De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing

Authors: Marcus Stoiber1, Joshua Quick2, Rob Egan3, Ji Eun Lee3, Susan Celniker1, Robert K. Neely4, Nicholas Loman2, Len A Pennacchio1,3, James Brown1,5,6,7

Affiliations:

1. Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology
2. University of Birmingham, Institute of Microbiology and Infection
3. Lawrence Berkeley National Laboratory, Joint Genome Institute
4. University of Birmingham, School of Chemistry
5. Centre for Computational Biology, School of Biosciences, University of Birmingham, UK
6. Department of Statistics, University of California, Berkeley, US
7. Preminon, LLC. A California Corporation

Abstract

Advances in single molecule sequencing technology have enabled the investigation of the full catalogue of covalent DNA modifications. We present an assay, Modified DNA sequencing (MoD-seq), that leverages raw nanopore data processing, visualization and statistical testing to directly survey DNA modifications without the need for a large prior training dataset. We present case studies applying MoD-seq to identify three distinct marks, 4mC, 5mC, and 6mA, and demonstrate quantitative reproducibility across biological replicates processed in different labs. In a ground-truth dataset created via *in vitro* treatment of synthetic DNA with selected methylases, we show that modifications can be detected in a variety of distinct sequence contexts. We recapitulated known methylation patterns and frequencies in *E. coli*, and propose a pipeline for the comprehensive discovery of DNA modifications in a genome without *a priori* knowledge of their chemical identities.

Introduction

DNA modifications are essential across the three kingdoms of life[1], and are used by cells for defense, gene regulation, cell differentiation, and the transmission of regulatory programs across generations. A host of assays have been developed to detect specific modified nucleotides, including and especially 5mC and 6mA, which are widely deployed by prokaryotes and eukaryotes[2-4]. Techniques exist to detect a diverse group of epigenetic modifications through the observation of DNA Pol II kinetics leveraging Single Molecule Real-Time sequencing (SMRT-seq) platform[5, 6]. In particular, the pioneering work of Clark *et al.*[7] demonstrated the capacity to identify DNA methylation marks via the comparison of native versus amplified DNA through supervised machine learning. The SMRT-seq platform, provides observations of DNA modifications through analysis of polymerase dynamics, which leads to the current requirement of deep read coverage in order to identify particular DNA modifications[5, 7].

Nanopore sequencing technology confers the opportunity to identify modified nucleotides through direct observations of single-molecules through monitoring electric current. Several pilot studies have demonstrated the feasibility of using nanopore-derived information to identify methylation marks in native DNA[8-10]. To date, such studies have used a highly processed form of the data generated by the nanopore platform. Further, no software packages have been developed to interrogate and visualize the raw data in a human-interpretable fashion. Here, we present software that implements visualization to enable direct exploration, and automated statistical procedures to discover DNA modifications of, in principle, any form, even when the chemical identity of the modification is not known *a priori*. That is, we utilize unsupervised, rather than supervised statistical learning. We demonstrate the efficacy of our approach for three known marks, 4mC, 5mC and 6mA, in an artificial “ground truth” setting, and also in a well-studied laboratory strain of *E. coli*.

Modified DNA sequencing (MoD-seq; Figure 1) requires the (nanopore) sequencing of a native and matched amplified DNA sample (where amplification is employed to produce unmodified DNA). These data are processed with the nanoraw software package (pypi.python.org/pypi/nanoraw; code repository https://github.com/marcus1487/nanoraw) and the re-processed raw signal is compared genome-wide leading to the identification of consistently modified bases, with discriminative power to accurately detect known marks in *E. coli*, and also the potential identification of new signals of unknown origin. Several similar approaches have been previously described[6, 8-11], but require large prior training datasets, and have not yet been conceptually packaged as an assay for the genome-wide discovery of DNA modifications. In our view, such technology may soon enable the description of the entire collection of modified nucleotides in a genome.

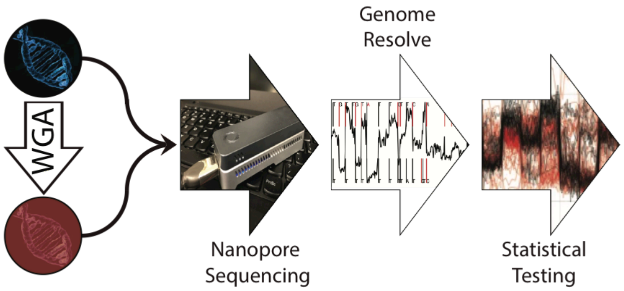


Figure 1. Modified DNA sequencing (MoD-seq) pipeline. Native and whole genome amplified (WGA) biological samples are processed using nanopore sequencing, raw signal is analyzed with *nanoraw* and statistical tests are performed to identify regions with modified bases.

We anticipate that this approach, enabled by the software we present, will play an important role in microbial, plant, and metazoan genomics, particularly and initially for non-model organisms. Many software packages exist[12, 13] to assemble complete genomes from nanopore data, and hence genome sequence and epigenetic modifications can be simultaneously obtained in a single assay without prior knowledge of the collection of extant epigenetic marks in an organism. Further, we point to future work, where coupling to mass spectrometry and NMR may provide a complete parts list of endogenous DNA modifications in any system. The implications of this technology are clear and wide reaching for cancer genomics, population genetics, studies of epigenetic heritability, and the environmental biosciences.

Results

Visualization of the raw output of nanopore sequencing

Base-calling in nanopore sequencing currently relies on treating signal as a locally stationary process, first involving segmentation into stationary regimes ("events"), and then kmer-calling within segments to assign putative kmers[14-16]. Initial assignments are then resolved to individual nucleotide calls by joint analysis of consecutive segments. Precision for the initial k-mer calls is relatively poor, and is improved upon reconciliation of neighboring regions. Individual "1D" nucleotide calls are now more than 90% accurate for single molecule reads[17], facilitating both de novo genome assembly[13, 18, 19] as well as the processing presented here.

We developed the *nanoraw* software package to resolve raw nanopore signal with genomic positions (Methods; implemented in the genome_resquiggle subcommand) and thus allow genome-browser style visualization. The alignment of raw signal to underlying genomic sequences constitutes a robust procedure applicable to current and foreseeable subsequent generations of the technology with little to no tuning of parameters. The *nanoraw* software allows the selection of genomic locations via a multitude of criterion enabling the visual identification of regions of consistent or inconsistent raw signal (Figure 2A, B), and, as a result gain insight and intuition into the process of nucleotide assignments.

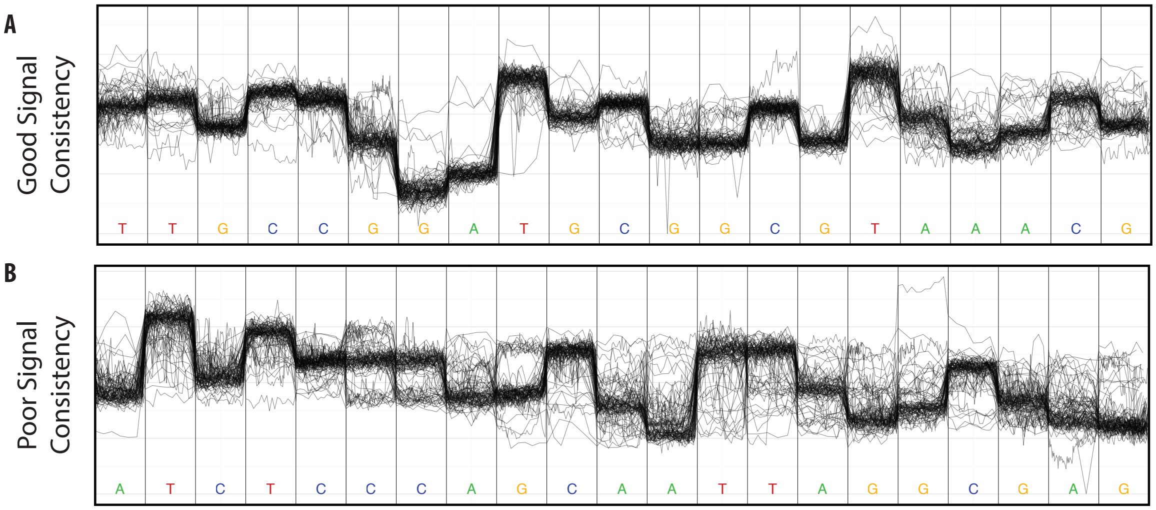


Figure 2. Selected regions of raw nanopore signal visualized with *nanoraw* to show good and poor raw signal consistency.

Identification of chemically modified nucleotides

We leverage the previously established[8, 10, 11] strategy comparing native to amplified DNA in the context of genome-guided analysis to discover modified nucleotides without *a priori* knowledge of their chemical composition or effect on the nanopore signal.

To demonstrate the feasibility of this approach for three major classes of DNA modifications, we constructed a ground truth dataset using seven purified methylases to introduce methylation to whole genome amplified *E. coli* DNA at known target sites (Table 1). These methylases catalyze the addition of a methyl group to the DNA to produce three distinct methylated bases: 4 methyl-cytosine (4mC; M.BamHI), 5 methyl-cytosine (5mC; M.HhaI, M.MpeI and M.SssI) and 6 methyl-adenine (6mA; M.TaqI, M.EcoRI and M.dam). Each of these samples and two control samples were processed by nanopore sequencing (Methods).

