



Detection and Identification of Base Modifications with Single Molecule Real-Time Sequencing Data

Patrick Marks, Onureena Banerjee, David Alexander

October 30, 2012

1 Introduction

Single Molecule Real-Time (SMRT[®]) DNA sequencing[1] data contains information about DNA base modifications imprinted in the kinetics of the polymerization reaction[2]. This data enables single-base resolution detection of ^m6A, ^m4C, and ^m5C[3]. We discuss the statistical approaches to the detection and identification of DNA modifications from SMRT sequencing data. Our implementation is available at <http://github.com/PacificBiosciences/kineticsTools>.

2 Preparation of IPD data

This analysis assumes that we are interested in locus-specific DNA modifications - that is, samples where the same modification is present at a given locus, in some fraction of the sample molecules.

We compute a vector of IPDs that map confidently to a given genomic location. We use the alignment tool BLASR [4] to generate alignments between reads and a reference template. BLASR maps each read to the reference sequence. BLASR generates a Mapping QV representing the confidence that the read maps uniquely to the selected genomic interval. Reads falling within repeats will have a low Mapping QV and are removed from the analysis to prevent incorrect modification calls inside repeats due to mismapped reads.

Sequencing errors in SMRT sequence data are predominately indels, which causes some ambiguity in the local placement of IPDs to genomic positions. To avoid most of these errors we only use IPDs if the SMRT sequence matches exactly for k bases around observed base. Currently we use $k = 1$.

These filtering steps yield a vector of IPDs confidently assigned to each genomic position.

The distribution of IPDs observed at a genomic location is roughly exponential, with a long tail caused by polymerase pausing inherent to SMRT sequencing. Information about base modifications is contained in the main distribution, while the contaminating pauses just contribute noise. We employ a simple capping procedure to reduce the influence of these outliers: $I_i^{(l)} = \min(R_i^{(l)}, Q_{99})$ where $I_i^{(l)}$ is the i th capped IPD at genomic location l , $R_i^{(l)}$ is the i th raw IPD and Q_{99} is the 99th percentile IPD over all IPD observations.

We summarize the capped IPDs at each positions with a sample mean and standard deviation of the mean:

$$\mu_l = \frac{1}{n_l} \sum_i R_i^{(l)} \quad (1)$$

$$\sigma_l^2 = \frac{1}{n_l} \sqrt{\sum_i (R_i^{(l)} - m_l)^2} \quad (2)$$

3 Control IPDs

Modification detection proceeds by comparing the observed mean IPD m_l with the expected mean IPD for unmodified DNA at that location. In case-control mode we sequence a sample of DNA from the same organism that has been *whole genome amplified* (WGA) to preserve the DNA sequence while removing all base modifications. In this mode of operation we summarize the IPDs observed in the control sample in the same way as the case sample, then proceed to the modification detection step. If a control sample is used we compute $\mu_l^{(c)}$ and $\sigma_l^{(c)}$ for each position on the control sample data as in Section 2.

3.1 *in-silico* Control

An alternative to the *case-control* paradigm is to construct an *in-silico* model to predict the mean IPD unmodified DNA given the local sequence context. We use the Gradient Boosting Machines [5] to construct a function $Q(\mathbf{c})$ that returns an estimate of the mean IPD given a DNA context $\mathbf{c} = \{c_{-8}, c_{-7}, \dots, c_0, c_1, \dots, c_3\}$ where $c_i \in \{A, C, G, T\}$ encodes the DNA sequence surrounding c_0 . The IPD prediction $Q(\mathbf{c})$ corresponds to the IPD of base c_0 . Initially we generated models spanning -15 to $+15$ base pairs to determine the appropriate size of the context vector. Figure 1 shows the GBM variable influence measure relative to the cognate base. We chose a context window of -8 to 3 by cutting off the window when the influence becomes small. The extent of the window agrees with the set of bases interacting with the polymerase in crystal structures. Each base c_i in the context vector becomes a feature available to GBM learning machinery. We chose GBM because it naturally handles categorical feature variables, and it automatically discovers complicated variable interactions. We a shrinkage (or learning rate) of $\lambda = 0.15$. We fit $T = 60000$ trees in our ensemble, each with a maximum interaction depth of $K = 12$. We used the `gbm` package[6] from R for initial experiments, and a custom port of the code for production training. The *in-silico* IPD model $Q(\mathbf{c})$ will have some bias compared to the true mean at some genomic location due to noise in the training dataset that was not generalized out in the training procedure, in or due to the residual influence of sequence outside of the context window.

