Spring 2019 – Epigenetics and Systems Biology
Discussion Session (Epigenetics)
Michael K. Skinner – Biol 476/576
Week 7 (February 21)

Epigenetics (History / Molecular Processes / Genomics)

Primary Papers

1. Haussmann, et al. (2016) Nature 540:301
2. Booth, et al. (2012) Science 336:934
3. Kelsey, et al. (2017) Science 358:69

Discussion

Student 18 – Ref #1 above

- What epigenetic mark was identified?
- What was the technology used?
- What function does the epigenetic mark have?

Student 19 – Ref #2 above

- What is hydroxymethylcytosine and how distinct from 5mC?
- What technology was used?
- What is the function of 5hmC and where expressed?

Student 20 – Ref #3 above

- What epigenetic marks were identified?
- What technology was used?
- How did the genomic profiling correlate with cellular differentiation?

LETTER

doi:10.1038/nature20577

m^{6}A potentiates *Sxl* alternative pre-mRNA splicing for robust *Drosophila* sex determination

Irmgard U. Haussmann[1,2], Zsuzsanna Bodi[3]*, Eugenio Sanchez-Moran[1]*, Nigel P. Mongan[4]*, Nathan Archer[3], Rupert G. Fray[3] & Matthias Soller[1]

N^6-methyladenosine (m^{6}A) is the most common internal modification of eukaryotic messenger RNA (mRNA) and is decoded by YTH domain proteins[1–7]. The mammalian mRNA m^{6}A methylosome is a complex of nuclear proteins that includes METTL3 (methyltransferase-like 3), METTL14, WTAP (Wilms tumour 1-associated protein) and KIAA1429. *Drosophila* has corresponding homologues named Ime4 and KAR4 (Inducer of meiosis 4 and Karyogamy protein 4), and Female-lethal (2)d (Fl(2)d) and Virilizer (Vir)[8–12]. In *Drosophila*, *fl(2)d* and *vir* are required for sex-dependent regulation of alternative splicing of the sex determination factor *Sex lethal* (*Sxl*)[13]. However, the functions of m^{6}A in introns in the regulation of alternative splicing remain uncertain[3]. Here we show that m^{6}A is absent in the mRNA of *Drosophila* lacking *Ime4*. In contrast to mouse and plant knockout models[5,7,14], *Drosophila Ime4*-null mutants remain viable, though flightless, and show a sex bias towards maleness. This is because m^{6}A is required for female-specific alternative splicing of *Sxl*, which determines female physiognomy, but also translationally represses *male-specific lethal 2* (*msl-2*) to prevent dosage compensation in females. We further show that the m^{6}A reader protein YT521-B decodes m^{6}A in the sex-specifically spliced intron of *Sxl*, as its absence phenocopies *Ime4* mutants. Loss of m^{6}A also affects alternative splicing of additional genes, predominantly in the 5′ untranslated region, and has global effects on the expression of metabolic genes. The requirement of m^{6}A and its reader YT521-B for female-specific *Sxl* alternative splicing reveals that this hitherto enigmatic mRNA modification constitutes an ancient and specific mechanism to adjust levels of gene expression.

In mature mRNA the m^{6}A modification is most prevalently found around the stop codon as well as in 5′ untranslated regions (UTRs) and in long exons in mammals, plants and yeast[2,3,6,7,15]. Since methylosome components predominantly localize to the nucleus, it has been speculated that m^{6}A localized in pre-mRNA introns could have a role in alternative splicing regulation in addition to such a role when present in long exons[9–12,16]. This prompted us to investigate whether m^{6}A is required for *Sxl* alternative splicing, which determines female sex and prevents dosage compensation in females[13]. We generated a null allele of the *Drosophila METTL3* methyltransferase homologue *Ime4* by imprecise excision of a *P* element inserted in the promoter region. The excision allele Δ22-3 deletes most of the protein-coding region, including the catalytic domain, and is thus referred to as *Ime4^{null}* (Fig. 1a). These flies are viable and fertile, but both flightless and this phenotype can be rescued by a genomic construct restoring *Ime4* (Fig. 1a, b). *Ime4* shows increased expression in the brain and, as in mammals and plants[17], localizes to the nucleus (Fig. 1c, d).

Following RNase T1 digestion and ^{32}P end-labelling of RNA fragments, we detected m^{6}A after guanosine (G) in poly(A) mRNA of adult flies at relatively low levels compared to other eukaryotes (m^{6}A/A ratio: 0.06%, Fig. 1g)[2,3,5], but at higher levels in unfertilized eggs (0.18%, Extended Data Fig. 1). After enrichment with an anti-m^{6}A antibody, m^{6}A is readily detected in poly(A) mRNA, but absent from *Ime4^{null}* flies (Fig. 1h–j).

As found in other systems, and consistent with a potential role in translational regulation[18–21], m^{6}A was detected in polysomal mRNA (0.1%, Fig. 1k), but not in the poly(A)-depleted rRNA fraction. This also confirmed that any m^{6}A modification in rRNA is not after G in *Drosophila* (Fig. 1l).

Consistent with our hypothesis that m^{6}A plays a role in sex determination and dosage compensation, the number of *Ime4^{null}* females was reduced to 60% compared to the number of males ($P < 0.0001$), whereas in the control strain female viability was 89% (Fig. 2a). The key regulator of sex determination in *Drosophila* is the RNA-binding protein Sxl, which is specifically expressed in females. Sxl positively auto-regulates expression of itself and its target *transformer* (*tra*) through alternative splicing to direct female differentiation[13]. In addition, *Sxl* suppresses translation of *msl-2* to prevent upregulation of transcription on the X chromosome for dosage compensation (Fig. 2b); full suppression also requires maternal factors[22]. Accordingly, female viability was reduced to 13% by removal of maternal m^{6}A together with zygotic heterozygosity for *Sxl* and *Ime4* (*Ime4$^{\Delta 22\text{-}3}$* females crossed with *Sxl^{7B0}* males, a *Sxl* null allele, $P < 0.0001$). Female viability of this genotype is completely rescued by a genomic construct (Fig. 2a) or by preventing ectopic activation of dosage compensation by removal of *msl-2* (*msl-2^{227}/Df(2L)Exel7016*, Fig. 2a). Hence, females are non-viable owing to insufficient suppression of *msl-2* expression, resulting in upregulation of gene expression on the X chromosome from reduced Sxl levels. In the absence of *msl-2*, disruption of *Sxl* alternative splicing resulted in females with sexual transformations (32%, $n = 52$) displaying male-specific features such as sex combs (Fig. 2c–e), which were mosaic to various degrees, indicating that Sxl threshold levels are affected early during establishment of sexual identities of cells and/or their lineages[13]. In the presence of maternal Ime4, *Sxl* and *Ime4* do not genetically interact (*Sxl^{7B0}/FM7* females crossed with *Ime4^{null}* males, 103% female viability, $n = 118$). In addition, Sxl is required for germline differentiation in females and its absence results in tumorous ovaries[23]. Consistent with this, we detected tumorous ovaries in *Sxl^{7B0}*/+;*Ime4^{null}*/+ daughters from *Ime4^{null}* females (22%, $n = 18$, Extended Data Fig. 2), but not in homozygous *Ime4^{null}* or heterozygous *Sxl^{7B0}* females ($n = 20$ each).

Furthermore, levels of the *Sxl* female-specific splice form were reduced to approximately 50%, consistent with a role for m^{6}A in *Sxl* alternative splicing (Fig. 2f and Extended Data Fig. 3a). As a result, female-specific splice forms of *tra* and *msl-2* were also significantly reduced in adult females (Fig. 2f and Extended Data Fig. 3b, c).

To obtain more comprehensive insights into *Sxl* alternative splicing defects in *Ime4^{null}* females, we examined splice junction reads from

[1]School of Biosciences, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [2]School of Life Science, Faculty of Health and Life Sciences, Coventry University, Coventry CV1 5FB, UK. [3]School of Biosciences, Plant Science Division, University of Nottingham, Sutton Bonington, Loughborough LE12 5RD, UK. [4]School of Veterinary Medicine and Sciences, University of Nottingham, Sutton Bonington, Loughborough LE12 5RD, UK.
*These authors contributed equally to this work.

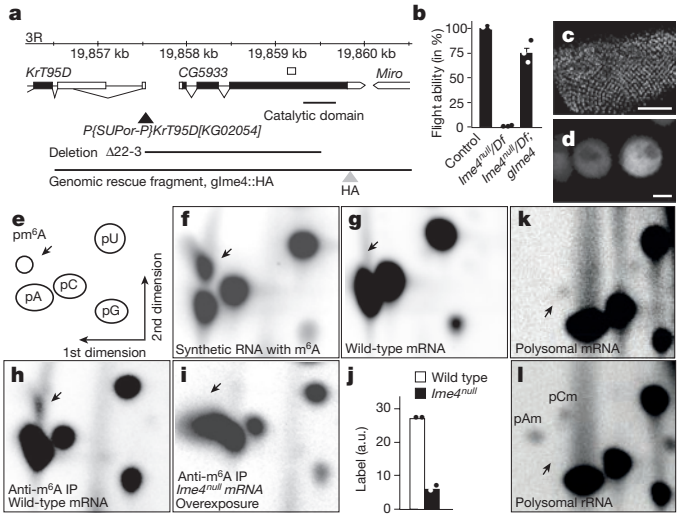


Figure 1 | Analysis of *Ime4* null mutants and m^{6}A methylation in *Drosophila*. **a**, Genomic organization of the *Ime4* locus depicting the transposon (black triangle) used to generate the deletion Δ22-3, which is a *Ime4* null allele and the hemaglutinin (HA)-tagged genomic rescue fragment. **b**, Flight ability of *Ime4*^{null}/*Df(3R)Exel6197* shown as mean ± s.e.m. of $n = 3$ (dots). *gIme4*, genomic rescue construct. **c**, **d**, Nuclear localization of Ime4::HA in eye discs (**c**) and brain neurons (**d**) expressed from *UAS*. Scale bars, 50 and 1 μm in **c** and **d**, respectively. **e**, Schematic diagram of 2D thin-layer chromatography (TLC). **f**, TLC from an *in vitro* transcript containing m^{6}A. **g**, TLC from mRNA of adult flies. **h**, **i**, TLC of fragmented mRNA after enrichment with an anti-m^{6}A antibody from wild-type (**h**) and *Ime4*^{null} flies (**i**, overexposed). IP, immunoprecipitation. **j**, Quantification of immunoprecipitated ^{32}P label shown as normalized mean of $n = 2$ (dots). a.u., arbitrary units. **k**, **l**, TLC from mRNA (**k**) or rRNA (**l**) from polysomes from wild-type flies. pAm, 2′-*O*-methyladenosine; pCm, 2′-*O*-methylcytidine.

RNA-seq. Besides the significant increase in inclusion of the male-specific *Sxl* exon in *Ime4*^{null} females (Fig. 2f–h and Extended Data Fig. 3a), cryptic splice sites and increased numbers of intronic reads were detected in the regulated intron. Consistent with our reverse transcription polymerase chain reaction (RT–PCR) analysis of *tra*, the reduction of female splicing in the RNA sequencing is modest, and as a consequence, alternative splicing differences of Tra targets *dsx* and *fru* were not detected in whole flies, suggesting that cell-type-specific fine-tuning is required to generate splicing robustness rather than being an obligatory regulator (Extended Data Fig. 4a–c). In agreement with dosage-compensation defects as a main consequence of *Sxl* dysregulation in *Ime4*^{null} mutants, X-linked, but not autosomal, genes are significantly upregulated in *Ime4*^{null} females compared to controls ($P < 0.0001$, Extended Data Fig. 4d, e).

Furthermore, *Sxl* mRNA is enriched in pull-downs with an m^{6}A antibody compared to m^{6}A-deficient yeast mRNA added for quantification (Fig. 2i). This enrichment is comparable to what was observed for m^{6}A-pull-down from yeast mRNA[24].

To map m^{6}A sites in the intron of *Sxl*, we employed an *in vitro* m^{6}A methylation assay using *Drosophila* nuclear extracts and labelled substrate RNA. m^{6}A methylation activity was detected in the vicinity of alternatively spliced exons (Fig. 2j, RNAs B, C, and E). Further fine-mapping localized m^{6}A in RNAs C and E to the proximity of Sxl-binding sites (Extended Data Fig. 5). Likewise, the female-lethal single amino acid substitution alleles *fl(2)d*^{1} and *vir*^{2F} interfere with Sxl recruitment, resulting in impaired *Sxl* auto-regulation and inclusion of the male-specific exon[25]. Female lethality of these alleles can be rescued by *Ime4*^{null} heterozygosity ($P < 0.0001$, Fig. 2k), further demonstrating the involvement of the m^{6}A methylosome in *Sxl* alternative splicing.

Next, we globally analysed alternative splicing changes in *Ime4*^{null} females compared to the wild-type control strain. As described earlier (Fig. 2h), a statistically significant reduction in female-specific alternative splicing of *Sxl* (ΔPSI (difference in percentage spliced in) = 0.34, $q = 9 \times 10^{-8}$) was observed. In addition, 243 alternative splicing events in 163 genes were significantly different in *Ime4*^{null} females ($q < 0.05$, ΔPSI > 0.2), equivalent to around 2% of alternatively spliced genes in *Drosophila* (Supplementary Table 1). Six genes for which the alternative splicing products could be distinguished on agarose gels were confirmed by RT–PCR (Extended Data Fig. 6). Notably, lack of *Ime4* did not affect global alternative splicing and no specific type of alternative splicing event was preferentially affected. However, alternative first exon (18% versus 33%) and

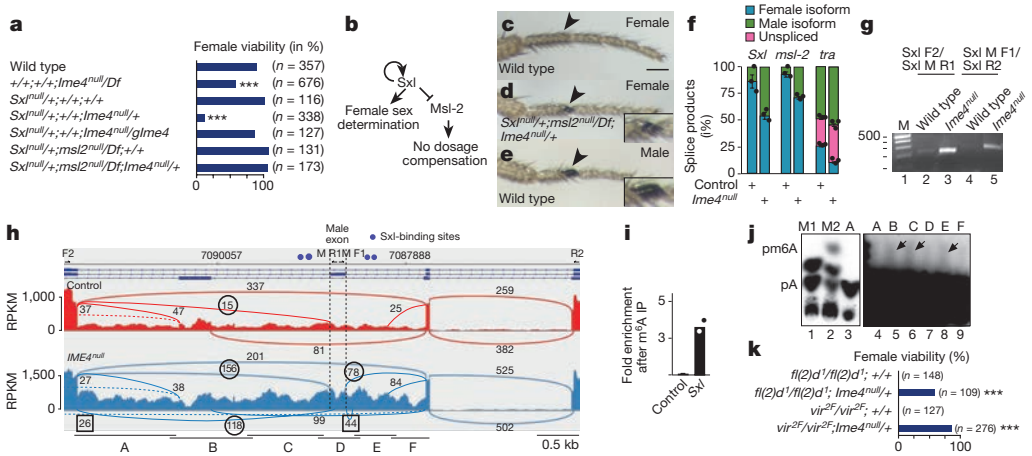


Figure 2 | m^{6}A methylation is required for *Sxl* alternative splicing in sex determination and dosage compensation. **a**, Female viability of indicated genotypes devoid of maternal m^{6}A (*n*, total number of flies, ***$P < 0.0001$). **b**, Schematic depicting Sxl control of female differentiation. **c**–**e**, Front legs of indicated genotypes. Scale bar, 100 μm. The arrowhead points towards the position of the sex comb normally present only in males (magnified in insets). **f**, Ratio of sex-specific splice isoforms from adult females from RT–PCR shown as mean ± s.e.m. ($n = 3$, $P < 0.01$ for the change in female isoforms). **g**, RT–PCR for male-specific *Sxl* splicing in control and *Ime4*^{null} females. Lanes are numbered at the bottom. M, DNA size marker. **h**, Sashimi plot depicting Tophat-mapped RNA sequencing reads and exon junction reads from control and *Ime4*^{null} females below the annotated gene model. Male-specific splice junction reads are circled and cryptic splice sites are boxed. RNA fragments used for m^{6}A *in vitro* methylation assays are indicated at the bottom (A–F). Primers are indicated on top of exons with arrows. **i**, Presence of m^{6}A in *Sxl* transcripts detected by m^{6}A immunoprecipitation followed by qPCR from nuclear mRNA of early embryos (shown as mean of $n = 2$, dots). **j**, One-dimensional TLC of *in vitro*-methylated, [^{32}P]-ATP-labelled substrate RNAs shown in **g**. Nucleotide markers from *in vitro* transcripts in the absence (M1) or presence (M2) of m^{6}A. The right image shows an overexposure of the same TLC, lanes 3 and 4 show the same sample. **k**, Rescue of female lethality of female-lethal *fl(2)d*^{1} and *vir*^{2F} alleles by removal of one copy of *Ime4*.

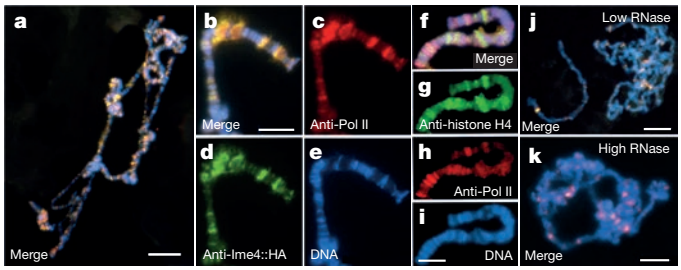


Figure 3 | Ime4 co-localizes to sites of transcription. **a**–**e**, Polytene chromosomes from salivary glands expressing IME::HA stained with anti-Pol II (red, **c**), anti-HA (green, **d**) and DAPI (DNA, blue, **e**), or merged (yellow, **a** and **b**). **f**–**i**, Polytene chromosomes stained with anti-Pol II (red, **h**), anti-histone H4 (green, **g**) and DAPI (DNA, blue, **i**), or merged (yellow, **f**). **j**, **k**, Polytene chromosomes treated with low (**j**, 2 μg ml^{-1}) and high (**k**, 10 μg ml^{-1}) RNase A concentration before staining with anti-Pol II, anti-histone H4 and DAPI. Scale bars, 20 μm (**a**, **j**, **k**) and 5 μm (**e**, **i**).

mutually exclusive exon (2% versus 15%) events were reduced in *Ime4*^{null} compared to a global breakdown of alternative splicing in wild-type *Drosophila*, mostly to the extent of retained introns (16% versus 6%), alternative donor (16% versus 9%) and unclassified events (14% versus 6%) (Extended Data Fig. 7a). Notably, the majority of affected alternative splicing events in *Ime4*^{null} were located to the 5′ UTR, and these genes had a significantly higher number of AUG start codons in their 5′ UTR compared to the 5′ UTRs of all genes (Extended Data Fig. 7b, c). Such a feature has been shown to be relevant to translational control under stress conditions[26].

The majority of the 163 differentially alternatively spliced genes in Ime4 females are broadly expressed (59%), while most of the remainder are expressed in the nervous system (33%), consistent with higher expression of *Ime4* in this tissue (Extended Data Fig. 7d). Accordingly, Gene Ontology analysis revealed a highly significant enrichment for genes involved synaptic transmission ($P < 7 \times 10^{7}$, Supplementary Table 1).

Since the absence of m^{6}A affects alternative splicing, m^{6}A marks are probably deposited co-transcriptionally before splicing. Co-staining of polytene chromosomes with antibodies against haemagglutinin (HA)-tagged Ime4 and RNA Pol II revealed broad co-localization of Ime4 with sites of transcription (Fig. 3a–e), but not with condensed chromatin—visualized with antibodies against histone H4 (Fig. 3f–i). Furthermore, localization of Ime4 to sites of transcription is RNA-dependent, as staining for Ime4, but not for RNA Pol II, was reduced in an RNase-dependent manner (Fig. 3j, k).

Although m^{6}A levels after G are low in *Drosophila* compared to other eukaryotes, broad co-localization of Ime4 to sites of transcription suggests profound effects on the gene expression landscape. Indeed, differential gene expression analysis revealed 408 differentially expressed genes (≥2-fold change, $q \leq 0.01$) where 234 genes were significantly upregulated and 174 significantly downregulated in neuron-enriched head/thorax of adult *Ime4*^{null} females ($q < 0.01$, at least twofold, Supplementary Table 2). Cataloguing these genes according to function reveals prominent effects on gene networks involved in metabolism, including reduced expression of 17 genes involved in oxidative phosphorylation ($P < 0.0001$, Supplementary Table 2). Notably, overexpression of the m^{6}A mRNA demethylase FTO in mice leads to an imbalance in energy metabolism resulting in obesity[27].

