

SINGLE NUCLEOTIDE POLYMORPHISM (SNP) DISCRIMINATIONS
BY NANOPORE SENSING

A Thesis
presented to
the Faculty of the Graduate School
at the University of Missouri-Columbia

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

by
RUICHENG SHI

Dr. Liqun Gu, Thesis Supervisor

July 2018

© Copyright by Ruicheng Shi 2018

All Rights Reserved

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

**SINGLE NUCLEOTIDE POLYMORPHISM (SNP) DISCRIMINATIONS
BY NANOPORE SENSING**

presented by Ruicheng Shi,

a candidate for the degree of Master of Science

and hereby certify that, in their opinion, it is worthy of acceptance.

Li-Qun Gu, Ph.D., Biological Engineering

Sheila Grant, Ph.D., Biological Engineering

Xiaoqin Zou, Ph.D., Biochemistry

ACKNOWLEDGMENT

Firstly, I would like to express my deepest gratitude to my advisor Dr. Li-Qun Gu, for his constant support of my research, for his persistence, enthusiasm, and broad knowledge. His guidance is available at all times during my research and writing of this thesis. I could not ask for a better thesis advisor and research mentor for my master study. I could also thank Dr. Sheila Grant and Dr. Xiaoqin Zou for taking their precious time out to serve as my committee members.

Secondly, I would like to thank my fellow lab mates, Xinyue Zhang, Kai Tian and Yong Wang, for their valuable insights and encouragements (both mentally and physically), and for all the fun we had in the last three years. Also, I thank my friends from Duan Lab, Michael and Ping, for their caring when I was a volunteer there.

Last but not the least, I would like to show gratitude to my parents for their unconditional love and support across the Pacific.

Table of Contents

ACKNOWLEDGMENT..... ii

LIST OF FIGURES v

LIST OF TABLES vii

ABSTRACT..... viii

CHAPTER 1 INTRODUCTION 1

 Overview of Nanopore Sensing 1

 Typical Nanopore detection scenario 5

 Overview of Single Nucleotide Polymorphism 7

 Types of SNPs 7

 SNPs and Point Mutations..... 8

 Importance of SNPs 9

 Detections of SNPs..... 11

CHAPTER 2. SEQUENCE - SPECIFIC COVALENT CAPTURE COUPLED WITH HIGH - CONTRAST NANOPORE DETECTION OF A DISEASE - DERIVED NUCLEIC ACID SEQUENCE..... 16

 Abstract 16

 Introduction 16

 Discussion 17

 Materials and General Procedures..... 24

CHAPTER 3. AN INTERNAL RNA BARCODE STRATEGY FOR LABEL-FREE NANOPORE MULTIPLEX SNP DISCRIMINATION 33

 Abstract: 33

 Introduction 34

 Results and discussions 35

 Further Improvements: 41

Discussion	43
Material and methods	46
References:.....	52

LIST OF FIGURES

Figure 1: A typical wild-type α -HL Nanopore and its corresponding dimension.....	1
Figure 2: General channel formation of alpha-hemolysin	2
Figure 3: Cross section view of a typical nanopore detection scenario	5
Figure 4: Typical melting curve scenario for 100-bp homozygotes and their SNP containing counterparts.....	12
Figure 5: Covalent capture of specific nucleic acid sequences by interstrand crosslink formation.....	18
Figure 6: Ap-containing probes selectively crosslink with the 1799 T→A mutant BRAF kinase gene sequence.	19
Figure 7: Crosslink yield as a function of the amount of mutant BRAF sequence present in mixtures of mutant and wild - type duplexes	20
Figure 8: Crosslinked DNA generated from the mutant BRAF - probe duplex E can be readily detected by its unique current signature in the α - HL nanopore.....	22
Figure 9: Incremental current decreases induced by sequential, irreversible blocking of individual α - HL pores by crosslinked duplexes in an experiment with multiple channels embedded in the lipid bilayer.....	24
Figure 10: DNA duplexes used in these studies.....	27
Figure 11: Ap-containing probes selectively cross-link with the 1799 T→A mutant BRAF kinase gene sequence.	28
Figure 12: Iron-EDTA footprinting defines the cross-link location in duplex E	29
Figure 13: Quantitative detection of cross-linked duplex E prepared for nanopore experiment by gel analysis.....	30
Figure 14: The effect of storage condition on the cross-link yield.	30
Figure 15: Continuous recording of the block by cross-linked mutant target/probe duplex E ..	31
Figure 16: Histograms showing the current-blocking levels (left) and dwell times (right) for the uncross-linked duplex F.....	31
Figure 17: Current traces for mixtures of cross-linked duplex E and uncross-linked duplex F.	32

Figure 18: Plot of detected ratio of persistent current blocking events versus short-duration current blocks as a function of the ratio of cross-linked duplex E.....	32
Figure 19: Validation histogram indicating the level differences of rA, rC and their DNA counterparts.....	36
Figure 20: MT•P duplexes generate unique step patterns that not only can discriminate between wild-type and mutants, but also different mutant types.	37
Figure 21: In the absence of Hg ²⁺ , both MT•P and WT •P duplexes generate spike events indistinguishable from other types of events, such as ssDNA translocation	49
Figure 22: A close look at the pattern difference between WT•P and MT•P duplexes.	43
Figure 23: Histograms showing S1 and S2 dwell time differences between KRAS G12D and TP53 R172H duplex	50
Figure 24: Traces recorded using the automated protocol.	51

LIST OF TABLES

Table 1: Sequences of probes and targets used in main study	43
Table 2: Duplexes illustrated in complimentary form	43
Table 3: Sequences used for validation.....	43

SINGLE NUCLEOTIDE POLYMORPHISM (SNP) DISCRIMINATIONS BY
NANOPORE SENSING

Ruicheng Shi

Dr. Li-Qun Gu, Thesis Supervisor

ABSTRACT

Single Nucleotide Polymorphisms (SNPs) are a common type of nucleotide alterations across the genome. A rapid but accurate detection of individual or SNP panels can lead to the right and in-time treatments which possibly save lives. In one of our studies, nanopore is introduced to rapidly detect BRAF 1799 T→A mutation (V600E), with the help of an Ap-dA cross-link right at the mutation site. These sequence-specific crosslinks are formed upon strong covalent interactions between probe based abasic sites (Ap) and target based deoxy-adenosine (dA) residues. Duplexes stabilized by the crosslink complexes create indefinite blocking signatures when captured in the nanopore, creating a high contrast compared to the “spike-like” translocations events produced by the uncrosslinked and wildtype duplexes. Those consistent blocking events couldn't be resolved unless an inverted voltage is applied. In a 1:1 BRAF mutant-wildtype mixture, the nanopore can successfully discriminate between the two sequences in a quantitative manner. In summary, nanopore paired with sequence-specific crosslink can detect a specific type of SNP with a high contrast manner.

In another study, nanopore sensing is modified to be capable of detections with multiple SNPs in a single detection mix. To achieve this, an RNA homopolymer barcode is integrated into the probe sequence so nanopore can read out a distinctive level signature

when the target-probe duplex is de-hybridizing through the pore. Since different RNA homopolymers (e.g. Poly rA and Poly rC) can generate signature levels distinctive from each other and other DNA sequences, they can be applied to generate characteristic patterns that simultaneously highlight multiple SNPs in the mixture. In this study, we assigned two different RNA barcodes (Poly rA and Poly rC) to label KRAS G12D and Tp53 R172H SNPs (both T→A mutations) in the solution. During nanopore readout, the KRAS G12D containing duplex generates a “downward” step pattern but Tp53 R172H always has an “upward” step pattern, the high contrast between those two patterns makes recognition easy enough with naked eyes, and further statistical analysis is unnecessary. Theoretically, at least four different barcodes can be implemented at the same time, furthermore, the length of the barcode can also affect the barcode pattern. Thus, in theory, a panel of more than 10 SNPs can be identified simultaneously.

CHAPTER 1 INTRODUCTION

Overview of Nanopore Sensing

Nanopores, in general, refer to pores with diameters in nanometer scales. They could be biological (proteins such as α -HL, aerolysin and MspA), artificial (solid state silicon nitride or graphene) and hybrid. Biological nanopores, mostly referred to as α -hemolysin (abbreviated as α -HL), were first discovered as exotoxins secreted by *Staphylococcus aureus*. In general, they are usually depicted as cross-membrane proteins that form channels through the lipid bilayer, allowing nano-scale and angstrom-scale particles such as biomolecules (e.g. RNA, DNA and proteins) and ions (e.g. Na^+ , K^+ and Cl^-) to be transported to the other side (from *cis* to *trans*).² A complete α -HL nanopore is a

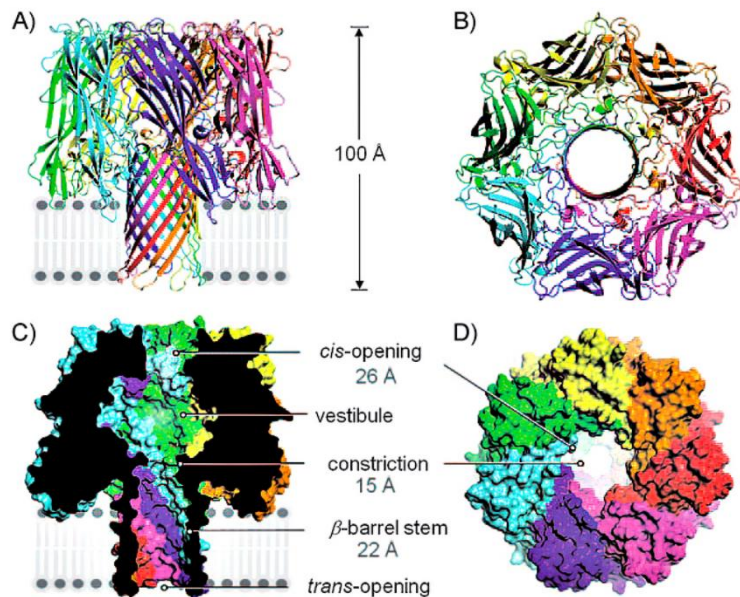


Figure 1. A typical wild type α -HL Nanopore and its corresponding dimension. Generally, seven identical monomers arrange coaxially to form a heptamer.

A and C: Cross section view;

B and D: Top-down view.

heptamer consists of seven identical monomers. Each monomer 1) starts with an amino

latch at the N terminus that plays a crucial part in heptamer formation and cell lysis³; 2) has a stem region consists of mainly beta sheets for stem formation, and 3) has a rim domain helps with establishing the correct orientation on the lipid bilayer. (Trp¹⁷⁹ Tyr¹⁸² Trp¹⁸⁷ Arg²⁰⁰ and Met²⁰⁴ are responsible for attachment to lipid bilayer/phospholipids.)

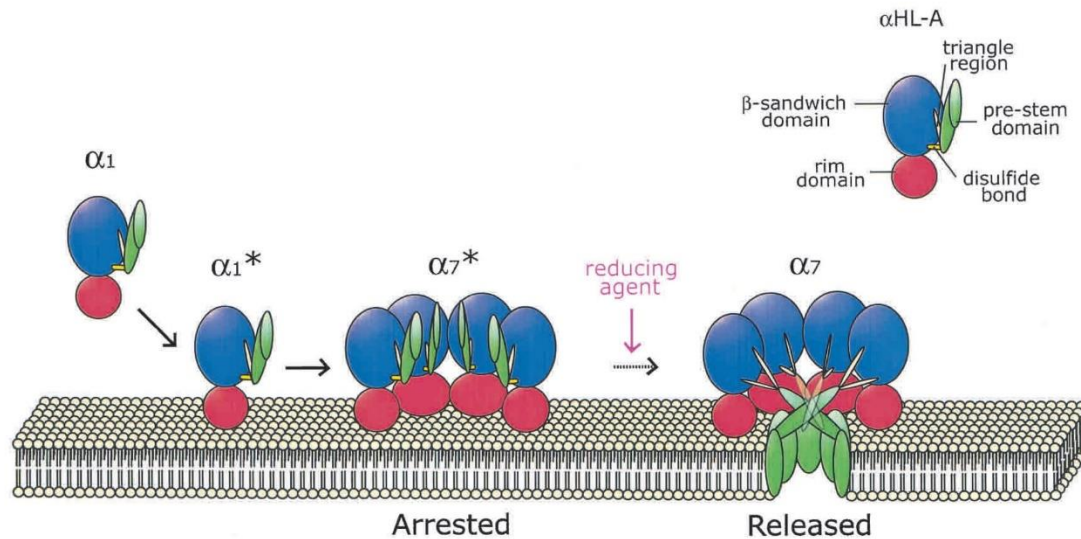


Figure 2. General channel formation of alpha-hemolysin. Notice the folded stem domain before penetration through the lipid membrane (adapted from Kawate et.al⁴)

Based on the properties of the monomer, several channel-forming mechanism was proposed^{3,5-7}. One of the popular mechanisms involves⁴ 1) the binding of the correct oriented monomer (alpha 1*, with a folded-up pre-stem domain) on the lipid surface through the rim domain; 2) The complete aggregation of seven monomers and the formation of the pre-pore alpha 7*; 3) The arrangement of the stem domain under reducing conditions and the formation of the assembled Alpha haemolysin channel. Thus, the complete structure of an assembled α -HL consists of a “mushroom” like cap-rim complex sitting on the top of the bilayer and a beta sheet stem barrel that buried under the lipid bilayer. Once a nanopore settles on the membrane surface, a 10 nanometer long

asymmetrical channel will form with an unevenly distributed diameter ranging from 1.5-2.6 nm (Figure 1). The narrowest section of the channel (1.5nm) is termed the constriction site. It is one of the main recognition sites for nanopore sensing, only allowing for single-stranded oligonucleotide to thread through. The opening located on the cap side is a little bit wider, which can merely incorporate double-stranded oligonucleotides. The other opening on the stem side also has a width only permitting for single-stranded-oligonucleotide translocation. Due to the distinguished asymmetrical configurations of the nanopore on two sides of the membrane, the side mushroom-like cap is facing can be defined as “*cis*”; the other side however, to which the stem is pointing, as “*trans*”.

Not until 1990s had scientists discovered its potentials acting as biosensors²: Since the unique nanometer-wide channel in α -HL could let ions diffuse through to produce a Pico-scale current under high salt condition and voltage gradient, (In fact, it is the only known nanopore that could remain open at neutral pH and high ionic strength)⁸, it is then possible to generate an electrostatic field adjacent to the pore to attract and pull (or push) charged molecules through the channels, and during this process, the ion current flowing through the channel is modulated specifically and measurably, which enables the scientists to reveal the presence and position of the blockades inside the channel. By measuring parameters such as dwell time and blocking level, certain molecules can be categorized and identified rapidly.

Such properties have guaranteed its feasibility for detecting particles of different sizes, especially for ultrafast DNA sequencing. Due to its rapid translocation speed (estimated 1000-10,000 nucleotides per second⁹), nanopore sequencing will outperform any other DNA sequencing strategies at the moment. If it is the case, an era of personalized whole genome sequencing under \$1,000 will be soon around the corner¹⁰.

