Improving Electronic Genome Mapping Genome Coverage With CRISPR/Cas9

Introduction

Genome mapping of ultra-long, single-molecule DNA has become an important adjunct to sequencing for discovering and characterizing long-range DNA structure and for assisting in genome assembly and structural variant detection.[1, 2] Initial mapping efforts focused primarily on optical methods, starting with examination of restriction site digestion following DNA combing.[3, 4] Later, imaging fluorescent labels on DNA in nanochannels was commercialized and used for examining many different genomes.[5-7] In the initial protocol, fluorescent nucleotides were incorporated into sequence-specific nick sites. Those labels were then visualized by laser excitation of DNA in nanochannels and their relative positions determined using high-resolution imaging systems. The flexibility of the system was extended by using directed binding of fluorescently labeled CRISPR/Cas complexes to mark regions that had a lack of nick sites.[8] In addition to just binding to DNA, CRISPR/Cas was also used to nick and incorporate fluorescent labels.[9, 10] With multiple methods for incorporating labels, it was possible to use multiple colors to distinguish different features on the DNA though the increase in colors significantly increased costs and lowered throughput.[11] Because of these limitations, the product was not commercially viable.

While the optical methods have been successful in mapping complex genomes and structural variants, physical constraints caused by the wavelength of light limiting tag resolution and the expense of purchasing and maintaining expensive lasers and imaging systems have been problematic. To provide an alternative to optical genome mapping (OGM), methods for electronic detection of ultra-long DNA were also developed.[12-14]

Electronic Genome Mapping (EGM) uses the same detection principle as nanopore sequencing where the physical presence of DNA or tagged DNA in a nanochannel causes a measurable current blockade. EGM has the advantage that DNA and site-specific tags can be detected at higher resolution and at lower cost than optical methods.

With recent advances in throughput and ease of use, EGM is now commercially available for examining human genomes. The standard protocol to attach tags to DNA in EGM is similar to what had been used initially in OGM, nicking the DNA site-specifically with commercially available nickases and incorporating modified nucleotides for tagging at those sites. As with OGM, these sites are useful for allowing analysis of up to ~90% of the human genome. However, there are some regions where the density of nick sites is too low to allow discrimination of molecules and there are other regions where the density of nick sites is too high and double-strand breaks may occur as a result. Both these issues can be ameliorated through appropriate use of CRISPR/Cas activities. We will describe the use of the CRISPR/Cas system for the sequence-specific blockade or the creation of nick sites as needed.

Results

The site-specific restriction enzymes BssSI and BspQI have non-palindromic recognition sites of 6 and 7 base pairs, respectively (CACGAG and GCTCTTC). The enzymes have been mutated to allow single-strand nicking of DNA. When used in combination on human genomic DNA, the enzymes cut or nick approximately once every 4000 bp. This creates a broad range of intervals between sites that can be as small as a few bp or as large as more than 100,000 bp. This distribution of separations between nick sites for the human genome is shown in Figure 1.

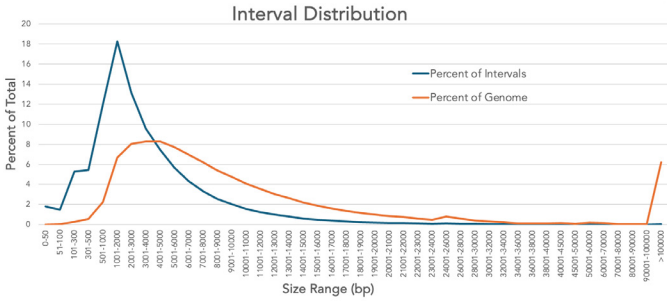

Figure 1: Size distribution of intervals between nick sites. The percent of nick sites falling into a given size range is shown in blue. The percent of the genome encompassed by these intervals is shown in red. While there is a higher fraction of small intervals, they encompass a very small part of genomic space. In contrast, even though there are very few intervals >100,000 bp, they collectively include more than 5% of the genome.