Next, we tested whether either of the two substantially divergent YTH proteins, YT521-B and CG6422 (Fig. 4a), decodes m^{6}A marks in *Sxl* mRNA. When transiently transfected into male S2 cells, YT521-B localizes to the nucleus, whereas CG6422 is cytoplasmic (Fig. 4b–d, Extended Data Fig. 8). Nuclear YT521-B can switch *Sxl* alternative splicing to the female mode and also binds to the *Sxl* intron in S2 cells (Fig. 4e, f). *In vitro* binding assays with the YTH domain of YT521-B demonstrate increased binding of m^{6}A-containing RNA (Extended Data Fig. 9). *In vivo*, YT521-B also localizes to the sites of transcription (Extended Data Fig. 10).

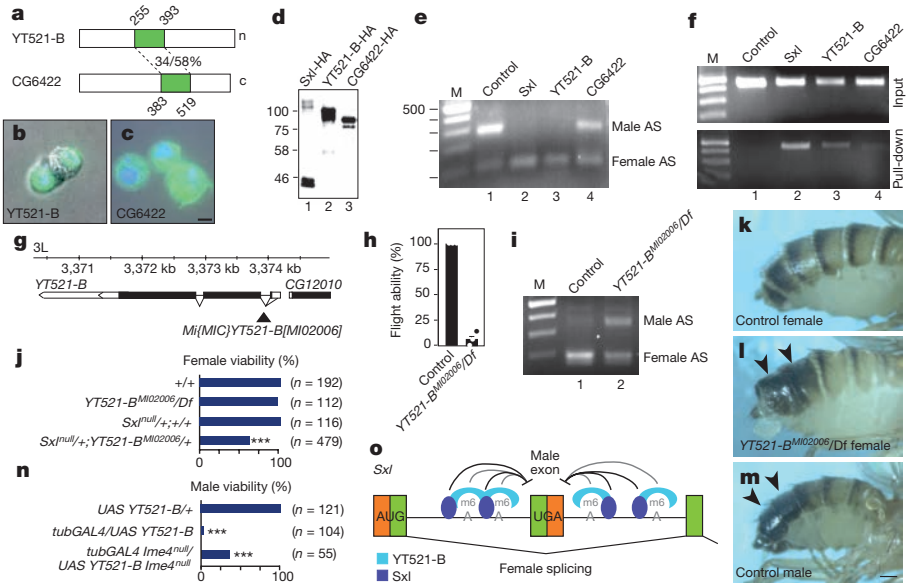


Figure 4 | YTH protein YT521-B decodes m^{6}A methylation in *Sxl*. **a**, Domain organization of *Drosophila* YTH proteins (YTH domain in green). n, nuclear; c, cytoplasmic. **b**–**d**, Cellular localization and size of HA-tagged YT521-B and CG6422 in S2 cells. Scale bar, 1 μm. **e**, Suppression of male-specific *Sxl* alternative splicing (AS) upon expression of Sxl and YT521-B, but not CG6422 in male S2 cells. **f**, Binding of YT521-B to pre-mRNA of the regulated *Sxl* intron. **g**, Genomic organization of the *YT521-B* locus depicting the transposon (black triangle) disrupting the ORF. **h**, Flight ability of *YT521-B*^{MI02006}/*Df(3L)Exel6094* shown as mean ± s.e.m. ($n = 3$). **i**, *Sxl* alternative splicing in female wild-type and *YT521-B*^{MI02006}/*Df(3L)Exel6094* flies. **j**, Female viability of indicated genotypes (*n*, total number of flies) reared at 29 °C. **k**–**m**, Abdominal pigmentation of indicated genotypes reared at 29 °C. The arrowheads point towards the position of the dark pigmentation normally present only in males. Scale bar, 100 μm. **n**, YT521-B was overexpressed from a *UAS* transgene with *tubulinGAL4* (2nd chromosome insert) in wild-type or *Ime4*^{null} flies at 27 °C. **o**, Model for female-specific *Sxl* alternative splicing by Sxl, m^{6}A and its reader YT521-B in co-operatively suppressing inclusion of the male-specific exon.

To further examine the role of YT521-B in decoding m^{6}A we analysed *Drosophila* strain *YT521-B*^{MI02006}, where a transposon in the first intron disrupts *YT521-B*. This allele is also viable (*YT521-B*^{MI02006}/*Df(3L)Exel6094*; Fig. 4g, h, j), and phenocopies the flightless phenotype and the female *Sxl* splicing defect of *Ime4*^{null} flies (Fig. 4h, i). Likewise, removal of maternal *YT521-B* together with zygotic heterozygosity for *Sxl* and *YT521-B* reduces female viability ($P < 0.0001$, Fig. 4j) and results in sexual transformations (57%, $n = 32$) such as male abdominal pigmentation (Fig. 4k–m). In addition, overexpression of YT521-B results in male lethality, which can be rescued by removal of *Ime4*, further reiterating the role of m^{6}A in *Sxl* alternative splicing ($P < 0.0001$, Fig. 4n). Since *YT521-B* phenocopies *Ime4* for *Sxl* splicing regulation, it is the main nuclear factor for decoding m^{6}A present in the proximity of the Sxl-binding sites. YT521-B bound to m^{6}A assists Sxl in repressing inclusion of the male-specific exon, thus providing robustness to this vital gene regulatory switch (Fig. 4o).

Nuclear localization of m^{6}A methylosome components suggested a role for this "fifth" nucleotide in alternative splicing regulation. Our discovery of the requirement of m6A and its reader YT521-B for female-specific *Sxl* alternative splicing has important implications for understanding the fundamental biological function of this enigmatic mRNA modification. Its key role in providing robustness to *Sxl* alternative splicing to prevent ectopic dosage compensation and female lethality, together with localization of the core methylosome component Ime4 to sites of transcription, indicates that the m^{6}A modification is part of an ancient, yet unexplored mechanism to adjust gene expression. Hence, the recently reported role of m^{6}A methylosome components in human dosage compensation[28,29] further support such a role and suggests that m^{6}A-mediated adjustment of gene expression might be a key step to allow for the development of the diverse sex determination mechanisms found in nature.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 March; accepted 25 October 2016.
Published online 30 November 2016.

1. Luo, S. & Tong, L. Molecular basis for the recognition of methylated adenines in RNA by the eukaryotic YTH domain. *Proc. Natl Acad. Sci. USA* **111,** 13834–13839 (2014).
2. Meyer, K. D. *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* **149,** 1635–1646 (2012).
3. Dominissini, D. *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485,** 201–206 (2012).
4. Perry, R. P. & Kelley, D. E. Existence of methylated messenger RNA in mouse L cells. *Cell* **1,** 37–42 (1974).
5. Zhong, S. *et al.* MTA is an *Arabidopsis* messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *Plant Cell* **20,** 1278–1288 (2008).
6. Schwartz, S. *et al.* High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* **155,** 1409–1421 (2013).
7. Ke, S. *et al.* A majority of m6A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes Dev.* **29,** 2037–2053 (2015).
8. Liu, J. *et al.* A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.* **10,** 93–95 (2014).
9. Horiuchi, K. *et al.* Identification of Wilms' tumor 1-associating protein complex and its role in alternative splicing and the cell cycle. *J. Biol. Chem.* **288,** 33292–33302 (2013).
10. 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).
11. Penalva, L. O. *et al.* The *Drosophila fl(2)d* gene, required for female-specific splicing of *Sxl* and *tra* pre-mRNAs, encodes a novel nuclear protein with a HQ-rich domain. *Genetics* **155,** 129–139 (2000).
12. Niessen, M., Schneiter, R. & Nothiger, R. Molecular identification of virilizer, a gene required for the expression of the sex-determining gene Sex-lethal in *Drosophila melanogaster*. *Genetics* **157,** 679–688 (2001).
13. Schütt, C. & Nöthiger, R. Structure, function and evolution of sex-determining systems in Dipteran insects. *Development* **127,** 667–677 (2000).
14. Geula, S. *et al.* Stem cells. m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* **347,** 1002–1006 (2015).
15. Luo, G. Z. *et al.* Unique features of the m6A methylome in *Arabidopsis* thaliana. *Nat. Commun.* **5,** 5630 (2014).
16. Xiao, W. *et al.* Nuclear m^{6}A reader YTHDC1 regulates mRNA splicing. *Mol. Cell* **61,** 507–519 (2016).
17. Hongay, C. F. & Orr-Weaver, T. L. *Drosophila* Inducer of MEiosis 4 (IME4) is required for Notch signaling during oogenesis. *Proc. Natl Acad. Sci. USA* **108,** 14855–14860 (2011).
18. Bodi, Z., Bottley, A., Archer, N., May, S. T. & Fray, R. G. Yeast m^{6}A methylated mRNAs are enriched on translating ribosomes during meiosis, and under rapamycin treatment. *PLoS One* **10,** e0132090 (2015).
19. Wang, X. *et al.* N^6-methyladenosine modulates messenger RNA translation efficiency. *Cell* **161,** 1388–1399 (2015).
20. Meyer, K. D. *et al.* 5′ UTR m^{6}A Promotes Cap-Independent Translation. *Cell* **163,** 999–1010 (2015).
21. Zhou, J. *et al.* Dynamic m^{6}A mRNA methylation directs translational control of heat shock response. *Nature* **526,** 591–594 (2015).
22. Zaharieva, E., Haussmann, I. U., Bräuer, U. & Soller, M. Concentration and localization of co-expressed ELAV/Hu proteins control specificity of mRNA processing. *Mol. Cell. Biol.* **35,** 3104–3115 (2015).
23. Salz, H. K. Sex, stem cells and tumors in the *Drosophila* ovary. *Fly (Austin)* **7,** 3–7 (2013).
24. Bodi, Z., Button, J. D., Grierson, D. & Fray, R. G. Yeast targets for mRNA methylation. *Nucleic Acids Res.* **38,** 5327–5335 (2010).
25. Hilfiker, A., Amrein, H., Dübendorfer, A., Schneiter, R. & Nöthiger, R. The gene virilizer is required for female-specific splicing controlled by Sxl, the master gene for sexual development in *Drosophila*. *Development* **121,** 4017–4026 (1995).
26. Starck, S. R. *et al.* Translation from the 5′ untranslated region shapes the integrated stress response. *Science* **351,** aad3867 (2016).
27. Church, C. *et al.* Overexpression of Fto leads to increased food intake and results in obesity. *Nat. Genet.* **42,** 1086–1092 (2010).
28. Moindrot, B. *et al.* A pooled shRNA screen identifies Rbm15, Spen, and Wtap as factors required for Xist RNA-mediated silencing. *Cell Reports* **12,** 562–572 (2015).
29. Patil, D. P. *et al.* m^{6}A RNA methylation promotes XIST-mediated transcriptional repression. *Nature* **537,** 369–373 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Horabin, N. Perrimon and the Bloomington, Harvard and Kyoto stock centres for fly lines, BacPAc for DNA clones, E. Zaharieva and M. L. Li for help with imaging, W. Arlt and R. Michell for comments on the manuscript, and J.-Y. Roignant for communication of results before publication. We acknowledge funding from the BBSRC (BB/M008606/1) to R.F.

Author Contributions I.U.H. and M.S. performed biochemistry, cell biology and genetic experiments, E.S.M. stained chromosomes, and Z.B., N.A. and R.F. performed biochemistry experiments. N.M. analysed sequencing data. I.U.H., R.F. and M.S. conceived the project and wrote the manuscript with help from N.M. and Z.B.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.S. (m.soller@bham.ac.uk).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

***Drosophila* genetics, generation of constructs and transgenic lines.** The deletion allele *Ime4*$^{\Delta 22\text{-}3}$ was obtained from imprecise excision of the transposon *P{SUPor-P}KrT95D* and mapped by primers 5933 F1 (CTCGCTCTA TTTCTCTTCAGCACTCG) and 5933 R9 (CCTCCGCAACGATCACAT CGCAATCGAG). To obtain a viable line of *Ime4*^{null}, the genetic background was cleaned by out-crossing to *Df(3R)Exel6197*. Flight ability was scored as the number of flies capable of flying out of a Petri dish within 30 s for groups of 15–20 flies for indicated genotypes. Viability was calculated from the numbers of females compared to males of the correct genotype and statistical significance was determined by a χ^2 test (GraphPad Prism). Unfertilized eggs were generated by expressing sex-peptide in virgin females as described[30].

The genomic rescue construct was retrieved by recombineering (Genebridges) from BAC clone *CH321-79E18* by first cloning homology arms with *Spe*I and *Acc*65I into *pUC3GLA* separated by an *Eco*RV site for linearization (CTCCGCCGCCGG AACCGCCGCCTCCTCCGCCACTTTGCAGGTTGAGCGGACCGCCTCCAG GGCCGCTGCCGCCGGTGCCGCTGATATCCCAGCATGGTAGCTGCGGCC ACTCCTAGTCCCGCCTTTAACCACAGCTTGGGGTCCTCCGTCATCAGG CCGAATTGCCTCGAG). An HA-tag was then fused to the end of the ORF using two PCR amplicons and *Sac*I and *Xho*I restriction sites. This construct was the inserted into *PBac{y+-attB-3B}VK00002* at 76A as described[31].

The *Ime4 UAS* construct was generated by cloning the ORF from fly cDNA into a modified *pUAST* with primers Adh dMT-A70 F1 EI (GCAGAATTCGAG ATCtAAAGAGCCTGCTAAAGCAAAAAAGAAGTCACCATGGCAGATGCGT GGGACATAAAATCAC) and dMT-A70 HA R1 Spe (GGTAACTAGTCTTTTG TATTCCATTGATCGACGCCGCATTGG) by adding a translation initiation site from the *Adh* gene and two copies of an HA tag to the end of the ORF. This construct was then also inserted into *PBac{y+-attB-3B}VK00002* at 76A.

For transient transfection in S2 cells, *YT52B-1* and *CG6422* ORFs were amplified from fly cDNA by a combination of nested and fusion PCR incorporating a translation initiation site from the *Adh* gene using primers CG6422 Adh F1 (GCCTGCTAAAGCAAAAAAGAAGTCACCACATGTCAGGCGTG GATCAGATGAAAATACCAG), pACT Adh CG6422 F1 (CCAGAGACCCCGGA TCCAGATATCAAAGAGCCTGCTAAAGCAAAAAAGAAGTCACCAC), CG 6422 Adh R1, (GATTCCTGCGAACAGGTCCCGTGGGCGAAAC) and CG6422 3′ F1 (CCCACGGGACCTGTTCGCAGGAATCTAG), CG6422 3′ R1 (CATTGC TTCGCATTTTATCCTTGTCCGTGTCCTTAAAGCGCACGCCGATTTTAAT TTG), pACT CG6422 3×HA R1 (GTGGAGATCCATGGTGGCGGAGCTCGA GGAATATTCATTGCTTCGCATTTTATCCTTGTC) for *CG6422* and primers YT521 Adh F1, (AAGCAAAAAAGAAGTCACATGCCAAGAGCAGCCCGTA AACAAACGCTGCCGATGCGCGAG), pACT Adh YT521 F1 (CCAGAGACC CCGGATCCAGATATCAAAGAGCCTGCTAAAGCAAAAAAGAAGTCACAT GCC), YT521 Adh R1 (TGCCATCCGGGCGAATCCTGCAAATTTACC ACTCTCGTTGACCGAGAAAATGAGCAGGAC) and YT521 3′ F1(GC AGGATTCGCCCGGATGGCAGCCCCCTCAC), pact YT521 R1 (GGTGGAG ATCCATGGTGGCGGAGCTCGAGCGCCTGTTGTCCCGATAGCTTCGCTG) for *YT521-B*, and cloned into a modified *pACT* using Gibson Assembly (NEB) also incorporating HA epitope tags at the C terminus. Constructs were verified by Sanger sequencing. The Sxl-HA expression vector was a gift from N. Perrimon[32].

The *YT521-B UAS* construct was generated by sub-cloning the ORF from the pACT vector into a modified *pUAST* with primers YT521 Adh F1 (AAGCAAAA AAGAAGTCACATGCCAAGAGCAGCCCGTAAACAAACGCTGCCGATGCG CGAG), YT521 Adh F2 (TAGGGAATTGGGAATTCGAGATCTAAAGAGCCT GCTAAAGCAAAAAAGAAGTCACATGCC) and YT521 3′ R1 (GGGCACGT CGTAGGGGTACAGACTAGTCTCGAGGCGCCTGTTGTCCCGATAGCTTC GCTG) by adding a translation initiation site from the *Adh* gene and two copies of an HA tag to the end of the ORF. This construct was then also inserted into *PBac{y+-attB-3B}VK00002* at 76A.

Essential parts of all DNA constructs were sequence-verified.

Cell culture, transfections and immune-staining of S2 cells. S2 cells (ATCC) were cultured in Insect Express medium (Lonza) with 10% heat-inactivated FBS and 1% penicillin/streptomycin. The Drosophila S2 cell line was verified to be male by analysing *Sxl* alternative splicing using species-specific primers Sxl F2 (ATGTACGGCAACAATAATCCGGGTAG) and Sxl R2 (CATTGTAACCACGACGCGACGATG) to confirm species and gender (Extended Data Fig. 8). Transient transfections were done with Mirus Reagent (Bioline) according to the manufacturer's instruction and cells were assayed 48 h after transfection for protein expression or RNA binding of expressed proteins. To adhere S2 cells to a solid support, Concanavalin A (Sigma) coated glass slides (in 0.5 mg ml^{-1}) were added 1 day before transfection, and cells were stained 48 h after transfection with antibodies as described. Transfections and follow up experiments were repeated at least once.

RNA extraction, RT–PCR, qPCR, immunoprecipitation and western blots. Total RNA was extracted using Tri-reagent (SIGMA) and reverse transcription was done with Superscript II (Invitrogen) according to the manufacturer's instructions using an oligodT17V primer. PCR for *Sxl*, *tra*, *msl2* and *ewg* was done for 30 cycles with 1 μl of cDNA with primers Sxl F2, Sxl R2 or Sxl NP R3 (GAGAATGGGACATCCCAAATCCACG), Sxl M F1 (GCCCAGA AAGAAGCAGCCACCATTATCAC), Sxl M R1 (GCGTTTCGTTGGCGAG GAGACCATGGG), Tra FOR (GGATGCCGACAGCAGTGGAAC), Tra REV (GATCTGGAGCGAGTGCGTCTG), Msl-2 F1 (CACTGCGGTCA CACTGGCTTCGCTCAG), Msl-2 R1 (CTCCTGGGCTAGTTACCTGCAATTC CTC), Ewg 4F and Ewg 5R and quantified with ImageQuant (BioRad)[22]. Experiments included at least three biological replicates.

For qPCR, reverse transcription was carried out on input and pull-down samples spiked with yeast RNA using ProtoScript II reverse transcriptase and random nanomers (NEB). Quantitative PCR was carried out using 2× SensiMix Plus SYBR Low ROX master mix (Quantace) using normalizer primers ACT1 F1 (TTAC GTCGCCTTGGACTTCG) and ACT1 R1 (TACCGGCAGATTCCAAACCC) and for Sxl, Sxl ZB F1 (CACCACAATGGCAGCAGTAG) and Sxl ZB R1 (GGGGTT GCTGTTTGTTGAGT). Samples were run in triplicate for technical repeats and duplicate for biological repeats. Relative enrichment levels were determined by comparison with yeast *ACT1*, using the $2^{-\Delta\Delta C_t}$ method[33].

