THE EFFECT OF STRUCTURE IN SHORT REGIONS OF DNA ON MEASUREMENTS ON SHORT OLIGONUCLEOTIDE MICROARRAY AND ION TORRENT PGM SEQUENCING PLATFORMS

by

Saeed Khoshnevis

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2013

Approved by:

_____
Dr. Jennifer Weller

_____
Dr. Cynthia Gibas

_____
Dr. Jessica Schlueter

_____
Dr. Cory Brouwer

_____
Dr. Shan Yan

ABSTRACT

SAEED KHOSHNEVIS. The effect of structure in short regions of DNA on measurements on short-oligonucleotide microarray and Ion Torrent PGM sequencing platforms. (Under the direction of DR. JENNIFER WELLER)

Single-stranded DNA in solution has been studied by biophysicists for many years, as complex structures, both stable and dynamic, form under normal experimental conditions. Stable intra-strand formations affect enzymatic technical processes such as PCR and biological processes such as gene regulation. In the research described here we examined the effect of such structures on two high-throughput genomic assay platforms and whether we could predict the influence of those effects to improve the interpretation of genomic sequencing results.

Helical structures in DNA can be composed of interactions across strands or within a strand. Exclusion of the aqueous solvent provides an entropic advantage to more compact structures. Our first experiments were tested whether internal helical regions in one of the two binding partners in a microarray experiment would influence the stability of the complex. Our results are novel and show, from molecular simulations and hybridization experiments, that stable secondary structures on the boundary, when not impinging on the ability of targets to access the probes, stabilize the probe-target hybridization.

High-throughput sequencing (HTS) platforms use as templates short single-stranded DNA fragments. We tested the influence of template secondary structure on the fidelity of reads generated using the Ion Torrent PGM platform. It can clearly be seen for targets where hairpin structures are quite long (~20bp) that a high level of mis-calling

occurs, particularly of deletions, and that some of these deletions are 20-30 bases long. These deletions are not associated with homopolymers, which are known to cause base mis-calls on the PGM, and the effect of structure on the sequencing reaction, rather than the PCR preparative steps, has not been previously published.

As HTS technologies bring the cost of sequencing whole genomes down, a number of unexpected observations have arisen. An example that caught our attention is the prevalence of far more short deletions than had been detected using Sanger methods. The prevalence is particularly high in the Korean genome. Since we showed that helical structures could disrupt the fidelity of base calls on the Ion Torrent we looked at the context of the apparent deletions to determine whether any sequence or structure pattern discriminated them. Starting with the genome provided by Kim et al (1) we selected deletions > 2 bases long from chromosome I of a Korean genome. We created 70 nucleotide fragments centered on the deletion. We simulated the secondary structures using OMP software and then modeled using the Random Forest algorithm in the WEKA modeling package to characterize the relations between the deletions and secondary structures in or around them. After training the model on chromosome I deletions we tested it using chromosome 20 deletions. We show that sequence information alone is not able to predict whether a deletion will occur, while the addition of structural information improves the prediction rates. Classification rates are not yet high: additional data and a more precise structural description are likely needed to train a robust model. We are unable to state which of the structures affect in vitro platforms and which occur in vivo. A comparative genomics approach using 38 genomes recently made available for the

CAMDA 2013 competition should provide the necessary information to train separate models if the important features are different in the two cases.

ACKNOWLEDGMENTS

I would never have been able to finish my dissertation without the guidance of my advisor, help from friends, and support from my family and wife.

I would like to express my deepest gratitude in acknowledging my advisor, Dr. Jennifer Weller, for her excellent guidance, encouragement, caring, and affection she has showed on me during the course of my work. I appreciate all her contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating.

I would like to thank Dr. Christopher Overall, Dr. D. Andrew Carr, and Deepthi Chaturvedi who as good friends, were always willing to help and give their best suggestions.

I would also like to thank my mother, sister, brother, and brother-in-law. They were always supporting me and encouraging me with their best wishes.

Finally, I would like to thank my wife, Dr. Nahid Babaesfahani for her love, support, and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: BACKGROUND

1.1: Abstract

As nucleic acids fold their properties change. This is taken for granted with functional RNA molecules, but the implications for assays such as microarrays and sequencing are seldom considered. Since such assays are the fundamental data on which genomics and functional genomics studies are based, the implications when errors are present are large. A number of nucleic acid modeling platforms exist that allow one to predict the structures present under experimental conditions, but the predictions do not take into account adjacent larger structures, nor are they usually tested in the lab. In this work we prepared a number of DNA constructs containing specific structures adjacent to the sequence to be measured and tested their performance on 1) long-oligonucleotide microarrays and 2) short-read sequencers. Finally, to determine whether the effects have any bearing on measurements of the human genome 3) we modeled regions of the human genome that are stated to contain short deletions, to determine whether structural motifs might signal those events.

1.2: Introduction

The relative stability of a DNA duplex structure depends primarily on the interactions between nucleotides and other nucleotides and nucleotides and solvent constituents, including hydrogen bonds between bases and between bases and surrounding solution molecules, and base-stacking interactions between adjacent bases.

Breslauer el al (1986) (2) published the calorimetric measurement of entropy (ΔS) and enthalpy (ΔH) of all possible nearest-neighbor interactions of DNA/DNA duplexes, which facilitated the reliable predictions of the overall stability of any DNA duplexes (the free energy (ΔG)) from their primary sequence.

Factors which have a great influence on the stability of DNA duplexes can be classified into: a) DNA sequence, its length and fidelity of pairing, b) mispaired and mismatched pairs and their position in a given duplex (3) and 3) environmental factors such as cation concentration and pH. As expected, most of the mismatches and mispairs are destabilizing to the duplex formation, relative to standard pairing, and those located at the center of a duplex are more destabilizing to duplex formation than those located at the end of a duplex. Duplex stability increases with increasing salt concentration up to ~1M (4,5) and decreases with extreme values in pH (~< 5 and ~> 9) (6).

Nearest-neighbor interactions serve as the foundation of thermodynamic models of DNA secondary structure prediction in solution. To simulate the secondary structures of a given template, these models use parameters such as internal and terminal DNA stacking (7), hairpins with and without loops, the presence of mismatches (8), dangling ends (9) and mono and divalent cation concentrations along with temperature and solvent polarity (10).

Transcriptome comparisons and genome wide association assays depend on the accurate measurement of millions of polymorphic sites across a genome. They are performed on microarrays and high-throughput short-read sequencers and by nature the samples start as extremely complex solutions. The complexity arises not only from sequence variation but also from how that variation affects structures and, in turn, on how

structures alter measurements of the sequences. Despite efforts to standardize conditions and calibrate the responses of these platforms, the raw data remain highly variable and success has been quite low in finding loci responsible for complex diseases and phenotypes (11,12). This is certainly due in part to the commonly small contribution of individual genes to complex phenotypes, particularly those that can be overwhelmed by environmental influences. In addition, the prevailing 'common allele, common phenotype' model is now widely seen as mistaken (13), and in its place a model in which rare alleles converge on a common phenotype has been embraced (14). In either case, phenotype is now interpreted as the outcome arising from disrupting a gene network, whose component gene functions and interactions are all candidates for causality. Creating an accurate network model requires that we have accurate measurements of each component gene and therefore that genomics and transcriptomics platforms deliver such measurements. It also requires that the models we use capture multi-dimensional interactions. That is, to predict the behavior of complex systems we need to a) study them globally and dynamically, b) measure them as quantitatively as possible and, c) integrate across different levels of information. These have been defined as the attributes of the Systems Biology paradigm, as expressed by Hornberg and colleagues (15) in the study of cancer. Our focus has been to bring nucleic acid structure as well as sequence into the modeling environment, and to consider its influence on the assays platforms as well as biology. Briefly, since the signal strength is used as a proxy for the concentration of target in microarray studies, if structure affects that estimate in unexpected ways the outcome of the gene level is likely to be incorrectly classified. Similarly, if structure alters the apparent base order in sequencing studies then the assigned genetic variance will be

incorrect, and correlations in the change of gene variance with phenotype will also be incorrect.

1.3: Data Generation Platforms Geared for Systems Biology Approaches

Unlike traditional biology, in which a small number of genes or gene products are studied at a time, systems biology focuses on complex interactions within biological systems and investigates the behavior and relationships across all of the elements (usually of one molecular type but increasingly across types as well) in that system (16,17). The goal of systems biology is to uncover the interactions of multiple components that lead to emergent properties characteristic of biological systems, develop predictive models and eventually formulate biological ‘laws’ that parallel those of physics.

Systems biology is a technology-driven discipline: the ‘-omics’ technologies, such as genomics, transcriptomics, proteomics, and metabolomics, are driving the acquisition of sufficient data to feed the models that describe how biological systems operate. These high throughput technologies not only report on each element but also allow profiling across many conditions and time intervals, and permit resolution to single-cell levels of discrimination (18). Results have included the identification of missing data in the form of new genes and gene functions (19,20), but more importantly have helped us to reconstruct gene networks, which are the means for characterizing the genotype to phenotype relationships (21), and improved our understanding of many genomic loci involved in the pathogenesis of human diseases (22).

In such bottom-up modeling, the quality of the data is of paramount concern: the accuracy, coverage, sensitivity and specificity of the measurements must be rigorously controlled since misleading and missing data could have a great impact on our

interpretations, particularly as we characterize the biological networks (23,24).

The following experiments are designed to investigate how structure in nucleic acids affects the interpretation of output from microarray and short-read sequence data, and the extent to which apparent short deletions in human sequence data might be related to specific types of structure. Two of the studies require bench work to construct and test hypotheses about the role of structure in signal while the third is a computational study correlating structure with the appearance of a short deleted region in the target.

1.4: Aims: Background and Significance

1.4.1: Aim 1: Microarrays

1.4.1.1: Background

The DNA microarray is the original example of the 'enabling' high throughput technologies; this family of platforms has been used to identify and quantify the mRNA transcripts present in samples, to perform re-sequencing, to identify single-nucleotide polymorphisms, copy number variations, and sequence variants (25-28). In the abstract, a microarray consists of a solid surface on which strands of short polynucleotides, called probes, have been anchored. The local region in which all of the strands are identical is called a spot. There can be millions of 'spots' on the array surface, each querying a distinct genomic target sequence. The assay is indirect: the sequence of the deposited probe is associated with the location of the spot, and the identity of complementary target is inferred based on complementarity to the probe. Sample preparation includes purification of the intended nucleic acid, possible conversion to a stable form, amplification, fragmentation and labeling. A solution of labeled targets is deposited on the array surface and incubated for some time, allowing targets to hybridize to

sufficiently complementary probes. Subsequent to hybridization the array is washed to eliminate nonbinding and unstable duplexes. Although detection methods vary, the most common chemistry is to use a fluorescent dye attached to the target along with a laser and detector tuned to that dye to produce photons. An image of the array is captured in which photons emitted lead to an 'exposure' level in a spot, it is assumed that this level correlates with the number of target molecules bound to probes in the region, and that it correlates in much the same way for all such pairs. That is, the spot intensity is transformed into a target concentration that is subsequently used for statistical and data mining analyses (29,30).

1.4.1.2: Microarray Interpretation Issues

Although this technology has had a great impact on biological and biomedical research, with myriad published achievements in gene expression analysis (12,31-34), genome-association (35,36) , genetic linkage (37,38), and network inference studies (39), it has also been shown that results derived from similar studies can be highly inconsistent (40-42). Although the issues are not unique to microarrays, the high-throughput nature and involved technical steps of the assays throw into strong relief the four sources of experimental variance: a) sample characteristics from inherent biological properties, b) experimental design weaknesses of high-throughput platforms, c) technical issues due to assay complexity, d) physical characteristics due to innate probe and target differences.

a) Biological variance: Biological differences are the result of real variances between samples. Individual cells may simply respond differently to different levels to the same input, or there may be single-nucleotide polymorphisms (SNPs), copy number variations (CNVs) (43) or different splice forms present in transcripts (44),

that lead to differences.

b) Experimental variance: High-throughput assays have the inherent flaw that there are far more measurements than samples. While there are some designs such as a common reference pool that can mitigate the problem they are not always used (45-47). Unfortunately calibration standards, while provided by some suppliers and embraced by qPCR users, were never widely used by the microarray research community (48).

c) Technical variance: A large number of artifacts arise from sample handling and array manufacture processes. Numerous investigations have been conducted to evaluate the influence of these factors, including batch effects (49), dye effects (50), post hybridization wash effects (51), platform-specific effects (52-55), and how statistical approaches weight assumptions inherent in experimental designs (56,57).

d) Physical variance: Probes and targets are physical molecules with structural properties that are affected by the assay environment - their thermodynamic and biochemical characteristics must be considered. A well-known example of such properties is the secondary structures which can exist in the probe (23,58,59). Much less consideration has been given to the structural properties of the targets (60)

1.4.1.3: Current Status and Outstanding Questions

Although microarray technology has been widely used the interpretation of signal intensities is not an easy task. While some sources of variance result in noise, showing the characteristic random normal distribution, many of the factors listed above introduce

a specific bias that must be handled individually. Most of the current studies that consider structure explore the effect of experimental properties on probes, including melting temperature (T_m), free energy (ΔG)(61), probe secondary structure (49), and probe length on probe-target hybridization (41,42,43). The few studies which address the effect of target secondary structures on hybridization signal intensities all assume that such structures always destabilize probe-target hybridization (49, 50). Our own results from molecular simulations and experimental data indicate that if the target has secondary structures around the binding region in flanking sequences, these structures may stabilize the probe-target hybridization instead. So in the first project, we tested the following hypothesis:

1) Stable secondary structures on the boundary of, but not impinging upon, the probe-target binding site, causes no change in the signal detected for a probe-target interaction on a microarray.

1.4.1.5: Significance

From the intensity of the spot on a microarray the signal is converted to a concentration equivalent. Some studies use ratios to produce a purely relative value, but this precludes the use of meta-experiments, the combining of experiments from multiple labs that has been touted as an added value for the rather high cost of producing microarrays (62). Any uncorrected factor that alters the apparent concentration of a particular target but not others will bias the results of the experiment: since similar values are often binned together in data mining methods this can affect the interpretation of many genes and pathways.

Microarrays still continue to be used in large numbers (63,64), especially in

studies of human health given the current strong emphasis on translational medicine because it has a proven track record spanning more than two decades in the lab, its limitations and possible pitfalls are quite well known, and there is general consensus on the methods for analyzing the results . This means that methods to better understand and correct for bias on microarrays continue to be an important focus of research.

1.4.2: Aim 2: Sequencing

1.4.2.1: Background

The advent of automated sequencers in the 1990's, based on the Sanger sequencing concept but using specialized chemistry and robotics, enabled routine and large-scale sequencing. The volume of data and the strategies required to optimize sample and data handling drove some of the first serious bioinformatics developments. However the costs were prohibitive except for large teams and consortia. The challenge to drive costs down to $1000 for a complete human genome was accepted by a number of companies, and, although not quite realized, we are approaching the point at which routine sequencing is affordable for biologists running single labs (65,66). Current technologies all use some variant of sequencing-by-synthesis, detecting the incorporation of each nucleotide by some change in chemistry (65,66). To achieve high throughput the purified nucleic acid is transformed if necessary (to cDNA if RNA is the original substance), fragmented into small pieces that are then modified to allow amplification and priming of the sequencing reaction, attached to the substrate used by the platform, and then sequenced in parallel while signal is collected (67-69). Once the signal has been collected it is processed, such that the base present at each position can be inferred, along with an associated quality value (70). Data analysis methods center on assembling these

short reads in the correct order and then identifying frequencies of occurrence of subsets of the data, followed by identifying unique features of the sequence (65,71).

1.4.2.2: Sequence Interpretation Issues

Although the processing and detection methods differ, the same factors that affect interpretation of microarray data must be taken into account when analyzing sequence data. Selection of the sample preparation technique greatly influences the success of subsequent data analysis methods. Accurate interpretation requires good experimental design, in this case the proper marriage of preparation technique and platform.

a) Biological variation: The number of ways in which samples can be prepared has proliferated, allowing discrimination of allelic differences, modified bases, splice variation, small and non-coding RNAs and others (72,73).

b) Experimental variation: The primary factor considered in this category is the depth of sequencing achievable by a given platform and chemistry (74). Another factor contributing to the experimental design is whether it will be necessary to use multiple platforms in order to bridge regions of sequence that one platform cannot handle with another, the most common example being the use of the GS FLX™ technology to generate reads that span repeat regions of a genome that the standard Illumina and Ion Torrent PGM™ platforms cannot bridge (75).

c) Technical variation: Library preparation introduces a wide range of bias, not all of which will be discussed here. One example is the method for processing bulk samples which requires first fragmenting the material to a uniform size. All such methods have a certain amount of sequence bias (76); the subsequent addition of adaptors that create amplification and sequencing-ready templates are also

inefficient and subject to bias (77). Multiplex PCR amplification has well-known problems (78). Since the commercial sequencing platforms do not release all of the details of their sequencing chemistries, it is difficult to state what buffer and enzyme-related factors are present, but these have certainly been characterized in related assays, in particular Sanger sequencing (79) with electrophoresis separation and fluorescent product detection (80). Similar to microarray platforms, no calibration standards exist to allow independent and objective reporting of instrument behavior independent from the production of internal sequences used to calibrate signal processing software. It has been a source of frustration to the sequence analysis community that the 'quality scores' produced by vendor software are not standardized to some external, verifiable metric (81). For those platforms that produce image files at each cycle, studies indicate that some part of the image creation or data-extraction process introduces variation that affects the overall read's sensitivity and accuracy (70,82-84).

d) Physical variation: As mentioned above, secondary structure is an integral characteristic of a nucleic acid. The nature and stability of such structures is highly dependent on the environment. The equilibrium between the hairpin and random coil conformation of a nucleic acid molecule not only depends on the composition and the number of residues participating in the stem and loop, but also depends on the ionic strength and the temperature of the solution (85). While microarray assay conditions were designed to minimize such structure, reactions involving enzymes have much less leeway, as PCR assay designers know too well. Some of this structure is biologically important in the context of an intact cell such as gene

expression regulation through protein binding to structures in untranslated regions (UTRs) (86), and some arise only in the context of the laboratory preparation steps. It has been noted that different high-speed sequencing platforms have different characteristic errors, some of which have been correlated with high GC-content or stable hairpin structures as has been shown by Dr Lin Liu: in Illumina HiSeq 2000 the average sequencing depth dropped ~1X when GC content increased from 60% to 70% (87). No systematic study of structure effects on sequencing fidelity has been carried out, probably in part because of the proprietary nature of the reagents.

1.4.2.3: Current Status and Outstanding Questions

Similar to microarrays, NGS technologies are considered transformative for today's biomedical research, but several studies have revealed problems with data reliability and reproducibility among NGS platforms. For example, Dohm and coworkers found that, in the reads generated by a Solexa platform, A to C base substitution errors were 10 times more frequent than the C to G substitutions (82). Similar artifacts were observed by Bravo and Irizarry who reported that, in the reads generated by the Illumina ChIP-seq experiment, A to T miscalls were the most common error (83). Finally, Oshlack and Wakefield used the Aggregated Tag Counts technique to identify differentially expressed genes in datasets generated by a number of different platforms and found that the ability to correctly call differential expression is strongly associated with the length of the transcript (84) and not simply the number of tags in a specific region. There is little published work exploring what template-related factors affect read accuracy; the current push is to increase read length for sequences accessible to the methodology. Since some of the structure-related issues were addressed for earlier generations of sequencers it may

be possible to adapt those methods to the new platforms, thereby recovering usable sequence. So in this project, we tested the following hypothesis:

2) Stable secondary affects the fidelity of read-through on an available short-read high-throughput platform, the Ion Torrent Personal Genome Machine (PGM)

1.4.2.4: Significance

Developers of biomedical applications are embracing high-speed sequencing platforms at an unprecedented rate, with consequences that can be immediate (determining what drug to prescribe) and long-term (development of new druggable targets) (88). Knowing what features lead to particular types of errors will help both those choosing the method for generating data and analysts developing methods for best analyzing the data to partition their selections correctly.

1.4.3: Aim 3: Computational Study of Deleted Human Sequences

1.4.3.1: Background

There are publicly available datasets from each of the major NGS platforms on reference genomes, particularly the HapMap samples originally shared across international institutions to produce human variation estimates (89). The outcomes of these profiling experiments are described in survey articles describing differences such as where errors accumulate and what types of errors are most commonly seen. An error that caught our attention was the reported prevalence of short deletions in the human genome (90,91). Ahn et al. 2009 examined 342,965 indels (<= 20bp) which they reported in the Korean individual genome (SJK) against dbSNP and they found that only 247 indels (0.1%) were validated and 113,287 (33.0%) non-validated and the remaining 229,431 (66.9%) indels were not found in dbSNP. They also compared SJK indels (< 4bp) with

those of Han Chinese (YH), HuRef (Venter), Watson, and Yoruba and reported that between SJK and YH genomes only 7.8% of the indels had the same genomic positions, size and type, between SJK and Venter genomes only 10.2% of the indels had the same genomic positions, size and type, between SJK and Watson genomes only 2% of the indels had the same genomic positions, size and type, and between SJK and Yoruba genomes only 49.4% of the indels had the same genomic positions, size and type.

Since preliminary data in our lab from the sequencing of constructs with strong hairpins resulted in apparent short deletions (unpublished data), this seemed a promising direction to pursue: did some fraction of the apparent deletions lie in highly structured regions that might have lead to sequencing errors. By comparing randomly selected sequences that match the reference genome as a training set and using regions apparently subject to deletions relative to the reference genome as our test set, the goal is to identify sequence/structural features that distinguish the sets. Because chemistries differ, the sensitivity of the different platforms to structure may well vary. Identifying signatures difficult for particular platforms to accurately produce will allow researchers to correctly pair the method and the target. Although not covered in this dissertation research, the long-term goal of the lab is to identify conditions on the Ion Torrent PGM^{TM} sequencer that allow accurate sequencing through highly structured templates.

1.4.3.2: Interpretation Issues

It is well recognized the sequencing errors create a barrier to correct correlation of genotype and phenotype in association studies. The assumption is that these errors result from mis-incorporation of nucleotides presumably arising from either slippage of short repeat regions or inability of the platform to maintain a signal difference in

homopolymeric regions (92,93). While slippage in repeat regions could create the appearance of a short deletion many of the regions containing ostensible deletions do not contain simple sequence repeats or homopolymeric regions. Kim et al. 2009 examined the genome of a Korean individual known as AK1 and reported 170,202 indels, from which just 60 indels were confirmed using the Sanger sequencing assay. The presence of such a very large number of not validated indels may cause one to consider whether all of these reported indels are truly present or whether some of them resulted from the assays' conditions.

1.4.3.3: Current Status and Outstanding Questions

Many large sequencing projects have been carried out on human samples using the various high-throughput short-read platforms. Unfortunately most of the data is not available even in the Short-Read Archive, so one must rely on summary statistics and previous analyses. We successfully identified one project that made the raw data available and used it as the basis for a structural modeling assessment and then we used the random forest algorithm implemented in the machine-learning environment (WEKA) to identify relevant features. In this project, we tested the following hypothesis:

3) The sequence context of short deletions has no structural context that discriminates them from similar sequences that are successfully sequenced.

1.4.3.4: Significance

If it is true that structure plays a significant role in the accuracy with which a particular platform reads out a target, then we want to predict those regions of the human genome with characteristics making them prone to experimental errors. Even where deletions are of biological rather than technical origin a structural context may correlate

to an important regulatory phenotype.

In summary, the effect of structure within the probe-binding interface of heteroduplex formation is accepted, but the effect of adjacent structures has not been reported. A significant change in binding stability would alter the interpretation of many microarray experimental results. Similarly, the effect of structure within the sequencing template of HTS platforms could lead to a number of types of read errors, and if long deletions are one such error the outcome is likely misinterpretation of genome or transcript structure. Finally, in a HTS experiment that reports on a very high frequency of deletion changes in a genome, we investigated whether a structural component might predict the appearance of the deletion.

CHAPTER 2: THE EFFECT OF TARGET STRUCTURE ON MICROARRAY HYBRIDIZATION

2.1: Overview

Studies that investigate the effects of secondary structure(s) on the rate and efficiency of the probe-target duplex formation on microarray platforms can be divided into two groups. One group focuses (94-98) on how the formation of secondary structure leads to a reduction of hybridization sensitivity and specificity. For example Mehlmann and Liu have shown that for perfectly complementary probe-target sequences, the presence of stable monomer structures at hybridization equilibrium significantly decreases the rate and efficiency of duplex formation. This is expected since it decreases the concentration of one of the reactants. The effect is a signal that is too low, a false negative in analysis terms. The other group of studies (99-101) has shown that the formation of secondary structure sometimes leads to unexpectedly high hybridization signals, such as that published by Trapp (2011) in which non-complementary target-probe sequences formed stable heterodimers with an internal bulged loop. A special class of structures called G-quadruplexes are also known to create duplex signal higher than the concentration of reactant would predict (101). Thus although the effect of structure can vary, it is widely acknowledged that the presence of structure in either probe or target can lead to signals that do not accurately reflect concentration, and structure must be considered in order to accurately analyze and interpret microarray data.

In most microarray experiments the question asked is how well the probe

hybridization discriminates between a perfect match and mismatches of varying degree (102,103). Sequence extending beyond the duplex region is considered irrelevant, except so far as it affects diffusion rates (104) or competes for the probe binding region, as indicated above. Indeed, solution thermodynamic theory states that only the N+ 1 base will affect the hybrid formation barring the existence of a competing structure (9).

Testing of structured templates is complicated by preparation challenges. A common source of known and highly structured sequences is the ribosomal RNA gene family, which has extensive experimental evidence from cross-linking and other types of assays that report on the major folded forms. Amplified fragments of 16S rDNA have been used to test probe responses on microarrays, results consistently show less signal than the added concentration would have predicted (94). Reducing targets to a size that eliminates the possibility that internal binding can be stable under hybridization conditions has been recommended (105), but under random shearing protocols this is also likely to disrupt the probe-target binding site at a fairly high frequency, which will also cause a decrease in signal compared to the input concentration. Very long targets diffuse slowly in hybridization solutions, and it has been shown that the rate of reaching equilibrium is considerably slower than many hybridization protocols permit (96), although those experiments did not consider secondary structure as a factor. None of these studies considered the effect of hairpins in the target adjacent to the heteroduplex region on binding stability. The competing models for outcomes when such structure is present include: the folded structure creates steric hindrance to a probe-target interaction leading to a diminished signal; the overall thermodynamic effect of total entropy from the exclusion of solvent will lead to a more stable complex and possibly an enhanced signal

relative to length matched probe-target pairs.

2.2: Methods - Computational

2.2.1: Target Construction to Test Computational Predictions

To investigate the effects of boundary sequences on the stability of probe-target duplexes, we selected two 33mer probes (SNP_A-8475541, SNP_A-8477444) from the Affymetrix SNP6.0 Array which are annotated to chromosome Y (human genome reference build version 36.3). They are among the probes having the highest fraction of partial alignment with sequences along chromosome Y, which means that stabilized partial hybrids could have a significant effect on interpreting the data.

Since full-length complements bind 100%, we could not use them to investigate the significance of stabilizing boundary structures, therefore we identified 603554 and 624697 partial alignments for SNP_A-8475541, and SNP_A-8477444 probes along chromosome Y using the SeqNFind™ platform with the following input parameters open gap=-3, extending gap=3 and word size of 6, with the goal of identifying those with significant but not complete binding so that differences could be observed.

To construct extended targets we used the complements of the partial alignments obtained from the alignment tool as probe-target binding cores and designed a nested set of sequences around them, such that increasing length gives rise to structure on either side. The probe-target binding may be longer or shorter than 33nt in length: a longer partial match simply extends over more bases, a shorter uses only a subset of the total primer length. Each set includes 10 nested targets. The smallest target in each set complements the core probe binding sequence and the remaining members of the set are longer by 1, 5, 10, 15, 20, 25, 30, 35, and 45 nucleotides to both sides of the core,

designated by the core label '+N' , as sown in Figure 1. In the following pages we refer to these as target-sets.

2.2.2: Molecular Simulations

We used the Oligonucleotide Modeling Platform (OMP DE™) (106), with parameters matching Affymetrix SNP6.0 array hybridization conditions (see Table 1), to model all of the optimal and suboptimal heteroduplex structures (targets and selected probes).

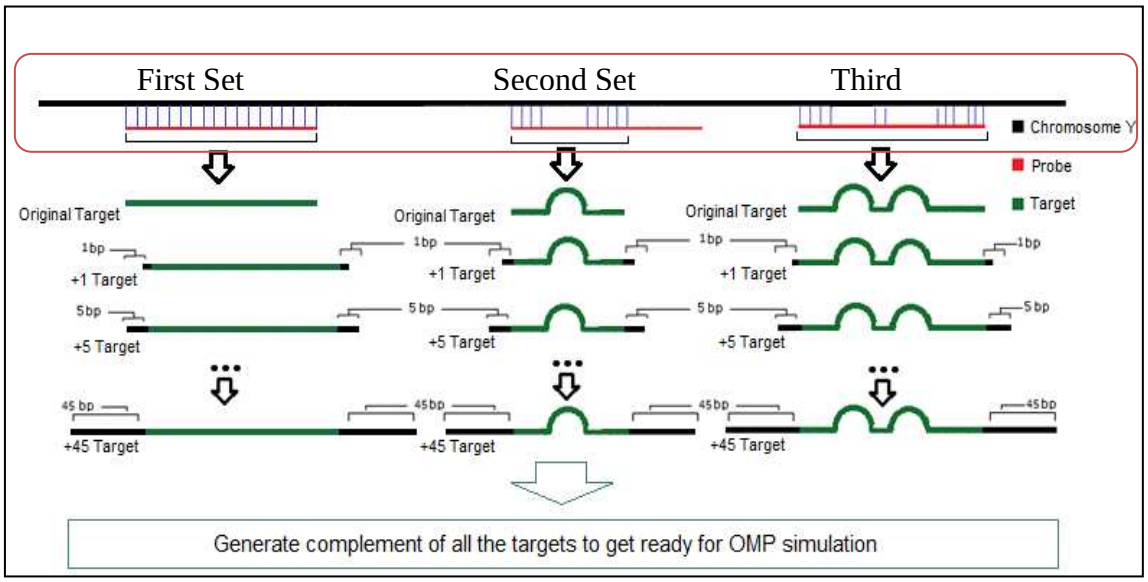


FIGURE 1: Schematic of the target-set design process. Highlighted in the red box are examples of 3 types of alignments of one probe to 3 sites on chromosome Y. The farthest left shows complete and perfect complementarity, the second shows an internal gap in complementarity and the third is an example where there are several internal gaps. Because gap lengths and the extent of complementarity vary, target length does not correlate directly with probe length.

2.2.3: ΔG Cutoff Calculation

A novel method was used to estimate the boundary condition for stable binding of the $\Delta G_{heterodimer}$, explained below (R code indicated in supplements corresponds to Figure

4).

To calculate ΔG cutoff for each probe, we used all optimal $\Delta G_{heterodimer}$ obtained from OMP (603554 and 624697 optimal $\Delta G_{heterodimer}$ for SNP_A-8475541and SNP_A-8477444 probes) and calculated The Probability Density Function (PDF) of $\max(\Delta G_{heterodimer}) - \Delta G_{heterodimer}$, where $\max(\Delta G_{heterodimer})$ reflects the value reported for the less stable conformation, and $\Delta G_{heterodimer}$ reflects any other conformation returned by the modeling software. The Chi-square statistical test was used to identify the critical value of this distribution for an $\alpha = 0.05$ and degree of freedom (df) = 1. Then, we found the maximum $\Delta G_{heterodimer}$ from the probability density function which had for its corresponding value on the x-axes a value equal to or greater than this critical value. We then considered all any duplex structures with $\Delta G_{heterodimer}$ less than this critical value to be stable, meaning that is it predicted to return a measurement higher than baseline on our microarray platform, and hence potentially useful for our study.

2.2.4: Target-Set Selection Criteria

Several measures are used to predict probe-target binding, including $\Delta G_{heterodimer}$, the total number of H-bonds, a minimum nucleation length and the OMP-calculated percent bound (PB). All of these values were calculated for each member of each target set, as described below.

From the work of others we know that continuously complementary heterodimer structures having $\Delta G \leq -10$ kcal/mol persist through the wash steps under commonly used conditions (107), although a variety of factors can modulate this cut-off, as discussed by Xia et al (108) Targets useful for comparison then require changes in ΔG resulting in structures at least that stable, so this represents one selection criterion. That is, we

retained in our target sets only members with predicted increased stability beyond that threshold as the length increases.

The Tm and percent bound (PB) value reported by the DNASoft OMP application have been reported in some of the literature (109,110) to be a reliable indicator for the amount of duplex formed. Given the sensitivity of the microarray scanning platform, a 10 % change in percent bound is readily measured, so we required that difference when selecting targets to compare. That is, we retained members in target-sets that were predicted to have $\Delta PB \geq 10\%$ when the length changed, excluding the bottom 10% and top 90% signal saturation.

2.2.5: Examine the Effects of the Target Length and Secondary Structures on Probe-Target Hybridization

To investigate whether structures that surround (and don't occlude) the probe-target binding site may stabilize the heterodimers, we investigated the result of following three experiments:

1) We gradually increased the target length (symmetrically centered on the probe binding site) from 1 to 45nt and counted all the heterodimer structures which satisfied our target-set selection criteria (a) $\Delta G_{heterodimer} \leq$ -10kcal/mol and b) the predicted target-percent bound increased at least 10%) and then plotted the result. Note: For each duplex, we obtained 1 optimal and 9 sub-optimal structures therefore the $\Delta G_{heterodimer}$ used in this part of analysis, was the weighted-average of the optimal and suboptimal $\Delta G_{heterodimers}$, and the target-percent bound was the summation of optimal and suboptimal heterodimers' target-percent bounds.
2) Since increasing the target length may generate a more stable probe-target binding

site, for the second experiment, we gradually increased the target from 1 to 45nt and counted all the heterodimer structures which not only satisfied a) $\Delta G_{heterodimer} \leq$ -10kcal/mol and b) the predicted target-percent bound increased at least 10%, but also c) keep the same base complementarily between the two strands and then plotted the result.

3) After applying base complementarily filter, the number of heterodimer structures reached a maximum at extensions of 15 and 20nt for probes 858_T and 850_T and then began to decline. To show that even though, the number of heterodimer structures decreased, their stability continued to increase , we compared the ΔG distribution of +45 with +15 targets for probe 858_T and +45 with +20 targets for probe 850_T. To do this comparison we subtracted the number of heterodimer structures of length +45 from those at length +15 and +20 for probe 858_T and850_T consecutively and then we plotted the ΔG distributions for all heterodimers containing from 4 to 12 complementary bases.

2.3: Methods - Experimental

2.3.1: Target and Probe Design

In this part of our study the goal was to experimentally validate the results obtained from the computational modeling described above, which indicated that the presence of a boundary structure stabilizes rather than destabilizes the probe-targets interactions.

From the set of possible target sets we selected 3 pairs for experimental testing. Each pair includes one target that is the same length or slightly longer than the 35nt probe (40-50nt) and one that is considerably longer (130-150nt) and includes hairpin structures

in the regions adjacent to the probe binding site. Criteria are described in more detail below.

Common criteria, applied to all 3 pairs include:

1) All probe-target binding site complementarity is imperfect (non-continuous) so all binding will fall below 100%, allowing competitive differences to be observed.
2) Factors contributing to duplex stability include the total number of H-bonds, a 'minimum nucleation length' of consecutive H-bonds, the ΔG of the duplex and the percent bound (PB). Figure 2 shows the pairs, which include:
    a) Target set 1571-150 and 1571-50, which focused on the number of H-bonds and the presence of a 'minimum nucleation length'.
    b) Target set 857-150 and 857-50, which focused on the total ΔG of the duplex.
    c) Target set 643-130 and 643-40, which focused on the duplex ΔG and the percent bound.
3) A design constraint was that the heterodimer portion of each structure (the probe-target interface that forms a duplex) was predicted to be more stable than any adjacent structure in the target or any alternative folded monomeric structure of the probe or target, or possible homodimers.
4) Note on experimental methods: because it has been proposed that aqueous hybridization wash conditions remove properly bound material we used the isopropanol conditions described by Pozhitkov and Noble (2006), although their more recent publications indicate that this extra care may not have been required (51).

