Mapping recently identified nucleotide variants in the genome and transcriptome

Chun-Xiao Song¹, Chengqi Yi²,³ & Chuan He¹

Nucleotide variants, especially those related to epigenetic functions, provide critical regulatory information beyond simple genomic sequence, and they define cell status in higher organisms. 5-methylcytosine, which is found in DNA, was until recently the only nucleotide variant studied in terms of epigenetics in eukaryotes. However, 5-methylcytosine has turned out to be just one component of a dynamic DNA epigenetic regulatory network that also includes 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine. Recently, a study has shown that the enzymatic demethylation of 5-carboxylcytosine (5caC) in DNA is achieved through a stepwise oxidation of 5-methylcytosine (5mC) by the ten-eleven translocation (TET) family dioxygenases (Fig. 1 and Table 1)⁴,⁶-⁸. These newly revealed DNA base modifications immediately drew broad attention from the research community and have been extensively reviewed⁹-¹¹. It is generally viewed as a ‘silencing’ epigenetic mark because of the hydrophobic recruitment of methyl-CpG-binding proteins². 5mC, although carrying a hydrophilic modification, is not simply an ‘activating’ epigenetic mark; it is regarded as an intermediate in an enzymatic demethylation pathway⁴,⁶,⁷,¹⁴-¹⁶ (Fig. 1) and appears to play complex roles in gene regulation¹⁷-²². In certain cells or tissues in which 5hmC accumulates to relatively high abundance, it may also have unique functions of its own that directly affect gene expression.

The mammalian genome possesses much more information than nucleotide sequences. Each adult human body contains >200 distinct cell types; yet despite their marked differences in phenotype and function, these cell types share an almost identical genome sequence. Epigenetic modifications play a major role in this diversity. An important epigenetic modification in mammalian genomic DNA is the nucleotide variant 5-methylcytosine (5mC); 5mC regulates gene expression, determines cell development and affects disease pathogenesis¹²-¹³. But 5mC is not the only nucleotide variant.

During the past three years, three additional cytosine variants were identified in the mammalian genome. In 2009, 5-hydroxymethylcytosine (5hmC) was shown to be enzymatically produced in relatively high abundance in certain mammalian cells and tissues³,⁴. Following this discovery, 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) were found in mouse embryonic stem cells (ESCs) and mouse tissues⁵-⁷. These cytosine derivatives are produced from a stepwise oxidation of 5mC by the ten-eleven translocation (TET) family dioxygenases (Fig. 1 and Table 1)⁴,⁶-⁸. These newly revealed DNA base modifications immediately drew broad attention from the research community and have been extensively reviewed⁹-¹¹. It is generally viewed as a ‘silencing’ epigenetic mark because of the hydrophobic recruitment of methyl-CpG-binding proteins². 5mC, although carrying a hydrophilic modification, is not simply an ‘activating’ epigenetic mark; it is regarded as an intermediate in an enzymatic demethylation pathway⁴,⁶,⁷,¹⁴-¹⁶ (Fig. 1) and appears to play complex roles in gene regulation¹⁷-²². In certain cells or tissues in which 5hmC accumulates to relatively high abundance, it may also have unique functions of its own that directly affect gene expression.

Current progress in understanding these enzymatic demethylation intermediates has been made by the development of 5hmC-detection and sequencing methods²⁶. In this Review, we discuss methods for detecting nucleotide modifications of potential functional importance, such as 5hmC, 5fC and 5caC in DNA, and we briefly summarize other interesting modifications, such as N⁶-methyladenosine (m⁶A) and 5mC in RNA. To fully describe new technologies and the biological insights revealed through their application to the study of a single nucleotide variant, we focus on 5hmC. This Review will not cover 5mC in DNA, which is the subject of many other articles²⁷,²⁸. Although 5-hydroxymethyluracil (5hmU) is a suggested intermediate in active demethylation mediated through deamination and then base excision repair (Fig. 1)²⁹,³⁰, it is essentially undetectable in the mammalian genome¹⁵, and a recent study has demonstrated that the feasibility of the deamination step⁵¹. Nevertheless, 5hmU behaves similarly to 5hmC in certain aspects. Therefore, many 5mC detection and profiling methods that we discuss can potentially be applied to future 5hmU detection and profiling, if it does indeed play functional roles in certain biological pathways³²,³³.

New nucleotide variants bring technological challenges

The nucleotide variants discussed here generally exist in very low abundance in the genome, ranging from several parts per million...
The pattern of DNA methylation is established and maintained by DNA methyltransferases. Demethylation can be passive (e.g., during replication) or active. TET family proteins can oxidize 5mC to 5hmC, 5hmC to 5fC, and then 5fC to 5caC. The oxidation products 5fC and 5caC can be removed by TDG to generate an abasic site. This abasic site can be repaired to a cytosine by the base excision repair (BER) pathway. Alternatively, 5hmC may be deaminated by AID or APOBEC to 5hmU, which can subsequently be removed and repaired by TDG or SMUG1 and then base excision repair, respectively. 5caC may also be removed in a decarboxylation pathway. Solid arrows indicate biochemically validated pathways whereas dashed arrows are pathways yet to be confirmed biochemically. 5hmU has not been detected in any mammalian genome so far. Blue boxes, nucleotide variants; orange boxes, enzymes.

