer, a measure of intratumor heterogeneity based on the distribution of mutant allele fraction across mutated loci within a tumor was associated with shorter survival time, even when controlling for critical risk factors such as age, HPV status, tumor grade and nodal involvement and presence of TP53 mutation (Mroz et al. 2015). In 149 patients with chronic lymphocytic leukemia, the number of sSNVs increased both with treatment and with number of treatments. This enriched subclonal evolution led to new driver mutations and decreased progress-free survival (Landau et al. 2013.) In summary, increased intratumor genetic heterogeneity is a poor prognostic factor for multiple cancer types and may warrant enhanced surveillance or modified treatment. Clinical care could be improved if actionable mutations were known to be present as clonally dominant truncal events or a set of comprehensive targets that cover detectable subclones.

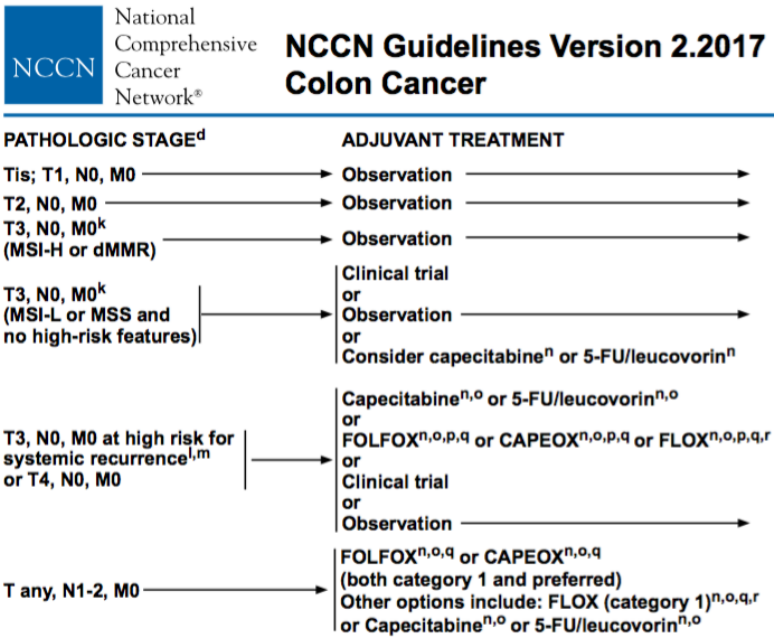


Figure 1: Colon cancer treatment guidelines: MSI-H: MSI high, equivalent to dMMR (deficient mismatch repair) MSI-L: MSI low, equivalent to MSS (microsatellite stable) Source: "Clinical Practice Guidelines in Oncology: Colon Cancer." NCCN. Version 2.2017- March 13, 2017. Accessed at: https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf

2.2 Definition of intratumor genetic heterogeneity

Genetic heterogeneity can be understood as any genomic alteration between cells including single nucleotide variants, copy number alterations (CNAs) and insertion/deletions (indels). One way to interpret genomic heterogeneity is to define subclones by the unique sSNVs that differentiate one sample of tumor cells from another. The number of subclones that arise over time and the number of sSNVs that define those subclones can characterize the genetic diversity of a tumor. The phylogenies of tumor clones can be derived from differences in sSNVs across the cancer samples. Zare et al. propose that determining the number of distinct subclones from multiple samples from a tumor provides a framework for linking a tumor's molecular anatomy to its structural anatomy and phylogenetic evolution.

