Design, Analytics and Quality Assurance for Emerging Personalized Clinical Diagnostics Based on Next-Gen Sequencing

by

Scott Morris

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2014 by the
Graduate Supervisory Committee:

Esma Gel, Co-Chair
George Runger, Co-Chair
Joseph Paulauskis
Ronald Askin

ARIZONA STATE UNIVERSITY

May 2014

ABSTRACT

Major advancements in biology and medicine have been realized during recent decades, including massively parallel sequencing, which allows researchers to collect millions or billions of short reads from a DNA or RNA sample. This capability opens the door to a renaissance in personalized medicine if effectively deployed. Three projects that address major and necessary advancements in massively parallel sequencing are included in this dissertation. The first study involves a pair of algorithms to verify patient identity based on single nucleotide polymorphisms (SNPs). In brief, we developed a method that allows de novo construction of sample relationships, *e.g.*, which ones are from the same individuals and which are from different individuals. We also developed a method to confirm the hypothesis that a tumor came from a known individual. The second study derives an algorithm to multiplex multiple Polymerase Chain Reaction (PCR) reactions, while minimizing interference between reactions that compromise results. PCR is a powerful technique that amplifies pre-determined regions of DNA and is often used to selectively amplify DNA and RNA targets that are destined for sequencing. It is highly desirable to multiplex reactions to save on reagent and assay setup costs as well as equalize the effect of minor handling issues across gene targets. Our solution involves a binary integer program that minimizes events that are likely to cause interference between PCR reactions. The third study involves design and analysis methods required to analyze gene expression and copy number results against a reference range in a clinical setting for guiding patient treatments. Our goal is to determine which events are present in a given tumor specimen. These events may be mutation, DNA copy number or RNA expression. All three techniques are being used in major research and diagnostic projects for their intended purpose at the time of writing this manuscript. The SNP matching solution has been selected by The Cancer Genome

Atlas to determine sample identity. Paradigm Diagnostics, Viomics and International Genomics Consortium utilize the PCR multiplexing technique to multiplex various types of PCR reactions on multi-million dollar projects. The reference range-based normalization method is used by Paradigm Diagnostics to analyze results from every patient.

ACKNOWLEDGEMENTS

I would like to thank all of those who helped develop and refine this dissertation. First, I would like to thank the members of the committee: Esma Gel, George Runger, Joseph Paulauskis and Ronald Askin for their assistance in developing the mathematical representations of these problems, and refining the ideas. I would like to thank Robert Penny for providing initial vision for the Paradigm test, as well as access to laboratory space, funding and patient specimens required to develop and verify these methods.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Nucleic acid polymers, including deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), play a critical role in life. DNA contains the genetic information and instructions required to run cells. RNA plays a key role in coding, regulation and expression of genes. In the most basic model of molecular biology, RNA is transcribed from DNA and is initially identical to gene regions in the DNA. RNA processing then occurs, which removes certain regions called introns to create the final messenger RNA (mRNA). The mRNA is then translated into proteins, which in turn conduct cellular processes (Krebs 2009). For example, proteins may be structural units of cells, enzymes that conduct chemical reactions, transporters that move other molecules throughout the cell, or may participate in signal transduction pathways that regulate other cellular processes. While many exceptions to this model have been identified, it is sufficient to explain the bulk of processes within the cell (Krebs 2009).

Both DNA and RNA are linear polymers that consist of nucleosides and a backbone. There are four different nucleosides that may make up the polymer, and the order of these nucleosides when read from one end to the other determines which protein is created. The sequence of the RNA, which is in turn determined to a large extent by the sequence of the DNA region that encodes it, determines the initial sequence of a nascent peptide. Many changes occur while processing a nascent peptide into a protein, but alterations in the DNA are usually found in the final protein in a manner that is highly predictable (Krebs 2009). This, combined with the high level of stability and ease of sequencing DNA, leads to DNA sequencing being the primary method used in clinical practice to determine alterations in the code of protein (Slebos 1990). mRNA serves as the messenger between the DNA and

protein, making the level of mRNA present (often referred to as expression) an indicator of how much protein is present (Gry 2009), although this not always true (Gygi 1999). One could ask why the surrogates of DNA and RNA are used rather than analyzing protein directly. While this would be ideal, protein is very heterogeneous, and thus, it is hard to measure accumulation (Chandra 2011) or sequence (Wang 2011) of many proteins simultaneously, while the homogenous nature of DNA (Davey 2011) and RNA (Martin 2011) makes it easy to analyze millions or billions of molecules simultaneously.

Many human medical conditions, including cancer, are caused by altered expression or coding of proteins. Current theory states that cancer is driven by a breakdown of signal transduction pathways that are responsible for important cellular "decisions" (Krebs 2009). For example, the decision of when to proliferate into more cells may be changed to allow growth of healthy cells into a tumor (DeBerardinis 2008), and the decision to initiate apoptosis (Ouyang 2012), or programmed death of the cell, may be prevented. For example, the KRAS gene is an "on/off" switch that controls the cellular decision to proliferate (Bryant 2014). While other signals usually control KRAS, certain mutations in this gene create a version of the protein that is always switched "on", leading to cellular proliferation (Lievre 2006), which is a hallmark of cancer. Another gene, ERBB2, is a sensor for signals that lead to proliferation (Liu 2011). If this gene is overexpressed, i.e., there is excessive ERBB2 present, the cell becomes highly sensitive to growth signals, leading the cell to respond to normal levels of growth factors as if there were a signal to begin proliferation (Menard 2004). As before, this leads to proliferation of the cell.

Within recent years, methods with the ability to detect DNA and RNA with a high level of accuracy have been developed. One technology called nCounter (Nanostring Technologies, Seattle, WA) can detect molecules directly from purified

isolates (Malkov 2009). PCR allows the quantity of a pre-specified sequence of DNA or RNA to be doubled in a sequential process (Mullis 1987). Once the number of molecules present reaches a detectable level, the original amount can be estimated by considering the number of times the quantity was doubled (Taylor 2010). For example, if 10,240 molecules are present after being doubled 10 times, it is clear that the original number present was $\frac{10240}{2^{10}} = 10$. Next-gen sequencing includes a wide variety of technologies that can sequence individual strands of DNA (Hou 2010). While these were originally used to ascertain the sequence of DNA strands and determine whether mutations existed (Schuster 2007), it quickly became evident that the relative amounts of various molecules, such as DNA or RNA, could also be determined (Ekblom 2011). In brief, many molecules are sequenced, the identity of each one is determined by comparing its sequence to a list of known sequences, then relative differences between abundances of different molecules are examined. From these results, one can determine which genes have altered expression (Martin 2011).

For example, if one gene shows much higher relative expression than the others and is known to drive cancer when overexpressed, one can speculate that the gene is driving the cancer and targeting it with a drug may reduce tumor size or growth (Von Hoff 2010). This is a form of the *affirming the consequent* logical fallacy, and thus the use of these results is not perfect. Despite this fallacy, when a potent oncogene is active, it frequently is driving the cancer.

Molecular biology and next-gen sequencing provide powerful tools to understand cancer. Well-designed mathematical tools can process data, determine the underlying drivers of cancer, and determine which drug a patient should receive to treat their cancer.

CHAPTER 2

TWO ALGORITHMS FOR BIOSPECIMEN COMPARISON AND DIFFERENTIATION USING SNP GENOTYPES

Abstract

Aims: Biobanks are frequently required to verify specimen relationships. We present two algorithms to compare single nucleotide polymorphism genotype patterns that provide an objective, high-throughput tool for verification. Methods: The first algorithm allows for comparison of all holdings within a biobank, and is well suited to construct sample relationships *de novo* for comparison to assumed relationships. The second algorithm is tailored to oncology, and allows one to confirm that paired DNAs from malignant and normal tissues are from the same individual in the presence of copy number variations. To evaluate both algorithms, we used an internal training data set (n=1504) and an external validation data set (n=1457). Results: In comparison to the results from manual review and *a priori* knowledge of patient relationships, we identified no errors in interpreting sample relationships within our validation data set. Conclusion: We provide an efficient and objective method of automated data analysis that is lacking for establishing and verifying specimen relationships in biobanks.

Introduction

Biobanks play a critical role in large-scale genomics projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). A primary responsibility of biobanks is to ensure proper chain of custody for specimens, and maintain detailed information about specimen relationships (Hirtzlin 2003). For example, a biobank must know which specimens came from a particular

patient, and must ensure that specimens derived from a single patient are not tracked as originating from different patients.

As the era of "big science" in genomics matures and collaborative efforts involving hundreds of institutions with varying protocols become common (The International Cancer Genome Consortium 2010), many biobanks are no longer able to maintain full chain of custody for their samples and thus are often unaware of errors in specimen relationships. During our experience in the TCGA project (McLendon 2008) (Cancer Genome Atlas Research Network 2011), we found approximately two percent of samples are not derived from the individual identified. For example, a tumor specimen may be incorrectly diagnosed upon pathological review or a diseased/non-diseased pair may not originate from the same individual upon genotyping. Common genomic analysis techniques such as DNA/RNA sequencing are part of many projects and can be used to detect errors in specimen relationships with a high level of accuracy. However, DNA/RNA sequencing is expensive and analyzing results can be time consuming. It is highly desirable to detect inconsistencies in specimen relationships prior to conducting expensive analysis such as sequencing (Glenn 2011).

Two genotyping methods are commonly available to establish and verify specimen relationships: short tandem repeats (STR) and single nucleotide polymorphisms (SNPs). STR has high discriminatory power for each locus (Rosenberg 2002) (Moretti 2001) (Lin 1995), and is available commercially for forensic analysis (AmpFISTR; Applied Biosystems, Carlsbad, CA). However, STRs are susceptible to microsatellite instability in many cancers, making them less suitable for doing comparisons that involve malignant specimens (Vauhkonen 2004). SNPs tend to have a lower discriminatory power per locus (Sanchez 2006), but are less likely to experience changes in cancer. Both SNPs and STRs can be impacted by copy

number events in DNA, especially loss of heterozygosity (LOH) in which one allele is lost (Bignell 2010). Because LOH occurs frequently in cancers, standard comparison methods may flag properly-paired specimens (i.e., specimens from the same individual) as not matching.

At the time of writing, there is no simple, inexpensive, and rapid method available to conduct biobank-wide comparisons to construct specimen relationships *de novo* or to compare diseased/non-diseased specimen pairs from the same individual in the presence of significant LOH. Laboratories specializing in malignancies often find that genotypes with LOH can only be effectively compared via manual review. This is not scalable as a bank containing just 1,000 specimens requires almost 500,000 pair wise comparisons to identify errors in specimen relationships such as those due to unexpected duplication of samples from a patient.

We have developed and evaluated the performance of two algorithms to determine specimen relationships. The first method provides a global comparison of specimen SNP results in order to establish *de novo* relationships between samples. The second method provides a tool that is less sensitive to copy number abnormalities but is well-suited to confirm that a given pair of malignant and disease-free tissues is indeed from the same patient. The aim of this study is to offer an efficient method to perform specimen relationship verification based on SNP results in fairly large datasets (with up to 100,000 specimens) on a common laptop or desktop computer.

Methods

Specimens were procured by the National Cancer Institute for use in the TCGA project from multiple biobanks worldwide, and were shipped to either the International Genomics Consortium (IGC; Phoenix, AZ) or Nationwide Children's

Hospital (NCH; Columbus, OH). DNA was extracted from whole blood and frozen tissue. Unamplified DNA (10 ng) was used in the genotyping process. SNP profiles were collected using the iPLEX Sample Identification panel (Seq ID) (Sequenom; San Diego, CA), a subset of the SNPforID panel (Jin 1995), per manufacturer's instructions. 45 multiplexed polymorphic *loci* were interrogated using Sequenom's Spectrotyper software, which flags results as conservative, moderate, or aggressive, based on the level of confidence in the genotype determination per locus. Data collected at IGC was immediately available for algorithm development (referred to as the training set; n=1504); whereas, data collected at NCH (referred to as the validation set; n=1457) was quarantined until the algorithms had been finalized.

For both algorithms, we made the assumptions that all SNPs are biallelic; each allele occurs with a probability of 0.5, alleles are in Hardy-Weinberg equilibrium, and all SNPs are independent of each other.

Algorithm 1: Global Comparison of Specimens

The global comparison algorithm allows biobank-wide comparisons (i.e., pair wise comparisons of all specimens in the bank). We conducted all possible pair wise comparisons of non-diseased specimens, yielding a total of $\binom{n}{2} = n(n-1)/2$ comparisons, where n is the number of non-diseased specimens in the bank.

For each comparison, the algorithm generates a total score that measures the amount of similarity between two specimens. In particular, we obtain this score by assigning predetermined values to different types of matches between the SNPs of the two specimens on a per-SNP basis. In the following, the symbol τ denotes the number of SNPs assayed in the panel (i.e., 45 for the Seq ID panel). For each SNP, we assign a "SNP Match Score" based on the similarity of the specimens with respect

to this SNP, and calculate an aggregate score of similarity by summing the match scores for each of the τ SNPs.

In our experimental case using the Seq ID panel that assays 45 SNPs, we used the scoring rules defined below. We note that the overall accuracy of the algorithm can be tuned for other purposes by adjusting the values of each parameter, and we select values for comparison of non-diseased specimens that are not expected to experience copy number variants.

Heterozygous SNP Match Score is assigned for comparisons where two specimens have matching heterozygous genotypes (e.g., AT and AT) for a particular SNP. This is given a value of +1 in the present study.

Homozygous SNP Match Score is assigned for comparisons where two specimens have matching homozygous genotypes (e.g., AA and AA) for a particular SNP. This is given a value of +2 in the present study. This value is based on the fact that a homozygous SNP match will occur by random chance with one half of the probability of a heterozygous SNP match, and thus, provides twice as much confidence as a heterozygous SNP match.

Missing SNP Score is assigned for comparisons where the genotype is missing from at least one of the specimens (i.e., no genotype call was made for the SNP). This score is given a value of 0 in this study, because it essentially provides no information for comparison.

Heterozygous SNP Mismatch Score is assigned for comparisons where one specimen is homozygous, and the other specimen is heterozygous (e.g., AA and AT)

for a particular SNP. This score is given a value of -2 in this study. This value was determined experimentally, using the training set as described in the results section. We note that it is possible, albeit rare, that two alleles are present, yet only one is detected.

Homozygous SNP Mismatch Score is assigned for comparisons where both specimens have homozygous genotypes, but they are different (e.g., AA and TT). This is assigned a value of -20. It is unlikely that a new allele would appear while the original allele is lost completely. The score of -20 allows a specimen match to occur with such an event only when the other SNPs show high likelihood of a specimen match.

Box 1 provides a pseudocode for the algorithm that calculates the score for each pair wise comparison. The algorithm, after calculating the score, terminates with the "match" decision, if the obtained score is higher than the threshold set by the user. Otherwise, the specimens are deemed to "not match".

```
score=0
for each SNP
        if (one sample has no result for SNP i) then
                score <-score+ missing Score
        else if (both alleles homozygous AND match) then {i.e., AA and AA}
                score <-score + Homozygous Match Score;
        else if (both alleles heterozygous AND match) then {i.e., AT and AT}
                score <-score + Heterozygous Match Score
        else if (one homozygous, one heterozygous) then {i.e., AA and AT}
                score <-score + Heterozygous Mismatch Score
        else if (both homozygous AND mismatch) then {i.e., AA and TT}
                score <-score + Homozygous Mismatch Score
        end if
 end for
         if (score ≥threshold) then
                cases match
        else
                cases don't match
end if
```

Box 1: The program code for the scoring strategy of the global comparison algorithm.

To better understand the distribution of the total scores we conduct a few theoretical calculations. If we consider a perfectly matching pair wise comparison of specimens (i.e., genotypes for all SNPs are identical between the two specimens) and assume that all alleles amplify so that no data is missing, we can obtain the distribution of the total score, *S*, using the binomial distribution since we can model the total score as the number of "successes" (i.e., matches) in τ trials (i.e., each SNP is one trial). Hence, the random variable that represents the total score, *S*, is a random variable that changes between τ and 2τ (i.e., 45 and 90 for the Seq ID panel).

Recall that a heterozygous SNP match is a "failure" with value +1 (heterozygous SNP match score) and a homozygous SNP match is a "success" with value +2 (homozygous SNP match score). Hence, the random variable $S' = S - \tau$ is binominally distributed with parameters τ and 0.5.

The maximum total score attainable is 2τ= 90 for our panel, and the expected value is $\frac{3}{2}\tau = 67.5$. Assuming there are no missing alleles (i.e. did not amplify), it is impossible to obtain a total score lower than 45 when two specimens are identical. However, if we allow up to five SNPs to be missing due to poor amplification, the critical threshold can be set at 40 when determining if two samples are derived from the same patient (i.e., matching specimens). The expected value of the total score from a pair of specimens derived from different patients (i.e., non-matching specimens) can be calculated to be -135 by conditioning arguments.

Threshold values were set to define multiple confidence levels for a specimen match. The scores for SNP comparison were summed to a single total score that represents the similarity between the two specimens. A total score of 40-49 was considered to be "marginal", 50-59 was considered to be "low", 60-69 was considered to be "medium", and greater than 70 was considered to be "high" likelihood of similarity. Alternatively, one could simply have a single threshold of matching set at 50. A total score value below 39 was considered to be a non-match situation, and hence, was not reported in the output.

Next, we use some basic graph theory principles to look for unexpected matching structures in results. This analysis was conducted both qualitatively by making data easily visualized and quantitatively. In order to visualize the results of all comparisons with matching scores (i.e., total score of at least 40) were exported to a text file that was loaded onto a network visualization tool (Shannon 2003). Figure 1 shows a scaled-down example of such an output, for a dataset with six

specimens, A through F. We use edge colors to show the confidence of each match on a green-yellow-red scale corresponding to the thresholds defined above; with red depicting comparisons that meet the highest threshold. In Figure 1, on right, the pair wise comparisons of specimens A, B and C were all matches, resulting in a clique in the network visualization. Cliques are important to observe, since they provide an easy method to check for consistency of results, i.e., if A matches B and B matches C, we would expect that A also matches C. In this case, one can reliably state that these three specimens are indeed from the same person. The second group on the left of the figure is not a clique; D matches E and E matches F, but D does not match F. This would indicate a situation that should be checked manually. Multiple methods are available in graph theory to verify that these sub networks are cliques quantitatively. Any network that is not a clique is flagged for more detailed manual review.