Specific criteria used in selecting the second and third target pairs include:

1) Under the hybridization conditions shown in Table 1, each member of the pair has same nucleotides complementary between the probe and two targets (see Figure 2).

TABLE 1: Hybridization conditions used in the OMP simulation

Assay Temp	45C
Monovalent	0.056M
DMSO	0.96%
TMAC	3.68M
PH	6.6

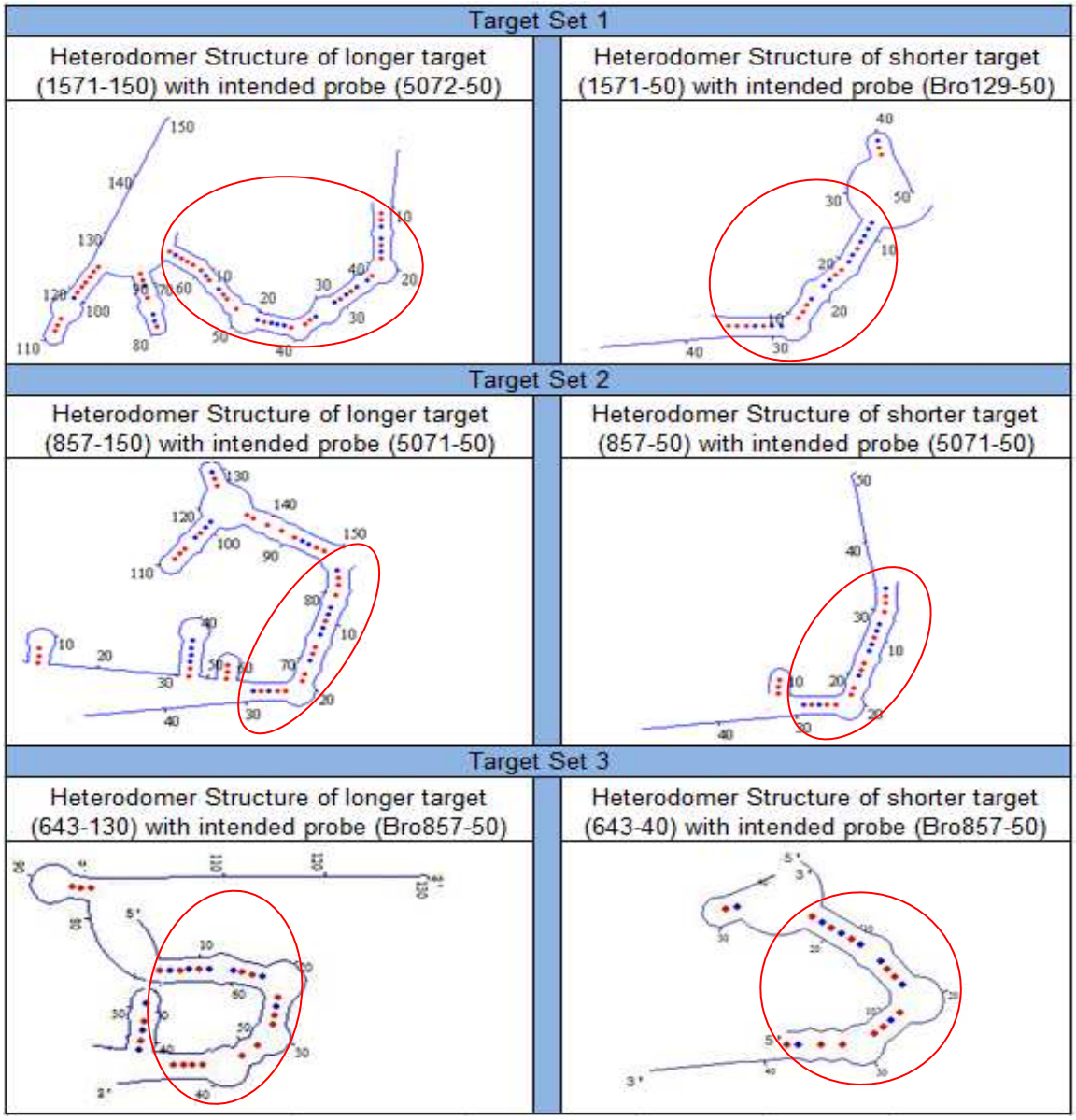


FIGURE 2: Optimal duplex structures of the three target-pairs which were selected for testing. As shown, each pair consists of a longer (~140nt) and shorter (~45nt) target. In the figures the red oval indicates the hybridization site. The pattern of complementary bases in the duplex is the same for both members of pairs 2 and 3.

2.3.2: Target Construction

Targets were assembled using overlapping oligonucleotides (111,112) which were designed to span the entire length of each target. The 3' overlaps were 15-35 nucleotides in length (Figure3). Target assembly and amplification was performed in three steps:

annealing, extension, and full-length PCR. Annealing was carried out in a volume of 30 µl, with 0.2 µM of each oligonucleotide in a buffer containing 1.5 mM $MgCl_2$ and 1X HF buffer (Phusion high-fidelity buffer from Promega). After mixing, the solution was heated to 95°C for 5 minutes, followed by gradual cooling to 37°C (60 minutes in a 100ml beaker of water heated to 95°C). To each reaction was added 200 µM (final) of each dNTP and 0.4U of Phusion polymerase, followed by incubation for 60 minutes at 37 ºC. The full-length construct was then amplified from the mixture of products using primers to the ends alone. These primers were modified such that the final targets had a Cy3 label on the 5' side of the strands that hybridize to the probes and a biotin on 5' side of the complementary strand. Biotin-streptavidin binding of the complementary strand was performed to remove the complementary strand, to prevent competitive binding of this strand to target when hybridized to the microarray. This PCR reaction was carried out in a 100 µl reaction containing 10 µl of re-amplified and gel-purified full-length target, 200 µM of each dNTP, 0.4 U of Phusion polymerase, 0.2 µM of terminal primers, 1.5 mM $MgCl_2$ and 1X HF buffer. PCR cycling was: 95º C for 3 min followed by 30 cycles at 95º C for 30 s, 58 ºC for 30 s and 72 ºC for 30 s, and terminated by 3 min extension at 72ºC. Correct modification was verified by analyzing 5ng of each target on 8% polyacrylamide gels (Figure 4).

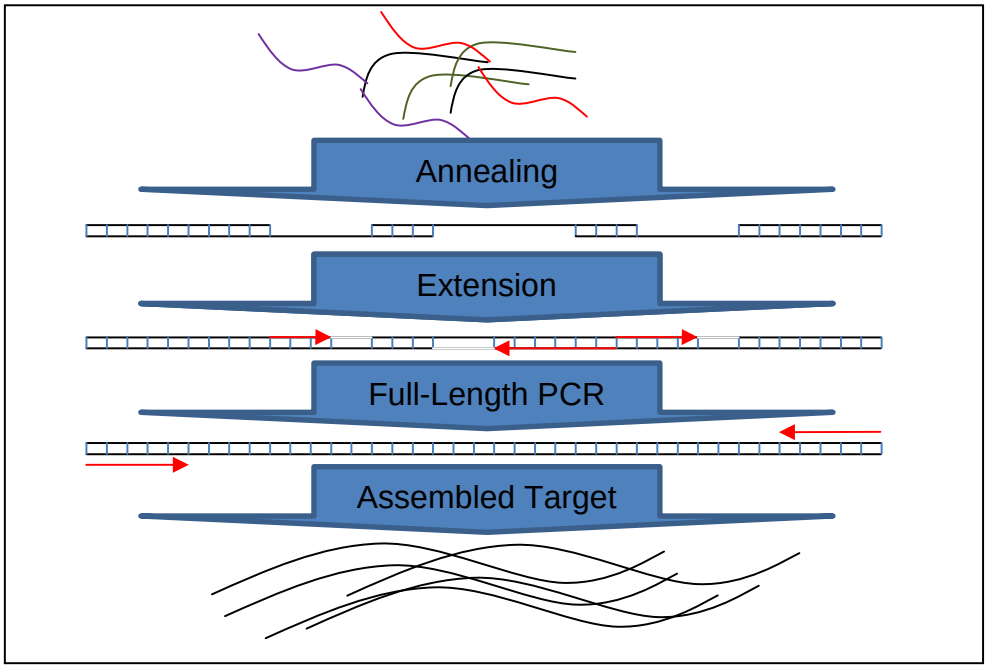


FIGURE 3: Schematic representation of steps in the template assembly process.

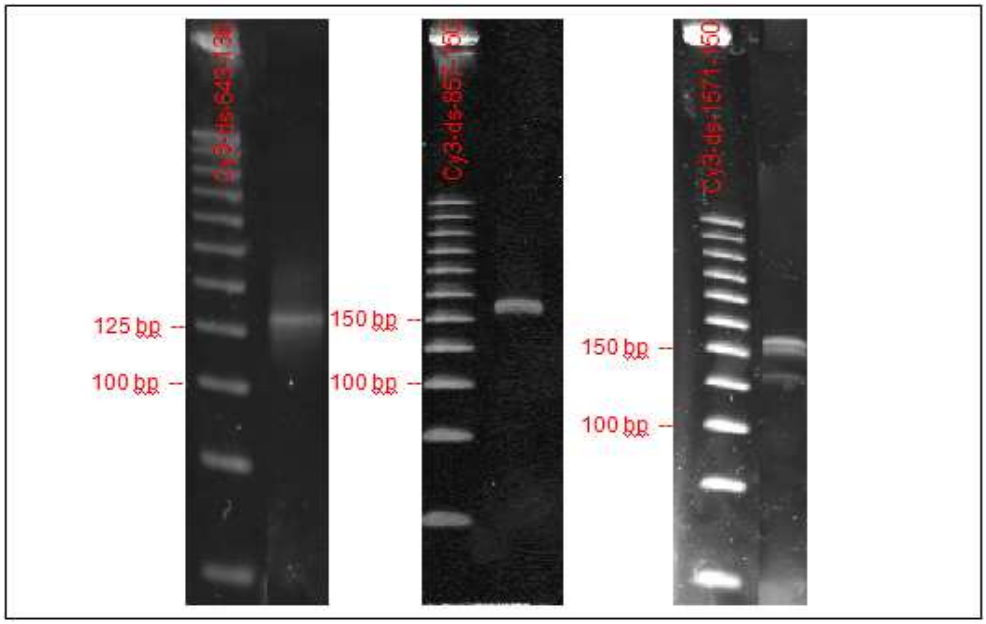


FIGURE 4: Gel picture of three cy3 labeled double stranded targets. This gel stained with syber-gold and NEB 25bp step ladder was used as the size standard.

2.3.3: Purification of Single-Stranded Targets

Labeled double-stranded targets were ethanol precipitated. After resuspension the desired 5'-Cy3 probe-complementary strands were isolated using Dynabeads® M-270 Streptavidin (from Invitrogen) to remove the biotin-labeled strand. Cy3-labeled single-stranded targets were assessed for length and purity by analyzing them on 8% polyacrylamide gels and visualizing them using the Tecan ReLoaded scanner (Figure 5), following the manufacturers gel visualization protocol.

As highlighted by blue oval in Figure 5, small portion of double stranded targets remained in the final isolated single stranded solutions which must be considered in assessing the final concentration of our single stranded targets.

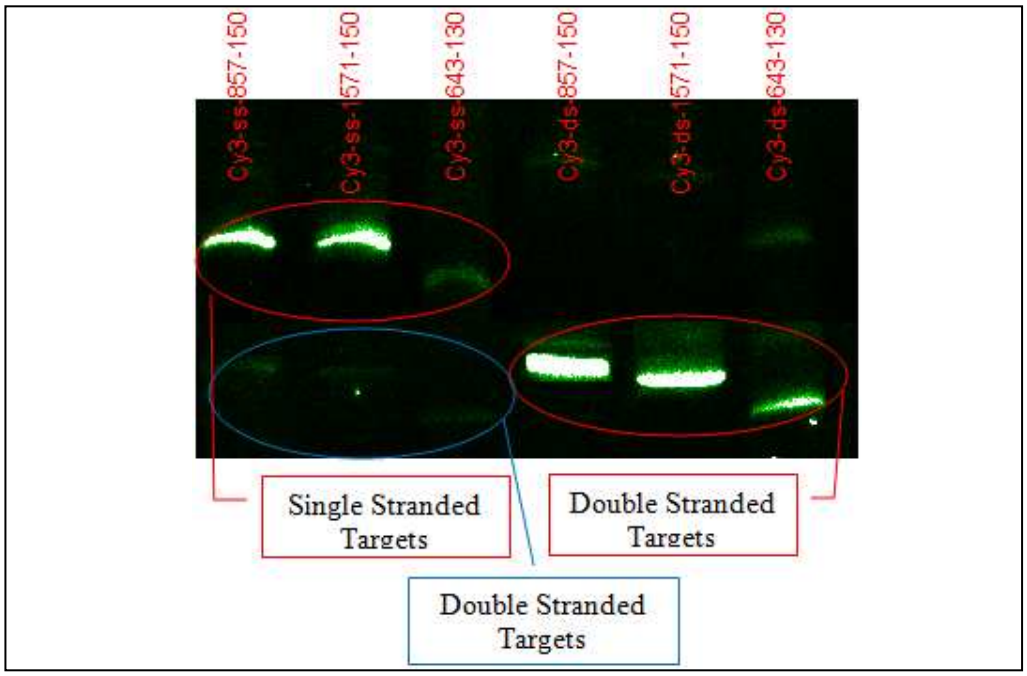


FIGURE 5: Gel image of single and double stranded target visualized using Tecan ReLoaded scanner. Blue oval highlighted double stranded target which remained in the isolated single stranded solutions.

2.3.4: Single-Stranded Targets: Concentration Calculation

To measure the concentration of our Cy3 –labeled single stranded targets, we first built a standard curve following these steps:

a) A Cy3 labeled oligonucleotide (100uM) was 5-fold serially diluted to create a calibration set (Table 2).

b) Each dilution was measured with a NanoDrop ND-3000 spectrophotometer to acquire RFU values, with three-fold replication.

c) The standard curve was created by plotting the known concentrations on the x-axis and measured RFUs associated to each concentration on the y-axis.

The RFU values of the targets were measured using the NanoDrop ND-3000 spectrophotometer. We note that there is likely still a small amount of double-stranded target (visible on the acrylamide gels – see Figure 5) remained in the isolated single stranded targets, because the Dynabead purification step is not completely efficient.

To determine the fraction of each RFU value that belonged to the single stranded targets we ran each target solution on an 8% polyacrylamide gel. Lanes were cut out and imaged in the Tecan Reloaded scanner for the Cy3 signal. Band intensities were measured for the single and double stranded targets, from which we calculated the portion of RFU values which belonged to each, and then the fraction of signal belonging to the single-stranded target available to bind to the probes on the microarray Finally, we used the RFU values of the single stranded targets to interpolate our targets' concentrations using the standard curve. The example below illustrates this process in detail.

To calculate the concentration of single stranded target (Cy3-1571-150), we measured: a) The total RFU value, which was 4452 f.u., b) The proportion of intensities of the single to double stranded bands was 4.5 (Figure 6) therefore, by solving 4.5X+X = 4452 equation, we found that the RFU values associated to the single stranded target was 3642.5. Using the standard curve (built using data in Table 2), we found the concentration of single stranded target was ~ 430 nM (Figure 6).

To make the hybridization buffer (60 µl), 14.6 µlof target was used. That is, target was diluted 60 / 14.6 = 4.1 folds, therefore for the above target the final concentration was 430 nM / 4.1 = ~ 104.6 nM

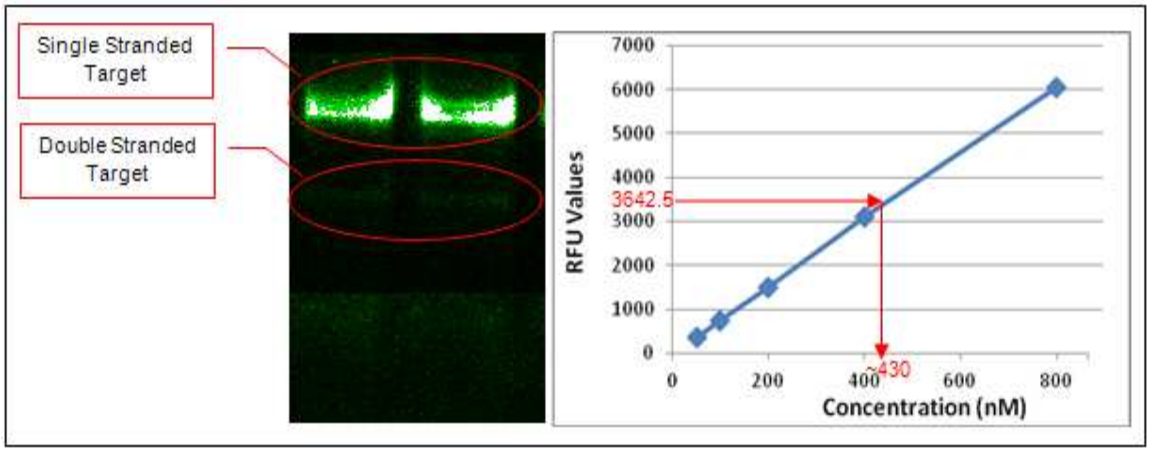


FIGURE 6: This figure illustrates the process of calculating the target concentration.

TABLE 2: Concentration series and associated RFU values used to build standard curve.

Concentration (nM)	RFU Values
50	370
100	730
200	1500
400	3100
800	6050

2.3.5: Array Design Specifications

The microarray slides were printed in-house using 110 µm quill pins on the BioRad Calligrapher according to the supplier's instructions. Probe concentration was 5 µM, slides were SuperChip Epoxy Slides (Erie Scientific through VWR).

As Figure 7 indicates, the array contains 4 rows and 4 columns. The first row contains 4 spots of buffer, the second row contains 4 spots of 5 µM Intended probes (against which targets were designed). The third row contains 4 spots of 5 µM unlabeled probes which were used as negative control to make sure our targets did not hybridize to the sentinel probes, and the fourth row contains 5 µM 'sentinel' probes, which contain a Cy3 label and were used to identify the position of the spots on the slide and to verify that the attachment chemistry was successful. Each slide contains two such arrays.

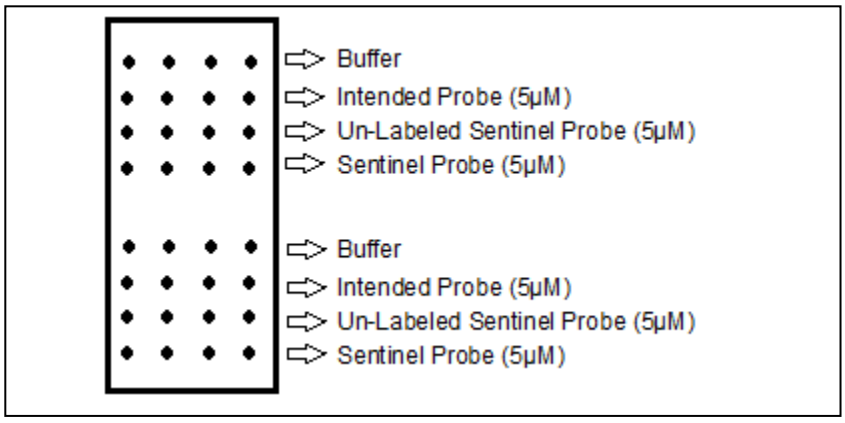


FIGURE 7: Slide layout.

2.3.6: Array Hybridization

Slides were placed in an HS 4800 Pro Hybridization Station (Tecan, Mannedorf, Switzerland), then they were blocked with BlockIt solution (ArrayIt, Sunnyvale, CA) for 30 minutes. Next, 60µL of hybridization solution containing 44.16 µL of 5M TMAC

(final concentration 3.68 M), 0.617 µL of 100% formamide (final percentage 0.96%), 0.672 µL of 5 M sodium chloride (final concentration 0.56M), and 14.6 µL target (50-100nM) was added to each array (two arrays per slide). Sides were incubated for 18 hours at 45 C. During this period they were subjected to mechanical agitation at medium intensity (1.1 minutes agitation with 3.5 minutes break). After hybridization, slides were washed with 99% isopropanol for 2 minutes (113) and then they were dried and scanned using Tecan ReLoaded scanner.

2.3.7: Image Acquisition and Data Analysis

Slides were scanned with the following parameter settings: 532nm laser, a 575nm filter, Hs Autofocus, small pinhole, 6µm resolution, and a 160 PMT gain in the LS Reloaded Scanner (Tecan, Mannedorf, Switzerland).

Images were saved in the Tagged Image File format (tif) and then analyzed using ImaGene software (Biodiscovery, Inc, Proteigene, Saint Marcel, France) with the parameters for segmentation option and set to seeded region growing. Each spot's intensity was transformed by subtracting the background intensities from the respective raw intensities, and then plotted. Figure 8 shows one example of images of the array associated to each target set before and after hybridization.

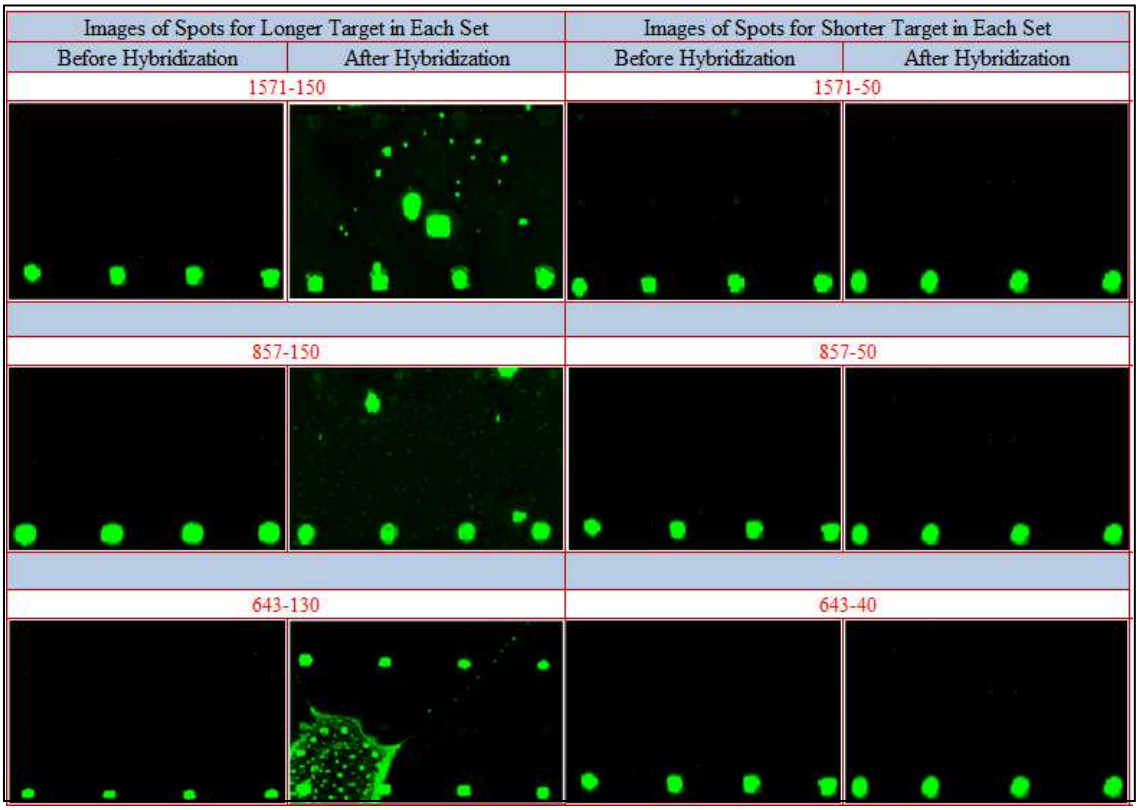


FIGURE 8: One example of scanned images before and after hybridization for longer and shorter target in each target set.

2.4: Results

2.4.1: Computational Predictions of the Constructed Targets

Table 3 indicates the number of locations which selected probes were aligned (from Affymetrix SNP 6) on chromosome Y using the Smith-Waterman algorithm as implemented on the SeqNFind™ platform. For each probe, as described in the Methods section, these aligned locations were used to generate a series of potential targets.

TABLE 3: Probes and number of aligned positions on the specified chromosome

Probe Name	Length	Chromosome	Number of aligned locations
SNP_A-8475541	33	Y	604487
SNP_A-8477444	33	Y	633188

Note: for simplicity of notation, throughout this study we labeled probe SNP_A-8475541 as T-850 and probe SNP_A-8477444 as T-858.

2.4.2: Results of ΔG Cutoff Calculation

To estimate a cutoff value for $\Delta G_{heterodimer.}$, we used all $\Delta G_{heterodimer}$ associated to optimal heterodimer structures and applied the method described above to determine the cut off values for $\Delta G_{heterodimer.}$ (Figure 9). As Figure 8 shows, the cutoff $\Delta G_{heterodimer}$ values for the stable duplex structures were around -10 kcal/mol.

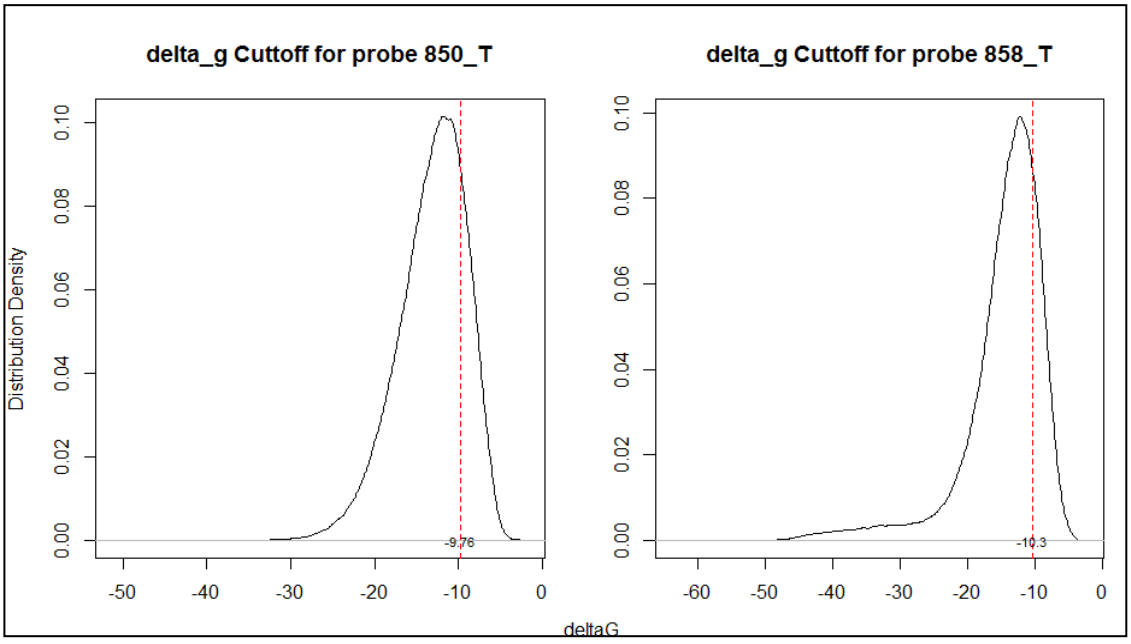


FIGURE 9: $\Delta G_{heterodimer}$ cut off values. a) $\Delta G_{heterodimer}$ cut off values for probe 858_T hybridized to the targets generated based on chromosome Y. b) $\Delta G_{heherodimer}$ cut off values for probe 850_T hybridized to the targets generated based on chromosome Y

2.4.3: Predicting the Effects of the Target Length and Secondary Structures on Probe-Target Hybridization

2.4.3.1: First Experiment

The results of this experiment are summarized in Figure 10. As the results

indicate, increasing the target length increased the number of duplex structures which satisfied our criteria ($\Delta G_{heterodimer} \leq$ -10kcal/mol and the predicted target-percent bound increased at least 10%). For example, when the length of targets associated with 858_T increased just by 1nt from each side, 13561 heterodimer structures which did not meet our criteria would satisfy them now, but when the length of targets for this probe increased by 45nt , a total of 169,036 heterodimer structures which previously did not meet our criteria would satisfy them now.

Interpretation of this result is not simple because, as target lengths get longer they may provide additional probe binding sites. Thus we had to filter the results to look only at those sequences that preserved the same pattern of base complementarily between the two strands. That is why, we conducted the second experiment.

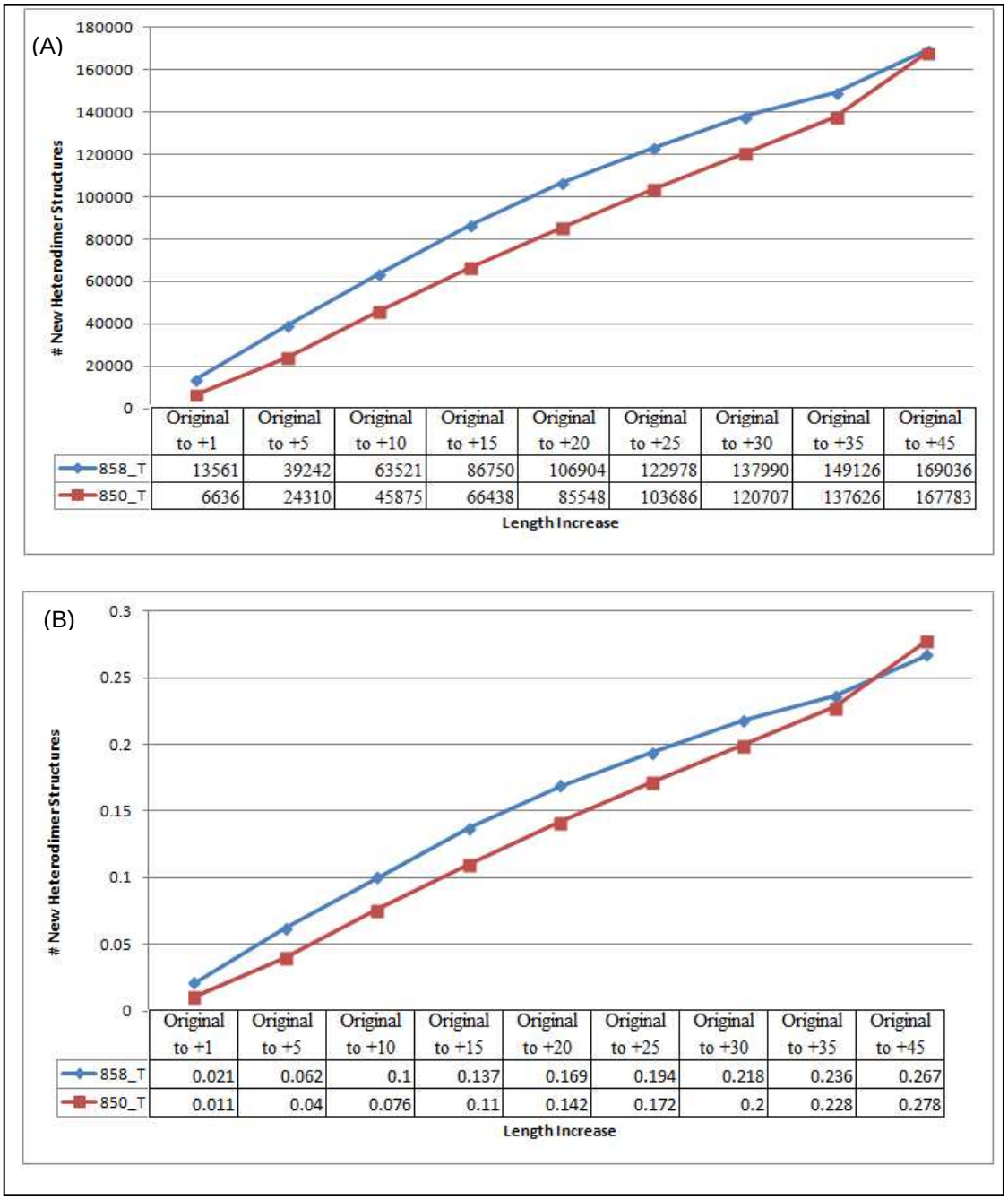


FIGURE 10: A summary of the effect of increasing length on heterodimer stability, where the core duplex complementarity is retained. In the top panel the y-axis has the actual number of structures while in the bottom panel the y-axis shows the percent increase over baseline instead. In both panels the x-axis indicates the increment in target length. The value in the table below shows the actual number of structures compared to the base targets (in panel A) and the percent increase in the number of structures compared to the base targets (in panel B).

2.4.3.2: Second Experiment

The results of this experiment are summarized in Figure 11. As we expected, by adding a new filter to only look for those heterodimer structures that preserved the same base complementarily between the two strands, the number of heterodimer structures which met our criteria was significantly reduced. For example, when the length of targets associated to 858_T increased by 45nt from both sides, using this filter resulted in only 15,530 heterodimer structures meeting our criteria, while in the absence of this filter 169,036 heterodimer structures would meet our criteria.

Comparing the results summarized in Figure 10 with those from Figure 11 showed: Within each target-set, there was a linear increase in the number of stable structures, but when the binding position was restricted the number reached a maximum at extensions of 15 and 20nt for probes 858_T and 850_T consecutively, and then began to decrease again.

By using the base complementarity restriction (Figure 11), we filtered those target structures that occlude the probe binding site, because increasing target length without constraining the sequence produced internal structures that blocked the probe binding site. Thus the fraction of stable duplexes using the same bases to form a heterodimer decreased.

