Mapping Base Modifications in DNA by Transverse-Current Sequencing

Jose R. Alvarez, Dmitry Skachkov, Steven E. Massey, Alan Kalitsov, and Julian P. Velev

Phys. Rev. Applied **9**, 024024 — Published 23 February 2018

DOI: 10.1103/PhysRevApplied.9.024024

Mapping DNA base modifications by transverse current sequencing

Jose R. Alvarez,[1,2] Dmitry Skachkov,[1] Steven E. Massey,[3] Alan Kalitsov,[1] and Julian P. Velev[1,4†]

[1] *Department of Physics, University of Puerto Rico, San Juan, PR 00931-3344, USA*
[2] *Escuela de Ciencias Naturales y Exactas, Pontificia Universidad Católica Madre y Maestra, Campus Santo Tomas de Aquino, Santo Domingo, DN 2748, RD*
[3] *Department of Biology, University of Puerto Rico, San Juan, PR 00631-3360, USA*
[4] *Department of Physics, University of Nebraska, Lincoln, NE 68588-0111, USA*

[†] Corresponding author: julian.velev@upr.edu

Sequencing DNA modifications and lesions, such as methylation of cytosine and oxidation of guanine, is even more important and challenging than sequencing the genome itself. The traditional methods for detecting DNA modifications are either insensitive to these modifications or require additional processing steps to identify a particular type of modification. Transverse current sequencing in nanopores can potentially identify the canonical bases and base modifications in the same run. In this work, we demonstrate that the most common DNA epigenetic modifications and lesions can be detected with any predefined accuracy based on their tunneling current signature. Our results are based on simulations of the nanopore tunneling current through DNA molecules, calculated using non-equilibrium electron transport methodology within an effective multi-orbital model derived from first-principles calculations, followed by base-calling algorithm accounting for neighbor current-current correlations. This methodology can be integrated with existing experimental techniques to improve base calling fidelity.

I. INTRODUCTION

The structure, development, and function of living organisms is encoded on several informational levels. All cells in an organism share the same genome, which is inherited via the germline and remains unchanged during the lifetime of the organism. At the same time, gene expression can be influenced by additional modifications to the genome, such as cytosine methylation, collectively referred to as the epigenome [1]. The epigenetic modifications are chemical changes of the canonical DNA bases, which occurs enzymatically after DNA replication. Thus, epigenetic modifications can differ in different cells and can change over time. A number of nucleotide modifications with biological significance have been identified in eukaryotic and prokaryotic cells. In addition to purposeful modifications, DNA base modifications can results from damage, such as oxidation [2].

In higher eukaryotes, the most common modification is the methylation of cytosine, 5-methylcytosine (^{m}C) [2]. Subsequent oxidation of ^{m}C produces 5-hydroxymethylcytosine (^{h}C). DNA methylation inhibits gene expression and is important for a variety of processes such as cell differentiation, parent-offspring imprinting, X-chromosome inactivation, and transposon repression [3]. Conversely, methylation abnormalities are associated with cancer and other diseases [4]. Furthermore, adenine methylation, N^6-methyladenine (^{m}A), is the most common DNA modification in prokaryotes and it is also thought to be biologically significant in eukaryotic DNA [5] and mRNA [6]. The most common type of DNA damage is guanine and adenine oxidation, 8-oxoguanine (^{o}G) and 8-oxoadenine (^{o}A), respectively [7,8]. Base oxidation is related to DNA deterioration with age.

Base modification is a dynamic process, which depends of the type of cell and its stage of development. Thus, the epigenome rather than the genome, is a better indicator of the cell's health [9]. Using epigenetic information for clinical diagnosis, however, involves the enormous task of genome-wide mapping of modifications on different types of cells over time. Furthermore, first and second generation sequencing methods are insensitive to base modifications. The most developed specialized technique is the bisulfite treatment for mapping cytosine methylation of genomic DNA [10,11]. Despite reports of successful mapping genome-wide methylation [12,13,14], the method suffers from limitations such as quality degradation of the primary DNA, misidentification between ^{m}C and ^{h}C, high cost and long processing times due to the number of steps involved [15,16,17].

A third generation of sequencing techniques is under development, featuring single-molecule methods [18]. These methods do not require complex chemical treatment or PCR amplification, thus, they are capable of very long read length with minimal setup sample

preparation. Single-molecule real-time sequencing (SMRT) [19] and single-molecule nanopore (SMNP) [20,21] sequencing have emerged as the most promising contenders. SMRT sequencing uses fluorescent labels to monitor polymerase kinetics during elongation of a DNA daughter strand [19]. It was reported that the fluorescent pulse width and inter-pulse duration are correlated with the type nucleotide, which allows for cytosine to be distinguished from ^{m}C and ^{h}C, and DNA methylation to be mapped [22]. This approach is utilized by the PacBio sequencing platform [23,24]. The technology is being extensively used to study DNA modifications in bacteria [25,26,27].

SMNP sequencing does not require fluorescent labeling and optical detection. In this approach a single-strand DNA or RNA is driven through a biological [28] or solid-state [29] nanopore. In the two distinct variants of the method, the modulation of the longitudinal ionic current through the nanopore [30], or the transverse tunneling current [31], is monitored as the molecule translocates through the nanopore. The nucleobases have distinct electric signatures due to their atomic or electronic structure, respectively. The ionic current methodology relies on the different shape of the nucleotides to produce distinct blockage of the ionic current. It has been used successfully in sequencing of homopolymers or oligomers of DNA [30,32], cytosine methylation in DNA [33,34,35,36,37,38] and RNA [39,40,41] and guanine oxidation in DNA [42,8]. This approach has been commercialized into the portable MinION sequencing platform [43,44]. Nevertheless, the similarity in geometry of the purine (A, G) and pyrimidines (T, C) bases is a substantial cause of errors. Introduction of base modifications aggravates the problem. The alternative is to measure the electronic current via electrodes embedded into the nanopore. This variant is in principle superior, since it is sensitive to both the electronic and atomic structure of the bases, however, the elaboration of the experimental setup is more challenging. The method has been demonstrated on homopolymers and short strands of DNA [45,46,47], methylated DNA [48], RNA [49], and even proteins [50].

The Achilles heel of all these methods are the high error rates. It has been recognized that all single-molecule methods exhibit inherent noise arising from correlations between the neighbors [51,52,53], albeit due to different physical mechanisms. In SMRT, the noise comes from the influence of the neighbors on the polymerase reaction speed [22]. In the ionic-current SMNP, the neighbors affect the vibrational modes of the nucleotide affecting its blocking properties, moreover, typically several nucleotides are accommodated in the nanopore [54,55,56]. In the transverse-current SMNP the noise arises from modifications of the electronic structure of the nucleotide due to hybridization with its neighbors [51,52,53].

In addition to the intrinsic noise there is extrinsic noise arising from the temperature induced molecule motion and interaction with the environment. This type of noise is relatively well studied in the context of transverse current SMNP sequencing [51,53,57,58]. It leads to a Gaussian spread of the current readings around the zero temperature value. Since this noise is not correlated, it could be averaged out and the spread can be controlled by improvements in the experimental setup [59].

