CHARACTERIZING NEXT-GENERATION SEQUENCING (NGS) PLATFORMS FOR MULTIPLEXED BIOSENSING AND APTAMER DISCOVERY: G-QUADRUPLEX APTAMERS AS A CASE STUDY

by

Sushant Patil

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
Nanoscale Science

Charlotte

2016

Approved by:

_____
Dr. Jennifer Weller

_____
Dr. Jerry Troutman

_____
Dr. Joanna Krueger

_____
Dr. Ed Stokes

_____
Dr. Valery Grdzelishvili

ABSTRACT

SUSHANT PATIL. Characterizing next-generation sequencing (NGS) platforms for multiplexed biosensing and aptamer discovery: G-quadruplex aptamers as a case study. (Under the direction of DR. JENNIFER WELLER)

Aptamers are single-stranded DNA or RNA molecules forming unique three dimensional structures that can bind to their targets (generally proteins or metabolites or other nucleic acids) with high specificity and affinity. Screening for aptamers requires filtering 10^{12} possibilities and enriching for successful binding partners– using high throughput sequencing platforms to sort out the possibilities in a single step presents clear advantages. However, aptamers are highly structured and DNA polymerases are known to have trouble processing structures *in vitro*. A prominent feature of several of the best-characterized aptamers is a planar structure called G-quadruplex (GQ); these structures can stack and are particularly stable. Far from being a purely synthetic construct, GQs are prevalent in many genomes, as well, and their involvement in modulating various cellular processes has been demonstrated. Before using NGS platforms for aptamer characterization it is important to find out whether NGS platforms sequence through and accurately report the location and neighboring sequence of GQs. We compared the performance of the two most popular NGS platforms, Ion Torrent and Illumina MiSeq, in sequencing systematically varied GQ templates: each platform has difficulty with templates whose melting temperature is above the instrument operating temperature. Each uses a different DNA polymerase and chemistries, so other factors may contribute. Only with the Ion Torrent is it possible to readily manipulate the reaction conditions – the addition of single stranded binding protein improved but did not eliminate the

accumulated base calling errors and low sequence quality of GQ containing templates. The Ion Torrent Hi-Q Sequencing Kit did not significantly alter the results: the level of base calling error remained essentially the same or increased in some cases. Further experiments suggest that the Ion Torrent signal acquisition and processing pipeline probably fails to capture and integrate the signal from a slower than expected polymerase rate. The alternate hypothesis, of replication slippage, was tested and discounted. A potential molecular biology approach to resolving the problem has been proposed that creates a double-stranded template that the polymerase can consistently process through and the resulting slower, yet consistent, nucleotide incorporation rate can be accounted for in signal processing pipeline.

Given a suitable NGS platform that is resilient to DNA structure, the end goal is to develop a protocol for multiplexed biosensing: here we propose an experiment utilizing the principle of an exonuclease I protection assay, with case studies based on characterized a thrombin protein-binding aptamer and an acetylated histone H4 peptide-binding aptamer.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ELISA	enzyme-linked immunosorbent assay
MMP-3	matrix metalloproteinase-3
HIV	human immunodeficiency virus
NGS	next-generation sequencing
PGM	Personal Genome Machine
bp	base pair
PCR	polymerase chain reaction
qPCR	quantitative PCR
emPCR	emulsion PCR
ISP	Ion sphere particle
EMSA	electrophoretic mobility shift assay
CD	circular dichroism
RPA	replication protein A
MSA	multiple sequence alignment
ExoI	Exonuclease I
EPA	Exonuclease protection assay

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 Multiplexed Biosensing

Proteins are biological macromolecules made up of amino acids monomers. Based on the composition and sequence of amino acids, their properties differ, the driving trait being the structure since the unique three-dimensional structure of each protein renders its particular function, essential to cellular processes. Probably the two most important breakthroughs in protein science have been the sequencing of insulin by Frederick Sanger[1] in 1949 and the structure determination of hemoglobin and myoglobin, by Max Perutz[2] and John Cowdery Kendrew[3], respectively in 1958. Both discoveries were rewarded with the Nobel Prize in Chemistry. Since then, driven by the need to understand more about these vital biomolecules and greatly facilitated by the advances in computational and structure determination techniques (e.g. NMR, X-ray crystallography etc.), the field of protein chemistry has flourished with the results that, as of March 2016, the Protein Data Bank holds >117,000 biological macromolecular structures[4]. With the vast number of protein molecules studied over the years, our understanding of their mechanisms and pathways has matured as well. Today, we know that proteins seldom function individually but generally operate within complexes, pathways and networks and their functions are dictated by the changes in the dynamics of these associations[5]. An example of this is the chemokine family of proteins. There has been increasing evidence that the biological activity of a chemokine is dictated by its interaction with other

chemokines[6,7]. For example, the synergistic interaction between chemokines CXCL4 and CCL5 was found to exacerbate atherosclerosis in mice[8]. Hence, in the current proteomics era, there is increasing emphasis on studying multiple proteins simultaneously in experimental samples. This complicates the development of assays aimed at detecting 'functional' protein components, since the functional unit may be a complex.

Reliable and quantitative protein assays are benchmarked against the ELISA, or enzyme-linked immunosorbent assay, first described in 1971[9,10]; the technique has remained the gold standard for quantitative determination of a protein biomolecule in a sample. Despite having superior accuracy and reproducibility compared to biochemical counterparts, the ELISA has become less relevant in addressing contemporary clinical needs, owing to its inability to measure more than a single analyte at a time. Diagnostic studies in patients often require complete profiling of protein complexes, when the presence of a single biomarker is not sufficient to confirm a particular disease[11]. Such protein profiles predict and characterize disease onset more accurately and may also serve as a biomarkers for tracking therapeutic responses. As consensus builds in the scientific community that early detection and personalized medicine is the way forward in fighting complex diseases, such as cancer or cardiovascular diseases, having a complete protein profile to monitor disease and treatment progression will be critical[12]. For these reasons, multiplexed biosensing- the analysis of many biomolecules simultaneously - is the need of the hour and is a hotly pursued area of research in the last decade[13]. In addition to addressing diagnostic needs for definitive panels, multiplexing platforms should offer high speed of detection (rapidly acquired signal in parallel assays) and high sensitivity (requiring only a small sample), which are particularly important in clinical settings.

Multiplexing should lower costs per assay and labor per data point (due to higher throughput platforms, automated sample preparation), and result in greater consistency in the data obtained as all the samples are compared against same controls[13,14].

Currently available multiplexed biosensing technologies can be broadly divided into three categories viz. mass spectroscopy, planar arrays and bead-based systems. In mass spectroscopy, intact protein molecules are enzymatically fragmented into peptides, the digested mixture is then volatilized and fragment mass-to-charge (m/z) ratios and intensities are determined. This data is compared against peptide databases wherein the peptide sequences are identified, followed by peptide grouping to recognize the original protein. Although fast and universal to many protein types, low sensitivity, extensive sample preparation and significant probability of false positive and false negative identification of protein from incomplete fragment profiles and database limitations has restricted the applicability of this platform to hypothesis-generating rather than hypothesis-validation studies[15]. In planar arrays such as protein microarrays, proteins are identified by their interaction with antibodies, and such interactions are detected by fluorescent dyes conjugated either to the protein target, the recognition antibody or a detection reagent in the form of an anti-antibody (Figure 1.1). Recently, a label-free detection approach, based on surface plasmon resonance (SPR) imaging to detect binding events, is gaining popularity[16]. In either format, the identity of different analytes is inferred from their spatial locations on a two dimensional array surface. In bead-based systems, capture antibodies are attached to optically encoded microspheres and binding events are detected by a fluorescently labeled antigen or antibody, similar to reagents used with planar arrays. However, unlike planar arrays, the identity of multiple analytes

is not traced by their spatial locations but by the unique optical signature of each microsphere. Because of their higher throughput capability over planar arrays, most of the commercial clinical analysis platforms utilize bead-based systems. For instance, in many clinical cancer tests, the BeadArray Microarray Technology[17] by Illumina, Inc (CA, USA), with its Veracode optical signature technology[18] is used. Because of the limited optical combinations of such beads, a maximum of 384 bead types have been achieved so far, and, as this number has not increased over a decade so, likely the physical limitation of this platform has been reached. Other examples include the Luminex xMAP (Multi-Analyte Profiling) platform by Luminex (TX, USA) and the CBA (Cytometric Bead Array) platform from BD Biosciences (CA, USA)[19]. Among service providers, Myriad RBM (TX, USA) is frequently cited, offering a multiplex immunoassay service for hundreds of proteins using just a microliter quantities of blood or serum. In all such protein to protein multiplex analysis systems, the major challenge lies in uniquely and sensitively identifying the capture molecules ('probes') that may be arrayed on a surface or mixed in a suspension, and ensuring that the probes are specific to the target that is clinically important. Thus the limit of these assays lies in the low 100's despite ever-increasing information about clinically significant protein variants, that number in the thousands. In addition, the clinical assays described have been painstakingly developed to identify a known variant, but none are discovery techniques for capturing new variants. Given the ten to one hundred fold gap in the numbers of specific sensors needed and the numbers currently in the repertoire, it is clear that a new approach to multiplex biosensor discovery is needed.

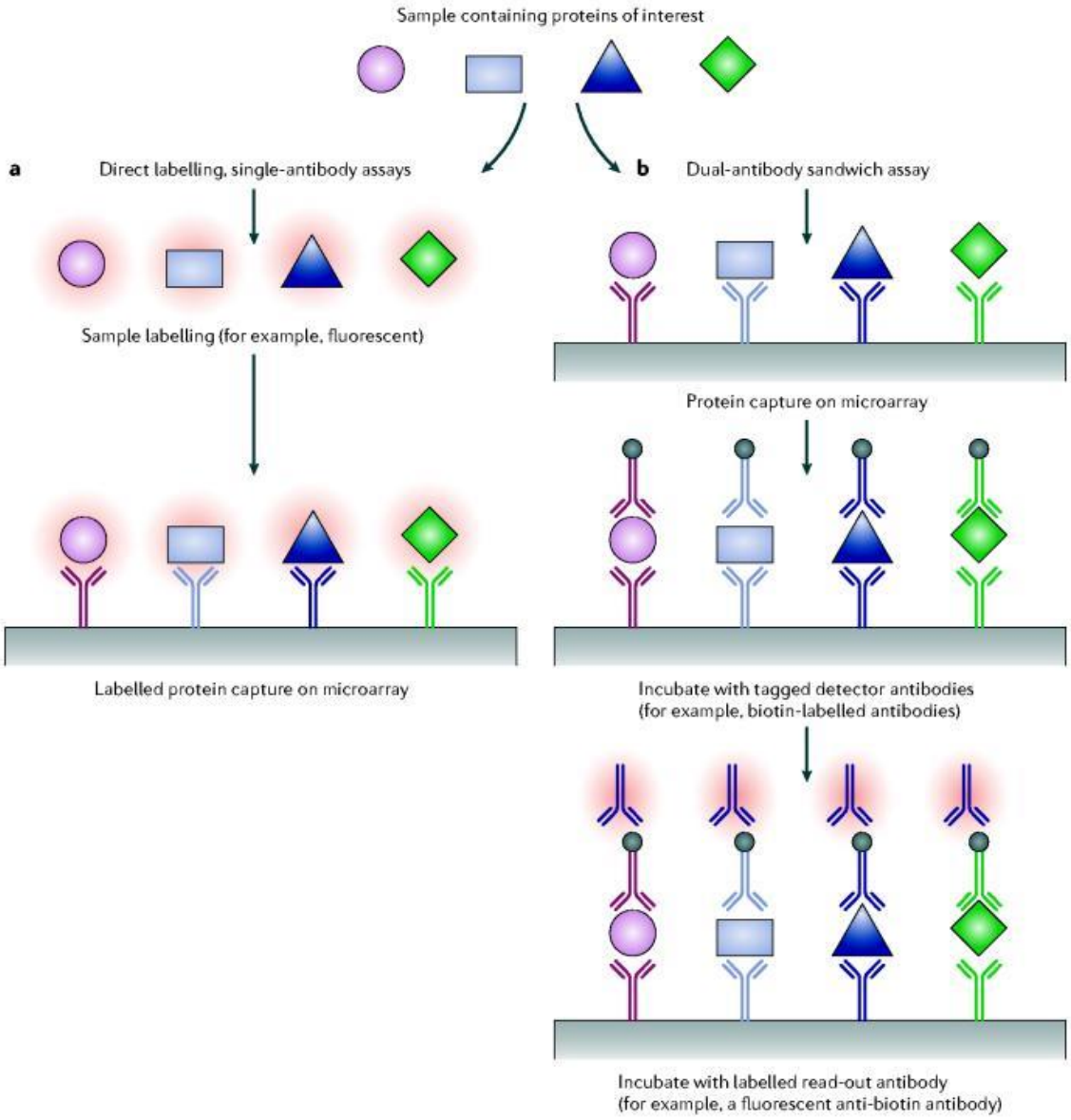


Figure 1.1: Schematic representation of common multiplexed protein detection approaches in planar arrays. Reproduced from reference[15].

Multiplexed biosensing is feasible: it has been demonstrated for a number of applications in various clinical disciplines including biomarker discovery, clinical diagnostics, tissue engineering and drug discovery. For example, by screening 204 analytes from total 294 individuals, Bertenshaw *et al.*[20] narrowed down eight potential

serum biomarkers for diagnosis of stage II ovarian cancer. Similarly, Kim *et al.*[21] were able to identify combinatorial biomarkers for early diagnostics of breast cancer by screening 35 analytes from total 194 subjects. Autoantibody profiling against various antigens for identifying biomarkers has been used in the classification of autoimmune diseases[19]. There are even some commercial products, including the AtheNA Multi-Lyte® Test Systems by ZEUS Scientific, Inc. (NJ, USA) and the BioPlex® 2200 ANA Screen by Bio-Rad Laboratories, Inc. (CA, USA). Both are FDA-approved assays for autoimmune and infectious diseases. In diagnostic microbiology, Opalka *et al.* simultaneously measured neutralizing antibodies against four different types of human papillomavirus[22]. Likewise, Bellisario *et al.*[23] were able to measure antibodies against three HIV-1 antigens using newborn dried blood-spot specimens. In tissue engineering, Pelttari *et al*[24] discovered that the loss of ectopic cartilage formation capacity of human articular chondrocytes was correlated only with a drop in the secretion of matrix metalloproteinase-3 (MMP-3), among 34 different analytes studied, and indicated the desirability of monitoring MMP-3 levels as a quality control measure for chondrocytes used for cell therapeutics. In drug discovery, the anti-inflammatory efficacy of Besifloxacin, a novel fluoroquinolone for ophthalmic infections by Bausch & Lomb (NY, USA) was assessed by studying LPS-induced cytokine expression in human THP-1 monocytes *in vitro* multiplexed assay[25]. Indeed, from early disease diagnosis to better disease characterization and thereby the most appropriate treatments, multiplexed immunoassays are revolutionizing clinical proteomics and driving us towards personalized medicine. However, to fully realize their potential a much richer source of protein variant-specific reagents is needed.

1.2 Aptamers

Aptamers are single-stranded DNA or RNA molecules forming unique three dimensional structures that can bind to their targets (generally proteins or metabolites) with high affinity. The concept of using nucleic acids for protein binding originated in the 1980s from studies on adenovirus and the human immunodeficiency virus (HIV), when it was learned that small structured RNAs encoded by these organisms bind to cellular or viral proteins with high specificity and affinity[26]. In 1990, Sullenger *et al*. demonstrated the first therapeutic application of aptamers, when inhibition of HIV replication was observed by the addition of an aptamer against the viral RNA genome element essential for its replication[27]. Since then, the field of aptamers has grown, catalyzed largely by the development of cheap mass production chemistry for DNA and RNA molecules and an *in vitro* aptamer discovery/selection process, called 'systematic evolution of ligands by exponential enrichment' (SELEX)[28,29] which has contributed the majority of the aptamers discovered to date. Briefly, SELEX involves selection of a final set of high-specificity aptamers starting from a library of 10^{12}- 10^{15} molecules. Multiple rounds of positive selection are carried out against a targeted cell component or purified biomolecule, followed by enrichment through PCR amplification (Figure 1.2). The ligand affinity constants (K_d) of aptamers typically range from picomolar to micromolar levels, with the lowest reported to date being 38 fM for an aptamer against interferon-γ (IFN-γ)[30]. Another key characteristic of aptamers is that stringent specificity can be attained, arising because of the counter selection step in SELEX. For example, an aptamer against the human immunodeficiency virus reverse transcriptase (RT) was found ineffective against feline or murine immunodeficiency virus RTs[31] and, *vice versa*, an aptamer against the

feline immunodeficiency virus RT could not bind the HIV RT[32]. In another example, an aptamer selected against a eukaryotic RNase H1 displayed no interaction with the structurally similar E.coli RNase H, or the HIV RT RNase H[33].

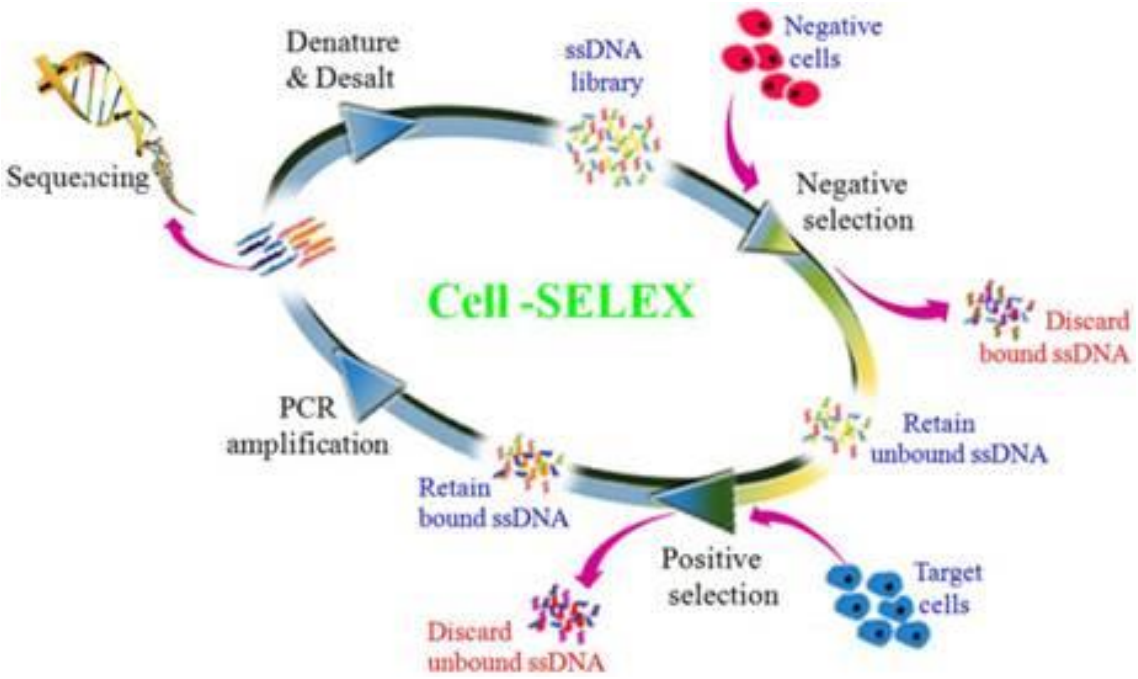


Figure 1.2: Schematic representation of cell-SELEX workflow. SELEX for protein or metabolite targets is performed in a similar way. Reproduced from reference[34].

While functionally similar to antibodies in terms of tight binding and variant-specific discrimination, aptamers offer many advantages. Compared to antibodies, aptamers are less laborious to create, less expensive to produce and more flexible to modify for detection methods in assays (generally this means conjugating dyes or functional groups to one end of the aptamer). Apart from being much quicker to produce than antibodies, there is little or no batch-to-batch variation and fewer concerns about

potentially toxic contaminants from the production method. Aptamers offer flexibility of selection against a wide variety of targets, and are especially desirable against cell-surface targets (membrane-associated proteins) that are not available in a functional recombinant form against which antibodies can be raised. Aptamers exhibit a high degree of thermal stability, can be dried down and shipped and stored at ambient temperature, and can be used under wide-ranging assay conditions where antibodies easily undergo irreversible thermal denaturation. Because of their smaller size, aptamers exhibit better pharmacokinetic and pharmacodynamic properties than antibodies – in addition, unlike antibodies, aptamers exhibit little to no immunogenicity.

The above mentioned characteristics of aptamers make them ideal candidates for applications in drug delivery, therapeutics, bio-imaging, disease diagnosis and so on[35,36]. The number of publications aptamers for such interesting applications keep on rising every year. Despite that, compared to antibodies, aptamers haven't received much commercial success so far. In 2004, Pegaptanib (trade name Macugen), a vascular endothelial growth factor (VEGF)-specific aptamer, for the treatment of wet age-related macular degeneration, became the first aptamer to be approved as a drug by the Food and Drug Administration (FDA)[37]. Currently, only a small number of aptamers are being tested in various phases of clinical trials[36]. The reasons for this low market realization include a strong industrial investment, and therefore commitment, in antibodies and also a general ignorance and thus indifference to aptamer science among the applied research and development communities. However, given some of the exclusive properties of aptamers to address specific needs in therapeutics or diagnostics coupled with increasing awareness, a bright commercial future is in sight for aptamers[38]. Since aptamers are

relatively short nucleic acid templates that form structures, and sorting through hundreds of millions of them is the first stage in the biosensor discovery process, marrying aptamer discovery to a high-throughput sequencing platform, discussed next, would seem to offer real advantages in developing clinical assays in a short time frame.

1.3 Next-generation Sequencing (NGS)

In 1977, Dr. Frederick Sanger devised a pioneering method for DNA sequencing[39], now known as Sanger sequencing, for which he jointly won the Nobel Prize in Chemistry in 1980. Sanger sequencing made use of strand-terminating dideoxynucleotides and radiolabeled dATP molecules, combined in a standard 'primer extension' reaction mix effected in the sequencing -by -synthesis strategy. Because of the dideoxynucleotides' reaction- terminating properties, DNA fragments of different lengths are produced. When length-resolved on polyacrylamide gels and visualized (originally by the virtue of ^{32}P in the dATP molecule), a sequence complementary to the input template can be determined. In 1986, Applied Biosystems, Inc. (MA, USA) introduced a fluorescent DNA sequencing instrument in which radiolabeled dATP was replaced by fluorescently labelled primers[40]. The raster scanning laser beam for gel visualization of DNA fragments made the process of detecting and size-ordering fragments less erroneous and laborious. Subsequently, the introduction of fluorescent dideoxynucleotides instead of fluorescent primers made the process more sensitive and cheaper. In the 1990s, the rate-limiting and error-prone manual step of slab gel electrophoresis was replaced by capillary electrophoresis, sample loading was simplified using automatic electro-kinetic injection, and changes in polymer characteristics significantly reduced the run times. All these refinements to Sanger sequencing were implemented in the labs involved in the

Human Genome Project[41] which still took more than 10 years to finish (1990-2003) and costed about $3 billion. Today, the same task can be achieved within a day or two, and a cost of under $5000 (or $1000 if a reference is available), with the application of Next Generation Sequencing (NGS) methods and instruments.

There exist some major differences between Sanger sequencing and NGS methods, which in turn contribute to the speed and affordability of the latter. The Human Genome Project adopted a "top down' approach in the sense that the genome was first fragmented into large pieces (100-500kbp), cloned in bacterial artificial chromosome (BAC), amplified in bacterial culture, further sheared into smaller fragments and then sub-cloned into plasmid vectors for amplification before re-purification and sequencing. After signal acquisition the data was processed and base-calls were made. Further analysis steps were required to order the fragments into gene and chromosome sized segments. NGS methods get rid of the lengthy and laborious large- and small-fragment cloning steps by directly fragmenting a genome to read-length size, and directly adding adaptor sequences for the library construction step. In Sanger sequencing, a 96-well or 384-well microtiter plate scale is used to systematically control individual plasmid-derived templates, whereas NGS methods employ a semiconductor chip or a glass slide which serves as a platform for millions of sequencing reactions to be carried out in parallel, using spatial resolution on a surface or beads to discriminate the templates, thereby greatly improving throughput. Also, unlike Sanger sequencing, base incorporation and detection (base calling) may be nearly simultaneous on NGS instruments[42]. Although Sanger sequencing holds a performance edge in terms of

producing about 20% longer average read lengths of fragments, the paired end-sequencing or long single molecule sequencing methods are now approaching that limit.

In research labs today, there are three primary NGS platforms in use: the Ion Torrent sequencing by synthesis based method [43] by Thermo Fisher Scientific (MA, USA), a sequencing by synthesis technology[44] by Illumina (CA, USA) and a method that detects each nucleotide as it passes through a pore, called single molecule real time sequencing[45] by Pacific Biosciences (CA, USA). In the context and relevance of this dissertation, the first two technologies will be discussed in detail below.

1.3.1 Ion Torrent - Ion Semiconductor Sequencing

The Ion Torrent sequencing protocol can be divided into four main steps viz. library preparation, template preparation, sequencing and data analysis. Library preparation starts with fragmentation of the target DNA to a size suitable to be sequenced, performed by sonication or enzymatic methods. Ion Torrent standard adaptors are ligated to the sheared fragments, which make them compatible for polymerase chain reaction (PCR) amplification and sequencing downstream. The fragments are generally size-selected at this stage for a maximum size of either 200 bp or 400 bp depending upon the amplification and sequencing kits to be used later. The library is quantified at this stage by either quantitative PCR (qPCR) or the Agilent 2100 Bioanalyzer gel electrophoresis system (Agilent Technologies, CA, USA) to estimate the dilution required to achieve the desired library concentration (in the picomolar range).

Template preparation is done through emulsion PCR (emPCR) and is performed using the Ion OneTouch™ 2 System. The automated process consists of mixing an aqueous phase with an oil phase, resulting in tiny aqueous droplets, which serve as micro-

reactors for millions of distinct PCRs. Each droplet ideally contains a micron sized polystyrene bead, called an Ion Sphere Particle (ISP), one DNA fragment, primers complementary to the adaptors added (one is biotinylated) and PCR reagent mix. The ISP has one million covalently linked complementary sequences to adaptor sequences on the DNA fragments, and thus during thermocycling, in each the microreactor, the PCR process coats each ISP with about one million copies of its unique template – the end result being millions of beads each with one million copies of just its own fragment. An enrichment step is performed after emPCR using magnetic Streptavidin beads and the Ion OneTouch™ ES (enrichment system), which breaks the microreactors and allows enrichment of beads containing templates that can be sequenced.

The sequencing primer is annealed to the enriched ISPs, DNA polymerase is bound to the template-primer complex, and the loaded ISPs are allowed to flow across the surface of the semiconductor sequencing chip, to be randomly deposited in individual wells on the chip (Figure 1.3). During the sequencing process itself, the chip is sequentially flooded with each one of the four nucleotides at a time. If the nucleotide is complimentary to the next nucleotide on template strand, it is incorporated in the growing strand and a proton is released during the reaction. The released proton alters the pH of the solution in that particular well, which is measured by an ion-sensitive layer beneath that well and converted into a voltage indicating that the base in that flow cycle was incorporated. So, in essence, the chemical information from DNA sequencing is captured form the CMOS chip layer, and translated into digital information which is interpreted as base calls by instrument signal detection software. This process occurs simultaneously in

millions of wells on the chip, generating an equal number of sequence ‘reads’ at the end of the process.

After filtering out low-quality and polyclonal reads (interpreted as reads from ISPs with more than one starting template, resulting in mixed nucleotide incorporation at most steps), reads are assembled into a contiguous sequence representing the full length of the original template, which may be a genome or large genomic region. Two assembly approaches can be adopted viz. comparative assembly and *de novo* assembly. If a reference genome is available, the easier approach is to use comparative assembly, which consists of mapping the reads to the reference genome. A reference could be either the genome of the same organism, or it may be stretched to include a genetically closely related organism. In the *de novo* assembly approach, overlapping reads are identified and merged, creating ever-longer ‘contigs’ (contiguous sequence). As contigs are linked through overlapping sequences, longer ‘scaffolds’ result. Finally, scaffolds are assembled into super-scaffolds and the consensus sequence for the entire targeted genomic region is derived. Genome assemblies are often evaluated based on the number of contigs and scaffolds required to connect the entire region targeted (smaller is better) and also on the lengths of individual contigs and scaffolds (larger is better)[46].

