Austin Journal of Proteomics, Bioinformatics & Genomics

Open Access

Austin Publishing Group

Special Issue - DNA Sequencing Technology

A Non-Traditional Approach to Whole Genome Ultra-Fast, Inexpensive Nanopore-Based Nucleic Acid Sequencing

Kanavarioti A*
Yenos Analytical LLC, USA

***Corresponding author:** Anastassia Kanavarioti, Yenos Analytical LLC, El Dorado Hills, CA, USA

Received: November 02, 2015; **Accepted:** November 19, 2015; **Published:** November 24, 2015

Abstract

DNA sequencing claims responsibility for breakthroughs in understanding the molecular basis of life, and improving quality of life through advances in prognosis, diagnosis, treatment, and cure of disease. The last 30 years have seen an exponential improvement onto the original Sanger sequencing by synthesis, as well as the emergence of new technologies. Still the mandate for cheaper, faster, longer, and more accurate reads hasn't been satisfied. We are proposing a single molecule approach by combining unassisted nanopore-based sequencing with labeled DNA, where the ion-channel readout of current *vs*. time (i-t) may represent base sequence. Pyrimidines on DNA are labeled selectively with Osmium tetroxide 2,2'-bipyridine (OsBp) ahead of sequencing. The OsBp label slows down the translocation to detectable levels, and provides base discrimination between labeled deoxythymidine, labeled deoxycytidine, and an intact base. This technology promises to sequence DNA with no limit in strand length, without amplification, and without the use of a processing enzyme; it requires consensus building, but no assembly and no scaffolding. To facilitate consensus building for a human chromosome long DNA, highly repetitive DNA sequences, such as the Alu repeats, may serve as markers. Observed translocation times of a series of osmylated oligos via the wt α-Hemolysin nanopore are exploited to estimate the time it takes (1 hour) to sequence a 100,000,000 bp genome at 128x coverage using one MinION™ device from Oxford Nanopore Technologies. This technology has the potential for mapping protein bound regions in dsDNA, sequencing RNA, as well as identifying methylated and other rare bases.

Keywords: α-Hemolysin Nanopore; DNA sequencing technology; Ion-channel measurements; Osmium tetroxide bipyridine; Whole genome

Abbreviations

C: Cytidine; T: Thymidine; ds: Double Stranded; ss: Single Stranded; i-t: Current *vs.* Time; OsBp: Osmium Tetroxide 2,2'-Bipyridine; bp: Base-Pairs; nt: Nucleotides; α-HL: α-Hemolysin; SBS: Sequencing By Synthesis

Introduction

DNA sequencing was enabled by the Sanger approach exploiting the enzymatic synthesis of the complementary of a target DNA strand using deoxynucleotide triphosphates and a small amount of the dideoxynucleotides that serve as chain terminators [1]. Each dideoxynucleotide carried a different fluorescent label, so that it was optically identifiable and distinguishable from the others. Since the seminal paper of Sanger in 1977 [1] remarkable technical progress resulted in the 2001 sequencing of the human genome, valued at 2.7 billion in FY 1991 dollars, that took over a decade to complete [2,3]. Cost and analysis time have dramatically decreased since then, but expensive instrumentation, consumables, as well as analysis time are still prohibiting the effort from being routinely implemented [4-6]. The progress is primarily due to engineering advances in miniaturization, parallelization, and computing speed. Most commercial DNA sequencers use the Sequencing by Synthesis (SBS) approach [7-9]. Besides much leverage gained by bioinformatics, the issues still are: (i) the four chemistries behind labeling nucleotides are less than 100% efficient, introducing insertion and deletion errors [9]; (ii) the amplification process of the target DNA, if required ahead of sequencing, has its own limits in amplifying sequence repeats [7]; (iii) library construction, primer incorporation, and amplification add complexity and expense to sample preparation, (iv) the polymerase enzymes typically synthesize up to a few thousand bases of the complementary strand, and then dissociate [8], limiting the maximum output length, and (v) the enzymatic synthesis is a relatively slow process [10] that is further slowed down by the change of reactants between mononucleotide additions. These issues are overshadowed by the fact that the SBS approach yields lengths that are a miniscule fraction of a whole genome, and presents challenges in the analysis of the data by requiring error correction, assembly, and scaffolding bioinformatics [11]. Read length in sequencing directly impacts sequencing accuracy. Specifically a read should be long enough to span a repetitive region in the genome. Practically speaking it takes a village, i.e., a number of scientists with completely different background and skills, to sequence a small genome and,

Citation: Kanavarioti A. A Non-Traditional Approach to Whole Genome Ultra-Fast, Inexpensive Nanopore-Based Nucleic Acid Sequencing. Austin J Proteomics Bioinform & Genomics. 2015; 2(1): 1012.

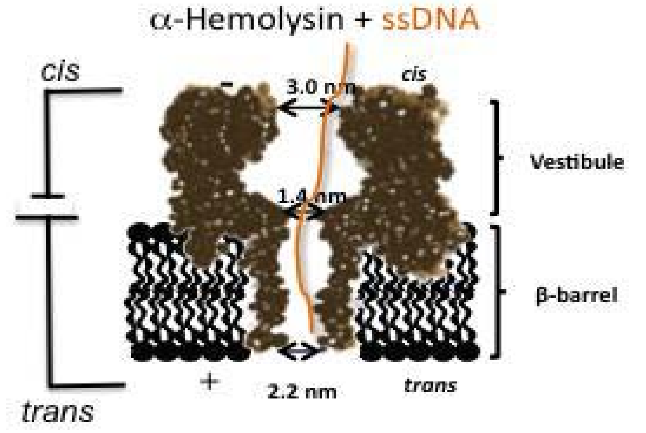

Figure 1A: Translocation of ssDNA via the α-HL nanopore showing the 1.4 nm constriction zone and the rather long but confined b-barrel; voltage (positive, trans to cis) across the insulated nanopore leads to ion current via the pore and threading of the ssDNA, which obstructs the current when inside the pore.