$$\mu_l^{(c)} = Q(\mathbf{Ctx}(l)) \quad (3)$$

$$\sigma_l^{(c)} = f(\mu_l^{(c)}) \quad (4)$$

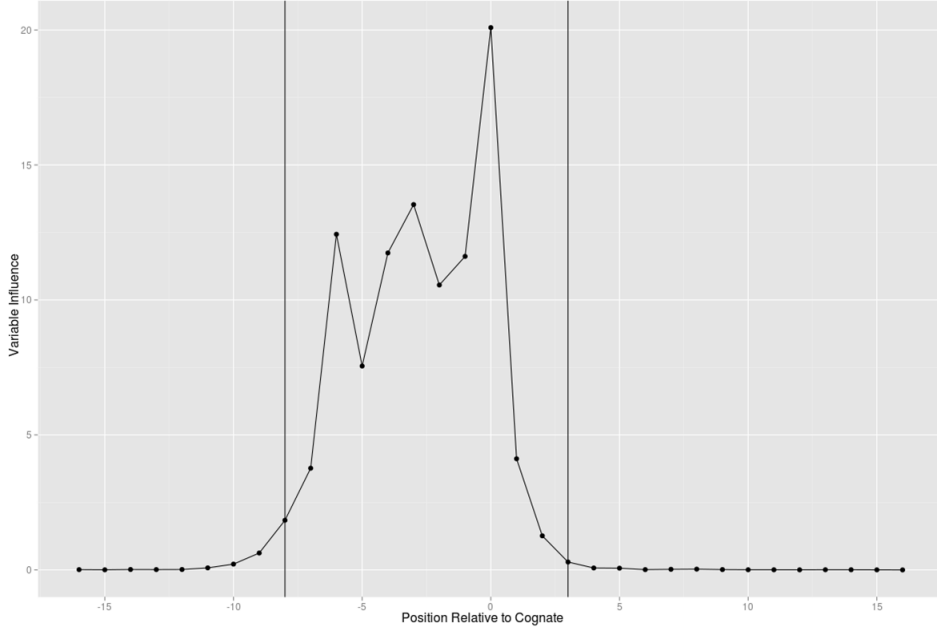


Figure 1: Influence of Base in Context Window

4 Modification Detection

We pose the modification detection problem as the detection of genome locations whose mean IPD differs significantly from the control mean. We use a Welch's t -test to test for differences in the means between the case sample and the control, derived from either a control sample or the *in-silico* model above.

The t -statistic is defined as:

$$s_l = \sqrt{\sigma_l^2 + \sigma_l^{(c)2}} \quad (5)$$

$$T_l = \frac{\mu_l - \mu_l^{(c)}}{s_l} \quad (6)$$

and we compute the p-value of T_l under the t -distribution, and report a Phred-transformed QV as well:

$$p = \Pr(t > T_l) \quad (7)$$

$$\text{QV} = -10 \log_{10} p \quad (8)$$

5 Modification Identification

5.1 Positive Control Model

We extend the *in-silico* model from an alphabet over DNA bases $c_i \in \{A, C, G, T\}$ to an alphabet that includes the *modified* bases we aim to identify: $c_i \in \{A, C, G, T, {}^{\text{m}6}\text{A}, {}^{\text{m}5}\text{C}, {}^{\text{m}4}\text{C}\}$. To train this model we require labeled examples of the modified bases in a reasonable diversity of background contexts. Base modifications appearing in bacterial restriction-modification systems provide an excellent source of training data that is fairly straightforward to label. A typical bacteria will express several (generally between 2 and 7, going as high as 20) methyltransferases that specifically methylate a particular sequence context in the host genome. The expression of the methyltransferase is generally paired with the expression of a restriction enzyme whose recognition site matches that of the methyltransferase. These *active* RM systems must methylate the genome nearly completely to prevent cutting of the host genome. Occasionally bacteria will carry an *orphan* methyltransferase that is not paired with a restriction enzyme, which often leads to weaker activity for that methyltransferase. We identify the *active* and *orphan* methyltransferases via a manual curation and use genomic positions modified by *active* RM systems in our training set. For example, the common 'GATC' motif, which carries an ${}^{\text{m}6}\text{A}$ modification on the A will lead to training context of the form $c = \dots, c_{-1} = G, c_0 = {}^{\text{m}6}\text{A}, c_1 = T, c_2 = T, c_3 = C$.