For immunoprecipitations of *Sxl* RNA bound to Sxl or YTH proteins, S2 cells were fixed in PBS containing 1% formaldehyde for 15 min, quenched in 100 mM glycine and disrupted in IP-Buffer (150 mM NaCl, 50 mM Tris–HCL, pH 7.5, 1% NP-40, 5% glycerol). After IP with anti-HA beads (Sigma) for 2 h in the presence of Complete Protein Inhibitor (Roche) and 40 U RNase inhibitors (Roche), IP precipitates were processed for *Sxl* RT–PCR using gene-specifc RT primer SP NP2 (CATTCCGGATGGCAGAGAATGGGAC) and PCR primers Sxl NP intF (GAGGGTCAGTCTAAGTTATATTCG) and Sxl NP R3 as described[31]. Western blots were done as described using rat anti-HA (1:50, clone 3F10, Roche) and HRP-coupled secondary goat anti-rat antibodies (Molecular Probes)[34]. All experiments were repeated at least once from biological samples.

Analysis of m^{6}A levels. Poly(A) mRNA from at least two rounds of oligo dT selection was prepared according to the manufacturer (Promega). For each sample, 10–50 ng of mRNA was digested with 1 μl of Ribonuclease T1 (1,000 U μl^{-1}; Fermentas) in a final volume of 10 μl in polynucleotide kinase buffer (PNK, NEB) for 1 h at 37 °C. The 5′ end of the T1-digested mRNA fragments were then labelled using 10 U T4 PNK (NEB) and 1 μl [γ-^{32}P]-ATP (6,000 Ci mmol^{-1}; Perkin-Elmer). The labelled RNA was precipitated, resuspended in 10 μl of 50 mM sodium acetate buffer (pH 5.5), and digested with P1 nuclease (Sigma-Aldrich) for 1 h at 37 °C. Two microlitres of each sample was loaded on cellulose TLC plates (20 × 20 cm; Fluka) and run in a solvent system of isobutyric acid: 0.5 M NH_4OH (5:3, v/v), as the first dimension, and isopropanol:HCl:water (70:15:15, v/v/v), as the second dimension. TLCs were repeated from biological replicates. The identification of the nucleotide spots was carried out using m^{6}A-containing synthetic RNA. Quantification of ^{32}P was done by scintillation counting (Packard Tri-Carb 2300TR). For the quantification of spot intensities on TLCs or gels, a storage phosphor screen (K-Screen; Kodak) and Molecular Imager FX in combination with QuantityOne software (BioRad) were used.

For immunoprecipitation of m^{6}A mRNA, poly(A) mRNA was digested with RNase T1 and 5′ labelled. The volume was then increased to 500 μl with IP buffer (150 mM NaCl, 50 mM Tris–HCL, pH 7.5, 0.05% NP-40). IPs were then done with 2 μl of affinity-purified polyclonal rabbit m^{6}A antibody (Synaptic Systems) and protein A/G beads (SantaCruz).

Polysome profiles. Whole-fly extracts were prepared from 20–30 adult *Drosophila* previously frozen in liquid N_2 and ground into fine powder in liquid N_2. Cells were then lysed in 0.5 ml lysis buffer (0.3 M NaCl, 15 mM $MgCl_2$, 15 mM Tris-HCl pH 7.5, cycloheximide 100 μg ml^{-1}, heparin (sodium salt) 1 mg ml^{-1}, 1% Triton X-100). Lysates were loaded on 12 ml sucrose gradients and spun for 2 h at 38,000 r.p.m. at 4 °C. After the gradient centrifugation 1-ml fractions were collected and precipitated in equal volume of isopropanol. After several washes with 80% ethanol the samples were resuspended in water and processed. Experiments were done in duplicate.

Nuclear extract preparation and *in vitro* m^{6}A methylation essays. *Drosophila* nuclear extracts were prepared from Kc cells as described[35]. Templates for *in vitro* transcripts were amplified from genomic DNA using the primers listed below and *in vitro* transcribed with T7 polymerase in the presence of [α-^{32}P]-ATP. DNA templates and free nucleotides were removed by DNase I digestion and Probequant G-50 spin columns (GE Healthcare), respectively. Markers were generated by

using *in vitro* transcripts with or without m^6ATP (Jena Bioscience), which were then digested with RNase T1, kinased with PNK in the presence of [γ-^{32}P]-ATP. After phenol extraction and ethanol precipitation, transcripts were digested to single nucleotides with P1 nuclease as above. For *in vitro* methylation, transcripts ($0.5–1 \times 10^6$ c.p.m.) were incubated for 45 min at 27 °C in 10 µl containing 20 mM potassium glutamate, 2 mM $MgCl_2$, 1 mM DTT, 1 mM ATP, 0.5 mM *S*-adenosylmethionine disulfate tosylate (Abcam), 7.5% PEG 8000, 20 U RNase protector (Roche) and 40% nuclear extract. After phenol extraction and ethanol precipitation, transcripts were digested to single nucleotides with P1 nuclease as above, and then separated on cellulose F TLC plates (Merck) in 70% ethanol, previously soaked in 0.4 M $MgSO_4$ and dried[36]. *In vitro* methylation assays were done from biological replicates at least in duplicates.

Primers to amplify parts of the *Sxl* alternatively spliced intron from genomic DNA for *in vitro* transcription with T7 polymerase were Sxl A T7 F (GGAGCTAATACGACTCACTATAGGGAGAGGATATGTACGGCAACAATAA TCCGGGTAG) and Sxl A R (CGCAGACGACGATCAGCTGATTCAAAGTGA AAG), Sxl B T7 F (GGAGCTAATACGACTCACTATAGGGAGAGCGCTCG CATTTATCCCACAGTCGCAC) and Sxl B R (GGGTGCCCTCTGTGGCTG CTCTGTTTAC), Sxl C T7 F (GGAGCTAATACGACTCACTATAGGGGTCGT ATAATTTATGGCACATTATTCAG) and Sxl C R (GGGAGTTTTGGTTC TTGTTTATGAGTTGGGTG), Sxl D T7 F (GGAGCTAATACGACTCACTA TAGGGAGAAAACTTCCAGCCCACACAACACACAC) and Sxl D R (GCATATCATATTCGGTTCATACATTTAGGTCTAAG), Sxl E T7 F (GGAG CTAATACGACTCACTATAGGGAGAGGGGAAGCAGCTCGTTGTAA AATAC) and Sxl E R (GATGTGACGATTTTGCAGTTTCTCGACG), Sxl F T7 F (GGAGCTAATACGACTCACTATAGGGAGAGGGGGATCGTT TTGAGGGTCAGTCTAAG) and Sxl NP2, Sxl C T7 F and Sxl C1 R (GTAG TTTTGCTCGGCATTTTATGACCTTGAGC), Sxl C2 F (GGAGCTAATACG ACTCACTATAGGGAGACTCTCATTCTCTATATCCCTGTGCTGACC) and Sxl C2 R (CTAATTTCGTGAGCTTGATTTCATTTTGCACAG), Sxl C3 F (GGAGCTAATACGACTCACTATAGGGAGACTGTGCAAAATGAAATCAAGC TCACGAAATTAG) and Sxl C R, Sxl E T7 F and Sxl E1 R (AAAAAAATCAAA AAAATAATCACTTTTGGCACTTTTTCATCAC), Sxl E2 F (GGAGCTAATAC GACTCACTATAGGGAGATGAAAAAGTGCCAAAAGTGATTATTTT TTTG), Sxl E2 R (AAAAGCATGATGTATTTTTTTTTTTTTTGTACTTTCG AATCACCG), Sxl E3 F (GGAGCTAATACGACTCACTATAGGGAGAC GGTGATTCGAAAGTACAAAAAAAAAAAAAATAC) and Sxl E R, Sxl C4 F (GAGCTAATACGACTCACTATAGGGAGAAATACTAAAACATCA AACCGCAAGCAGAGCAGC) and Sxl C4 R (GAGTGCCACTTCAAAAT CTCAGATATGC), Sxl C5 F (CTAATACGACTCACTATAGGGAGACTCTTT TTTTTTTTCTTTTTTTTACTGTGCAAAATG) and Sxl C5 R (AAAAAAATAT GCAAAAAAAAAAAGGTAGGGCACAAAGTTCTCAATTAC), Sxl C6 F (GAGCTAATACGACTCACTATAGGGAGACTGTGCAAAATGAAATCAAGC TCACGAAATTAG) and Sxl C6 R (CAATTTCACTATATGTACGAAA ACAAAAGTGAG), Sxl E4 F (GGAGCTAATACGACTCACTATAGGGA GAACCAAAATTCGACGTGGGAAGAAAC) and Sxl E4 R (TAATCACT TTTGGCACTTTTTCATCACATTAAC), Sxl E5 F (GGCTAATACGACTCACTAT AGGGAGATTTTTTTGATTTTTTTAAAGTGAAAATGTGCTCC) and Sxl E5 R (CACCGAAAAAAAATAAAAAAAAAATAATCATGGGACTATACTAG), Sxl E6 F (GGCTAATACGACTCACTATAGGGAGACTTAAGTGCCAATATTTAAAGT GAAACCAATTG) and Sxl E6 R (CCCCCAGTTATATTCAACCGTGAAAT TCTGC).

Illumina sequencing and analysis of differential gene expression and alternative splicing. Total RNA was extracted from 15 pulverized head/thoraces previously flash-frozen in liquid nitrogen, using TRIzol reagent from *white* (*w*) control and *w*;*Ime4*$^{\Delta 22\text{-}3}$ females that have been outcrossed for several generations to *w*; *Df(3R)Exel6197* to equilibrate genetic background. Total RNA was treated with DNase I (Ambion) and stranded libraries for Illumina sequencing were prepared after poly(A) selection from total RNA (1 µg) with the TruSeq stranded mRNA kit (Illumina) using random primers for reverse transcription according to the manufacturer's instructions. Pooled indexed libraries were sequenced on an Illumina HiSeq2500 to yield 40–46 million paired-end 100 bp reads, and in a second experiment 14–19 million single-end 125-bp reads for three controls and mutants each. After demultiplexing, sequence reads were aligned to the *Drosophila* genome (dmel-r6.02) using Tophat2.0.6 (ref. 37). Differential gene expression was determined by Cufflinks-Cuffdiff and the FDR-correction for multiple tests to raw *P* values with $q < 0.05$ considered significant[38]. alternative splicing was analysed by SPANKI[39] and validated for selected genes based on length differences detectable on agarose gels. Illumina sequencing, differential gene expression and alternative splicing analysis was done by Fasteris (Switzerland). For dosage compensation analysis, differential expression analysis of X-linked genes versus autosomal genes in *Ime4*^{null} mutant was done by filtering Cuffdiff data by a *P* value expression difference significance of $P < 0.05$, which corresponds to a false discovery rate of 0.167 to detect subtle differences in expression consistent with dosage compensation. Visualization of sequence reads on gene models and splice junctions reads in Sashimi plots was done using Integrated Genome Viewer[40]. For validation of alternative splicing by RT–PCR as described above, the following primers were used: Gprk2 F1 (CCAACCAGCCGAAACTCACAGTGAAGC) and Gprk2 R1 (CAGGGTCTCGGTTTCAGACACAGGCGTC), fl(2)d F1 (GCAGCAAACGA GAAATCAGCTCGCAGCGCAG) and fl(2)d R1 (CACATAGTCCTGGAATTCTT GCTCCTTG), A2bp1 F3 (CTGTGGGGCTCAGGGGCATTTTTCCTTCCTC) and A2bp1 R1 (CTCCTCTCCCGTGTGTCTTGCCACTCAAC), cv.-c F1 (GGGTT TCCACCTCGACCGGGAAAAGTCG) and cv.-c R1 (GCGTTTGCGG TTGCTGCTCGCGAAGAGAG), CG8312 F1 (GCGCGTGGCCTCCTTCTT ATCGGCAGTC) and CG8312 R1 (GCGTGGCCACTATAAAGTCCACCTCATC), Chas F2 (CCGATTCGATTCGATTCGATCCTCTCTTC) and Chas R1 (GTCGGTGTCCTCGGTGGTGTTGGTGGAG). GO enrichment analysis was done with FlyMine. For the analysis of uATGs, an R script was used to count the uATGs in 5′ UTRs in all ENSEMBL isoforms of those genes which are differentially spliced in Ime4 mutants, that were then compared to the mean number of ATGs in all *Drosophila* ENSEMBL 5′ UTRs using a *t*-test. Gene expression data were obtained from flybase.

R script.

```
library(seqinr)
library(Biostrings)
>fasta_file <-read.fasta("Soller_UTRs.fa", as.string = T)# read fasta file
>pattern <-"atg" # the pattern to look for
>dict <-PDict(pattern, max.mismatch = 0)#make a dictionary of the pattern to look for
>seq <- DNAStringSet(unlist(fasta_file)[1:638])#make the DNAstringset from the DNAsequences that is, all 638 UTRs related to the 156 genes identified in spanki
>result <-vcountPDict(dict,seq)#count the pattern in each of the sequences
>write.csv2(result, "result.csv")
>fasta_file <-read.fasta("dmel-all-five_prime_UTR-r6.07.fa", as.string = T)# read fasta file
>pattern <-"atg" # the pattern to look for
>dict <-PDict(pattern, max.mismatch = 0)#make a dictionary of the pattern to look for
>seq <- DNAStringSet(unlist(fasta_file)[1:29822])#make the DNAstringset from the DNAsequences that is, all UTRs
>result <-vcountPDict(dict,seq)#count the pattern in each of the sequences
>write.csv2(result, "result_allutrs.csv")
```

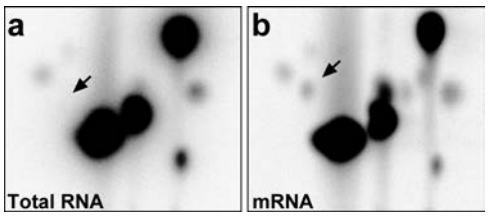
Polytene chromosome preparations and stainings. Ime4 or YT521-B were expressed in salivary glands with *C155-GAL4* from a *UAS* transgene. Larvae were grown at 18 °C under non-crowded conditions. Salivary glands were dissected in PBS containing 4% formaldehyde and 1% Triton X-100, and fixed for 5 min, and then for another 2 min in 50% acetic acid containing 4% formaldehyde, before placing them in lactoacetic acid (lactic acid:water:acetic acid, 1:2:3). Chromosomes were then spread under a siliconized cover slip and the cover slip removed after freezing. Chromosome were blocked in PBT containing 0.2% BSA and 5% goat serum and sequentially incubated with primary antibodies (mouse anti-PolII H5, 1:1000, Abcam, or rabbit anti-histone H4, 1:200, Santa-Cruz, and rat anti-HA monoclonal antibody 3F10, 1:50, Roche) followed by incubation with Alexa488- and/or Alexa647-coupled secondary antibodies (Molecular Probes) including DAPI (1 µg ml^{-1}, Sigma). RNase A treatment (4 and 200 µg ml^{-1}) was done before fixation for 5 min. Ovaries were analysed as previously described[41].

RNA binding assays. The YTH domain (amino acids 207–423) was PCR-amplified with oligos YTHdom F1 (CAGGGGCCCCTGTCGACTAGTCCCGGGAA TGGTGGCGGCAACGGCCG) and R1 (CACGATGAATTGCGGCCGCTCTAGA TTACTTGTAGATCACGTGTATACCTTTTTCTCGC) and cloned with Gibson assembly (NEB) into a modified pGEX expression vector to express a GST-tagged fusion protein. The YTH domain was cleaved while GST was bound to beads using Precession protease. Electrophoretic mobility shift assays and UV cross-linking assays were performed as described[35,42]. Quantification was done using ImageQuant (BioRad) by measuring free RNA substrate to calculate bound RNA from input. All binding assays were done at least in triplicates.

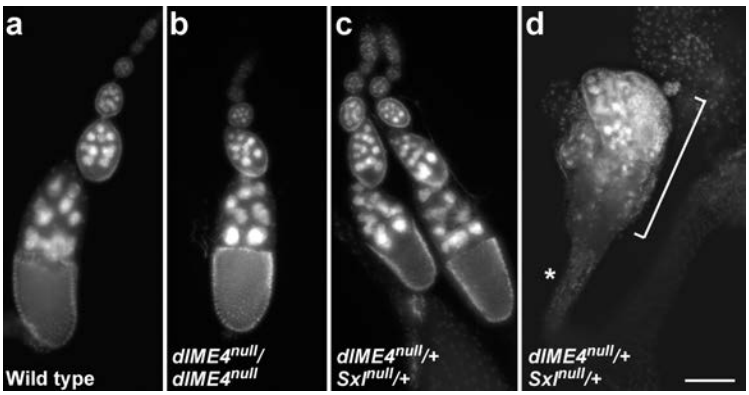
Data availability statement. RNA-seq data that support the findings of this study have been deposited at GEO under the accession number GSE79000, combining the single-end (GSE78999) and paired-end (GSE78992) experiments. All other data generated or analysed during this study are included in this published article and its Supplementary Information.

30. Haussmann, I. U., Hemani, Y., Wijesekera, T., Dauwalder, B. & Soller, M. Multiple pathways mediate the sex-peptide-regulated switch in female *Drosophila* reproductive behaviours. *Proc. R. Soc. Lond. B* **280,** 20131938 (2013).

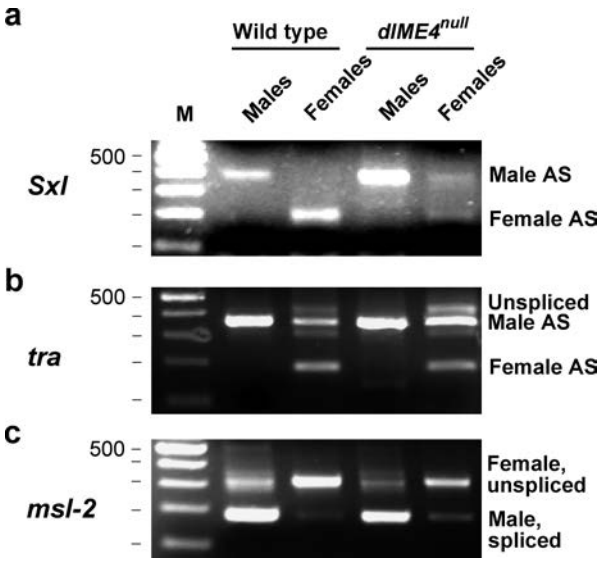
31. Haussmann, I. U., Li, M. & Soller, M. ELAV-mediated 3′-end processing of *ewg* transcripts is evolutionarily conserved despite sequence degeneration of the ELAV-binding site. *Genetics* **189,** 97–107 (2011).
32. Yan, D. & Perrimon, N. *spenito* is required for sex determination in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **112,** 11606–11611 (2015).
33. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25,** 402–408 (2001).
34. Haussmann, I. U., White, K. & Soller, M. Erect wing regulates synaptic growth in *Drosophila* by integration of multiple signaling pathways. *Genome Biol.* **9,** R73 (2008).
35. Soller, M. & White, K. ELAV inhibits 3′-end processing to promote neural splicing of *ewg* pre-mRNA. *Genes Dev.* **17,** 2526–2538 (2003).
36. Harper, J. E., Miceli, S. M., Roberts, R. J. & Manley, J. L. Sequence specificity of the human mRNA N6-adenosine methylase *in vitro*. *Nucleic Acids Res.* **18,** 5735–5741 (1990).
37. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
38. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7,** 562–578 (2012).
39. Sturgill, D. *et al.* Design of RNA splicing analysis null models for post hoc filtering of *Drosophila* head RNA-seq data with the splicing analysis kit (Spanki). *BMC Bioinformatics* **14,** 320 (2013).
40. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).
41. Soller, M., Bownes, M. & Kubli, E. Control of oocyte maturation in sexually mature *Drosophila* females. *Dev. Biol.* **208,** 337–351 (1999).
42. Soller, M. & White, K. ELAV multimerizes on conserved AU4-6 motifs important for *ewg* splicing regulation. *Mol. Cell. Biol.* **25,** 7580–7591 (2005).