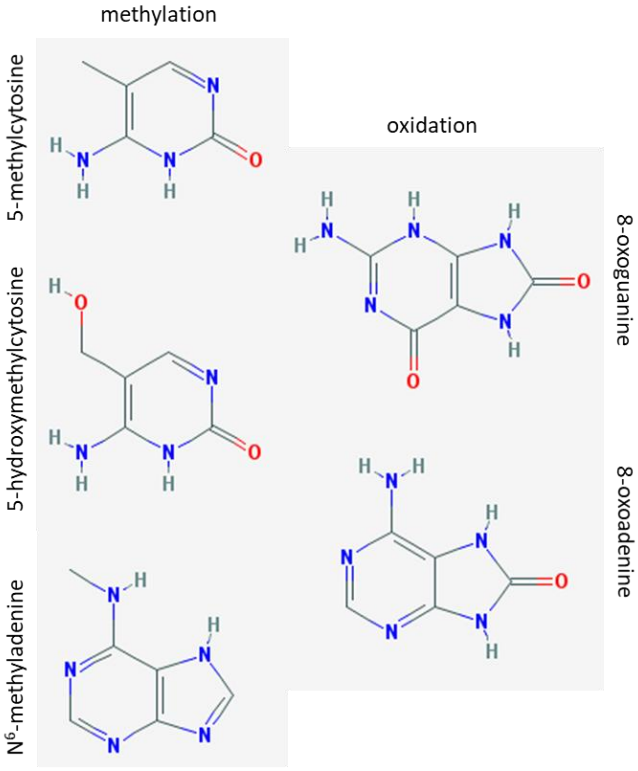

Fig. 1: Common DNA base epigeniomic modifications and lesions. Methylation of cytosine and adenine (left) and oxidation of guanine and adenine (right).

In previous works, we showed that correcting for the intrinsic noise permits base calling in arbitrary long DNA or RNA sequences with a precision comparable to commonly used next generation sequencing methods [51,52]. Epigenetic modifications, however, complicate base calling dramatically by introducing a number of modified bases with very similar electronic and atomic

structure. There is some evidence that single-nucleotide epigenetic medications can be distinguished via transverse current [60,61], however, it is not clear if they would be statistically distinguishable from the canonical bases when embedded in long chains. In this work, we investigate the possibility of mapping a number of common DNA modifications, such as methylation of cytosine (^{m}C and ^{h}C) and adenine (^{m}A) and oxidation of guanine (^{o}G) and adenine (^{o}A), via the transverse-current SMNP technique. Our conclusions are based on large-scale numerical simulations of the current readings through DNA molecules in nanopores. The error correction method we use is very general and could be implemented with the experimental SMNP techniques. Moreover, it could in principle be applied to the other single-molecule sequencing techniques which also suffer from correlated errors.

II. METHODOLOGY

In the transverse-current SMNP sequencing setup a single-strand DNA molecule translocates through the nanopore between two tapered electrodes, which are assumed to make contact with one nucleotide at a time. Since DNA are large molecules, full first-principles calculations of electronic structure of a polynucleotide molecule is still a very demanding task, especially in the case when large statistics are required. Therefore, to represent the single-strand DNA molecule we use an effective multi-orbital tight-binding (TB) Hamiltonian derived from first-principles calculations [52]. This approach represents a combination of a fragment molecular orbital (FMO) scheme [62,63,64] with a projection of the fragment Hamiltonian on a small set of orbitals active in the transport [65,66,67]. The first-principles calculations are performed within the density functional theory (DFT) as implemented in the Amsterdam Density Functional (ADF) package [68,69]. We use the Perdew-Burke-Ernzerhof (PBE) exchange and correlation functional with triple-zeta polarized (TZP) basis set.

A. Hamiltonian of single-strand DNA chain

Within the FMO framework [62,63,64], the single-strand DNA molecule is represented as a set of fragments $F = F_1\, F_2 \ldots F_{i-1}F_iF_{i+1} \ldots F_{N-1}F_N$ where each fragment F_i is a nucleotide. The interaction between fragments decays very rapidly with the distance. If the interaction is truncated to the first nearest neighbor, FMO allows us to construct the Hamiltonian

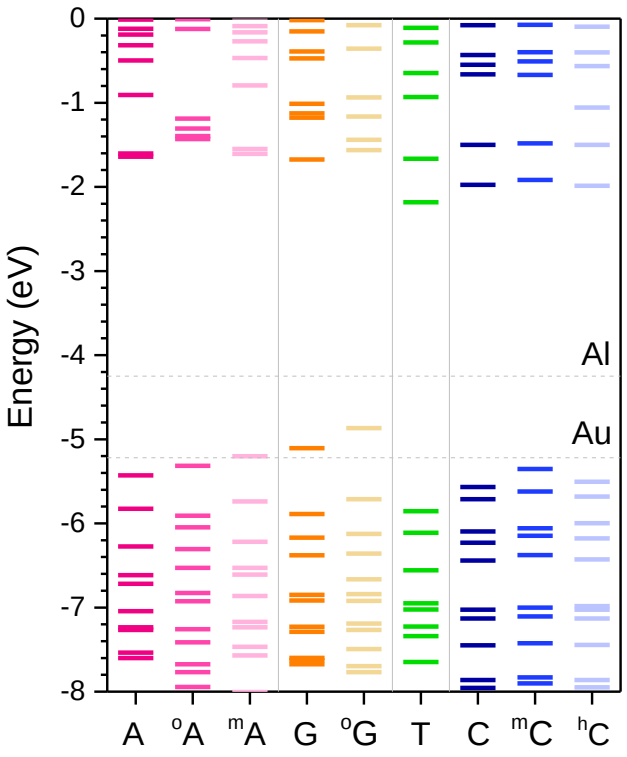

Fig. 2: Molecular energy levels calculated from first principles of the four canonical DNA bases (A, G, C, T) compared to the modified bases methylcytosize (^{m}C), hydroxymethylcytosine (^{h}C), and methyladenine (^{m}A) and oxoguanine (^{o}G) and oxoadenine (^{o}A). The Fermi levels of the metal electrodes, Al and Au, are indicated by horizontal dashed lines.

of an arbitrary long polynucleotide chain from the Hamiltonians of single fragments and pairs of fragments as $F = F_1F_2 + \cdots + F_{i-1}F_i + \cdots + F_{N-1}F_N - F_2 - \cdots - F_i - \cdots - F_{N-1}$.