Our positive control training set incorporates data from 11 bacteria, with the following number of unique methylated motifs for each modification type – ${}^{\text{m}6}\text{A}$:36, ${}^{\text{m}4}\text{C}$:7, ${}^{\text{m}5}\text{C}$:7.

5.2 Viterbi Decoding

Let $\mathbf{S} = s_1, s_2, s_3, \dots, s_n$ be the unmodified DNA sequence ($s_i \in \{A, C, G, T\}$). Let $\mathbf{M} = m_1, m_2, m_3, \dots, m_n$ be a DNA sequence carrying modifications ($m_i \in \{A, C, G, T, {}^{\text{m}6}\text{A}, {}^{\text{m}5}\text{C}, {}^{\text{m}4}\text{C}\}$). A modification removal $R(m)$ operation maps a modification to its unmodified base - so $R({}^{\text{m}6}\text{A}) = A, R({}^{\text{m}5}\text{C}) = C, R({}^{\text{m}4}\text{C}) = C$. The modified sequence \mathbf{M} is constrained to maintain the base identity of the sequence: $R(m_i) = s_i$. We can model the likelihood of each observed IPD data point along the sequence μ_i in the same manner:

$$\log \Pr(\mathbf{O} \mid \mathbf{M}) = \sum_i \log \Pr(O_i \mid \mathcal{C}(\mathbf{M}, i))$$

where $\mathcal{C}(\mathbf{M}, i)$ is the *context* function that snips out the -8 to +3 bp context from sequence \mathbf{M} around position i . IPD observations O_i are assumed independent given the context $\mathcal{C}(\mathbf{M}, i)$. We seek to find the modification sequence $\hat{\mathbf{M}}$ that maximizes the likelihood of IPD observations:

$$\hat{\mathbf{M}} = \underset{\mathbf{M}}{\operatorname{argmax}} \sum_i \log \Pr(O_i \mid \mathcal{C}(\mathbf{M}, i))$$

Again we use the t-distribution to model the likelihood of the observed IPDs, given a mean prediction generated by the *in-silico* IPD model:

$$T_i = \frac{\mu_i - Q(\mathcal{C}(\mathbf{M}, i))}{s_i} \quad (9)$$

$$\Pr(O_i \mid \mathcal{C}(\mathbf{M}, i)) = f(T_i) \quad (10)$$

where $f(T)$ is the t -distribution PDF.

We find the maximum-likelihood modification sequence by applying the Viterbi algorithm[7]. At each position i we define $\mathbf{H}^{(i)}$, the set of all possible modification contexts centered at i with a unmodified sequence that matches the reference sequence. In order to reduce the algorithm run time, we reduce the size of $\mathbf{H}^{(i)}$ by only considering alternatives that have some supporting evidence in the nearby single-site p -values. Here we show the general formulation:

$$\mathbf{H}^{(i)} = \{\mathcal{C}(\mathbf{M}, i) \mid \mathbf{M}, R(m_i) = s_i\}$$

The Viterbi forward matrix $\alpha(H_j^{(i)}, i)$ is defined recursively: The first argument is the current state $H_j^{(i)}$ drawn from the possible modification configurations $\mathbf{H}^{(i)}$; the second argument is the position i .

$$\alpha(H_j^{(i)}, i) = \max_{K \in \mathbf{H}^{(i-1)}} \alpha(K, i-1) \Pr(O_i \mid H_j^{(i)}) \mathbf{SM}(K, H_j^{(i)})$$

$\mathbf{SM}(K, L) = \mathbf{1}\{K_1 = L_2, \dots, K_{11} = L_{12}\}$ is the *context matching function*, where K and L are 12 base context strings. It returns 1 if the last 11 bases of K match the first 11 bases of L , and 0 otherwise. This enforces the constraint that modification sequence contexts are self-consistent for all paths through the Viterbi matrix.