Methylase	Known Recognition Site	Average Depth	Methylase Class	Sites in *E. coli* Genome
TaqI	TCG**A**	22	6mA	30914
BamHI	GGAT**C**C	36	4mC	988
EcoRI	GA**A**TTC	27	6mA	1290
HhaI	G**C**GC	50	5mC	65566
MpeI	**C**G	39	5mC	693340
SssI	**C**G	19	5mC	693340
dam	G**A**TC	33	6mA	38240

Table 1. Tested methylases with known recognition site (methylated base underlined), depth of sequencing, methylation class, and number of sites within the *E. coli* genome.

We align individual reads, corresponding to single molecules, to a reference genome (Methods). We note that the *nanoraw* pipeline can easily be applied to a genome derived directly from the same nanopore data using established pipelines[12] when analyzing organisms without a reference genome. Nanopore assemblies of *E. coli* have yielded accurate single-contig genomes (99.5% nucleotide identity,[20-22]). After alignment, raw signal was re-segmented using *nanoraw* to map raw signal events onto the genome. Given two collections of corrected events, one corresponding to native DNA and another to amplified, the identification of DNA modifications is reduced to a statistical testing problem. This approach contrasts with previous DNA modification identification algorithms which model signal shifts and require new training datasets for each modification and genomic sequence context[8, 11]. To identify genomic positions with shifted electric current, as compared to an amplified sample, the *nanoraw* pipeline employs the Mann-Whitney U-test[23]. As modified bases consistently shift the electric current at several bases surrounding the modified base (Supp. Figure 1), Fisher's method[24] was applied across a moving window to produce final significance tests (Methods). Bases admitting statistically significant tests indicated regions with modified nucleotide(s).

For each chemical modification (4mC, 5mC and 6mA) *nanoraw* discovered the known sequence specificity of each enzyme based solely on shifted signal levels (Figure 3A, Supp. Figure 2; the signal and p-value distribution figures are immediately produced by the *nanoraw* software). Dam methylase shows expected motif degeneracy[5, 25] and comparatively weak specificity. Fisher's method p-value distributions for the top 1,000 most significant regions containing the known motif (Figure 3A, Supp. Figure 2, lower panel) indicate that globally the highest significance values centered on or immediately before the known modified base. Genome-wide, each methylase shows strong preference for the known sequence motif with variable levels of accuracy (Figure 3B; area under the curve (AUC) from 0.59 to 0.86). For M.BamHI the AUC can be improved from 0.66 to 0.75 by including the discovered degeneracy at the fourth position of the motif (Figure 3A) indicating a lack of precision in the known motif. Statistical power for the detection of modified bases also scales as expected with sequencing depth (Supp. Figure 3).

Endogenous modifications in laboratory strain of *E. coli*

To simultaneously assess our capacity to identify endogenous modifications along with the biological and technical reproducibility of our approach, we applied the MoD-seq pipeline to *E. coli* (strain K-12 MG1655), one of the best-studied genetic model organisms, independently in two laboratories (Experiment A: Pennacchio Lab, LBNL, USA, and Experiment B: Loman Lab, Univ. Birmingham, UK). Due to inclusion of two additional bacterial samples the coverage of the *E. coli* genome from experiment B samples was substantially lower (8X native and 11X amplified average strand-specific coverage) than the experiment B samples (21X native and 156X amplified average coverage). In both experiments the expected[26, 27] modifications catalyzed by M.dam (6mA) and M.dcm (5mC) were identified as the top hits (Figure 4A and Supp. Figure 4) and ubiquitously throughout the genome with similar specificity to *in vitro* methylation (Figure 4B).

Reproducibility of the top sites identified within both experiments was strong (Supp. Figure 5) with over 18% of modified bases shared in the top 1% of identified sites in both experiments. To assess this correspondence we created pseudo-experiments by downsampling reads from the experiment B samples to achieve 10X average strand-specific coverage (approximately equal to read depth in experiment B). The

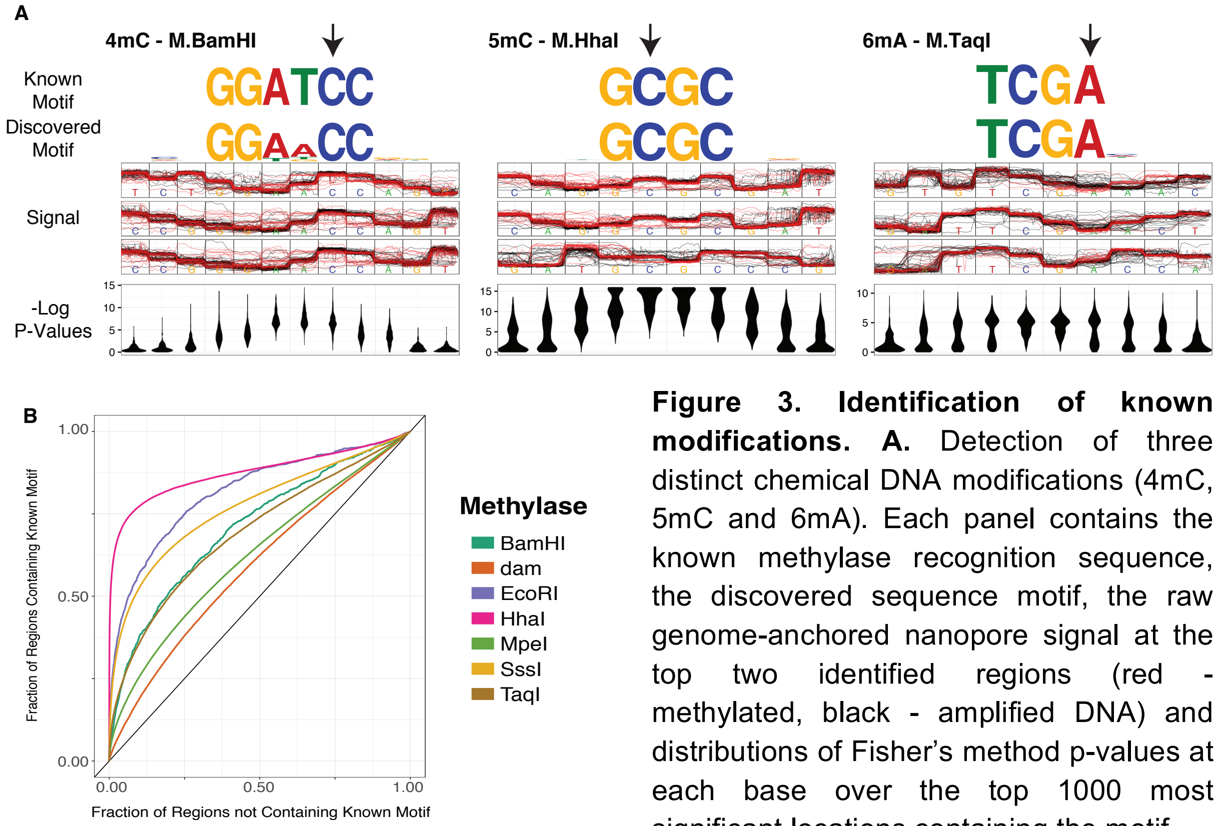


Figure 3. Identification of known modifications. A. Detection of three distinct chemical DNA modifications (4mC, 5mC and 6mA). Each panel contains the known methylase recognition sequence, the discovered sequence motif, the raw genome-anchored nanopore signal at the top two identified regions (red - methylated, black - amplified DNA) and distributions of Fisher's method p-values at each base over the top 1000 most significant locations containing the motif. Additional tested methylases are shown in Supp. Figure 2 **B.** A range of capacity across seven methylases to discriminate regions with the known motif (y-axis) from regions without the known motif (x-axis).

correspondence between these pseudo-experiments show ~20% overlap between the top 1% of calls from the full experiment B indicating that the replicability between the two experiments on different continents presented here is comparable to technical replicates (Supp. Figure 6). We observe that reproducibility is strongly affected by the depth of coverage over a site. When only sites with strand-specific coverage greater than 12X across the four samples were considered (two experiments with native and amplified sequencing samples) we observed a marked increase in overlap at the top of the rank lists (32% overlap between top 1% of sites from the two experiments; Supp. Figure 7 and 8). An extended discussion of factors affecting reproducibility of identification of DNA modifications can be found in the supplemental text.

Motif analysis reveals that 96% and 68% of modifications we detect in experiments A and B respectively are attributable to known methylase based on sequence context. Downsampling indicates that the lower percentage from experiment A is due to increased statistical power provided by deeper sequencing (Supp. Figure 9). Additionally, downsampling analysis suggests that greater than 10X strand-specific coverage is needed to achieve optimal consistency of identified sites, with some additional power achieved out to 15X coverage (Supp. Figure 9; "Fraction of Top 2k Sites Containing Motif" panel).