Most of the genome can be readily mapped and examined using these nickases. However, it is beneficial to be able to place additional nick sites in intervals of >50,000 bp so that molecules covering those regions are more readily discriminated in the size range routinely analyzed on the OhmX system (65,000 – 500,000 bp). In addition, nicks on opposite strands in intervals less than 50 bp can be problematic. When there are closely spaced nicks on opposite DNA strands, it is possible for the double stranded region between the nicks to melt, generating a double strand break that creates an interruption in the assembly. These are referred to as Proximity Breaks (PB). Nicks on the same strand are not an issue. Each PB has a characteristic likelihood of being problematic based on the base composition and how the DNA is handled affecting the frequency with which a site breaks. However, it is a common enough occurrence that minimizing PBs improves the quality of the DNA analysis. Preventing at least one of the closely spaced nicks would be advantageous for eliminating PBs.

The Cas9 component of CRISPR has been well characterized with mutations known that will lead to either nicking of a single strand of DNA or no cutting at all when the CRISPR complex binds specifically to DNA.[15] That binding has been shown to be specific and long lasting in normal conditions.[16] Here, we have evaluated both dead (dCas9 which binds but does not cut) and nicking (D10ACas9) versions with multiple sgRNAs to introduce nicks where needed and eliminate nicks where a resulting PB is problematic. The expected experimental flow for nicking and blocking is shown in Figure 2.