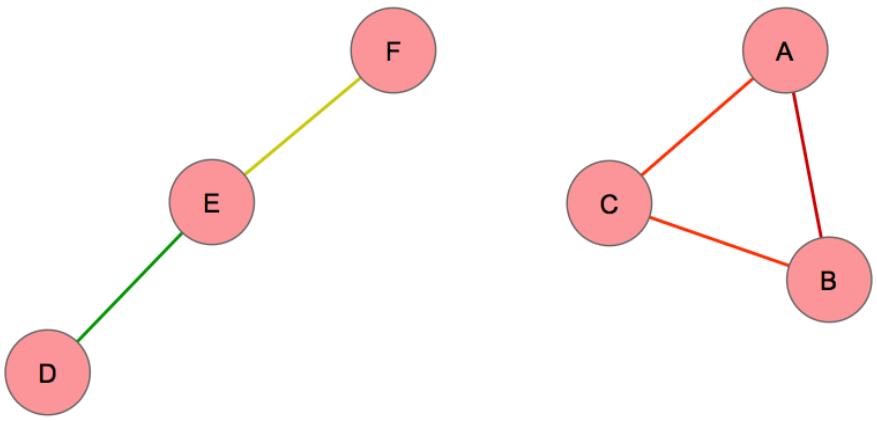

Figure 1: The results for tissue matching comparisons can be visualized in network diagrams. Related samples appear in clique networks with edge colors showing the confidence of each identification match on a green-yellow-red scale; red being the highest.

In order to provide an estimate of false positive rate, we simulated SNP results randomly. Because these results were generated randomly, there should be no matches found and any match found indicates a false positive. These simulations were conducted with different assumptions for minor allele frequency (MAF), starting at 50% and decreasing in increments of 5%. In these simulations, we assumed that all SNPs had the same MAF. Additionally, we calculated the observed MAF for each SNP individually, and randomly generated results using the non-constant observed MAF for each SNP. For each simulation, 50,000 patients were simulated, which resulted in approximately 1.25 billion pairwise comparisons.

Algorithm 2: Comparison of diseased and non-diseased sample pairs

When SNP results are obtained from oncology specimens, there is a higher chance that there will be LOH and such genetic abnormalities can complicate tissue matching. We have developed a method that is insensitive to LOH and provides a systematic, repeatable approach to confirm that a diseased/non-diseased pair is indeed from the same patient. Unlike the previous method, this algorithm is not designed to conduct a global (i.e., everything to everything) comparison, but rather to confirm the hypothesis that a given pair of specimens is derived from the same patient.

This algorithm eliminates sensitivity to LOH in specimen matching by using the non-diseased specimen as the basis of comparison. Consider a homozygous SNP in the non-diseased sample; it should also appear as homozygous in the diseased sample, even if LOH has occurred. However, a heterozygous SNP in the non-diseased tissue can appear as homozygous in the diseased tissue due to LOH (i.e., if one allele is lost), and such an event can mislead inferences on specimen

relationships. Therefore, the algorithm evaluates information only from homozygous SNPs in non-diseased specimens and ignores the heterozygous SNPs that may have been altered by LOH in the diseased specimen. A key benefit to this approach is that no assumptions are made as to the rate of LOH.

Since analysis of diseased specimens can be tricky due to genetic abnormality, the systematic approach provided by this algorithm is important. Because this method is fairly sensitive to incorrect reads, we exclude SNPs with calls flagged as "aggressive" by the Sequenom software. The algorithm, for a given pair of diseased and non-diseased specimens, can be summarized as follows.

1. Ignore SNPs that did not amplify (or had aggressive calls) in either the diseased or non-diseased specimen, and retain only those SNPs that are present for both.
2. Obtain SNP results for the non-diseased specimen. Ignore all SNPs that are heterozygous and retain only those that are homozygous.
3. Obtain SNP results for the diseased specimen and retain results from only those SNPs that were homozygous in the non-diseased sample.
4. Assign N_τ to be the number of SNPs retained after steps 1, 2, and 3.
5. Compare the results from the retained SNPs for the diseased and non-diseased samples, and assign the number of matching SNPs to Z
6. If $N_\tau = Z$ and $N_\tau \geq \kappa$, where κ is a threshold, then the specimens match (i.e. derived from the same patient); otherwise, they do not match. In cases where $N_\tau - Z$ is low, the results can be flagged for manual review.

Table 1 shows each of the steps on a simple example with seven SNPs. The first three rows depict the number of SNPs retained after the first three steps, which

result in $N_\tau = 3$ SNPs. We then observe that S=2 of these remaining SNPs match. Because one of the remaining three SNPs doesn't match between the diseased and non-diseased specimen (i.e. $N_\tau \neq S$), we conclude that this sample pair does not match.

Data	Sample	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7
Initial	Diseased	AA		CC	CC	AT	CT	GG
	Nondiseased	AA	TC	CT		TT	CT	AA
After elimination of SNPs with missing data	Diseased	AA		CC		AT	CT	GG
	Nondiseased	AA		CT		TT	CT	AA
After elimination of SNPs that are heterozygous in the nondiseased sample	Diseased	AA				AT		GG
	Nondiseased	AA				TT		AA

Of seven SNP genotypes reviewed, only three are considered in the comparison of identity since the homozygous genotypes of the nondisease specimen serve as the basis of comparison.
Three SNPs retained (Nτ) and two SNPs matched (total score); as there was one mismatch, these cases are determined not to match.
Nτ: Number of SNPs.

Table 1: A simplistic representation of the scoring strategy for the tissue matching comparison.

The basic process outlined is extremely simple and easy to make calls either manually or programmatically. Thousands of diseased/non-diseased specimen pairings can be easily done in a spreadsheet application, or the algorithm can be coded using common programming languages. However, before this comparison can be done, the correct value of κ, the threshold, must be identified. There are many ways to determine the appropriate value of κ. For example, κ may be chosen such that the false positive rate or false negative rate matches a predefined threshold.

We decided to determine κ to minimize the overall project costs related to false results. In this context, we will define a failing result as a determination that a diseased and non-diseased specimen pair do not match, and a passing result as a determination that the specimens match (i.e., the specimen pair is from the same patient). The derivation below shows a method for determining the optimal value for κ to minimize cost. We note that the same equations can be used to determine a

value of κ that meets some criteria on the false failing or false passing rate. This derivation is conducted once and the same value of κ can be maintained as long as assumptions are unchanged.

For the cost formulation, we use the following notation to represent the two types of costs. C_{FP} is the cost of a false passing result (in dollars). This should include the cost of conducting downstream analysis before discovering the error or the cost of including incorrect data in the downstream analysis. C_{FF} is the cost of a false failing result. In the case of a false failing result, the specimen will be excluded from further analysis and all costs incurred up to this point will be lost. This should include the cost of tissue collection and all lab work conducted prior to and including the SNP analysis but no downstream analyses.

If a total of τ SNPs are being tested, then the number of SNPs that amplify in both diseased and non-diseased specimens (denoted by N_τ) will be a random variable that is binomially distributed with parameters τ and p_a, where p_a denotes the probability that each SNP will amplify in both diseased and non-diseased specimens. This value can be determined by examining historical data and calculating the fraction of SNPs for which a result was obtained. We assume that p_a is the same for all SNPs and that the SNPs amplify independently of each other. That is,

$$P(N_\tau = n) = \binom{\tau}{n} p_a^n (1 - p_a^n)^{\tau - n}, \quad 0 \le n \le \tau$$

Equation 1

Then, we can calculate the distribution for M_τ, the number of homozygotes among the N_τ SNPs that amplify in both diseased and non-diseased specimens. Given the assumption that there are two alleles for each SNP, each with 50% probability of occurring, the probability of homozygous genotypes for each SNP of the diseased and non-diseased specimen pair is equal to 0.5. Then, the conditional distribution of the random variable M_τ given that $N_\tau = n$, is binomial. That is,

$$P(M_\tau = m|N_\tau = n) = \binom{n}{m}(0.5)^m(1-0.5)^{n-m} = \binom{n}{m}(0.5)^n, \quad \text{for} \quad 0 \le m \le n$$

Equation 2

This allows us to calculate the distribution of M_τ for m=0,1,…, τ as:

$$P(M_\tau = m) = \sum_{n=m}^{\tau} P(M_\tau = m|N_\tau = n)P(N_\tau = n)$$

$$= \sum_{n=m}^{\tau} \binom{n}{m}(0.5)^n \binom{\tau}{n} p_a^n (1-p_a^n)^{\tau-n}$$

Equation 3

In this context, we define a false failing result (denoted by FF below) to be the event where less than κ SNPs amplify and are homozygous. This is the only scenario where two correctly paired diseased and non-diseased specimens would obtain a false failing result. Note that our false failing definition contains both those pairings that are correct and those that are not. Thus, it is an upper bound as to the rate of pairings that are correctly paired but were unverifiable. Note that we assume that the risk of a passenger mutation causing a new allele in the tumor is trivial. The false failure probability can be expressed as the probability of obtaining an M_τ strictly less than κ. That is,

$$P_\kappa(FF) = P(M_\tau < \kappa) = \sum_{m=1}^{\kappa-1} P(M_\tau = m)$$

Equation 4

A false passing result (FP) occurs when $M_\tau \ge \kappa$ homozygous alleles are present and they all match by random chance. We define R to represent the event that any diseased specimen SNP will match the non-diseased specimen SNP by random chance, given that the non-diseased SNP is a homozygote. The value of $P(R)$ is calculated as shown below. The generic notation of a SNP with alleles *A* and *a* is used. H is the condition that alleles are homozygous for a given SNP in the non-

diseased specimen. L_n is defined as the normal allele, and L_t is defined as the tumor allele. Then, the probability of the event *R* can be calculated as

$$P(R) = P(L_n = A|H)P(L_t = A) + P(L_n = a|H)P(L_t = a)$$

Equation 5

Hence,

$$P(R) = (0.5)(0.25) + (0.5)(0.25) = 0.25$$

Equation 6

The false passing probability can be found by calculating the probability that there will be zero mismatches among the $M_\tau = m$ SNPs given that the samples are from different individuals. Conditioning on the distribution of M_τ and taking a sum over all values of M_τ greater than or equal to κ we obtain

$$P_\kappa(FP) = \sum_{m=\kappa}^{\tau} \big(P(R)\big)^m P(M_\tau = m)$$

Equation 7

where:

$$P(\text{zero mismatches}|M_\tau = m \text{ and different individuals}) = (P(R))^m$$

Equation 8

Now the expected cost, *C* can be calculated for each value of κ using:

$$C_\kappa = C_{FP}P_\kappa(FP) + C_{FF}P_\kappa(FF), \quad 0 \le \kappa \le \tau$$

Equation 9

Because τ is small and integer (e.g., $\tau = 45$ for the Seq ID panel), we can enumerate the cost for each value of κ and select the lowest cost. In our case, we found the minimum value of C_κ by enumerating all 46 possible values $\kappa = 0, 1, \dots, 45$. If the values for C_{FF} and C_{Fp} are not readily available, researchers can choose a given

P_{FP} or P_{FF} or can construct ROC curves to determine the appropriate value of κ. For our assumptions, we found that $\kappa = 10$ was optimal.

Results

The training set included results from 1504 specimens with results for at least 40 SNPs obtained at IGC during routine processing of specimens for TCGA. These 1504 specimens included 285 patients for which at least one pair of diseased and non-diseased specimens were available (and thus expected to match). Prior to evaluation of the algorithm, specimen IDs were manually examined to determine the specimen relationships so that these can be compared to the ones indicated by the two algorithms developed. Our goal was to determine the effectiveness of the two algorithms we have developed for match testing.

We first tested whether the assumptions of the model are met in our training set. To verify that Hardy-Weinberg equilibrium is present, we determined the fraction of SNPs that were homozygous. If the allele frequency is 50% and Hardy-Weinberg equilibrium is present, we would expect 50% of SNP reads to be homozygous. We selected the first allele listed for each SNP in the panel, and determined its frequency, which one expected to be 50%. Note that the other allele will have a frequency of one minus this value. Figure 2 shows the percent of homozygous genotypes and the frequency of the first allele, along with the 95% confidence interval for the points assuming all assumptions were met. We observed that neither assumption was fully met, since a majority of the points do not fall within the 95% confidence interval. Even though the two assumptions are not met, we tested our algorithms to observe the performance.

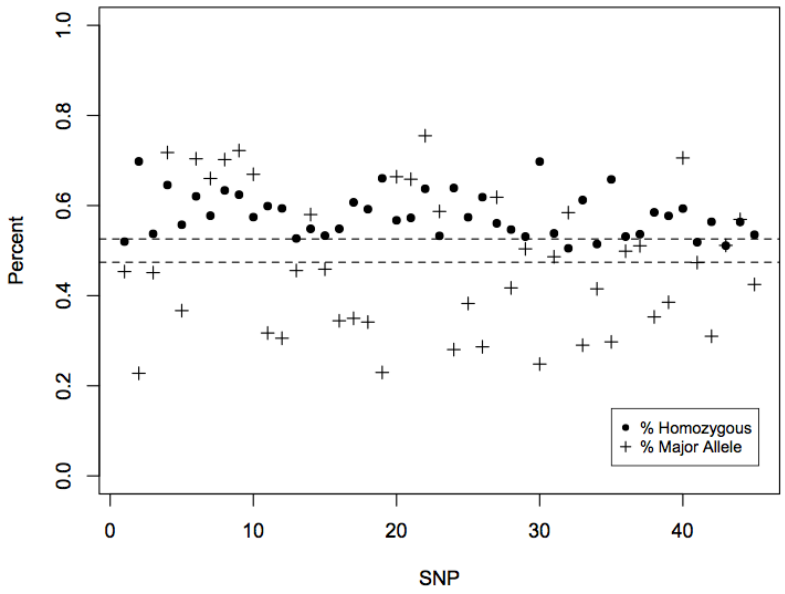

Figure 2: An examination of algorithm assumptions. The percentage of homozygous genetypes and allele frequencies for each SNP. Assumptions are not met perfectly, but are followed approximately.

Performance of Algorithm 1 for Global Comparison

First, the training set was used in the development of the scoring scheme for the global algorithm. We used the following criteria to determine the heterozygous mismatch score:

1. The results should match the results obtained by manual review.
2. When more than two samples match, the resulting network should be a clique (i.e., every specimen matches every other specimen in the group).
3. The scores should be bimodal with clear separation between the matching and mismatched specimen total scores.
4. The heterozygous mismatch score should be as low as possible while satisfying the other conditions.

Based on these criteria, we incrementally increased heterozygous mismatch scores starting with a value of -1 and decreasing it by 1 per iteration. A value of -1 was not sufficient, and allowed samples that had been manually determined to be from two different patients to have a small number of matches between them. These matches violated condition 1, as they did not match in manual review. These matches also violated condition 2, because every sample from one patient did not match every sample from the other patient, and the group consisting of both patients was not a clique in many instances. We next attempted a value of -2 for the heterozygous mismatch score and found all conditions were met.

When all samples are compared, an extremely small proportion is expected to match because there are only a few samples from the same patients. The mean score of all comparisons conducted in our training set (i.e., -114) is only slightly higher than the theoretical mismatch score of -135. Figure 3 is a histogram of all pair wise comparison scores in the training set. We observe that only a single mode is visible and all comparisons appear to be non-matches. This is expected since the majority of the non-diseased specimens at the bank are expected to be from different donors. In Figure 4, we show a similar histogram, but only include scores greater than 0 so the non-matching mode doesn't dominate. From this view, we also observe a clear separation between the two modes, with the tail of the non-matching group on the left and the small number of matches on the right (i.e., matching versus non-matching scores). The mean score for matches (i.e., score ≥ 40) was found to be 68.8, which is consistent with the predicted value of 67.5. Also, the predicted threshold of 40 accurately defines a good threshold between the two modes in the observed bimodal distribution (Figure 6). The algorithm identified a single pairing that was not previously identified. Further review by our panel of

geneticists determined that the sample pair was indeed a match and the algorithm was correct.

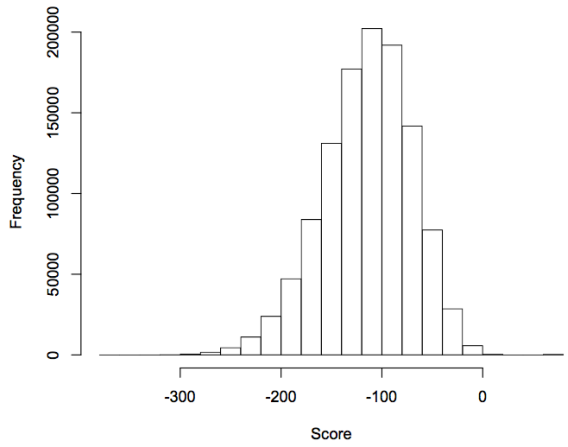

Figure 3: Distribution of scores generated by the global algorithm using the training data set (n=1504)

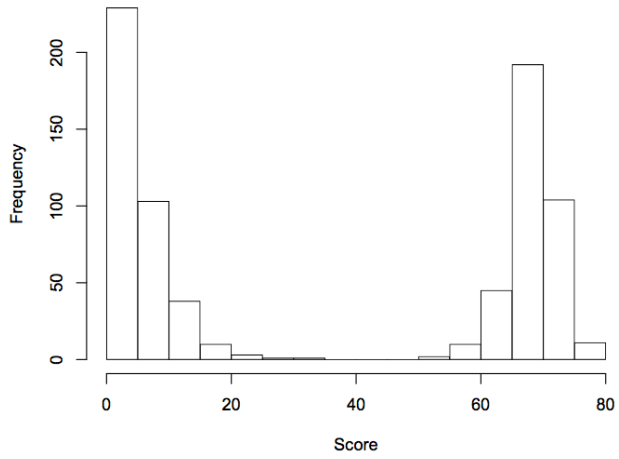

Figure 4: Distribution of scores greater than zero generated by the global algorithm using the training set (n=1504)

The validation set consisted of 1457 samples obtained at NCH during routine processing for TCGA. This number does not include samples with IDs having less than 40 SNP calls. These specimens were quarantined at NCH until all scoring and

decision criteria were finalized on the training set. Using the scores and decision criteria obtained from the validation set, we implemented our global comparison algorithm on this data set. We observed a perfect concordance between the relationships identified by our algorithm and the relationships that were previously identified by geneticists at NCH. Figure 5 and Figure 6 report the same information for the validation set as Figure 3 and Figure 4 described for the training set. Comparing the validation and training sets, we observed that the distribution of scores were nearly identical in both cases. This indicates that our algorithm for global comparison is highly likely to be useful for different data sets.