To discover novel modifications, we employed an unsupervised dimension reduction approach. We associate each modification with a 5-vector where values are the statistically normalized deviations in signal intensity between the native and amplified libraries (Methods). We then projected these vectors into a 2-dimensional plane using Multidimensional Scaling (MDS) for visualization and clustering (Figure 4C). The dominant clusters correspond to the positional offsets (centered on U-test p-values; Supp. Figure 1) from the known 4mC and the 5mC methylases (DCM and DAM methylase respectively). Other sites contain the potential for modifications of unknown origin -- however these are clearly rare. Hence, unsurprisingly, the vast majority of epigenetic modifications in *E. coli* are of known origin, and our pipeline detects them and provides a clustering that contains information about the underlying modification. This

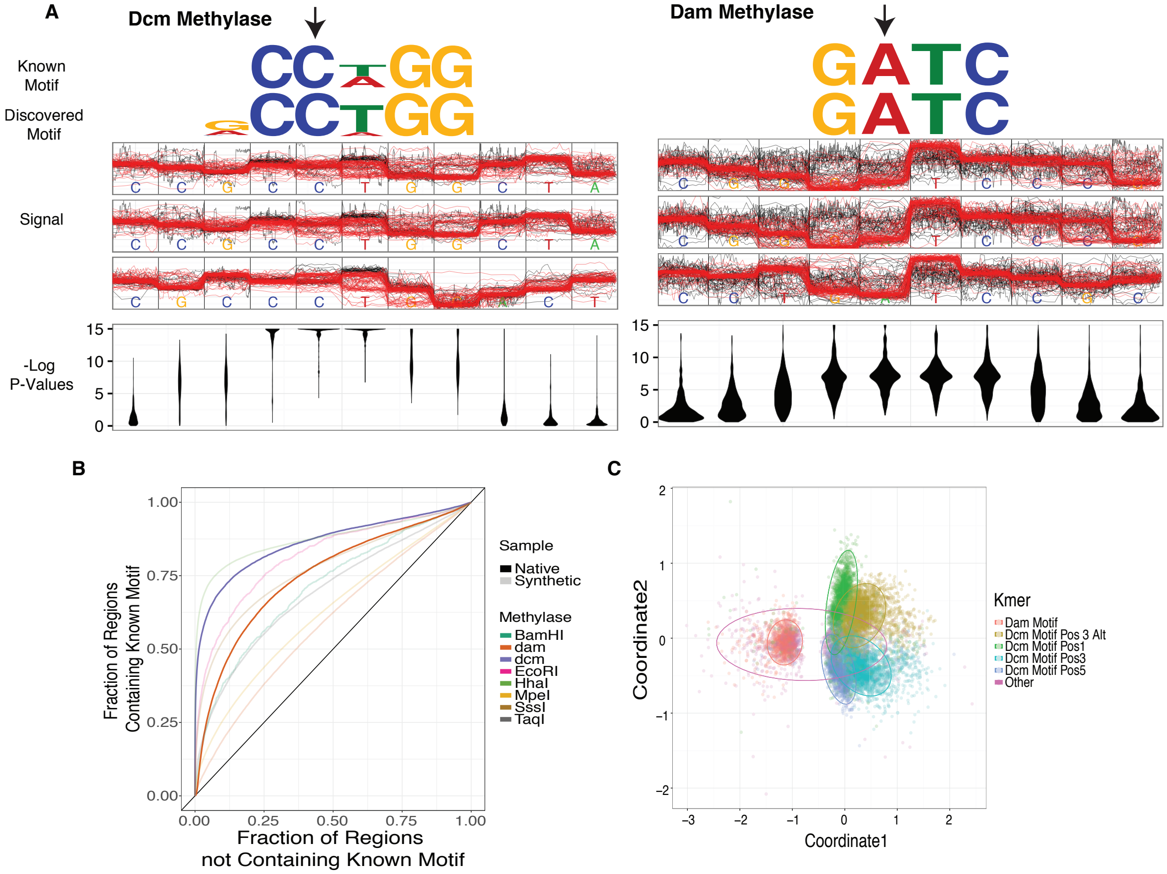


Figure 4. Identification of DNA Modifications using MoD-seq A. Examples for the two major classes of modifications found in native E. coli (dcm and dam methylase). Panels are as in Figure 3A (red - native, black - amplified DNA). **B.** Capacity to discriminate known native modified base motifs using MoD-seq as in Figure 3B. **C.** Clustering of individual modified locations based solely on raw genome-anchored signal. Coloring indicates which motif the region matches (PosN indicates matching to one of the several modified positions within the motif).

analysis demonstrates that the raw signal can be used to visualize, cluster, and detect distinct DNA modifications genome-wide.

A note on signal normalization

In addition to resolving raw signal with a genomic alignment and appropriate statistical testing, raw signal normalization is key to the accurate identification of modified bases. Many current nanopore signal processing applications[14-16] utilize picoamp estimations provided by Oxford Nanopore Technologies. However, this strategy leaves considerable and systematic variation in the signal. Extant pipelines rely on picoamp levels and model specific parameters to discover epigenetic modifications[14-16]. To discover modification directly from raw signal, we apply a median normalization strategy (Methods) based solely on the raw signal (without use of segmentation, called bases or picoamp normalization parameters) to greatly reduce this variation. We view a large fraction of the variance in picoamp measurements as bias, likely due to pore-specific and time-specific fluctuations in current level. Median normalization increases the percent of variance explained by 4-mer sequence context from 68.9% (using picoamp measurements) to 96.7%, and reveals remarkable reproducibility in the raw signal (Figure 5). K-mers with high residual variance after this normalization remain an intriguing furrow for future studies. We propose that median signal normalization constitutes a useful default for raw signal processing, and that k-mer centric conditional variance be adopted as the metric for the assessment of future signal normalization procedures, provided that the normalization procedure does not take called sequence context into account.

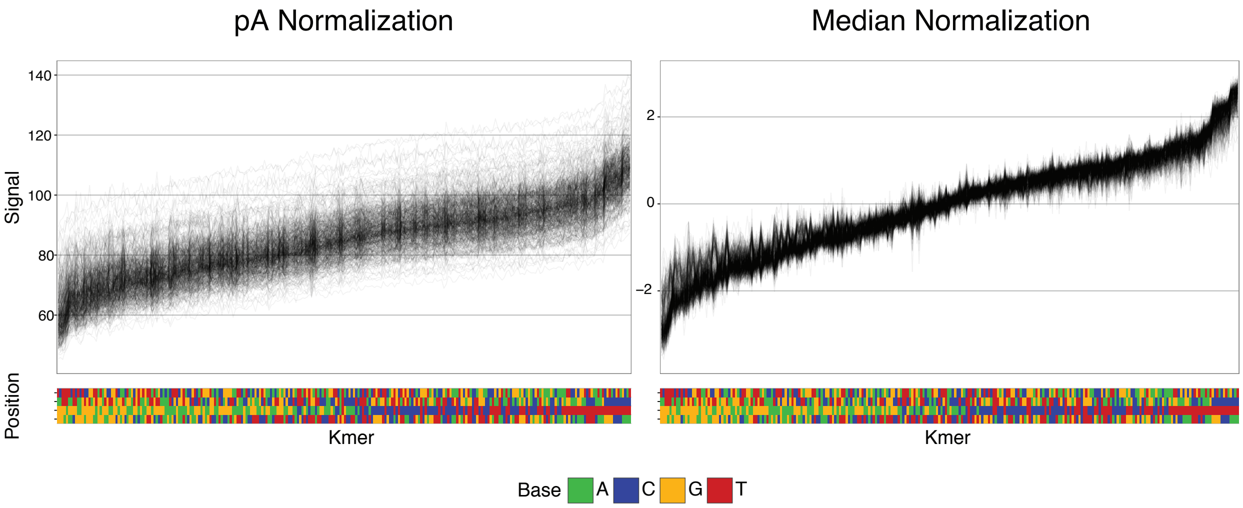


Figure 5. Oxford Nanopore Technologies picoamp normalization (left) versus median normalization (right). Left to right are 4-mers (from bottom to top position: one base already passed through the pore, the base at the center of the pore and two bases that have not yet passed though) ordered by mean signal across all reads. Each line represents the mean signal of one read across all 4-mers.

Discussion

The MoD-seq assay identifies DNA modifications using raw nanopore data and statistical analysis. At present, we require repeated observations of modifications for detection, meaning that individual modified nucleotides need to be consistently present in multiple copies of the genome, e.g. in multiple cells or plasmids in the sample. Once identified, modeling of the consistent signal shifts could conceivably generate phased, single-stranded modification sites in individual, single-molecule reads. As nanopore median and maximum read lengths continue to increase, these calls could be used to phase an epigenome with single-strand resolution. In diploid organisms, such as human, this would be a considerable advance over bisulfite sequencing and antibody-based approaches, and will enable the study of "population epigenetics".

Iterating the base-calling procedure to explicitly take into account modifications during sequencing, progressively enlarging the chemical vocabulary of the base caller, will, at least in part, ameliorate the combinatorial complexity associated with this task. Indeed, one of the most obvious applications of *nanoraw* is the generation of training sets for base-calling algorithms that seek to improve the state of the art in this sequencing platform. Iteratively refined base-calls may improve resolution and power for the detection of modifications. Importantly, *nanoraw* does not require *de novo* knowledge of modification identity or sequence specificity. Thus training data sets can be produced wherein modified and unmodified bases exist within very short distances. As most current base calling methods "remember" a local state, such a training set will likely prove useful to increase the accuracy of nanopore data.