Other than ultrafast DNA sequencing, nanopore is also applicable in other fields of studies. Nanopore detection, for example, can be applied to several other applications other than rapid DNA sequencing. 1) For example, Cheley et al.¹¹ reported an engineered nanopore that had a potential in fabrication of multianalyte biosensors and controlling chemical reactions. 2) Zhang et al.¹², reported a novel method using nanopore and Biotin-labeled miRNA in revealing protein/miRNA unfolding process. 3) Wang et al.¹³ and Zhang et al.¹⁴ implemented nanopore with chemically modified DNA duplexes for rapid SNP detections. 4) Jayawardhana, et al.¹⁵, converted nanopore into a useful weapon for explosives detections.

Besides natural nanopores, nano-scale artificial pores have also been fabricated to resemble the function of its biological counterparts. Solid-state nanopore, in specific, has been developed recently to overcome the drawbacks such as instability and restricted dimensions of the existing biological pores.² Solid-state nanopores could be made by dedicatedly guided etching through insulating layers or glass. During this process, pores

at controllable sizes could be made on treated surfaces, granting both with high stability and flexibility under extreme conditions. Benefiting from these properties, solid-state nanopores may have the potential to be integrated into portable devices and arrays. However, due to the limitation of engraving techniques, solid-state nanopores could not be as uniform as the biological nanopores that have gone through the process of self-assembly. In general, solid-state nanopores could provide an equally promising prospect of application as the biological ones.

Typical Nanopore detection scenario

In general, the nanopore experiment is carried out in a vacuum grease sealed, detachable device with two chambers (cis side and trans side) separated by a thin Teflon film (25 μm in thickness). Both cis and trans side are connected with each other through a 100-125 μm orifice on the film.

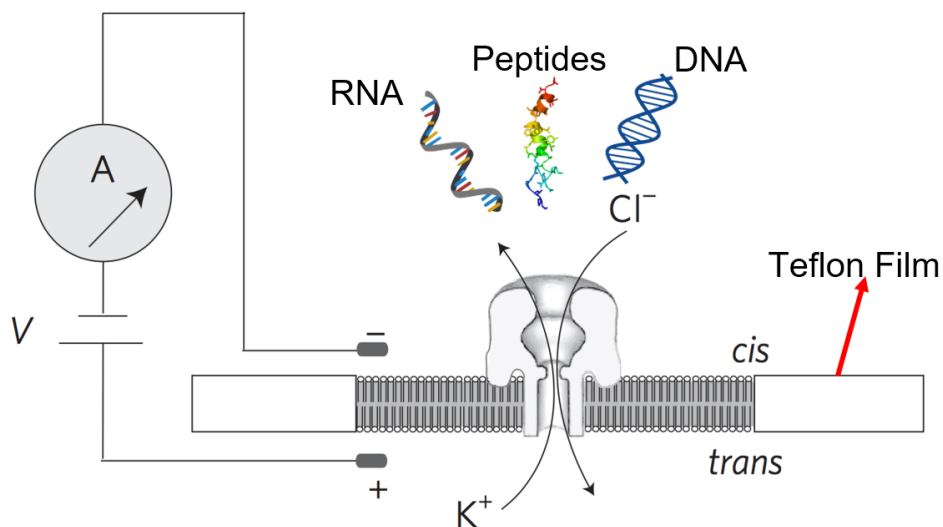


Figure 3. Cross section view of a typical nanopore detection scenario. Usually, samples are added at the cis side, but locations may vary based on the sample charge type.

Artificial bilayer formation: During experiments, artificial phospholipid bilayer was formed on the orifice by Montal-Mueller method: First, a fixed volume (1 ml) of buffer (usually 1 M KCL, 10 mM Tris, pH = 7.2, combinations may vary according to the experiment) is injected into both cis and trans chambers. Second, 1 dip (approximately 0.1 ml) of pretreat and lipid are added on the film and buffer on each side and a five-minute interval was introduced to assure that the lipid was spread evenly and completely on the liquid surface. Third, additional 1 ml buffer was added on each side, bringing up the lipid surface and forming a lipid bilayer through the orifice. Normally, due to the capacitive property of the artificial membrane, a successfully induced bilayer could generate a typical square wave peaks at 150-200 pA under a triangular wave input.

Signal recording and data analysis: At the start of the recording, α -HL nanopores (roughly 1ul) and samples are added respectively to cis and trans chambers under an applied potential of 100-180mV after the bilayer formation. Pico-ampere scale currents (pA) are recorded by an Axopatch 200B amplifier and filtered with a built-in 4-pole low-pass Bessel filter at 5 kHz. Digital signals are then transmitted into the terminal using a DigiData 1440A A/D converter (Molecular Devices) at a sampling rate of 20 kHz. The data recording and acquisition processes are controlled through a Clampex 10.4 (Molecular Devices). During the final stage, nanopore traces are examined on Clampfit 10.4 to generate event scatter plots and duration/level histogram for further analysis.

Overview of Single Nucleotide Polymorphism

Types of SNPs

SNPs, or Single Nucleotide Polymorphisms, are one type of genetic alterations that only involve the change of a single nucleotide within the sequence. It can be derived from the misincorporation of an undesired base (into a specific location) by DNA polymerases during sequence replications, or by other external factors. Theoretically, SNP can appear anywhere randomly across the sequence. However, it is not the case. A study led by Barreiro et al.¹⁶ showed a clear bias towards SNP occurrence at the non-coding regions (or intergenic regions) than the coding regions. This is because a portion of SNPs (termed “nonsynonymous SNPs”) appear at the coding regions will more likely to alter the amino acid sequences than the others (synonymous SNPs), which further affect protein structure and biological functions. Haplotypes with such dysfunctional proteins are more likely to be repressed during selection. Thus, in general, it would appear to us that fewer SNPs in the coding region could be inherited. Nonsynonymous SNPs can also be further divided into two subcategories, **nonsense and missense SNPs**. **Nonsense SNPs** often introduce a premature stop codon to the transcript, resulting in truncated peptides or no protein production at all (point-nonsense mediated mRNA decay)¹⁶; **Missense SNPs** modify peptide information by changing codons mapping to a different amino acid. Those alterations usually result in protein property and/or structural changes, undermining its normal functions. One of the well-known cases is Sickle Cell Anemia^{17,18}. This inheritable disease originates from a single A to T transversion at position r334, leading to a deformed hemoglobin (HB.s) almost incapable of Oxygen transportation. Besides

Nonsynonymous SNPs, SNPs in coding regions can also be synonymous. However. Due to codon degeneracy, peptide sequences are immune to those genetic alterations and they can be descended to the progenies with ease.

Story of the SNPs in the non-coding regions are a little bit different since they cast their influences on a transcriptional/post-transcriptional level though reaching the same end goal. Non-coding regions are usually responsible for protein expression regulations, Pre-mRNA splicing, mRNA and non-coding RNA regulations, etc¹⁹⁻²². Mutations in those regions usually hamper normal DNA transcriptions and result in abnormalities in matured mRNAs and altered protein expression levels, which similarly, tamper normal protein functions. However, due to the nature of gene transcriptions/transcription modifications, those alterations are not constantly under selection pressure thus enjoy a higher occurrence rate than those in the coding regions.

SNPs and Point Mutations

Though both SNPs and point mutations are dealing with genetic alterations on a single-nucleotide level, there are some basic differences²³. First, point mutations have a much wider definition. It often stands for single nucleotide alterations that are both inheritable and non-inheritable, while SNPs specifically referring to point mutations that are **inheritable**. Mutations will occur over time, but they are constantly monitored and corrected by the self-repairing mechanisms which are effectively eliminating anomalies out of the sequences. Only single mutations hidden from the surveillance will be preserved and passed down to the progeny cells, transforming into SNPs.

Second, unlike mutations, most SNPs are bi-allelic. That means at a specific mutation site there will only be two possibilities, for example, a mutated locus at position 500 only allows the sequence to be either A or G. The reason behind this phenomenon is unclear. One possibility is that transitions (which is bi-allelic) are more likely to be preserved than transversions (which is multi-allelic) due to fewer lesions to the sequences, thus, they are more likely to go undetected and consequently survive self-corrections. In fact, approximately 2/3 of the SNPs are transitions.²⁴

Third, SNPs have stronger statistical significance. SNPs are defined as single nucleotide variations with a >1% detection rate in a population²⁵. While variations with occurrence less than 1% can only be defined as mutations. With approximately 1% of the population bearing the same genetic variation, it is safe to regard SNPs as an important marker in not only genetic, but also demographic studies. SNP can accurately summarize the historical genetic change of the entire population, including phenotypic variations and evolutionary divergence among human races²⁶.

Importance of SNPs

As described above, one of the important roles SNPs serve is to accurately reflect human genetic evolution progressions and divergence, however, the significance of SNP is far beyond this. The significance of SNP should be viewed on a larger scale. In general,

SNPs should be treated as a powerful tool in human health improvement on all three different levels: Individual genetics, Family genetic inheritance and Population genetics.

Individual Level: SNP panels are unique among individuals. Sequence variations are constantly influencing protein sequences and their expression levels, which consequently shape one's responses towards outside stimulations such as diseases. Since future medicines emphasize more on personally-tailored diagnosis and treatments. Gaining a deeper knowledge of individual disease susceptibility will enhance treatment qualities and save lives. With the thorough but in-depth individual genetic profile provided by the selected SNP panels, physicians can accurately determine the subtype of the disease and perform the most effective treatment according to patient's genotypes and health conditions. Doctors can also predict patient's susceptibility to certain disease subtypes and take necessary precautions to prevent further deteriorations.

Family genetic inheritance: Though SNPs are highly unique among individuals, family members somehow share the same patterns. By analyzing those patterns scientists can have a relatively clear image on the family inheritance map. E.g. the origins of the family, possible family ancestors, potential family members, potential disease susceptibilities. In short, SNPs can effectively determine family inheritance both biologically and demographically.

Population genetics: SNPs also play important roles on the population level. Through high-throughput genotyping and analysis, scientists can have a brief overview of the genetic history of the population²⁶. They can also acquire information on how each race has evolved and separated over time. Based on large-scale SNP mapping and comparisons, a clear evolution course of each race can be drawn.

Detections of SNPs

Like other sequence-based detections, typical approaches for SNP detections are rather traditional. Up to date, the major routine techniques consist of: High-throughput sequencing^{27,28}, hybridization-based genotyping²⁹, and enzyme-based methods^{17,30}.

High-throughput sequencing: Sequencing has been and will always remain a powerful tool for sequence interrogation. Its widely used sequencing-by-synthesis mechanism is the most direct, accessible and high-throughput way to acquire genetic codes at a specific location. Among all three generations of sequencing approaches, the second-generation sequencing (454 pyrosequencing, Illumina sequencing, SOLiD sequencing, etc.) are the most suitable methods in the current. Because their relatively short read length compared to the first-generation Sanger sequencing, they are not suitable for reviewing genome-wide novel SNPs. However, they are perfect for verifying a specific set of known SNPs at a designated location. Combined with the high-throughput read numbers (454 pyrosequencing can achieve 1 million reads per run), second generation sequencing can identify multiple panels of SNPs at a small region or increase readout accuracy through deeper sequencing depth.

However, it is far from enough after acquiring the sequencing data, another major task is to identify the SNP from the acquired sequences. To tackle this predetermined computer algorithms are employed to “call” potential single nucleotide variants (SNVs), referring to previously sequenced, genome-wide SNP databases^{31,32}. Up to date, Probabilistic based³³ and Heuristic based algorithms³⁴ are the two most used methods.

Hybridization-based genotyping: DNA duplexes base their stabilities on the hydrogen bonds from perfectly matched base pairs and base stacking forces. The introduction of a mismatch will partially undermine the consistency of base stacking and joint forces from other hydrogen bonds, which in turn destabilize duplex thermal stability, such as melting temperatures.

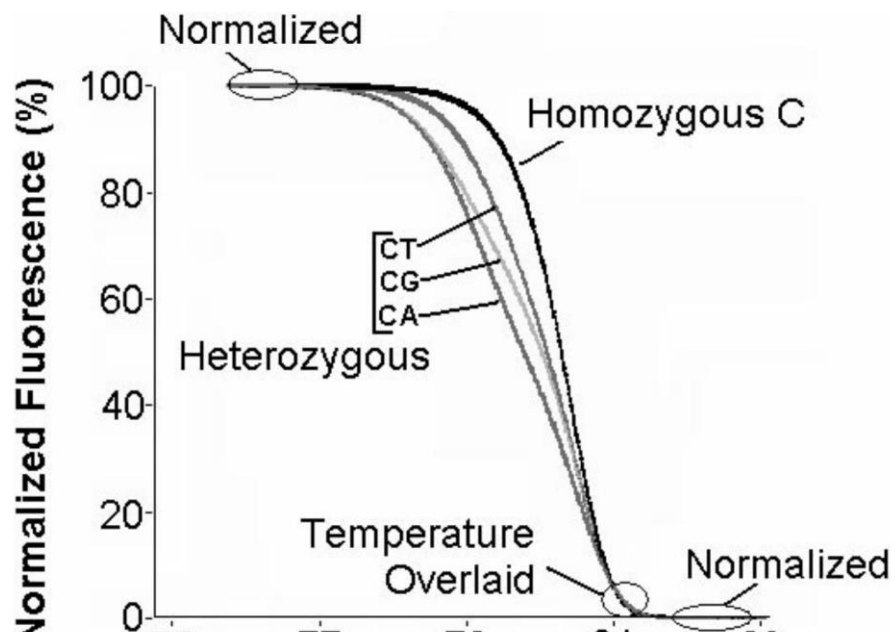


Figure 4 Typical melting curve scenario for 100-bp homozygotes and their SNP containing counterparts (adapted from Reed et al.¹)

Specific techniques employing melting temperature varies. One of the approaches involving biotin enrichment of samples on the beads after PCR. After necessary washing steps to get rid of extra samples. Samples attached to the beads are forced to annealed to pre-designed oligonucleotides that would only perfectly matched to one type of the samples (either wildtype or mismatched). After annealing, temperature is swept across the potential region and a melting temperature curves are generated and determined. By analyzing vertical differences at specific temperature points, one can easily pick up and categorize each genotype. Usually, an SNP containing sample will result in lower-than-usual melting temperatures. Another method involves a predetermined probe, which has a fluorophore and a quencher on each end. Initially, the probe has a chunk of matching sequences on either side so it can self-hybridize into a stem-loop formation. In this configuration, quencher and fluorophore are in close proximity to each other so no fluorescence is emitted. After annealing to the target samples, the complementary sequences in the middle will form duplexes structures with the probe and “stretches” the probe into a linear configuration. In this case, the quencher is far away from the fluorophore and unable to inhibit fluorescence emission. By measuring the emission intensity differences across the targeted temperature range, SNPs can be identified easily²⁹. In general, methods based on hybridization are relatively accurate but with low sensitivity. This is due to the limited vertical curve distance between two different genotypes at a given temperature point.

Enzyme-based methods: Mostly, enzyme-based methods generally refer to any SNP identification techniques involving using an enzyme. Thus, universal methods such as

PCRs and RFLPs (Restriction fragment length polymorphisms) can be counted into this category.

RFLPs. Some SNPs are situated in a location whose sequences can be recognized and cleaved by a specific type II restriction enzyme. Those cleaved fragments usually form a distinctive band pattern on the gel¹⁷. By analyzing specific patterns digested by the restriction enzymes, SNP containing alleles can be discovered with high accuracy. However, this method is only applicable when SNP is in a recognizable sequence, which limiting its range of applications. Moreover, extra labors are required for digestion and gel electrophoresis, which further inhibits high-throughput analysis.