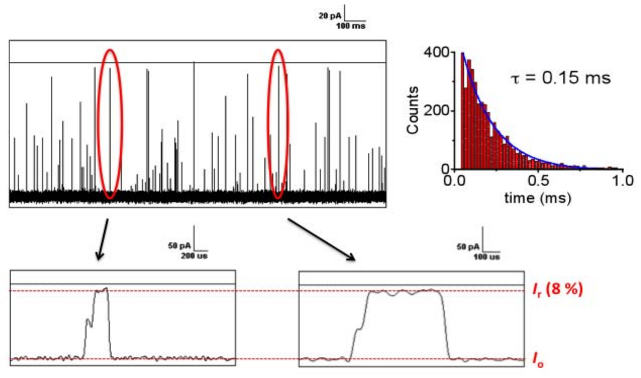

Figure 1B (from [29]): Top Left, Observed conductance measurement current vs time (i-t in ms) profile shown for dA$_{10}$dT(OsBp)dA$_9$ via the α-HL nanopore at 120mV in 1M KCl, pH 7.4 with 10mM PBS, at 22±1°C. Top Right, Single molecule counts as a function of translocation with average dwell time τ=t=0.15ms. Bottom: Two translocation events selected and shown magnified (time in μs) to show current obstruction at a relative residual current I_r/I_o = 8%, with I_o ≈ 120μA. Events with much higher relative residual current are attributed to events other than complete translocation of a molecule.

despite all of the *de novo* assembly efforts, *de novo* sequencing is still a challenge [12,13].

To move forward a non-traditional approach may be more suitable. One such strategy is using nanopores, with sub 2nm diameter, located within an isolated membrane that separates two compartments filled with electrolyte (Figure 1A). Applying a voltage across the two compartments leads to a constant flow of ions via the nanopore; this flow is partially blocked by the occasional passage of a single molecule through the pore [14]. Numerous studies have explored translocation of single stranded (ss) nucleic acids via the α-Hemolysin pore (α-HL) and show that conductance measurement (i-t) yield current modulation and translocation time (Figure 1B) with sequencing information attributed to nanopore/nucleobase interactions [15-20]. Even though this strategy avoids synthesis of the complementary strand, it is still being explored assisted by an enzyme to slow down the otherwise too fast translocation, and act as a motor to move the DNA strand one base at a time [21]. In this report we review DNA translocation via α-HL, in the absence of an enzyme, and the use of labeled DNA to differentiate the bases, and simultaneously slow down translocation [22,23]. We will discuss what are the potential gains from this "nanopore/labeled DNA" approach, and outline the requirements to make it a preferable alternative to current technologies.

The mandate for improvements in DNA sequencing technologies is cheaper, faster, longer, and more accurate reads [4,18]. Recent literature reveals that combination of highly accurate short reads from one technology, such as Illumina Seq, with low accuracy long reads from another technology, such as Oxford Nanopore (ONT) or Pacific Biosciences (PacBio) platforms yields markedly improved draft genomes [24,25]. One such combination yielded accurate (99.88%) read for a contig (resulting sequence) less than 6% of the Saccharomyces cerevisiae genome which is only 10% of the length of an average human chromosome, indicating that *de novo* sequencing is still a challenge [24]. What ONT and PacBio technologies have in common is massive single molecule parallelization/detection accompanied by reads whose length is limited by the enzyme's dissociation rate to a range of 5 to 35 kbases for PacBio and 5 to 20 kbases for ONT. Both technologies share the "enzyme assistance", and also share a relatively high number of errors in base calling, about 12% errors due to the synthesis of the complementary in the PacBio case [25], and about 30% errors due to the nanopore that senses a short sequence of bases and not a single base with ONT [24]. Nevertheless ONT enjoys a distinct advantage in claiming the capability to sequence a sample using a small device the size of a USB drive (MinION™) in any location equipped with a laptop computer, Internet access, and cloud service [19]. In order to improve beyond what those two technologies are offering, the development of a faster, enzyme-free technology, with unlimited length reading capability, and better accuracy in base calling, is required; such technology would easily fulfill the above mandate.

The PacBio technology uses a polymerase, because it is a SBS technology. In contrast to PacBio, the ONT technology is not based on synthesis, so it doesn't need the polymerase per se. However the observation that translocation of ssDNA via nanopores is too fast for detection [10,15], led to the use of DNA polymerase to slow it down, and also to move the strand forward one base at a time. Due to the proof reading function of the polymerase, DNA processing is not always at a constant speed, and not always in the forward direction. This interrupted movement misleads the reading process [26] and has led to the exploration of other enzymes that act as molecular motors [27]. Most importantly the movement of the strand led by an enzyme is 100 to 1000 times slower compared to state-of-the-art detection speed [10], causing undesirable delay when reading millions of bases at high coverage is the task at hand. Other alternatives to the PacBio and ONT single molecule technologies are being explored, and have been discussed elsewhere [28]. One specific approach, nanopore-based, will be reviewed here. It was recently proposed and has had limited experimental evaluation; notwithstanding exhibited very promising results [29]. It is based on labeling selectively one or more of the nucleobases of the target DNA. The function of the label is to gain base differentiation, and provide bulkiness to slow down the translocation. The labeling approach is not new; it was successfully

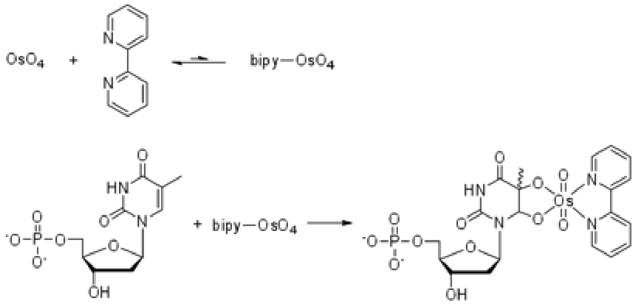

Figure 2: Reaction of osmium tetroxide with 2,2'-bipyridine forms a reactive complex (bipy-OsO4 or OsBp), which in a second step reacts with the C5-C6 double bond of a pyrimidine (thymidine monophosphate (dTMP) shown here) to form the osmylated-thymidine, dT(OsBp). OsBp adds from either side to the double bond of the pyrimidines, shown by Capillary electrophoresis analysis, and yields two topoisomers, one from either side of the double bond at a ratio of about 2:1. Notably addition from either side is not inhibited in a sequence of Ts, as found experimentally [22,33]. One way to illustrate the difference between osmylated and intact bases is to compare (molecular weight) of each: dC (111), dT (126), dA (135), dG (151); dC-OsBp (521), dT-OsBp (536), i.e. osmylation adds about 400% mass to the reactive base compared to an unreactive one.

implemented using specific peptidic or pegylated attachments to the bases [30-32], but direct labeling of a genomic length DNA is not feasible using these methods. There is a clear distinction between labeling the mononucleotides versus labeling the target DNA strand. Even though the chemistry behind mononucleotide labeling has improved over the years, it is still not 100% perfect. However scientists feel comfortable using it, because of the easiness in analyzing the labeled mononucleotide product and assessing purity and impurities. In contrast to labeling the mononucleotides, labeling the bases in the target DNA has seen limited exposure [22,33]. This is due to the absence of analytical methods to evaluate reactivity and product purity at the single base level in a polymer composed of a few million of bases. The discussion of a case study below will hopefully change this attitude.