In organisms and systems with more diverse DNA modifications, or less sequence specificity at modified residues, our clustering approach provides an avenue for the systematic discovery of any modified nucleotides, even when we do not know the identity of the modifications *a priori*. For instance, one could couple MoD-seq to mass spectrometry (MS) or nuclear magnetic resonance (NMR) to discern specific moieties. One approach would be to fragment native DNA and use biotinylated oligonucleotide probes to pull down suspect regions of native DNA identified by *nanoraw* e.g. via sonication or the use of restriction enzymes, and then subject the precipitate to tandem mass spectrometry – a procedure we call MoD-MS/MS. In this way, a complete survey of DNA modifications could be derived *de novo* without the need for antibodies to specific moieties or enzymes. Given recent descriptions of the likely importance of 6mA modifications in metazoans, and the opportunistic (antibody enabled) nature of such discovers to date, it seems unlikely that a complete vocabulary of endogenous DNA modifications exists for complex organisms.

As the consistency of raw nanopore signal improves, it may become possible to identify modified bases from individual molecules without the need for repeated observations. This would open new furrows in exposure biology, where adducts are distributed stochastically in the genome due to non-endogenous chemical activities. More than 200 different types of DNA adduct resulting from exposure to exogenous and endogenous DNA binding compounds have been described[28]. These observations could be correlated with patterns of mutation in tumors and cell lineages within tumors to study the mechanisms of DNA repair underlying individual and environmentally-induced cancer susceptibility. Ultimately, such technologies may enable new diagnostic and therapeutic strategies in precision medicine.

Additionally, we acknowledge the enormous power of the human end-user for the detection of interesting patterns in complex data. The effective visualization of raw nanopore signal in genomic contexts may yield unexpected dividends as biologists browse signal-level information in regions containing genes or genomic elements of interest. For instance, we anticipate the generation of “ChIP-nano*”* assays to discover patterns of epigenetic marks associated with transcription factor binding sites. Such correspondences seem likely given a recent report that at least 70% of transcription factors in *Arabidopsis thaliana* have differential binding affinities at 5mC sites[29].

Lastly, direct RNA sequencing will likely soon be possible on the Oxford Nanopore platform, and the approaches we present here may be useful for the study of RNA modifications. Combining MoD-seq with a variety of pull-down assays for DNA and RNA has the potential to transform our understanding of the molecular codes of life.

Methods

DNA Sample Preparation and Nanopore Sequencing

Standard sample preparation including DNA extraction, whole genome amplification and nanopore library preparation (all samples presented here are “2D” reads) are described in detail in the supplementary methods. *In vitro* DNA methylation procedures are described in full in the supplemental methods.

Resolve Indels Using Raw Nanopore Observations

Raw nanopore signal is processed first by segmentation into “events”, followed by assigning bases to those segments and joining of selected neighboring segments. This produces estimated base calls that contain errors. We then align individual reads to a reference genome (based on the sample) or a *de novo* assembled genome. We then resolve differences between the reads and the underlying genome sequence by first re-segmenting the raw nanopore signal at genomic insertions and deletions, and then correcting for miscalled bases. Resolving the segmentation of the raw signal to match the known bases constitutes the base algorithm used for all analyses presented in this manuscript. The full algorithm description can be found in supplemental methods and is implemented and publicly available in the open source python package *nanoraw* via the *genome_resquiggle* subcommand (pypi.python.org/pypi/nanoraw; code repository https://github.com/marcus1487/nanoraw).

Statistical Testing for the Identification of Modified Nucleotides

With raw nanopore signal assigned to each genomic base, comparison of raw signal levels between two samples is reduced to a testing problem. In order to test the difference between two samples the mean signal for each read at a genomic base are computed. We propose the Mann-Whitney U-test[23] to test for differences in median signal intensity between two samples of interest. A robust order statistic-based approach is chosen as signal shifts near modified bases appear consistent, but not necessarily large in scale which other tests (e.g. t-test) have increased power to detect. The U-test is applied at every position across the genome with sufficient coverage (at least 5 reads in both samples). Since signal is affected at several bases surrounding a modified base, Fisher's method[24] for combining p-values is computed on a moving window (of 5 bases for all tests in this paper) to produce final p-values.

Signal Level Normalization

Nanopore sequencing is originally recorded as raw signal intensity values as a digital integer measuring the electric current across the nanopore. When the raw nanopore signal is converted to events (estimated base locations) Oxford Nanopore Technologies converts this integer signal value to an estimated picoamp (pA) level. Thus far, these measurements have been the gold standard for downstream analysis of nanopore signal levels.

To investigate and compare alternative normalization methods the following metric is proposed: fraction of variance unexplained by k-mer (here we use 4-mer). We use a median normalization procedure as the default for downstream signal processing. For a read with N raw signal observations (where R_i is the raw signal level at the i^{th} observation) we define the median normalized signal (M_i) as $M_i = (R_i - median_{j\in[1,N]}(R_j))/MAD$ where $MAD = \sum_{i=1}^{N} \left|R_i - median_{j\in[1,N]}(R_j)\right|$. All signal measurements presented in this manuscript have been median normalized (with the one exception for the comparison to pA normalization). In addition we winsorize the signal to clip aberrant spikes at plus and minus five MAD.

Modification Based Sequence Motifs

To identify the sequence preference for the sites identified from a given MoD-seq experiment we first identify the top 1,000 unique genomic locations based on computed Fisher's method p-values. Fifteen bases of context are included up and downstream around each identified location. The meme algorithm[30] is then applied to these sequences in the ZOOPS model. The top hit based on value is reported.

Clustering

Each region centered on a base with significantly deviated current (based on U-test p-value) is identified. All pairwise Euclidian distances between identified regions are computed based on a 5-vector of differences between native and amplified signal levels. The dimension reduction algorithm MDS is applied in order to visualize clustering of the data.

Acknowledgments

We thank Kenneth H. Wan, LBNL, for helpful comments during manuscript preparation. Research was conducted at the E.O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California. RN would like to acknowledge the EPSRC (EP/N020901/1) and the European Union’s Horizon 2020 research and innovation program under grant agreement No 634890, “BeyondSeq”.

Author Contributions

MS organized the project. MS conceived the analysis procedures with input from JB. MS developed and implemented the *nanoraw* software package and nanopore analysis pipelines. MS, RE and JB reviewed processing pipelines. RN completed *in vitro* methylation reactions. JQ prepared samples and libraries for all Univ. of Birmingham experiments. JL prepared samples and libraries for all LBNL experiments. NL managed work and study design at the Univ. Birmingham, and LP managed work and study design at LBNL. MS and JB prepared the manuscript with input from all authors. All authors reviewed and edited the manuscript.

References

1. Alberts, B. Molecular biology of the cell, Edn. Sixth edition. (Garland Science, Taylor and Francis Group, New York, NY; 2015).
2. Heyn, H. & Esteller, M. An Adenine Code for DNA: A Second Life for N6-Methyladenine. *Cell* **161**, 710-713 (2015).
3. Sun, Q. et al. N6-methyladenine functions as a potential epigenetic mark in eukaryotes. *Bioessays* **37**, 1155-1162 (2015).
4. Luo, G.Z., Blanco, M.A., Greer, E.L., He, C. & Shi, Y. DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? *Nat Rev Mol Cell Biol* **16**, 705-710 (2015).
5. Feng, Z. et al. Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput Biol* **9**, e1002935 (2013).

6. Beaulaurier, J. et al. Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat Commun* **6**, 7438 (2015).
7. Clark, T.A. et al. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res* **40**, e29 (2012).
8. de Souza, N. Protein nanopores to detect DNA methylation. *Nat Methods* **11**, 8 (2014).
9. Schreiber, J. et al. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proc Natl Acad Sci U S A* **110**, 18910-18915 (2013).
10. Laszlo, A.H. et al. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc Natl Acad Sci U S A* **110**, 18904-18909 (2013).
11. Simpson, J.T. et al. Detecting DNA Methylation using the Oxford Nanopore Technologies MinION sequencer. *bioRxiv* (2016).
12. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R. & Phillippy, A.M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv* (2016).
13. Sovic, I., Krizanovic, K., Skala, K. & Sikic, M. Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. *Bioinformatics* **32**, 2582-2589 (2016).
14. David, M., Dursi, L.J., Yao, D., Boutros, P.C. & Simpson, J.T. Nanocall: An Open Source Basecaller for Oxford Nanopore Sequencing Data. *Bioinformatics* (2016).
15. Boža, V.B., Broňa; Vinař, Tomáš DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads. *arXiv* **1603.09195** (2016).
16. Timp, W., Comer, J. & Aksimentiev, A. DNA base-calling from a nanopore using a Viterbi algorithm. *Biophys J* **102**, L37-39 (2012).
17. Technologies, O.N. (2016).
18. Judge, K. et al. Comparison of bacterial genome assembly software for MinION data. *bioRxiv* (2016).
19. Jain, M. et al. Improved data analysis for the MinION nanopore sequencer. *Nat Methods* **12**, 351-356 (2015).
20. Loman, N.J., Quick, J. & Simpson, J.T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* **12**, 733-735 (2015).
21. Quick, J., Quinlan, A.R. & Loman, N.J. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *Gigascience* **3**, 22 (2014).
22. Quick, J., Quinlan, A.R. & Loman, N.J. Erratum: A reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer. *Gigascience* **4**, 6 (2015).
23. Mann, H.B.W., D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18**, 50-60 (1947).
24. Fisher, R.A. Statistical methods for research workers, Edn. 7th. (Oliver and Boyd, Edinburgh,; 1938).
25. Fang, G. et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat Biotechnol* **30**, 1232-1239 (2012).
26. Marinus, M.G. & Morris, N.R. Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. *J Bacteriol* **114**, 1143-1150 (1973).