PCR. General PCR methods can be modified to interrogate single SNPs. The merit of this method utilizes the need of a perfect match at 3' end of the primers for successful elongations³⁰. Based on this mechanism, two pairs of primers could be designed, with one of the primer pairs target at the WT sequences and the other aiming at the SNP containing allele. Each primer has a 3' end aligning precisely at the potential SNP sites. If the samples only consist of SNP containing sequences, the PCR products will be exclusively from the primer pair with the perfect match. If the sample is a WT/SNP mixture, both primer pairs will elongate and produce PCR products. PCR SNP detection is relatively efficient in interrogating specific known SNPs. However, they are not every proficient when facing high-throughput demands. Moreover, like other detection approaches, PCR is labor intensive and can't achieve real-time interrogation.

Compared with nanopore detection. In general, methods described above are the most commonly used approaches for SNP detections. However, most of them are suffering from intensive labor demands, or not suitable for high-throughput integrations, or not very sensitive to low trace of positive samples. On the other hand, nanopore detection is ultrafast, label free and not labor demanding. However, when it comes to SNP detection, no specific methods have been developed accordingly. Existing methods^{13,14,35,36} are not suitable for multiple SNP detections or still need extra labeling process. In this study, new SNP detection schemes are proposed and investigated with the aim of improving 1) mono SNP digitized detections, and 2) multiple SNP discriminations. With the help of crosslink reaction and integrated barcode designs, we are one step further towards nanopore digital analysis and future high-through sensing.

CHAPTER 2. SEQUENCE-SPECIFIC COVALENT CAPTURE COUPLED WITH HIGH-CONTRAST NANOPORE DETECTION OF A DISEASE- DERIVED NUCLEIC ACID SEQUENCE

Abstract

Hybridization-based methods for the detection of nucleic acid sequences are important in research and medicine. Short probes provide sequence specificity, but do not always provide a durable signal. Sequence-specific covalent crosslink formation can anchor probes to target DNA and might also provide an additional layer of target selectivity. Here, we developed a new crosslinking reaction for the covalent capture of specific nucleic acid sequences. This process involved reaction of an abasic (Ap) site in a probe strand with an adenine residue in the target strand and was used for the detection of a disease-relevant T→A mutation at position 1799 of the human BRAF kinase gene sequence. Ap-containing probes were easily prepared and displayed excellent specificity for the mutant sequence under isothermal assay conditions. It was further shown that nanopore technology provides a high contrast—in essence, digital—signal that enables sensitive, single-molecule sensing of the cross-linked duplexes.

Introduction

Methods for the detection of DNA and RNA sequences are important in research and medicine, and many different approaches will undoubtedly be required to meet various needs.³⁷⁻⁴⁰ Logically, many strategies for the detection of nucleic acid sequences rely on Watson–Crick hybridization of a probe strand to target DNA or RNA in samples. However, the noncovalent, inherently reversible nature of nucleic acid

hybridization presents challenges, because the signal can be compromised by partial denaturation of the probe–target duplex during analysis (e.g., washing). The use of longer (>20 nt) probes increases the stability of probe–target complexes but degrades sequence specificity.⁴¹⁻⁴³

Covalent crosslinks can be used to stabilize target–probe complexes,^{44,45} and, in some cases, can provide an additional layer of target selectivity beyond that afforded by Watson–Crick hybridization.⁴⁶⁻⁴⁸ In the work described here, we developed a new crosslinking reaction that might be useful for the covalent capture of specific nucleic acid sequences. The crosslinking probes used in these studies were prepared in a one-step procedure from inexpensive commercial reagents and achieved excellent sequence specificity under isothermal assay conditions. Crosslinked DNA duplexes generated in these studies were quantitatively measured by using denaturing gel electrophoresis and a protein nanopore.

Discussion

The crosslinking process developed here involved covalent reaction of an abasic (Ap) site in the probe strand with a deoxyadenosine (dA) residue in the target strand (Figure 5 A).5Importantly, Ap-containing probe strands were easily generated by treatment of the corresponding 2'-deoxyuridine-containing oligo deoxyribonucleotide with the enzyme uracil DNA glycosylase (UDG).⁴⁹⁻⁵¹ We set out here to determine whether the dA–Ap crosslinking reaction could be exploited for selective detection of a single-nucleotide polymorphism (SNP) in a human gene sequence. SNPs are the smallest differences that

can exist in nucleic acid sequences yet have immense importance in biology and medicine.^{37-40,52-54} Seeking proof-of-concept, we focused on detection of a T→A

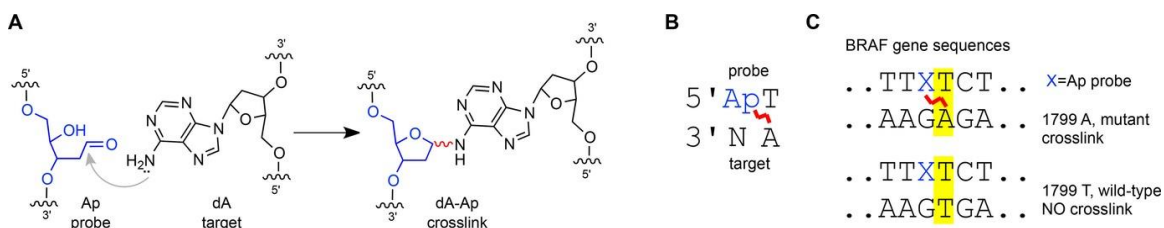


Figure 5. Covalent capture of specific nucleic acid sequences by interstrand crosslink formation. A) Covalent crosslink formation by reaction of an Ap aldehyde residue in the probe strand with an adenine residue in the target sequence. B) Sequence motif for dA–Ap crosslinking reaction. C) Sequence-specific covalent capture of the mutant BRAF gene sequence by an Ap-containing probe strand.

mutation at position 1799 of the BRAF kinase gene sequence that encodes a oncogenic V600E substitution in the protein.^{55,56} The anticancer drug vemurafenib (Zelboraf) specifically inhibits the V600E kinase.^{55,56}

We designed an Ap-containing oligonucleotide probe to crosslink with A1799 in the mutant BRAF sequence (Figure 5 C). Formation of the dA–Ap crosslink has previously been observed⁴⁹⁻⁵¹ when an adenine residue was positioned 1 nt to the 3'-side of the Ap site on the opposing strand (Figure 5 B), but, until now, the sequence specificity of this crosslinking reaction has not been characterized. Incubation of the mutant BRAF target–probe duplex A in HEPES buffer (50 mM, pH 7) containing NaCl (100 mM) at 37 °C gave a 7.3±2.0 % yield of a slowly migrating band on a denaturing polyacrylamide gel, consistent with that expected⁴⁹⁻⁵¹ for the crosslinked duplex (Figure 6 A, lane 1, and Figure 11). In contrast, the wild-type target–probe duplex B gave a relatively low yield of a slowly migrating band (2.3±0.6 %, Figure 6 A, lane 5).

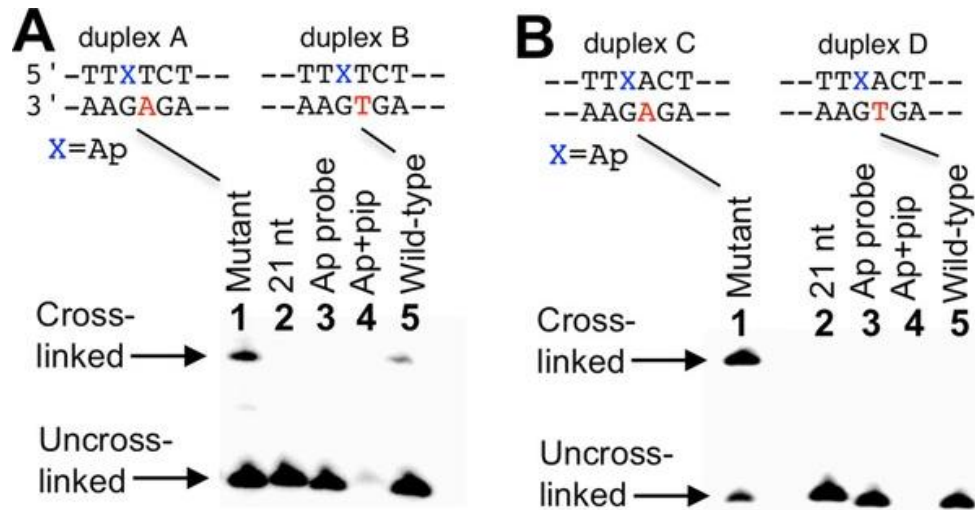


Figure 6. Ap-containing probes selectively crosslink with the 1799 T→A mutant BRAF kinase gene sequence. A) Gel electrophoretic analysis of crosslink formation in 21 bp duplexes containing the first-generation Ap-containing probe and either the mutant (lane 1) or wild-type (lane 5) BRAF sequence. B) Crosslinking by a second-generation Ap probe containing an adenine residue on the 3'-side of the Ap site is completely selective for the mutant BRAF sequence. Complete probe and target sequences are shown in Figure S1.

Encouraged by the selective crosslinking of the Ap-containing probe with the mutant BRAF target sequence, we sought a second-generation Ap-containing probe that would decrease the background signal associated with crosslinking to the wild-type BRAF sequence. The exact location of the crosslink generated between our first-generation probe and the wild-type sequence was uncertain, but we suspected that flexibility introduced by the T–T mismatch⁵⁷ enabled crosslink formation between the Ap site and the directly opposing guanine residue.⁵⁸⁻⁶⁰ Accordingly, we prepared a new probe strand containing an adenine residue on the 3'-side of the Ap site, such that the probe was complementary to the wild-type BRAF sequence (duplex D, Figure 6 B). We were

gratified to find that background crosslink formation between the second-generation probe and the wild-type BRAF sequence in duplex D decreased to undetectable levels (Figure 6 B, lane 5). We were further pleased to find that the desired crosslink formation between the second-generation probe and the mutant BRAF target sequence in duplex C increased dramatically to $85\pm 3\%$ (Figure 6 B, lane 1 and Figure 11). Iron-EDTA footprinting experiments confirmed that crosslink attachment in duplex C was to the adenine residue at position 1799 of the mutant sequence (Figure 12).

We next set out to determine whether the crosslinking reaction could be used to quantitatively measure the fraction of mutant versus wild-type BRAF sequence present in a sample. In this experiment, mixtures containing various proportions of mutant and wild-type duplexes were denatured by warming to $70\text{ }^{\circ}\text{C}$ in the presence of the second-generation Ap probe, cooled, incubated at $37\text{ }^{\circ}\text{C}$, and the crosslink yield was assessed by gel electrophoretic analysis. A clear connection between crosslink yield and the fraction of mutant duplex in the samples was observed (Figure 7).

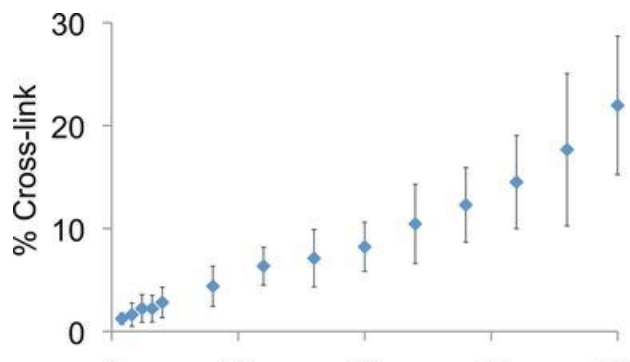


Figure 7. Crosslink yield as a function of the amount of mutant BRAF sequence present in mixtures of mutant and wild-type duplexes. Samples run in triplicate containing various proportions of mutant and wild-type BRAF duplexes (21 bp) were denatured by warming at $70\text{ }^{\circ}\text{C}$ in the presence of the second-generation probe, cooled, incubated at $37\text{ }^{\circ}\text{C}$, and assessed by gel electrophoretic analysis to determine the yield of interstrand crosslink.

We then examined use of the α -hemolysin (α -HL) protein nanopore for single-molecule detection of this crosslinked probe–target duplex. The α -HL ion channel can be used to create a device in which a nanoscale pore (1.4nm wide)⁶¹ spans a lipid bilayer that separates two chambers of aqueous electrolyte.^{62,63} Application of an electric potential induces a readily measured ion current, and the sequence and structure of nucleic acids can be analyzed based upon the characteristic current blocks produced when they are driven into the pore by the electrophoretic potential.⁶⁴⁻⁶⁷ For the nanopore experiments, we prepared a third-generation Ap-containing probe strand with a dC30 overhang on the 3'-end (Figure 8). The poly-dC extension was employed to increase the rate at which the α -HL nanopore captures the duplexes and to facilitate rapid unzipping of(un-crosslinked)

duplexes in the nanopore. ^{14,64-67}Separate gel electrophoretic analysis demonstrated that cross-link yields were not affected by the dC30 overhang (Figures 13 and 14).

In a device employing a single α -HL nanopore embedded in the lipid bilayer, analysis of the mixture generated by combination of the Ap probe with the mutant BRAF target sequence (duplex E) revealed several distinct current signatures. We observed very short current blocks, consistent with the translocation of single strands ($I/I_0 = 13.2 \pm 0.3\%$; $t = 150 \pm 30$ ms) and un-crosslinked duplexes ($I/I_0 = 12.2 \pm 0.9\%$; $t = 12 \pm 0.3$ ms; Figure 8). ⁶⁴⁻

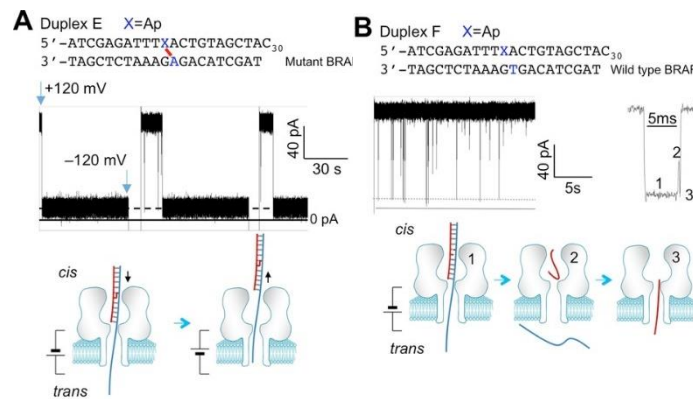


Figure 8. Crosslinked DNA generated from the mutant BRAF–probe duplex E can be readily detected by its unique current signature in the α -HL nanopore. The mixtures of species generated by incubation of the third-generation probe with either mutant or wild-type BRAF gene sequences were analyzed by using a single α -HL ion channel embedded in a lipid bilayer. Current traces were recorded at +120 mV in Tris (10 mM, pH 7.4) containing KCl (1 M) at 22 °C. A) Analysis of the mixture generated by hybridization of the mutant BRAF sequence with the Ap-containing probe strand. The current block was recorded for 1 min, then voltage polarity was reversed to translocate the crosslinked duplex back to the cis solution. Current trace showing persistence of the current block by crosslinked duplex E for 30 min is provided in Figure S6. B) Wild-type BRAF sequence does not generate crosslinked DNA when hybridized with the Ap-containing probe strand. Short current blocks are consistent with translocation of single-stranded DNA and un-crosslinked duplex F. The illustration depicts the three-step unzipping/translocation process for duplex DNA.