While working with metalorganic molecules to label ssDNA, we evaluated Osmium tetroxide 2,2'-bipyridine (OsBp) [33]. OsBp is known to add to the C5-C6 double bond of the pyrimidine ring (Figure 2). Because Osmium is a good contrast agent for imaging by electron microscopy (EM), osmylated DNA (DNA(OsBp), Note) was proposed 60 years ago and exploited in attempts to obtain DNA sequence information by EM imaging [34-36]. The more recent advancement of nanopores as single molecule detection devices, and the corresponding progress in manufacturing, parallelization, and commercialization of such platforms [37], supported the idea of testing DNA(OsBp) as a surrogate for intact DNA (Figure 3). The proposition was motivated by studies showing that the osmylation is a remarkably clean reaction yielding products in practically 100% yield with no detectable side-reactions. The selectivity of this reaction for one of the pyrimidines over the other, Deoxythymidine (dT) over Deoxycytidine (dC), is 30-fold, and easily leads to labeling of either dT only, or dT+dC [22,33].

5'....CCG GGA GCT GCA TGT GTC AGA GG....3' the target strand
5'.... 000 000 001 000 101 010 000 00....3' Protocol A
5'.... 220 000 021 020 101 012 000 00....3' Protocol B

5'....CCT CTG ACA CAT GCA GCT CCC GG....3' complementary
5'....001 010 000 001 000 001 000 00....3' Protocol A
5'....221 210 020 201 020 021 222 00....3' Protocol B

Figure 3: Sequencing strategy where 1=dT(OsBp) and 2=dC(OsBp). All sequences shown refer to deoxybases; for simplicity d is left out. For successful sequencing both the target strand and its complementary should be sequenced. Sequencing of the complementary strand is necessary so that dA and dG in the target strand can be identified via the corresponding dT and dC in the complementary. (i) Protocol A requires 60 min incubation at room temperature, and yields 90% dT(OsBp) and 6.5% dC(OsBp); Protocol B yields practically 100% osmylated pyrimidines [22,23]. As shown experimentally α-HL discriminates by both, relative current levels and translocation times, between dA, dT(OsBp), and dC(OsBp), shown here as 0, 1, 2, respectively [29]. Presumably discrimination between dA and dG, if any, is not detectable.

Unpublished data of the osmylation reaction suggest false positives and false negatives to be below 1/10,000, a remarkable feat for any modification reaction. False positive refers to the undetectable reactivity of OsBp with the purines, and false negative refers to unreacted pyrimidines. Development work led to pseudo-first order conditions that render percent of unmodified pyrimidine a function of OsBp concentration and incubation time [33]. Extensive studies with short, specifically designed, oligos, for which all products could be identified analytically, led to the conclusion that OsBp labeling is independent of sequence, length, and composition [22,33]. Critically important for sequencing is the observation that the reactivity is not altered within long sequences of pyrimidines. This was evidenced by the rate for complete osmylation of dT_{15} that is, within experimental error, comparable to the rate of monomer osmylation (dTTP to dTTP (OsBp)) [22,33]. It turns out that the same protocol – in the absence of any denaturing agents - works predictably and reproducibly for short and long oligos, as well as for M13mp18, a circular ssDNA with 7249 bases and secondary structure [22]. The success in using the same protocol for long ssDNA with secondary structure and for short oligos is attributed to the hydrophobicity of the OsBp moiety that disrupts base-pairing and base-stacking. This feature implies that any ssDNA of unknown sequence, presumably including any type of DNA repeats, can be predictably osmylated. Most importantly, analytical methods were developed so that the extent of labeling can be assessed independently by a simple UV-Vis assay after removal of the excess label. Specifically, the pyrimidine/OsBp adduct exhibits a new chromophore in the range 300 to 320 nm where DNA does not absorb [22,33]. This chromophore was the basis for developing the UV-Vis assay to quantitatively measure extent of osmylation, and facilitate the development of two protocols (Figure 3). Protocol A exploits low concentration of OsBp and short incubation, and yields primarily dT(OsBp), and Protocol B uses higher concentration with longer incubation, and yields practically 100% (dT+dC)(OsBp); both protocols work at room temperature. The UV-V is assay serves as a quality control assay (±3%) to confirm extent of osmylation. Since osmylation is not inhibited in a 6 M urea solution, long DNA with secondary structure can be osmylated both in the presence/absence of 6 M urea, and then determined, using the UV-Vis assay, whether osmylation extent is comparable under the two conditions.

Table 1: Observed dwell times (t) of oligos via α-HL at two voltages [29], and calculated t per pyrimidine(OsBp) unit. From [29] conditions: 1M KCl, 10mM PBS buffer, pH 7.4 with 10μM oligo at 22±1°C.

	120 mV			140 mV		
Oligo (*=OsBp)	Oligo, observed translocation speed t (μs)	Total intact bases (μs)	Osmylated unit, corrected translocation speed t (μs)	Oligo, observed translocation speed t (μs)	Total intact bases (μs)	Osmylated unit, corrected translocation speed t (μs)
dA$_{20}$	50	50	-	30	30	-
dA$_{10}$dT*dA$_{9}$	150	47.5	102.5	100	28.5	71.5
dA$_{10}$5-MedC*dA$_{9}$	310	47.5	262.5	240	28.5	211.5
dA$_{10}$dC*dA$_{9}$	360	47.5	312.5	260	28.5	231.5
dA$_{10}$dU*dA$_{9}$	470	47.5	422.5	360	28.5	331.5
pGEX3’ (4T*)[1]	890	47.5	842.5/4=211	490	28.5	461.5/4=115
pGEX3’ (4T*+5C*)[2]	4200	35	(4200-842.5)/5= 665[2]	3500	21	(3479-461.5)/5= 603[2]

[1,2] pGEX3’ is a 23nt long PCR primer with sequence: 5’CCG GGA GCT GCA TGT GTC AGA GG3’

[1] Calculated contribution per dT(OsBp) within a sequence of OsBp/total bases=4/9

[2] Calculated contribution per dC(OsBp) within a sequence of OsBp/total bases=9/18, after subtracting the contribution of the 4dT(OsBp) (see text).