27. Geier, G.E. & Modrich, P. Recognition sequence of the dam methylase of Escherichia coli K12 and mode of cleavage of Dpn I endonuclease. *J Biol Chem* **254**, 1408-1413 (1979).
28. Hemminki, K., Koskinen, M., Rajaniemi, H. & Zhao, C. Dna adducts, mutations, and cancer 2000. *Regul Toxicol Pharmacol* **32**, 264-275 (2000).
29. O'Malley, R.C. et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280-1292 (2016).
30. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).
31. Kahramanoglou, C. et al. Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. *Nat Commun* **3**, 886 (2012).
32. Ip CLC, L.M., Tyson JR et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis [version 1; referees: 2 approved]. *F1000Research* **4**, 1075 (2015).
33. Laver, T. et al. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* **3**, 1-8 (2015).
34. Sovic, I. et al. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun* **7**, 11307 (2016).
35. Venta, K. et al. Differentiation of short, single-stranded DNA homopolymers in solid-state nanopores. *ACS Nano* **7**, 4629-4636 (2013).
36. Strimmer, K. A unified approach to false discovery rate estimation. *BMC Bioinformatics* **9**, 303 (2008).
37. Strimmer, K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* **24**, 1461-1462 (2008).

De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing

Supplemental Material

Supplemental Text

Overlap and reproducibility of identified modifications between replicates

Measuring and quantifying reproducibility of assays which identify modified nucleotides is inherently challenging. Unlike many biological assays where the top hits consistently show effect sizes much larger than the majority of other sites (such as ChIP-seq), all modified sites throughout the genome are essentially equally statistically powered (modulo a few factors). This means that if for example 50,000 sites within the *E. coli* genome are truly modified in 100% of tested DNA fragments then these 50,000 will be randomly re-ordered in terms of any statistical test from one replicate to the next. Additionally, the null distribution contains all other sites in the genome. For the relatively small *E. coli* genome this elicits ~8.5M strand specific tested sites. Under the null distribution p-values will randomly distribute even between 0 and 1. Thus by random chance, we would expect to see within the *E. coli* genome at least one non-modified base with a p-value as low as 10e-7. As the statistical test for the identification of truly modified sites is likely not powered to to this level, the 50,000 truly modified sites will be mixed within these randomly identified sites from the null distribution. As shown in the main text and Supp. Figures 7 and 8 strand-specific coverage has a strong effect on the reproducibility as increased coverage increases the statistical power.

In addition to these factors, methylation is well documented to change with cell growth phase[31]. Given that no attempt was made to synchronize the growth phase in either lab, this and other biological effects may contribute to the discovery of modified sites unique to one of the two experiments. Using a mixture model approach (Methods) to account for differential coverage, we estimate that genome-wide 67% of dam and dcm recognition sites are methylated within the Pennacchio lab data while only 41% of such sites are methylated within the Loman lab data (Supp. Figure 10). Globally we estimate 30% of sites within the Pennacchio lab sample show a shift in signal (indicating that they are in close proximity to a modified nucleotide) as compared to 6% of sites from the Loman lab (Supp. Figure 10).

Finally, given that the nanoraw MoD-seq pipeline does not have single base precision in the current implementation, sites within several bases of a modified base may all obtain significant p-values. These sites contribute additionally to a lack of overlapping significant sites between two replicates.

Supplemental Methods

Experiment Library Preparation (LBNL)

Total genomic DNA from *E. coli* str. K-12 substr. MG1655 was extracted using previously described methods[32]. In brief, DNA was extracted from approximately 4 x 10^9 log-phase cells using the QIAGEN Genomic-tip 20/G according to the manufacturer's instructions (Qiagen, Valencia, California). Whole genome amplification of *E. coli* total genomic DNA was performed using the QIAGEN REPLI-g Single Cell Kit according to the manufacturer's instructions (Qiagen). DNA was quantified using Qubit dsDNA BR assay (Life Technologies, Grand Island, New York).

2D sequencing libraries were prepared from native and amplified E. coli DNA according to the ONT recommended protocol (SQK-NSK007). In summary, DNA was fragmented using a Covaris g-TUBE (Covaris, Ltd., Brighton, United Kingdom). The fragmented DNA then underwent DNA damage repair using the FFPE DNA Damage Repair Kit (NEB, Ipswich, Massachusetts) and AMPure XP bead clean-up (Beckman Coulter, Brea, California). The DNA was end-repaired and A-tailed using the NEBNext Ultra II End Prep Kit (NEB). Following AMPure XP bead clean-up, adapters were ligated onto the DNA using Blunt/TA Ligase Master Mix (NEB). Libraries underwent a clean-up step using MyOne C1 Streptavidin beads (Life Technologies) and were quantified using Qubit dsDNA HS assay (Life Technologies). All sequencing runs were performed using R9 flow cells and MinION Mk1b devices with the standard MinKNOW 48-hour sequencing protocol. Metrichor was used to perform basecalling using the 2D Basecalling for FLO-MIN105 250bps workflow.

Experiment Library Preparation (University of Birmingham)

Total genomic DNA was isolated from three bacterial cell pellets (*S. aureus*, *M. smegmatis* and *E. coli* K-12) using the genomic buffer set and 500/G genomic tips (Qiagen) following the manufacturer's instructions and mixed in equal amounts. DNA was fragmented using a Covaris g-TUBE in a centrifuge at 5000 rpm. Part of the material was end-repaired and A-tailed using the NEBNext Ultra II End Prep Kit. Following AMPure XP bead clean-up, PCR adapters provided in the SQK-NSK007 kit (ONT) were ligated onto the fragments using Blunt/TA Ligase Master Mix (NEB). 10 ng of the cleaned-up, adapted fragments were PCR amplified using LongAmp Taq 2x Master Mix (NEB) and the primers provided in the SQK-NSK007 kit. Following 18 cycles of PCR fragments were cleaned-up and sequencing libraries were prepared for both PCR amplified and the native DNA set aside earlier according to the ONT recommended protocol (SQK-NSK007) described above. Sequencing runs were performed using R9 flow cells and MinION Mk1b devices with the standard MinKNOW 48-hour sequencing protocol. Metrichor was used to perform basecalling using the 2D Basecalling for FLO-MIN105 250bps workflow.

PCR amplified DNA that underwent methylase treatment were barcoded using the native barcoding kit (EXP-NBD002) so multiple treatments could be multiplexed on one flowcell. Approximately 200 ng input DNA for each treatment was barcoded and pooled according to the 2D Native barcoding genomic DNA protocol. A library was prepared from the pooled, barcoded fragments using the SQK-LSK208 kit according according to the ONT recommended protocol. Two sequencing runs were performed using R9.4 flow cells and MinION Mk1b devices with the standard MinKNOW 48-hour sequencing protocol. Metrichor was used to perform basecalling using the 2D Basecalling for FLO-MIN106 250bps workflow.

Synthetic DNA Modification

DNA methyltransferases were purchased from New England Biolabs and used according to the manufacturer's instructions. The exceptions are the M.MpeI (Chrometra, Belgium) and M.TaqI, which was expressed and purified by the Protein Expression Facility, Birmingham. DNA methylation was performed in vitro by incubating DNA (60 ng/uL) with 1 uL of methyltransferase and 80uM S-adenosyl-L-methionine in 50 uL of aqueous solution containing the appropriate methyltransferase buffer (NEB Cutsmart Buffer was used for both the M.MpeI and M.TaqI). Reactions were incubated at 37C for 1h (60C for 1h for M.TaqI) and then purified directly for sequencing using SPRI magnetic beads.

Resolve Indels Using Raw Nanopore Observations

Raw nanopore data produced by the Oxford Nanopore Technologies MinoION device is stored as a digital integer value that represent a measure of electric current as DNA passed through a nanopore (at a current rate of 4000 observations per second). As DNA passes through a nanopore this signal changes as some function of the local base pair composition of that DNA molecule. For DNA this function has been resolved with considerable accuracy by Oxford Nanopore technologies, but significant errors remain in the data (between 70%-90% accuracy reported though this depends strongly on the version of pore used[19, 33, 34]). These errors can make it difficult to process or interpret the signal associated with a particular position of interest on the genome as is common practice in genomic sciences. Thus a key step to more exact and confident interpretation is to resolve base calls made from this raw nanopore signal with a known or discovered consensus genome.

In order to address this problem the following algorithm is proposed and implemented in the *nanoraw* software package to assign contiguous segments of raw nanopore signal with genomic positions. Starting from the Oxford Nanopore Technologies base calls the first step is to align base calls to a provided genome (this genome could even have been discovered and assembled *de novo* from the same run[12, 13, 18]). Currently, *nanoraw* uses the graphmap algorithm[34], but any long read aligner could be used. Then stretches of correctly mapped regions (including matching and mismatched base pairs) are used to anchor the called nanopore segments to genomic bases. Then insertion and deletions (indels) from the genome to the base calls must be resolved to assign raw nanopore signal to the assigned genomic bases from the read alignment.

For insertions into the genome, there are segments of the raw nanopore signal that are assigned to base(s) that do not exist in the genome. When such a region is encountered the region is extended out to the neighboring segments and one new segment is determined from the raw signal (using the process described below). Conversely for deletions there are genomic bases that have no assigned signal. The region defined by events surrounding these deleted base calls are selected and the number of deleted base pairs plus one (for the two correctly aligned neighboring bases) segments are then identified from the raw signal in this region.

