

Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine

Chun-Xiao Song^{1,3}, Tyson A Clark^{2,3}, Xing-Yu Lu¹, Andrey Kislyuk², Qing Dai¹, Stephen W Turner², Chuan He¹ & Jonas Korlach²

We describe strand-specific, base-resolution detection of 5-hydroxymethylcytosine (5-hmC) in genomic DNA with single-molecule sensitivity, combining a bioorthogonal, selective chemical labeling method of 5-hmC with single-molecule, real-time (SMRT) DNA sequencing. The chemical labeling not only allows affinity enrichment of 5-hmC-containing DNA fragments but also enhances the kinetic signal of 5-hmC during SMRT sequencing. We applied the approach to sequence 5-hmC in a genomic DNA sample with high confidence.

The base 5-hydroxymethylcytosine (5-hmC) is a newly discovered DNA modification in mammalian cells and, along with 5-methylcytosine (5-mC), is believed to be an important epigenetic mark involved in many critical cellular functions, including embryonic stem cell differentiation, normal myelopoiesis as well as zygotic development^{1–3}. Understanding the biological functions of 5-hmC requires the development of sensitive sequencing methods to reveal locations of this base modification in the genome, as existing sequencing methods, such as bisulfite sequencing, cannot be used to differentiate 5-hmC from 5-mC (ref. 4).

Recently, we developed a selective chemical labeling technology for 5-hmC, in which 5-hmC is first modified with an azide-substituted glucose using β -glucosyltransferase followed by a click chemistry reaction to install a biotin tag⁵. Using this method, 5-hmC in genomic DNA has been enriched for deep sequencing to provide the genomic distribution of this base modification. However, this and other methods^{5–11} do not currently give information about the exact genomic locations of 5-hmC. Progress in the understanding of 5-hmC biology has been hampered by the lack of a method for high-throughput, strand-specific, base-resolution sequencing of 5-hmC.

Single-molecule, real-time (SMRT) DNA sequencing is a third-generation sequencing technology that uses individual DNA polymerase molecules to perform DNA synthesis, monitoring the continuous incorporation of phospholinked nucleotides¹². The real-time recording of nucleotide incorporations, detected as fluorescent pulses, generates not only the sequence

readout but also valuable information about the polymerase kinetics, which can be used to identify DNA base modifications. Typically, the polymerase rate at and around the modified base position in the DNA template is slowed compared to unmodified DNA. This can be expressed quantitatively by comparing the time between incorporation events, the interpulse duration (IPD), for each template position. Using this method, we have demonstrated that SMRT sequencing can be used to directly detect DNA methylation including N⁶-methyladenine, 5-mC and 5-hmC (ref. 13).

We combined the selective chemical labeling of 5-hmC and SMRT sequencing to provide a high-throughput, base-resolution 5-hmC detection method. The selective chemical labeling enables enrichment of 5-hmC-containing DNA to reduce the amount of sequencing required. In addition, the larger size of the tag on 5-hmC yields larger kinetic signals in SMRT sequencing for confident assignments of the modified base at lower sequencing coverage.

To achieve targeted enrichment and SMRT sequencing of 5-hmC (Fig. 1), we first used β -glucosyltransferase to transfer azide-glucose to 5-hmC, yielding β -6-azide-glucosyl-5-hydroxymethyl-cytosine (N₃-5-gmC), as described previously⁵. For the second step, we developed a new cleavable biotin-containing

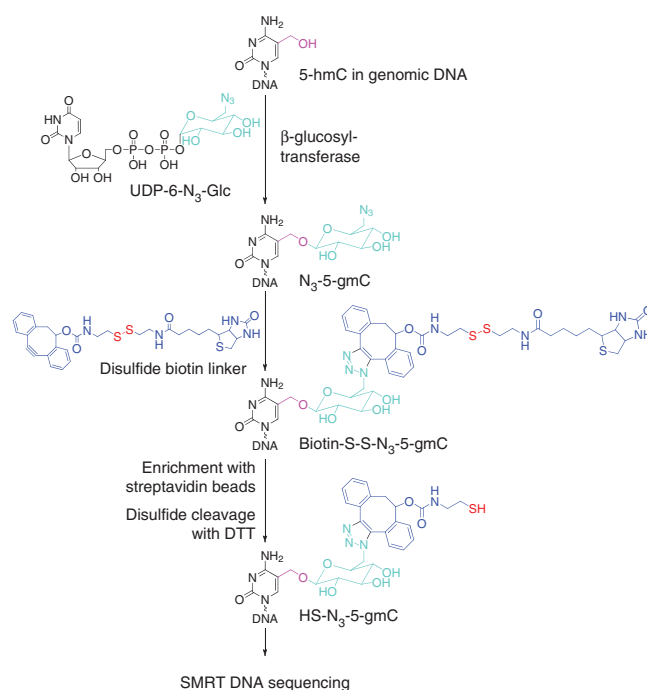


Figure 1 | Principle of selective chemical labeling of 5-hmC followed by SMRT DNA sequencing.

¹Department of Chemistry and Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois, USA. ²Pacific Biosciences, Menlo Park, California, USA.

³These authors contributed equally to this work. Correspondence should be addressed to C.H. (chuanhe@uchicago.edu) or J.K. (jkorlach@pacificbiosciences.com).

RECEIVED 5 APRIL; ACCEPTED 28 SEPTEMBER; PUBLISHED ONLINE 20 NOVEMBER 2011; DOI:10.1038/NMETH.1779

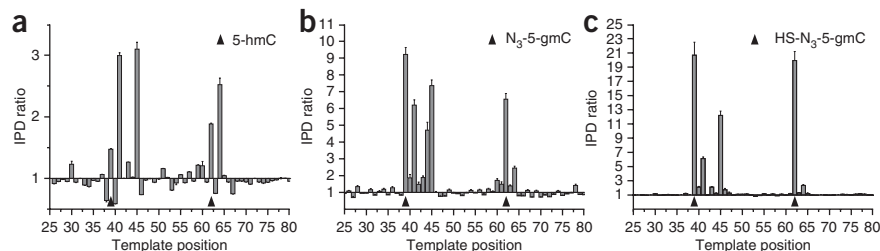


Figure 2 | Effects of 5-hmC modifications on polymerase kinetics in SMRT DNA sequencing. (a–c) Using synthetic DNA templates with known positions of 5-hmC (triangles), the graphs show IPD ratios, at each template position, of the modified DNA template over a control DNA template of identical sequence but lacking 5-hmC. Conditions were 5-hmC untreated (a), upon coupling with glucose azide (b) and upon additional coupling with the disulfide-containing biotin linker, followed by cleavage of the disulfide bond (c).

capture agent with a disulfide linker as the click reaction partner to form biotin-S-S-N₃-5-gmC to accommodate the sequencing protocol, as that method also uses biotin (for immobilization of DNA polymerase). After selectively capturing 5-hmC-containing DNA fragments from genomic DNA by streptavidin beads, a simple dithiothreitol (DTT) treatment releases the bound DNA fragments of interest, with 5-hmC modified as HS-N₃-5-gmC. Using synthetic DNA templates, we found this DTT-mediated cleavage to be quantitative and confirmed efficient conversion of all other reaction steps, as described previously⁵ (Supplementary Fig. 1). The pulldown yield for DNA fragments containing only a single 5-hmC was ~50%, consistent with the much lower density-dependence of biotin-based pulldown

methods, compared to antibody-based immunoprecipitation⁸. The approach is highly specific to 5-hmC, as no 5-mC-containing DNA was pulled down as measured by UV absorbance (Supplementary Table 1). The disulfide-reduction strategy to release desired DNA fragments was also less time-consuming and more efficient than the previous monomeric avidin column-based purification method⁵, increasing the pulldown efficiency by two- to threefold (Supplementary Table 1). To test the method, we subjected synthetic DNA templates with known 5-hmC positions to this selective chemical labeling protocol and tested the effect of the various modifications on kinetic signatures during SMRT sequencing. The 5-hmC itself gave an increase in IPD values of about two- to threefold compared to an unmodified control template (Fig. 2a), as observed previously¹³. The addition of azide-glucose to form N₃-5-gmC resulted in a substantial increase in the kinetic signature with IPD ratios of ~7–9 (Fig. 2b). The cleaved biotin linker adduct HS-N₃-5-gmC resulted in an even stronger kinetic signature with IPD ratios of ~7–25 (Fig. 2c), allowing the clearest detection of the modification. As previously described, the kinetic signatures extended over a region around the 5-hmC position (over a range starting ~1 base before the modified base and ending ~7 bases after it) and were sequence context-dependent¹³. For the final