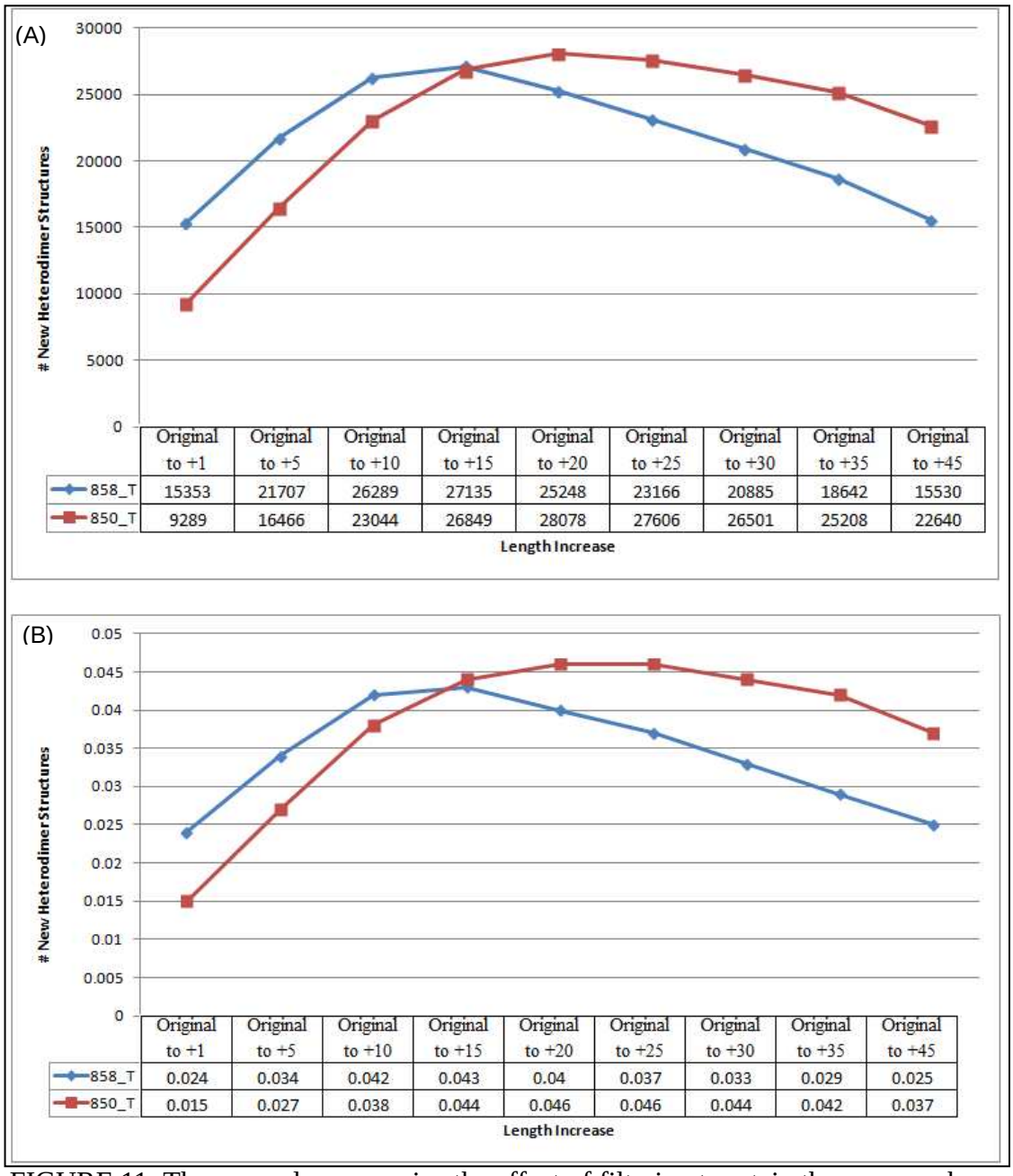
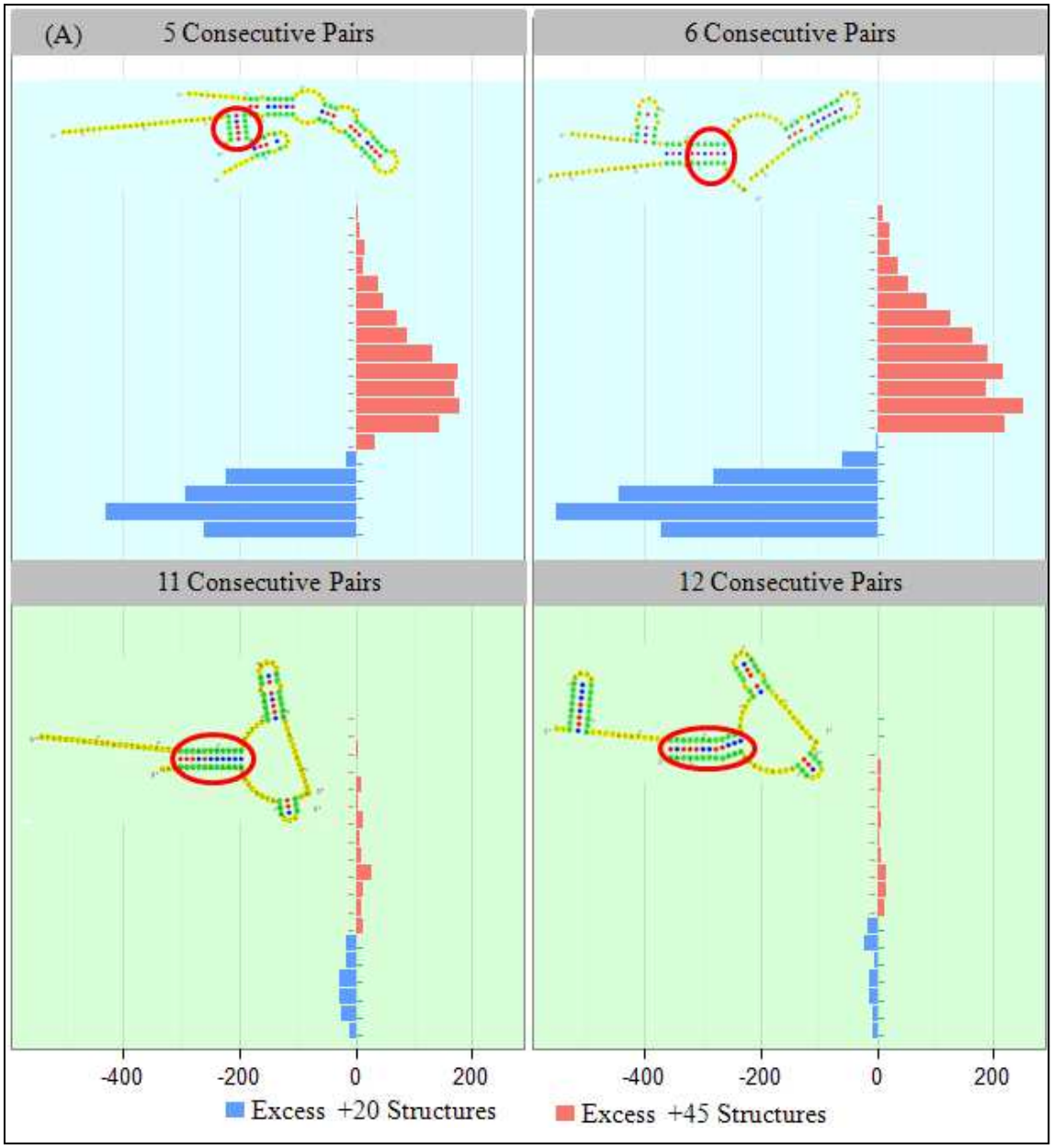


FIGURE 11: These panels summarize the effect of filtering to retain the same probe-binding core from the base heterodimer across all longer targets. In the top panel the x-axis has the actual number of structures while in the bottom panel the x-axis shows the percent increase over baseline instead. In both panels the y-axis indicates the increment in target length. The value in the table below shows the actual number of structures (in panel A) and percent increase in the number of structures (in panel B) for each of the base targets.

2.4.3.3: Third Experiment

Results summarized in Figure 12 indicated although the number of alternate heterodimer structures for both probes began to decrease after some point, the stability of the remaining structures continued to increase.



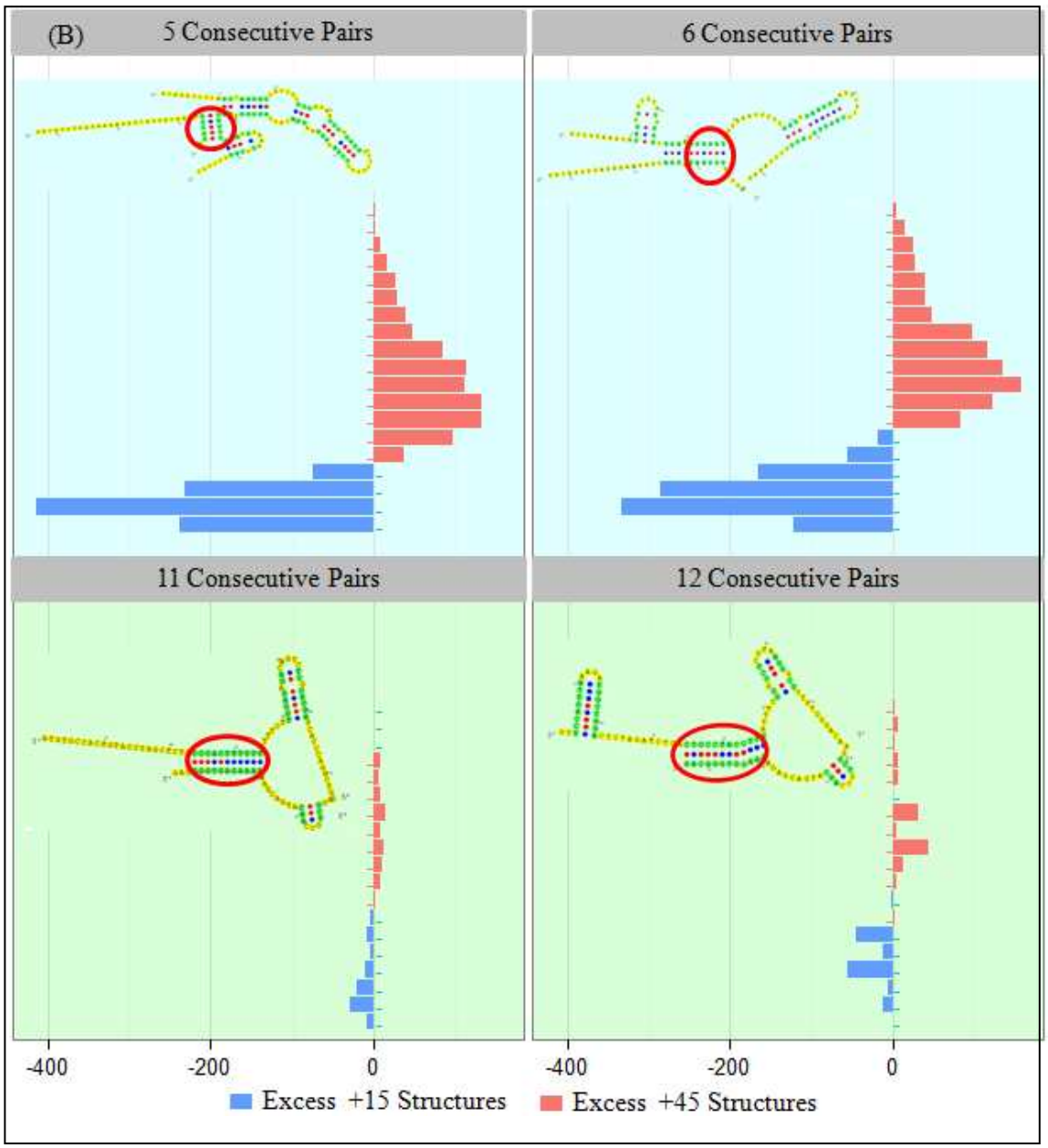


FIGURE 12: In panel (A) for probe T-850, we binned together duplex structures with the same ΔG. Within a bin, we then subtracted the number of heterodimer structures of length base +45 from those of length base +20 and then plotted the ΔG distributions for all heterodimers which had 5, 6, 11, and 12consecutive complementary bases. In panel (B) for probe T-858, we carried out the same process but in this case the lengths were base +45 and base +15.

2.4.4: Results of Hybridization

For targets of each set we ran two hybridization experiments. The targets' concentrations used for each of these experiments are indicated in Table 4.

TABLE 4: The final target concentrations used in each experiment.

Target	Experiment #	Final Concentration (nM)
1571-150	1	~104.6
1571-50	1	100
1571-150	2	~120
1571-50	2	150
857-150	1	~89
857-50	1	~100
857-150	2	~125
857-50	2	~150
643-130	1	~60
643-40	1	50
643-130	2	~95
643-40	2	100

TABLE 5: List of the all predicted $\Delta G_{heterodimer}$, number of H-bonds, Percent bound (PB), and minimum nucleation length for each heterodimer structure under the hybridization conditions (Table 1).

Thermodynamic Values for Heterodimer Structures					
Targets	Probes	ΔG°	H-bonds	PB	MN
1571-150	5072	-26.75	36	99	1 set of 6 nt
1571-50	Bro129	-15.55	22	95	2 sets of 7 nt
857-150	5071	-30.68	19	99.73	2 sets of 5nt
857-50	5071	-16.07	19	98.77	2 sets of 5nt
643-130	Bro857	-13.19	19	0	1 set of 6 nt
643-40	Bro857	-7.36	19	0	1 set of 6 nt

2.4.4.1: Results of Hybridization for Target Set 1 (1571-150 and 1571-50)

OMP predicted the information summarized in Table 5, showing a) at equilibrium

both targets in this set bound >=95 %, b) minimum nucleation length is longer for the short target (2 sets of 7nt) in comparison with the long target (1 set of 6nt), and c) the number of H-bonds involved in the heterodimer structure is greater for the longer target (36 versus 22). However the results of both hybridization experiments (Figure 13, 14) show a hybridization signal was only detected for the longer target (1571-150); therefore, the percent bond and minimum nucleation length could not be a driving force for this hybridization, because if they were, hybridization signal must be detected for the shorter target instead.

We believe the number of H-bonds involved in forming the heterodimer structure was the dominant factor stabilizing this hybridization.

Results of the first and second hybridization experiments for the target 1571-150:

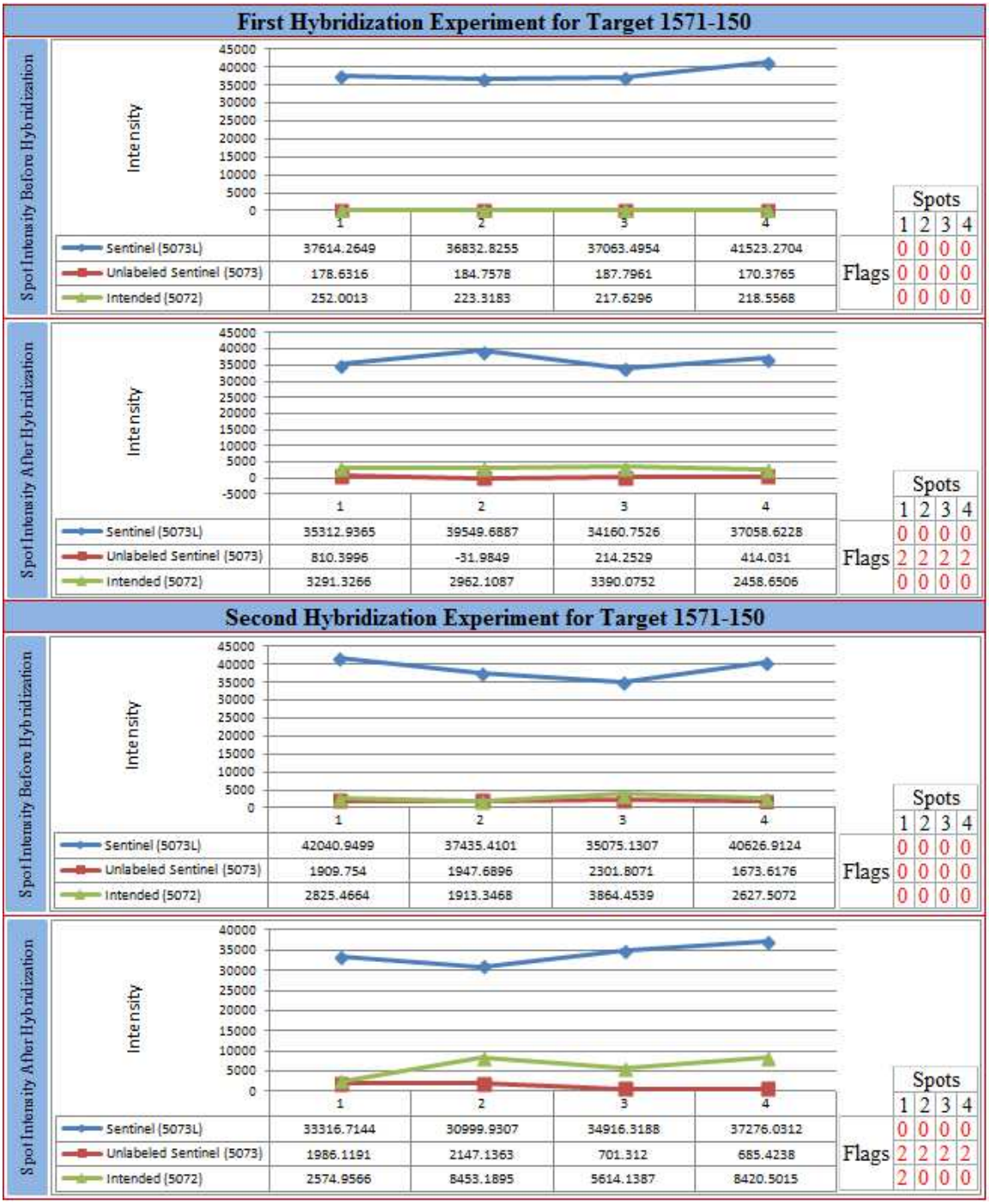


FIGURE 13: This figure summarizes the results of two hybridization experiments for target 1571-150 and contains 1) plots of spots intensities before and after hybridization, 2) the intensity values for each spot located in a table at the bottom of each graph, and 3) spot quality flags which are located in a table at the bottom right corner of each graph. In the spot quality table, flag 0 means the spot has a good quality, flag 2 means empty spots and flag 3 means poor quality spots.

Results of the first and second hybridization experiments for the target 1571-50:

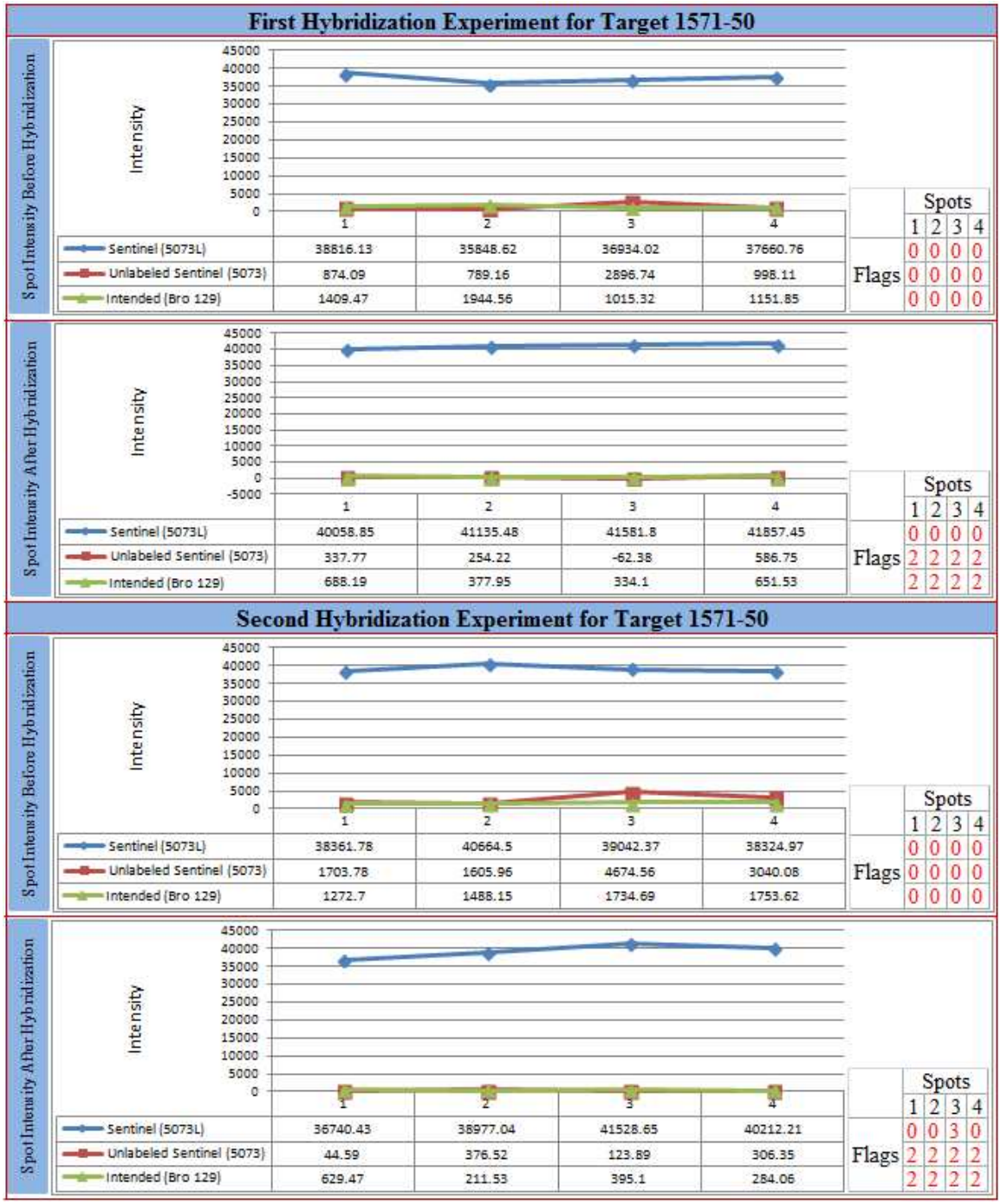


FIGURE 14: This figure summarizes the result of two hybridization experiments for target 1571-50 and contains 1) plots of spot intensities before and after hybridization, 2) the intensity values for each spot located in a table at the bottom of each graph, and 3) spot quality flags which are located in a table at the bottom right corner of each graph. In the spot quality table, flag 0 means the spot has a good quality, flag 2 means empty spots and flag 3 means poor quality spots.

2.4.4.2: Results of Hybridization for Target Set 2: (857-150 and 857-50)

The results of the first and second hybridization experiments, which are summarized in Figures 15 and16, indicated that hybridization signal was only detected for the longer target (857-150) in this set. If the number of H-bonds, PB or minimum nucleation length, or a combination of these factors, were driving stable hybridization, the hybridization signal should be detected for both, or neither, because both duplexes have the same number of H-bonds (19 nt), very similar percent bound levels (~98%), and identical base complementary between the two strands. Therefore in this case there must be another factor(s) which stabilizes this hybridization.

Examining the optimal heterodimer structures associated to the members of this set (Figure 1) shows that the heterodimer structure of the longer target had some secondary structures adjacent to the probe-target binding interface while the heterodimer structure associated to the shorter target did not have such structures. We believe these surrounding structures stabilized the duplex formed by the longer target by exclusion of solvent (entropy-driven) but it is also possible that, once formed, it diffused away more slowly, shortening the time to re-form the complex.

Results of the first and second hybridization experiments for the target 857-150:

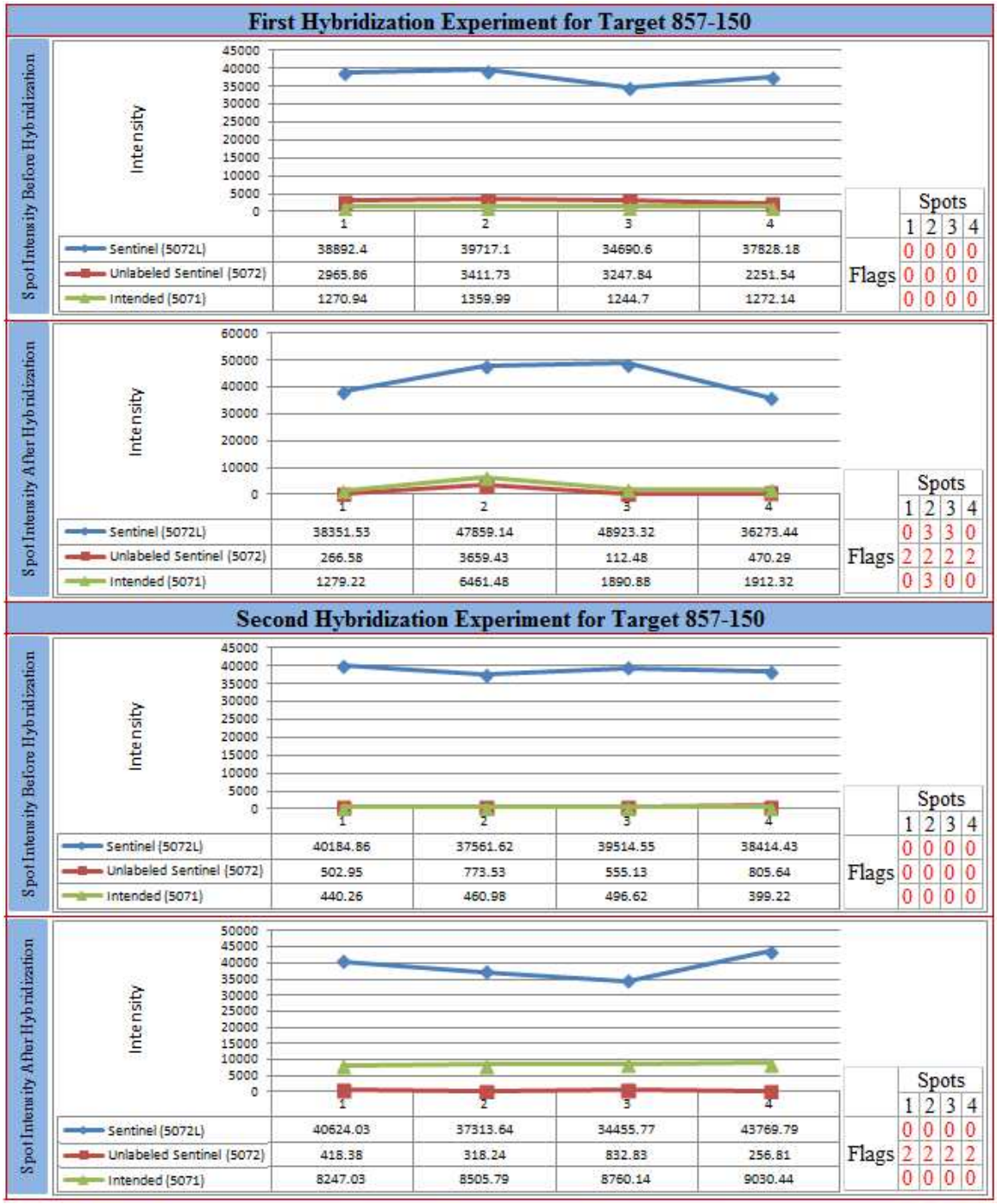


FIGURE 15: This figure summarizes the results of two hybridization experiments for target 857-150 and contains 1) plots of spot intensities before and after hybridization, 2) the intensity values for each spot located in a table at the bottom of each graph, and 3) spot quality flags which are located in a table at the bottom right corner of each graph. In the spot quality table, flag 0 means the spot has good quality, flag 2 means empty spots and flag 3 means poor quality spots.

Results of the first and second hybridization experiments for the target 857-50:

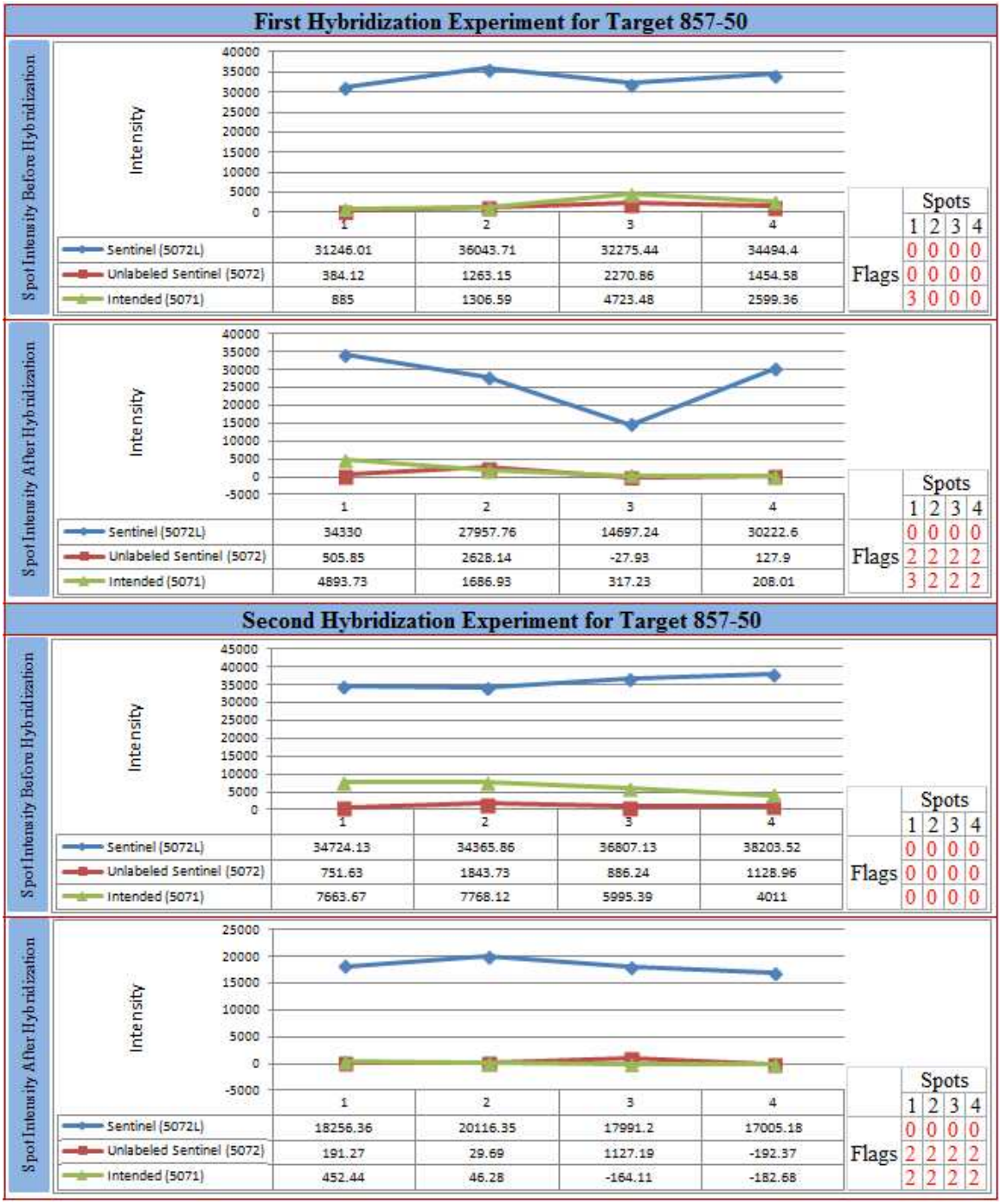


FIGURE 16: This figure summarizes the result of two hybridization experiments for target 857-50 and contains 1) plots of spot intensities before and after hybridization, 2) the intensity values for each spot located in a table at the bottom of each graph, and 3) spot quality flags which are located in a table at the bottom right corner of each graph. In the spot quality table, flag 0 means the spot has good quality, flag 2 means empty spots and flag 3 means poor quality spots.

2.4.4.3: Results of Hybridization for Target Set 3: (643-130 and 643-40)

The result of the first and second hybridization experiments, which are summarized in Figures 17 and 18, indicated that hybridization signal was only detected for the longer target (643-130) in this set. If the number of H-bonds, PB or minimum nucleation length, or a combination of these factors, were driving stable hybridization, then the hybridization signal should be detected for both, or neither, of them, because both duplexes have the same number of H-bonds (19 nt), very close to the same percent bound (~0%), and identical base complementary between the two strands. Therefore, there must be another factor(s) that is stabilizing this hybridization.

Examining the optimal heterodimer structures associated with the members of this set (Figure 1), it can be seen that the heterodimer structure of the longer target had some secondary structures adjacent to the probe-target binding interface while the heterodimer structure associated to the shorter one did not have. We believe these surrounding structures stabilized the duplex formed by the longer target by exclusion of solvent (entropy-driven) but it is also possible that, once formed, it diffused away more slowly, shortening the time to re-form the complex.

Results of the first and second hybridization experiments for the target 643-130:

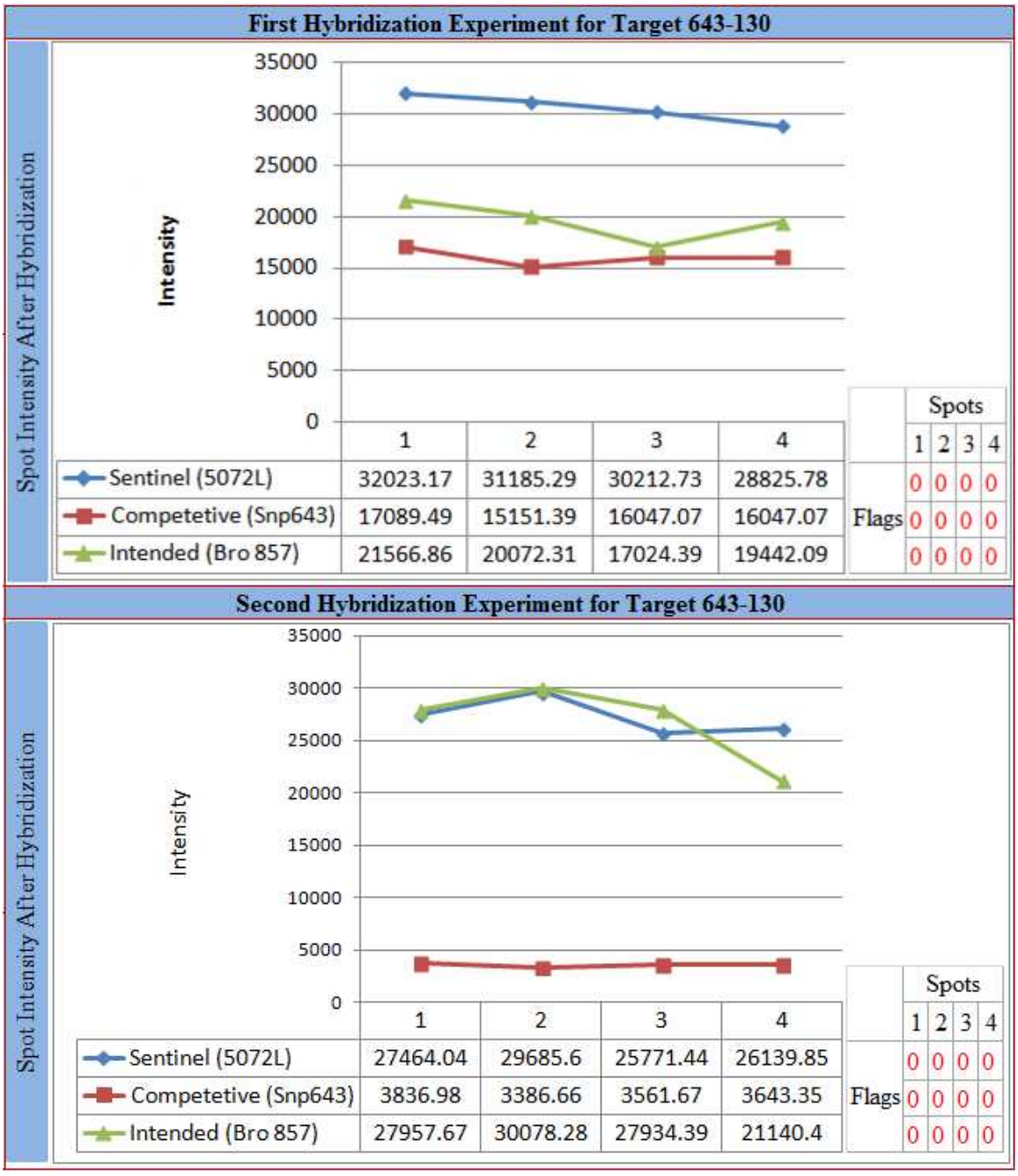


FIGURE 17: This figure summarizes the result of two hybridization experiments for target 643-130 and contains 1) plots of spot intensities after hybridization, 2) the intensity values for each spot located in a table at the bottom of each graph, and 3) spot quality flags which are located in a table at the bottom right corner of each graph. In the spot quality table, flag 0 means the spot has good quality, flag 2 means empty spots and flag 3 means poor quality spots.