For both insertions and deletions the final stage is to identify a specified number of new segments within a stretch of raw signal. In order to accomplish this the running difference between the mean signal of neighboring regions (currently using 4 observation windows) are computed. The site with the largest difference in signal level is called as the first segment. Then the next highest site is chosen unless it is within 4 observations of a previously added segmentation site and this process is repeated until the requested number of segments are identified. It is possible that the requested number of segments cannot be identified, and in this case the neighboring correctly called segments are included into the region. If extending this indel region intersects another indel these entire regions are merged and re-segmented together.

This algorithm is currently implemented in the *genome_resquiggle* sub-command of the *nanoraw* software package.

Identify Regions of Interest

In addition to the Mann-Whitney U-test, the t-test is a supported test in the *nanoraw* software for convenience though we found better identification of known methylated site with the robust Mann-Whitney U-test. Additional regions of interest that are query-able by the *nanoraw* software are regions of maximal coverage, regions centered on a k-mer of interest (e.g. homopolymers have proven to be difficult to process in nanopore data[14, 35]), and regions with the largest raw difference in signal means between two groups of samples. These collection of region identification tools gives the nanopore investigator incredible power to interpret and further develop the potential for this technology.

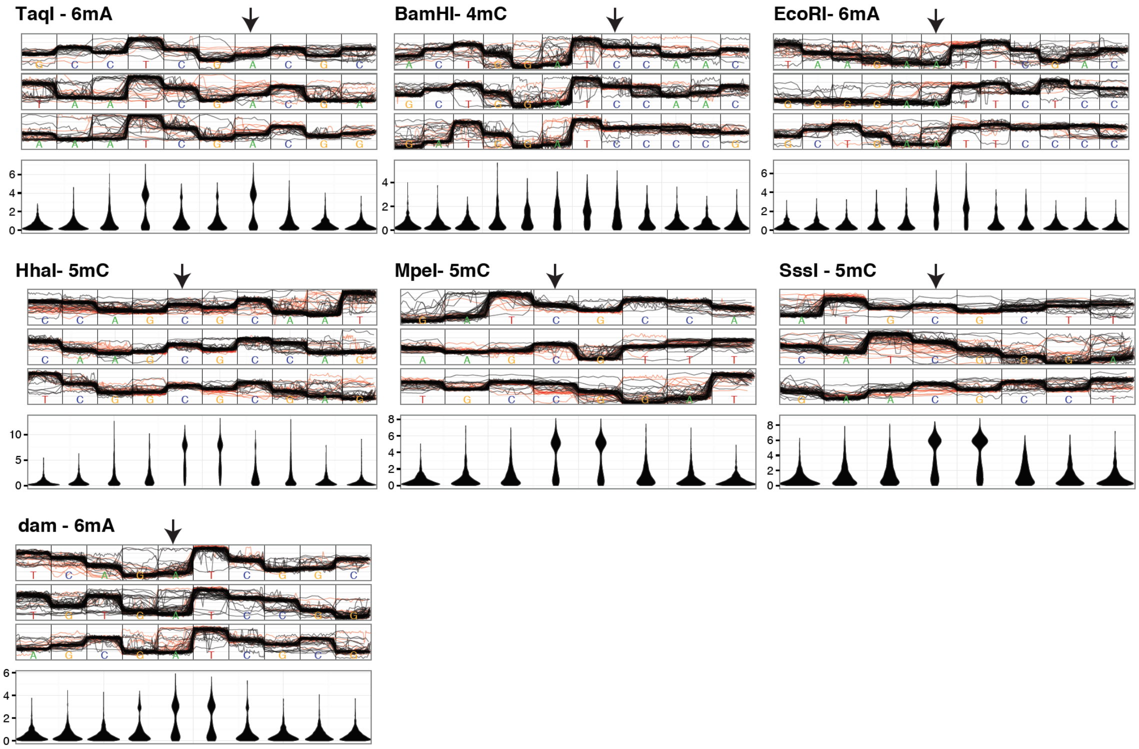
Filter to identify reads with most consistent and biologically relevant signal

In order to remove reads that appear to be of low quality we have developed a filter based on the number of observations per base (event). Given the resolution of the raw signal to the genomic alignment, we have much more accurate picture of how many observations are made per genomic base. Bases that contain many more observations indicate a “stuck” base. Sometimes this may be of biological interest, but signal level variance analysis indicate that the majority of reads with many “stuck” bases do not provide signal levels matching the trends for reads that pass through the pore at a consistently fast rate. We recommend a filter to remove any reads with greater than 5,000 observations at a single base or greater than 200 observations in more than 1% of the bases within a read and this filter is applied to all analyses presented in this paper.

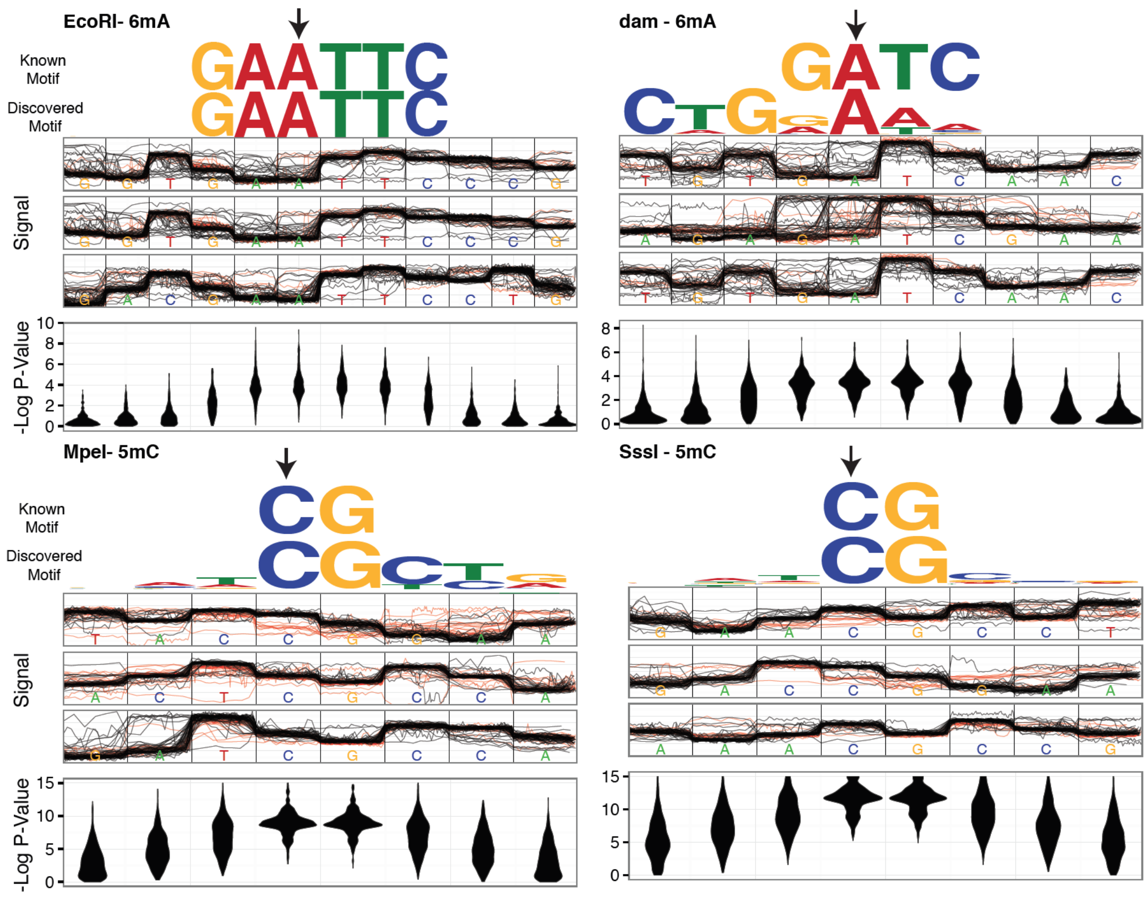
Mixture Model Percent Modified Bases Estimation

In order to estimate the fraction of positions with signal affected by a DNA modification we employ a mixture model implemented in the R package fdrtools[36, 37]. This model attempts model a distribution of p-values with a uniform component (which represents the false tests) and a monotonically increasing component (representing the true positive tests). Here we report one minus the estimated fraction composed within the null (uniform) component as the fraction of sites affected by modified bases.