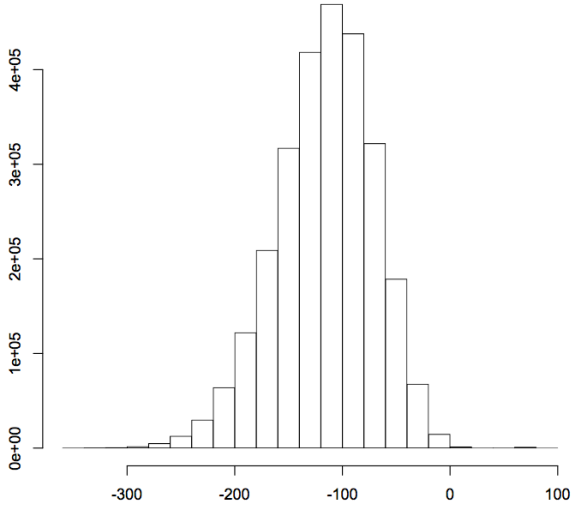

Figure 5: Distribution of scores generated by the global algorithm using the validation set (n=1457).

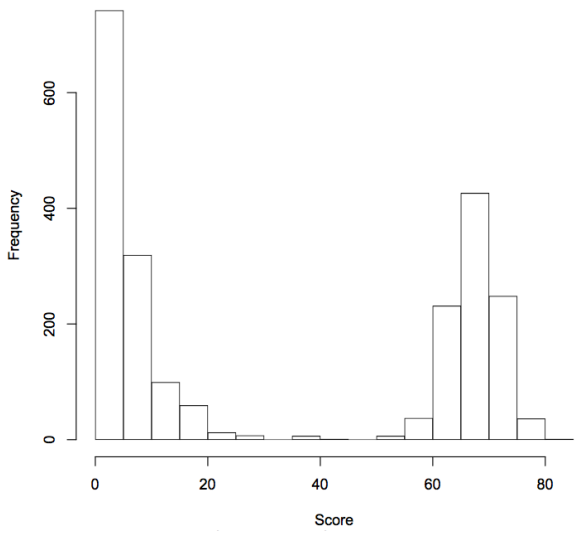

Figure 6: Distribution of scores greater than zero generated by the global algorithm using the validation set (n=1457).

Table 2 shows the average results of the three replicates during simulation. It should be noted that each of these numbers is from a total of 1.25 billion comparisons. It is noted that there is not a major increase the number of marginal results until the MAF drops below 30-35%, and false positive results don't start to occur until MAF below 25-30%. Using the individually calculated MAF values (mean = 0.365, minimum=0.226) for each SNP in our dataset ("Observed"), we find a very manageable number of marginal results and no false positive results.

MAF	# Marginal (score 40-49)	# Positive (Score 50+)
50%	0	0
45%	0	0
40%	0	0
35%	2	0
30%	77	0.3
25%	9632	70
Observed	6	0

Table 2: Marginal and positive results in randomly simulated trials of 50,000 samples for various minor allele frequencies.

Performance of Algorithm 2 for Comparison of Diseased and Non-diseased Specimen Pairs

The method for comparing diseased and non-diseased specimen pairs was derived without the use of data, thus there is no need for the use of a training set. The results from 1504 specimen pairs obtained at IGC during routine processing of specimens for TCGA was used for validation. Since the algorithm is designed for hypothesis testing rather than a global comparison to determine specimen relationships, we only conducted comparisons on paired diseased and non-diseased samples that were believed to come from the same patient. We validated the model by verifying that the results follow the distributions that we identified and have provided above. There is no "absolute truth" available regarding this comparison, as all cases in the data set were originally "believed" to match.

A summary of the results of our tissue matching comparison shows that specimen pairs fall into two separate groups: matching and nonmatching pairs. Figure 7 shows the number of matching and nonmatching SNPs present for each of the 1504 paired specimens. A SNP was only counted if the disease-free specimen was homozygous for that SNP. We observe from the figure that there are two clear

groups, as expected. The first group contains 0 or 1 non-match, and an average of about 25 matches. The second group contains many mismatches with a much smaller number of matches. This result is what was expected. There are three comparisons that have 5 to 10 matches and 0 or 1 nonmatches that do not fall into the expected groups. These comparisons were in fact ambiguous to geneticists as well.

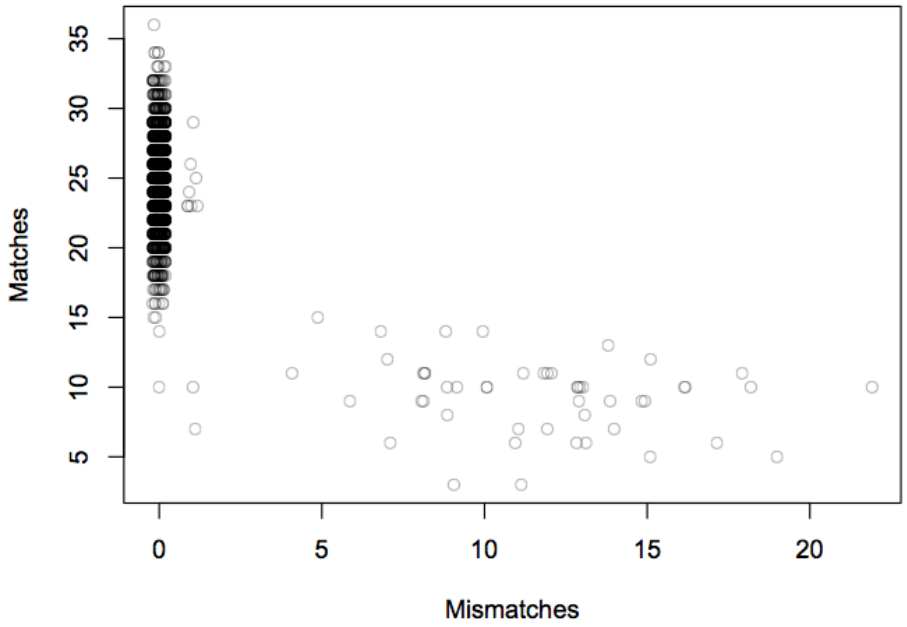

Figure 7: Mismatching vs. matching SNPs for tissue comparisons.

Discussion

We have developed two intuitive and efficient algorithms to analyze SNP genotypes. Although the derivation of these algorithms was mathematically complex, the actual execution is simple. The global algorithm assigns a distance metric to comparisons in a straightforward manner with simple thresholds to decide matching versus nonmatching relationships for the entire database of specimens. The algorithm for paired diseased and non-diseased specimen comparisons, on the other hand, is designed to overcome genetic abnormalities due to LOH and

mathematically assesses the similarity between two specimens using the information from a predetermined number of SNPs. Either of these algorithms can be implemented manually or using a worksheet application or script.

The global comparison algorithm was found to be highly robust in its ability to accurately determine specimen relationships *de novo* even when the assumptions were not perfectly met. In particular, we found that results obtained from the validation set matched the distributions obtained from the training set. The threshold value of 40 that we determined and verified in the training set was also found to provide effective separation between matching and nonmatching specimens in the validation set, as was seen in Figure 8. All known matching specimens were confirmed to match by our algorithm. Furthermore, the algorithm found a previously undiscovered additional set of matching specimen IDs in our training set that was confirmed to be a true match after further manual review.

We observed a small number of cases that were in the region between the two modes (match and non-match) in our validation set (Figure 8) that justified additional review as no scores fell within this range in our training set (Figure 6). Upon review, we found that the one case with a score of 44 (which was deemed to be a marginal match according to the identified decision rules) has a simple substitution of CC to TT in a single SNP, while all other SNPs matched. In this case, we decided that the marginal match result was actually appropriate, because it flagged the result for manual review. Three comparisons had a score of 36, which was just below the threshold for a marginal match. Manual review of these specimens indicated that they did not match; suggesting that the score derived by the algorithm correctly categorized these three suspicious points. The other comparisons had scores that fell between 20-30 (clearly within the non-match range) were confirmed as being non-matches by subsequent manual review. When results

were simulated, we found that our results are fairly robust to assumptions being approximately met. It was noted that with the 1000 Genomes data was deposited into DBSNP after completion of this study, and one could fairly easily find SNPs that match the assumptions in a near perfect manner with minimal effort.

The algorithm for verifying tissue matching between diseased and non-diseased specimens was also very robust (Figures 6 to 8). Of all the comparisons conducted, only three results were ambiguous. In this case, there were only 10 SNP matches and either zero or one SNP mismatch. Although this result is possible, it is highly unlikely that a result this extreme would be attained by our sampling of about 1500 cases. Otherwise, we tended to get more matching SNPs than expected when the data is viewed as a whole, which is likely due to the assumptions of the model not being met. Specifically, allele frequencies that are not 50% and alleles that are not in Hardy-Weinberg equilibrium would be expected to make more matching alleles whether specimens match or do not match.

One weakness of our approach is that we compared our experimental algorithm to the calls made by manual review. It is conceivable that two specimens could have highly similar results and would erroneously be called as a match by our algorithm and manual review when they were in fact not matching. Although our model suggests the risk of such an event is trivial, such a possibility cannot be eliminated. Indeed, all specimens used in this study were expected to have known relationships, and this study would have been irrelevant had they been correct.

In general, we observed more matching SNPs than were expected for both matching and non-matching specimen comparisons for both algorithms. Again, this is most likely caused by some assumptions not being completely satisfied. While it is possible to modify our methods to match the data distribution exactly, it would defeat our main goal of developing a simple, intuitive and efficient algorithm.

Verifying specimen relationships is an important quality control process for biobanking. However, no simple, automated and high-throughput data analysis tools are commonly available. In this study, we provided two algorithms that establish specimen relationships, which can be customized to evaluate results from a variety of SNP assays. The first algorithm provides a global comparison of all specimens in the database and is highly suitable for identification of unexpected sample duplication. The second algorithm is tailored for tissue matching between diseased and corresponding non-diseased tissue that is complicated by LOH. Both algorithms provide rapid and easily interpretable results and can be performed on a common laptop computer. These algorithms are important for automating error detection in sample IDs that could otherwise compromise the quality and effectiveness of downstream processes. Moving forward we hope the integrity and reliability of the biobanking industry is improved by use of quality control tools such as described here.

CHAPTER 3

OPTIMAL HIGH DENSITY MULTIPLEXING OF QUANTITATIVE PCR

Abstract

Aims: It is highly desirable to multiplex PCR reactions, especially ones that are already validated individually. It is rarely possible to multiplex large numbers of PCR reactions by trial-and-error methods due to the large number of combinations possible. Methods: We identify and quantify the consequences of phenomena that cause conflicts between multiplexed PCR reactions, and use an integer programming model to partition existing qPCR reactions into multiplexes while minimizing conflicts. Results: We simulated a variety of scenarios, and determined that it was feasible to multiplex many assays into a small number of multiplexes even when extremely high dimer conflicts existed. For a case study, we used two sets of reactions, one contained 56 quantitative PCR (qPCR) reactions for RNA and one contained 86 DNA assays. The RNAs were successfully multiplexed into four multiplexes with an average of 14 reactions per multiplex, and the DNAs were successfully multiplexed into four multiplexes with and average of 21.5 reactions per multiplex. Conclusion: We provide a reliable method for multiplexing existing quantitative PCR assays into a small number of multiplexes. We determined a method to successfully multiplex PCR reactions while reducing the relative abundance of dimers to desired PCR product.

Introduction

Multiplexed PCR is highly desirable because it allows multiple PCR assays to be run in a single tube (Henegariu 1997) (Edwards 1994). When multiple PCR assays are multiplexed, the primers for all assays are put into the same tube, and each target is amplified independently of the others simultaneously. This reduces assay

setup complexity by reducing the number of reactions run, which leads to reduced costs and consumption of the sample being analyzed. The main limit of multiplexing has been the ability to analyze the results for each reaction separately until recently. For example, when samples are analyzed by gel electrophoresis, the main limit is the number of fragments that can be resolved from each other on a gel (Edwards 1994). Taqman probes allow the amplification level of each reaction to be monitored in real-time via fluorescence (Life Technologies 2013), but this process is typically limited to four reactions per multiplex due to limitations on the number of fluoresce markers that can be distinguished (Applied Biosystems 2013). With the widespread deployment of next-gen sequencing methods (Schuster 2007) (Bybee 2011), it is now possible to individually sequence DNA strands, making it possible to identify each DNA molecule individually with near perfect accuracy. This allows a near unlimited number of molecules to be distinguished.

A method to multiplex very high numbers of molecules is highly desirable. However, this cannot be accomplished by simply putting many PCR assays into a single tube. When one attempts this, there is a high probability that dimers will form (Brownie 1997). Because these dimers occur in a template-independent manner and primers are present at concentrations many orders of magnitude greater than template, they rapidly outcompete the intended reaction. It is fairly straightforward to predict whether a given pair of reactions will generate dimers (Figure **8**). We noted two types of dimers: amplifiable ones that are amplified during the PCR reaction, rendering the entire assay useless by consuming the reaction components, and non-amplifiable ones that cause a lesser loss in quality by rendering fewer primers available to prime template. It should be noted that DNA polymerases used in PCR can only extend DNA in a 5’-3’ direction, and this is the primary distinction between an amplifiable and non-amplifiable dimer.

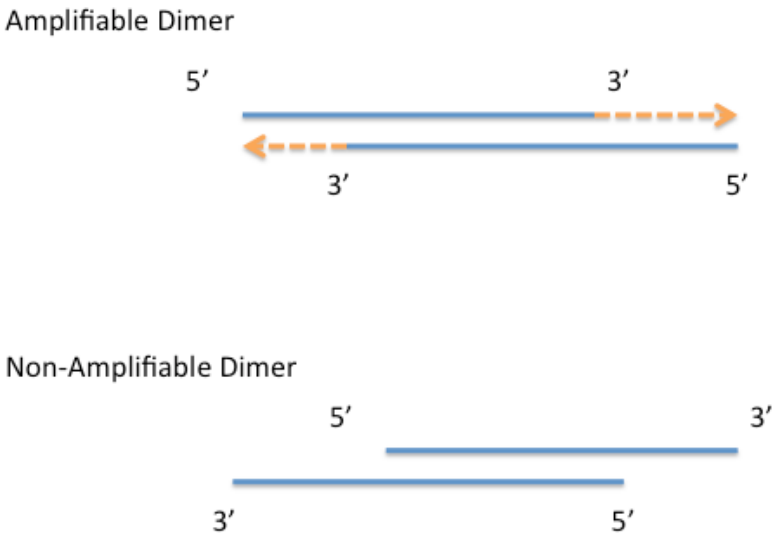

Figure 8: Illustration of amplifiable dimers (top) and non-amplifiable dimers (bottom). Note that the amplifiable dimer can be extended in the 5’-3’ direction, but the non-amplifiable dimer cannot.

Previous work focuses on making multiple assays distinguishable from each other (Rachlin, 2005; Konwar, 2005). Previously, gel electrophoresis was used to discriminate between amplicons, which can only discriminate between different sizes of PCR products. Neither of these studies attempted, however, to prevent the formation of dimers. Thus, the goal of these projects was to create PCR reactions of various sizes that could be easily discriminated by electrophoresis. It is likely that dimers did not present a major issue as the number of assays multiplexed was fairly small. Additionally, dimers were probably not a major issue as they do not resolve well on the types of gels used in this study, thus they could be readily ignored. Fortunately, designing assays to allow separation on gel electrophoresis is no longer necessary with modern next-gen sequencing technologies, as the identity of PCR products can be obtained directly from the sequence result regardless of the PCR

reaction size. Additionally, the types of assays being optimized in these studies involved a qualitative outcome: either a PCR fragment was present, or it wasn't. It was not necessary to measure fragments in a quantitative fashion, or in other terms, preserving relative differences between different amplicons. In order to accomplish this, these studies generated many sets of primers *in silico*, and then determined which ones work best together.

Another method described by Shen *et al.* (Shen 2010) focuses on minimizing the formation of dimers. Much like our team, the authors observed the negative impact of dimers and the need to prevent their formation during PCR. They generated many primers for each target region, and then selected a subset that would not conflict with each other in a single reaction. A greedy graph-expanding algorithm was used. In many regards, this is the work most similar to ours. It is different in the respect that it focuses on creating many assays and selecting only those that are compatible rather that our strategy of multiplexing existing assays without excluding any.

Our proposed method provides three clear advantages over the state-of-the-art. First, it allows us to handle the two types of dimers (amplifiable and non-amplifiable) separately. This is desirable because amplifiable dimers cause substantially more issues than non-amplifiable ones and thus it may be desirable to allow several non-amplifiable dimers rather than accepting a single amplifiable one. Second, it allows the use of existing assays that have already been validated. This allows conversion of legacy PCR assays to next-gen sequencing multiplexes, and also overcomes the recurring issue that only 80-90% of RNA assays will work on difficult specimens by allowing all assays to be individually designed and validated prior to multiplexing. Third, it allows a group of existing PCR assays to be partitioned into multiple multiplexes. This may be desirable when there is a conflict between two

required assays that prevents them from being run in the same multiplex or when there is a limit on the maximum number of assays per multiplex.

Methods

Our goal is to organize a set of PCR reactions into a pre-defined number of multiplexes. This organization must prevent any two assays that would form an amplifiable dimer from being put in the same multiplex. The grouping of PCR reactions must also minimize the sum of the non-amplifiable pairwise dimers present in each multiplex. We assume that dimers only form in a pairwise manner between assays. We do not anticipate or assume that higher order dimers will occur. For example, we assume that putting three assays into a multiplex will not cause a dimer to form if none of the three possible pairings of results in a dimer. Additionally, we want to limit the total number of assays present in each multiplex to a pre-defined number to control the depth of multiplexing. Unlike other methods, we only focus on existing primer sets, and do not make attempts during oligo design to minimize the risk of dimers forming between different reactions.

We define an integer program as follows:

Parameters:

$N \in Z_+ :=$ the number of assays to be multiplexed

$n \in Z_+ :=$ the maximum number of assays to be put in one multiplex

$M \in Z_+ :=$ the number of groups (multiplexes)

We define the decision variable, $x_{ij} := 1$ if assay i is in group j and $x_{ij} := 0$ otherwise, for all $i \in I$ and $j \in J$. Note that we use the term "group" here to represent a subset of assays multiplexed together into a single reaction.