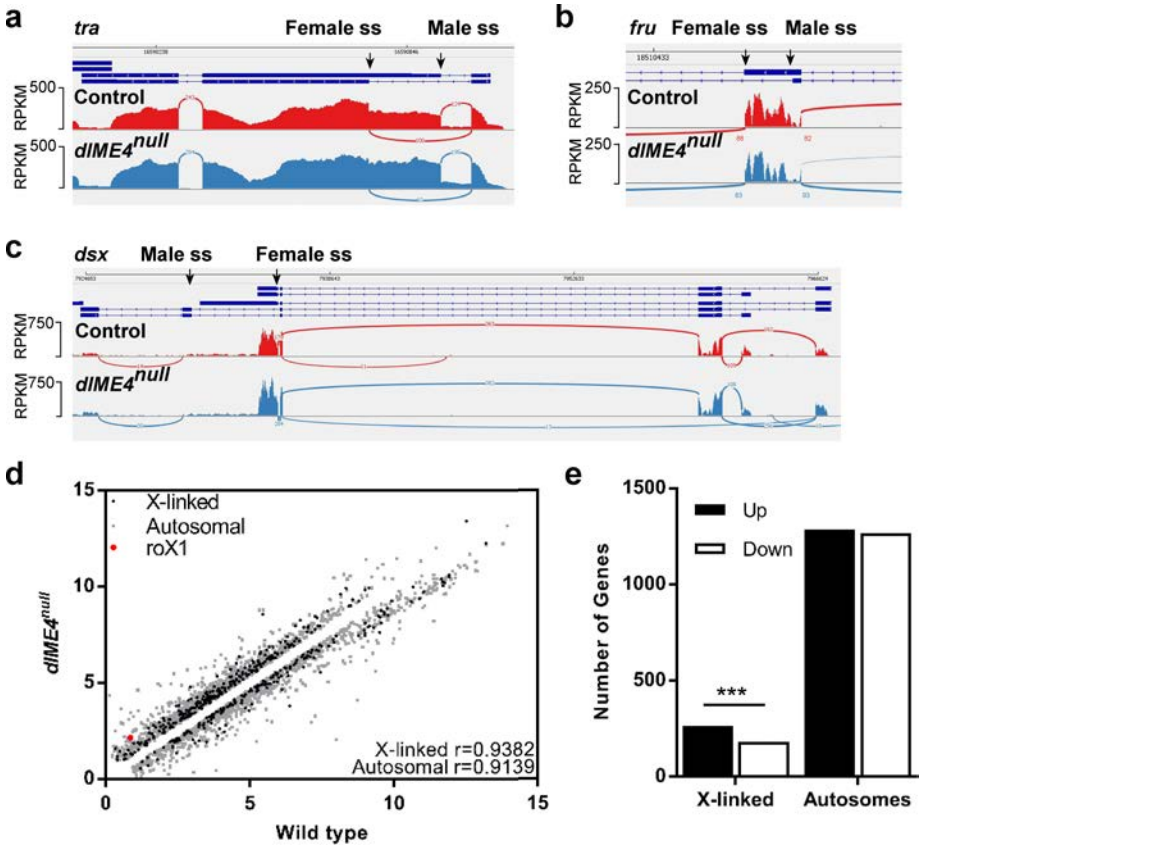
Extended Data Figure 1 | m^{6}A levels in unfertilized eggs. a, **b**, Thin-layer chromatography from maternal total RNA (**a**) and mRNA (**b**) present in unfertilized eggs. The arrow indicates m^{6}A.



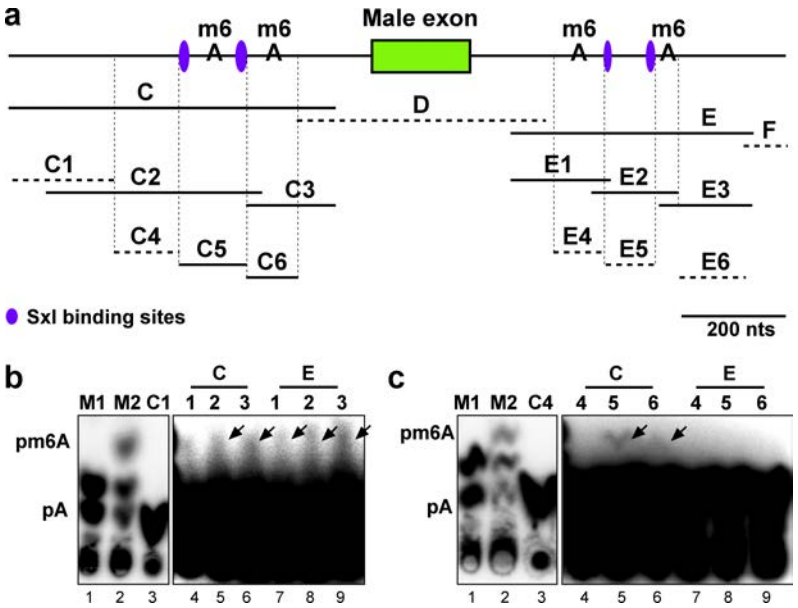
Extended Data Figure 2 | Ime4 supports Sxl in directing germline differentiation. a–**c**, Representative ovarioles of wild-type (**a**), *Ime4*^{null}/*Ime4*^{null} (**b**) and *Sxl*/+;*Ime4*^{null}/+ females (**c**), and a tumerous ovary of a *Sxl*/+;*Ime4*^{null}/+ female (**d**). The tumorous ovary consisting mostly of undifferentiated germ cells in **d** is indicated with a bracket and the oviduct with an asterisk. Scale bar, 100 μm (applies to all panels).



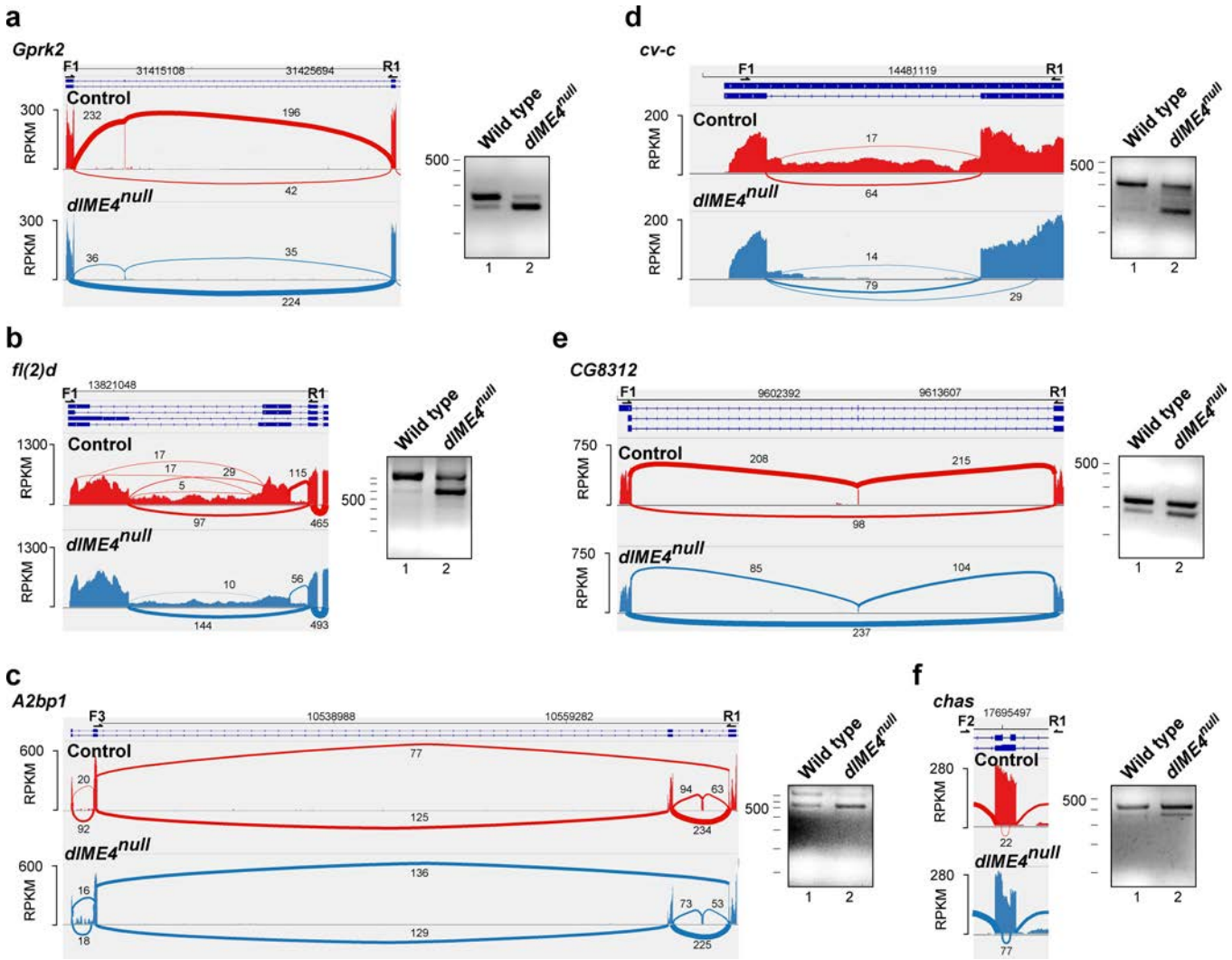
Extended Data Figure 3 | *Ime4* is required for female-specific splicing of *Sxl*, *tra* and *msl-2*. **a**–**c**, RT–PCR of *Sxl* (**a**), *tra* (**b**) and *msl-2* (**c**) sex-specific splicing in wild-type males and females, and *Ime4*^{null} males and females. 100-bp markers are shown on the left. AS, alternative splicing.



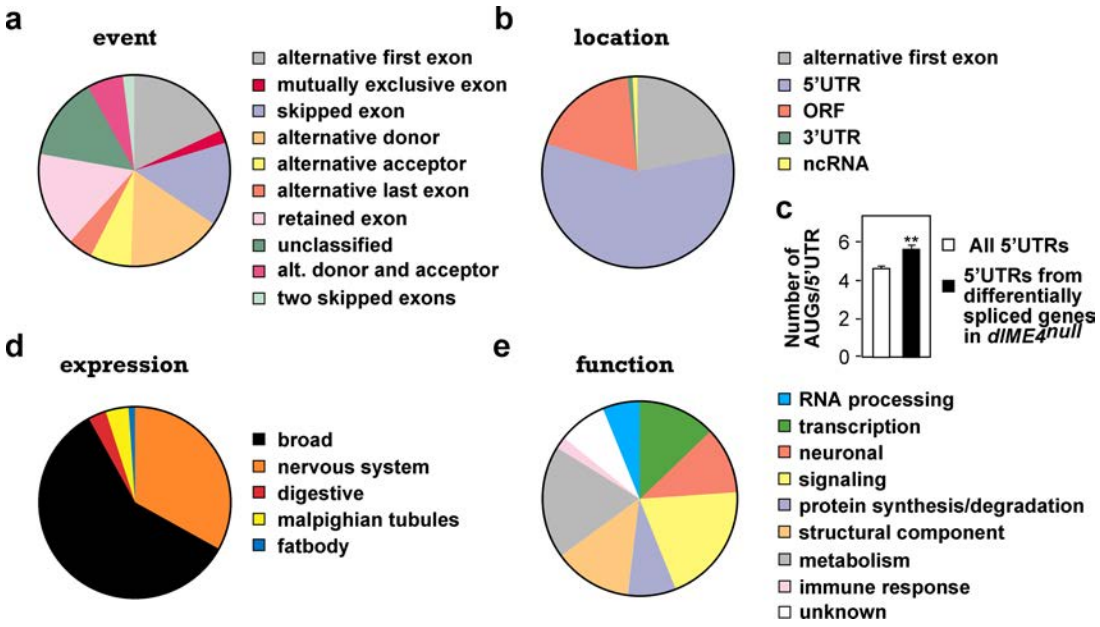
Extended Data Figure 4 | Alternative splicing of sex-determination genes and differential expression of X-linked genes in *Ime4*^{null} females. **a**–**c**, Sashimi plot depicting Tophat-mapped RNA sequencing reads and exon junction reads below the annotated gene model for sex-specific alternative splicing of *tra*, *fru* and *dsx*. The thickness of lines connecting splice junctions corresponds to the number of junction reads also shown. ss, splice site. **d**, Significantly ($P < 0.05$, $q < 0.166853$) differentially expressed gene expression values expressed as reads per kb of transcript per million mapped reads (RPKM) were $\log[x + 1]$-transformed and Spearman *r* correlation values determined for X-linked and autosomal genes in wild-type and *Ime4*^{null} *Drosophila*. **e**, The proportion of autosomal and X-linked genes that were significantly either up- or downregulated in *Ime4*^{null} as compared to wild-type *Drosophila* were statistically compared using χ^2 with Yates' continuity correction. GraphPad Prism was used for statistical comparisons. Similar results as for the single-read RNA-seq experiment were obtained for the paired-end RNA sequencing experiment.



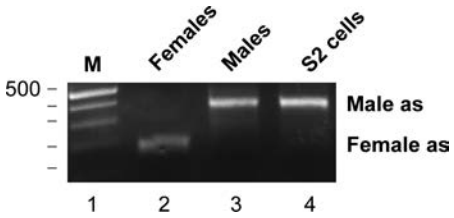
Extended Data Figure 5 | m^{6}A methylation sites map to the vicinity of *Sxl* binding sites. a, Schematic of the *Sxl* alternatively spliced intron around the male-specific exon depicting substrate RNAs used for *in vitro* m^{6}A methylation. Solid lines depict fragments containing m^{6}A methylation and dashed lines indicate fragments where m^{6}A was absent. **b**, **c**, 1D-TLC of *in vitro* methylated [^{32}P]-ATP-labelled substrate RNAs shown in **a**. Markers are *in vitro* transcripts in the absence (M1) or presence (M2) of m^{6}A ^{32}P-labelled after RNase T1 digestion. The right panels in **b** and **c** show an overexposure of the same thin-layer chromatography.



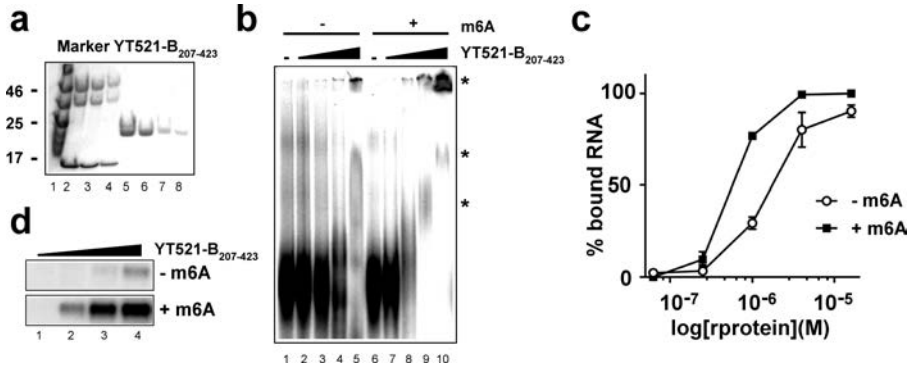
Extended Data Figure 6 | RT–PCR validation of differential alternative splicing in *Ime4*^{null} flies. a–f, Sashimi plots depicting Tophat-mapped RNA sequencing reads and exon junction reads below the annotated gene model of indicated genes on the left, and RT–PCR of alternative splicing shown on the right using primers depicted on top. The thickness of lines connecting splice junctions corresponds to the number of junction reads also shown.



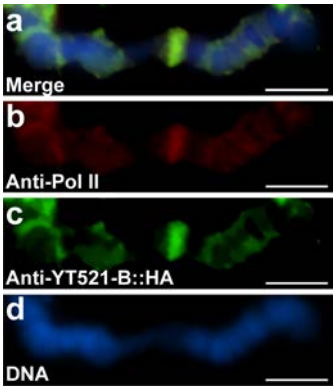
Extended Data Figure 7 | *Ime4* affects alternative splicing predominantly in 5′ UTRs in genes with a higher than average number of upstream start codons. a, **b**, Classification of differential alternative splicing in *Ime4*^{null} according to splicing event (**a**) and location of the event in the mRNA (**b**). **c**, Quantification of upstream start codons (AUGs) in all annotated 5′ UTRs (white) or in alternative isoforms differentially spliced between wild-type and *Ime4*^{null} insects. All *Drosophila* UTRs were accessed in fasta format from Flybase (version r6.07), (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/current/fasta/). An R script was used to count the number of ATG sequences in all *Drosophila* 5′ UTRs and from the genes identified by the Spanki analysis comprising 638 5′ UTRs. A *t*-test was then used to statistically compare the number of ATGs present in the 638 5′ UTRs of the differentially spliced genes as compared to all 29,822 *Drosophila* 5′ UTRs. **d**, **e**, Classification of differentially alternatively spliced genes in *Ime4*^{null} according to expression pattern (**d**) or function (**e**).



Extended Data Figure 8 | *Drosophila* S2 cells are male. RT–PCR of *Sxl* alternative splicing in females, males and S2 cells. 100-bp markers are shown on the left.



Extended Data Figure 9 | Preferential binding of the YTH domain of YT521-B to m^{6}A-containing RNA. a, Coomassie-stained gel depicting the recombinant YTH domain (amino acids 207–423) of YT521-B. **b**, **c**, Electrophoretic mobility shift assay of YTH domain binding to *Sxl* RNA fragment C with or without m^{6}A (50% of adenosine in the transcript methylated) and quantification of RNA bound to the YTH domain shown as mean ± s.e.m. ($n = 3$). Note that the YTH domain does not form a stable complex with RNA (asterisk) and that this complex falls apart during the run or forms aggregates in the well. **d**, UV cross-linking of the YTH domain to *Sxl* RNA fragment C at 0.25 μM, 1 μM, 4 μM and 16 μM (lanes 1–4).



Extended Data Figure 10 | YT521-B co-localizes to sites of transcription. a–**d**, Polytene chromosomes from salivary glands expressing YT521-B::HA stained with anti-Pol II (red, **b**), anti-HA (green, **c**) and DNA (DAPI, blue, **d**), or merged (yellow, **a**). Scale bars, 5 μm.

3. W. Kim, S. Kook, D. J. Kim, C. Teodorof, W. K. Song, *J. Biol. Chem.* **279**, 8333 (2004).
4. V. Giambra *et al.*, *Mol. Cell. Biol.* **28**, 6123 (2008).
5. F. E. Garrett *et al.*, *Mol. Cell. Biol.* **25**, 1511 (2005).
6. W. A. Dunnick *et al.*, *J. Exp. Med.* **206**, 2613 (2009).
7. M. Cogné *et al.*, *Cell* **77**, 737 (1994).
8. J. P. Manis *et al.*, *J. Exp. Med.* **188**, 1421 (1998).
9. A. G. Bébin *et al.*, *J. Immunol.* **184**, 3710 (2010).
10. E. Pinaud *et al.*, *Immunity* **15**, 187 (2001).
11. C. Vincent-Fabert *et al.*, *Blood* **116**, 1895 (2010).
12. R. Wuerffel *et al.*, *Immunity* **27**, 711 (2007).
13. Z. Ju *et al.*, *J. Biol. Chem.* **282**, 35169 (2007).
14. H. Duan, H. Xiang, L. Ma, L. M. Boxer, *Oncogene* **27**, 6720 (2008).
15. M. Gostissa *et al.*, *Nature* **462**, 803 (2009).
16. C. Chauveau, M. Cogné, *Nat. Genet.* **14**, 15 (1996).
17. C. Chauveau, E. Pinaud, M. Cogne, *Eur. J. Immunol.* **28**, 3048 (1998).
18. M. A. Sepulveda, F. E. Garrett, A. Price-Whelan, B. K. Birshtein, *Mol. Immunol.* **42**, 605 (2005).
19. E. Pinaud, C. Aupetit, C. Chauveau, M. Cogné, *Eur. J. Immunol.* **27**, 2981 (1997).
20. A. A. Khamlichi *et al.*, *Blood* **103**, 3828 (2004).
21. R. Shinkura *et al.*, *Nat. Immunol.* **4**, 435 (2003).
22. A. Yamane *et al.*, *Nat. Immunol.* **12**, 62 (2011).
23. M. Liu *et al.*, *Nature* **451**, 841 (2008).
24. J. Stavnezer, J. E. Guikema, C. E. Schrader, *Annu. Rev. Immunol.* **26**, 261 (2008).
25. S. Duchez *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3064 (2010).
26. T. K. Kim *et al.*, *Nature* **465**, 182 (2010).