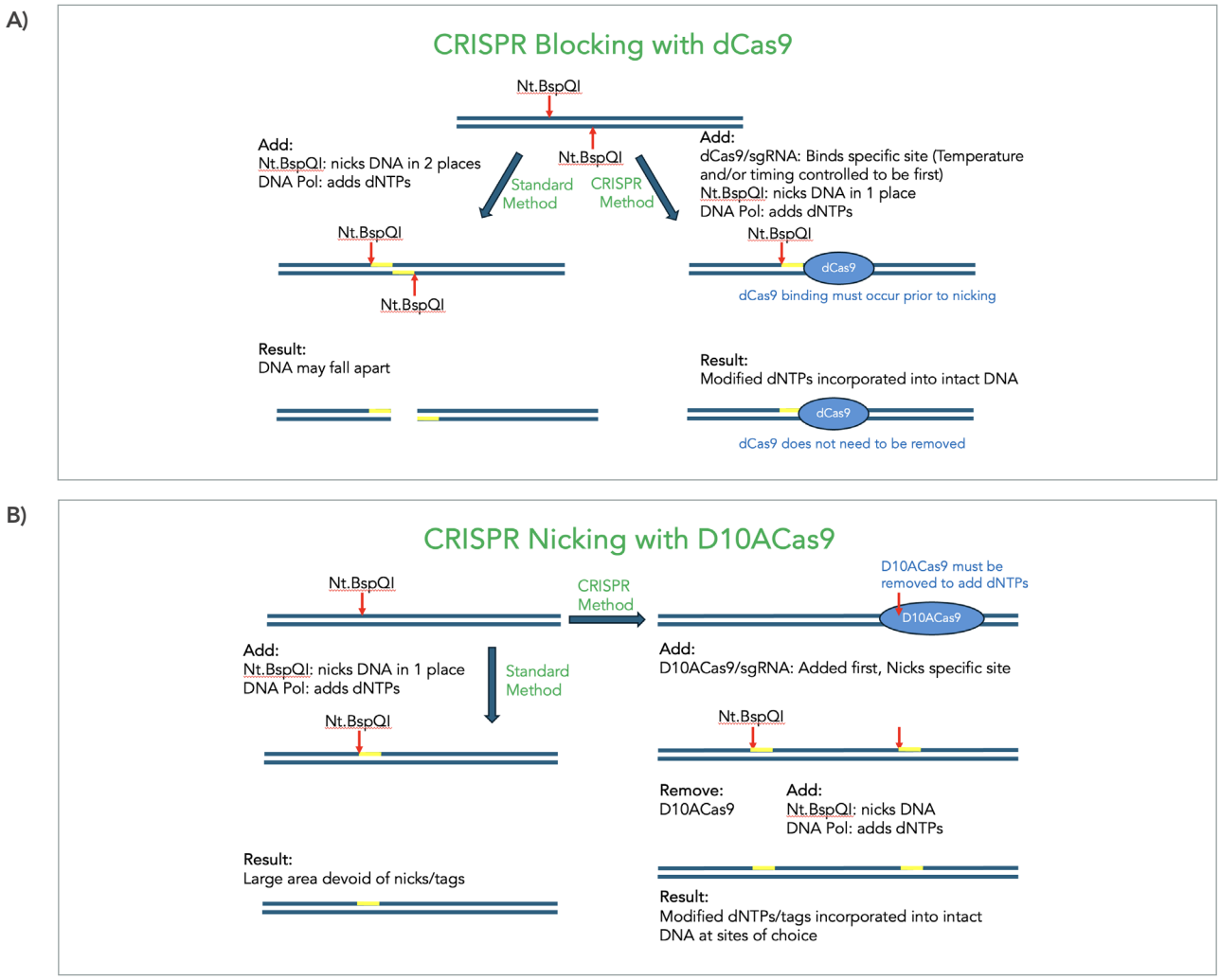

Figure 2: Outline of CRISPR approaches. The timing of addition and inactivation of protein components must be carried out in a specific order to allow proper functioning of each component. The expected conditions for blocking (A) and nicking (B) are shown.

Site	Sequence	Fragment Ends	Length
Nt.BspQI-1	Left End - GAAGAGC	1-2392	2392
Nt.BspQI-2	GAAGAGC - GAAGAGC	2393-6484	4092
Nt.BspQI-3	GAAGAGC - GAAGAGC	6485-8697	2213
Nt.BspQI-4	GAAGAGC - GAAGAGC	8698-10365	1668
Nt.BspQI-5	GAAGAGC - GAAGAGC	10366-13281	2916
gRNA1	TCGTTTCTGGAAGAGCACGG	13277-13296	
Nt.BspQI-6	GAAGAGC - GAAGAGC	13282-24764	11483
Nt.BspQI-7	GAAGAGC - GCTCTTC	24765-27241	2477
Nt.BspQI-8	GCTCTTC - GCTCTTC	27242-34334	7093
gRNA2	CTGTGCTTTCAGTGGATTTC	31235-31254	
Nt.BspQI-9	GCTCTTC - GAAGAGC	34335-34795	461
Nt.BspQI-10	GAAGAGC - GAAGAGC	34796-47707	12912
Nt.BspQI-11	GAAGAGC - End	47708-48502	795
Nb.BssSI-1	End - CACGAG	1-20356	20356
Nb.BssSI-2	CACGAG - CACGAG	20357-25572	5216
gRNA3	CTTAGGTGTTTTAACTCGTG	25572-25591	
Nb.BssSI-3	CACGAG - CTCGTG	25573-27956	2384
Nb.BssSI-4	CTCGTG - CTCGTG	27957-29425	1469
Nb.BssSI-5	CTCGTG - CTCGTG	29426-34430	5005
Nb.BssSI-6	CTCGTG - CTCGTG	34431-35219	789
Nb.BssSI-7	CTCGTG - CACGAG	35220-42416	7197
Nb.BssSI-8	CACGAG - CTCGTG	42417-42737	321
Nb.BssSI-9	CTCGTG - End	42738-48502	5765

Table 1: Locations of BspQI nick sites and gRNAs. The positions within lambda DNA where Nt.BspQI and Nb.BssSI will nick and the gRNAs will bind are listed. The distances between Nt.BspQI and Nb.BssSI nick sites are shown in the right column. The fragments and gRNA involved in or affected by blocking are shaded in gray. The nick recognition sites on the gRNA are shown in red.