2.3 Current Methodology

There are three distinct areas in which researchers are using tumor phylogenetics to try to improve clinical treatment: cross-sectional tumor phylogenetics, regional bulk tumor phylogenetics and single-cell tumor phylogenetics (Schwartz and Schaffer, 2017). Cross-sectional tumor phylogenetics consists of analyzing tumors of multiple individuals for the purpose of determining if sets of mutations cluster at different stages of the cancer (e.g. one could infer if a certain mutation is a seminal event found in in situ carcinomas that is followed by an orderly process of subsequent mutations as tumor stage progresses). Regional bulk tumor phylogenetics (the focus of this paper) focuses on multiple samples from within one patient and constructing a phylogenetic tree for one patient (e.g. looking at samples from within a primary tumor and a metastasis and forming a tree). Often, within this problem, another embedded problem is to infer whether there are multiple distinct clones within one bulk genomic sample (i.e. determining the clones within a sample and then using the inferred clones across samples to build a tree). If clonal inference is not done, one is assuming that each sample is clonally homogeneous. Single-cell tumor phylogenetics is the idea of constructing a phylogenetic tree given single-cell data from an individual (basing inference on individual tumor cells). Currently, there exist methods implemented for each of the three different phylogenetic problems as well as for various input data types.

3 Methods

With our method, we focus on regional bulk tumour phylogenetics, and attempt to extend the Zare et al. model. Zare et al. focus only on SNVs (i.e. they disregard copy number), as do we in our implementation. However, Zare et al. ignore within sample heterogeneity - they assume each sample consists of a distinct clone. We attempt to extend this to allow for intra-sample heterogeneity - or that a sample can consist of multiple distinct clones.

3.1 Setting

We are in the setting where we have C, the matrix of mutational cancer cell fraction (CCF) and we wish to output Z and P: the tree and clonal CCF. We know $C = ZP$, and so for a given number of clones c we wish to find clusters (columns of Z) and optimize such that P is valid (all columns in P must sum to 1). In our method we focus only on SNVs and ignore CNAs and indels.

More specifically, we have,

$$s = \#samples; \qquad n = \#loci;$$

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1s} \\ c_{21} & c_{22} & \cdots & c_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{ns} \end{pmatrix}_{loci \times samples}$$

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1c} \\ z_{21} & z_{22} & \cdots & z_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nc} \end{pmatrix}_{loci \times clones} \qquad P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{c1} & p_{c2} & \cdots & p_{cs} \end{pmatrix}_{clones \times samples}$$

Z is a binary matrix where a 1 at z_{ij} denotes having mutation at locus i in clone j.

Say we fix the number of clones at c.

$$C_i = \begin{pmatrix} c_{1i} \\ c_{2i} \\ \vdots \\ c_{ni} \end{pmatrix} = p_{1,i} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + p_{2,i} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + p_{2^n,i} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

We wish to optimize the above for $p_{j,i}$ with the constraint that only c $p_{j,i}$ values can be non-zero (only have c clones) across the C_i, $i = 1, \ldots, s$, where the non-zero $p_{j,i}$ can be (and will be) different across the C_i, but for the $2^n - c$ $p_{j,i} = 0$, these must be the same j across the i. Note that $p_{j,i}$ is the percent of sample i made up of clone j and therefore we have the constraint: $\sum_j p_{j,i} = 1 \forall i$.

3.2 Basic Example

Let $s = \#samples = 2, n = \#loci = 3$. Then we have:

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix}_{loci * samples}.$$

Then we find we are optimizing the below:

$$\begin{pmatrix} c_{11} \\ c_{21} \\ c_{31} \end{pmatrix} = p_1 \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + p_2 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + p_3 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + p_4 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + p_5 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + p_6 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + p_7 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + p_8 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix};$$

$$\begin{pmatrix} c_{12} \\ c_{22} \\ c_{32} \end{pmatrix} = p_1^* \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + p_2^* \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + p_3^* \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + p_4^* \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + p_5^* \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + p_6^* \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + p_7^* \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + p_8^* \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

under the constraint that $\sum p_i^* = \sum p_i = 1$ and, given c clones, only c p_i, p_i^* values are allowed to be non-zero.