Discussion

Since osmylation increases the mass of the reacting base by 4-fold (see Figure 2, caption), it fueled the speculation that any size-suitable nanopore could discriminate between osmylated and native base. Preliminary experiments to assess pore size suitability using solid-state silicon nitride (SiN) nanopores showed that 3nm long and 1.6nm wide SiN pores permit translocation of 80-mer long osmylated oligos, and exhibit dramatic tranlocation slowdown with increasing osmylation [38]. These observations led us to undertake experiments with wt α-HL. Personnal communication from Professor Mark Akeson of the Genomic Institute of the University of California in Santa Cruz indicated that 80-mers with a consecutive sequence of 19 osmylated pyrimidines successfully translocate via the α-HL nanopore, albeit very slowly. Additional experiments with short, specifically designed oligodeoxynucleotides (oligo) dA$_{10}$XdA$_{9}$ via α-HL showed slow and distinct translocation times for different X=deoxypyrimidine (see Table 1, and in [29]. Relative residual current values I_r/I_o are also distinct, but not as dramatically different as the corresponding dwell times (t). For X = dA, dT(OsBp), dC(OsBp), dU(OsBp), or 5-MedC(OsBp) I_r/I_o are equal to 0.14, 0.08, 0.11, 0.12 and 0.12, respectively [29], accurate to ±0.01. These data indicate that the α-HL constriction zone/β-barrel interacts strongly with both the OsBp and the base moieties. The dwell times listed in Table 1 are well above detection limits, and show proof-of-concept for base-differentiation that may lead to nanopore-based sequencing using osmylated DNA.

DNA (OsBp) Structural Insights

Inspection of Figure 3 suggests that if there was one perfect label for each base, and if a nanopore could clearly discriminate among them, then sequencing of the target strand would be sufficient. This scenario is not very different from the scenario where a perfect nanopore discriminates among dA, dG, dT and dC; both are too good to be true. The possibility of finding a second label, comparable to OsBp, but reactive towards the purines and selective for one over the other is likely [39,40]. In the presence of such two labels sequencing of target strand would be sufficient. It is anticipated that two labels will be required for direct RNA sequencing, where typically only one strand is present. In the presence of one label, such as OsBp, it is necessary to sequence both strands at two levels of osmylation, in order to assign all four bases (see discussion below). Actually the data listed in Table 1 are consistent with OsBp not “hiding” the base it is attached to, since electrophoretic values indicated differences based on nucleobase identity. Structural considerations indicate that OsBp lines up parallel to the strand direction, extends all the way to the second adjacent neighbor, and most likely “hides” the immediately adjacent base (see Figure 4, and discussion below). “Hiding” of the adjacent base by OsBp may not be an issue based on the following considerations: If the adjacent base is a purine, then its identification is done by sequencing the complementary strand. If the adjacent base

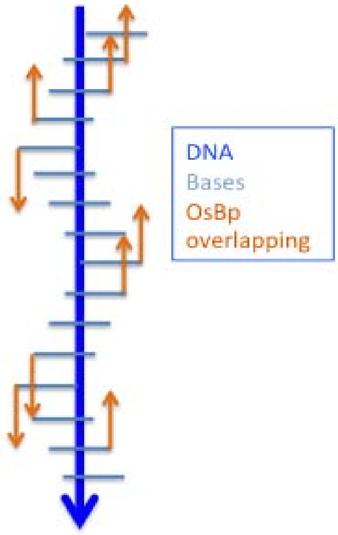

Figure 4: Osmylated DNA strand representation to show the approximately parallel line up of OsBp moieties along the strand, the top or bottom conjugation of OsBp with the nucleobase (see Figure 2 caption), the extension of OsBp to obscure next-door neighbor, and the plausible overlap of two OsBp moieties. The later is consistent with the observed twice as slow translocation time within sequences with multiple pyrimidines (see Table 1). In this two-dimensional representation some interactions appear artificially close and others apart. Please note that in ssDNA, adjacent bases can take positions practically across from each other in order to minimize next-door neighbor OsBp interactions.

is another pyrimidine, then its identification may be based on the high selectivity of OsBp for dT over dC as well as on the electrophoretic properties of the combined duet, when fully osmylated. Experiments are necessary to show that osmylated dTdT translocates with different properties compared to fully osmylated dTdC, or dCdT, and that fully osmylated dT *vs.* dTdT or *vs.* dTdTdT or *vs.* dTdTdTdT are distinct and hence detectable; similar considerations apply to osmylated dC *vs.* dCdC *vs.* dCdCdC *vs.* dCdCdCdC, etc. Discussion of the data in Table 1 (see below) clearly suggests that the above discriminations are plausible.

Proposed Methodology for Sequencing Osmylated DNA

Counting bases between markers

Central to our model are three requirements: (i) to identify every single position of dT(OsBp), (ii) to identify every single position of dC(OsBp), and (iii) to determine the number of intact bases between modified positions. Based on Table 1 dwell times at 120mV per unit for dA, dT(OsBp) and dC(OsBp) are 2.5µs, 102.5µs and 312.5µs, respectively. For the sake of this discussion we assume that any unmodified base translocates like dA, i.e. with 2.5µs dwell time. These data suggest that the first two requirements are met, i.e. identification of intact base, and osmylated dT or dC can be done with marked discrimination using dwell times compared to relative residual current, in contrast to what is suitable for intact DNA [10,20,37]. Earlier studies showed that translocation duration is proportional to the number of bases [15]. Hence osmylated bases, identified as inter-event current obstructions "spikes" may act as primary markers, so that bases in between can be counted. In order for the number of intact bases to be determined from the time interval between spikes, the duration needs to be 50µs or longer, and hence intact bases need to be many more compared to the labeled ones. A large number of bases remain intact only when the extent of osmylation is low. Because the selectivity of OsBp for dT over dC is 30-fold, at low osmylation levels practically most modified bases will be dT(OsBp) [33]. Since labeling is random, i.e., length, sequence, and composition independent, then even at a low level of osmylation all the dT positions will be osmylated within a population of strands, but only a few dTs will be osmylated per strand leaving long stretches of intact bases that may be counted; accuracy should be determined experimentally. Once all the strands are lined up along each other, a consensus strand should form (see more below), and in this consensus strand every single dT will be positioned in its proper position among the other three bases. dC(OsBp) at this level will be infrequent, and easy to trace due to its different electrophoretic properties from dT(OsBp).