In practice, we first solve in ADF for the electronic structure of each fragment, F_i, to obtain the fragment Hamiltonian $\boldsymbol{h}_{F_iF_i}$ and molecular orbitals (MOs) $\boldsymbol{\phi}_{F_i}$. The single fragment energy levels are plotted in Fig. 2 for all canonical and modified bases we are considering. Then we calculate in ADF the electronic structure for all pairs of fragments, F_iF_j. These were generated by taking the standard DNA backbone of length two and attaching to it all the possible combinations of two bases. The MOs of the individual fragments are used as a new basis set, $\boldsymbol{\phi}_{F_iF_j} = \{\boldsymbol{\phi}_{F_i}, \boldsymbol{\phi}_{F_j}\}$. The wave function of the pair is a linear combination of this fragment orbital basis, $\boldsymbol{\psi} = \boldsymbol{C}\boldsymbol{\phi}_{F_iF_j}$. The matrix of expansion coefficients, $\boldsymbol{C}$, is obtained by solving the generalized eigenvalue equation $\boldsymbol{HC} = E\boldsymbol{SC}$. In addition to the wave function and the energy levels of the fragmetn pair, ADF yeilds the

Hamiltonian matrix in the fragment orbial basis, $\boldsymbol{H}_{F_iF_j} = \boldsymbol{\phi}_{F_i}\hat{H}\boldsymbol{\phi}_{F_j}^{\dagger}$ and the overlap matrix between the orbitals, $\boldsymbol{S}_{F_iF_j} = \boldsymbol{\phi}_{F_i}\boldsymbol{\phi}_{F_j}^{\dagger}$.

The dimension of the matrices $\boldsymbol{H}$ and $\boldsymbol{S}$ is determined by the number of MOs in the fragments, which is fairly large. We can reduce the size of these matrices by projecting the complete space of MOs on the MOs active in the transport, residing within the bias window between the Fermi energies of the electrodes. This is achieved by using established projector operator techniques [65,66,67]. The space of the fragment MOs is split into two complementary subspaces $\{\boldsymbol{\phi}_i\} \in \mathcal{P}$ and $\{\boldsymbol{\phi}_j\} \in \mathcal{Q}$, where $\mathcal{P}$ consists of the MOs in the active region and $\mathcal{Q}$ is the rest of the system. The projection operator on state $\boldsymbol{\phi}_i$ is $\boldsymbol{P}_i = \boldsymbol{\phi}_i \sum_j [\boldsymbol{S}^{-1}]_{ij} \boldsymbol{\phi}_j^{\dagger}$ and $\boldsymbol{P} = \sum_{p\in\mathcal{P}} \boldsymbol{P}_p$ is the projection operator on subspace $\mathcal{P}$. Thus, we can project the Schrödinger equation for the fragment, $\boldsymbol{H}\boldsymbol{\psi}_F = E\boldsymbol{\psi}_F$, on the subspace $\mathcal{P}$ using the Hermitian projection $\boldsymbol{P}^{\dagger}\boldsymbol{H}\boldsymbol{P}$. This results in an effective Schrödinger equation, $\boldsymbol{H}_{eff}\boldsymbol{P}\boldsymbol{\psi}_F = E\,\boldsymbol{P}\boldsymbol{\psi}_F$, where $\boldsymbol{H}_{eff}$ is the energy dependent effective Hamiltonian projected on the $\mathcal{P}$ region

$$\boldsymbol{H}_{eff}(E) = \boldsymbol{H}_{PP} + \boldsymbol{V}_{PP}(E) \tag{1}$$

where $\boldsymbol{V}_{PP}(E) = \left(E\boldsymbol{S}_{PQ} - \boldsymbol{H}_{PQ}\right)\boldsymbol{G}_{QQ}\left(E\boldsymbol{S}_{QP} - \boldsymbol{H}_{QP}\right)$ and $\boldsymbol{H}_{XY} = \boldsymbol{X}^{\dagger}\boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{S}_{XY} = \boldsymbol{X}^{\dagger}\boldsymbol{Y}$ with $\boldsymbol{X},\boldsymbol{Y} \in \{\boldsymbol{P},\boldsymbol{Q}\}$. The first term is simply the Hamiltonian of the subspace $\mathcal{P}$ and the second term is the perturbation acounting for the interaction between the $\mathcal{P}$ and $\mathcal{Q}$ regions. In our case, the effective Hamiltonian is obtained after projection on the five MOs closest to the bias window – the lowest unoccupied MO (LUMO) and the highest four occupied MOs (HOMO, HOMO$_{-1}$,…, HOMO$_{-3}$).

Finally, using the effective Hamiltonian in conjunction with the FMO method, we can construct a manageable size Hamiltonian matrix for very long DNA chains as

$$\begin{aligned} \boldsymbol{H}_{ii} &= \boldsymbol{h}_{F_iF_i} + \left[\boldsymbol{H}_{F_{i-1}F_i}\right]_{F_iF_i} + \left[\boldsymbol{H}_{F_iF_{i+1}}\right]_{F_iF_i} \\ \boldsymbol{H}_{ij} &= \boldsymbol{H}_{F_iF_j}\delta_{j,i\pm1} \end{aligned} \tag{2}$$

where the second and third terms in the first expression have the meaning of a renormalization factor on the onsite energy $\boldsymbol{h}_{F_iF_i}$ Hamiltonian at site i coming from the neighbors fragments on sites $i \pm 1$.

B. Tunneling current through a nucleotide

Having constructed the Hamiltonian of long single-strand DNA chains, we use the non-equilibrium Green's function (NEGF) method [70] to obtain the transverse tunneling current through each nucleotide n by connecting consecutive nucleotides of the molecule the electrodes

$$I^n = \frac{2e}{h}\int dE\left(f_L(E) - f_R(E)\right)T^n(E) \tag{3}$$

where the semi-infinite left/right electrodes are in equilibrium with chemical potentials μ_α and f_α (α = L/R) are the Fermi-Dirac distribution functions. $T^n(E)$ is the transmission probability through the junction. Within the NEGF method it is given as $T^n(E) = \mathrm{Tr}\left[\boldsymbol{\Gamma}_L\boldsymbol{G}\boldsymbol{\Gamma}_R\boldsymbol{G}^{\dagger}\right]$ where $\boldsymbol{G} = (E\boldsymbol{S} - \boldsymbol{H} - \boldsymbol{\Sigma}_L - \boldsymbol{\Sigma}_R)^{-1}$ is the retarded Green function of the molecule connected to the electrodes and $\boldsymbol{\Sigma}_\alpha$ are the self-energies due to the connection of the electrodes, which are added to the n-th nucleotide. Then $\boldsymbol{\Gamma}_\alpha = -2\mathrm{Im}\boldsymbol{\Sigma}_\alpha$ are the electron escape rates to the electrodes.

The escape rates are proportional to the surface density of states (DOS) in the electrodes. In s-type metals, such as Au, DOS is approximately constant around the Fermi energy, which allows us to treat the electrodes on the level of the wide-band approximation (WBA). WBA consists in taking $\boldsymbol{\Gamma}_\alpha$ to be an energy independent function. Thus, by neglecting the level-shift (the real part of the self-energy), within WBA the self-energy can be given by $\boldsymbol{\Sigma}_\alpha = -(i/2)\boldsymbol{\Gamma}_\alpha$, with $\boldsymbol{\Gamma}_\alpha = \Gamma_\alpha\boldsymbol{S}_\alpha^n$ and $\boldsymbol{S}_\alpha^n$ is an overlap matrix between the nucleotide n and electrode α. Since in the nanopore setup the nucleotides are not chemically bonded to the electrodes and the contact is weak, we can set $\boldsymbol{S}_\alpha^n = \boldsymbol{I}$ for all nucleotides. In our setup we assume that the molecule is in the center of the nanopore, thus the overlap of both ends of the nucleotide with the electrode is the same (taken to be $\Gamma_\alpha = 10^{-3}$ eV).