Acknowledgments: We thank T. Honjo for providing AID$^{-/-}$ mice and F. Lechouane for sorted B cells DNA samples. We are indebted to the cell sorting facility of Limoges University for excellent technical assistance in cell sorting. This work was supported by grants from Association pour la Recherche sur le Cancer, Ligue Nationale contre le Cancer, Cancéropôle Grand Sud-Ouest, Institut National du Cancer, and Région Limousin. The data presented in this paper are tabulated here and in the supplementary materials.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1218692/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 and S2
References (*27*–*30*)

4 January 2012; accepted 27 March 2012
Published online 26 April 2012;
10.1126/science.1218692

Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution

Michael J. Booth,[1]* Miguel R. Branco,[2,3]* Gabriella Ficz,[2] David Oxley,[4] Felix Krueger,[5] Wolf Reik,[2,3]† Shankar Balasubramanian[1,6,7]†

5-Methylcytosine can be converted to 5-hydroxymethylcytosine (5hmC) in mammalian DNA by the ten-eleven translocation (TET) enzymes. We introduce oxidative bisulfite sequencing (oxBS-Seq), the first method for quantitative mapping of 5hmC in genomic DNA at single-nucleotide resolution. Selective chemical oxidation of 5hmC to 5-formylcytosine (5fC) enables bisulfite conversion of 5fC to uracil. We demonstrate the utility of oxBS-Seq to map and quantify 5hmC at CpG islands (CGIs) in mouse embryonic stem (ES) cells and identify 800 5hmC-containing CGIs that have on average 3.3% hydroxymethylation. High levels of 5hmC were found in CGIs associated with transcriptional regulators and in long interspersed nuclear elements, suggesting that these regions might undergo epigenetic reprogramming in ES cells. Our results open new questions on 5hmC dynamics and sequence-specific targeting by TETs.

5-Methylcytosine (5mC) is an epigenetic DNA mark that plays important roles in gene silencing and genome stability and is found enriched at CpG dinucleotides (*1*). In metazoa, 5mC can be oxidized to 5-hydroxymethylcytosine (5hmC) by the ten-eleven translocation (TET) enzyme family (*2*, *3*). 5hmC may be an intermediate in active DNA demethylation but could also constitute an epigenetic mark per se (*4*). Levels of 5hmC in genomic DNA can be quantified with analytical methods (*2*, *5*, *6*) and mapped through the enrichment of 5hmC-containing DNA fragments that are then sequenced (*7*–*13*). Such approaches have relatively poor resolution and give only relative quantitative information. Single-nucleotide sequencing of 5mC has been performed by using bisulfite sequencing (BS-Seq), but this method cannot discriminate 5mC from 5hmC (*14*, *15*). Single-molecule real-time sequencing (SMRT) can detect derivatized 5hmC in genomic DNA (*16*). However, enrichment of 5hmC-containing DNA fragments is required, which causes loss of quantitative information (*16*). Furthermore, SMRT has a relatively high rate of sequencing errors (*17*), and the peak calling of modifications is imprecise (*16*). Protein and solid-state nanopores can resolve 5mC from 5hmC and have the potential to sequence unamplified DNA (*18*, *19*).

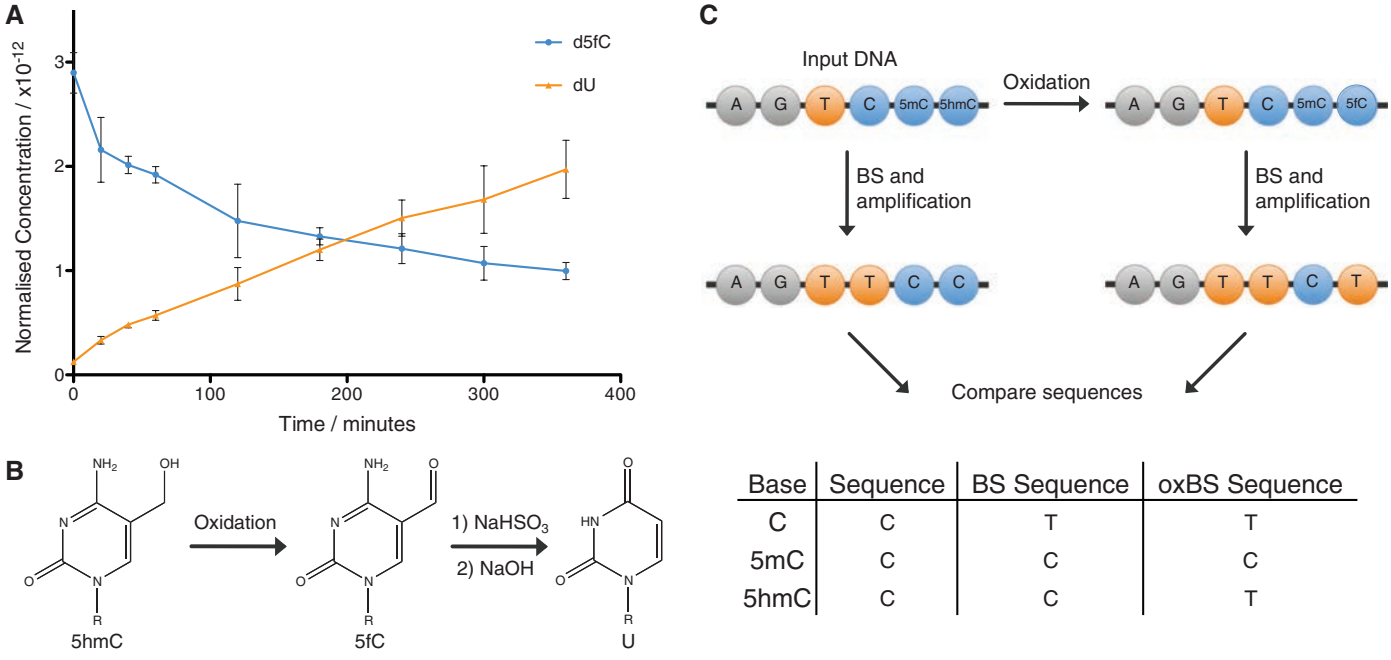
We observed the decarbonylation and deamination of 5-formylcytosine (5fC) to uracil (U) under bisulfite conditions that would leave 5mC unchanged (Fig. 1A and supplementary text). Thus, 5hmC sequencing would be possible if 5hmC could be selectively oxidized to 5fC and then converted to U in a two-step procedure (Fig. 1B). Whereas BS-Seq leads to both 5mC and 5hmC being detected as Cs, this "oxidative bisulfite" sequencing (oxBS-Seq) approach would yield Cs only at 5mC sites and therefore allow us to determine the amount of 5hmC at a particular nucleotide position by subtraction of this readout from a BS-Seq one (Fig. 1C).

Specific oxidation of 5hmC to 5fC (table S1) was achieved with potassium perruthenate ($KRuO_4$). In our reactivity studies on a synthetic 15-nucleotide oligomer single-stranded DNA (ssDNA) containing 5hmC, we established conditions under which $KRuO_4$ reacted specifically with the primary alcohol of 5hmC (Fig. 2A). Fifteen-nucleotide oligomer ssDNA that contained C or 5mC did not show any base-specific reactions with $KRuO_4$ (fig. S1, A and B). For 5hmC in DNA, we only observed the aldehyde (5fC) and not the carboxylic acid (*20*), even with a moderate excess of oxidant. The $KRuO_4$ oxidation can oxidize 5hmC in samples presented as double-stranded DNA (dsDNA), with an initial denaturing step before addition of the oxidant; this results in a quantitative conversion of 5hmC to 5fC (Fig. 2B).

To test the efficiency and selectivity of the oxidative bisulfite method, three synthetic dsDNAs containing either C, 5mC, or 5hmC were each oxidized with $KRuO_4$ and then subjected to a conventional bisulfite conversion protocol. Sanger sequencing revealed that 5mC residues did not convert to U, whereas both C and 5hmC residues did convert to U (fig. S2). Because Sanger sequencing is not quantitative, to gain a more accurate measure of the efficiency of transforming 5hmC to U, Illumina (San Diego, California) sequencing was carried out on the synthetic DNA containing 5hmC (122-nucleotide oligomer) after oxidative bisulfite treatment. An overall 5hmC-to-U conversion level of 94.5% was observed (Fig. 2C and fig. S14). The oxidative bisulfite protocol was also applied to a synthetic dsDNA that contained multiple 5hmC residues (135-nucleotide oligomer) in a range of different contexts that showed a similarly high conversion efficiency (94.7%) of 5hmC to U (Fig. 2C and fig. S14). Last, the $KRuO_4$ oxidation was carried out on genomic DNA and showed through mass spectrometry a quantitative conversion of 5hmC to

[1]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK. [2]Epigenetics Programme, Babraham Institute, Cambridge CB22 3AT, UK. [3]Centre for Trophoblast Research, University of Cambridge, Cambridge CB2 3EG, UK. [4]Proteomics Research Group, Babraham Institute, Cambridge CB22 3AT, UK. [5]Bioinformatics Group, Babraham Institute, Cambridge CB22 3AT, UK. [6]School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SP, UK. [7]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, UK.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: wolf.reik@babraham.ac.uk (W.R.); sb10031@cam.ac.uk (S.B.)



Base	Sequence	BS Sequence	oxBS Sequence
C	C	T	T
5mC	C	C	C
5hmC	C	C	T

Fig. 1. A method for single-base resolution sequencing of 5hmC. (**A**) Reaction of 2′-deoxy-5-formylcytidine (d5fC) with $NaHSO_3$ (bisulfite) quenched by NaOH at different time points and then analyzed with high-performance liquid chromatography (HPLC). Data are mean ± SD of three replicates. (**B**) Oxidative bisulfite reaction scheme: oxidation of 5hmC to 5fC followed by bisulfite treatment and NaOH to convert 5fC to U. The R group is DNA. (**C**) Diagram and table outlining the BS-Seq and oxBS-Seq techniques.

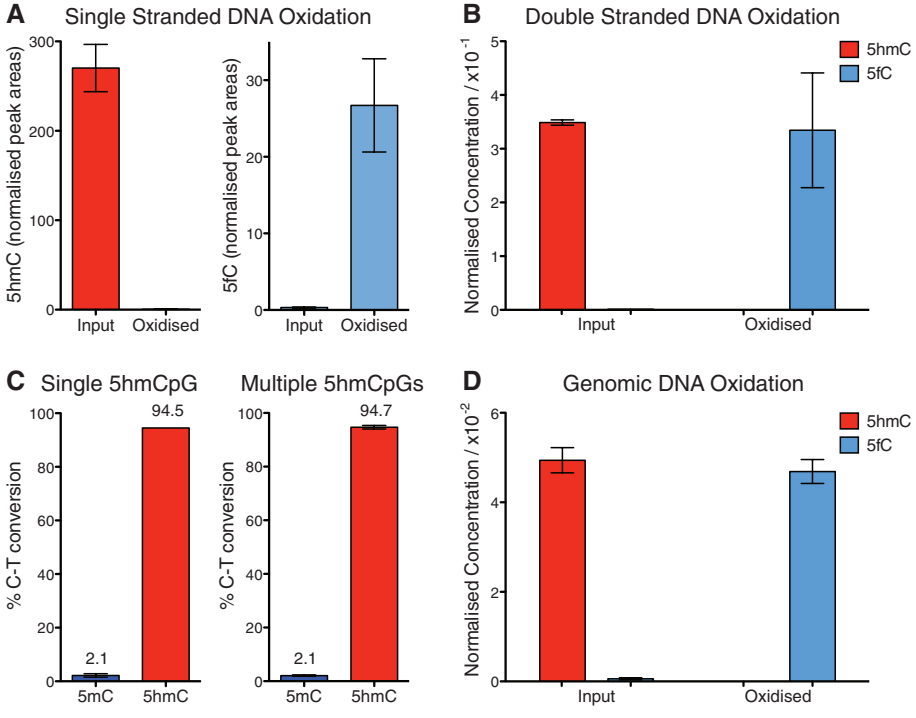


Fig. 2. Quantification of 5hmC oxidation. (**A**) Levels of 5hmC and 5fC (normalized to T) in a 15-nucleotide oligomer ssDNA oligonucleotide before and after $KRuO_4$ oxidation, measured with mass spectrometry. (**B**) Levels of 5hmC and 5fC (normalized to 5mC) in a 135-nucleotide oligomer dsDNA fragment before and after $KRuO_4$ oxidation. (**C**) C-to-T conversion levels as determined by means of Illumina sequencing of two dsDNA fragments containing either a single 5hmCpG (122-nucleotide oligomer) or multiple 5hmCpGs (135-nucleotide oligomer) after oxidative bisulfite treatment. 5mC was also present in these strands. (**D**) Levels of 5hmC and 5fC (normalized to 5mC in primer sequence) in ES cell DNA measured before and after oxidation. Data are mean ± SD.

5fC (Fig. 2D), with no detectable degradation of C (fig. S1C). Thus, the oxidative bisulfite protocol specifically converts 5hmC to U in DNA, leaving C and 5mC unchanged, enabling quantitative, single-nucleotide-resolution sequencing on widely available platforms.

We then used oxBS-Seq to quantitatively map 5hmC at high resolution in the genomic DNA of mouse embryonic stem (ES) cells. We chose to combine oxidative bisulfite with reduced representation bisulfite sequencing (RRBS) (*21*), which allows deep, selective sequencing of a fraction of the genome that is highly enriched for CpG islands (CGIs). We generated RRBS and oxidative RRBS (oxRRBS) data sets, achieving an average sequencing depth of ~120 reads per CpG, which when pooled yielded an average of ~3300 methylation calls per CGI (fig. S3). After applying depth and breadth cutoffs (supplementary materials, materials and methods), 55% (12,660) of all CGIs (*22*) were covered in our data sets.

To identify 5hmC-containing CGIs, we tested for differences between the RRBS and oxRRBS data sets using stringent criteria, yielding a false discovery rate of 3.7% (supplementary materials, materials and methods). We identified 800 5hmC-containing CGIs, which had an average of 3.3% (range of 0.2 to 18.5%) CpG hydroxymethylation (Fig. 3, A and B). We also identified 4577 5mC-containing CGIs averaging 8.1% CpG methylation (Fig. 3B). We carried out sequencing on an independent biological duplicate sample of the same ES cell line but at a different passage

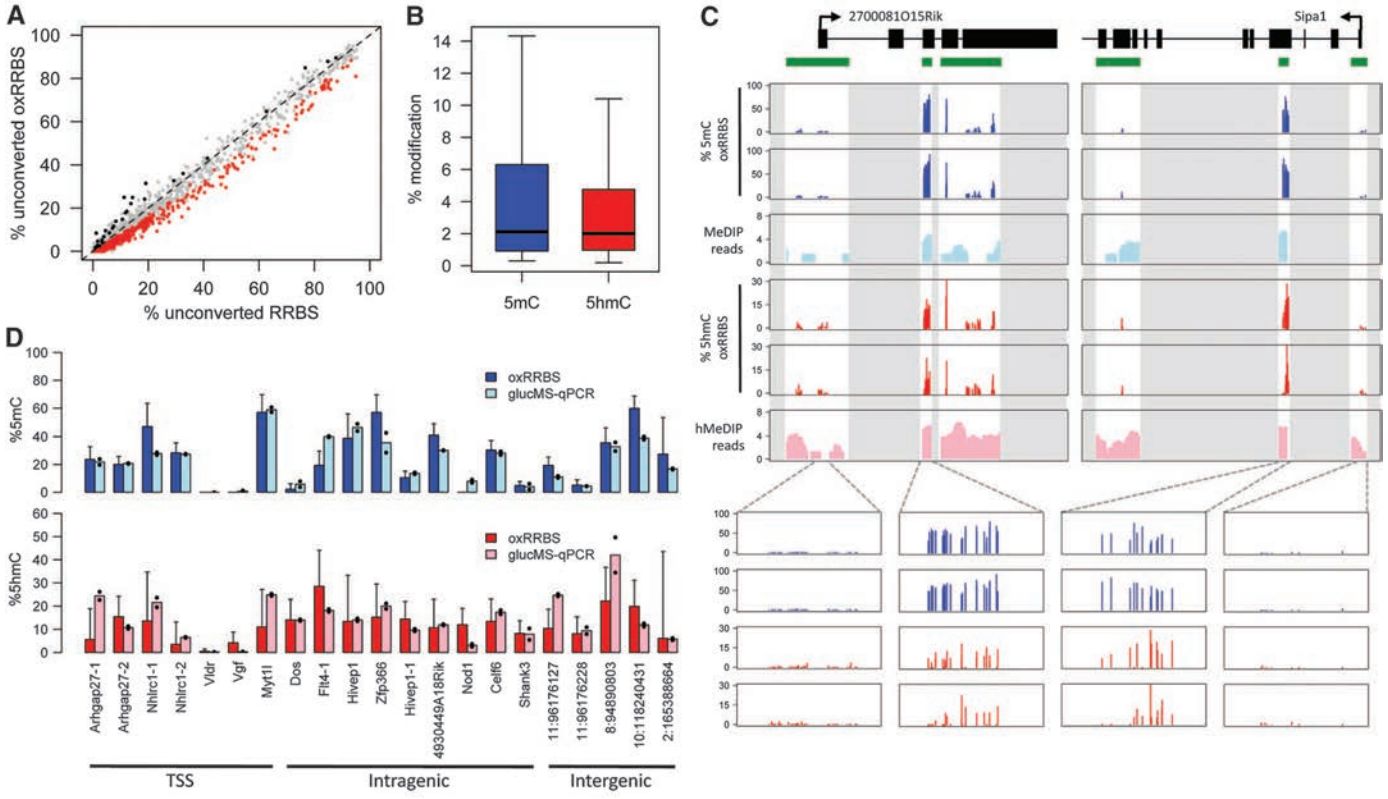


Fig. 3. Quantification of 5mC and 5hmC levels at CGIs by means of oxRRBS. (**A**) Fraction of unconverted cytosines per CGI; 5hmC-containing CGIs (red) have a statistically significant lower fraction in the oxRRBS data set; a false discovery rate of 3.7% was estimated from the CGIs with the opposite pattern (black). (**B**) 5mC and 5hmC levels within CGIs with significant levels of the respective modification. (**C**) Examples of genomic RRBS and oxRRBS profiles overlapped with (h)MeDIP-Seq profiles (*7*). Green bars represent CGIs; data outside CGIs were masked (gray areas). Each bar in the oxRRBS tracks represents a single CpG (in either DNA strand). (**D**) 5mC and 5hmC levels at selected MspI sites were validated through glucMS-qPCR. OxRRBS data are percentage ± 95% confidence interval. Mean glucMS-qPCR values are shown, with the black dots representing individual replicates.

number, which according to mass spectrometry had reduced levels of 5hmC (0.10 versus 0.16% of all Cs), and consistently we found fewer 5hmC-containing CGIs (supplementary text). 5hmC-containing CGIs present in both samples showed good quantitative reproducibility (fig. S5). In non-CpG contexts, we found very few CGIs (71) with levels of 5mC above the bisulfite conversion error (0.2%) (fig. S9) and no CGIs with detectable levels of 5hmC.