We define the first constraint to ensure each assay is assigned to one and only one group:

$$\sum_{j \in J} x_{ij} = 1 \quad \text{for all } i \in I$$

Equation 1

We next define a constraint to ensure that no more than n assays are assigned to any group:

$$\sum_{i \in I} x_{ij} \leq n \quad \text{for all } j \in J$$

Equation 2

Note that it was previously stated that amplifiable dimers were not allowed within a multiplex. Rather than writing a constraint to prevent this, we instead penalize this occurrence in the objective function. In order to assign penalties to pairings of assays, we need to define a new binary variable. We let $y_{ikj} = 1$ if i and k are both in group j, and $y_{ikj} = 0$ otherwise, for all $i, k \in I$ and $j \in J$. We write the constraint as:

$$y_{ikj} \geq x_{ij} + x_{kj} - 1, \quad \text{for all } j \in J, \quad i, j \in I, i \neq k$$

Equation 3

We assume that a penalty matrix $R_{i \times k}$ provides the penalty of putting reactions i and k together, with each element being r_{ik}. If putting reactions i and k together does not result in any amplifiable or non-amplifiable dimers, then r_{ik} is set to 0. If there is a non-amplifiable dimer, then r_{ik} is set to an arbitrary value of 1 to slightly penalize this undesirable situation. If there is an amplifiable dimer, on the

other hand, r_{ik} is set to a large number, for example N^2, to ensure that it is always preferable to eliminate one amplifiable dimer over any number of non-amplifiable dimers. Finally, we write the mathematical programming formulation as:

$$\text{minimize } \sum_{j\in J}\sum_{i\in I}\sum_{k>i} r_{ik} y_{ijk}$$

Subject to:

$$\sum_{j\in J} x_{ij} = 1 \quad \text{for all } i \in I$$

$$\sum_{i\in I} x_{ij} \leq n \quad \text{for all } j \in J$$

$$y_{ikj} \geq x_{ij} + x_{kj} - 1, \quad \text{for all } j \in J, \quad i,k \in I, i \neq k$$

$$x_{ij} \in \{0,1\} \quad \text{for all } i \in I, j \in J$$

$$y_{ijk} \in \{0,1\} \quad \text{for all } i,k \in I, j \in J$$

Equation 6

This problem was coded into OPL and was solved with CPLEX (IBM, Armonk, New York). We optimized multiple problem instances involving the partitioning og 60-80 reactions into four groups, and the runtime was typically 1 second to 15 minutes on a single 2.4gHz core system with 4GB memory available on Windows 7. These problems were all real problems being solved, and thus, are a good sample of problems likely to be experienced in the future.

Determining the optimal number of multiplexes

Rather than specifying a value of N, the number of multiplexes, in advance, one may wish to optimize this value. Lower values of N result in less work in the lab each time the assay is run, but higher values of N may have the advantage of

resulting in less dimers. There is no direct method to quantify the trade-off between number of multiplexes and the amount of resultant dimers. It was assumed that even one amplifiable dimer was unacceptable, but non-amplifiable dimers may be tolerable.

In order to facilitate decision-making, we create a chart showing the trade-off between the number of multiplexes and the number of dimers. It describes the best possible solution for each value of N. If we set the penalty for an amplifiable dimer to N^2, we can ensure that a higher priority is put on removing amplifiable dimers than non-amplifiable ones. From a given value of the objective function, Z^*, we can determine exactly how many amplifiable $\lfloor Z^*/N^2 \rfloor$ and non-amplifiable dimers ($Z^* \bmod N^2$) are present based on each objective value, where $\bmod$ is the modulo operator.

In order to determine the trade-off, we need to explore three criteria. Each of these values may be determined readily.

- The number of multiplexes, N
- Amplifiable dimers, $\lfloor Z^*/N^2 \rfloor$
- Non-amplifiable dimers, $Z^* \bmod N^2$

For example, if $N = 5$ and we obtain $Z^* = 58$, we can determine that there are two amplifiable dimers and 8 non-amplifiable dimers in this problem.

To simplify somewhat, we can state that there must never be any amplifiable dimers, so the first feasible value of N is the lowest value for which there are no amplifiable dimers. At this point, we need to decide on the tradeoff between non-amplifiable dimers and number of multiplexes. For larger problems, these results could be plotted as shown in Figure **9**. This chart allows the scientist to observe the trade-off between dimers and the number of multiplexes. This also provides an

important secondary function of allowing the scientist to make the final decision based on observation of the results, which is critical for acceptance. This provides additional utility as scientists are trained to scrutinize results to make a decision and are unlikely to trust a method that gave a solution they could not review.

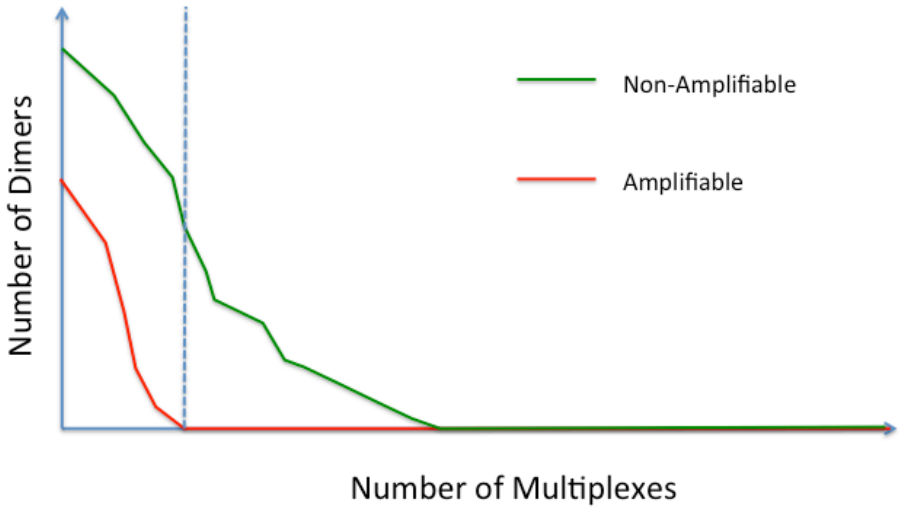

Figure 9: Mock-up of dimers vs. Multiplexes chart. This chart shows how adjusting the number of mulitplexes affects the number of dimers. The dotted line shows the value where no amplifiable dimers are present.

Simulation

In order to understand the effects of various parameters on the performance of this multiplexing strategy, we simulated various penalty matrices, $\boldsymbol{R}$. Many scenarios were studied, but the most valuable studies were the ones where $\boldsymbol{R}$ matrices were randomly generated, with a given probability that any given pair of reactions would cause a dimer. The number of assays to be multiplexed was varied. In each case, the number of multiplexes to be created was varied as shown in Figure **9**.

For the first simulation, a total of 100 assays were multiplexed. Rather than simulating primer sequences, we simulated the values of the penalty matrix, $\boldsymbol{R}$. In

real-world scenerios, this is a deterministic process as the primer sequences are defined, and generating the penalty matrix is a deterministic process. For simulation purposes, we selected 10% probability that any given pair would result in an amplifiable dimer, and a 10% chance it would result in an non-amplifiable dimer. The variance parameter, $\boldsymbol{q}$, was set to 10%. The results are shown in Figure **10**. In this example, amplifiable dimers were eliminated with 8 multiplexes, and all dimers were eliminated by 10 multiplexes. It was noted that the number of amplifiable dimers was non-increasing with respect to the number of multiplexes, but the number of non-amplifiable dimers was not. This is because an amplifiable dimer is penalized substantially higher than a non-amplifiable one, and there are scenarios where allowing several dozen non-amplifiable dimers would allow elimination of an amplifiable one. In this case, the objective function is still non-increasing.

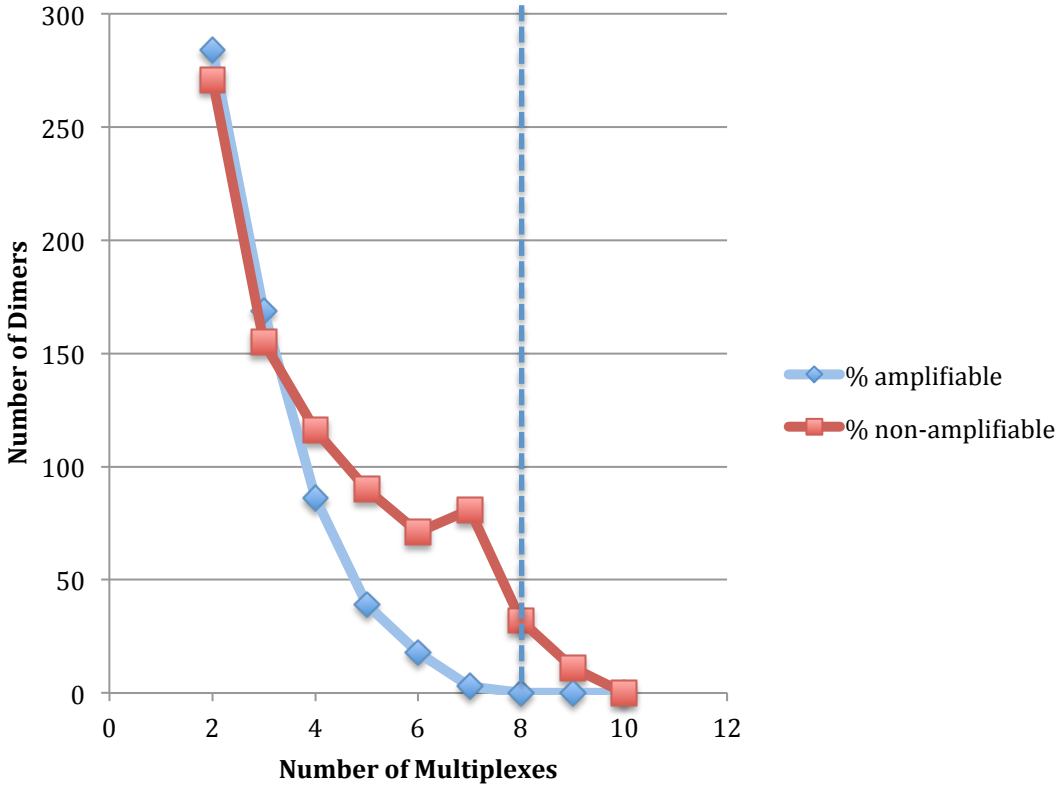

Figure 10: Plot of multiplexes vs. dimers for a simulated scenario involving 100 reactions where the rate of amplifiable and non-amplifiable dimers is 10%.

The same situation was simulated again, with the rate of both non-amplifiable and amplifiable dimers decreased to 5%. The result is shown in Figure **11**. As expected, the number of multiplexes required is lower, with only 5 being needed to remove amplifiable dimers and 6 being needed to remove all dimers.

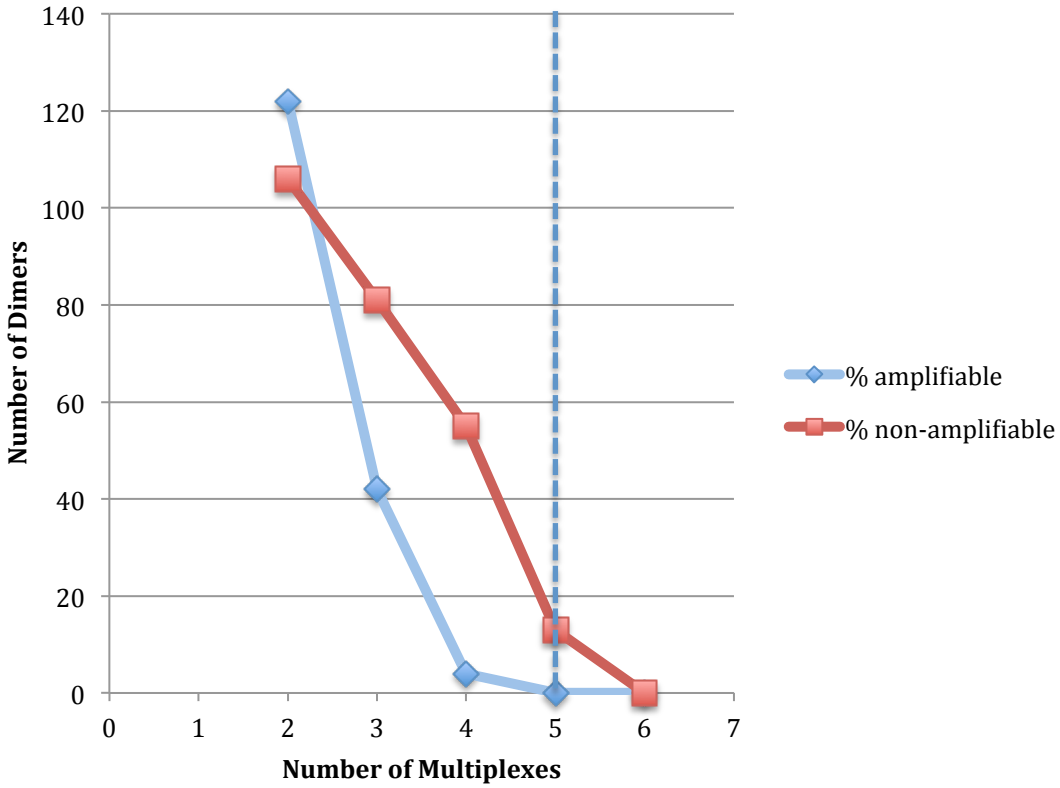

Figure 11: Plot of multiplexes vs. dimers for a simulated scenario involving 100 reactions where the rate of amplifiable and non-amplifiable dimers is 5%.

The same situation was simulated again, with the rate of both non-amplifiable and amplifiable dimers increased to 15%. At this point, the situation is so extreme it is beyond what would be experienced in any realistic scenario with a total of 30% of pairings generating a dimer of some type. The result is shown in Figure **12**. As expected, the number of multiplexes required is higher, with 10 being needed to remove amplifiable dimers and 14 being needed to remove all dimers. This is still fairly low considering that many consider it difficult to multiplex even 4 assays into a single reaction. If we only attempt to remove only amplifiable dimers, we can still average 10 assays per multiplex even with a 15% amplifiable dimer rate.

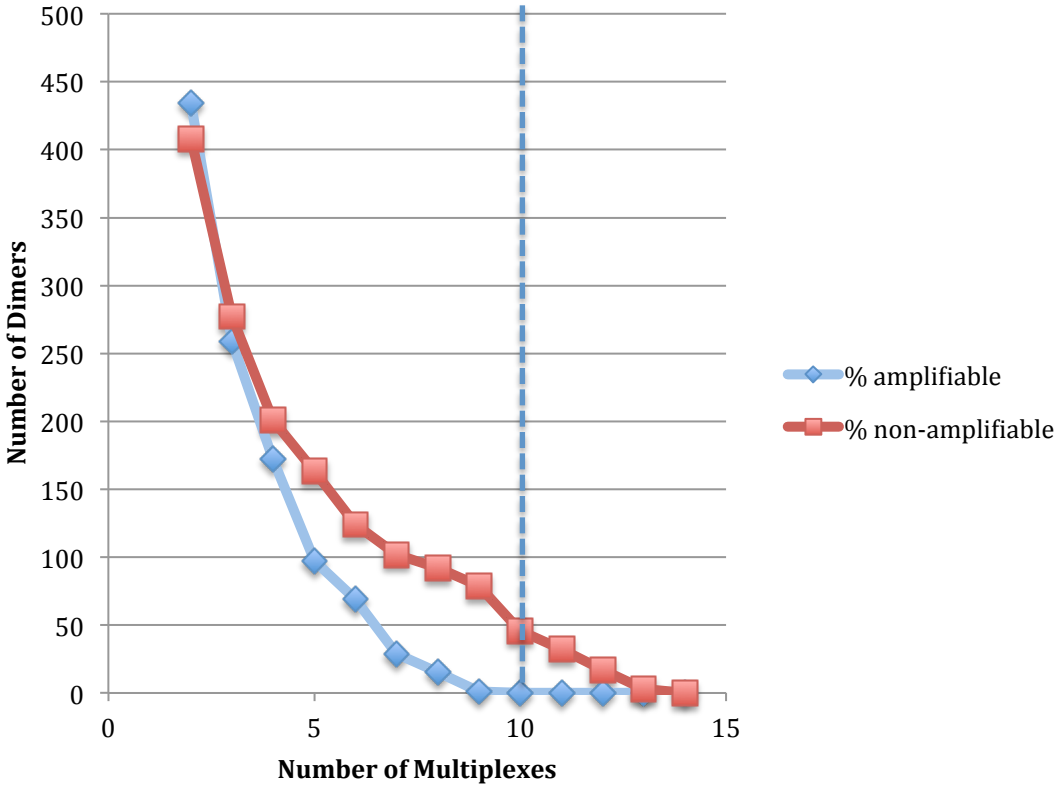

Figure 12: Plot of multiplexes vs. dimers for a simulated scenario involving 100 reactions where the rate of amplifiable and non-amplifiable dimers is 15%.

Case Study

In order to verify the effectiveness of this method with respect to its stated purpose of creating multiplexed PCR assays without interfering dimers, we created PCR assays, verified them, designed a multiplex strategy with the method outlined, then ran the obtained PCR assay to verify dimers were not present. Primers were designed using RealTimeDesign software (Biosearch Technologies 2014). Each primer pair was synthetized and run through a variety of tests to verify it met the quality metrics individually. Reactions were re-designed as needed. Dimers were detected via AutoDimer software (Butler 2004), and a program was written to insert the AutoDimer results into the penalty matrix used by the optimization program. In

brief, this program parses the output from AutoDimer, determines which assays are involved in each dimer, determines whether the dimer is amplifiable or not, then outputs the penalty matrix, $\boldsymbol{R}$, in a format that can be readily read by OPL.

In order to demonstrate the improvement obtained with our method, we started by arbitrarily multiplexing primers into two groups as a non-optimized method. This process was conducted in the laboratory and not *in silico*. This attempt was conducted before we understood how severely dimers affected our process, and only one attempt was conducted as it was immediately clear how poorly multiplexing performed when the grouping of the reactions was done arbitrarily. Figure **13** shows the results of this experiment. It is clear to a molecular biologist that the majority of product present is dimer, especially in the second multiplex where almost no true amplicon is present. One can observe the dimers in as the shorter fragments in the region indicated by "dimer" and the true PCR product as the longer fragments in the region above the top blue line. These were next-gen sequenced, but little meaningful data were obtained, as nearly all reads were the unwanted dimer products.