Determination of all Ts and Cs using protocol A

Since the aim of this technology is to sequence whole genomes, it will be of practical use to have consensus strands that are much shorter than a whole genome. To get short consensus strands one needs secondary markers along the genome. We propose to use sequences of consecutive dTs, perhaps 4 or 5dTs, as such markers. Five consecutive dTs may be rare but abundant enough to be used as markers to split the target strand into consensus sequences of more manageable length. The assumption is made that the dwell times of 2, 3, 4 or 5 consecutive dTs will substantially increase as a function of the # of dTs. Protocol A yields 90% of osmylated dTs and 6.5% of osmylated dCs [33]. With sufficient coverage to account for the 90%, instead of 100% labeling, all dTs will be identified. Once a consensus strand is marked front and back by, let us say, fully osmylated dTdTdTdTdT (secondary marker), then the number of bases in between can be obtained from the time interval between the markers, by subtracting out the dwell times of every spike in between the markers, and dividing by the dwell time of the intact base as estimated using standard homopolymers under the same experimental conditions. The series of consensus strands, between these secondary markers, can now be used as a continuous ladder to line up the strands obtained with the lower level osmylation protocol, and redo the calculation of the intact bases while having much fewer osmylated dTs to subtract. This comparison should lead to an improved determination of the number of overall bases between secondary markers. Other secondary markers may be used as well: For example, highly repetitive DNA sequences, such as minisatellites or the Alu repeats in the human chromosomes [41,42], may be exploited due to the anticipated easily identifiable i-t pattern. For example, Alu repeats typically appear with an average length of 3000 bases in between. It is rather ironic that the one feature, i.e. DNA repeats, that has hindered completion of most genomes, is the feature anticipated to build consensus in the proposed nanopore-based/ osmylated DNA technology. Using Protocol A osmylated strands can also yields positions of dCs, easily detectable due to the different properties. Not every dC in a strand will be osmylated, but all dCs will appear osmylated with sufficient coverage. Identification of all dCs is expected due to the proven independence of dC(OsBp) from length, sequence, and composition [22].

Determination of all As and all Gs of the target strand using the above two protocols

dA and dG of the target strand will be obtained from the complementary strand after conducting both, low level and Protocol A osmylation experiments, as described above. Taken together these readings should provide the final draft base sequence. It is anticipated that for each experiment there will be two families, one family of strands that translocated from the 3'end and a second family of strands that translocated from the 5'end [43]. Evidently the sequence obtained for one family, if read from the back, should be identical to the sequence of the other family. The higher the coverage, the higher the confidence/accuracy will be for the draft sequence of the target DNA. It is worth noting that experiments at different levels of osmylation for the two strands, target and complementary, could be conducted simultaneously using a MinION™ device, that contains up to 512 functional α-HL nanopores, assuming that the software can handle developing two separate sequences at two directions each. Unpublished data suggest that dsDNA can be denatured/osmylated with Protocol A, and the resulting osmylated product is in the form of ssDNA(OsBp). This product typically does not hybridize back. However with low osmylation levels partial hybridization may be possible. When strands, target and complementary, at low osmylation levels are present in the same solution, a 4 M urea is recommended in order to prevent partial rehybridization; 4 M urea was shown to be compatible with the stability of the α-HL nanopore as well as with the translocation of ssDNA [44]. Using the MinION™ with the 512 α-HL nanopores for sequencing a target dsDNA leads to theoretical 128x coverage for both strands at both low level and Protocol A osmylations.

How to read a sequence resulting from protocol A osmylation?

One may envision reading a sequence from i-t traces from a low level osmylated ssDNA, let us say 2 to 3 osmylated Ts per 100 nucleotides, as follows. Profiles i-t will be governed by unmodified base translocation (I_r/I_o = 0.14), interrupted by a number of spikes attributed to dT(OsBp) obstructions. As seen in Table 1 at 120mV translocation of a single dA is estimated at 2.5μs, whereas translocation of a single dT(OsBp) is estimated at 102.5μs, i.e. 41 times longer compared to dA. Assuming a 2.5% dT(OsBp) over total number of nucleotides, the average duration of one dT(OsBp) translocation and the average duration of the translocation of a sequence of 41 intact bases might be about equal. Reading a sequence from a Protocol A osmylated ssDNA is somewhat more complex due to the large number of dT(OsBp), about 25% of total bases for a typical DNA, and the more frequent appearance of dTs that are adjacent or close to each other. The calculated data on Table 1 provide an estimate for the i-t traces from Protocol A readings.

Columns 2 and 5 in Table 1 are the experimentally determined translocation times, t, after analysis of the histograms for the listed oligos. Oligo dA_{20} exhibits t=50μs at 120 mV and t=30μs at 140 mV. These values provide unit dA translocation t=2.5 and 1.5μs at 120 and 140 mV, respectively. Since all the oligos were tested under identical conditions, we use these unit dA values to correct the observed oligo translocation and calculate corrected single osmylated base translocation for the oligos where the middle X= dT(OsBp), dC(OsBp), 5-MedC(OsBp), or dU(OsBp). Notably the correction is small and barely changes the observed translocations. As expected the values are faster at 140 mV compared to 120 mV. Specifically, we found t in the order of osmylated dT<5-MedC<dC<dU with values at 120 mV at 102.5, 262.5, 312.5 and 422.5μs, respectively. These values are quite distinct from each other, and clearly detectable by state-of-the-art patch-clamp amplifiers.

Inspection of the dwell times determined for a regular oligo, a PCR primer, pGEX3', provides a deeper insight (Table 1). The corrected translocation time for pGEX3' with 4 osmylated dTs, divided by four, provides a unit value for dT(OsBp) which is double compared to the value obtained from dA_{10}dT(OsBp)dA_9 (211 vs. 102.5μs). Similarly, from Table 1 the experimentally observed value for pGEX3' translocation (with 4T(OsBp) and 5C(OsBp) corrected for the intact bases and for the presence of the 44T(OsBp)) yields unit t per dC(OsBp) that is about double compared to the one obtained from dA_{10}dC(OsBp)dA_9 (t=665 *vs.* 312.5μs). Similar quantitative comparisons obtained from the data at 140mV indicate that within a sequence with about 50% of pyrimidines (4/9 for T(OsBp), and 9/18 for (T+C)OsBp in pGEX3'), OsBp moieties may extend up to the second next neighbor, as well as overlap with another OsBp moiety in a way that shields two bases per OsBp instead of one (Figure 4). The calculations in Table 1 are suggestive of a substantial effect of proximate pyrimidines on translocation duration. Data are required for pyrimidine (Py) translocation of sequences Py-Xn-Py, where X is intact base and n=0, 1, 2, 3, or 4.