Figure 1 DNA nucleotide variants, including 5hmC, 5fC and 5caC. The pattern of DNA methylation is established and maintained by DNA methyltransferases. Demethylation can be passive (e.g., during replication) or active. TET family proteins can oxidize 5mC to 5hmC, 5hmC to 5fC, and then 5fC to 5caC. The oxidation products 5fC and 5caC can be removed by TDG to generate an abasic site. This abasic site can be repaired to a cytosine by the base excision repair (BER) pathway. Alternatively, 5hmC may be deaminated by AID or APOBEC to 5hmU, which can subsequently be removed and repaired by TDG or SMUG1 and then base excision repair, respectively. 5caC may also be removed in a decarboxylation pathway. Solid arrows indicate biochemically validated pathways whereas dashed arrows are pathways yet to be confirmed biochemically. 5hmU has not been detected in any mammalian genome so far. Blue boxes, nucleotide variants; orange boxes, enzymes.

Table 1 Proteins that deposit, bind to, modify or remove nucleotide variants, and the known genomic locations of some of these nucleotide variants

| Modification | Proteins that deposit the modification | Proteins that modify, remove or bind the modification | Genomic or transcriptomic location |
|-------------|--------------------------------------|------------------------------------------------------|----------------------------------|
| 5hmC        | TET1-3 (refs. 4,8)                    | TET1-3 (refs. 6,7)                                    | With affinity-based profiling, it is shown to be enriched at transcription start sites, promoters, exons, CTCF-binding sites and enhancers. With single-base resolution sequencing, it shows highest enrichment at distal regulatory regions, near but not on transcription factor-binding sites. |
| 5fC         | TET1-3 (ref. 6)                       | TET1-3 (ref. 6), TDG                                    | Unknown                          |
| 5caC        | TET1-3 (refs. 6,7)                    | TDG,23,24                                             | Unknown                          |
| m^-A in mRNA| MT-A70 (A 70 kD subunit protein in a 200-kD protein complex) | FTO37, YTHDF2-3 and ELAVL1 (binding proteins) | Enriched around stop codons, in 3' UTRs and within long internal exons. |
| 5mC in RNA  | NSUN2 (ref. 92)                      | Unknown                                               | Enriched in untranslated regions (both 5' and 3' UTRs) and near Argonaute-binding regions within mRNA. |
Box 1 Biological functions of new DNA nucleotide variants

Emerging data depict 5hmC as an intermediate involved in DNA 5mC demethylation\(^7,10,29,41\) as well as a functional epigenetic mark involved in gene regulation\(^17\text{-}22,115,116\). 5hmC was shown to exist in the mammalian genome in the past\(^117,118\). However, only in 2009 was 5hmC found to be oxidized from 5mC by TET family iron(II)/α-ketoglutarate-dependent dioxygenases\(^7,10\) (Fig. 1), which play important roles in ESC maintenance and differentiation\(^11\text{-}13,17\), normal hematopoiesis and malignancies\(^58,109,110\), and zygote development\(^111\text{-}113\). The level of 5hmC varies significantly in different tissues and cells (Table 2). 5hmC accumulates with age and is most enriched in brain tissues (0.4\text{-}0.7\% of cytosine)\(^6,15\), where it is suspected to play regulatory roles in neurodevelopment and aging, as well as in neurological diseases\(^39,57,65\). It exists in intermediate amounts (0.1\% of cytosine) in mouse ESCs\(^6\) where it may have dual roles of being an DNA demethylation intermediate as well as exhibiting gene regulation functions\(^11\text{-}13,17\). Proteins that bind 5hmC have been reported, although they are not specific to 5hmC\(^115\text{-}116\). The exact mechanisms of 5hmC-mediated regulation still require further investigation, in particular to identify and characterize potential 5hmC-specific binding proteins. 5hmC is strongly depleted in human cancer cells compared with normal tissue (0.03\text{-}0.1\% of guanine), which indicates its roles in aberrant DNA demethylation in oncogenic processes\(^40,119\text{-}121\).

5fC and 5caC are continuous products of TET protein-mediated oxidation of 5hmC and they can be removed by TDG (Fig. 1a). The level of 5fC is about tenfold lower than that of 5hmC in mouse ESCs, and the level of 5caC is about tenfold lower than 5fC (Table 2)\(^5,6\). In contrast to 5hmC, 5fC and 5caC have not been found to accumulate in brain tissues. In fact, 5caC has been detected only in mouse ESCs, but not yet in other tissues\(^5,6,15\). Currently, 5fC and 5caC are thought to be intermediates of active 5mC demethylation, either in a replication-dependent manner in zygotic development\(^41,42\), or in a replication-independent manner through TDG-mediated base excision repair in specific loci (Fig. 1)\(^7,23,30\), although other pathways such as decarboxylation are possible\(^122\). However, 5fC and 5caC may also have regulatory roles, which will be confirmed if specific binding proteins can be identified. For example, 5fC and 5caC were recently shown to slow down RNA polymerase II transcription, which may suggest possible functional interplay between transcription and these epigenetic modifications\(^123\).

derivative that is highly immunogenic (Fig. 2a)\(^38\). The resulting anti-CMS antibody is very specific with much less density bias compared with anti-5hmC antibodies\(^22,38\).