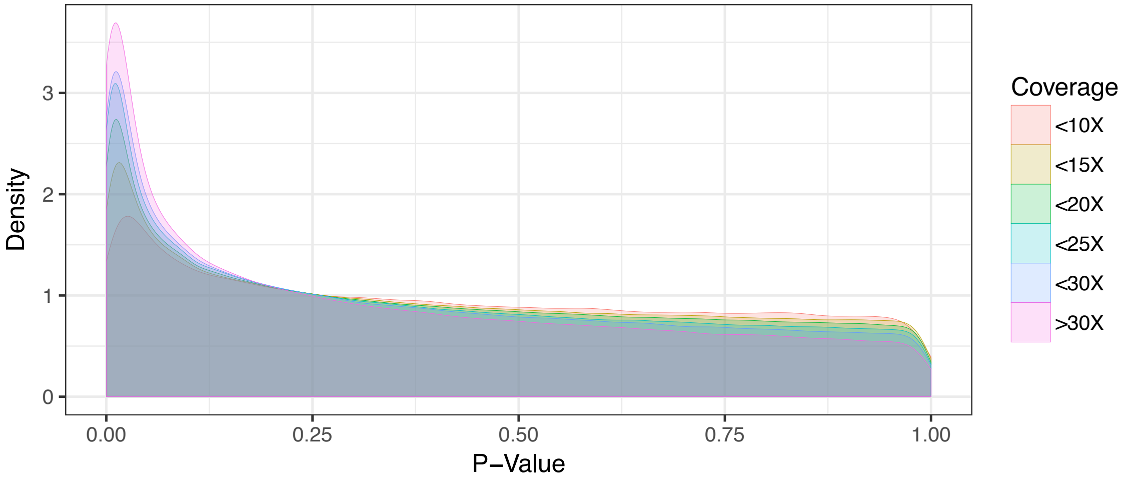
Supplemental Figures



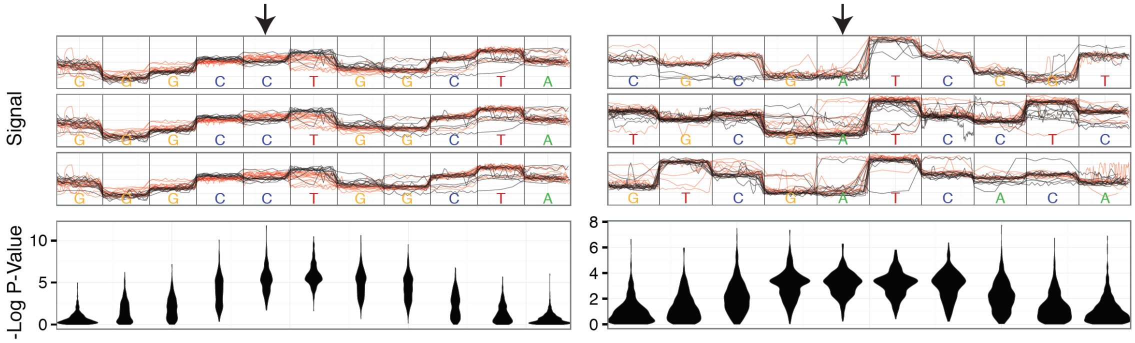
Supplemental Figure 1. U-test p-values for each known modification show significance at bases around the known modified base. Arrows indicate known methylation site.



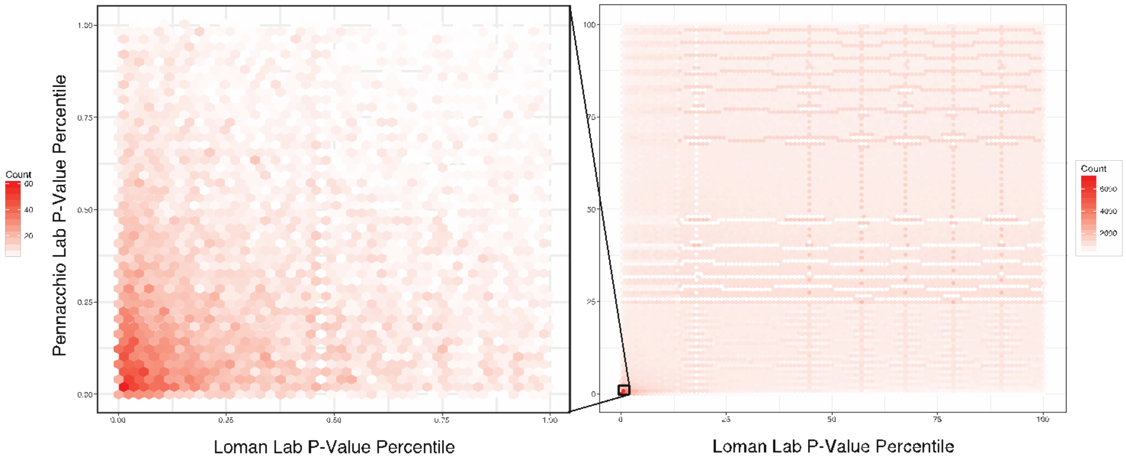
Supplemental Figure 2. Additional known methylase signal plots (as in Figure 3). Arrows indicate known methlyation site



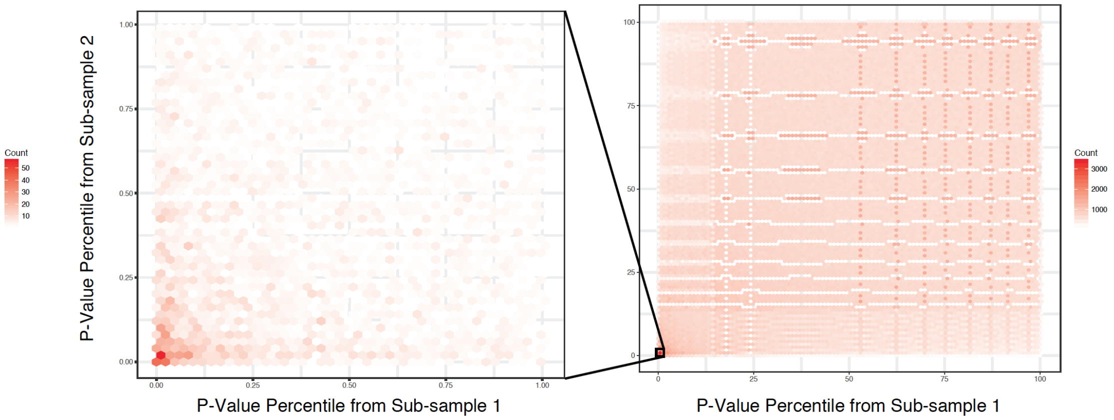
Supplemental Figure 3. Distribution of p-values at given thresholds of minimum (between native and amplified) coverage at a site. Pennacchio lab data used for this analysis.



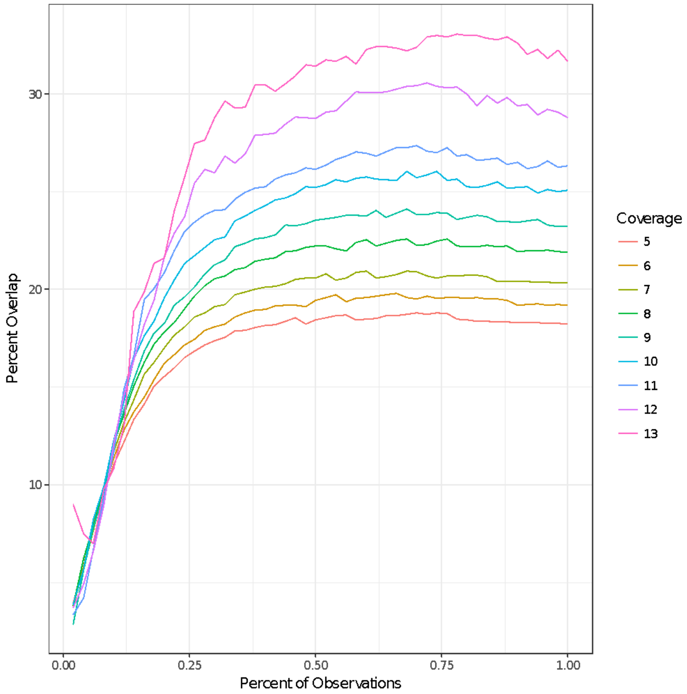
Supplemental Figure 4. Examples for two major classes of modifications from lower coverage data from Loman lab as well as p-value distributions for 1,000 most significant sites within the known motif (as in Figure 4A). Arrows indicate known methylation site.