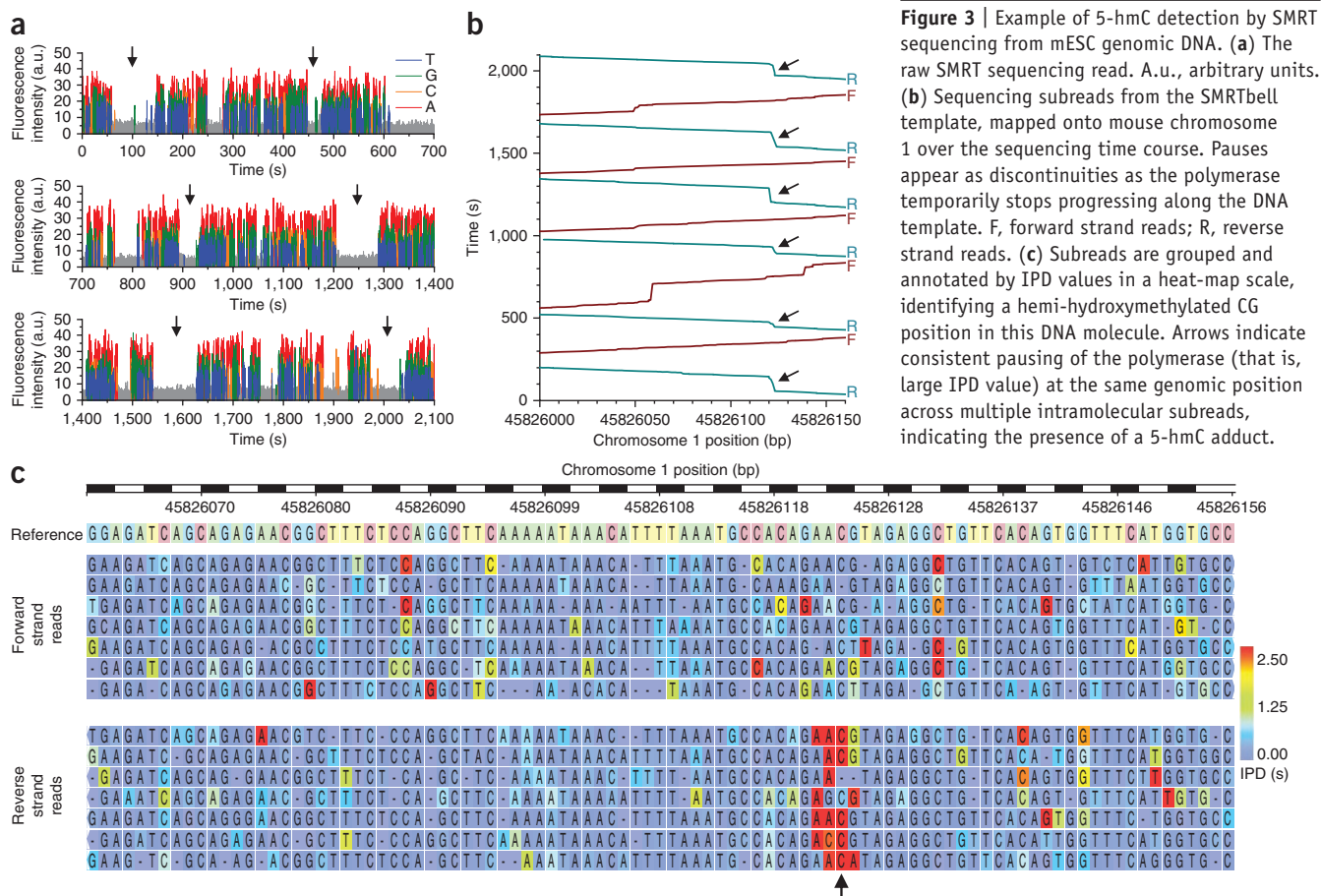


Figure 3 | Example of 5-hmC detection by SMRT sequencing from mESC genomic DNA. (a) The raw SMRT sequencing read. A.u., arbitrary units. (b) Sequencing subreads from the SMRTbell template, mapped onto mouse chromosome 1 over the sequencing time course. Pauses appear as discontinuities as the polymerase temporarily stops progressing along the DNA template. F, forward strand reads; R, reverse strand reads. (c) Subreads are grouped and annotated by IPD values in a heat-map scale, identifying a hemi-hydroxymethylated CG position in this DNA molecule. Arrows indicate consistent pausing of the polymerase (that is, large IPD value) at the same genomic position across multiple intramolecular subreads, indicating the presence of a 5-hmC adduct.

adduct, HS-N₃-5-gmC, we observed characteristic secondary peaks for most sequence contexts 2 and 6 bases downstream of the 5-hmC position. This information can be used algorithmically to increase the confidence of 5-hmC assignments.

Because the forward and reverse strand of a double-stranded DNA template molecule are interrogated independently in a SMRT sequencing read, our approach can differentiate hemi- and fully hydroxymethylated DNA. To demonstrate this, we sequenced synthetic DNA templates with known 5-hmC modifications on either or both strands and found that kinetic signals were strictly limited to sequencing reads obtained from template strands carrying the modification (**Supplementary Fig. 2**).

To test the feasibility of this approach for biological samples, we applied this method to detect 5-hmC in genomic DNA from mouse embryonic stem cells (mESC), one of the first cell types in which 5-hmC had been detected by bulk methods^{2,5}. After selectively labeling and enriching 5-hmC-containing DNA fragments, we prepared them into SMRTbell DNA libraries¹⁴ for shotgun SMRT sequencing and analyzed them for kinetic signatures. We detected clear signals for HS-N₃-5-gmC (**Supplementary Table 2**), as shown by the example of a hemi-hydroxylated position on mouse chromosome 1 (**Fig. 3**). Because of the greatly increased magnitude of the kinetic signal, assignments of 5-hmC positions can be made for single-molecule reads and do not require comparison with an amplified, unmodified control sample.

Positions of HS-N₃-5-gmC are visible as extended pauses in a raw sequencing read (**Fig. 3a**). Pauses corresponding to 5-hmC positions can be distinguished from occasional stochastic pauses¹⁵ by taking advantage of circular consensus sequencing¹⁴, in which the same base of a DNA molecule is interrogated multiple times in consecutive subreads on the topologically circular SMRTbell template. This allows for multiple occurrences of 5-hmC detection for high-confidence assignments; in the example shown, the polymerase pauses consistently at a particular template position in all reverse-strand subreads (**Fig. 3b**). The absence of pausing at this CG position in the corresponding forward strand subreads implies that this genomic location was hemi-hydroxymethylated. Grouping of forward- and reverse-strand subreads facilitates strand-specific 5-hmC identification by kinetic consensus and eliminates the effects of raw subread errors by virtue of generating a circular consensus sequence¹⁴ (**Fig. 3c**). From initial sequencing of this sample, we found high-confidence genomic positions displaying 5-hmC signatures across many different sequence contexts containing CG (**Supplementary Table 2**). The sequenced DNA fragment sizes ranged from ~100 bases to ~400 bases, allowing the observation of multiple 5-hmC occurrences on the same DNA molecule (**Supplementary Table 2**). We are now refining the 5-hmC detection algorithm by characterizing sequence-context influences and effects of closely spaced 5-hmC positions on the kinetic signatures.

We also validated four high-confidence hits in the CCGG context as well as an instance of the CCGG context that we observed to have no 5-hmC using a low-throughput commercial 5-hmC detection assay. The assay is based on PCR amplification of a specific sequence after glucosylation-sensitive restriction digestion (**Supplementary Fig. 3**).

In conclusion, we presented to our knowledge the first high-throughput sequencing method for 5-hmC by taking advantage of

selective chemical labeling of 5-hmC and SMRT sequencing. We anticipate that the method can be combined with the detection of 5-mC, which also causes a kinetic signature in SMRT sequencing, albeit currently at a smaller amplitude so that more fold coverage is required for high-confidence assignments. We are now applying this method to sequence 5-hmC across the entire mESC genome. With an estimated abundance of ~1,500,000 5-hmC positions (~0.05% of all nucleotides) and considering ~100-bp DNA fragments sequenced, the total genomic fraction amounts to ~150 Mb. This compares well with a current SMRT sequencing throughput of ~50–100 Mb per run. The above estimate does not take into account likely clustering of 5-hmC, which would reduce the overall sequencing required. Additional improvements in the throughput of sequencing and increases in read length and accuracy are also expected to reduce the number of sequencing runs. We anticipate that this technology will be a powerful method to explore the biology of 5-hmC.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturemethods/>.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This study was supported partly by US National Institutes of Health grants GM071440 (to C.H.) and 1RC2HG005618-01 (National Human Genome Research Institute to S.T.) and The University of Chicago. We thank K. Spittle, M. Boitano, J. Eid, J. Wegener and K. Luong for assistance in data acquisition and figure generation.

AUTHOR CONTRIBUTIONS

C.H., C.-X.S., J.K. and S.W.T. designed experiments. C.-X.S. performed the labeling and pulldown of synthetic template and mESC sample. T.A.C. prepared library constructs and conducted the sequencing experiments. A.K. analyzed data. Q.D., C.-X.S. and X.-Y.L. carried out the chemical synthesis. X.-Y.L. validated mESC hits. C.H., C.-X.S. and J.K. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturemethods/>.

Published online at <http://www.nature.com/naturemethods/>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Kriaucionis, S. & Heintz, N. *Science* **324**, 929–930 (2009).
- Tahiliani, M. *et al. Science* **324**, 930–935 (2009).
- Bhutani, N., Burns, D.M. & Blau, H.M. *Cell* **146**, 866–872 (2011).
- Huang, Y. *et al. PLoS ONE* **5**, e8888 (2010).
- Song, C.X. *et al. Nat. Biotechnol.* **29**, 68–72 (2011).
- Wu, H. *et al. Genes Dev.* **25**, 679–684 (2011).
- Ficz, G. *et al. Nature* **473**, 398–402 (2011).
- Pastor, W.A. *et al. Nature* **473**, 394–397 (2011).
- Williams, K. *et al. Nature* **473**, 343–348 (2011).
- Xu, Y. *et al. Mol. Cell* **42**, 451–464 (2011).
- Munzel, M., Globisch, D. & Carell, T. *Angew. Chem. Int. Ed.* **50**, 6460–6468 (2011).
- Eid, J. *et al. Science* **323**, 133–138 (2009).
- Flusberg, B.A. *et al. Nat. Methods* **7**, 461–465 (2010).
- Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S. & Turner, S.W. *Nucleic Acids Res.* **38**, e159 (2010).
- Korlach, J. *et al. Methods Enzymol.* **472**, 431–455 (2010).



ONLINE METHODS

Preparation of genomic DNA, synthetic DNA templates and β -glucosyltransferase. Mouse feeder-free E14Tg2A embryonic stem cell (mESC) genomic DNA was prepared as previously described⁵. Oligonucleotides were synthesized and synthetic DNA templates were prepared as previously described^{13,16}. β -glucosyltransferase (β -GT) was prepared as previously described⁵.

The 5-hmC labeling reaction and chemistry. For synthetic DNA, the 5-hmC labeling reactions were performed in a 20- μ l solution containing 50 mM HEPES buffer (pH 7.9), 25 mM $MgCl_2$, 2 μ g DNA template, 100 μ M UDP-6- N_3 -Glc and 1 μ M β -GT. The reactions were incubated for 1 h at 37 °C. After the reaction, the DNA substrates were purified by Qiagen Nucleotide removal kit and eluted in H_2O . For mESC DNA, the 5-hmC labeling reactions were performed in a 400- μ l solution containing 50 mM HEPES buffer (pH 7.9), 25 mM $MgCl_2$, 200 μ g sonicated genomic DNA (100–500 bp), 100 μ M UDP-6- N_3 -Glc and 2 μ M β -GT. The reactions were incubated for 1 h at 37 °C. After the reaction, the DNA substrates were purified by Bio-Rad Micro Bio-Spin 6 spin column and eluted in H_2O . The click chemistry was performed with addition of 150 μ M dibenzocyclooctyne with disulfide biotin linker into the DNA solution, and the reaction mixture was incubated for 2 h at 37 °C. The DNA samples were then purified by Qiagen nucleotide removal kit and eluted in H_2O .