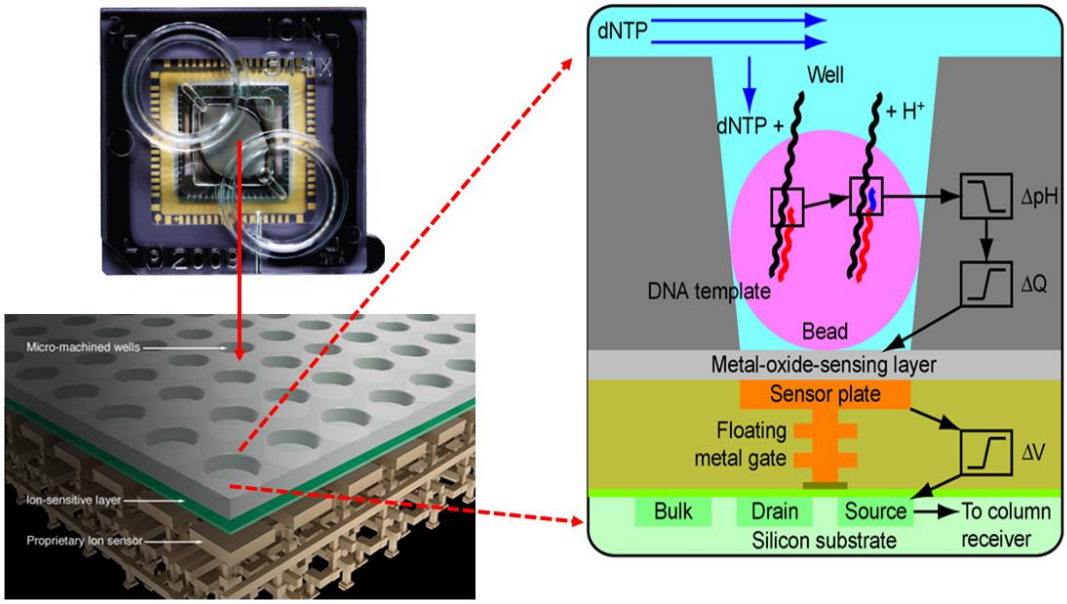


Figure 1.3: (Top Left) Top view of Ion Torrent 314^{TM} chip (Bottom left) Representation of semiconductor wafer of the chip having about a million of micro machined wells. (Right) Representation of a single well on chip. Each well contains a polymer bead populated with ssDNA template. After adding primer and polymerase, each of the dNTPs is sequentially flowed over the chip. When a nucleotide is incorporated into a growing DNA strand, a proton is released. The resultant drop in pH is detected by the sensor underneath the chip and then that nucleotide is noted as the next base in the sequenced strand. Adapted from reference[43,47]

1.3.2 Illumina - Sequencing by Synthesis with Optical Detection

The Illumina platform differs from the Ion Torrent mainly in the template preparation steps and sequence detection technology. The targeted DNA is still fragmented and adaptors are ligated to the fragments to create the template library. The template library is then distributed across a surface that is covalently modified with sequences complementary to the adaptors. This template deposition step is followed by a PCR step that yields clusters, and is carried out in a proprietary flow cell (Figure 1.4). Clusters are generated because propagation of single template molecules must be carried

out to generate enough copies for detection. First, the compliment of the hybridized fragment is created, the strands are denatured to wash-off original template strand and then the remaining strands are PCR amplified by a process called ‘bridge amplification’. The strand folds over toward the surface, to hybridize to the sequences complementary to more of the library adapters anchored on the flow cell surface. The complimentary strand is then generated, forming a ‘bridge’, which is again denatured to generate two single-stranded copies covalently attached to the flow cell. Multiple iterations of this process result in millions of clusters across the surface, each of one type of fragment, now ready to be sequenced.

Sequencing is done by reversible dye terminator technology (Figure 1.5). Following primer annealing, all four nucleotides, each conjugated with a distinct fluorescent dye molecule, flow together over the cell surface. The 3’-OH position of each nucleotide is blocked by a cleavable terminator group, which ensures that only one nucleotide is added at a time to the growing strand. Unbound nucleotides are washed away and an image of flow cell is recorded, the color at each cluster location is used to identify the fluorescent dye and thus infer the corresponding nucleotide (the base call). After unblocking 3’-OH position so that the growing strand can be extended, the cycle is repeated, the number of flow cycles will complete all molecules at or below the size selected for library fragments during the preparation step. Data analysis is performed the similar way as that of Ion Torrent.

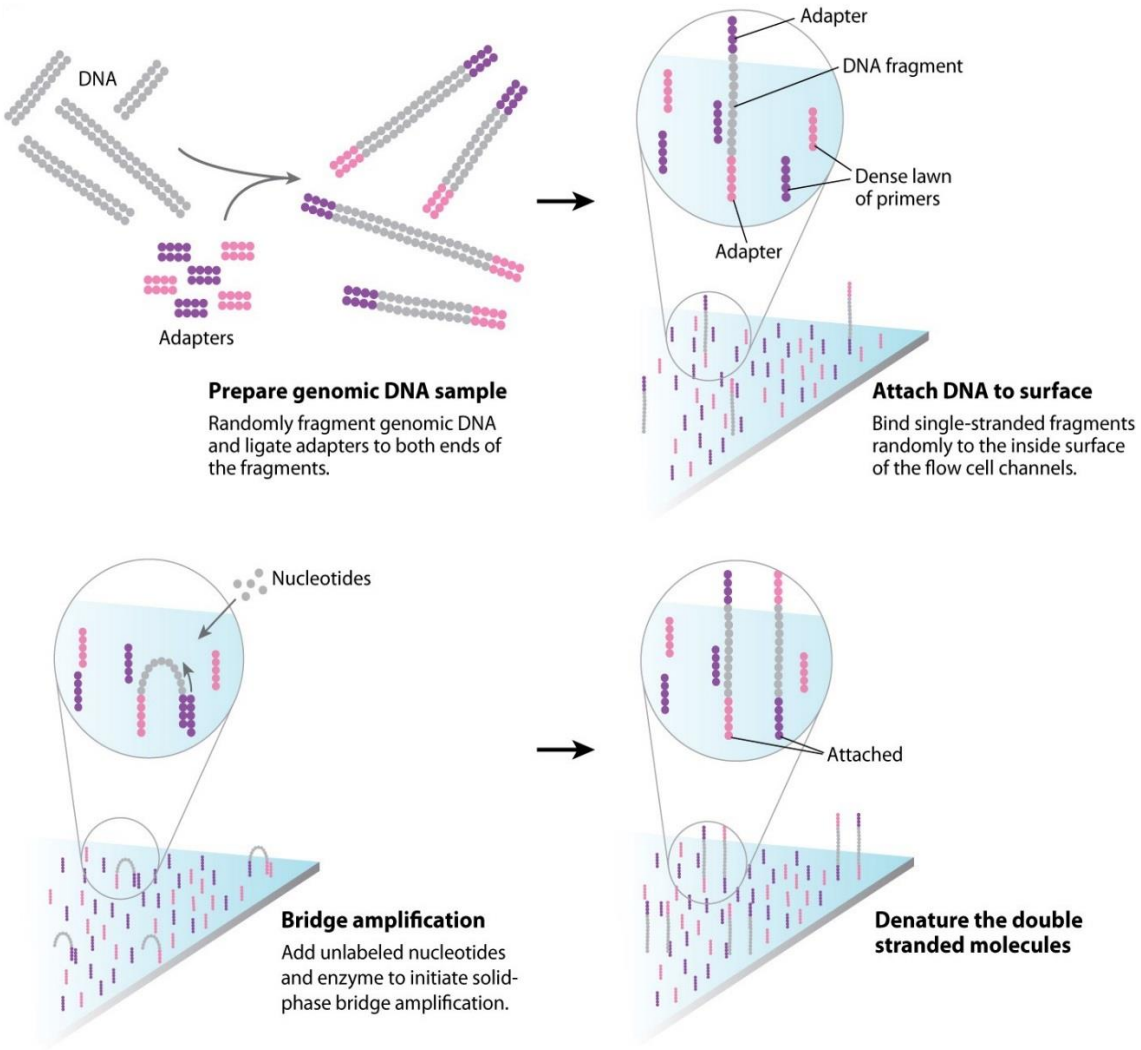


Figure 1.4: Cluster generation on Illumina flow cell by bridge amplification process. Reproduced from reference[48].

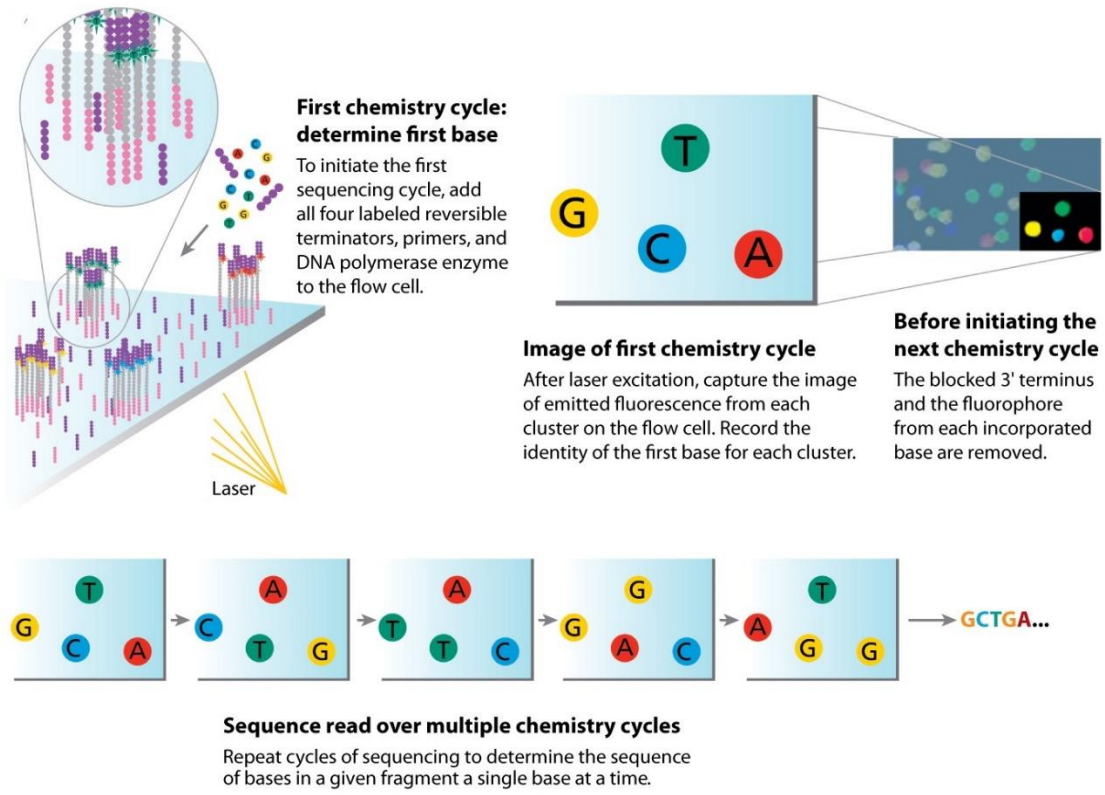


Figure 1.5: Sequencing by reversible dye terminator technology in Illumina. Reproduced from reference[48].

1.4 Dissertation Objectives and Outline

The overall goal of the dissertation is to determine whether we can design a process in which NGS platform sequence sorting capabilities converge with large-scale DNA aptamer characterization, to help provide a more efficient aptamer discovery scheme, and if so to determine whether we can establish a novel, multiplexed biosensing method. We aim to test how accurately DNA aptamers can be sequenced on the Ion Torrent and Illumina platforms, determine whether the sequencing accuracy be improved by common methods developed on other platforms to handle structures templates and,

finally, whether aptamer-protein interactions can be detected using one or both of these NGS platforms. The following specific aims will be addressed:

Paralleling the advent of NGS platforms, there have been efforts to find alternatives to the SELEX process of aptamer selection. Although the possibility of a new NGS-based aptamer discovery method that is much faster and cheaper than SELEX is compelling, the underlying requirement is that NGS platforms should be able to accurately sequence complex DNA structures. Sanger sequencing platforms were known to yield artifacts when presented with certain sequence motifs, and the NGS literature also records some examples in which sequences with some particular characteristics are either absent or erroneously sequenced. Since a prominent feature of several of the best-characterized aptamers is a structure called the G-quadruplex (GQ), we focused our sequencing tests on a family of these structures. In Chapter 2, a comparative analysis of Ion Torrent and Illumina platform and chemistry performances in sequencing these GQ templates is performed.

Having discovered the poor sequencing performance of GQ-containing templates, particularly by the Ion Torrent, the causes and possible solutions for Ion Torrent GQ sequencing issues are addressed in Chapter 3. To understand the characteristics that make GQ structures difficult to sequence, the sequencing output of systematically varied GQ structures is studied as a function of structural stability, represented by the melting temperature, T_m. The effectiveness of two reagents that are often added to sequencing mixtures, viz. *E. coli* Single-stranded DNA Binding (SSB) protein and reagents in the Ion PGM Hi-Q Sequencing Kit were tested with respect to improving sequencing accuracy. Since the errors observed on both platforms could be due to either replication slippage or

instrument signal acquisition and processing pipeline, we also tested the replication slippage hypothesis, using Ion Torrent reagents and conditions, in Chapter 3.

Given that some of our sequencing problems can only be resolved by instrument modification, our final goal in characterizing the feasibility of the Ion Torrent instrument for aptamer discovery was to develop a universal protocol for multiplexed biosensing, where the outcome is independent of the structural complexity of aptamers but still uses the native capabilities of the NGS platform. Chapter 4 proposes an exonuclease I protection assay based protocol for testing this objective. Thrombin aptamer is a good candidate for a proof-of-concept experiment, which is described in detail, and acetylated histone H4 is suggested as a second test case. Finally, our conclusions are summarized, highlighting the global significance of the research, and suggestions for additional avenues of further research that we consider most promising.

CHAPTER 2: COMPARATIVE ANALYSIS OF ION TORRENT AND ILLUMINA PERFORMANCES IN SEQUENCING G-QUADRUPLEX (GQ) TEMPLATES

2.1 Rationale

Over the past two decades, systematic evolution of ligands by exponential enrichment (SELEX) has remained the method of choice for selecting aptamers with high affinity and specificity. Although effective against a wide variety of targets, SELEX[28,29] has a few weaknesses: a typical SELEX experiment involves 8-15 cycles of positive selection-negative selection-PCR amplification before sequencing the enriched target pool, and the protocol can take several weeks to complete, which adds to the labor costs associated with the process. The process of progressive enrichment reduces a pool of 10^{12}- 10^{15} possible target candidates to a small number (ideally fewer than 100), each of which is then sequenced- historically using Sanger sequencing chemistry. One consequence of carrying out multiple rounds of PCR amplification is that the evolution of the library is partly driven by the presence of easily amplified sequences, which compete with the often highly structured high affinity sequences. Because of the structural complexity inherent in providing specific recognition, high-affinity sequences may have poor PCR template properties that artificially reduce their representation. That is, sequences with complex motifs that form structured regions cause polymerases to traverse more slowly, and thus amplify poorly. More recently developed protocols of aptamer selection have replaced Sanger chemistry with one of the NGS-associated platforms, which circumvents some of the shortcomings of conventional SELEX. In the

case of high-throughput NGS platforms, the number of enrichment cycles does not have to be nearly as great: instead of sequencing each to be nearly as great: instead of sequencing each target individually one can sequence 100 million simultaneously. Identifying high-affinity targets is still required but decreasing the background through PCR enrichment, with the associated risks of false positives, can be greatly reduced (from 10^{12} to 10^{8} on Illumina, for example, well within one-two PCR enrichment steps). Researchers have been able to generate high affinity aptamers after just a few rounds of SELEX, followed by sequencing[49-51]. In fact, Hoon *et al.*[52] identified anti-thrombin aptamers with dissociation constants in the nanomolar range from just one round of selection, followed by sequencing and a rigorous informatics analysis. Not only does such NGS-aided aptamer selection avoid PCR bias but it shortens the overall time to carry out the protocol from weeks to hours, thereby reducing the associated costs[53,54]. In this regard, it is absolutely evident that efficient aptamer selection protocols should use NGS platforms, assuming bias is not evident, for their ability to rapidly sequence tens of millions of aptamer candidate molecules in a cost-effective manner.

While not using Sanger dideoxynucleotide chemistry, the NGS platforms still perform sequencing by synthesis, and require a DNA polymerase. As mentioned above, some DNA structures will cause DNA polymerases to pause, stall or fall off the template, with varying effects on the efficiency and accuracy of the sequence readout. These effects can include truncation, mis-incorporation errors (apparent SNPs), insertions, deletions, and random addition of bases at the terminus. Aptamers by definition include a structure used in recognition of a second molecule – to what extent are the NGS polymerases affected by these structures? A prominent feature of several of the best-characterized

aptamers is a structure called the G-quadruplex (GQ). In a GQ, four guanines form a square planar structure called a tetrad that is stabilized by Hoogsteen hydrogen bonding. Two or more such tetrads can then stack on top of each other, favored by pi-pi stacking interactions, to form a three-dimensional structure. A GQ can adopt either parallel (four strands in the same direction) or antiparallel (one or more strands in opposite direction than rest) topology. The central channel of a GQ is electronegative because of the orientation of the carboxyl groups and hence, such structures are further stabilized by monovalent cations (e.g. K^+) that intercalate between the tetrad layers[55](Figure 2.1). In general, a GQ sequence can be defined as d($G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$), where N can be any nucleotide[56]. Currently, there are over 20 characterized GQ-containing aptamers[57]. These provide a ready-made test case for assessing whether such complex, stable structural motifs are sequenced accurately by the state of the art sequencers. The NGS-based aptamer screening protocols that are being developed will not be productive if the sequencing is error-prone.

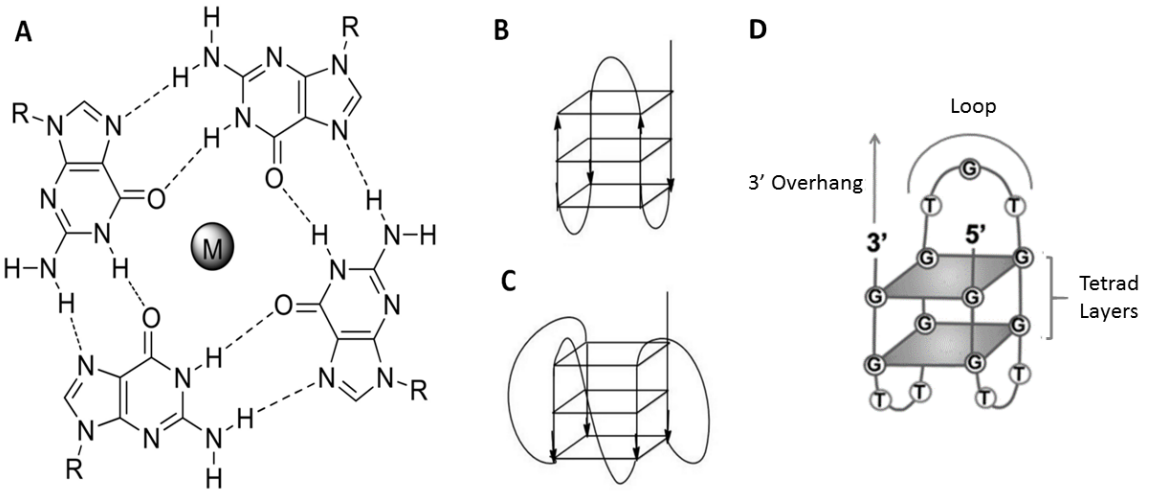


Figure 2.1: (A) A single layer of a GQ, showing how four guanines interact by Hoogsteen hydrogen bonding and are further stabilized by a monovalent cation in the central electronegative pore. (B) Antiparallel GQ topology (C) Parallel GQ topology (D) A three-dimensional representation of a GQ with tetrad layers, loop and 3' overhang regions indicated. Adopted from reference[58].

The other, and perhaps more interesting, reason for studying the sequencing characteristics of GQs is that they have intrinsic biological relevance. The human genome is rich in putative quadruplex sequence motifs[56,59], and their structures have also been detected[60] proving their *in vivo* existence. Probably the most widely studied role of GQs is in gene regulation, due to their prevalence in promoter regions across both prokaryotic and eukaryotic genomes[61-63]. Additionally, sequences containing GQ cores are implicated to play important roles in telomere maintenance[64], DNA replication[65], recombination[66], initiation of double strand break[67] and protein binding[68]. While the number of genomes sequenced every year is rising exponentially, and huge amounts of money and effort are being invested, it should be noted that most of these genomes are done at a very low coverage with a limited ability to detect more difficult regions. Failure to correctly sequence GQs may introduce significant biases during genomic assembly, annotation,

and downstream network inference analyses of promoters and transcription factor binding sites, which can have serious implications, particularly in clinical studies.

In order to determine the efficiency and accuracy of NGS platforms when challenged with structured templates, we studied the sequencing accuracy of systematically varied GQ structures using the Ion Torrent PGM and Illumina MiSeq platforms (chemistry plus detection instrument).

2.2 Methods

2.2.1 GQ Design and Synthesis

The two most prominent structural variables affecting a GQs' stability are the number of tetrad layers[69] and the loop length[69-74]. Empirically, it has been observed that the stability increases with the number of layers and decreases when there is an increase in loop length. To observe the effect of these two parameters on sequencing output, GQ templates with systematically varying stability were designed (Figure 2.2). In Set I, the position of the GQ (with three layers and three nucleotides in loops) with respect to the sequencing primer binding site was increased in three nucleotide increments, for a set of four templates. This set tests whether the lead distance for polymerase registration is important. Set II included four GQs with, respectively, two, three, four and five tetrad layers with the loop length of three nucleotides (Figure 2.2B). This set tests the point at which the polymerase cannot unstack the layers. In Set III, the GQ loop length of three-layered template was varied such that there were one, three, five and seven nucleotides in the loop (Figure 2.2C). This set tests whether loop length affects how well a polymerase can unstack three-layered GQs. Set IV consisted of two well-characterized thrombin-binding aptamers (29-mer[75] and 15-mer[76]). This set tests whether the synthetic sets

provide good predictions for characterized aptamers. As a platform performance control, an unstructured template was synthesized, to serve as a positive control for sample preparation and sequencing accuracy. All of these templates were incorporated in approximately equal proportions into each sequencing library. As a final note, another significant variable affecting stability is the loop sequence[71,77]; however, given the large number of possible sequences formed for a given loop length, that variable has not been addressed in the current sequence-independent experiments. The loop sequence consisted of thymine nucleotides to provide a balanced GC-content for the templates, except in the case of Loop7 where an adenine was inserted in the middle to circumvent Ion Torrent's known homopolymer sequencing problem[78].

The sequences of all the templates are listed in Table 2.1. Each template is bounded by Ion Torrent PGM adapter sequences (one on each side), as: 5'-CCTCTCTATGGGCAGTCGGTGAT - (Target Sequence) - CTGACTGAGTCGGAG ACACGCAGGGATGAGATGG -3'. In the target sequence, a distinct two nucleotide recognition "key" was incorporated for each template; this served as a bar code to sort the sequencing output for each different template in the library. All oligonucleotides were purchased from Bio Basic Inc. (ON, Canada); templates were HPLC-purified. Each lyophilized template sample was reconstituted with the manufacturer's specified amount of DNA suspension buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0) to produce a 100 uM stock solution.

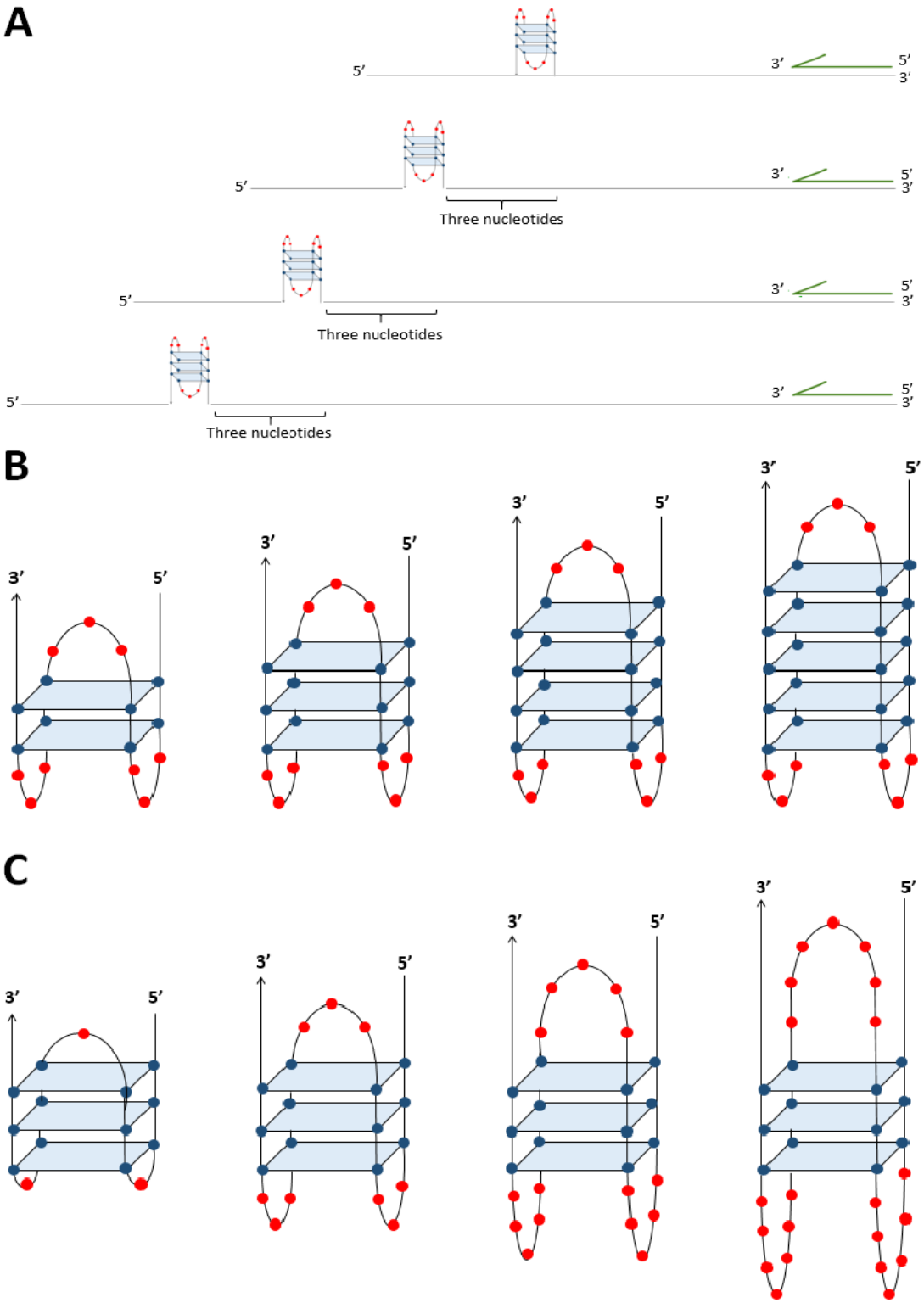


Figure 2.2: (A) The Set I of templates in which the GQ position is successively moved three nucleotides downstream from the sequencing primer (green arrow). (B) The Set II of templates with GQs having two, three, four and five tetrad layers (left to right) (C) The Set III of templates with GQs having one, three, five and seven nucleotides in loop (left to right).

Table 2.1 Library Composition – Oligonucleotide Sequences

Variable	Abbreviation	Sequence of Variable Region (5’ → 3’)	Number of GQ Layers	GQ Loop Length (Nucleotides)
3’ Overhang Length (Set I)	OH	GGGTTTGGGTTTGGGTTTGGGCT	3	3
	+3OH	GGGTTTGGGTTTGGGTTTGGGATTAA	3	3
	+6OH	GGGTTTGGGTTTGGGTTTGGGATTATTAT	3	3
	+9OH	GGGTTTGGGTTTGGGTTTGGGATTATTATTAC	3	3
Number of GQ Layers (Set II)	Layer2	GGTTTGGTTTGGTTTGGCG	2	3
	Layer3	GGGTTTGGGTTTGGGTTTGGGCT	3	3
	Layer4	GGGGTTTGGGGTTTGGGGTTTGGGGCC	4	3
	Layer5	GGGGGTTTGGGGGTTTGGGGGTTTGGGGGCA	5	3
GQ Loop Length (Nucleotides) (Set III)	Loop1	GGGTGGGTGGGTGGGTA	3	1
	Loop3	GGGTTTGGGTTTGGGTTTGGGCT	3	3
	Loop5	GGGTTTTTGGGTTTTTGGGTTTTTGGGTT	3	5
	Loop7	GGGTTTCTTTGGGTTTCTTTGGGTTTCTTTGGGTC	3	7
Set IV	T29*	AGTCCGTGGTAGGGCAGGTTGGGGTGACTAG	2	2-4
	T15*	GGTTGGTGTGGTTGGTG	2	2-3
-	Control	AAACCGCCGGAGTGGTGGGAACCGCAGTTGTGT TAGTTGTAGCGCGACGG	0	0

* Thrombin aptamers

The Illumina MiSeq chemistry uses different adapter sequences than the Ion Torrent PGM. Hence, to make the library compatible with the Illumina platform, the above Ion Torrent templates were extended using 'bridge' primers to the Illumina adapter sequences, both primers were incorporated to produce Illumina templates using a two-step protocol. The 'bridge' primers, shown below, are complimentary on one side to the Ion Torrent adapter sequences on the templates, while on the other end they are complementary to the Illumina adapter sequences to be ligated in the second step.