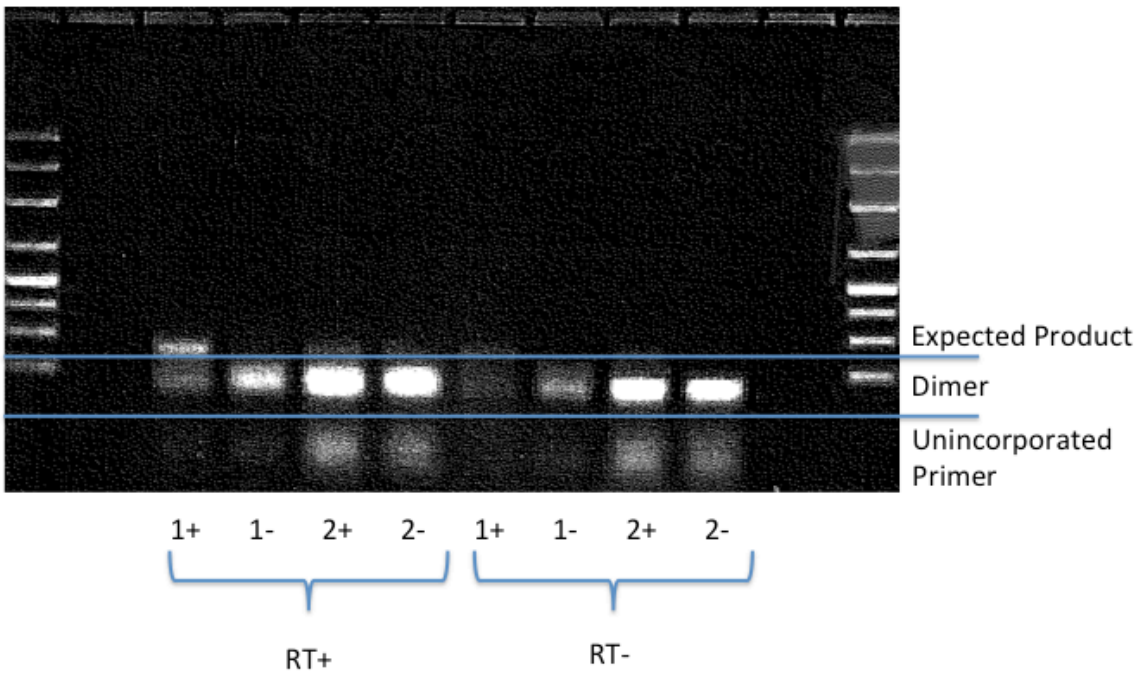

Figure 13: Non-optimally multiplexed PCR. Two multiplexes (1,2) run. +/- indicates whether positive control (with template) or negative control (no template). RT+/RT- indicates whether reverse transcriptase was added. Different multiplexes are run on the horizontal axis, and the direction of the electrophoresis is from the top to the bottom, such that smaller fragments appear lower on the gel. The smallest band on the ladder is 50bp, and the second is 100 bp. Dimers appear as small fragments between the two blue lines. Based on this figure, it is clear that nearly all product of the reaction is dimer.

We then used our optimization method to multiplex the reactions, and sequenced them using a custom Ampliseq panel on Ion Torrent PGM. Two reactions had to be removed after viewing results. Two reactions caused an apparent single-primer dimer, which likely occurred during the emulsification PCR step and thus would not be captured by our method.

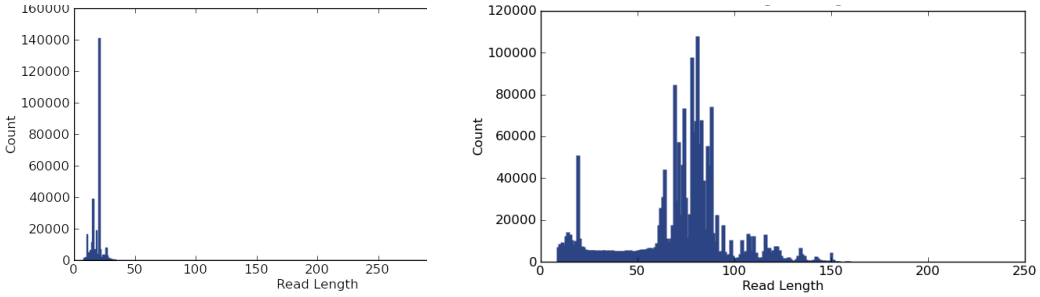

Figure 14: Comparison of read length with random multiplexing (left) and multiplexing via the discussed method (right). Dimers are less than 50bp long, and amplicon is larger than 50bp. With random multiplexing, nearly all reads are dimer, while less than 5% of the reads are dimer after using the discussed method.

After using our multiplexing method, no dimers were visible by gel (Figure **14**). When the same set of primers were randomly assigned to groups and sequenced on the Ion Torrent PGM, nearly all reads were dimer. Dimers can be readily identified by size alone as it is impossible for a dimer to be larger than 50 bp, and it is impossible for amplicon to be smaller than 50bp.

Discussion

We have successfully modeled risk of primer dimer formation and developed a model to prevent their formation during multiplexing. This model was tested in a case study and performed well. Our method is unique in two regards. First, it multiplexes existing assays rather than creating new ones. This allows easier multiplexing of assays that are difficult to create, because it is known in advance that the assays are able to detect their target prior to multiplexing. Second, we focus on preventing dimer formation. Unlike other methods, next-gen sequencing is particularly sensitive to dimers. Preventing dimer formation substantially increases

the number of usable reads by reducing the number of unusable dimer reads. Our method provides an effective solution to an emerging problem.

The use of an integer program substantially decreases the required effort when compared to enumeration of all possible solutions. In the example of 100 assays being multiplexed into 4 multiplexes of 25 reactions each, a total of $\binom{100}{25}\binom{75}{50}\binom{50}{25} = 1.6\times10^{57}$ possible combinations exist. The world's most powerful computer, Tianhe-2, is capable of 38.86 petaflops/s. Using this computer, it would take 1.3×10^{35} years.

The ability to multiplex existing assays is very valuable for those planning to convert existing assays from legacy analysis methods to next-gen sequencing. For example, Genomic Health may benefit by converting their Oncotype Dx panel to a next-gen sequencing assay by this method. This method is also useful for those designing particularly difficult assays that are unlikely to work on the first attempt. For example, Paradigm Diagnostics and Viomics use this method to multiplex complex RNA assays that only have an 80% success rate. This allows assays to be optimized individually, then multiplexed. A third use of this method is to multiplex assays where only a small number of possible designs work for a given target. For example, certain mutations may only be targeted by a single assay design, so methods that design many assays until they find one that can readily be added to a multiplex are not useful. Paradigm Diagnostics uses this multiplexing strategy for this purpose.

This strategy relies upon an integer program, which typically produces a single, unique solution; however, in this case, a series of solutions are produced, allowing the researcher to select the one that best fits their situation. It is hard to anticipate all factors that will influence the day-to-day use of an assay, and scientists generally like to be involved in this decision-making process. For example, a machine

may run 96 reactions at a time in a PCR plate with 8 rows and 12 columns. In this case, a researcher given the option between seven or eight multiplexes may select eight because it allows each patient to have a full row in the PCR plate. Paradigm Diagnostics uses the Rotor-Gene Q instrument (Qiagen), which runs tubes that come in sets of four, so four multiplexes are preferred. All such circumstances cannot be anticipated in advance, and the plot of multiplexes vs. dimers provides a valuable tool for decision making by experts.

In conclusion, we have developed a method to multiplex existing PCR assays without creating dimers. This method is proven experimentally to substantially decrease the number of dimers present while increasing the relative proportion of usable reads.

CHAPTER 4

ROBUST NORMALIZATION OF MULTIPLEXED QUANTITATIVE MOLECULE COUNTING ASSAYS AGAINST A KNOWN REFERENCE RANGE

Abstract

We created a method to normalize gene expression and copy number results obtained from next-generation sequencing (NGS) that is suitable for medical diagnostics. The method allows direct comparison of a patient specimen to a small collection of similar but disease-free tissues. A robust normalization method scales a new sample in such a way that it can be directly compared to the reference range by eliminating certain competitive effects that are unique within NGS. Unlike other methods, ours does not require subjective tuning of parameters. This method allows robust normalization of samples with high levels of expression alteration, and from samples that are highly degraded. This method was tested via simulation and is used by Paradigm Diagnostics, inc. for analyzing patients in their PCDx test.

Introduction

One of the most difficult computational tasks in modern molecular biology is converting RNA expression data into meaningful and useful data (Schlitt 2004). When microarrays, the first major such technology, were introduced, a full understanding and subsequent cure of cancer appeared eminent (Perez-Diez 2000). However, it rapidly became evident that these results required a different type of analysis compared to other biological results (Allison 2006; Ioannidis 2009). The field of bioinformatics rapidly developed around these new assays, but tended to focus primarily on datamining-type techniques such as clustering (e.g., K-means) rather than statistical methods (Allison 2006).

By the time of this writing, a wide variety of biomarkers that can predict response to a particular treatment are known (Von Hoff 2010). Most of these markers were not discovered by meticulous mining of well-curated datasets, but rather by deliberately designing completely new types of pharmaceuticals that target known biomarkers (Moulder 2001; Baselga 2001; Abe 1998). For example, it has long been known that estrogen receptor is present at elevated levels in some cancers, leading them to be excessively sensitive to estrogen (Jensen 1971). Because estrogen is a growth factor, this stimulated cells to proliferate. The drug tamoxifen was developed to block the estrogen receptors, and thus halt the growth of tumors fueled by estrogen receptors, but was completely ineffective against tumors driven by other factors (Abe 1998).

The field of targeted medicine (a.k.a. precision medicine) involves screening patients for factors known to predict drug response prior to treatment (Von Hoff 2010). This field focuses primarily on oncology due to the severe side effects of the drugs and the limited remaining expected. In other forms of medicine, a trial-and-error approach to trying different drugs may still be the norm. Tests to predict treatment response may involve administering a single test for a single drug (e.g., testing for KRAS mutations prior to administering cetuximab) (Baselga 2001), a cancer-specific panel of a few tests (e.g., ER/PR/HER2 for breast cancer) (Bauer 2007), or a large panel that involves hundreds of markers broadly suited for many cancer types (Von Hoff 2010). Many scientists have conducted small-scale tests for a small subset of these markers, and these tests are often effective for a given cancer (Baselga 2001). Few have attempted studies that utilize broad-scale panels (Mook 2007), and even fewer have designed panels that can be used in a number of different cancers (Von Hoff 2010). Such a panel is highly desirable for a person who has exhausted all options for their cancer, especially if it is an uncommon cancer.

Such a panel may discover a treatment that is known to be effective against a marker found in the patient's tumor but is typically used on a different cancer type. For example, ERBB2 gene amplifications are occasionally found in testicular cancer (unpublished results), and the drug Herceptin is known to be effective against breast cancers with ERBB2 amplification (Bauer 2007).

In this study, we aim to analyze a broad-scale panel containing a diverse set of nucleic acid biomarkers that is relevant to multiple types of cancer. This panel will measure four main types of nucleic acid events: mutations, mRNA expression, DNA copy number, and chromosomal abnormalities (i.e., gene fusions). This current study will only focus on the mRNA expression and DNA copy number components. This collection of assays will have higher depth (with respect to the total number of markers) and breadth (with respect to the types of markers) than most tests currently available. The primary goal of using a broad panel is to maximize the probability that an actionable marker will be found. For our purposes, we define "actionable" to be a marker that allows clinical intervention with the intent of extending life or improving quality of life.

This type of test has unique design features that are different than those found in typical tests. Specificity, or 1 minus the type I error, is very important in a test of this type. Patients usually have little time left to live and the drugs have major side effects. The drugs recommended are also extremely expensive. The risk of finding at least one false positive is increased when compared to a test that examines a single marker due to the multiple comparison problem. It is essential that the Positive Predictive Value (PPV), or the proportion of positive calls that are true positives, of the test is also high. It is acceptable to have a fairly poor sensitivity (*i.e.*, 1 minus the type II error). Poor sensitivity is acceptable as the current option of doing nothing effectively has 0% sensitivity; and hence, any level of sensitivity is

considered to be an improvement over the current state of medicine. This is especially true in RNA expression, where no other alternative is available. In the case of DNA mutation (not the focus of this paper), other sensitive methods exist so a higher level of sensitivity is required. The goal of this analysis is not to be perfect, but rather to present patients and their physicians with a better option than what is currently available.

In expression analysis, and to a lesser extent, DNA copy number analysis, changes in relative quantities can be substantial- for example, it is not uncommon for a gene to be expressed in amounts hundreds of fold higher in a tumor specimen than in a cancer-free one (Gordon 2002, Notterman 2001). These outliers consume sequencing wells and suppress reads from other genes. Because there are a finite number of wells for detection within the sequencing chip, we are effectively measuring contrasts. This concern was first reported and addressed by Robinson and Oshlack (2010). For example, suppose we have two genes, Gene 1 and Gene 2 (Figure **15**). If Gene 2 is over expressed and Gene 1 remains constant, it will appear that Gene 2 is more prolific and Gene 1 is less prolific because the higher numbers of Gene 2 compete with Gene 1 for sequencing wells, leading to a downward ”shift” in Gene 1.

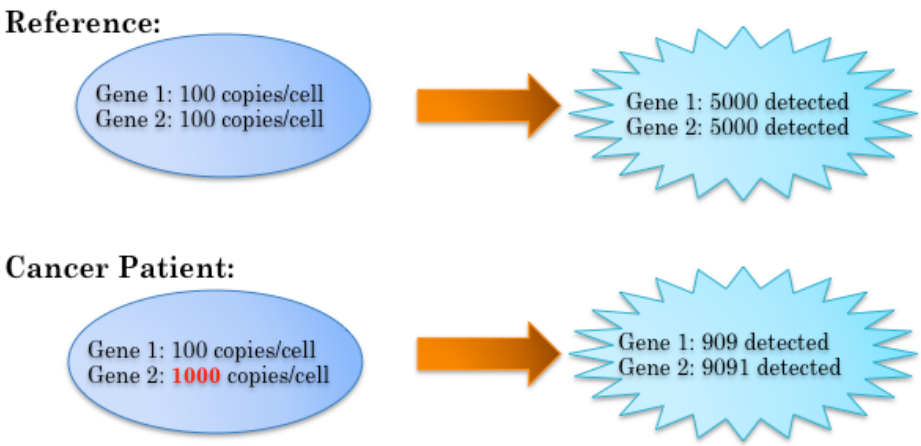

Figure 15: Overexpression on gene 2 causes apparent decrease in the detected levels (left) gene 1 in a cancer patient. Actual values are shown in circles, observed values are shown in star shapes.

To explain this further, we again examine Figure **15**. This figure shows a reference sample that has equal amounts of gene 1 and gene 2 in each cell (circle on left), leading to equal amounts observed in sequencing (star shape on right). However, when the same process is run on a cancer cell (bottom of figure), which has increased expression of gene 2, it leads to an apparent reduction in gene 1 by overcrowding the sequencing wells (in this example, there are only a total of 10,000 sequencing wells present). In this situation, we can multiply the 909 counts of gene 1 by 5.5 to get the corrected value of 5000. Likewise, for gene 2, we can multiply the observed 9091 counts by 5.5 to get a corrected value of 50,000 counts. When compared to the reference sample, we can see that gene 1 expression is not altered (5000 vs. 5000 counts), but gene 2 has experienced a 10-fold increase in expression (50,000 counts in cancer vs. 5000 counts in the reference). Ascertaining the multiplier needed to normalize the result is the primary topic of this study.

As more genes are available, the effect and a solution become more obvious. Figure **16** contains both a "normal" specimen, which does not experience alterations, and a "cancer" specimen, which contains one overexpressed gene. This picture assumes that 1000 reads are sampled from each specimen, and that there is no variance. The "true" value is some representation of the actual number of molecules in the specimen, and the "reads" value is what is observed from sequencing. These assumptions are not realistic but simplify interpretation. It is clear that all of the genes that are not overexpressed are decreased to about 1/4 their original expression. If one knew the "true" number of genes, correcting for this would be trivial. However, this information is not available in real-world situations. Rather, the sampled data must be examined. One could look at each read in the cancer sample to determine which number it must be multiplied by to equal the normal read. In this case, one would obtain: 3.7, 4.0, 4.0, 4.0, 0.2. It is clear that 4 is the correct number, and 0.2 is the outlier. Using a robust loss function achieves the same purpose: it looks for groups of genes for which a similar multiplier is ideal. We explore various types of loss functions throughout this paper to determine how different loss functions affect the ability to normalize.

"Normal" sample:

Gene	True	Reads*
1	46	26
2	672	382
3	195	111
4	568	323
5	278	158

*assume 1000 reads are sampled during sequencing

"Cancer" sample:

Gene	True	Reads*
1	46	7
2	672	95
3	195	28
4	568	81
5	**5560***	790

X 4 =

Corrected
28
380
112
324
3160

Figure 16: Example of the effects of an overexpressed gene in a five-gene panel. The top "normal" sample contains no overexpression, and the bottom "cancer" sample has overexpression of gene 5. In this simple example, there is no variance and the "reads" results are the ones obtained by sequencing. Multiplying the sampled result in the cancer sample by 4 corrects for competition by the overexpressed gene.

This can also be viewed as a robust regression problem (Figure **17**). For this type of problem, we are trying to determine the slope with a known zero intercept. In this instance, the standard least squares method is influenced heavily by the outlier as expected, leading to a substantial difference between the slope estimated and the known correct value. It is clear that, if robust regression were conducted

with a suitable loss function, the resulting slope estimate would be more accurate than the least squares answer.

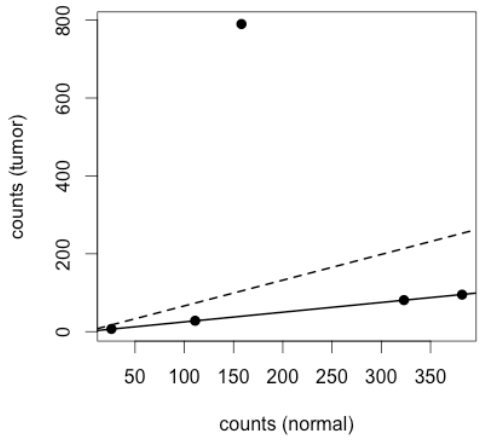

Figure 17: Sample results of a tumor and normal specimen drawn as a regression problem. The solid line represents the known correct solution, and the dotted line represents the solution obtained by standard least squared regression with an intercept of 0. This is the same data shown in Figure 16.