Assume we have 4 clones. Then $2^n - c = 2^3 - 4 = 4$ of the ps will be 0. Assume our optimization found $p_1 = p_1^* = p_2 = p_2^* = p_3 = p_3^* = p_6 = p_6^* = 0$. Then we find:

$$\begin{pmatrix} c_{11} \\ c_{21} \\ c_{31} \end{pmatrix} = p_4 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + p_5 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + p_7 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + p_8 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} p_5 + p_8 \\ p_5 + p_7 + p_8 \\ p_4 + p_7 + p_8 \end{pmatrix};$$

$$\begin{pmatrix} c_{12} \\ c_{22} \\ c_{32} \end{pmatrix} = p_4^* \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + p_5^* \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + p_7^* \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + p_8^* \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} p_5^* + p_8^* \\ p_5^* + p_7^* + p_8^* \\ p_4^* + p_7^* + p_8^* \end{pmatrix}.$$

Or, succinctly,

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} p_4 & p_4^* \\ p_5 & p_5^* \\ p_7 & p_7^* \\ p_8 & p_8^* \end{pmatrix} = ZP.$$

Zare et al. noted that if we let clone number exceed number of samples it is not guaranteed that the conditions for perfect phylogeny are met. We let clone number (3) exceed sample number (2) in this simple case. By letting clone number exceed number of samples, we are forcing at least one sample to be composed of 2 or more distinct clones. This allows us to account for intra-sample heterogeneity.

Our big idea behind optimization is to enumerate all possible Z matrices (given a fixed number of clusters, c) and then to fix P for a given Z. Then we determine which of these Z, P pairs result in the closest value to the original C. Enumeration of the Z is computationally intensive.

3.3 Fixing of P

For each Z matrix we find the best P by obtaining the least squares estimate of P and then linearly transform the P so it fits the restriction underlying the P matrix that each column of P must sum to 1.

3.4 Enumeration of Z

3.4.1 Method 1

Our first method to generate Z generated all possible combinations of columns of zeros and ones (where no columns were repeated). This method of enumeration does not incorporate the assumption of perfect phylogeny, and as such, becomes intractable when we have a large number of clones or loci. This method is $O(\binom{2^n}{c})$ where n is the number of loci and c is the number of clones.

3.4.2 Method 2

We use the assumption of perfect phylogeny in order to make our problem tractable. Therefore we consider how to generate only Z from which the perfect phylogeny assumption is held. We get this general methodology from Gusfield (1991).

We will consider a new matrix, M, with dimensions number of clones by number of loci, where the matrix is sorted such that the clone with the greatest number of mutations is the top row of the matrix. We can obtain the matrix from Z by first transposing Z and letting each mutated loci be denoted by a specific number (i.e. if mutated at loci 1, denoted by a 1, loci 2 with a 2, and so on). We then sort the matrix such that the clone with the maximum number of mutations is the top row of the matrix, and then we shift all numbers in the row to the left such that all 0s are to the right of the matrix. A tree with perfect phylogeny is one where in this matrix M, all of the same number must be in the same column. If there are two numbers that cannot be in the same column, this corresponds to a Z that doesn't fit the perfect phylogeny assumption. Note that the mutations that occur in the most number of samples appear on the left-hand side of the M-matrix - this ordering is important.

Unfortunately, we were unable to determine all the restrictions on M to generate only Z from perfect phylogeny- we did however eliminate the generation of many Z that do not follow perfect phylogeny.

For example, if we have the following Z we obtain the following M.

$$Z = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 0 \\ 1 & 2 & 3 & 4 & 0 & 0 \\ 1 & 2 & 3 & 0 & 0 & 6 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 0 \\ 1 & 2 & 3 & 4 & 0 & 0 \\ 1 & 2 & 3 & 6 & 0 & 0 \end{pmatrix} = M.$$

We can note this Z corresponds to a tree that fits the assumption of perfect phylogeny: see the corresponding tree where the branches are labeled with the loci that was mutated (Figure 2 A.).