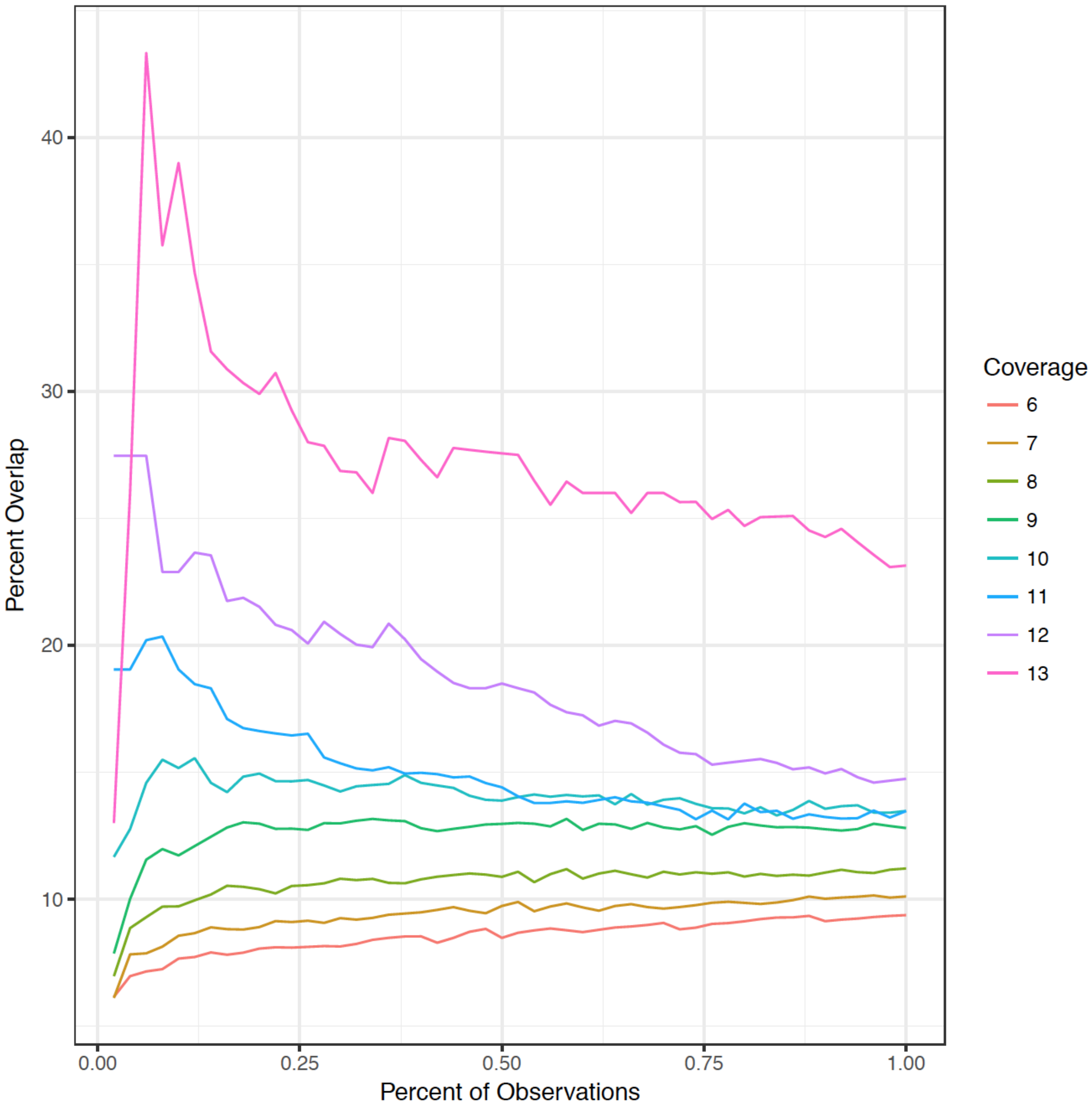
Supplemental Figure 5: Density of sites showing correspondence between two labs across rank lists (by p-value) from Loman and Pennacchio labs. Right panel is zoomed in to the top 1% of both lists.



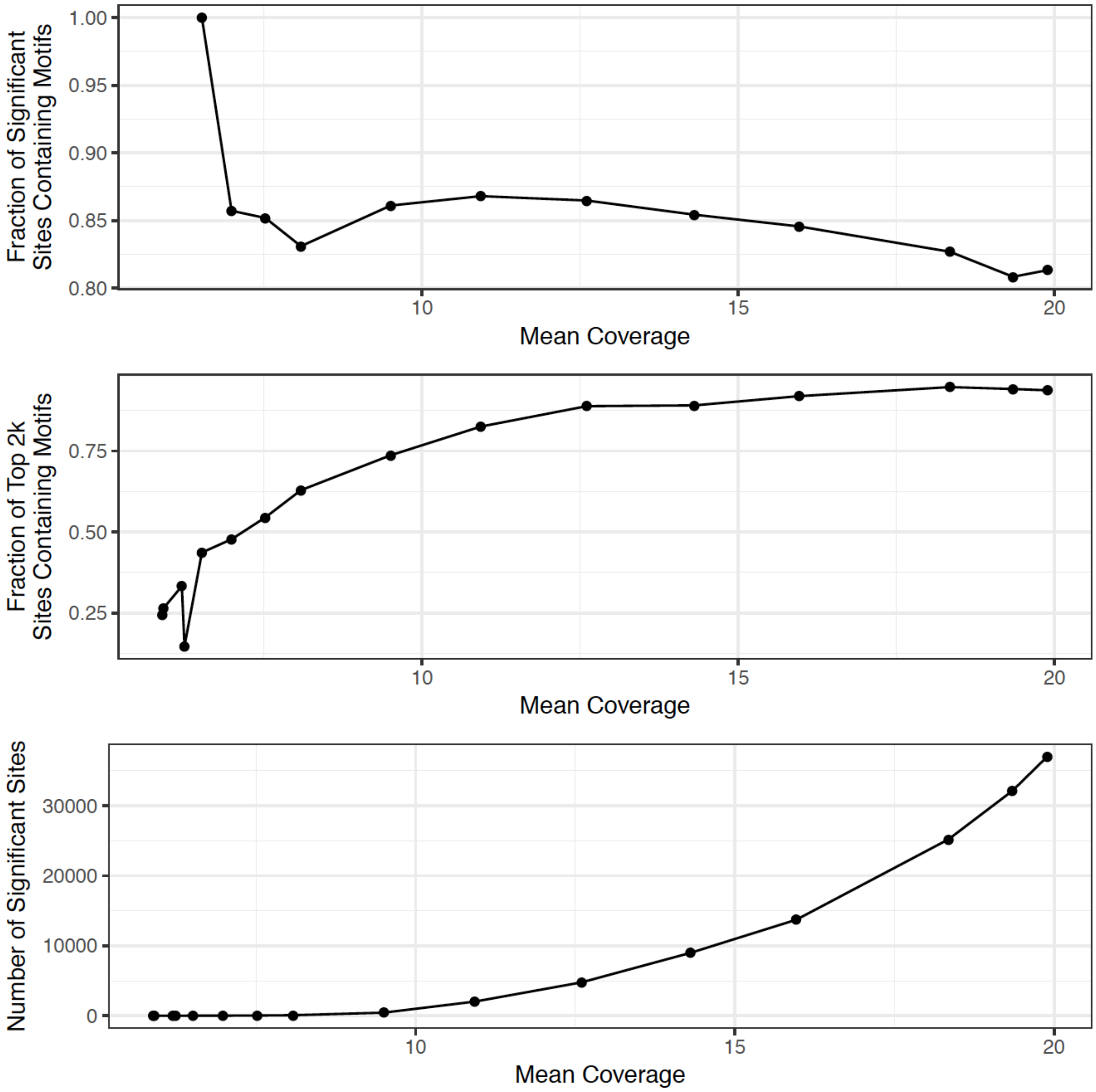
Supplemental Figure 6: Density of sites showing correspondence (as in Supp. Figure 5) between two random samples from the same experiment to achieve 10X strand-specific coverage across rank lists (by p-value). Right panel is zoomed in to the top 1% of both lists.



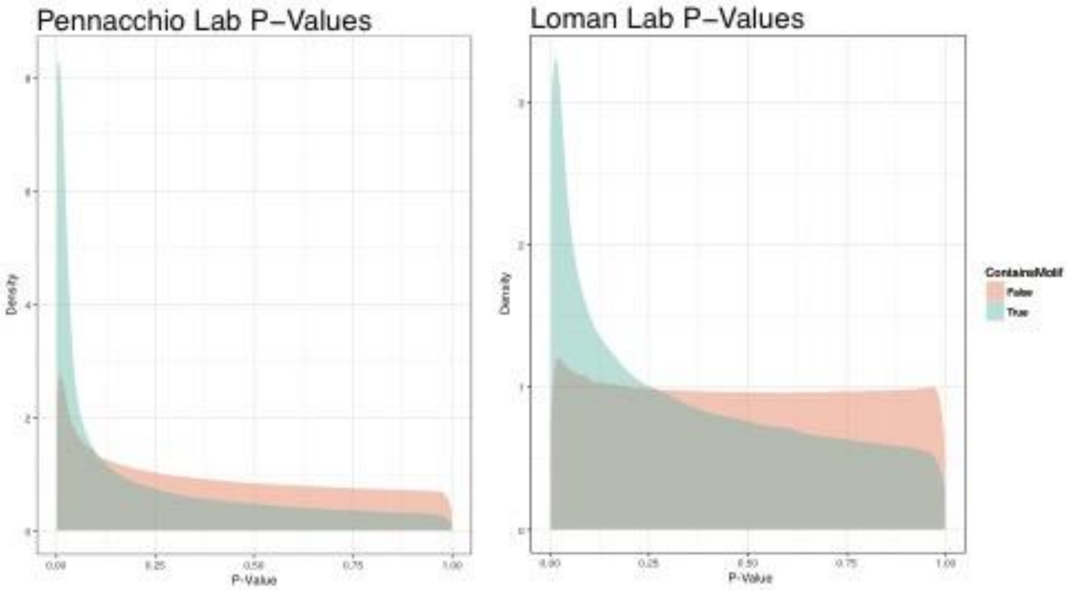
Supplemental Figure 7. Comparison of modified bases identified by entirely independent processing pipelines including labs, technicians and sources. The percentage of observations from the rank list produced from each lab (x axis) compared to the percentage of overlap between the two experiments (y axis). Different lines indicate the minimal coverage filter applied to test each base across all four sequencing experiments (native and amplified from both experiments).



Supplemental Figure 8. Comparison of modified bases identified (as in Supp. Figure 7) from two sub-samples from the same experiment to achieve ~10X strand-specific coverage. The percentage of observations from the rank list produced from each sub-sample (x axis) compared to the percentage of overlap between the two sub-samples (y axis). Different lines indicate the minimal coverage filter applied to test each base across all four sequencing experiments at each tested site (native and amplified from both sub-samples).



Supplemental Figure 9. Relationship between statistical power (depth of coverage) and fraction of identified sites with known motifs in native *E. coli* samples over range of down-sampling to achieve different levels of strand-specific coverage (x-axis).



Supplemental Figure 10. P-value distributions for Pennacchio (left) and Loman (right) labs across both regions that contain either dam or dcm motif and those regions that do not contain a motif.