⁶⁷More importantly, we observed persistent current blocks ($I/I_0 = 13.2 \pm 0.3\%$, Figure 8A), consistent with capture of a crosslinked duplex in the nanopore. Following capture of the crosslinked duplex, current flow could be restored only when the voltage polarity of the

nanopore device was reversed; this caused the crosslinked duplex to back out of the channel. When the voltage polarity was reset, the open pore was again able to record current signatures associated with the nucleic acid species in the bulk mixture (Figure 8A). Analysis of the mixture generated by combination of the Ap probe with the wild-type BRAF gene sequence (duplex F) revealed no persistent current blocks, only short current blocks consistent with the translocation of single strands and un-crosslinked duplexes (Figures 8B and 16). The current signature of a crosslinked duplex is unmistakably different from that of the un-crosslinked DNA, thus providing a high-contrast signal for detection of the BRAF mutation. When multiple α -HL ion channels were embedded in the lipid bilayer, the analysis of mixtures derived from mutant duplex E revealed a series of incremental current decreases consistent with sequential, irreversible blocking of individual pores by the crosslinked duplex (Figure 9).¹⁴ By counting the number of events with each type of current signature, the nanopore could be used for the quantitative analysis of mixtures containing both mutant and wildtype BRAF sequences (Figures 17 and 18).

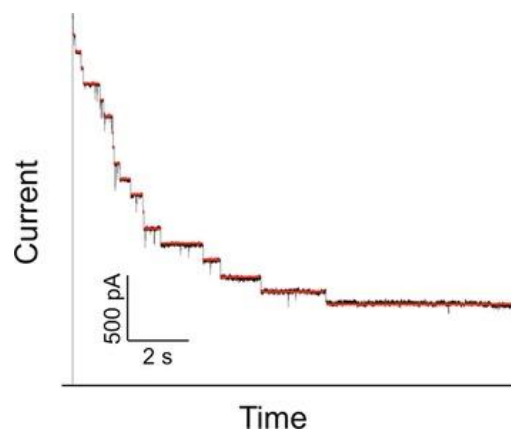


Figure 9. Incremental current decreases induced by sequential, irreversible blocking of individual α -HL pores by crosslinked duplexes in an experiment with multiple channels embedded in the lipid bilayer. The mixture contained crosslinked duplex, un-crosslinked duplex, and single strands. The analysis was carried out at 120 mV in Tris buffer (10 mM, pH 7.4) containing KCl (1 M) at 22 °C. The trace shown was low-pass filtered at 1 kHz.

Our results introduce a new hybridization-induced, programmable crosslinking reaction that can be used for the sequence-specific covalent capture of nucleic acids. The probe–target complexes generated in this manner could be detected by typical fluorescence,^{41–43} colorimetric⁶⁸, or electrochemical methods;⁶⁹ However, we showed here that nanopore technology, combined with sequence-specific crosslinking chemistry, has the potential to provide a high contrast—in essence, digital—signal for single-molecule sensing of nucleic acid sequences.

Materials and General Procedures

Reagents were purchased from the following suppliers and were of the highest purity available: oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA). Uracil DNA glycosylase (UDG), and T4 DNA polynucleotide kinase (T4 PNK) were from New England Biolabs (Ipswich, MA). [γ 32P]-ATP (6000 Ci/mmol)

was purchased from PerkinElmer. C-18 Sep-Pak cartridges were purchased from Waters (Milford, MA), and BS Poly prep columns were obtained from BioRad (Hercules, CA). Acrylamide/bis-acrylamide 19:1 (40% Solution/Electrophoresis) was purchased from Fisher Scientific (Waltham, MA). Quantification of radioactivity in polyacrylamide gels was carried out using a Personal Molecular Imager (BIORAD) with Quantity One software (v.4.6.5). Preparation of Cross-Linked DNA Substrates. The complementary oligonucleotides for each duplex were annealed¹ at a 1:1 molar ratio and treated with the enzyme UDG (50 units/mL, final concentration) to generate Ap sites. The enzyme UDG was removed by phenol-chloroform extraction and the DNA ethanol precipitated and the pellet washed with 80% EtOH-water.¹ The resulting Ap-containing DNA duplexes were re-dissolved in a buffer composed of HEPES (50 mM, pH 7) containing NaCl (100 mM) and incubated at 37 °C for 120 h. Reaction mixtures were then analyzed in the nanopore experiments. In some cases, parallel denaturing polyacrylamide gel electrophoretic analysis of the cross-linking reaction mixture was carried out as previously described.⁴¹⁻⁴⁵ Briefly, the DNA was ethanol precipitated and 5'-³²P-labeled using standard procedures. After ³²P-labeling, the protein was removed by phenol- chloroform extraction and the sample was desalted by passage through sephadex G-25. The samples were then mixed with formamide loading buffer, loaded into the wells of a 20% denaturing polyacrylamide gel, and gel electrophoresed for 4 h at 1600 V. The amount of radiolabeled DNA in each band from the gel was measured by phosphorimager analysis. In these experiments, 0.5% yields of cross-linked DNA are easily detectable. This provides a discrimination factor of ≥ 160 for the second-generation probe shown in Figure 6B of the main manuscript (cross-link yield target / cross-link yield non-target). S3

Hydroxyl radical footprinting of the dA-Ap cross-linked duplex. We employed literature protocols to footprint cross-link duplex E.⁴⁶⁻⁵⁰ In this experiment, the strand opposing the Ap containing oligonucleotide was 5'-labeled using standard procedure.³⁷ Labeled DNA was annealed with the uracil-containing complement and treated with UDG to generate the abasic site as described above. The Ap-containing double stranded DNA (~400,000 cpm) was incubated in HEPES buffer (50 mM, pH 7) containing NaCl (100 mM) at 37 °C for 120 h. The DNA was ethanol precipitated, suspended in formamide loading buffer and separated on a 2-mm thick 20% denaturing polyacrylamide gel. The slow-forming cross-link duplex band was visualized using X-ray film, the band cut out of the gel, and the gel slice crushed, and the gel pieces were vortexed in elution buffer (NaCl, 200 mM; EDTA, 1 mM) at room temperature for at least 1 h. The mixture was filtered through a Poly-Prep column to remove gel fragments, and the residue was ethanol precipitated and re-dissolved in water and diluted with 2x oxidation buffer (10 µL of a solution composed of sodium phosphate, 20 mM, pH 7.2; NaCl, 20 mM sodium ascorbate, 2 mM; H₂O₂, 1 mM). To this mixture was added a solution of iron-EDTA (2 µL, EDTA, 70 mM; (NH₄)₂Fe(SO₄)₂·6H₂O, 70 mM) to start the reaction, and the mixture vortexed briefly and incubated at room temperature for 5 min before addition of thiourea stop solution (10 µL of a 100mM solution in water). Hydroxyl radical footprinting reactions, Maxam-Gilbert G reactions, and Maxam-Gilbert A+G reactions were performed on the labeled duplex to generate marker lanes.⁶ The resulting DNA fragments were analyzed using gel electrophoresis as described above. Electrophysiology measurements. A membrane of 1,2-diphytanoyl-sn-glycero-3-phosphocholine was formed on a small orifice of approximately 150 µm diameter in a Teflon partition that separates two identical Teflon

chambers. Each chamber contained 2 mL of electrolyte solution (1 M KCl, 10 mM Tris-HCl, pH 7.4). Less than 1 μ L of α -hemolysin was added to the cis chamber with stirring, after which, a conductance increase indicated the formation of a single channel. For multichannel recording, 2 to 5 μ L of α -hemolysin was added. The ionic current through the α -hemolysin protein nanopore was recorded by an Axopatch 200B amplifier (Molecular Devices Inc., Sunnyvale, CA), filtered with a built-in 4-pole low-pass Bessel Filter at 5 kHz, and finally acquired into the computer using a DigiData 1440A A/D converter (Molecular Devices) at a sampling rate of 20 kHz. All the data recording and acquisition including single channel, S4 multichannel and persistent blocking recording of DNA cross-links were controlled through a Clampex program (Molecular Devices) and the analysis of nanopore current traces was performed using Clampfit software 10.4 (Molecular Devices).

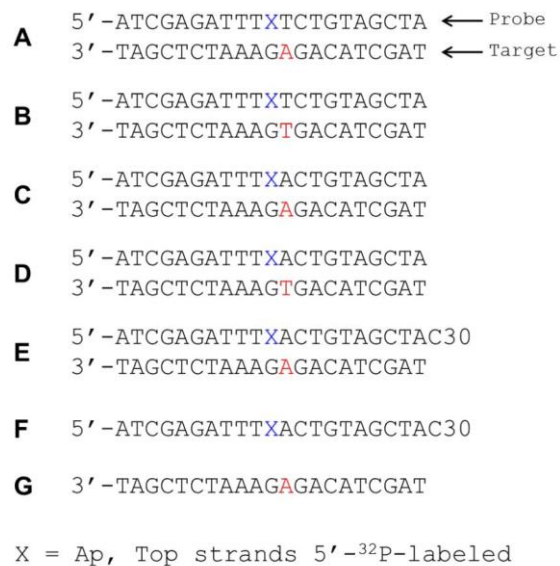


Figure 10. DNA duplexes used in these studies.

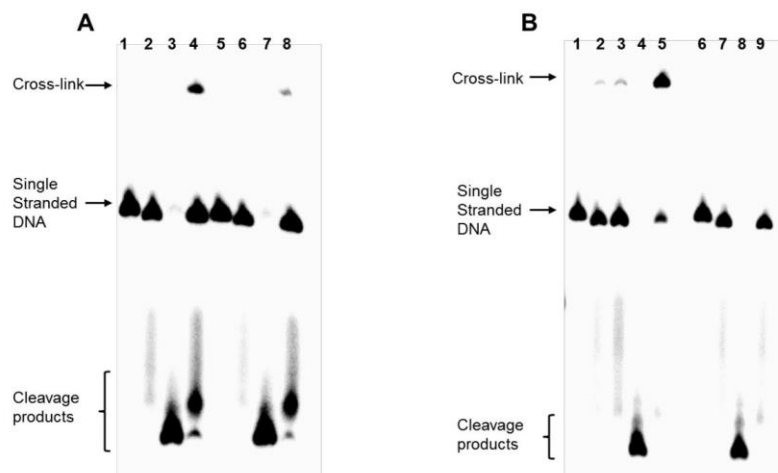


Figure 11 . Ap-containing probes selectively cross-link with the 1799 T→A mutant BRAF kinase gene sequence. Panel A: Gel electrophoretic analysis of cross-link formation between first generation Ap-containing probe in a 21nt duplex and, panel B: cross-linking by second generation Ap-probe that introduces a mismatch adjacent to the Ap residue. The middle bands correspond to the ^{32}P -labeled full length labeled 2'-deoxyoligonucleotides and the upper bands cross-linked DNA. Ap sites were generated by treatment of the corresponding 2'-deoxyuridinecontaining duplex with UDG. The Ap-containing duplexes were incubated in HEPES buffer (50 mM, pH 7 containing 100 mM NaCl) at 37 °C. After 120 h, the loading dye was added to the reaction mixture for gel analysis. The ^{32}P -labeled 2'-deoxyoligonucleotides were resolved on a denaturing 20% polyacrylamide gel and the radioactivity in each band was quantitatively measured by phosphorimager analysis. Panel A: Lane 1 dU-containing duplex A, lane 2 Ap-containing duplex A, lanes 3 piperidine work up of duplex A, Lane 4 duplex A after 120 h incubation in the buffer, Lane 5 dU-containing duplex B, lane 6 Ap-containing duplex B, lanes 7 piperidine work up of duplex B, and Lane 8 duplex B after 120 h incubation in the buffer. Panel B: Lane 1 dU-containing duplex C, lane 2 Ap-containing duplex C, lanes 3 Ap-containing duplex C, lane 4 piperidine work up of duplex C, Lane 5 duplex C after 120 h incubation in the buffer, Lane 5 dU-containing duplex D, lane 6 Ap-containing duplex D, lanes 7 piperidine work up of duplex D, and Lane 8 duplex D after 120 h incubation in the buffer.

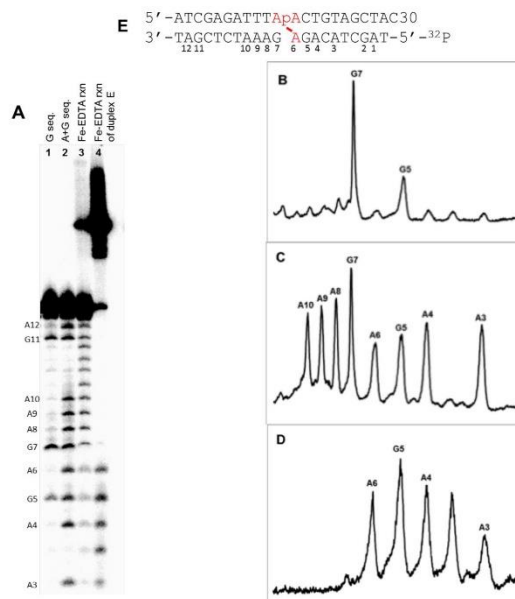


Figure 12. Iron-EDTA footprinting defines the cross-link location in duplex E. Panel A: Lane 1 is a Maxam-Gilbert G-lane of the labeled 2'-deoxyoligonucleotide strand in duplex E. Lane 2 is an A+G lane of the labeled 2'-deoxyoligonucleotide strand in duplex E. Lane 3 is the iron EDTA cleavage reaction on the labeled 2'-deoxyoligonucleotide duplex E. Lane 4 is the iron EDTA footprinting on the cross-linked duplex E. Panel B, C and D are densitometry traces of lanes 1, 2 and 4 on the sequencing gel (panel A) where each peak represents a band on the gel.

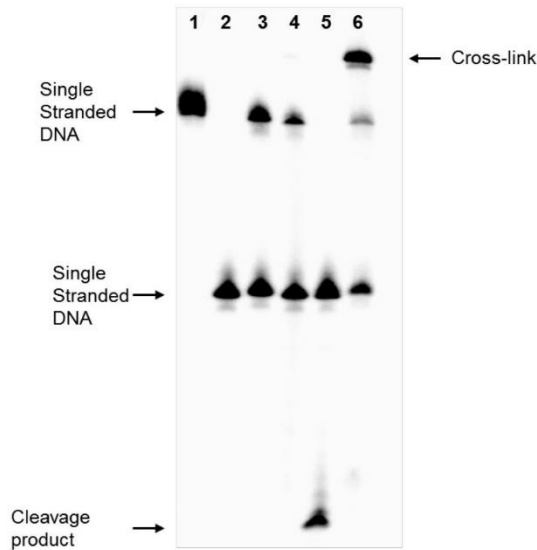


Figure 14. Quantitative detection of cross-linked duplex E prepared for nanopore experiment by gel analysis. After generation of dA-Ap cross-link for nanopore experiment under standard condition, the DNA was radiolabeled and analyzed by gel electrophoresis. The radioactivity in each band was quantitatively measured by phosphorimager analysis. Lane 1 dU-containing strand F, Lane 2 strand G, lane 3 dU-containing duplex E, lane 4 Ap-containing duplex E, lane 5 piperidine work up of duplex E, and Lane 6 Ap-containing duplex E after 120 h incubation in HEPES buffer (50 mM, pH 7 containing 100 mM NaCl) at 37 °C.