Results of the first and second hybridization experiments for the target 643-40:

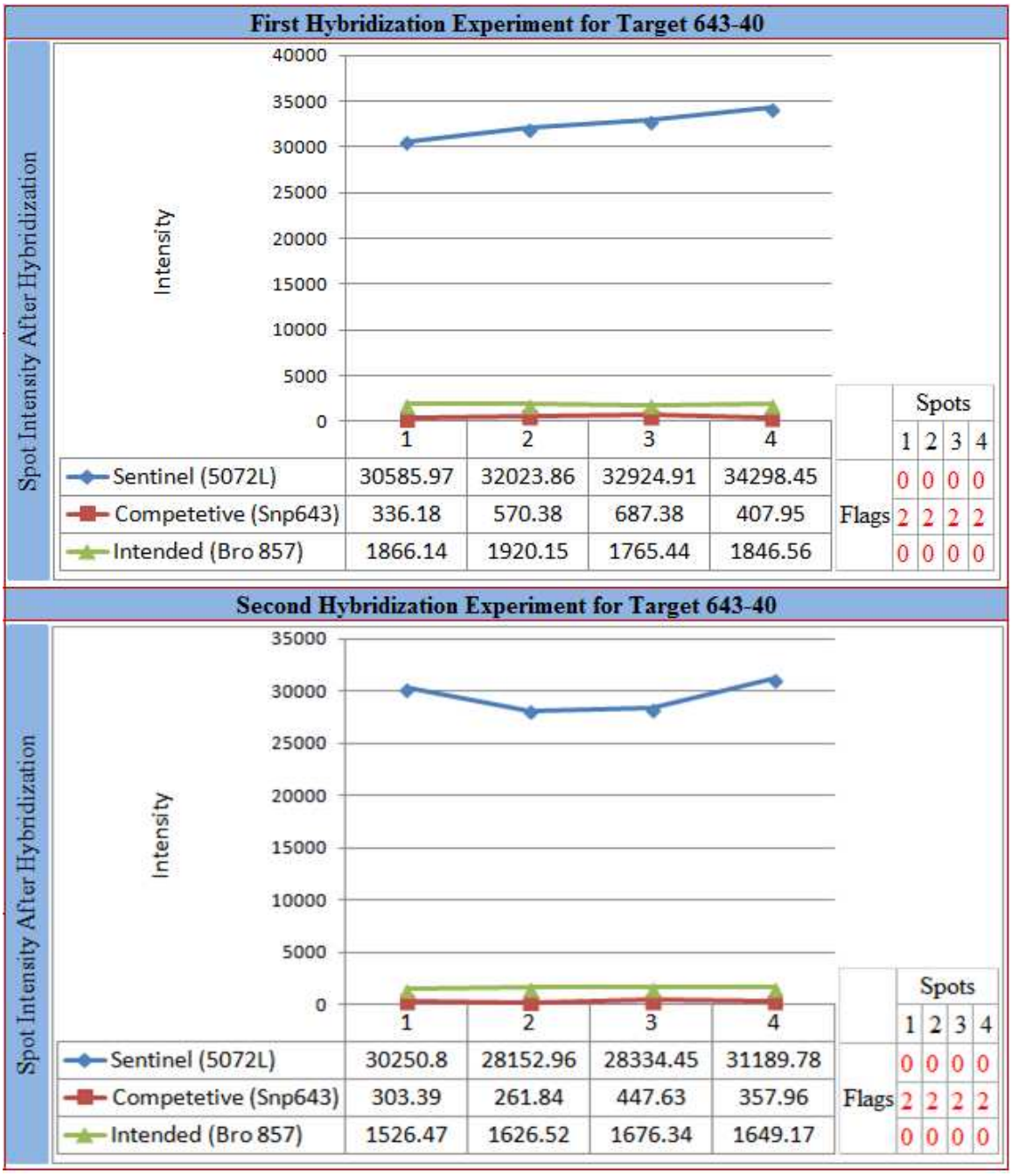


FIGURE 18: This figure summarizes the result of two hybridization experiments for target 643-130 and contains 1) plots of spot intensities after hybridization, 2) the intensity values for each spot located in a table at the bottom of each graph, and 3) spot quality flags which are located in a table at the bottom right corner of each graph. In the spot quality table, flag 0 means the spot has good quality, flag 2 means empty spots and flag 3 means poor quality spots.

2.5: Discussion

In these experiments we first modeled and then designed experimental targets that had partial but measureable binding to the probes, so we could discriminate the effect of secondary structure and investigate whether secondary structures stabilize or de-stabilize the binding of targets to probes when they are adjacent to the probe-target binding interface. This is important to hybridization technologies in which the target is of variable length (the result of random shearing) or longer than the probe complement for other reasons (as are most amplicons).

To investigate this matter, we 1) designed several series of nested sets of sequences around the common heteroduplex forming region and modeled them, 2) designed 3 target pairs to have differing degrees of secondary structure external to the probe-binding region and performed microarray hybridization experiments on them.

Our results from molecular simulations indicated that stable secondary structures on the boundary, when not impinging on the ability of targets to access the probes, stabilized the probe-target hybridization. The results summarized in Figure 11 show that for ~ 5% of those structures which had $\Delta G_{heterodimer}$ value equal to -10 kcal/mol, an increase in the target length from 33 to 70b which preserved the same probe-target base complementarity, resulted in a more negative overall $\Delta G_{heterodimer}$. In fact we modeled beyond 70nt length, but the number of heterodimer structures which satisfied the above conditions decreased to the small number, although for those structures the stability of the product increased. This is summarized in Figure 12.

Our results from the experimental data, in particular the hybridization results obtained from second and third target sets (target sets 643 and 857) confirm the

prediction that as a target gets longer and this sequence allows the formation of secondary structure in the regions adjacent to the target-probe binding site, duplex formation is stabilized relative to a target having the same duplex forming pattern but no such adjacent structures. This has implications for the analysis of microarray data when partially matching targets with lengths longer than the duplex are in the mixture. While a perfect match will dominate and, barring significant internal structure, is likely to yield a reasonably accurate measurement, when no such competition is in place, imperfectly matched targets can bind quite stably.

The results of experimental data for all three target sets show that a hybridization signal was only detected for the longer target in each set. To investigate what factor(s) was the driving force for this hybridization, we have examined all the features (the total number of H-bonds, a minimum nucleation length of consecutive H-bonds, the ΔG of the duplex and the percent bound) considered in our design and found out: a) number of H-bonds and , minimum nucleation length, and percent bound were not this driving force, because both targets in set 2 and 3 (857 and 643) have the same H-bonds, a minimum nucleation length, and percent bound (Table 5) and the results (Figure 15,16, 17, and 18) shown that hybridization signal was only detected for the longer target in each set, b) the ΔG of the duplex could be this driving force, because it is consistently lower in the longer target in compare with the shorter target in each set.

Comparing the duplex structures (Figure 2) and thermodynamic parameters associated with such complexes (Table 5) shows that given both a short and long target in that preserved the same base complementary between the probe and target, the longer target gave considerably more signal, which is best accounted for by the folding of

adjacent regions and exclusion of solvent, since the ΔG of the heteroduplex regions remained unchanged.

CHAPTER 3: THE EFFECT OF STRUCTURE ON SEQUENCEING FIDELITY ON THE ION TORRENT PGM

3.1: Overview

While it is accepted that high GC- regions may affect the ability of a DNA polymerase to process, so that highly structured templates are difficult to copy faithfully in PCR reactions (114-117) and may be difficult to sequence in Sanger sequencing reactions (118), there has been little attention paid to the relationship between structural features of templates and measurement errors in high throughput sequencing (HTS) platforms. On the other hand, considerable attention has been given to the problems created by the various chemistries: the homopolymer problem on the 454 and Ion Torrent platforms are well documented (119,120) as is the apparent sensitivity of the Illumina chemistry to high AT regions (78,121,122).

To test the hypothesis that structure affects the fidelity of read-through on the short-read high-throughput sequencing platform in our lab, the Ion Torrent Personal Genome Machine (PGM), we have used 10 synthetic constructs, which were initially designed for microarray studies, to investigate the effects of structures (hairpins) at or around probe-target binding sites on probe-target hybridization.

In our design, we considered the following three aspects of the hairpin structures in the templates: 1) the lengths, 2) the frequency, and 3) the location of each relative to the sequencing adaptor. The length and number of hairpins was considered because biophysical studies showed that the transition from a folded to coiled structure (opening

of hairpin structures) depends on both the size and the number of the hairpins (123,124). The location of the hairpin was included because the PGM software returns no data if there are 8 or fewer bases past the key sequence (personal communication, Ion Torrent training course). Thus a very stable hairpin right on the boundary with the adaptor might appear to return no sequence when in fact a small number of bases had been read. We placed some structures near that boundary in order to investigate the interference of the adjacent hairpin structure on primer-target binding or polymerase attachment to the duplex region.

Each construct contained first, a core 50mer segment (derived from the sequence of a Brucella gene) elongated from one or both sides by adding oligonucleotides in a self-complementary segment that can self-hybridize to create a range of stable secondary structures (Figure 19), and second, sequencing adaptors needed for the platform - for some targets both template orientations were created to see if this changed the outcome of sequencing.

Since the PGM creates amplified copies of one target on each bead of a chip, this platform yields the sequences of individual input molecules rather than the bulk sequence property characterized by the Sanger sequencing with gel electrophoresis methods. The sequence derived from these beads was assembled using the AbySS (125) software package with a Chastity filter option ‘on’. The Chastity filter (126,127) is a base call quality control filter which is defined by the ratio of the highest of the four (base type) intensities divided by the sum of the two highest intensities.

3.2: Type of Sequencing Errors

Sequence quality has a direct impact on the usefulness and biological relevance of the data (128), any excessive errors may have significant effect on our interpretation of the results. The primary source of these errors can be from sequencing, assembly or the alignment processes.

Several variables account for the sequence read quality. For example DNA extraction and library preparation may yield chimeric sequences. Sequencing errors at the reagent flow level may cause loss of base resolution. There are a range of potential sequencing errors that can be introduced in the sample preparation steps, such as the PCR amplification bias observed in Illumina data (78), polyclonal errors observed in SOLiD data (129), or homopolymer sequencing errors observed in PGM reads (130). Library preparation can limit sequencing coverage that allows the full length of template molecules to be inferred. This last factor is important because lack of base coverage uniformity may cause variation in a poorly covered region to be mis-called or even omitted. PGM coverage is known to be biased against sequences with very low (< 20%) or high (> 80%) GC rich regions (131).

Another source of error may arise during read assembly. The accuracy of assembly mostly depends on the software and its parameters (132). To reduce the computational effort required to assemble millions of reads (133), most of the assemblers for next-generation sequencing break the reads into smaller sequences called k-mers (k defines the size of the sequence to be matched) and then links k-mers sharing k-1 nucleotides to build a de Bruijn graph. The value of the parameter k has significant influence on the quality of the assembly (132).

Another source of error may arise during read alignment (134). Alignment to a known genomic scaffold is one the fundamental analysis step undertaken once the DNA sequence has been produced. It is often preferable to *de novo* assemblies due to the increased speed and reduced memory requirements entailed, but like *de novo* assemblies the accuracy of alignments varies considerably depending on the software and the parameters chosen (135).

In this project, since we were in the position of knowing the correct outcome, we optimized the parameters of the *de novo* assembler and alignment tool in order to maximize our ability to achieve individual target reconstruction. Aligning the resulting assemblies to their known targets allowed us to investigate 1) whether there was any association between the secondary structures and the sequence coverage, 2) the effects of k-mer size on contig assemblies, and 3) the effects on contig assembly of using a low-quality filter in addition to the Chastity filter.

3.3: Material and Methods

3.3.1: Reagent Acquisition

Oligonucleotides were obtained from Operon (all HPLC grade, integrity validation was carried out using polyacrylamide gels) and PCR reagents were obtained from New England BioLabs.

3.3.2: Overview

To carry out this study, we followed these steps: 1) Computational modeling of the targets' s structures under Ion-Torrent sequencing conditions, 2) Construction of target templates and sequencing libraries, 3) Verification of target templates by performing Sanger sequencing, 4) Preprocessing and analysis of the results.

1) Target Modeling Under Ion Torrent Platform Conditions:

We used the Oligonucleotide Modeling Platform^{TM} software (Visual OMP v7, DNA Software) to model ten constructs used in microarray study under the physical conditions prevailing on the Ion Torrent platform, as follows: temperature at 42 - 45 ºC, [Na^{+}]=~40mM, [Mg^{++}]=6.3mM, ph= 7.5. Figure 19 depicts the most stable secondary structures for each target under the sequencing conditions.

Based on the results obtained from OMP modeling we classified the targets into the following groups, also shown in Figure 19. The Group 1 templates contain structures with either very small hairpins (3 to 6bp) or with loops that interrupt the hairpin. Group 2 templates had a longer hairpin (~11bp). Group 3 targets contained one or two very long hairpin structures (~20bp). Group 4 targets had 5 or 6 small hairpins in close proximity to each other (separated by 5 to 10 nucleotides). The majority of the structures within each group had the hairpin occurring at approximately the middle of the sequence.

Note: The modeling parameters reflected the sequencing reaction conditions, which are quite distinct from most microarray hybridization conditions, and the duplex region in this sequencing experiment is at one end, where the sequencing primer binds, and is contiguous, again in distinction to the microarray hybridization experiments.

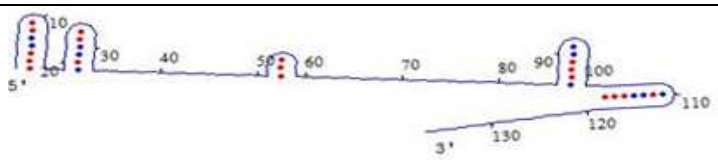
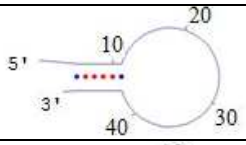
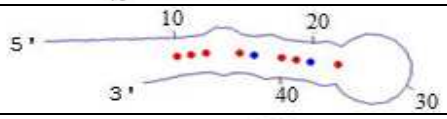
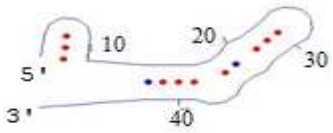
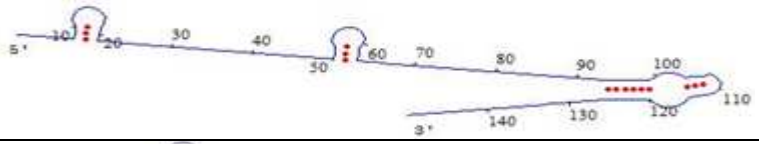
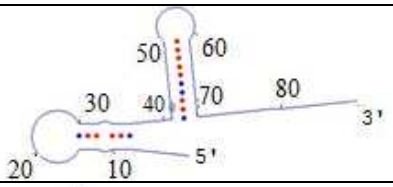
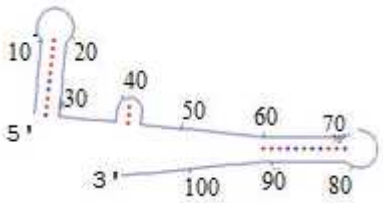
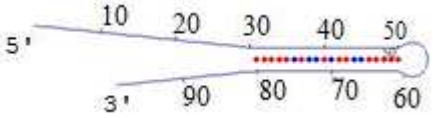
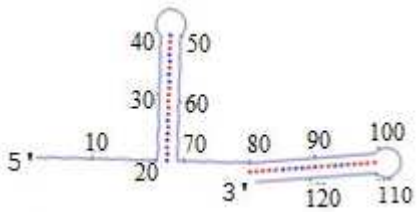
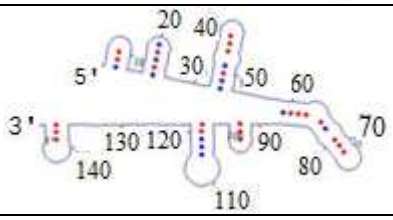
Groups	Target Name	Visual OMP Images
1	1981-137	
1	129-50	
1	1571-50	
1	857-50	
1	1571-150	
2	1981-89	
2	1981-109	
3	1981-99	
3	1981-129	
4	857-150	

FIGURE 19: Predicted secondary structures of all targets under conditions present during

Ion Torrent PGM sequencing. Images were generated using Visual OMP.

3.3.3: Sequencing Library Construction

Targets were constructed using overlapping oligonucleotides which were designed to span the entire length of each target with overlaps of 15-35 nucleotides at the 3' ends (Figure 20). This assembly was performed in three steps, annealing, extension, and full-length PCR. Annealing was carried out in a volume of 30 µl, using 0.2 µM of each oligonucleotide in a buffer containing 1.5 mM $MgCl_2$ and 1X HF buffer (Phusion high-fidelity buffer, Promega Corp.). After mixing, the solution was heated to 95°C for 5 minutes, followed by gradual cooling (60 minutes in water 150ml of initially 100°C) to 37°C. Each reaction was continued by adding 200 µM of each dNTP and 0.4U of Phusion polymerase (Promega), followed by incubation for 60 minutes at 37 ºC. After all components were added the full-length construct was amplified using primers to the ends alone. This was done in a 50 µl reaction containing 5 µl of assembled target, 200 µM of each dNTP, 0.4 U of Phusion polymerase, 0.2 µM of terminus primers, 1.5 mM $Mgcl_2$ and 1X HF buffer. PCR cycling was: 95º C for 3 min followed by 30 cycles at 95º C for 30 s, 58 ºC for 30 s and 72 ºC for 30 s, and terminated by 3 min extension at 72 ºC.

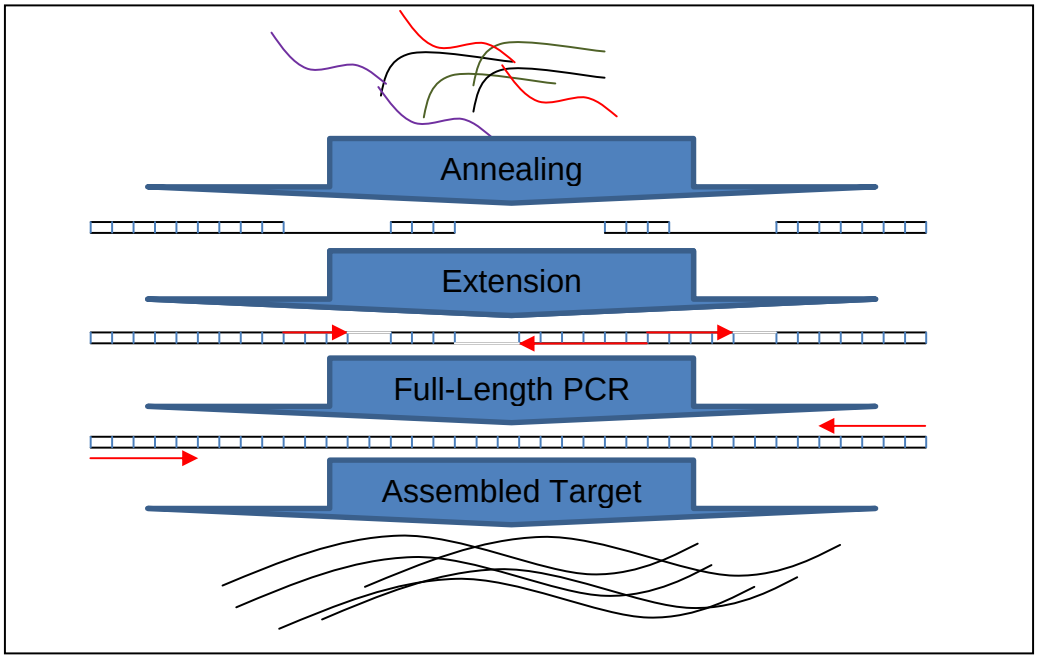


FIGURE 20: Schematic representation of steps in the template assembly process.

3.3.4: Template Modification for the Ion-Torrent Platform

The validated templates were next modified to be suitable for sequencing on the Ion-Torrent by performing standard PCR with fusion primers. PCR was carried out in a 50 µl reaction containing 5 µl of assembled target, 200 µM of each dNTP, 0.4 U of Phusion polymerase, 1.5 mM $MgCl_2$, 1X HF buffer and 0.2 µM of the Life Technologies-specified forward and reverse fusion primers for the PGM (ordered from Operon MWG). At the time the reactions were perform, the 5' region of one adaptor was biotinylated (adaptor A) while the other primer was not (adaptor P1). That the expected, correct modification had occurred was verified by analyzing 5ng of each target on 8% polyacrylamide gels (Figure 21).

Due to the Ion-Torrent read length limitation (~100 bases at the time of this study), and the length and location of secondary structures on some of our targets, a bidirectional sequencing approach was performed for 6 out of 10 targets, while for the other four targets sequencing was carried out from only one orientation. That is, in total

we created 16 (6 * 2 + 4 = 16) distinct, structured amplicons if the orientation is considered distinct. Targets were prepared for sequencing according to the Ion Template 314 kit User Protocol (Life Technologies, Ion Community resources for PGM Users).

Since no protocols were available for performing paired-end sequencing on the Ion-Torrent platform at the time of this study, we created amplicon libraries for both strands for those targets requiring bidirectional sequencing. Because these are not truly "paired-end" targets, in the analyses we refer to them as paired-targets, to emphasize that the pairs do not originate from the same ISP. For the other four targets, we produced amplicon libraries for one strand only and in the analysis we refer to each as a single-target.

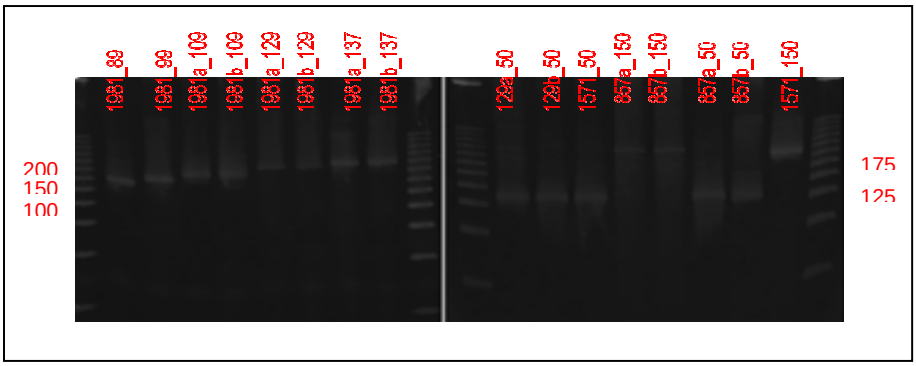


FIGURE 21: Gel picture of 16 Ion Torrent targets which have adaptor A (30 nt) on the 5' side and adaptor P1 (30 nt) on the 3'side.

3.3.5: Template Verification

Sanger sequencing on an ABI 3130 Genetic Analyzer was performed to verify that PCR errors had not corrupted the majority of our input sequences. The lengths of all targets were first assessed by analyzing 5ng of each assembled target on 8% polyacrylamide gels. Templates of the expected length were purified using Ampure XP™

beads (Agencourt) according the suppliers protocol, and then sequenced using standard Sanger Big Dye (v3) sequencing reactions on an ABI 3130 sequencer using the suppliers protocol (Life Technologies/ABI).

3.3.6: Ion Torrent Run

We obtained the same concentration of our templates and combined them to prepare the concentration needed for emulsion PCR (according to the manual), then we followed the instructions for the emulsion PCR and sequencing according to the manuals for kit version 1. After the sequencing run finished, we used the fastq outputs for our analysis.

3.3.7: Preprocessing and Analysis of the Results

3.3.7.1: Classification and Alignment of Ion Torrent Reads

The first stage of analysis followed a 3-step method (outlined in Figure 22) comprising classification, pairwise alignment and multiple alignment. Based on the known target signature (that is, we have unique keys for each target) the reads associated to each amplicon were separated into individual groups. Within each separated group, pairwise global alignment, using the Biopython Emboss suite (136) with a gap penalty of 50 and gap extension penalty of 0.5 was carried out. As a last step, in the multiple alignment process (from a python script, available in the supplementary materials for this chapter), gap(s) were introduced as needed to maintain sequence concordance in the set in the following manner: when a gap in the target alignment pattern was found the gap was introduced to all target and read alignment patterns except the read associated to the target which had a gap in that position. Figure 23 illustrates this process in detail.

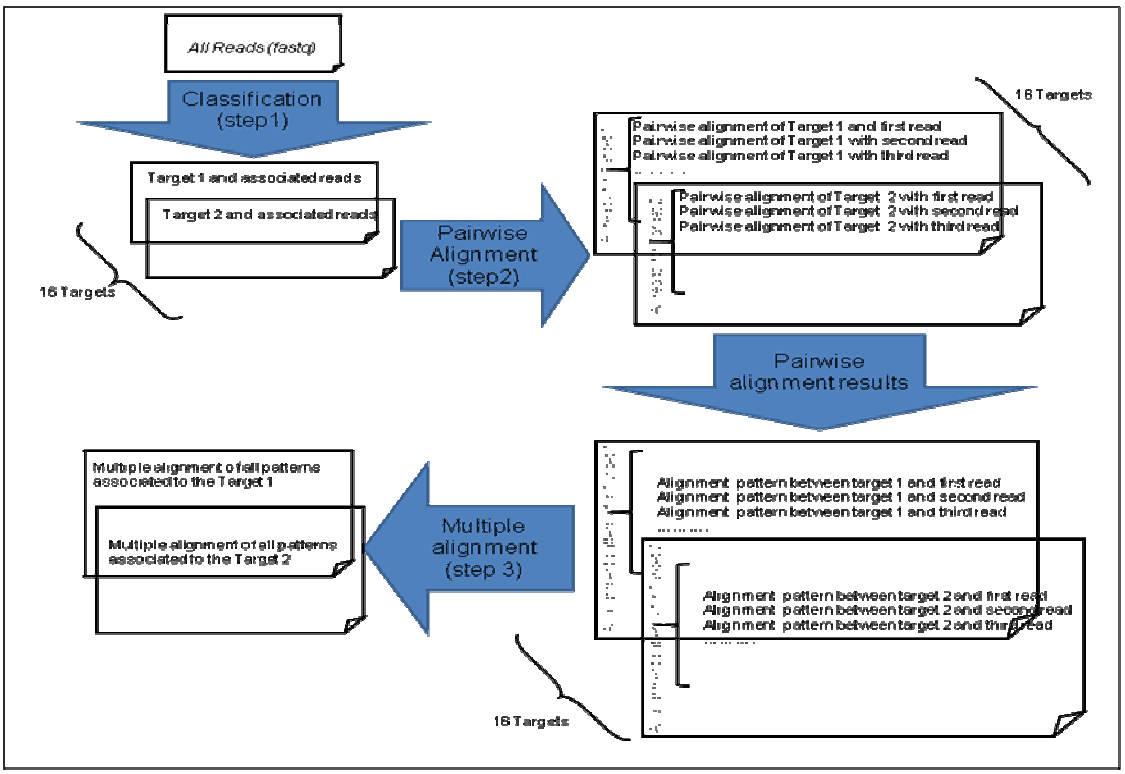


FIGURE 22: Schematic representation of the steps for aligning the reads to the original template.

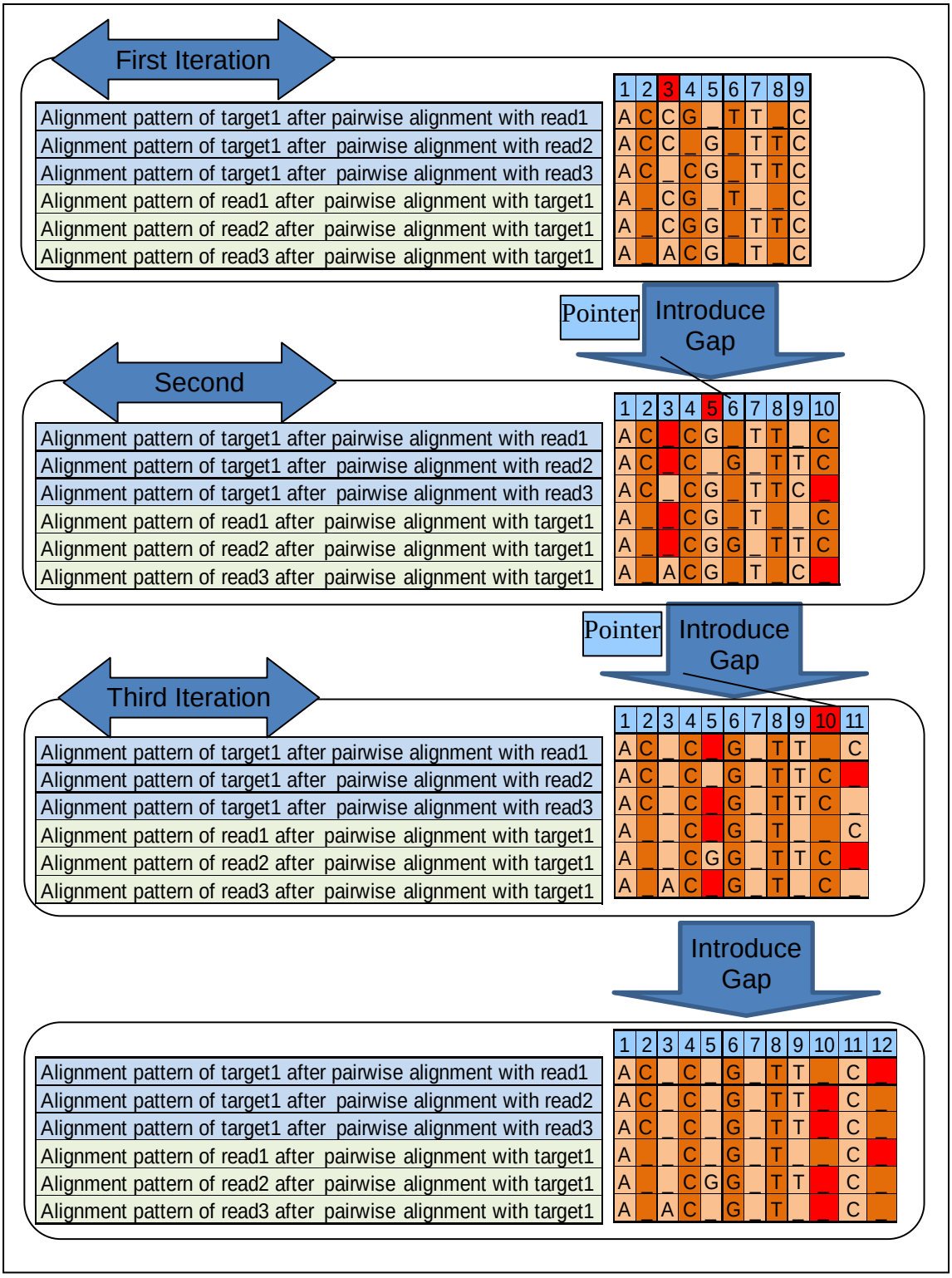


FIGURE 23: Illustration of step 3 of the process which was used to align reads to the associated target.

3.3.7.2: *De Novo* Assembly

Since in this experiment, prior to sequencing, we knew the sequences of our targets, we used this fact to investigate whether secondary structures on the sequencing reads affect the quality of assembled sequences. To investigate this we performed *de novo* assembly on the reads associated to each target and compared the result of assembly with the known sequences. Prior to assembly, a python script were used to remove the adapter sequences after which AbySS 1.3.0 (137) was used to assemble the contigs

Based on a Technical Note by Illumina, the only quality filter that definitely improves an overall assembly on their platform is the Chastity filter (126). To investigate this matter for the Ion-Torrent platform, two assemblies were performed in parallel using the filters provided in the Abyss assembly tool, in the first assembly the Chastity and 'end-trimming of low-quality base calls' filters were used, and in the second assembly just the Chastity filter was used. The low-quality trimming filter, which trims bases from the ends of reads, was set to a cut-off value of 20 for all assemblies. Comparison of the results of these two assemblies is discussed in the Results section titled 'Results of *De Novo* Assembly'.

Since the target sequences are known, for each target a k-mer that maximized the correct target reconstruction was determined. For the paired-targets, contig assemblies for each strand were conducted separately, then the resulting contigs were combined and the final sequence was aligned to the known target. For the single-target products clearly only the single-direction contig was available to align with the known target.

3.4: Results

3.4.1: Alignment of the Reads to Designated Target

Sequencing on an Ion Torrent PGM 314 chip produced 162,032 reads. Table 6 shows the distribution of these sequences across the 16 targets. After applying the described alignment methods, we detected substitutions, insertions, deletions, and sequence matches in every position for all associated reads and generated the graphs of incident rate of matches and deletions.

TABLE 6: Distribution of Ion-Torrent sequencing reads across the 16 targets

Target Name	Number of Reads
129a-50	19,972
129b-50	17,636
1981a-109	1,952
1981b-109	2,268
1981a-129	1,387
1981b-129	7,525
1981a-137	1,105
1981b-137	787
857a-150	4,662
857b-150	1,056
857a-50	1,791
857b-50	10,802
1571-150	20,810
1571-50	34,529
1981-89	2,434
1981-99	3,768
Not Found	29,548
Total	162,032

Figure 24 illustrates one example of an incident rate graph for a member of our structure types Group 2 and Group 4 (1981a-129, 857a-150 respectively). In each graph, x-axes indicate the target position (target length) and y-axes indicate the number of reads

which have deletions or matches at every position. The target sequence is given along the bottom part of the graph. Secondary structure positions are highlighted in dark gray and regions are demarcated by vertical red dashed lines. The horizontal red dashed line indicates the threshold imposed to eliminate noise: each position in a given target must be observed in at least 50 reads to be included in the summary. The graphs for all targets were generated and are available in Appendix.

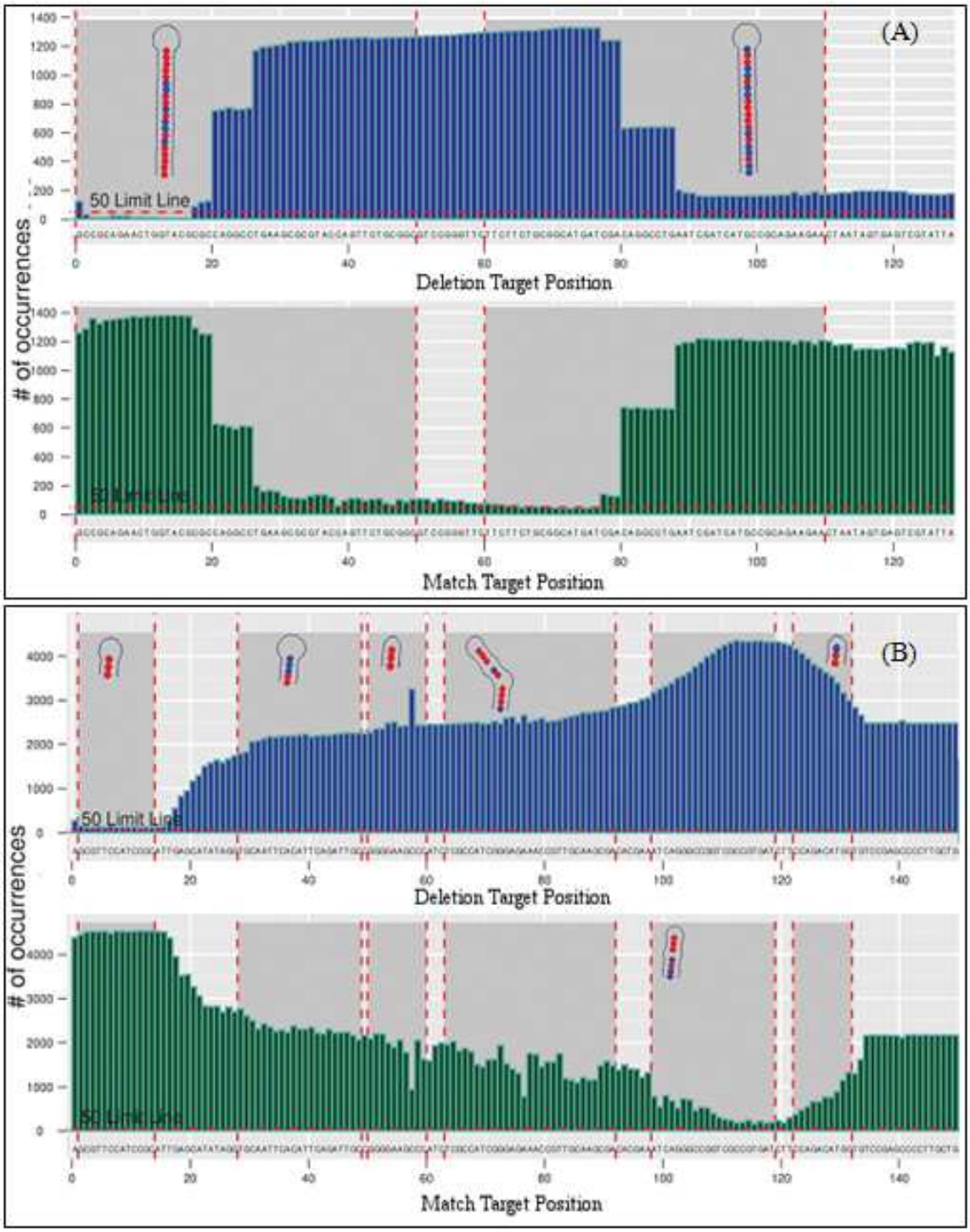
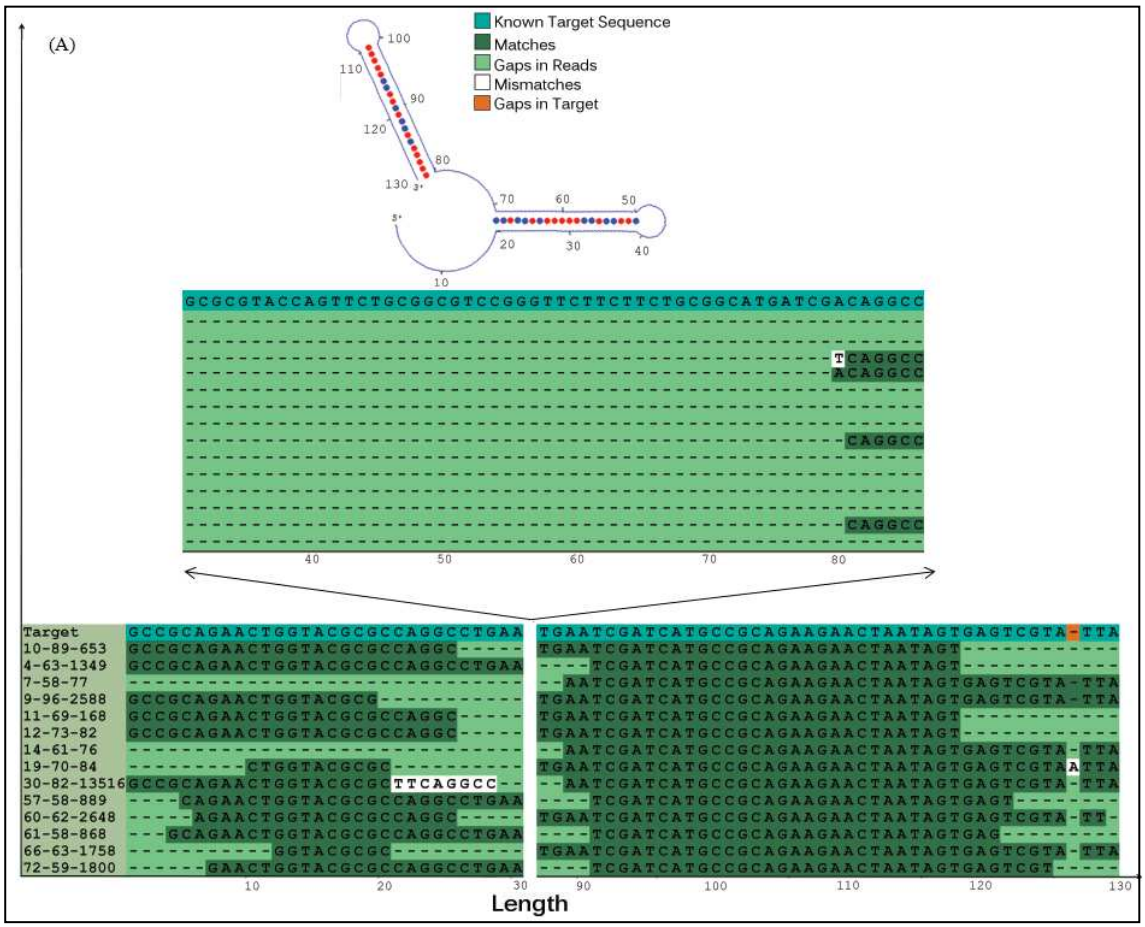


FIGURE 24: Graphs A, and B illustrate the deletion and sequence match distributions, respectively, for targets 1981a_129, and 857a_150 which are representatives of groups 2 and group 4. The structure contributing to deletions is shown in the relevant part of the deletion graph.