Besides antibodies, enzymes that specifically recognize and react with the nucleotide of interest have proved to be extremely valuable. In this regard, the T4 bacteriophage enzyme β-glucosyltransferase (βGT) has become a critical tool for specifically modifying 5hmC for subsequent detection and sequencing (Fig. 2b)\(^43\). This bacteriophage enzyme has long been known to transfer a glucose moiety from UDP-glucose to 5mC, which also exists in the T4 phase genome\(^44\). Because the glucosylation reaction is specific to 5mC, it was used to transfer a radioactively labeled glucose to 5mC for quantification, which is more sensitive and accurate than the antibody-based detection\(^32,45\).

Restriction endonucleases sensitive to methylation have long been used to detect DNA methylation\(^46\). Recently, several restriction endonucleases including MspI\(^47,48\), TaqI\(^49\), MspII\(^50,51\), PvuRts1I\(^52,53\) and SauUSI\(^54\) have been employed to detect 5hmC at specific loci and potentially for genome-wide analysis as well. These enzymes are either selective to 5hmC or blocked by glucose-modified 5hmC (from βGT-catalyzed glucosylation), thus providing a sequence-dependent interrogation of 5hmC. Methyltransferase\(^55\) and exonuclease\(^46\) have also been used for the detection of 5hmC. Equivalent methods have yet to be reported for 5fC and 5caC.

For a more accurate method, researchers turn to the gold standard in quantifying low levels of modified nucleotides: liquid chromatography (LC)-mass spectrometry (MS), which separates and identifies hydrolyzed nucleotides. Carell and colleagues coupled LC with high-resolution MS and used isotope-labeled internal standards to achieve accurate quantification of 5hmC\(^57\). Later, they applied the same approach to the detection of 5fC in genomic DNA isolated from mouse ESCs\(^5\). They reacted the formyl group of 5fC with

Table 2 Relative abundance and known tissue locations of new nucleotide modifications

| Modification | Tissues and cell lines | Relative abundance | Genome- or transcriptome-wide profiling methods applied |
|--------------|------------------------|--------------------|-----------------------------------------------------|
| 5hmC in DNA  | Mouse ESC              | 0.1\% of cytosine\(^6\) | hMeDIP\(^18\), GLIB\(^22\), Anti-CMS\(^22\) TAB-Seq\(^9\) (single-base resolution) |
|              | Human ESC              | Not available      | hMeDIP\(^62\), hMe-Sea\(^16\), TAB-Seq\(^9\) (single-base resolution) |
|              | Mouse brain tissue     | 0.4\text{-}0.7\% of cytosine\(^6,15\) | hMeDIP\(^35\), hMe-Sea\(^39\) |
|              | Human brain tissue     | Not available      | Not available                                      |
|              | Other mouse tissue     | 0.02\text{-}0.3\% of cytosine\(^6,15\) | hMeDIP\(^12\) |
|              | Human cancer cells     | 0.03\text{-}0.1\% of guanine\(^19\) | hMeDIP\(^63\) |
|              | Mouse P19 and 3T3-L1 cells | Not available | Not available                                      |
| 5fC in DNA   | Mouse ESC              | 20 p.p.m. of cytosine\(^6\) | Not available                                      |
|              | Mouse tissues          | 3\text{-}20 p.p.m. of cytosine\(^6\) | Not available                                      |
| 5caC in DNA  | Mouse ESC              | 3 p.p.m. of cytosine\(^6\) | Not available                                      |
| m^6_A in mRNA | Human HepG2 cells     | Not available      | Anti-m\(^6\)A antibody\(^101\) |
|              | Mouse liver            | Not available      | Anti-m\(^6\)A antibody\(^101\) |
|              | Mouse brain            | Not available      | Anti-m\(^6\)A antibody\(^102\) |
|              | Human HEK293T cells    | Not available      | Bisulfite sequencing\(^92\) |
| 5mC in RNA   | HeLa cells             | Not available      | Anti-m\(^6\)A antibody\(^102\) |

For a more accurate method, researchers turn to the gold standard in quantifying low levels of modified nucleotides: liquid chromatography (LC)-mass spectrometry (MS), which separates and identifies hydrolyzed nucleotides. Carell and colleagues coupled LC with high-resolution MS and used isotope-labeled internal standards to achieve accurate quantification of 5hmC\(^57\). Later, they applied the same approach to the detection of 5fC in genomic DNA isolated from mouse ESCs\(^5\). They reacted the formyl group of 5fC with

© 2012 Nature America, Inc. All rights reserved.
Recent advances in applying triple quadrupole mass spectrometer for LC–tandem mass spectrometry (MS/MS) detection of rare base modifications, such as 5hmC, 5fC, 5caC, m6A, and other nucleotide variants, further improve the detection limits and allow for quantification. In fact, LC-MS/MS is the only reported method so far that can quantify 5caC, the scarcest cytosine derivative in ESC genomic DNA, at the level of ~3 p.p.m. of that of cytosine. The relative abundance of these new nucleotide variants determined by LC-MS are summarized in Table 2.