While the ultimate goal is to use the various CRISPR complexes for the improved analysis of human DNA, initial experiments were carried out with the 48.5 kb lambda DNA that provides a simpler model system that can be more easily studied. To test the use of dCas9 for blocking a Nt.BspQI site, an sgRNA was synthesized corresponding to sequence coordinates 13,277-13,296, covering the Nt.BspQI nick site at position 13,281. Nick and binding site locations for Nt.BspQI, Nb.BssSI, and gRNAs with lambda DNA are listed in Table 1. The synthesized sgRNA was bound to dCas9 as recommended by the vendor. The complex was added to the lambda DNA prior to the addition of Nt.BspQI nickase and incubated for 1 hr at 37 °C. The lambda DNA was then treated using the standard nicking/labeling protocol including DNA purification on a Nanosep column to remove excess unincorporated nucleotides used in the labeling. No special treatment to remove dCas9 was attempted. Following the standard tagging and RecA coating of the DNA, the tagged DNA was injected into the OhmX instrument and run to collect >100,000 molecules. Typical molecular traces for tagged lambda are shown in Figure 3.

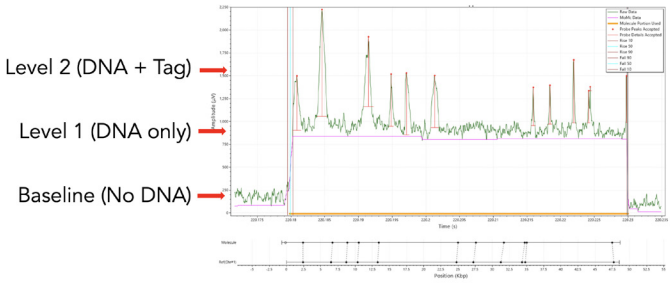

Figure 3: Voltage patterns for lambda DNA. Either end of a molecule may enter the channel first. When the DNA enters the detector region, the voltage changes from baseline due to the current blockade (Level 1). It changes again whenever there is a tag present (Level 2). The time between tags is measured and converted to base pairs using a signal processing algorithm.

As shown in Figure 4, the frequency of tag 5 is substantially reduced with most DNA molecules showing an apparent false negative. It is noteworthy that the rate of false positives in the neighboring region goes down and it is likely that some of those are actually misassigned as true positives at tag 5.