Genes associated with 5mC-containing CGIs included *Dazl*, which is known to be methylated in ES cells (fig. S7) (*23*). Similarly, we found that *Zfp64* and *Ecat1* had significant levels of 5hmC (*7*). Genes with >5% 5hmC at transcription start site (TSS) CGIs were associated with gene ontology terms related to transcription factor activity—and in particular were enriched in developmentally relevant genes encoding for Homeobox-containing proteins (such as *Irx4*, *Gbx1*, and *Hoxc4*). To validate our method, we quantified 5hmC and 5mC levels at 21 CGIs containing MspI restriction sites by means of glucosylation-coupled methylation-sensitive quantitative polymerase chain reaction (glucMS-qPCR) (Fig. 3D) (*24*). We found a good correlation between the quantification with oxRRBS and glucMS-qPCR [correlation coefficient (r) = 0.86,

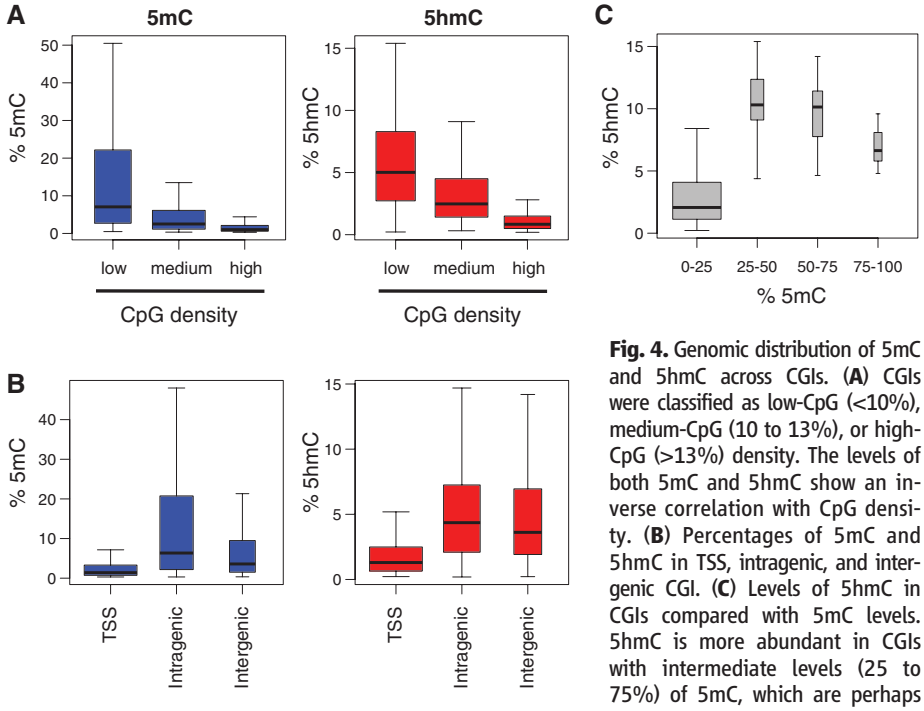


Fig. 4. Genomic distribution of 5mC and 5hmC across CGIs. (**A**) CGIs were classified as low-CpG (<10%), medium-CpG (10 to 13%), or high-CpG (>13%) density. The levels of both 5mC and 5hmC show an inverse correlation with CpG density. (**B**) Percentages of 5mC and 5hmC in TSS, intragenic, and intergenic CGI. (**C**) Levels of 5hmC in CGIs compared with 5mC levels. 5hmC is more abundant in CGIs with intermediate levels (25 to 75%) of 5mC, which are perhaps more epigenetically plastic. For all boxplots, the width of the box is proportional to the amount of data within that group.

$P = 5 \times 10^{-7}$ and $r = 0.52$, $P = 0.01$ for 5mC and 5hmC, respectively], showing that oxRRBS reliably measures 5hmC at individual CpGs. We also found a good correlation between oxRRBS and our previously published (hydroxy)methylated DNA immunoprecipitation sequencing [(h)MeDIP-Seq] data sets (fig. S8) (*7*).

Across CGIs, both 5mC and 5hmC levels are inversely correlated with CpG density, and intragenic and intergenic CGIs contain higher levels of either modification than those overlapping TSSs (Fig. 4, A and B, and fig. S6) (*13*, *22*). TET1 is enriched at TSSs, and thus, a high turnover of 5mC and 5hmC that would keep the steady-state levels low at these sites has been suggested (*9*). Non-TSS CGIs, however, appear to accumulate substantial amounts of both marks, suggesting reduced turnover in these regions. We find that the highest levels of 5hmC are found at CGIs with intermediate levels (25 to 75%) of 5mC (Fig. 4C and fig. S6). Although low-5mC CGIs have reduced potential for 5hmC generation and/or are subjected to a high turnover, high-5mC CGIs are perhaps protected from extensive TET-mediated oxidation, thus stabilizing methylation. Intermediate-5mC CGIs are therefore potentially more epigenetically plastic, given the relatively high abundance of both marks.

Most TSS CGIs (98%) have less than 10% 5mC, as well as low 5hmC, and these are associated with higher transcription levels than average (fig. S10). Within this narrow window, we find a mild negative correlation between transcription and both 5mC and 5hmC levels (fig. S10). At higher 5mC levels, there are insufficient CGIs to obtain a statistically significant result, and it remains possible that here the epigenetic balance between 5mC and 5hmC plays an important transcriptional role, as we previously suggested (*7*).

Last, we quantified 5mC and 5hmC levels at two classes of retrotransposons [long interspersed nuclear element–1 (LINE1) and intracisternal A-particle (IAP)] using two approaches: aligning the oxRRBS reads to the respective consensus sequences and combining oxidative bisulfite with MassARRAY technology (Sequenom, San Diego, California) (fig. S11). We find that LINE1 elements display a considerable amount of 5hmC (approximately 5%), as previously suggested through (h)MeDIP-Seq (*7*). IAPs, on the other hand, have low or no 5hmC. Because LINE1 elements are reprogrammed during preimplantation development whereas IAPs are resistant to this process (*25*), this suggests a possible involvement of 5hmC in the demethylation of specific repeat classes.

The oxBS-Seq method reliably maps and quantifies both 5mC and 5hmC at the single-nucleotide level. Owing to the fundamental mechanism of oxBS-Seq, the approach is compatible with any sequencing platform. In ES cells, we found that in CGIs 5hmC is exclusive to CpG dinucleotides and that it accumulates at intragenic, low-CpG-density CGIs, which tend to have intermediate levels of 5mC and may be particularly epigenetically plastic.

References and Notes

1. A. M. Deaton, A. Bird, *Genes Dev.* **25**, 1010 (2011).
2. M. Tahiliani *et al.*, *Science* **324**, 930 (2009).
3. S. Ito *et al.*, *Nature* **466**, 1129 (2010).
4. M. R. Branco, G. Ficz, W. Reik, *Nat. Rev. Genet.* **13**, 7 (2012).
5. S. Kriaucionis, N. Heintz, *Science* **324**, 929 (2009).
6. M. Münzel *et al.*, *Angew. Chem. Int. Ed.* **49**, 5375 (2010).
7. G. Ficz *et al.*, *Nature* **473**, 398 (2011).
8. W. A. Pastor *et al.*, *Nature* **473**, 394 (2011).
9. H. Wu *et al.*, *Genes Dev.* **25**, 679 (2011).
10. S. G. Jin, X. Wu, A. X. Li, G. P. Pfeifer, *Nucleic Acids Res.* **39**, 5015 (2011).
11. C. X. Song *et al.*, *Nat. Biotechnol.* **29**, 68 (2011).
12. K. Williams *et al.*, *Nature* **473**, 343 (2011).
13. Y. Xu *et al.*, *Mol. Cell* **42**, 451 (2011).
14. Y. Huang *et al.*, *PLoS ONE* **5**, e8888 (2010).
15. C. Nestor, A. Ruzov, R. Meehan, D. Dunican, *Biotechniques* **48**, 317 (2010).
16. C. X. Song *et al.*, *Nat. Methods* **9**, 75 (2012).
17. J. Eid *et al.*, *Science* **323**, 133 (2009).
18. E. V. Wallace *et al.*, *Chem. Commun. (Camb.)* **46**, 8195 (2010).
19. M. Wanunu *et al.*, *J. Am. Chem. Soc.* **133**, 486 (2010).
20. G. Green, W. P. Griffith, D. M. Hollinshead, S. V. Ley, M. Schroder, *J. Chem. Soc. Perkin Trans. 1* **1**, 681 (1984).
21. A. Meissner *et al.*, *Nature* **454**, 766 (2008).
22. R. S. Illingworth *et al.*, *PLoS Genet.* **6**, e1001134 (2010).
23. J. Borgel *et al.*, *Nat. Genet.* **42**, 1093 (2010).
24. S. M. Kinney *et al.*, *J. Biol. Chem.* **286**, 24685 (2011).
25. N. Lane *et al.*, *Genesis* **35**, 88 (2003).

Acknowledgments: We thank T. Green and R. Rodriguez for helpful discussions and J. Webster for help with mass spectrometry. We thank the Biotechnology and Biological Sciences Research Council (BBSRC) for a studentship (M.J.B.). The W.R. lab is supported by BBSRC, Medical Research Council, the Wellcome Trust, European Union EpiGeneSys, and BLUEPRINT. The S.B. lab is supported by core funding from Cancer Research UK. M.J.B. and S.B. are inventors on provisional applications filed for U.S. patents on oxBS-Seq (patent applications US61/605702; US61/641134; US61/623461; and US61/513356). OxRRBS data are deposited in the European Molecular Biology Laboratory–European Bioinformatics Institute ArrayExpress Archive (http://www.ebi.ac.uk/arrayexpress) under the accession number E-MTAB-1042. S.B. is an advisor to Illumina.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1220671/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 and S2
References (*26*–*40*)

16 February 2012; accepted 13 April 2012
Published online 26 April 2012;
10.1126/science.1220671

Science
AAAS

Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution

Michael J. Booth, Miguel R. Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik and Shankar Balasubramanian (April 26, 2012)

Science **336** (6083), 934-937. [doi: 10.1126/science.1220671] originally published online April 26, 2012

Editor's Summary

Distinguishing Epigenetic Marks

Methylation of the cytosine base in eukaryotic DNA (5mC) is an important epigenetic mark involved in gene silencing and genome stability. Methylated cytosine can be enzymatically oxidized to 5-hydroxymethylcytosine (5hmC), which may function as a distinct epigenetic mark––possibly involved in pluripotency––and it may also be an intermediate in active DNA demethylation. To be able to detect 5hmC genome-wide and at single-base resolution, **Booth *et al.*** (p. 934, published online 26 April) developed a 5hmC sequencing chemistry that selectively oxidizes 5hmC to 5-formylcytosine and then to uracil while leaving 5mC unchanged. Using this method, mouse embryonic stem cell genomic D NA was sequenced to reveal that 5hmC is found enriched at intragenic CpG islands and long interspersed nuclear element–1 retrotransposons.

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

REVIEW

Single-cell epigenomics: Recording the past and predicting the future

Gavin Kelsey,[1,2]*† Oliver Stegle,[3,4]*† Wolf Reik[1,2,5]†

Single-cell multi-omics has recently emerged as a powerful technology by which different layers of genomic output—and hence cell identity and function—can be recorded simultaneously. Integrating various components of the epigenome into multi-omics measurements allows for studying cellular heterogeneity at different time scales and for discovering new layers of molecular connectivity between the genome and its functional output. Measurements that are increasingly available range from those that identify transcription factor occupancy and initiation of transcription to long-lasting and heritable epigenetic marks such as DNA methylation. Together with techniques in which cell lineage is recorded, this multilayered information will provide insights into a cell's past history and its future potential. This will allow new levels of understanding of cell fate decisions, identity, and function in normal development, physiology, and disease.

The discovery and description of individual cells in the body has fascinated biologists and pathologists since the cell was discovered (*1*). With the advent of molecular cell biology, methods have been developed for measuring properties and functions of single cells at increasing resolution. This includes, among others, fluorescent protein reporters and single-molecule detection of RNA or DNA. Only recently however, have high-throughput sequencing methods allowed us more comprehensive access to genomic information in single cells. Hence, single-cell RNA sequencing has revealed how heterogeneous the transcriptome of individual cells can be within a seemingly homogeneous cell population or tissue, providing insights into cell identity, fate, and function in the context of both normal biology and pathology [Stubbington *et al.* (*2*) and Lein *et al.* (*3*)]. A few years from now, we likely will have access to total RNA, small and long noncoding RNA, and transcriptional initiation output of the transcriptome (in addition to the stable cytoplasmic component). The development of single-cell RNA sequencing was followed by single-cell genome sequencing, which has provided new insights into genomic stability and genomic variations that occur in physiology and in disease—for example, in cancer, reproductive medicine, or microbial genetics (*4*).

Epigenetics connects the genome with its functional output (Fig. 1). Various epigenetic marks have been described, ranging from DNA (such as DNA methylation) to histone modifications, which can affect the way the cell reads its genome and hence its transcriptional output. Transcription factors that bind to DNA can create or alter epigenetic states (e.g., open or closed chromatin and higher-order chromatin conformation), or their binding can be sensitive to preexisting epigenetic states. Some epigenetic marks can also be heritable from one cell generation to the next (during mitosis) or from one organism generation to the next [intergenerational or transgenerational epigenetic inheritance (*5*)]. However, there are key questions in epigenetics that can only be addressed by determining the epigenome in single cells. For example, how is transcriptional heterogeneity between cells connected with epigenetic heterogeneity (if it is), do changes in transcription precede or follow epigenetic marks when cells change their fate or function, and are epigenetic states better or worse identifiers of rare cell populations and transitional states than the transcriptome? The recent development of single-cell epigenomics methods is beginning to allow us to address these fundamental questions.

> ***"[T]oday we can probe the majority of epigenetic dimensions with single-cell resolution."***

Single-cell epigenome methods can identify open or closed chromatin, including nucleosome positioning (*6*–*11*). From these, one can infer the likelihood of certain transcription factors to bind or not bind to specific DNA sequences within individual cells, and methods are being developed that allow for assaying transcription factor binding directly—for example, single-cell chromatin immunoprecipitation sequencing (ChIP-seq). Thus, one can currently measure (albeit imperfectly) the heterogeneity in a cell population of key histone marks associated with transcriptional states, such as H3K4me3, which indicates active transcription, or H3K27me3, which is found on genes with a repressed transcriptional state (*12*). Functional states (such as transcriptional output) of the genome are also guided by the way the DNA in each cell is organized into higher-order chromatin, which can be determined by single-cell high-throughput chromosome conformation capture (Hi-C) (*13*). Finally, various DNA modifications—such as methylation (5mC), hydroxymethylation (5hmC), and formylcytosine (5fC)—can be located at the single-cell level by sequencing in most areas of the genome, including at single-nucleotide resolution (*14*–*18*). These modifications are part of the biological turnover of DNA methylation and are associated, for example, with transcriptional repression (5mC) or enhancers, including active ones (5hmC and 5fC). Hence, today we can probe the majority of epigenetic dimensions with single-cell resolution.

The techniques described above have been combined into single-cell multi-omics (*19*), which can reveal new connections between regulatory principles that operate in the individual layers (Figs. 1 and 2). Hence, genome sequencing together with transcriptome sequencing can reveal how genetic variation is related to transcriptional variation (*20*, *21*). Furthermore, genome-scale methylome sequencing coupled with the transcriptome (*22*, *23*) has identified widespread associations between epigenetic marks and transcriptional heterogeneity. The latest incarnation, triple-omics, combines genome, methylome, and transcriptome (*24*) assays and can reveal methylome, chromatin accessibility, and the transcriptome (*11*). Together with the development of multidimensional computational methods (*22*, *25*), these techniques are beginning to tease out intricate and unique cell- and locus-specific relationships between, say, methylation and nucleosome accessibility of a gene promoter and the transcriptional output of the gene (*11*).

Single-cell profiling of DNA modifications

Because epigenetic information comes in multiple forms—covalent modifications on DNA, posttranslational modifications of histones, chromatin accessibility and compaction, and higher-order conformation of chromosome domains—each layer of information requires a different biochemical approach to profile it. This has implications for the nature and quality of the information generated from single cells and for the ability to combine multiple measures from the same single cell in multi-omic applications. Depending on the type of question, it will be necessary to determine whether depth or breadth (many, many cells) is required for any specific study (Fig. 2).

Technically, DNA methylation has been the easiest to assay, building on well-established bisulphite chemistry (*26*). However, bisulphite treatment degrades DNA, preventing full-genome coverage and requiring an adaptation of bisulphite sequencing (BS-seq) to the single-cell level (*14*–*16*). BS-seq, by which unmodified cytosine is converted to thymine but 5mC remains unconverted (*26*), yields single-base precision in principle, with the advantage that both modified and unmodified sites are identified (*26*). Therefore, sites without

[1]Epigenetics Programme, Babraham Institute, Cambridge CB22 3AT, UK. [2]Centre for Trophoblast Research, University of Cambridge, Cambridge CB2 3EG, UK. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD Hinxton, Cambridge, UK. [4]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg 69117, Germany. [5]Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK.
*These authors contributed equally to this work. †Corresponding author. Email: gavin.kelsey@babraham.ac.uk (G.K.); oliver.stegle@ebi.ac.uk (O.S.); wolf.reik@babraham.ac.uk (W.R.)

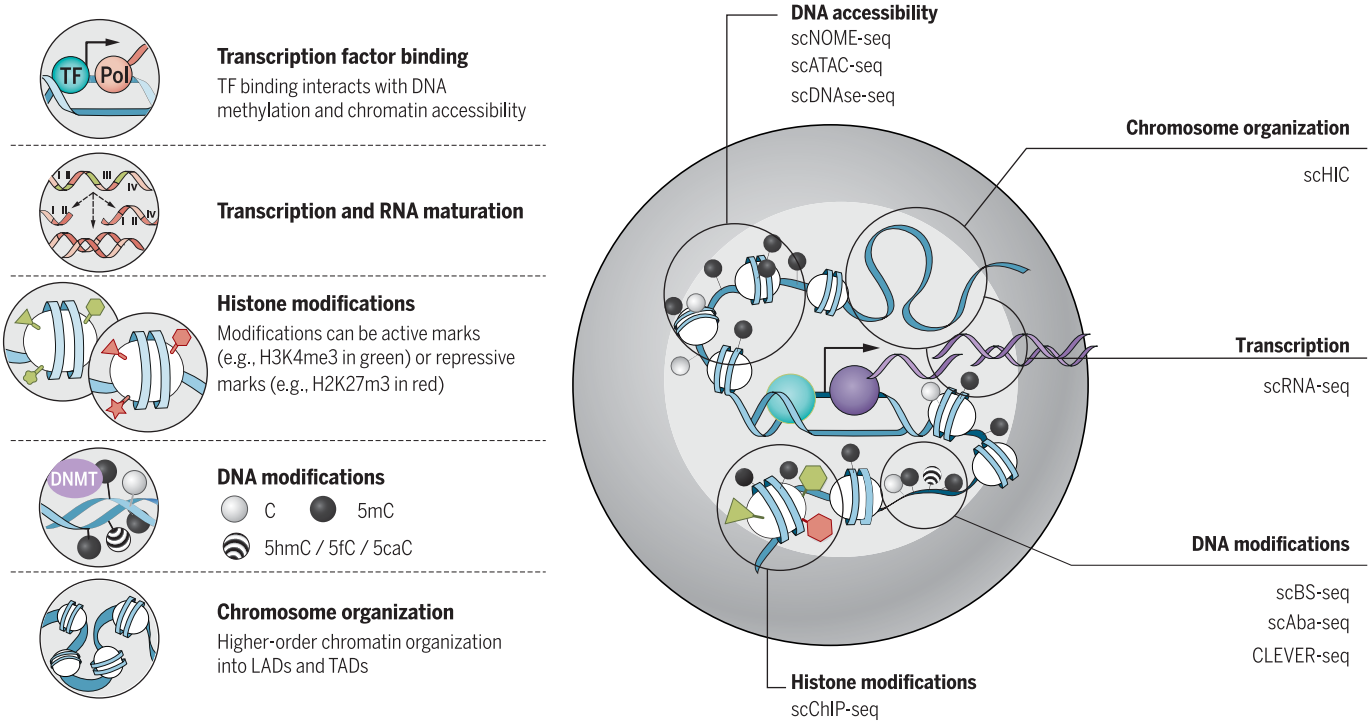


Fig. 1. Single-cell methods and heterogeneity of different molecular layers. (**Left**) Overview of different molecular layers that can be assayed using single-cell protocols. (**Right**) A cell with different layers of multi-omics measurements, as defined on the left. Concordance or heterogeneity respectively may exist between the different layers, and this can be recorded by single-cell sequencing and computationally evaluated.

information are not falsely assigned as unmethylated and, because of the general congruence of methylation over consecutive CpGs in many genomic contexts, missing sites can be imputed from relatively sparse data.

Current single-cell BS-seq (scBS-seq) protocols achieve a coverage up to ~40% (*15*), which means that for most loci the observed sequence reads will originate from only one chromosomal copy. Recent advances in performing single-cell methylation profiling with combinatorial indexing (*27, 28*) may mitigate some of these limitations while simultaneously offering scalability to thousands of cells in a single experiment (Fig. 2). Alternatively, because methylation state can determine whether particular restriction enzymes cleave their recognition sites, methods that use methylation-sensitive or dependent restriction enzymes could present an alternative to bisulphite-based methods (*29*).