Affinity enrichment of the modified 5-hmC DNA fragments (HS-5- N_3 -gmC). Invitrogen Dynabeads MyOne Streptavidin C1 was used to pull down the biotinylated DNA following the manufacturer's instructions with minor modifications. Binding and washing buffer for coupling of nucleic acids was used with the addition of 0.01% Tween-20. We used 100 μ l of beads, following the general washing procedure and immobilization of nucleic acids from the manufacturer's manual. To cleave the disulfide bonds and release the 5-hmC-containing fragments, beads were incubated in 100 μ l of 50 mM DTT in H_2O for 2 h at room temperature (25 °C) using gentle rotation. Beads were segregated with a magnet, leaving the 5-hmC-containing DNA in solution. The released DNA solution was then applied to Bio-Rad Micro Bio-Spin 6 spin columns to remove DTT. Finally, a Qiagen MinElute PCR Purification kit was used to purify the DNA solution and elute into EB buffer. From 160 μ g labeled mESC genomic DNA, 1.3 μ g pulldown DNA was obtained for SMRTbell library construction.

Pull-down specificity test. Synthetic 32-mer double-stranded DNA bearing 5-mC or 5-hmC was treated with the labeling process. We purified 2.2 μ g of each labeled DNA by 50 μ l Invitrogen Dynabeads MyOne Streptavidin C1 as described above. The pull-down DNA was purified using the Qiagen Nucleotide Removal Kit and eluted in 30 μ l H_2O . We pulled down 1,080 ng of 5-hmC-containing DNA as determined by NanoDrop (pulldown yield was ~50%). No 5-mC-containing DNA was pulled down as measured by NanoDrop. Sequences of the oligonucleotides are listed in **Supplementary Table 3**.

SMRT DNA sequencing. Synthetic 5-hmC-containing SMRTbells were made by annealing complementary oligonucleotides with specific four-nucleotide overhangs. Annealed oligonucleotides

were ligated at 25 °C for 60 min with T4 DNA Ligase (New England Biolabs). Incompletely ligated SMRTbells were degraded with 16 U μ g⁻¹ exonuclease III (New England Biolabs) and 1 U μ g⁻¹ exonuclease VII (USB) at 37 °C for 30 min. Synthetic SMRTbells were purified using the QIAquick PCR Purification kit (Qiagen). Sequences of the oligonucleotides are listed in **Supplementary Table 3**. All oligonucleotides contained a 5' phosphate. Control oligonucleotides contained a standard cytosine in place of the 5-hmC modification. Genomic SMRTbell libraries were generated using the A-tailing method as previously described¹⁴. Average insert size was ~200 bp. SMRT DNA sequencing was carried out on the PacBio RS using standard protocols for small insert libraries (Pacific Biosciences). Reads were processed and mapped to the mouse genome (mm9, retrieved from the University of Santa Cruz (UCSC) Genome Browser (<http://www.ncbi.nlm.nih.gov/pubmed/20959295>) using the basic local alignment with successive refinement (BLASR) mapper (<http://www.pacbiodevnet.com/SMRT-Analysis/Algorithms/BLASR>) and the Pacific Biosciences SMRTAnalysis pipeline (<http://www.pacbiodevnet.com/SMRT-Analysis/Software/SMRT-Pipe>) using the standard no-filter mapping protocol. IPDs were measured as previously described¹³ for all pulses aligned to each position in the genome. For each position with per-strand coverage exceeding 5 $\times$, the distribution of IPDs was compared against the expected IPD distribution using a likelihood ratio goodness-of-fit test, applied separately for forward and reverse strand reads. Positions in the genome were ranked according to the outcome of the statistical test on the forward strand and reverse strand (hemi-hydroxymethylation) as well as both strands together (full hydroxymethylation). Results were analyzed visually in the SMRTView genome browser, and a scoring cutoff was selected to generate a set of high-confidence single-nucleotide 5-hmC predictions. The sequencing data with associated kinetic parameters, the analysis tools used to extract 5-hmC positions and detailed instructions on how to use the kinetic analysis tools are available at <http://pacificbiosciences.com/devnet/files/datasets/publications/dna-modification-hmc/1.0/index.html>.

To validate the consistency of our high-confidence 5-hmC predictions with previously published methods, we surveyed the recent publications on 5-hmC sequencing and compared the identified regions of interest with our reported 5-hmC positions. Hydroxymethylated DNA immunoprecipitation (hMeDIP) sequencing of mouse E14 embryonic stem cells has been published in reference 9, reporting the targets of the *TET1* gene and its partner SIN3A complex followed by Illumina sequencing, with a total of ~500 Mb of the mouse genome loci tags, and in reference 6, using ChIP-chip followed by Affymetrix arrays, covering approximately 80 Mb of mouse genome sequence. We cross-checked our predictions with the set of all genetic loci identified in reference 9 and found that 52% of our 5-hmC position predictions overlapped with this set. By comparison, only 35% of the sequence regions identified in reference 6 were covered in reference 9.

Combined glucosylation and restriction analysis validation of 5-hmC hits. One microgram unsonicated mESC genomic DNA was glucosylated with regular glucose using β -GT. MspI digestion was performed following the manufacturer's manual (New England Biolabs). PCRs were performed in a 50- μ l solution containing 50 ng genomic DNA, 2.5 units REDTaq DNA polymerase (Sigma-Aldrich), 1 $\times$ reaction buffer, 200 μ M each dNTP



and 0.2 μM of each primer (primer sequences in **Supplementary Table 3**). PCR cycling conditions were as follows: 94 °C for 3 min and then 25 cycles of PCR at 94 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s, followed by a final extension step at 72 °C for 3 min. The PCR products were separated by electrophoresis on 3% NuSieve 3:1 agarose gels (Lonza).

Chemical synthesis. UDP-6- N_3 -Glc was prepared as described previously⁵. The disulfide biotin linker was chemically synthesized from biotinyl cystamine (Santa Cruz Biotech) as described previously⁵. Chemical shifts (δ) were reported in p.p.m. All coupling constants (J values) were reported in hertz. Split pattern abbreviations are as follows: singlet (s), doublet (d), triplet (t), multiplet (m) and doublet of doublets (dd). ^1H NMR (500 MHz,

CD_3OD): δ 7.50 (1 H, aromatics), 7.34–7.23 (7 H, aromatics), 5.35, (1 H, s, CHO), 4.36, 4.17 (m, 2 H, CHNH), 3.39 (m, 1 H, CHS), 3.36 (t, 2 H, $J = 6.75$ Hz, CH_2S_2), 3.23 (m, 4 H, CH_2NH), 3.15, (1 H, dd, $J_1 = 15$ Hz, $J_2 = 2$ Hz, CH_2), 3.08 (m, 1 H, CHHexoS), 2.81, (1 H, dd, $J = 4.5, 12.5$ Hz, CH_2), 2.74 (t, 2 H, S_2CH_2), 2.60 (d, 1 H, $J = 12.5$ Hz, CHHendoS), 2.10 (t, 2H, $J = 7.5$ Hz, CH_2CO), 1.5 (m, 6 H, biotin- CH_2). ^{13}C NMR (125 MHz, CD_3OD): δ 174.8, 164.7, 156.6, 152.2, 151.0, 129.7, 128.0, 127.9, 126.9, 126.6, 125.8, 125.5, 123.6, 121.0, 112.4, 109.6, 76.7, 61.9, 60.2, 55.6, 45.8, 39.8, 39.6, 38.0, 37.6, 37.4, 35.3, 28.3, 28.0, 25.4; high-resolution mass spectroscopy (HRMS): m/z 625.1990 [$\text{M} + \text{H}^+$]. Calculated for $\text{C}_{31}\text{H}_{36}\text{N}_4\text{O}_4\text{S}_3$ m/z 624.1899.

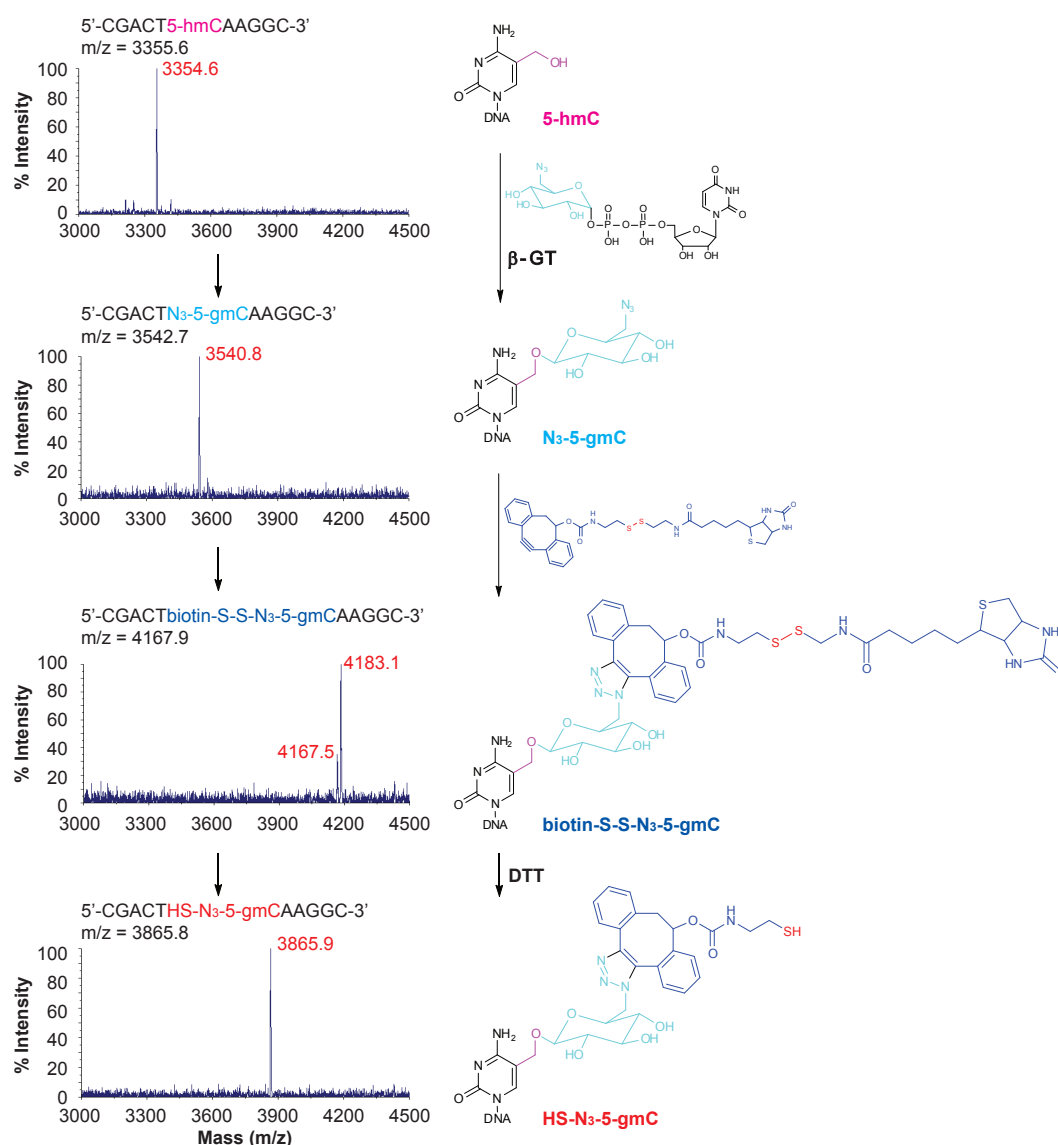
16. Dai, Q., Song, C.X., Pan, T. & He, C. *J. Org. Chem.* **76**, 4182–4188 (2011).

Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine

Chun-Xiao Song, Tyson A Clark, Xing-Yu Lu, Andrey Kislyuk, Qing Dai, Stephen W Turner, Chuan He & Jonas Korlach

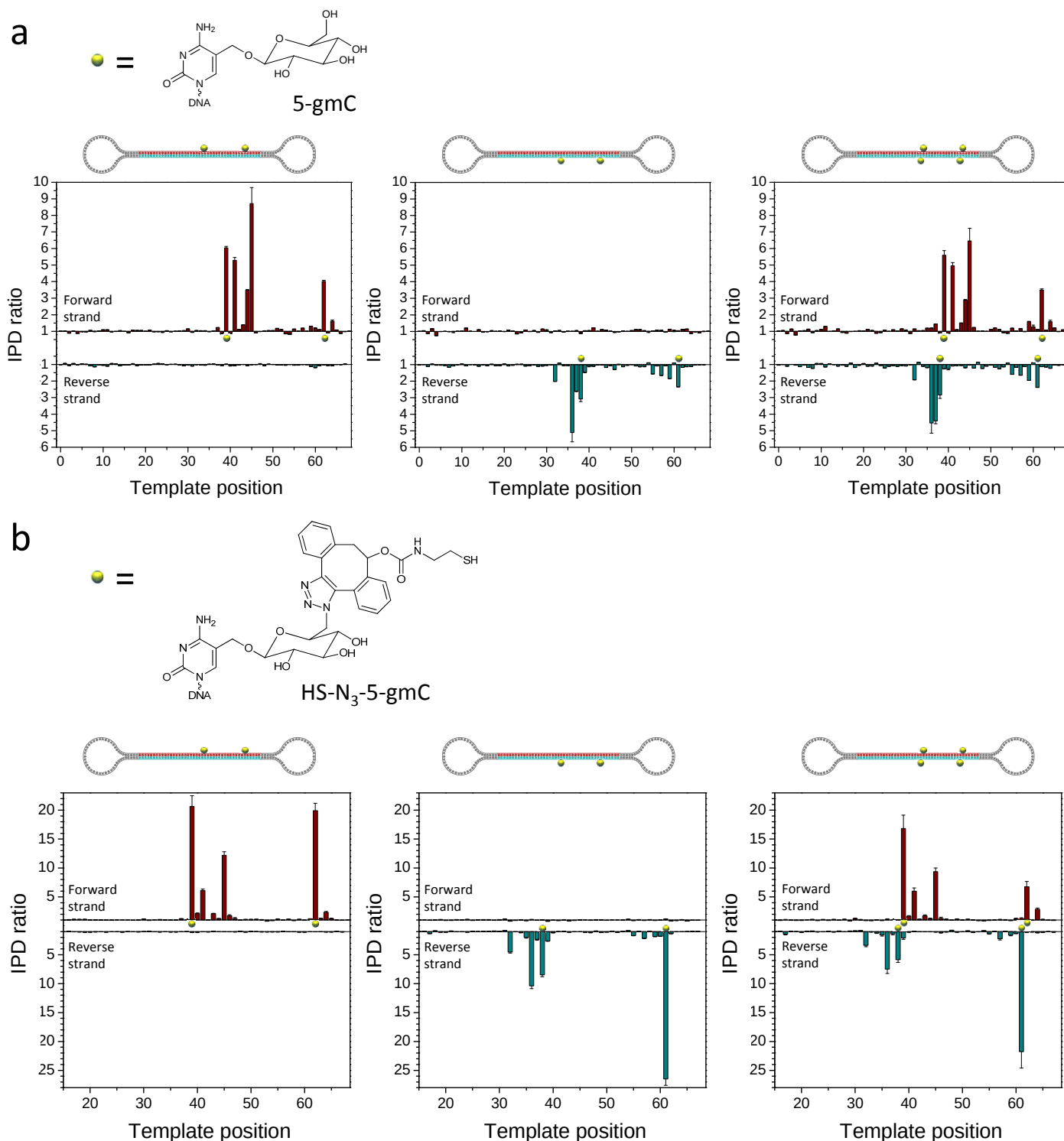
Supplementary Figure 1	Efficiency of the selective chemical labeling protocol of 5-hmC.
Supplementary Figure 2	SMRT sequencing of selectively labeled 5-hmC differentiates hemi-and fully-hydroxymethylated DNA.
Supplementary Figure 3	Validation of 5-hmC SMRT sequencing by a low-throughput restriction digestion assay.
Supplementary Table 1	Enrichment, yield and specificity of the selective labeling of 5-hmC.
Supplementary Table 2	Selected genomic locations of 5-hmC in mouse ES cells from initial shotgun SMRT DNA sequencing.
Supplementary Table 3	Primer list and sequences.

Supplementary Fig. 1. Efficiency of the selective chemical labeling protocol of 5-hmC.



Molecular structures of each step in the workflow are shown next to the corresponding MALDI spectra. The depicted mass spectra were performed using source material taken directly from the reaction mixture, without prior purification, thereby reflecting both a validation of the mass of each intermediate as well as the high efficiency of each conversion step.

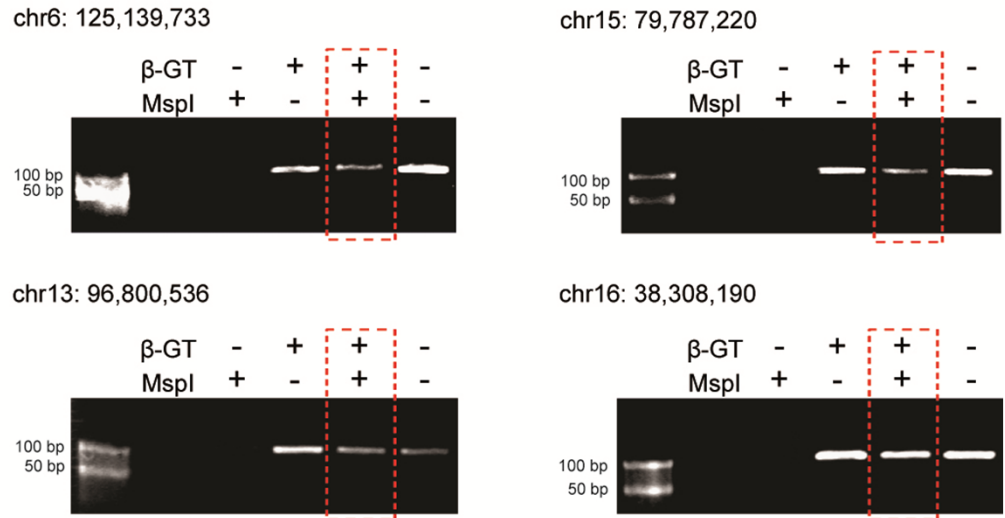
Supplementary Fig. 2. SMRT sequencing of selectively labeled 5-hmC differentiates hemi- and fully hydroxymethylated DNA.



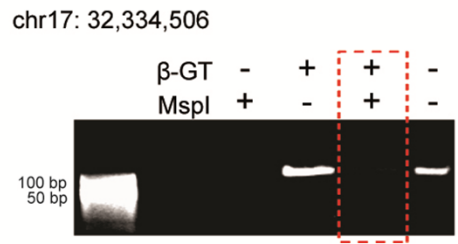
IPD values from synthetic DNA SMRTbell templates containing two 5-hmC positions in CpG sequence contexts on the forward strand only (left panel), the reverse strand only (middle panel), and on both strands (right panel) were compared to control DNA templates lacking 5-hmC. **(a)** 5-hmC upon coupling with glucose, the natural substrate of β -glucosyltransferase (5-gmC), and **(b)** 5-hmC upon coupling with the disulfide-containing biotin linker, followed by disulfide cleavage (HS-N₃-5-gmC). The full sequences of the DNA templates are provided in the Methods section.

Supplementary Fig. 3. Validation of 5-hmC SMRT sequencing by a low-throughput restriction digestion assay.

Positive CCGG site:



Negative CCGG site:



In this assay, MspI will recognize and cleave CCGG when the inner C is cytosine, 5-mC or 5-hmC, but not 5-gmC. After MspI digestion of unmodified and glucose-modified genomic DNA, PCR amplification of target DNA with primers flanking a CCGG site of interest is performed to detect whether cleavage has taken place, and thus the hydroxymethylation status of the inner C on the CCGG site.

Supplementary Table 1. Enrichment, yield and specificity of the selective labeling of 5-hmC.

	5-hmC level (% of total nucleotides)	Pull-down yield of original approach	Pull-down yield of current approach
Synthetic 32mer double strand DNA with one 5-mC	-	Not tested	Undetectable by Nanodrop
Synthetic 32mer double strand DNA with one 5-hmC	-	35%	50%
Mouse postnatal day 7 cerebellum genomic	0.1	0.46%	1.3%
Mouse adult cerebellum genomic DNA	0.4	1.5%	3.1%
Wild-type Drosophila melanogaster genomic DNA	<0.004	Not tested	Undetectable by Nanodrop

The current labeling method using a disulfide-containing linker and DTT elution was compared to the original method by Song et al. (2011) Nature Biotechnology 29, 68-72 in synthetic and genomic DNA samples.