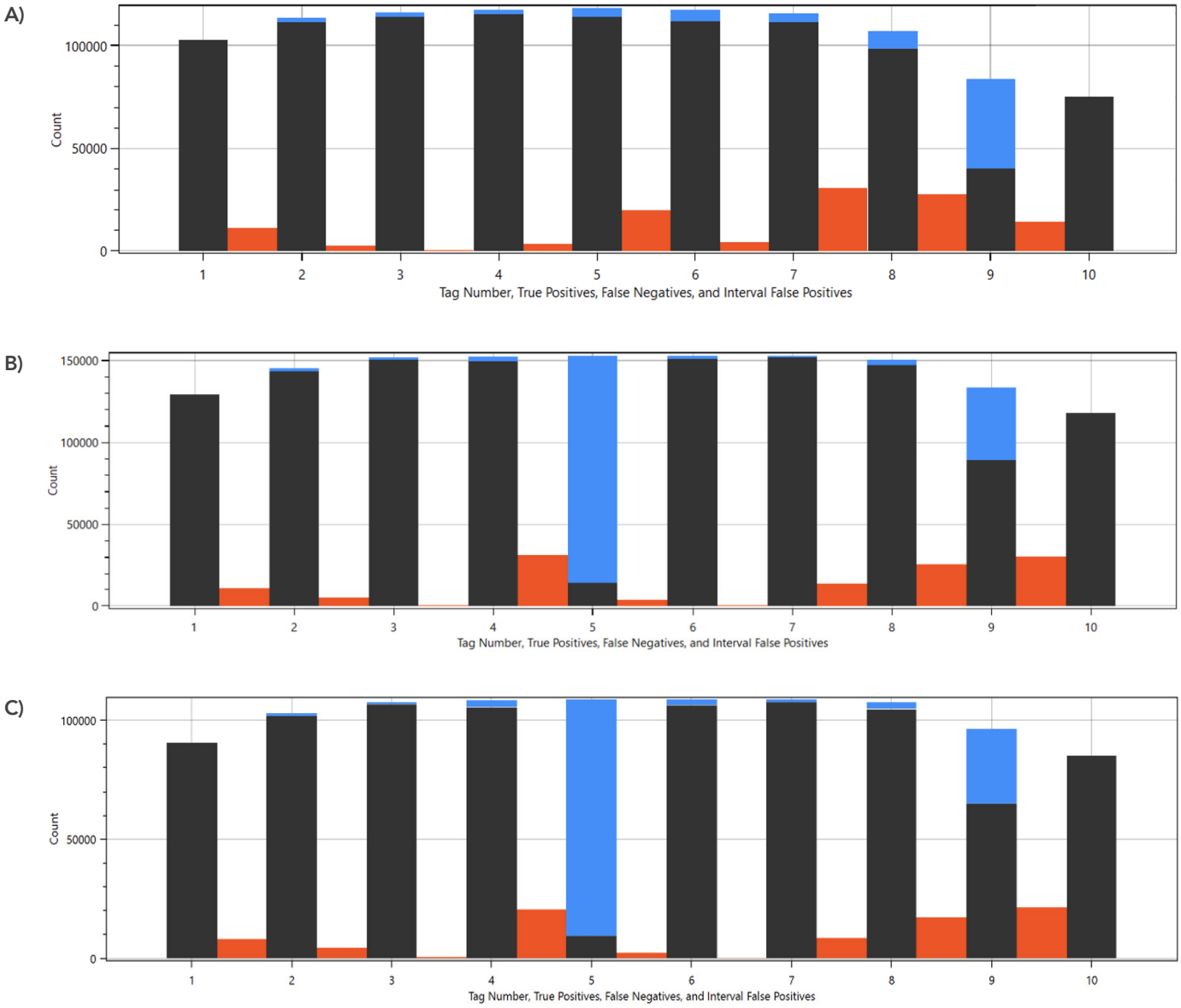

Figure 4: Histogram of tag frequency for control and blocked lambda. The frequency of detection for each of the ten tags is shown in black. The blue bars indicate that there was no tag present in a region in which there was DNA in the nanochannel. The red bars correspond to tags that were found in a region that should not have had a tag. Tag 5 is not present if blocked by CRISPR gRNA1. Panel A has no blocker present while panels B and C were blocked at 37 °C and 25 °C, respectively.

To minimize the number of addition and mixing events required for blocking, a different method of adding the CRISPR complex was attempted. Rather than adding them sequentially as in Figure 4B, both the blocking and nicking components were added simultaneously with relative activity controlled by temperature in Figure 4C. The CRISPR binding occurs at lower temperatures[16] than the nicking activity so the initial incubation was at 25 °C for CRISPR binding followed by an incubation at 37 °C for nicking. The results with binding at 25 °C and 37 °C to block nick sites were indistinguishable (Figure 4). To ensure that this result was generalizable with other nickases, another gRNA was designed to cover the BssSI site at position 25,572 and tested in the same manner as with the BspQI blocker. Blocking with the BssSI directed gRNA also blocked nicking.