**Genome-wide profiling methods**

Compared to simple detection and quantification, genome-wide profiling methods that combine affinity-based enrichment and high-throughput sequencing to yield a genome-wide distribution map of the modified base can provide much-needed biological insights. Figure 2 and Table 3 summarize and compare all reported profiling methods for 5hmC. Antibody is usually the first thing to come to mind (Fig. 2a). Traditional antibody-based captures, such as methylated DNA immunoprecipitation–sequencing (MeDIP-seq) and the related methyl-binding protein-sequencing (MBD-seq), have been used extensively to map methyleomes. Similarly, several groups have simultaneously developed hydroxymethyl-DNA immunoprecipitation–sequencing (hMeDIP) using antibodies raised against 5hmC. However, careful analyses reveal the tendency of these anti-5hmC antibodies to recognize modification-dense regions as well as CA repeats. Such biases, together with high background noise and inferior reproducibility when using antibodies from different lots, can pose problems in data analysis.

Nevertheless, valuable information on the genome-wide distribution of 5hmC has been gained. As mentioned above, the anti-CMS antibody showed substantially improved performance over the anti-5hmC antibody in genome-wide pull-down and sequencing, with less bias and lower background noise. Based on this body of work, 5hmC has been shown to be enriched at transcription start sites, promoters, gene bodies (exons), CCCTC-binding factor (CTCF)-binding sites and enhancers in ESCs, thus suggesting roles for 5hmC in DNA methylation fidelity, pluripotency and lineage commitment balance (Table 1).

We took advantage of the βGT-catalyzed 5hmC glucosylation reaction and developed a selective chemical labeling–based method, hMe-Seal (Fig. 2b). Like unmodified glucose, azide-modified glucose is well tolerated by βGT and efficiently transferred to 5hmC. A biotin can be subsequently installed onto the azido group. Relying on the extremely tight and specific binding between biotin and streptavidin, we can in principle label every 5hmC and perform selective pull-down for genome-wide profiling or loci-specific analysis of 5hmC distribution. Thanks to the use of a disulfide linker, the enriched product can be readily released from streptavidin by means of reduction with...
Table 3 Advantages and disadvantages of current 5hmC sequencing methods

| Methods       | Advantages                                                                 | Disadvantages                                                                 |
|---------------|----------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Affinity-based methods |                                                                                     |                                                                                      |
| hMe-DIP18-20,35,61-63    | Antibody readily available; relatively straightforward procedure                   | Biased to heavily modified regions and CA-repeats; high background; antibody lot-to-lot inconsistency |
| Anti-CMS22             | Much less bias and lower background compared to anti-5hmC antibody                | PCR bias after bisulfite treatment                                              |
| hMe-Seal39,65-67       | Highly efficient, specific and unbiased labeling; built-in disulfide bond for easy pull-down | Requires synthesis of azide-modified glucose (now commercially available)        |
| GLIB22                 | Highly specific biotin-based pull-down; readily available materials               | Sodium periodate oxidation introduces high background; comparison to a nontrivial negative control is necessary |
| JBP-1 (refs. 33,68)    | Highly efficient one-step βGT labeling; readily available materials               | No published genome-wide profiling data for comparison                           |
| Single-base resolution methods |                                                                                     |                                                                                      |
| SMRT67             | Single-molecule sequencing, no PCR amplification required; strand-specific 5mC sequencing | Loss of quantitative information due to prior enrichment; higher sequencing capacity needed |
| oxB-Seq78           | Low-cost and readily available materials; simple procedure                         | Oxidation degradation of DNA; repeated bisulfite treatments to fully deaminate 5fC |
| TAB-Seq69          | Measure 5mC directly; readily deaminate 5caC under traditional bisulfite treatment | Requires highly active TET enzymes for high conversion rate of 5mC to 5caC           |

dithiothreitol67. hMe-Seal is robust with extremely low background and no bias64. It should also be noted that the glucose modification on the enriched DNA fragments does not interfere with polymerases employed regularly for library preparation in Illumina sequencing. Only with densely modified sites do we observe pausing with Taq polymerase. In such cases other polymerases can be used to replace Taq. Using hMe-Seal, we have done whole-genome profiling of 5hmC in mouse and human brain tissues. We found distinct age-dependent distribution of 5hmC in brain tissues as compared with ESCs. Specifically, we saw enrichment within gene bodies of expressed genes and upstream of the transcription start sites, but we observed depletion at the transcription start sites, suggesting a unique function of 5hmC in neurodevelopment39,65.