5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCATCTCATCCCTGCGTGT-3'

5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTCTCTATGGGCAGTCGG-3'

For the first step, the bridge conversion reaction mix contained: 26 nM Ion Torrent Library (an equimolar mixture of all templates that was quantified by qPCR, described below), 5 uM each of the forward and reverse bridge primers and 1X KAPA HiFi HotStart ReadyMix (Kapa Biosystems, MA, USA) in a 0.2 mL thin-walled PCR tube. The thermocycling profile was: 3 min at 95°C, 30 cycles of 30s at 95°C, 30s at 55°C, and 30s at 72°C, followed by a final extension step of 5 min at 72°C. The PCR product was purified using the Agencourt AMPure XP system (Beckman Coulter, Inc., CA, USA). Briefly, the sample was incubated with 2.5X volume paramagnetic AMPure XP beads for 30 minutes, the supernatant was removed, followed by washing with 80% ethanol and then DNA elution in DNA suspension buffer. In the second step for conversion to Illumina templates, Illumina adapters, sequences that include primer binding sites, indices to track samples, and terminal sequences complimentary to the

sequencing flow cell, were ligated to the template created in first step. The reaction mix consisted of: 5 uL DNA product (155.8 ng/µL) from first step, 5 uL Nextera XT Index Primer 1 (N707), 5 uL Nextera XT Index Primer 2 (S507), 25 uL 2X KAPA HiFi HotStart ReadyMix (Kapa Biosystems, MA, USA) and 10 uL PCR-grade water. Reduced- cycle PCR amplification was carried out using a thermocycling profile of: 3 min at 95°C, 8 cycles of 30s at 95°C, 30s at 55°C and 30s at 72°C followed by a final extension step of 5 min at 72°C. The product obtained was purified again using AMPure XP beads as described above. The Illumina templates are thus longer than the Ion Torrent templates by the length of the Ion Torrent primers, and basically functioned as longer overhangs to the GQ structures.

2.2.2 Library Quantification

Accurate library quantification is of paramount importance to obtaining sequencing data, both for the Ion Torrent and the Illumina platforms. Libraries with lower concentration lead to a low number of reads and poor error statistics, whereas higher concentrations result in mixed, uninterpretable signals, called either polyclonal ISPs (Ion Torrent) or overlapping clusters (Illumina). The library quantification protocol and kit from Kapa Biosystems (MA, USA, KAPA Library Quantification Kit) was used, as per the manufacturer's guidelines for the type of platform and type of qPCR instrument available. Briefly, 12 µL of the KAPA SYBR® FAST qPCR Master Mix was mixed with 4 µL oligonucleotide solution and 4 µL PCR-grade water. Similar reaction mixtures were prepared for each of the six, 10-fold serially diluted, standard solutions provided in the kit. Each solution was prepared in triplicate for both the standard and test solutions. Solutions were added into the wells of a qPCR plate (Biorad, CA, USA), the plate was

sealed with an optical tape (Biorad, CA, USA) and loaded in a BioRad MyiQ™ Single-Color Real-Time PCR Detection System (Biorad, CA, USA). The thermocycling profile was: 5 min at 95°C, then 35 cycles of 30s at 95°C, 45s at 60°C was used. This was followed by melt-curve analysis wherein the temperature was raised from 55°C to 95°C at 0.5°C/min.

The MyiQ™ software was used to derive the threshold cycle for crossing the response detection limit (Ct) and, based on the standards, dilutions for making test solution concentrations correctly for sequencing. Melt curve profiles were examined to ensure that we obtained a single, sharp peak, which implies the absence of any reaction by-products. All of the templates were then pooled together, in equimolar proportions, to produce a final library having a concentration of 26 pM (Ion Torrent) and 4nM for the Illumina library.

2.2.3 Ion Torrent PGM Sequencing

Template-positive ISPs were prepared with the Ion PGM™ Template OT2 400 Kit using Ion OneTouch™ 2 System (Thermo Fisher Scientific, MA, USA). The OneTouch2 was set up according to the manufacturer's protocol. At the end of the run (~ 8 hours), both Ion OneTouch™ Recovery Tubes were taken out of the instrument and all but 100 μL of the Recovery Solution was removed from each tube. The ISPs in the remaining 100 μL solution were dispersed in the solution and processed according to the manufacturer's instructions. The recommended Ion Sphere™ Quality Control assay was performed at this stage to ensure that the proportion of templated ISPs are in the acceptable range of 10–30%.

Enrichment of the ISPs was done using the Ion OneTouch™ ES (enrichment system), according to the manufacturer's directions. At the end of the run, enriched ISPs were collected in the PCR tube with Neutralization Solution and stored until ready for sequencing.

Before sequencing, the PGM was cleaned and, then the run was set up and initialized, as per the manufacturer's protocol for using the Ion PGM™ Sequencing 200 Kit v2. The internal control requires adding 5 µL of Control Ion Sphere™ Particles (these contain calibration sequences used by the software to determine accuracy of base calls and loading statistics) to the entire volume of ISPs obtained after the enrichment step, for approximately 5% representation. The solution was processed according to the manufacturer's instructions, then 12 µL of the Sequencing Primer was added, annealed and then 3 µL of Ion PGM™ Sequencing 200 v2 Polymerase was added and incubated at room temperature for 5 minutes. An Ion 316™ Chip was loaded with the 30 µL of the above solution, centrifuged as per the manufacturer's guidelines to ensure uniform loading of the chip with ISPs. The excess solution was then removed from the loading port and the chip was placed in the chip socket of PGM. Sequencing was carried out for 250 flows, for which the expected mean read-length is 130 bases- enough to fully cover the lengths of all expected templates in the library (40- 73 bases).

2.2.4 Illumina MiSeq Sequencing

For Illumina MiSeq processing we used the MiSeq Reagent Kits v3, paired-end protocol. Reagents were handled according to the manufacturer's protocol unless otherwise specified. The flow cell was cleaned and the instrument was set up as per the manufacturer's instructions; bidirectional sequencing run was started for 251 cycles.

2.2.4 Data Analysis

Both Ion Torrent and Illumina sequencing data are provided in FASTQ file format,[79] which contains information about the location of the read, the read sequence and quality score for each base in the read. Each FASTQ file was processed to extract the needed information for statistical analysis using the in-house programs written in Python and R[80] languages (Appendix A).

Routinely, the low quality reads and/or base-calls are omitted from the sequencing output by the platform-supplied software, before further assembly and annotation. However, for the purpose of this study, such low-quality base-calls were retained. The sequences for each template were binned based on the starting two nucleotide sequence, unique for each template. For analysis of both Ion Torrent and Illumina sequencing data, only the part of the sequence that contained the GQ and the downstream 23-nucleotide truncated P1 (trP1) adapter was used, unless indicated otherwise. The full sequences and lengths of templates used for Illumina and Ion Torrent platforms are listed in Appendix B.

2.3 Results and Discussion

Sequencing accuracy has been assessed using two metrics, viz. base-call quality scores and base-calling errors (called % mismatches) to the known template. During sequencing, whenever a base is identified (called) in a read, the sequencer also provides an associated quality score (Q)[81,82] which is a measure of the projected accuracy in calling that base. Mathematically, the quality score is defined as:

$$Q = -10\log_{10}(e)$$

where e is the estimated probability of an incorrect base call. Note that e is defined somewhat differently by Ion Torrent and by Illumina because they have different characteristic errors and so use different error functions, therefore quality scores are not directly comparable although the overall rate of error is benchmarked to be the same. A higher quality score implies a low probability of an incorrect base call and *vice versa* (Table 2.2). For the percent miscalls, because the sequence of each HPLC-purified templates under study is known, the reads are aligned to the actual sequence using the first two "key" nucleotides in the sequence, and from there the percent mismatches that accumulate along the read are quantified.

Table 2.2 Quality Scores and Error Probabilities

Quality Score	Probability of Incorrect Base Call	Accuracy of Base Call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

The number of reads obtained for each template, after removing the polyclonal and low quality reads for Ion Torrent and number of passing filter reads for Illumina platform are listed in Table 2.3

Table 2.3: Distribution of Ion Torrent and Illumina sequencing reads across the templates

	Ion Torrent	Illumina
Layer2	186,206	814,662
Layer3/Loop3/ OH	44,688	947,769
Layer4	35,357	737,541
Layer5	1,859	436,972
Loop1	26,844	553,291
Loop5	19,765	499,051
Loop7	291,491	1,157,366
+3OH	19,837	455,732
+6OH	103,122	507,501
+9OH	86,434	427,539
T15	505,467	1,471,511
T29	849,630	2,279,909
Control	252,587	1,041,380

2.3.1 Results for Target Set I: Tracking GQ Induced Effects on Sequencing with Varying 3’ Overhang Lengths

The quality score trends of Ion Torrent and Illumina platforms in sequencing templates with varying 3’ overhang length are shown in Figure 2.3. For the Ion Torrent, it can be seen that the quality scores for the unstructured, control template remain consistent and high across the sequence, while those for GQ-containing templates rapidly drop below 20, soon after beginning of sequencing (Figure 2.3A). A closer look at the trends in Figure 2.3B reveals that the position where quality scores start to decline for GQ-containing templates coincides with the GQ start site, i.e., where there is a longer 3’

overhang length there is also more high-quality sequence before the drop-off. This supports the inference that the GQ structure causes the quality score decay. On the Illumina platform, on the other hand, sequence quality continues to be high for both the control template and GQ containing 3’ overhang templates (Figure 2.3C). Stable quality scores are seen throughout until at the end where scores naturally drop when the polymerase reaches the end of a template, including on the control template. Figure 2.3D reveals no striking differences in quality score trends either.

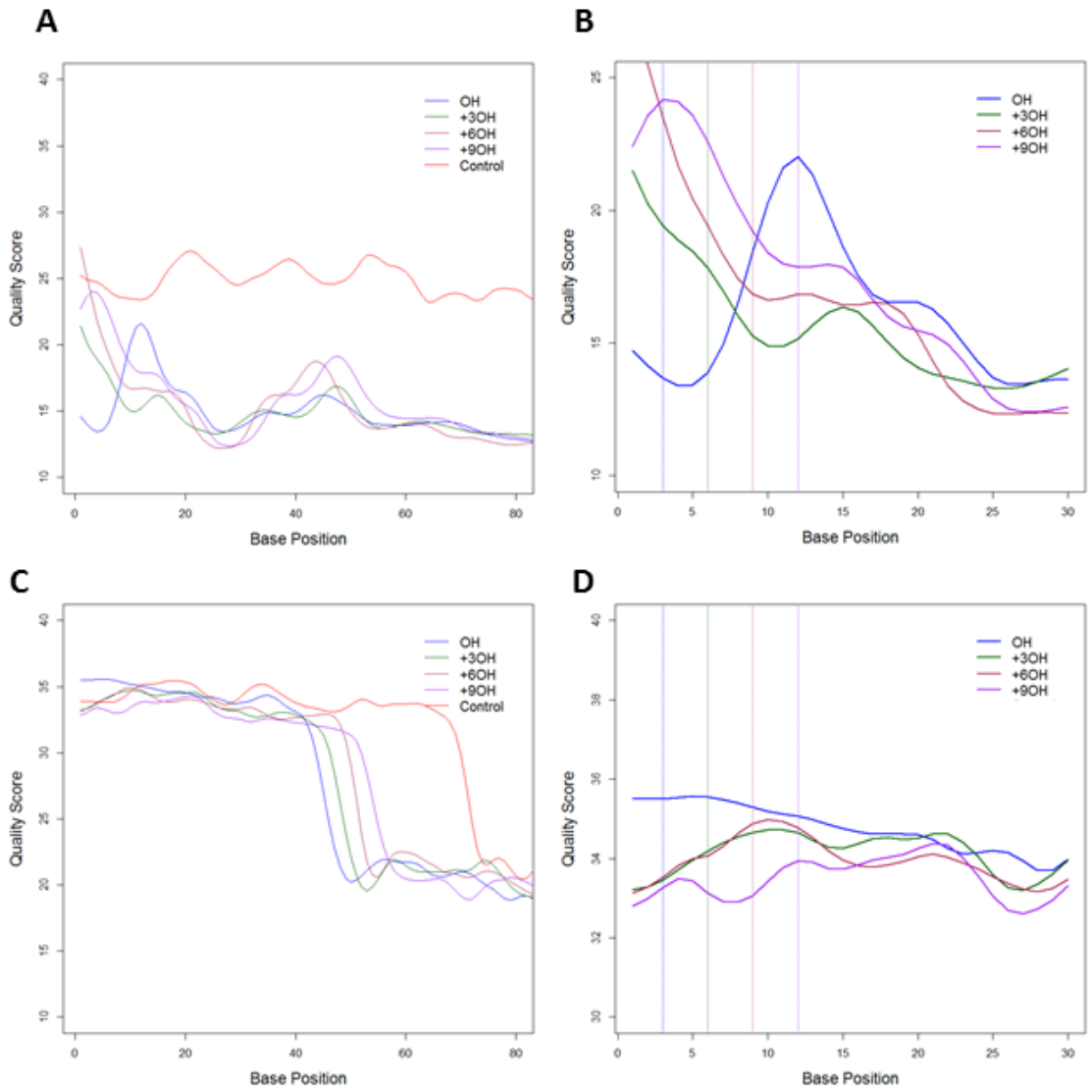


Figure 2.3: (A) and (C) Quality score trends of templates with different 3' overhang length, by Ion Torrent and Illumina respectively. (B) and (D) Enlarged views of plots (A) and (C) respectively, focused on GQ start sites- indicated by the vertical dotted lines. OH, +3OH, +6OH and +9OH represent templates in which GQ position was successively moved three nucleotides downstream (OH = Overhang).

In another visualization of the results, Figure 2.4 is a heat-map that displays position-wise base mismatches incurred during the sequencing. For sequences obtained on the Ion Torrent, the mismatches become prominent within one nucleotide immediately after the first GQ start site and, absent a recalibration step allowing gaps (not a standard

procedure for resequencing projects), continue to accumulate throughout the reads till the end, making the plot progressively more red from left to right (start to end of the sequence). For data acquired using the Illumina MiSeq, no noticeable accumulation of errors can be seen across the length of the sequences.

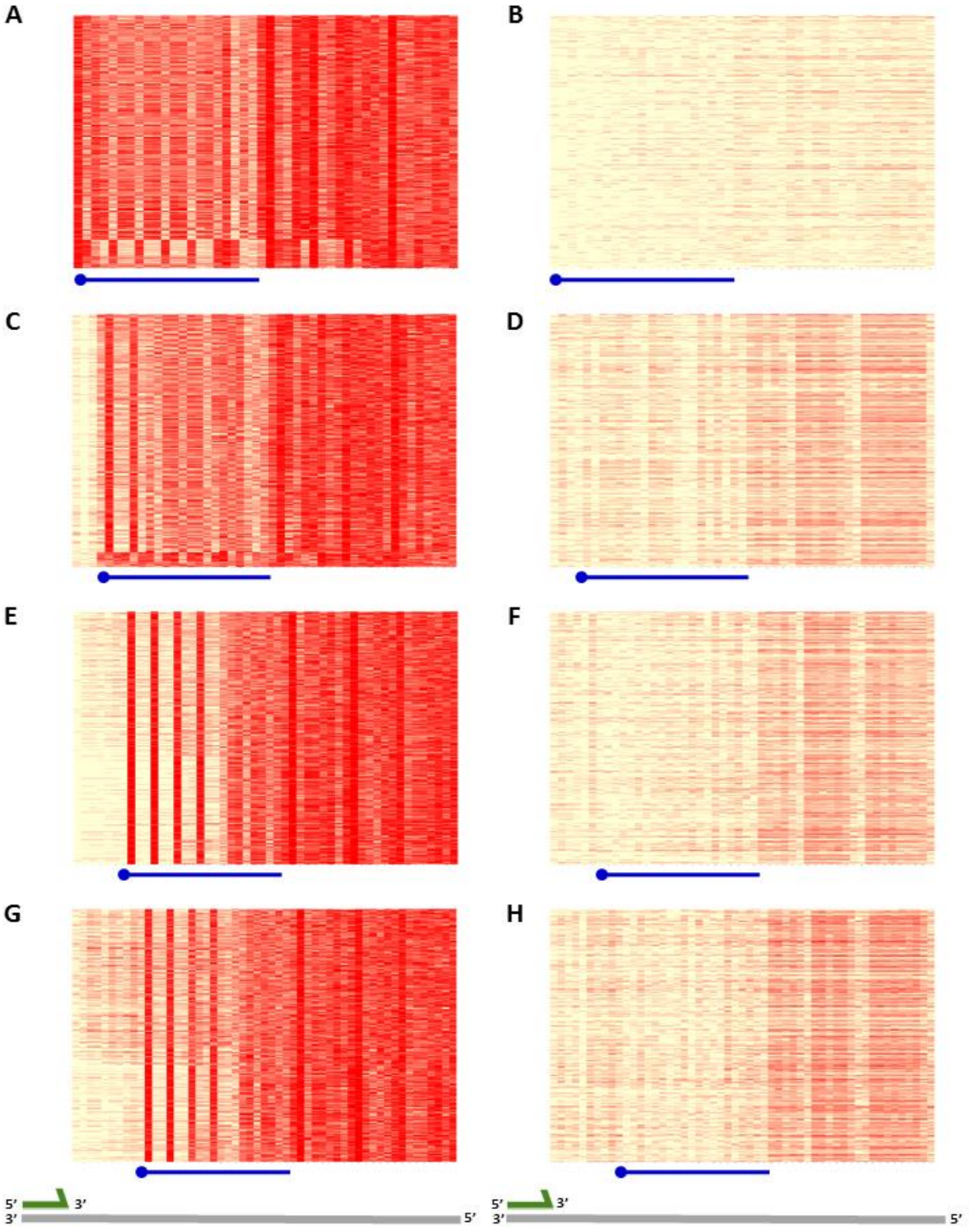


Figure 2.4: Position-wise base call errors, shown as a heatmap plot, with red indicating incorrect base-calls. (A), (C), (E), (G) represent plots for OH, +3OH, +6OH, +9OH templates sequenced by Ion Torrent, while (B), (D), (F), (H) represent the same respective templates sequenced by Illumina. Each horizontal line represents an independent read. Each column is a base position in the template. The horizontal blue line below each plot indicates the part of the sequence that forms GQ, with the first GQ base shown as a circle.

The above base mismatches results are quantitatively represented as a boxplot in Figure 2.5. As Figure 2.5A shows, the overall sequencing accuracy of the Ion Torrent platform is severely affected by the polymerase's encounter with a GQ structure, whereas the Illumina polymerase appears to be able to process through GQ structures. While a gradual increase in errors is expected due to a relatively longer 'After GQ' region than 'Before GQ' region, when there is an intervening structured region the increase is both abrupt and large. Given the large sample sizes, *p* values calculated by the two-sample t-test are <0.001, for all the comparisons performed in this work, indicating the differences are statistically significant.

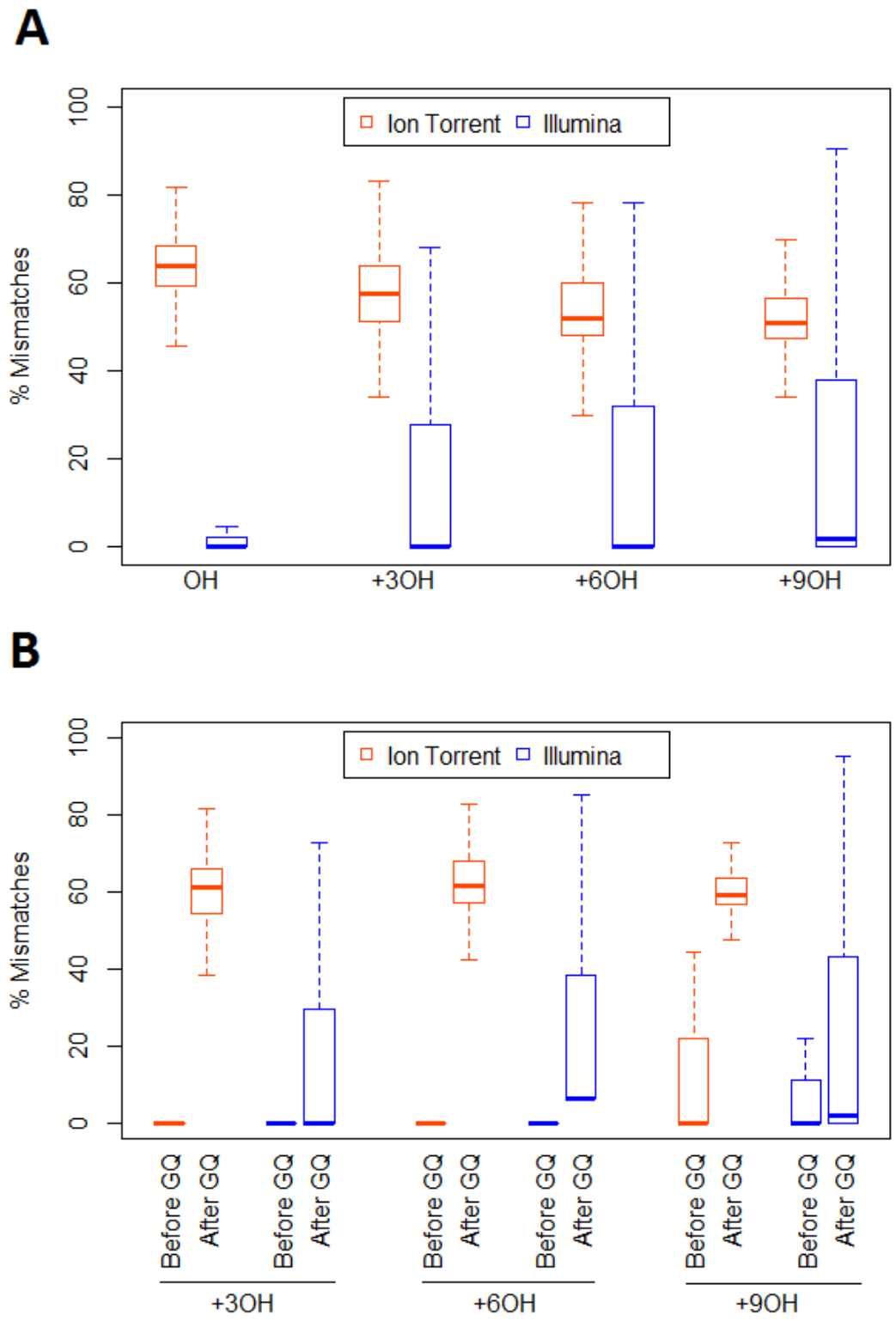


Figure 2.5 Boxplots representing (A) Percent base call errors by Ion Torrent and Illumina for the entire templates. (B) Percent base call errors before and after start of a GQ-sequence in a template. 'After GQ' sequence includes the GQ-forming sequence in addition to the downstream sequence.

2.3.2 Results for Target Set II: Effect of GQ Layers on Sequencing Accuracy

As mentioned above, the software provided with the Ion Torrent and Illumina systems calculates the respective quality scores, but because the chemistries and detection signal characteristics are different the models and parameters are not identical. Thus, although the overall report for errors is the same for the same score, the detailed interpretation as to the type of error is not the same. There is also some agreement that the Ion Torrent algorithm tends to report lower quality scores than does the Illumina algorithm (biolektures.wordpress.com/2011/09/05/ion-torrent-qv-prediction-algorithm/). In the following summary graphs, the quality scores for each of the platforms were normalized using the respective quality score for the unstructured control template. The absolute quality score values are shown in Appendix C. In Figure 2.6 the overall effect of increasing numbers of GQ layers on sequence platform ability is shown.

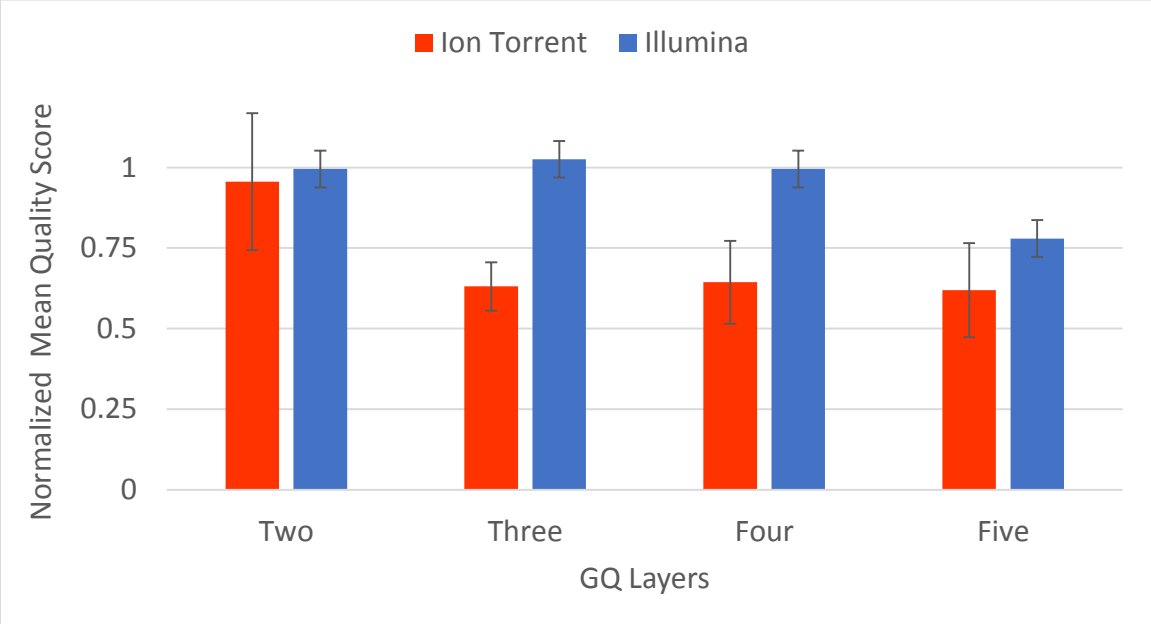


Figure 2.6: Normalized mean quality scores of Ion Torrent and Illumina for templates with increasing numbers of GQ layers.

It is apparent from the figure that on the Illumina platform the templates are readily sequenced until there are five layers of GQs, at which point the average quality score decreases. On the Ion Torrent platform, on the other hand, acceptable quality scores can be obtained only for two-layered GQ set members. Thus, some aspect of the chemistry (perhaps the polymerase used, or the presence of a denaturant) or platform (perhaps the temperature) allows the Illumina system to better process templates containing these types of structures.