Blocking nick sites with CRISPR is straightforward because the bound CRISPR/Cas9 complex does not need to be removed from the DNA for proper downstream sample processing. However, if new tags are to be introduced via nicking, the CRISPR complex must be removed from the nick site prior to incorporation of modified nucleotides because the bound protein prevents access to the nick site by the DNA polymerase. When the D10ACas9 nicks the DNA but is not removed, the modified nucleotides that are used in the labeling reaction cannot be incorporated so no tagging occurs (data not shown). When the complex is removed via heat or proteinase K, modified nucleotide incorporation can occur at the Cas9 nick identically to nicks generated by Nt.BspQI (Figure 5).

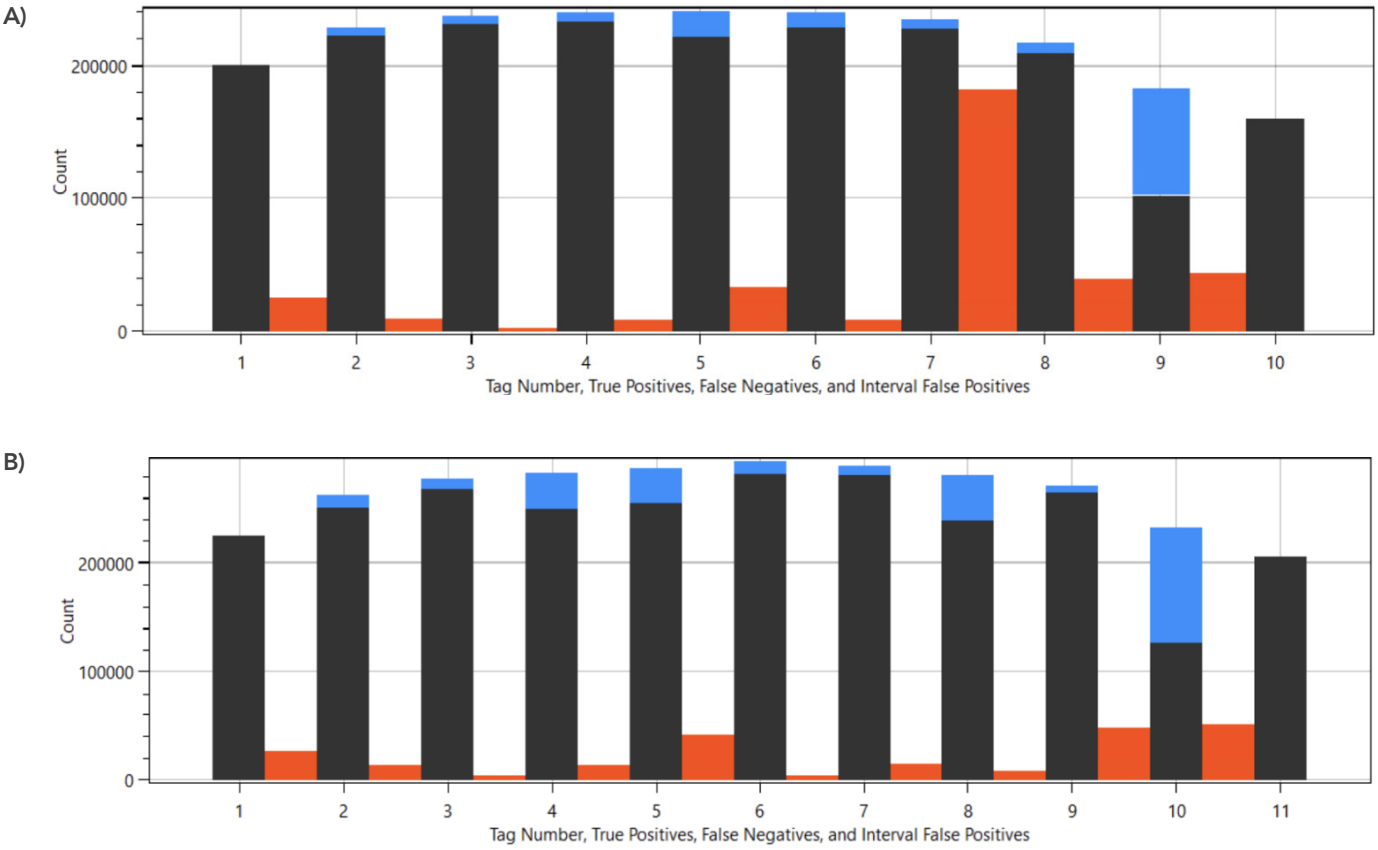

Figure 5: Histogram of tag frequency for control and gRNA nicked lambda. The frequency of detection for each of the ten tags is shown in black. The blue bars indicate that there was no tag present in a region in which there was DNA in the nanochannel. The red bars correspond to tags that were found in a region that should not have had a tag (false positives). The top histogram matches tags to the standard set of 10 Nt.BspQI tags and has a significant apparent false positive between tags 7 and 8 corresponding to the site where gRNA2 should nick. In the bottom histogram, an additional site is added to the reference corresponding to where gRNA2 should nick and the false positive is now scored correctly as a true positive.

Discussion

The results provided here demonstrate the power and added flexibility that CRISPR functions can add to EGM. By tuning the number of nick sites, improved accuracy and additional genomic regions can be analyzed. However, there are also a number of other functions that can be envisioned by using other CRISPR capabilities. This work was limited to using the pre-existing nicking protocol, but it is also feasible to attach large tags directly to the CRISPR complex such that nicking, labeling, and tagging are not necessary. The bound CRISPR complex could be sufficiently large for detection on the OhmX Platform, and it could be added after the standard process to minimize handling steps or replace them entirely. The CRISPR/Cas tag could be made distinguishable from the nicking tag by incorporating any desired characteristics. While gRNAs cannot be made for all potential sequences of interest, the requirements are highly flexible, and it is rare not to find a suitable target within a region of thousands of bases.