A related biotin-based 5hmC-profiling method is referred to as glucosylation, periodate oxidation, biotinylation (GLIB). It utilizes βGT to transfer an unmodified glucose to 5hmC, followed by cleavage of the vicinal hydroxyl groups in the glucose by sodium periodate to generate reactive aldehyde groups, which can then be biotinylated using an aldehyde-reactive hydroxyamine-biotin probe for further enrichment22. However, the sodium periodate oxidation may cause DNA damage and introduce high background. Nevertheless, with appropriate controls the GLIB method provides an alternative approach. Applying this method and the anti-CMS antibody-based enrichment, the Rao group revealed the distribution of 5hmC in ESCs as described above22.

After treating 5hmC with βGT, Klungland and colleagues showed that the J-binding protein 1 (JBP-1), which is known to interact with glucosylated 5hmU in certain ketoplastids, can also bind and therefore enrich glucosylated 5hmC for specific 5hmC profiling43,68. Thus, JBP-1 works as a naturally existing ‘antibody’ for glucosylated 5hmC.

Single-base-resolution sequencing methods

Although valuable, affinity-based genome-wide profiling methods have several disadvantages. First, these methods generate distribution maps with poor resolution as a result of the size limitation of the nucleic acid fragmentation and capture technology. Second, enrichment renders it impossible to measure the absolute abundance of the nucleotide modification. Third, the propensity of affinity-based methods to amplify frequent but weak signals may impose biases69. In contrast, a single-base-resolution mapping method, especially a whole-genome sequencing method without prior enrichment, could provide the most accurate and quantitative information regarding the modification.

The simplest way to achieve single-base resolution sequencing of a nucleotide variant would be to recognize its physical size or properties directly during sequencing. Unfortunately, the current second-generation sequencing technologies involve sample pre-amplification, which leads to the loss of base-modification information. Third-generation sequencing technologies that feature single-molecule sequencing and do not require sample pre-amplification may offer a solution70. The single-molecule, real-time (SMRT) sequencing developed by Pacific Biosciences records the incorporation of phospholinked nucleotides by individual DNA polymerase in real time71. By further monitoring the polymerase kinetics during replication, SMRT can directly detect DNA base modifications including 5mC and 5hmC, albeit with low confidence72. Through collaboration with Pacific Biosciences, we have successfully integrated hMe-Seal (Fig. 2) and SMRT sequencing to improve the polymerase kinetics for confident detection of 5hmC at single-base resolution67. Further technological advances are needed before this approach can be applied to the whole mammalian genome 5hmC sequencing. Other third-generation sequencing approaches, such as nanopore sequencing73, also have the potential to detect 5mC74 and 5hmC75,76 at the single-base level, but these applications are still in the early stages of development.

Bisulfite sequencing, the gold standard for single-base resolution sequencing of 5mC, can be adapted to essentially any sequencing platform. In this approach the distinct chemical reactions of cytosine and 5mC with sodium bisulfite (NaHSO3) (cytosine deaminates to uracil, whereas 5mC remains intact) are exploited to achieve single-base-resolution differentiation of cytosine from 5mC77. Complications arise, however, with all of the newly discovered cytosine derivatives. Under bisulfite conditions, cytosine, 5fC (which requires harsher conditions for complete deamination78 and 5caC76,69 undergo...
Figure 3 OxBS-Seq and TAB-seq for single-base resolution sequencing of 5hmC. (a) OxBS-Seq requires two bisulfite sequencings. In the first sequencing, 5mC in genomic DNA is oxidized to 5fC by KRuO₄, and subsequently converted into T by bisulfite treatment and PCR. In the second sequencing, genomic DNA is subjected to bisulfite treatment and PCR without KRuO₄ treatment. The first sequencing reveals sites of 5mC; this information is subtracted from the 5mC plus 5hmC sites provided by the second traditional bisulfite sequencing. (b) TAB-Seq directly reads out 5hmC in one bisulfite sequencing. 5hmC is protected from TET-mediated oxidation and bisulfite conversion by βGT-catalyzed glucosylation. Next, 5mC is oxidized by TET to 5caC, and subsequently converted into T after bisulfite treatment and PCR. Therefore, TAB-Seq reveals sites of 5hmC in genomic DNA with absolute abundance at each modification site. g5hmC, glucosylated 5-hydroxymethylcytosine.

RNA modifications
Chemical modifications (e.g., methylation) on DNA and histones have been widely accepted as key processes that regulate gene expression. In contrast to the limited types of modifications found in DNA, cellular RNAs, including mRNA and noncoding RNA, contain >100 structurally distinct post-transcriptional modifications at thousands of sites (http://rna-mdb.cas.albany.edu/RNAmods/). We have hypothesized that some of these RNA modifications can also be dynamic and reversible and may play regulatory roles analogous to reversible DNA and protein modifications. Traditional methods to determine the localization of RNA modifications such as TLC, primer extension, ligation, microarray or mass spectrometry are low throughput, laborious, time-consuming and especially difficult for low-abundance cellular RNAs such as mRNA. As a result, the functions of potential dynamic RNA modifications, especially those on low-abundance mRNA and noncoding RNA that will be discussed in this Review, remained largely unexplored owing to the lack of large-scale sequencing methods and lack of RNA demodification enzymes. In fact, before 2011, there were
Figure 4 New sequencing methods for RNA modifications. (a) Bisulfite sequencing can be used to map transcriptome-wide 5mC in RNA. (b) Inosine chemical erasing can be used to sequence inosine in a mammalian transcriptome. In the control group, inosine is converted into G by reverse transcription and PCR amplification. In the acrylonitrile treatment group, reverse transcription is blocked at the modified inosine site, which leads to identification of inosines on RNA. (c) m6A as a reversible RNA modification. m6A is generated by RNA methyltransferase(s) and removed by demethylases such as FTO. It further interacts with binding proteins and may regulate various biological functions. Its genomic distribution can be determined by antibody-based immunoprecipitation (IP). No known reversible chemical modifications on RNA that could affect gene expression.