When NGS first arrived on the market, new algorithms to analyze the expression data they generated were not immediately available. As a result, many turned to methods designed for microarray. These include Lowess normalization (Yang 2002) and quantile normalization (Irizarry 2003). However, microarray is substantially different from next-gen sequencing when it comes to measuring expression. In microarray, each gene can be read individually. Thus, the competitive effects observed in next-gen sequencing are not a concern.

The reads per kilobase per million mapped (RPKM) method (Wagner 2012) was one of the first methods designed specifically for next-gen sequencing. It is often still used on NGS data for its simplicity- basically, the number of reads obtained for a given gene are divided by the total number of reads obtained. RPKM makes no effort to compensate for competitive effects. This method works

reasonably well for large panels (for example, 20,000 genes) where only a few hundred experience altered expression levels, because the number of altered genes is very small in comparison to the number of unaltered genes, and thus, they don't occupy enough sequencing wells to alter apparent gene expression of others substantially.

Robinson & Oshlack (2010) proposed the Trimmed Mean of M values (TMM) normalization method to overcome this. They assumed that the majority of genes are not differentially expressed. They log transformed data to decrease variance, especially from very high values. They then compared gene ratios between a new specimen and a reference, removing the upper and lower x% of the ratios. Using the remaining data, they estimated a scaling factor that can be used to adjust a sample to align it to the reference.

The state of the art in this field is tailored to large-scale research projects. Our aim in this project is to design a method that is useful in real-world clinical situations. Medical diagnostics typically focus on small subsets of genes (20 to 200) that are expected to have altered expression. In these cases, the assumption that the vast majority of genes will be unaltered is not valid. Unlike the TMM, we don't need to identify genes with altered expression, nor do we need to assume that the majority of genes are not altered. Our only assumptions are: 1) when genes experience altered expression, the magnitude of expression alteration is independent for each altered gene and 2) there is a subset of genes that are not altered. While our method does assume that there are unaltered genes that can be used for normalization, it does not have hard limits on what proportion of genes are unaltered nor does it require identification and exclusion of altered ones. Our proposed method also accounts for variance in gene expression in addition to the mean, providing a model of greater fidelity to describe this complex process.

Methods

We want to develop a method to analyze next-gen sequencing data that will allow normalization and identification of significant events in both RNA and copy number. We assume that DNA library construction (copy number and mutation) is independent of RNA library construction (expression). We additionally assume that each library made from RNA is independent of the other RNA libraries, and thus, must be normalized separately.

Our primary goal in this normalization process is to center the data correctly without the need to identify outliers (i.e., genes that experience significant alterations upward or downward in expression). As shown in Figure **15**, once the scaling factor is known, one can simply multiply the expression results by this number to get the data adjusted so that it can be directly compared to the values obtained in cancer-free tissues. Thus, we want to use a loss function that focuses on groups of values that are similar to cancer-free tissues and tends to ignore extreme values. This is effectively the opposite of the standard least squares loss function, which puts high weight on extreme points.

We calculate $\boldsymbol{a}$ as the input that yields the minimum value over a loss function (see example in Figure **16** above). The loss function yields a value that estimates how well a given value of $\boldsymbol{a}$ serves to align a cancer specimen to a previously determined reference range. A well-selected loss function will have a high probability of yielding a minimum at the correct value of $\boldsymbol{a}$. While the correct value of $\boldsymbol{a}$ is usually not known in advance, simulated data based on known values can be used to determine the accuracy of different loss functions. Additionally, paired samples

handled in different ways can be sequenced and normalized by a given loss function, with correlation of the final result being a measure of loss function accuracy.

Although these loss functions are typically not convex, our method only requires finding a minimum over a single parameter with a narrow range of values. Indeed, a value of $\boldsymbol{a}$ less than one would imply that genes experienced decreased expression. After running over 100 actual cases, it became apparent that values of $\boldsymbol{a}$ greater than five are not observed. Thus, simply enumerating values over a given range, for example between 0.5 and 5 with increments of 0.01, can be done within a second on a modern computer.

This is a novel method for comparing one sample to a reference range while being robust to outliers. In this application, outliers are the most interesting results, so a method that allows normalization in their presence is highly desirable. This method does not require the comparison of results to additional pre-defined reference genes under the (often false) assumption that the reference genes are not altered. Once normalization is complete, standard statistical tests can be done, such as a t-test, to determine if individual genes are altered.

Derivation of loss function

We want to find the value of $\boldsymbol{a}$, a scalar multiple that corrects for well competition, i.e., the apparent decrease in reads in response to an overexpressed gene. To generate a custom loss function, we start with a logistic function (Equation 1). This function was selected because it is robust to outliers, i.e., it gives higher weight to groups of data points that are similar and lower weight to individual outlier points. For now, we only assume that $\boldsymbol{x}$ is related to $\boldsymbol{a}$. The exact nature of this relationship will be explored later.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Equation 1

We then subtract ½ and square the function to make it two-tailed (Equation 2). This is necessary to penalize both overestimation and underestimation of x.

$$f(x) = \left(\frac{1}{1 + e^{-x}} - \frac{1}{2}\right)^2$$

Equation 2

The plots of the two functions can be seen in Figure **18**. It is clear that modifying the function in this way converts a one-tailed function to a two-tailed one. We also note that as the value of x moves away from zero, the slope of the line approaches zero. This tendency gives less weight to extreme points when we use this as a loss function.

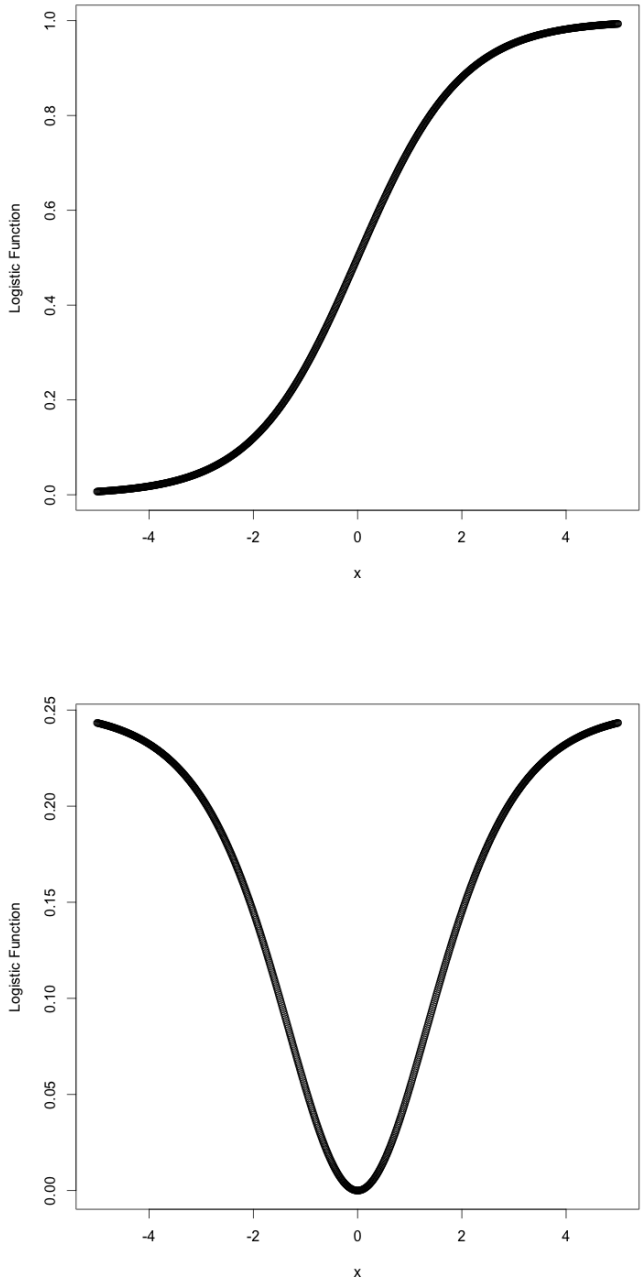

Figure 18: Logistic function (top) and the modified function created by subtracting 1/2 then squaring the logistic function (bottom).

We then make a few modifications to make this function work in this application. We define the following parameters:

$\boldsymbol{m}$ ≔ The number of genes in the panel

$\boldsymbol{I} = \{\mathbf{1}, \mathbf{2}, \ldots \boldsymbol{m}\}$ ≔ The set of genes in the panel

$\boldsymbol{\mu_i}$ ≔ Estimated mean expression from the reference range

$\boldsymbol{\sigma_i}$ ≔ Estimated standard deviation of expression from the reference range

$\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}\}$ ≔ An array of observations for each gene in the new sample to be normalized

$\boldsymbol{a}$ ≔ The normalization parameter

We replace the x term with the statistical distance and take the sum over all genes to find the overall penalty for a given value of $\boldsymbol{a}$.

$$\boldsymbol{L(x, a)} = \frac{\boldsymbol{\mu_i}}{\boldsymbol{\sigma_i}} \sum_{\boldsymbol{i \in I}} \left(\frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-(x_i - a\mu_i)/\sigma_i}} - \frac{\mathbf{1}}{\mathbf{2}} \right)^{\mathbf{2}}$$

Equation 3

The minimum value of $\boldsymbol{L(x, a)}$ solves this problem by determining the best value of $\boldsymbol{a}$. This is true because the equation was written to assign lower values of $\boldsymbol{L(x, a)}$ to better solutions as defined by our model, i.e., where the reads of many genes in the new sample are similar to the reference range. That is, $\boldsymbol{a^*} = \mathbf{argmin}_{\boldsymbol{a \in A}} \boldsymbol{L(x, a)}$.

Logarithmic scale

This method may also be performed on the logarithmic scale. This may particularly useful for RNA sequencing, where one may consider a decrease to 1/10

of the original expression level being of the same magnitude as an increase to 10 times the original expression. This case is handled in a similar manner as the previous description involving a linear scale. All values in both the reference range and the sample are first log-transformed, and then used to calculate the mean and standard deviation.

$$\boldsymbol{L(x,a) = \frac{\mu_i}{\sigma_i} \sum_{i \in I} \left(\frac{1}{1 + e^{-(x_i - \mu_i - \log a)/\sigma_i}} - \frac{1}{2} \right)^2}$$

Equation 4

Note that the only difference in this equation is the substitution of $(\boldsymbol{x_i - \mu_i - a})$ for the distance term rather than the original $(\boldsymbol{x_i - a\mu_i})$. The value of $\boldsymbol{a}$ is then added to all the log-transformed values from the original sample, and then they are transformed back to natural units.

Other Loss Functions

In addition to our proposed loss function, we explore three other loss functions: Huber's T, Hampel's 17A, and least squares. Each of these functions has a different set of properties, and represent different strategies to handle loss. Least squares is by far the most common loss function. It has a parabolic shape, and gives increasingly higher weight to outliers.

Huber's T function acts like least squares when $|\boldsymbol{z}| \leq \boldsymbol{k}$, then increases linearly beyond this range (Figure **19**). While this function does assign more loss to outliers, the loss increases linearly. We selected a value of $\boldsymbol{k = 1}$ for our simulations and found that varying this parameter to values of 2 or 3 had little impact on results. Hampel's 17A function is similar to our function in respect to its tails, which become flat when

$|z|$ becomes large (Figure **19**). We used $k = 1$, $b = 2$ and $c = 3$ for Hampel's 17A simulations. In all cases, we calculated z as the statistical distance in the same manner as before, i.e., $z = (x_i - a\mu_i)/\sigma_i$ for a given value of a.

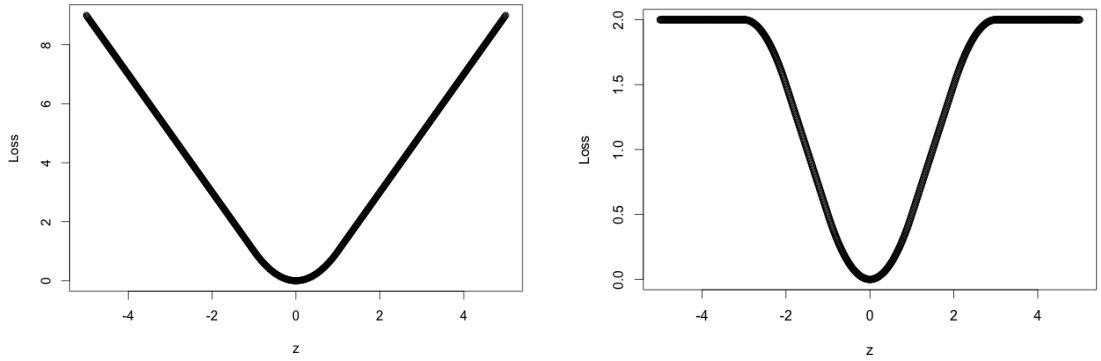

Figure 19: Plot of Huber's T function with k=1 (left) and Hampel's 17A function with parameters k=1, b=2, c=3 (right).

Pre-normalization

In practice, it was observed that the total number of reads obtained from a specimen varied widely. It was not uncommon to observe two similar samples that obtained a total number of reads that varied by more than ten fold. This may be due to many factors, but the major culprit is likely the quantity of amplifiable template present. Calculating the value of a for a given sample will correct for this; however the possible values of a would span a wide range of values in this case.

We decide to pre-normalize samples so that the sum of all reads is the same for all samples. We calculate the normalized value, x_i', as:

$$x_i' = \frac{x_i}{\sum_{j \in I} x_j} \qquad \text{for all } i \in I$$

Equation 5

This results in the sum of all x_i' being 1 regardless of the total number of reads obtained. This also insures that the values of a fall over a much smaller range.

Pseudocode

```
1. Create a vector a.vec of enumerated values of a
2. Obtain raw.counts vector from the new specimen
3. Obtain ref.raw.counts matrix (genes x samples) from the reference range
4. Pre-normalize raw.counts and each column of ref.raw.counts as:
5.      norm.counts[i] = raw.counts[i]/sum(raw.counts) for i=1..gene.count
6.      save result as norm.counts and ref.norm.counts
7. Calculate the mean and s for each row of ref.norm.counts
8. for each value i in a.vec
9.      Loss.net=0
10.    for (each gene j)
11.             Loss.net += mean[j]/s[j]*(1/(1+exp(-(norm.counts [j]-
                  a*mean[j])/sqrt(norm.counts [j])))-1/2)^2
12.             Loss[i]=Loss.net
13.     end for
14. end for
15. Determine which i results in the lowest L(x,a) (Equation 4), and select a[i]
16. For visual inspection, create a plot with a.vec on the x-axis and Loss on the y-
    axis
17. Repeat above for each library if multiple libraries where used
```

Table 3: Pseudocode used calculate a.

Simulation

We assume that there are a fixed number of reads, n, obtained from each sequencing run. In reality, various numbers of reads are obtained, but by pre-normalizing the raw results we can achieve the same effective result. We define r_i to be the average number of counts obtained for gene i in the reference sample, and c_i

to be the number of counts obtained for gene $\boldsymbol{i}$ in a patient sample. $\boldsymbol{I}$ is the set of all genes in the panel. If we ignore stochastic effects and rounding, the number of reads detected, $\boldsymbol{d_i}$, for each gene is:

$$\boldsymbol{d_i} = \frac{\boldsymbol{n}}{\sum_{i \in I} \boldsymbol{c_i}} \boldsymbol{c_i} \qquad \text{for all } \boldsymbol{i} \in \boldsymbol{I}$$

Equation 6

Three major assumptions need to be made, regarding the mean, the standard deviation, and the number of genes in the panel. The mean and standard deviation describe the true distribution from which both the reference range and unaltered sample are drawn. For this purpose, we randomly generated reference range means from U(100,10000) and coefficient of variance was sampled from a uniform distribution with minimum and maximum values of 0.1 and 0.2 (i.e., U(0.1,0.2)). All variances/means were re-generated for each sample of this simulation. For each of the six samples in the reference range and the simulated patient sample, a value was generated by the distribution of that gene's reference. If it is selected as an overexpressed gene, the simulated value was multiplied by the selected value to achieve the "overexpressed" read in the simulated patient sample.

The calculation for the actual (correct) value of $\boldsymbol{a}$, which we call $\boldsymbol{a}'$, can be determined when data is simulated, allowing us to determine the accuracy of our algorithm when compared to the true value. By solving Equation **6** for c_i, we realize that a' is simply the inverse of the ratio:

$$c_i = \frac{\sum_{i \in I} c_i}{n} d_i \text{ Therefore}$$

$$a' = \frac{\sum_{i \in I} c_i}{n}$$

Equation 6

We have developed an R code to vary the two variables requested (proportion overexpressed and number of genes), and run multiple replicates such that we can obtain multiple estimates of a for similar sets of conditions. Using these values, we can determine the standard deviation of the value of a obtained. The pseudocode for this program is shown in Table **4**.

```
1. Create a vector of overexpression levels (single value or sampled for
distribution)
2. Create a vector of gene counts to simulate (e.g., 10,20,30,…)
3. Create a vector of proportion of genes overexpressed (e.g., 0, .1,..., 1)
4. Define the number of reference samples (i.e., 6) and replicates (i.e., 100)
5. for each value of i in the counts vector
6.    for each value of j in the proportion vector
7.         for each value of k in the replicates vector
8.              sample mean from U(100,10000) for each gene (counts total genes)
9.              sample sd from U(.1,.2)*mean for each gene
10.             sample values for each of 6 reference range samples from N(mean, sd)
11.              calculate mean (ref.mean) and standard deviation (ref.sd) from the
                 reference range
12.              sample a new result for each gene (new.sample) from N(mean, sd).
                (This is our new sample that will be normalized)
13.              select proportion*counts genes randomly from new.sample,
                  then multiply each one by a value in levels, sampling a new value
                  from levels for each one. (This simulates overexpression)
14.              Calculate a.e, the expected value of the normalization parameter a,
                 as sum(ref.means)/sum(new.sample)
15.              Pre-normalize: new.sample=new.sample*sum(ref.means)/sum(new.sample)
16.              determine a as the minimum value of the loss function
17.         end for
18.         calculate the mean value of (a.e-a)/a.e
19.         calculate a.sd as the standard deviation of a over the replicates
20.     end for
21. end for
```

Table 4: Pseudocode used to simulate values of a.

In the first run of this simulation, all genes designated as altered experienced an overexpression of 5-fold. In Figure **20**, we explore the difference between the value of a we estimated and the true value of a. Note that our method is nearly

perfect until the proportion of genes overexpressed reaches 40%, then the accuracy drops substantially around 50%. We note no relationship between the number of genes and the variance of our estimate of a, indicating that our assay is suitable for both large and small panels of genes.