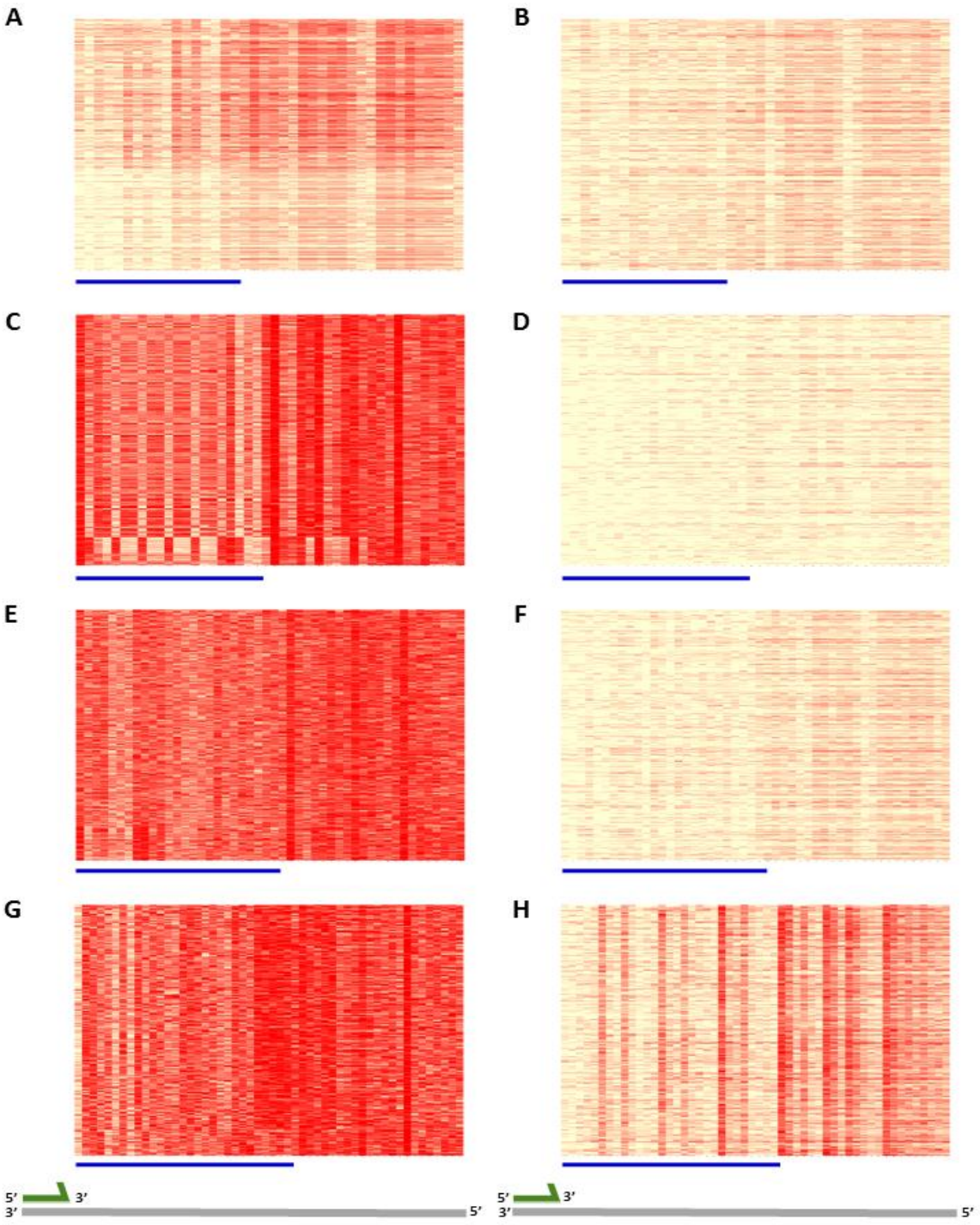


Figure 2.7: Position-wise base call errors, shown as a heatmap plot, with red indicating incorrect base-calls. (A), (C), (E), (G) represent plots for two, three, four and five layered GQ templates sequenced by Ion Torrent. (B), (D), (F), (H) represent the equivalent template plots sequenced by Illumina. Each horizontal line represents an independent read. Each column is a base position in the template. Mismatches are shown in red. The horizontal blue line below each plot indicates total GQ forming sequence.

The occurrence of mismatches at each position in individual reads stacked as a heat map have been depicted in Figure 2.7. As a general observation there is always more red (base call error) in the Ion Torrent images than for the equivalent Illumina images. After encountering a GQ with more than two layers, base mismatches appear at a high frequency in Ion Torrent data. Another observation is that there is a slight difference in where base-calling errors start to accumulate for recalcitrant templates: they start with the first base of a GQ for Ion Torrent data, but downstream by a couple of bases in Illumina data. These base mismatches results are quantitatively represented in the boxplots shown in Figure 2.8.

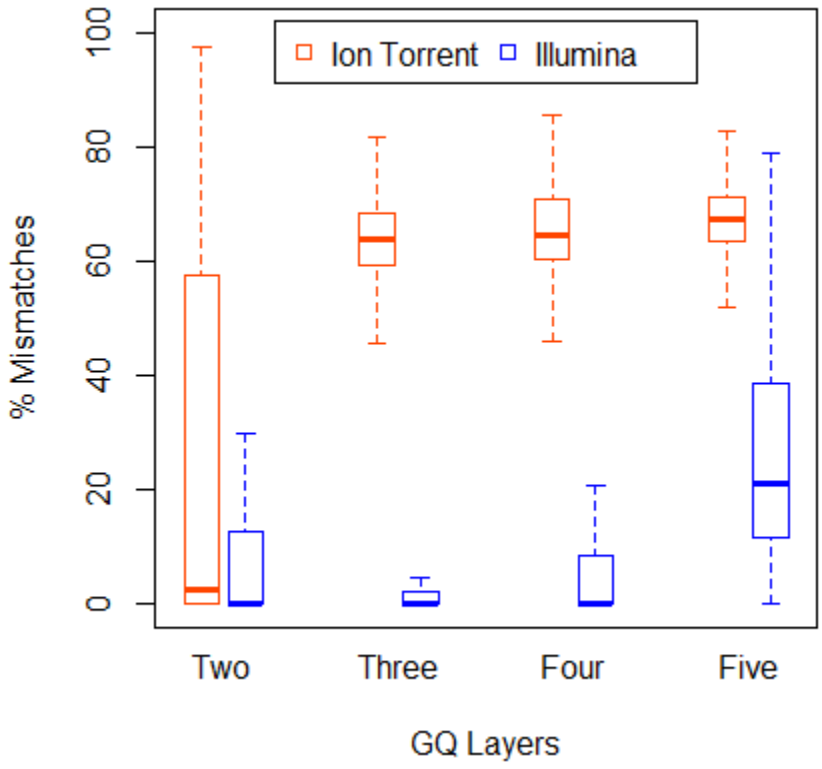


Figure 2.8: Boxplots representing percent total base calling errors for data acquired using the Ion Torrent (red) or the Illumina (blue) platform, for targets having increasing numbers of GQ layers, as shown.

2.3.3 Results for Target Set III: Effect of GQ Loop Length on Sequencing Accuracy

The loop is the sequence that connects the GQ layers, and the length may constrain the flexibility of the structure. As a reminder, all of the templates in this set contain three GQ layers. For data acquired using the Illumina platform, the mean quality score is indistinguishable from the control sequence means when the loop has three, five and seven nucleotides; the mean quality score drops only with the most constrained structure, in which there is only one nucleotide in the loop (Figure 2.9). For data acquired using cognate targets on the Ion Torrent platform, on the other hand, only the most flexible, seven-nucleotide, loop has an acceptable mean quality score. Overall, there are considerable differences in two platforms' mean quality scores when loops are three, five and seven nucleotides in length.

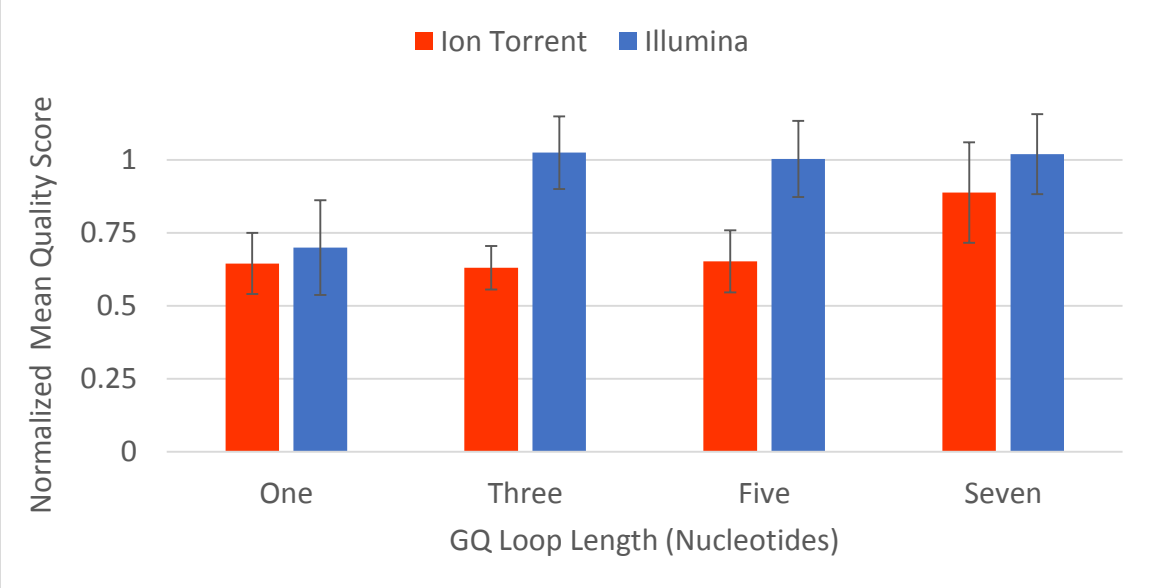


Figure 2.9: Normalized mean quality scores for Target Set III, based on data acquired using the Ion Torrent (red) and Illumina (blue) platforms, where targets have three GQ layers and loops of length one, three, five and seven nucleotides.

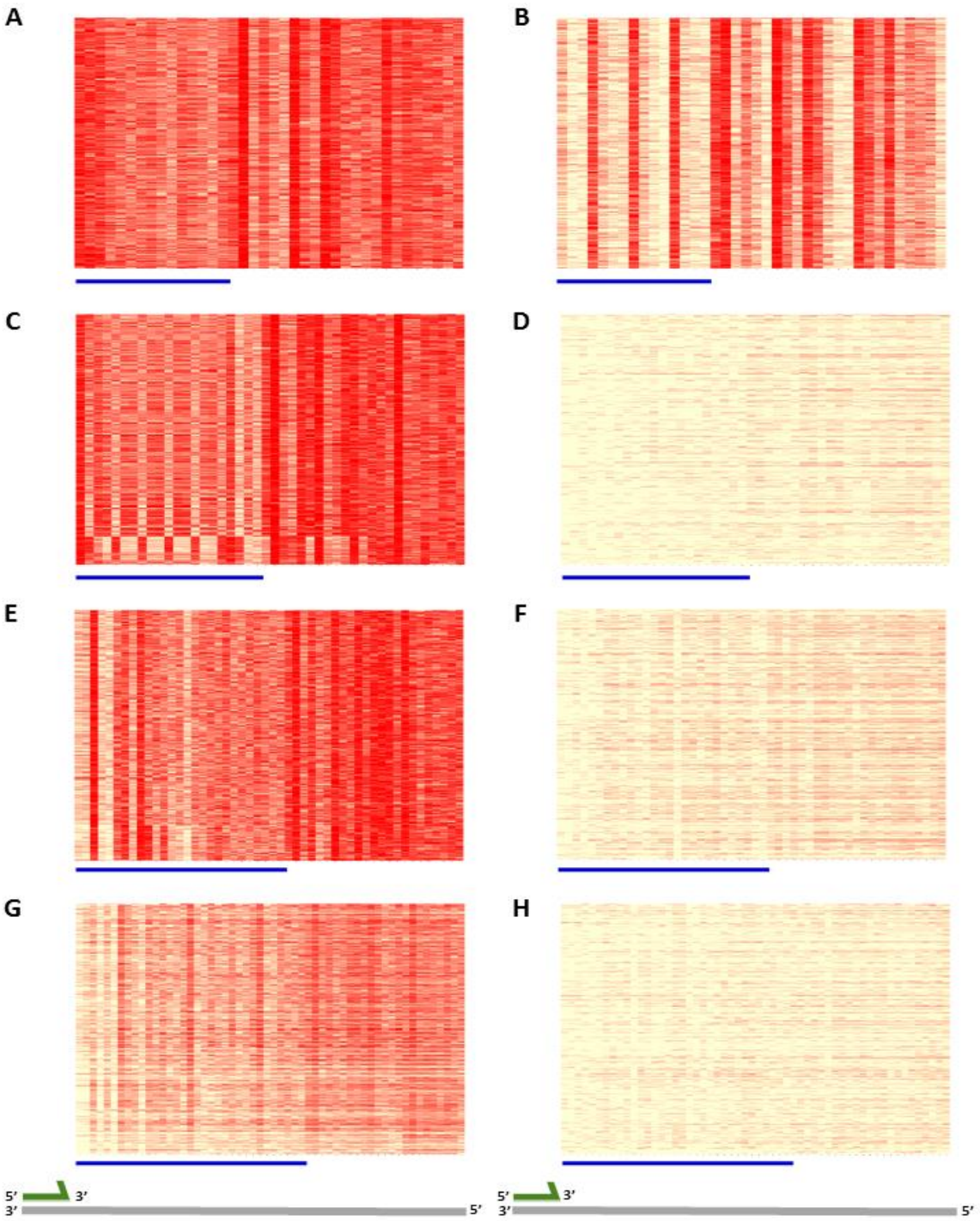


Figure 2.10: Position-wise base call errors, shown as a heatmap plot, with red indicating incorrect base-calls. (A), (C), (E), (G) represent plots for three-layered GQ templates with one, three, five and seven nucleotides in the loops, sequenced by the Ion Torrent. (B), (D), (F), (H) represent the equivalent templates sequenced by the Illumina MiSeq. Each horizontal line represents an independent read. Each column is a base position in the template. Base call errors are shown in red. Horizontal blue line below each plot indicates the three-layered GQ forming region.

Heat maps showing position-wise base calling errors for all of the sequences in the data set have been depicted in Figure 2.10. Similar to the trends for Set II, Ion Torrent heat maps are redder overall than those depicting data acquired with Illumina. Once the polymerase encounters a GQ structure, base calling errors begin to appear in Ion Torrent data. Longer loops are correlated with fewer errors on both systems. As seen with template Set II, the base-calling errors start appearing from the first base of the first GQ layer on the Ion Torrent and a few bases downstream from that for Illumina. The mean base call error results are quantitatively represented in the boxplot in Figure 2.11.

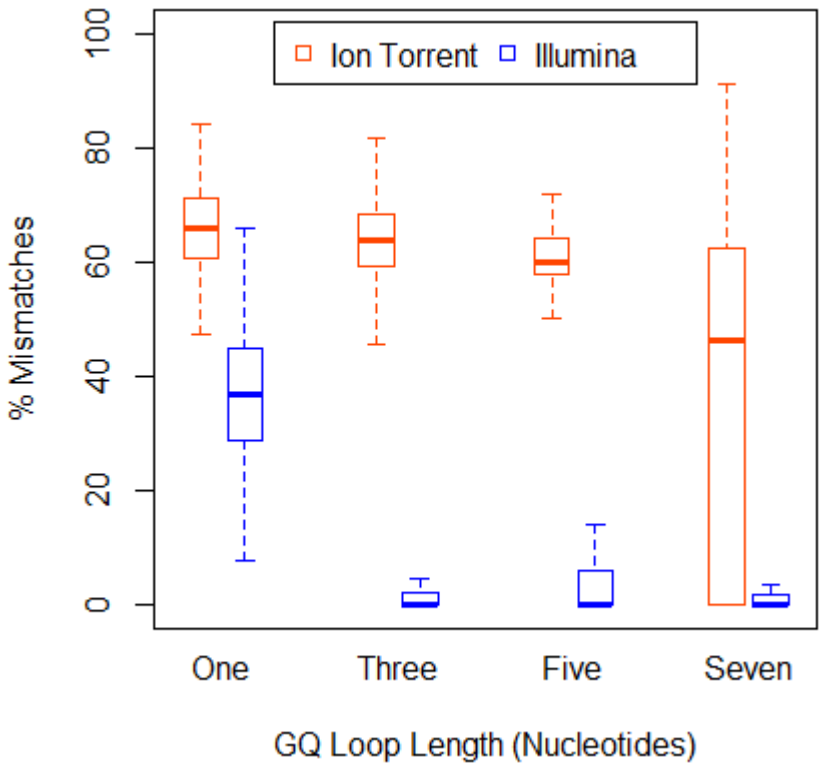
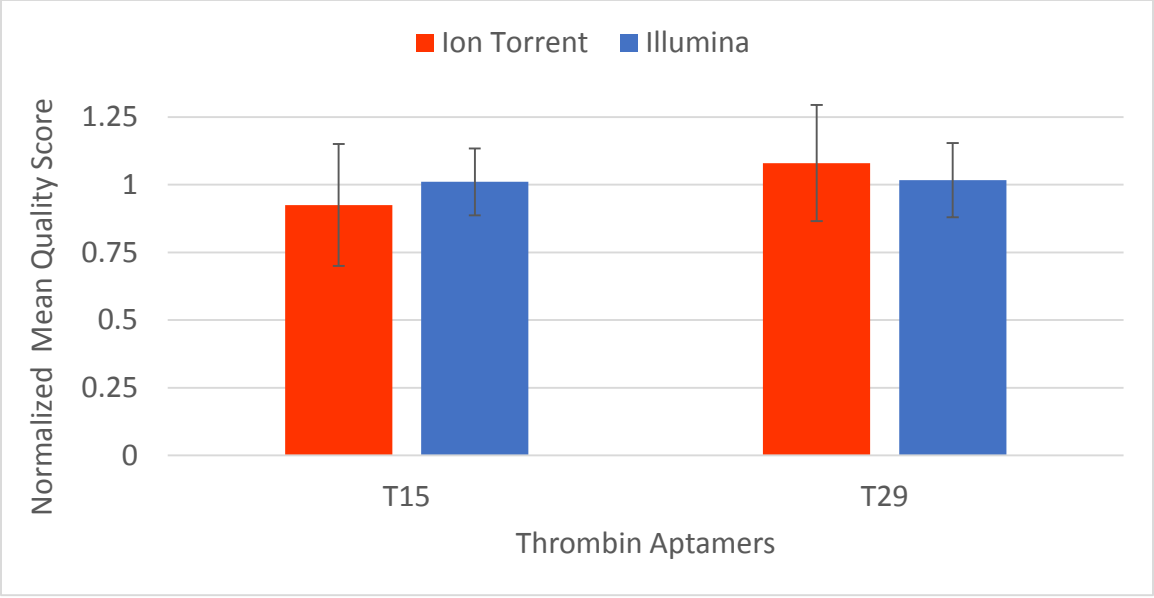


Figure 2.11: Boxplots representing mean percent base error calls for each target in Set III, comparing Ion Torrent data (red) and Illumina data (blue) for three-layered GQs with the indicated number of nucleotides in the loops.

2.2.4 Results for Target Set IV: Sequenceability for Thrombin Aptamers

The previous experiments used systematically varied templates to assess how general template structures limit data quality on these two sequencing systems. However, these structured templates are not known to bind any proteins and hence are not candidates for multiplex assays, which is our end goal. Therefore we added two sequences containing a GQ that have been well characterized in binding the small protein thrombin. The thrombin aptamer 'T15' has two GQ layers and loop lengths of two to three nucleotides, while the aptamer 'T29' has two GQ layers but loop lengths of two to four nucleotides. As shown in Figure 2.12, for both targets the Illumina quality scores are identical to the unstructured control sequence. For these templates the Ion Torrent platform seems to perform well (although it appears to do even better than the control sequence in one case, likely due to shorter template length than control sequence). These results are consistent with the results found using our test set templates, with two GQ layers and loops greater than one nucleotide.



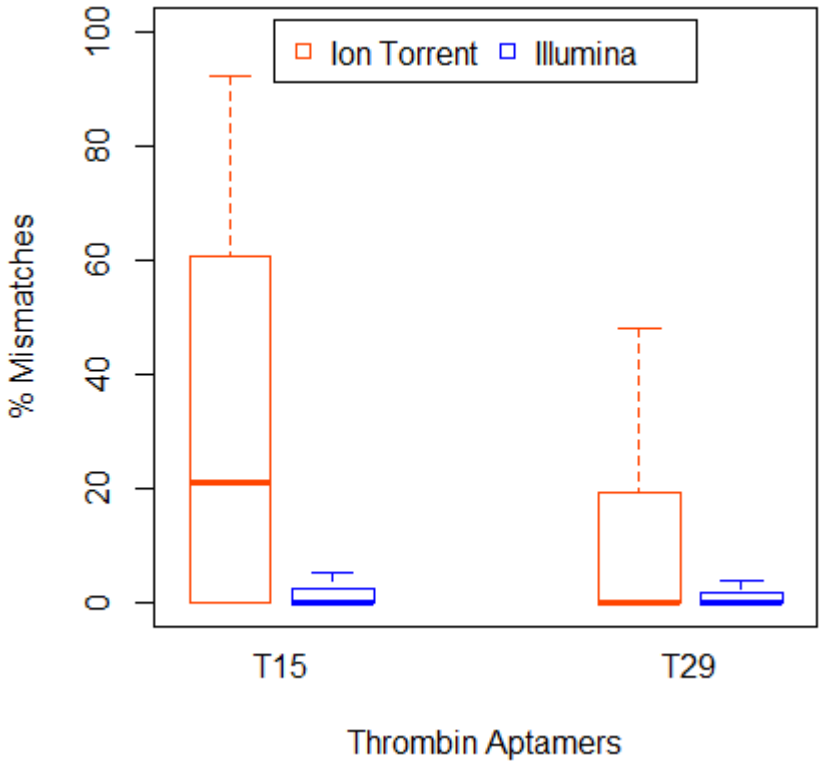


Figure 2.12: (Top) Normalized mean quality scores of thrombin aptamers T15 and T29, data acquired on either the Ion Torrent (red) or Illumina (blue) platforms. (Bottom) Boxplot representing mean percent mismatches for thrombin aptamers T15 and T29, data acquired using Ion Torrent (red) or Illumina (blue) platforms.

2.2.5 Template Structural Complexity- A Reason for Polyclonal ISPs?

The Ion Torrent data analysis software is an algorithm pipeline that performs a filtering process on reads before generating the final sequencing data set that is provided as output. An important step in the process is to recognize and remove those reads that contain more than one sequence: in this case there will be very high base call errors because two different bases are being incorporated in the same flow set (which should only happen for homopolymers), which muddles the result. The physical process of sampling dictates that a certain number of ISPs will be captured in emulsion reactors with zero, one, two or more template molecules, but the reagent proportions and subsequent clean-up steps aim to optimize the number with only one template per ISP. It is known that using a higher ratio of template library to beads will force the distribution to having more polyclonal ISPs.

In the sample to library preparation steps carried out in these experiments, we have consistently observed that the software recognizes a very high percentage of polyclonal ISPs (>60%) in the sequencing output, a number that does not decrease even using lower concentrations of library input for emPCR (attaching templates to ISPs). There is an initial control step for this process, (percent templated ISPs are calculated by an Ion Sphere™ Quality Control assay) for which our results consistently fell between 10-20%., indicating that only a small proportion of polyclonal ISPs in the final data set is expected, contrary to observations. In trying to determine why it was so difficult to get an adequate amount of data on the Ion Torrent for these experiments, one possibility is that the structures are creating polyclonal signals even though the ISP templates themselves are not polyclonal. Figure 2.13 shows heat map of the base calls accumulating across a

set of reads for a template in the set that contains a GQ (+9OH) compared to the unstructured control sequence where different bases are represented by different colors. For the control sequence, base calls are made with high fidelity throughout the reads, and the proportion of ISPs reflects the fraction added to the library. Remember that, for a GQ forming template with more than one layer, the Ion Torrent reads out 'scrambled' sequence from the start of the GQ structure onwards. As with the other heat maps, each horizontal line represents an independent read, which is the combined output of the template on an ISP rendered as a base call – the expected result of emulsion PCR is that about one million polymerization reactions occurring on a single ISP (which ends up in a chip well). If different bases are incorporated at identical positions on an ISP with the same base sequence but a structure that inhibits consistent polymerase processing, then different base calls may be made at identical positions. This will lead to that ISP being labelled as polyclonal, and the data will be filtered out. Providing stronger support for this hypothesis would require performing sequencing runs with libraries composed of just one template, with libraries of GQ-forming templates compared to sequence-balanced and length balanced but unstructured templates. Then the percentage of polyclonal ISPs in each sequencing output could be quantified and a correlation between the number of polyclonal ISPs and the structural complexity in the template could be analyzed; GQ-forming templates are expected to yield a higher polyclonal ISP percentage. However, given the large costs and labor needed for such a study, this problem has not been further explored in this dissertation.

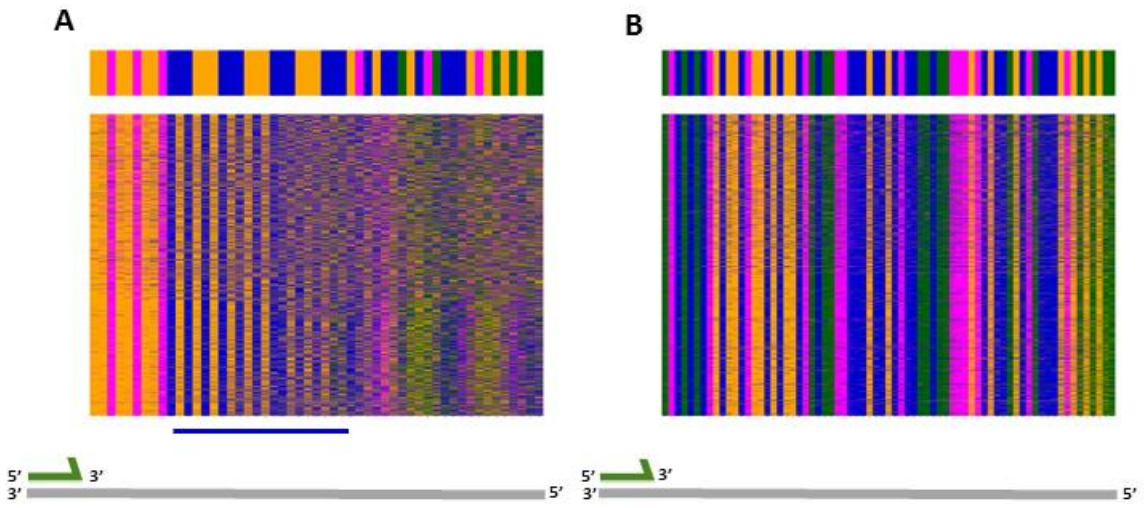


Figure 2.13: A heat map in which bases as called in an Ion Torrent sequencing run are individually colored for independent reads (horizontal lines) with a top horizontal bar indicating the true (expected) template sequence- each column is a base position in the template (A) the GQ structured template +9OH (B) The unstructured Control template. Colors are A= orange, T= magenta, G= green, C= blue. The first orange column in +9OH corresponds to two 'A' nucleotides (total 53 nucleotides) while the first green column in Control template corresponds to one 'G' nucleotide (total 71 nucleotides). The horizontal blue line below plot (A) indicates the location of GQ forming sequence. The gray line below each plot shows the full template length with the green band indicating the sequencing primer-binding site (where incorporation does not occur).

This hypothesis is further supported by the % representation of different templates in the final sequencing output (Table 2.4). The % representation was calculated as a percentage of a template to the total unfiltered reads (including polyclonal and low-quality reads). Because all the templates were in approximately equimolar proportions in the library, theoretically the % representation of each template should be ~7.5% (After quantitation, each template was added to the library mixture in approximately equimolar amounts: because concentrated samples are highly diluted in this step fairly large errors are expected, but not order of magnitude differences). A trend can be seen where the % representation decreases with increasing structural stability of the template.

Table 2.4 Representation of different templates in final sequencing output.

Template	% of total ISPs
Layer2	1.58
Layer3/Loop3/ OH	0.38
Layer4	0.30
Layer5	0.02
Loop1	0.23
Loop5	0.17
Loop7	2.47
+3OH	0.17
+6OH	0.87
+9OH	0.73
T15	4.29
T29	7.20
Control	2.14

2.4 Conclusion

The performance of the two most popular NGS platforms, Ion Torrent and Illumina, in sequencing systematically varied GQ-structured templates has been compared by studying mean base-call quality scores and individual base call errors. A trend was observed that, as the number of GQ layers increases, and hence the overall thermodynamic stability of the template, the more poorly the sequencing performance, as reflected in the accuracy and base quality after a certain structural stability break point. The Illumina MiSeq was able to accurately sequence most structured templates, while the Ion Torrent was less effective in nearly all cases.

The results obtained raise questions about the suitability of using Ion Torrent platforms in general aptamer selection protocols, for which the individual occurrence of structure is unknown or known only in a very general way. The sequencing software may remove ISPs with structured sequences, making them effectively invisible, while those structured sequences that survive the filtering may not have accurate base calls or at best report low quality scores that would lead to sequence deprecation.

Although our focus is on aptamers and their use in assays, there are also implications of this work for genome sequencing. Protein-binding nucleic acids often form a structure, and some are known to contain GQ layers. If such sequences are routinely problematic for the platform they will not appear in the data set and therefore cannot be included in an assembly data set. This will cause important regulatory sequences to be missing when the sequences connecting regulatory networks are sought. While using multiple platforms and reads is a recommended process for many reasons, it is important that researchers using these platforms remain aware of characteristically missing data elements and their likely importance. While the complementary strand to a GQ-structured sequence may be less refractory to the polymerase, in our lab other types of structured sequence have not proved to be more amenable to sequencing as the stability of the structure is not greatly different (Khoshnevis dissertation, unpublished data). In any case, the Ion Torrent does not provide a paired-end sequence method so it would be difficult to devise an approach that would allow every strand to be sequenced in both orientations in equal numbers and be sure of correctly pairing them given a high number of errors in one direction.