Several recently developed high-throughput sequencing methods specific for RNA modifications have rekindled interest in functional dynamics of RNA modifications, in particular those in mRNA and noncoding RNA. For example, bisulfite sequencing was applied to map transcriptome-wide 5mC in RNA and reveal that 5mC exists not only in tRNA and rRNA as previously known, but also in mRNA and certain noncoding RNAs. A chemical method, termed inosine chemical erasing, which involves cyanoethylation combined with reverse transcription, was developed to sequence inosine (I) in mammalian transcriptomes (Fig. 4b). RNA editing converts A to I and C to U, and I may play regulatory roles. Although sequencing of RNA editing events is straightforward using current RNA-Seq technology, caution should be exercised when analyzing sequencing data so as to avoid errors that arise from copy number variants or sequencing errors. More comprehensive analyses and orthogonal approaches such as inosine chemical erasing should facilitate the discovery of additional RNA-editing events.

In 2011, our laboratory showed that m6A, the most prevalent internal mRNA modification, is a major substrate of the fat mass and obesity-associated protein FTO both in vitro and inside cells (mRNA was isolated by poly(T) oligo with subsequent removal of rRNA). Raising the possibility that this reversible RNA nucleotide modification could serve as a chemical mark to tune gene expression analogous to methylated nucleotides observed in DNA. Recently, antibodies raised against m6A were used to enrich m6A-containing RNA fragments for high-throughput sequencing (Fig. 4c). This m6A-Seq approach was applied to human and mouse samples, and revealed that the transcriptome-wide m6A distribution was dynamically modulated and preferentially enriched around stop codons, in 3'UTRs and within long internal exons (Tables 1 and 2). In addition, several m6A-binding proteins have been identified, suggesting a function for m6A in regulating cellular dynamics. This field of reversible RNA modifications holds great promise in uncovering new biology associated with RNA metabolism, localization and translation.

Perspective
The rapid progress of research on 5hmC has benefited from the rapid development of methods for 5hmC detection, profiling and now quantitative base-resolution mapping. These advances may guide studies of other nucleotide variants, especially the recently discovered 5fC and 5caC in mammalian DNA. The current lack of methods to reliably profile and quantitatively assess the location and abundance of these further oxidized 5mC derivatives substantially limits further research on these nucleotide variants. Antibodies against 5fC and 5caC are available for immunostaining, but given the low number of 5fC and 5caC nucleotides in mammalian genomic DNA (only p.p.m. of cytosine in mouse ESC), similar to the levels of DNA damage, it can be very challenging to apply an antibody-based capture strategy, which tends to favor densely populated modifications. Even if the antibodies can pull down certain genomic regions, such an approach will still have very limited coverage. One solution to this problem is to selectively label 5fC or 5caC with biotin. The high-affinity interaction between biotin and streptavidin can in principle capture every modification with no density or sequence-dependent bias, which is extremely important for reliable enrichment of scarce modifications. Chemical transformations are available for the aldehyde group in 5fC and the carboxylate group in 5caC, such as hydroxylamine–aldehyde condensation for 5fC (right at the time this paper was accepted a method describing the hydroxylamine-based profiling of 5fC was published online, which showed the enrichment of 5fC in CpG islands of promoters and exons (103) and EDC-mediated coupling for 5caC, which can be used to introduce a biotin group. However, both approaches can introduce high background noise as a result of side reactions of hydroxylamine and EDC with other functionalities on DNA (104,105). Therefore, careful tuning of the reaction conditions and appropriate controls are necessary. Besides chemical transformation, enzymatic approaches are also attractive if selective 5fC and 5caC enzymes can be developed. TDG is a good starting point because it can remove 5fC and 5caC and generate abasic sites for further labeling. TDG also recognizes T/G and U/G mismatches, which have to be repaired first.
Another possibility is to evolve an engineered βGT that can selectively label these modifications, especially 5caC. Compared to limited DNA modifications, hundreds of RNA modifications present an even greater technological challenge owing to the huge pool of structural and functional diversity. For instance, although m5A has been known for decades as an internal mRNA modification, it has only recently been recognized as another reversible nucleotide modification. Although distribution of m6A has been known for decades as an internal mRNA modification, although m6A has been known for decades as an internal mRNA modification, the huge pool of structural and functional diversity. For instance, cancer and early zygotes, where sample amounts can continue as refined or completely new methods are developed. In addition to high-resolution sequencing, methods are needed to analyze nucleotide variants in rare cells and in living cells. 5hmC and related nucleotide variants may play roles in the development of cancer and early zygotes, where sample amounts can be very limited. Therefore, sequencing methods that can deal with hundreds to thousands of cells or even single-cell sequencing will have a profound impact on fundamental biological understanding as well as diagnostics. Understanding the dynamics of these modifications in living biological systems would benefit from methods for high-resolution, single-molecule imaging. Recent discoveries of new nucleotide variants with epigenetic functions have stimulated the development of methods to detect, profile, and sequence these base modifications in the genome and transcriptome. In turn, the technological advances accelerate research to understand the biology of these nucleotide variants. This trend will continue as refined or completely new methods are developed.