Supplementary Table 2. Selected genomic locations of 5-hmC in mouse ES cells from initial shotgun SMRT DNA sequencing.

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
1	9,503,685	-	AGCACGTTG	447
1	9,503,686	+	GCACGTTGT	447
1	13,194,033	-	AAAACGCCC	370
1	24,618,689	+	AAGCGTCTA	523
1	35,900,822	-	CATACGTTT	151
1	36,296,282	+	AAGCGTCTC	145
1	36,703,311	-	ATGACGTCT	139
1	37,017,369	-	TTGTCGAAG	185
1	52,225,773	+	ATGCGGCAA	163
1	64,278,647	+	CGTCGCTAA	130
1	74,135,048	+	TCACGGGTC	178
1	78,448,250	+	AAACGTGCC	162
1	83,106,446	+	CATCGTACA	157
1	89,105,617	+	CACCGACCG	250
1	92,173,220	-	TATACGTGT	286
1	99,808,187	+	TACCGTGCT	366
1	106,821,713	-	TGTCCGTAT	291
1	121,775,893	+	ATGCGTCCT	176
1	134,794,940	-	CTGCCGTGA	220
1	135,168,144	-	AAGACGCAG	169
1	138,029,197	-	CTCCCGACT	181
1	140,395,359	-	ATTACGAAT	285
1	154,419,309	+	AGACGAGGG	380
1	163,194,570	+	CACCGACTA	206
1	163,194,587	+	TTCCGTTGT	206
1	163,194,597	+	TCCCGTAAA	206
1	181,354,263	-	ACTACGTAA	194
1	181,354,264	+	CTACGTAAG	194
1	182,273,412	-	CAGACGTGG	404
1	182,273,427	-	TTCCGCTG	404
1	182,273,454	+	GGCCGGCCT	404
1	182,273,474	-	GGTACGTGC	404
1	182,273,475	+	GTACGTGCT	404
1	182,821,950	-	GTATCGCAA	188
1	183,310,912	-	CCTACGTCT	133
1	183,750,864	+	AGTCGCGGG	111
1	183,750,866	+	TCGCGGGTT	111
1	183,782,096	-	TTAGCGAGC	212
2	18,775,203	+	AGGCGTTAG	186
2	30,408,655	-	CTACCGCAT	116
2	31,332,202	+	GGACGTGTG	118
2	32,538,807	-	AGCTCGTCC	415

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
2	32,791,439	+	CAGCGACGT	189
2	32,862,581	+	AATCGTGGG	104
2	35,978,383	-	TGAGCGTGG	226
2	57,187,898	-	ATCTCGCTC	174
2	59,112,939	+	CCACGTGGG	162
2	59,140,804	+	AAACGGGTG	292
2	59,701,759	+	CTGCGTGGG	337
2	60,634,191	-	AGAACGGCA	285
2	74,919,376	+	ACACGTGTC	361
2	90,561,454	-	TTACGGCT	132
2	90,561,476	-	TATGCGTGG	132
2	90,561,483	-	GGATCGTAC	132
2	100,608,418	+	CCACGACAA	368
2	115,865,535	-	GTGCCGCAC	242
2	115,865,539	-	CGCACGCAT	242
2	115,865,584	+	GGTCGTGCG	242
2	115,865,588	+	GTGCGTCCT	242
2	118,730,324	-	TAAACCGT	169
2	118,730,326	-	AACGCGTGT	169
2	122,172,119	+	GTTTCGACCT	214
2	122,172,247	+	GAGCGGCAC	214
2	124,958,879	-	CTAACGTGA	114
2	130,489,161	-	TAGACGGCT	170
2	130,908,800	-	TCTACGTAA	404
2	130,908,801	+	CTACGTAAA	404
2	135,217,864	+	TTTCGTAGC	146
2	135,368,273	+	ATGCGCAGA	144
2	147,411,853	+	CTCCGATA	122
2	153,446,675	+	TTGCGTGGC	217
2	154,711,754	-	TAAACGCCA	207
2	160,511,339	-	TTGTCGTTG	246
2	160,511,441	-	CACACGACA	246
2	161,040,533	+	GCACGGGTA	280
2	161,040,564	+	AGACGTAGA	280
2	164,677,072	-	TAATCGATG	103
2	165,872,071	-	TTATCGTCT	266
2	167,143,228	+	ACACGGACC	120
2	169,796,481	+	ATGCGGTTG	163
2	170,101,238	+	CAACGTGTG	141
2	172,109,145	-	CTCACGAGA	364
2	172,263,288	-	CTTACGTTC	573
2	179,936,763	-	TACCCGTAG	175
2	179,936,764	+	ACCCGTAGT	175
3	9,673,361	+	AATCGCTGA	164
3	10,310,622	-	CCTGCGTGC	153
3	27,645,064	+	AACCGAGCC	272

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
3	35,886,187	+	TCCCGACTG	310
3	36,397,780	+	TTACGTGGG	223
3	57,933,994	+	CCACGTAAG	305
3	82,169,243	-	TCTCCGTAG	207
3	86,409,235	+	CCACGGAGA	138
3	86,720,917	+	GTACGGTTC	385
3	86,721,052	+	TTCCGTAGA	385
3	95,303,789	+	TTACGGAAT	175
3	101,828,888	+	GAGCGTGTG	212
3	103,152,318	+	TTTCGGTGA	240
3	105,768,086	+	TGACGTACC	191
3	107,744,623	+	GACCGTGAG	347
3	115,521,853	+	TTCCGTGTT	191
3	115,718,928	-	AGTACGCTA	236
3	127,850,831	+	AATCGATAA	180
3	130,898,406	+	AAGCGTAGA	224
3	135,233,390	-	CTGACGGAT	248
3	137,568,903	-	ATTACGTCA	175
3	142,554,657	-	CTTACGTAA	216
3	152,095,882	+	CATCGCATG	235
4	10,663,714	-	TCCACGACA	261
4	23,682,457	+	GAACGAGTA	244
4	34,835,057	-	TATACGCCC	168
4	40,882,085	-	CAACCGCAG	161
4	44,893,154	+	AGCCGTGGG	145
4	46,572,658	+	AGACGTGAA	346
4	58,690,663	+	CTGCGGTTC	68
4	62,115,070	+	AGGCGCTCG	145
4	83,834,796	+	AAGCGATTG	257
4	83,834,803	-	TGCACGATT	257
4	83,834,804	+	GCACGATTA	257
4	106,033,820	+	TCCCGTATG	251
4	116,672,074	+	GAACGGACT	257
4	117,824,762	-	TGCACGAGC	380
4	117,824,845	-	GATGCGCTT	380
4	126,048,931	-	TACTCGAGA	227
4	129,164,678	-	AAGACGACA	202
4	129,734,489	+	GATCGTATG	216
4	133,289,803	+	AGCCGTAGA	327
4	134,979,554	-	GACTCGGGT	224
4	134,979,555	+	ACTCGGGTG	224
4	135,377,417	+	CACCGAGGT	275
4	142,811,206	+	CTACGCACA	103
4	147,540,166	-	GGTTCGGGG	207
4	147,540,191	+	TAGCGTGTG	207
4	147,540,211	+	GAACGACCT	207

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
4	148,431,093	+	ACACGCTAC	247
4	150,173,775	-	GGCCCGTCC	193
5	8,527,886	+	TCCCGCTAT	136
5	8,553,137	-	TCCACGACA	327
5	13,402,119	+	CTCCGTAAG	117
5	22,999,162	-	ATACCGCCA	380
5	22,999,163	+	TACCGCCAT	380
5	28,720,600	+	AAGCGCCTG	280
5	29,747,487	+	GCACGTGAA	148
5	31,592,801	-	ATTGCGTAT	149
5	36,817,688	-	CATACGCTA	264
5	38,979,310	+	TTCCGAGGC	737
5	44,103,551	+	CTTCGTCTC	287
5	53,198,643	+	CAACGAGTG	202
5	53,198,658	+	GTGCGCGGA	202
5	53,198,660	+	GCGCGGATG	202
5	64,611,780	+	AGACGGAGA	131
5	64,716,536	+	ATGCGCCTG	204
5	65,692,407	+	CTCCGACAC	295
5	65,692,415	+	CCACGCGCT	295
5	65,692,456	+	GCACGTCTG	295
5	66,417,627	+	TCACGGCTA	233
5	74,031,038	+	AGTCGATGC	289
5	77,299,505	+	TTACGTGGC	308
5	92,735,489	-	CCCACGGAA	231
5	97,459,157	-	ACTGCGAAT	110
5	107,480,568	-	TACACGTCC	228
5	112,027,294	+	TCACGTCTG	224
5	116,927,134	+	TGTCGGGTC	167
5	117,493,748	+	CTACGTGAC	129
5	124,269,216	-	TGGCCGCTA	243
5	124,269,235	-	TGGCCGTAG	243
5	124,269,241	-	TAGCCGCCT	243
5	126,392,882	+	AAACGCACA	211
5	128,684,837	+	AAGCGCTGG	152
5	139,854,355	-	TCAACGTGG	174
5	141,199,530	+	ATGCGTGGA	138
5	148,479,970	-	TGAACGGGA	205
5	148,897,758	-	TTTGCCTTT	297
5	148,897,779	-	TACCCGTAT	297
5	150,056,633	-	TCTGCGCCG	96
5	150,056,636	-	GCGCCGGGC	96
5	151,084,080	+	ACACGTGAA	139
5	151,234,503	+	CTCCGGTAC	115
6	5,702,451	+	TCCCGTCGG	122
6	6,603,497	+	CATCGAATG	254