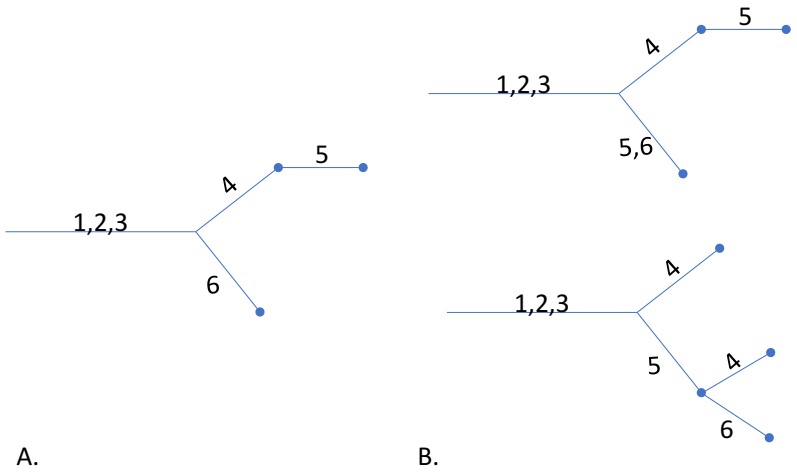


Figure 2: A. presents a tree that adheres to perfect phylogeny whereas B. depicts two trees that fail to meet the assumption of perfect phylogeny as a distinct mutation arises twice.

We can now consider if the tree does not fit under the assumption of perfect phylogeny. If we have the following Z we obtain the following M.

$$Z = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 0 \\ 1 & 2 & 3 & 4 & 0 & 0 \\ 1 & 2 & 3 & 0 & 5 & 6 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 0 \\ 1 & 2 & 3 & 4 & 0 & 0 \\ 1 & 2 & 3 & 5 & 6 & 0 \end{pmatrix} = M.$$

We see that the tree depicted in Z does not adhere to the perfect phylogeny assumption because a distinct mutation arises twice (see Figure 2 B.). We see this reflected in M by the inability to have all of the same numbers appear in the same column (namely 5).

3.4.3 Comparison

Usually $c << n$ (that is we have much fewer clones than number of loci) which makes method 2 a much more computationally efficient method than method 1. As such, we felt it was reasonable to assume perfect phylogeny and use the much more computationally efficient method of enumerating Z.

3.5 Final note on implementation

Finally, rather than enumerate all possible M (or equivalently all possible Z which fit the perfect phylogeny assumption) we generate a random number of M. If our error from the best Z, P pair is no longer minimized upon adding more M, then we assume our Z has converged and set the Z equal to the Z, P pair which minimizes the error. For example, given 5 mutations and 4 samples, generating at least 5,000 M (and thus 5,000 Z) usually results in convergence and therefore in order to improve efficiency we generate only 5,000 M (and therefore only 5,000 Z) when we have 5 mutations and 3 samples.

In our example, given the small size of the parameter space, we converged to a global minimum. However, for larger parameter spaces, one should demonstrate that convergence to a local minimum error given a certain number of iterations is producing a tree very similar to the Z that minimizes the global error minimum. In our simulations, converging to low error (local minima) implies similar tree structure although not always the correct tree structure, that is, some local minima have slightly increased likelihood compared to the global minimum. This save on computation time may result in an imperfect tree with similar structure to the correct tree.