Most genomic regions, including centromeres, telomeres, duplications, and other troublesome genomic regions, have been found to have sequences that can produce distinctive nicking or binding patterns. For example, the highly repetitive Dux4 region near the end of chromosome 4 can be involved in a pathogenic repeat contraction that causes facioscapulohumeral muscular dystrophy (FSHD). It is nearly identical for over 160 kb to the same region of chromosome 10. The standard nicking enzymes alone have difficulty distinguishing between the two regions due to > 99% identity. Additionally, there are a small number of variations that lead to the 4qA and 4qB haplotypes, which are critical for distinguishing a pathogenic from a non-pathogenic repeat contraction.[17] These cannot be readily identified with nickases, but the custom design of gRNAs allows for the generation of easily distinguished tagging patterns. Similarly, the repeat sequences on the two chromosomes are not 100% identical and can be selectively tagged if appropriate gRNAs are used. Similarly, centromeres are poorly tagged and cannot be uniquely identified via nickases. gRNAs can be designed to allow unique tagging patterns for each centromere commonly involved in Robertsonian translocations.

If multiple gRNAs are used in an assay, they can be uniquely labelled to allow different tags for each gRNA complex. Because the magnitude of the current blockade in the nanochannels depends on tag size and shape, the same detector can be used for all tags. Use of multiple means of detection is not required for the equivalent of multi-color signals. There are many genomic features arising from base modifications or protein binding whose positions on the genome are valuable to understand. A single OhmX Detector could be used to determine the relative positions of multiple analytes on the tagged genome background.

The ability of CRISPR complexes to bind DNA at lower temperatures than those at which many enzymes are active provides flexibility in when components can be added and simplifies the protocol. When blocking nickases, CRISPR can be added at the same time as the nickases tested here because they are not active at 25°C, so binding occurs before nicking can take place. The temperature can be shifted without additional mixing steps.