Enhanced confidence of sequencing by using protocol B or an extended protocol A

As tested experimentally, labeling a purine, dA or dG, with OsBp is undetectable up to 1/10,000 (unpublished results). Sequencing with high coverage of both, target strand and complementary, after osmylating at two levels, low and Protocol A, should be sufficient to provide high accuracy in base calling. Protocol B was developed so that unmodified dT is much below 1/10,000 and unmodified dC at 1/10,000. However its use is questionable, due to the expected high overlap of OsBp moieties and the expected very slow translocations. Only practice will show if Protocol B provides real value, or if it should be replaced by a 2- or 3-times longer incubation of Protocol A, to produce approximately 6.5x2=13% or 6.5x3=18.5% of osmylated dC in addition to fully osmylated dT.

Direct dsDNA sequencing after denaturation and osmylation

dsDNA may be denatured and the required amount of label added to reach the desired concentration of OsBp. Please note that it is the final concentration of OsBp that is responsible for percent labeling and not the molar ratio of OsBp to DNA; molar ratio of OsBp to DNA should be over 25-fold [33].

Sequencing in the presence of short homopyrimidines to mark locations on the target strand

Additional help in facilitating the development of a consensus sequence can be accomplished by exploiting short oligos as follows: In the absence of urea short oligos, such as an oligocytidylate or oligothymidylate may be added in the nanopore experiment. These oligos will hybridize with sequences of consecutive dGs or dAs on either one of the DNA strands. The likelihood for multiple consecutive dGs or dAs is small, so it will be a relatively rare event that a region of the DNA strand is hybridized. When the hybridized part of the strand comes to the constriction point of the nanopore, the translocation will stop, and resume only after the oligo dissociates [45,46]. Events due to strand dissociation may be distinguished from events due to the presence of OsBp, and can be used as tertiary markers, to facilitate consensus strand construction.

Identification of 5'Me-dC (OsBp) and 5'OHMe-dC (OsBp)

Distinguishing the different forms of methylated dC by nanopore-based sequencing [47] using osmylated DNA is another application of the proposed technology. Unpublished results show that the selectivity for dT over dC is high, the selectivity for 5'Me-dC lies in between, and the selectivity for 5'OHMe-dC is 2-fold higher compared to dC. OsBp selectivity for the different methylated dCs will determine their relative distribution after Protocol A osmylation. A similar kinetic approach for detection of the different Cs is used successfully by Pac Bio's SMRT technology [48,49]. Discrimination based on residual current as well as dwell time will be additional parameters to facilitate identification. Determination of methylation levels is expected to be concomitant with the basic sequencing described above, and will not require additional experimentation.

Proposed Methodology for Direct RNA Sequencing

Earlier studies showed that RNA successfully traverses the α-HL nanopore [15,16]. Nanopore-based unassisted sequencing of osmylated RNA may follow the above protocols with some notable differences. For DNA we use dsDNA, label both strands, and sequence both strands, because sequencing the complementary strand provides

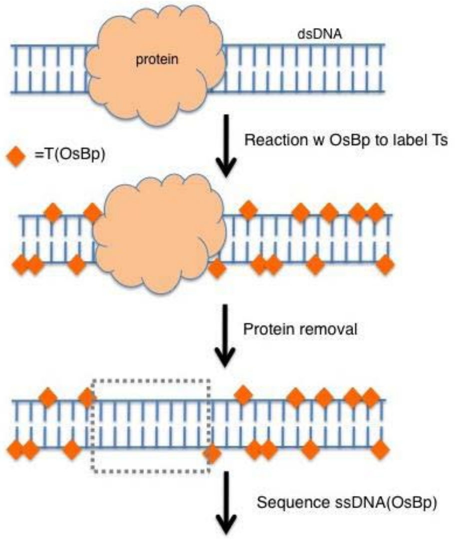

Figure 5: Strategy for mapping of Protein-bound dsDNA. Protocol C, 110 min incubation at room temperature with 5.25mM OsBp (prepared from a 15.75mM OsBp stock solution saturated in 2,2'-bipyridine [56]) results in a dsDNA with overall 50% of osmylated T. Regions of dsDNA bound to a protein remain intact/unlabelled.

identification of A and G on the target strand. In order to conduct ssRNA sequencing directly, identification of A and G will require the use of a purine-specific label. It turns out that the selectivity of OsBp for U over C is a mere 4-fold, so discrimination will depend heavily on dwell times and residual current and much less on distribution. Comparison of dU(OsBp) with dC(OsBp) from Table 1 also shows that their dwell times are disparate. Assuming that this observation transfers to the ribonucleotide derivatives, optimization of conditions may further enhance theses differences. An additional challenge will be to discriminate among the rare bases included in the tRNAs [50]. Monofunctional platinators are well known to coordinate with the N7 position of the purines [39,51,52], and could be used in addition to OsBp.

Proposed Methodology for Mapping Protein Bound DNA Regions

Technologies to probe protein/dsDNA interactions are sparse [53,54]. To label dsDNA has proven challenging. Osmium tetroxide, 1,10-phenanthroline (Os,phen) was reported to label dsDNA, but severely deforms the ds structure [55]. Preliminary results show that Protocol B labels dsDNA, but at the same time denatures it onto two strands of osmylated ssDNA. If the ds structure is distorted, the protein will likely dissociate, and the mapping of its position on the dsDNA will be unsuccessful. In this context we have developed a novel methodology, Protocol C, whereas OsBp labels only 50% of Ts in dsDNA without disrupting the ds structure [38,56]. After denaturing away the protein, dsDNA can be probed to find the region(s) where the protein was bound (Figure 5) by using the nanopore-based sequencing methods put forward in this report.

How Long does it Take to Sequence a 100,000,000 bp Genome?

Nanopore-based unassisted sequencing with osmylated DNA using an ONT-type of platform may fulfill the mandate of faster, cheaper, yielding longer and more accurate reads [4]. The cost is practically the amortization cost of the MinION™ device with consumables such as the flow cell, the α-HL protein, OsBp, an inexpensive chemical, and a resin to remove excess label. No expensive labels, no enzymes, and no primers are used. False positives are better than 1/10,000, i.e., orders of magnitude better than with any other labeling reaction. The length of the read is theoretically expected to be as long as the translocating molecule, as going back against an applied voltage is highly unlikely. Due to the read being as long as the actual target molecule, assembly and scaffolding are avoided. Bioinformatic tools may be available to provide real-time read-out, once the first consensus strand has been identified, and reading proceeds to the second in line consensus strand. Therefore analysis of data is not expected to add any delay to the actual sequencing experiment.