CHAPTER 3: STUDY OF CAUSES AND POSSIBLE SOLUTIONS FOR ION TORRENT GQ SEQUENCING ISSUES

3.1 Rationale

The suitability of an NGS platform for selecting DNA-based sensors rests on the ability to know directly the DNA sequence of successful baits. The poor sequencing performance seen in previous experiments, particularly by the Ion Torrent, in sequencing GQ-containing templates prompted us to investigate the causes for the phenomenon. It is a well-known fact that the polymerization efficiency and fidelity of many DNA polymerases is negatively affected by templates having a high GC content[83]. To determine if a similar trend exists with templates containing stable GQs and to understand which characteristics make a GQ structure difficult to sequence, the sequencing output of systematically varied GQ structures has been studied as a function of GQ melting temperature, T_m. Empirically, the GQ stability increases with additional tetrad layers and reducing the GQ loop size. To quantify this trend, and to confirm the formation of a GQ structure itself, circular dichroism experiments have been performed.

The Ion PGM Sequencing 200 Kit v2 includes a DNA polymerase from *Bacillus stearothermophilus* (*Bst*). *Bst* polymerase possesses a strong strand displacement activity[84], which means it has the ability to displace any downstream DNA, the polymerase encounters during complementary strand synthesis (Figure 3.1). This type of activity is usually reported in terms of the displacement of a strand already bound to the

template strand by Watson-Crick base pairing, where the polymerase 'peels off' the non-template strand as it processes. *Bst* polymerase has been demonstrated to unfold and process hairpin structures[85,86], one of the most stable structural features adopted by nucleic acids. Considering this, the error-prone results obtained by this polymerase in processing relatively less stable GQ templates, as shown in Chapter 2, are highly unexpected.

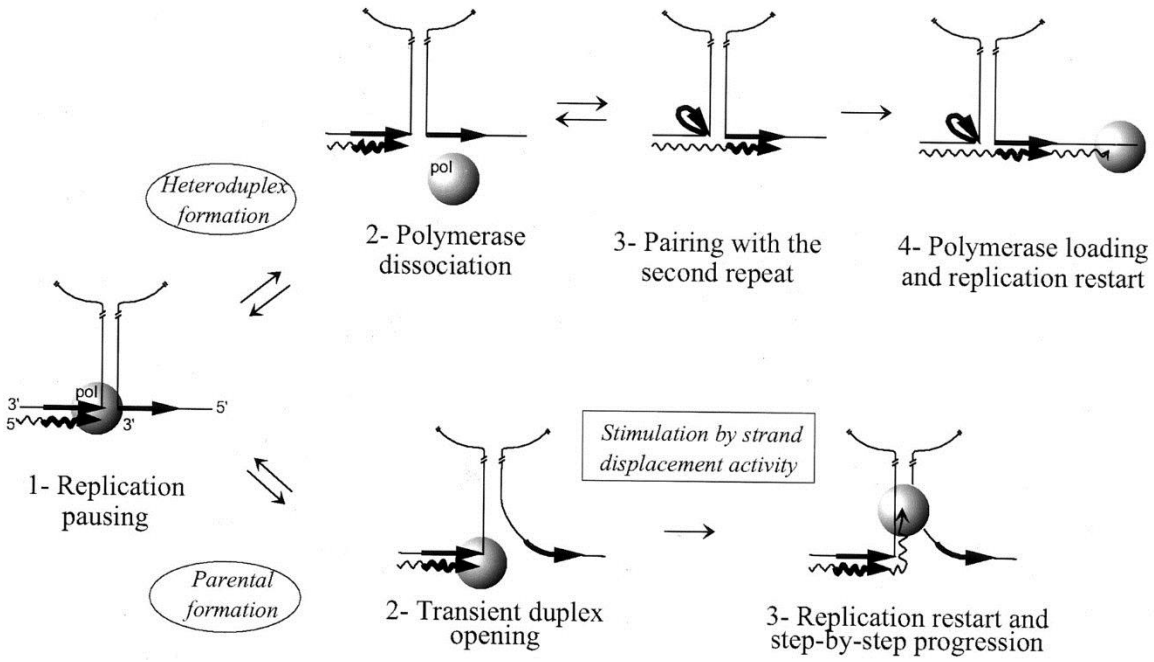


Figure 3.1: Model depicting two different paths (replication slippage or strand displacement) adopted by polymerase upon encountering a structure. Reproduced from reference[85].

We performed multiple sequence alignment (MSA) for all reads of the templates, using the actual sequence as a reference, to see if it gave insight into the mechanism. A sequence gap in the GQ region was a consistent feature observed for any template with a

T_m above the operating temperature of the particular platform, but the reads aligned with high consensus for sequence downstream of the GQ. This could result either from the phenomenon called 'replication slippage' or from a slowing of the polymerization rate to allow melting of the impeding structure. Each of these hypotheses are addressed in subsequent experiments.

Because templates difficult to sequence *in vitro* have been noted over many years, several approaches to ameliorate the problem have been documented, both chemical (such as denaturing agents like DMSO) and biochemical (proteins and metabolites). Single-stranded binding (SSB) proteins refer to a class of proteins that bind to single stranded DNA (ssDNA), found in both prokaryotes and eukaryotes, including humans. These are shown to play important roles in modulating various cellular processes by binding to ssDNA substrates and melting any secondary structures in the vicinity, and then preventing re-formation of the structures through their binding. Several extensive reviews of SSB proteins functions and modes of action are available[87-89]. SSB protein has demonstrated the ability to interact with and unfold GQ structures[69,90] and therefore has been tested as the first of the two compensatory additives to the reaction solutions, modifications aimed at improving sequencing accuracy. It was only possible to test this on the Ion Torrent platform, which has a more open architecture, allowing chemistry modifications.

In addition to the SSB protein approach, another modification utilized the Ion Torrent Hi-Q Sequencing Kit. In 2014, after screening over 10,000 potential polymerases mutants, for their ability for correct variant calling (measured in terms of sensitivity and positive predicate value, PPP), Ion Torrent introduced the Hi-Q sequencing polymerase.

The kit claims to offer "Robust variant detection- equivalent or greater insertion and deletion (indel) sequencing accuracy and reduced false positives compared to previous Ion Torrent™ chemistries observed across targeted re-sequencing panels." and "Higher quality *de novo* assemblies- up to 90% decrease in indel error rates for microbial sequencing" as advertised on the product's webpage. Although the polymerase is proprietary and hasn't been characterized independently, improved template processivity tuned to the detection frequency of the Ion Torrent platform appeared to be a promising improvement.

3.2 Methods

3.2.1 Circular Dichroism Spectroscopy and Thermal Denaturation

A JASCO J-720 spectropolarimeter equipped with a Peltier temperature control system (JASCO Corporation, TYO, Japan) was used to record spectra. The samples consisted of 5-10 μM of a GQ template in 1X buffer (pH adjusted Ion PGM™ Sequencing 200 v2 W2 Solution) remaining from a sequencing run. All the sample were incubated at 95 °C for 5 minutes and then gradually cooled down to room temperature over an hour, to induce GQ formation. For each template under study, a baseline subtracted spectrum was recorded, from 240 to 320 nm with a scan rate of 0.5 nm/s at 20 °C. The reported spectrum represent the average of three wavelength scans, processed by the Savitzky-Golay smoothing method[91]. The spectrum was examined for a positive maxima at 263 nm or 292 nm, a feature indicating the presence of a parallel or antiparallel GQ structure, respectively. Thermal denaturation experiments were carried out by heating the sample from 20 °C to 90 °C at the rate of 1 °C/ min and monitoring the

change in the ellipticity at a specified wavelength (263 nm for parallel GQ or 292 for antiparallel GQ).

3.2.2 Multiple Sequence Alignment (MSA) of Reads

For each template, either all the reads or a randomly selected subset of 10,000 reads (using the sample() function in R), were considered for MSA. MSA was performed using MAFFT (version 7) program[92]. Because the actual sequence was known, the following arguments were used to align the sequencing reads to the actual sequence:

% mafft --add sequencing_reads –keeplength actual_sequence > output

Additionally, a four nucleotide sequencing key 'TCAG', following by a two nucleotide template-specific key was added at 5' end, to each of the reads and the actual sequence, in order to "force" the MSA algorithm to align the reads to the actual sequence starting from the first nucleotide, thereby mimicking the actual, experimental scenario. It must also be noted that, because the template length gets extended by 18 nucleotides due to an extended adapter (P1 *vs* trP1) on ISPs, the following 18-nucleotide sequence was incorporated at 3' end of the actual sequence while performing MSA.

5' AAAGCGGAGGCGTAGTGG 3'

3.2.3 Primer Extension Assay

1 µL of 10 µM template was annealed with 1 µL of 100 µM of 5' 6-FAM labelled sequencing primer by heating the mixture at 95 °C for 2 minutes and then cooling to 37 °C for 2 minutes. To it, 1 µL 10 mM dNTPs, 1 µL of Ion Torrent Polymerase and 6 µL of Ion Torrent W2 buffer were added and the solution was incubated at 50 °C for 5 minutes and then at 80 °C for 20 minutes to inactivate the polymerase. 10 µL of Gel Loading Buffer II (Ambion- Thermo Fisher Scientific Inc, MA, USA) was added, the

solution was heated at 95 °C for 5 minutes and run on 8M Urea polyacrylamide gel (15%) until a marker dye reached the bottom. The gel was visualized with the GelDoc-It® Imaging System (UVP, CA, USA).

3.2.4 Ion Torrent PGM Sequencing

3.2.4.1 Using Single Stranded Binding (SSB) Protein

To determine an effective concentration ratio of SSB protein to DNA template, an electrophoretic mobility shift assay (EMSA) was performed. Starting solutions containing 20 µL of 31 nM Layer5 template in DNA suspension buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0) were made; the template was chosen, because this is one of the most stable and longest templates under study. To the DNA was added 20 uL of *E. coli* SSB protein (MCLAB, CA, USA), in concentrations ranging from 31 nm to 31 mM; the protein was in a solution containing 50 mM Tris-HCl, 200 mM NaCl, 1.0 mM DTT, 0.1 mM EDTA, 50% Glycerol at pH 7.5. Incubations were carried out at room temperature for 30 minutes. 20 µL of the reaction mixtures in 1X DNA Loading Buffer Blue (Bioline, LDN, UK) were run on a 20% native polyacrylamide gel under a constant voltage of 120 V, in 1X TBE buffer, at 4 °C, until a marker dye reached the bottom. The gel was removed from the plates and stained with SYBR® Gold Nucleic Acid Gel Stain (Thermo Fisher Scientific, MA, USA) according to the supplier's recommendations, then visualized with the GelDoc-It®^{e} Imaging System (UVP, CA, USA) (Figure 3.1). The molar ratio of 10^4 $SSB_{tetramer}$/ ssDNA was found to be minimum to induce a quantitative mobility shift.

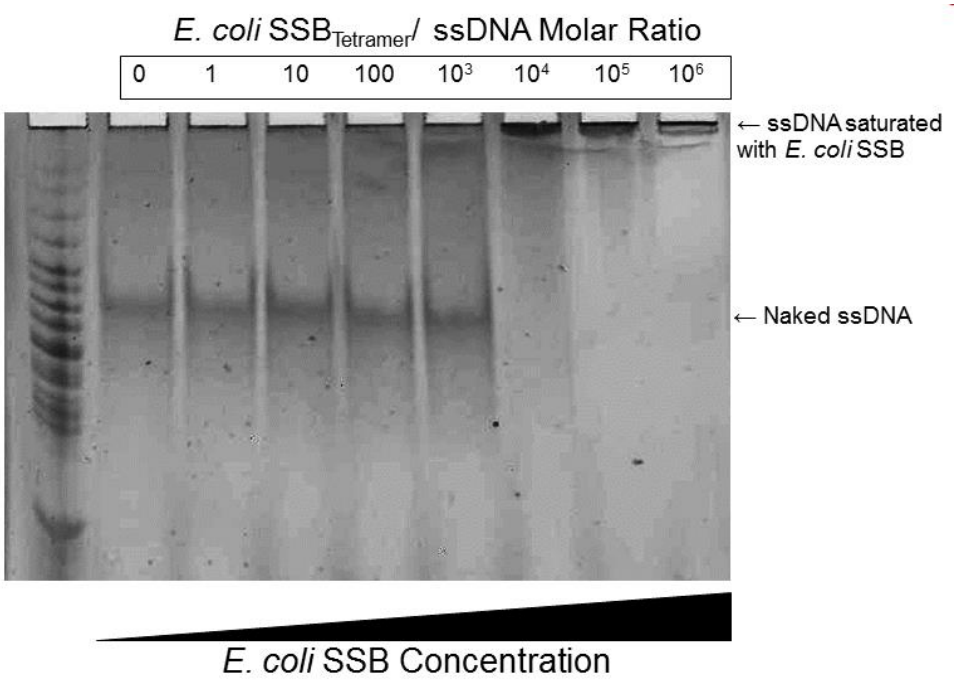


Figure 3.1 EMSA gel image (inverted greyscale) of SSB protein binding to Layer5 template. Ten-fold increases in protein:DNA molar ratios are shown, left to right. A small amount of shifter product is seen at the 10^3 lane, and the shift is quantitative in the 10^4 lane and thereon.

Although no quantitative method has been published for determining accurate molar concentrations of the amount of template on ISPs loaded on an Ion Torrent chip, an estimation can be made, as follows. An Ion 316™ chip can accommodate a maximum of ~6.1 million ISPs; each ISP is manufactured to allow it to be coated with about one million copies of a template. That implies a maximum of ~6.1 trillion copies (~10 picomoles) of ssDNA templates on the chip, although loading is never complete. Therefore, based on EMSA results, a minimum of 1 micromole of $SSB_{tetramer}$ is needed if all templates are about the length of our test case. However, excess sequencing primer molecules remaining in the loading solution will also serve as binding substrates for SSB protein. In addition, there may be decreased mixing in the region of an ISP that is close to the well surfaces, making it harder to get protein in contact with the DNA. Considering

this, the amount of $SSB_{tetramer}$ added was twice the amount suggested by the gel (1.98 micromoles).

The same type of chip and sequencing reagents were used to carry out this test as in the initial experiment described in Chapter 2. After the sequencing run, the chip was taken out and the accumulated liquid in the chip was pipetted off. To remove the complementary (non-covalently attached) strand created in the sequencing by synthesis step, a solution of 30 µL of 100mM NaOH at room temperature was pipetted into the loading port, allowed to sit for 10 minutes and then pipetted off through the loading port. The chip was further treated with a continuous addition of 200 µL of 100mM NaOH solution through the loading port at a rate of ~1 µL per second, letting the solution flow out from the other port of the chip. Residual NaOH solution was removed from the loading port and wiped off from the other port with a Kimwipe, and the chip was centrifuged to remove the leftover NaOH solution from chip, if any. To neutralize the solution surrounding the ISPs, three 200 µL portions of Annealing Buffer were pipetted in through the loading port in the same way as above and the buffer solution was removed as described above. To reload the sequencing reagents, 12 µL of the Sequencing Primer was mixed with 18 µL of Annealing Buffer, the solution was heated at 95°C for 2 minutes and was quickly added into the chip, followed by centrifugation to evenly disperse it over the chip. The chip was incubated at room temperature for 5 minutes to allow the primer to anneal to the template. The excess primer solution was removed from the loading port. Next, 30 µL of 5.0 mg/mL solution (1.98 micromoles) of SSB protein (in 50 mM Tris-HCl, 200 mM NaCl, 1.0 mM DTT, 0.1 mM EDTA, 50% Glycerol at pH 7.5) was added through the loading port at a rate of ~1 µL per second and the chip was

centrifuged for uniform dispersion. The chip was incubated at room temperature for 30 minutes, after which the excess, unbound protein solution was pipetted away through the loading port. Finally, the polymerase solution (3 µL of Ion PGM™ Sequencing 200 v2 Polymerase mixed with 27 µL of Annealing Buffer) was added through the loading port at a rate of ~1 µL per second and the chip was centrifuged for uniform dispersion. The excess solution was then removed from loading port and the chip was placed in the chip socket of PGM. Sequencing was carried out for 250 flows. The sequencing data was analyzed in the same manner as standard sequencing, using programs provided in Appendix A.

3.2.4.2 Ion PGM™ Hi-Q™ Sequencing Kit

The Ion PGM™ Sequencer was cleaned and initialized using Ion PGM™ Hi‑Q™ Sequencing Kit reagents. The same Ion 316™ Chip used as in the standard Ion Torrent sequencing experiment described in Chapter 2. The chip preparation (denaturing the synthesized complimentary strand, washing it away, primer annealing and Ion PGM™ Hi‑Q™ Sequencing Polymerase binding) was performed the same way as for the SSB protein experiment. Sequencing was carried out for 250 flows. The sequencing data was analyzed in the same manner as standard sequencing, using programs provided in Appendix A.

3.3 Results and Discussion

3.3.1 CD Data – Addressing the Cause of GQ Sequenceability Issue

3.3.1.1 Melting Points (T_ms) of GQ Templates

As the temperature increases, the GQ structure starts to unfold and the ellipticity goes down (Figure 3.2). A denaturation curve of the change in ellipticity as a function of

temperature was then plotted, and the T_m was calculated by applying a sigmoidal fit with Boltzmann function to the denaturation curve.

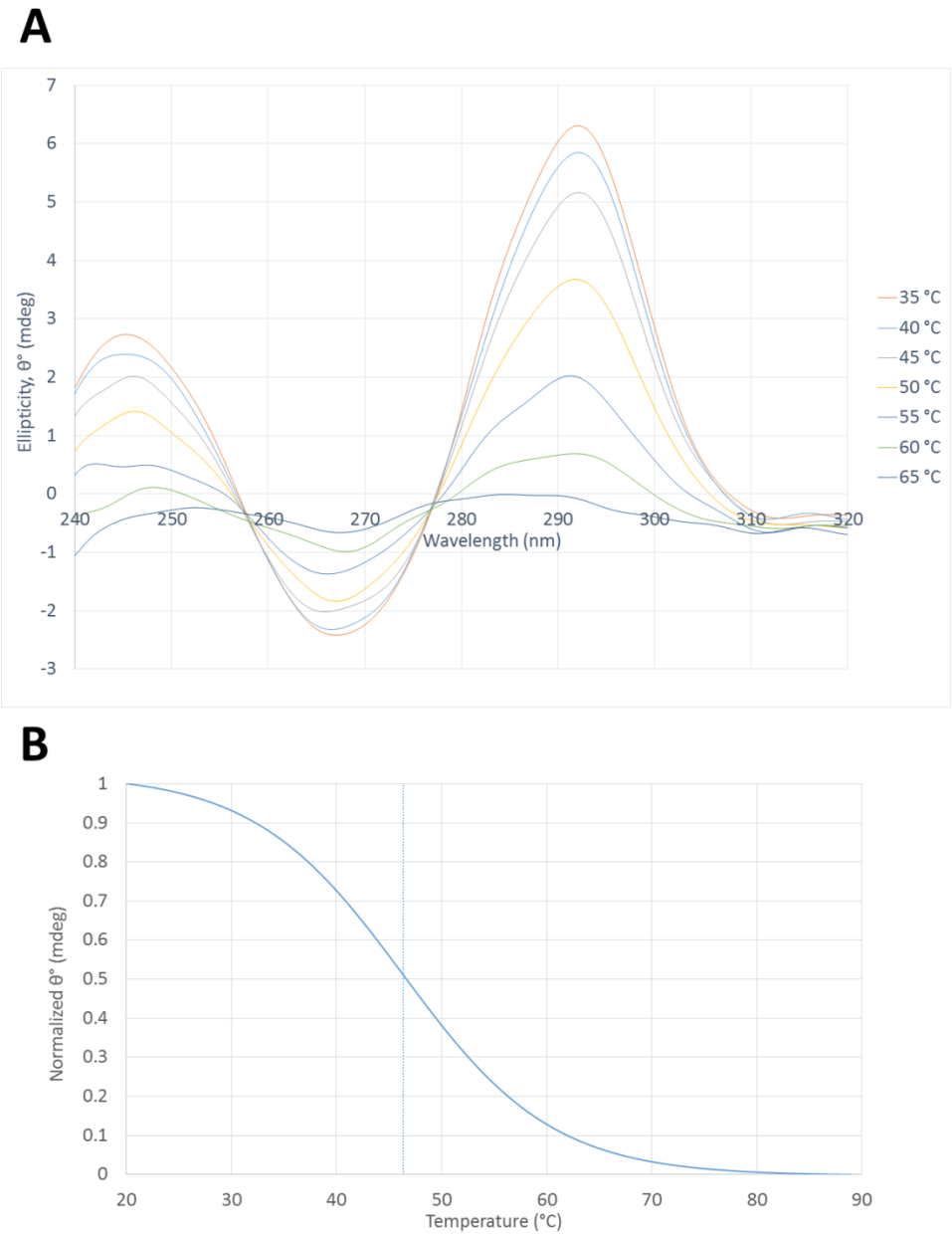


Figure 3.2: (A) Example CD spectra of T15 (thrombin aptamer) recorded at seven different temperatures. A decrease is observed at 292nm with increasing temperature (B) Thermal denaturation curve of T15 recorded at 292 nm.

The T_ms for all the templates studied and the GQ topologies adopted (parallel or antiparallel) by them are listed in Table 3.1. Melting temperatures increased with the increase in the number of tetrad layers, except that the three layered GQ had a T_m slightly higher than the four layered template. Also, consistent with other reports[70,71], an increase in melting temperatures was recorded when there was a decrease in loop length. Variations in the 3' overhang length do not seem to affect the GQ melting temperatures.

Table 3.1 GQ topologies and T_ms of templates under study

Abbreviation	Melting Temperature (T_m)	Topology
Layer2	46.3	Antiparallel
Layer3	63.6	Parallel
Layer4	61.0	Parallel
Layer5	71.2	Parallel
Loop1	82.2	Parallel
Loop3	63.6	Parallel
Loop5	54.4	Parallel
Loop7	53.2	Parallel
OH	63.6	Parallel
+3OH	65.7	Parallel
+6OH	63.4	Parallel
+9OH	62.9	Parallel
T15	46.6	Antiparallel

Representative CD spectra for parallel and antiparallel GQ topologies are shown in Figure 3.3, along with the spectrum of the thrombin aptamer T29. As can be seen, T29 did not form a GQ structure in the buffer used. Nagatoishi *et al.*[93] reported that GQ structure formation in the T29 aptamer, when buffer conditions are not favorable, is induced by presence of the thrombin protein, so this observation is thus consistent with that report. Note that strong binding was observed between T29 aptamer and the protein in later experiments (Chapter 4), which shows that the GQ structure is induced by the protein. All plotted data can be found in Appendix D.

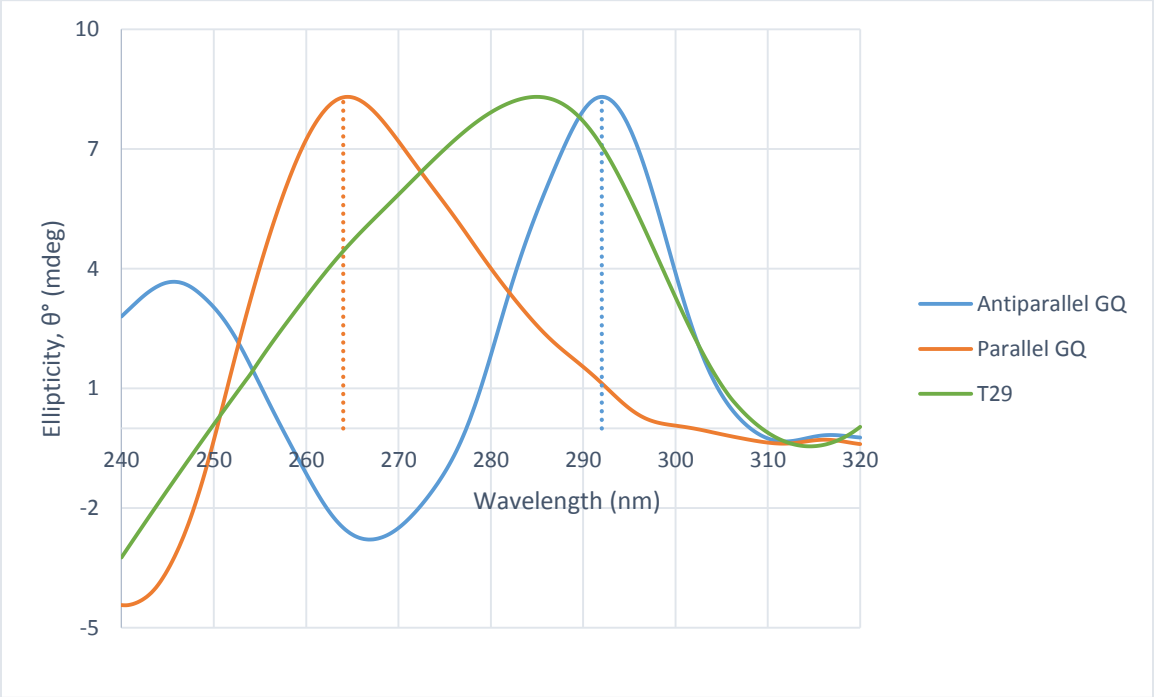


Figure 3.3: Representative CD spectra for parallel and antiparallel GQ topologies and spectrum of T29 template.

3.3.1.2 Correlation between GQ T_m and Sequencing Accuracy

Quality scores and percent base calling errors for the templates in the study were plotted against the melting temperatures obtained above (Figure 3.4). We know each sequencer's operating temperature, 48-51 °C for the Ion Torrent PGM and 65° C (no range provided) for the Illumina MiSeq. We then determined how the data points in each of the plot group. For the Ion Torrent, only the templates Layer2, Loop7 and T15 have T_ms within or below the operating range, and the sequence data show higher mean quality scores and lower base calling errors. For the Illumina MiSeq, with its higher operating temperature, poorer sequence quality is seen only for templates Layer5 and Loop1, which have higher melting temperatures than the platform's 65 °C. T29, as documented in Chapter 2, had impeccable quality scores and few mismatches (at par or even better than unstructured control template) on both sequencers which can be attributed to the absence of a GQ structure. While polymerase activity and differences in buffer components undoubtedly contribute to the differences, based on the observed negative relationship between the instrument's operating temperature and the sequencing accuracy, the most obvious way for Ion Torrent to improve the sequencing performance for structured templates would be to raise the operating temperature. However, this would likely change the processivity rate of the polymerase and hence the rate of release of hydrogen ion, so signal detection parameters would likely have to be adjusted as well. The elevated temperature might affect polymerase stability over many cycles, so this is another factor that would have to be tested. Whether the CMOS and electronics would be affected by this difference in temperature would also need to be evaluated.

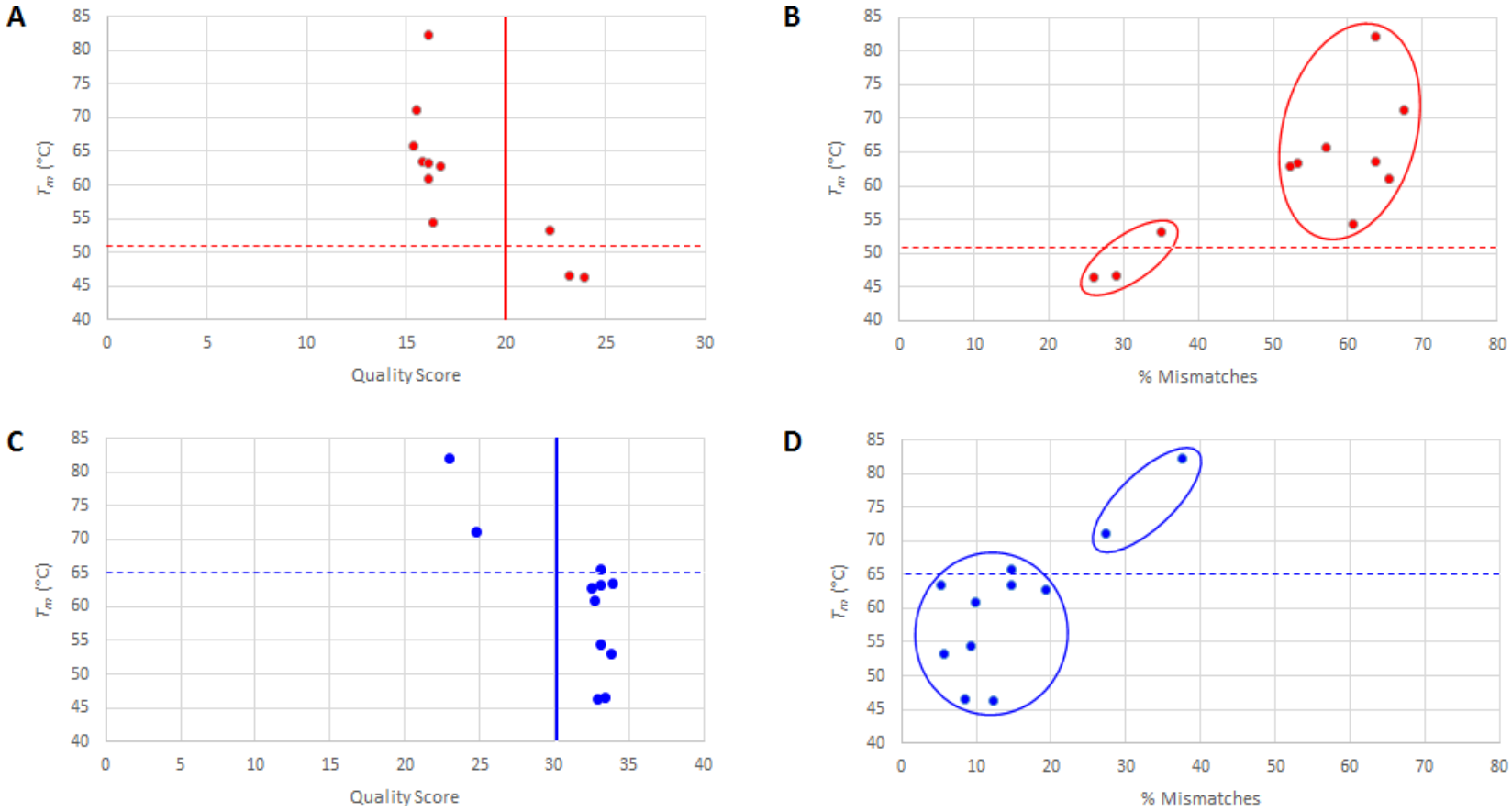


Figure 3.4: Correlation between GQ T_m and sequencing accuracy. The overall mean quality scores of templates are plotted against melting temperatures for Ion Torrent (A) and Illumina (C). Percent base calling errors of templates are plotted against melting temperatures for the Ion Torrent (B) and Illumina (D) platforms. The dotted horizontal lines indicate the instrument's operating temperature. The vertical lines in (A) and (C) are the standard quality score cut-off used by the platform.