CRISPR also has other potential applications that enhance the utility of EGM. CRISPR has been used for targeted selection and similar techniques could be used for EGM.[18] Targeted selection and creation of sample-specific tagging that would allow multiplexing and fragment identification would be enabled. The ability of CRISPR/Cas9 and similar complexes to be modified in a variety of ways—either to bind, bind and nick, or bind and cleave—provides many possible enhancements to EGM.

Learn more at nabsys.com

References

1. Logsdon, G.A., M.R. Vollger, and E.E. Eichler, *Long-read human genome sequencing and its applications.* Nat Rev Genet, 2020. 21(10): p. 597–614.
2. Zook, J.M., et al., *A robust benchmark for detection of germline large deletions and insertions.* Nat Biotechnol, 2020. 38(11): p. 1347–1355.
3. Schwartz, D.C., et al., *Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping.* Science, 1993. 262(5130): p. 110–4.
4. Michalet, X., et al., *Dynamic molecular combing: stretching the whole human genome for high-resolution studies.* Science, 1997. 277(5331): p. 1518–23.
5. Lam, E.T., et al., *Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly.* Nat Biotechnol, 2012. 30(8): p. 771–6.
6. Levy-Sakin, M. and Y. Ebenstein, *Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy.* Curr Opin Biotechnol, 2013. 24(4): p. 690–8.
7. Cao, H., et al., *Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology.* Gigascience, 2014. 3(1): p. 34.
8. Zhang, D.C., Saki; Sugerman, Kenneth; Lee, Joyce; Lam, Ernest T.; Bocklandt, Sven; and H.H. Cao, Alex R. , *CRISPR-bind: a simple, custom CRISPR/dCas9-mediated labeling of genomic DNA for mapping in nanochannel arrays.* bioRxiv, 2018(371518).
9. McCaffrey, J., et al., *CRISPR-CAS9 D10A nickase target-specific fluorescent labeling of double strand DNA for whole genome mapping and structural variation analysis.* Nucleic Acids Res, 2016. 44(2): p. e11.
10. Abid, H.Z., et al., *Customized optical mapping by CRISPR-Cas9 mediated DNA labeling with multiple sgRNAs.* Nucleic Acids Res, 2021. 49(2): p. e8.
11. Jeffet, J., et al., *Single-molecule optical genome mapping in nanochannels: multidisciplinarity at the nanoscale.* Essays Biochem, 2021. 65(1): p. 51–66.
12. Thompson, J.F. and J.S. Oliver, *Mapping and sequencing DNA using nanopores and nanodetectors.* Electrophoresis, 2012. 33(23): p. 3429–36.
13. Kaiser, M.D.D., J. R.; Grinberg, B. S.; Oliver, J. S.; Sage, J. M.; Seward, L.; Bready, B., *Automated Structural Variant Verification in Human Genomes using Single-Moleucle Electronic DNA Mapping.* bioRxiv, 2017. 140699.
14. Oliver, J.S.C., A.; Davis, J. R.; Grinberg, B. S.; Hutchins, T. E.; Kaiser, M. D.; Nurnberg, S.; Sage, J. M.; Seward, L.; Simelgor, G.; Weiner, N. K.; Bready, B., *High-Definition Electronic Genome Maps from SIngle Molecule Data.* bioRxiv, 2017. 139840.
15. Jiang, F. and J.A. Doudna, *CRISPR-Cas9 Structures and Mechanisms.* Annu Rev Biophys, 2017. 46: p. 505–529.
16. David, S.R., et al., *Temperature dependent in vitro binding and release of target DNA by Cas9 enzyme.* Sci Rep, 2022. 12(1): p. 15243.
17. Lemmers, R.J., et al., *Worldwide population analysis of the 4q and 10q subtelomeres identifies only four discrete interchromosomal sequence transfers in human evolution.* Am J Hum Genet, 2010. 86(3): p. 364–77.
18. Malekshoar, M., et al., *CRISPR-Cas9 Targeted Enrichment and Next-Generation Sequencing for Mutation Detection.* J Mol Diagn, 2023. 25(5): p. 249–262.