3.4.2: *De Novo* Assembly

To investigate whether the structures had any effect on the sequencing process, we used ABySS to assemble the contigs, choosing parameters that maximized each target's correct reconstruction. Figure 25 illustrates the assembly results generated by ABySS for the 1981_129, and 1981_99 templates respectively (the graphs for all targets can be found in Appendix II). In this example the Chastity filter was used and the k-mer size was set to 58. For the first target (A) which is a paired-target, contigs were generated using reads associated with both strands, while for the second target (B), which is a single-target, the reads from the single available strand were used to generate all contigs.



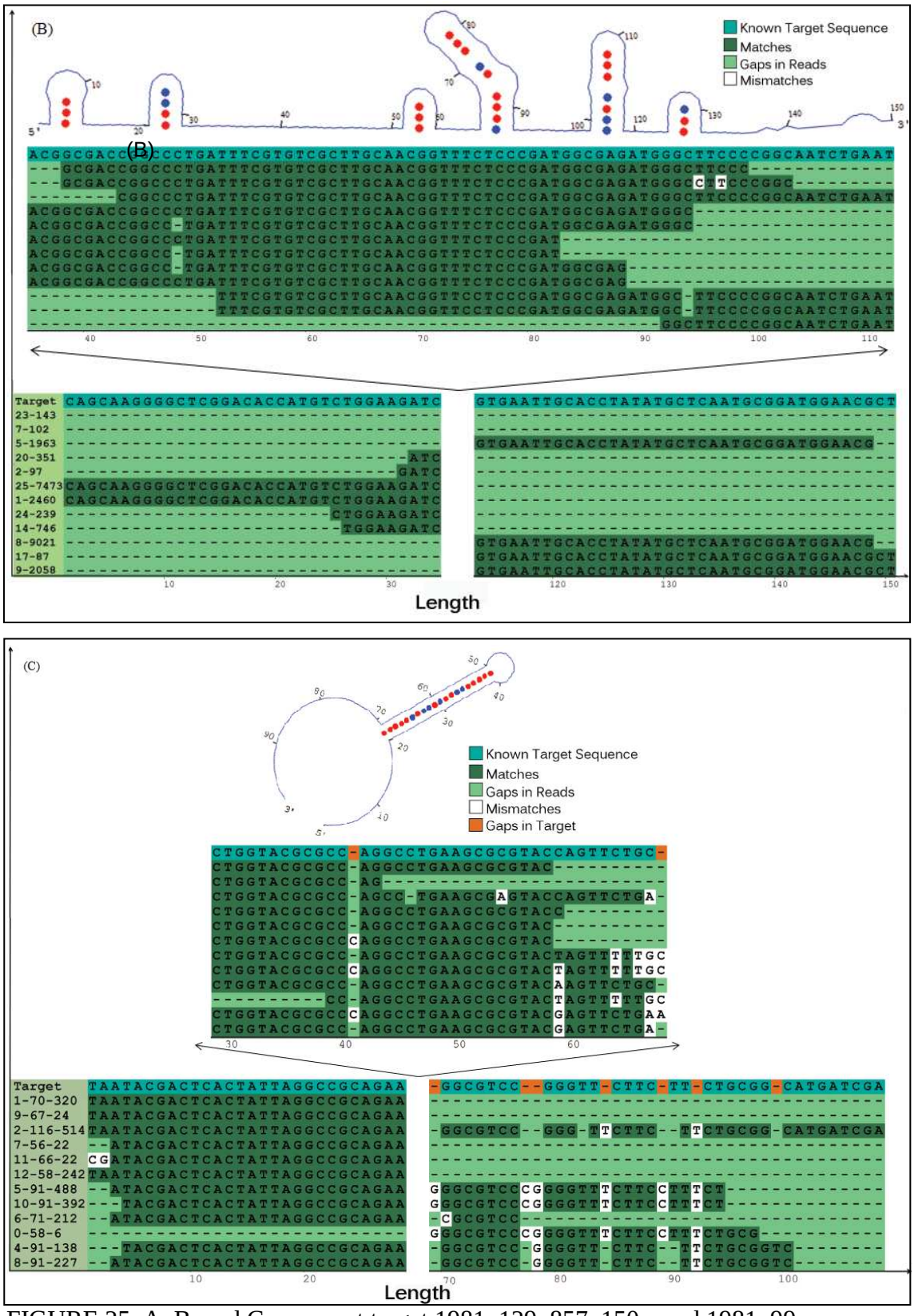


FIGURE 25: A, B, and C represent target 1981_129, 857_150a, and 1981_99 respectively. Each sub-figure has two parts: 1) a model generated by Visual OMP

software which illustrates the type and location of secondary structures, and 2) the multiple sequence alignment (MSA) representation of the contigs and original target sequences. To maximize the use of figure space, the middle part of each MSA (bases 31 to 90 for target A and base 27 to 70 for target B and bases 24 to 68 for target C) was shifted up. In each MSA representation the x-axis indicates the target length and the y-axis shows the template label. In each MSA, sequences highlighted in light blue identify the original sequence, dark green indicates the individual contig sequences, light green indicates the gaps, white indicates parts of the contig sequences which are not matched with target sequence and orange indicates gaps introduced into original target as the result of the MSA process because a de novo approach was taken. Since, during alignment, gaps were introduced into the original template sequences, the maximum length indicated on the x-axis may be longer than the actual length of the original template; therefore, to find the counterpart area between a structural model and MSA, gaps must be included.

3.5: Discussion

The results presented above clearly demonstrate that there is a strong association between sites of indels (although deletions were observed far more often than insertions, which are not shown here) and the location of secondary structures on the target. As hairpin structures get longer, as shown for targets 1981-129 (Figure 24a), or the distance between hairpins decreases, as shown in target 857-150 (Figure 24b), the sequencing reaction is subject to more mistakes, both as an increased rate of indels and as mis-incorporation errors (data not shown). These indels can be as small as a 1-nt deletion or insertion events, or as relatively long (20 bases or more) gaps in the assembled contigs (see Figure 25). The first type of error (small indels) are compensated for during the assembly process, in which the fully assembled contigs cover almost the entire length of the known targets, but contigs containing large insertions and deletions result in a large divergence from the known target. As shown in Figure 25, generated contigs for target 1981-129 (Figure 25a) are completely missing in the region between bases 31 to 80 (highlighted by light green area) while there is only one contig (0_115_546) generated for

target 1981-99 (Figure 24C) that almost covers entire length of the target; the majority of the contigs have errors between bases 28 to 74. Comparison with the folded structures of these two targets suggests a strong association between missing sequence information and the presence of hairpin structures. This is not a surprising result since there is a large body of evidence showing that during DNA replication secondary structures may cause DNA polymerase fork-pausing which as a result creates a high-frequency site for indels (again, mostly deletions) (138).

To illustrate the importance of this phenomenon, consider the processes of a typical RNA-Seq experiment. First isolated RNA will be converted to short cDNA fragments which are used as templates for a given NGS sequencing technology. After sequencing is conducted, reads are typically mapped to a reference genome, transcript library or exon-exon junction library to identify novel gene models, or refine existing gene models, or determine the gene expression level from read count statistics. If, as indicated by our results, some of the sequencing templates (fragmented cDNAs) have structures similar to those illustrated in Figure 24a, the sequencing reads may have missed a big portion of the actual sequence, thereby leading to a result that is incorrectly identified as a novel gene or a novel splicing variant. Hairpin structures are particularly common in untranslated regions (UTRs) and other regulatory sequences, where they have a functional role, exacerbating the interpretation issues.

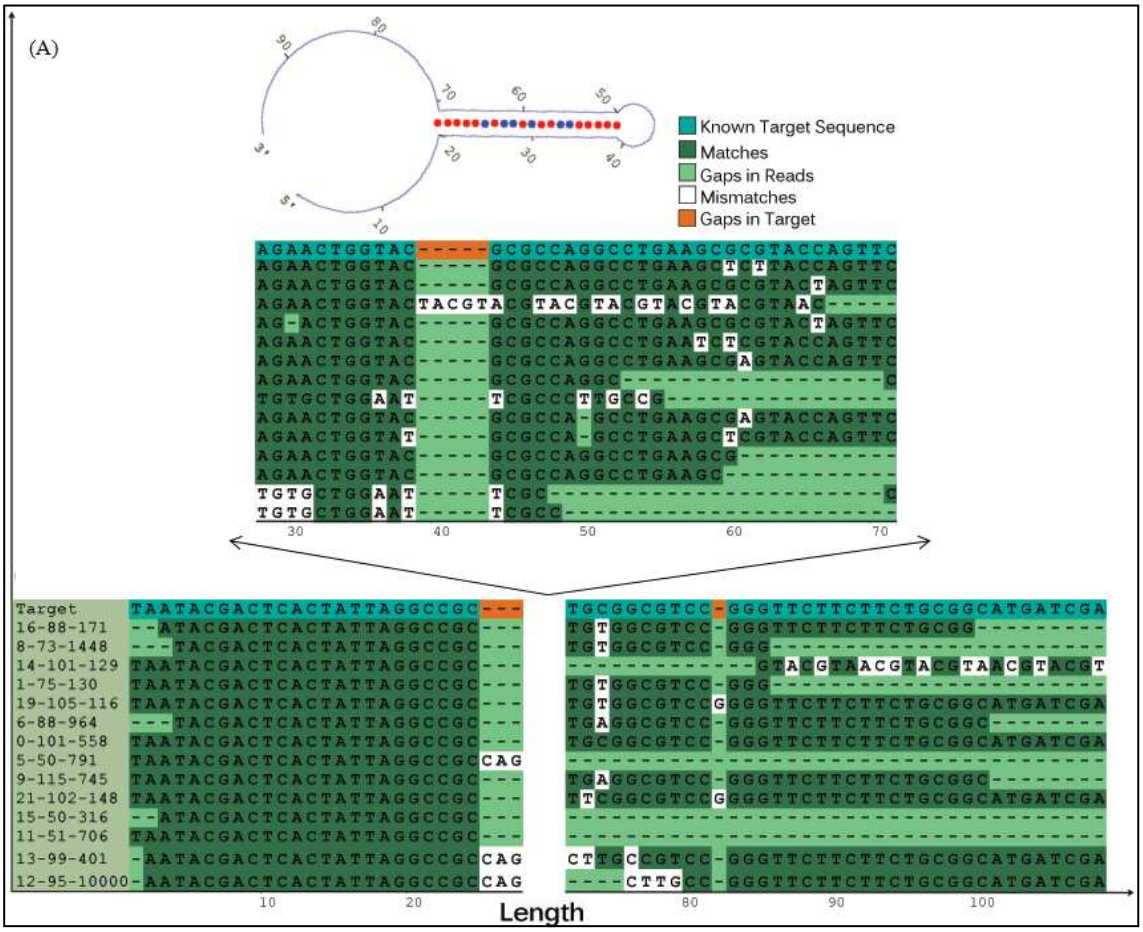
Another important message we obtained from our results is related to the size of the k-mer chosen for a *de novo* assembly. In all of the assemblers' algorithms, which are based on de Bruijn graphs (139), reads are decomposed into smaller sub-reads of length k, called k-mers. Our assembly results show that the length of the k-mers affects the

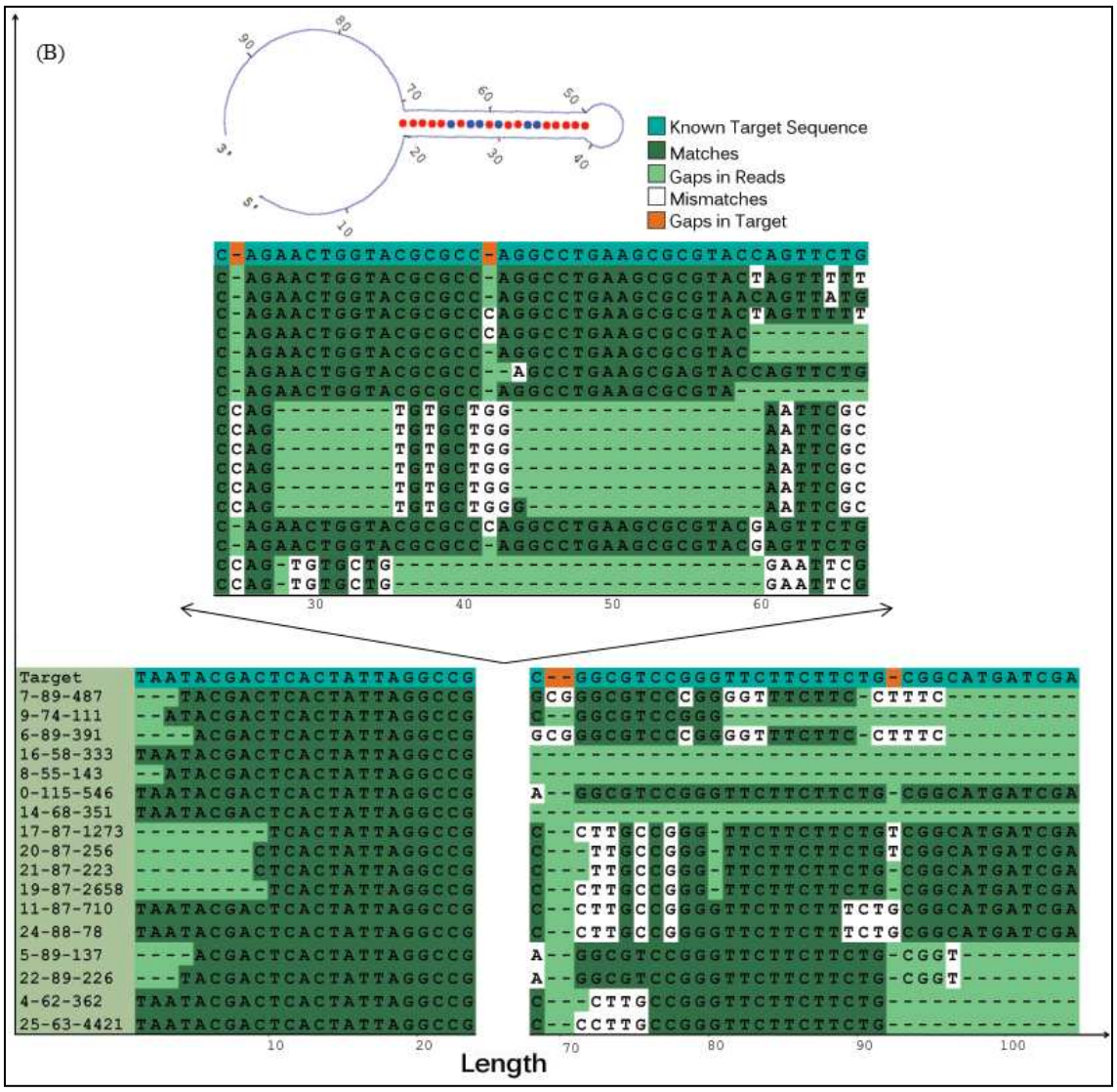
assembly results. If the selected k-mer size is longer than many of the reads that should map to a target, the short reads would be dropped from the assembly process and the resulting contigs will lose that part of the sequence.

The mechanism leading to an apparent deletion when structure is present has not been demonstrated, but we can suggest several possibilities. For targets that contain a long hairpin structure, when the polymerase reaches the structure, the resulting pause may lead to release of the polymerase from the DNA. If the polymerase falls off the target, the duplex may partially melt and the polymerase may re-bind further back, introducing repeats (which we observed, data not shown) or bind to a region that is apparently primed by the hairpin, placing it much further down the linear sequence (as shown above). Depending on the position of the structure and length of the target, many short reads may result. If the product just preceding a hairpin has many short reads, and the selected k-mer size is longer than those reads, the short reads would be dropped from the assembly process and the resulting contigs will lose that part of the sequence. If the region contains repeated elements and the selected k-mer size is smaller than those reads, then the short reads would be used in constructing the contigs with insertions close to the hairpins. To overcome these problems for contigs assemblies where strong structure is expected, we suggest the use of multiple k-mers, weighting the contigs according to the number of k-mers used to construct them. Previous studies have also showed that using multiple k-mers clearly improved the quality of *de novo* assembly of a transcriptome (140-142) although no rationale was given. Figure 19 shows an example of constructing contigs using multiple sequence alignment and multiple k -mers, from target 1981_99. For these assemblies, k-mer sizes of 50, 55, and 58 were used successively (Table 6). As

expected, and shown in the Figure 26 , with k-mer sizes of 50 and 55 more contigs were generated, some of which do cover the structured region (bases 19 to 69), but with a k-mer length of 58 the number of contigs decreased significantly while the length of repeated sequence leading in to the structured region increased.

We can conclude from the results that as the hairpin structures get longer and more condensed: 1) the distribution of deletions becomes more prominent, and the deletions mostly appear in the assembled contigs at the positions covered by the structures, 2) full reconstruction of the original sequence may not be completely obtainable by *de novo* assembly approaches, and 3) adding a low quality filter in addition to the Chastity filter will not improve the overall contig assemblies.




(B)
Known Target Sequence
Matches
Gaps in Reads
Mismatches
Gaps in Target
Target
7-89-487
9-74-111
6-89-391
16-58-333
8-55-143
0-115-546
14-68-351
17-87-1273
20-87-256
21-87-223
19-87-2658
11-87-710
24-88-78
5-89-137
22-89-226
4-62-362
25-63-4421
Length

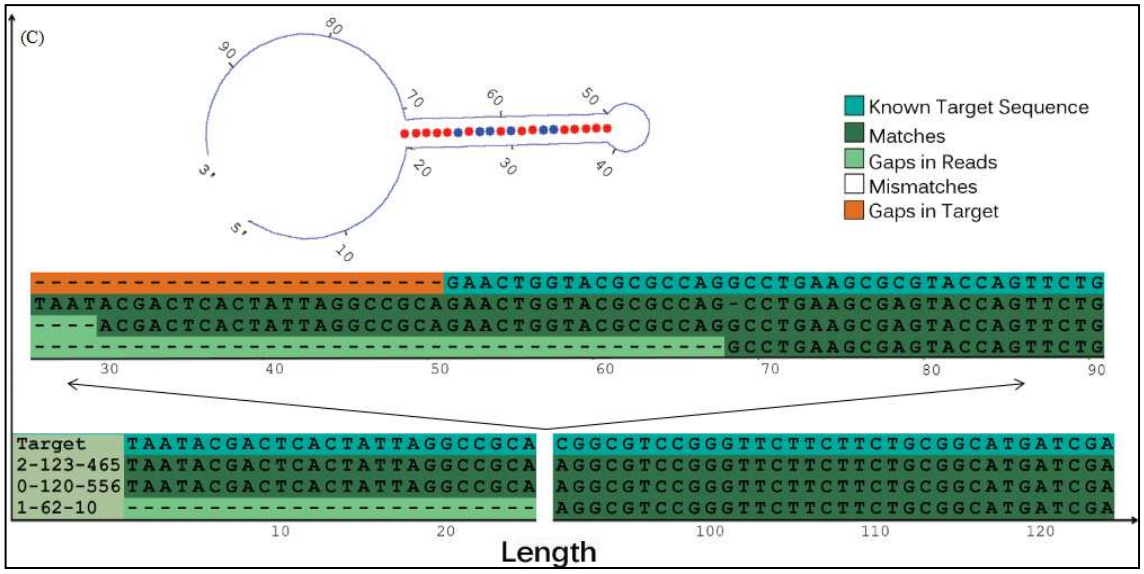

(C)
Known Target Sequence
Matches
Gaps in Reads
Mismatches
Gaps in Target
Target
2-123-465
0-120-556
1-62-10
Length

FIGURE 26: sub-figures A, B, and D contain: a) an MSA of target 1981_99 with contigs generated by using ABySS; k was set to 50, 55, and 58 in successive iterations. To maximize the use of figure space, the middle part of each MSA (bases 28 to 71 for target A and base 24 to 67 for target B and bases 26 to 90 for target D) was shifted up. In each MSA x-axes indicates the target length and the y-axes indicates the contig names. In each MSA, sequences highlighted in light blue indicates the original target sequence, dark green indicates the individual contig sequences, light green indicates the gaps, white indicates parts of the contig sequences which are not matched with target sequence and orange indicates gaps introduced into original target as the result of MSA . Since, during alignment, gaps were introduced into target sequences the max. length indicated on the x-axes may be longer than the actual length of the target; therefore, to find the counterpart area between structural model and MSA, gaps must be included.

CHAPTER 4: STUCTURE PATTERNS CHARACTERISTIC OF SHORT DELETIONS

4.1: Overview

Small insertions and deletions (INDELs) have been discovered in all human genomes that have been sequenced (90,91), but their location and extent depends on the sequencing platforms employed, the analysis approaches, and validation methods. A recent comparison of 5 sets of genome sequencing data, generated by different sequencing platforms (Figure 27) (143), indicates that there is a surprising level of variation in the form of INDELs (limited to events < 4bp) compared to SNP levels. We questioned this rate of INDEL variation, in part based on our experience with the behavior of highly structured targets in microarray and the Ion Torrent PGM platforms, in which stem-loop structures (internal folding of single-stranded DNA) produced high levels of apparent deletions. Mechanistically, it has been shown *in vivo* that formation of stem-loop hairpins interferes with DNA replication, repair, and translation (144,145). Tri-nucleotide repeats have been studied specifically as they are the basis of several genetic diseases and a number of forensic identification tests. Such sequences have been shown to fold into a stem-loop hairpin when part of *in vitro* assays (146) as well as *in vivo* (147). Because polymerase slippage on such sequences should lead to repeats appearing more frequently than is observed, a repair mechanism was sought. Recent studies (148,149) have revealed that human cells possess a DNA hairpin repair mechanism which can efficiently remove DNA hairpins containing 20 or 25 repeats, thus limiting

rapid changes, which is especially important in coding regions where such changes are likely to be deleterious.

Since in the sequencing process, regardless of the platforms, these editing systems are not present to prevent folding of DNA strands, we hypothesize that some of these reported deletions are related to the secondary structures that form under the conditions of the assays, and cause relatively high levels of skips or other types of errors.

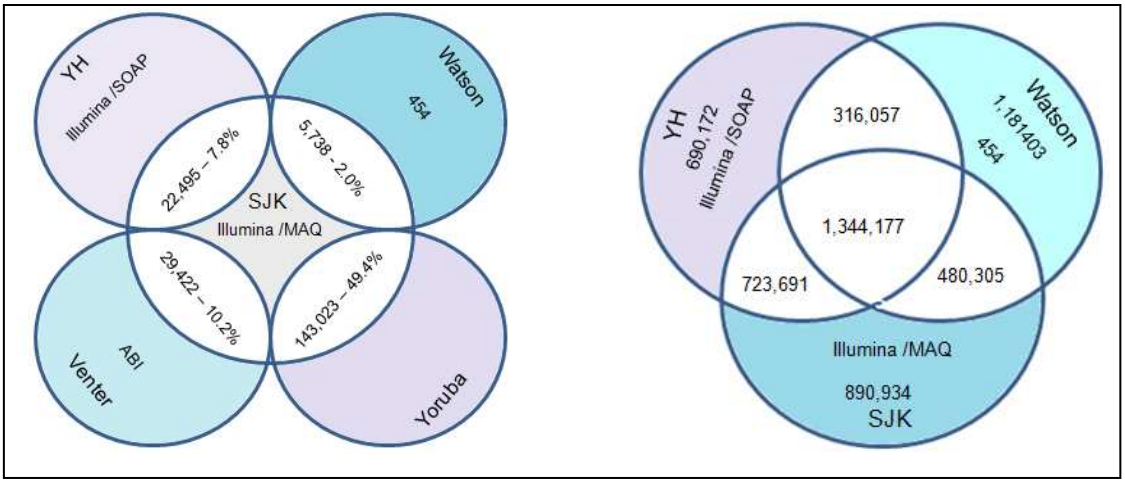


FIGURE 27: Comparison of INDEL (< 4b) and SNP frequency across different genomes.

In this study, our interest was to look for enrichment of structural motifs in or around deleted segments that are independent of the sequence itself. Although there could be an *in vivo* effect, we modeled using conditions that prevail in the sequencing platform rather than within cells, since this process most directly affects results. In the absence of a specific mechanism relating structural features to a deletion event, and because structural features are cardinal in nature, we used random forest modeling, a structured machine learning approach, to identify relevant features.

4.2: Characteristic of the Dataset Used in This Study

To investigate this hypothesis, we used the genomic sequence dataset created by Kim et al (1), one of those reported to have a high incidence of short deletions. We obtained the sequence reads and INDEL calls from the supplemental material provided by the authors. The original study design used the following set of steps in obtaining the data elements to analyze:

1) Genomic DNA samples were obtained from an anonymous healthy Korean adult male known as AK1.
2) Paired-end and singleton reads were generated using the Illumina GA and GAII Instruments with standard protocols. Reformulated cleavage reagent was used to generate sequence reads of up to 2 X 106, much longer than ordinary read length at the time of publication, which was (2 X 36). Longer reads were used to identify INDELs up to 29 bases in length.
3) High-quality reads were aligned to human reference genome build 36.3 using the GSNAP alignment tool and allowing up to 5% mismatches.
4) SNPs and INDELs were identified using the Alpheus^{TM} software system.
5) For validation, 67 Putative SNPs, indels and deletions were validated by targets Sanger sequenced using ABI 3730xl DNA analyzer and ABI BigDye Terminator cycle sequencing. The final data set included 95,143 small deletions, of which 3603 (length>=3b) map to chromosome 1 (homo sapiens). Of the 67 selected variants, all were confirmed by Sanger sequencing. These variants were distributed over all 23 chromosomes.

4.3: Materials and Methods

4.3.1: Part I: Investigation into the Presence of a Structural-Dependent Pattern That Predicts the Presence of a Short Deletion, Based on the Base Content and Helical Regions in the Neighborhood of the Deletion Sites.

4.3.1.1: Fragment Set Construction

To investigate our hypothesis, we carried out the following steps to construct fragments surrounding the regions of interest and to simulate their secondary structures:

1) Assemble human chromosome one using contigs reported on NCBI map viewer for build 36.3.
2) Construct 70 base length fragments, centering on the deleted segment, using the physical locations of all deletions greater >=3 nucleotides, as reported in the Kim et al. paper for chromosome 1 (3603). The sequence was generated using an Illumina instrument and recommended library kits, which at the time of publication produced 36-base read lengths. The authors used modified conditions to generate longer reads, averaging 106 bases. The data was obtained from the Short Read Archive (SRA), using identification number XXX. Given sequence read lengths between 36 and 106 nucleotides, we used the average of these two numbers (~70) to construct the fragments.

   To calculate the fragment boundaries, we subtracted the length of each short deletion from 70 (total fragment length) and divided the result by 2. The deletion was centered and the upstream and downstream boundaries were determined based on the calculated value. For example, if the length of a given deletion is 12, we build our 70 bases fragment by concatenating the 29 ((70-

12)/2=29) upstream bases to the start of the deletion region (12 bases) and the 29 downstream bases (29+12+29=70) to the end of the deletion region. Throughout this study, these fragments are called true deletion (TDEL) fragments.

3) Add Illumina sequencing adaptors to each side of every fragment.

Note: In the OMP™ modeling software (from DNA Software) these TDEL fragments were designated as 'probe' sequences, while the sequencing primer was designated as the 'target' sequence. The presence of a surface (the flow-cell) and a double-stranded segment of template (the bound sequencing primer) can both change the predicted folding of the target.

4) Model the optimal heterodimer structures of all of the fragments using the developers edition of Oligonucleotide Modeling Platform (OMP DE™) (150), under the conditions reported in Table 4 (sequencing conditions). OMP uses nearest-neighbor model with empirical data to determine a set of thermodynamic parameters for all optimal (heteroduplex structures which are energetically most likely to appear) and sub-optimal (heteroduplex structures which are less energetically favorable) heteroduplex structures For this study, we selected just the optimal structures.

Note: We used sequencing adaptors and primer (Figure 28) to model fragments but since in this study we were investigating the effect of secondary structures and assay conditions on skips or other types of error in reads, we used only part of the structure which formed as a result of folding fragment to itself and not the part which was in duplex form.

5) Assign a code to the type of structure a base is involved in which include:

hydrogen bonded hairpins (H), loops (L), bulges (B), or none of the above (F =free) (Figure 28).

These are standard structure representations (151) when crystallographic coordinates are not available.

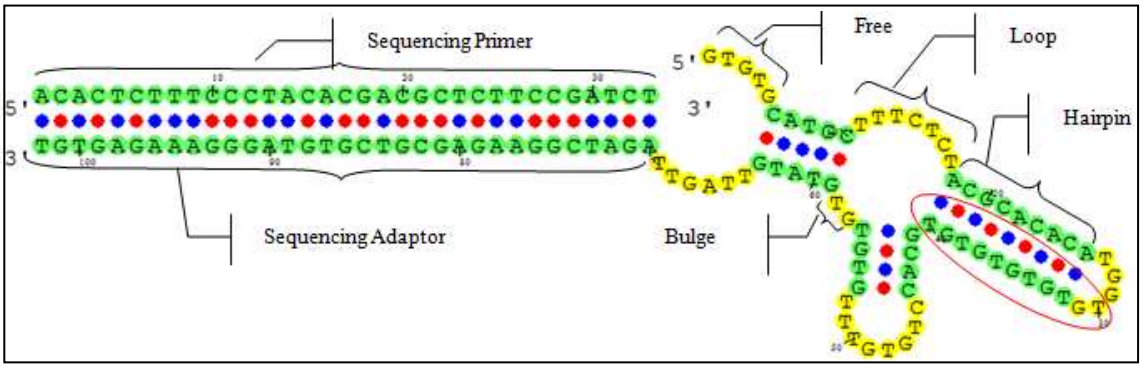


FIGURE 28: This is the heteroduplex structure generated for one of the probe-target used in this study. Four different structure types identified by OMP are marked as free, loop, bulge, and hairpin. The red oval indicates the location of deletion in this fragment.

TABLE 7: Conditions used for OMP modeling

Temperature	28 ºC
Monovalent concentration [Na^+]	~40mM
Divalent concentration	6.3mM
PH	7.5

4.3.1.2: Control Fragment Set Construction

For each deleted segment we identified up to 1001 sequences containing the deleted region in different sequence contexts that are present elsewhere on chromosome 1 that were successfully sequenced and then performed the same structural identification process which was carried out for TDEL fragments. The actual number found ranged from (1 to 1001) and we did not do an exhaustive search (past 1001). The process included these steps:

1) Construct a 70 base fragment centering on the deleted core.

Note: Throughout this study, these fragments are called Non deletion (NDEL) fragments.

2) Model the formation of the secondary structures using the same conditions and software.

3) Assign a code to the type of structure a base is involved in (described above).

4.3.1.3: Investigate the Likelihood That the Deleted Segment on TDEL Fragments Had a Structure Typical of the NDEL Group

To determine the likelihood that structure is associated with a deletion event, we stratified the TDEL fragments into 8 groups (Table 8) based on the involvement of their deletion cores in hairpin structure and used Fisher's exact test. For member of each group, we calculated two fractions: a/b and c/d.

For each group, a is 1, b is the total number of TDEL fragments, c is the number of NDEL fragments which satisfied the same conditions as the related TDEL fragments, and d is the total number of NDEL fragments in that group. For example. in Group one, defined as fragments having a hairpin structure of length 0-10 (Table 8) there are 1788 TDEL fragments. If for a given TDEL fragment in this group, we examined all NDELs (for the total found, up to 1001 sequences) and found that the deletion segment is involved in structure in between 0 and 10 bases for 200 of the 1001 reference fragments then the two fractions passed to Fisher's exact test would be 1/1788 (a/b) and 200/1001 (c/d). The Fisher's test used the above fractions and equation below to calculate the probability and p-value

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

TABLE 8: The percentage of deletion segments in a hairpin structure and the number of instances for each group.

Groups	% of Deletion segment in hairpin structure	# TDEL instances
1	0 to 10	1788
2	11 to 20	198
3	21 to 30	223
4	31 to 40	274
5	41 to 50	259
6	51 to 60	173
7	61 to 70	192
8	71 to 80	211
9	>=81	285

4.3.2: Part II: Train Predictive Models Using the Random Forest Algorithm Implemented in the Machine-Learning Environment WEKA.

There are a large number of machine-learning algorithms to select from in WEKA. All of them are well tested and widely accepted. Our choice was guided by the work of Hooghe and colleagues (152) who were looking for similar sequence/structure features that predict transcription factor binding sites. The authors provide guidelines for determining whether the random forest method is an appropriate choice, supporting our selection of it for these experiments..