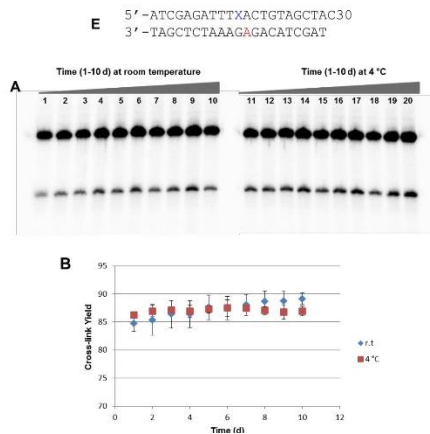


Figure 13. The effect of storage condition on the cross-link yield. After the cross-link generation using standard procedure, the dA-Ap cross-link was stored in 2 different conditions, room temperature (r.t) and 4 °C for 10 days. At specified time points, aliquots were removed and frozen at -20 °C before gel analysis. The ³²P-labeled 2'-deoxyoligonucleotides were resolved on a denaturing 20% polyacrylamide gel and the radioactivity in each band was quantitatively measured by phosphorimager analysis



Figure 16. Continuous recording of the block by cross-linked mutant target/probe duplex E for 30 min at +120 mV, in 1M KCl, 10 mM Tris, pH 7.4 at 22 °C. The result demonstrates the permanent trapping and current blocking of the cross-linked DNA duplex in the nanopore.

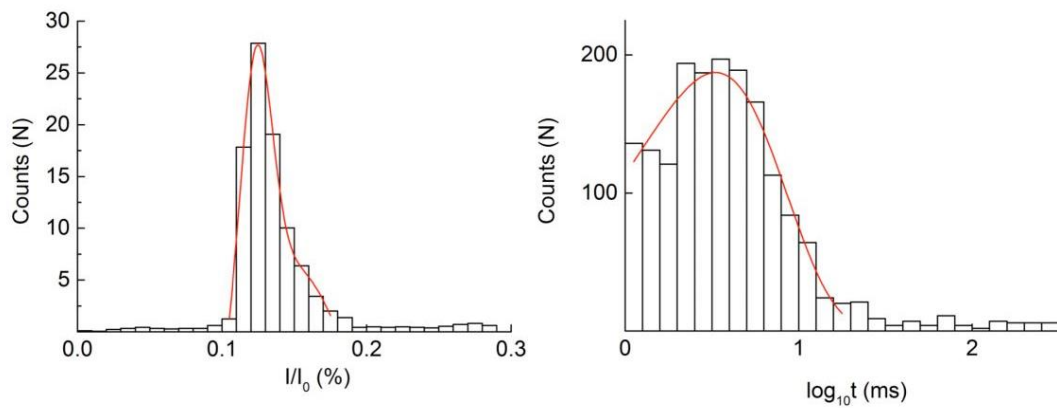


Figure 15. Histograms showing the current-blocking levels (left) and dwell times (right) for the uncross-linked duplex F.

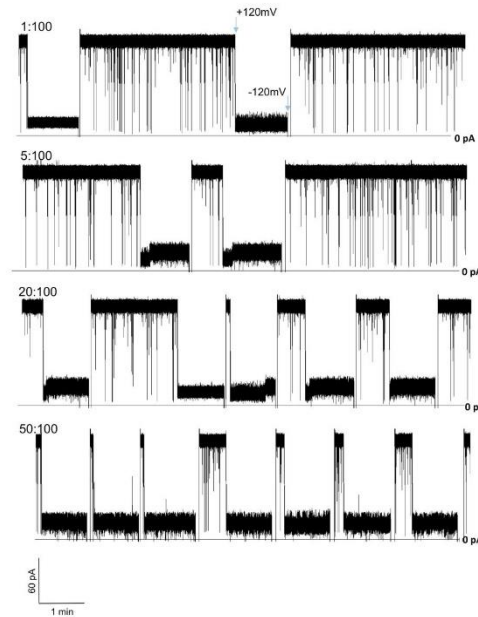


Figure 18. Current traces for mixtures of cross-linked duplex E and uncross-linked duplex F. These are single-channel experiments recorded at +120 mV, for 10 min in 1M KCl, 10 mM Tris, pH 7.4 at 22 °C. When irreversible trapping of a cross-linked duplex was recorded, the voltage polarity was reversed to -120 mV to clear the nanopore and then reset to +120 mV to resume sampling the nucleic acids in the mixture. Samples containing larger fractions of cross-linked duplex (lower traces) yield more frequent irreversible current blocks.

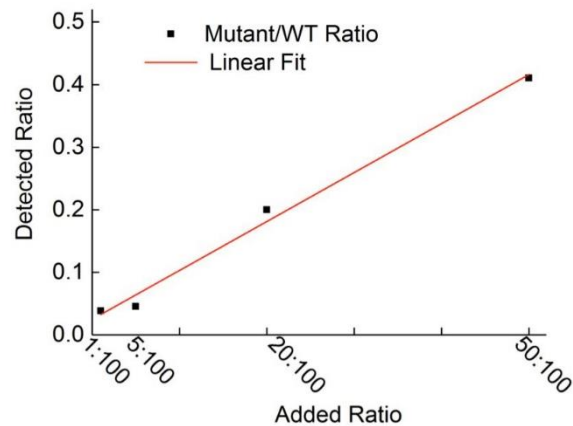


Figure 17. Plot of detected ratio of persistent current blocking events versus short-duration current blocks as a function of the ratio of cross-linked duplex E: uncross-linked duplex F in sample (based on data shown in Fig. S8).

CHAPTER 3. AN INTERNAL RNA BARCODE STRATEGY FOR LABEL-FREE NANOPORE MULTIPLEX SNP DISCRIMINATION

Abstract:

Single Nucleotide Polymorphisms (SNPs) are common yet significant biomarkers for precision medicine, which contribute collectively in disease development. In clinical detection, multiple SNPs should be analyzed in a timely and sensitive manner. Here, we have developed a label-free nanopore protocol for multiple SNP detections. By employing certain types of RNA as internal barcodes and the help of the nanolock, probes with chimeric RNA homopolymers can generate distinctive step patterns linking to individual SNPs. Those stepwise dehybridization kinetics can serve as a qualitative nanopore SNP identifier to unmistakably spot various SNP types with great precision and ease, no complicated statistical analysis is necessary. Paired with automated protocol and asymmetrical configurations, this nanopore detection mechanism can be further modified to meet the needs for the rapid and easy-to-use clinical investigations. We have examined the current configuration with KRAS G12D and Tp53 R172H mutants mixed within the same sample. Expectedly, sufficient number of signals from both mutated targets can be recognized and picked up within 20 minutes. Theoretically, in the future, such a nanolock-based approach can be adapted and applied to a wide range of disease panels with various SNP combinations, further help advancing diagnosis and treatment for personal medicine.

Nanopore, Barcode, SNP detection, nanolock, single molecule detection, multiplex detection, BRAF, KRAS, Precision Medicine

Introduction

The merit of nanopore approach comes from the distinctive electric patterns generated when a single molecule translocates through the pore (constriction site). Up to date, numerous efforts have been made to further advance this manner (to suit various detection schemes) into an all-round ultrasensitive detection platform^{11,12,35,70-80}; including SNP detections^{13,14,81,82}. Single nucleotide polymorphisms (SNPs) are common (estimated 1 in every 1kb^{83,84}) site-specific single nucleotide alterations across the genome. SNP clustering not only leads to complex diseases (e.g. CVDS and cancers⁸⁵, but also governs individual disease susceptibilities and treatment efficiencies⁸⁶. Thus, due to the sequence-specific and abundance nature of SNPs, the corresponding detection methods should be both high-throughput and sensitive. Currently, mainstream techniques such as sequencing⁸⁷, enzyme-based⁸⁸ and hybridization-based methods⁸⁹ can be high-yield, but most of them suffer from low sensitivity and have to rely on quantitative and laborious approaches in the determination. Nanopore sensing, on the contrary, is capable of low-limit detections but not proficient in multiplex screening. To date, efforts have been reported using nanopore for multiplex sensing, however, most of them are either not suitable for SNP screening, or need extra labeling steps.^{13,14,36,81,90-96} Due to their minimal sequence variances, SNPs cannot be discriminated from each other by normal detection signatures (e.g., peak shift). This calls for new strategies for nanopore multiplex detections.

Here, an integrated barcode design is first introduced and tested. A chimeric nucleotide capture probe has been designed to incorporate a trunk of barcode-acting RNA sequence to label different SNPs in the sample. The distinctive ionic pattern is generated by the intrinsic level difference between RNA barcode and DNA in the pore, and the overall dwell time is prolonged by the Hg-DNA base pair interactions (nanolock). With rationally designed chimeric probes, at least two different SNPs could be recognized simultaneously without further analysis.

Results and discussions

The general idea of this approach is to assign a unique “barcode” to each SNP types in samples, so multiple mutants can be screened and sorted simultaneously based on barcode features. (e.g. level differences) Previous studies^{36,95} have already shown that probes labelled with polymers, such as PEGs, can be utilized for nanopore multiplex miRNA detection. However, those existing “external” labeling methods demand additional procedures for label attachment, which may undermine yield and waste labor. RNA homopolymers, on the other hand, can serve as an “internal” barcode for nanopores. First, qualified as a barcode, RNA homopolymers can generate contrasting signals from their DNA counterparts in a sequence specific manner, which means that characteristic barcode signals are adjustable by the sequence composition and length. Second, they can be integrated with DNA sequence as DNA-RNA chimeras, which are comparatively convenient and commercially available, leading to great extra labor saving. Given the

relatively simple reaction in labeling the PEG barcodes by click chemistry, at least an hour is still needed to complete the labelling procedure.

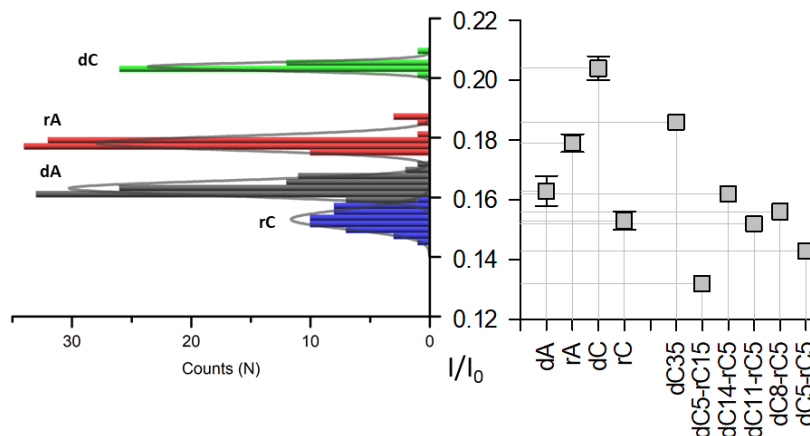


Figure 19. (Left) Validation histogram indicating the level differences of rA , rC and their DNA counterparts. A biotin conjugated chimeric probe thread through the pore from 3’ end and immobilized at the RNA site. I_{dA}/I_0 : (16.30±0.24)%; I_{rA}/I_0 : (17.81±0.23)%; I_{dC}/I_0 : (20.38±0.14)%; I_{rC}/I_0 : (15.26±0.33)%. All data are presented in (mean ± S.D.)%. Histograms were constructed with a Bin=0.002 and fitted to the Gaussian distribution. (Right) Comparisons with previous result, which entered the pore from 5’ end. **Notice that the I/I_0 varies with the length of the RNA.**

However, determining the barcode is far from enough, its signal should also be sufficiently “visible” to detection. It is widely recognized that single-stranded oligonucleotides translocate through the nanopore at a speed which is too high to distinguish the sequence details.⁹⁷ Thus, for a single-stranded barcode to provide “useful” label information, the translocation velocity must be reduced. Existing study from Wang et. al¹³ provides a feasible solution featuring T-Hg-T (Thymine-Hg-Thymine) nanolocks. The non-covalent interaction between Thymidine-Thymidine mismatch and an Hg²⁺ ion stabilizes the duplex specifically at the mismatch site and prolongs the dwell time. In this

case, a T:T nanolock can be rationally implemented as a “break” to introduce a long enough pause for steady barcode readouts.

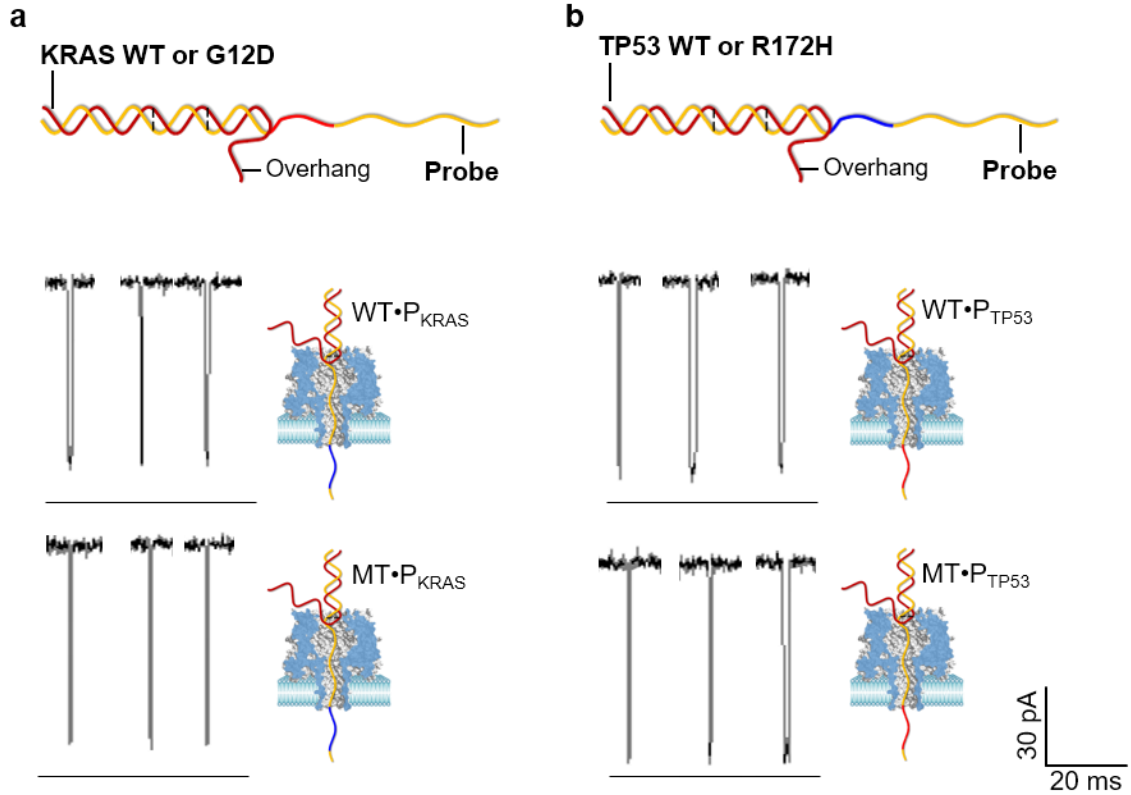


Figure 20. MT•P duplexes generate unique step patterns that not only can discriminate wild type and mutant, but also different mutant types. (a,b) Cartoons depicting the general structure of WT/MT•P duplexes. All duplexes can form at least one nanolock (Break) with a corresponding leading step (S1, c, d, e, f). Duplexes with the correct SNPs can form a second nanolock (Mutant Identifier) and develop a lagging step (S2, e, f). Sequence differences between barcodes lead to unique comparisons between S1 and S2. (c,d) With the presence of Hg^{2+} , wild type duplex show one step pattern. (e, f) On the contrary, mutated duplex show signature step patterns, either a downward pattern (KRAS G12D) or upward pattern (Tp53 R172H).