The standard Viterbi traceback procedure yields the maximum-likelihood modification state at each genomic position. We compute a Modification QV for each modification in the ML configuration by comparing the likelihood of the best modification sequence to the likelihood with a given modification set back to the canonical base:

$$p_i = \frac{\Pr(\mathbf{O} \mid R(\hat{\mathbf{M}}, i))}{\log \Pr(\mathbf{O} \mid R(\hat{\mathbf{M}}, i)) + \log \Pr(\mathbf{O} \mid \mathbf{M}')} \quad (11)$$

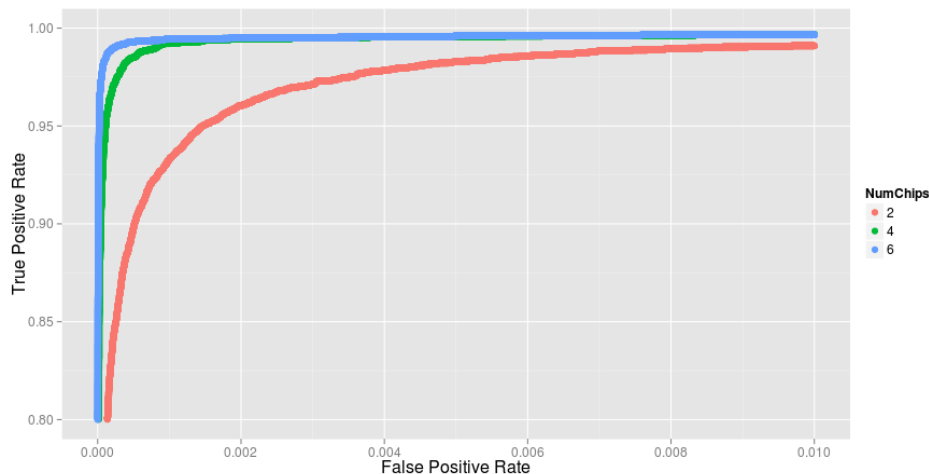
$$\text{QV}_i = -10 \log_{10} p_e \quad (12)$$

where $R(\hat{\mathbf{M}}, i)$ denotes the $\hat{\mathbf{M}}$ with base m_i converted to the unmodified base $R(m_i)$.

6 Results

6.1 Detection Performance

Figure 2 shows the ROC curve for single-site detection of m^6A in *E.coli*, for data from varying numbers of SMRT Cells. In these data we get the following coverage levels: 2 cells: 32x per strand, 4 cells: 65x per strand, 6 cells: 95x per strand. The *E.coli* strain tested here has three methylated motifs: GATC, GCACNNNNNGTT, AACNNNNNNGTGC, with a total of 39430 sites matching one of those motifs. Commonly, we observe a handful of sites that match a methylated motif without any detected methylation. Generally both strands of paired motif will be unmethylated. We don't account for this effect, causing the true positive rate to saturate below 100%.

Figure 2: ROC for m^6A detection in *E.coli* genome

Motif	Not Detected	m^4C	m^5C	m^6A	modified_base
CACC	3830	1222	10	0	114
GGCC	235	0	1065	0	4
GATC	66	0	0	4836	0
None		85	98	93	21799

Table 1: Modification Identification Confusion Matrix

6.2 Identification Performance

We tested the modification identification capability on the bacteria *Desulfurobacterium thermolithotrophum*, which carries RM systems using m^6A , m^5C , and m^4C modifications. The motif CACC is modified with m^4C , GGCC with m^5C , and GATC with m^6A . Table 1 shows that we can accurately separate m^4C and m^5C modifications, while accurately calling m^6A . The TET treatment protocol that amplifies the m^5C signal appears to negatively impact the strength of the m^4C signal. Work is in progress to mitigate this effect during the sample preparation.

References

- [1] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.
- [2] B.A. Flusberg, D.R. Webster, J.H. Lee, K.J. Travers, E.C. Olivares, T.A. Clark, J. Korlach, and S.W. Turner. Direct detection of dna methylation during single-molecule, real-time sequencing. *Nature methods*, 7(6):461–465, 2010.

- [3] T.A. Clark, I.A. Murray, R.D. Morgan, A.O. Kislyuk, K.E. Spittle, M. Boitano, A. Fomenkov, R.J. Roberts, and J. Korlach. Characterization of dna methyltransferase specificities using single-molecule, real-time dna sequencing. *Nucleic Acids Research*, 40(4):e29–e29, 2012.
- [4] Mark Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): Theory and application. *BMC Bioinformatics*, 13(1):238, 2012.
- [5] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29:1189–1232, 2001.
- [6] Greg Ridgeway. *gbm: Generalized Boosted Regression Models*, 2010. R package version 1.6-3.1.
- [7] G.D. Forney Jr. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

For Research Use Only. Not for use in diagnostic procedures. Copyright 2012, Pacific Biosciences of California, Inc. All rights reserved. Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at <http://www.pacificbiosciences.com/licenses.html>. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT and SMRTbell are trademarks of Pacific Biosciences in the United States and/or certain other countries. All other trademarks are the sole property of their respective owners.