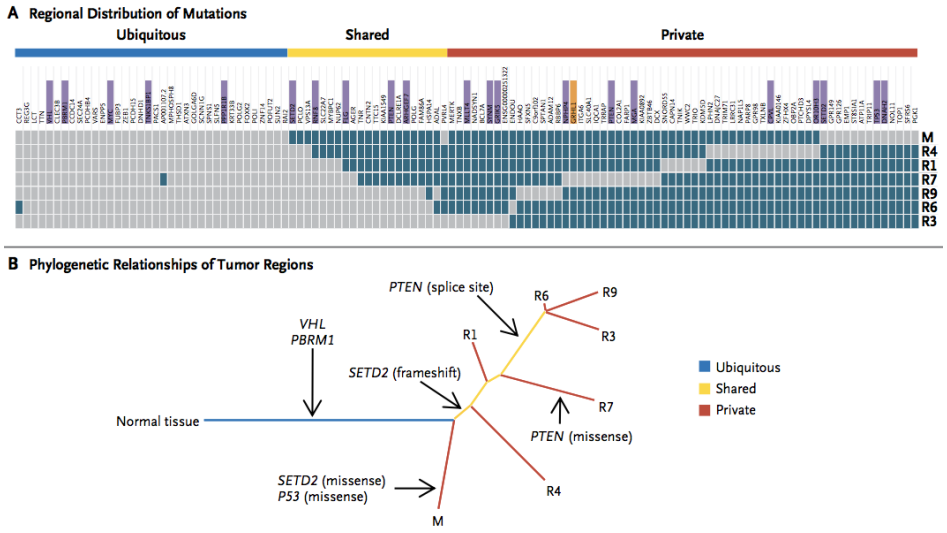


Figure 3: Figure 4 found in the paper by Gerlinger et al. (2012) that shows the mutation pattern and inferred phylogenetic tree in patient 2.

3.6 Finding number of clones

Our algorithm requires that the number of clones be fixed. As such, we needed to determine the optimal number of clones. Increasing the number of clones will always reduce the amount of difference between ZP and C. As such, in order to determine optimum c one needs to weight the reduction in error to the number of clones: c. We can determine c by using AIC.

3.7 Implementation

The implementation of our methodology (specifically only method 2's approach to enumerate Z) is found at `https://github.com/munozalexander/Tumor-Heterogeneity`. Moreover, the implementation is attached as an Appendix.

4 Results

We applied our method to an example found in literature. Gerlinger et al. produced a phylogenetic tree given mutations found in different tumors within a patient (see Figure 3; Gerlinger et al. 2012). Given the breakdown of whether a mutation was present in a sample or not (part A. of Figure 3) we could construct a matrix Z. Due to the computational intractability as our method as it is currently implemented, we only looked at 6 clinically relevant mutations which corresponds to 3 distinct primary samples and 1 distinct metastatic sample. The Z matrix corresponding to this simplified tree from Gerlinger et al. is found in Figure 4 as well as the code used to generate this Z.

We then stochastically determined a valid P matrix and created a C by letting $C = ZP$. We then applied our method to this C matrix and compared our inferred Z matrix to the true Z. We generated data (a C matrix) from these 4 samples assuming 4 clones and using a stochastic P matrix. We varied the number of clones (c) and selected the "optimum" number using AIC - 4 clones was the optimal number of clones (which is the number of clones assumed to generate the data - see Figure 7). We then fit our model and looked at the resulting Z. Looking at the output in Figure 4, we see our obtained optimal Z is the same as inputted tree: our method found the correct tree. See Figures 5-7 for more details.

```
In [448]: Gerlinger_data = pd.read_csv('Data/Gerlinger.csv')
          loci = [0,1,2,4,6]
          Gerlinger_Z = Gerlinger_data.ix[loci,[3,6,7,9]]
          opt = Z_P_Optimizer(loci_num=len(loci), samples_num=3, clones_num=4)
          P = opt.draw_P_matrix()
          C = np.dot(Gerlinger_Z, P)
          print('Loci genotyped:', list(Gerlinger_data.ix[loci,1]))
          Gerlinger_Z

          Loci genotyped: ['PBRM1', 'VHL', 'PTENsplice', 'PTENmis', 'TP53']
```

Out[448]:

	R3	R7	R1	M
0	1	1	1	1
1	1	1	1	1
2	1	0	0	0
4	0	1	0	0
6	0	0	0	1

```
In [449]: opt_z, opt_p, opt_err = opt.optimal_Z_P(C_in=C, Zs_count=5000)
          opt_z

Out[449]: array([[ 1.,  1.,  1.,  1.],
                 [ 1.,  1.,  1.,  1.],
                 [ 1.,  0.,  0.,  0.],
                 [ 0.,  1.,  0.,  0.],
                 [ 0.,  0.,  1.,  0.]])
```

Figure 4: Simplified tree from Gerlinger at al. as well as the code to find the optimal tree.