Therefore, in order to meet those requirements, the corresponding probe be designed to contain a pre-defined RNA homopolymer as an internal barcode (see Table 1). This barcode is flanked by an SNP capture sequence and a 3'-poly(dA)₁₅ lead. Poly dA is believed to have less secondary structures and produce a “cleaner” recording trace

(unpublished data). The capture sequence has been designed to (1) anneal & capture SNP containing /WT target sequences and form T:T mismatches; (2) establish a T:T nanolock (#1 pause, barcode amplifier) to ensure a prolonged RNA barcode coverage at the constriction site, which produces the leading step S1; and (3) generate a second T:T nanolock (#2 pause, mutant identifier) at the SNP site that produces the lagging step S2. Since random sequence will be residing at the constriction site, S2 will hold a different level in contrast to S1, hence forming a unique step pattern.

Additionally, unique overhangs were introduced to enable rapid duplex dehybridization outside the pore. According to ding et.al⁹⁸, hairpins with overhangs being longer than 12nt tend to unzip themselves outside the pore, and consequently produce a burst-like, shortened dwell pattern. Therefore, the overhang on either side of the duplex should at least longer than 12nt. Overall, Such a bi-overhang design may result in the effective elimination of the extra dwell time brought by the perfectly matched duplex used by Wang et.al¹³ and drastically enhance the dwell time comparison after nanolock formation (Table 1&2). Probes with different barcodes and capture arms were constructed to hybridize and distinguish multiple SNP/WT target sequences.

Since the present basic design is based on the hypothesis that RNA and DNA behaves differently inside the pore in a sequence-specific manner, additional experiments were needed for confirmation. To validate this, we first need to know how well sequence the

specific RNA homopolymers can be separated inside the pore. We first constructed a set of streptavidin-biotin conjugated nucleotides that contain either a 15nt barcode sequence (RNA) or a DNA counterpart (see Table 1&2). The probe was immobilized in the pore upon 3' entry and the level differences were revealed. Upon all the nucleotides have been tested (dA vs. rA; dC vs. rC) (Fig19), we found that dC always occupied a higher level than rC, ($\Delta I_{RES}^{\text{poly(dC)-poly(rC)}}=5.1\%$), while dA and rA tended to have a reverted situation, with dA slightly below rA ($\Delta I_{RES}^{\text{poly(dA)-poly(rA)}}=-1.5\%$). In terms of sequence differences within one nucleic acid type (either DNA or RNA), a moderate difference was found between either rC vs rA or dC vs dA ($\Delta I_{RES}^{\text{poly(rA)-poly(rC)}}=2.6\%$, $\Delta I_{RES}^{\text{poly(dA)-poly(dC)}}=-4.1\%$). The level displacements are in approximate agreement with Butler et.al^{99,100}, Akeson et.al¹⁰¹ and Purnell et.al¹⁰², but in contradictory to Stoddart et.al¹⁰ under similar conditions. This finding suggests the feasibility to use integrated RNA barcode to “label” different probes in a multiplex detection scenario.

We constructed two probes with different RNA barcodes integrated. The first probe (P_{KRAS}) targets an antisense KRAS fragment with a G12D (G-to-A) mutation, in which a 15nt poly(rA)₁₀(rG)(rA)₄ is integrated as the barcode. The second probe (P_{TP53}) targets an antisense R172H (G-to-A) carrying Tp53 fragment with a poly(rA) barcode. (rG and rC were introduced to reduce complementary structure between overhangs). The reason we choose rA and rC as barcodes is that they could be fairly distinguished from other nucleic acids (RNA or DNA), which was also confirmed by other similar studies^{99,100,102}.

When the probe-target complexes approach the pore from the *cis* opening, the longer probe overhang will thread through the constriction site first and initialize the unzipping outside the pore. (Fig 20 c. d. e. f.) The shorter target overhang will be protruding outside the pore as a “support”. If no Hg^{2+} was present, the fast and transient unzipping process would produce burst-like translocation patterns (Fig 22)⁹⁸. However, an Hg^{2+} overdose would trigger nanolock formations and produce halt(s) during unzipping. In the case of WT•P complex, only a single nanolock (barcode amplifier) would form, causing a pause when the RNA barcode occupied the construction site. Subsequently, only a unique barcode level would be revealed (Fig 20 c&d). However, for MT•P duplexes carrying two nanolocks in the sequence, a second pause could be observed due to the incorporation of the second nanolock (mutant identifier). The constriction site would be covered by random DNA sequences between the first and second nanolock, producing an altered, “generalized” DNA level different from the first barcode level (Fig 20 e&f).

With a proper amount of Hg^{2+} in the solution in our experiment, the dwell pattern changed sharply from the burst-like pattern (dwell time) (Fig 21) to the step pattern that is 100X prolonged. Typical signs of WT•P and MT•P complexes are shown in Fig 20. The WT•P signal displays a single step pattern. The most significant indicator for an MT•P is the two-step pattern, due to the additional nanolock at the SNP site. On the contrary, WT•P and other semi nanolocked (with only one Hg^{2+} bounded) duplexes will only display a one-step pattern, which can be well distinguished from the two-step pattern (Fig 22).

However, the successful multiplex detection necessitates different probe-containing duplexes having distinctive and comparable patterns. In the present case, the poly rA barcodes in the MT•P_{KRAS} duplexes produce a “downward” step pattern, with a leading step (S1) slightly higher than that of the lagging step (S2). Comparatively, the poly rC barcodes produce an “upward” step pattern, indicating that an MT•P_{Tp53} duplex has de-hybridized through the pore. Those patterns can be well recognized without any level of analysis (Fig 22). In a 1:1 mixed situation (KRAS MT•P_{KRAS}: MT•P_{Tp53} = 1:1, 1M KCl 10mM Tris DEPC), those two different patterns could still be distinguished and their corresponding counts aligned well with the theoretical ratio.

We also swapped the barcodes between two probes. To our surprise, the swapped barcodes no longer produce distinctive step patterns, which, however, can be well-understood: rA always occupied a slightly higher level than that of rC (Fig 19), thus, if rC was to replace rA as barcode, there would be a large possibility that the slightly higher level from rA become lower (rC) and consequently indistinguishable with the lagging step (S2) from the random DNA sequence. Vice Versa, the lower level of rC (Fig 19) will also “blend in” with the lagging step(S2) and appear as a single step pattern.

Further Improvements:

The real application of the nanopore multiplex detection requires a much faster analysis. To improve this multiplex system, sample capture rates must be improved. Wanunu et

al.¹⁰³ reportedly increased double stranded DNA capture rates by establishing a salt gradient across a solid state nanopore. Likewise, an asymmetrical KCl solution (*cis:trans*=0.5M:3M) setup was used to promote the target pattern occurrences (lower the time for recording). The asymmetrical solution setup produced much faster capture rate with still recognizable step patterns. However, the frequency of blocking unwanted events also sees a remarkable increase. In this scenario, the system can be further improved with a simple built-in episodic protocol. With this protocol over-long blocking events can be well circumvented by automatically inverting voltage after a fixed period (Fig 24). In a typical run (15 min), roughly 400 inverts were introduced and approximately 200 distinctive step patterns can be well-recorded. In a typical nanopore detection, 150-200 events will be large enough for fitting a histogram with a 95% confidence level. (unpublished data). In view of this, one can collect almost all the required data in just 2-3 runs (run numbers may vary based on specific scenarios).

We also utilized this protocol to rapidly generate a standard ratio curve overnight (~4 hrs), ranging from MT•P_{KRAS}: MT•P_{Tp53}= 0 to 1:1. Further analysis of the dwell time of each step from the step patterns demonstrated some interesting results. Fitting of the dwell logarithmic histogram shows varied biases towards a particular step. For example, KRAS G12D•P rA duplexes produce a much longer S2 than S1, ($\tau_{\text{off}}=150.55\text{ms}$ (S2) Vs. 68.72ms (S1), S.D.) However, in Tp53, S1 is longer than S2 ($\tau_{\text{off}}=64.58\text{ms}$ (S2) Vs. 196.43ms (S1), S.D.) (Fig 23) This is due to the duplex length differences behind the nanolock. Generally, the longer the sequence is, the longer it will take for the sequence after the nanolock to disassociate¹⁴.

Discussion:

In all, this work provides a novel direct approach that shows the promising potential for extended identification of multiple known SNPs in a mixture. Integrating the T-Hg-T interaction motif reported by Wang et. al¹³ with the integrated ssRNA barcode design, we have succeeded in further clarifying the detection mechanism, and more importantly, making the approach be capable of simultaneous SNP detection in a given sample with ease: The RNA integrated DNA probe can be ordered through company and used in a plug-and-play manner, which eliminates the laborious labeling procedures as used in previous studies^{14,36,81}; The multiple step events are only limited to the SNP containing duplexes, thus, false positive results can be largely minimized; The unique trends (features?) of the step patterns are closely related to the corresponding SNP types, so they can be distinguished without the need of further statistical analysis. The asymmetrical solution setup and episodic protocol demonstrate the large possibility to further upgrade this method to automate YES/NO detection in the future. Besides, the RNA barcode types can be further explored rationally with varied compositions and lengths to establish various detection combinations. Theoretically, for RNA barcode with a fixed length, three types of barcodes (Poly A, T and C) can be utilized to generate varied levels of S1. (Poly G is not recommended because of possible unexpected G-quadruplex formation) Additionally, difference in RNA barcode length (5, 10, and 15 mer) will also affect S1 levels (unpublished data). Therefore, in theory, with the correct combinations between

poly nucleotide type and length, almost (approximately) ten different SNPs can be distinguished from the mix.

In this study, the T-Hg-T nanolock is capable of detecting X-to-A/T mutations. This system can also be tailored for X-to-C/G mutations: C-Ag-C interaction is a similar nanolock to T-Hg-T that can be directly incorporated into the probe. It is believed that only a few bases need to be altered for detecting X-to-C/G substitutions such as MTHFR Prostate Cancer-derived mutation A1298C.

Nevertheless, though highly promising and easy to operate, care should be taken in conducting the detections following the present design. First, as RNA barcodes are relatively fragile, our data indicated possible degradations and the loss of distinctive patterns after one-month storage at -20 °C, and frequent freezing-thawing cycles would deprive RNA concentrations furthermore, which means the limited storage duration and freezing-thawing cycles of the RNA barcodes. In addition, newly received probes should be handled with care and subjected to detections immediately.

It is expected that the present study can be more complete by testing Poly rT in the scheme, and the built-in protocol should be more efficient to handle extremely high number of blocking events in the asymmetrical setup. Additional approaches such as

auto-blocking detection, voltage reverting mechanisms and AI (artificial intelligence) image recognition could be integrated as further improvements.

In 2000, Deamer et.al⁹ first proposed the concept of “target molecular barcode” (TMBC) that exploits unique nanopore signals from series of DNA segments to mark designated agents. To our knowledge, our internal barcode system can be categorized as one type of TMBCs. Overall, these internally-integrated RNA barcode systems, together with DNA-metal base pair-based interactions, will ultimately permit the nanopore to detect several mutations simultaneously, moving us a little bit closer for clinical parallel diagnosis and personalized therapies.

Material and methods

1) Sample preparation (duplex denaturing & annealing)

All chemicals, including KCl, 3-(N-morpholino) propanesulfonic acid (MOPS) and DEPC were purchased from Sigma-Aldrich (St. Louis, MO) unless stated otherwise, DNA-RNA chimeric probe and its partially complementary target DNA was obtained from IDT (Integrated DNA Technologies). Commonly, Equal molar DNA-RNA chimeric probe and target DNA is mixed under appropriate salt conditions. The mixture is then subjected to 95°C denaturing for 3-5mins and followed by a slow cool-down process.

2) Single channel recording


Briefly, a small pre-drilled orifice in a Teflon sheet accommodated the formation of a diphytanoyl-sn-glycerol-3-phosphocholine membrane, which electrically barricaded two symmetrical Teflon chambers (*cis* and *trans*) from the middle. 2ml electrolytes were introduced into both chambers as needed. In symmetrical setups, both chambers were filled with recording solution containing 1 M KCl, 25 mM MOPS, pH 7.4, DEPC; In asymmetrical setups, *cis* chamber solution contains 1 M KCl, 25 mM MOPS, pH 7.4, DEPC and *trans* solution contains 3 M KCl, 25 mM MOPS, pH 7.4, DEPC in *trans*). Voltage was applied from the *trans* chamber. 1µl α-hemolysin stock solution was added at near-membrane location and a single channel formation could be observed by a stepwise conductance increase. Approximately 50 µM HgCl₂ was introduced into the *cis* chamber where 150nM sample was already premixed.

3) Data processing

An Axopatch 200B was applied to record and amplify the pico-ampere current through the pore. The initial recordings were then filtered with a built-in 4-pole low-pass Bessel filter at 5 kHz, and finally captured and converted by DigiData 1440A A/D converter into the computer at 20 kHz sampling rate. The entire process was monitored and controlled through an on-board Clampex 10.4 system. Analysis of the recorded trace, including amplitude histogram analysis and duration histogram analysis, was performed using Clampfit 10.7. (All equipment and software used above were purchased from Molecular Devices Inc., Sunnyvale, CA, USA unless stated otherwise). Results were presented in mean \pm standard deviation. Experiments were performed at $25^{\circ}\text{C} \pm 1^{\circ}\text{C}$.

Table 1. Sequences of probes and targets used in the main study. The underlined letter indicates the mutation site. For wild-type KRAS, it is a G in the sense strand and a C in the antisense strand; for the mutated KRAS G12D, it is an A in the sense strand and a T in the antisense strand. Similarly, the wild-type Tp53 contains a G in the sense strand and a C in the antisense strand, and the mutated Tp53 R172H contains an A in the sense strand and a T in the antisense strand. The red letter shows the position where the nanolock will be formed. The rG among Poly rA barcode and rAs among poly rC barcode are designed to avoid unwanted base pairing while forming the duplex. Both RNA barcodes are painted in yellow.

SNP	Name	Sequence 5'→3'
KRAS G12D	P _{KRAS}	GTATAAACTTGTGGTGGTTGGAGCTG <u>T</u> TGGCGTAGGC <u>T</u> AGAGCGrArArArArArArArArArArGrArArArA AAAAAAAAAAAAAAAAAAAA
	WT _{KRAS}	AGCTGTATCGTCAAGGCGCTCT <u>T</u> GCCCTACGCCA <u>C</u> CAGCTCCAACCACCACAAGTTTATAC
	MT _{KRAS}	AGCTGTATCGTCAAGGCGCTCT <u>T</u> GCCCTACGCCA <u>T</u> CAGCTCCAACCACCACAAGTTTATAC
Tp53 R172H	P _{Tp53}	ACGGAGGTCGTGAGAC <u>T</u> CTGCCCCACCATG <u>T</u> GCGCTGrArCrArArCrCrCrCrCrCrCrCrCrCrCrCrC CAAAAAAAAAAAAAAAAAA
	WT _{Tp53}	CATCACCATCGGAGCAGCGC <u>T</u> CATGGTGGGGGCAG <u>C</u> GTCTCAGACCTCCGT
	MT _{Tp53}	CATCACCATCGGAGCAGCGC <u>T</u> CATGGTGGGGGCAG <u>T</u> GTCTCAGACCTCCGT

Table 3. Duplexes illustrated in complimentary form. The sequence used in the main study in duplex form. underlined letter indicates the mutation site. The red letter shows the position where the nanolock will be formed.  represents a nanolock.