ACKNOWLEDGMENTS

This study was supported by US National Institutes of Health (HG006827 to C.H.). We thank S.F. Reichard, for editing the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nbt.2398. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Klose, R.J. & Bird, A.P. Genomic DNA methylation: the mark and its mediators. Trends Biochem. Sci. 31, 89–97 (2006).
2. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13, 484–492 (2012).
3. Kriauciunas, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science 324, 929–930 (2009). Authors show that 5hmC is present in high levels in Purkinje cells, suggesting a role for 5hmC in neuronal function.
4. Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324, 930–935 (2009).

This paper showed that 5hmC is present in ESCs and is enzymatically produced through TET-catalyzed oxidation of 5mC.

5. Pfaffender, T. et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. Angew. Chem. Int. Ed. Engl. 50, 7008–7012 (2011).
6. Ito, S. et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 333, 1300–1303 (2011).

In addition, authors quantified 5fC and 5caC using LC-MS/MS in mouse ESCs and tissues.

7. He, Y.F. et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science 333, 1303–1307 (2011).

This paper revealed the first biochemically validated active DNA demethylation pathway.

8. Ito, S. et al. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature 466, 1129–1133 (2010).
9. Munozel, M., Globisch, D. & Carell, T. 5-Hydroxymethylcytosine, the sixth base of the genome. Angew. Chem. Int. Ed. Engl. 50, 6460–6468 (2011).

10. Bhutani, N., Burns, D.M. & Blau, H.M. DNA demethylation dynamics. Cell 146, 866–872 (2011).

11. Kang, K.P. et al. Genome-wide 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell 8, 200–213 (2011).

12. Williams, K. et al. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell 8, 200–213 (2011).

13. Tan, L. et al. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev. 25, 679–684 (2011).

14. Wu, H. et al. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature 473, 389–393 (2011).

15. Pastor, W.A. et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature 473, 394–397 (2011). Authors developed anti-CMS and the GLIB method for genome-wide 5hmC profiling and also demonstrated advantages of the biotin-based enrichment methods.

16. Maiti, A. & Drohat, A.C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine, potential implications for active demethylation of CpG sites. J. Biol. Chem. 286, 35334–35338 (2011).

17. Zhang, L. et al. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. Nat. Chem. Biol. 8, 328–330 (2012).

18. Metzker, M.L. Sequencing technologies—the next generation. Nat. Rev. Genet. 11, 31–46 (2010).

19. Song, C.X. & He, C. The hunt for 5-hydroxymethylcytosine: the sixth base. Epigenetics 3, 521–523 (2011).

20. Sulewski, A. et al. Detection of DNA methylation in eucaryotic cells. Folia Histochem. Cytobiol. 45, 315–324 (2007).

21. Fouse, S.D., Nagarajan, R.O. & Costello, J.F. Genome-scale DNA methylation analysis. Epigenetics 2, 105–117 (2010).

22. Gu, J., Su, Y., Zhang, C., Ming, G. & Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell 145, 423–434 (2011).

23. Cortellino, S. et al. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. Cell 146, 67–79 (2011).

24. Nabel, C.S. et al. AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. Nat. Chem. Biol. 8, 751–758 (2012).

25. Terragni, J., Bitinaite, J., Zheng, Y. & Pradhan, S. Biochemical characterization of recombinant beta-glucoyltransferase and analysis of global 5-hydroxymethylcytosine in unique genomes. Biochemistry 51, 1009–1019 (2012).

26. Robertson, A.B. et al. A novel method for the efficient and selective identification of 5-hydroxymethylcytosine in genomic DNA. Nucleic Acids Res. 39, e55 (2011).

27. Grippo, P., Iaccarino, M., Rossi, M. & Scarrano, E. Thin-layer chromatography of nucleotides, nucleosides and nucleic acid bases. Biochim. Biophys. Acta 95, 1–7 (1965).

28. Jin, S.G., Wu, X., Li, A.X. & Pfeifer, G.P. Genome mapping of 5-hydroxymethylcytosine in the human brain. Nucleic Acids Res. 39, 5015–5024 (2011).

29. Li, W. & Liu, M. Distribution of 5-hydroxymethylcytosine in different human tissues. J. Nucleic Acids 2011, 870726 (2011).