5 Discussion and Future Work

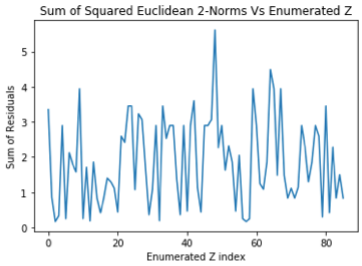
To optimize clinical care in the setting of intratumor genetic heterogeneity, therapy would ideally target clonally dominant truncal somatic events or adopt multiple targeted therapies in a combinatorial approach such that all detectable subclones were covered by at least one therapy. Computational methods such as ours that rely on physically dividing samples could at the very least provide evidence of subclones that would be resistant to a targeted therapy better than one biopsy alone, but our method would certainly not guarantee that shared mutations are pan-tumor truncal events given the impossible resolution of mapping an entire tumor. As methods progress, a multi-biopsy approach would require an estimation of how to maximize genetic information with the least number of samples. In addition to taking multiple samples for genetic profiling, increasing library complexity and read depth could also aid detection of low allelic fraction events that would confer resistance to putative therapy. This is also a costly requirement that would require optimization and rely on continued decline in the costs of sequencing to enable clinical use.

While we improved the computational efficiency of our method by attempting to take advantage of the perfect phylogeny assumption, there is still room for improvement of computational efficiency. Our current method generates all possible Ms given a certain rule (eliminates a portion of Z which violate perfect phylogeny) and then filters these M to remove offenders of the perfect phylogeny assumption. A quicker, more computationally efficient method, would be to only generate M that result in Z following the perfect phylogeny assumption. This would remove the computational burden of generating unnecessary M as well as the burden associated with filtering. Or, one could optimize Z, P jointly with a method other than enumerating all possible Z. For instance, Zare et al. (2014) computed Z, P by maximizing the likelihood of Z, P using an expectation maximization algorithm.

Moreover, our method only looked at SNVs. Future methods could attempt to combine not only SNV information, but also copy number alterations as well as insertions and deletions.

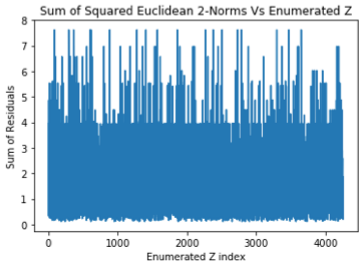
A.

```
In [74]: opt.plot_errors(C_in=C, Zs_count=100)
```



B.

```
In [75]: opt.plot_errors(C_in=C, Zs_count=5000)
```



C.

```
In [462]: return_dict = opt.optimal_Z_P(C_in=C, Zs_count=5000, return_dict=True)
          set_Zs, set_idx = np.unique([np.array_str(a) for a in return_dict['Zs']], return_index=True)
          top_errs = np.argsort(np.array(return_dict['errors'])[set_idx])[:5]
          for Z_curr in set_Zs[top_errs]:
              print(Z_curr)
              print('\n')
```

```
[[ 1.  1.  1.  1.]
 [ 1.  1.  1.  0.]
 [ 0.  1.  0.  0.]
 [ 0.  0.  1.  0.]
 [ 1.  0.  0.  0.]]


[[ 1.  1.  1.  1.]
 [ 1.  1.  1.  0.]
 [ 1.  0.  0.  0.]
 [ 0.  1.  0.  0.]
 [ 0.  0.  1.  0.]]


[[ 1.  1.  1.  1.]
 [ 1.  1.  1.  1.]
 [ 0.  1.  1.  0.]
 [ 0.  0.  0.  1.]
 [ 1.  0.  0.  0.]]


[[ 1.  1.  1.  1.]
 [ 1.  1.  1.  1.]
 [ 1.  1.  0.  0.]
 [ 0.  0.  1.  0.]
 [ 0.  0.  0.  1.]]


[[ 1.  1.  1.  1.]
 [ 1.  1.  1.  1.]
 [ 0.  0.  1.  1.]
 [ 1.  0.  0.  0.]
 [ 0.  1.  0.  0.]]
```

Figure 5: Plotting the sum of squared Euclidean 2-norms vs stochastically enumerated Z index shows multiple troughs (A,B). These multiple local minima arise via the lack of solely linearly independent Z enumeration, for Z matrices with column swaps are included multiple times. Furthermore, structurally similar trees yield local minima. While the true, correct tree is located at the global minimum, local minima were either column swaps of this true Z or were structurally similar trees (C).