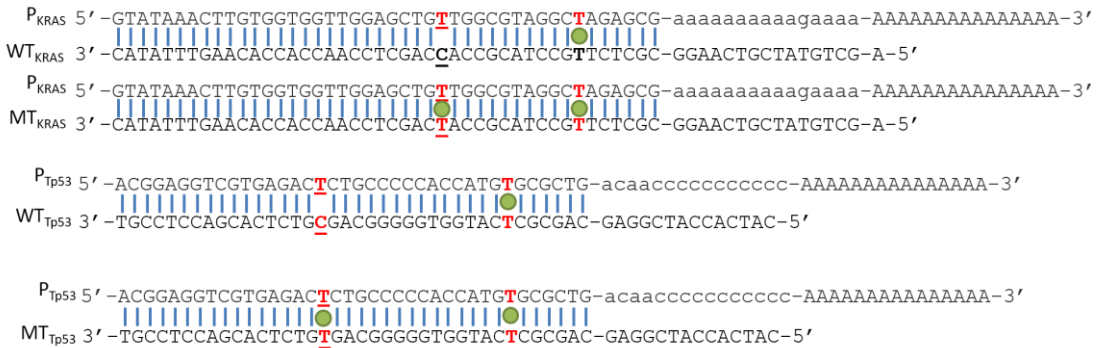


Table 2. Sequences used for validation. Oligos used for RNA and DNA level validation. All sequences have 5' biotin modification to react with Streptavidin in the solution. The strong interaction between Biotin-Streptavidin immobilizes the sequence and generate a steady readout

Name	Sequence 5'→3'
Chimera rC15	/5BiosG/CATCArCrCrCrCrCrCrCrCrCrCrCrCrCrCrCrCACCCCCCCCCCCCC
Chimera dC15	/5BiosG/CATCACCCCCCCCCCCCCCACCCCCCCCCCCCC
Chimera rA15	/5BiosG/CATCArArArArArArArArArArArArArArArAACCCCCCCCCCCC
Chimera dA15	/5BiosG/CATCAAAAAAAAAAAAAAAAAACCCCCCCCCCCCC

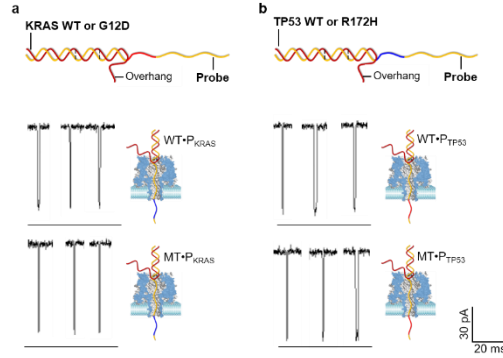


Figure 22 . In the absence of Hg^{2+} , both MT•P and WT•P duplexes generate spike events indistinguishable with other types of events, such as ssDNA translocation.

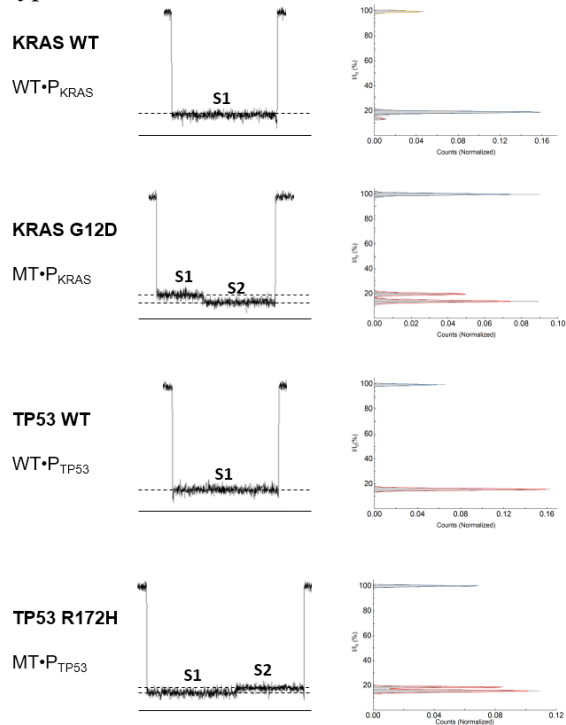


Figure 21. A close look at the pattern difference between WT•P and MT•P duplexes. The corresponding level histograms are displayed on the right. Notice single step pattern generated by the WT•P is on the same level with the leading step (S1) by MT•P, which implies 1) the step pattern is in close correspondence with the location of the two nanolocks; 2) the S1 pattern in WT•P is the same S1 step pattern in MT•P generated by the RNA barcode.

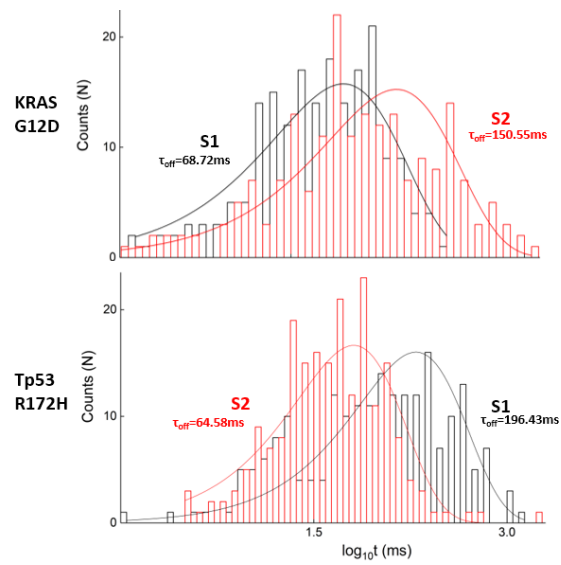


Figure 23. Histograms showing S1 and S2 dwell time differences between KRAS G12D and TP53 R172H duplex. Histograms were fitted to a log probability exponential distribution.

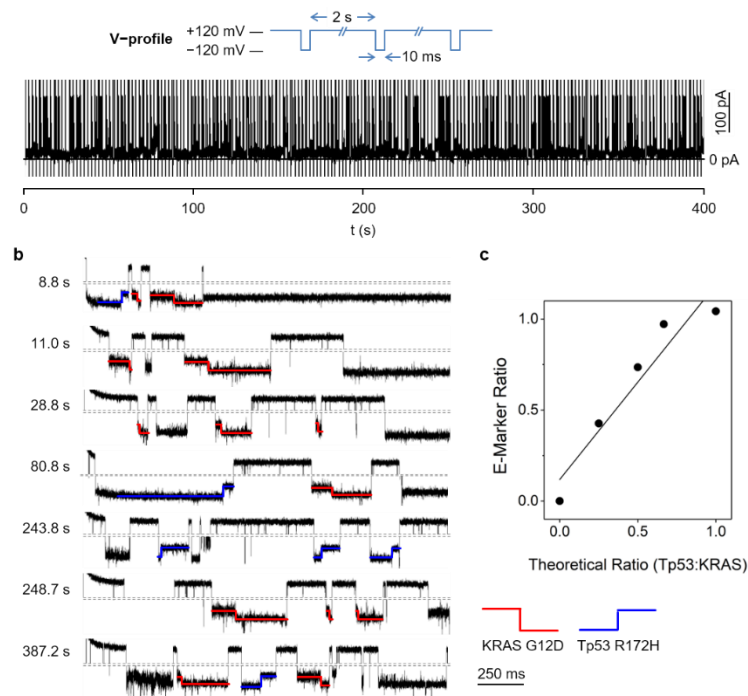


Figure 24. Traces recorded using the automated protocol. The protocol inverts voltage roughly every 2 seconds (a). (b) representative recordings truncated at 8.8s, 11.0s, 28.8s, 80.8s, 243.8s, 248.7s, 387.2s showing patterns generated by KRAS G12D (red) and Tp53 R172H (blue). (c) Plot of detected ratio of Tp53: KRAS versus the equivalent theoretical ratio of the duplexes.

References

- 1 Reed, G. H. & Wittwer, C. T. Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. *Clin Chem* **50**, 1748-1754, doi:10.1373/clinchem.2003.029751 (2004).
- 2 Dekker, C. Solid-state nanopores. *Nature Nanotechnology* **2**, 209-215, doi:<http://dx.doi.org/10.1038/nnano.2007.27> (2007).
- 3 Song, L. *et al.* Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science* **274**, 1859-1866 (1996).
- 4 Kawate, T. & Gouaux, E. Arresting and releasing Staphylococcal α -hemolysin at intermediate stages of pore formation by engineered disulfide bonds. *Protein Science : A Publication of the Protein Society* **12**, 997-1006 (2003).
- 5 Olson, R., Nariya, H., Yokota, K., Kamio, Y. & Gouaux, E. Crystal structure of Staphylococcal LukF delineates conformational changes accompanying formation of a transmembrane channel. *Nature Structural Biology* **6**, 134, doi:10.1038/5821 (1999).
- 6 Walker, B., Krishnasastri, M., Zorn, L. & Bayley, H. Assembly of the oligomeric membrane pore formed by Staphylococcal alpha-hemolysin examined by truncation mutagenesis. *The Journal of biological chemistry* **267**, 21782-21786 (1992).
- 7 Yamashita, D. *et al.* Molecular basis of transmembrane beta-barrel formation of staphylococcal pore-forming toxins. *Nature Communications* **5**, 4897, doi:10.1038/ncomms5897
<https://www.nature.com/articles/ncomms5897#supplementary-information> (2014).
- 8 Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 13770-13773 (1996).
- 9 Deamer, D. W. & Akeson, M. Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends in Biotechnology* **18**, 147-151, doi:[http://dx.doi.org/10.1016/S0167-7799\(00\)01426-8](http://dx.doi.org/10.1016/S0167-7799(00)01426-8) (2000).
- 10 Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G. & Bayley, H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7702-7707, doi:10.1073/pnas.0901054106 (2009).
- 11 Gu, L.-Q., Braha, O., Conlan, S., Cheley, S. & Bayley, H. Stochastic sensing of organic analytes by a pore-forming protein containing a molecular adapter. *Nature* **398**, 686, doi:10.1038/19491
<https://www.nature.com/articles/19491#supplementary-information> (1999).
- 12 Zhang, X. *et al.* Nanopore electric snapshots of an RNA tertiary folding pathway. *Nature Communications* **8**, 1458, doi:10.1038/s41467-017-01588-z (2017).
- 13 Wang, Y. *et al.* Nanolock–Nanopore Facilitated Digital Diagnostics of Cancer Driver Mutation in Tumor Tissue. *ACS Sensors* **2**, 975-981, doi:10.1021/acssensors.7b00235 (2017).
- 14 Zhang, X. *et al.* Characterization of Interstrand DNA–DNA Cross-Links Using the α -Hemolysin Protein Nanopore. *ACS Nano* **9**, 11812-11819, doi:10.1021/acsnano.5b03923 (2015).
- 15 Dilani A. Jayawardhana, J. A. C., Qitao Zhao, Daniel W. Armstrong, and Xiyun Guan*. Nanopore-stochastic-detection-of-a-liquid-explosive-component-and-sensitizers-using-boromycin-and-an-ionic-liquid-supporting-electrolyte_2009_Analytical-Chemistry. *Analytical-Chemistry* (2009).

- 16 Barreiro, L. B., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. *Nature genetics* **40**, 340-345, doi:10.1038/ng.78 (2008).
- 17 Saiki, R. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350-1354, doi:10.1126/science.2999980 (1985).
- 18 Wu, D. Y., Ugozzoli, L., Pal, B. K. & Wallace, R. B. Allele-specific enzymatic amplification of beta-globin genomic DNA for diagnosis of sickle cell anemia. *Proceedings of the National Academy of Sciences* **86**, 2757-2760, doi:10.1073/pnas.86.8.2757 (1989).
- 19 Eddy, S. R. The C-value paradox, junk DNA and ENCODE. *Current Biology* **22**, R898-R899, doi:10.1016/j.cub.2012.10.002 (2012).
- 20 Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 5294-5300, doi:10.1073/pnas.1221376110 (2013).
- 21 Palazzo, A. F. & Gregory, T. R. The Case for Junk DNA. *PLoS Genetics* **10**, e1004351, doi:10.1371/journal.pgen.1004351 (2014).
- 22 Graur, D. *et al.* On the Immortality of Television Sets: “Function” in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biology and Evolution* **5**, 578-590, doi:10.1093/gbe/evt028 (2013).
- 23 Karki, R., Pandya, D., Elston, R. C. & Ferlini, C. Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC Medical Genomics* **8**, 37, doi:10.1186/s12920-015-0115-z (2015).
- 24 Collins, D. W. & Jukes, T. H. Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence. *Genomics* **20**, 386-396, doi:<https://doi.org/10.1006/geno.1994.1192> (1994).
- 25 Keats, B. J. B. & Sherman, S. L. in *Emery and Rimoin's Principles and Practice of Medical Genetics* (eds David Rimoin, Reed Pyeritz, & Bruce Korf) 1-12 (Academic Press, 2013).
- 26 Norrgard, K., Schultz, J. Using SNP data to examine human phenotypic differences. *Nature Education* **1**, 85 ((2008)).
- 27 Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513, doi:10.1038/35035083 (2000).
- 28 Van Tassel, C. P. *et al.* SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5**, 247, doi:10.1038/nmeth.1185
<https://www.nature.com/articles/nmeth.1185#supplementary-information> (2008).
- 29 Howell, W. M., Jobs, M., Gyllensten, U. & Brookes, A. J. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. *Nat Biotechnol* **17**, 87-88, doi:10.1038/5270 (1999).
- 30 Newton, C. R. *et al.* Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Research* **17**, 2503-2516, doi:10.1093/nar/17.7.2503 (1989).
- 31 Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**, 443, doi:10.1038/nrg2986 (2011).
- 32 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