C. Base calling from current signature

The outcome of the sequencing is a series of the tunneling current readings through successive nucleotides. A crucial step of the process is the identification of the base from its tunneling current signature. The usual approach is to use a maximum likelihood base-calling algorithm [53]. Namely, given the current reading I, we pick the base X which has the

maximum probability to produce it. This probability is given by the Bayes' formula, $P(X|I) = P(I|X)P(X)/[\sum_X P(I|X)P(X)]$, which is essentially the overall probability that the base is of type X and the current through X is equal to I. However, since in natural DNA the bases appear with approximately the same frequency, the algorithm simplifies to $\max_X P_X(I)$ where $P_X(I) = P(I|X)/\sum_X P(I|X)$ is the probability distribution function (PDF) of the tunneling current through the different bases. This procedure we call simple or zeroth order in the current correlations base-calling algorithm and we have shown that it performs poorly in the case of intrinsic noise because the PDFs in that case are not simple Gaussian distributions [51,52].

An improved version of the algorithm uses a Bayesian improvement strategy, where on each subsequent step we include the information contained in the higher order joint current PDFs to improve on the sequence obtained at the previous step. The maximum likelihood base-calling algorithm provides the initial sequence [51,52]. For the purpose of this base-calling procedure, during the calibration step, the joint current PDFs of up to order n, $P_{X_1, X_2,\dots,X_n}(I_1, I_2, \dots, I_n)$, are constructed to measure a sequence of currents $I_1, I_2, \dots, I_n$ through a sequence of bases $X_1, X_2, \dots, X_n$. These PDFs contain information not only about the transmission through individual bases, but also about the correlations between the currents through neighboring bases. The PDF used in the maximum likelihood approach is simply the first order PDF, $P_{X_1}(I_1)$, which contains no information about correlations. In principle, we can compute joint PDFs of any order given large enough statistics. For a molecule of length N, if we had all the PDFs of order N, picking the correct sequence would amount to simply finding $\max_{X_1, X_2,\dots,X_N} P_{X_1, X_2,\dots,X_N}(I_1, I_2, \dots, I_N)$. However, the number of such PDFs and the size of the sample necessary to construct them makes this prohibitively expensive. Instead, we calculate the lowest order PDFs and use them in an iterative improvement procedure. Since the strength of the correlations decreases with the distance between the nucleotides, we can reconstruct the full N-base PDF by using a few of the lowest order PDFs.

As a first step of the procedure, we construct the initial sequence from the first order (maximum likelihood) PDF as $\tilde{X}_k^{(1)} = \max_{X_1} P_{X_1}(I_k)$ for $k = 1 \dots N$. On each subsequent step n, we have the sequence from the previous step $\tilde{\boldsymbol{X}}^{(n-1)}$ inferred using PDFs of order up to $(n-1)$. Then we introduce the new information contained in the PDF of order n. First, we check the sequence for consistency by comparing base $\tilde{X}_k^{(n-1)}$ at position k with the base at the same position calculated using the higher order PDF, $\tilde{X}_k^{(n)} = \max_{X_1 \dots X_n} P_{X_1 \dots X_n}$ subject of the constraint that all the bases except the one in position k are taken from the $\tilde{\boldsymbol{X}}^{(n-1)}$ sequence. This check implies n comparisons, because there are n ways to predict base X_k at position k using the $(n-1)$ neighbors. If $\tilde{X}_k^{(n-1)} = \tilde{X}_k^{(n)}$, then the nucleotide $\tilde{X}_k^{(n)}$ is assumed certain and it is incorporated in the new sequence.

The nucleotides that fail the consistency test are then regenerated with the help of $P_{X_1 \dots X_n}$. For each position k the nucleotide with maximum probability to yield the sequence of n current is chosen as $\tilde{X}_k^{(n)} = \max_{X_1 \dots X_n} P_{X_1 \dots X_n}$ subject to constraint that some of the neighbors of X_k are fixed (certain) i.e. the maximization is performed over the subspace of uncertain bases. The sequence thus generated is n-th order consistent. This process continues until a certain n_{max} order of the PDF. Since the influence of the neighbors further away is bound to be smaller, it is feasible to construct enough high order PDFs to reduce the error rates below a desired threshold.

Alternative approaches have been proposed to deal with the problem of correlated noise. For example, in the context of ionic current nanopore sequencing, correlations arise from the fact that multiple nucleotides block the current in the same time [71]. To disambiguate the current reading for a particular multiplet, a Viterbi algorithm is used which consists in calculating the maximum likelihood that a current state is a transition from previous states. In our case, we assume that the electrodes make contact with only one nucleotide at a time and this type of correlation is not present. Instead, currents are correlated due to chemical bonding of the nucleotide with its neighbors. Nevertheless, the Viterbi algorithm can be adapted to our problem as well.

III. RESULTS AND DISCUSSION

We use this technique to simulate the outcome of the sequencing of long strands of DNA containing epigenetic modifications. In particular, we aim to evaluate the possibility of mapping the canonical genome and epigenome on the same run. The process

comprises of two stages: calibration and measurement. In the *calibration* stage, a very large number of current measurements through known single-strand DNA sequences is collected for the particular nanopore setup. From these measurements the joint PDFs of the current through single nucleotide, pairs of nucleotides, and triples of nucleotides are constructed. In the *measurement* stage, the PDFs constructed in the calibration stage are used to call the bases of an unknown sequence based on current readings through the base and its neighbors.

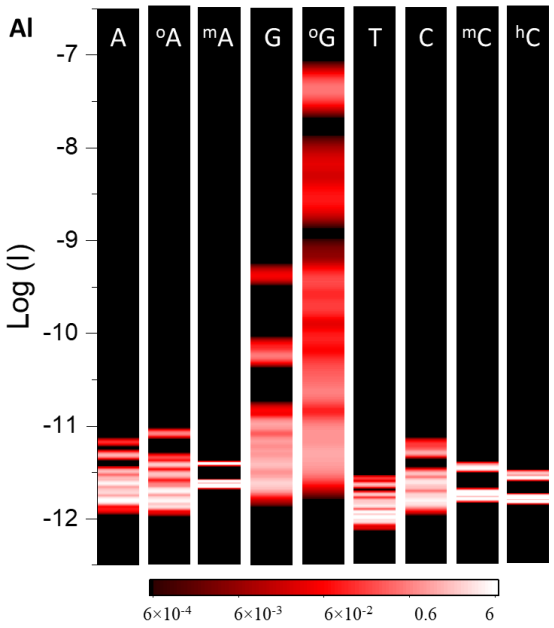

Fig. 3: Current probability distibution functions in the tunneling regime. The four canonical DNA bases (A, G, T, C) and modifications (^{m}C, ^{h}C, ^{m}A, ^{o}G, and ^{o}A) are shown. The current is calculated with Al electrodes at 0.1V bias voltage and low temperature.