In [155]:
```
errors = []
for i in range(10000):
    z_opt, p_opt, err_opt = opt.optimal_Z_P(Zs_count=1)
    if len(errors) > 0:
        errors.append(min([err_opt, errors[-1]]))
    else:
        errors.append(err_opt)
plt.figure()
plt.plot(errors)
plt.xlabel('Number of Z matrices generated')
plt.ylabel('Minimum Euclidean 2-Norm Error')
plt.title('Error vs Number of Zs')
plt.show()
```

Figure 6: When minimum Euclidean 2-norm error is calculated over iterations, convergence is quickly attained for the Gerlinger dataset. After 5000 Z enumeration steps, there is a negligible decrease in error with increased iterations.

```
In [320]: errors = []
clone_range = np.arange(2,11,1)
for clone_count in clone_range:
    opt = Z_P_Optimizer(loci_num=5, samples_num=3, clones_num=clone_count)
    z_opt, p_opt, err_opt = opt.optimal_Z_P(C_in=C, Zs_count=5000)
    errors.append(err_opt)
    print(clone_count, end=", ")
print("done.")

2, 3, 4, 5, 6, 7, 8, 9, 10, done.

In [322]: plt.figure()
aic = 100*np.log(np.array(errors))+2*clone_range
plt.plot(clone_range, aic)
plt.ylabel('Akaike\'s Information Criterion')
plt.xlabel('Number of clones')
plt.title('AIC vs Number of Clones')
plt.show()
```

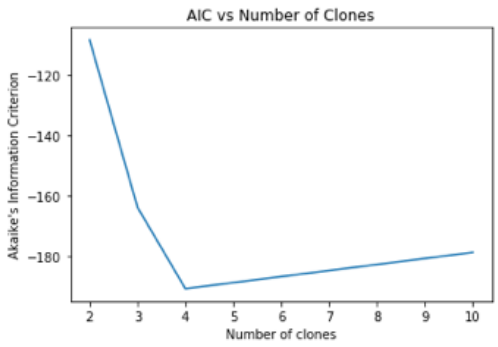


Figure 7: Akaike's information criterion (AIC) was calculated for clone numbers ranging from 2 to 10 for the Gerlinger dataset (where the true number of clones is 4). Increasing from 2 to 4 clones results in a substantial increase in likelihood that outweighs the penalty for the increased degrees of freedom from the additional parameters. However, after 4 clones, increasing the clone number no longer results in an increased likelihood and AIC increases with the penalty from the increased parameters. Our minimum AIC at a clone count of 4 matches the true value for the Gerlinger dataset.

6 References

Gerlinger M, Rowan AJ, Horswell S, et al. "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing." *New England Journal of Medicine* March 2012: 366(10): 883-92.

Gusfield, Dan. "Efficient Algorithsm for Inferring Evolutionary Trees". *Networks* Vol. 21 (1991) p. 19-28. http://csiflabs.cs.ucdavis.edu/~gusfield/nperfect.pdf

Landau DA, Carter SL, Stojanov P, et al. "Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia." *Cell* 152, 714–726, February 14, 2013

Makohon-Moore AP, Zhang M, Reiter JG. "Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer". *Nature Genetics*. 49.3 March 2017.

Mroz EA, Tward AM, Hammon RJ, et al. "Intra-tumor Genetic Heterogeneity and Mortality in Head and Neck Cancer: Analysis of Data from The Cancer Genome Atlas ". *PLoS Medicine* (2015).

Samuel H, Hudson TJ. "Translating Genomics to the Clinic: Implications of Cancer Heterogeneity". *Clinical Chemistry* 59:1 127–137 (2013)

Schwartz, R, Schaffer AA. "The Evolution of Tumour Phylogenetics: Principles and Practice." *Nature Reviews Genetics* 18.4 (2017): 213-29.

Zare H, Wang J, Hu A, et al. "Inferring Clonal Composition from Multiple Sections of a Breast Cancer." *PLoS Computational Biology* 10.7 (2014).