4.3.2.1: Data Preparation for WEKA

To use WEKA, data must be in a single flat file format, where each data point is described by a fixed number of attributes, of which the last attribute is usually the class characteristic we desire to predict (in this case whether a segment is deleted or not).

4.3.2.2: Experiment 2-1: The Complete Deleted Fragment and Control Set

In this experiment, we wanted to test whether we could train a model that correctly classifies TDEL sequences given the complete data set. To perform this experiment we followed these steps:

1) We selected 1794 significant sequences among all the TDEL pool which have p-value <= 0.01 based on the result of the Fisher test obtained from the previous section.
2) From the NDEL pool we randomly selected 5 corresponding sequences. We note that the total NDEL pool for individual TDEL fragments varies: one TDEL fragment had (#NDEL<10), four TDEL fragments had (10< #NDEL< 100), 50 TDEL fragments had (10<#NDEL<100) and the remaining 1414 TDEL fragments all had (#NDEL≥ 1001) fragments.
3) We generated two data matrices:

   The extended data matrix, shown in Table 9, contains the following information

   a) At each position, the nucleotide present.
   b) At each position the type of structure predicted.
   c) A structural encoding of the sequence that is deleted in the TDEL group (the sequence is also present in the NDEL group, of course), independent of the location in the fragment.

   In Table 9 (a truncated version is shown below), the odd-numbered columns contain the nucleotide identity at the given location (using the 5’ to 3’ numbering convention for representing a single stranded nucleic acid), and even-numbered columns label the type of structure in which that nucleotide is predicted to

participate, with categories that include hydrogen bonds (H), loops (L), bulges (B), or (F) none of the above. The complete table includes 142 data columns and the class attribute column (whether or not a deletion was observed for this fragment).

TABLE 9: Sample of dataset using in WEKA generated based on the extended format.

id	1	2	3	4	5	6	7	…	135	136	137	138	hairpin	bulge	loop	free	deletion
TP-1	A	F	C	H	A	H	A	…	A	H	T	B	5	1	1	1	Yes
TN-1	T	F	A	F	G	F	C	…	G	F	A	F	5	3	0	0	No
TN-2	G	H	G	H	T	H	T	…	A	F	G	F	6	0	0	1	No
TN-3	A	F	G	F	G	F	G	…	A	F	G	F	6	0	0	0	No
TN-4	A	F	G	F	G	F	C	…	G	F	T	F	4	0	1	0	No
TN-5	T	F	C	F	T	F	T	…	G	F	A	F	3	0	0	0	No

In the second data matrix (Table 10), we generated a more condensed structural encoding as follows:

1) Each sequence was segmented into neighborhoods of contiguous nucleotides that are in the same type of structure; the number of such nucleotides per segment was counted. An average base content was calculated. That is, as shown in Table 10, each fragment is described by a string that includes:
    a) The number of bases involved in a specific structure and the structure label.
    b) The proportion (fraction relative to the length of the segment) of AT of the segment. Because Illumina sequencing chemistry is known to be less accurate in AT-rich regions (78,121,153), we wanted to retain some composition information without retaining the complete sequence string.

TABLE 10: Examples of condensed structure encoding for several fragments. The pound sign (#) indicates a number counting the nucleotides in a given segment and the fraction of AT.

Segments	Composition
1	#F#AT#H#B#AT#F#AT#H#B#AT#L#AT
2	#F#AT#H#B#AT#F#AT#H#B#AT#L#AT
3	#F#AT#H#B#AT#F#AT#H#B#AT#L#AT
4	#H#B#AT#F#AT#H#B#AT#F#AT#H#B#AT#L#AT
5	#H#B#AT#F#AT#H#B#AT#F#AT

2) Full structure and AT proportion encoding that provides a nucleotide-by-nucleotide description of the structure and the fraction of AT present at 5-nucleotide intervals, aligned to the TDEL fragments.

   a) Since this set is calibrated to structure in the TDEL fragments, some corresponding TN fragments (the abbreviation of NDEL used in the table) do not contain some structures, indicated by a 0.

   b) Since the deletions are centered but of different lengths we did not label the start and stop positions of the deleted segments.

TABLE 11: Sample of dataset using in WEKA generated based on the condensed format.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14		119	120	121	122	hairpin	bulge	loop	free	deletion
TP1	2	F	1	AT	3	H	1	B	1	AT	3	L	0	AT	...	1	F	1	AT	5	1	1	1	Yes
TN1	0	F	0	AT	0	H	0	B	0	AT	0	L	0	AT	...	6	F	0.5	AT	5	3	0	0	No
TN2	9	F	1	AT	2	H	0	B	1	AT	4	L	1	AT	...	0	F	0	AT	6	0	0	1	No
TN3	0	F	0	AT	1	H	0	B	0	AT	0	L	0	AT	...	10	F	0.8	AT	6	0	0	0	No
TN4	0	F	0	AT	0	H	0	B	0	AT	0	L	0	AT	...	4	F	1	AT	4	0	1	0	No
TN5	0	F	0	AT	0	H	0	B	0	AT	0	L	0	AT	...	15	F	0.7	AT	3	0	0	0	No

To clarify how these descriptions were determined and formatted, several examples are given in Figure 29.

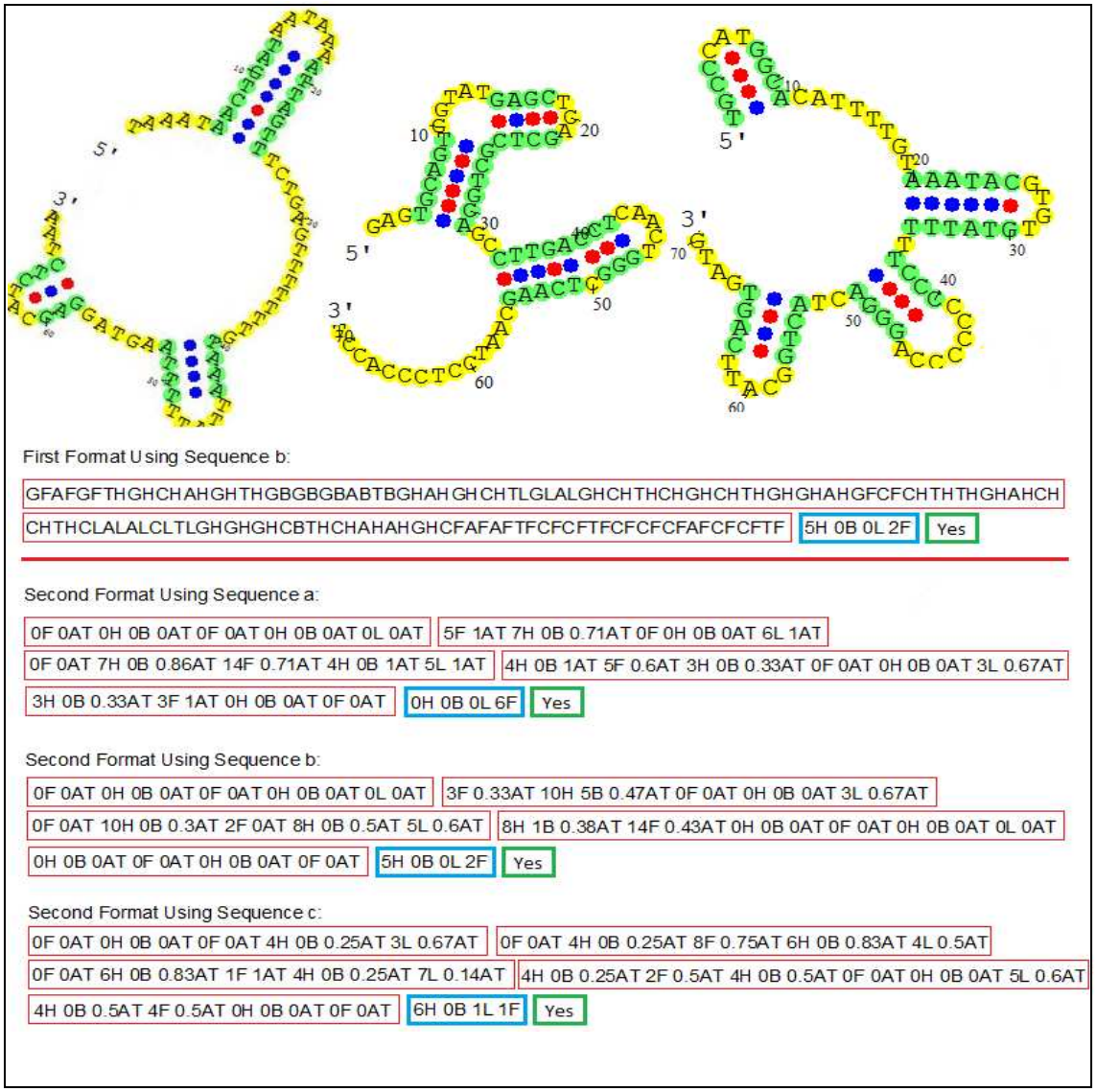


FIGURE 29: Three structures (a, b and c) and their respective data matrices in two formats described above. The extended format is shown for the shortest fragment (b) as the first example. Red blocks contain nucleotides and their structure assignment, blue blocks contain the structural composition of the deleted core ,green blocks show class attributes. The condensed format is shown for each of the three structures (a, b, and c). A red block indicates segments including 1)the number of consecutive bases involved in a given structure followed by 2) the AT-composition of each structure, blue blocks contain the structural composition of the deleted core and green blocks show the class attribute. Structures are denoted as described in the text.

4) We used a Random Forest classification algorithm in WEKA, with parameters maxDepth: 0 (for unlimited), numFeatures: all attributes, numTrees: 124, seed: 1 and cross-validation fold: 10, to predict the output of the last column.

4.3.2.3: Experiment 2-2: Structurally Stratified Deletion/Control Groups

As shown in the Results section (see Table 14), we were unable to train a model that successfully predicted the class attribute (presence or absence of a deletion) with high precision or sensitivity. Since it appeared that some structural motifs might be more significant than others we stratified the data into 8 subsets (summarized in Figure 30) as follows:

1) Using the deleted region as the reference points, indicate whether there is a hairpin structure within, to the right or to the left of the deleted segment (or any combination of these). Presence and absence are labeled as ‘0’ and ‘1’ respectively, and the order is left, right, center. So each fragment has a 3-numeral code of zeros and ones, and fragments are sorted into groups that share that label, which is also used as the group label for simplicity. For example, in Group 001 there are no structure regions on the right and left of the deleted segment (hence ‘00’) and there is a hairpin structure that encompasses the center of the TDEL fragment (thus the final ‘1’). Figure 30 gives an illustrative example for each of the 8 groups.

   Note: Group 000 is a special case having no stable structures, which is not useful for this study. We have 142 TDEL fragments in this group.

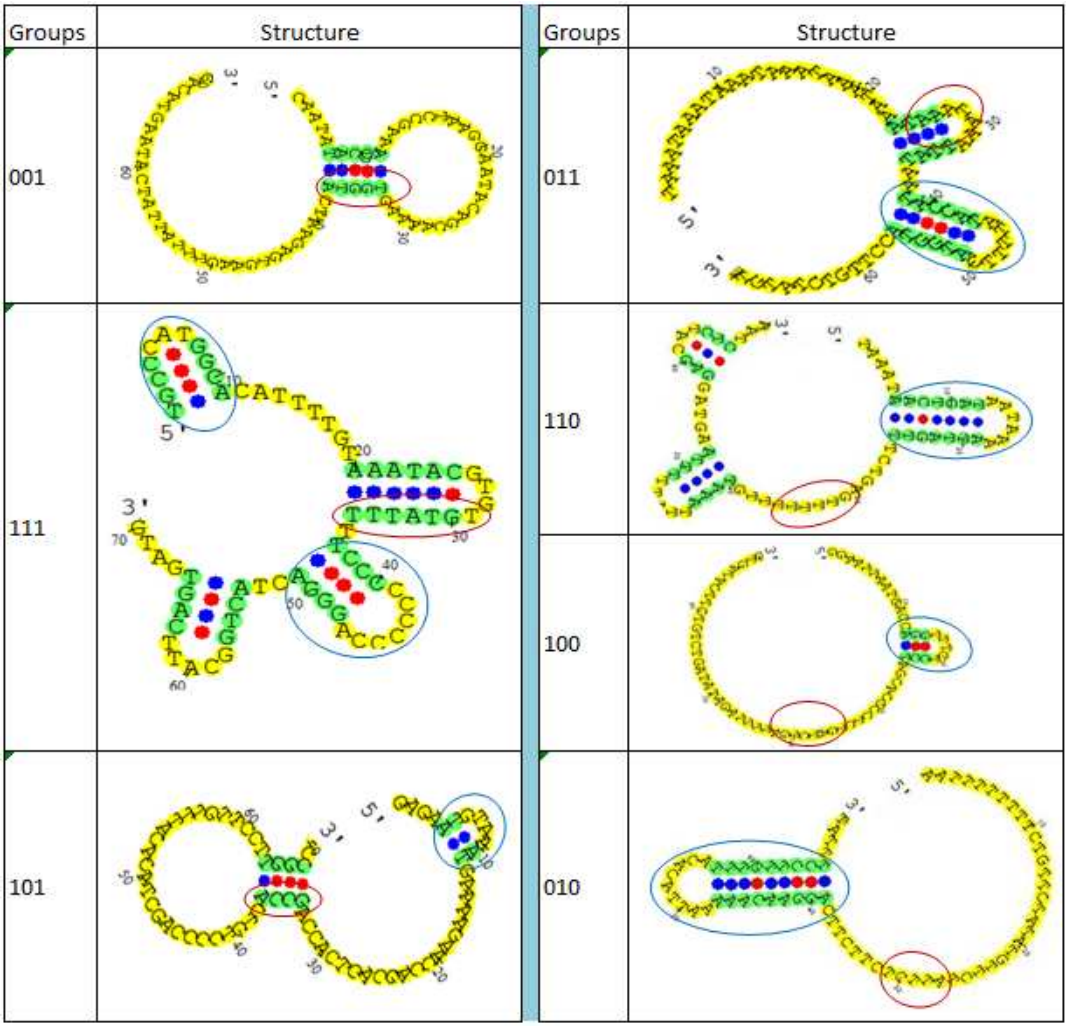


FIGURE 30: OMP predicted structures for 7 example sequences and their labels. In all of the sub- figures the deletion in the fragment is marked by a red oval and structures to the left and right of the deletion segment, where present, are marked by blue ovals. In group 001 there is at least one structure which encompasses the deletion segment. In group 111 there are structures to the left, right and in deletion segment. In group 101, there is structure to the left and also in deletion segment. In group 011, there is structure to the right and also deletion segment. In group 110 there is structure to the left and right of deletion segment but the deletion segment is free of structure. In group 100 there is structure to the left of the deletion segment and the deletion segment itself is free of structure. In group 010, there is structure to the right of the deletion segment but the deletion segment is free of structure.

2) We generated the same two types of data matrices for each group that are described above. Briefly, for each TDEL sequence we selected 5 NDEL sequences at random from the corresponding NDEL pool, but now the pool includes not just

the same deleted core but belongs to the same structural group. Taking Group 110 as an example, for each TDEL we have selected 5 NDEL sequences from the portion of its NDEL pool that includes the same structural elements (a hairpin to the right and left but not including the deleted segment). Summaries of the number of fragments in each pool are given in the Results.

3) We used the matrices as input to the Random Forest classification algorithm, with the following parameter values: maxDepth: 0 (for unlimited), numFeatures: all attributes, numTrees: 124, seed: 1 and cross-validation fold: 10. The class attribute was: (Deletion: Yes or No). Classification was performed using:
    a) Both sequence and structure features.
    b) Just structure features.
    c) Just sequence features.
4) We compared the classifications results obtained from both formats in terms of True Positive (TP), False Negative (FN), True Negative (TN), False Positive (FP), and receiver-operator curve (ROC)

4.3.2.4: Experiment 2-3: Balancing Group Sizes

Upon stratification, the distribution of structure neighborhoods that include that sequence deleted in our targeted fragments can be highly skewed. Recall that the exact deleted sequence is used to identify sequences in which the same nucleotides were successfully sequenced, the goal being to sample a large number of contexts for those nucleotides. Imposing a common structure filter on the available pool of NDELs results in very different sizes of the sets of non-deleted reference fragments for some of the deletion-containing sequences, varying from 5-200 sequences in 80% of the groups. This

is shown in Figure 31. For large sets, selecting 5 NDEL sequences per tree may not sample the distribution sufficiently, while for very small sets there may not be a large enough group to train on. We used several strategies to see how important this effect may be. Because inspection of the two sets of results of the stratification experiment (see Results) showed that the more condensed format yielded better classification outcomes than the full sequence + structure encoding format, we proceeded using just this format in the following experiments.

1) For each group we have iteratively generated datasets using random selection of 5 NDELs for each TDEL sequence, with replacement at each iteration, over 40 iterations.
2) We used the Random Forest classification algorithm with default settings (detailed above) to classify TDEL form NDEL sequences.
3) For each group, we averaged the results of the 40 trials to generate the output, which includes scores for the following: true positive, false positive, true negative, false negative, true positive precision, true negative precision and receiver-operator curve (ROC) rates.

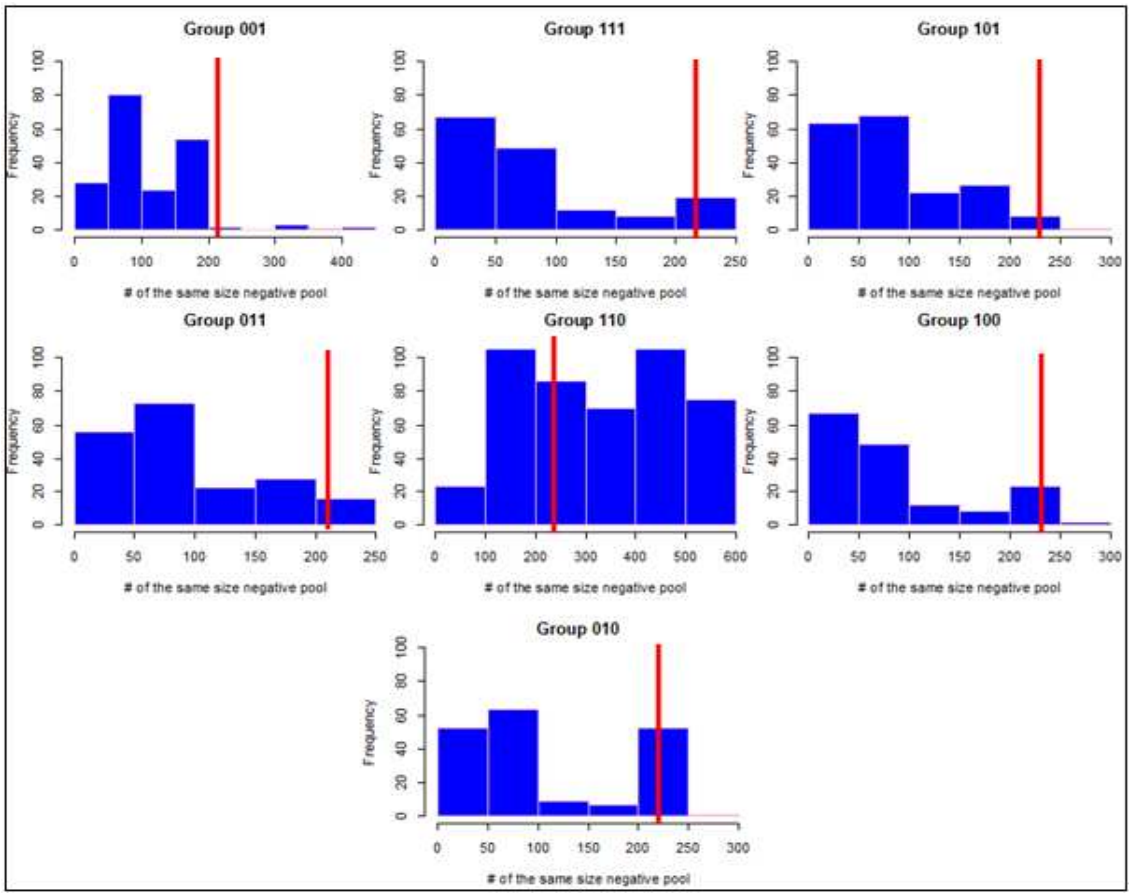


FIGURE 31: Distributions of negative pools for the seven structure groups. In all groups, except 110, the size of negative pools for >=80% of TDEL sequences is between 5 – 200 fragments (indicated by a red bar to indicate the disparity in different groups).

4.3.2.5: Experiment 2-4: Stability of the Structures

As shown in the Results, the classification performance of our models remained poor. We know that local structures vary in stability, and we did not use a cut-off to declare that a structure actually exists under the conditions present in a sequencing reaction: that is, a hairpin containing 2 bases was not discriminated from a hairpin with 6 bases. We know from designing PCR primers that polymerases are well able to melt less stable hairpins, and a rule of thumb in PCR primer design, whose reaction conditions are similar to those in sequencing, is to avoid primers that can form a hairpin in which more

than 6 bases can form hairpins are avoided. In case this creates sufficient noise to confound our models, we tested whether a ‘stability filter’ should be used, we carried out the following steps:

1) Re-stratify the 7 structural neighborhood groups to create a set of ‘stability bins’ for hairpin structures. The bins were selected to balance the size of each sub-group against the number in the base group (0-5 sided hairpins) as much as we could.

   Table 12 shows the resulting numbers once this operation has been performed, and also shows the stability of the groups we created: For example, group 001, which has only one structure at the center where the core deleted sequence is, was divided into 4 sub-groups comprised of hairpins with 1 - 5bp, 6 - 10bp, 11 - 15bp, and 16 – 30bp.

2) For each TDEL sequence we selected 5 NDEL sequences from the appropriate sub-group, randomly with replacement, over 10 iterations.

3) The Random Forest classification algorithm with default parameters (described above) was employed to classify sequences.

TABLE 12: This table indicates how we separated each group to sub-groups.

Groups	interval size of structure on left of core	interval size of structure on right of core	interval size of core	# of positive instances		Groups	interval size of structure on left of core	interval size of structure on right of core	interval size of core	# of positive instances
001	0	0	1-6	48		110	1-3	1-3	0	45
001	0	0	6-10	48		110	3-4	1-3	0	45
001	0	0	10-16	64		110	4-30	1-3	0	58
001	0	0	16-35	53		110	1-3	3-4	0	48
111	1-30	1-4	1-5	48		110	3-4	3-4	0	31
111	1-30	1-4	5-35	54		110	4-30	3-4	0	64
111	1-30	4-30	1-5	31		110	1-3	4-6	0	53
111	1-30	4-30	5-35	27		110	3-4	4-6	0	34
101	1-3	0	1-8	26		110	4-30	4-6	0	56
101	3-4	0	1-8	29		110	1-3	6-30	0	34
101	4-30	0	1-8	39		110	3-4	6-30	0	29
101	1-3	0	8-35	32		110	4-30	6-30	0	39
101	3-4	0	8-35	31		100	1-3	0	0	43
101	4-30	0	8-35	43		100	3-4	0	0	41
011	0	1-3	1-7	38		100	4-6	0	0	48
011	0	3-5	1-7	32		100	6-30	0	0	29
011	0	5-30	1-7	27		010	0	1-3	0	54
011	0	1-3	7-35	29		010	0	3-4	0	45
011	0	3-5	7-35	43		010	0	4-6	0	47
011	0	5-30	7-35	27		010	0	6-30	0	38

4.3.2.6: Experiment 2-5: Weighting the TDEL and NDEL Pools

We originally selected a very large number of control fragments (1000 times more in almost all cases). However the structure stratification process resulted in uneven distribution of those controls across the different groups and may have introduced bias, since we do not know if our groups are appropriate. As an alternative approach, we have pooled the NDEL sequences and weighted their contributions to account for the difference in the number of samples.

To generate a data matrix for each sub-group we followed these steps:

1) For each TDEL sequence, we sampled the entire NDEL pool in the dataset. For example; for Group 001 that has deletions of 1-6 nucleotides (Table 12), we had 48 TDEL sequences and 1418 NDEL sequences. Thus the data matrix will be for 1466 fragments (48+1418=1466).
2) Weight each instance by its relative contribution to the structure sub-group.
    a) Each TDEL fragment has the same weight, assigned as 1 over the total number of TDEL sequences in that group.
    b) Each NDEL fragment in a sub-group carries the same weight, assigned as the fraction of instances in the structural sub-group over the total number of NDELs. In each sub-group, the weight for all NDEL instances associated to a given TDEL instance is the same and calculated by dividing the size of NDEL pool by the summation of the sizes of all NDEL pools in that sub-group. Figure 32 illustrates our method for weighting TDEL and NDEL instances for sub-group 001.

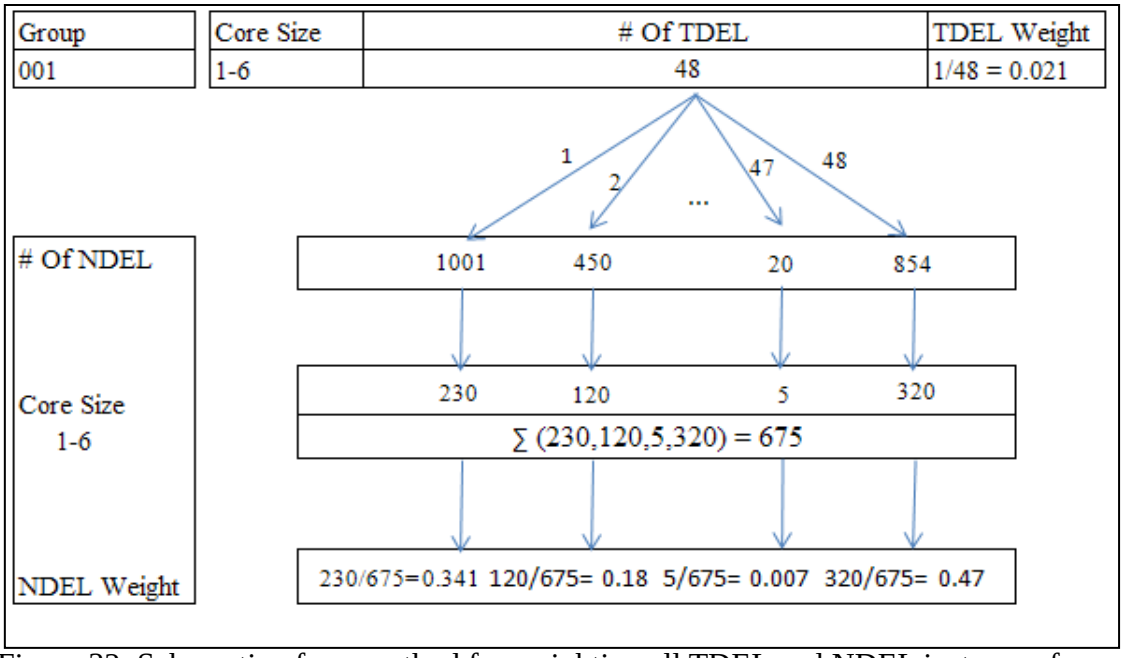


Figure 32: Schematic of our method for weighting all TDEL and NDEL instances for sub-group 001.

3) We used the Random Forest classification algorithm with default parameters (described above) to classify fragments as TDELs or NDELs.

4.3.3: Part III - Testing

After training the model on sequences showing deletions and sequences without those deletions found on chromosome 1, we used the model to test its ability to predict deletion-containing sequences found on chromosome 20. The steps are identical to those described for preparing the chromosome 1 datasets, briefly described below. The condensed format was used.

1) Construct 70 base length fragments, centering on the deleted segment, using the physical locations of all deletions greater >=3 nucleotides, as reported in the Kim et al. paper for chromosome 20 (262). These fragments are called true deletion (TDEL) fragments.

2) For each TDEL segment we identified up to 1001 identical sequences elsewhere on chromosome 20 that were successfully sequenced and then performed the same process which was carried out for TDEL fragments.

3) Construct WEKA data matrices, as described above, to perform following two experiments.

4.3.3.1: Experiment 3-1: Testing the Model on Sequences from Chromosome 20

The data set contained 262 TDEL sequences and 2 randomly selected sequences for each from the corresponding NDEL pool. We used the model constructed for Experiment 2-1 to reevaluate it using this test set.

4.3.3.2: Experiment 3-2: Testing the Model on Chromosome 20 Using Stratified Groups

We constructed our seven training data matrices, using the condensed format, and then used the corresponding models constructed in Experiment 2-2 to reevaluate the predictions using these test sets.

4.4: Results

4.4.1: Part I

4.4.1.1: Investigate the Likelihood That the Deleted Segment on TDEL Fragments Had a Structure Typical of the NDEL Group

Out of 3603 small deletions reported for chromosome 1, 1794 of them have P-value <= 0.01 which indicates the structures in which these deletion regions participated were not formed by chance. Table 13 indicates a representative portion of these results. The complete list of significant sequences can be found in Appendix II)

TABLE 13: Some representative sequences found to have significant structure associated with the deletions on chromosome 1. P-values were obtained using Fisher's exact test.

Target ID	Deletion Sequence	Deletion Length	P-value
target_chr1_novel_179479103	CGCGCGC	7	9.11E-170
target_chr1_rs28544222_59206042	TATATATAT	9	5.70E-148
target_chr1_novel_246849533	ATATATATA	9	6.69E-136
target_chr1_novel_21191432	ATATATA	7	1.81E-126
target_chr1_novel_241174531	GCGCGC	6	3.79E-85
target_chr1_novel_64861749	GCCTGTG	7	1.40E-80
target_chr1_novel_237210353	ATATAT	6	4.26E-58
target_chr1_novel_177611384	ATATAT	6	4.26E-58
target_chr1_novel_25056563	GGGGG	5	2.17E-50
target_chr1_novel_110868684	GGGGG	5	2.17E-50
target_chr1_novel_65367560	GGGGG	5	2.17E-50
target_chr1_novel_244921250	GGGGG	5	2.17E-50
target_chr1_novel_242453173	GGGGG	5	2.17E-50
target_chr1_novel_182018970	GGGGG	5	2.17E-50
target_chr1_novel_118894853	CATGC	5	5.60E-46
target_chr1_novel_233779552	CTGCT	5	6.69E-45
target_chr1_novel_156920258	AAAAAAAAAAA	11	2.20E-43
target_chr1_rs5775307_76030123	AAAAAAAAAAA	11	2.20E-43

4.4.2: Part II

4.4.2.1: Experiment 2-1: The Complete Deleted Fragment and Control Set

Results are summarized in Table 14. Examination of the true positive rate, ratio of true positive to false negative, and ratio of true negative to false positive all indicated that, regardless of the formats, using all of the TDEL sequences did not allow us to train the classifier algorithm to predict the class with precision or sensitivity.

TABLE 14: This table contains the results of classification for the expanded (First) and condensed (Second) formats. Columns indicates true positive (TP) rate, false positive (FP) rate, true negative (TN) rate, false negative (TN) rate, true positive precision, area under the receiver-operator curve (ROC), number of true positives over false negatives, and number of true negatives over false positives, from left to right.

	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area	TP/FN	TN/FP
Extended Format	0.135	0.022	0.978	0.865	0.549	0.793	223/1429	8077/183
Condensed Format	0.30	0.05	0.95	0.61	0.619	0.848	644/1008	7599/397

4.4.2.2: Experiment 2-2: Structurally Stratified Deletion/Control Groups

In Tables 15, 16, and 17 below, we show the results of classification for all seven groups, with two formats side by side. For each group, the tables consecutively shown results for a) both sequence and structure data, b) just structure data, and c) just sequence data.

TABLE 15: WEKA model accuracy results by group, using structure and sequence data.

Results for the Extended Format							Results for the Condensed Format					
Groups	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area
001	0.25	0.03	0.97	0.75	0.61	0.84	0.52	0.04	0.96	0.48	0.71	0.89
111	0.15	0.02	0.98	0.85	0.61	0.78	0.38	0.05	0.95	0.62	0.59	0.85
101	0.21	0.03	0.97	0.79	0.59	0.84	0.38	0.05	0.95	0.62	0.6	0.87
011	0.17	0.03	0.97	0.83	0.54	0.84	0.39	0.05	0.95	0.61	0.61	0.88
110	0.06	0.03	0.98	0.94	0.32	0.72	0.23	0.05	0.95	0.77	0.47	0.79
100	0.09	0.03	0.97	0.91	0.41	0.71	0.17	0.05	0.95	0.83	0.4	0.72
010	0.04	0.03	0.98	0.96	0.26	0.74	0.27	0.05	0.95	0.73	0.52	0.82

TABLE 16: WEKA model accuracy results by group, using structure data.

Results for the Extended Format							Results for the Condensed Format					
Groups	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area
001	0.3	0.05	0.95	0.7	0.53	0.81	0.51	0.05	0.95	0.48	0.67	0.87
111	0.13	0.05	0.95	0.87	0.34	0.77	0.38	0.06	0.94	0.62	0.57	0.84
101	0.24	0.05	0.95	0.76	0.49	0.81	0.41	0.06	0.94	0.59	0.58	0.84
011	0.16	0.06	0.94	0.84	0.34	0.8	0.36	0.06	0.94	0.64	0.54	0.84
110	0.09	0.04	0.96	0.91	0.29	0.65	0.2	0.04	0.96	0.8	0.5	0.74
100	0.12	0.08	0.92	0.88	0.23	0.62	0.14	0.08	0.93	0.86	0.28	0.63
010	0.12	0.09	0.91	0.88	0.21	0.62	0.3	0.07	0.93	0.7	0.45	0.72

TABLE 17: WEKA model accuracy results by group using sequence data.

Results for the Extended Format							Results for the Condensed Format					
Groups	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area
001	0.21	0.02	0.98	0.79	0.69	0.82	0.41	0.06	0.94	0.59	0.58	0.82
111	0.15	0.03	0.97	0.85	0.48	0.82	0.33	0.06	0.94	0.67	0.51	0.82
101	0.22	0.02	0.98	0.78	0.65	0.83	0.44	0.07	0.93	0.56	0.55	0.84
011	0.14	0.03	0.97	0.86	0.52	0.83	0.34	0.07	0.93	0.66	0.49	0.85
110	0.1	0.02	0.98	0.91	0.46	0.74	0.27	0.06	0.94	0.73	0.46	0.78
100	0.05	0.03	0.97	0.95	0.27	0.73	0.24	0.08	0.92	0.76	0.37	0.76
010	0.07	0.02	0.98	0.94	0.4	0.78	0.25	0.07	0.93	0.75	0.4	0.76

A number of trends can be observed in the above results. Because there are so many True Negatives, the True Positive and False Negative rates, along with the Precision, were the values we monitored most closely in assessing model strength.

1) Comparing the classification results for a complete data set and the stratified data sets (Tables 14, 15) , when both sequence and structure information is included in the data matrix, the condensed format always improved scores for TP, FN, Precision and ROC. Stratification gives similar or improved scores for those

metrics for Groups 001, 111, 101 and 011 (structures in which the deletion is in a helical region) in both formats and worse scores for those metrics for Groups 110, 100 and 010 (structures in which the deletion has neighboring helices) except for Precision in the condensed format which improves.

2) Across these experiments comparing the results from data matrices in which the sequence and structure, structure alone and sequence alone, the condensed format (Format 2) always gave a higher TP score, lower FN, equal or higher Precision and higher ROC than did the extended format (Format 1) when structural features were included (Tables 15 and 16). When only sequence was used to build the model, the TP and FN still improve with the condensed format, while for Precision and ROC the values are usually similar and in a few cases the extended format performed better (Table 17). We note that although the information encoded in the sequence-only experiment did not explicitly include structure, we were implicitly including structural information because groups were formed on that basis.
3) The best scores seen included a rate of 0.52 for the TP, 0.48 for FN, 0.83 for precision and 0.89 for the ROC, all in results for the combined sequence and structure data (Table 15). Including both sequence and structure information improved all of the results for Groups 001, 111 and 011 and improved the performance with respect to the Precision and ROC for 101, 110 and 010 and improved the Precision for 100. The model using structure gave the best TP and FN for 010 while the model using sequence gave the best TP and FN for Groups 110 and 100.