As an example we estimate here how long it will take to sequence a 100,000,000 bp, the size of human chromosome 15, assuming A=C=25% to facilitate calculations. We presume osmylation by the low level osmylation protocol as well as by Protocol A, conduct sequencing in a mixture of four-type of strands using a single α-HL nanopore and allow for up to 128x coverage of each of the four strands. To this end, we will use estimates for translocation time (t) obtained at the conditions listed in Table 1 at 120mV; these values are 102.5, 312.5, and 2.5μs for osmylated dT, dC and intact dA/dG, respectively. It is sufficient to calculate how long it will take for the slowest family, i.e., the one with 90% dT-osmylation and only 6.5% dC-osmylation (Protocol A). Then t is given from eq 1.

$$t = \{0.90 \times 0.25 \times 102.5 + 0.065 \times 0.25 \times 312.5 + (0.50+(0.10+0.935) \times 0.25) \times 2.5\} \times 100{,}000{,}000 = 3004\ s \quad (eq\ 1)$$

In addition to the 3004 s, extra time should be added for (i) waiting until a strand finds the pore and/or intervals between translocations, (ii) hybridized sequences to dissociate and get through the pore, and (iii) a portion of osmylated pyrimidines that are slowed down by OsBp overlap (pGEX3', Table 1). The conjecture is that all these extra processes may add less than 600 s to the 3004 s calculated already from the contribution of the four bases and provide an overall estimate of 3600 s or 1 hour. This t=1h corresponds to one translocation and in order to account for a coverage of 128 for each of the two strands at two levels of osmylation each, a device such as the MinION™ with 512 nanopores, or another suitable device, can be used to provide the required coverage. Notably the wt α-HL that was used for the data on Table 1 is not the same as the proprietary α-HL in MinION™, but translocation features may not be very different between the two nanopores since it is the constriction site that appears to dominate interactions. Hence one hour may be all it takes to sequence the 100,000,000 bp DNA using a single MinION™ with no need for assembly and scaffolding. This tentative speed translates onto a theoretical 512x0.1x24=1229 gigabase per day compared to the 1 gigabase per day claim for the same device, while using current ONT protocols [57]. To sequence the human genome at about 3 billion bp, one MinION™ could complete the job in 30 hours at a cost of replacing the flow cell and amortization of the device. Sample preparation amounts to 60 minutes for the longest osmylation process (Protocol A), and 15 min for purification. A cost estimate can be calculated by (i) amortization of the initial fee of $ 1,000 for using the device to sequence 20 human genomes, (ii) requiring two flow cells at $ 900 each with an estimated life of 24 hours each to complete the 30 hour task of sequencing one whole human genome, and (iii) an additional $ 150 for purification minicolumns and the purchase of OsO4 and 2,2'-bipyridine. This calculation brings the cost for sequencing one human genome to $ 2,000, and the average cost per chromosome to $ 87.

Conclusion

In conclusion, the potential advantages of using nanopores and osmylated nucleic acids compared to the most advanced current sequencing technologies are: (i) labeling chemistry with 1/10,000 false positives, (ii) simplest and cheapest technology regarding consumable reagents, (iii) no known bias for sequence repeats, (iv) identification of rare RNA bases, or dC-methylated bases along with dC, (v) unlimited length of reads with theoretically expected length as long as the target strand, (vi) least ambiguous technology due to the fact that there is no assembly and no scaffolding necessary, and (vii) two to three orders of magnitude the fastest and cheapest technology envisioned.

Acknowledgements

We thank Dr. Eric Ervin (Electronic Biosciences, San Diego, CA, USA) for reviewing this manuscript and providing helpful comments.

Note: Osmylation is used in the literature to describe the reaction of OsO4. For simplicity we use the same term here to describe the reaction of the complex osmium tetroxide 2,2'-bipyridine (OsBp). Hence osmylated oligos/DNA are the products carrying the OsBp moiety.

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. USA. 1977; 74: 5463-5767.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001; 291: 1304-1351.

3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409: 860-921.

4. National Institute of Health. THE $1000 GENOME.

5. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

6. For participants of the Personal Genome Project.

7. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, et al. The challenges of sequencing by synthesis. Nat Biotechnol. 2009; 27: 1013-1023.

8. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009; 323: 133-138.

9. Chen CY. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. Front Microbiol. 2014; 5: 305

10. Maglia G, Heron AJ, Stoddart D, Japrung D, Bayley H. Analysis of single nucleic acid molecules with protein nanopores. Methods Enzymol. 2010; 475: 591-623.

11. Liao YC, Lin SH, Lin HH. Completing bacterial genome assemblies: strategy and performance comparisons. Sci Rep. 2015; 5: 8747.

12. Fierst JL. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Front Genet. 2015; 6: 220.

13. Kisand V, Lettieri T. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. BMC Genomics. 2013; 14: 211.

14. Bayley H, Martin CR. Resistive-pulse sensing – From microbes to molecules. Chem Rev. 2000; 100: 2575-2594.

15. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci. USA 1996; 93: 13770-13773.

16. Akeson M, Branton D, Kasianowicz JJ, Brandin E, Deamer DW. Microsecond Time-Scale Discrimination Among Polycytidylic Acid, Polyadenylic Acid, and Polyuridylic Acid as Homopolymers or as Segments Within Single RNA Molecules. Biophys. J. 1999; 77: 3227-3233.

17. Meller A, Nivon L, Brandin E, Golovchenko J, Branton D. Rapid nanopore discrimination between single polynucleotide molecules. Proc Natl Acad Sci. USA 2000; 97: 1079-1084.

18. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nat. Biotechnol. 2008; 26: 1146-1153.

19. Bayley H. Nanopore sequencing: from imagination to reality. Clin Chem. 2015; 61: 25-31.

20. Wolna AH, Fleming AM, An N, He L, White HS, Burrows CJ. Electrical Current Signatures of DNA Base Modifications in Single Molecules Immobilized in the alpha-Hemolysin Ion Channel. Isr J Chem. 2013; 53: 417-430.

21. Lieberman KR, Cherf GM, Doody MJ, Olasagasti F, Kolodji Y, Akeson M. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. J Am Chem Soc. 2010; 132: 17961-17972.

22. Kanavarioti A. Osmylated DNA, a novel concept for sequencing DNA using nanopores. Nanotechnology. 2015; 26: 134003.

23. Kanavarioti A. Patent pending #62/083,256. Labeled Nucleic Acids: A surrogate for Nanopore-based Nucleic Acid Sequencing.

24. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz M, McCombie WR. Oxford Nanopore Sequencing, Hybrid Error Correction, and de novo Assembly of a Eucaryotic Genome.Genome Res. 2015; 25: 1750-1756.

25. Hackl T, Hedrich R, Schultz J, Förster F. Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics. 2014; 30: 3004-3011.

26. Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, et al. Decoding long nanopore sequencing reads of natural DNA. Nat Biotechnol. 2014; 32: 829-833.