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
6	7,358,521	-	AGAACGTGG	124
6	24,774,226	+	CTACGGATA	246
6	28,740,433	+	AGGCGTGCT	268
6	30,219,668	-	GGCACGAGG	238
6	30,387,538	-	TAGACGTGT	294
6	30,387,539	+	AGACGTGTA	294
6	30,387,575	-	TCTGCGAGG	294
6	30,419,550	+	CTGCGCGTG	192
6	30,419,552	+	GCGCGTGCA	192
6	31,171,925	+	TCTCGTGGA	217
6	31,283,390	+	CGGCGGCAA	127
6	32,969,520	-	CAAACGTGT	343
6	47,722,733	-	CGAACGCCT	178
6	48,558,749	+	CTCCGTTAG	250
6	48,992,949	-	CCACCGCTG	227
6	49,044,219	-	AGGGCGGGA	339
6	52,170,688	+	ATGCGAGTC	105
6	52,170,716	+	GGCCGAGAC	105
6	64,749,756	+	TGCCGCGCG	273
6	64,749,758	+	CCGCGCGTG	273
6	64,749,760	+	GCGCGTGTA	273
6	72,158,107	+	TTTCGGGCA	112
6	72,158,114	+	CACCGCTAC	112
6	82,693,484	+	CCACGCATG	153
6	83,314,238	+	AGACGGGGC	369
6	83,934,454	+	CTTCGACGT	189
6	83,934,457	+	CGACGTATA	189
6	85,899,623	-	AGTACGGTG	336
6	85,899,734	-	GCTCCGTAG	336
6	88,708,798	+	ACACGTGTG	148
6	92,385,191	+	TTACGTTGT	158
6	92,414,495	-	GATTCGCTA	233
6	99,218,342	+	GCTCGTCCA	135
6	99,976,554	+	AGGCGTCAG	194
6	99,976,847	-	GACCCGTGA	224
6	99,976,852	-	GTGACGTTT	224
6	99,976,853	+	TGACGTTTG	224
6	100,400,684	-	AGCCCGTTA	127
6	108,417,331	+	GTACGGAGG	248
6	108,616,096	-	TCTACGGCA	284
6	113,250,627	+	CTCCGACAT	289
6	113,671,466	+	GCCCGTAGC	111
6	115,569,873	+	GTACGAGTA	216
6	119,063,964	+	AAACGTACG	229
6	119,063,968	+	GTACGTCTA	229
6	119,593,678	+	CTTCGTGAG	307

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
6	122,343,787	+	CTACGCACG	256
6	124,349,353	+	GGCCGGTGA	203
6	124,349,380	+	CAACGTCTC	203
6	125,139,734	+	TCCCGGGGC	126
6	125,452,234	+	TTGCGGAAG	184
6	134,402,108	-	GTGCCGTGG	120
6	134,986,592	+	GGCCGTGCC	427
6	134,986,629	-	CGCACGCCT	427
6	135,682,619	-	AAGGCGGCC	299
6	135,960,650	+	AGACGCTGT	162
6	137,232,436	-	AAGACGTGT	233
6	140,484,916	+	AGCCGAGTG	244
6	140,484,932	+	ACCCGAGA	244
6	140,484,985	-	CACACGCAA	244
6	140,485,050	+	GTGCGTTAG	244
6	142,968,031	-	ATGACGTAA	264
6	143,155,868	+	GATCGATAA	263
6	143,165,297	-	TACACGTCG	238
6	143,181,857	-	GCTCCGTCT	199
6	143,181,894	+	AGTCGAACC	199
6	143,181,900	-	ACCCGAAG	199
6	143,181,901	+	CCCCGAAGG	199
6	143,181,909	+	GACCGTATC	199
6	143,336,315	-	ACAACGATG	250
6	144,324,037	-	GAGCCGTAT	146
6	144,324,065	-	GGTGCGTGG	146
6	146,331,247	+	ACACGGTAG	152
7	4,423,179	+	GAACGTGGC	194
7	4,423,186	+	GCACGTCTA	194
7	6,183,004	-	AAGCCGAG	197
7	6,183,027	-	CATACGGAC	197
7	6,183,028	+	ATACGGAAT	197
7	28,150,550	-	AGGACGGTG	257
7	29,024,885	-	TACACGTAC	118
7	52,155,343	-	TGCACGTTA	314
7	52,315,998	+	AAGCGTCCG	217
7	71,678,434	+	CAGCGCTCC	165
7	73,194,394	+	AGCCGGAGG	250
7	75,313,194	+	ACACGGGTG	122
7	75,452,715	-	GTCCCGCAG	215
7	80,596,495	-	CTTACGCAG	156
7	80,960,915	-	TGAACGAAG	189
7	80,974,029	+	TTACGTGAG	161
7	86,858,929	-	CACACGCAG	304
7	103,856,740	+	GGGCGTATG	166
7	116,330,545	+	ATTCGTTCC	276

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
7	123,249,918	+	GGGCGCTCC	213
7	123,249,932	-	CATACGTAA	213
7	123,249,933	+	ATACGTAAA	213
7	123,679,511	+	GCACGACCA	141
7	134,703,291	+	CCCCGGCCC	298
7	134,703,418	+	AGACGGCGT	298
7	134,703,421	+	CGGCGTGGG	298
7	142,244,449	-	CCCACGTGT	157
7	146,863,614	+	ACCCGGAGA	138
7	150,684,819	+	AGGCGAGGG	151
8	11,631,171	+	GAGCGTAGC	176
8	11,631,176	+	TAGCGGGGA	176
8	12,150,991	+	TCACGTCAC	277
8	12,238,603	+	TCACGTGGA	182
8	12,895,503	+	CAGCGCTCT	164
8	14,962,537	+	AACCGGTAC	322
8	14,962,653	+	CACCGCCAG	400
8	14,962,701	+	CAGCGACAC	400
8	14,962,706	+	ACACGCTGT	400
8	19,982,500	-	CTTCCGGGT	237
8	19,990,675	-	GACCCGGA	152
8	25,281,345	-	CACACGTAT	125
8	25,526,824	+	CTGCGTCTA	242
8	28,024,457	+	CTCCGTCTG	499
8	28,267,600	+	GAACGCTTA	204
8	28,267,618	+	AGTCGTTCT	204
8	28,267,631	+	ATACGCTAG	204
8	33,320,750	+	AGGCGTCTC	247
8	34,687,538	+	AGGCGTTAA	220
8	34,687,578	-	TCCACGTAT	220
8	36,210,894	-	CCAGCGTGT	259
8	36,210,923	-	CCCACGGAT	259
8	36,210,924	+	CCACGGATG	259
8	36,210,944	-	GTACCGAAT	259
8	37,582,807	-	AAAGCGACA	272
8	37,623,283	+	ACACGCAGG	269
8	42,094,716	-	TCTACGTAG	201
8	42,094,717	+	CTACGTAGG	201
8	42,418,988	+	GTGCGTTGT	235
8	42,633,779	+	CCACGTGAC	174
8	42,633,809	+	ATGCGGCTG	174
8	45,329,831	+	CCACGACAA	481
8	45,329,848	+	TCACGCATT	481
8	49,017,466	-	GACTCGCTG	279
8	49,026,474	-	TGGTCGTGC	225
8	49,511,415	-	TCACCGCAT	304

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
8	49,511,416	+	CACCGCATG	304
8	49,511,441	-	TCTACGTGT	304
8	49,511,442	+	CTACGTGTG	304
8	49,511,455	+	CGCCGATGC	304
8	49,511,474	-	ATGACGGCT	304
8	49,552,187	-	TTTCCGTAG	203
8	63,102,771	+	GCCCGTGAG	295
8	67,464,673	+	TCACGTGAG	194
8	69,004,255	+	CTTCGTACT	95
8	71,632,381	-	CATACGAGT	247
8	71,632,382	+	ATACGAGTA	247
8	71,827,863	+	GGCCGTCAC	239
8	71,827,891	+	GTGCGTGTG	239
8	73,269,911	-	GGTCCGTAA	254
8	73,269,919	-	ACAACGTGG	254
8	73,305,104	+	TGGCGCGCA	513
8	73,305,106	+	GCGCGCAGG	513
8	77,524,843	+	CTTCGTGAG	227
8	80,546,962	+	CCACGCCTG	298
8	81,255,959	-	TCACCGAGT	240
8	81,887,539	-	CAAACGCAG	75
8	93,506,927	+	GTGCGTCTG	220
8	94,328,709	-	GCTGCGTAC	335
8	94,328,713	-	CGTACGCTC	335
8	95,393,649	+	AAGCGCCCA	329
8	96,348,181	+	CAGCGTTAG	159
8	97,934,849	+	ATACGCCGG	258
8	97,934,913	-	ACTCCGTCT	258
8	108,195,787	+	TCTCGCTTC	223
8	109,223,029	+	GAGCGAGCT	167
8	109,223,052	+	TACCGTGCA	167
8	110,498,517	+	GATCGTTCC	311
8	110,498,522	+	TTCCGTAAA	311
8	110,498,551	+	GATCGCTGC	311
8	110,498,567	+	AAGCGGTCT	311
8	111,473,204	-	CATTCGGAC	204
8	111,473,208	-	CGGACGAGC	204
8	111,771,029	+	CACCGCAAT	400
8	111,771,054	-	GCAGCGGAT	400
8	111,771,055	+	CAGCGGATG	400
8	114,439,197	+	GGTCGTAGT	127
8	114,511,122	+	GGACGGCTG	273
8	114,511,182	+	AACCGGTAG	273
8	114,511,196	+	TGACGTAGC	273
8	117,541,483	-	CTTGCGTGA	325
8	119,192,946	+	ATGCGTGAG	137

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
8	120,314,809	-	GAGTCGGCA	195
8	120,932,522	+	GTCCGTAGA	178
8	120,932,549	+	GAACGTAAA	178
8	122,997,614	-	ACTCCGTCT	393
8	122,997,639	-	TGGTCGCAA	393
8	122,997,690	-	GTCTCGAAC	393
8	123,078,662	+	GGCCGAGTA	181
8	124,165,263	+	GAGCGCTGA	132
8	124,179,264	+	CTGCGTAAG	322
8	124,179,314	+	AAACGCAGC	322
8	124,179,340	+	ATGCGTATG	322
8	125,040,076	+	TTACGTACA	161
8	125,237,640	-	ATCACGTGG	352
8	126,431,359	+	ATCCGAGGT	335
8	126,694,528	+	CAACGTTAT	259
8	127,054,749	+	CTGCGTGAC	288
8	127,054,759	-	GCAGCGGGT	288
8	127,922,410	-	ACTACGCAT	292
8	128,518,023	-	ACTACGTCA	362
8	128,526,676	+	TCTCGCTAT	195
8	128,884,013	+	TTGCGTGTC	417
8	128,884,119	-	CATTCGCGC	417
8	128,884,121	-	TTCGCGCCT	417
8	128,884,139	-	TGCACGCTG	417
8	128,884,140	+	GCACGCTGT	417
9	8,633,875	+	ATACGTAGA	100
9	13,509,426	+	GTGCGGTAG	235
9	21,570,948	+	GTGCGTCTC	138
9	37,347,652	-	GGACCGAGG	159
9	42,778,221	+	TTCCGTCTGA	198
9	42,778,224	+	CGTCGACAC	198
9	42,778,231	-	ACTACGTGT	198
9	42,778,232	+	CTACGTGTC	198
9	44,502,170	-	TCCACGATT	407
9	48,597,905	+	ATACGGGGA	173
9	51,891,338	-	TGGACGTTA	299
9	51,891,339	+	GGACGTTAT	299
9	51,891,382	+	GACCGGCTA	299
9	51,891,403	+	CCTCGTCTT	299
9	51,891,497	-	GCTACGCTT	299
9	60,253,583	+	AGACGTTTG	103
9	61,550,623	+	AGACGTACA	250
9	62,094,233	+	AGACGCACA	198
9	62,753,955	+	ATCCGGCAG	98
9	63,626,603	-	GGTACGTAG	252
9	63,626,604	+	GTACGTAGA	252