A. Calibration

We perform the calibration by simulating the current readings through long single-strand DNA sequences using the following procedure: First, we perform first-principles calculation of the canonical nucleotides A, G, C, T and the modified variants ^{m}C, ^{h}C, ^{m}A, ^{o}G, and ^{o}A, as well as all the possible pairings of the nucleotides. Second, applying the projection on the active energy window for transport, we obtain the effective Hamiltonian representation of the single nucleotides and nucleotide pairs. Third, using the FMO prescription, we construct the effective Hamiltonian of long single-

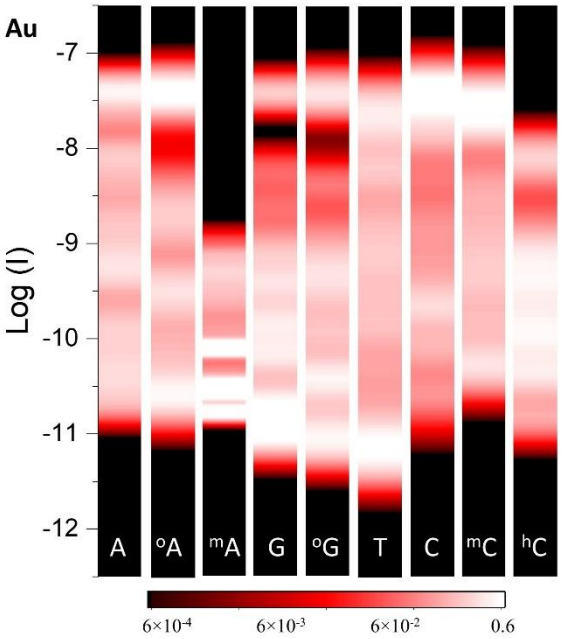

Fig. 4: Current probability distibution functions in the resoannt regime. The four canonical DNA bases (A, G, T, C) and modifications (^{m}C, ^{h}C, ^{m}A, ^{o}G, and ^{o}A) are shown. The current is calculated with Au electrodes at 0.1V bias voltage and low temperature.

strand DNA chains containing modified bases. Fourth, we calculate the currents through each nucleotide as it passes between the electrodes. Finally, based on these data we construct the current distributions through individual nucleotides and the joint current distributions through pairs and triples of nucleotides. In a laboratory setting the calibration step could be performed by collecting the current readings from a sequencer for a large set of known DNA sequences. Generally, the precision of the base calling procedure will increase by employing a larger training set during the calibration step, which would produce higher resolution PDFs.

The first-principles-calculated MOs of the four DNA nucleotides, A, G, C, T, and the modified variants ^{m}C, ^{h}C, ^{m}A, ^{o}G, and ^{o}A are displayed in Fig. 2. They are compared to the work functions of two electrodes calculated at the same level of the theory [72]. Overall, the calculated HOMO and LUMO levels of the DNA nucleotides and methylated cytosine are consistent with previous DFT calculations [73,74,75]. Also, the alignment of the DNA levels with the metal work functions agrees well with photoemission data [76]. We notice that although the modified nucleotides have similar atomic structure to the canonical nucleotides, their electronic structure differ substantially due to different electron donor properties of the -H, -CH_3, -

CH_2OH, and -O functional groups. Both the methyl and the oxygen groups act as acceptors which raises in energy the HOMO level in comparison to the canonical nucleotides.

From Fig. 2 it becomes clear that there are two distinct transport regimes: tunneling regime when the electrode Fermi level and the entire bias window is in the gap of the nucleotides (Al) and resonant regime when the Fermi level of the electrode is close to the HOMO levels of the nucleotides and some of the MOs of the nucleotides fall within the bias window (Au). In the tunneling case, the main contribution of transport come from the frontier MOs of the nucleotides, i.e. HUMO and LUMO. While, in the resonant case, several MOs contribute to electric current which mandates the multi-orbital representation for the bases.

Next, we construct non-parametric joint PDFs for the current distributions through single, $P_{X_1}(I_1)$, pairs $P_{X_1X_2}(I_1, I_2)$, and triples $P_{X_1X_2X_3}(I_1, I_2, I_3)$, of nucleotides, where P_1 is the probability of measurement the current I_1 thought the nucleotide X_1. P_2 the join probability of measurement the pairs current I_1, I_2, through the pairs nucleotides X_1X_2, etc. where $X_i \in (A, G, C, T, {}^mC, {}^hC, {}^mA, {}^oG, {}^oA)$. This is done by generating a number of long single-strand DNA chains and calculating the currents through each nucleotide. In this case, 2000 chains of 200 bases each were used. The currents through each base, pair, and triple of bases are collected together to construct the PDFs. The single-nucleotide PDFs for the Al and Au electrodes at zero temperature and finite bias are shown in the Figs. 3 and 4 respectively. The most prominent feature of the PDFs is the orders of magnitude spread in the tunneling current in both cases and the corresponding large overlap among them, despite the lack of any environmental noise. The shape of the PDFs corresponds to a multimodal Gaussian mixture, with each of the modes corresponding to a particular configuration of the neighboring nucleotides. As we have discussed previously, this intrinsic noise arises from the influence on the electronic structure of the nucleotide of the neighboring nucleotides [51,52]. This noise is very large and clearly distinct from the environmental noise produced by environment- and temperature-driven shifts of the position of the molecule with respect to the electrodes. Nevertheless, as seen in Figs. 3,4, the PDFs of the canonical bases and their modifications are statistically different, which is a result of the influence of the end group on the electronic structure of the base [48].

Another observation is that the spread and overlap are much larger in the resonant regime. In the tunneling regime, the electron transmission probability decreases exponentially with the energy difference between the MO level and the Fermi level, which causes the transmission to be dominated by the HOMO. Correspondingly, the magnitudes of the currents in Fig. 3 follow the order of the HOMO levels in Fig. 2, with the smallest expected value of the current for T and the largest for G. For the same reason, the contribution of the satellite levels translates into smaller satellite contributions to the current and smaller spreads.

Conversely, in the resonant regime at finite bias several MOs, including satellite levels induced by neighboring nucleotides, contribute to the transmission with unitary probability. This leads to the almost complete smearing out of the current because of the randomness of the chain, as illustrated by the PDFs in Fig. 4. In this case, the multi-orbital model is essential to obtain correct results, because lower-lying orbitals also give large contributions to the current.

B. Base calling

After the PDFs have been constructed, we can use them to call the bases of unknown DNA sequences based on the current readings through each nucleotide and its neighbors. The algorithm can be tested by running it on a known DNA sequence, calling the bases based on their current signature, and comparing the called sequence with the original. As a measure of the *fidelity* of the method we use the ratio of the number of correctly identified bases to the total number of bases. We can also introduce *partial fidelities* for each nucleotide as the ratio of the correctly called bases of type X to the total number of such bases in the sequence. We test the fidelity of this base calling procedure on a set of 20 random 200-base DNA sequences which include canonical bases as well as base modifications.