3.3.2 Ion Torrent's Poor Performance in Structured Regions: A Signal-processing Issue

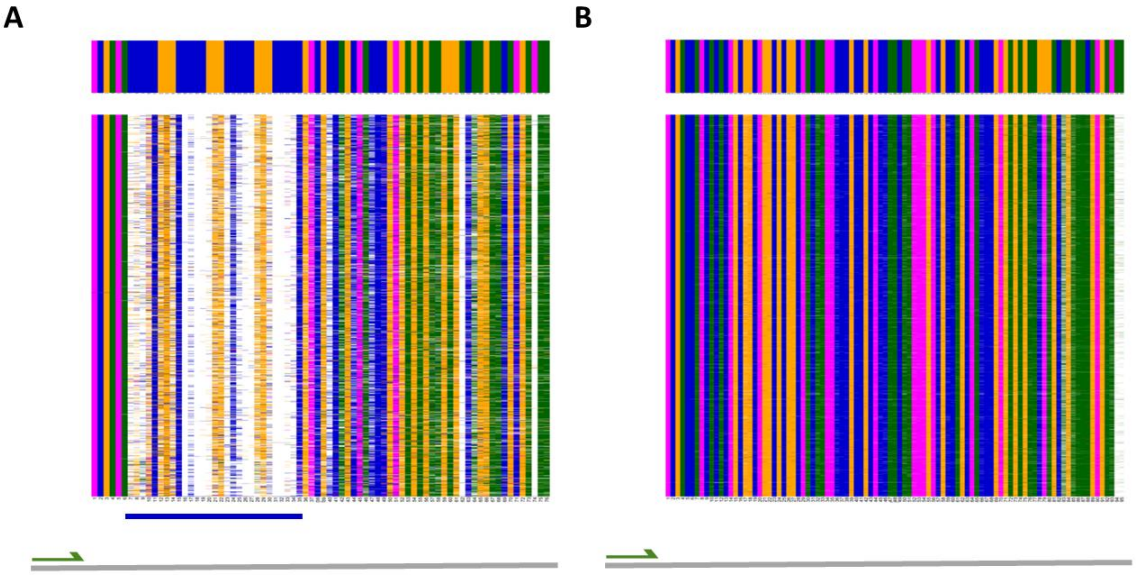


Figure 3.6: (A) MSA of Layer5 template (B) MSA of Control template. Bases are individually colored for independent reads (horizontal lines) with a top horizontal bar indicating the true (expected) template sequence- each column is a base position in the template. Colors are A= orange, T= magenta, G= green, C= blue Gap= white. The horizontal blue line below plot (A) indicates the location of GQ forming sequence. The gray line below each plot shows the full template length with the green bard indicating the sequencing primer-binding site (where incorporation does not occur).

Figure 3.6 depicts the MSA of a GQ template (Layer5) and an unstructured Control template. It is evident that there is a gap in the MSA figure during the GQ-forming sequence, after which base-calls are again made with high consensus and fidelity. This could imply either a phenomenon called 'replication slippage' or a signal-processing issue. During replication slippage, the polymerase dissociates from the template, 'hops' over the (in this case) GQ structure, and then associates with the template to continue polymerization for the downstream sequence (Figure 3.1). In rate slowing, the polymerase is still attached to the template, but it slows down, so in

processing the GQ sequence the events are not properly paced for the instrument to identify individual steps correctly.

A ‘replication slippage assay’ was performed: if the polymerase dissociates and then re-associates downstream of the double-stranded region, the resulting complementary product will be shorter in length. The results of the primer extension assay are shown in (Figure 3.7). The presence of only full-length sequencing by synthesis products indicates that the polymerase does not dissociate and then re-associate. This implies that the poor sequencing output obtained for GQ-templates is not due to replication slippage.

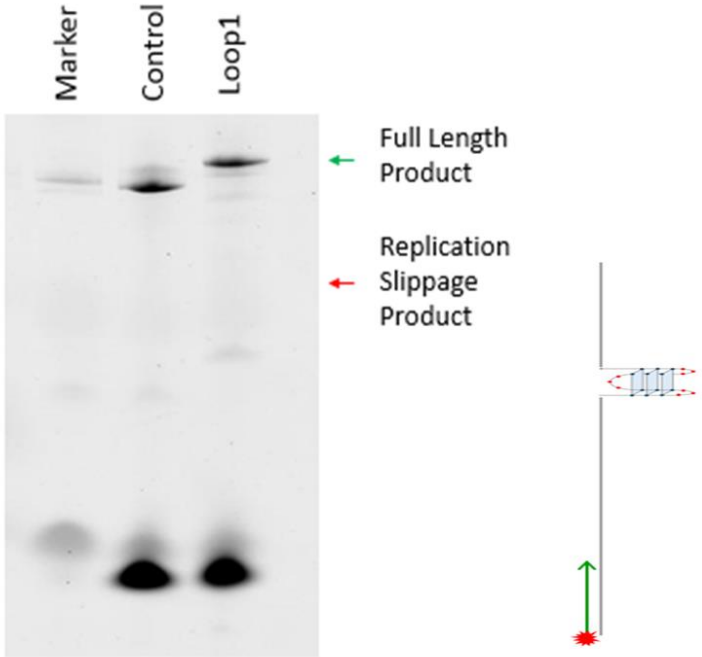


Figure 3.7: Replication slippage assay. A full-length product is seen for both Loop1 and Control templates.

An alternative mechanism is that the instrument cannot capture signal if the enzyme rate is too slow. To further explore this signal-processing issue, understanding of

how Ion Torrent processes the raw data is necessary. One of the main assumptions made in signal processing and subsequent base-calling is that the nucleotide incorporation by polymerase occurs in the same time interval of a flow cycle, for all the templates on a bead and all the beads on a chip. The instrument times the signal detection interval accordingly. In other words, if the rate of nucleotide addition by polymerase is slower than the signal detection time frame, that nucleotide incorporation event is likely to be missed. Based on this and on the evidence that the polymerase indeed sequences the entire GQ template accurately, it may be the case that upon encountering a GQ, the polymerase slows down in its rate of nucleotide addition, as it has to first 'displace' the GQ structure, thereby affecting its overall kinetics and this eventually leads to the base-calling error. As shown in Figure 3.8, a slower nucleotide incorporation event will result in a broader signal peak, for which the base-call will not be made, leading to gaps in MSA in that region. Many independent studies have conclusively demonstrated the 'slowing down' of polymerases upon encountering specific DNA sequences such as palindromic DNA that forms the classic hairpin secondary structure in single stranded nucleic acids[94-96], as well as in high GC regions such as those in trinucleotide repeats of (CGG)n/(CCG)n or (CTG)n/(CAG)n[97,98] and other types of 'slow zones'[99]. As pointed out previously, it is difficult to prove this hypothesis due to the lack of access to the raw data.

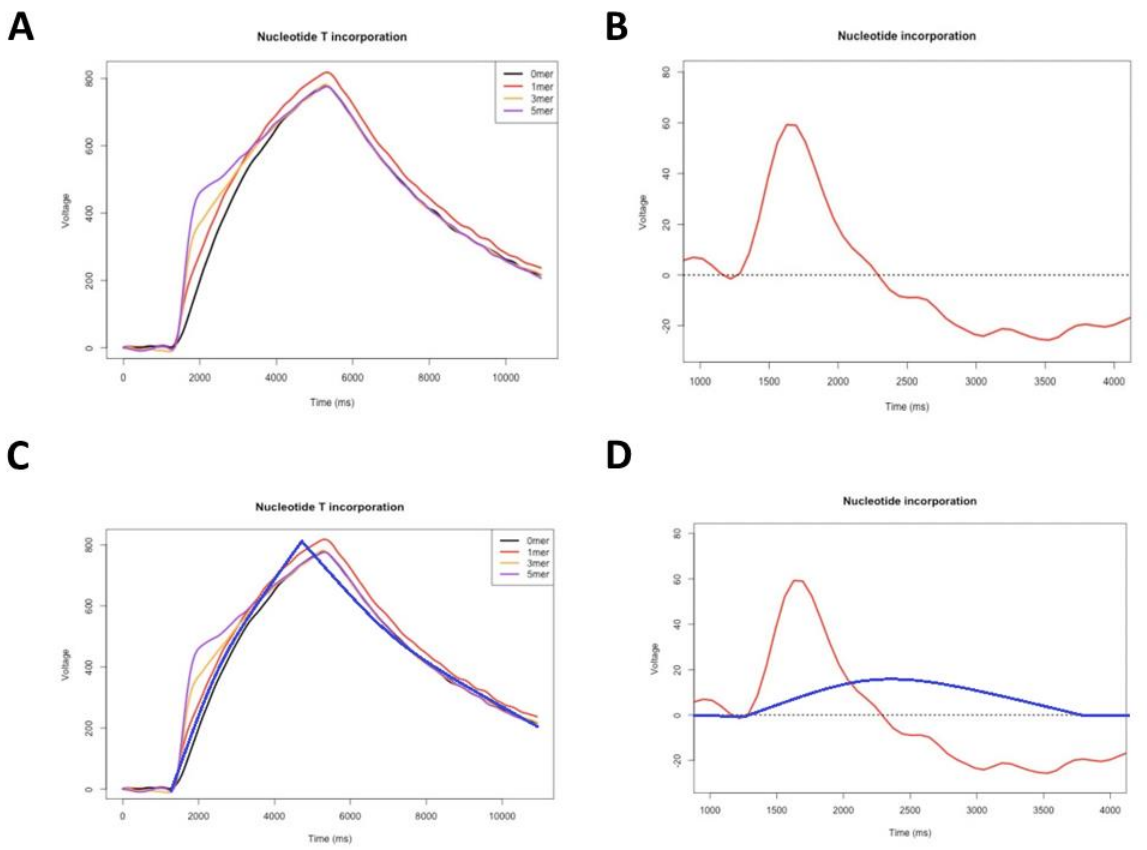


Figure 3.8: Change in the waveform (A) and signal peak (B) during standard nucleotide incorporation event. Change in the waveform (C) and signal peak (D) during slower nucleotide incorporation event, indicated by a blue curve. Adopted from biolektures.wordpress.com/2011/08/22/challenges-in-improving-ion-torrent-raw-accuracy-part-3/, biolektures.wordpress.com/2011/08/30/ion-torrent-signal-processing-part-1-background-and-nucleotide-incorporation-models/.

3.3.3 Effect of SSB on Sequencing Accuracy

The number of reads obtained for each template with standard Ion Torrent sequencing and with SSB protein incubation are listed in Table 3.2

Table 3.2: Distribution of Ion Torrent and Illumina sequencing reads across the templates

	Standard Ion Torrent Sequencing	With SSB Protein
Layer2	186,206	196,346
Layer3/Loop3/ OH	44,688	19,570
Layer4	35,357	26,380
Layer5	1,859	1,858
Loop1	26,844	36,865
Loop5	19,765	71,790
Loop7	291,491	381,820
+3OH	19,837	9,016
+6OH	103,122	53,465
+9OH	86,434	47,601
T15	505,467	497,684
T29	849,630	764,062
Control	252,587	213,677

The consequences of adding SSB to the sequencing reactions are displayed in Figure 3.9. The two templates showing the highest improvements in sequencing accuracy are Loop5 and T15; whereas for templates +6OH, +9OH and control the accuracy drops, as revealed by both the quality scores and base-call errors. For all other templates, there is an insignificant change in accuracy.

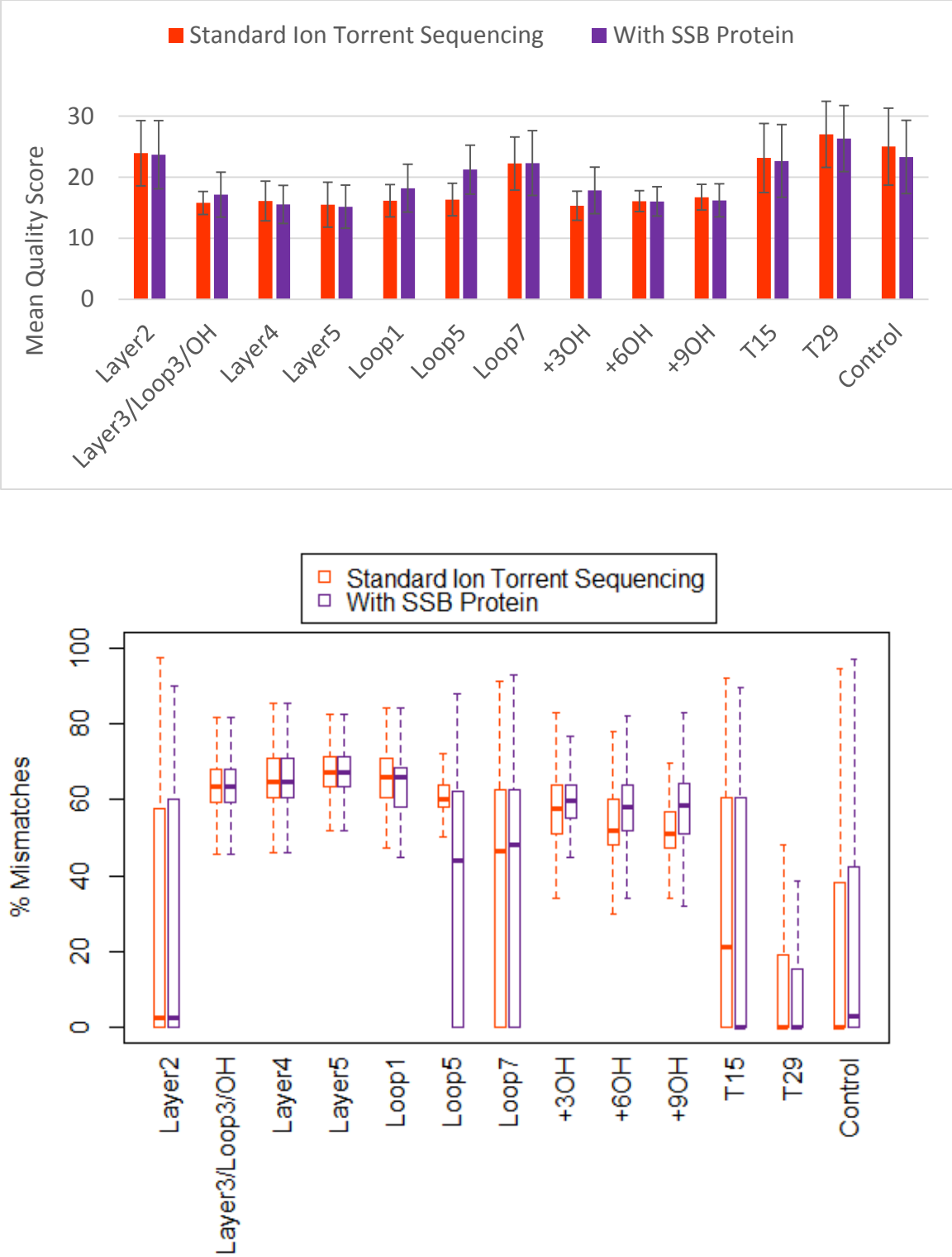


Figure 3.9: Mean quality scores for all sequences identified for each template (top) and the percent base calling error (bottom), obtained using the standard Ion Torrent sequencing conditions and those for which SSB protein was added.

3.3.4 Effect of Ion PGM Hi-Q Sequencing Kit on Sequencing Accuracy

The number of reads obtained for each template with Ion Torrent and Illumina sequencing ae listed in Table 3.3

Table 3.3: Distribution of Ion Torrent and Illumina sequencing reads across the templates

	200 Sequencing Kit v2	Hi-Q Sequencing Kit
Layer2	186,206	82,175
Layer3/Loop3/ OH	44,688	5,986
Layer4	35,357	13,426
Layer5	1,859	722
Loop1	26,844	11,804
Loop5	19,765	1,815
Loop7	291,491	31,836
+3OH	19,837	8,811
+6OH	103,122	55,305
+9OH	86,434	46,160
T15	505,467	305,932
T29	849,630	592,362
Control	252,587	189,052

As shown in Figure 3.10, the effects of using the Hi-Q Sequencing Kit on GQ templates are adverse. Layer2 and Loop7 templates displayed the highest decline in sequencing accuracy, as indicated by both quality scores and base-call errors. Other templates show no change or slight decrease in accuracy. With the exact mechanism of action of the Hi-Q polymerase or the reagents being unavailable, it is not possible to offer

a mechanistic explanation for the data obtained. While the Ion Torrent platform is open with respect to reagent addition, unlike the Illumina cartridges, experiments using other DNA polymerases were even less successful than the one shown, and this direction was abandoned (data not shown).

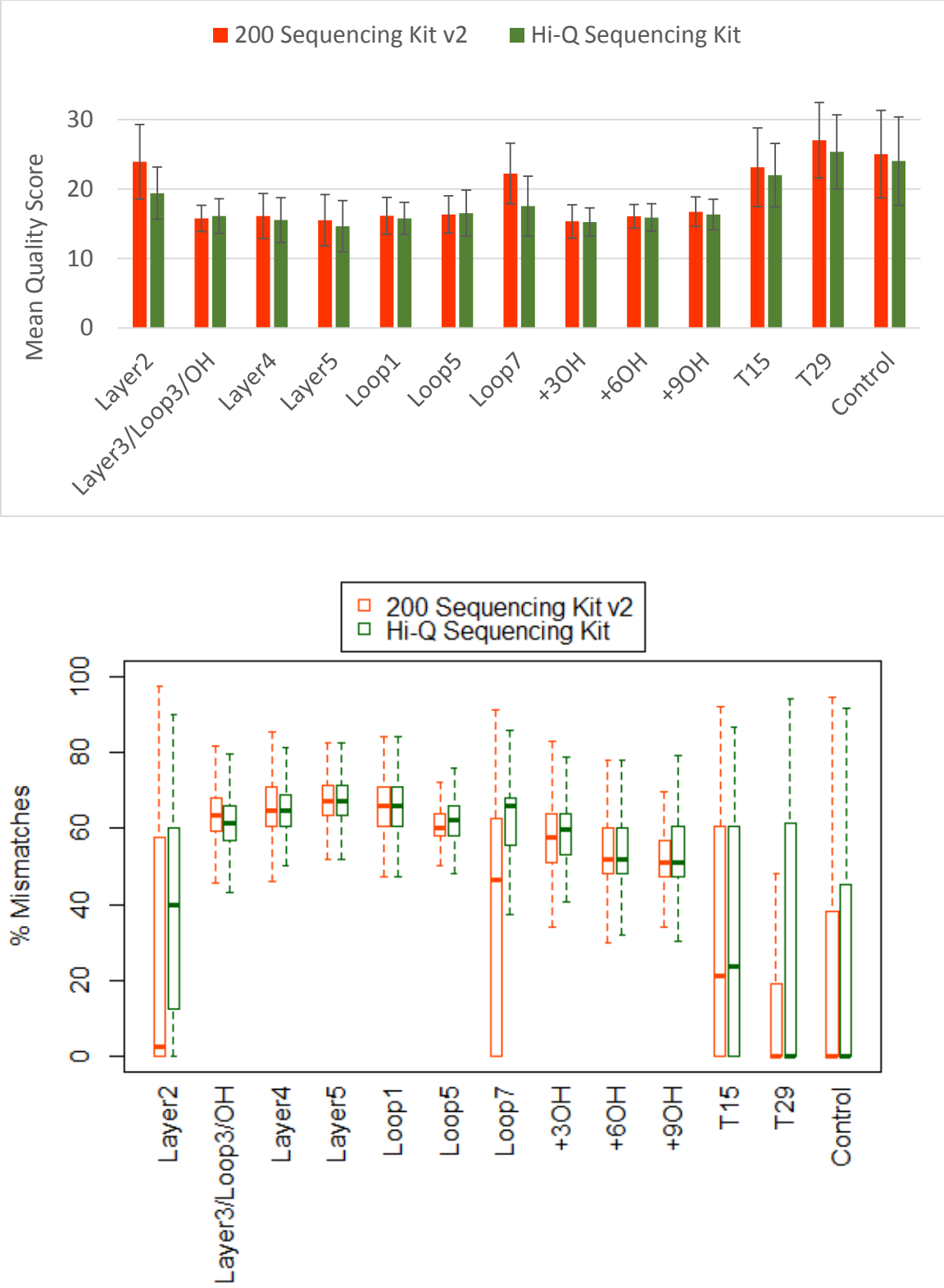


Figure 3.10: Mean quality scores (top) and percent base calling errors (bottom) obtained with Ion Torrent 200 Sequencing Kit v2 and Hi-Q Sequencing Kit.

3.4 Conclusion

The T_ms for the GQ templates studies were found to be in accordance with the trends reported elsewhere: template stability increased with additional GQ layers and reduction of loop lengths. A negative correlation between a template's T_m and sequencing accuracy has been shown, conditioned on the operating temperature of the platform used.

While modification of the Ion Torrent instrument settings may allow data collection rates tuned to polymerase kinetics, changing instrument settings usually invalidates service contracts and agreements with the source company. While basic research scientists may embrace this, clinical scientists cannot, as part of the certification process for their results rests on instrument calibration. Since in other respects the Ion Torrent seems the most affordable and most easily reconfigurable NGS platform for developing nucleic-acid based multiplex sensors, different approaches to regularizing the kinetics were tried.

Incubation with SSB protein was an effective additive for improving sequencing accuracy for a couple of templates, but was detrimental for another three templates. The Ion Torrent Hi-Q sequencing kit reagents and polymerase produced no significant enhancements in performance over standard reagents provided in the 200 Sequencing Kit v2.

CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

4.1 Multiplexed Biosensing - A Proof-of-principle with Thrombin and Acetylated Histone H4 Peptide as Examples

Exonuclease I (ExoI) is an enzyme belonging to the class of nucleases that specifically catalyzes the degradation of ssDNA by cleaving nucleotides in the 3' to 5' direction. Although used routinely for degrading excess primer after PCR, novel applications of the enzyme are starting to emerge[100-104]. All such applications are based on the principle of the Exonuclease Protection Assay (EPA) wherein the degradation of an otherwise ssDNA substrate by ExoI is inhibited by a ssDNA-target complex formation. The target molecule can be a metabolite, another nucleic acid, or a protein. When a complex is present, the ExoI approach site, at the 3' end of ssDNA, is sterically hindered by the bound target, leading to ssDNA protection against the cleavage activity.

Based on the principle of EPA, and coupled with high throughput sequencing capabilities, a novel multiplexed biosensing method is proposed (Figure 4.1). The protocol employs an aptamer sequence extended by a 'barcode sequence' at its 5' end. The barcode sequence is bounded by Ion Torrent PGM adapter sequences (one on each side), as: 5'- CCTCTCTATGGGCAGTCGGTGAT - (Barcode Sequence) - CTGACTGAGTCGGAG ACACGCAGGGATGAGATGG -3' and the barcode sequence is ten nucleotides in length, unique for each aptamer. In this proof-of-concept study, two protein-aptamer pairs are used, for which quantitative protein:aptamer binding ratios are

determined (e.g. using the gel migration assay) viz. thrombin and acetylated histone H4 peptide, chosen due to the detailed characterization of the complexes, moderate to strong stabilities, ease of availability and low costs. Aptamer incubation with the target protein/peptide leads to part of the DNA being protected from ExoI cleavage, thus it can be, subsequently, PCR amplified and detected as sequencing output. Sequencing data of different aptamers is binned based on the characteristic barcode sequences. Quantitation can be achieved by performing serial dilutions of protein: aptamer, which leads to proportionally less binding of protein to aptamer, thus less protection of that aptamer by ExoI, which is finally reflected in sequencing yield. A control consists of incubating the uncomplexed aptamer with ExoI, to make sure the structured portion is not inherently resistant to enzymatic digestion. The number of copies of aptamer represented in the sequencing output are plotted against the protein concentration, and the equation of best-fit line can be used for quantitative determination of protein of interest.

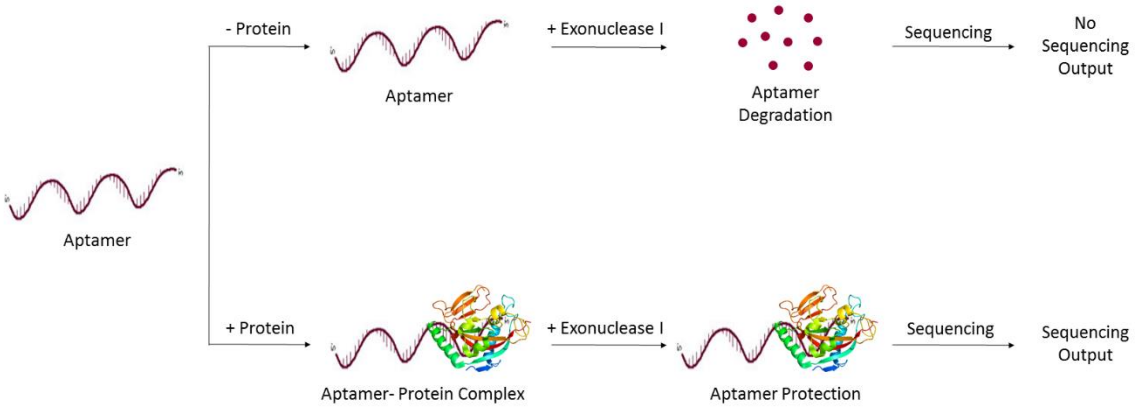


Figure 4.1: Schematic representation of multiplexed biosensing method based on EPA and NGS.

4.2 Sequencing by Strand Displacement Synthesis

A molecular biology solution to removing the folded structure of a single-stranded template is to create the much more regular structure of double-stranded DNA and use a displacement polymerase with no exonuclease activity. In brief, the approach consists of annealing a primer downstream of the sequencing primer binding site on templated ISPs, leaving a gap upstream, and carrying out synthesis of the complimentary strand at an elevated temperature (above the T_m of the anticipated structure), or with reaction additives that promote melting of highly structured templates. A thermostable DNA polymerase can be used in this step, and the reaction can be carried out on a template already attached to ISPs, in bulk. The reaction product is washed under mild conditions, to remove unused primer, nucleotides and the DNA polymerase without separating the strands of DNA. In the second step, the sequencing primer is annealed to the template and the *Bst* polymerase is bound. Beads are applied to the Ion Torrent chip and the sequencing reaction is performed according to the standard protocol. In this case, for the length of the entire template sequence that has a complement, for each nucleotide addition, the polymerase has to first displace a portion of the complementary strand before synthesizing the new strand. This will ensure a uniform, albeit slightly slower, rate of nucleotide addition throughout the length of the template, irrespective of its structural complexity.