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
9	75,084,160	-	TCAGCGTCC	162
9	77,772,710	-	GGCTCGAAT	307
9	90,308,680	-	CAAGCGTGC	182
9	95,685,741	-	GCAGCGCTG	333
9	95,685,750	-	AACTCGAGA	333
9	102,738,682	+	TTACGCCTG	219
9	102,738,718	-	ATAACGATT	219
9	102,738,730	-	AGTGCGGCT	219
9	102,738,741	-	TAGGCGTGC	219
9	102,738,742	+	AGGCGTGCT	219
9	102,979,351	+	AGTCGAGTA	228
9	103,376,665	-	CTCCCGCTA	270
9	107,849,235	+	GGACGATAT	198
9	114,895,701	-	TGCACGTGA	194
9	115,154,478	-	GCCACGGTG	153
9	118,343,295	+	TTACGGAGA	125
9	121,307,727	-	CTCTCGATC	184
10	4,751,048	+	CAGCGCGTA	222
10	4,751,050	+	GCGCGTATC	222
10	5,816,634	+	CCTCGTAAA	158
10	13,410,091	+	AAACGTAGC	218
10	19,999,141	-	CTGTCGTAT	197
10	19,999,159	-	GCAGCGAGG	197
10	21,601,793	+	ATCCGTTGT	266
10	21,871,213	+	GTCCGTGGC	215
10	21,871,240	+	GGGCGCCGG	215
10	21,871,242	-	GCGCCGGCC	215
10	21,871,243	+	CGCCGGCCT	215
10	21,871,288	+	GTTGCCCC	215
10	38,868,916	-	CCCACGTAC	129
10	42,840,935	+	GATCGTTCG	159
10	42,840,939	+	GTTGTTTC	159
10	42,840,976	-	ACTACGAAT	142
10	57,694,662	+	CTGCGTTGC	189
10	58,317,769	+	CTACGTACG	207
10	59,186,973	+	ATGCGCACA	298
10	62,429,686	+	ATACGTACG	95
10	67,350,898	+	AACCGAGGT	342
10	67,352,696	-	TATACGTAC	273
10	67,732,716	-	TGACCGTTT	120
10	67,897,338	+	AGCCGGGGG	380
10	68,671,428	+	ACTCGTAAT	262
10	69,040,512	+	AATCGGGGC	206
10	69,577,081	-	AGCCCGGCA	426
10	76,605,024	-	ACAACGCAG	105
10	77,097,031	+	ACTCGTATG	198

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
10	83,889,057	-	GACACGGGC	370
10	93,030,005	+	AGCCGGGGC	142
10	93,030,010	+	GGGCGTGTT	142
10	93,030,019	-	TTCACGAGA	142
10	93,044,689	-	TCCACGTTC	201
10	104,950,330	-	CTTCCGTGC	198
10	110,211,162	+	ACGCGTGTG	111
10	110,487,322	-	ACAGCGCCC	174
10	110,593,725	+	TAACGCGTC	248
10	110,593,727	+	ACGCGTCTT	248
10	116,554,991	-	TTCACGCAC	186
10	116,840,312	+	CGTCGCTAC	332
10	118,716,125	-	GTCACGTGG	233
10	119,847,914	+	ACACGTAGC	82
10	124,817,075	+	ACACGTGTA	163
11	3,084,994	+	ACACGTATA	278
11	3,085,001	+	TACCGATGT	278
11	3,085,010	+	GTACGGGCT	278
11	3,093,180	-	TGCTCGCAC	401
11	3,093,344	+	AGGCGCTAG	401
11	3,093,359	+	CCTCGTACC	401
11	3,093,461	+	CCCCGTCTC	374
11	3,093,469	+	CCACGGCAA	374
11	3,093,474	-	GCAACGCGA	374
11	3,093,486	-	GCGCCGAA	374
11	3,093,513	+	GTGCGCTCT	374
11	3,093,599	-	GAAGCGGGT	374
11	3,093,663	-	ATTGCGTGG	374
11	3,093,675	-	CTCACGTCG	276
11	3,093,702	+	AGCCGAATC	276
11	5,469,590	-	CATGCGTGC	230
11	6,521,908	+	TCACGTCCT	286
11	6,521,947	+	TAGCGTAGA	286
11	30,351,520	+	CTTCGCTGT	113
11	30,473,398	+	AGTCGCAGC	337
11	35,821,062	-	CTCACGTGT	184
11	44,356,138	+	CTACGTAAG	174
11	45,871,669	+	TGACGCTCC	387
11	45,871,749	+	TTACGTTGC	387
11	49,009,589	+	GCACGTAGC	321
11	49,009,597	-	CCCCGTGG	321
11	59,370,639	-	GTAACGGCA	214
11	60,058,634	+	CAGCGGGGG	413
11	60,058,679	+	CTTCGCTTA	413
11	60,058,766	-	CTCACGAGG	413
11	60,058,767	+	TCACGAGGA	413

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
11	62,412,938	+	AGACGTGGG	151
11	62,875,093	+	ACCCGCTAT	184
11	66,321,679	-	CTAACGTAT	354
11	67,067,610	-	GGCACGCTT	290
11	68,009,268	+	GAGCGACTG	234
11	69,205,410	-	GAAACGGCT	148
11	74,295,018	+	GCACGCCAT	183
11	74,640,668	+	GGACGTAAG	118
11	76,141,833	-	TTACCGACA	425
11	76,141,834	+	TACCGACAG	425
11	76,189,644	-	CATACGCCC	169
11	84,216,737	-	CTCTCGAAA	368
11	87,543,226	-	ATATCGTTG	249
11	88,805,767	+	TTGCGTAGG	170
11	95,006,517	+	CTTCGAGCT	202
11	95,260,876	-	TACACGTGT	196
11	97,608,173	-	ACAGCGTAA	331
11	97,608,323	+	CACCGTCCA	331
11	100,649,390	-	AGGACGGAA	263
11	101,653,559	-	TCCACGTTA	188
11	106,368,444	-	CACACGTGA	212
11	106,368,445	+	ACACGTGAG	212
11	108,845,344	+	TTACGTGAC	248
11	109,262,911	-	TCTGCGTCT	301
11	109,532,659	+	GTGCGTAAG	466
11	113,693,650	+	CCTCGTAGA	244
11	113,694,678	-	GCTCCGTGG	362
11	115,309,711	+	ATGCGTTAA	124
11	117,174,069	-	GGACCGGCC	276
11	117,174,116	+	CTCCGGAAG	276
11	117,504,761	+	GCTCGCTCG	165
11	117,504,765	+	GCTCGGGAC	165
11	117,504,798	+	ACACGAAAG	165
11	119,496,179	+	AGGCGTGCA	143
12	3,237,657	+	CCACGGAGG	316
12	4,008,539	+	GTACGTATC	501
12	60,022,012	+	GACCGAAAC	211
12	70,728,271	-	ATAACGAAC	277
12	71,144,008	-	CTCTCGGAA	366
12	72,211,282	+	CTACGTCTA	257
12	72,974,141	+	CCACGGCTA	152
12	73,163,457	+	CAGCGTAGA	243
12	77,941,744	-	TGTTTCGTTT	386
12	80,298,443	+	TTACGCATA	181
12	84,798,521	+	ATGCGCCAT	347
12	85,276,883	+	AGACGGGTA	276

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
12	86,012,497	+	AGGCGTAGC	203
12	87,007,228	-	ACTGCGGAG	160
12	87,007,229	+	CTGCGGAGG	160
12	109,129,161	-	GGAACGCAG	212
12	109,464,356	+	CTTCGCAGG	213
12	112,065,056	+	GGGCGTGCT	205
12	112,065,068	+	TGGCGGCAC	205
12	113,577,143	-	GTCACGGAA	165
12	113,577,144	+	TCACGGAAC	165
12	113,577,148	-	GGAACGTGG	165
12	113,577,149	+	GAACGTGGT	165
13	9,184,304	-	GGCTCGGCC	157
13	38,322,811	+	ACCCGGGTC	326
13	39,462,075	-	GTAGCGAAA	127
13	41,035,513	-	AAGTCGAGA	320
13	41,035,552	-	AAAGCGTCA	320
13	41,035,553	+	AAGCGTCAG	320
13	41,035,584	-	GTGACGTCA	320
13	41,035,700	-	CCTACGGAA	320
13	44,293,458	-	TCCCCGCA	145
13	47,019,899	+	TTACGCTCA	187
13	48,450,693	-	GCTACGAAT	125
13	53,432,947	-	CTGTCGGGT	263
13	55,104,726	-	CCCTCGAGT	190
13	72,440,270	+	ATGCGCAA	619
13	94,967,030	+	CTTCGTGAT	249
13	96,284,085	-	CGACCGGAA	361
13	96,284,124	-	TGGACGCAG	361
13	96,800,528	+	CCCCGCTTC	173
13	100,278,383	+	GTACGTGCA	264
13	100,903,845	+	TTACGTATT	433
13	101,286,235	-	CTTACGTTT	156
13	101,578,020	-	TAGGCGACT	393
13	114,364,632	+	ATCCGGTGA	276
13	114,906,368	+	ATGCGGTAG	141
14	8,979,849	-	GCTCCGATG	191
14	8,979,850	+	CTCCGATGT	191
14	8,979,866	+	CATCGTCTC	191
14	10,999,953	-	TTGACGTAA	791
14	20,971,835	+	TATCGCTTG	145
14	20,971,870	-	ACTACGTCT	145
14	20,971,871	+	CTACGTCTC	145
14	22,438,195	+	AGACGTGCA	186
14	22,522,458	-	CACACGGGT	360
14	24,684,773	-	GAGCCGAGT	211
14	31,997,973	-	GCCACGTGC	278