The most basic base-calling algorithm uses the information contained in the single nucleotide (maximum likelihood) PDF. In essence, a base at position k is assigned based on the maximum probability to measure this current I_k through any of the nucleotides $\tilde{X}_k = \max_{X_1} P_{X_1}(I_k)$. The fidelities in the two transport regimes are shown in the first column of Table 1. The error rates are clearly unacceptably high, which is a consequence of the large overlap between the PDFs. However, the multimodal structure of the noise in the current PDFs is evidence that the currents through

neighboring bases are correlated. As we have discussed before, the information in the joint PDFs can be used in a Bayesian improvement scheme to increase the base calling accuracy [51,52]. In essence, using the current through the neighbors to explain the current through a particular base removes the contributions of satellite peaks from the PDF and reduces the effective PDF overlap.

	Al		Au	
	P_{X_1}	$P_{X_1X_2X_3}$	P_{X_1}	$P_{X_1X_2X_3}$
A	75 (6)	98 (2)	39 (7)	64 (7)
^{o}A	46 (6)	99 (1)	59 (7)	56 (11)
^{m}A	100 (0)	99 (1)	73 (13)	85 (13)
G	61 (7)	99 (1)	53 (7)	64 (7)
^{o}G	76 (6)	96 (3)	24 (11)	35 (11)
T	76 (5)	99 (1)	22 (5)	57 (7)
C	37 (7)	98 (2)	39 (6)	60 (6)
^{m}C	70 (12)	99 (2)	23 (10)	69 (16)
^{h}C	97 (7)	99 (1)	78 (20)	67 (22)
DNA	62(4)	99 (1)	38 (3)	59 (4)

Table 1: Calculated overall fidelity and partial fidelity for each type of base for 20 × 200-base long randomly generated DNA sequences containing epigenetic modifications and lesions. Fidelity is given in percent and the standard deviation of the fidelity between the samples is given in the parentheses. Results with and without taking into account current-current correlations are compared. Calculations are performed with Al and Au electrodes at 0.1V applied bias and zero temperature.

The results of the test are listed in Table 1. The main observation, arising from the partial fidelity numbers, is that the DNA modifications can clearly be distinguished from the canonical bases in the same run without any special processing. As before, we observe that the fidelities in the tunneling regime are consistently higher than those in the resonant regime. In this regime, the transport is dominated by the HOMO level which derives from the base itself and the contributions of the satellite levels from the neighboring bases are exponentially smaller. Curiously, both methylation and oxidation improve the base-calling fidelity because the higher HOMO levels give rise to narrower PDFs. Despite the small PDF spreads though, the simple base-calling algorithm still misidentifies the nucleotides in the regions where the PDFs overlap. Accounting for current correlations, in this case, essentially fully disambiguates current distributions and the fidelity is raised to 100%.

In contrast, in the resonant regime the contributions from satellite peaks give rise to five orders of magnitude spread in the PDFs and essentially complete overlap between among the PDFs. Accounting for current correlations improves the fidelity, however, the error rates remain unacceptably high. Thus, low error rates are intrinsically linked to the nanopore setup working in the tunneling regime. While this condition can be achieved by an appropriate choice of the electrode material, there are limited options and this approach is not tunable. A much more flexible option would be to elaborate a gate electrode on the nanopore, such as that applying a gate voltage would shift the DNA levels away of the electrode Fermi level. In this case, the calibration would be dependent not only on the nanopore geometry and electrode material, but also on the gate voltage.

IV. CONCLUSIONS

In summary, we demonstrate that nanopore transverse current sequencing can in principle identify not only the canonical DNA bases but also all the DNA epigenetic modifications and lesions in the same run with the same precision and without any special preparation. The changes to the electronic structure of the bases due to chemical modifications are comparable to the differences between the canonical bases themselves, thus, each modified base can be treated as a distinct type of nucleotide in the calibration step. Once the DNA base modifications are included in the calibration, they can be called in equal footing with the canonical bases and the error rates for all bases can be reduced under a desired precision by including higher order current-current correlations in the procedure.

Acknowledgments

The work at the University of Puerto Rico was supported by NSF (Grants Nos. EPS-1002410, EPS-1010094, and DMR-1105474). The calculations were performed at the Holland Computing Center of the University of Nebraska.

[1] V. Marx, Epigenetics: *Reading the second genomic code*, Nature **491**, 143 (2012)

[2] J. Korlach and S. W. Turner, *Going beyond five bases in DNA sequencing*, Curr. Opin. Struct. Biol. **22**, 251 (2012)