Since every NGS template has to be modified by the addition of adaptors for carrying out emulsion PCR and ISP capture, it is possible to design a universal primer to create the double-stranded downstream template as well. While this would add an additional step, it should be straightforward as the ISPs can be handled in bulk for this

step, without even a requirement for emulsion reactors. Since the time frame for carrying out Ion Torrent sequencing is ~7 hours (for 300nt templates) while the Illumina platform requires ~50 hours for 300nt templates, adding 2-3 hours to the sample preparation time does not lose the competitive edge for the Ion Torrent technology.

4.3 Summary

A biosensor is an analytical device, used for the detection of an analyte that combines a biological component with a physicochemical detector. A multiplex biosensor is a design that allows the simultaneous detection of multiple analytes. In biomedical applications the analyte is a biological molecule for which quantitative information is sought, generally assumed to be a protein. For clinical tests it is frequently true that subtle structural differences must be detected, as they have significant biological implications, such as single-amino acid changes that affect the activity or regulation of an enzyme. To capture small structural differences requires a highly configurable reagent base, something with an enormous possible repertoire. The structural possibilities of nucleic acid aptamers have attracted attention on this account. Once a protein binds to an aptamer a selection and enrichment process must be carried out in order to identify the aptamer. The SELEX method is currently used, but it leads to many false positives and false negatives from non-selective binding and the very high background created by the initial pool of 10^{12} or so candidates. Since NGS platforms routinely display 10^{9} – 10^{12} nucleic acid fragments, we investigated whether the characteristics of the platform would allow high-confidence determination of the sequence of structured DNA templates, and whether the platform design was open enough to allow on-chip detection of a protein-aptamer complex.

The Thermo Fisher Ion Torrent PGM sequencer and the Illumina MiSeq sequencer with associated chemistries were both tested using related template families (differing only by sequencing primer extensions) containing systematically varied GQ layers, loops and overhanging 3' strands, as well as two known thrombin-binding aptamers (each aptamer binds a GQ but they bind to different sites on thrombin). Conventional sequencing approaches show that neither platform can accurately report the sequence of the most stable structures, but that the Illumina platform could accurately process more of them. This is important from a genome sequencing perspective, since GQ and other highly folded structures appear at sites where proteins such as transcription factors bind – mistakes in identifying these sequences have consequences for genome assembly, annotation and the interpretation of gene regulatory networks, among others. We suggest that additional structured template families be produced and sequenced on the Illumina platform to determine whether there is a characteristic structural signature or stability that defines where a structure that defies processing on this platform exists and, hence, where other approaches, such as Sanger sequencing and specialized reagents, should be used. It is not clear to us that the Illumina platform is a good candidate as an aptamer discovery platform, despite its greater ability to process through stable sequence structures, as it uses a closed reagent cassette, dedicated fluorescent chemistry and reagents requiring very specialized enzymes. Stripping away the complementary strand, flowing in protein candidates and then repeating a sequencing reaction are not operations one can easily design around the reagent cartridges. The Ion Torrent is a much more open platform, with readily accessible ports for adding additional reagents, the use of standard nucleotides and beads in individual wells, offer a much better chance for downstream

manipulation. Despite these attractive features, sequencing on the Ion Torrent gave poorer results that those obtained on the Illumina platform, not surprising since it operates at a lower temperature and does not use denaturants in its standard reagent cocktail. The addition of single-stranded binding protein marginally improved sequence calling accuracy, but the use of a kit supposedly optimized to handle structure in microbial DNA samples offered no improvement. However, this did demonstrate that the platform could handle the addition of reagents and re-sequencing of templates, so some feasibility for adaptation was demonstrated. Multiple alignment analysis of the output sequence data suggested that either the polymerase was skipping over the structured portion, or that the rate of polymerase procession was sufficiently altered that the instrument could not capture the data. A test of whether Ion Torrent reagents showed replication by slippage errors was negative, lending weight to the hypothesis that it is the signal capture step that is misaligned to the polymerase rate under these circumstances. Our preferred next step would be to work with Ion Torrent engineers to modify the data collection intervals on the instrument to allow direct reading without additional steps. However, it is not clear that the company is interested, and for applied researchers the prospect of modifying an instrument away from certification specs would render any assay useless. Therefore a molecular biological approach was used on the GQ template family, to see whether ‘straightening’ the template by making it double stranded and then using the strand displacement property of the provide enzyme would provide a consistent rate that the instruments current settings could accept. This experiment is less ideal than an instrument approach, as it involves extra time and reagents. However, it may yet be useful for those performing genomic sequencing studies, for reasons similar to those

discussed for the Illumina platform. An earlier study of a set of structured templates (hairpins of graduated stability rather than GQ layers) showed results similar to those found here on the Ion Torrent (Khoshnevis PhD dissertation, unpublished data) and we suggest that a systematic study indicating at what point the platform can no longer provide reliable data, perhaps with a diagnostic signature for problem regions, would be a useful calibration tool.

The end goal is to bind proteins to pre-sequenced aptamers and detect those events on the chip. The Ion Torrent chip does allow stripping of complementary strands, the inflow of additional reagents to promote aptamer folding, addition of proteins, and we have shown that binding of those proteins does occur using a fluorescent microscopy approach for initial detection (data not shown). We used dye- labeled proteins and microscopic imaging, but the hydrophobic dyes lead to a lot of non-specific binding to chip components. It is preferable to use an assay that the CMOS sensor of the chip is already set up to respond to: an increase in pH. Another approach is to attach a distinct long oligonucleotide to each protein with its own primer for sequencing, and use that sequence, obtained from a second sequencing step, to determine those sites that have a protein attached. This simply makes use of the same detection method as the instrument is already set up to handle, but does require a separate handling step for the proteins. The downside is the extensive handling and modification of the proteins, which will undoubtedly lead to some loss of native structures. As a simpler approach, an exonuclease I protection assay has modified and is here proposed for thrombin and acetylated histone H4 peptide.

As a protein discovery tool, we would like to have available robotic retrieval systems sensitive enough to retrieve individual beads from the wells on a chip (a 5um displacement in the x and y directions), and determine how many ISPs would be required to provide sufficient material for effective mass spectrophotometric determination of the bound protein. Theoretically, 100 beads should provide sufficient material, but this supposes complete binding of every aptamer on a bead, an unlikely scenario. This would allow deposition of genomic sequences on beads; after incubation with cell lysates those proteins binding to the sequences, single or as complexes, could be retrieved and analyzed. This platform then becomes a protein discovery tool for cell biology as well as a tool for identifying protein aptamers for technology applications.

REFERENCES

1. Sanger, F. The terminal peptides of insulin. *Biochemical Journal* **45**, 563 (1949).

2. Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H. & Will, G. Structure of hæmoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. *Nature* **185**, 416-422 (1960).

3. Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662-666 (1958).

4. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235-242 (2000).

5. Petricoin, E. F. *et al.* Mapping molecular networks using proteomics: a vision for patient-tailored combination therapy. *Journal of clinical oncology* **23**, 3614-3621 (2005).

6. Koenen, R. R. & Weber, C. Therapeutic targeting of chemokine interactions in atherosclerosis. *Nature reviews Drug discovery* **9**, 141-153 (2010).

7. Weber, C. & Koenen, R. R. Fine-tuning leukocyte responses: towards a chemokine 'interactome'. *Trends in immunology* **27**, 268-273 (2006).

8. Koenen, R. R. *et al.* Disrupting functional interactions between platelet chemokines inhibits atherosclerosis in hyperlipidemic mice. *Nature medicine* **15**, 97-103 (2009).

9. Engvall, E. & Perlmann, P. Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. *Immunochemistry* **8**, 871-874 (1971).

10. Van Weemen, B. & Schuurs, A. Immunoassay using antigen—enzyme conjugates. *FEBS letters* **15**, 232-236 (1971).

11. Balboni, I., Limb, C., Tenenbaum, J. D. & Utz, P. J. Evaluation of microarray surfaces and arraying parameters for autoantibody profiling. *Proteomics* **8**, 3443-3449 (2008).

12. Osterfeld, S. J. *et al.* Multiplex protein assays based on real-time magnetic nanotag sensing. *Proceedings of the National Academy of Sciences* **105**, 20637-20640 (2008).

13. Spindel, S. & Sapsford, K. E. Evaluation of optical detection platforms for multiplexed detection of proteins and the need for point-of-care biosensors for clinical use. *Sensors* **14**, 22313-22341 (2014).

14. Balboni, I. *et al.* Multiplexed protein array platforms for analysis of autoimmune diseases. *Annu. Rev. Immunol.* **24**, 391-418 (2006).

15. Kingsmore, S. F. Multiplexed protein measurement: technologies and applications of protein and antibody arrays. *Nature reviews Drug discovery* **5**, 310-321 (2006).

16. Kodoyianni, V. Label-free analysis of biomolecular interactions using SPR imaging. *Biotechniques* **50**, 32-40, doi:10.2144/000113569 (2011).

17. Oliphant, A., Barker, D. L., Stuelpnagel, J. R. & Chee, M. S. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* **32**, 56-58 (2002).

18. Lin, C. H., Yeakley, J. M., McDaniel, T. K. & Shen, R. in *DNA and RNA Profiling in Human Blood* 129-142 (Springer, 2009).

19. Hsu, H. Y., Joos, T. O. & Koga, H. Multiplex microsphere-based flow cytometric platforms for protein analysis and their application in clinical proteomics–from assays to results. *Electrophoresis* **30**, 4008-4019 (2009).

20. Bertenshaw, G. P. *et al.* Multianalyte profiling of serum antigens and autoimmune and infectious disease molecules to identify biomarkers dysregulated in epithelial ovarian cancer. *Cancer Epidemiology Biomarkers & Prevention* **17**, 2872-2881 (2008).

21. Kim, B. K. *et al.* The multiplex bead array approach to identifying serum biomarkers associated with breast cancer. *Breast Cancer Research* **11**, R22 (2009).

22. Opalka, D. *et al.* Simultaneous quantitation of antibodies to neutralizing epitopes on virus-like particles for human papillomavirus types 6, 11, 16, and 18 by a multiplexed luminex assay. *Clinical and diagnostic laboratory immunology* **10**, 108-115 (2003).

23. Bellisario, R., Colinas, R. J. & Pass, K. A. Simultaneous measurement of antibodies to three HIV-1 antigens in newborn dried blood-spot specimens using a multiplexed microsphere-based immunoassay. *Early human development* **64**, 21-25 (2001).

24. Pelttari, K. *et al.* Secretion of matrix metalloproteinase 3 by expanded articular chondrocytes as a predictor of ectopic cartilage formation capacity in vivo. *Arthritis & Rheumatism* **58**, 467-474 (2008).

25. Zhang, J.-Z. & Ward, K. W. Besifloxacin, a novel fluoroquinolone antimicrobial agent, exhibits potent inhibition of pro-inflammatory cytokines in human THP-1 monocytes. *Journal of antimicrobial chemotherapy* **61**, 111-116 (2008).

26. Dollins, C. M., Nair, S. & Sullenger, B. A. Aptamers in immunotherapy. *Human gene therapy* **19**, 443-450 (2008).

27. Sullenger, B. A., Gallardo, H. F., Ungers, G. E. & Gilboa, E. Overexpression of TAR sequences renders cells resistant to human immunodeficiency virus replication. *Cell* **63**, 601-608 (1990).

28. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *nature* **346**, 818-822 (1990).

29. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-510 (1990).

30. Kimoto, M., Yamashige, R., Matsunaga, K.-i., Yokoyama, S. & Hirao, I. Generation of high-affinity DNA aptamers using an expanded genetic alphabet. *Nature biotechnology* **31**, 453-457 (2013).

31. Tuerk, C., MacDougal, S. & Gold, L. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proceedings of the National Academy of Sciences* **89**, 6988-6992 (1992).

32. Chen, H., McBroom, D. G., Zhu, Y.-Q., Gold, L. & North, T. W. Inhibitory RNA ligand to reverse transcriptase from feline immunodeficiency virus. *Biochemistry* **35**, 6923-6930 (1996).

33. Pileur, F. *et al.* Selective inhibitory DNA aptamers of the human RNase H1. *Nucleic acids research* **31**, 5776-5788 (2003).

34. Wu, X. *et al.* DNA Aptamer Selected against Pancreatic Ductal Adenocarcinoma for in vivo Imaging and Clinical Tissue Recognition. *Theranostics* **5**, 985 (2015).

35. Song, K.-M., Lee, S. & Ban, C. Aptamers and their biological applications. *Sensors* **12**, 612-631 (2012).

36. Keefe, A. D., Pai, S. & Ellington, A. Aptamers as therapeutics. *Nature Reviews Drug Discovery* **9**, 537-550 (2010).

37. Gragoudas, E. S., Adamis, A. P., Cunningham Jr, E. T., Feinsod, M. & Guyer, D. R. Pegaptanib for neovascular age-related macular degeneration. *New England Journal of Medicine* **351**, 2805-2816 (2004).

38. Bruno, J. G. Predicting the uncertain future of aptamer-based diagnostics and therapeutics. *Molecules* **20**, 6866-6887 (2015).

39. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463-5467 (1977).

40. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (1985).

41. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).

42. Mardis, E. R. Next-generation sequencing platforms. *Annual review of analytical chemistry* **6**, 287-303 (2013).

43. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352 (2011).

44. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *nature* **456**, 53-59 (2008).

45. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138 (2009).

46. Baker, M. De novo genome assembly: what every biologist should know. *Nature methods* **9**, 333 (2012).

47. Merriman, B., Torrent, I., Rothberg, J. M. & Team, D. Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis* **33**, 3397-3417 (2012).

48. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387-402 (2008).

49. Schütze, T. *et al.* Probing the SELEX process with next-generation sequencing. *PLoS One* **6**, e29604 (2011).

50. Cho, M. *et al.* Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proceedings of the National Academy of Sciences* **107**, 15373-15378 (2010).

51. Kupakuwana, G. V., Crill II, J. E., McPike, M. P. & Borer, P. N. Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. *PLoS One* **6**, e19395 (2011).

52. Hoon, S., Zhou, B., Janda, K. D., Brenner, S. & Scolnick, J. Aptamer selection by high-throughput sequencing and informatic analysis. *Biotechniques* **51**, 413-416 (2011).

53. Blind, M. & Blank, M. Aptamer selection technology and recent advances. *Molecular Therapy—Nucleic Acids* **4**, e223 (2015).

54. Song, Y., Zhang, H., Zhu, Z. & Yang, C. in *Aptamers Selected by Cell-SELEX for Theranostics* 339-352 (Springer, 2015).

55. Balagurumoorthy, P. & Brahmachari, S. K. Structure and stability of human telomeric sequence. *Journal of Biological Chemistry* **269**, 21858-21869 (1994).

56. Todd, A. K., Johnston, M. & Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic acids research* **33**, 2901-2907 (2005).

57. O Tucker, W., T Shum, K. & A Tanner, J. G-quadruplex DNA aptamers and their ligands: structure, function and application. *Current pharmaceutical design* **18**, 2014-2026 (2012).

58. Ma, D.-L. *et al.* Structure-based approaches targeting oncogene promoter G-quadruplexes. (2012).

59. Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K. & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic acids research* **34**, 5402-5415 (2006).

60. Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nature chemistry* **5**, 182-186 (2013).

61. Rawal, P. *et al.* Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome research* **16**, 644-655 (2006).

62. Baral, A., Kumar, P., Pathak, R. & Chowdhury, S. Emerging trends in G-quadruplex biology–role in epigenetic and evolutionary events. *Mol. Biosyst.* **9**, 1568-1575 (2013).

63. Bugaut, A. & Balasubramanian, S. 5′-UTR RNA G-quadruplexes: translation regulation and targeting. *Nucleic acids research* **40**, 4727-4741 (2012).

64. Paeschke, K., McDonald, K. R. & Zakian, V. A. Telomeres: structures in need of unwinding. *FEBS letters* **584**, 3760-3772 (2010).

65. Paeschke, K., Capra, J. A. & Zakian, V. A. DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell* **145**, 678-691 (2011).

66. Mani, P., Yadav, V. K., Das, S. K. & Chowdhury, S. Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination. *PLoS One* **4**, e4399-e4399 (2009).

67. De, S. & Michor, F. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nature structural & molecular biology* **18**, 950-955 (2011).

68. Sissi, C., Gatto, B. & Palumbo, M. The evolving world of protein-G-quadruplex recognition: a medicinal chemist's perspective. *Biochimie* **93**, 1219-1230 (2011).

69. Ray, S. *et al.* RPA-mediated unfolding of systematically varying G-quadruplex structures. *Biophysical journal* **104**, 2235-2245 (2013).

70. Guédin, A., Gros, J., Alberti, P. & Mergny, J.-L. How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic acids research* **38**, 7858-7868 (2010).

71. Smirnov, I. & Shafer, R. H. Effect of loop sequence and size on DNA aptamer stability. *Biochemistry* **39**, 1462-1468 (2000).

72. Zhang, A. Y., Bugaut, A. & Balasubramanian, S. A sequence-independent analysis of the loop length dependence of intramolecular RNA G-quadruplex stability and topology. *Biochemistry* **50**, 7251-7258 (2011).

73. Bugaut, A. & Balasubramanian, S. A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry* **47**, 689-697 (2008).

74. Smargiasso, N. *et al.* G-Quadruplex DNA Assemblies: Loop Length, Cation Identity, and Multimer Formation†. *Journal of the American Chemical Society* **130**, 10208-10216 (2008).

75. Tasset, D. M., Kubik, M. F. & Steiner, W. Oligonucleotide inhibitors of human thrombin that bind distinct epitopes. *Journal of molecular biology* **272**, 688-698 (1997).

76. Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H. & Toole, J. J. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. (1992).

77. Olsen, C. M., Lee, H.-T. & Marky, L. A. Unfolding Thermodynamics of Intramolecular G-Quadruplexes: Base Sequence Contributions of the Loops†. *The Journal of Physical Chemistry B* **113**, 2587-2595 (2008).

78. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology* **30**, 434-439 (2012).

79. Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research* **38**, 1767-1771 (2010).

80. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *Journal of computational and graphical statistics* **5**, 299-314 (1996).

81. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research* **8**, 175-185 (1998).

82. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **8**, 186-194 (1998).

83. Arezi, B., Xing, W., Sorge, J. A. & Hogrefe, H. H. Amplification efficiency of thermostable DNA polymerases. *Analytical biochemistry* **321**, 226-235 (2003).

84. Thomas, D. C., Nardone, G. A. & Randall, S. K. Amplification of padlock probes for DNA diagnostics by cascade rolling circle amplification or the polymerase chain reaction. *Archives of pathology & laboratory medicine* **123**, 1170-1176 (1999).

85. Canceill, D., Viguera, E. & Ehrlich, S. D. Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency. *Journal of Biological Chemistry* **274**, 27481-27490 (1999).

86. Viguera, E., Canceill, D. & Ehrlich, S. D. In vitro replication slippage by DNA polymerases from thermophilic organisms. *Journal of molecular biology* **312**, 323-333 (2001).

87. Dickey, T. H., Altschuler, S. E. & Wuttke, D. S. Single-stranded DNA-binding proteins: multiple domains for multiple functions. *Structure* **21**, 1074-1084 (2013).

88. Ashton, N. W., Bolderson, E., Cubeddu, L., O'Byrne, K. J. & Richard, D. J. Human single-stranded DNA binding proteins are essential for maintaining genomic stability. *BMC molecular biology* **14**, 1 (2013).

89. Marceau, A. H. Functions of single-strand DNA-binding proteins in DNA replication, recombination, and repair. *Single-Stranded DNA Binding Proteins: Methods and Protocols*, 1-21 (2012).

90. Budhathoki, J. B. *et al.* RecQ-core of BLM unfolds telomeric G-quadruplex in the absence of ATP. *Nucleic acids research* **42**, 11528-11545 (2014).

91. Savitzky, A. & Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* **36**, 1627-1639 (1964).

92. Katoh, K., Misawa, K., Kuma, K. i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30**, 3059-3066 (2002).

93. Nagatoishi, S., Tanaka, Y. & Tsumoto, K. Circular dichroism spectra demonstrate formation of the thrombin-binding DNA aptamer G-quadruplex under stabilizing-cation-deficient conditions. *Biochemical and biophysical research communications* **352**, 812-817 (2007).

94. Bedinger, P., Munn, M. & Alberts, B. M. Sequence-specific pausing during in vitro DNA replication on double-stranded DNA templates. *Journal of Biological Chemistry* **264**, 16880-16886 (1989).

95. LaDuca, R. J., Fay, P. J., Chuang, C., McHenry, C. S. & Bambara, R. A. Site-specific pausing of deoxyribonucleic acid synthesis catalyzed by four forms of Escherichia coli DNA polymerase III. *Biochemistry* **22**, 5177-5188 (1983).

96. Lemoine, F. J., Degtyareva, N. P., Lobachev, K. & Petes, T. D. Chromosomal translocations in yeast induced by low levels of DNA polymerase: a model for chromosome fragile sites. *Cell* **120**, 587-598 (2005).

97. Kang, S., Ohshima, K., Shimizu, M., Amirhaeri, S. & Wells, R. D. Pausing of DNA synthesis in vitro at specific loci in CTG and CGG triplet repeats from human hereditary disease genes. *Journal of Biological Chemistry* **270**, 27014-27021 (1995).

98. Samadashwily, G. M., Raca, G. & Mirkin, S. M. Trinucleotide repeats affect DNA replication in vivo. *Nature genetics* **17**, 298-304 (1997).

99. Cha, R. S. & Kleckner, N. ATR homolog Mec1 promotes fork progression, thus averting breaks in replication slow zones. *Science* **297**, 602-606 (2002).

100. Zheng, D., Zou, R. & Lou, X. Label-free fluorescent detection of ions, proteins, and small molecules using structure-switching aptamers, SYBR gold, and exonuclease I. *Analytical chemistry* **84**, 3554-3560 (2012).

101. Wu, Z., Zhen, Z., Jiang, J.-H., Shen, G.-L. & Yu, R.-Q. Terminal protection of small-molecule-linked DNA for sensitive electrochemical detection of protein binding via selective carbon nanotube assembly. *Journal of the American Chemical Society* **131**, 12325-12332 (2009).

102. Lv, Y., Xue, Q., Gu, X., Zhang, S. & Liu, J. A label-free fluorescence assay for thrombin based on aptamer exonuclease protection and exonuclease III-assisted recycling amplification-responsive cascade zinc (ii)-protoporphyrin IX/G-quadruplex supramolecular fluorescent labels. *Analyst* **139**, 2583-2588 (2014).

103. Yao, Y., Wang, Q., Hao, Y.-h. & Tan, Z. An exonuclease I hydrolysis assay for evaluating G-quadruplex stabilization by small molecules. *Nucleic acids research* **35**, e68 (2007).

104. Wang, X.-L. *et al.* Ultrasensitive detection of protein using an aptamer-based exonuclease protection assay. *Analytical chemistry* **76**, 5605-5610 (2004).

APPENDIX A: SEQUENCING DATA ANALYSIS SCRIPTS

The FASTQ file generated at the end of a sequencing run contains quality scores in ASCII format along with some extraneous details, irrelevant for the present analysis. The following python script restructures the FASTQ file to a .csv file that contains only the sequences of reads and the numerical values of base-call quality scores.

```
import math

class MSA:
   def __init__(self, filename=None):

      file = open("seq_and_scores_csv.txt", "w") # Open output file for printing results
      with open(filename,"r") as source: # Open input file for reading lines
         rdr= iter(source)
         for line in rdr: # Read one line at a time
            if line.rstrip('\n').split(':')[0] == '@NOER9': # Search for '@', which indicates
start of a seq-read block

               seq = next(rdr).rstrip('\r\n') # Read next line, which is a sequence
               file.write(seq) # Print sequence in output file

               next(rdr).rstrip('\r\n') # Read next line, which has nothing but a '+' and should
be ignored

               score_line = next(rdr).rstrip('\r\n') # Read next line, which has scores in ascii
format
               for q in score_line: # Convert these scores one at a time and print them in
output file
                  Q = ord(q)-33
                  file.write(','+str(Q))

               file.write('\n') # Print start of a new line
      file.close() # Close output file


if __name__=='__main__':
   # Enter file name here
   sequences = MSA('R_2014_10_06_18_54_31_user_UNC-81-GQ_200bp-316Chip-
1_12March2016.fastq')
```

Further data analysis was carried out in R using the above Python generated .csv file.

```
#Import python generated CSV file
data<- read.csv("seq_and_scores_csv.txt", header = FALSE, col.names=c(1:252))
data$X1<- as.character(data$X1)

#Sort reads into different GQs based on two-letter starting key
Layer2 <- data[substr(data$X1,1,2) == "CG",]
Layer3_Loop3_OH <- data[substr(data$X1,1,2) == "AG",]
Layer4 <- data[substr(data$X1,1,2) == "GG",]
Layer5 <- data[substr(data$X1,1,2) == "TG",]
Loop1 <- data[substr(data$X1,1,2) == "TA",]
Loop5 <- data[substr(data$X1,1,2) == "AA",]
Loop7 <- data[substr(data$X1,1,2) == "GA",]
OH3 <- data[substr(data$X1,1,2) == "TT",]
OH6 <- data[substr(data$X1,1,2) == "AT",]
OH9 <- data[substr(data$X1,1,2) == "GT",]
T29 <- data[substr(data$X1,1,2) == "CT",]
T15 <- data[substr(data$X1,1,2) == "CA",]
Control <- data[substr(data$X1,1,2) == "CC",]

# From hereon, only Layer2 is shown as an expample
Layer2Seq<-
strsplit("CGCCAAACCAAACCAAACCATCACCGACTGCCCATAGAGAGG",split="
)[[1]]

# Error Heatmap
Layer2Mismaches<- matrix(0,nrow = nrow(Layer2), ncol = length(Layer2Seq))

for(i in 1:nrow(Layer2)){
  if(nchar(Layer2$X1[i]) <= length(Layer2Seq)){
    Layer2Mismaches[i,1:nchar(Layer2$X1[i])]<-
as.numeric(strsplit(as.character(Layer2$X1[i]),split="")[[1]] ==
Layer2Seq[1:nchar(Layer2$X1[i])])
  }
  if(nchar(Layer2$X1[i]) > length(Layer2Seq)){
    Layer2Mismaches[i,]<-
as.numeric(strsplit(as.character(Layer2$X1[i]),split="")[[1]][1:length(Layer2Seq)] ==
Layer2Seq)
  }
}

Layer2Mismaches<- Layer2Mismaches[,3:ncol(Layer2Mismaches)]

tiff(file="Heatmap_2Layer.tiff",width=15000, height=15000, res = 100)
heatmap(Layer2Mismaches, Rowv = NA, Colv = NA, cexCol = 4, cexRow = 0, scale =
"none")
```

```
dev.off()

gc()

sink('% Mismatches.txt')
#Layer2
cat("Layer2: Mean =", mean((1- (rowSums(Layer2Mismaches)/
ncol(Layer2Mismaches)))*100), "\t","SD =", sd((1- (rowSums(Layer2Mismaches)/
ncol(Layer2Mismaches)))*100), "\t", "SE =", sd((1- (rowSums(Layer2Mismaches)/
ncol(Layer2Mismaches)))*100)/ sqrt(nrow(Layer2Mismaches)), "\n")

sink()

#Phreds

sink('Phred Scores.txt', append=TRUE)

cat("Layer2", mean(rowMeans (Layer2[,4:(length(Layer2Seq)+1)], na.rm = TRUE, dims
= 1)), sd(rowMeans (Layer2[,4:(length(Layer2Seq)+1)], na.rm = TRUE, dims = 1)), "\n")

sink()

sink('Phred Scores_Normalized.txt', append=TRUE)

cat("Layer2", mean((rowMeans (Layer2[,4:(length(Layer2Seq)+1)], na.rm = TRUE, dims
= 1))/mean(rowMeans (Control[,4:(length(ControlSeq)+1)], na.rm = TRUE, dims = 1))),
sd((rowMeans (Layer2[,4:(length(Layer2Seq)+1)], na.rm = TRUE, dims =
1))/mean(rowMeans (Control[,4:(length(Layer2Seq)+1)], na.rm = TRUE, dims = 1))),
"\n")

sink()


## Generate heatmap of colored bases
data2 <- data
data2$X1<- gsub("A", "1", data2$X1)
data2$X1<- gsub("T", "2", data2$X1)
data2$X1<- gsub("G", "3", data2$X1)
data2$X1<- gsub("C", "4", data2$X1)

Layer2_2 <- data2[substr(data2$X1,1,2) == "43",]

Layer2_2_matrix <- matrix(0, nrow = nrow(Layer2_2), ncol =length(Layer2Seq) )
for(i in 1:nrow(Layer2_2)){
  if(nchar(Layer2_2$X1[i]) <= length(Layer2Seq)){
```