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
14	34,192,194	-	GCTACGGGT	150
14	40,903,215	+	AGACGCTAG	283
14	41,718,656	-	TTAACGATT	187
14	41,718,666	-	GTCACGTTA	187
14	49,459,320	+	GGACGGCGG	414
14	49,459,323	+	CGGCGGACG	414
14	49,459,364	+	CCACGTCTC	414
14	61,928,120	+	GGTCGTA	78
14	78,049,177	-	GAGGCGCAT	114
14	87,188,193	+	ATACGTATG	293
14	99,768,850	+	ACCCGATTT	270
14	99,768,882	+	GACCGTGCT	270
14	104,758,731	+	AAACGAGAC	542
14	104,758,739	+	CCTCGGAGA	542
14	104,758,912	-	CAGGCGTCA	542
14	118,982,526	+	TGCCGGTTT	189
14	118,982,586	-	AGCACGCTT	189
15	12,568,959	+	CTGCGTGAA	218
15	36,050,804	+	CCACGACCA	321
15	36,050,812	+	ATTCGCTCT	321
15	37,777,293	+	CAGCGGGTG	222
15	58,788,039	+	CTTCGATGA	75
15	66,773,540	+	CTCCGTGAG	248
15	76,870,583	+	ATGCGTATA	220
15	76,989,087	-	TACACGGAG	435
15	76,989,092	-	GGAGCGACT	435
15	76,989,105	-	TTAGCGCAA	435
15	76,989,112	-	AACTCGGTT	435
15	77,874,382	+	TGGCGTATA	308
15	81,108,294	+	AGCCGTGAG	125
15	81,544,745	+	TACCGTGAG	210
15	81,544,809	+	ACACGGAGG	210
15	83,373,006	-	TCTCCGGGA	180
15	84,098,380	-	CGCACGCAG	169
15	85,255,587	+	AGACGAGGC	208
15	85,255,592	+	AGGCGGTAG	208
15	86,108,029	-	GTGCCGGGC	426
15	86,151,789	-	TACACGCAT	175
15	94,411,651	-	CCTACGGAC	165
15	95,692,807	+	AGACGTGGC	133
15	96,489,674	-	GCAACGAAT	224
15	97,014,742	+	CTACGTCCC	291
15	97,577,358	+	CAACGAACT	315
15	97,905,500	+	AAGCGGCAG	268
15	97,905,527	+	GAACGTCCA	268
15	97,905,542	+	CTACGATCC	268

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
15	99,448,241	+	CCTCGAGTG	265
16	3,115,966	+	TCACGTTTT	306
16	4,483,046	-	GTTCCGTGA	444
16	8,664,468	+	ATGCGTGGG	223
16	14,129,704	+	ACTCGTTTC	105
16	17,638,599	-	TTCACGCCT	140
16	22,730,857	-	GCCACGTAG	252
16	22,800,640	-	AGTGCGCAG	386
16	23,968,508	-	ACAGCGATA	201
16	23,968,509	+	CAGCGATAG	201
16	23,968,540	-	AAGACGAAT	201
16	28,840,541	+	ACACGTCAC	318
16	28,840,574	-	TATCCGGAT	318
16	28,859,764	+	TTGCGTGAG	156
16	29,679,849	-	CCTACGCAC	148
16	31,100,548	+	ACTCGTATC	274
16	32,474,332	-	AGGCCGCTC	363
16	33,026,049	+	TGACGTTGA	278
16	33,446,870	+	GCACGTTGT	126
16	35,458,774	-	GTCACGGGT	197
16	37,565,662	-	ATTACGCCT	159
16	48,298,795	+	AGTCGTTCT	87
16	49,635,332	-	CAATCGAAG	114
16	50,458,978	-	GGATCGTGG	175
16	56,153,748	+	AGGCGTTGA	334
16	57,240,380	-	GAGACGTGT	145
16	72,990,634	+	CTACGTAGG	289
16	72,990,641	+	GGCCGAGAG	289
16	91,558,334	+	GAACGTGGG	189
16	92,485,236	-	CCCACGTGG	350
16	93,889,751	+	TTACGTGGG	102
16	94,914,821	+	TTGCGTGTG	396
16	96,383,282	+	ACCCGCAGA	119
17	5,768,711	+	CTCCGAAA	240
17	10,665,995	-	ATTCCGCCA	247
17	13,206,170	+	TTGCGGTGG	205
17	13,206,177	-	GGGACGTAG	205
17	23,956,602	+	TCTCGACC	313
17	23,956,647	+	AGGCGTAGA	313
17	23,956,824	+	AGGCGATCT	313
17	25,366,835	+	GGACGTGTC	105
17	26,761,967	+	TTACGGAGA	145
17	26,835,577	+	TGCCGTCTG	374
17	30,200,355	-	CTTACGATT	203
17	30,247,514	+	TGGCGTTTA	111
17	30,604,996	+	TGTCGCAAG	311

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
17	30,950,019	+	TATCGCGGA	356
17	30,950,021	+	TCGCGGACT	356
17	30,950,031	+	TCTCGAGGA	356
17	35,153,670	-	GTGACGCCA	186
17	56,892,789	+	TAGCGTCTG	332
17	64,711,551	+	AGACGCTGT	153
17	68,343,502	+	CCTCGTGTA	185
17	69,602,088	+	CTCCGTGTC	386
17	80,432,017	+	TTACGTGTC	472
18	33,100,625	-	GTTGCGAGG	514
18	35,830,696	-	TCCACGTGT	288
18	38,492,811	-	CCTCCGACA	321
18	38,492,823	-	AGTTCGATC	321
18	42,667,319	+	TAACGTGGG	237
18	42,667,413	-	ACAGCGCTA	237
18	42,667,421	-	AGGTCGGGA	237
18	54,981,379	+	GCACGGGGA	165
18	55,304,676	-	TTACGAAG	149
18	56,900,400	+	AGTCGCAGT	227
18	56,900,426	-	AAAGCGGAA	227
18	56,900,504	+	GAACGTGTG	227
18	60,895,911	-	TACACGTAT	164
18	64,117,072	+	GCTCGTACA	401
18	65,101,890	+	CCTCGTCAG	143
18	69,553,424	-	CTATCGGGG	130
18	69,802,506	+	TACCGTGAC	298
18	69,802,584	+	CTCCGATGT	298
18	75,337,941	+	ATGCGTGAG	223
18	77,437,475	+	CTGCGTCGG	241
18	77,437,478	+	CGTCGGGCC	241
18	80,656,810	-	TGGACGGAC	219
18	81,827,000	+	TTACGTATT	127
19	4,774,811	-	GTAACGAAG	215
19	4,909,780	-	GAACCGATT	294
19	4,909,963	-	CCCTCGTAC	294
19	16,675,077	+	CTGCGTAAT	250
19	25,420,318	-	TGGACGTTA	166
19	25,420,350	-	GAAGCGTTT	166
19	33,719,095	+	CCACGAATC	379
19	37,098,880	-	GCTACGTCG	308
19	37,098,883	-	ACGTCGGCC	308
19	37,098,892	-	TCTACGTAC	308
19	42,277,189	+	ATCCGTGGC	279
19	43,739,612	-	GAAGCGCTG	437
19	45,178,696	-	CTCACGGAT	130
19	45,830,543	+	AAACGTTGG	103

Chromosome	Position	Strand	Sequence context (-4..+4)	DNA insert length (bases)
19	46,533,497	+	GTACGTAGG	356
19	47,226,380	+	GCACGTAGC	234
19	53,424,321	-	TCTCCGTTG	253
19	55,300,546	-	GTATCGCTA	232
19	57,186,131	+	GGTCGAGCA	149
19	58,620,459	+	ATGCGCCCC	126
19	58,620,464	+	CCCCGTAAT	126
19	58,736,817	+	CTTCGAGAC	157
X	28,108,211	+	CATCGACAC	166
X	76,275,185	+	TATCGGATA	294
X	78,483,244	+	GCACGTAGA	93
X	104,773,271	+	CTTCGTCAT	147

Assignments were filtered from a total of 191,745 sequencing reads mapped to the mouse genome (version ncbi 37/mm9) for genomic locations which were covered by high-confidence reads displaying the kinetic signature for 5-hmC adducts, along with sequenced DNA fragment sizes and sequence contexts. Pairs of adjacent CG positions indicating fully hydroxymethylated sites are indicated by boxes.

Supplementary Table 3. Primer list and sequences.

Pull-down specificity test:

Primer	Sequence
Forward	GGTACCATCACAGTATxGAATCTCAGAAGCTT
Reverse	GAAAGCTTCTGAGATTxGATACTGTGATGGTA

x indicates 5-mC, or 5-hmC

Synthetic 5-hmC containing SMRTbell templates:

Primer	Sequence
TopLeft_hmC	CCCGAxGCATGATCTGTACTTGATCGACxGTGCAAC
BottomLeft_hmC	CxGGTCGATCAAGTACAGATCATGxGT
TopRight	GCATGATCTGTACTTGATCGACCGTGCA
BottomRight	GAAGTGCACGGTCGATCAAGTACAGATCATGCGTTG
Left_Hairpin	CGGGCTCGGAACGAAAGTTCCGAG
Right_Hairpin	CTTCTCTCTCTTTTCTCTCTCTCCGTTGTTGTTGTTGAGAGAGA

x indicates 5-hmC

Combined glycosylation restriction analysis validation of 5-hmC hits:

Genomic Location and Strand	Sequence
Chr6: 125139690-125139793 Forward	GGCAGGGACTAGGAAAGACC
Chr6: 125139690-125139793 Reverse	TCCCAGCTTAGAGGTGAGAGAC
Chr13: 96800484-96800578 Forward	GCCAATCATCCAGAACATGC
Chr13: 96800484-96800578 Reverse	CCATTCCCACTGTGTTCTCC
Chr15: 79787144-79787248 Forward	CCCAACAAGAGGCAAGTGAG
Chr15: 79787144-79787248 Reverse	ATTGGAGCCTGCTTGTCATT
Chr16: 38308165-38308262 Forward	CACCATGTTCTGCACGACTC
Chr16: 38308165-38308262 Reverse	TGCTCCCCTCTCTGTCTCTG
Chr17: 32334426-32334530 Forward*	CCAGCTGGCACTTGACTCAT
Chr17: 32334426-32334530 Reverse*	AGCCCATCTGCCTTACAGC

*negative control