The models were better able to fit some groups than others. Group 001 had the best scores in all three data sets, with Groups 111, 101 and 011 having scores similar to each other but lower than for Group 001. Groups having helices adjacent to the deletion region (110, 100, 010) were modeled less accurately in all cases, and no particular model did best in all data sets for these groups.

4.4.2.3: Experiment 2-3: Balancing Group Sizes

In Table 18 we summarize the average of the results obtained after performing 40 iterations of the classification model for each group, using the data matrix containing both structure and sequence information and the condensed format. Comparing the results to those seen in Table 15, right side, we observe that for most groups and most metrics the values remain relatively unchanged.

TABLE 18: WEKA model results on the training data for all structural groups. Each number in this table is the average of 40 iterations across the data matrix used to train the model.

Groups	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area	TP/FN	TN/FP
001	0.5	0.03	0.97	0.51	0.77	0.91	108/105	1034/31
111	0.34	0.03	0.97	0.66	0.73	0.89	53/105	769/20
101	0.37	0.03	0.97	0.63	0.72	0.91	75/127	981/29
011	0.39	0.03	0.97	0.61	0.72	0.91	76/120	949/29
110	0.26	0.02	0.98	0.74	0.74	0.86	138/398	2627/49
100	0.19	0.04	0.96	0.81	0.52	0.8	33/130	776/29
010	0.27	0.03	0.97	0.73	0.64	0.86	50/134	889/28

4.4.2.4: Experiment 2-4: Stability of the Structures

We summarized the results of classification for all sub groups as in Table 19.In Table 19 we summarize the metrics obtained when helical structures are filtered for stability under sequencing reaction conditions. We note that this further decreases the size of the training sets – this can be seen by examining the TP/FN and TN/FP columns, which show the actual numbers of samples in each class. The following general trends can be observed.

1) Groups in which the deletion is part of a helical structure (001, 111, 101, and 011) show improved classification rates for the more stable structures when the additional level of stratification is applied.
2) In this experiment we attempted to keep sample groups of a similar size so models would be comparable: this is why helix lengths vary in the different classes shown in the table. In some groups increasing helix length corresponds to improved classifier results - for example in Group 111 there is improved classifier rates when a helix adjacent to a central helix exceeds a 4bp length: between the TP increases from 0.34 to 0.47 as the helices all become more stable, and the FN and Precision similarly improve in the series. For Group 001 the classifier improves up to a point, where the helical length is 10-16bp and then falls off slightly as the helix is even longer.

TABLE 19: WEKA fully stratified model results on the training data for all helical-stability subgroups within structural groups.

Groups	size of structure at the left side of core	size of structure at the right side of core	size of structure at core	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area	TP/FN	TN/FP
001	0	0	1-6	0.40	0.05	0.95	0.60	0.62	0.83	20/28	228/12
001	0	0	6-10	0.40	0.05	0.95	0.60	0.62	0.84	20/28	226/12
001	0	0	10-16	0.58	0.04	0.96	0.42	0.76	0.92	38/26	308/12
001	0	0	16-35	0.50	0.07	0.93	0.50	0.64	0.84	27/26	202/15
111	1-30	1-4	1-5	0.34	0.05	0.95	0.66	0.56	0.84	16/32	227/13
111	1-30	1-4	5-35	0.3	0.05	0.95	0.7	0.53	0.81	16/37	250/15
111	1-30	4-30	1-5	0.44	0.06	0.94	0.56	0.61	0.89	17/13	141/9
111	1-30	4-30	5-35	0.47	0.07	0.93	0.53	0.63	0.80	13/14	98/7
101	1-3	0	0-9	0.57	0.07	0.93	0.43	0.70	0.92	18/14	112/8
101	3-4	0	0-9	0.55	0.08	0.92	0.45	0.67	0.89	18/15	107/9
101	4-30	0	0-9	0.36	0.05	0.95	0.64	0.61	0.83	16/29	182/10
101	1-3	0	9-36	0.27	0.07	0.93	0.73	0.43	0.80	7/19	120/10
101	3-4	0	9-36	0.30	0.05	0.95	0.70	0.61	0.75	9/20	118/6
101	4-30	0	9-36	0.53	0.07	0.93	0.47	0.69	0.88	20/17	132/9
011	0	1-3	0-7	0.35	0.05	0.95	0.65	0.58	0.83	13/25	176/10
011	0	3-5	0-7	0.37	0.09	0.91	0.63	0.57	0.79	12/20	90/9
011	0	5-30	0-7	0.50	0.07	0.94	0.50	0.61	0.89	14/13	124/8
011	0	1-3	7-36	0.47	0.06	0.94	0.53	0.65	0.87	14/15	115/7
011	0	3-5	7-36	0.30	0.06	0.94	0.70	0.51	0.82	13/30	202/13
011	0	5-30	7-36	0.31	0.08	0.92	0.69	0.48	0.82	9/19	112/9
110	1-3	1-3	0	0.30	0.08	0.92	0.70	0.44	0.81	14/31	205/18
110	3-4	1-3	0	0.35	0.05	0.95	0.65	0.59	0.85	16/29	214/11
110	4-30	1-3	0	0.28	0.04	0.96	0.72	0.58	0.79	16/42	278/12
110	1-3	3-4	0	0.29	0.05	0.95	0.71	0.55	0.80	14/34	206/12
110	3-4	3-4	0	0.27	0.06	0.94	0.73	0.48	0.76	8/23	143/9
110	4-30	3-4	0	0.38	0.06	0.94	0.63	0.58	0.88	24/40	299/18
110	1-3	4-6	0	0.22	0.06	0.94	0.78	0.43	0.8	12/41	245/15
110	3-4	4-6	0	0.3	0.06	0.94	0.7	0.49	0.79	10/24	155/11
110	4-30	4-6	0	0.24	0.05	0.95	0.76	0.48	0.78	13/43	261/14
110	1-3	6-30	0	0.23	0.05	0.95	0.77	0.5	0.71	8/26	145/8
110	3-4	6-30	0	0.16	0.06	0.94	0.84	0.37	0.64	5/24	124/8
110	4-30	6-30	0	0.31	0.05	0.95	0.69	0.57	0.88	12/27	185/10
100	1-3	0	0	0.19	0.08	0.92	0.81	0.33	0.67	8/35	187/17
100	3-4	0	0	0.31	0.06	0.94	0.69	0.51	0.8	13/28	193/12
100	4-6	0	0	0.18	0.06	0.94	0.83	0.38	0.69	8/40	221/14
100	6-30	0	0	0.28	0.05	0.95	0.72	0.53	0.82	8/21	138/7
010	0	1-3	0	0.3	0.06	0.94	0.7	0.49	0.79	16/38	249/17
010	0	3-4	0	0.35	0.06	0.94	0.65	0.55	0.85	16/29	204/13
010	0	4-6	0	0.29	0.06	0.94	0.71	0.48	0.77	14/33	218/15
010	0	6-30	0	0.34	0.07	0.93	0.66	0.51	0.82	13/25	171/13

4.4.2.5: Experiment 2-5: Weighting the TDEL and NDEL Pools

A large improvement was seen with classifier scores when training fragments were weighted by frequency within the class. Table 20 summarizes these results. Some trends that can be observed follow.

1) For all groups, we found greater better model rates after adding weights.

2) The trends observed for helix stability influence conditions were preserved, as were the relative strength to discriminate particular structural groups.

TABLE 20: WEKA model results on the training data, with helix stability subgroups on the structural groups, weighted by fraction of TDELs and NDELs in total group.

Groups	size of structure at the left side of core	size of structure at the right side of core	size of structure at core	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area	TP/FN	TN/FP
001	0	0	6-10	0.905	0.005	0.995	0.095	0.975	0.997	43/5	236/1
001	0	0	10-16	0.939	0.006	0.995	0.061	0.971	0.997	60/4	318/2
001	0	0	16-35	0.898	0.012	0.988	0.102	0.951	0.989	48/5	214/3
111	1-30	1-4	1-5	0.877	0.010	0.990	0.123	0.945	0.991	42/6	237/3
111	1-30	1-4	5-35	0.864	0.008	0.992	0.136	0.959	0.993	46/7	263/2
111	1-30	4-30	1-5	0.834	0.017	0.983	0.167	0.926	0.979	23/5	103/2
111	1-30	4-30	5-35	0.913	0.015	0.985	0.087	0.927	0.992	27/3	148/2
101	1-3	0	1-8	0.939	0.027	0.974	0.061	0.905	0.990	24/2	96/3
101	3-4	0	1-8	0.876	0.029	0.971	0.124	0.900	0.987	25/4	96/3
101	4-30	0	1-8	0.823	0.010	0.991	0.177	0.954	0.986	32/7	170/2
101	1-3	0	8-35	0.866	0.020	0.981	0.135	0.900	0.984	28/4	157/3
101	3-4	0	8-35	0.864	0.010	0.990	0.137	0.954	0.990	29/5	138/1
101	4-30	0	8-35	0.886	0.009	0.991	0.114	0.963	0.991	38/5	165/2
011	0	1-3	1-7	0.850	0.008	0.993	0.150	0.959	0.993	32/6	184/1
011	0	3-5	1-7	0.841	0.026	0.974	0.159	0.915	0.979	27/5	96/3
011	0	5-30	1-7	0.841	0.019	0.981	0.159	0.903	0.987	23/4	129/3
011	0	1-3	7-35	0.801	0.021	0.979	0.199	0.904	0.981	23/6	119/3
011	0	3-5	7-35	0.828	0.010	0.990	0.172	0.945	0.989	36/7	213/2
011	0	5-30	7-35	0.797	0.015	0.985	0.204	0.925	0.986	22/6	118/2

110	1-3	1-3	0	0.881	0.013	0.987	0.119	0.932	0.991	40/5	220/3
110	3-4	1-3	0	0.862	0.005	0.995	0.138	0.970	0.994	39/6	224/1
110	4-30	1-3	0	0.859	0.007	0.994	0.141	0.963	0.992	50/8	288/2
110	1-3	3-4	0	0.861	0.009	0.991	0.139	0.954	0.992	41/7	216/2
110	3-4	3-4	0	0.855	0.001	0.999	0.145	0.993	0.991	27/5	152/0
110	4-30	3-4	0	0.919	0.009	0.991	0.081	0.956	0.996	59/5	313/3
110	1-3	4-6	0	0.838	0.011	0.989	0.162	0.939	0.990	44/9	257/3
110	3-4	4-6	0	0.844	0.005	0.995	0.156	0.969	0.989	29/5	165/1
110	4-30	4-6	0	0.873	0.008	0.992	0.127	0.957	0.994	49/7	273/2
110	1-3	4-30	0	0.788	0.010	0.990	0.212	0.948	0.985	27/7	150/2
110	3-4	4-30	0	0.759	0.010	0.990	0.241	0.941	0.981	22/7	133/1
110	4-30	4-30	0	0.867	0.014	0.986	0.133	0.928	0.992	34/5	192/3
100	1-3	0	0	0.877	0.008	0.993	0.123	0.962	0.990	38/5	201/2
100	3-4	0	0	0.885	0.010	0.991	0.115	0.951	0.993	36/5	203/2
100	4-6	0	0	0.844	0.001	0.999	0.156	0.993	0.994	41/8	235/0
100	4-30	0	0	0.862	0.011	0.990	0.138	0.946	0.990	25/4	144/2
010	0	1-3	0	0.922	0.005	0.995	0.078	0.975	0.996	50/4	265/1
010	0	3-4	0	0.933	0.008	0.992	0.067	0.962	0.997	42/3	215/2
010	0	4-6	0	0.891	0.004	0.996	0.109	0.979	0.990	42/5	232/1
010	0	4-30	0	0.926	0.006	0.994	0.074	0.969	0.996	35/3	183/1

4.4.3: Part III: Testing

4.4.3.1: Experiment 3-1: Testing with Chromosome 20 Sequences Against the Unstratified Model.

For 262 TDELs and corresponding but randomly selected NDELs (2 per TDEL) the ability of the unstratified and unweighted model to classify the samples was tested. The results are summarized in Table 21.

TABLE 21: WEKA un-stratified model for Chromosome 20 TDELs and corresponding NDELs.

TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area	TP/FN	TN/FP
0.21	0.02	0.981	0.79	0.846	0.622	55/207	508/10

Overall these results show a lower TP than the training set, a higher FN, a better Precision and lower ROC (see Table 14).

4.4.3.2: Experiment 3-2: Testing with Chromosome 20 Sequences Against the Stratified Model.

For the same set of TDEL and NDEL sequences derived from chromosome 20, the ability of the unweighted, stratified model to classify the samples was test. The results are summarized in Table 22.

TABLE 22: WEKA structure-stratified model results for Chromosome 20 TDELs and corresponding NDELS, by group.

Groups	TP Rate	FP Rate	TN Rate	FN Rate	Precision	ROC Area	TP/FN	TN/FP
001	0.12	0.08	0.92	0.88	0.42	0.53	5/38	79/7
111	0.07	0.05	0.95	0.93	0.42	0.56	5/71	144/7
101	0.10	0.06	0.94	0.90	0.44	0.58	4/37	77/5
011	0.11	0.14	0.29	0.86	0.89	0.46	6/47	91/15
110	0.09	0.06	0.43	0.94	0.91	0.49	3/29	59/4
100	0.17	0.08	0.50	0.92	0.83	0.65	1/5	11/1
010	0.10	0.10	0.33	0.90	0.90	0.50	1/9	18/2

In this case the classification of samples is considerably worse than the training set, except for the Precision for Groups 011 and 110.

The combined results suggest results suggest that the stratification model is over-trained on the chromosome I data. We note that the model correctly classified 55 out of 262 TDEL and 508 out of 518 NDEL sequences from the chromosome 20 test set. Data associated to these 55 True Positive (TP) and 207 false negative (FN) sequences (see Appendix III and IV) indicated:

1) Of the 55 fragments correctly classified, 45 belonged to Groups 110, 100 and 010 (Appendix III), for which our training set showed the poorest performance. This corresponds to fragments in which the deletion core is not part of a helix but for which there is a helix on at least one side.
2) All TDEL sequences for which the deleted core is composed only of thymidine were correctly classified.

4.5: Discussion

Training the model against chromosome 1 sequences in which we encoded sequence and structural context both explicitly per base and in a condensed format by structural region lead to improved training set predictoin rates, indicating that the condensed format derived model eliminated some noise in the data matrix. This was true even when only sequence information was encoded (see Tables 14 and 17). However, the sequence information may be too condensed, as in the condensed version we simply retained the fraction of AT per region based on a known limitation of the Illumina technology. We used only the condensed format in the further experiments.

Dividing the training set into groups based on the presence of a predicted helical structural element, in any of 3 locations on a fragment, led to improved performance on the training data for each of the 7 groups. The best results were obtained in the training set when a helix was present only in the center of the fragment, coinciding with the position of the deletion. Precision ranged from 0.59 – 0.83 across the groups and the ROC from 0.72 – 0.89. Using only structural information did not lead to the same improvements in either metric (see Table 16) so clearly there is some important

information in the sequence context as well. Sequence information alone (see Table 17) did not perform as well on either metric, indicating that it lacks important information.

When we attempted to further categorize the model by imposing a threshold for helix stability under sequencing conditions there did seem to be a trend to better performance in some structural groups: longer helices or adjacent long-enough helices improve performance on the training set. However, the trade-off in separating helix lengths by the need to retain sample groups of similar size for comparison became impossible to manage. To carry out this part of the study would require a much larger data set. In addition, some of the fragments are subject to variant structures of similar stability, and it is unclear how to handle that level of complexity in this type of model. We did try weighting the final, fully subdivided model to compensate for the small sample sizes, and this improves the model performance on the training set considerably, across all groups (see Table 20). We were not convinced that the final model was robust, so we began using the chromosome 20 test data by starting with the un-stratified and the stratified models (Tables 21 and 22). While the un-stratified model shows improved precision and similar ROC to the training data, the stratified model only shows improved precision in 4 of the groups and the ROC has a poorer score in all of the groups. To our surprise, the groups best classified in the un-stratified model (and those with improved precision in the stratified model) are those lacking a central helix, the opposite of what one would predict from the performance of the model on the training data. Some of the groups had very few members, and the chromosome 20 deletions were not screened first for statistical significance. Additional data will be necessary to pursue the stratified

model, ideally from additional chromosomes and individuals. While the un-stratified model has reasonable performance, the stratified model is likely over-trained.

What is the interpretation of these deletions? While we began this study under the influence of a technical error we had found in our lab, in fact many of these deletions are likely of biological origin. The original producers of the data sets demonstrated that a small number of the deletions could be verified in the sample. Thus we may be attempting to identify two separate mechanisms that do not have the same responses, both of which are important. To differentiate technical deletions from true biological deletions we need additional information. Lacking the genomic material or a budget to re-sequence a genome in multiple ways, what prospect is there for obtaining such information? We note that one of the challenges in the CAMDA 2013 contest is to infer the presence of structural variants including deletions but also copy number variants and to determine how they can be distinguished from systematic sequencing errors, with particular emphasis on the Korean genome. The contest organizers have provided genome sequencing data from 38 individuals who are part of the Korean Personal Genome Project (KPGP), among them there are genomic sequencing data for two twin pairs and one Caucasian female individual – the reason for inclusion being detection of systematic sequencing errors. That is, variants that appear only in one Korean sample should not be present in the sample from the Caucasian female, and variants that appear in one of a pair of twins should be present in the other, else these variants would be characterized as arising from sequencing or data preprocessing errors.

With respect to the structural features we included in the stratification scheme, it is possible that over-simplification of the structure has eliminated much of the signal.

Groups using the Random Forest strategy to identify transcription factor binding sites use near-crystallographic levels of resolution. There are known DNA structures not related to protein binding interactions that also required a high level of spatial resolution, including expansion of some DNA repeat sequences in the human genome, which underlie several human disorders (147). Most models for repeat expansion agree that expansion occurs through the formation of structures with B and non-B conformations (152-154). Having three-dimensional (spatial) information about these structures was essential in allowing researchers to understand the expansion mechanism. The structural information used for our classification was based on the OMP application, which predicts two-dimensional structure by modeling a thermodynamic minimum for a stable form, based on the calculated Gibbs Free Energy. The available structural motifs include: 1) Hairpin, 2) Bulge, 3) Loop, and 4) none of the above (Free) (see Figure 28) but does not include proximity, twist, roll and similar spatial values. By using a 3D structural prediction tool such as 3DNA, a given base pair can be classified across 16 parameters. Having this additional structural information may be required for us to improve the stratified model. Thus, while our simple model does have predictive value, additional data and more three-dimensional structural information are both needed to make significant improvements.

CHAPTER 5: CONCLUSIONS

5.1: Chapter 2

5.1.1: Hypothesis

In our first experiments we investigated whether the presence of helical structures adjacent to the probe-target duplex formation region affects the stability of the heteroduplex on the microarray surface, and thus might affect the interpretation of microarray results. In addition we investigated the utility of a number of biophysical properties and modeling methods in predicting the results that we did see.

5.1.2: Results

Our results show that secondary structures adjacent to the heteroduplex region in a probe bound to a microarray surface stabilizes the duplex, leading to a higher signal than is seen when the cognate target without such structures bound. This would be interpreted as an increased concentration of the target in the mixture. Since most microarray hybridizations add randomly sheared target, whose mean length is longer than the probe, there is the potential for considerable mis-interpretation of results. Available modeling tools do not take such structures into account. We were unable to identify a single thermodynamic property that correctly predicts the observed effect.

There are two possible explanations for the observed effect, not necessarily acting independently, discussed below.

5.1.3: Open Questions

1) More highly structured targets diffuse very slowly in hybridization solutions so they remain in proximity to the probe when they detach and thus are more likely to re-bind in a short amount of time.
2) The folded structure is more entropically favored when bound to the probe, since more solvent is excluded, and thus it has a more favorable binding constant than a simple heteroduplex. It is important to remember that the binding event occurs in three dimensions, so the duplex may fold in complex ways.

5.2: Chapter 3

5.2.1: Hypothesis

Having observed that helical structures adjacent to a heteroduplex affected the behavior of the microarray platform, we next tested whether such structures would affect the read-through fidelity of a polymerase on an HTS platform, in this case the Ion Torrent Personal Genome Machine (PGM).

5.2.2: Results

Our results demonstrate that there is a strong association between the site and length of a variety of base read errors and the location of secondary structures on a sequencing template. As a hairpin structure gets longer the sequencing reaction is subject to more mistakes, both as an increased rate of indels and as mis-incorporation errors. We controlled for a variety of known nucleotide composition sensitivities with this platform, such as tracts of homopolymer. The effect of structure should be considered as one source of sequencing errors.

5.2.3: Open Questions

We were only able to test the templates using the PGM, with validation on an ABI 3130 capillary gel system. Structure sensitivity may vary under conditions used with other HTS platforms, since sequencing conditions differ. The availability of a set of structured test constructs to test both chemistries and algorithms in every sequencing platform would greatly assist in determining what types of structures are likely to cause significant errors, and to develop sequencing conditions and chemistries that could overcome particular problems, similar to what was accomplished with the Sanger chemistry and capillary sequencing platforms in the past.

5.3: Chapter 4

5.3.1: Hypothesis

The availability of inexpensive HTS platforms has led to an explosion of available human genomes. The Thousand Genomes Project has been working progressively through a list of features, starting with single nucleotide polymorphisms and copy number variations. Structural variation, in the form on deletions and insertions has now become a focus, as evidenced by the current CAMDA 2013 competition, one of the main questions of which is to understand the presence of a large number of short deletions. These do not appear to affect the health of individuals, since none of the 38 genomes made available suffer from clinical symptoms of known genetic origin. Before the announcement of the competition we had become aware of the deletion rate, and had begun to study it from the perspective of secondary structure.

5.3.2: Results

Fragments twice the length of the sequencing reads and centered on the deletion were collected from chromosome I of the first Korean genome to be made available. These were matched to successfully sequenced fragments in which the deleted core was present in a different context. We modeled the structure of all fragments and trained a Random Forest model to classify fragments in this training set as either likely to contain a deletion or not. We next tested the model against similar fragments from chromosome 20 of the same genome, and achieved similar ROC rates between the test and training sets. Although the model does not classify fragments with high precision, we were able to show that including the context of both structural information and sequence composition greatly improved the performance of the model.

5.3.3: Open Questions

There are three elements that should be explored in continuing this research. The first has to do with the resolution of our structural model. We used a simple secondary structure encoding, but three-dimensional relationships may be required to resolve all of the necessary features. This would create 16 features per sequence rather than the 4 that we used, and will greatly expand the time and computational resources needed to carry out the modeling. Secondly, as we stratified the data set according to structural families, the size of each family became quite small, from hundreds of examples to tens. We concluded that in our most stratified models we had over-trained on the available sequences, and the next step should be to cull all of the genome for the deletion fragments, with the goal of sufficiently populating all downstream sub-groups. We would then require data from an additional genome for the test set, and the recently released

CAMDA competition data makes this possible. Finally, we are not able to discriminate the cause of the deletions in our data set: some are clearly biological while others are likely to arise from technical sources. These may require separate models, but first we need to clearly discriminate them The CAMDA data set includes one Caucasian genome and two genomes from identical twins, run by the same team on the same instruments and chemistry, which should allow discrimination of both types of deletion.

Structure is an implicit property of nucleic acids in solution, and is known to affect both technical assays and biological activities. Data modeling and analysis methods should always consider both immediate and neighboring structure when seeking to interpret measurements that use hybridization as part of the platform.

REFERENCES

1. Kim JI, J.Y., Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Lee S, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park WY, Kim H, Church GM, Lee C, Kingsmore SF, Seo JS. (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, 460(7258), 1011-1015.

2. Breslauer KJ, F.R., Blocker H, Marky LA. (1986) Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A*, 83, 3746–3750.

3. Gil Tae Hwang, Y.H.a.F.E.R. (2009) The effects of unnatural base pairs and mispairs on DNA duplex stability and solvation. *Nucleic Acids Research*, 14, 4757–4763.

4. De Costa NT, H.J. (2013) Evaluating the Effect of Ionic Strength on Duplex Stability for PNA Having Negatively or Positively Charged Side Chains. *PLoS ONE*, 8(3), e58670.

5. Williams, A.P., Longfellow, C. E., Freier, S. M., Kierzek, R., and Turner, D. H. (1989) Laser temperature-jump, spectroscopic, and thermodynamic study of salt effects on duplex formation by dGCATGC. *Biochemistry*, 28, 4283–4291.

6. Richard Owczarzy, B.G.M., Yong You, Mark A. Behlke, and Joseph A. Walder. (2008) Predicting Stability of DNA Duplexes in Solutions Containing Magnesium and Monovalent Cations. *Biochemistry*, 47, 5336–5353.

7. SantaLucia, J., Jr. (1998) A unified view of polymer, dumbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, 95, 1460–1465.

8. Allawi, H.T.a.S., J.,Jr. (1998) Nearest neighbor thermodynamic parameters for internal G□A mismatches in DNA. *Biochemistry,*, 37, 2170–2179.

9. Bommarito, S., Peyret,N. and SantaLucia,J.,Jr. (2000) Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res*, 28, 1929–1934.

10. Peyret, N. (2000), Prediction of Nucleic Acid Hybridization: Parameters and Algorithms PhD Dissertation, Wayne State University, Detroit, M. PhD.

11. Page, G.P., George, V., Go, R.C., Page, P.Z. and Allison, D.B. (2003) Are we there yet?’’: deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am J Hum Genet*, 73(4), 711-719.

12. Aguan K, C.J., Thompson LP, Weiner CP. (2000) Application of a functional genomics approach to identify differentially expressed genes in human myometrium during pregnancy and labour. *Mol Hum Reprod* 6, 1141-1145.

13. Cox, J.K.P.a.N.J. (2002) The allelic architecture of human disease genes: common disease – common variant...or not? *Human Molecular Genetics*, 11, 2417–2423.

14. Gibson, G. (2012) Rare and common variants: twenty arguments. *Nat Rev Genet.* , 13(2).

15. Jorrit J. Hornberg, F.J.B., Hans V. Westerhoff, Jan Lankelma. (2006) Cancer: A Systems Biology disease. *BioSystem*, 83, 81-90.

16. Yang L, G.S., Li Y, Zhou S, Tao S. (2011) Protein microarrays for systems biology. *Acta Biochim Biophys Sin (Shanghai).* 43(3), 161-171.

17. Ideker T, G.T.a.H.L. (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2, 343–372.

18. Singh, M.W.a.A.K. (2011) Single-cell protein analysis. *Current Opinion in Biotechnology*, 23, 1-6.

19. Rod K. Nibbe, M.K., Mark R. Chance. (2010) An Integrative -omics Approach to Identify Functional Sub-Networks in Human Colorectal Cancer. *PLoS Comput Biol*, 6(1), e1000639.

20. Troy Hawkins, M.C.a.D.K. (2008) New paradigm in protein function prediction for large scale omics analysis. *Molecular BioSystems*, 4, 223–231.

21. Wang YC, H.S., Lan CY, Chen BS. (2012) Prediction of phenotype-associated genes via a cellular network approach: a Candida albicans infection case study. *PLoS One*, 7(4), e35339.

22. Novelli G, P.I., Mango R, Romeo F, Mehta JL. (2010;), *World J. Cardiol*. 436 ed, Vol. 2, pp. 428-436.

23. Kevin J Thompson†, H.D., Jeffrey L Solka and Jennifer W Weller. (2009) A white-box approach to microarray probe response characterization: the BaFL pipeline. *BMC Bioinformatics*, 10, 449.

24. McClintick JN, E.H. (2006) Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinformatics*, 7, 49.

25. Konings P, V.E., Jackmaert S, Ampe M, Verbeke G, Moreau Y, Vermeesch JR, Voet T. (2012) Microarray analysis of copy number variation in single cells. *Nat Protoc*, 7(2), 281-310.

26. de Leeuw N, H.-K.J., Simons A, Geurts van Kessel A, Smeets DF, Faas BH, Pfundt R. (2011) SNP array analysis in constitutional and cancer genome diagnostics--copy number variants, genotyping and quality control. *Cytogenet Genome Res*, 135(3-4), 212-221.

27. Huang J, W.W., Zhang J, Liu G, Bignell GR, Stratton MR, Futreal PA, Wooster R, Jones KW, Shapero MH. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics.*, 1(4), 287-299.

28. Xiang Guo, Q.-R.C., Young K Song, Jun S Wei and Javed Khan. (2011) Exon array analysis reveals neuroblastoma tumors have distinct alternative splicing patterns according to stage and MYCN amplification status. *BMC Medical Genomic*, 4, 35.

29. Downward, A.S.a.J. (2001) Navigating gene expression using microarrays — a technology review. *Nat Cell Biol*, 3(8), E190-195.

30. Lee NH, S.A. (2007) Microarrays: an overview. *Methods Mol Biol.*, 353, 265-300.

31. Schena M, S.D., Davis RW, Brown PO. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 647-670.

32. Alizadeh AA, E.M., Davis RE, Ma C, Lossos IS, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403, 503-511.

33. Bethin KE, N.Y., Sladek R, Asada M, Sadovsky Y, Hudson TJ, et al. (2003) Microarray analysis of uterine gene expression in mouse and human pregnancy. *Mol Endocrinol* 17, 1454-1469.

34. Hoshida Y, V.A., Kobayashi M, Peix J, Chiang DY, et al. (2008) Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *N Engl J Med*, 359, 1995–2004.

35. David W Craig, M.J.H., Diane Hu-Lince, Victoria L Zismann, Michael C Kruer, Anne M Lee, Erik G Puffenberger, John M Pearson, Dietrich A Stephan. (2005) Identification of disease causing loci using an array-based genotyping approach on pooled DNA. *BMC Genomics*, 6, 138.

36. Szelinger S, P.J., Craig DW. (2011) Microarray-based genome-wide association studies using pooled DNA. *Methods Mol Bio*, 700, 49-60.

37. MiddletonFA, P., Gentile KL, Morley CP, Zhao X, Eisner A, Brow A, Petryshen TL, Kirby AN, Medeiros H, Carvalho C, Macedo A, Dourado A, Coelho I, Valente J, Soares MJ, Ferreira CP, Lei M, , Azevedo MH, Kennedy JL, Daly MJ, Sklar P, Pato CN. (2004) Genome-wide linkage analysis of bipolar disorder using high density single nucleotide polymorphism (snp) genotyping arrays: A comparison

with microsatellite markers and finding of significant linkage to chromosome 6q22. *Am J Hum Genet*, 74(5), 886-897.

38. Pato CN, M.F., Gentile KL, Morley CP, Medeiros HM, Macedo A, Azevedo MH, Pato MT. (2005) Genetic linkage to chromosome 6q22 is a consistent finding in Portuguese subpopulations and may generalize to broader populations. *Am J Med Genet B Neuropsychiatr Genet*, 134, 119-121.

39. Taylor RC, S.M., Weller J, Khoshnevis S, Shi L, McDermott J. (2009) A network inference workflow applied to virulence-related processes in Salmonella typhimurium. *Ann N Y Acad Sci.*, 1158, 143-158.

40. Tan PK, D.T., Spitznagel EL, Xu P, Fu D, et al. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*, 31, 5676–5684.

41. Michiels S, K.S., Hill C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365, 488–492.

42. Ein-Dor L, K.I., Getz G, Givol D, Domany E. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21, 171–178.

43. Heidenblad M, L.D., Veltman JA, Jonson T, Mahlamaki EH, et al. (2005) Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene*, 24, 1794–1801.

44. Yan H, Y.W., Velculescu VE, Vogelstein B, Kinzler KW. (2002) Allelic variation in human gene expression. *Science*, 297, 1143.

45. Kerr MK, C.G. (2001) Experimental design for gene expression microarrays. *Biostatistics*, 2, 183–201.

46. Kerr MK, C.G. (2003) Design considerations for efficient and effective microarray studies. *Biometrics*, 59, 822–828.

47. Vinciotti V, K.R., D'Alimonte D, Liu X, Cattini N, et al. (2005) An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics*, 21, 492–501.

48. ES., K. (2006) The End of the Microarray Tower of Babel: Will Universal Standards Lead the Way? *J Biomol Tech*, 17(3), 200-206.

49. Leek JT, S.R., Bravo HC, Simcha D, Langmead B, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11, 733–739.

50. Liang M, B.A., Rute E, Greene AS, Cowley AW. (2003) Quantitative assessment of the importance of dye switching and biological replication in cDNA microarray studies. *Physiol Genomics*, 14, 199–207.

51. Pozhitkov AE, B.I., Brouwer MH, Noble PA. (2010) Beyond Affymetrix arrays: expanding the set of known hybridization isotherms and observing pre-wash signal intensities. *Nucleic Acids Res*, 38(5), e28.

52. Larkin JE, F.B., Gavras H, Sultana R, Quackenbush J. (2005) Independence and reproducibility across microarray platforms. *Nat Methods*, 2, 337–344.

53. Irizarry RA, W.D., Spencer F, Kim IF, Biswal S, et al. (2005) Multiplelaboratory comparison of microarray platforms. *Nat Methods*, 2, 345–350.

54. Shi L, R.L., Jones WD, Shippy R, Warrington JA, et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24, 1151–1161.

55. Severgnini M, B.S., Mangano E, Scarlatti F, Mezzelani A, et a. (2006) Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment. *Anal Biochem*, 353, 43–56.

56. Jeffery IB, H.D., Culhane AC. (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7, 359.

57. Jeanmougin M, d.R.A., Marisa L, Paccard C, Nuel G, et al. (2010) Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PLoS ONE*, 5, e12336.

58. Anthony RM, S.A., Chan AB, Boender PJ, Klatser PR, Oskam L. (2003) Effect of secondary structure on single nucleotide polymorphism detection with a porous microarray matrix; implications for probe selection. *Biotechniques*, 34(5), 1082-1086.

59. Yang Gao, L.K.W.a.R.M.G. (2006) Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nucleic Acids Research*, 34, 11.

60. Ratushna VG, W.J., Gibas CJ. (2005) Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genomics*, 6:31.

61. Yilmaz, L.S., and D. R. Noguera. (2004) Mechanistic approach to the problem of hybridization efficiency in fluorescent in situ hybridization. *Appl.Environ. Microbiol*, 70, 7126–7139.

62. Allen Day, M.R.C., Jun Dong, Brian D O'Connor and Stanley F Nelson. (2007) Celsius: a community resource for Affymetrix microarray data. *Genome Biology* 8:R112.

63. Sunitha Kogenaru, Q.Y., Yinping Guo and Nian Wang. (2012) RNA-seq and microarray complement each other in transcriptome profiling. *BMC Genomics*, 13:629.

64. Marino Marinković, W.C.d.L., Mark de Jong, Michiel H S Kraak, Wim Admiraal, Timo M Breit, Martijs J Jonker. (2012) Combining Next-Generation Sequencing and Microarray Technology into a Transcriptomics Approach for the Non-Model Organism Chironomus riparius. *PloS one*, 10, e48096.

65. Alberto Magi, M.B., Alessia Gozzini, Francesca Girolami, Francesca Torricelli and Maria Luisa Brandi. (2010) Bioinformatics for Next Generation Sequencing Data. *Genes*, 1, 294-307.

66. Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. *New Biotechnology*, 25, 4.

67. Voelkerding KV, D.S., Durtschi JD. (2009) Next-generation sequencing: from basic research to diagnostics. *Clin Chem*, 55(4).

68. Liu L, L.Y., Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. (2012) Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012, 251364.

69. Medvedev P, S.M., Brudno M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, S13-20.

70. Susmita Datta, S.D., Seongho Kim, Sutirtha Chakraborty, and Ryan S. Gill. (2010) Statistical Analyses of Next Generation Sequence Data: A Partial Overview. *J Proteomics Bioinformatics*, 3(6), 183–190.

71. Schadt EE, T.S., Kasarskis A. (2010) A window into third-generation sequencing. *Hum Mol Genet*, 19(R2), R227-240.

72. Atlasi Y, M.S., Ziaee SA, Gokhale PJ, Andrews PW. (2008) OCT4 spliced variants are differentially expressed in human pluripotent and nonpluripotent cells. *Stem Cells*, 26(12), 3068-3074.