Mapping the derivatives of 5mC in single cells has been particularly useful in preimplantation embryos, in which oxidation of 5mC contributes to the active demethylation of the paternal chromosomes (*30*). The pronounced strand bias in distribution between sister cells of these modifications along the same chromosome has provided high-resolution analysis of sister-chromatin exchange (*31*) and has been used as a lineage reconstruction tool (*17*), as well as mapping active demethylation in advance of expression at the promoters of developmentally important genes (*18*). Such advances have required alternative approaches, because 5mC cannot be discriminated from the less abundant 5hmC after bisulphite treatment, and the rarer derivatives 5fC and 5-carboxycytosine (5caC) are indistinguishable from unmodified cytosine.

Treatment with the CpG methylase M.SssI [methylase-assisted bisulphite sequencing, (MAB-seq)] (*31*) allows indirect detection of 5fC, together with 5caC, due to their retention as the only sites remaining susceptible to C to T conversion after bisulphite treatment. Careful control of the methylation reaction is needed to minimize false-positive calls, particularly for a rare modification such as 5fC, which is present at most at tens of thousands of CpG sites, compared with millions of CpGs modified by 5mC. 5hmC can be profiled in single cells by glucosylating 5hmC positions to generate recognition sites for the restriction endonuclease AbaS1 (scAba-seq) (*17*). This provides a positive readout of 5hmC, but, with the inclusion of multiple enzymatic reactions, there is an unknown false-negative rate, which might contribute to a range in the number of 5hmC positions recorded in single cells. 5fC can be detected in single cells by direct chemical labeling with the specific reactivity of malononitrile [chemical-labeling-enabled C-to-T conversion sequencing (CLEVER-seq)] (*18*). The adduct produced prevents normal pairing with G, such that labeled 5fC sites are read as T during polymerase chain reaction (PCR) amplification. In theory, this approach may allow for robust detection of modified bases on single-molecule sequencing platforms.

Combining methylation profiling into multi-omics approaches

scBS-seq can be combined with scRNA-seq through separation of nuclei from cell cytoplasm, separation of RNA and DNA for separate downstream reactions, or preamplification of RNA and DNA in the same cell lysate before splitting and parallel processing for genomic DNA amplification and cDNA library preparation (*22–24*). BS-seq coverage is sufficiently uniform to permit identification of chromosome aneuploidies or large CNVs from regional variations in read depth (*24*). Of note, similar to scRNA-seq protocols that use plate-based methods, scBS-seq can in principle be coupled with profiling of up to tens of cell-surface markers that can be assayed using fluorescence-activated cell sorting, an approach that has been applied in immunology [see Stubbington *et al.*, (*2*)].

Bisulphite sequencing also underlies the nucleosome occupancy and methylome (NOME) sequencing method, which enables information on nucleosome positioning and accessible chromatin to be inferred simultaneously with DNA methylation (*9–11*). Individual lysed cells are treated with M.CviPI, which methylates GpC sites in accessible DNA; then, following bisulphite treatment, methylated cytosines in a GpC context demarcate accessible DNA (linker regions and nucleosome-free DNA), while methylation is read from conversion events of CpGs. Because

both accessible and nonaccessible states are reported, missing information is not falsely assigned, which provides an advantage over other methods for chromatin accessibility. On the other hand, as a method that sequences the genome with no selectivity for open chromatin, high levels of sequencing may be needed to guarantee coverage of elements of interest.

Another potential limitation is the need to filter out C-C-G and G-C-G positions from the methylation data, which reduces the number of genome-wide cytosines that can be assayed compared with scBS-seq by ~50%. However, despite this filter, a large proportion of the loci in genomic regions with important regulatory roles, such as promoters and enhancers, can still be profiled using scNOME-seq–based methods (*11*). scNOME-seq has identified chromatin remodeling dynamics on the two parental alleles during preimplantation development, discriminating cis-regulatory elements open in all cells and promoters that diverge in accessibility between individual blastomeres, these being relatively enriched in gene ontology (GO) terms related to developmental processes and cell differentiation (*9*). Further enhancements of these data can be provided by incorporating transcriptome information from the same cell (Fig. 2) (*11*) to query the strength of coupling between DNA methylation, open chromatin, and transcriptional output.

Mapping functional chromatin states in single cells

A variety of assays have been adapted to profile chromatin states in single cells; these are predicated on enrichment-based strategies; thus, in principle, they have a lower sequencing overhead than scNOME-seq. Open chromatin can be identified by deoxyribonuclease I (DNase I) sensitivity, which was first adapted to the single-cell level in a low-throughput application able to detect an average of ~40,000 DNase I hypersensitive sites (DHSs) per cell (*6*). However, due to nonspecific signals throughout the genome, the false-discovery rate is high. Thus, previous knowledge of DHSs from bulk experiments is required to identify genuine DHSs, with the confidence of detection of proximal regulatory elements scaling with expression level of associated genes.

Higher-throughput applications have been developed for the assay for transposase-accessible chromatin sequencing (ATAC-seq), in which DNA accessibility is probed by the ability of the prokaryotic Tn5 transposase to insert sequencing adapters into accessible regions of the genome, in contrast to regions that are inaccessible, such as those interacting with a nucleosome. These approaches have used microfluidics to process single cells and introduce cell-identifying barcodes as part of the tagging process (*7*) or by combinatorial-cell barcoding (*8*) (Fig. 2), allowing parallel processing of a large number of samples (>10,000).

Throughput levels face a cost of reduced depth, as typically <10% of known promoters are represented in an individual scATAC-seq library. Sparseness of data limits analysis of cellular variation at individual regulatory elements. This may preclude ab initio identification of open chromatin sites, and the absence of open chromatin at a locus of interest in a single cell may reflect missing data. As well as reporting active regulatory elements governing hematopoietic differentiation, scATAC-seq has identified the evolution of regulatory elements during disease progression in acute myeloid leukemia (*32*). In addition, the ability of scATAC-seq to delineate the cis-regulatory landscapes of constituent cell types from a complex solid tissue has been demonstrated by isolating single nuclei from frozen samples of mouse forebrain (*33*).

"Technological advances for assaying epigenetic diversity at the single-cell level have gone hand-in-hand with computational methods for interpreting the data generated."

Posttranslational modifications of histones that correlate with chromatin activity states are conventionally mapped by ChIP-seq. Adapting ChIP-seq to extract this information from single cells presents additional problems of specificity and sensitivity, because it is dependent on antibody binding to pull down modified histones with associated DNA. Droplet approaches and cellular barcoding to label nuclei individually at the stage of micrococcal nuclease digestion (which fragments chromatin into nucleosomes) with immunoprecipitation on pools of cells and subsequent deconvolution of single-cell data after multiplex library sequencing allow thousands of single cells to be processed in single experiments (*12*) (Fig. 2). Yet, although ~50% of sequencing reads may fall within known peaks of H3K4me3 enrichment (the archetypal mark of active promoters), only ~5% of known peaks are detected per cell, with data too sparse for productive de novo peak calling.

We shall inevitably see technical improvements in each of these chromatin profiling methods, as well as incorporating them into multi-omic approaches. A challenge is to extract RNA from cell lysates in a way that preserves both chromatin state and RNA integrity, but with the sparsity of data from current scATAC-seq, scDNase-seq, or scChIP-seq methods, attainment of parallel data on gene expression and chromatin state at specific loci is challenging, and processing increasing numbers of cells may be necessary to obtain sufficient convergent information. Any of the above methods in theory could be combined with bisulphite sequencing to investigate DNA methylation state, which is not to underestimate the technical challenges that may need to be overcome in adding the chemical steps involved in bisulphite treatment.

Readouts of gross chromatin organization in single cells

Higher orders of chromosome organization in interphase nuclei are represented by a number of configurations: topologically associated domains (TADs) divide the genome into structurally separate segments contained in loops and constrained by boundary elements, and lamin-associated domains (LADs) occupy the nuclear periphery. LADs have been probed at the single-cell level by Dam-ID, in which the Dam adenosine methyltransferase is fused with lamin B1 (a constituent of the nuclear lamina) and expressed in cells so that sites of interaction are mapped from sequence tags after DpnI digestion (*34*). Because LADs are megabase-scale chromosome domains, with 1100 to 1400 domains present in a typical cell, only a low rate of false negatives is expected. The extent of heterogeneity between cells thus allows a good measure of the numbers of constitutive and facultative LADs, as well as cooperativity between LADs; such data are not accessible from population-based approaches. Dam-ID methodology could be applied to any other protein interacting with DNA, such as chromatin remodelers and transcription factors. One caveat is that the false-negative rate will increase as the domain of interaction diminishes, or for proteins with very transient interactions.

Hi-C data measures the proximity of DNA sequences in three-dimensional (3D) space on the basis of ligation events in fixed nuclei. A variety of optimizations have been introduced to increase resolution of the data (*35*), as well as throughput (*36*, *37*), since the first report of a single-cell Hi-C method (*13*). Using haploid cells, single-cell Hi-C has allowed modeling of the 3D organization of all chromosomes in individual cells (*38*) and revealed how bulk-cell data obscures the dynamic reorganization of chromosome compartments during the cell cycle (*36*). Despite recent advances, the resolution of scHi-C methods remains insufficient to interrogate contacts between specific promoters and their enhancers, which awaits progress in miniaturizing approaches to promoter-capture Hi-C or complementation with functional experiments, such as epigenome editing (*39*).

Scalability and limitation of current methods

There are common challenges and limitations that apply to several single-cell epigenome methods. An important bottleneck is the currently limited capture rate (e.g., up to ~40% for scBS-seq), which means that even if libraries are sequenced to saturation, missing values are unavoidable (Fig. 3). Other potential drawbacks are low mappability rates (~20 to 30%) and high levels of PCR duplicates (*15*), in particular for deeply sequenced libraries (*16*), which need to be considered when analyzing the resulting data.

So far, epigenome-based methods tend to offer lower throughput than scRNA-seq, which can already be scaled to tens or hundreds of thousands of cells. Recent advances to perform single-cell methylation profiling, ATAC-seq, and Hi-C using combinatorial indexing (*8*, *28*, *37*) have narrowed

this gap. However, in particular, multi-omics methods that require a physical separation step of the RNA and DNA remain limited to medium-throughput analyses of hundreds of cells (Fig. 2). Another current challenge is to estimate and control for technical sources of variation. In single-cell transcriptomics, the level of technical noise can be estimated with spike-in standards, but such normalization strategies are not established for epigenome sequencing. A general strategy that can be useful are negative and positive controls—e.g., diluted bulk material used to create "pseudo cells" or control wells that combine one cell each from different species (*16*), which can be processed alongside each batch of single cells.

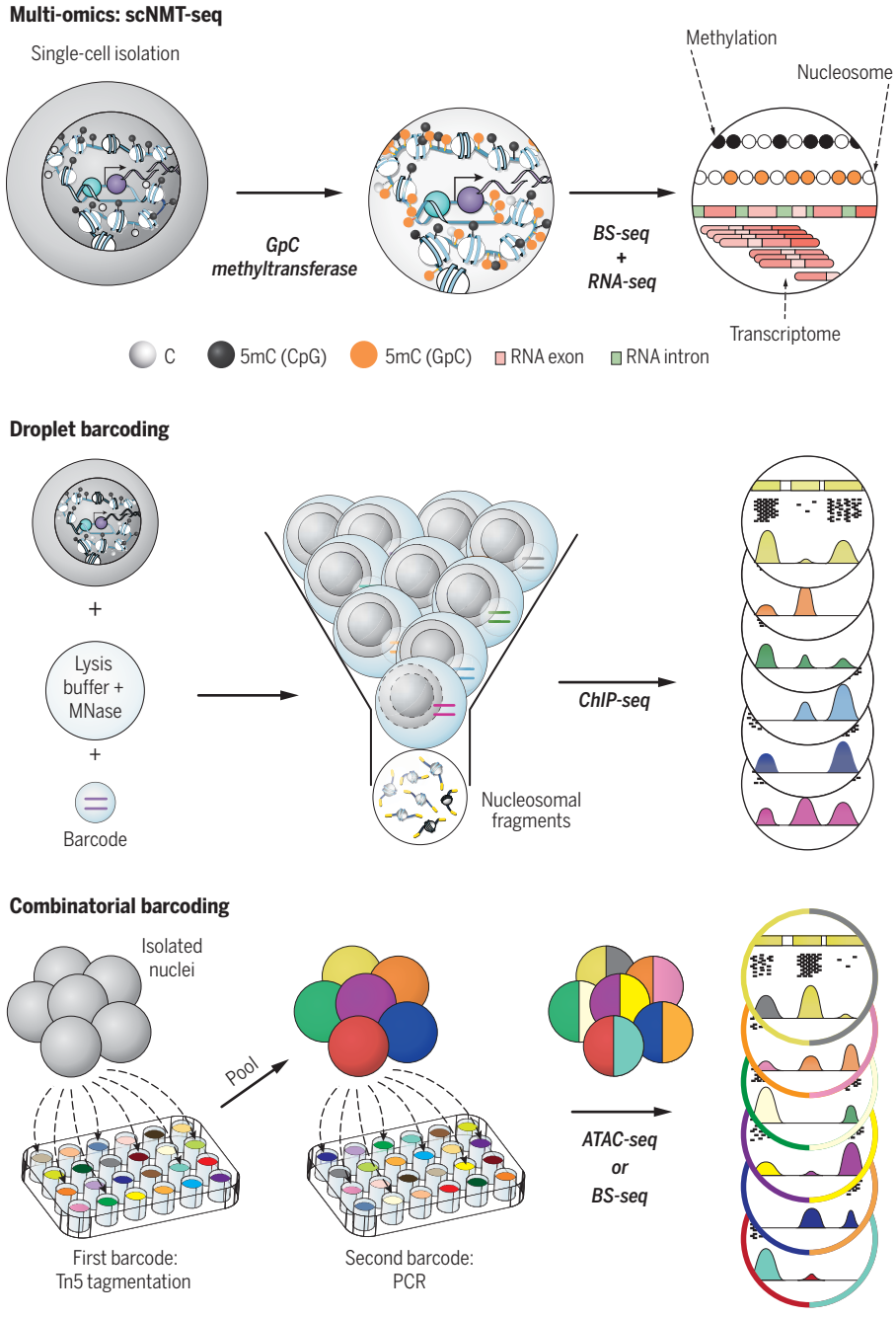


Fig. 2. Depth versus breadth: Multi-omics and cell-barcoding methods. Examples of different technical approaches are shown. (Top) Single-cell nucleosome, methylation, and transcription sequencing (scNMT-seq) (*11*) by which nucleosome accessibility, DNA methylation, and the transcriptome are read simultaneously at considerable depth in each cell; however, with individual cells processed in parallel but separately, cell numbers that can be currently analyzed in this way are limited to hundreds or thousands. (Middle) Barcoding chromatin in individual cells encapsulated in oil droplets, followed by pooling to bulk up material, enables thousands of cells to be processed while seeking to preserve signal-to-noise ratio (*12*). (Bottom) Combinatorial-cell barcoding (*8*, *64*), where readouts can be identified as coming from individual cells by unique combinations of barcodes present in each cell. This approach can be carried out on large numbers of cells (millions), but the depth of information per cell is limited.

Computational analysis to account for missing information using pooling strategies and imputation

Technological advances for assaying epigenetic diversity at the single-cell level have gone hand-in-hand with computational methods for interpreting the data generated (Fig. 3). A first critical step in the computational analysis is the appropriate normalization of the sequencing data while accounting for the typically high levels of noise observed. The sparse coverage of processed single-cell epigenome data sets requires careful consideration in downstream analyses.

Protocols vary in their coverage and whether missing data can be identified directly. For methods that use a bisulphite conversion step, the read coverage is independent of variation in DNA methylation, and hence missing data can be readily identified. For other methods, such as single-cell ATAC-seq, this can be more difficult because the absence or presence of sequence reads is the primary readout of the assay. Different strategies to address the low coverage in these data, such as aggregating read information within regions, by combining reads in consecutive sequence windows (*15*, *16*, *40*) or in annotated genomic contexts, such as promoter regions, enhancers and the like have been proposed. However, there are trade-offs between spatial resolution and coverage, parameters that may greatly affect downstream analyses.

Depending on the question, it may be advantageous to adjust for differences in global methylation, either at the whole-cell level or stratified by genomic context (*16*). A second strategy is to pool cells with similar epigenetic profiles, such as with an initial clustering step to then aggregate read information across cells within each cluster (*27*). These average profiles can offer high spatial resolution, however, at the cost that epigenetic diversity can only be studied at the level of the identified cell clusters (*24*). A third strategy comprises model-based approaches to impute missing information with predictive models. Such strategies have been proposed in the context of bulk epigenome profiles (*41*, *42*) and most recently have been generalised for imputing single-cell DNA methylation data (*25*). Additionally, we note that parallel data from multi-omics experiments will be associated with different patterns of missing data. Because of cost and experimental limitations, not all molecular layers will be assayed in each cell, and hence new computational methods need to handle heterogeneous designs to impute entire molecular layers.

Interrogating single-cell epigenome variation

Depending on the biological question at hand, several downstream analyses can be considered. Caution is required to consider the biological

sources of variation that one may expect in a given study. For example, the cell cycle is a dominant driver of gene expression variation in single cells (*43*) but also manifests at other molecular layers, including copy-number states and DNA methylation (*9*). Also, DNA replication dynamics need to be taken into consideration during experimental design and data analysis.

A starting point for many analyses can be tests for differential epigenetic profiles between different cell clusters—for example, to identify differentially methylated regions between cell types or states (*16*). In cell populations without strong substructure, it may be advantageous to quantify the epigenetic diversity of individual loci with the pairwise distance of global methylome (*16*) or estimates of epigenetic variability between cells at individual loci (*15*).

As multi-omics protocols become more widely accessible, there are also exciting opportunities to interrogate associations between different epigenetic layers and to examine associations with the transcriptome. This allows the strength of coupling between different regulatory layers to be probed in great detail. Variation in coupling strength—for example, between DNA methylation and transcription—is known from bulk analyses, comparing pluripotent to somatic cell types (*44*).

However, the variation in coupling strength can be investigated with single-cell techniques for classes of loci or individual loci between cells or between different loci within the same cell. Such variation has already been identified at different levels, including individual loci such as gene promoters and enhancers with epigenetic variation associated with expression levels of individual genes, as well as global genome-wide couplings between different layers (*22*). If multi-omics methods are applied to hybrids or outbred individuals, it may be possible to assess allele-specific methylation and expression, thereby aligning regulatory differences across molecular layers (*23*). For other analyses, it remains an open question how to best integrate data across different molecular layers. Tying together different data modalities will improve cell clustering, and the use of epigenetic information in tandem with transcriptional data will aid in reconstructing pseudotemporal orderings of cells (Fig. 4).

Adding a temporal dimension in single-cell studies

Putting multidimensional information together for each single cell gives insights not only into cell identity and function but also, through the use of different layers of the epigenome, into past history and future potential (Fig. 4A). Imagine that an otherwise stable DNA methylation mark (for example, in an imprinted gene) has changed at a specific developmental time point, which can be recorded through lineage tracing by CRISPR scarring (*45*–*49*) (Fig. 4B). This is an example in which past history is recorded. Conversely, characteristic DNA methylation patterns in induced pluripotent stem cells (iPSCs) can be predictive of the differentiation potential of these iPSCs (*50*), an example of an epigenetic state revealing future potential.

Different epigenetic marks have different stabilities in time, providing the potential to record various biological time scales. An extracellular signal acting via intracellular signaling pathways will affect transcription factor binding and thus transcription. Because transcription factor binding can be highly dynamic and nonprocessed transcripts are usually short-lived, such signals may reflect the shortest possible biological response time scale. Conversely, though, some transcription factors may bind throughout cell division (*51*) and transmit epigenetic information to the next cell generation. Different binding time scales and their functional consequences may be revealed by coupling the analysis to cell cycle state through the transcriptome (*43*). Similarly, nucleosome accessibility in promoters (or other regulatory sequences) may occur before the chromatin opening up (as may be the case with pioneer factors) or, more conventionally, allow access to transcription factors. Within one cell cycle, therefore, we can reconstruct a signaling response at its cognate promoter, giving rise to transcriptional initiation followed by the processed transcript in the cytoplasm. We can discover multiple genomic dimensions in which this signaling response plays out within this single cell. It is currently possible to reconstruct such multidimensional responses in highly synchronized tissue culture systems but not in the natural setting in vivo, let alone in complex disease situations.