[3] T. Kouzarides, *Chromatin modifications and their function*, Cell **128**, 693 (2007)
[4] K. D. Robertson, *DNA methylation and human disease*, Nat. Rev. Genet. **6**, 597 (2005)
[5] G.-Z. Luo, M. Andres Blanco, E. Lieberman Greer, C. He, and Y. Shi, *DNA N6-methyladenine: a new epigenetic mark in eukaryotes?* Nat. Rev. Mol. Cell Biol. **16**, 705 (2015)
[6] K. D. Meyer and S. R. Jaffrey, *The dynamic epitranscriptome: N6-methyladenosine and gene expression control*. Nat. Rev. Mol. Cell Biol. **15**, 313 (2014)
[7] S. S. David, V. L. O'Shea, and S. Kundu, *Base-excision repair of oxidative DNA damage*. Nature **447**, 941 (2007)
[8] N. An, A. M. Fleming, H. S. White, and C. J. Burrows, *Nanopore detection of 8-oxoguanine in the human telomere repeat sequence*. ACS Nano **9**, 4296 (2015)
[9] H. Heyn and M. Esteller, *DNA methylation profiling in the clinic: applications and challenges*, Nat. Rev. Genet. **13**, 679 (2012)
[10] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul, *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands*, Proc. Natl. Acad. Sci. **89**, 1827 (1992)
[11] S. J. Clark, J. Harrison, C. L. Paul, and M. Frommer, *High sensitivity mapping of methylated cytosines*, Nucleic Acids Res. **22**, 2990 (1994)
[12] A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander, *Genome-scale DNA methylation maps of pluripotent and differentiated cells*, Nature **454**, 766 (2008)
[13] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen, *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning*. Nature **452**, 215 (2008)
[14] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker, *Highly integrated single-base resolution maps of the epigenome in Arabidopsis*, Cell **133**, 523 (2008)
[15] S. G. Jin, S. Kadam, and G. P. Pfeifer, *Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine*, Nucleic Acids Res. **38**, 1 (2010)
[16] Y. Huang, W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu, and A. Rao, *The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing*. PLoS One **5**, 1 (2010)
[17] K. R. Pomraning, K. M. Smith, and M. Freitag, *Genome-wide high throughput analysis of DNA methylation in eukaryotes*, Methods **47**, 142 (2009)
[18] H. Bayley, *Sequencing single molecules of DNA*, Curr. Opin. Chem. Biol. **10**, 628 (2006)
[19] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, *Real-time DNA sequencing from single polymerase molecules*, Science **323**, 133 (2009)
[20] Y. Wang, Q. Yang, and Z. Wang, *The evolution of nanopore sequencing*, Front. Genet. **5**, 1 (2015)
[21] M. Fyta, *Threading DNA through nanopores for biosensing applications*, J. Phys.: Condens. Matter **27**, 273101 (2015)
[22] B. A. Flusberg, D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares, T. A. Clark, J. Korlach, and S. W. Turner, *Direct detection of DNA methylation during single-molecule, real-time sequencing*, Nat. Methods **7**, 461 (2010)
[23] B. M. Davis, M. C. Chao, and M. K. Waldor, *Entering the era of bacterial epigenomics with single molecule real time DNA sequencing*. Curr Opin Microbiol. **16**, 192 (2013)
[24] R. J. Roberts, M. O. Carneiro, and M. C. Schatz, *The advantages of SMRT sequencing*. Genome Biol. **14**, 405 (2013)
[25] C.-X. Song, T. A. Clark, X.-Y. Lu, A. Kislyuk, Q. Dai, S. W Turner, C. He, and J. Korlach, *Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine*. Nat. Meth. **9**, 75 (2012)
[26] M. J. Blow, T. A. Clark, C. G. Daum, A. M. Deutschbauer, A. Fomenkov, R. Fries, J. Froula, D. D. Kang, R. R. Malmstrom, R. D. Morgan, J. Posfai, K. Singh, A. Visel, K. Wetmore, Z. Zhao, E. M. Rubin, J. Korlach, L. A. Pennacchio, R. J. Roberts, *The epigenomic landscape of prokaryotes*. PLoS Genetics, **12**, e1005854 (2016)
[27] S. J. Mondo, R. O. Dannebaum, R. C. Kuo, K. B. Louie, A. J. Bewick, K. LaButti, S. Haridas, A. Kuo, A. Salamov, S. R. Ahrendt, R. Lau, B. P. Bowen, A. Lipzen, W. Sullivan, B. B. Andreopoulos, A. Clum, E. Lindquist, C. Daum, T. R. Northen, G. Kunde-Ramamoorthy, R. J. Schmitz, A. Gryganskyi, D. Culley, J. Magnuson, T. Y. James, M. A. O'Malley, J. E. Stajich, J. W. Spatafora, A. Visel, and I. V. Grigoriev, *Widespread adenine N6-methylation of active genes in fungi*. Nat Genet. **49**, 964 (2017)
[28] M. Wanunu, *Nanopores: A journey towards DNA sequencing*, Phys. Life Rev. **9**, 125 (2012)

[29] C. Dekker, *Solid-state nanopores*, Nat. Nanotech. **2**, 209 (2007).
[30] J. J. Kasianowicz, E. Brandin, D. Branton, and D. W. Deamer, *Characterization of individual polynucleotide molecules using a membrane channel*, Proc. Natl. Acad. Sci. **93**, 13770 (1996)
[31] M. Zwolak and M. Di Ventra, *Electronic signature of DNA nucleotides via transverse transport*, Nano Lett. **5**, 421 (2005)
[32] E. Shapir, H. Cohen, A. Calzolari, C. Cavazzoni, D. A. Ryndyk, G. Cuniberti, A. Kotlyar, R. Di Felice, and D. Porath, *Electronic structure of single DNA molecules resolved by transverse scanning tunnelling spectroscopy*, Nat. Mater. **7**, 68 (2008)
[33] A. H. Laszlo, I. M. Derrington, H. Brinkerhoff, K. W. Langford, I. C. Nova, J. M. Samson, J. J. Bartlett, M. Pavlenok, and J. H. Gundlach, *Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA*, Proc. Natl. Acad. Sci. **110**, 18904 (2013)
[34] J. Schreiber, Z. L. Wescoe, R. Abu-Shumays, J. T. Vivian, B. Baatar, K. Karplus, and M. Akeson, *Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands*, Proc. Natl. Acad. Sci. **110**, 18910 (2013)
[35] J. Shim, G. I. Humphreys, B. M. Venkatesan, J. M. Munz, X. Zou, C. Sathe, K. Schulten, F. Kosari, A. M. Nardulli, G. Vasmatzis, and R. Bashir, *Detection and quantification of methylation in DNA using solid-state nanopores*, Sci. Rep. **3**, 1389 (2013)
[36] E. V. B. Wallace, D. Stoddart, A. J. Heron, E. Mikhailova, G. Maglia, T. J. Donohoe, and H. Bayley, *Identification of epigenetic DNA modifications with a protein nanopore*, Chem. Commun. **46**, 8195 (2010)
[37] M. Wanunu, D. Cohen-Karni, R. R. Johnson, L. Fields, J. Benner, N. Peterman, Y. Zheng, M. L. Klein, and M. Drndic, *Discrimination of methylcytosine from hydroxymethylcytosine in DNA molecules*, J. Am. Chem. Soc. **133**, 486 (2011)
[38] Z. L. Wescoe, J. Schreiber, and M. Akeson, *Nanopores discriminate among five C5-cytosine variants in DNA*, J. Am. Chem. Soc. **136**, 16582 (2014)
[39] M. Ayub, S. W. Hardwick, B. F. Luisi, and H. Bayley, *Nanopore-based identification of individual nucleotides for direct RNA sequencing*, Nano Lett. **13**, 6144 (2013)
[40] J. A. Cracknell, D. Japrung, and H. Bayley, *Translocating kilobase RNA through the staphylococcal alpha-hemolysin nanopore*, Nano Lett. **13**, 2500 (2013)
[41] M. Ayub and H. Bayley, *Individual RNA base recognition in immobilized oligonucleotides using a protein nanopore*, Nano Lett. **12**, 5637 (2012)
[42] A. E. P. Schibel, N. An, Q. Jin, A. M. Fleming, C. J. Burrows, and H. S. White, *Nanopore detection of 8-oxo-7,8-dihydro-2'-deoxyguanosine in immobilized single-stranded DNA via adduct formation to the DNA damage site*. JACS **132**, 17992 (2010)
[43] A. C. Rand, M. Jain, J. Eizenga, A. Musselman-Brown, H. E. Olsen, M. Akeson, and B. Paten, *Cytosine variant calling with high-throughput nanopore sequencing*. bioRxiv. 2016. doi:10.1101/047134
[44] J. T. Simpson, R. Workman, P. C. Zuzarte, M. David, L. J. Dursi, W. Timp. *Detecting DNA methylation using the Oxford Nanopore Technologies MinION sequencer*. bioRxiv. 2016. doi:10.1101/047142
[45] S. Chang, S. Huang, J. He, F. Liang, P. Zhang, S. Li, X. Chen, O. Sankey, and S. Lindsay, *Electronic signatures of all four DNA nucleosides in a tunneling gap*, Nano Lett. **10**, 1070 (2010)
[46] A. P. Ivanov, E. Instuli, C. M. McGilvery, G. Baldwin, D. W. McComb, T. Albrecht, and J. B. Edel, *DNA tunneling detector embedded in a nanopore*, Nano Lett. **11**, 279 (2011)
[47] M. Tsutsui, M. Taniguchi, K. Yokota, and T. Kawai, *Identifying single nucleotides by tunnelling current*, Nat. Nanotech. **5**, 286 (2010)
[48] M. Tsutsui, K. Matsubara, T. Ohshiro, M. Furuhashi, M. Taniguchi, and T. Kawai, *Electrical detection of single-methylcytosines in a DNA oligomer*, J. Am. Chem. Soc. **133**, 9124 (2011)
[49] T. Ohshiro, K. Matsubara, M. Tsutsui, M. Furuhashi, M. Taniguchi, and T. Kawai, *Single-molecule electrical random resequencing of DNA and RNA*, Sci. Rep. **2**, 501 (2012)
[50] Y. Zhao, B. Ashcroft, P. Zhang, H. Liu, S. Sen, W. Song, J. Im, B. Gyarfas, S. Manna, S. Biswas, C. Borges, and S. Lindsay, *Single-molecule spectroscopy of amino acids and peptides by recognition tunneling*, Nat. Nanotech. **9**, 466 (2014)
[51] J. R. Alvarez, D. Skachkov, S. E. Massey, J. Lu, A. Kalitsov, and J. P. Velev, *Intrinsic noise from neighboring bases in the DNA transverse tunneling current*, Phys. Rev. Appl. **1**, 34001 (2014)
[52] J. R. Alvarez, D. Skachkov, S. E. Massey, A. Kalitsov, and J. P. Velev, *DNA/RNA transverse current sequencing: intrinsic structural noise from neighboring bases*, Front. Genet. **6**, 213 (2015).
[53] J. N. Pedersen, P. Boynton, M. Di Ventra, A.-P. Jauho, and H. Flyvbjerg, *Classification of DNA nucleotides with transverse tunneling currents*. Nanotechnology **28**, 015502 (2016)
[54] T. Szalay and J. A. Golovchenko, *De novo sequencing and variant calling with nanopores using PoreSeq*, Nat. Biotechnol. **33**, 1 (2015)
[55] J. Comera and A. Aksimentiev, *DNA sequence-dependent ionic currents in ultra-small solid-state nanopores*, Nanoscale **8**, 9600 (2016)
[56] S. Carson, J. Wilson, A. Aksimentiev, P. R. Weigele, and M. Wanunu, *Hydroxymethyluracil modifications enhance the flexibility and hydrophilicity of double-stranded DNA*, Nucleic Acids Res. **44**, 2085 (2016)