27. Derrington IM, Craig JM, Stava E, Laszlo AH, Ross BC, Brinkerhoff H, et al. Subangstrom single-molecule measurements of motor proteins using a nanopore. Nat Biotechnol. 2015; 33: 1073-1075.

28. Huang S, Romero-Ruiz M, Castell OK, Bayley H, Wallace MI. High-throughput optical sensing of nucleic acids in a nanopore array. Nat Nanotechnol. 2015; 10: 986-991.

29. Ding Y, Kanavarioti A. Single Pyrimidine Discrimination during Voltage-driven Translocation of Osmylated Oligodeoxynucleotidesα via the alpha-Hemolysin Nanopore. Beilstein. J Nanotechnology.

30. Mitchell N, Howorka S. Chemical tags facilitate the sensing of individual DNA strands with nanopores. Angew Chem Int Ed Engl. 2008; 47: 5565-5568.

31. Borsenberger V, Mitchell N, Howorka S. Chemically labeled nucleotides and oligonucleotides encode DNA for sensing with nanopores. J Am Chem Soc. 2009; 131: 7530-7531.

32. Kumar S, Tao C, Chien M, Hellner B, Balijepalli A, Robertson JW, et al. PEG-Labeled Nucleotides and Nanopore Detection for Single Molecule DNA Sequencing by Synthesis. Sci Rep. 2012; 2: 684.

33. Kanavarioti A, Greenman KL, Hamalainen M, Jain A, Johns AM, Melville CR, et al. Capillary electrophoretic separation-based approach to determine the labeling kinetics of oligodeoxynucleotides. Electrophoresis. 2012; 33: 3529-3543.

34. Beer M, Moudrianakis EN. Determination of Base Sequence in Nucleic Acids with the Electron Microscope: Visibility of a Marker. Proc Natl Acad Sci USA. 1962; 48: 409-416.

35. Chang CH, Beer M, Marzilli LG. Osmium-Labeled Polynucleotides. The reaction of osmium Tetroxide with Deoxyribonucleic Acid and Synthetic polynucleotides in the presence of tertiary Nitrogen Donor Ligands. Biochemistry 1977; 16: 33-38.

36. Palecek E. Probing DNA structure with Osmium Tetroxide Complexes in Vitro. Methods Enzymol 1992; 212: 139-155.

37. Haque F, Li J, Wu HC, Liang XJ, Guo P. Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA. Nano Today 2013; 8: 56-74.

38. Henley RY, Vazquez-Pagan AG, Johnson M, Kanavarioti A, Wanunu M. Osmium-based pyrimidine contrast tags for enhanced nanopore-based DNA Base discrimination. PLOS One.

39. Saenger W. Principles of Nucleic Acid Structure, Cantor CR, editor. By Springer-Verlag. 1984; 210.

40. Andrepont C, Marzilli PA, Pakhomova S, Marzilli LG. Guanine nucleobase adducts formed by a monofunctional complex: [Pt(N-(6-methyl-2-picolyl)-N-(2-picolyl)amine)Cl]Cl. J Inorg Biochem. 2015.

41. Human Genome.

42. Strachan T, Read A. Human Molecular Genetics, chapter 9. P 292. ISBN: 9780815341499.

43. Mathé J, Visram H, Viasnoff V, Rabin Y, Meller A. Nanopore unzipping of individual DNA hairpin molecules. Biophys J. 2004; 87: 3205-3212.

44. Japrung D, Henricus M, Li Q, Maglia G, Bayley H. Urea facilitates the translocation of single-stranded DNA and RNA through the alpha-hemolysin nanopore. Biophys J. 2010; 98: 1856-63.

45. An N, Fleming AM, Burrows CJ. Interactions of the human telomere sequence with the nanocavity of the alphahemolysin ion channel reveal structure-dependent electrical signatures for hybrid folds. J Am Chem Soc. 2013; 135: 8562-8570.

46. Ding Y, Fleming AM, He L, Burrows CJ. Unfolding Kinetics of the Human Telomere i-Motif Under a 10 pN Force imposed by the alpha-Hemolysin Nanopore Identify Transient Folded-State Lifetimes at Physiological pH. J Am Chem Soc. 2015; 137: 9053-9060.

47. Schreiber J, Wescoe ZL, Abu-Shumays R, Vivian JT, Baatar B, Karplus K, et al. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. Proc Natl Acad Sci USA. 2013; 110: 18910-18915.

48. Davis BM, Chao MC, Waldor MK. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. Curr Opin Microbiol. 2013; 16: 192-198.

49. Lee WC, Anton BP, Wang S, Baybayan P, Singh S, Ashby M, et al. The complete methylome of Helicobacter pylori UM032. BMC Genomics. 2015; 16: 424.

50. Saenger W. Principles of Nucleic Acid Structure. Cantor CR, editor. Springer-Verlag. 1984; 180.

51. Johnstone TC, Park GY, Lippard SJ. Understanding and improving platinum anticancer drugs--phenanthriplatin. Anticancer Res. 2014; 34: 471-476.

52. Andrepont C, Marzilli PA, Pakhomova S, Marzilli LG. Guanine nucleobase adducts formed by a monofunctional complex: [Pt(N-(6-methyl-2-picolyl)-N-(2-picolyl)amine)Cl]Cl. J Inorg Biochem. 2015.

53. Thompson JF, Oliver JS. Mapping and sequencing DNA using nanopores and nanodetectors. Electrophoresis. 2012; 33: 3429-3436.

54. Furey TS. ChIP-seq and Beyond: new and improved methodologies to detect and characterize protein-DNA interactions Nat Rev Genet. 2012; 13: 840-852.

55. Palecek E, Vlk D, Vojtísková M, Boublíková P. Complex of osmium tetroxide with 1,10-phenanthroline binds covalently to double-stranded DNA. J Biomol. Struct Dyn. 1995; 13: 537-46.

56. Kanavarioti A. Patent pending, # 62/212,072. Mapping protein bound regions on dsDNA by reaction of unblocked Thymidines with Osmium tetroxide 2,2'-bipyridine.

57. Oxford Nanopore Technologies.

Austin J Proteomics Bioinform & Genomics - Volume 2 Issue 1 - 2015
ISSN : 2471-0423 | www.austinpublishinggroup.com
Kanavarioti. © All rights reserved

Citation: Kanavarioti A. A Non-Traditional Approach to Whole Genome Ultra-Fast, Inexpensive Nanopore-Based Nucleic Acid Sequencing. Austin J Proteomics Bioinform & Genomics. 2015; 2(1): 1012.