30. Jia, G. et al. N6-methyladenosine in nuclear RNA is a major substrate of the ubiquitously associated FTO. Nat. Chem. Biol. 7, 885–887 (2011). This paper reported FTO as the first known RNA demethylase that mediates m6A demethylation. Authors also quantified m6A by antibody and LC-MS/MS.
38. Ko, M. et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with 5-hmc-mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nat. Neurosci. 14, 1607–1616 (2011).
39. Haffter, M.C. et al. Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. Oncotarget 2, 627–637 (2011).
40. Beck, S. & Rakyan, V.K. The methylome: approaches for global DNA methylation landscape of human embryonic stem cells. Nat. Rev. Genet. 12, 632–645 (2011).
41. Xu, Y. et al. Development of oxBS-Seq for single-base resolution 5hmC sequencing and performed the whole-genome mapping of 5hmC sites in mouse and human ESCs.
42. Nienhuis, J. & Turner, S.W. Going beyond five bases in DNA sequencing. Curr. Opin. Struct. Biol. 22, 251–261 (2012).
43. Pickrell, J.K. et al. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. 40, 5023–5033 (2012).
44. Durairaj, A. & Limbach, P.A. Improving CMC-derivatization of pseudouridine in RNA-Seq data. Methods Enzymol. 501, 193–207 (2012).
45. Song, C.X. et al. Detection of 5-hydroxymethylcytosine in DNA by transferring a beta-glucosyl transferase from T4-infected Escherichia coli. J. Biol. Chem. 237, 1968–1976 (1962).
46. Song, C.X. et al. Detection of 5-hydroxymethylcytosine using JBP1-coated magnetic beads. Nat. Protoc. 7, 340–350 (2012).
47. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell 149, 1368–1380 (2012).
102. Meyer, K.D. et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3′-UTRs and near stop codons. Cell 149, 1635–1646 (2012). This paper reported m^6^A distribution in HEK cells and mouse brain. It also suggested a correlation between m^6^A deposition and microRNA binding on mRNA.

103. Raiber, E.A. et al. Genome-wide distribution of 5-formylcytosine in ES cells is associated with transcription and depends on thymine DNA glycosylase. Genome Biol. 13, R69 (2012).

104. Münzel, M., Lercher, L., Müller, M. & Carell, T. Chemical discrimination between dC and 5MedC via their hydroxylamine adducts. Nucleic Acids Res. 38, e192 (2010).

105. Chu, B.C., Wahl, G.M. & Orgel, L.E. Derivatization of unprotected polynucleotides. Nucleic Acids Res. 11, 6513–6529 (1983).

106. Matsuo, K., Nishikawa, K. & Shindo, M. Stereoselective synthesis of beta-glycosyl esters of cis-cinnamic acid and its derivatives using unprotected glycosyl donors. Tetrahedr. Lett. 52, 5688–5692 (2011).

107. Horowitz, S., Horowitz, A., Nilsen, T.W., Munns, T.W. & Rottman, F.M. Mapping of N6-methyladenosine residues in bovine prolactin mRNA. Proc. Natl. Acad. Sci. USA 81, 5667–5671 (1984).

108. Behm-Ansmant, I., Heilm, M. & Motorin, Y. Use of specific chemical reagents for detection of modified nucleotides in RNA. J. Nucleic Acids 2011, 408053 (2011).

109. Moran-Crusio, K. et al. Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. Cancer Cell 20, 11–24 (2011).

110. Quivoron, C. et al. TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer Cell 20, 25–38 (2011).

111. Lodhi, K., Jin, S.G., Pfeifer, G.P. & Szabo, P.E. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc. Natl. Acad. Sci. USA 108, 3642–3647 (2011).

112. Wossidlo, M. et al. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat. Commun. 2, 241 (2011).

113. Gu, T.P. et al. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature 477, 606–610 (2011).

114. Bokar, J.A., Shambaugh, M.E., Polayes, D., Matera, A.G. & Rottman, F.M. Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase. RNA 3, 1233–1247 (1997).

115. Frauer, C. et al. Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain. PLoS ONE 6, e21306 (2011).

116. Yildirim, O. et al. Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. Cell 147, 1498–1510 (2011).

117. Penn, N.W., Suwalski, R., O’Riley, C., Bojanowski, K. & Yura, R. The presence of 5-hydroxymethylcytosine in animal deoxynucleic acid. Biochem. J. 126, 781–790 (1972).

118. Valinluck, V. et al. Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). Nucleic Acids Res. 32, 4100–4108 (2004).

119. Jin, S.G. et al. 5-Hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with IDH1 mutations. Cancer Res. 71, 7360–7365 (2011).

120. Kudo, Y. et al. Loss of 5-hydroxymethylcytosine is accompanied with malignant cellular transformation. Cancer Sci. 103, 670–676 (2012).

121. Lian, C.G. et al. Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. Cell 150, 1135–1146 (2012).

122. Schiesser, S. et al. Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing. Angew. Chem. Int. Ed. Engl. 51, 6516–6520 (2012).

123. Kellinger, M.W. et al. 5-formylcytosine and 5-carboxycytosine reduce the rate and substrate specificity of RNA polymerase II transcription. Nat. Struct. Mol. Biol. 19, 831–833 (2012).