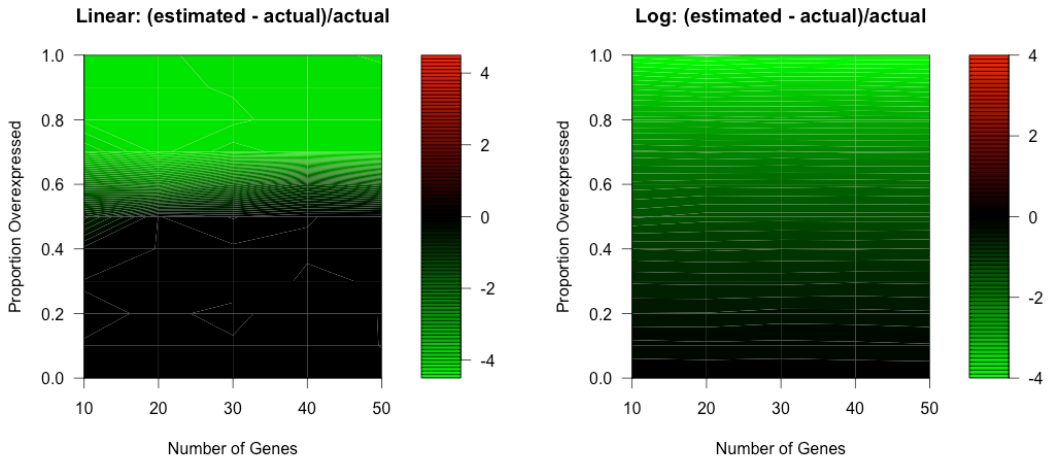

Figure 20: Contour plot of the difference between the estimated value of a and the actual value, a', where all overexpressed genes experience a 5 fold increase.

Figure **21** is interesting at it shows a massive loss in precision at the edge of the plateau. After some consideration, this loss of precision at the edge of the plateau makes sense. At this boundary, either group of points (the unaltered or overexpressed) may define the minimum, and very small stochastic effects determine which set (reference or overexpressed) achieves the lowest value.

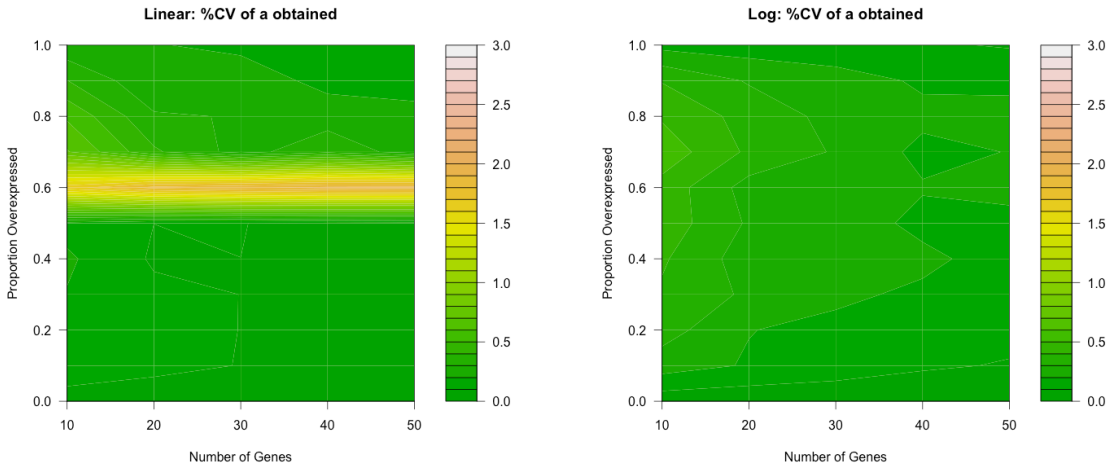

Figure 21: Contour plot of standard deviation of the a estimator as the proportion of 5-fold overexpressed genes and the number of genes is varied.

This is a fairly artificial situation as all genes are altered to the same extent. In practice, genes will be altered to varying levels. The strong dividing line we see here as the algorithm switches from treating the unaltered genes correctly to treating the overexpressed genes as unaltered would not occur in a real-life scenario, as genes are typically altered at varying levels. Additionally, stated assumption of this model was that genes would be altered to varying extents when they are overexpressed.

The simulation was repeated under a more complex yet realistic scenario. As before, a proportion of genes was selected to be altered. Rather than setting all genes to be overexpressed 5 fold, we tried two other scenarios. First, we sampled the fold change from the normal distribution with mean 5 and standard deviation 3 (Figure **22**). Second, we tested a more skewed distribution by sampling the fold change from e^x, where x is distributed normal(0,3). The result from the second, skewed distribution is shown in Figure **23**.

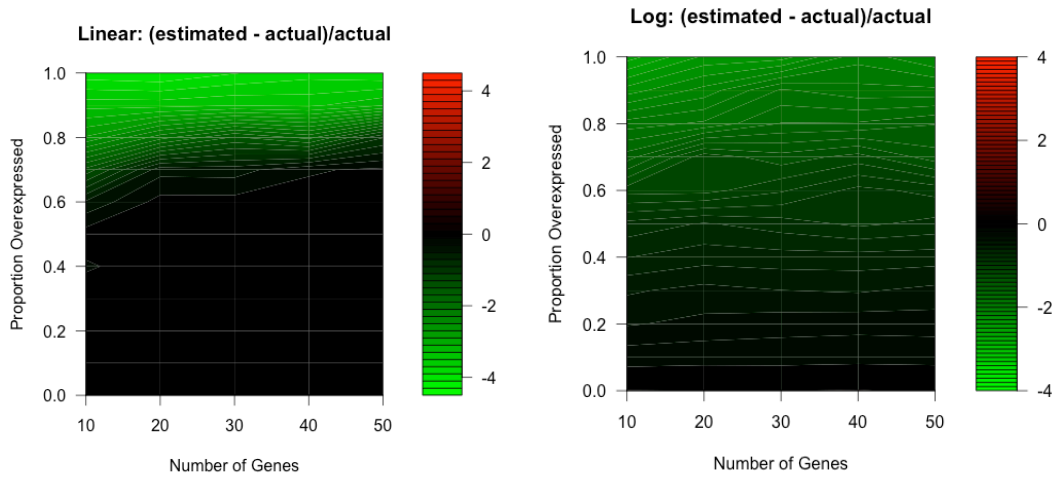

Figure 22: Contour plot of the difference between the estimated value of a and the actual value, a', where overexpressed genes experience a fold increase from *N(5,3)*.

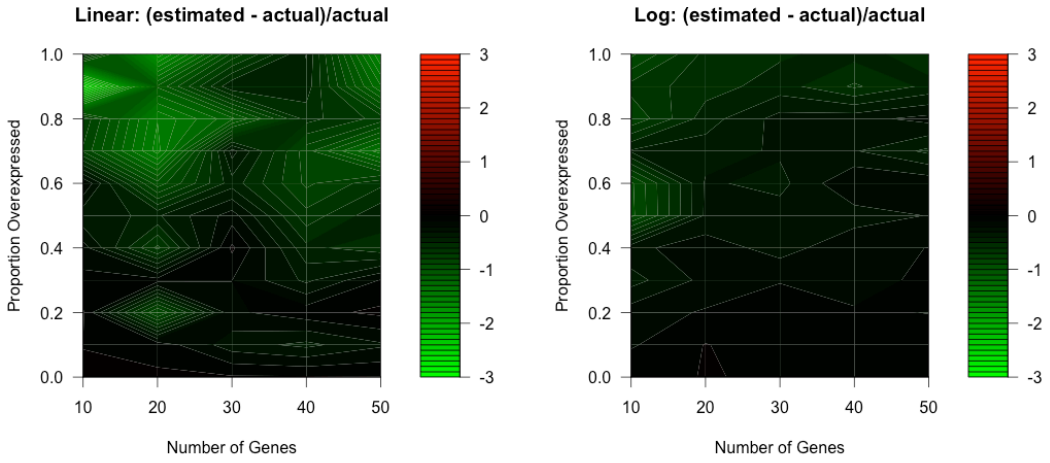

Figure 23: Contour plot of the difference between the estimated value of a and the actual value, a', where overexpressed genes experience a fold increase from $e^x, x \sim N(0,3)$.

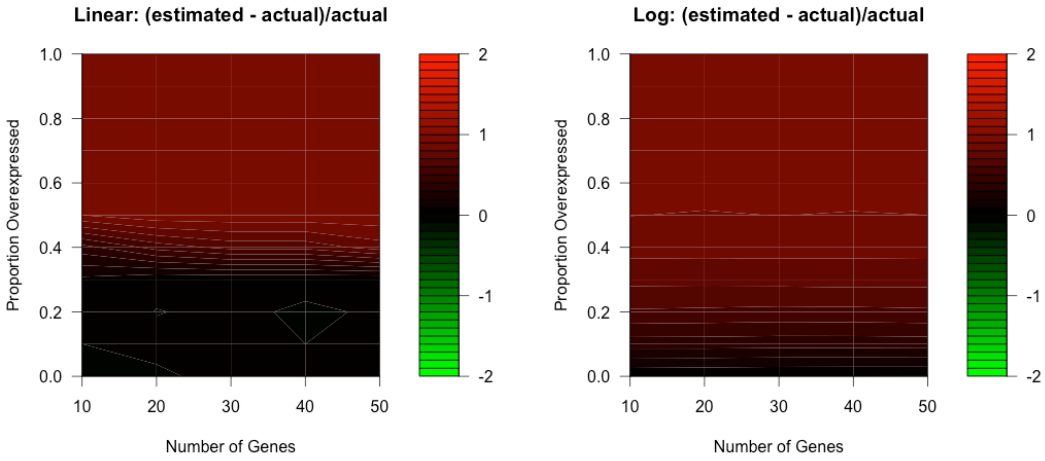

Figure 24: Contour plot of the difference between the estimated value of a and the actual value, a', where genes are underexpressed to 0.01 fold the original level.

In order to better understand the distribution of results obtained, we used the simulation to generate histograms of the values of a obtained, expressed as (actual-expected)/expected. Figure **25** shows these histograms, with various values of proportion overexpressed and gene counts explored. Note that only three different values of gene counts were explored, as this factor doesn't appear to affect the results. 100 replicates were used for each histogram, and genes were overexpressed exactly 5 fold. These results further demonstrate that the method is fairly stable even under a worst-case scenario such as this one.

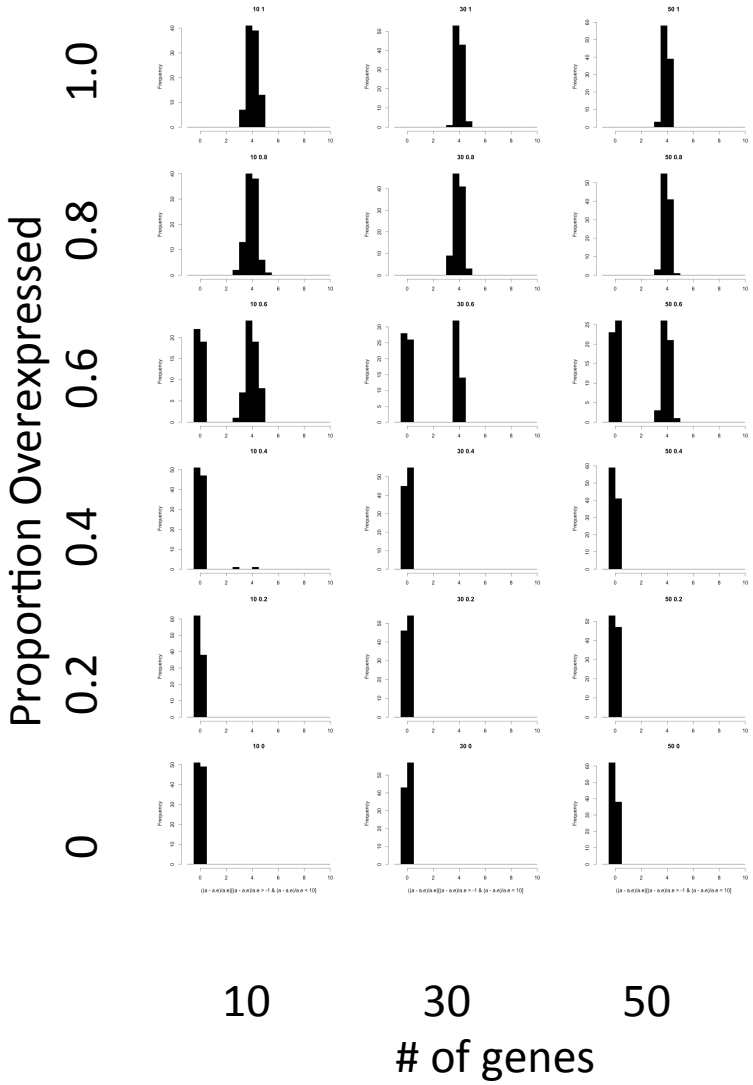

Figure 25: Histograms of (actual-expected)/expected values of a for various combinations of genes and proportion overexpressed, where genes are overexpressed 5 fold. The linear method was used for this figure.

The effects of various loss functions were explored. The results of the scenario where overexpression is distributed as normal (5,3) was selected for this simulation because we decided this scenario was the most representative of the types of data we would experience in routine clinical testing. The results are shown in Figure **26**. An interesting trend is immediately observable: the behavior at the

extremes of a loss function determines how well the function can normalize. As expected, functions that give increasing weight to outliers, such as least squares, perform poorly. Functions that become flat at the extremes, which effectively give less weight to outliers, perform the best. Our method and Hampel's 17A function fall into this category, and they show similar and effective normalization. Huber's T function has linear behavior at the extremes, and as expected performs at a level better than least squares but inferior to functions with flatter tails.

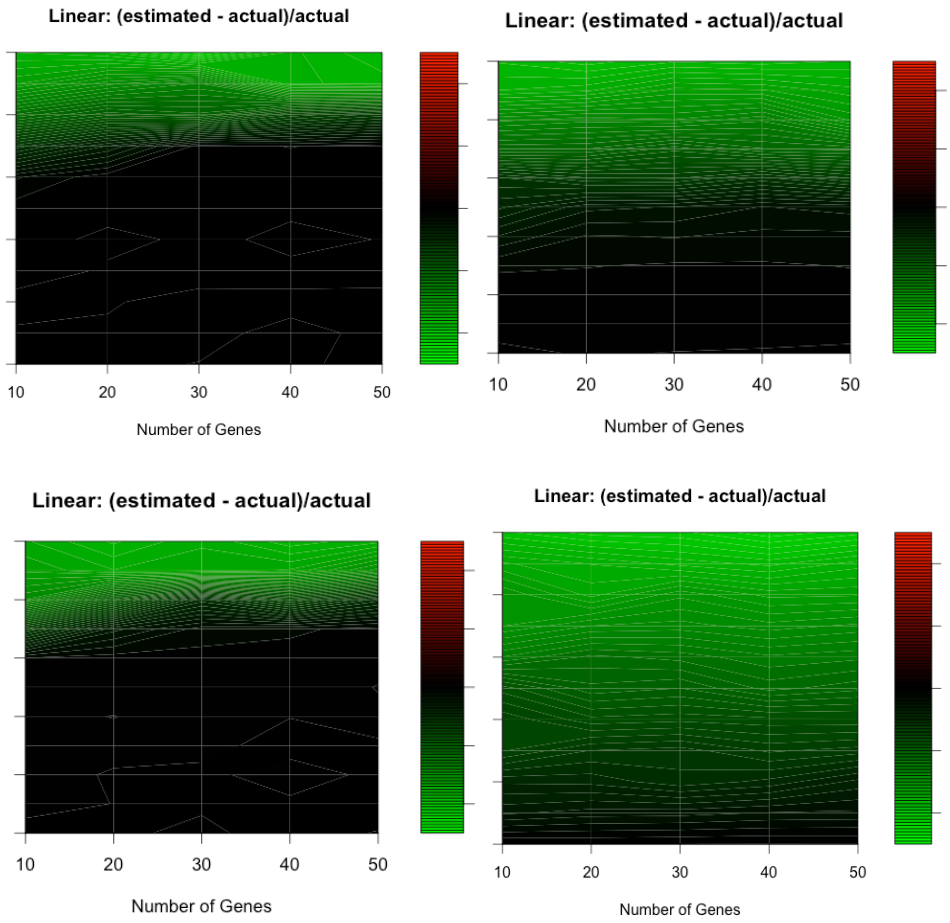

Figure 26: Contour plot of the difference between the estimated value of a and the actual value, a', where overexpressed genes experience a fold increase distributed *N(5,3)* using different loss functions: our proposed method (top left), Huber's T (top right), Hampel's 17A (bottom left), and least squares (bottom right). The color scale ranges from -7 to +7.

Application

Six colon and six lung cases were processed. Each case included both tumor and adjacent normal specimen, and both specimens were provided in both frozen and formalin-fixed, paraffin-embedded (FFPE) format. Thus, each case provided four specimens. We selected both FFPE and frozen specimens because they are

completely different formats of the same specimen. If we can analyze both and get similar results, we can demonstrate that our method is consistent. While the raw results initially look completely different between FFPE and frozen specimens, especially in respect to mean and variance of gene expression, we should detect similar results after normalizing as both specimens are experiencing the same underlying molecular events. Upon analysis of the results, it was clear that the set of four specimens provided for three of the lung cases did not match each other based on germline SNPs, and thus, these were eliminated from the analysis yielding a final set of six colon and three lung cases.

Analytes were extracted, then quantified and normalized. The Highpure FFPE RNA micro kit (Roche 04823125001) was used. The protocol was modified slightly to use a heptane/methanol precipitation to deparaffinize. The DNA was purified with the QiaAmp DNA FFPE tissue kit (Qiagen 56404). The quantity of RNA and DNA was quantified with the Qubit fluorometer, then adjusted to a fixed concentration. The volumes were reduced to a minimum of 100uL for DNA and 20uL for RNA in attempts to reach the target concentration.

Libraries were created via PCR. RNA was converted to cDNA via a proprietary process. Libraries were then built via PCR, with four replicates being created for each library. The ExoSAP-IT reagent was used to remove unincorporated primers, then libraries from a single patient were mixed.

Sequencing occurred via the standard Ion Torrent sequencing protocol provided by Life Technologies using 318 chips. On some of the colon samples, the RNA needed to be repeated. In this case, the RNA only was run on 314 chips, which resulted in similar coverage to that obtained by combined RNA and DNA on a 318 chip.