- 33 Martin, E. R. *et al.* SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics* **26**, 2803-2810, doi:10.1093/bioinformatics/btq526 (2010).
- 34 Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, doi:10.1101/gr.129684.111 (2012).
- 35 Wang, G., Zhao, Q., Kang, X. & Guan, X. Probing Mercury(II)-DNA Interactions by Nanopore Stochastic Sensing. *The Journal of Physical Chemistry B* **117**, 4763-4769, doi:10.1021/jp309541h (2013).
- 36 Zhang, X., Wang, Y., Fricke, B. L. & Gu, L.-Q. Programming Nanopore Ion Flow for Encoded Multiplex MicroRNA Detection. *ACS Nano* **8**, 3444-3450, doi:10.1021/nn406339n (2014).
- 37 Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187, doi:10.1038/nature09792
<https://www.nature.com/articles/nature09792#supplementary-information> (2011).
- 38 McCarthy, J. J., McLeod, H. L. & Ginsburg, G. S. Genomic Medicine: A Decade of Successes, Challenges, and Opportunities. *Science Translational Medicine* **5**, 189sr184-189sr184, doi:10.1126/scitranslmed.3005785 (2013).
- 39 Gibbs, P. E. M. & Lawrence, C. W. Novel Mutagenic Properties of Abasic Sites in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **251**, 229-236, doi:<https://doi.org/10.1006/jmbi.1995.0430> (1995).
- 40 M, G. K. *et al.* The Pharmacogenetics Research Network: From SNP Discovery to Clinical Drug Response. *Clinical Pharmacology & Therapeutics* **81**, 328-345, doi:10.1038/sj.clpt.6100087 (2007).
- 41 Hu, L. *et al.* Fluorescence in situ hybridization (FISH): an increasingly demanded tool for biomarker research and personalized medicine. *Biomarker Research* **2**, 3, doi:10.1186/2050-7771-2-3 (2014).
- 42 Silverman, A. P. & Kool, E. T. Detecting RNA and DNA with Templated Chemical Reactions. *Chemical Reviews* **106**, 3775-3789, doi:10.1021/cr050057+ (2006).
- 43 Silverman, A. P. & Kool, E. T. in *Advances in Clinical Chemistry* Vol. 43 79-115 (Elsevier, 2007).
- 44 French, C. *et al.* Detection of the Factor V Leiden Mutation by a Modified Photo-Cross-Linking Oligonucleotide Hybridization Assay. *Clinical Chemistry* **50**, 296-305, doi:10.1373/clinchem.2003.023556 (2004).
- 45 Viereg, J. R., Nelson, H. M., Stoltz, B. M. & Pierce, N. A. Selective Nucleic Acid Capture with Shielded Covalent Probes. *Journal of the American Chemical Society* **135**, 9691-9699, doi:10.1021/ja4009216 (2013).
- 46 Fujimoto, K., Yamada, A., Yoshimura, Y., Tsukaguchi, T. & Sakamoto, T. Details of the Ultrafast DNA Photo-Cross-Linking Reaction of 3-Cyanovinylcarbazole Nucleoside: Cis-Trans Isomeric Effect and the Application for SNP-Based Genotyping. *Journal of the American Chemical Society* **135**, 16161-16167, doi:10.1021/ja406965f (2013).
- 47 Nishimoto, A. *et al.* 4-vinyl-substituted pyrimidine nucleosides exhibit the efficient and selective formation of interstrand cross-links with RNA and duplex DNA. *Nucleic Acids Research* **41**, 6774-6781, doi:10.1093/nar/gkt197 (2013).
- 48 Peng, X. & Greenberg, M. M. Facile SNP detection using bifunctional, cross-linking oligonucleotide probes. *Nucleic Acids Research* **36**, e31-e31, doi:10.1093/nar/gkn052 (2008).
- 49 Price, N. E., Catalano, M. J., Liu, S., Wang, Y. & Gates, K. S. Chemical and structural characterization of interstrand cross-links formed between abasic sites and adenine residues in duplex DNA. *Nucleic Acids Research* **43**, 3434-3441, doi:10.1093/nar/gkv174 (2015).

- 50 Price, N. E. *et al.* Interstrand DNA–DNA Cross-Link Formation Between Adenine Residues and Abasic Sites in Duplex DNA. *Journal of the American Chemical Society* **136**, 3483-3490, doi:10.1021/ja410969x (2014).
- 51 Yang, Z., Price, N. E., Johnson, K. M. & Gates, K. S. Characterization of Interstrand DNA–DNA Cross-Links Derived from Abasic Sites Using Bacteriophage ϕ 29 DNA Polymerase. *Biochemistry* **54**, 4259-4266, doi:10.1021/acs.biochem.5b00482 (2015).
- 52 Knez, K., Spasic, D., Janssen, K. P. F. & Lammertyn, J. Emerging technologies for hybridization based single nucleotide polymorphism detection. *Analyst* **139**, 353-370, doi:10.1039/C3AN01436C (2014).
- 53 Shen, W., Tian, Y., Ran, T. & Gao, Z. Genotyping and quantification techniques for single-nucleotide polymorphisms. *TrAC Trends in Analytical Chemistry* **69**, 1-13, doi:<https://doi.org/10.1016/j.trac.2015.03.008> (2015).
- 54 Katsanis, S. H. & Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nature Reviews Genetics* **14**, 415, doi:10.1038/nrg3493 (2013).
- 55 Sun, C. *et al.* Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma. *Nature* **508**, 118, doi:10.1038/nature13121
<https://www.nature.com/articles/nature13121#supplementary-information> (2014).
- 56 Bollag, G. *et al.* Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nature Reviews Drug Discovery* **11**, 873, doi:10.1038/nrd3847 (2012).
- 57 Rossetti, G. *et al.* The structural impact of DNA mismatches. *Nucleic Acids Research* **43**, 4309-4321, doi:10.1093/nar/gkv254 (2015).
- 58 Johnson, K. M. *et al.* On the Formation and Properties of Interstrand DNA–DNA Cross-Links Forged by Reaction of an Abasic Site with the Opposing Guanine Residue of 5'-CAp Sequences in Duplex DNA. *Journal of the American Chemical Society* **135**, 1015-1025, doi:10.1021/ja308119q (2013).
- 59 Catalano, M. J. *et al.* Chemical Structure and Properties of Interstrand Cross-Links Formed by Reaction of Guanine Residues with Abasic Sites in Duplex DNA. *Journal of the American Chemical Society* **137**, 3933-3945, doi:10.1021/jacs.5b00669 (2015).
- 60 Dutta, S., Chowdhury, G. & Gates, K. S. Interstrand Cross-Links Generated by Abasic Sites in Duplex DNA. *Journal of the American Chemical Society* **129**, 1852-1853, doi:10.1021/ja067294u (2007).
- 61 Song, L. *et al.* Structure of Staphylococcal α -Hemolysin, a Heptameric Transmembrane Pore. *Science* **274**, 1859-1865, doi:10.1126/science.274.5294.1859 (1996).
- 62 Shi, W., Friedman, A. K. & Baker, L. A. Nanopore Sensing. *Analytical Chemistry* **89**, 157-188, doi:10.1021/acs.analchem.6b04260 (2017).
- 63 Wanunu, M. Nanopores: A journey towards DNA sequencing. *Physics of Life Reviews* **9**, 125-158, doi:<https://doi.org/10.1016/j.plrev.2012.05.010> (2012).
- 64 Kong, J., Zhu, J. & Keyser, U. F. Single molecule based SNP detection using designed DNA carriers and solid-state nanopores. *Chemical Communications* **53**, 436-439, doi:10.1039/C6CC08621G (2017).
- 65 Zhang, X. *et al.* Mimicking Ribosomal Unfolding of RNA Pseudoknot in a Protein Channel. *Journal of the American Chemical Society* **137**, 15742-15752, doi:10.1021/jacs.5b07910 (2015).
- 66 Tian, K., Decker, K., Aksimentiev, A. & Gu, L.-Q. Interference-Free Detection of Genetic Biomarkers Using Synthetic Dipole-Facilitated Nanopore Dielectrophoresis. *ACS Nano* **11**, 1204-1213, doi:10.1021/acsnano.6b07570 (2017).
- 67 Wang, Y., Zheng, D., Tan, Q., Wang, M. X. & Gu, L.-Q. Nanopore-based detection of circulating microRNAs in lung cancer patients. *Nature Nanotechnology* **6**, 668, doi:10.1038/nnano.2011.147
<https://www.nature.com/articles/nnano.2011.147#supplementary-information> (2011).

- 68 Alexandre, I. *et al.* Colorimetric Silver Detection of DNA Microarrays. *Analytical Biochemistry* **295**, 1-8, doi:<https://doi.org/10.1006/abio.2001.5176> (2001).
- 69 Li, X., Lee, J. S. & Kraatz, H.-B. Electrochemical Detection of Single-Nucleotide Mismatches Using an Electrode Microarray. *Analytical Chemistry* **78**, 6096-6101, doi:10.1021/ac060533b (2006).
- 70 Wang, Y., Zheng, D., Tan, Q., Wang, M. X. & Gu, L. Q. Nanopore-based detection of circulating microRNAs in lung cancer patients. *Nat Nanotechnol* **6**, 668-674, doi:10.1038/nnano.2011.147 (2011).
- 71 Ayub, M., Stoddart, D. & Bayley, H. Nucleobase Recognition By Truncated α -Hemolysin Pores. *ACS nano* **9**, 7895-7903, doi:10.1021/nn5060317 (2015).
- 72 Butler, T. Z., Pavlenok, M., Derrington, I. M., Niederweis, M. & Gundlach, J. H. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proceedings of the National Academy of Sciences* **105**, 20647-20652, doi:10.1073/pnas.0807514106 (2008).
- 73 Cao, C. *et al.* Discrimination of oligonucleotides of different lengths with a wild-type aerolysin nanopore. *Nature Nanotechnology* **11**, 713, doi:10.1038/nnano.2016.66 <https://www.nature.com/articles/nnano.2016.66#supplementary-information> (2016).
- 74 Wang, S., Haque, F., Rychahou, P. G., Evers, B. M. & Guo, P. Engineered Nanopore of Phi29 DNA-Packaging Motor for Real-Time Detection of Single Colon Cancer Specific Antibody in Serum. *ACS Nano* **7**, 9814-9822, doi:10.1021/nn404435v (2013).
- 75 Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A* **93**, 13770-13773 (1996).
- 76 Nivala, J., Mulrone, L., Li, G., Schreiber, J. & Akeson, M. Discrimination among Protein Variants Using an Unfoldase-Coupled Nanopore. *ACS Nano* **8**, 12365-12375, doi:10.1021/nn5049987 (2014).
- 77 Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α -hemolysin nanopore. *Nature biotechnology* **31**, 247-250, doi:10.1038/nbt.2503 (2013).
- 78 Etedgui, J., Kasianowicz, J. J. & Balijepalli, A. Single Molecule Discrimination of Heteropolytungstates and Their Isomers in Solution with a Nanometer-Scale Pore. *Journal of the American Chemical Society* **138**, 7228-7231, doi:10.1021/jacs.6b02917 (2016).
- 79 Wang, Y. *et al.* Nanopore Sensing of Botulinum Toxin Type B by Discriminating an Enzymatically Cleaved Peptide from a Synaptic Protein Synaptobrevin 2 Derivative. *ACS Applied Materials & Interfaces* **7**, 184-192, doi:10.1021/am5056596 (2015).
- 80 Tian, K. *et al.* Single Locked Nucleic Acid-Enhanced Nanopore Genetic Discrimination of Pathogenic Serotypes and Cancer Driver Mutations. *ACS Nano* **12**, 4194-4205, doi:10.1021/acsnano.8b01198 (2018).
- 81 Imani, N. M., Ruicheng, S., Xinyue, Z., Li-Qun, G. & S., G. K. Sequence-Specific Covalent Capture Coupled with High-Contrast Nanopore Detection of a Disease-Derived Nucleic Acid Sequence. *ChemBioChem* **18**, 1383-1386, doi:doi:10.1002/cbic.201700204 (2017).
- 82 Chen, X. *et al.* Label-Free Detection of DNA Mutations by Nanopore Analysis. *ACS Applied Materials & Interfaces* **10**, 11519-11528, doi:10.1021/acsami.7b19774 (2018).
- 83 The International, S. N. P. M. W. G. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928, doi:10.1038/35057149 (2001).
- 84 Wang, D. G. *et al.* Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* **280**, 1077-1082, doi:10.1126/science.280.5366.1077 (1998).

- 85 Erichsen, H. C. & Chanock, S. J. SNPs in cancer research and treatment. *British Journal of Cancer* **90**, 747-751, doi:10.1038/sj.bjc.6601574 (2004).
- 86 Shastry, B. S. SNP alleles in human disease and evolution. *Journal Of Human Genetics* **47**, 561, doi:10.1007/s100380200086 (2002).
- 87 Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Research* **19**, 1124-1132, doi:10.1101/gr.088013.108 (2009).
- 88 Shagin, D. A. *et al.* A Novel Method for SNP Detection Using a New Duplex-Specific Nuclease From Crab Hepatopancreas. *Genome Research* **12**, 1935-1942, doi:10.1101/gr.547002 (2002).
- 89 Gaylord, B. S., Massie, M. R., Feinstein, S. C. & Bazan, G. C. SNP detection using peptide nucleic acid probes and conjugated polymers: Applications in neurodegenerative disease identification. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 34-39, doi:10.1073/pnas.0407578101 (2005).
- 90 Gao, C., Ding, S., Tan, Q. & Gu, L.-Q. Method of Creating a Nanopore-Terminated Probe for Single-Molecule Enantiomer Discrimination. *Analytical Chemistry* **81**, 80-86, doi:10.1021/ac802348r (2009).
- 91 Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**, 265-270, doi:10.1038/nnano.2009.12 (2009).
- 92 Braha, O. *et al.* Simultaneous stochastic sensing of divalent metal ions. *Nature Biotechnology* **18**, 1005, doi:10.1038/79275
https://www.nature.com/articles/nbt0900_1005#supplementary-information (2000).
- 93 Kasianowicz, J. J., Henrickson, S. E., Weetall, H. H. & Robertson, B. Simultaneous Multianalyte Detection with a Nanometer-Scale Pore. *Analytical Chemistry* **73**, 2268-2272, doi:10.1021/ac000958c (2001).
- 94 Borsenberger, V., Mitchell, N. & Howorka, S. Chemically Labeled Nucleotides and Oligonucleotides Encode DNA for Sensing with Nanopores. *Journal of the American Chemical Society* **131**, 7530-7531, doi:10.1021/ja902004s (2009).
- 95 Kumar, S. *et al.* PEG-Labeled Nucleotides and Nanopore Detection for Single Molecule DNA Sequencing by Synthesis. *Scientific Reports* **2**, 684, doi:10.1038/srep00684
<https://www.nature.com/articles/srep00684#supplementary-information> (2012).
- 96 An, N., White, H. S. & Burrows, C. J. Modulation of the current signatures of DNA abasic site adducts in the alpha-hemolysin ion channel. *Chem Commun (Camb)* **48**, 11410-11412, doi:10.1039/c2cc36366f (2012).
- 97 Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nature Biotechnology* **26**, 1146, doi:10.1038/nbt.1495 (2008).
- 98 Ding, Y., Fleming, A. M., White, H. S. & Burrows, C. J. Internal vs Fishhook Hairpin DNA: Unzipping Locations and Mechanisms in the α -Hemolysin Nanopore. *The Journal of Physical Chemistry. B* **118**, 12873-12882, doi:10.1021/jp5101413 (2014).
- 99 Butler, T. Z., Gundlach, J. H. & Troll, M. Ionic Current Blockades from DNA and RNA Molecules in the α -Hemolysin Nanopore. *Biophysical Journal* **93**, 3229-3240, doi:<https://doi.org/10.1529/biophysj.107.107003> (2007).
- 100 Butler, T. Z., Gundlach, J. H. & Troll, M. A. Determination of RNA Orientation during Translocation through a Biological Nanopore. *Biophysical Journal* **90**, 190-199, doi:10.1529/biophysj.105.068957 (2006).
- 101 Akeson, M., Branton, D., Kasianowicz, J. J., Brandin, E. & Deamer, D. W. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophysical Journal* **77**, 3227-3233 (1999).
- 102 Purnell, R. F., Mehta, K. K. & Schmidt, J. J. Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopore. *Nano Lett* **8**, 3029-3034, doi:10.1021/nl802312f (2008).

- 103 Wanunu, M., Morrison, W., Rabin, Y., Grosberg, A. Y. & Meller, A. Electrostatic Focusing of Unlabeled DNA into Nanoscale Pores using a Salt Gradient. *Nature nanotechnology* **5**, 160-165, doi:10.1038/nnano.2009.379 (2010).