The applications with the most fundamental potential for breakthroughs will also consider epigenetic memory in the system. Some epigenetic marks are heritable across cell divisions (more so in somatic cells than in early embryos), including 5mC DNA methylation, where the inheritance is very stable with a well-understood mechanism. Others, such as H3K27me3 and H3K9me2/me3, may also be inherited, although perhaps with less stability and less fidelity. Whether histone marks associated with transcriptional activation could also be heritable is an open question. A key question here is to what extent epigenetic marks are instructive (e.g., imprinting) or follow transcriptional activation or repression to lock in stabilization of cell fate decisions.

Lineage marking via single-cell sequencing methods will allow us to follow the timing of particular epigenetic changes with regard to the states before the initiation of, during, and post transcription. Furthermore, hairpin bisulphite sequencing (*52*, *53*) (in which methylation information is obtained from both DNA strands) in single cells will identify how heritable methylation is at individual loci and how heterogeneous or homogeneous such heritability is within a cell population. Measurements of 5hmC, 5fC, and 5caC across cell populations, together with mechanistic modeling approaches (*54*, *55*), will allow insights into the generation of epigenetic heterogeneity versus stable inheritance in early development, aging, and disease. The exciting prospect of single-cell epigenome editing (*39*) suggests that detailed

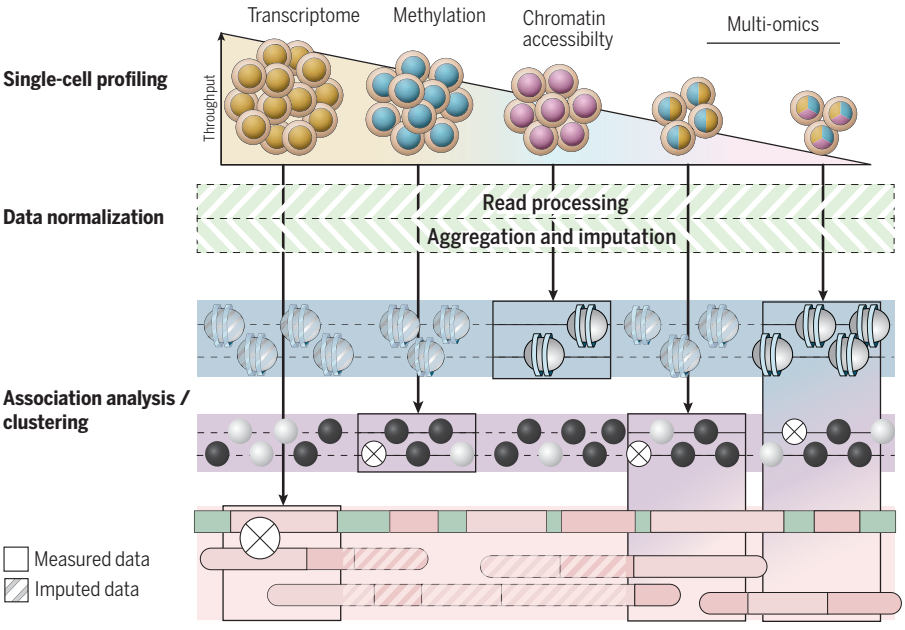


Fig. 3. Multi-omics and computational methods. Shown are typical trade-offs between single-cell RNA-seq, single-cell epigenome protocols, and multi-omics methods that provide readouts from multiple molecular layers in parallel. Consequently, it is commonly required to integrate data from different sequencing protocols. Raw sequence reads from these methods are deduplicated and aggregated into locus-specific readouts, with an optional imputation step to complete missing information. Associations between molecular layers can be used for completing missing data and allow for discovering regulatory associations.

functional testing of epigenetic marks in their various roles may soon become a reality too.

Epigenetic information may also be used to measure cell lineages (Fig. 4B). Lineage-tracing methods using CRISPR scarring have been devised (*45*–*49*), but it is not clear how accurately and reliably they work in different biological settings. Thus, DNA modifications may allow us to trace lineages by marking a particular chromosome or DNA strand, which is segregated into a particular cell type (*17*). This will be especially useful for DNA modifications that are not normally heritable (such as 5hmC, 5fC, or 5caC).

Some heritable epigenetic marks may be functionally neutral—i.e., set up in early development but simply mechanically copied at each cell division. Because the maintenance methylation machinery has a finite error rate [1 in 25 cell divisions per CpG, although this has only been measured in certain contexts (*56*)], every cell may harbor a unique code of methylation sites that would allow tracking of its developmental trajectory. This acts as if lineage were marked by DNA mutations (either natural ones or induced) (Fig. 4B). This may allow noninvasive lineaging in the future without genetic manipulation, which might be particularly useful in human studies.

We have highlighted the different time scales of variation of these different layers of the epigenome, as well as their interdependencies. It is important to recognize that most of these are from indirect measurements or inferences. In due course, we may connect epigenome dimensions by pseudotime measurements, allowing us to formulate temporal connections and dependencies. However, what is yet to materialize are real-time in vivo recording systems of epigenetic states, ideally at a single-locus level. Hence the single-cell epigenomics revolution has additional challenges to overcome. Our existing methods are already allowing us to zoom in on new concepts of "cell fate"—for example, in developmental systems where cell history can be recorded in epigenetic marks. Yet their actions at key decision points require yet unknown mechanisms (*57*, *58*). This presumably requires new epigenomic codes for cell plasticity and future potential. Deeper insights into these rules will provide not only a better understanding of living biological systems but also new tools and new ways of thinking about changing cell fate experimentally.

At the other end of the spectrum, we anticipate information regarding the presumed degradation of cell fate during aging. Models involve either clonal competition or exhaustion and hence a potential loss of cell heterogeneity in an aging tissue. Conversely, an increase in heterogeneity may occur with a concomitant loss of coherence of transcriptional networks (*59*). Interestingly, programmed changes of the epigenome during aging, particularly of the DNA methylome, accurately record chronological age. However, this "methylation aging clock" can be accelerated or decelerated by biological interventions that shorten or lengthen life span, respectively (*60*–*62*). It remains to be seen how this methylation clock plays out at the single-cell level. As many human adult diseases, including cancer, are associated with altered epigenome patterns, individual cells may gradually and in a potentially programmed way acquire disease risk via changes in epigenetic marks during aging. Conversely, single-cell multi-omics methods may identify hidden cell states with potential for tissue repair or rejuvenation.

As large-scale efforts are mapping all human cells transcriptionally and spatially [e.g., the

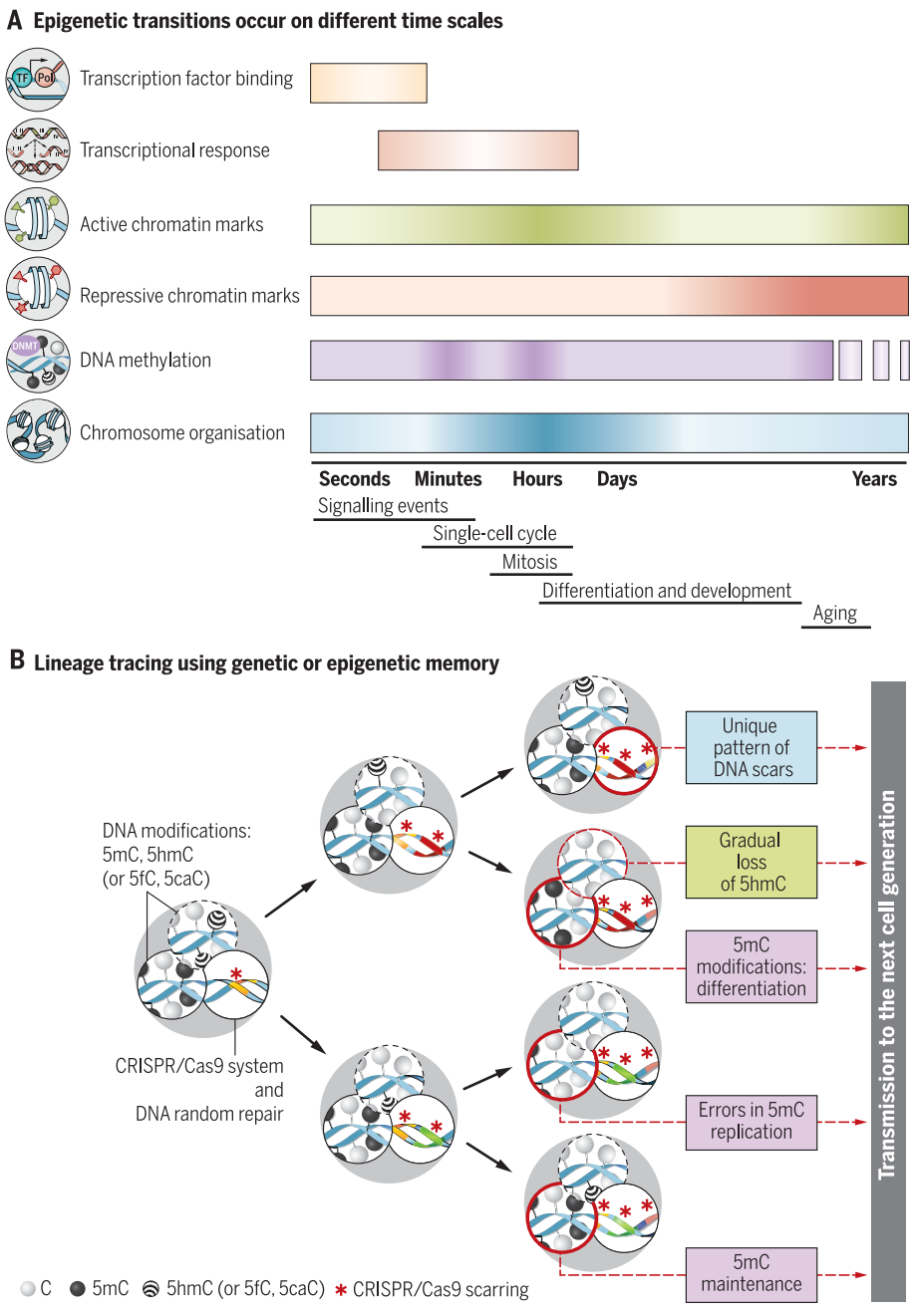


Fig. 4. Time scales of epigenetic heterogeneity at different layers and lineage tracing. (**A**) Shown are different layers of information that can be recorded at least in principle by single-cell multi-omics, from transcription factor binding and transcriptional responses to long-term epigenetic memory such as is possible with DNA methylation. Rough time scales are indicated by colored bars—with shading indicating transitions in information—and may range from seconds to years. With aging, fidelity of epigenetic information such as DNA methylation may degrade, leading to increased cell-to-cell heterogeneity. (**B**) Lineage tracing using genetic or epigenetic memory. Cell lineage can be traced by CRISPR scarring approaches in which each cell and its descendants within a lineage are linked by unique mutations or barcodes. DNA modifications may also be used to track lineage based on their inheritance and on errors in their maintenance at DNA replication. Nonheritable modifications (5hmC, 5fC, and 5caC) have a short-term lineaging potential, whereas heritable modifications (5mC) have long-term noninvasive lineaging potential.

Human Cell Atlas (*63*)], there is the prospect in the future that epigenomics measurements, in particular, will add unprecedented layers of information about memory of past experiences and about future potential of cells in the human body.

Outlook

Imagine that we had at our disposal the techniques for single-cell multi-omics, including the ability to identify all key epigenetic modalities, robustly and at an affordable cost. Imagine similarly that we had the computational tools to unravel and visualize connections between the different molecular layers within and between cells. From such advances, we anticipate answering many questions in embryonic development (including comparisons of various organisms). We would like to know any epigenetic determinants of cell fate and lineage decisions and their timing and/or memory of such decisions.

Travelling back in time (i.e., generating iPSCs) or across tissues (via transdifferentiation), we will be able to see how each cell responds in terms of erasing epigenetic memory and acquiring new cell fate trajectories, especially those not part of the normal developmental repertoire. We also anticipate unraveling tissue-level heterogeneity. Highly multiplexed methylome sequencing can already identify cell types in a complex tissue such as the brain with similar accuracy as transcriptome sequencing (*27*).

Finally, we aim to discover links between epigenetic and genetic heterogeneity, showing to what extent epigenetic change (particularly in disease) is driven by underlying changes in DNA sequence such as copy-number variation, mutations, and rearrangements in cancer, or the mobility of selfish DNA elements. Conversely, primary epimutations may underlie the initiation of some diseases but may subsequently elicit more permanent genetic change that stabilizes the disease phenotype.

These advances have implications for diagnosing and understanding disease progression. We envision that precancerous cell states may be detected at an early stage in tissues by their single-cell epigenome signatures, and other chronic diseases may also reveal unique signatures of progression. Single-cell epigenomic analyses might allow for a biopsy of only a few cells or by capturing small amounts of cell-free DNA in circulation. Such tools may also reveal cell populations in tissues with the greatest potential for regeneration and tissue repair.

REFERENCES AND NOTES

1. R. Hooke, *Micrographia: Or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses, with Observations and Inquiries Thereupon* (Courier Corporation, 2003).
2. M. J. T. Stubbington, O. Rozenblatt-Rosen, A. Regev, S. A. Teichmann, *Science* **358**, 58–63 (2017).
3. E. Lein, L. E. Borm, S. Linnarsson, *Science* **358**, 64–69 (2017).
4. C. Gawad, W. Koh, S. R. Quake, *Nat. Rev. Genet.* **17**, 175–188 (2016).
5. C. D. Allis, T. Jenuwein, D. Reinberg, *Epigenetics* (CSHL Press, 2007).
6. W. Jin *et al.*, *Nature* **528**, 142–146 (2015).
7. J. D. Buenrostro *et al.*, *Nature* **523**, 486–490 (2015).
8. D. A. Cusanovich *et al.*, *Science* **348**, 910–914 (2015).
9. F. Guo *et al.*, *Cell Res.* **27**, 967–988 (2017).
10. S. Pott, *eLife* **6**, e23203 (2017).
11. S. J. Clark *et al.*, *bioRxiv* 138685 [Preprint] (17 May 2017).
12. A. Rotem *et al.*, *Nat. Biotechnol.* **33**, 1165–1172 (2015).
13. T. Nagano *et al.*, *Nature* **502**, 59–64 (2013).
14. H. Guo *et al.*, *Genome Res.* **23**, 2126–2135 (2013).
15. S. A. Smallwood *et al.*, *Nat. Methods* **11**, 817–820 (2014).
16. M. Farlik *et al.*, *Cell Reports* **10**, 1386–1397 (2015).
17. D. Mooijman, S. S. Dey, J. C. Boisset, N. Crosetto, A. van Oudenaarden, *Nat. Biotechnol.* **34**, 852–856 (2016).
18. C. Zhu *et al.*, *Cell Stem Cell* **20**, 720–731.e5 (2017).
19. I. C. Macaulay, C. P. Ponting, T. Voet, *Trends Genet.* **33**, 155–168 (2017).
20. I. C. Macaulay *et al.*, *Nat. Methods* **12**, 519–522 (2015).
21. S. S. Dey, L. Kester, B. Spanjaard, M. Bienko, A. van Oudenaarden, *Nat. Biotechnol.* **33**, 285–289 (2015).
22. C. Angermueller *et al.*, *Nat. Methods* **13**, 229–232 (2016).
23. Y. Hu *et al.*, *Genome Biol.* **17**, 88 (2016).
24. Y. Hou *et al.*, *Cell Res.* **26**, 304–319 (2016).
25. C. Angermueller, H. J. Lee, W. Reik, O. Stegle, *Genome Biol.* **18**, 67 (2017).
26. M. Frommer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1827–1831 (1992).
27. C. Luo *et al.*, *Science* **357**, 600–604 (2017).
28. R. M. Mulqueen *et al.*, *bioRxiv* 157230 [Preprint] (2 June 2017).
29. L. F. Cheow, S. R. Quake, W. F. Burkholder, D. M. Messerschmidt, *Nat. Protoc.* **10**, 619–631 (2015).
30. J. R. Peat *et al.*, *Cell Reports* **9**, 1990–2000 (2014).
31. X. Wu, A. Inoue, T. Suzuki, Y. Zhang, *Genes Dev.* **31**, 511–523 (2017).
32. M. R. Corces *et al.*, *Nat. Genet.* **48**, 1193–1203 (2016).
33. S. Preissl *et al.*, *bioRxiv* 159137 [Preprint] (6 July 2017).
34. J. Kind *et al.*, *Cell* **163**, 134–147 (2015).
35. I. M. Flyamer *et al.*, *Nature* **544**, 110–114 (2017).
36. T. Nagano *et al.*, *Nature* **547**, 61–67 (2017).
37. V. Ramani *et al.*, *Nat. Methods* **14**, 263–266 (2017).
38. T. J. Stevens *et al.*, *Nature* **544**, 59–64 (2017).
39. J. van Arensbergen, B. van Steensel, *Mol. Cell* **66**, 167–168 (2017).
40. S. Gravina, X. Dong, B. Yu, J. Vijg, *Genome Biol.* **17**, 150 (2016).
41. W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, B. E. Engelhardt, *Genome Biol.* **16**, 14 (2015).
42. J. Ernst, M. Kellis, *Nat. Biotechnol.* **33**, 364–376 (2015).
43. F. Buettner *et al.*, *Nat. Biotechnol.* **33**, 155–160 (2015).
44. G. Ficz *et al.*, *Cell Stem Cell* **13**, 351–359 (2013).
45. A. McKenna *et al.*, *Science* **353**, aaf7907 (2016).
46. J. P. Junker *et al.*, *bioRxiv* 056499 [Preprint] (1 June 2016).
47. S. D. Perli, C. H. Cui, T. K. Lu, *Science* **353**, aag0511 (2016).
48. R. Kalhor, P. Mali, G. M. Church, *Nat. Methods* **14**, 195–200 (2017).
49. K. L. Frieda *et al.*, *Nature* **541**, 107–111 (2017).
50. M. Nishizawa *et al.*, *Cell Stem Cell* **19**, 341–354 (2016).
51. X. Huang, J. Wang, *Cell Stem Cell* **20**, 741–742 (2017).
52. C. D. Laird *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 204–209 (2004).
53. L. Zhao *et al.*, *Genome Res.* **24**, 1296–1307 (2014).
54. F. von Meyenn *et al.*, *Mol. Cell* **62**, 983 (2016).
55. P. Giehr, C. Kyriakopoulos, G. Ficz, V. Wolf, J. Walter, *PLOS Comput. Biol.* **12**, e1004905 (2016).
56. T. Ushijima *et al.*, *Genome Res.* **13**, 868–874 (2003).
57. H. J. Lee, T. A. Hore, W. Reik, *Cell Stem Cell* **14**, 710–719 (2014).
58. H. Mohammed *et al.*, *Cell Reports* **20**, 1215–1228 (2017).
59. C. P. Martinez-Jimenez *et al.*, *Science* **355**, 1433–1436 (2017).
60. S. Horvath, *Genome Biol.* **14**, R115 (2013).
61. G. Hannum *et al.*, *Mol. Cell* **49**, 359–367 (2013).
62. T. M. Stubbs *et al.*, *Genome Biol.* **18**, 68 (2017).
63. A. Regev *et al.*, *bioRxiv* 121202 [Preprint] (8 May 2017).
64. B. Lake *et al.*, *bioRxiv* 128520 [Preprint] (19 April 2017).

ACKNOWLEDGMENTS

W.R. thanks I. Herraez, T. Stubbs, S. Clark, C. Alda, H. Mohammed, M. Eckersley-Maslin, S. Rulands, W. Dean, J. Marioni, and B. Simons for discussions or comments on the manuscript. Thank you to V. Juvin (SciArtWork) for artwork. Work in W.R.'s laboratory is supported by the Wellcome Trust, the Biotechnology and Biological Sciences Research Council (BBSRC), and the Medical Research Council (MRC): G.K. is supported by the BBSRC and the MRC; O.S. is supported by European Molecular Biology Laboratory core funding, the Wellcome Trust, and the European Research Council. W.R. is a consultant and shareholder of Cambridge Epigenetix.

10.1126/science.aan6826

Single-cell epigenomics: Recording the past and predicting the future

Gavin Kelsey, Oliver Stegle and Wolf Reik

Science **358** (6359), 69-75.
DOI: 10.1126/science.aan6826