Sample Plots of a, the Normalization Parameter

Figure **27** shows loss plots (plots of a vs. $L(x,a)$) obtained from a tumor/normal pair from a lung adenocarcinoma patient. Note the smooth, inverted bell curve shape of the normal specimen, as expected. The x-axis is the value of $1/a$, and the y-axis is the loss. We expect smooth bell shaped curves for the normal as is seen in Figure **27**. When overexpressed genes are present in a tumor specimen (right), we observed skewedness towards higher levels of $1/a$ and multiple local minima. We choose to plot $1/a$ on the x-axis rather than a to make the chart more intuitive to view. The tumor specimen shows overexpression of genes in the library depicted by red, causing a skewed distribution shape and decrease in optimal $1/a$ to compensate. As discussed in Chapter 3, the RNA expression assays were split into four multiplexes. Because each multiplex is normalized separately, this process was run four times, yielding four independent optimal values of $1/a$.

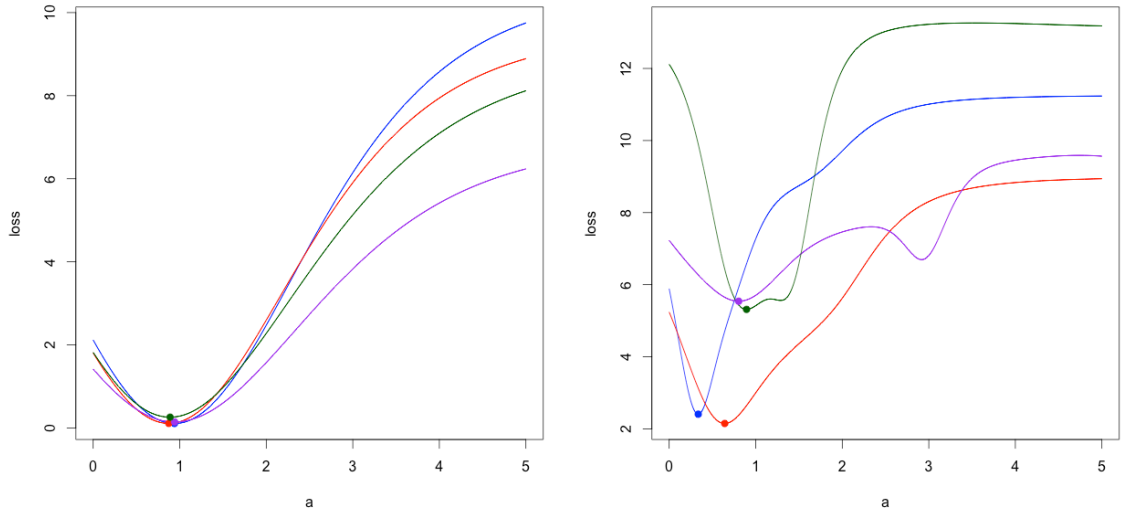

Figure 27: Expression normalization loss plots for various values of a in a healthy tissue (left) and a cancer tissue (right). The x-axis the value of $1/a$, and the y value is loss. Each color indicates a different library (set of multiplexed assays), and the dot on each line shows the optimal value found for $1/a$.

Discussion

We created a method to normalize new results against a reference range that is robust to outliers. Unlike other methods, our method does not require identification of outliers. This is a key distinction, because it eliminates the need to make an assumption that may not hold true for all instances and removes any subjectivity from the interpretation of results. We use a robust loss function in a manner similar to robust regression.

This strategy is novel and is robust to extreme outlines within the dataset. Data was simulated by adding random noise to known values, and the accuracy was measured by comparison of the normalized values to the known values that were used to simulate the data. This strategy is in many ways ideal because at the time of

writing, there is no accepted method to establish baseline truth in complex molecular expression results, making it impossible to have a comparator to determine the accuracy of our results.

We explored a variety of loss functions for this robust normalization method. In general, those that gave less weight to outliers performed better. Examples of functions that perform well are our proposed function and Hampel's 17A. Functions that give increasing weight to outliers, i.e., least squares, performed poorly. Huber's T function gives increasing weight to outliers on a linear basis and showed intermediate performance as expected.

We are currently running a series of patient samples in duplicate to better understand precision of the normalized result obtained with this method. We have shown that the same specimen can be processed in FFPE and frozen formats with similar results obtained. We are also running other methods to assess gene expression and copy number (such as FISH and IHC) to compare our method to the current standard. While FISH and IHC are only available for a small subset of the genes in our panel, this strategy provides additional evidence that our method works for at least some subset of the assays within the panel.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

We have successfully applied a variety of mathematical techniques to various aspects of NGS. While there are many components required to analyzed NGS data, we selected ones that posed major hurtles in regard to computational complexity, assay design and steps critical to accuracy that currently do not have viable solutions. Certain steps that are well defined with clear solutions were not altered, as suitable solutions already exist. In the second chapter, we described simple algorithm to analyze SNP results to determine whether two samples are from the same patient. In the third chapter, we describe a method to multiplex the assays that are used in NGS. In the fourth chapter, we describe a method to normalize results obtained from NGS, allowing direct comparison between a patient specimen and a reference range of disease-free specimens. All of these methods have a similar theme: they directly address a specific issue demanded by clinical medicine, and they all have very low computational complexity. These aspects are interrelated, as a clinical test offered to large numbers of patients must use a reasonable amount of computing power. If too much computing power were required, the test would not be feasible. Computational analysis is currently not reimbursable by insurance or Medicare and thus it would be impossible to fund capital or operational expenses of a large computing infrastructure.

At the time of writing, all of these projects are in use for their designed purpose. The specimen comparison algorithms discussed the second chapter are used routinely in TCGA project. They replaced algorithms that required nearly a full days' worth of computational power on a major hospital's network, and produced better sensitivity and specificity. The multiplexing method discussed in the third chapter is routinely used to design new multiplexes when assays are added to PCR-

based tests offered by Paradigm Diagnostics and Viomics Inc. The normalization method is used by Paradigm Diagnostics for analyzing every patient case, and is also used by International Genomics Consortium for various research projects.

This work can be extended in many aspects. More in-depth research into the current work may be conducted, or similar solutions may be created for other problems. For example, the multiplexing algorithm can be expanded to estimate the cost of the number of multiplexes into the objective function. While there is not a one-size-fits-all solution for this, *ad hoc* solutions can be made for specific projects. For example, in the Paradigm Diagnostics' test, the tubes used come in strips of four, so there is little cost to increase from three to four multiplexes, but substantial additional cost to move from 4 to 5. The normalization algorithm can be modified to account for Poisson variance, *i.e.*, variance that is related to the number of counts obtained. This was attempted in the current study, but was not needed because most genes had coverage (i.e., number of counts) greater than 500, so Poisson variance contributed a trivial amount of variance when compared to the average coefficient of variance of 40% observed. However, if genes frequently had lower coverage (increasing Poisson variance) or an application with less variance due to intrinsic biological factors was used (decreasing variance from other sources), it may become valuable to incorporate a calculation for Poisson variance into the variance calculation.

This research may also be expanded into similar project types. For example, understanding the subclonal makeup of tumors of major interest at the time of writing. Tumors often consist of multiple subclones, or distinctly different populations of cells rather than a single, uniform collection of cells. These subclones are usually progressive, *i.e.*, they occur in sequence, which each new subclone obtaining new events. Some events, such as copy number and mutation, occur in integral

quantities and thus may be well understood by a well-designed integer program. Other events, such as gene expression, are continuous and can be well understood by datamining-type techniques or mixed integer programs. There are likely applications for methods based on loss functions, similar to the strategy used in chapter 4. For example, such a loss function could be used as an objective function, although this would yield a non-convex problem that could only be practically solved with a small number of variables.

We have successfully created low-complexity models for complex biological phenomena directly related to clinical medicine by carefully tailoring models to specific purposes. While these methods are not as broadly applicable as the datamining methods used in medical research, they are highly suitable for clinical medicine. While the bioinformatics field as a whole is rapidly evolving to more complex and generalized methods, our strategy of efficient and specific methods is key to bringing discoveries from the intellectual to practical.

REFERENCES

Abe, O, K Abe, et al. "Tamoxifen for early breast cancer: an overview of the randomised trials." *Lancet* 351, no. 9114 (1998): 1451-1467.

Allison, David, Xiangqin Cui, et al. "Microarray data analysis: from disarray to consolidation and consensus." *Nature Reviews Genetics* 7, no. 1 (2006): 55-65.

Applied Biosystems. *Factors Influencing Multiplex Real-Time PCR.* 10 14, 2013. http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_076529.pdf (accessed 10 14, 2013).

Baselga, J. "The EGFR as a target for anticancer therapy—focus on cetuximab." *European Journal of Cancer* 37 (2001): 16-22.

Bauer, Katrina, Monica Brown, et al. "Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype." *Cancer* 109, no. 9 (2007): 1721-1728.

Bignell, Graham, Chris Greenman, et al. "Signatures of mutation and selection in the cancer genome." *Nature* 463, no. 7283 (2010): 893-898.

Biosearch Technologies. *Real Time Design.* 3 24, 2014. https://www.biosearchtech.com/realtimedesign (accessed 3 24, 2014).

Brownie, Jannine, Susan Shawcross, et al. "The elimination of primer-dimer accumulation in PCR." *Nucleic acids research* 25, no. 16 (1997): 3235-3241.

Bryant, Kirsten, Joseph Mancias, Alec Kimmelman, and Channing Der. "KRAS: feeding pancreatic cancer proliferation." *Trends in Biochemical Sciences*, 2014.

Butler, Peter Vallone and John. "AutoDimer: a screening tool for primer-dimer and hairpin structures." *BioTechniques* 37 (2004): 226-231.

Bybee, Seth, Heather Bracken-Grissom, et al. "Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics." *Genome biology and evolution* 3 (2011): 1312-1323.

Cancer Genome Atlas Research Network. "Integrated genomic analyses of ovarian carcinoma." *Nature* 474, no. 7353 (2011): 609-615.

Chandra, Harini, Panga Jaipal Reddy, and Sanjeeva Srivastava. "Protein microarrays and novel detection platforms." *Expert Review of Proteomics* 8, no. 1 (2011): 61-79.

Davey, John W., Paul A. Hohenlohe, et al. "Genome-wide genetic marker discovery and genotyping using next-generation sequencing." *Nature Reviews Genetics* 12, no. 7 (2011): 499-510.

DeBerardinis, Ralph, Julian Lum, Georgia Hatzivassiliou, and Craig Thompson. "The biology of cancer: metabolic reprogramming fuels cell growth and proliferation." *Cell metabolism* 7, no. 1 (2008): 11-20.

Edwards, Mary and Richard Gibbs. "Multiplex PCR: advantages, development, and applications." *Genome Research* 3, no. 4 (1994): S65-S75.

Ekblom, Robert, and Jordi Galindo. "Applications of next generation sequencing in molecular ecology of non-model organisms." *Heredity* 107, no. 1 (2011): 1-15.

Glenn, Travis C. ""Field guide to next-generation DNA sequencers." *Molecular Ecology Resources* 11, no. 5 (2011): 759-769.

Gordon, Gavin, Roderick Jensen, et al. "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma." *Cancer Research* 62, no. 17 (2002): 4963-4967.

Gry, Marcus, Rebecca Rimini, et al. "Correlations between RNA and protein expression profiles in 23 human cell lines." *BMC Genomics* 10, no. 1 (2009): 365.

Gygi, Steven, Yvan Rochon, et al. "Correlation between protein and mRNA abundance in yeast." *Molecular and Cellular Biology* 19, no. 3 (1999): 1720-1730.

Henegariu, O, N Heerema. "Multiplex PCR: critical parameters and step-by-step protocol." *Biotechniques* 23, no. 3 (1997): 504-511.

Hirtzlin, Isabelle, Christine Dubreuil, et al. "An empirical survey on biobanking of human genetic material and data in six EU countries." *European Journal of Human Genetics* 11, no. 6 (2003): 475-488.

Hou, XiaoGuang, LuFeng Ren, et al. "The next-generation sequencing technology: a technology review and future perspective." *Science China Life Sciences* 53, no. 1 (2010): 33-57.

Ioannidis, John, David Allison, et al. "Repeatability of published microarray gene expression analyses." *Nature genetics* 41, no. 2 (2009): 149-155.

Irizarry, R, B Hobbs, et al. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics* 4, no. 2 (2003): 249-264.

Jensen, Elwood, George Block, et al. "Estrogen receptors and breast cancer response to adrenalectomy." *National Cancer Institute Monographs* 34 (1971): 55-70.

Jin, Li, and Ranajit Chakraborty. "Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics." *Heredity* 74, no. 3 (1995): 274-285.

Konwar, Kishori, Ion Mandoiu, et al. "Improved algorithms for multiplex PCR primer set selection with amplification length constraints." *APBC*, 2005: 41-50.

Krebs, Jocelyn, Steven Kilpatrick. *Lewin's GENES X.* Sudbury, MA: Jones & Bartlett Learning, 2009.

Lievre, Astrid, Jean-Baptiste Bachet, et al. "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer." *Cancer research* 66, no. 8 (2006): 3992-3995.

Life Technologies. *TaqMan® Chemistry vs. SYBR® Chemistry for Real-Time PCR .* 10 14, 2013. http://www.lifetechnologies.com/us/en/home/life-science/pcr/real-time-pcr/qpcr-education/taqman-assays-vs-sybr-green-dye-for-qpcr.html (accessed 10 14, 2013).

Lin, Li, and Ranajit Chakraborty. "Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics." *Heredity* 74, no. 3 (1995): 274-285.

Liu, Bolin, Zeying Fan, et al. "Potent anti-proliferative effects of metformin on trastuzumab-resistant breast cancer cells via inhibition of erbB2/IGF-1 receptor interactions." *Cell Cycle*, 2011: 2959-2966.

Malkov, V. A., Serikawa, K. A., et al. "Multiplexed measurements of gene signatures in different analytes using the Nanostring nCounter™ Assay System." *MC research notes* 2, no. 1 (2009): 80.

Martin, Jeffrey A., and Zhong Wang. "Next-generation transcriptome assembly." *Nature Reviews Genetics* 12, no. 10 (2011): 671-682.

McLendon, Roger, Allan Friedman, et al. "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." *Nature* 455, no. 7216 (2008): 1061-1068.

Menard, S., P. Casalini, M, et al. "Oncogenic protein tyrosine kinases: Role of HER2/neu in tumor progression and therapy." *Cellular and Molecular Life Sciences* 61, no. 23 (2004): 2965-2978.

Mook, Stella, Lauta Vant Veer, et al. "Individualization of Therapy Using Mammaprint® ì: from Development to the MINDACT Trial." *Cancer Genomics-Proteomics* 4, no. 3 (2007): 147-155.

Moretti, Tamyra R., Anne L. Baumstark, et al. "Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples." *Journal of forensic sciences* 46, no. 3 (2001): 647-660.

Moulder, Stacy, Michael Yakes, et al. "Epidermal growth factor receptor (HER1) tyrosine kinase inhibitor ZD1839 (Iressa) inhibits HER2/neu (erbB2)-overexpressing breast cancer cells in vitro and in vivo." *Cancer research* 61, no. 24 (2001): 8887-8895.

Mullis, KB and FA Faloona. "pecific Synthesis of DNA in vitro via a Polymerase-Catalyzed Chain Reaction." *Methods in Enzymology* 155, no. F (1987): 335-350.

Notterman, Daniel, Uri Alon, Alexander Sierk, and Arnold Levine. "Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays." *Cancer Research* 61, no. 7 (2001): 3124-3130.

Ouyang, L, Z Shi, et al. "Programmed cell death pathways in cancer: a review of apoptosis, autophagy and programmed necrosis." *Cell proliferation* 45, no. 6 (2012): 487-498.

Perez-Diez, A, A Morgun, and N Shulzhenko. *Microarrays for Cancer Diagnosis and Classification.* Landes Bioscience. 1 1, 2000. http://www.ncbi.nlm.nih.gov/books/NBK6624/ (accessed 22 2014, 3).

Rachlin, John, Chunming Ding, et al. "MuPlex: multi-objective multiplex PCR assay design." *Nucleic acids research* 22, no. 2 (2005): 2005.

Robinson, Mark, and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data."." *Genome Biology* 11, no. 3 (2010): R25.

Rosenberg, N. A., Pritchard, J. K., et al. "Genetic structure of human populations." *Science* 298, no. 5602 (2002): 2002.

Sanchez, Juan J., Chris Phillips, et al. "A multiplex assay with 52 single nucleotide polymorphisms for human identification." *Electrophoresis* 27, no. 9 (2006): 1713-1724.

Schlitt, Thomas and Patrick Kemmeren. "From microarray data to results." *EMBO Reports* 5, no. 5 (2004): 459-463.

Schuster, Stephan C. "Next-generation sequencing transforms today's biology." *Nature* 200, no. 8 (2007): 2007.

Shannon, P, Markiel, A, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome Research* 13, no. 11 (2003): 2498-2504.

Shen, Zhiyong, Wubin Qu, et al. "MPprimer: a program for reliable multiplex PCR primer design." *BMC Bioinformatics* 11, no. 143 (2010): Online.

Slebos, Robert, Robert Kibbelaar, et al. "K-ras oncogene activation as a prognostic marker in adenocarcinoma of the lung." *New England Journal of Medicine* 323, no. 9 (1990): 561-565.

Taylor, Sean, Michael Wakem, et al. "A practical approach to RT-qPCR—publishing data that conform to the MIQE guidelines." *Methods* 50, no. 4 (2010): S1-S5.

The International Cancer Genome Consortium. "International network of cancer genome projects." *Nature* 464, no. 7291 (2010): 993-998.

Vauhkonen, Hanna, Minttu Hedman, et al. "Evaluation of gastrointestinal cancer tissues as a source of genetic information for forensic investigations by using STRs." *Forensic science international* 139, no. 2 (2004): 159-167.

Von Hoff, DD, Stephenson, JJ, et al. "Pilot Study Using Molecular Profiling of Patients' Tumors to Find Potential Targets and Select Treatments for Their Refractory Cancers." *Journal of Clinical Oncology* 28, no. 33 (2010): 4877–4883.

Wagner, Günter, Koryu Kin. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples." *Theory in Biosciences* 131, no. 4 (2012): 281-285.

Wang, Qing, Raghothama Chaerkady, et al. "Mutant proteins as cancer-specific biomarkers." *Proceedings of the National Academy of Sciences* 108, no. 6 (2011): 2444-2449.

Yang, Yee Hwa, et al. "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." *Nucleic acids research* 30, no. 4 (2002): e15