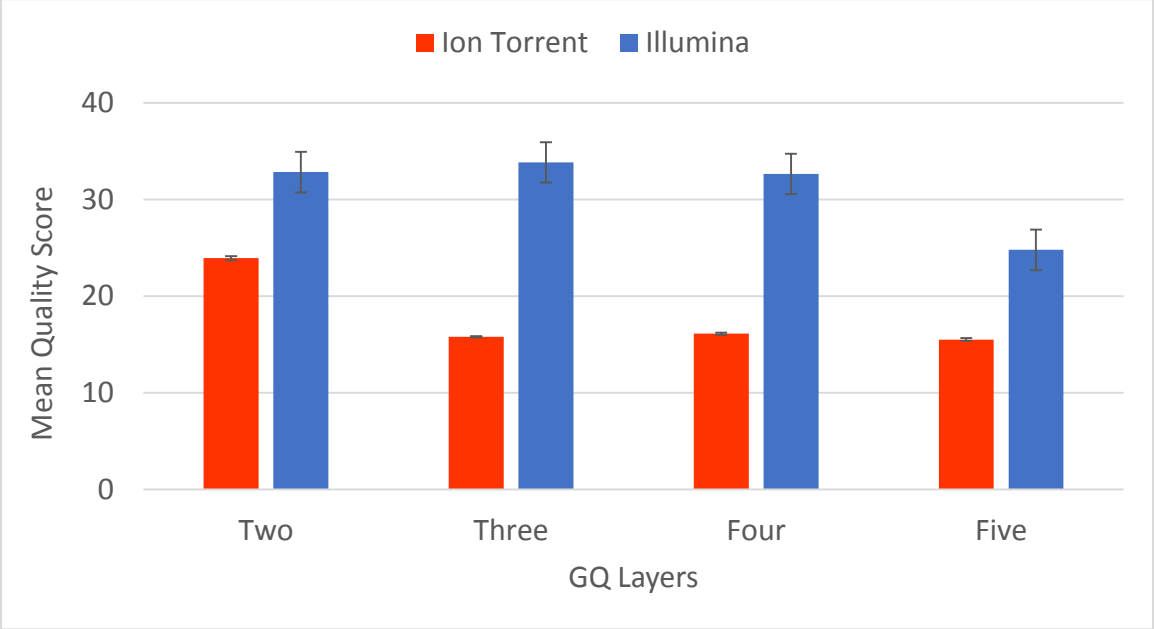
```
   Layer2_2_matrix[i,1:nchar(Layer2_2$X1[i])]<-
strsplit(as.character(Layer2_2$X1[i]),split='')[[1]]
  }
  if(nchar(Layer2_2$X1[i]) > length(Layer2Seq)){
   Layer2_2_matrix[i,]<-
strsplit(as.character(Layer2_2$X1[i]),split='')[[1]][1:length(Layer2Seq)]
  }
}

Layer2_2_matrix <-  matrix(as.numeric(Layer2_2_matrix), nrow = nrow(Layer2_2), ncol
=length(Layer2Seq))
Layer2_2_matrix <- Layer2_2_matrix[,3:ncol(Layer2_2_matrix)]

tiff(file="ColoredHeatmap_2Layer.tiff",width=15000, height=15000, res = 100)
heatmap(Layer2_2_matrix, Rowv = NA, Colv = NA, cexCol = 4, cexRow = 0, scale =
"none", col = c('white', 'orange','magenta', 'darkgreen', 'mediumblue'), xlab = "Base
Position", yaxt='n')
dev.off()
```

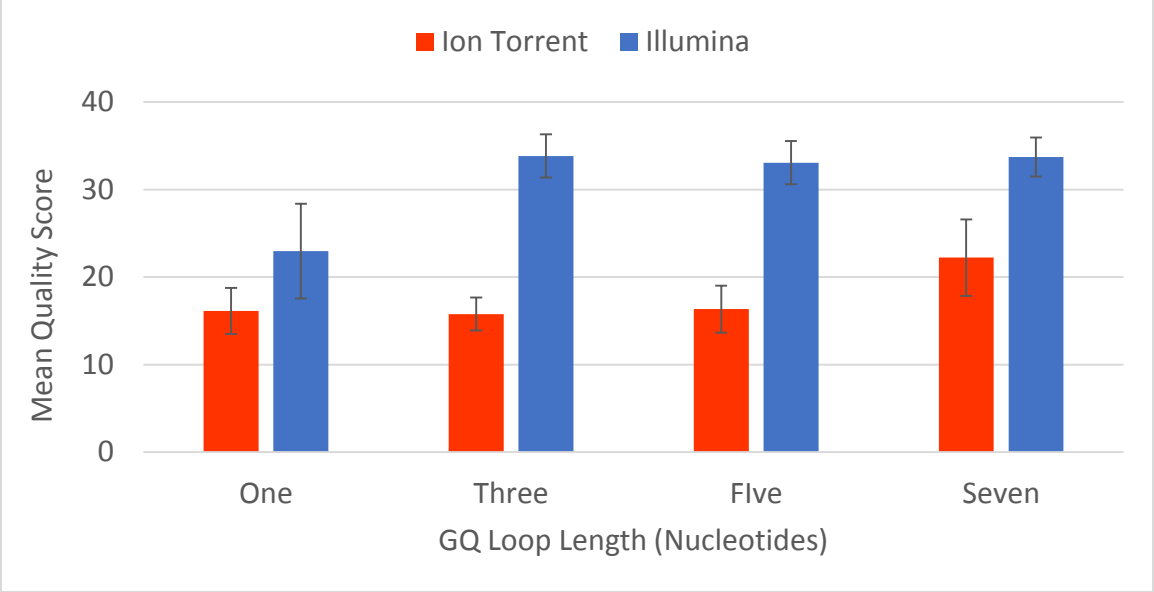
Full length template sequencesfor Illumina platform. The underlined part represents templates for Ion Torrent platform. Sequences in bold are used for data analysis on both Illumina and Ion Torrent platforms.

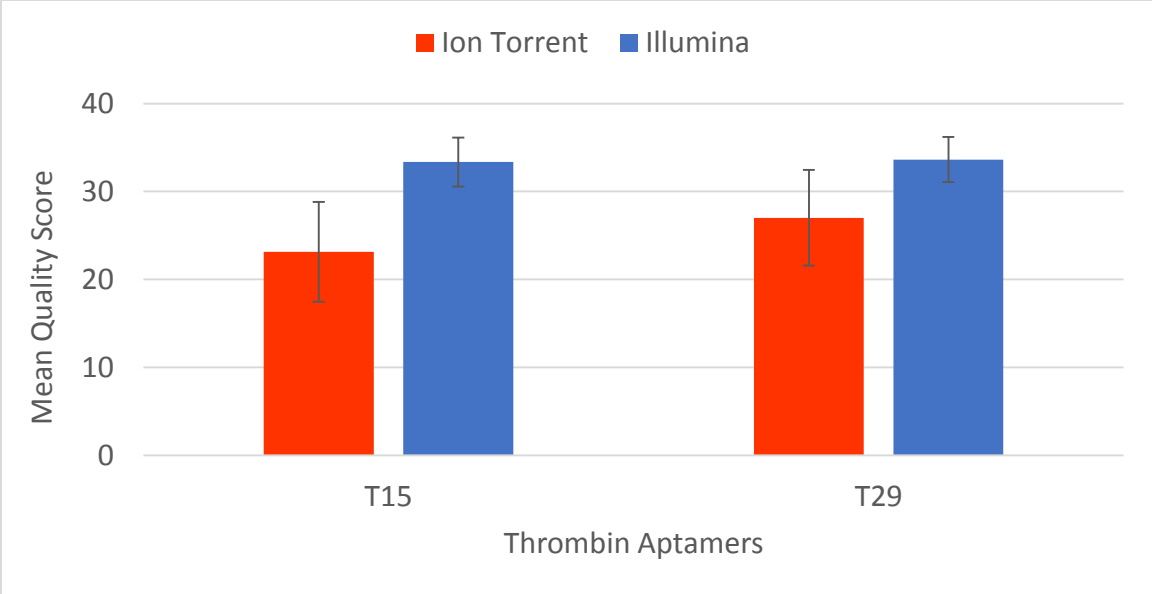
Abbreviation	Sequence (5' → 3')	Length of sequence considered for data analysis
OH/ Layer3/ Loop3	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGT ATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGGTTTGGGTTTGGGTTTGGG**CT CTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTCCGAGCC CACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	44
+3OH	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGT ATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGGTTTGGGTTTGGGTTTGGGAT TA**ACTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTCCGA GCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	47
+6OH	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGT ATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGGTTTGGGTTTGGGTTTGGGAT TATT**ATCTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTC CGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	50
+9OH	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGT ATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGGTTTGGGTTTGGGTTTGGGAT TATTATT**ACCTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACAT CTCCGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	53
Layer2	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGT ATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGTTTGGTTTGGTTTGG**CGCTGA CTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTCCGAGCCCACG AGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	40
Layer4	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGT ATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGGGTTTGGGGTTTGGGGTTTG GGG**CCCTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTC CGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	48
Layer5	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGT ATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGGGGTTTGGGGGTTTGGGGGT TTGGGGG**CACTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACA TCTCCGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	52

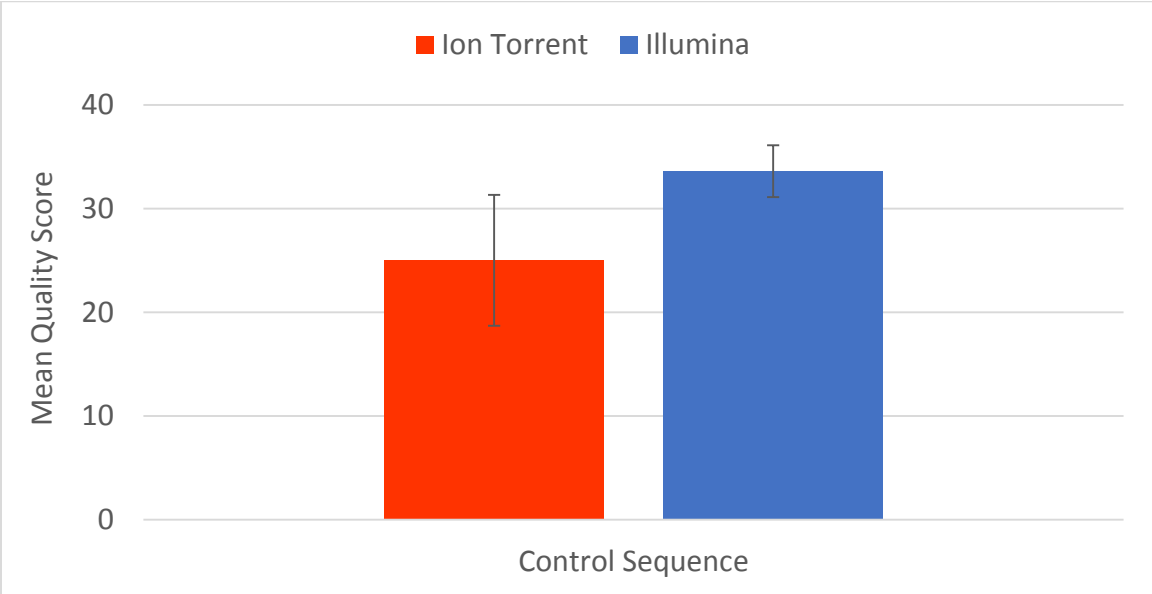
Abbreviation	Sequence (5’ → 3’)	Length of sequence considered for data analysis
Loop1	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGGTGGGTGGGTGGG**TACTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTCCGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	38
Loop5	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGGTTTTTGGGTTTTTGGGTTTTTGGG**TTCTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTCCGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	50
Loop7	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGGTTTCTTTGGGTTTCTTTGGGTTTCTTTGGG**TCCTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTCCGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	56
T29	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATAGTCCGTGGTAGGGCAGGTTGGGGTGACT**AGCTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTCCGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	52
T15	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATGGTTGGTGTGGTTGG**TGCTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTCCGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	38
Control	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**CCTCTCTATGGGCAGTCGGTGATAAACCGCCGGAGTGGTGGGAACCGCAGTTGTGTTAGTTGTAGCGCGAC**GGCTGACTGAGTCGGAGACACGCAGGGATGAGATGGCTGTCTCTTATACACATCTCCGAGCCCACGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG	71

APPENDIX C: ABSOLUTE QUALITY SCORES FOR ION TORRENT AND ILLUMINA

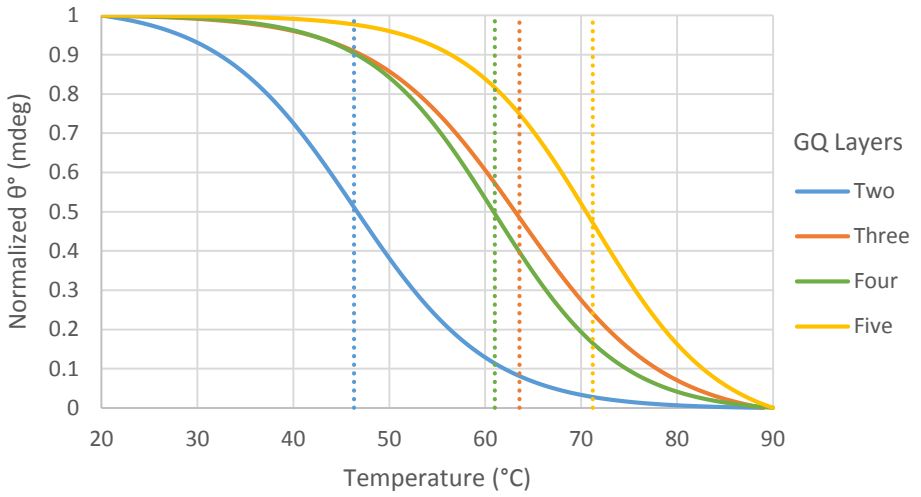




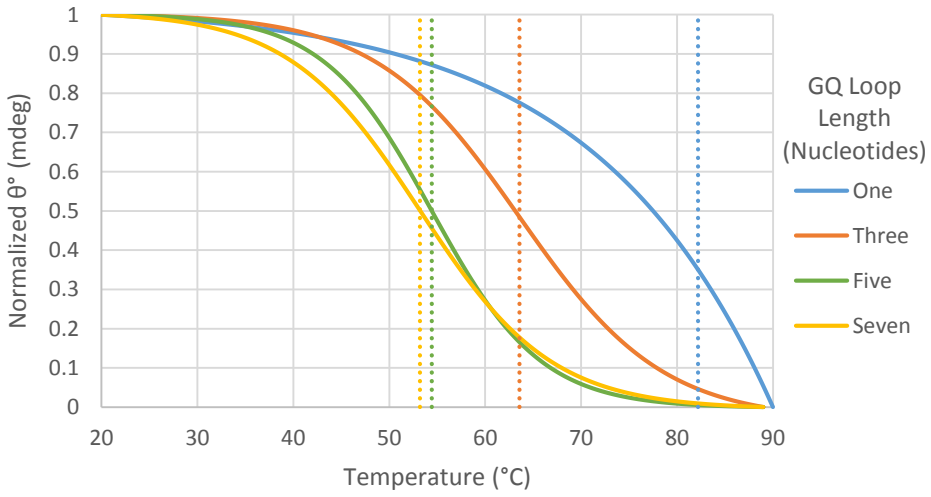

Ion Torrent
Illumina
40
30
20
10
0
Mean Quality Score
T15
T29
Thrombin Aptamers


Ion Torrent
Illumina
40
30
20
10
0
Mean Quality Score
Control Sequence

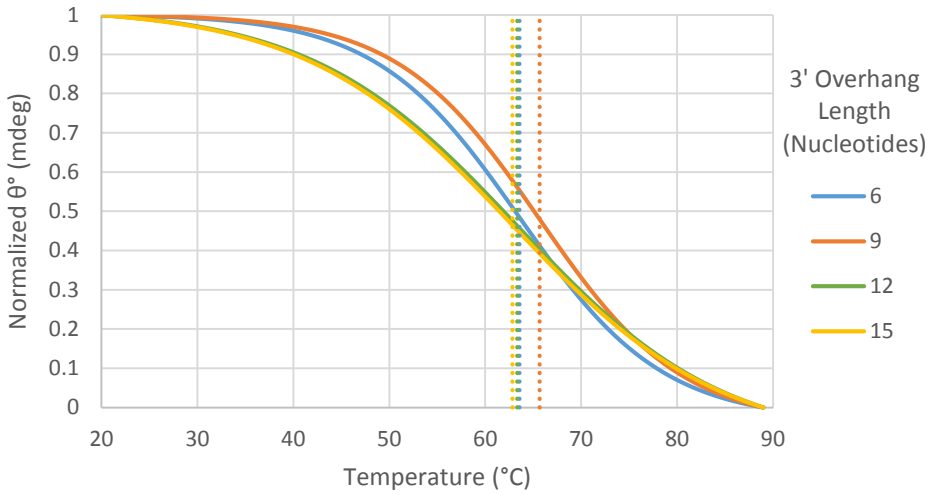
APPENDIX D: THERMAL DENATURATON CURVES



CD melting curves for templates with varying number of GQ layers. Vertical dotted lines indicate T_ms.

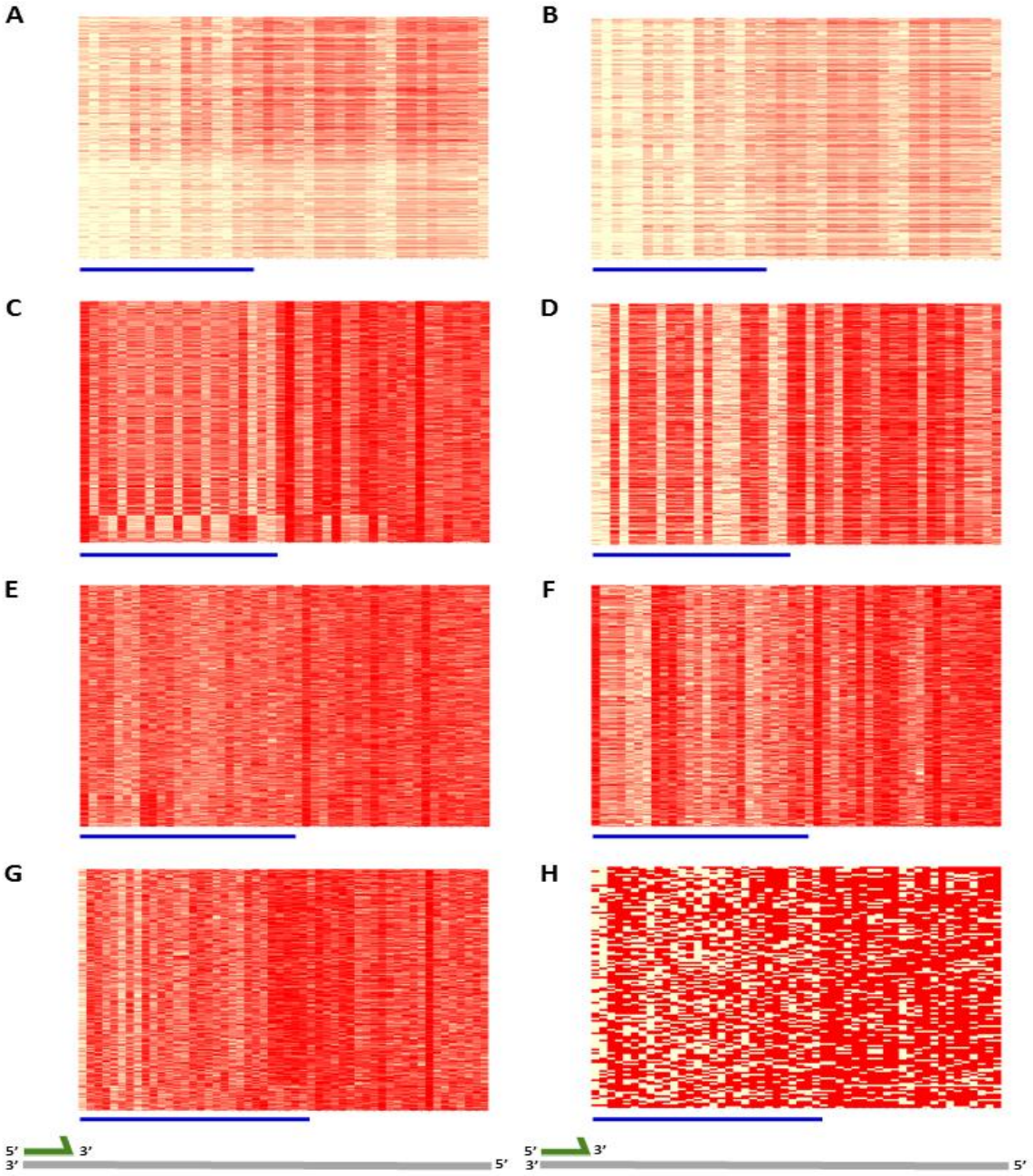


CD melting curves for templates with varying lengths of GQ loops. Vertical dotted lines indicate T_ms.

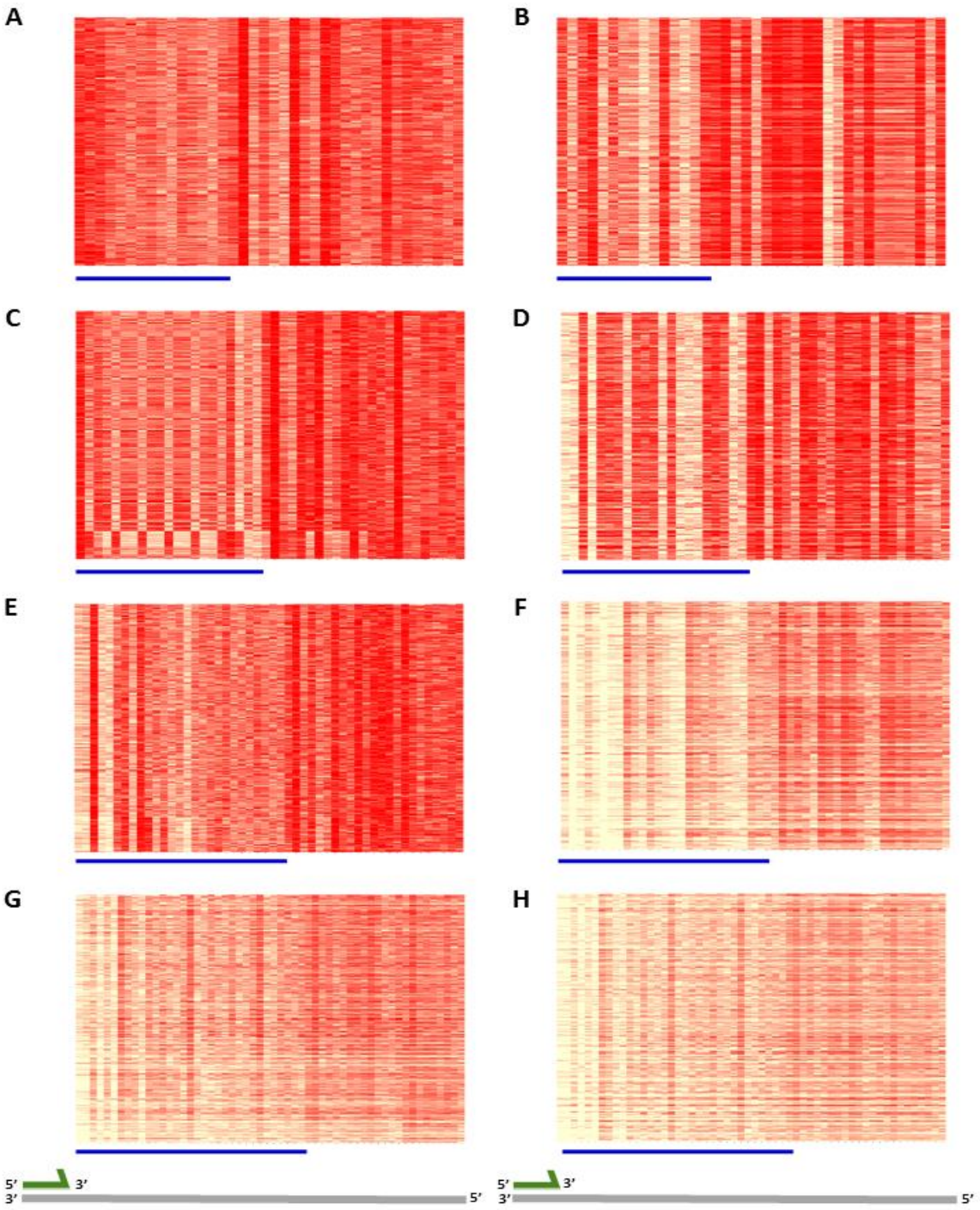


CD melting curves for templates with varying 3’ overhang lengths. Vertical dotted lines indicate T_ms.

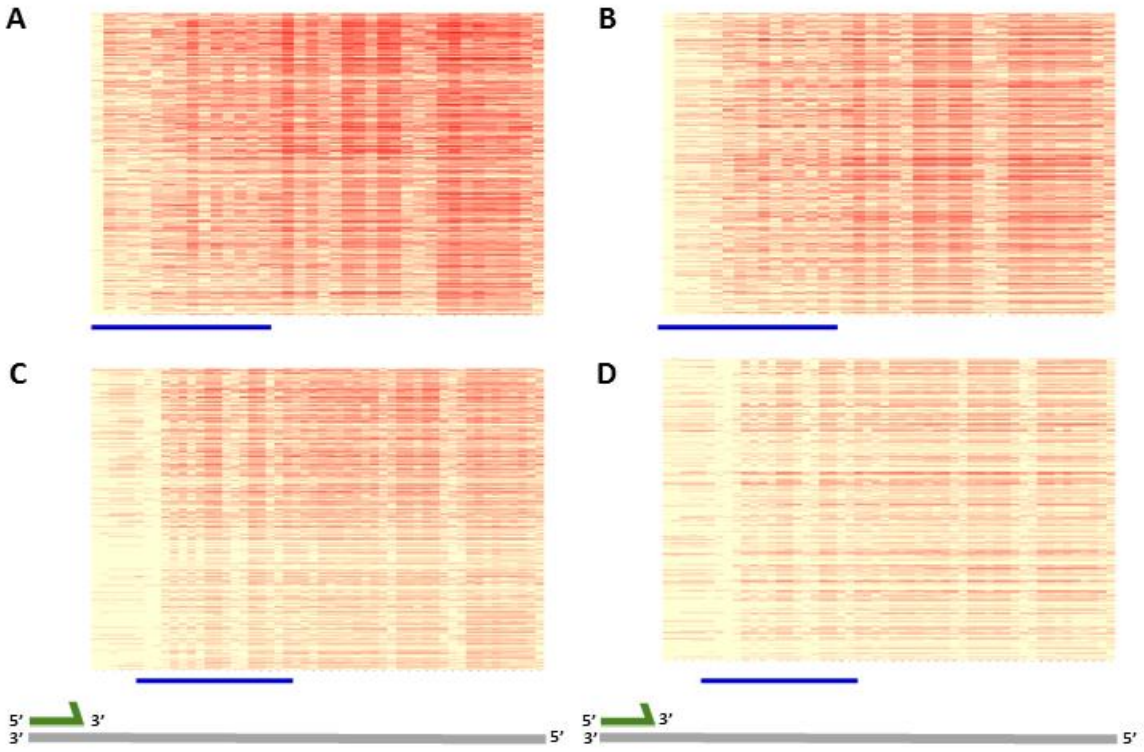
APPENDIX E: ERROR HEATMAPS FOR SEQUENCING WITH SSB PROTEIN



Position-wise base call errors, shown as a heatmap plot, with red indicating incorrect base-calls. (A), (C), (E), (G) represent plots for two, three, four and five-layered templates sequenced under standard sequencing conditions, while (B), (D), (F), (H) represent the same respective templates sequenced with SSB protein. Each horizontal line represents an independent read. Each column is a base position in the template. The horizontal blue line below each plot indicates the part of the sequence that forms GQ, with the first GQ base shown as a circle.



Position-wise base call errors, shown as a heatmap plot, with red indicating incorrect base-calls. (A), (C), (E), (G) represent plots for templates with one, three, five and seven nucleotides in loops, sequenced under standard sequencing conditions, while (B), (D), (F), (H) represent the same respective templates sequenced with SSB protein. Each horizontal line represents an independent read. Each column is a base position in the template. The horizontal blue line below each plot indicates the part of the sequence that forms GQ, with the first GQ base shown as a circle.



Position-wise base call errors, shown as a heatmap plot, with red indicating incorrect base-calls. (A), (C) represent plots for T15 and T29 templates under standard sequencing conditions, while (B), (D) represent the same respective templates sequenced with SSB protein. Each horizontal line represents an independent read. Each column is a base position in the template. The horizontal blue line below each plot indicates the part of the sequence that forms GQ, with the first GQ base shown as a circle.