73. Buckland, P.R. (2004) Allele-specific gene expression differences in humans. *Human Molecular Genetics*, 13, R255–R260.

74. John Archer, G.B., Simon J Watson, Paul Kellam, Andrew Rambaut and David L Robertson. (2012) Analysis of high-depth sequence data for studying viral diversity:

a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics*, 13, 47.

75. Jiang Du, R.D.B., Zhengdong D. Zhang, Yong Kong, Michael Snyder, Mark B.Gerstein. (2009) Integrating Sequencing Technologies in Personal Genomics: Optimal Low Cost Reconstruction of Structural Variants. *PLoS Computational Biology*, 5, e1000432.

76. Kasper D. Hansen , S.E.B.a.S.D. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* 38(12), e131.

77. Fanglei Zhuang, R.T.F., Zhiyi Sun, Yu Zheng and G. Brett Robb. (2012) Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Research*, 40, No 7, e54.

78. Daniel Aird, M.G.R., Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, Andreas Gnirke. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2), R18.

79. Sanger F, C.A. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Mol. Biol*, 94(3), 441-448.

80. guide, C. DNA Sequencing by Capillary Electrophoresis. *Applied Biosystems*.

81. Ledergerber C, D.C. (2011) Base-calling for next-generation sequencing platforms. *Brief Bioinform*, 12(5), 489-497.

82. Dohm JC, L.C., Borodina T, Himmelbauer H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, 36, e105.

83. Bravo HC, I.R. (2009) Model-Based Quality Assessment and Base-Calling for Second-Generation Sequencing Dat. *Biometrics*, 66(3), 665-674.

84. Oshlack A, W.M. (2009) Transcript length bias in RNA-sequencing data confounds systems biology. *Biol Direct*, 4:14.

85. Stellwagen E, A.A., Dong Q, Stellwagen NC. (2007) Electrophoretic mobility is a reporter of hairpin structure in single-stranded DNA oligomers. *Biochemistry*, 46(38), 10931-10941.

86. Honggang Wang, W.S., Zhu Li, Xiufang Wang, and Zhanjun Lv. (2011) Identification and characterization of two critical sequences in SV40PolyA that

activate the green fluorescent protein reporter gene. *Genet Mol Biol.*, 34(3), 396–405.

87. Lin Liu, Y.L., Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. (2012) Comparison of Next-Generation Sequencing Systems. *Biomedicine and Biotechnology*, 2012, 11.

88. Woollard PM, M.N., Vamathevan JJ, Van Horn S, Bonde BK, Dow DJ. (2011) The application of next-generation sequencing technologies to drug discovery and development. *Drug Discov Today*, 512-519.

89. Martin Shumway, G.C., and Hideaki Sugawara. (2010) Archiving next generation sequencing data. *Nucleic Acids Res*, 38, D870–D871.

90. Mullaney JM, M.R., Pittard WS, Devine SE. (2010) Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet*, 19(R2), R131-136.

91. Mills RE, L.C., Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*, 16(9), 1182-1190.

92. Santosh Kumar, T.W.B., and Sylvie Cloutier. (2012) SNP Discovery through Next-Generation Sequencing and Its Applications. *International Journal of Plant Genomics*, 15 pages.

93. Michael A Quail, M.S., Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow and Yong Gu. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 341.

94. Wen-Tso Liu, H.G., and Jer-Horng Wu. (2006) Effects of Target Length on the Hybridization Efficiency and Specificity of rRNA-Based Oligonucleotide Microarrays. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, 73–82.

95. Mehlmann, M., M. B. Townsend, R. L. Stears, R. D. Kuchta, and K. L.Rowlen. (2005) Optimization of fragmentation conditions for microarray analysis of viral RNA. *Anal. Biochem*, 347, 316–323.

96. Lima, W.F., B. P. Monia, D. J. Ecker, and S. M. Freier. (1992) Implication of RNA structure on antisense oligonucleotide hybridization kinetics. *Biochemistry*, 31, 12055–12061.

97. Southern, E., K. Mir, and M. Shchepinov. (1999) Molecular interactions on microarrays. *Nat. Genet*, 21, 5–9.

98. Armitage, B.A. (2003) The impact of nucleic acid secondary structure on PNA hybridization. *Drug Discov Today*, 8, 222–228.

99. Christian Trapp, M.S.a.A.O. (2011) Stability of double-stranded oligonucleotide DNA with a bulged loop: a microarray study. *BMC Biophysics*, 4:20.

100. Farhat N. Memon, A.M.O., Olivia Sanchez-Graillet, Graham J.G. Upton and Andrew P. Harrison. (2010) Identifying the impact of G-Quadruplexes on Affymetrix 3′ Arrays using Cloud Computing. *Integrative Bioinformatics*, 7(2):111.

101. Harrison, A.P., Johnston,C.E. and Orengo,C.A. (2007) Establishing a major cause of discrepancy in the calibration of Affymetrix GeneChips. *BMC Bioinformatics*, 8.

102. Hinanit Koltai1, a.C.W.-B. (2008) Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. *Nucl. Acids Res*, 36 (7), 2395-2405.

103. Draghici, S., Khatri,P., Eklund,A.C. and Szallasi,Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*, 22, 101–109.

104. Jon Lapham, J.P.R., Peter B. Moore and Donald M. Crothers. (1997) Measurement of diffusion constants for nucleic acids by NMR. *Biomolecular NMR*, 10, 255–262.

105. Nguyen, H.K., and E. M. Southern. (2000) Minimising the secondary structure of DNA targets by incorporation of a modified deoxynucleoside: implications for nucleic acid analysis by hybridisation. *Nucleic Acids Res*, 28, 3904–3909.

106. DNA Software from (DNA Software; http://www.DNAsofware.com)

107. Fish DJ, H.M., Brewood GP, Goodarzi JP, Alemayehu S, Bhandiwad A, Searles RP, Benight AS. (2007) DNA multiplex hybridization on microarrays and thermodynamic stability in solution: a direct comparison. *Nucleic Acids Res*, 35(21), 7197-7208.

108. Xia XQ, J.Z., Porwollik S, Long F, Hoemme C, Ye K, Müller-Tidow C, McClelland M, Wang Y. (2010) Evaluating oligonucleotide properties for DNA microarray probe design. *Nucleic Acids Res*, 38(11), e121.

109. Chizhikov V, W.M., Ivshina A, Hoshino Y, Kapikian AZ, Chumakov K. (2002) Detection and genotyping of human group A rotaviruses by oligonucleotide microarray hybridization. *J Clin Microbiol.*, 40(7), 2398-2407.

110. Lee I, A.S., Chen H, Maruyama A, Wang N, McInnis MG, Athey BD. (2008) Discriminating single-base difference miRNA expressions using microarray Probe Design Guru (ProDeG). *Nucleic Acids Res*, 36(5), e27.

111. Holowachuk, E.W.a.R., M.S. (1995) Efficient gene synthesis by Klenow assembly/extension-Pfu polymerase amplification (KAPPA) of overlapping oligonucleotides. *Genome Research*, 4, 299-302.

112. Czar, M., Anderson, C., Bader, J. and Peccoud, J. (2009) Gene synthesis demystified. *Trends in biotechnology*, 27, 63-72.

113. Alex E. Pozhitkov, I.B., Marius H. Brouwer and Peter A. Noble. (2009) Beyond Affymetrix arrays: expanding the set of known hybridization isotherms and observing pre-wash signal intensities. *Nucleic Acids Research*, 38, e28.

114. McDowell DG, B.N., Parkes HC. (1998) Localised sequence regions possessing high melting temperatures prevent the amplification of a DNA mimic in competitive PCR. *Nucleic Acids Res*, 26.

115. Viswanathan VK, K.K., Cianciotto NP. (1999) Template secondary structure promotes polymerase jumping during PCR amplification. *BioTechniques*, 27, 508 – 511.

116. Boynton, G.L.M.a.K.A. (1995) PCR bias in amplification of androgen receptor alleles a trinucleotide repeat marker used in clonality studies. *Nucl. Acids Res*, 23, 1411–1418.

117. Frey, U.H., Bachmann, H.S. (2008) "PCR-amplification of GC-rich regions: 'slowdown PCR'. *Nat Protoc*, 3(8).

118. Kieleczawa, J. (2006) Fundamentals of sequencing of difficult templates--an overview. *Biomol Tech* 17(3), 207-217.

119. Nicholas J Loman, R.V.M., Timothy J Dallman, Chrystala Constantinidou, Saheer E Gharbia, John Wain & Mark J Pallen. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*, 30(5), 434-439.

120. Susan M Huse, J.A.H., Hilary G Morrison, Mitchell L Sogin and David Mark Welch. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8:R143.

121. Nakamura K, O.T., Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, 39(13): e90.

122. Oyola SO, O.T., Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP, Quail MA. (2012) Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics.*

123. Kuznetsov, S., Ren, C.-C., Woodson, S. and Ansari, A. (2008) Loop dependence of the stability and dynamics of nucleic acid hairpins. *Nucleic acids research*, 36, 1098-1112.

124. Cirulli, E., Singh, A., Shianna, K., Ge, D., Smith, J., Maia, J., Heinzen, E., Goedert, J., Goldstein, D. and the Center for, H.I.V. (2010) Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biology*, 11, R57.

125. Simpson JT, W.K., Jackman SD, Schein JE, Jones SJM, Birol I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19(6), 1117-1123.

126. Illumina. (2009). Technical note: De Novo Assembly Using Illumina Reads.

127. Martin Kircher, P.H.a.J.K. (2011) Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*, 12, 382.

128. DePristo, M.A.e.a. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*, 43, 491–498.

129. Sasson, A.M., T. P. (2010) Filtering error from SOLiD Output. *Bioinformatics*, 26, 849–850.

130. Zhen Xuan Yeo, M.C., Yoon Sim Yap, Peter Ang, Steve Rozen, Ann Siew Gek Lee. (2012) Improving Indel Detection Specificity of the Ion Torrent PGM Benchtop Sequencer. *PLoS ONE*, 7(9), e45798.

131. Lauren M. Bragg, G.S., Margaret K. Butler, Philip Hugenholtz, Gene W. Tyson. (2013) Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comput Biol*, 9(4), e1003031.

132. Montoya-Burgos, Y.S.-G.a.J.I. (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res*, 20(10), 1432-1440.

133. Medvedev P, B.M. (2009) Maximum likelihood genome assembly. *J Comput Biol*, 16(8), 1101-1116.

134. Landan, G.G., D. (2009) Characterization of pairwise and multiple sequence alignment errors. *Genes*, 441, 141–147.

135. Frith, M.C., Hamada, M. & Horton, P. (2010) Parameters for accurate genome alignment. *BMC Bioinformatics*, 11, 80.

136. Rice, P.L., I. and Bleasby,A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16, 276-277.

137. Jared T. Simpson, K.W., Shaun D. Jackman. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Res*, 19, 1117-1123.

138. Junhua Zhao, A.B., Guliang Wang, and Karen M. Vasquez. (2009) Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci*, 67(1), 43-62.

139. Idury, R.M., Waterman, M.S. (1995) A new algorithm for DNA sequence assembly. *Computational Biology*, 2(2), 291-306.

140. Martin, J., Bruno, V., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M. and Wang, Z. (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, 11, 663.

141. Surget-Groba, Y.a.M.-B., J. (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome research*, 20, 1432-1440.

142. Crawford, J., Guelbeogo, W., Sanou, A., Traoré, A., Vernick, K., Sagnon, N.F. and Lazzaro, B. (2010) De Novo Transcriptome Sequencing in Anopheles funestus Using Illumina RNA-Seq Technology. *PLoS ONE*, 5, e14202.

143. Sung-Min Ahn, T.-H.K., Sunghoon Lee. (2009) The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res*.

144. Mirkin, S.M. (2007) Expandable DNA repeats and human disease. *Nature*, 447, 932–940.

145. McMurray, C.T. (2010) Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet*, 11, 786–799.

146. Gacy AM, G.G., Juranic N, Macura S, Mcmurray CT. (1995) Trinucleotide Repeats That Expand in Human-Disease Form Hairpin Structures in-Vitro. *Cell*, 81, 533–540.

147. Liu G, C.X., Bissler JJ, Sinden RR, Leffak M. (2010) Replication-dependent instability at (CTG) x (CAG) repeat hairpins in human cells. *Nat. Chem. Biol*, 6.

148. McMurray, C.T. (2010) Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet*, 11(11), 786–799.

149. Tianyi Zhang, J.H., Liya Gu, Guo-Min Li. (2012) In vitro repair of DNA hairpins containing various numbers of CAG/CTG trinucleotide repeats. *DNA Repair*, 201–209.

150. John SantaLucia, J. (2007) Physical Principles and Visual-OMP Software for Optimal PCR Design. *Methods Mol Biol*, 402, 3-34.

151. Hendrix DK, B.S., Holbrook SR. (2005) RNA structural motifs: building blocks of a modular biomolecule. *Q Rev Biophys*, 38(3), 221-243.

152. Bart Hooghe, S.B., Frans van Roy and Pieter De Bleser. (2012) A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic Acids Research*, 40, e106.

153. Oyola SO, O.T., Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP, Quail MA. (2012) Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics*.

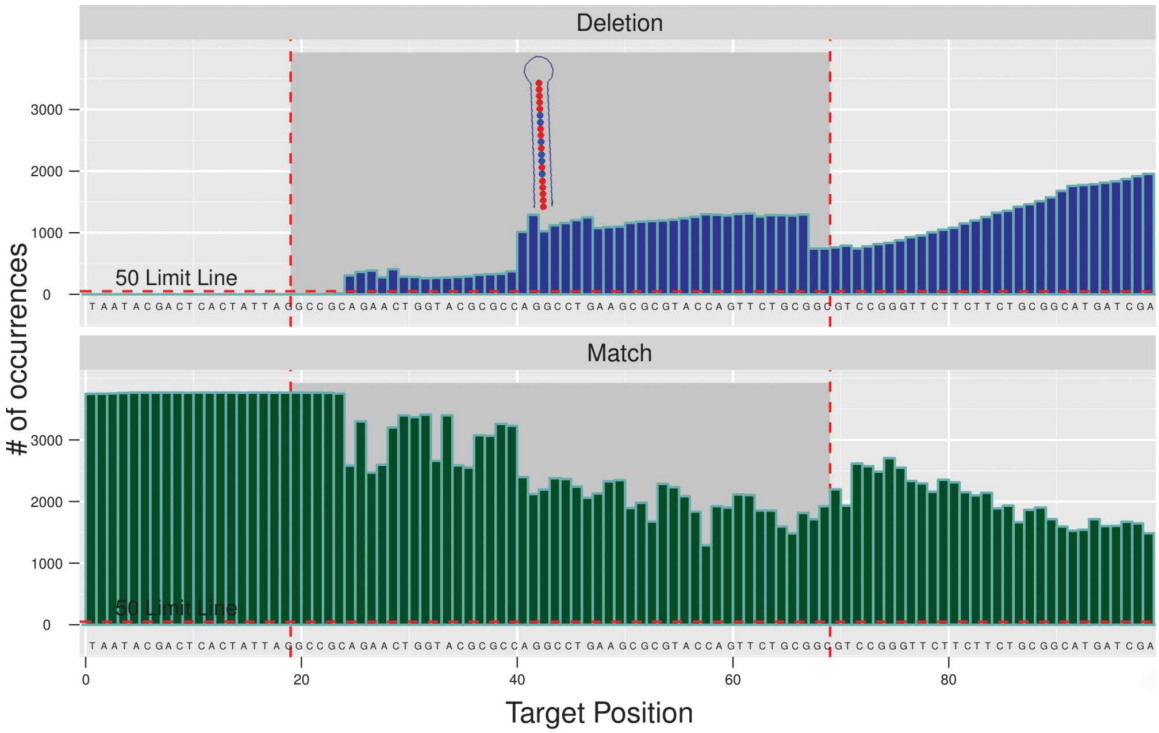
154. Jarem DA, H.L., Delaney S. (2010) AGG interruptions in (CGG)(n) DNA repeat tracts modulate the structure and thermodynamics of non-B conformations in vitro. *Biochemistry*, 49(32), 6826-6837.

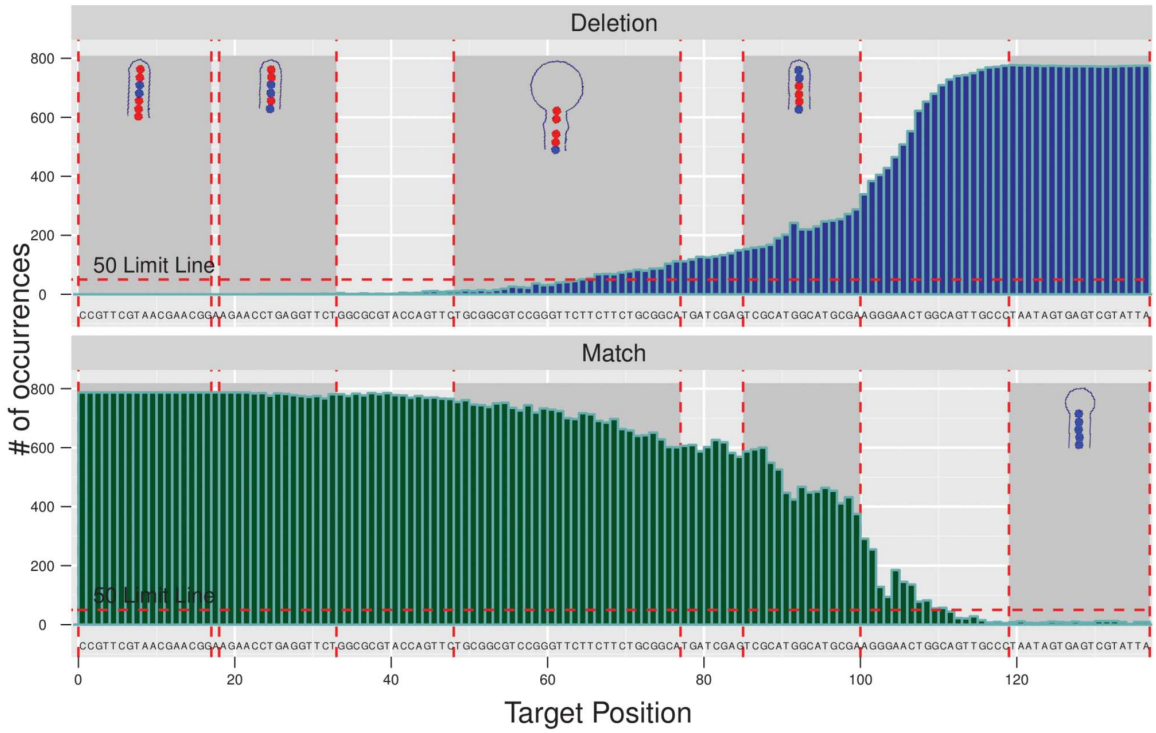
155. Jens Vo□lker, V.G., Horst H. Klump, G. Eric Plum, and Kenneth J. Breslauer. (2012) Energy Landscapes of Dynamic Ensembles of Rolling Triplet Repeat Bulge Loops: Implications for DNA Expansion Associated with Disease States. *J Am Chem Soc*, 134, 6033−6044.

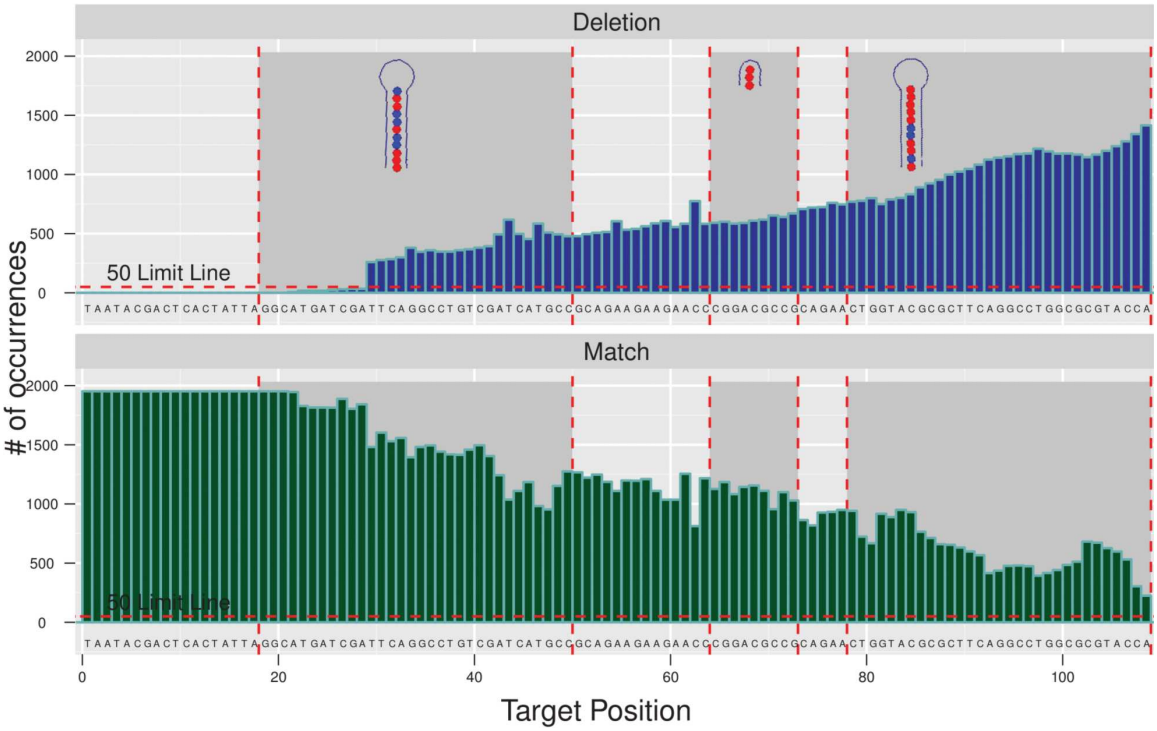
156. Kovtun IV, M.C. (2008) Features of trinucleotide repeat instability in vivo. *Cell Res*, 18, 198–213.

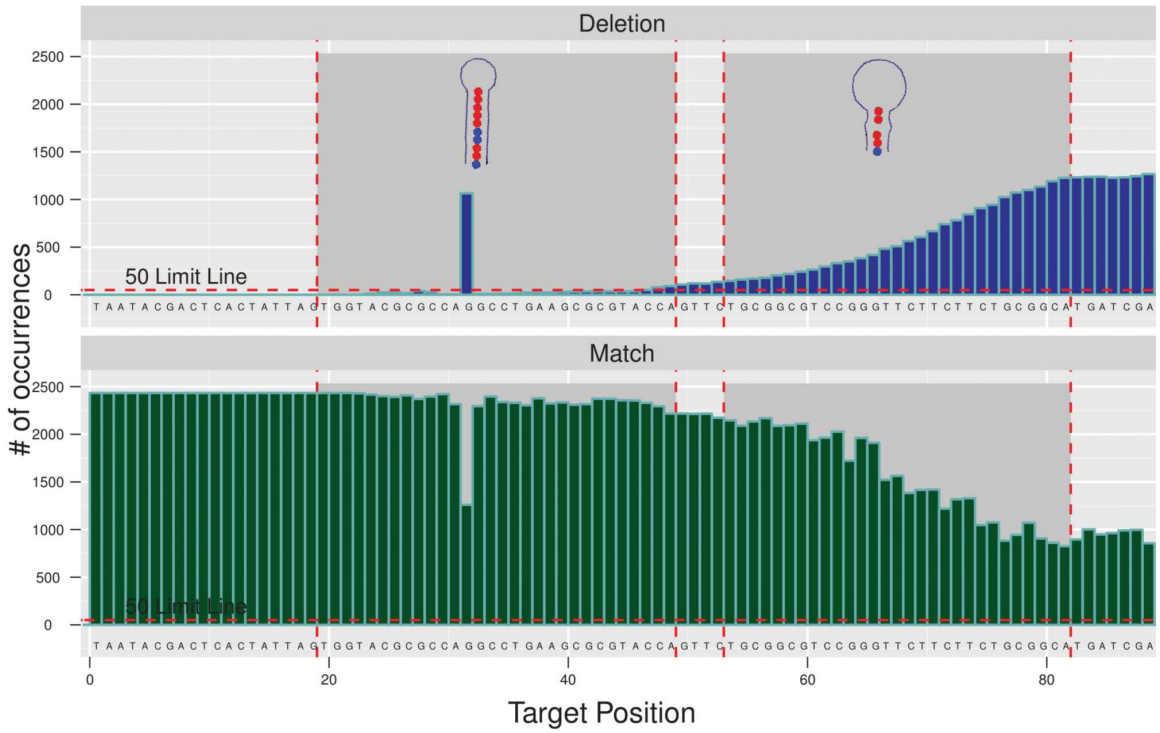
APPENDIX

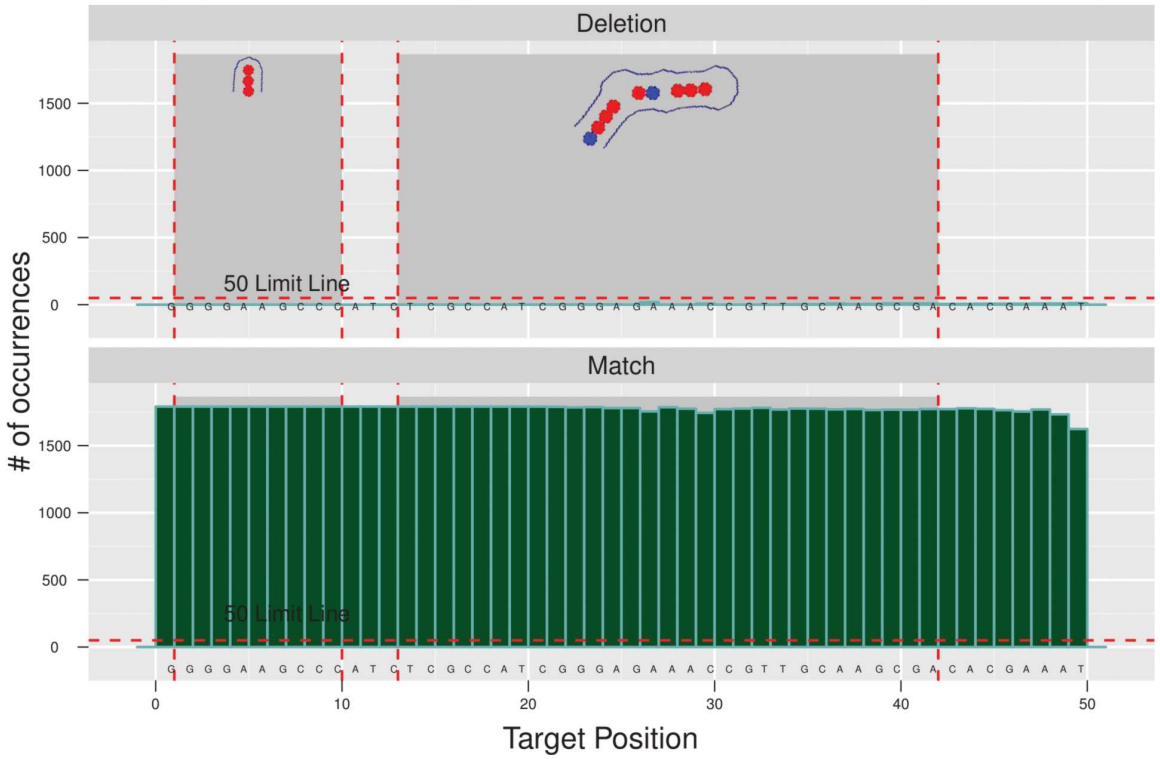
Graphs A, B, C, D, E, F, and G illustrate the deletion and sequence match distributions, respectively, for targets 1981_99, 1981_137, 1981_109, 1981_89, 857_50, 129_50, and 1571_50 which are representatives of group 3, group 1, group 2, group 2, group 1, group 1, and group 1 respectively. The structure contributing to deletions is shown in the relevant part of the deletion graph.

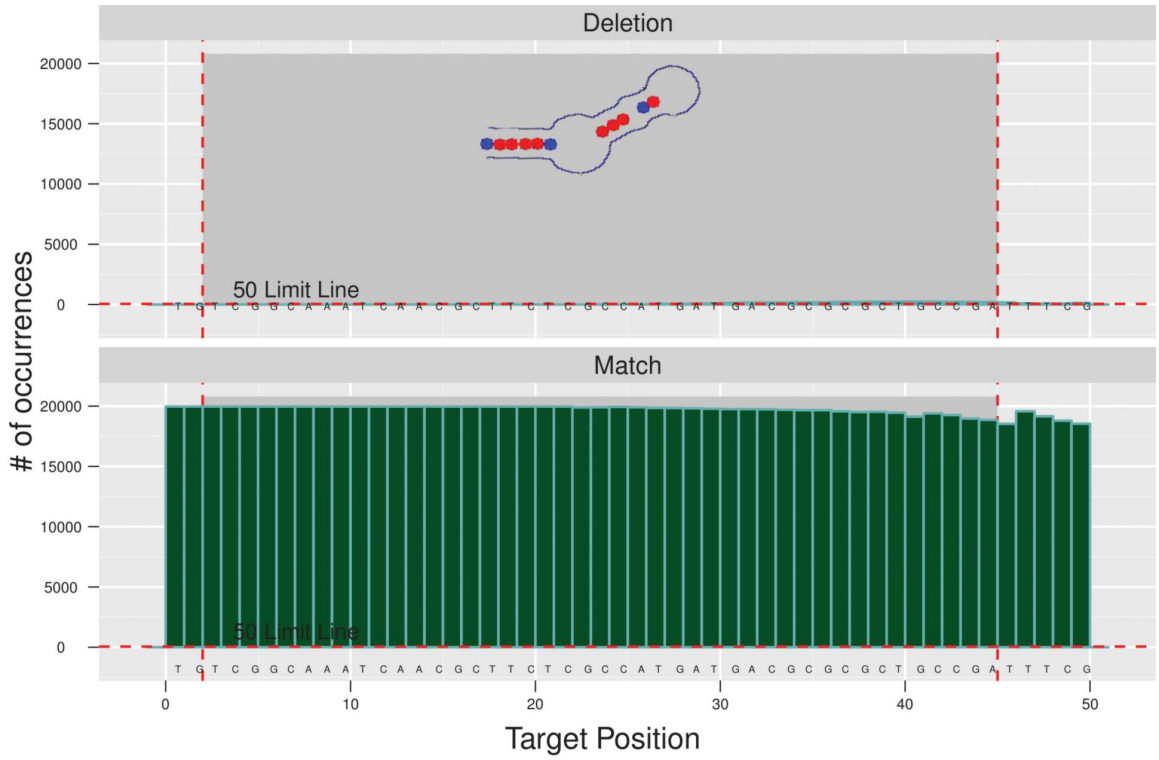


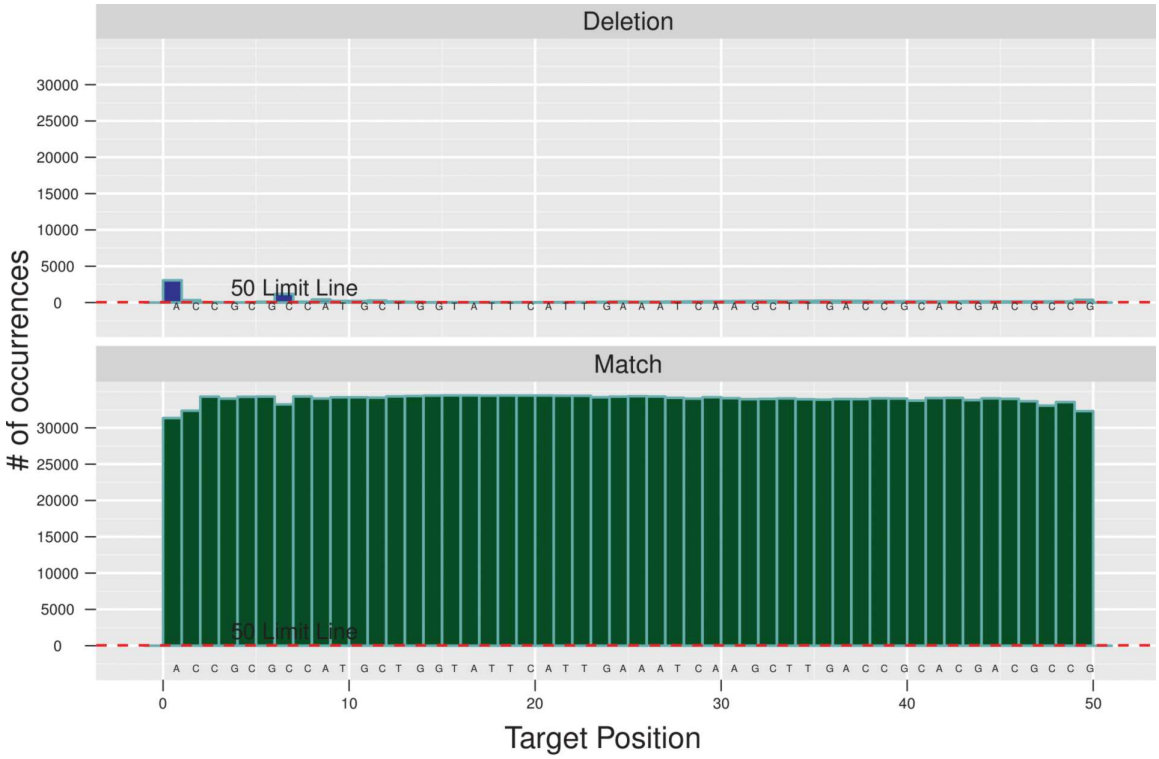

Deletion
Match
50 Limit Line
# of occurrences
Target Position
CCGTTCGTAACGAACGGAAGAACCTGAGGTTCTGGCGCGTACCAGTTCTGCGGCGTCCGGGTTCTTCTTCTGCGGCATGATCGAGTCGCATGGCATGCGAAGGGAACTGGCAGTTGCCGTAATAGTGAGTCGTATTA
0
20
40
60
80
100
120
200
400
600
800


Deletion
Match
50 Limit Line
# of occurrences
Target Position
TAATACGACTCACTATTAGGCATGATCGATTCAGGCCTGTCGATCATGCCGCAGAAGAAGAACCCGGACGCCGCAGAACTGGTACGCGCTTCAGGCCTGGCGCGTACCA
0
20
40
60
80
100
500
1000
1500
2000


Deletion
2500
2000
1500
1000
500
0
50 Limit Line
T A A T A C G A C T C A C T A T T A G T G G T A C G C G C C A G G C C T G A A G C G C G T A C C A G T T C T G C G G C G T C C G G G T T C T T C T T C T G C G G C A T G A T C G A
Match
# of occurrences
50 Limit Line
0
20
40
60
80
Target Position


Deletion
1500
1000
500
0
50 Limit Line
G G G G A A G C C C A T C T C G C C A T C G G G A G A A A C C G T T G C A A G C G A C A C G A A A T
Match
# of occurrences
50 Limit Line
0
10
20
30
40
50
Target Position


Deletion
20000
15000
10000
5000
0
50 Limit Line
T G T C G G C A A A T C A A C G C T T C T C G C C A T G A T G A C G C G C G C T G C C G A T T T C G
Match
# of occurrences
20000
15000
10000
5000
0
50 Limit Line
T G T C G G C A A A T C A A C G C T T C T C G C C A T G A T G A C G C G C G C T G C C G A T T T C G
0
10
20
30
40
50
Target Position


Deletion
30000
25000
20000
15000
10000
5000
0
50 Limit Line
A C C G C G C C A T G C T G G T A T T C A T T G A A A T C A A G C T T G A C C G C A C G A C G C C G
Match
# of occurrences
30000
25000
20000
15000
10000
5000
0
50 Limit Line
A C C G C G C C A T G C T G G T A T T C A T T G A A A T C A A G C T T G A C C G C A C G A C G C C G
0
10
20
30
40
50
Target Position