[57] J. Lagerqvist, M. Zwolak, and M. Di Ventra, *Classification of DNA nucleotides with transverse tunneling currents*, Nano Lett. **6**, 779 (2006)
[58] M. Krems, M. Zwolak, Y. V. Pershin, and M. DiVentra, *Effect of noise on DNA sequencing via transverse electronic transport*, Biophys. J. **97**, 1990 (2009)
[59] J. Lagerqvist, M. Zwolak, and M. DiVentra, *Influence of the environment and probes on rapid DNA sequencing via transverse electronic transport*. Biophys.J. **93**, 2384 (2007)
[60] G. Sivaraman, R. G. Amorim, R. H. Scheicher, and M. Fyta, *Diamondoid-functionalized gold nanogaps as sensors for natural, mutated, and epigenetically modified DNA nucleotides*, Nanoscale **8**, 10105 (2016)
[61] G. Sivaraman, R. G. Amorim, R. H. Scheicher, and M. Fyta, *Insights into the detection of mutations and epigenetic markers using diamondoid-functionalized sensors*, RSC Adv. **7**, 43064 (2017)
[62] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi, *Fragment molecular orbital method: an approximate computational method for large molecules*, Chem. Phys. Lett. **313**, 701 (1999)
[63] V. Deev and M. A. Collins, *Approximate ab initio energies by systematic molecular fragmentation*, J. Chem. Phys. **122**, 154102 (2005)
[64] O. R. Meitei and A. Heßelmann, *Molecular energies from an incremental fragmentation method*, J. Chem. Phys. **144**, 084109 (2016)
[65] P. C. P. de Andrade and J. A. Freire, *Effective Hamiltonians for the nonorthogonal basis set*, J. Chem. Phys. **118**, 6733 (2003)
[66] P. C. P. De Andrade and J. A. Freire, *Electron transfer in proteins: Nonorthogonal projections onto donor-acceptor subspace of the Hilbert space*, J. Chem. Phys. **120**, 7811 (2004)
[67] M. Soriano and J. J. Palacios, *Theory of projections with nonorthogonal basis sets: Partitioning techniques and effective Hamiltonians*, Phys. Rev. B **90**, 075128 (2014)
[68] C. Fonseca Guerra, J. G. Snijders, G. te Velde, and E. J. Baerends, *Towards an order-N DFT method*, Theor. Chem. Acc. **99**, 391 (1998)
[69] G. TeVelde, F. M. Bickelhaupt, E. J. Baerends,C. Fonseca Guerra, S. J. A. van Gisbergen, J. G. Snijders, and T. Ziegler, *Chemistry with ADF*. J. Comput. Chem. **22**, 931 (2001)
[70] Y. Meir and N. S. Wingreen, *Landauer formula for current through an interacting electron region*, Phys. Rev. Lett. **68**, 2512 (1992)
[71] W. Timp, J. Comer, and A. Aksimentiev, *DNA base-calling from a nanopore using a Viterbi algorithm*, Biophys. J. **102**, L37 (2012).
[72] N. E. Singh-Miller and N. Marzari, *Surface energies, work functions, and surface relaxations of low-index metallic surfaces from first principles*, Phys. Rev. B **80**, 235407 (2009)
[73] D. Roca-Sanjuán, M. Rubio, M. Merchán, and L. Serrano-Andrés, *Ab-initio determination of the ionization potentials of DNA and RNA nucleobases*, J. Chem. Phys. **125**, 37 (2006)
[74] T. Tsukamoto, Y. Ishikawa, Y. Sengoku, and N. Kurita, *A combined DFT/Green's function study on electrical conductivity through DNA duplex between Au electrodes*, Chem. Phys. Lett. **474**, 362 (2009)
[75] A. Okamoto, Y. Maeda, T. Tsukamoto, Y. Ishikawa, and N. Kurita, *A combined nonequilibrium Green's function/density-functional theory study of electrical conducting properties of artificial DNA duplexes*, Comput. Mater. Sci. **53**, 416 (2012)
[76] Y. Lee, H. Lee, S. Park, and Y. Yi, *Energy level alignment at the interfaces between typical electrodes and nucleobases: Al/adenine/indium-tin-oxide and Al/thymine/indium-tin-oxide*, Appl. Phys. Lett. **101**, 233305 (2012)